THE UNIVERSITY OF CHICAGO

IDENTIFYING AUTOREGULATORY PROPERTIES OF THE NUCLEIC ACID BINDING DOMAINS WITHIN THE YIN YANG 1 TRANSCRIPTION FACTOR AND INVESTIGATING THE REGULATORY IMPLICATIONS OF H3K4ME3 AND H3K27ME3 CHROMATIN MARKS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN CELL AND MOLECULAR BIOLOGY

BY

JIMMY ELIAS

CHICAGO, ILLINOIS

JUNE 2025

ABSTRACT

Gene expression is a tightly regulated system critical to defining cellular identity. Of the many regulatory components governing gene expression, this thesis looks at two specific components that can influence both transcriptional activation and repression.

First, I investigate the pleiotropic transcription factor (TF) Yin Yang 1 (YY1). Although this TF is developmentally essential and functions in many fundamental cell processes, there is still confusion regarding what contextual aspects dictate its function. I utilize protein biochemistry and biophysical assays to interrogate the interfaces responsible for RNA binding, a recently discovered capability of this TF. I reconcile previous studies claiming that different regions of the protein confer these nucleic acid interactions by demonstrating that there are multiple domains of YY1 that can bind RNA. My work also uncovers a previously unannotated intramolecular inhibitory mechanism that YY1's N-terminus can impose upon its zinc finger module.

Second, I look into the nucleosomal-existence of activating (H3K4me3) and repressive (H3K27me3) chromatin marks, in the context of the bivalency model. An entire field of chromatin biology has canonized bivalent chromatin as a functional co-occurence essential for cellular differentiation. In this thesis I address specific pitfalls of the conventional way that chromatin immunoprecipitation (ChIP) is performed and present methodologies developed by the Ruthenburg lab (ICeChIP and ReICeChIP) that rigorously address these drawbacks and provide quantitative insight to the distribution of these chromatin marks. With these observations, we directly call the bivalency model into question as we assess the chromatin states of mouse embryonic stem cells through differentiation to neuronal precursor cells.

Maintaining the balance between gene activation and repression requires multiple layers of regulation. This work provides frameworks for deeper investigations of gene regulation in the future.

A mis padres, Olivia y Jaime Elias. Por sus sacrificios y apoyo

que me han dado la oportunidad de vivir este sueño

To my brother, Ricardo, for his never-ending support, strength, and cheer

And to my nephew, Dominick, for his ability to always put a smile on my face

"He is beginning to believe."

—Morpheus, *The Matrix* (1999)

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I can easily say that it has taken more than a village to get me to this stage of my life. First, I would like to thank the mentors who introduced me to scientific research and fostered my initial interests and curiosities. I was exposed to wet lab research by the Center Scholars Program, an initiative proposed by Dr. Andrew Feinberg and funded by the National Human Genome Research Institute. I distinctly remember when Dr. Norann Zaghloul showed me how to pour agarose gels and run my first PCRs. These experiences and her mentorship have monumentally impacted the trajectory of my life. The Center Scholars Program also introduced me to my undergraduate research mentor, Dr. Nicholas Durham. Dr. Durham taught me about the rigors of science and the camaraderie required to excel in it. He poured immense amounts of time, energy, and kindness into my mentorship and truly made me feel that he believed in me. I would be remiss not to mention Vicky Schneider, the director of the Center Scholars Program during those years. She had a caring heart for every student and constantly ensured that everything ran smoothly. Thank you all for introducing me to the world of scientific research.

A central component to my scientific career are the communities I've been a part of. In this spirit, I'd like to thank the Center for Inherited Disease Research at The Johns Hopkins University and Dr. John Maris' group at the Children's Hospital of Philadelphia. These were necessary experiences in preparation for graduate school.

This brings me to the Ruthenburg lab. I am incredibly grateful to have been given the opportunity to work with such interesting, incredible, hard-working, and dedicated individuals. I do not believe that there is a peril or obstacle that we could not surmount, and I could not have wished for a better group of people to have spent my PhD with. I'd also like to extend a special thanks to my co-author: Dr. Rohan Shah, a tireless perfectionist through and through. Coupled to the Ruthenburg lab, the 8th floor community has been such a warm collection of laughs, stories, and espresso. It's been amazing to work with everyone on

the floor and even though members of labs have come and gone, I can confidently say that the incredible spirit of the 8th floor rings true with each new scientist that joins.

This brings me to my mentor, Dr. Alexander Ruthenburg. He has shaped the scientist I am today and I cherish the fact that at the onset of my projects, he not only guided me towards the larger questions we were asking but, he **joined** me at the bench to ensure my understanding and development. His enthusiasm and energy are contagious and he has supported me through the toughest trials of my life thus far. I'd also like to acknowledge the members of my thesis committee, Dr. Ilaria Rebay, Dr. Jonathan Staley, and Dr. Yang Li. Thank you for agreeing to be on my committee, providing your insights whenever we met, and ensuring that progress was made toward my professional goals.

The friends I have made throughout graduate school have been essential to my happiness during my PhD. Olivia, Jojo, Grace, Rob, Julio, Fernando, Jordan, Meike, Steven, Liz, Emily, Kourtney, Evan, Matt; we've made such formative memories over these years and I look forward to the many more to come. This extends to anyone who has come to a volleyball outing, grilling session, intramural team members, or pick-up soccer session. It has all been wonderful. And I cannot leave out my hometown or college friends. Tom, Joe, Dylan, Dhruv, Sterling, Dan, Nick, Irvin, Jon, Peven, Erica, and Shona, you have all played immense roles in getting me to where I am today.

Words cannot describe how thankful I am to my family for the constant support they've provided me with as I've delved into this mysterious world away from home. My parents left Peru with the dream that my brother and I would be able to pursue our dreams y ya estámos viviéndolos. My older brother Ricardo has always shown me the path forward and how to stay strong through any situation we've encountered and I know I've needed his strength and support to finish out this chapter of my life. As for my nephew Dominick, he is an inspiration for me to continue to do better. The world is in front of you Dominck and I know that you have the heart and positivity to make any change that you set your mind

to. You've always put a smile on my face and I'm so wildly proud of you, thank you for being as amazing as you are and helping me reach this milestone of my life. Y también me gustaria agradecer a mis tíos, tías, y primos que me han soportado desde mi juventud. Este logro es para todos nosotros.

Lastly, I need to thank my two support systems that keep me going day in and day out: my partner, Amanda Keplinger, and my lovely cat, Toast. Amanda has been my partner in lab, on the soccer field, and on countless other adventures. Her smile and positivity give me ease on any endeavor we embark upon. Toast reminds me of the finer necessities of life. She reminds me to calm down, enjoy the sun, and rest when needed. Regardless of how hard any day has been or how late any experiment has gone, these two put a smile on my face whenever I get home. Thank you all for being a part of my journey and I'm incredibly excited for what's to come.

CHAPTER 1

INTRODUCTION

1.1 Cellular identity and the roles of transcription factors

DNA encodes the inheritable genetic information necessary for cellular identity and function across biology. The central dogma of life is built upon the information flow of DNA to RNA to protein via the essential processes of transcription and translation. In eukaryotes, the genome is packaged into the nuclei of cells and organized to allow for specific gene expression profiles. The human genome encompasses 3 billion base pairs of DNA, which, if stretched end to end, would measure to approximately 2m in length. This genetic information is compacted into a nucleus that ranges between 5 to 10 micrometers. For reference, this type of compaction has been described as "packing 40 km (24 miles) of extremely fine thread into a tennis ball!" (1). The cell has developed mechanisms to organize the genome and ensure functional fidelity for the many essential nuclear processes that must occur for survival.

At the largest scale of genomic organization, chromosomes preferentially occupy certain 3D locations within the nucleus, denoted as "chromosome territories" (2). Chromosomes are macromolecular structures that can range in size between 240 Mbp and 19 Mbp of DNA (3; 4). These territories were first described and annotated via DNA fluorescent in situ hybridization (FISH) (2; 5). At this macro level, it has been observed that genomic regions typically associated with the nuclear periphery represent transcriptionally repressed regions of the genome, while regions that are transcriptionally active are localize towards the center of the nucleus (6). Distributed throughout the genome are gene regulatory elements, such as enhancers, promoters, and repressor regions. Since the discovery of the operon, gene regulatory elements have been thoroughly investigated to uncover mechanisms governing gene expression (7). Of note, upon completion of the Human Genome Project, mutations and aberrations in regulatory elements are most frequently attributed to disease phenotypes

(8; 3; 4).

Although chromosomes preferentially occupy specific regions of the nucleus, the organization of these structures are dynamic, leading to fluctuations in transcriptionally active or repressed regions. The human genome is currently classified into two "compartments", compartment A and compartment B. These classifications were first described within the seminal work of Lieberman et al (6), in which a genome-wide investigation of 3D organization was accomplished through the development of the chromosome conformation capture methodology, Hi-C. Previously, iterations of chromosome conformation capture methodologies, such as 3C, 4C, and 5C, could only assay the interactions of a specific locus of interest with either another single locus, such as in 3C, or eventually against the entire genome, as demonstrated in 5C. The development of Hi-C established the first all vs. all assay to assess which regions of the genome are in close proximity to each other within a population of cells. Briefly, these chromatin capture technologies utilize crosslinking agents to "lock" nuclei, and therefore the 3D organization of the genome. By permeablizing nuclei and digesting the genome with restriction enzymes, the entire genome can be fragmented. Following this digestion, the addition of biotinylated nucleotides fill the ends of digested DNA fragments that are close in proximity. Blunt ligation connects these chimeric dsDNA molecules which can then be isolated by streptavidin affinity purification. Mapping these isolated DNA fragments produces contact density maps which map all loci of the genome and the corresponding regions that were in close proximity upon initial nuclear crosslinking.

These contact density maps gave us our initial insights into the 3D organization of the human genome across multiple cell types. Transcriptionally active regions (compartment A) of the genome were associated with gene rich regions of the genome and activating epigenetic chromatin marks such as H3K4me3. The B compartment was decorated with repressive marks such as H3K27me3. Genomic regions that possess these characteristics are typically 1 megabase in length and illustrate a fundamental aspect of genome organization:

insulation. Genomic insulation is an organization principal promoting intracompartment associations in lieu of spurious intercompartment interactions. Through innovation of the Hi-C methodology and the utilization of different restriction enzymes, Rao et al. (9) were able to map finer interactions within these compartments; therefore, we began identifying and dissecting Topologically Associated Domains (TADs) and the chromatin loops within them.

TADs within the human genome are typically between 0.1 - 1Mb in length. Similar to compartments, the DNA sequences within TADs typically interact with one another more frequently than participating in inter-TAD contacts. TAD boundaries are demarcated with the presence of convergent CCCTC-binding factor (CTCF) binding sites. Known mainly as an insulating, or repressive, transcription factor, binding of CTCF to these boundary sites and interaction with the ring-like complex known as Cohesin, has established the foundational role that these architectural proteins play in human genome organization. The loop extrusion model, coupled to biochemical investigations dissecting the interface of the CTCF and Cohesin interaction elucidated the mechanism responsible for the formation of loops within TADs. These loops can mediate interactions between genes and regulatory elements that reside within these TADs. Importantly, when compared at the megabase scale, different cell lines exhibit similar contact densities of genomic compartments (9). This observation even extends to other species, as syntenic genomic regions between mice and humans demonstrated similar patterns of interaction within this study (9). These observations unveil common organization principles of closely related metazoans but, importantly, there are deeper, finer, contact differences present at the next level of genomic organization: chromatin loops.

Chromatin loops are typically sub 1-Mb yet greater than 100 kb, containing a few genes and the cis-regulatory elements that can act upon them (either enhancers or repressors). In a cell-type specific and gene expression dependent manner, increased contacts are seen

between regulatory elements and the promoters of genes that they regulate, demonstrating that these finer through-space interactions play a regulatory role for transcription and therefore, cell identity. This corroborates with the functional binding of transcriptional machinery to regulatory elements distributed throughout the genome. Cell-type specific transcription factors typically bind accessible cognate dsDNA sequences and coordinate the localization of downstream protein interaction partners. For transcriptional activation, TFs participate in "fuzzy" (not conformationally rigid or defined) binding interactions with the Mediator complex (10) and form a through-space bridge between enhancer elements and gene promoters, resulting in the increase in contact frequency observed from Hi-C. A similar mechanism is observed for repressive elements. TFs that belong to the Krüppel-like factor (KLF) family of TFs recognize their cognate DNA binding sites throughout the genome and mediate transcriptional repression via association with C-terminal binding domain protein (CtBP). This interaction mediates the further recruitment of histone deacetylase complexes (HDACs) which remove activating chromatin marks such as H3K27ac. The removal of this activating mark promotes the compaction of DNA and thus inhibits the association of transcriptional machinery necessary for gene activation. It is evident that genome organization and the association of nuclear proteins play a pivotal role in regulating gene expression.

There are many questions inherent to the governing principles of genome organization. Although we've been able to identify TADs and chromatin loops, and appreciate their insulatory capabilities, there is still the question of specificity between regulatory elements. How do enhancers associate with a specific gene within a TAD instead of any of the other genes that may also be in close proximity? Interactions across long distances (more than 50 kb) have also been observed. The same enhancer element can be active in more than one cell type yet activate a different subset of genes within these transcriptional programs. What are the factors that confer this kind of specificity?

It has been observed that the genome can undergo large-scale reorganization in response

to cellular stresses. The foundational example of such large scale reorganization is the heat shock response (11; 12). In this state of stress, Heat Shock Factor 1 (HSF1) is relieved of repression via HSP70 and binds to Heat Shock Reponse Elements (HREs) within the promoters of genes that are critical to respond to this environmental stress (13). Microscopy work has shown that there is a coalescence of genes across multiple chromosomes which establish a currently accepted structure of transcription — a "transcription factory" or hub (14). This colocalization of genes and transcriptional machinery induces the formation of phase-separated bodies, a result of liquid-liquid phase separation (LLPS). Although exemplified in the heat shock response, LLPS and the establishment of transcriptional hubs are a currently accepted occurrence within the field of transcription and correlate with robust transcriptional output (15; 16). Be that as it may, there is controversy surrounding the necessity for phase separation to facilitate transcriptional firing, the regulatory implications that phase separation can confer, the molecular machinery within these phase separated bodies, and the frequency in which they occur.

A cell's gene expression program defines its functional output and the way that the cell can maintain homeostasis by interpreting the various signals and stresses it may encounter. Therefore, investigation into the many layers of gene regulation that ensure proper transcriptional programming is paramount. While there are many types of genetic and epigenetic mechanisms at play within the field of chromatin biology, this thesis will delve into two regulatory components that play major roles in determining eukaryotic cellular identity: 1. transcription factors and 2. chromatin modifications.

Regarded as interpreters of the genome, transcription factors are seen as "master regulators" that can control cell fate (17; 18). Transcription factors are often modular: typically composed of characterized well-structured DNA binding domains coupled to disordered "effector" domains. The DNA binding domains of TFs recognize specific DNA sequence motifs distributed throughout their respective genomes, while their effector domains predominantly

participate in protein-protein interactions. This coupling endows TFs with a diverse spectrum of functionalities (19) (Figure 1.1). Although frequently associated with their roles in transcriptional activation, TFs can also mediate mechanisms of transcriptional repression. These capabilities delineate different classes of transcription factors and the roles they can play in gene regulatory networks. These classes can be defined as 1. general TFs 2. cell-type specific TFs, and 3. "house-keeping" TFs.

In the paradigm of transcriptional activation, transcription factors play an essential role by localizing and spatially coordinating the necessary components for proper transcriptional firing. Ubiquitously expressed general (or basal) TFs, such as TFIIB, TFIID, TFIIE, TFIIF, and TFIIH were among the first identified TFs due to their associations with eukaryotic RNA polymerase II after separation via phosphocellulose and ion exchange chromatography. (20; 21). The general TFs begin to encode levels of specificity and regulation for gene expression (22; 21). An example is seen with transcription factor TATA-binding protein (TBP), a component of TFIID, which recognizes TATA DNA elements that can be found within the promoters of expressed genes. This binding then enables association with TFIIB which will be the docking site for the rest of the assembly of the RNA polymerase II pre-initiation complex. While basal transcription factors are ubiquitously expressed within metazoan cells, many transcription factors are the downstream effectors of extracellular stimuli and can mediate transcriptional firing at cell-type specific subsets of genes containing their respective DNA-binding motifs.

Inherent to their name, lineage-specific TFs are typically only expressed within differentiated cell types and enable the transcription of genes necessary for the functional outputs of these cells types. It has been shown that lineage specific TFs typically occupy super-enhancers with other members of the transcriptional machinery (such as Mediator) within differentiated cells or tissues (8; 17). While some cell types have well characterized TFs associated with their gene expression programs (TBX5, GATA4, and TBX20 within cardiac

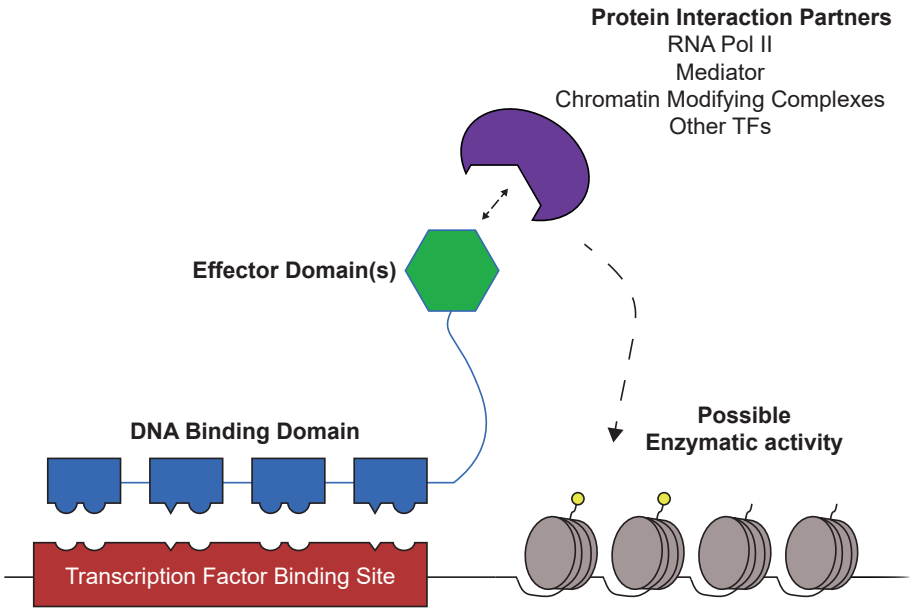


Figure 1.1: Transcription factors are modular in design and recognize specific DNA sequence motifs with their DNA binding domains. A cartoon depicting the canonical composition of transcription factors.

cells; or SOX2 and OLIG1 within neuronal cells found in the cerebral cortex), the combinatorial expression of TFs provides a vast diversity in possible gene expression programs. This transcriptional diversity makes it difficult to correctly predict how perturbations, both

within the amino acid sequence of the TF or within the regulatory sequences they identify, will ultimately influence cell fate (23; 24; 25; 26; 27; 10). A prime example of the power TFs can exert upon cellular identity is best demonstrated by the Yamanaka factors (18). The Yamanaka factors (c-Myc, Sox2, Oct4, and Klf4) can induce reprogramming of differentiated cells back to a more pluripotent state, enabling plasticity between cell states.

As for their repressive properties, TFs have multiple mechanisms by which they can influence transcriptional repression. This type of activity is essential in preventing spurious transcriptional activation after the establishment of a lineage-specific transcriptional program. A common mechanism for transcriptional repression is the prevention of the association of RNA polymerase to a downstream gene. A foundational example to our understanding of gene regulatory networks is the lac operon. A heterodimeric protein binds upstream genetic elements and prevents transcription of downstream genes required for the metabolic breakdown of lactose (28; 7). If the cell is exposed to an environment with lactose as its main energy source, the metabolic byproduct allolactose will bind to this inhibitory protein and induce a conformational change that relieves this repression. Another mechanism related to transcriptional silencing is the incorporation of TFs into histone modifying complexes. The *Drosophila* TF *Pleiohomeotic* (PHO), is a member of the Polycomb group complex and recognizes Polycomb Response Elements (PREs) to direct and install repressive chromatin marks at homeotic genes essential for proper development (29). An emphasis should be placed on the critical nature of the protein interaction partners for these transcription factors. TF functional activity can be modulated based on contextual cues within their local environments. Due to the complexity of these environments and the diversity of TF functionality, it is essential to identify and catalog the finer principles which govern TF activity.

This thesis delves into the regulatory principles of the TF, Yin Yang 1 (YY1). Discovered and cloned by three labs simultaneously in 1991, this highly pleiotropic TF can act as both a transcriptional activator or repressor (30; 31; 32). Human YY1 is composed of 414 amino

acids and contains four C_2H_2 zinc fingers towards its C-terminal end (AA 298-397). Its DNA binding domain recognizes the consensus sequence 5'-CGCCATNTT-3', which is found in the promoters of both metazoan and viral genes. YY1 has been shown to be essential for development in multiple capacities. Homozygous YY1 knock out (KO) experiments induce developmental lethality in mice, with YY1 KO cells capable of reaching the initial blastocyst stage though fail to develop in the gastrulation phase during implantation within the uterine tissue (33). Furthermore, these developmental dysregulations occur in a dosage dependent manner. By utilizing a Cre/LoxP system to attain stratified levels of YY1 expression and avoid embryonic lethality, (34) et al. observed dosage-dependent survivability across mouse cohorts. Mice that expressed 25% of WT YY1 expression levels died between 13.5 and 14.5 days post coitum. Those that survived past these time points were typically smaller in size and had issues in development of their lungs. These changes were attributed to the roles that YY1 can play in cytokinesis and cellular proliferation (34). YY1 haploinsufficiency has also been attributed to neuronal defects, such as Gabriel-de Vries (GADEVs) disease (35). Thus, understanding how YY1 is governed to perform such divergent transcriptional outcomes is important.

In terms of activation, YY1 can bind to Initiatior elements upstream of transcription start sites and induce transcriptional firing in the absence of general transcription factors such as TBP (36). As YY1 has been observed to typically bind cell-type specific enhancer and promoter elements, this type of transcriptional recruitment likely plays a role in setting cell-type specific expression programs (37). Direct interactions with core transcriptional machinery or general transcription factors are not the only means of transcriptional activation for YY1. It has also been observed that YY1 can associate and then direct chromatin modifying complexes, such as the BAF complex, to target genes to modify the local chromatin environment for transcriptional activation (38). Because YY1 is ubiquitously expressed across human cell lines, yet tends to occupy enhancers and promoters of cell-type specific

genes, hypotheses have claimed that YY1 is an architectural genome organization factor that promotes through-space interactions within a local chromatin environment (37; 39).

In terms of repression, YY1 has been shown to compete with other activating TFs for overlapping consensus sites (40; 41). This competition causes displacement of these activating factors and renders these specific genes transcriptionally silent. In direct opposition to the previously noted example involving the BAF complex, YY1 can associate with Polycomb Group proteins and recruit these repressive complexes to specific subsets of genes (42; 43; 44). Again, the nature of whether YY1 acts as a repressor or activator depends mainly on its protein-interaction partners, therefore the question of how does YY1 (or TFs as a whole) navigate through the complex environment of the nucleus and adhere to a proper functionality is a query worth investigating.

Although multiple methodologies have been developed to assay TF occupancy (ChIP-seq (45), ChIP-exo (46), CUT&RUN (47)) an open question in the transcription field still pertains to how TFs associate with their sites of occupancy within the nucleus. Despite cataloging cognate sequence motifs for a large number of TFs (19), chromatin accessibility, as well as DNA binding motifs, are not sufficient to provide the rationale for all observed binding events (48; 49). What are further elements of the nuclear environment that could aid in genomic occupancy?

Although the central dogma describes the relationship from DNA to RNA to protein, this information flow mainly describes the pathway for mRNA production to functional protein. With the advent of RNA-seq and the completion of the Human Genome Project (4; 3), we have become aware of the pervasive nature of noncoding RNA production (ncRNA). Only 1-3% of the human genome is comprised of protein coding genes and yet, 75-90% of the genome is transcribed into RNA (50). This eye-opening observation has launched an entire field of molecular biology characterizing the world of ncRNAs. Many noncoding RNA molecules had been identified before and are parts of essential cellular machinery, typically in ribo-

nucleic acid protein complexes (RNPs). Ribosomal RNAs (rRNAs), are components of the ribosome, which performs the essential process of protein synthesis (51; 52; 53). PIWI RNAs (piRNAs), are surveillance RNA molecules involved in PIWI mediated RNA degradation and are essential for distinguishing self vs non-self nucleic acid molecules within cells (54; 55; 56; 57). Small nucleolar RNAs (snRNAs) are components of the spliceosome, an RNP machine necessary for splicing pre-mRNA transcripts to prepare them for proper nuclear export and ultimately translation (58). In addition to these types of ncRNA production, long non-coding RNAs (lncRNAs) have also emerged as regulatory molecules within the nucleus.

lncRNAs are RNA molecules that are typically lowly expressed, >200 bp in length, under polyadenylated, and exhibit low amounts of sequence conservation between organisms (59; 60). Their mechanisms of transcriptional regulation revolve around modulation of local chromatin architecture, either by association with factors related to the act of transcription (TFs or chromatin modifying complexes), or by structural rearrangement of the genome to faciltate contacts betwen regulatory elements and genes (61; 62; 63; 64; 65). Characteristic and well-studied ncRNAs include XIST, Tsix, and HOTTIP (61; 65; 66). XIST is a central component of mammalian X chromosome inactivation. After its initial act of transcription, Xist spreads across the soon-to-be inactivated X chromosome and recruits the repressive histone modifying complex PRC2. Deposition of H3K27me3 drives the compaction of the inactive X chromosome into a heterochromatic, repressed, region of the genome that eventually relocates to the nuclear periphery (61; 67).

Tsix is a ncRNA antisense and downstream of Xist (66). Via CLIP-seq, Tsix was shown to bind the genome organizing TF, CTCF. This binding interaction has been shown to be necessary for "X chromosome pairing" an upstream process essential for proper X chromosome inactivation. Tsix is posited to help recruit CTCF and tether this TF to these genomic regions as shRNA knockdown and LNA gapmer directed knockdown of Tsix reduced the frequency of X chromosome pairing, concomittant with loss of CTCF binding at the X

chromosome pairing sites.

HOTTIP is a ncRNA produced near the developmentally essential homeobox cluster within metazoans (65). Through interactions with the WDR5 subunit of the MLL1 complex, HOTTIP induces deposition of H3K4me3 and activation of the Hox cluster in a temporally-sensitive manner in order to facilitate proper development of metazoans which possess the HOX cluster. Through these protein interactions, ncRNAs, and the widespread production of RNAs within the nucleus have demonstrated that they play a role in overall genome organization. Multiple techniques have been developed to map the interactome of these ncRNAS, such as RNA antisense purification (RAP) (68), Capture Hybridization Analysis of RNA Targets (CHART) (69), Mapping RNA-genome Interactions (MARGI) (70), GRID-seq (71), and SPRITE (72). These methodologies highlight the cis and trans interactions that these ncRNAs partake in and provide a basis for the inquiries assessing the regulatory roles that RNA can play within the nuclear environment. A subclass of lncRNAs are chromatin associated non-coding RNAs (cheRNAs) which have been observed to be cell-type specific RNA molecules that regulate nearby genes in *cis* (63; 62). In addition to these observations, TF-dependent cell-type specific cheRNA networks have also been described (64).

Recent work has shone a light on the prevalence of unannotated RNA binding domains pervasive throughout transcription factors (73; 74). The functional impacts of such binding events are not known but current studies are providing insights as to how transient RNA interactions may influence TF functionality (75; 76; 77; 78). Until there is delineation of the RNA binding interfaces and investigation into the principles regarding specificity for such RNA binding, sound conclusions cannot be drawn towards the modulatory properties of these previously unappreciated RNA binding capabilities. In this work I investigate the RNA binding properties of YY1 and identify aspects of these RNA interactions that may hold true to a larger set of TFs.

1.2 Histone modifications and their attributed roles in gene regulation

The basic unit of compaction within the eukaryotic genome is the nucleosome, the building block of chromatin. 147 base pairs of DNA wraps a hetero-octameric protein complex consisting of the four core histone proteins H2A, H2B, H3, and H4 (79) (Figure 1.2).

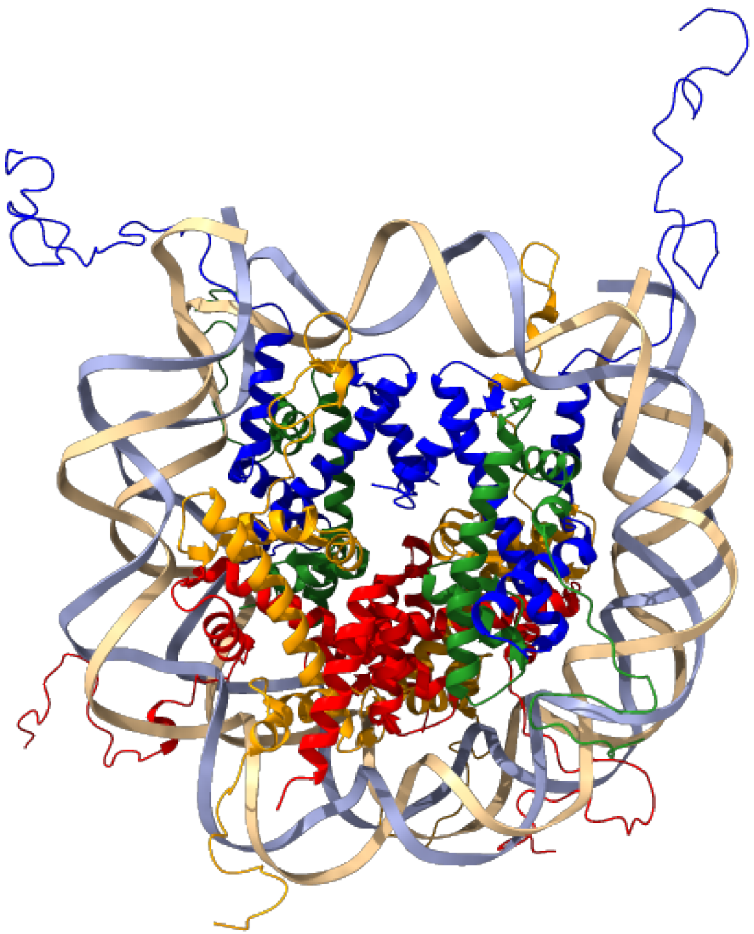


Figure 1.2: The nucleosome core particle. Crystal structure of the nucleosome core particle depicting the association of DNA around an octamer of the histone core proteins: H2A (in orange), H2B (in red), H3 (in blue), and H4 (in green). Adapted from Davey et al. (80) PDB:1KX5

The nucleosome core particle possesses another layer of regulation — the post-translation

modification of the histone octamer. A myriad of histone modifications have been detected and the functional impacts between the presence of these marks and the transcriptional outputs of their resident genomic regions have been documented (81; 82) (Figure 1.3). Of the milieu of histone modifications that have been observed, this thesis work looks into the interplay between the activating mark H3K4me3 and the repressive mark, H3K27me3.

Chromatin modifications are dynamically modulated via the interplay of readers, writers, and erasers. Readers are proteins, typically subunits of larger protein complexes, that are specifically recruited to epigenetic marks that they recognize for further catalytic activity. An example would be the interaction between H3K27me and Polycomb Repressive Complex 2 (PRC2) recruitment. PRC2 is the sole chromatin modifying complex responsible for deposition of H3K27 methylation states and forms a positive feedback loop of methylation (83). This methylation induces compaction of these genomic regions, resulting in transcriptional repression and the formation of heterochromatin.

H3K4me3 is typically found at active promoters and enhancers of cell-type specific genes (84; 85; 86; 87; 88; 89). This type of modification correlates with "open" chromatin, allowing the association of transcriptional machinery and the initiation of transcription. A cursory view of chromatin epigenetic landscapes would posit the deposition of "activating" or "repressive" chromatin marks to be binary, however, this is not the case. Given that each nucleosome contains two copies of each core histone, and that modifications can even be deposited along the same histone tail, it is possible for a nucleosome to possess both of these opposing marks. The co-deposition of activating and repressive chromatin marks has been coined as "bivalent" chromatin and has established an entire field of work that embraces two principles which govern this biological phenomenon (90; 91; 92).

The first aspect of bivalency is that activating and repressive marks, such as H3K4me3 and H3K27me3, mark a small subset of lineage-specific genes within pluripotent cells and render these genomic regions "poised" for transcriptional regulation. The second principle

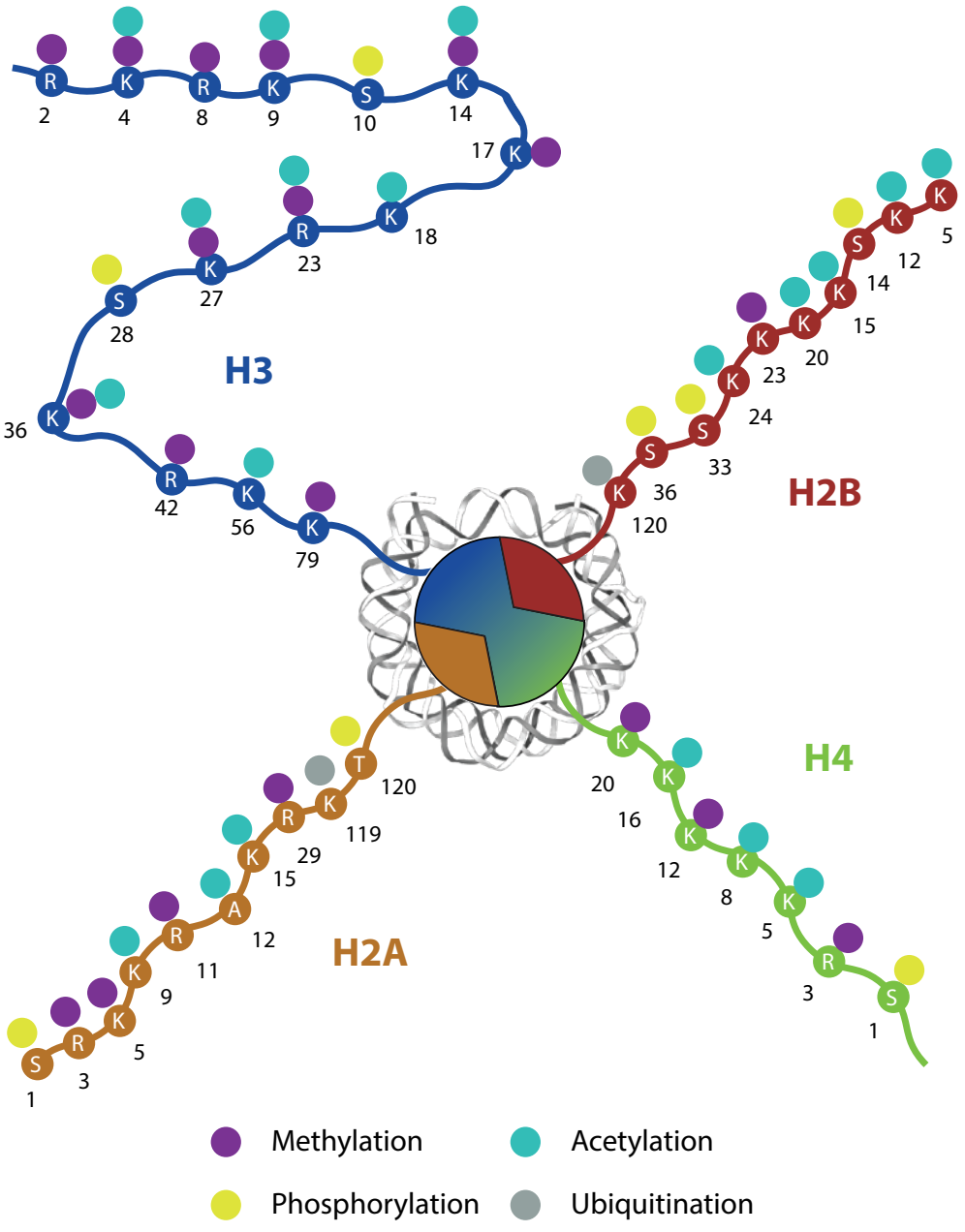


Figure 1.3: A sampling of histone post translational modifications. Cartoon depicting the variety of histone modifications that can be deposited on histone tails.

states that upon differentiation, these marks would "resolve" into a terminal transcriptional state, either active or repressed (based on the presiding chromatin mark), and commit the

undifferentiated cell type into a more terminal cell state.

While the bivalency model posits an intuitive explanation for lineage determination, the initial studies foundational to this model have many caveats that I address within this thesis. To understand and draw correlation between histone modifications and transcriptional activity, the primary methodology to assess modification presence through the genome has been chromatin immunoprecipitation, (ChIP) (45; 93). In this method, nuclei are isolated, chromatin is fragmented and incubated with an antibody specific for the modification of interest (or protein of interest as discussed earlier with TFs) in order to immunoprecipitate and ultimately identify the genomic regions that harbor your factor of interest. By comparing the abundance of retrieved genomic sequences to a proper control library, one can assess enrichment for specific regions of the genome and begin making claims denoting occupancy of your factor/moiety of interest.

Since its development, ChIP has become one of the most utilized methodologies in molecular biology. However, there are many caveats inherent to the methodology that have not been properly addressed across the field of chromatin biology. First, calculating enrichment of a genomic locus is a ratiometric readout that is dependent on the sequencing depth of your input material and abundance of your chromatin mark. This is demonstrated in (94) where two cell populations can have drastically different levels of the H3K79me2 chromatin mark yet, with traditional ChIP-seq data analyses, these quantitative differences are not captured when assessing genomic loci for enrichment. Inherent to the traditional way of analyzing ChIP-seq, normalization scales of input are not consistent between experiments, thus rendering quantitative comparison between experiments murky to assert.

These limitations to ChIP-seq pale in comparison to a fundamental oversight within an essential component of the protocol: antibody specificity. Histone modifications can constitute "minute" changes in chemical composition, an single methyl group differentiating H3K27 mono versus di versus trimethylation. Utilizing highly specific antibodies that can

distinguish between these modifications is essential in order to properly attribute organizational and functional outputs to the presence of these marks. It has been shown that many antibodies that were conventionally used for ChIP-seq experiments investigating the deposition of H3K4 methylation states, recognized off-target methyl forms which clouded our interpretation of the coexistence of these marks (95). Additionally, it was shown that peptide arrays, the "gold standard" for assessing whether antibodies used in ChIP experiments were highly specific, were a poor indicator of actual specificity in this experimental format.

The Ruthenburg lab has developed internally calibrated chromatin immunoprecipitation (ICeCHIP) to address the limitations of conventional ChIP(96). By generating semi-synthetic nucleosomes bearing on and off target modifications and spiking in this material at the start of the ChIP workflow, ICeChIP is capable of assessing antibody specificity as well as provide a quantitative measure of the abundance of a mark at a given genomic loci. Utilizing this technique, this thesis tests the central aspects of the bivalency model and assesses its merit wholistically.

1.3 Open questions and this thesis

Thus far, I have described how TFs and histone modifications can balance transcriptional activation and repression to ensure proper gene expression. Both layers of regulation have distinct open questions that I investigate within this thesis. Due to the recent elucidation of the pervasive nature of RNA binding within TFs, delineation of what domains confer this binding activity, the level of specificity of the interactions, and the overall functional implications of such binding events are currently unknown. In regards to histone modification distribution, the chromatin field is still turning a blind eye towards the oversights outlined above. These practices obfuscate the conclusions we can make regarding bivalent chromatin and thus require rigorous and quantitative investigation to reach a true understanding of the

regulatory mechanisms at play. My work provides frameworks for addressing these knowledge gaps and highlights the intricacies that can be uncovered by fundamental biological research.

In Chapter 2 I set out to understand how RNA binding influences the functionality of the ubiquitously and constitutively expressed TF, Yin Yang 1. Initially, conflicting pieces of literature ascribed RNA binding capabilities to different regions of the protein. To delineate if both or neither conclusions were true, I utilized a bacterial expression system and protein biochemistry to design, express, and purify full length YY1 and various other truncation and mutant protein fragments. Using biophysical assays to query nucleic acid binding interactions and protein secondary structure, I uncover a previously unannotated nucleic acid binding domain of YY1 as well as intrinsic autoregulatory characteristics for the full length TF that impact both of the nucleic acid binding domains.

In the third chapter of this thesis, I leverage the quantitative nature of ICeChIP to probe into the functional implications of bivalent domains. From the development of a sequential version of ICeChIP, which can quantify the abundance of nucleosomes containing both H3K4me3 and H3K27me3 chromatin marks, I look into the prevalence and fate of bivalent domains across a differentiation scheme taking E14 mouse embryonic stem cells to neuronal precursor cells. With these quantitative insights, we find that the central tenets of the bivalency model do not hold true for our differentiation path.

From these investigations, it is my hope that this work sheds light on mechanisms governing the fine line between transcriptional activation and repression. My next hope is that this work can provide generalizable and applicable insights into TF activity modulation and the coexistence of chromatin marks distributed throughout the genome for further dissection of gene regulatory networks.

CHAPTER 2

THE N-TERMINUS OF YY1 REGULATES DNA AND RNA BINDING AFFINITY FOR BOTH THE ZINC-FINGERS AND AN UNEXPECTED NUCLEIC ACID BINDING DOMAIN

2.1 Attributions

This chapter has been adapted from: Elias J. *et al.* The N-terminus of YY1 regulates DNA and RNA binding affinity for both the zinc-fingers and an unexpected nucleic acid binding domain. Preprint at *bioRxiv*, doi: 10.1101/2024.10.04.616721 (2024). Jane Rosin and Amanda Keplinger aided in purification of YY1 mutant protein constructs as well as fluorescence polarization assays presented in Figure S3B, S4, and S5A. All other experiments were conducted by the author.

2.2 Abstract

Transcription factors (TFs) play central roles in dictating cellular identity and function by regulating gene expression programs. Beyond their well-folded DNA binding domains (DBDs) which recognize cognate DNA elements in the genome, TFs are enriched for intrinsically disordered regions (IDRs), which have a host of proposed functions including facilitating protein-protein interactions, aiding in binding site search, and binding RNA. Defining intrinsic regulatory properties of TFs requires further mechanistic investigation. We chose to investigate the DNA and RNA binding properties of Yin Yang 1 (YY1), a ubiquitously expressed TF directly involved in transcriptional activation, repression and genome architecture. Through systematic in vitro nucleic acid binding experiments we resolve conflicting literature defining the RNA binding interface of YY1, demonstrating that there are two RNA binding domains within YY1: its canonical 4 zinc finger DBD and a previously unannotated

nucleic acid binding domain, which we term the REPO-NAB. Furthermore, we discover surprising autoinhibitory properties that the N-terminus of the protein imparts on each of these binding domains. Our results provide a new example of IDR-mediated regulation within TFs and enables future mechanistically precise functional investigations.

2.3 Introduction

Transcription factors (TFs) play a central role in dictating cellular identity (8; 18). Typically modular in composition, TFs are minimally composed of a DNA-binding domain, which imparts sequence-specific DNA recognition, and an activation/effector domain, which interfaces with nuclear protein complexes to recruit or stabilize their local activity (19; 10). Yet for a given TF, only a fraction of accessible cognate sites in the genome are occupied (19; 97; 98; 99). Precisely how TFs are localized to subsets of regulatory genetic elements distributed throughout the genome to define distinct gene expression programs remains unclear (99; 100; 101; 75). The prevailing explanation for this disconnect is that there are additional genomic interfaces, either through DNA, RNA, or co-factor proteins that further specify localization through multivalent energetics (19; 10; 102; 103; 104; 105; 106). Yet for most transcription factors, detailed mapping of these interfaces and their specificities is lacking such that the conventional explanation for site-specificity remains untested. To begin to explicitly test these ideas, we chose to biochemically dissect the nucleic acid binding interface(s) of the constitutively expressed transcription factor Yin Yang 1 (YY1) (30; 107; 37), and define the intrinsic autoregulatory features within the protein itself. YY1 derives its name from initial observations describing the TF's ability to be both a transcriptional activator and repressor of the adeno-associated virus P5 (AAVP5) element (30). While this distinction is attributed to the presence or absence of the E1A cofactor, how YY1 accomplishes these diametrically opposed transcriptional activities in other cases has been a matter of long-standing interest. Context-dependent mechanisms ranging from the ability to inter-

act with core transcriptional machinery (108), engage in DNA repair (109), recruit histone modifying complexes (42; 38; 110), and promote through-space contact of cell-type specific enhancers and promoters (37) have all been attributed to this multifunctional TF.

The canonical DNA binding domain of YY1 consists of 4 C2H2 zinc fingers (ZnFs) (111) situated at the C-terminal end of the protein (112), while the N-terminal portion harbours regions of the protein mapped to activation and/or repressive activities annotated by specific functional studies (30; 107; 108; 42; 38; 110) (Figure 2.1A). In addition to DNA binding, YY1 is capable of binding RNA (113; 39; 74; 114; 115; 61; 116). However, there is a standing controversy within the literature regarding the RNA binding interface of YY1 (113; 39; 74) rendering experiments that precisely perturb this interface to ascertain the regulatory impact fraught. Previous work has postulated that RNA binding plays a role in both YY1's homo-dimerization (37), as well as mediating a transcriptional positive feedback loop, where transcriptional machinery is maintained in the local vicinity of gene regulatory elements upon proper transcriptional firing (39). Unambiguous delineation of the impact that RNA binding can have upon YY1 cellular function requires systematic side-by-side comparisons of nucleic acid affinities of putative RNA binding domains compared to full length YY1, and such analyses have not yet been performed .

Although YY1 is ubiquitously expressed across human cell types and consistently occupies cell-type specific enhancers and promoters, the dominant factors which play a role in determining YY1 genomic occupancy have remained elusive. By taking a nucleic acid-centric perspective and evaluating YY1 binding events from multiple data sets (39; 117; 118), we observe that the genomic occupancy of YY1 cannot be explained by a combination of its cognate dsDNA binding motifs defined by SELEX nor its possible RNA binding motifs derived from CLIP-seq (Figure 2.1B). These observations prompted our investigation into previously unannotated characteristics of the TF.

In this study, we measure the nucleic acid binding capacity of purified YY1 and frag-

ments thereof using fluorescence polarization. We observe unexpected binding properties for multiple domains of YY1: (1) YY1's canonical DNA binding ZnF domain exhibits a nearly 10-fold higher affinity for single stranded RNA than double stranded DNA consensus motifs and both nucleic acid types share an overlapping interface; (2) YY1's Recruitment of Polycomb (REPO) domain (42) possesses previously unannotated tight binding capacity for both DNA and RNA with little apparent sequence specificity. (3) The N-terminus of YY1, which is predicted to be disordered, can inhibit nucleic acid binding to the ZnFs and REPO domain, suggesting an autoinhibitory intra/inter-molecular mechanism to provide fine-tuning of the protein's activity, and this property seems to account for the weaker apparent affinity for RNA displayed by the full length YY1 protein than displayed for the individual nucleic acid binding domains. Our study elucidates a previously unannotated nucleic acid binding domain of YY1, defines several surprising autoregulatory features intrinsic to the protein which will enable functional tests of their properties *in vivo*, and provides a template for further mechanistic dissection of transcription factors.

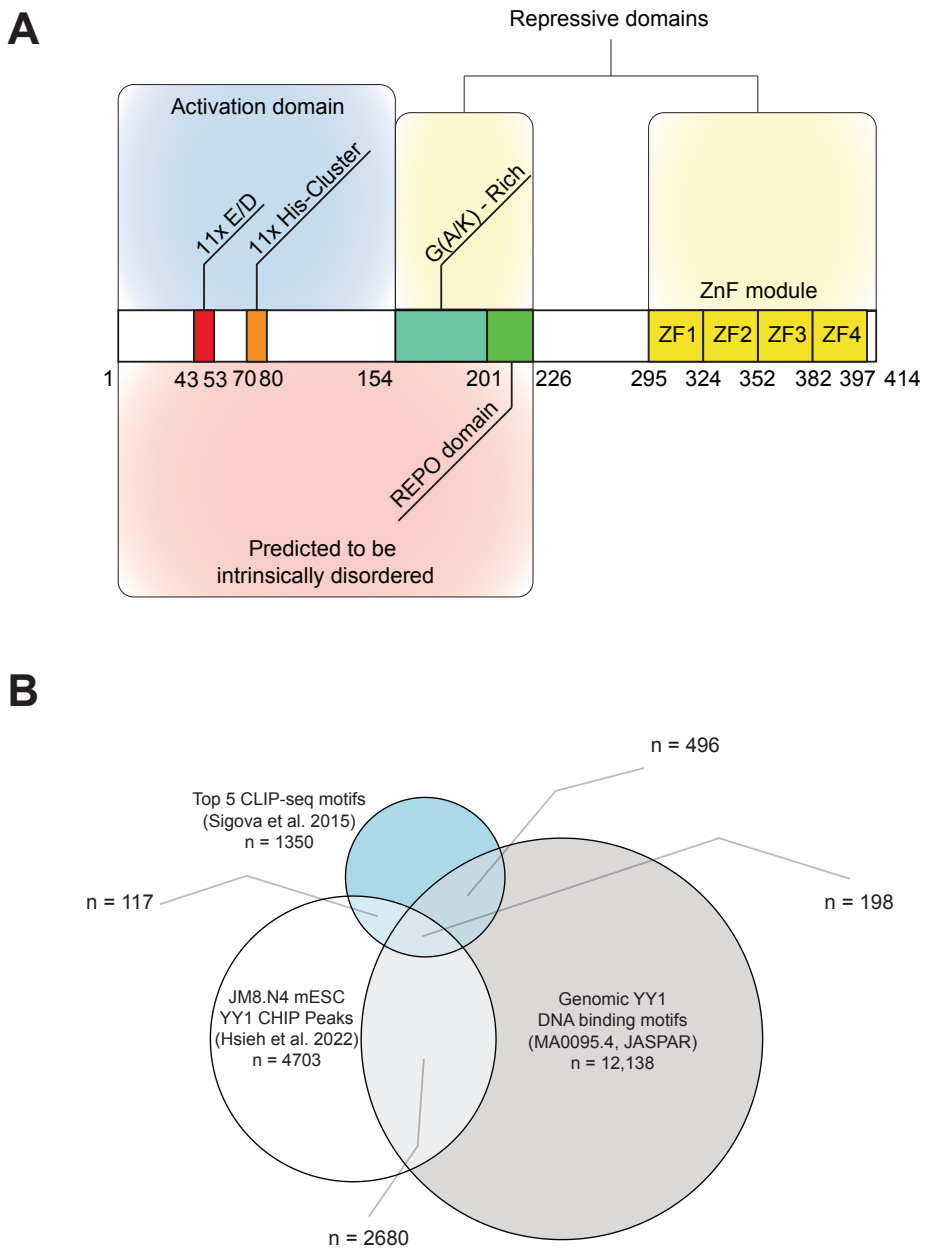


Figure 2.1: YY1 is a multi-functional transcription factor that adheres to the transcription factor occupancy paradox. **(A)** Schematic of YY1, highlighting domains and key functional features. **(B)** Venn diagram showing the overlap of murine YY1 ChIP-seq sites (118), cognate DNA binding sites (JASPAR, MA0095.4), and RNA sequence motifs (39), all within ATAC accessible loci (117)

2.4 Material and Methods

2.4.1 Cloning and purification of YY1 constructs

YY1 protein was purified using a method modified from (113). Briefly, a plasmid containing full length YY1 was a gift from Richard Young (Addgene plasmid # 104396; http://n2t.net/addgene:104396; RRID:Addgene_104396) and the full length protein sequence was inserted into a modified pGEX-6P vector using BamHI/XhoI restriction enzyme sites. All proteins were expressed as R3C-cleavable N-terminal fusions with glutathione-S-transferase (GST). Truncation mutants were generated using QuickChange and HiFi methodologies according to manufacturer's guidelines, with oligonucleotides described in Supplemental Table 1. Protein expression constructs were transformed into Rosetta 2 (DE3) pLysS competent cells and grown in 1L LB cultures containing 34 µg/mL chloramphenicol and 100 µg/mL carbenicillin at 37°C until an optical density at 600 nm (OD_{600nm}) measured ~0.6. Upon reaching this OD_{600nm} cultures were cooled to 25°C then induced with 1 mM isopropyl-β-D thiogalactopyranoside (IPTG) and supplemented to 50 µM $ZnSO_4$, then shaken at 25°C for 18 hours. Cultures were then harvested by centrifugation in a Thermo Sorvall RC 3BP+ at 4000g for 25 min at 4°C, and resuspended in 50 mLs of Buffer A (20 mM Tris·HCL pH = 8.0, 100 mM NaCl, 5% glycerol) supplemented with 5 mM β-mercaptoethanol, 2 mM dithiothreitol (DTT), 1 mM phenylmethylsulfonyl fluoride (PMSF), and 1250 U of Benzonase (Millipore, 71205-3). Lysis was accomplished via 4 passages through an American Laboratories French Pressure cell, then clarified by centrifugation on a Sorvall 3B+ at 25,000g for 25 min at 4°C. All proteins were purified to homogeneity by the sequence of Glutathione Sepharose 4FF (GE Healthcare, 25 mL bed in XJ-50 column) and Heparin HiTrap (GE Healthcare, 5 mL) chromatography, followed by anion/cation exchange chromatography dependent on the isoelectric point of the respective construct. All protein purification workflows are depicted in (Figure 2.2). HRV-3C protease was utilized to cleave the GST tag after the initial GST

column. The size and purity of purified proteins were monitored by SDS-PAGE and protein concentrations were determined by the RC DC Protein Assay (Bio-Rad, 5000122).

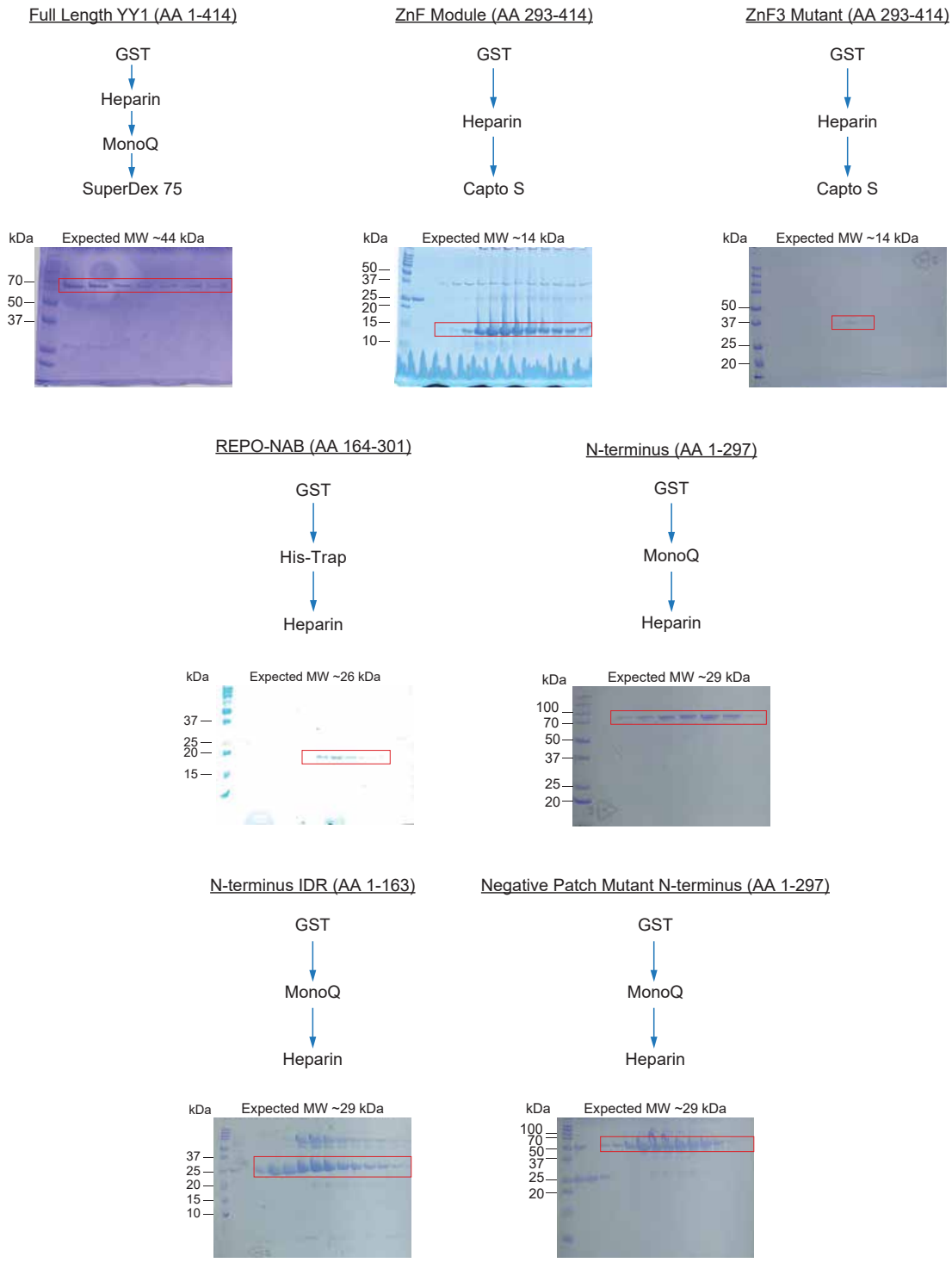


Figure 2.2: Purification schemes for full length YY1 and mutant constructs.

Figure 2.2 *(previous page)*: Depicted are the step-wise approaches for purification of protein constructs utilized within this paper. Expected MWs (kDa) of each construct is reported above each respective SDS-PAGE gel, which showcase the final purity of the protein purified and utilized in subsequent assays. We note that the N-terminus purification yields a protein that migrates at an elevated MW than anticipated and attribute this deviation to the inherent anomalous SDS-PAGE migration of full length YY1 (30) as well as previously observed peculiar migration of the N-terminus (119)

2.4.2 Fluorescence polarization

We used 5' FAM labeled oligonucleotides as the binding substrates in our fluorescence polarization (FP) assays (sequences are listed in Supplemental Table 1). The RNA and DNA oligonucleotides were resuspended in minimal YY1 binding buffer (10 mM HEPES·KOH [pH = 7.3], 50 mM NaCl, 50 mM KCl, 5 mM $MgCl_2$). Duplex DNA substrates were annealed by heating 1.1:1 molar ratio of unlabeled:fluorescently labeled ssDNA oligonucleotides in 1x NATE buffer (50 mM NaCl, 10 mM Tris·HCl [pH = 7.5], 1 mM EDTA) to 95°C followed by slow cooling to room temperature. Purified proteins were dialyzed into YY1 Binding Buffer (10 mM HEPES·KOH [pH = 7.3], 50 mM NaCl, 50 mM KCl, 5 mM $MgCl_2$, 100 µM $ZnSO_4$, 0.01% NP-40), quantified, and nucleic acid substrates were added, to a final concentration of either 10 or 20 nM, to the highest protein concentration within the experiment. This protein-nucleic acid solution was then serially diluted with YY1 binding buffer containing the same concentration of nucleic acid, to a final protein concentration of 10 nM. Concentration ranges are stated within the figure legends for each binding experiment. The reaction volumes for each serial dilution were 165 µL, and this reaction was split into three 50 µL technical replicates within black, nonbinding surface 384 well plates (Corning CLS3575). Fluorescence anisotropy was measured by a TECAN Infinite 200 Pro using excitation/emission wavelengths of 485 nm/535 nm with excitation/emission bandwidths of 25 nm/35 nm at 25°C. YY1 binding buffer was used as a blank and G-factor calibration was performed on YY1 binding buffer containing the proper concentration of labeled nucleic acid

to produce a fluorescence polarization reading of 20 mP. Anisotropy values were calculated using the following equation:

$$A = \frac{I^{par} - I^{cross}}{I^{par} + 2 * I^{cross}}$$

where I^{par} and I^{cross} represent parallel and perpendicular intensities, respectively. Anisotropy data was then analysed in R, normalized, and fit to nonlinear regressions derived from the Langmuir equation and the Hill equation.

Langmuir:

$$y = \frac{x}{x + K_d}$$

Hill:

$$y = \frac{x^n}{x^n + {K_d}^n}$$

y is the anisotropy measurement, x is the concentration of purified protein, and the parameters we are fitting for are K_d, the dissociation constant, and n, the Hill coefficient. The error in the fit is reported as uncertainty and all reported K_d values have correlation coefficients ≥ 0.9. We chose to present the Langmuir K_d values as the binding is anticipated to be monovalent based on substrate design.

2.4.3 Circular dichroism

Circular dichroism spectra were recorded on a Jasco J-1500 CD Spectrometer equipped with a thermoelectrically controlled cell holder using a quartz cell with a 1.0 mm optical path length. Purified proteins at 0.05 mg/mL were dialyzed into 10 mM Phosphate buffer pH = 7.3 with 150 mM NaCl and scanned between wavelengths from 170 to 260 nm at 25°C. The molar ellipticity from three scans were averaged to provide the final curves presented. For the melt curve analysis, the REPO-NAB protein was continuously scanned over a temperature range from 25°C to 98°C increasing at a rate of 0.1°C/min.

2.4.4 RoseTTAFold and AlphaFold3 structural predictions

YY1's full amino acid sequence was submitted to RoseTTAFold and AlphaFold3 with default parameters.

2.4.5 Computational analyses of YY1 genomic binding sites

Data from previous studies (39; 118; 117) and ENCODE were used to generate the Venn diagram depicting YY1's adherence to the transcription factor binding paradox. HOMER was used to identify the top CLIP-seq motifs for YY1 by using the command findMotifsGenome.pl with the following parameters: -size 100 bp -rna -noknown -nocheck -len 12,18, 25, 30 -mis = 3 -bg BirA_BothStrands_CTRL_CLIP with this file representing background CLIP-seq reads from a BirA pulldown from a cell line with untagged YY1. CompareMotifs.pl was then used to reach a final list of 22 motifs in which the top 6, which represented $\tilde{1}0\%$ of YY1 CLIP-seq binding sites, were used for further analysis. Two of these sequences are used within the project as "Bioinformatically Derived Sequence 1 and 2" (Figure 2.3 B). Next, we used the JASPAR web database (MA0095.4) to identify genome wide YY1 cognate binding motifs within mm10 and JM8.N4 mESC YY1 CHIP-seq data was obtained (118). Finally, we utilized ENCODE E14.5 ATAC data (117) to identify accessible regions of the mouse genome. We intersected all three previously mentioned datasets with this ATAC dataset to selectively choose motifs and binding sites that are within E14.5 mESC accessible regions. After these initial intersections, we used bedtools -intersect to identify overlapping regions across the datasets and plotted the number of overlapping regions in R using eulerR.

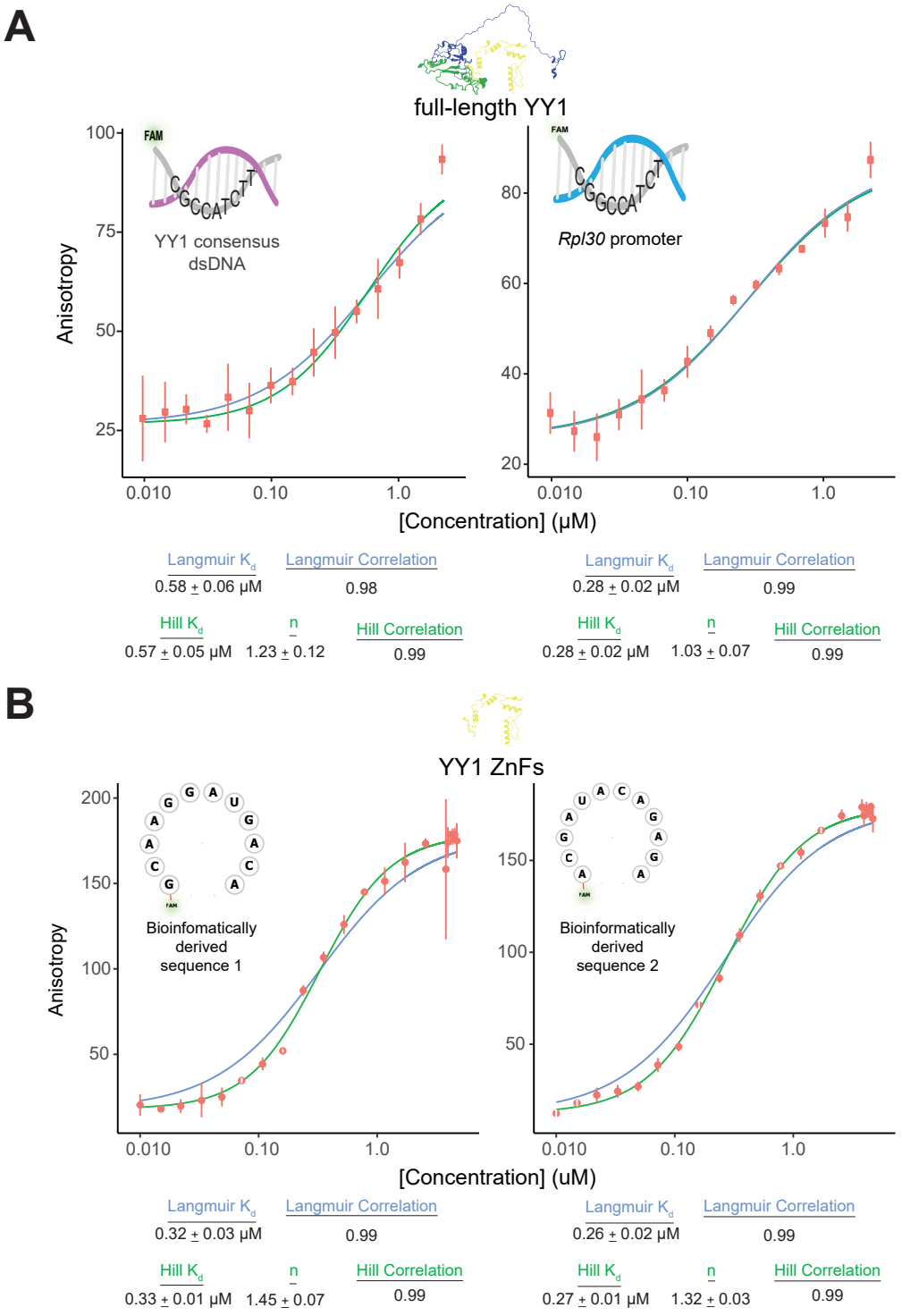


Figure 2.3: Full length YY1 and the ZnF module exhibit characteristic specific activities.

Figure 2.3 *(previous page)*: **(A)** Full length YY1 fluorescence polarization assays against the YY1 consensus dsDNA sequence and the *Rpl30* dsDNA promoter. Error bars represent standard deviation from three technical replicates and the nonlinear regression fits of both Langmuir (blue) and Hill (green) equations are depicted for each graph. Langmuir observed K_d measurements are reported below each binding curve, as well as Hill observed K_d and the Hill coefficient (n). Full length YY1 concentration range: 10 nM – 1.5 µM. **(B)** Fluorescence polarization assays utilizing the ZnFs against two bioinformatically derived ssRNA sequences from Sigova et al. (39) CLIP-seq data. The methodology to identify these sequences is outlined in the Methods section "Computational analyses of YY1 genomic binding sites". Error bars represent standard deviation from three technical replicates and the nonlinear regression fits of both Langmuir (blue) and Hill (green) equations are depicted for each graph. Langmuir observed K_d measurements are reported below each binding curve, as well as Hill observed K_d and the Hill coefficient (n).

2.5 Results

2.5.1 The canonical DNA binding domain of YY1 has a higher affinity for ssRNA than dsDNA

To understand the nucleic acid binding properties of YY1 and reconcile the divergent reports of which regions of the protein are responsible for RNA binding activity (113; 39), we designed, expressed, and purified YY1 fragments, then utilized them in quantitative fluorescence polarization assays with a panel of model nucleic acid binding partners. Although denaturing preparations of the protein have been widely used in the past (30; 112; 39; 120) we were concerned that specific activity variation from the refolding process could compromise the strength of conclusions that could be drawn (including incomplete binding saturation in affinity assays). To avoid these potential complications, we chose native purification conditions for full length and fragments of YY1 (Figure 2.2); only modest activity variation amongst preparations and complete saturation in binding affirmed consistent and high specific activity. Our panel of nucleic acid substrates (Figure 2.5 A) spans the range of strong to weak binding interactions for YY1 reported in the literature (112; 39; 121). For cognate dsDNA reported to have tight binding, we screened the YY1 consensus sequence defined by

SELEX (122), as well as the AAVP5 element which was co-crystallized with the sole known DNA binding domain of YY1 (112; 120), its 4 C-terminal C_2H_2 zinc finger (ZnF module) (Figure 2.1). As a further comparison point, we employed a mutated variant of a 30-bp duplex element from the *Rpl30* promoter, one of the many essential genes regulated by YY1 (37; 123). *Rpl30*'s promoter contains a consensus sequence identical to that defined by SELEX, and it has been previously shown that scrambling the YY1 binding site within this promoter dramatically erodes YY1 affinity (39). As the unmutated *Rpl30* duplex displayed binding affinity slightly tighter than the SELEX consensus sequence for full length YY1 (Figure 2.3 A, Figure 2.4 A), we present only the *Rpl30* scrambled mutant data in Figure 2.5. Both CLIP-seq (39) and CLAP-seq (124) indicate that there is little RNA-sequence motif selectivity in living cells, and prior *in vitro* qualitative studies detected little apparent specificity (113), apart from a bias for U- over C-rich sequences (39; 114). To sample sequence, structure, and a small range of lengths, we assembled a panel of RNA oligonucleotides to probe YY1 RNA-binding affinity. For most YY1 constructs, we relied on the previously described *ARID1A* RNA species as a positive control for RNA binding (39; 125) and as a putative negative control, we chose an RNA devoid of secondary structure and monotonic in sequence, polyuridine (24U), but other RNA oligonucleotides were screened with some protein fragments (Figure 2.3 B, Figure 2.6).

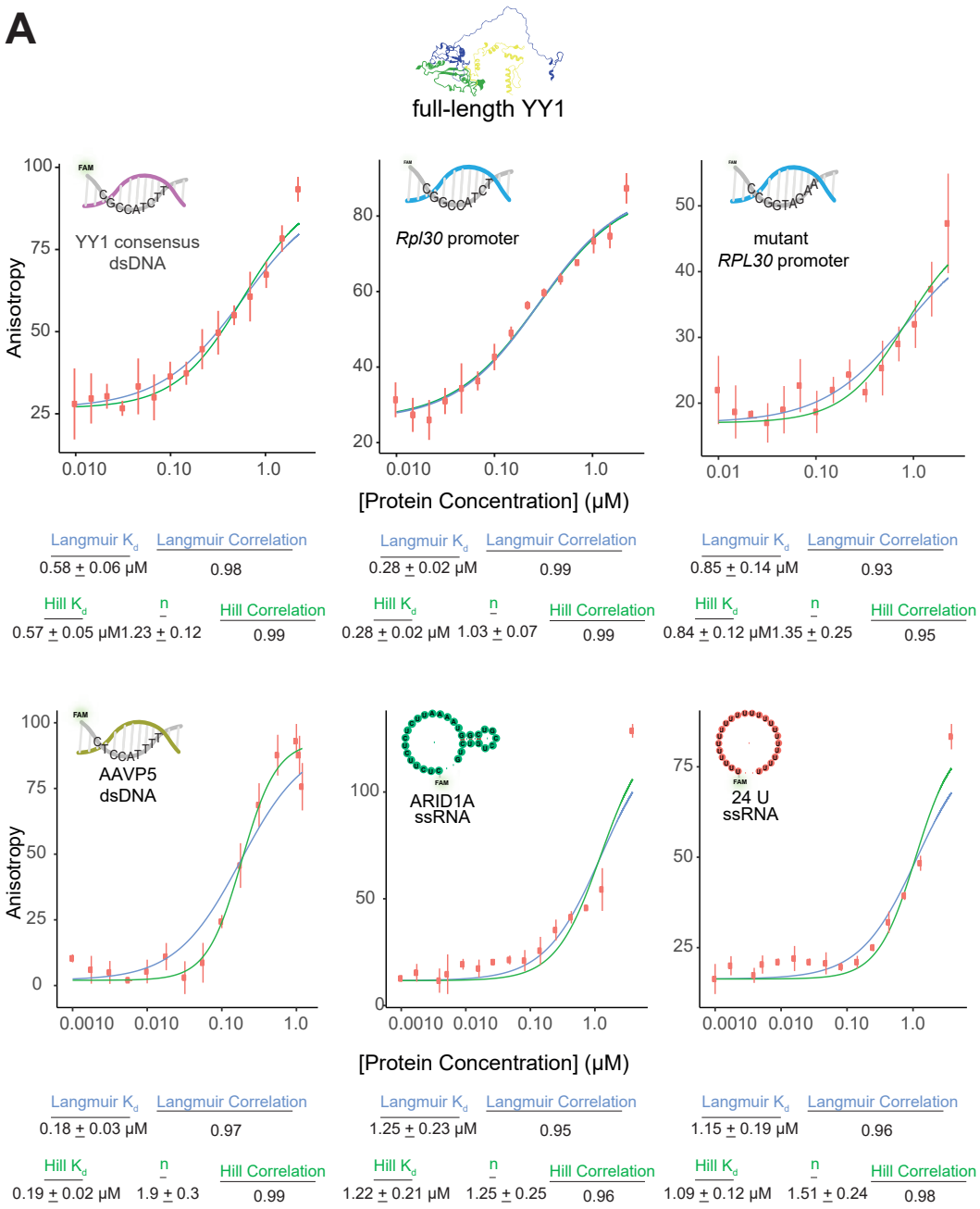


Figure 2.4: Purified full length YY1 exhibits characteristic binding behaviours to our full panel of nucleic acids.

Figure 2.4 *(previous page)*: **(A)** Fluorescence polarization assays against our panel of nucleic acids. Error bars represent standard deviation from three technical replicates and the nonlinear regression fits of both Langmuir (blue) and Hill (green) equations are depicted for each graph. Langmuir observed K_d measurements are reported below each binding curve, as well as Hill observed K_d and the Hill coefficient (n). Concentration ranges are available within the supplementary datasets uploaded to Zenodo.

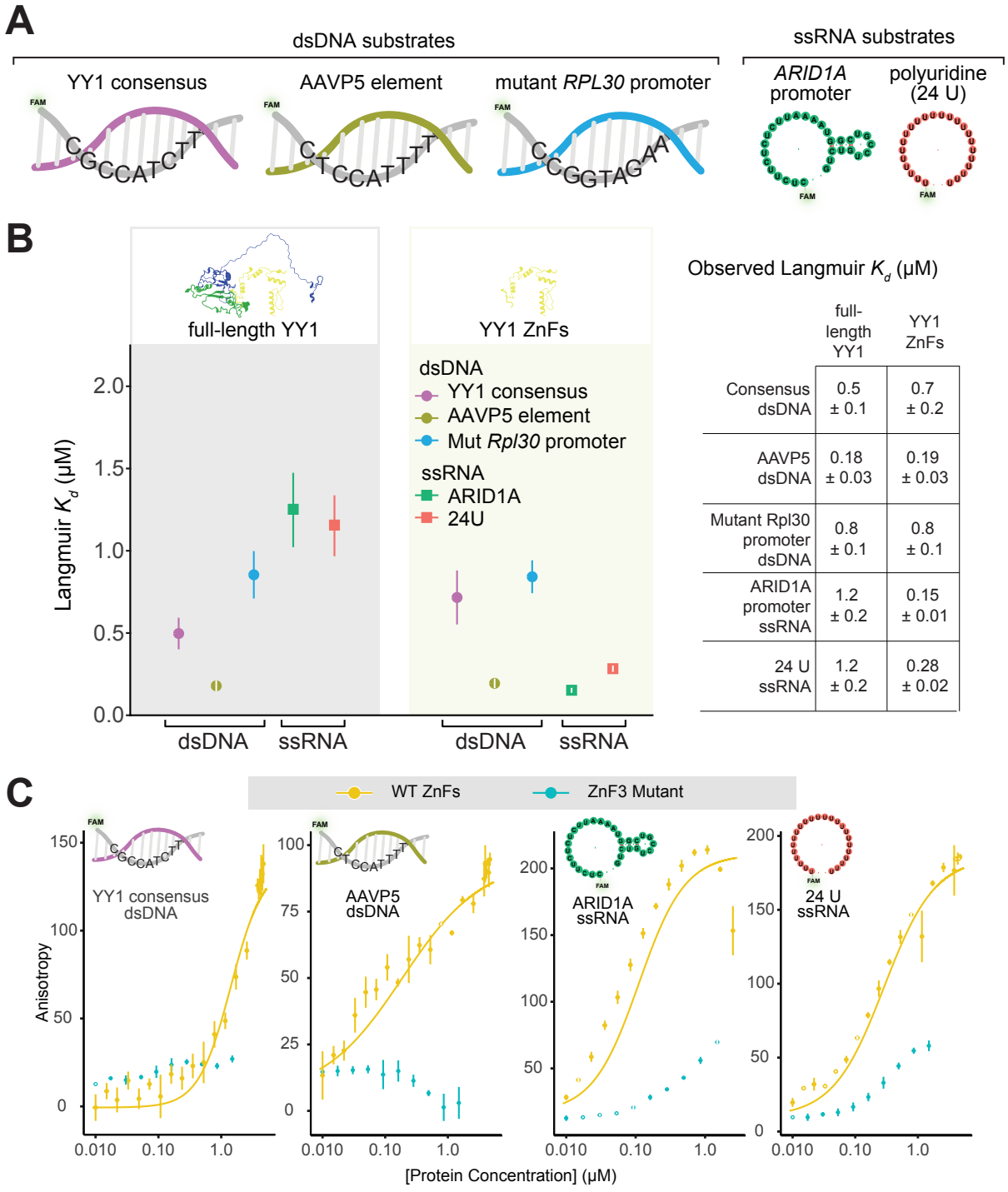


Figure 2.5: DNA and RNA binding of full length YY1 versus its zinc finger module.

Figure 2.5 *(previous page)*: **(A)** 5' FAM labeled nucleic acid panel used in fluorescence polarization assays. ssRNAs are depicted as their predicted RNA structure when input into RNAfold (126) with default parameters. **(B)** Plot of the measured binding affinities (observed Langmuir monovalent binding isotherm K_d) of full length YY1 and its 4 C_2H_2 zinc fingers (ZnFs, AA 293-414) for our panel of nucleic acids. Error reported represent the standard error within the fit. (Right) Table with numeric values for measured K_d values. **(C)** Fluorescence polarization binding curves comparing WT ZnFs to the ZF3 mutant (AA 365-369 to alanine) for our panel of nucleic acids. Error bars represent the standard deviation from 3 technical replicates, for clarity, only the Hill-equation fits are presented.

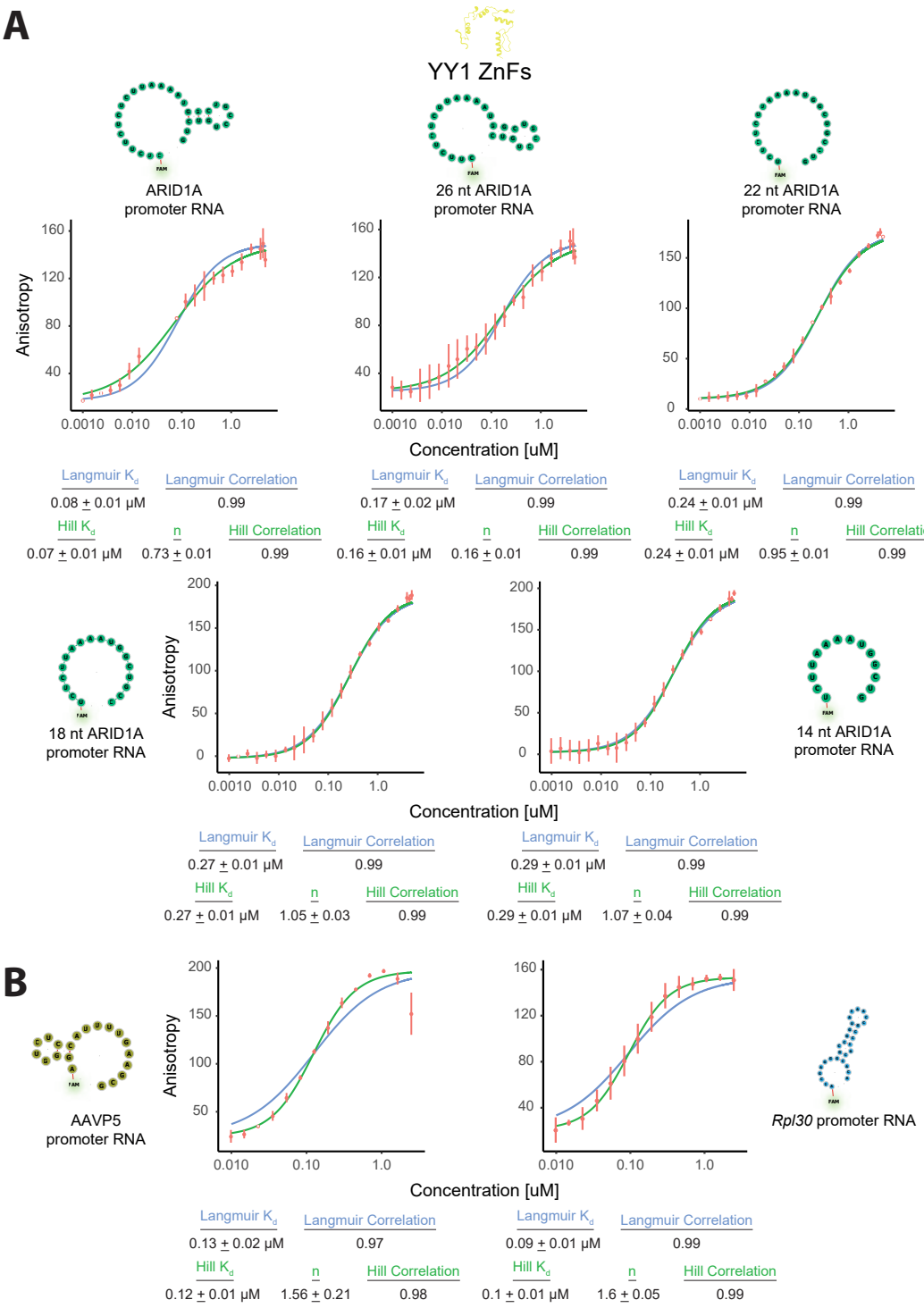


Figure 2.6: YY1's ZnF module displays little sequence, length, or secondary structure specificity in binding ssRNA.

Figure 2.6 *(previous page)*: **(A)** Fluorescence polarization binding curves utilizing YY1's ZnF module and a truncation series of *ARID1A* ssRNAs. The initial *ARID1A* ssRNA (top left) is the 30-nt nucleic acid substrate consistently used within our work, and subsequent truncations remove 2 nucleotides from both the 5' and 3' ends of the previous ssRNA, rendering shorter ssRNA molecules of lengths 26, 22, 18, and 14 nt. Error bars represent standard deviation from three technical replicates and the nonlinear regression fits of both Langmuir (blue) and Hill (green) equations are depicted for each graph. Langmuir observed K_d measurements are reported below each binding curve, as well as Hill observed K_d and the Hill coefficient (n). **(B)** Fluorescence polarization binding curves utilizing YY1's ZnF module and ssRNA species derived from their respective dsDNA substrates i.e. these ssRNAs are the transcriptional products of the AAVP5 dsDNA promoter and the *Rpl30* dsDNA promoter. Error bars represent standard deviation from three technical replicates and the nonlinear regression fits of both Langmuir (blue) and Hill (green) equations are depicted for each graph. Langmuir observed K_d measurements are reported below each binding curve, as well as Hill observed K_d and the Hill coefficient (n).

Purified full-length YY1 demonstrated expected binding affinities for our panel of nucleic acids (Figure 2.5A and B, Figure 2.4), in excellent accordance with previously observed K_d values (39; 120; 127), and overall displayed higher affinity for dsDNA substrates compared to ssRNA substrates. We then purified and assayed the ZnF module of YY1 against our panel of nucleic acids. Although it has been previously reported that the zinc fingers of YY1 can non-specifically bind RNA (113) we were surprised to discover the strength of these interactions exceeded the affinities for dsDNA substrates within our panel by several-fold (Figure 2.5 B). Direct comparisons of ssRNAs with the sequence and length corresponding to the transcribed products of various DNA template strands within our panel, also demonstrated several-fold higher affinity (Figure 2.6 B), excluding sequence- and length- effects as possible explanations for tighter RNA binding. Consistent with previous observations (113; 39) the ZnF module displayed little sequence- or length- specificity (Figure 2.6 A) for RNA, nor any apparent preference for the RNA's capacity to adopt stable structure (Figure 2.5 A, B, 2.3 B, and 2.6). We note that prior measurements did not perform direct comparisons of DNA and RNA binding of the ZnF module to the full-length YY1 protein, underscoring the value of our systematic comparisons to reveal potentially important differences.

The capacity of YY1 to bind both DNA and RNA raises the question of whether both nucleic acids bind a similar interface within the ZnF module. Others have noted competitive binding of RNA and dsDNA with the ZnF module (113) and full length YY1 (114; 115), but the details of this interface on the protein side remain obscure. To further probe whether DNA and RNA bind an overlapping interface, we mutated residues S365-N369 to alanine, targeting residues in ZnF3 that participate in both specific base and nonspecific phosphate backbone interactions with the central CATT motif of the consensus sequence (112)(25), while preserving cysteine and histidine residues requisite for C_2H_2 zinc finger folding. As anticipated, these mutations ablated apparent dsDNA binding (Figure 2.5 C), yet they also dramatically attenuated ssRNA binding (Figure 2.5 C) supporting the hypothesis that both types of nucleic acids may compete for the same interface within the ZnFs. We note that the complete ablation of dsDNA affinity by this pentamutant is distinct from the severe, but still detectable, erosion of RNA affinity, suggesting overlapping, but non-identical interfaces within the zinc finger module for DNA and RNA binding. Consonant with this interpretation, a previous NMR titration of ssRNA suggested that ZnF1-2 account for most of the RNA-binding chemical shift perturbation (113), and while there is some overlap with sidechains involved in DNA binding in these first two fingers (112), there are seemingly distinct surfaces involved in RNA binding within the ZnF module.

Intriguingly, while our DNA binding affinity measurements for the ZnF module are similar to those previously observed (39; 120; 127) (AAVP5 dsDNA K_d measurements range from 0.47 ± 0.05 µM (120) to 0.58 ± 0.04 µM (127), as compared to our value of 0.18 ± 0.03 µM), we find the RNA-binding affinity of full-length YY1 protein is an order of magnitude lower than that of its C-terminal ZnF-module. This suggests the remaining portion of the protein may in some way interfere with this domain's intrinsic RNA-binding properties.

2.5.2 The conserved REPO domain of YY1 binds nucleic acids when isolated from the rest of the N-terminal domain's repression

We sought to further biochemically dissect the regions of the protein N-terminal to the ZnFs in order to delineate binding activity modulation properties and/or other potential nucleic acid binding domains. We approached the design of our first N-terminal protein construct by simply dividing YY1 into two fragments: the ZnF module (Figure 2.5), and the remaining N-terminal portion of the protein (amino acids 1-297) (Figure 2.7 A, N-terminus). Previously Sigova et al. attributed RNA binding to this identical region of YY1 (39), which is consistent with previously noted flexible, nonspecific, electrostatically-driven interactions that IDRs can partake in (128; 129). This hypothesis is further supported by the presence of two RNA-binding arginine rich motifs (ARMs) within the N-terminus; an amino acid motif that has been proposed to be a general feature of TFs that bind RNA to promote their proper genomic localization (74). Yet after purifying and assaying the binding capacity of this N-terminal fragment of YY1, we were surprised to observe that this protein fragment had little apparent affinity for our panel of nucleic acids (Figure 2.7 B).

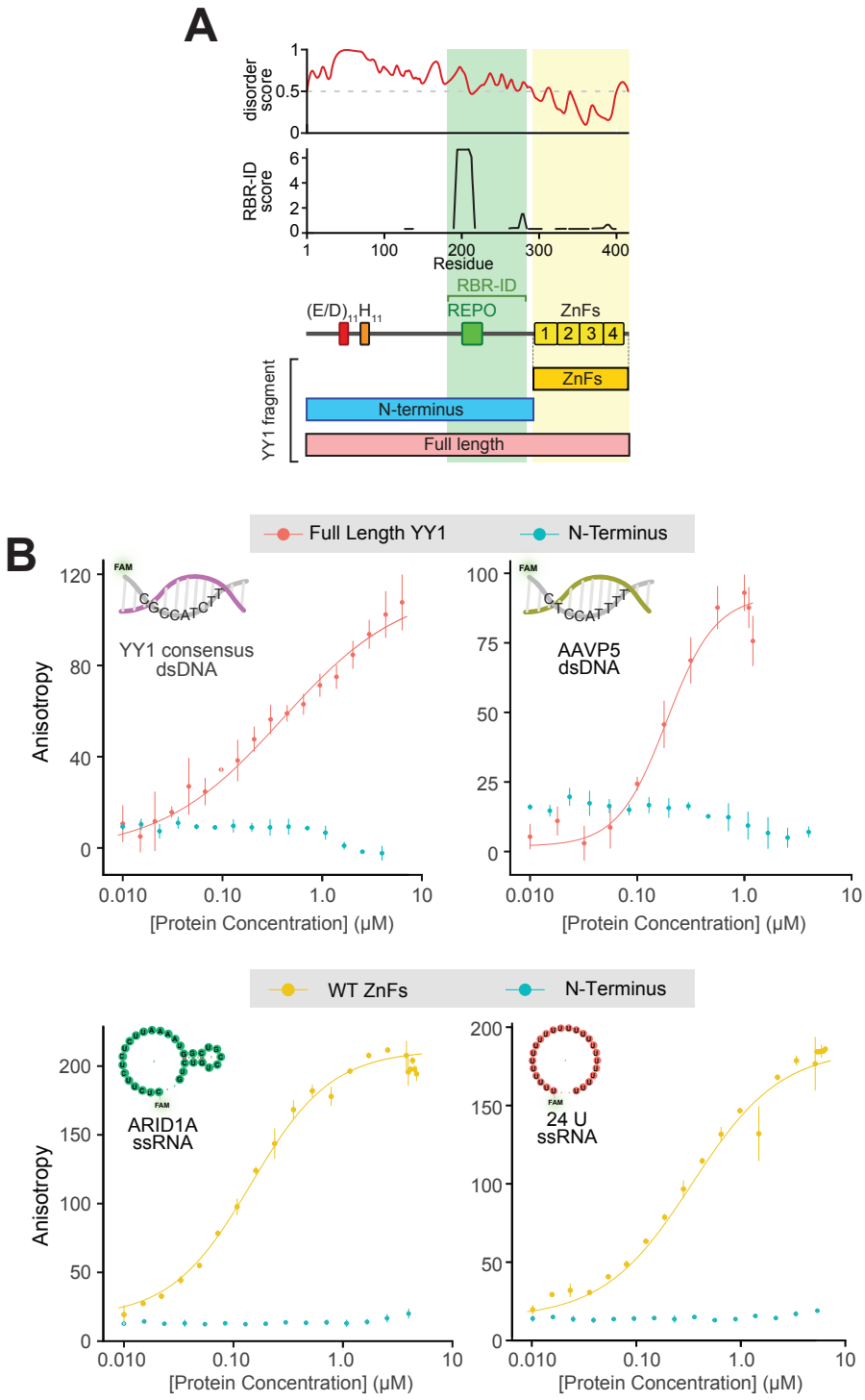


Figure 2.7: YY1's N-terminus (AA 1-297) does not bind nucleic acids.

Figure 2.7 *(previous page)*: **(A)** YY1 domain structure map. (Top) IUPRED3 disorder prediction spanning YY1. (Middle) RNA Binding Region (RBR)-ID score (46) displaying the amino acid residues significantly cross-linked to nuclear RNA within E14 mESCs. (Bottom) Protein constructs that have been purified and assayed in this figure. **(B)** Fluorescence polarization binding experiments for the indicated protein constructs and nucleic acids. Error bars represent standard deviation across three technical replicates.

To reconcile the absence of apparent nucleic acid binding by the N-terminus with the prior study that noted weak binding (39), we sought additional pieces of evidence that could aid the design of further subdivisions of the N-terminus. RBR-ID mass spectrometry data (73) suggests a region of the N-terminus roughly corresponding to the REPO domain of YY1 is significantly crosslinked to nuclear RNA within mouse embryonic stem cells (Figure 2.8 A). The REPO domain has been shown to be both necessary and sufficient for recruitment of Polycomb group proteins to DNA, resulting in transcriptional silencing at loci of recruitment (42) and is highly conserved at the sequence level (82% identity) to the *Drosophila* homolog of YY1, PHO. With this information, we expressed and purified a protein construct with 20 amino acids flanking the two main RBR-ID peaks (Figure 2.8 A, REPO-NAB, AA 164-301) and assessed whether it could bind our panel of nucleic acids. Surprisingly, we observed both DNA and RNA binding capacity, leading us to call this fragment the REPO-NAB for its *n*ucleic *a*cid *b*inding (Figure 2.8 B). REPO-NAB DNA binding was unanticipated, as this region of the protein lacks homology to DNA-binding folds and the only annotated DNA-binding portion of the protein is the C-terminal four C_2H_2 ZnF module.

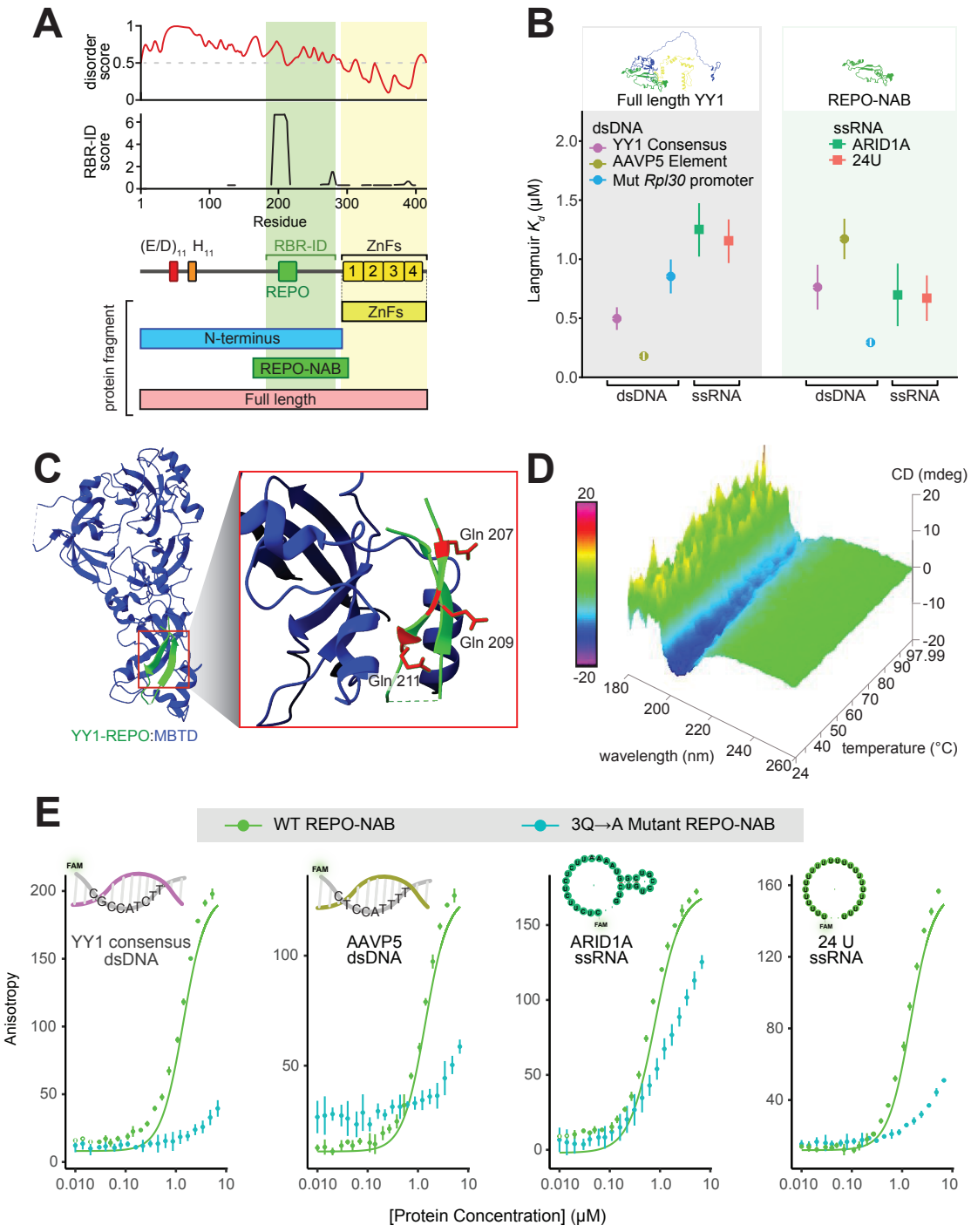


Figure 2.8: The REPO-NAB domain of YY1 maintains secondary structure in isolation and can bind nucleic acids.

Figure 2.8 *(previous page)*: **(A)** YY1 domain structure map displaying the spans of YY1 protein constructs used, including the REPO-NAB (AA 164-301; spanning two RBR-ID peaks with 20 amino acid cushion). (Top) IUPRED3 disorder prediction spanning YY1. (Middle) RNA Binding Region (RBR)-ID score (73) displaying the amino acid residues significantly cross-linked to nuclear RNA within E14 mESCs. (Bottom) Protein constructs that have been purified and assayed in this figure as compared to those presented in Figure 3. **(B)** Langmuir K_d plot displaying the measured binding affinities of full length YY1 and the REPO-NAB fragment for our panel of nucleic acids. Error bars represent the standard error of the fit. **(C)** YY1-REPO:MBTD crystal structure (PDB: 4C5I) (130). (Inset) Zoom in of the YY1:MBTD protein-protein interface with Glutamines 207, 209, and 211 depicted in red. **(D)** Temperature dependent circular dichroism of purified REPO-NAB domain indicative of β-sheet character. **(E)** Fluorescence polarization binding curves comparing WT REPO-NAB domain (red) to the 3Q→A Mutant (Q207A, Q209A, and Q211A) REPO-NAB domain (blue) for the indicated nucleic acid substrates. Error bars represent standard deviation across three technical replicates, for clarity, only the Hill-equation fits are presented, Langmuir fits are available in (Figure 2.9 A).

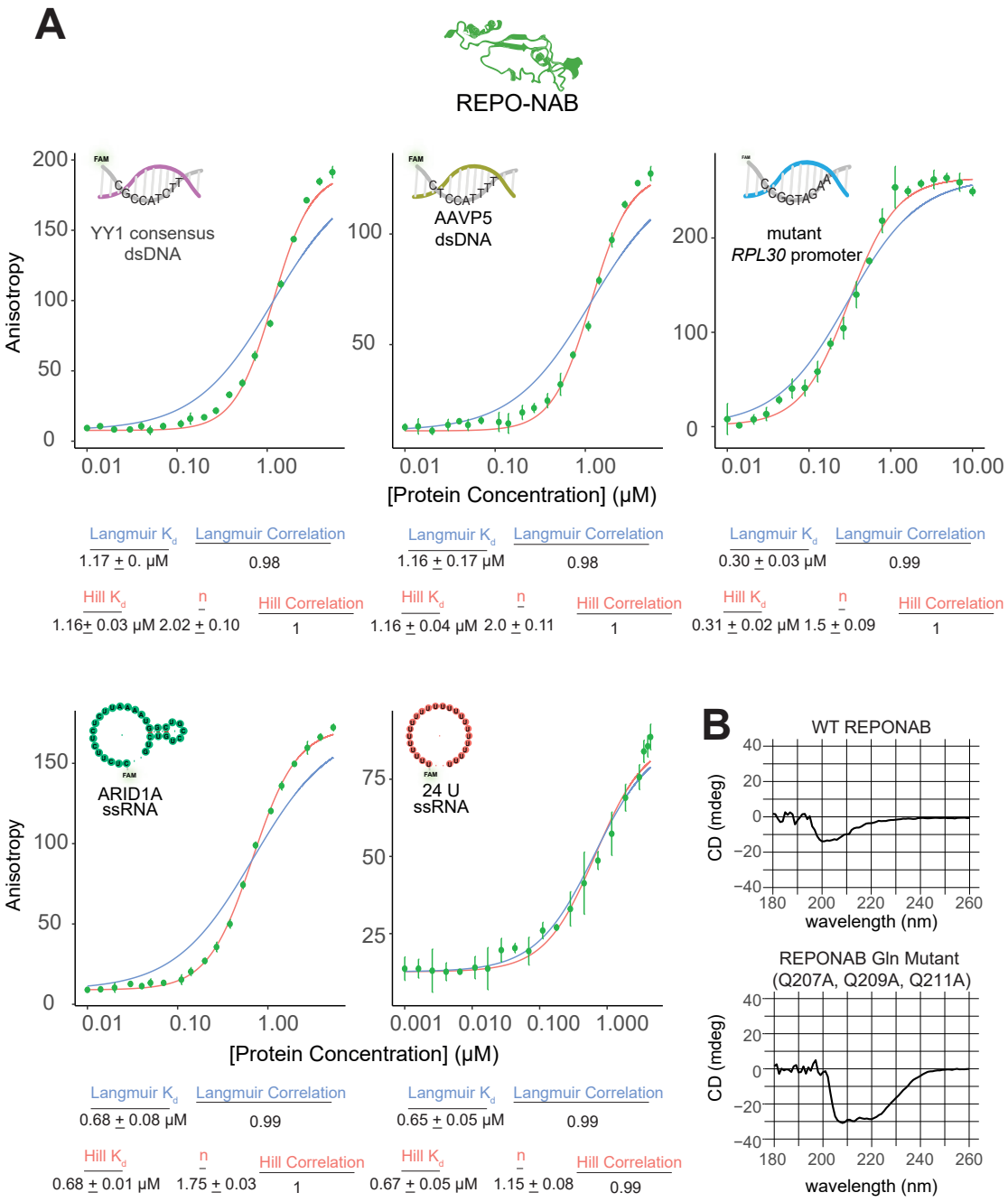


Figure 2.9: The REPO-NAB domain binds nucleic acids and maintains its β-sheet character in isolation.

Figure 2.9 *(previous page)*: **(A)** Fluorescence polarization assays against our panel of nucleic acids. Error bars represent standard deviation from three technical replicates and the nonlinear regression fits of both Langmuir (blue) and Hill (red) equations are depicted for each graph. Langmuir observed K_d measurements are reported below each binding curve, as well as Hill observed K_d and the Hill coefficient (n). Concentration ranges are available within the supplementary datasets uploaded to Zenodo. **(B)** CD spectroscopy of WT REPO-NAB and the 3Q→A mutant demonstrating both adopt similar β-sheet rich folds at room temperature. Note that the concentration of the 3Q→A mutant is somewhat higher.

The specificity properties of the REPO-NAB are intriguing: although there is little apparent specificity amongst our panel of RNA species, the DNA-binding selectivity is quite distinct from the ZnF module and the full protein. While the REPO-NAB and full length YY1 display similar affinity for the SELEX consensus motif, the AAVP5 sequence is bound >5 fold more weakly by the REPO-NAB (Figure 2.8 B). The REPO-NAB construct displays a several fold-higher affinity for the mutated *Rpl30* dsDNA promoter element (0.30 ± 0.03 µM), whereas 0.8 ± 0.1 µM for both the full-length protein and the ZnFs (Figure 2.8 B). Collectively these data suggest that the full-length protein suppresses these apparent DNA binding preferences of the REPO-NAB in addition to RNA-binding inhibition.

Although the REPO-NAB is predicted to be intrinsically disordered by IUPRED (131) (Figure 2.8 A), the apparent selectivity of its DNA-binding suggest some degree of structure. We noted that when generating *de novo* structural models of YY1, RoseTTAFold (132) and AlphaFold3 (133) consistently predicts that the REPO domain of YY1 maintains its anti-parallel β-sheet character within the overall structure of the protein. This is consistent with (and likely informed by) a previous crystal structure of the human YY1 REPO domain in complex with the 4MBT domain of human MBTD1, homologous to the Drosophila PhoRC-polycomb repressive complex (130) (Figure 2.8 C). However, it is unknown whether the REPO domain maintains this anti-parallel β-sheet character in isolation. To address this, we performed circular dichroism with biochemically purified REPO-NAB and observed that the anti-parallel β-sheet character of this domain is maintained in isolation although the

melting transition is not sharply defined (Figure 2.8 D). Similar β-sheet character, previously noted for the full N-terminus (134), is at least in part attributable to the REBO-NAB. We noted that three glutamines (Q207, Q209, and Q211) were not involved in the YY1:MBTD1 interface and therefore could play a role in the REPO domain's nucleic acid binding (Figure 2.8 C), given the propensity for glutamines to engage in nucleic acid-recognition (135; 136; 137; 138). Mutating each of these three residues to alanine, we achieved a partial separation-of-function mutant without impacting the β-character (Figure 2.9 B): dsDNA binding was ablated, whereas RNA binding attenuated within our focused panel of nucleic acids (Figure 2.8 E).

2.5.3 Defining the autoinhibitory function of YY1's N-terminus

A possible explanation for the distinct nucleic acid binding preferences of the REPO-NAB and ZnF module, relative to the full-length protein, is that the N-terminus of YY1 could regulate them. From our binding assays with the N-terminus (AA 1-297, Figure 2.7) and REPO-NAB (AA 164-301, Figure 2.8), we can conclude that when the REPO-NAB is covalently linked to the rest of the N-terminus it becomes inhibited from binding any nucleic acid species within our panel. Similarly, the ZnF module's RNA binding capacity is reduced in the context of the full protein (Figure 2.5 B). To test this hypothesis directly, we performed fluorescence polarization competition experiments assaying the affinity of the ZnFs for the *ARID1A* RNA and the YY1 consensus dsDNA sequence in both the presence and absence of purified N-terminus (Figure 2.10 A and B). For both nucleic acid types we observed decreased binding with the addition of the N-terminus as a function of its concentration. As the N-terminus does not display any detectable nucleic acid binding in the concentration regime (Figure 2.7 B), these observations lead us to believe that the N-terminus can inhibit nucleic acid binding of the ZnFs in *trans i.e.,* not covalently bound to the ZnFs. While Figure 2.10 A is consistent with the apparent *cis*-attenuation of RNA-binding of this protein region in

the context of the full-length protein (compared to ZnF module alone), the *trans* inhibition of DNA binding by the N-terminus is unexpected (Figure 2.10 B), as these properties did not markedly differ between full length YY1 and the ZnF fragment.

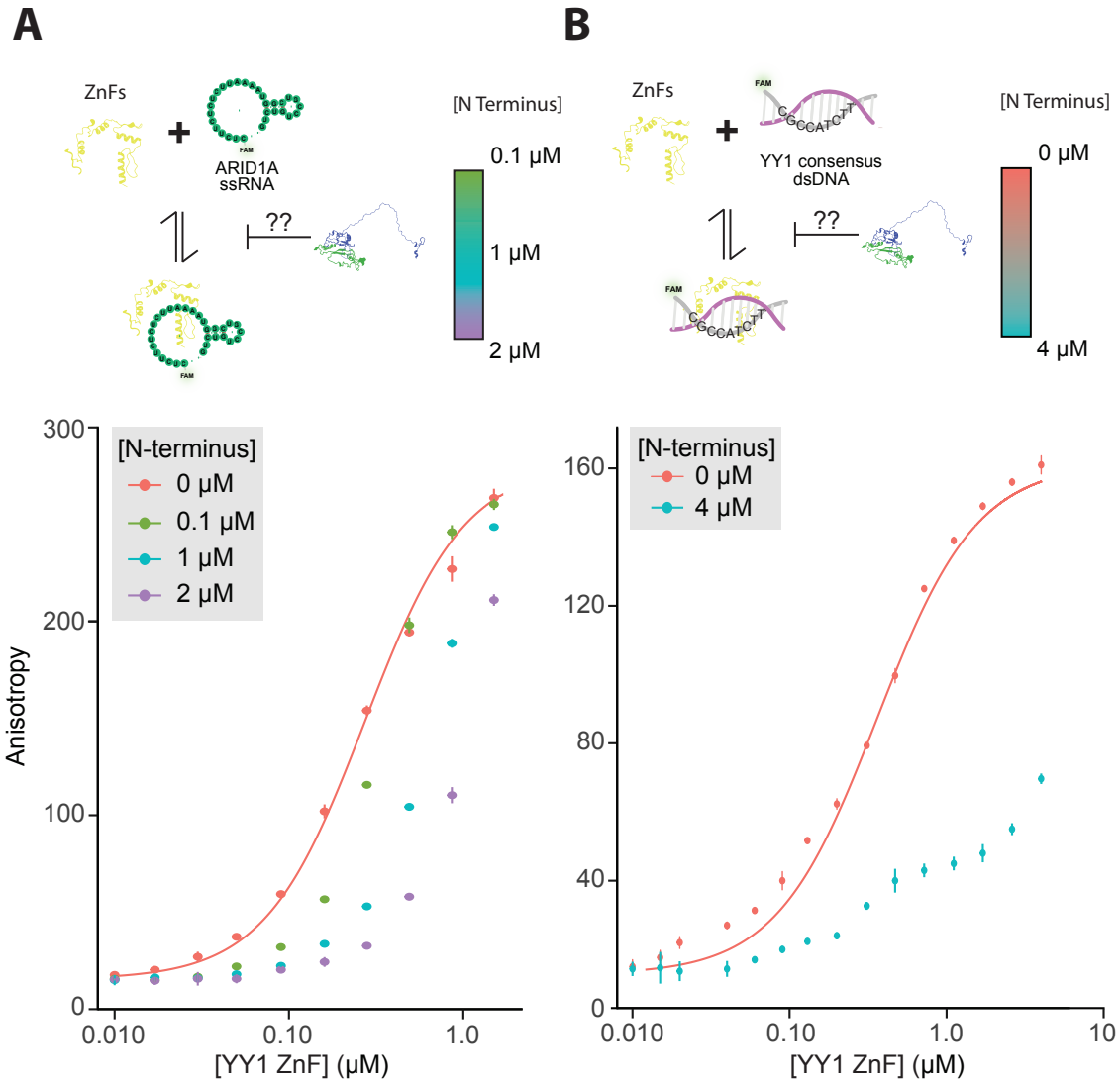


Figure 2.10: YY1's N-terminus modulates nucleic acid binding of its ZnF module.

Figure 2.10 *(previous page)*: **(A** and **B)** Fluorescence polarization competition experiments. The indicated amounts of purified N-terminus were added in *trans* to a ZnFs titration of each of the representative nucleic acid substrates indicated. Error bars represent standard deviation across three technical replicates.

We hypothesized that intramolecular inhibition could occur via charge complementarity between nucleic acid binding domains and a stretch of negatively charged amino acids (E/D) within the N-terminal domain of YY1. A portion of YY1's "core IDR" (cIDR) contains a consecutive stretch of 11 acidic amino acids (AA E43-D53) that are fully conserved across metazoans and have recently been shown to be a component of the cIDR which is both necessary and sufficient to drive phase separation of YY1 in both *in vitro* and cell-based assays (139). This stretch of negative amino acids has also been hypothesized to aid in YY1's genomic target search by autoinhibition of spurious nucleic acid binding (140). Upon generating an N-terminus construct with these 11 residues (E43-D53) mutated to alanine (Figure 2.11A), we observed that this mutant, like the wild type N-terminus could not bind our panel of nucleic acids (Figure 2.11 B). Although the REPO-NAB domain is not sensitive to regulation by the E/D patch in *cis*, we sought to perform similar *trans* complementation experiments with the ZnF module and the N-terminus bearing the E/D patch mutation. In this context as well, there was also little difference between the N-terminus and the mutant—both similarly impaired RNA binding of the zinc fingers in *trans* (Figure 2.11 C). Collectively these experiments suggest that rather than localized to a single cluster of the 11 E/D residues that we targeted, the nature of inhibition may be more complex. Charge complementarity-based regulation, if it indeed exists, must be dispersed across the IDR. This E/D cluster accounts for only 11 of the 38 acidic residues in this region of the protein and similar delocalized properties have been noted in other cases (97; 98; 101).

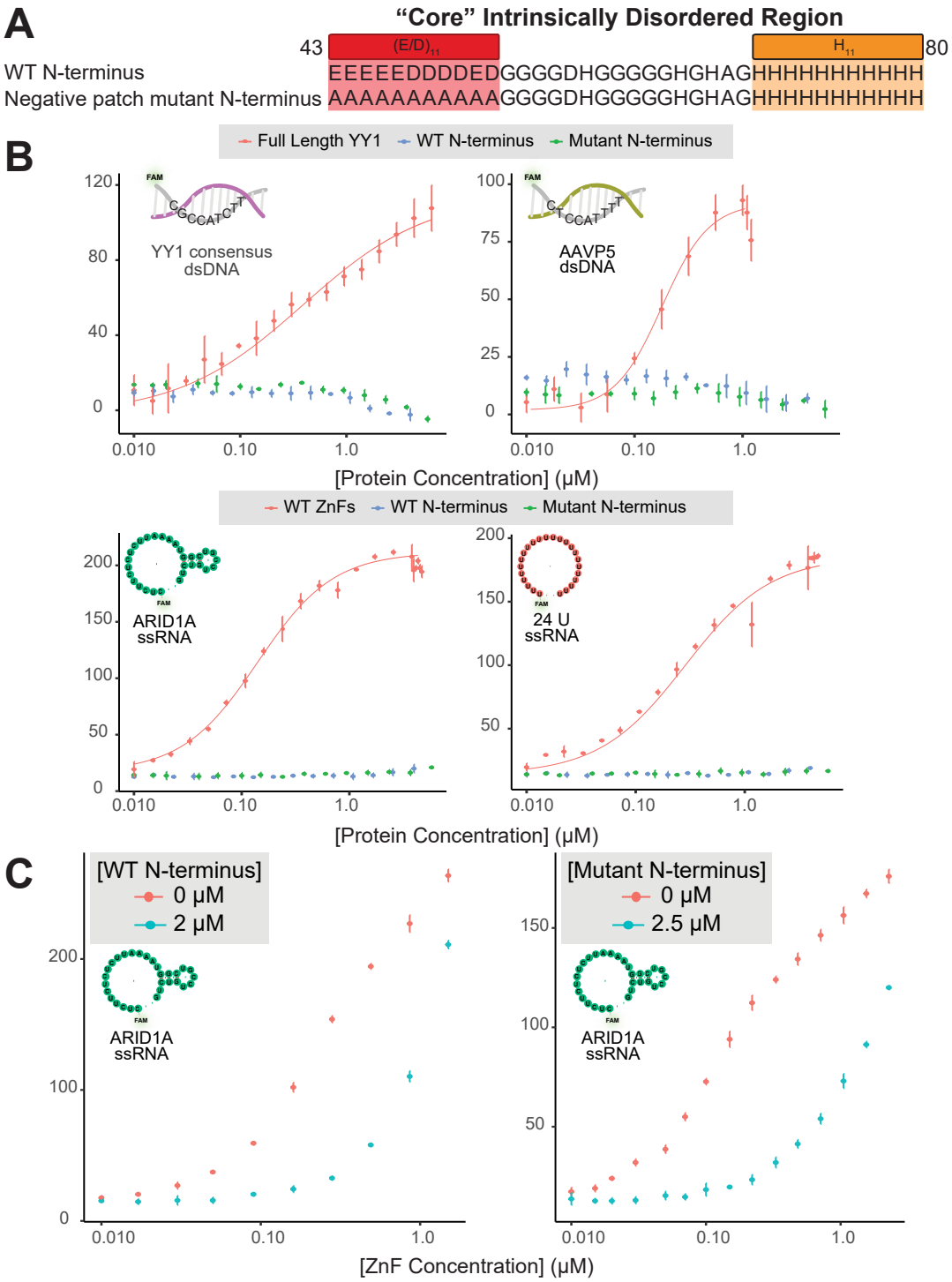


Figure 2.11: The acidic stretch (AA 43-53) of YY1's N-terminus is dispensable for N-terminus mediated autoregulation of ZnF nucleic acid binding.

Figure 2.11 *(previous page)*: **(A)** Single letter amino acid representation of both the wildtype (WT) and negative patch mutant of the "core Intrinsically Disordered Region (IDR)" (56) in YY1's N-terminus. **(B)** Fluorescence polarization binding experiments for the indicated protein constructs and nucleic acids. Error bars represent standard deviation across three technical replicates. **(C)** Fluorescence polarization competition experiments. (Left) Competition experiment from figure 5A, highlighting the highest concentration of WT N-terminus inhibiting ZnF nucleic acid binding. (Right) The indicated amounts of purified negative stretch mutant N-terminus were added in *trans* to purified ZnFs and binding to the representative nucleic acid substrates was measured. Error bars represent standard deviation across three technical replicates.

To further investigate this inhibition, we sought to purify a truncated N-terminus lacking the REPO-NAB domain (Figure 2.12A, AA 1-163) and assess its inhibitory properties. This fragment displayed several orders of magnitude weaker binding to nucleic acids (Figure 2.12 B)— although it was not as binding deficient as the full N-terminus construct, we were unable to measure dissociation constants for any of the nucleic acid species within our panel. Since we observed close to no change in anisotropy for the disordered N-terminus in the concentration range in which the ZnF module reaches saturation with our panel of nucleic acids (0 – 2 µM), we decided to stage competition experiments in order to assess whether the disordered N-terminus could regulate nucleic acid binding. We observed little to no inhibition of binding of either nucleic acid type when we added the disordered N-terminus *in trans* to the ZnFs at concentrations below (Figure 2.12 C,D), but approaching regimes where this module itself could directly bind (Figure 2.12 B). Considering the discrepancy between the affinities the ZnF module and full length YY1 exhibit for RNA binding, these observations suggests that direct linkage of the REPO-NAB is necessary to orient the N-terminus to inhibit RNA-binding of the ZnF module.

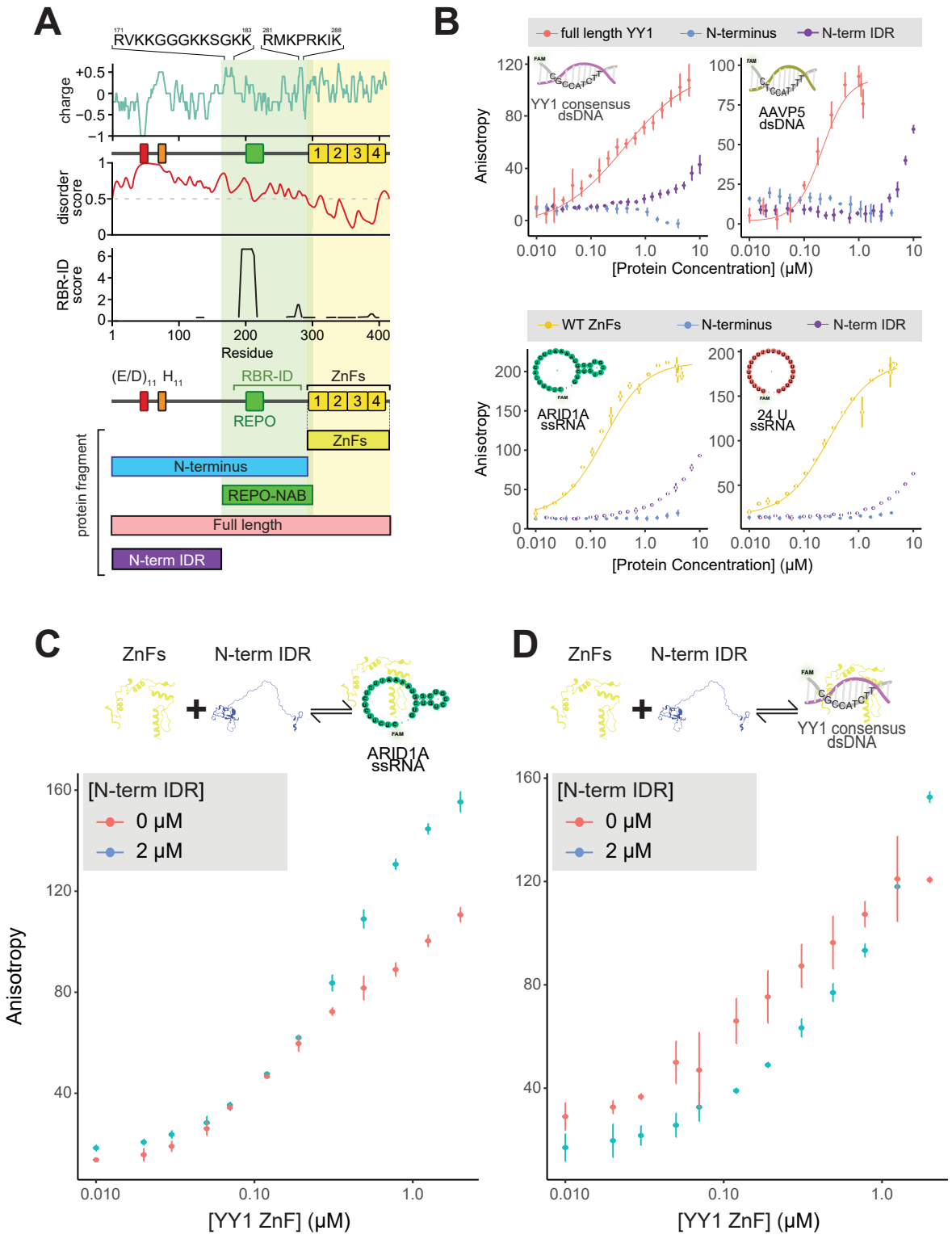


Figure 2.12: YY1's N-terminal IDR (N-terminus ΔREPO) has little nucleic acid binding capacity and does not inhibit ZnF nucleic acid binding.

Figure 2.12 *(previous page)*: **(A)** (Top) YY1 domain map. EMBOSS net charge for 5 amino acid sliding window, with the only two regions highlighted that are basic patches not in the ZnF module as preciously defined (5 or more consecutive resides that have K/R represented at a frequency >0.5, also termed "ARM-like" motifs (74). (Upper Middle) IUPRED3 disorder prediction spanning YY1. (Lower Middle) RNA Binding Region (RBR)-ID score (73) displaying the amino acid residues significantly cross-linked to nuclear RNA within E14 mESCs. (Bottom) Domain structure and pertinent protein constructs that have been purified and assayed. **(B)** Fluorescence polarization binding experiments for the indicated protein constructs and nucleic acids. Error bars represent standard deviation across three technical replicates. **(C)** and **(D)** Fluorescence polarization competition experiments. The indicated amounts of purified ΔREPO N-terminus were added in *trans* to purified ZnFs and binding to the representative nucleic acid substrates was measured. Error bars represent standard deviation across three technical replicates.

2.6 Discussion

2.6.1 Summary

In this work, we systematically interrogate the nucleic acid-binding domain structure, and intrinsic regulation thereof, by other portions of YY1, revealing a surprisingly complex architecture of autoregulatory logic in the absence of additional co-factors. First, we address the controversy regarding which portions of the YY1 protein are responsible for RNA-binding via systematic side-by-side equilibrium binding studies, whose quantitative nature proved essential to our proposed resolution. Semi-quantitative measurements have been used to suggest that the N-terminus of the protein, but not the C-terminal ZnF module, bears weak RNA-binding capacity relative to the full-length protein (39). Whereas more quantitative examinations of the C-terminal ZnFs in isolation, absent data presented for the full protein (113) or the N-terminal portion (113), argue that this module harbours RNA binding capacity thought to be important for function, without clear exclusion of the N-terminus contributions. Consonant with the latter two papers, we find that indeed the C-terminal ZnF module has both DNA- and RNA- binding capacity, but to our surprise, the RNA binding affinity of the ZnF module in isolation exhibits an order of magnitude tighter binding

than what we observed with the full-length protein, under identical conditions, whereas the dsDNA affinities are quite similar for both (Figure 2.5 B). We detect no RNA affinity for an identical N-terminal fragment of the protein at a concentration regime more than an order of magnitude higher than where Sigova and colleagues detected weak N-terminal binding (39) (Figure 2.7 B). Remarkably, dissection of the N-terminal domain into the N-terminal IDR (AA 1-163) and the REPO-NAB (AA 164-301) reveals the latter's marked affinity for RNA (as well as DNA) (Figure 2.8 B) is seemingly suppressed by the former (Figure 2.7 B). It is possible that in the prior work (39), proteolytic fragments bearing the REPO-NAB separate from the N-terminal IDR are present in the preparation of the N-terminal fragment which could account for apparent binding by EMSA and is consistent with the complex pattern of shifted bands observed. Our data suggests both seemingly divergent prior RNA-binding attributions to be true in the sense that there are regions in both fragments that have RNA-binding capacity, and in so doing, we reveal a potent RNA-binding inhibitory property of the N-terminal IDR. This inhibitory property of the N-terminus is not limited to the REPO-NAB– we have discovered a similar regulatory mechanism the N-terminus imposes upon the ZnF module as well. Although our study is limited to the scope of nucleic acid species surveyed and by the *in vitro* nature of our quantitative assays, these observations of the protein's intrinsic properties may provide insight into the highly context-dependent nature of YY1's activities *in vivo*.

2.6.2 The interfaces of RNA and DNA binding overlap in both nucleic acid binding domains.

With recent work shedding light on the capacity of TFs to bind both duplex DNA and RNA (74; 78; 76; 77), inquiry into the mechanisms and functional consequences of these properties are of great interest. We and others (37; 113; 39; 74; 114), have investigated whether dsDNA and ssRNA compete for a similar binding interface within full length YY1. In the most

mechanistically detailed prior study, Wai and colleagues (113) used ^{15}N-heteronuclear single-quantum-coherence (^{15}N-HSQC) NMR experiments to identify amino acid residues in the ZnF module that engaged in RNA binding, finding the bulk of RNA interactions localized to ZnF1 and ZnF2 (113). Although a number of residues that displayed significant chemical shift perturbations in the presence of RNA are involved in dsDNA contacts in the YY1 ZnF module crystal structure (112), the authors noted a contiguous surface of amino acids within ZnF2 (V324, H325, V326, L340, Q344) that was completely distinct from the dsDNA binding interface (113). Mutagenesis of these amino acids to alanine diminished apparent binding to RNA by at least an order of magnitude (113), consistent with this unique interface representing an energetically consequential portion of the RNA-binding capacity. However, retention of DNA-binding affinity of this pentamutant was not evaluated, so it remains uncertain whether it represents a *bona fide* separation of function mutant.

In an attempt to design the converse separation of function mutation, wherein DNA-binding affinity of the ZnF module was selectively perturbed with minimal RNA-impact, we targeted amino acids in a different segment of the ZnF module (AA 365-369, ZnF3, Figure 2.5 C). This mutant indeed displays no measurable DNA binding activity and its RNA-binding was severely attenuated, but still detectable (Figure 2.5 C). This impact on RNA-binding was unexpected, as these residues, and ZnF3, overall exhibit minimal chemical shift perturbations in the presence of RNA (113). One possible explanation for this apparent disparity in the importance of the ZnF3 for RNA binding is the identity of the RNA in our experiments is distinct—the K_d measurements for ssRNA in this prior work range from 20-60 fold weaker than our measurements. The Wai paper used a 14-mer of RNA isolated from a SELEX experiment for their affinity and structural studies ($K_{\mathrm{d}} = 3.8 \pm 0.6$ µM, although this was within error of a poly-A substrate of the same length in their experiments). Whereas the majority of our experiments were conducted with longer RNA species, we note serial truncation of the ARID1A 30-mer down to a 14-mer element leads to a $\tilde{3}$-fold affinity loss

monotonically across the series (Figure 2.6 A), and this $K_d = 0.29 \pm 0.1$ µM is comparable to two 12-mers of completely different sequence composition (Figure 2.3 B). Thus, size and sequence do not account for our order of magnitude higher affinity measurements for the ZnF module binding RNA in this regime. The methods of protein preparation are highly analogous, the binding buffers are only subtly different, and although the measurement methods differ, both microscale thermophoresis and fluorescence polarization are solution-based measurements free of confounding surface effects. It may be that for shorter ssRNA, the interface of the YY1 ZnF module species less extensively engages ZnF3 and ZnF4. This is consistent with the origin of the sequence used in NMR structural studies, which was selected for using a Zn-coordinating mutant in ZnF4 that should unfold it (113). Nevertheless, our data with longer RNA species and the ZnF3 mutant clearly indicates that this region of the protein can also be important for RNA binding.

We have identified a similar scenario in which DNA and RNA share an overlapping binding interface in the elucidation of the REPO-NAB's nucleic acid binding properties. Our mutant REPO-NAB, harbouring alanine substitutions at Q207, Q209, and Q211, loses its capacity to bind dsDNA substrates (Figure 2.8 E) but these mutations attenuate, yet do not completely diminish, the ability of this mutant to bind our ssRNA substrates. Recent work implicates ARM-motifs, basic residue rich patches directly adjacent to canonical DNA binding domains of several transcription factors, in promoting their chromatin association through RNA binding (74), a mechanism first proposed for YY1 (37). Although the REPO-NAB does contain a basic patch directly flanking the ZnF module (AA 281-288, Figure 2.12 A), this region is not able to bind nucleic acid in the context of the full N-terminus. While it is possible that this motif does participate in RNA binding as proposed for other ZnF TFs (74), our mutagenesis of the distinct REPO domain within this construct (AA 201-226) suggests that the ARM motif is not playing a dominant role in the nucleic acid binding we observe. Further *in vivo* experiments elucidating how the binding of one nucleic acid

type (in this case ssRNA) and the loss of binding for another (dsDNA in this case) can affect the context-dependent localization and function of full length YY1 can be undertaken with separation-of-function mutants of the sort we report here. The identification of the REPO-NAB's nucleic acid binding properties also affords another example to the growing list of cryptic/ancillary nucleic acid binding domains (75; 78; 76). Understanding what this additional binding capacity imparts upon their respective proteins in cellular contexts is the next key step to define their functional importance.

2.6.3 YY1's N-terminus tunes the nucleic acid binding affinity of the REPONAB and the ZnF module

Here we begin to dissect the autoinhibitory properties of the N-terminus of YY1, in both intra- and inter-molecular mechanisms (*cis* versus *trans*) on the two nucleic acid binding domains of YY1. Our measurements reveal how the N-terminus, either covalently attached in single polypeptide or added in *trans*, affects nucleic acid binding of the REPO-NAB and ZnF modules. We find that when the REPO-NAB is physically linked to the remainder of the N-terminus of YY1 (the N-terminal IDR), it is efficiently inhibited from binding both ssRNA and dsDNA (Figure 2.7). Similarly, in competition experiments assaying the ZnF module's capacity to engage nucleic acids in the presence/absence of the N-terminus we observe a drastic inhibition of either type of nucleic acid binding. The impact on DNA binding is unexpected as we envisioned that reconstituting the inhibitory interaction in *trans* would begin to recapitulate the properties of full length YY1—blunting of the intrinsic capacity of the ZnF module's capacity to bind RNA efficiently, while preserving the DNA-binding capacity. However, adding these two fragments in *trans* may fail to capture some aspect of orientation provided by direct linkage in the native full protein. The differences of the N-terminus added in *trans* versus *cis* could be interpreted as revealing the potential for regulation by YY1's reported capacity to dimerize/multimerize (107; 37; 109; 141) in a po-

tentially RNA-dependent manner (107; 37; 141). One possible function of these mechanisms is to keep YY1 in an inhibited state, awaiting further protein/nucleic acid interactions to ultimately impart context-dependent functionality. Further experiments will be needed to precisely map the interfaces between the REPO-NAB, the ZnF module and the N-terminus that impart this nucleic acid binding inhibition within YY1, and to determine how generalizable these regulatory features are to other transcription factors, that often have IDRs (97; 98; 100; 101) and poorly understood additional RNA-binding regions (74).

2.6.4 How the two newly identified YY1 activities may relate to its myriad functions

Given the wide spectrum of functions ascribed to YY1 and the large collection of well-established protein binding partners in each of these scenarios (37; 108; 109; 42; 38; 110), we propose that the intrinsic YY1 nucleic acid binding and autoregulation thereof could be key components of these diverse regulatory outcomes. The acidic transactivation domain of YY1 has been mapped in cell-based reporter assays to the N-terminus, with the precise boundaries of this region differing somewhat but generally including amino acids 1-99 (108). This constitutes the bulk of the N-terminal IDR construct (Figure 2.12 A) which inhibits nucleic acid binding of the REPO-NAB and is necessary but not sufficient for ZnF module RNA-binding suppression in *cis*. Thus, the act of participating in transcriptional activation interactions (10), could relieve nucleic-acid binding inhibition of the REPO-NAB and enable RNA-binding of the ZnFs. Similarly, a core region of the N-terminal IDR spanning amino acids 43-80 has recently been noted to drive phase separation *in vitro* with corresponding impact on cellular YY1 function (139). These properties require 11 consecutive histidine residues within this core IDR which have also been proposed to mediate the homodimerization/multimerization of YY1 via zinc coordination (119). Interestingly, RNA has been proposed to facilitate the YY1 dimerization (37; 141) thought to be essential for through

space architectural linkage of two YY1 DNA binding sites (37). Given our data that suggests this YY1 region is involved in suppressing RNA-binding of either of the two nucleic acid binding domains, we postulate that the YY1 protein-protein interface can either act to dimerize the protein or inhibit RNA-binding in *cis*, and that the addition of RNA competitively releases this N-terminal inhibitory region to engage in chromosome looping *trans*-YY1 interactions. This is consistent with a prior observation that Rbm25 stabilizes YY1 binding to chromatin via protein-protein interactions, and some of this stabilization could be accounted for by YY1's RNA binding (142).

More broadly, the putative interplay between protein-partner binding and RNA/DNA binding could account for the remarkable diversity of YY1 cellular functions. Specific and nonspecific nucleic acid binding by the domains that we have interrogated could play a role in YY1's genomic localization, supporting the observations of pervasive transcription factor "trapping" posited by others (39; 74) and implicating IDRs in proper TF binding site engagement (97; 100; 101; 139; 140). Our work has demonstrated that rigorous, systematic, biophysical approaches can uncover unannotated properties of very well-characterized proteins and can therefore guide further investigation to their function. We contend that defining the intrinsic properties of the YY1 polypeptide with respect to nucleic acid binding and its autoinhibition, represents a critical advance to elucidating YY1's precise molecular mechanisms and their functional impact in transcription, repression and genome architecture. In future studies, we hope to characterize the interplay of these features with the catalogue of known YY1 protein binding partners, as well as determine the functional consequences of these newly defined properties *in vivo*.

CHAPTER 3

RETHINKING THE ROLE OF NUCLEOSOMAL BIVALENCY IN EARLY DIFFERENTIATION

3.1 Attributions

This chapter has been adapted from: Shah, R. N. *et al.* Re-evaluating the role of nucleosomal bivalency in early development. Preprint at *bioRxiv*, doi: 10.1101/2021.09.09.458948. (2021). Asymmetric disulfide-linked H3K4me3-H3K27me3 were synthesized and provided by the Fierz Laboratory at École polytechnique fédérale de Lausanne, Switzerland. The 304M3B-1xHRV3C antibody was developed by the Koide Lab at New York University with Adrian Grzybowski, PhD'18. Dr. Grzybowski also developed the reICeChIP method, conducted reICeChIP-seq on naïve mouse embryonic stem cells, and conducted methyltransferase assays. Dr. Shah and Dr. Ruthenburg conducted reICeChIP for the different cell populations. Dr. Shah also performed the computational analyses necessary to calculate Histone Modification Densities (HMDs), generate metaprofile plots, alluvial plots, Receiver Operator Curves (ROCs), and draw comparisons between our datasets and previously published work (91; 143; 144). Growth of all cell populations (naïve mESCs, "primed mESCs", and NPCs) necessary for analysis were conducted by the author, as well as genome browser views displayed in Figures 1C and 2B .

3.2 Abstract

Nucleosomes, composed of DNA and histone proteins, represent the fundamental repeating unit of the eukaryotic genome (79); posttranslational modifications of these histone proteins influence the activity of the associated genomic regions to regulate cell identity (145; 146; 147). Traditionally, trimethylation of histone 3-lysine 4 (H3K4me3) is associated

with transcriptional initiation (84; 85; 86; 87; 88; 89), whereas trimethylation of H3K27 (H3K27me3) is considered transcriptionally repressive (148; 149; 150; 151; 152). The apparent juxtaposition of these opposing marks, termed “bivalent domains" (92; 153; 91), was proposed to specifically demarcate of small set transcriptionally-poised lineage-commitment genes that resolve to one constituent modification through differentiation, thereby determining transcriptional status (154; 155; 156; 90). Since then, many thousands of studies have canonized the bivalency model as a chromatin hallmark of development in many cell types. However, these conclusions are largely based on chromatin immunoprecipitations (ChIP) with significant methodological problems hampering their interpretation. Absent direct quantitative measurements, it has been difficult to evaluate the strength of the bivalency model. Here, we present reICeChIP, a calibrated sequential ChIP method to quantitatively measure H3K4me3/H3K27me3 bivalency genome-wide, addressing the limitations of prior measurements. With reICeChIP, we profile bivalency through the differentiation paradigm that first established this model (92; 91): from naïve mouse embryonic stem cells (mESCs) into neuronal progenitor cells (NPCs). Our results cast doubt on every aspect of the bivalency model; in this context, we find that bivalency is widespread, does not resolve with differentiation, and is neither sensitive nor specific for identifying poised developmental genes or gene expression status more broadly. Our findings caution against interpreting bivalent domains as specific markers of developmentally poised genes.

3.3 Introduction

In its original conception, the bivalency model posits that the combination of H3K4me3 and H3K27me3 represents a specific regulatory marker of developmentally staged genes. Specifically, lineage commitment genes are thought to be held in a low-expression, transcriptionally “poised” state by promoter nucleosomes bearing both H3K4me3 and H3K27me3 (92; 91; 156; 90). Upon differentiation, the bivalent domain “resolves” into a monova-

lent state, and the associated gene is either transcriptionally activated or terminally repressed if H3K27me3 or H3K4me3 is lost, respectively(92; 91; 156; 90). The elegance of this instructive model inspired a host of follow-on studies that have suggested that bivalency is important in differentiation (157; 158; 159; 160; 161; 162; 163; 164), embryogenesis (153; 165; 166; 167; 168), genome architecture (90; 143; 169; 170; 171), and oncogenesis (172; 173; 174; 175; 176).

In the absence of unambiguous biochemical or functional validation (153; 177; 178), these studies have largely relied upon ChIP, with the vast majority of studies defining loci with independent ChIP enrichment for H3K4me3 and H3K27me3 as bivalent domains. However, this analysis cannot distinguish whether the two modifications coexist or represent two distinctly marked subpopulations of alleles or cells. Further, because different ChIPs are normalized separately, they exist on separate relative scales and cannot be quantitatively compared without internal calibration (96; 95; 179). As such, it is impossible to quantify the extent of bivalency at a given locus or to measure its changes through differentiation.

To address the first problem, several studies have used sequential ChIP (92; 163; 180; 181; 182), measuring coexistence by using the eluent of an IP against H3K4me3 as the substrate for an IP against H3K27me3 (or vice versa). However, these experiments used antibodies of unknown specificity (92; 162; 180; 181; 182), were uncalibrated, and were often undersampled (181; 183), precluding quantification of the extent of modification. Moreover, many used relatively large chromatin fragments in their pulldowns, making it difficult to determine whether modifications coexisted on one nucleosome or discretely marked neighbouring nucleosomes (92; 163; 181; 182). The limitations of these sequential ChIP studies preclude accurate assessment of key properties of bivalency.

Our previous work introduced internally calibrated ChIP (ICeChIP), in which barcoded nucleosome internal standards are used to measure antibody specificity and as analytical calibrants that enable computation of the histone modification density (HMD), or the pro-

portion of nucleosomes at a given locus with the modification of interest (96; 95; 179). By identifying regions with high H3K4me3 and H3K27me3, we indirectly identified many promoters with a nonzero amount of bivalency, including those regulating developmental and metabolic genes (95). However, this analysis was limited; it was not sensitive for bivalency at less extensively modified loci, nor could it quantify the extent of bivalency. Here, we directly quantify this nucleosomal mark pattern by calibration of a modified sequential ChIP approach to critically evaluate the bivalency model in the differentiation system in which the foundational observations were made.

3.4 Measuring bivalency with reICeChIP

To directly measure bivalency and evaluate its role in differentiation, we first attempted to deploy our calibrants with published sequential ChIP methods. However, when evaluated with internal standards, these methods (92; 163; 180) displayed extremely low enrichment and variable specificity (Figure 3.1a), with common elution methods either failing to release most of the captured material (182) or compromising the specificity of the second IP (Figure 3.1b-c). With such heavy losses, we became concerned that we would undersample and potentially bias the measurement of bivalent nucleosomes. We sought a method of elution from the primary IP that was both more efficient and would preserve nucleosome integrity for the second IP. To that end, we modified a recombinant biotinylated Fab (304M3-B) specific for H3K4me3 (184) with an intervening HRV 3C endoprotease cleavage site to enable quantitative elution by enzymatic cleavage under mild conditions.

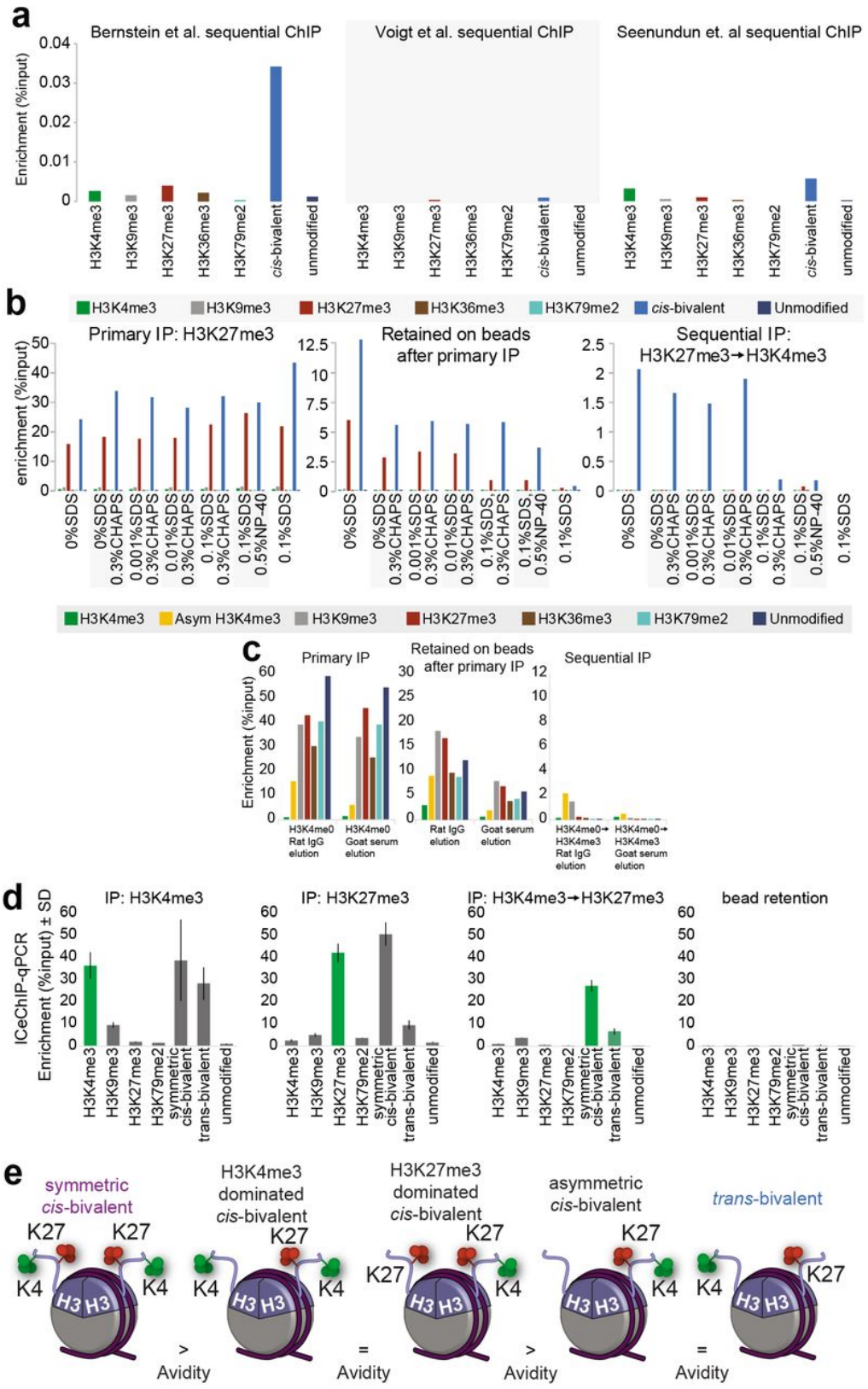


Figure 3.1: Evaluation of sequential ChIP methods.

Figure 3.1 *(previous page)*: **(a)** Enrichment of on- and off-target nucleosome standards under sequential ChIP protocols developed by Bernstein et al. (92), Voigt et al. (180), and Seenundun et al. (163). **(b-c)** Enrichment at different sequential ICeChIP steps with (b) chemical denaturant elution and (c) immunoglobulin and serum elution. **(d)** Enrichment of different nucleosome standards with ICeChIP-qPCR performed against H3K4me3, H3K27me3, and bivalency, with beads showing very little retention of chromatin (n=3 technical replicates). Error bars represent standard deviation. **(e)** Different configurations of bivalency on a single nucleosome. Of these, only trans-bivalency has been identified by mass spectrometry (180; 185).

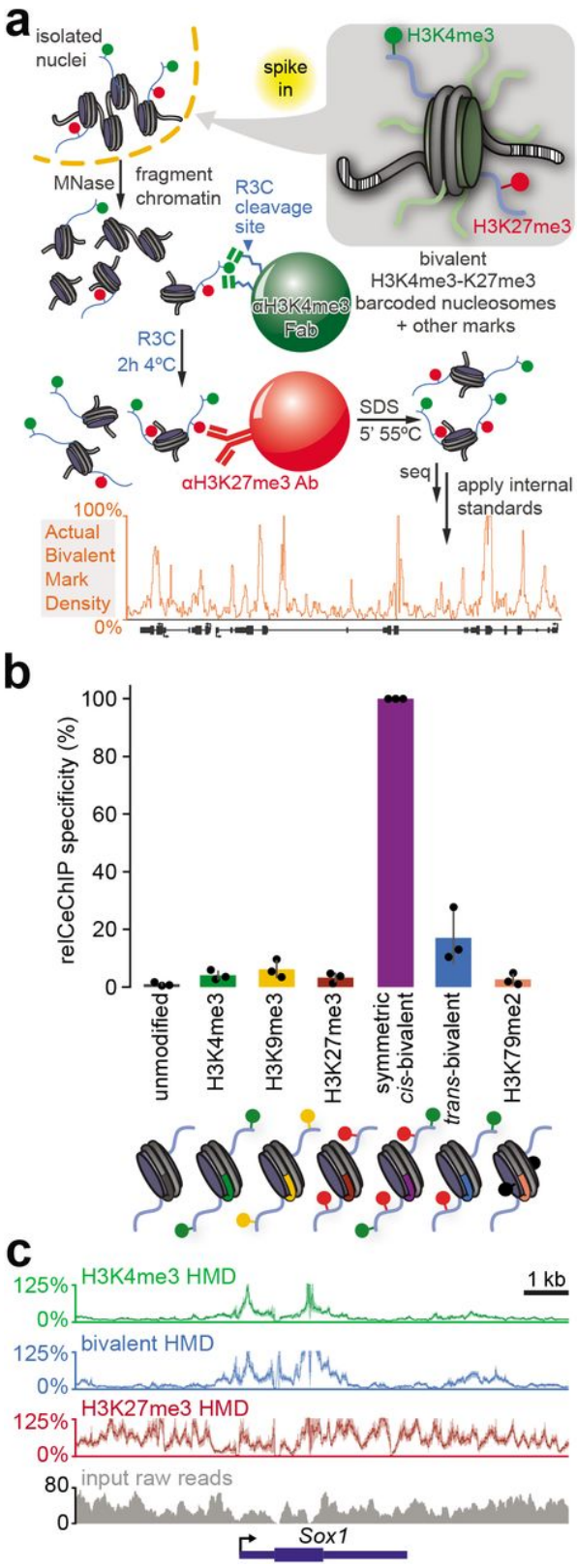


Figure 3.2: Workflow and evaluation of reICeChIP-seq.

Figure 3.2 *(previous page)*: **(a)** Schematic of reICeChIP-seq. The recombinant α-H3K4me3 Fab 304M3-B achieves high affinity by "clasping" the histone tail between two Fab molecules (184), a binding mode readily achieved by multiple copies of the Fab presented on a bead, but not by the Fab in solution. Thus, protease cleavage not only elutes nucleosomes from the beads but also likely from the Fab complex. **(b)** Enrichment of different barcoded nucleosomes in reICeChIP-seq (n=3 biological replicates). Error bars represent S.D. **(c)** Representative line plot showing histone modification density of H3K4me3, H3K27me3, and bivalency ICeChIP-seq presented with 95% confidence intervals (lighter shade) and input read depth in naïve mESCs. Bivalency is calibrated to the trans-bivalency nucleosome standard and corrected for off-target H3K9me3 pulldown.

We then leveraged this reagent to develop reICeChIP (Figure 3.2a). The first pulldown was conducted with the cleavable α-H3K4me3 Fab from native mononucleosomes (96; 186) spiked with nucleosome internal standards. We then eluted the captured nucleosomes from streptavidin resin by cleaving the antibody with HRV 3C endoprotease (187) and, with this eluent, conducted a second pulldown against H3K27me3 with a conventional antibody. This method eluted material from the primary pulldown more efficiently (Figure 3.1, resulting in 1000-2500x higher enrichment of the target over the published methods (Figure 3.2 b, Figure 3.1 a). This improvement enabled genome-wide measurement of bivalency HMD (Figure 3.2 c), representing the proportion of nucleosomes at a given locus modified with both H3K4me3 and H3K27me3, using the trans-bivalent nucleosome standards (188) as the calibrant (Figure 3.1 e, Figure 3.3; Supplementary Note 1).

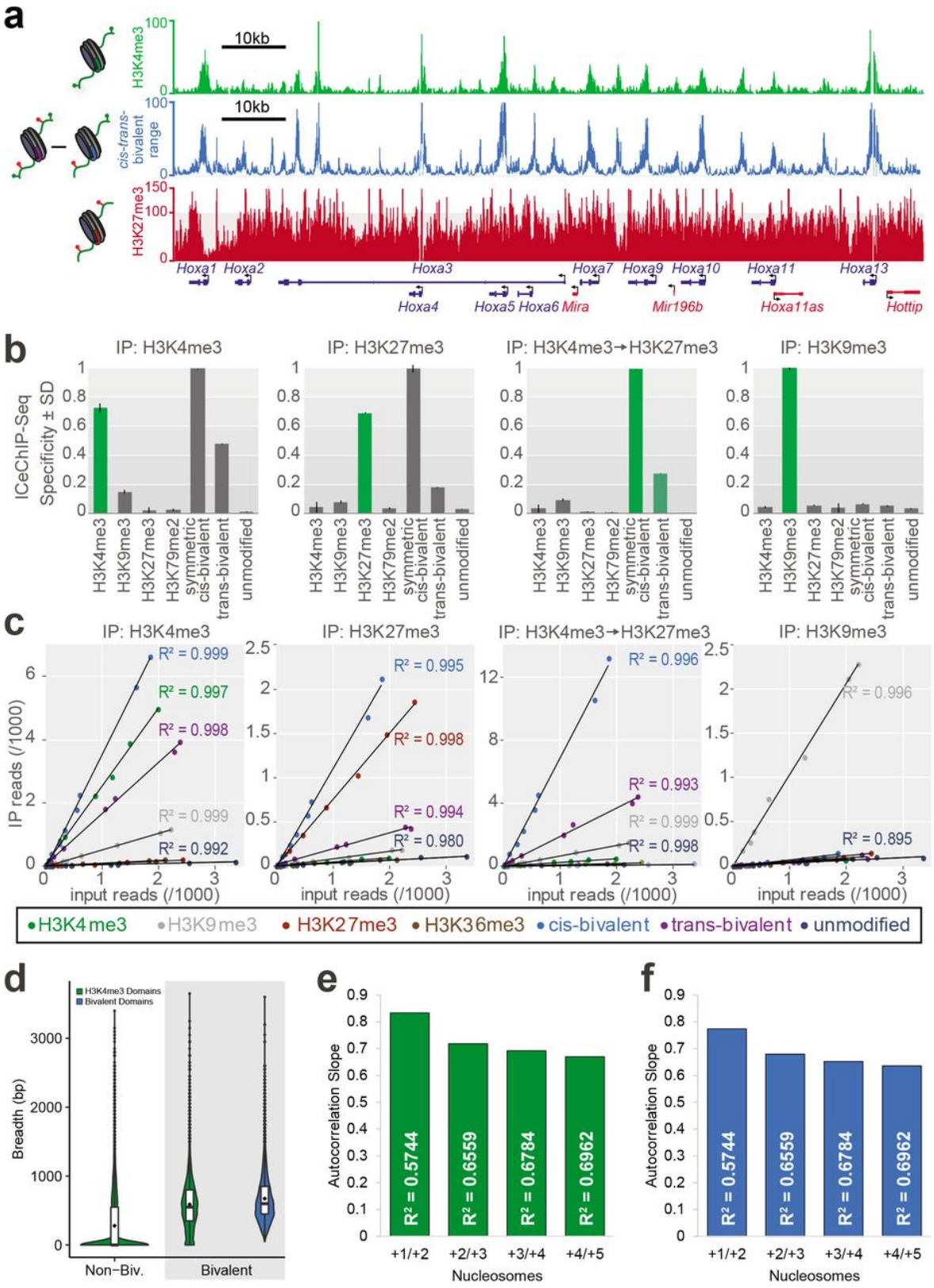


Figure 3.3: Evaluation of reICeChIP specificity and standards.

Figure 3.3 *(previous page)*: (a) Representative genome browser view of H3K4me3, H3K27me3, and bivalency, shown as a range of possible values by normalization to trans-bivalent (upper limit) or cis-bivalent (lower limit) nucleosome standards. (b) Relative pull-down of different nucleosome standards in ICeChIP-seq, normalized to the most-enriched standard. (c) Scatterplots of reads from DNA barcodes applied to nucleosome standards in ICeChIP-seq. (d) Violin plots of peak breadth (consecutive segment of 50bp windows overlapping promoter with >25% HMD) for H3K4me3 (green) and bivalency (blue) at non-bivalent and bivalent genes (>25% HMD) in naïve mESCs. (e-f) Autocorrelation of (e) H3K4me3 and (f) bivalency HMDs between nucleosomes in naïve mESCs. Nucleosomes are defined as sequential 200bp windows from the TSS.

3.5 Bivalency through differentiation

With reICeChIP, we sought to study the role of bivalency in development by tracking its changes across a differentiation pathway that was used in several classic studies of bivalency (92; 91; 189): differentiation from naïve mESCs (189) through the primed mESC state (189) to NPCs. In naïve mESCs, we noted that bivalency was far more widespread than previously reported (Figure 3.4 a,b); rather than $\tilde{1}$000 bivalent genes in naïve mESCs (189), we observed at least 10% bivalency HMD at most promoters (25768/42622), with almost 5000 promoters bearing bivalency at more than 50% of their nucleosomes (Figure 3.4 a, c; Supplementary Notes 2, 3). This trend is recapitulated with primed mESCs, with the consensus set of bivalent promoters representing fewer than 2000 genes (91; 90; 143; 144), as compared to more than 25,000 that are >25% bivalent in our analysis.

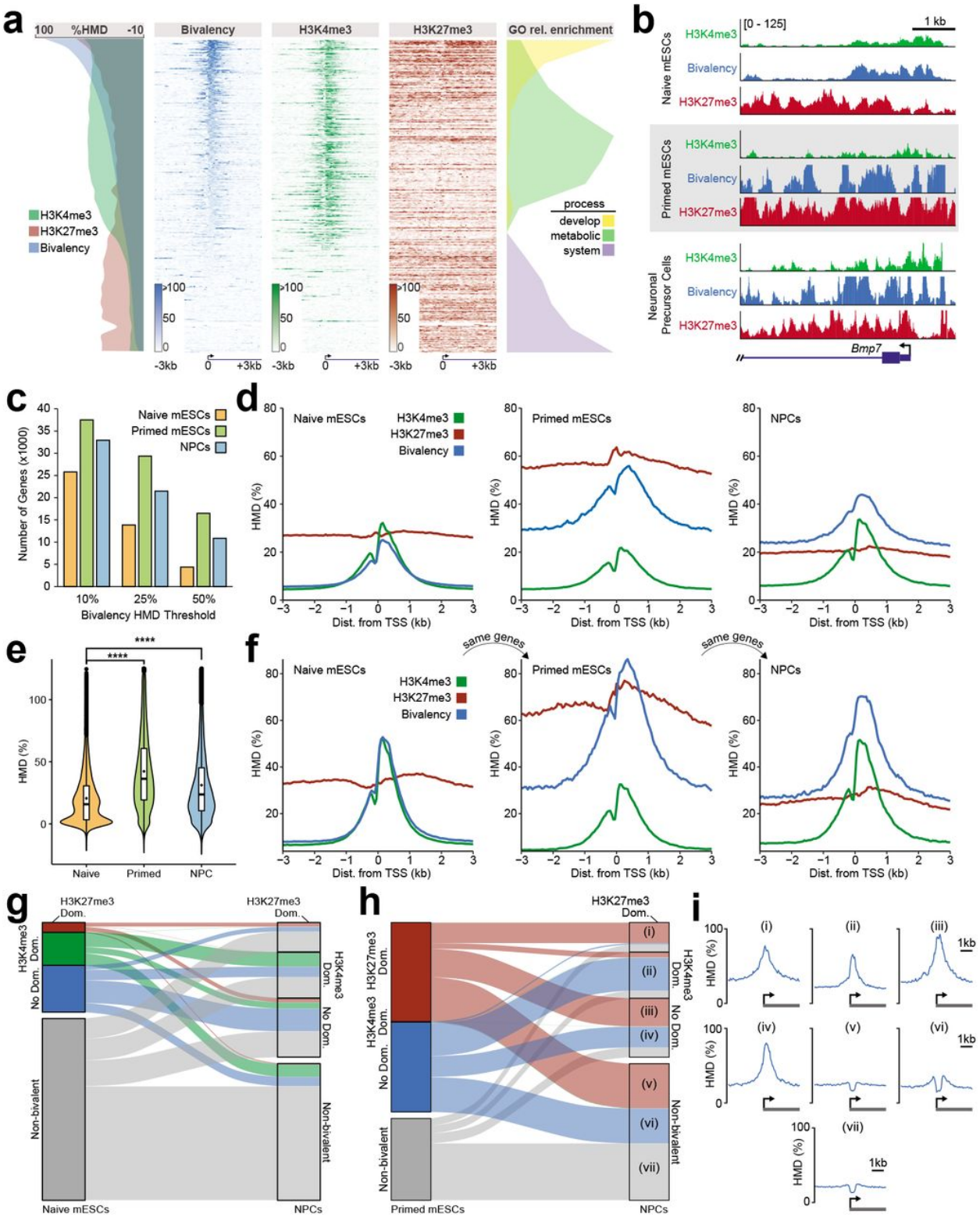


Figure 3.4: Bivalency is widespread and does not resolve over differentiation.

Figure 3.4 *(previous page)*: (a) Bivalency, H3K4me3, and H3K27me3 at all Refseq promoters in naïve mESCs, with relative enrichment of GO terms. Genes are rank ordered by bivalency HMD at promoter, defined as the region from 0 to +400 bp relative to the TSS. (b) Representative locus view of H3K4me3, H3K27me3, and bivalency at promoters in naïve mESCs (top), primed mESCs (centre), and NPCs (bottom), presented on the same scale of 0-125% HMD. (c) Number of promoters with bivalency HMDs above the given thresholds in each cell type out of a total of 42,622 Refseq promoters. (d) Metaprofiles of H3K4me3, H3K27me3, and bivalency at all promoters in naïve mESCs, primed mESCs, and NPCs. Heatmaps for primed mESCs and NPCs are presented in Extended Data Fig. 3b. (e) Distribution of bivalency HMDs at all Refseq promoters in three cell states, zoomed to below 125% HMD. Overall, 99.5% of naïve promoters, 87.3% of primed promoters, and 91.6% of NPC promoters have an HMD below 100%. Full plot in Extended Data Fig. 3a. (f) Metaprofiles of H3K4me3, H3K27me3, and bivalency at promoters identified as bivalent in naïve mESCs (25% HMD threshold), tracked from naïve mESCs to primed mESCs to NPCs. Heatmaps for bivalency are presented in Extended Data Fig. 3f. (g-h) Alluvial plots of dominance and bivalency of genes from (g) naïve mESCs to NPCs or (h) primed mESCs to NPCs. Bivalency [>25% HMD] can be subcategorized into dominance classes by independent ICeChIP for the constituent marks, with H3K27me3 in excess (H3K27me3/H3K4me3 $> e^1$), H3K4me3 dominant (H3K27me3/H3K4me3 $< e^{-1}$), or intermediate ratios (no dominance). (i) Bivalency metaprofiles for gene subsets indicated in panel (h) from −3kb to +3kb relative to the TSS. **** $p < 2.2 \times 10^{16}$.

Even more striking were the changes in bivalency across this differentiation scheme. Previous studies suggested that bivalency largely disappears upon differentiation to NPCs (92; 91; 154; 155). However, we found the opposite; promoter bivalency *increases* upon differentiation (Figure 3.4 d-e, Figure 3.5, with thousands more genes meeting bivalency HMD thresholds relative to naïve mESCs (Figure 3.4 c). Similarly, we find that bivalent domains do not resolve upon differentiation; tracking bivalent genes from naïve mESCs through differentiation, we observe that bivalency is higher at these same promoters in primed mESCs and NPCs (Figure 3.4 f, Figure 3.5 c-f). As previously reported, primed mESCs have the most bivalency, likely related to the high level of promoter H3K27me3 in this state (189) (Figure 3.4 d). Accordingly, there are 27% fewer bivalent genes in NPCs than in primed mESCs (Figure 3.4 e). However, this decrease is nowhere near the previously reported decrease of 92% (91), and bivalent genes from primed mESCs remain highly bivalent in NPCs

(Figure 3.5 g-h). Collectively, these data suggest that bivalency is far more widespread in this system than previously appreciated and remains elevated through differentiation, rather than resolving to one of the two monovalent states.

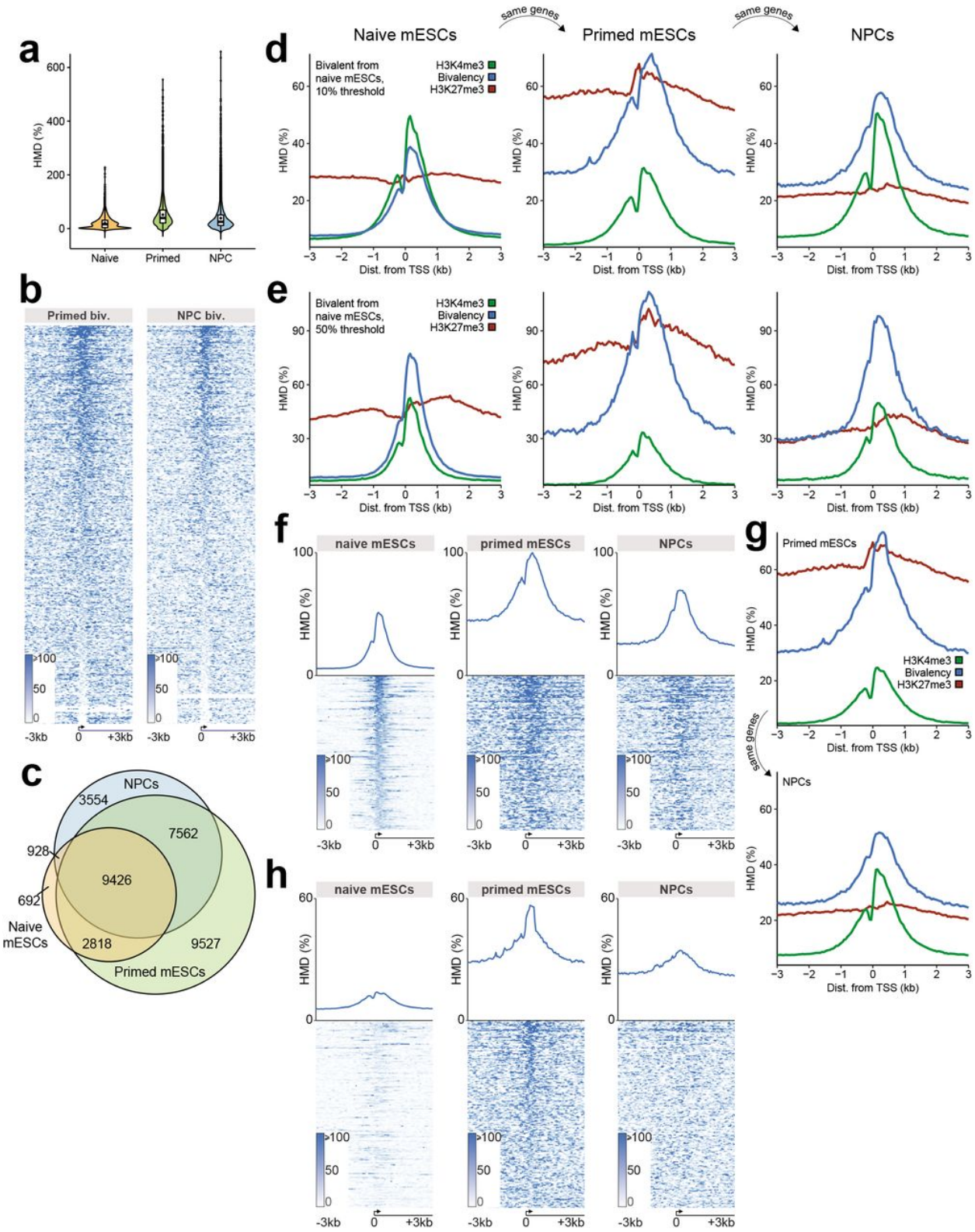


Figure 3.5: Tracking bivalent genes through differentiation.

Figure 3.5 *(previous page)*: (a) Distribution of bivalency HMDs at all Refseq promoters in three cell states. (b) Heatmaps of bivalency at all Refseq promoters in primed mESCs and NPCs. Genes are ordered by bivalency HMD at the promoter. (c) Venn diagram showing overlap of bivalent genes (25% HMD threshold) in naïve mESCs, primed mESCs, and NPCs. (d-e) Metaprofiles of H3K4me3, H3K27me3, and bivalency for bivalent genes in naïve mESCs with a (d) 10% or (e) 50% HMD threshold. (f) Heatmaps and metaprofiles of bivalent genes from naïve mESCs. (g) Metaprofiles of H3K4me3, H3K27me3, and bivalency at genes tracked from primed mESCs to NPCs for bivalent genes in primed mESCs (>25% HMD). (h) Heatmaps and metaprofiles of bivalent genes in primed mESCs that are not bivalent in naïve mESCs.

To investigate this discrepancy with the literature, we compared promoters identified as bivalent by other studies (91; 155; 143) to ours. The previously identified genes had 50-100% more H3K27me3 than do most bivalent genes in our set (Figure 3.6 a-b), suggesting that the previous studies undersampled H3K27me3 and thus could only identify regions with high H3K27me3 as bivalent. Accordingly, H3K27me3 dominant bivalent genes had the greatest proportional overlap with these canonical bivalent loci compared to other dominance classes (i.e. whether the bivalent genes have excess H3K27me3, excess H3K4me3, or roughly equal levels as measured by independent ICeChIP experiments for these two marks; (Figure 3.6 c). The common practice of measuring bivalency as regions of overlapping H3K4me3 and H3K27me3 is also problematic, even with calibrated data (96); many promoters with high H3K4me3 and H3K27me3 bear less than 25% bivalency (Figure 3.6 d). Notably, even for the previously identified bivalent genes, bivalency still increases relative to naïve mESCs upon differentiation. And in our datasets, this holds true across modification dominance classes – even the H3K27me3 dominant bivalent genes, which most closely resemble the canonically bivalent loci (Figure 3.6, 3.7). To the extent that any bivalency class resolves from naïve mESCs to NPCs, the largest set of genes is from the H3K4me3 dominant bivalent genes ($p = 1.78 \times 10^{-133}$; Figure 3.4 g), despite its minimal overlap with the canonical bivalent loci (Figure 3.6 c).

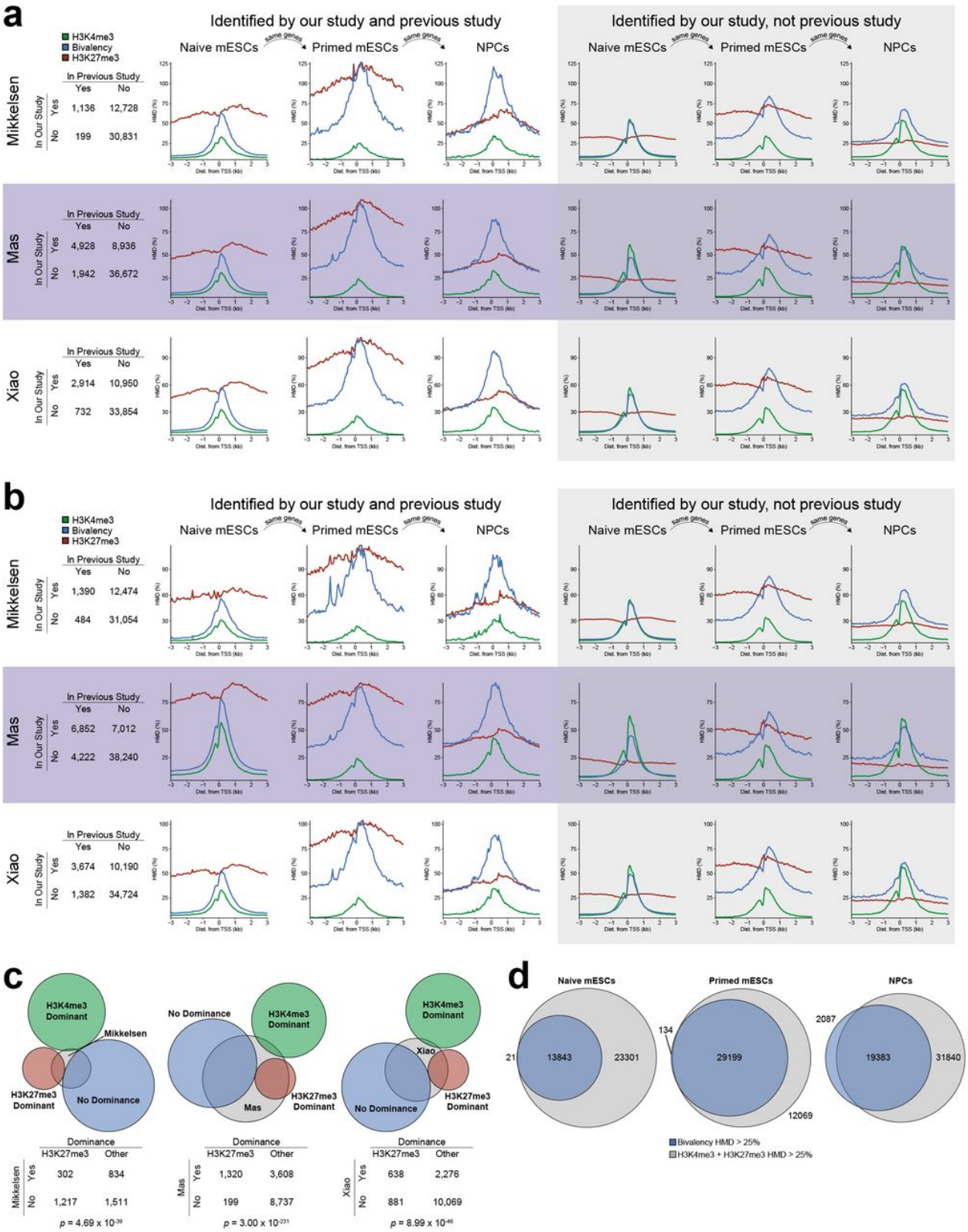


Figure 3.6: Comparing our bivalent genes to other studies.

Figure 3.6 *(previous page)*: (a-b) Contingency tables and metaprofiles for genes that are identified as >25% bivalent in our study and by Mikkelsen et al.18, Mas et al.35, and Xiao et al.58, wherein: (a) gene is identified as bivalent in the external study if overlapping H3K4me3 and H3K27me3 peaks overlap the 0 to +400bp region of a gene relative to the TSS, or (b) gene is identified as bivalent in the external study if overlapping H3K4me3 and H3K27me3 peaks overlap the region from 2.5kb upstream of the TSS to the end of the gene22. (c) Overlap of bivalent genes from external datasets (as defined in part a) with each of our bivalent gene dominance classes in naïve mESCs. Significance computed by two-tailed Fisher hypergeometric test. (d) Overlap of genes with bivalency HMD $>$ 25% and with H3K4me3 + H3K27me3 HMD $>$ 25% in all three cell states.

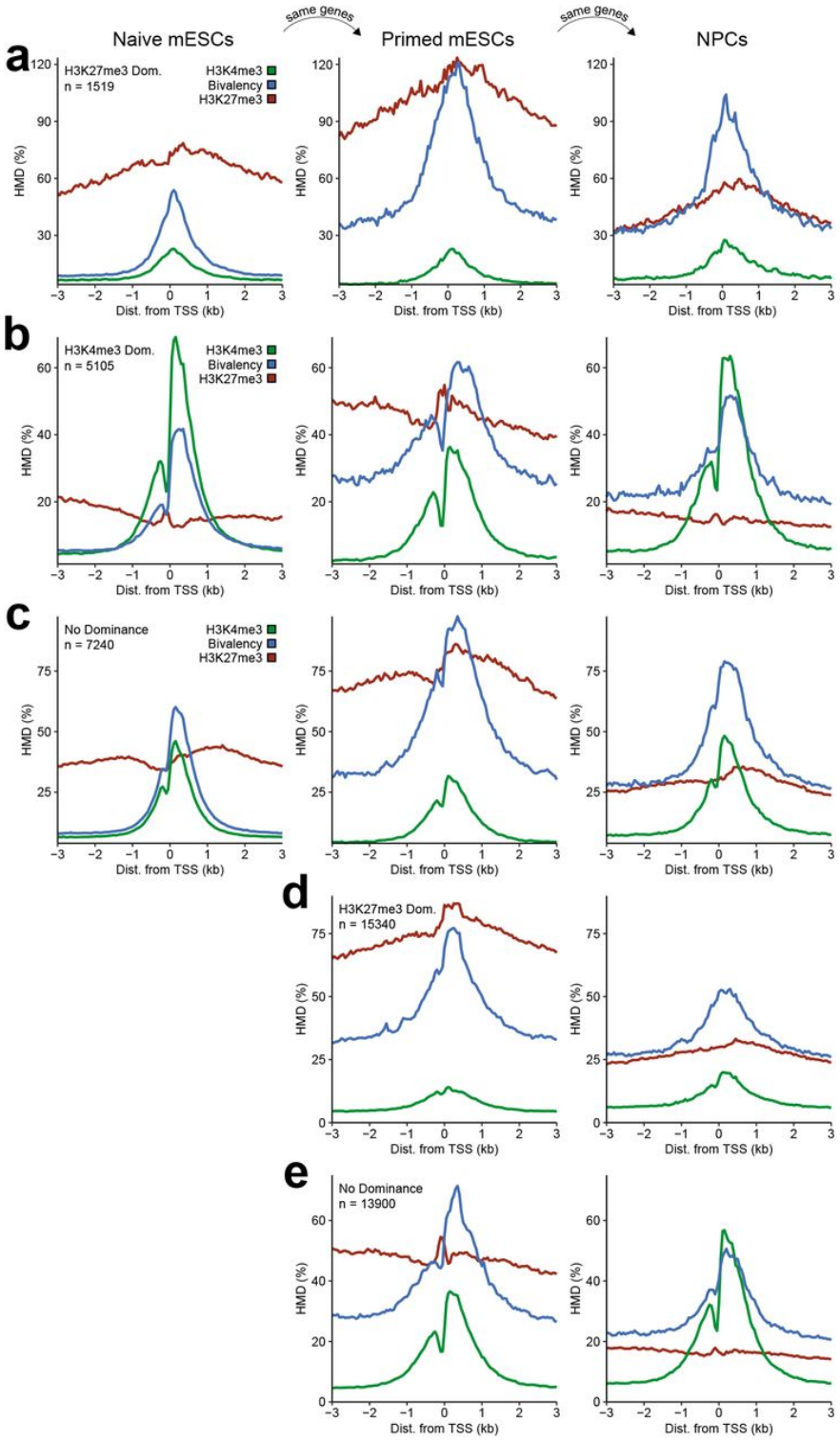


Figure 3.7: Bivalency changes across differentiation by modification dominance class.

Figure 3.7 *(previous page)*: (a-c) Metaprofiles of H3K4me3, H3K27me3, and bivalency for bivalent genes (>25% HMD) in naïve mESCs that are (a) H3K27me3 dominant (H3K27me3/H3K4me3 $> e^1$), (b) H3K4me3 dominant (H3K27me3/H3K4me3 $< e^1$), or (c) have no dominance in naïve mESCs, tracked through three cell states. (d-e) Metaprofiles of H3K4me3, H3K27me3, and bivalency for bivalent genes (>25% HMD) in primed mESCs that are for indicated dominance classes tracked from primed mESCs to NPCs.

Having found that bivalency is unexpectedly common and persistent in early differentiation, we investigated the enzyme complexes that could potentially account for this ubiquity. Previous work suggested that H3K27me3 and H3K4me3 each inhibit deposition of the other (180; 188; 190; 191), particularly when symmetric (Supplementary Note 4), raising questions as to whether the pervasive bivalency we observe is plausible. To address this concern, we performed histone methyltransferase (HMTase) assays with Set1B and the full panel MLL-family core complexes (MLL1, MLL2, MLL3, MLL4), which collectively account for the bulk of H3K4 methylation in humans (192). We find that these complexes all tolerate a wide spectrum of H3K27me3-decorated nucleosomes (Figure 3.8, indicating that the formation of bivalent nucleosomes is not precluded by allosteric modulation of H3K4me3 installation by core factors. Although it has been suggested that Set1a (193), Mll2 (194), Ezh1 (195), and Ezh2 (196) are all important for establishing bivalency, only Mll2 appears to be sensitive for identifying bivalent promoters in naïve mESCs, with none showing high specificity for the same (Figure 3.9). Together, these data support the proposed specialized role for Mll2 in bivalency (194), indicate a pleiotropic role for PRC2 beyond its role in establishing bivalency, and provide plausible enzymatic avenues to the prevalent bivalency we observe by reICeChIP.

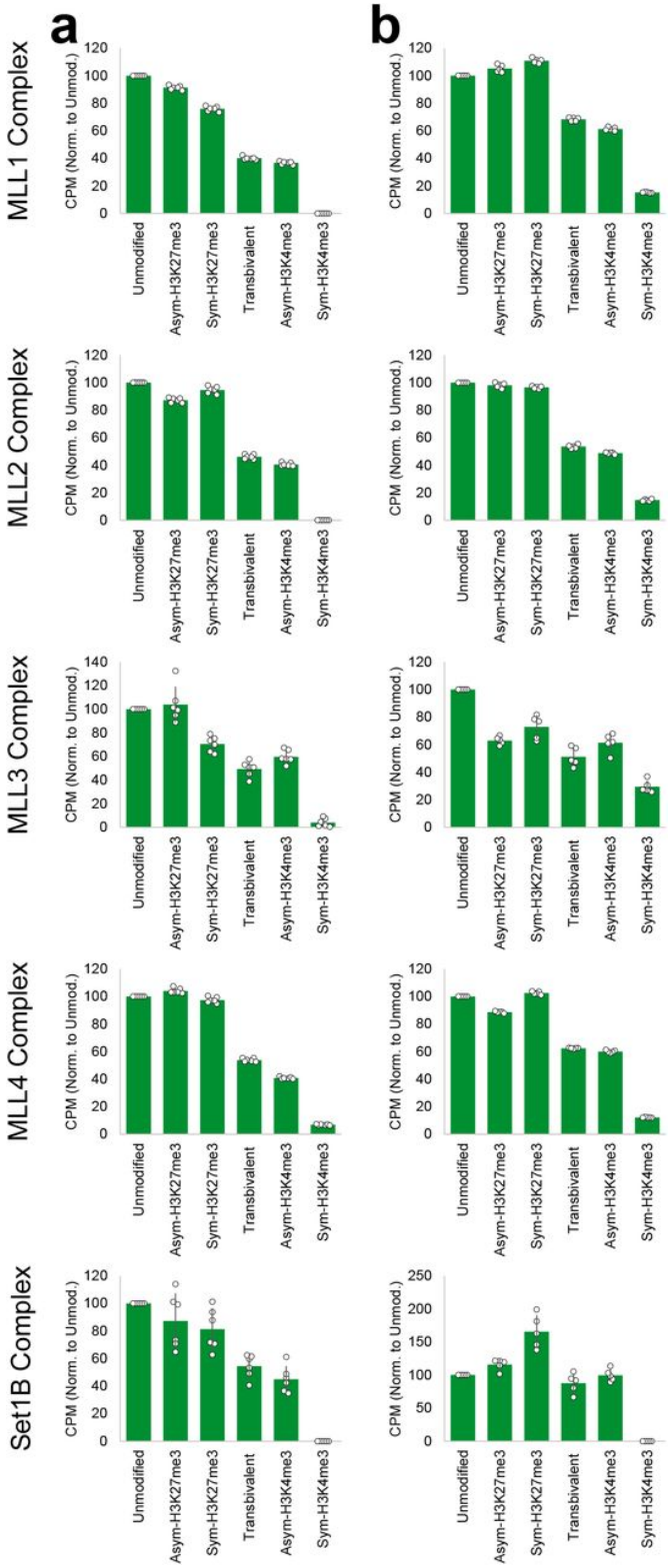


Figure 3.8: Methyltransferase assays identifying potential pathways for establishment of bivalency.

Figure 3.8 *(previous page)*: (a-b) Methyltransferase assays for MLL1, MLL2, MLL3, MLL4, and Set1B core HMTase complexes using (a) 15 ng/uL (n=6) and (b) 20 ng/uL (n=5) semisynthetic nucleosomes as substrates for methylation. Endpoints were established at 180 min by kinetic evaluation to be sensitive to difference in activity for this panel. Signal is corrected for background and no nucleosome substrate activity. Error bars represent standard deviation.

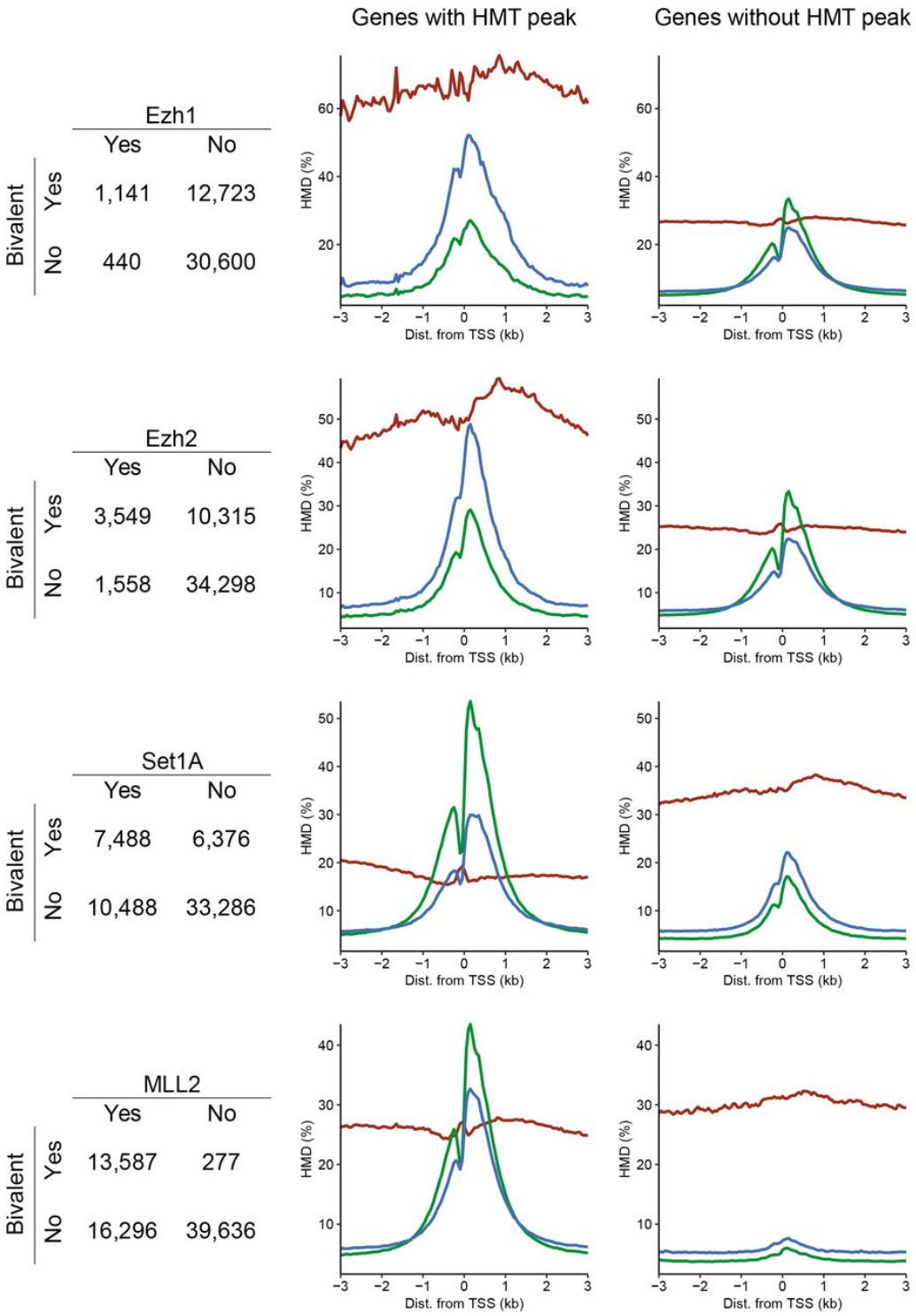


Figure 3.9: HMTase peaks and bivalency.

Figure 3.9 *(previous page)*: Contingency tables and metagene profiles in naïve mESCs for genes with and without overlapping HMT peaks. Ezh1 and Ezh2 peaks were identified as Suz12 peaks lost upon Ezh1 or Ezh2 knockout (197). Set1A peaks were identified by ChIP against Set1A (193). Mll2 peaks were identified by ChIP against Mll2 (178).

3.6 Bivalency, gene expression, and ontology

A key pillar of the bivalency hypothesis is that bivalent promoters are associated with transcriptionally repressed genes poised to either be activated or terminally silenced upon differentiation (92; 91; 156; 90). However, bivalency is not solely found at genes with low expression in any of our measurements (Figure 3.10 a; Figure 3.11 a-b). Rather, bivalent genes had higher average expression than did non-bivalent genes or the set of all genes, and these genes display modestly higher average expression through differentiation (Figure 3.10 a), with bivalency remaining similar across most gene expression deciles (Figure 3.11 c). Bivalency associated similarly with bulk gene expression (Figure 3.10 b) and the proportion of cells expressing the associated transcripts in single cell RNA-seq (Figure 3.10 c), suggesting that the association of bivalency with higher-expressed genes is not solely driven by intercellular heterogeneity. Consistent with previous observations(91), bivalency was higher at promoters with high CpG content (Figure 3.11 d) and associated with lower DNA methylation compared to non-bivalent genes (Figure 3.11 e, also holds for each dominance class). These data all suggest that bivalent genes are more highly expressed than non-bivalent genes as a whole, and this latter class is seemingly more subject to regulation by DNA methylation.

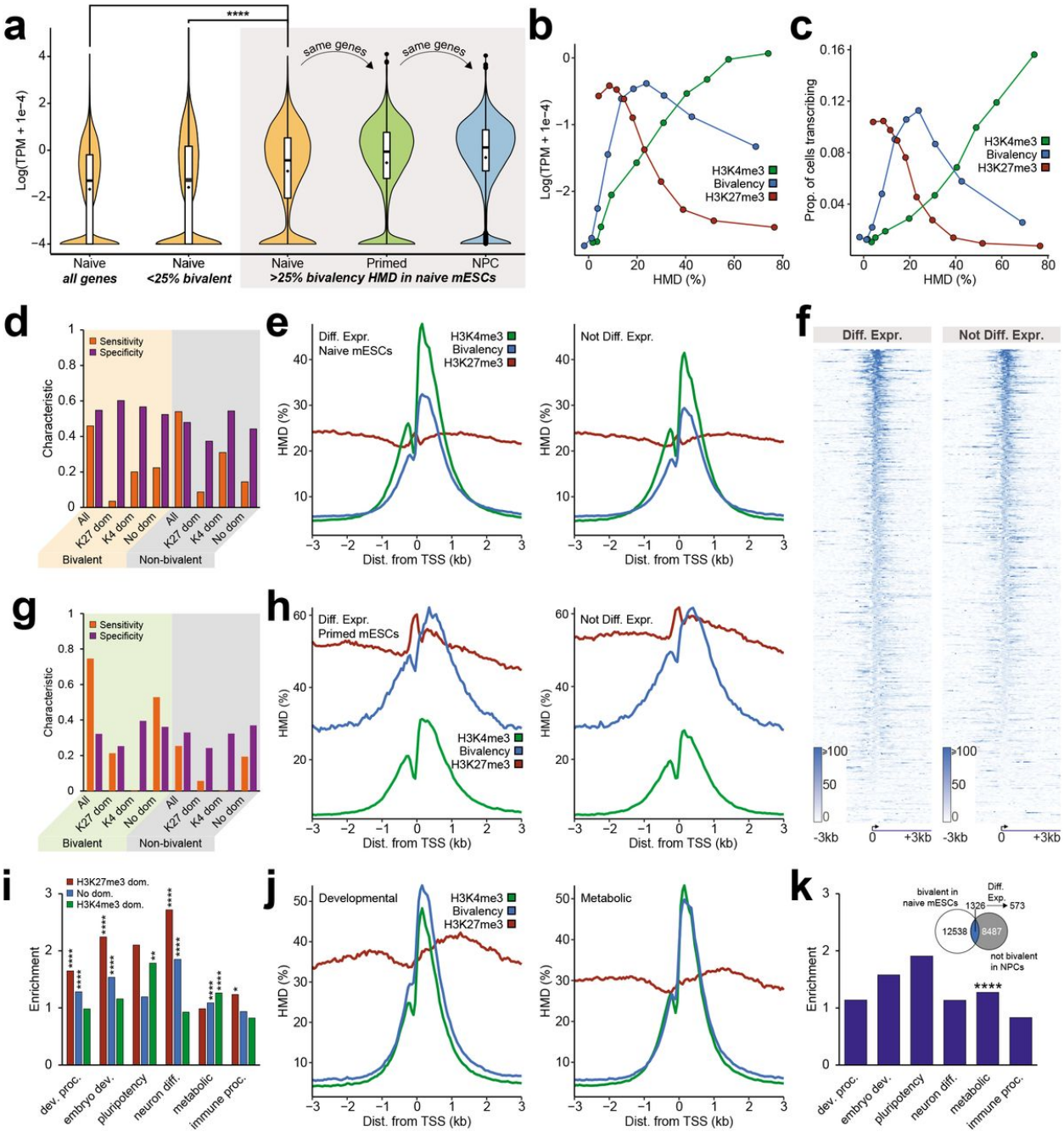


Figure 3.10: Bivalency is neither sensitive nor specific for poised nor developmental genes.

Figure 3.10 *(previous page)*: (a) Violin plots of gene expression79 for all genes in naïve mESCs, non-bivalent genes (<25% HMD) in naïve mESCs, and bivalent genes (>25% HMD) tracked from naïve mESCs to the same genes in the indicated lineages. Significance computed by Welch's two-tailed t-test. (b) Gene expression vs. HMD for H3K4me3, H3K27me3, and bivalency (genes are binned into HMD deciles). (c) Proportion of actively transcribing cells by single-cell RNA-seq80 vs. HMD for H3K4me3, H3K27me3, and bivalency (genes are binned into HMD deciles). (d) Sensitivity and specificity (Supplementary Note 5) of bivalent and non-bivalent genes in naïve mESCs identifying differentially expressed genes (DEGs) from the naïve state to the NPC state. (e) Metaprofiles of H3K4me3, H3K27me3, and bivalency and (f) heatmaps of bivalency in naïve mESCs at DEGs and non-DEGs relative to NPCs. (g) Sensitivity and specificity of bivalent and non-bivalent genes in primed mESCs identifying DEGs from the primed state to the NPC state. (h) Metaprofiles of H3K4me3, H3K27me3, and bivalency in primed mESCs at DEGs and non-DEGs. (i) Gene ontology term enrichment of H3K27me3-dominant bivalent genes, H3K4me3-dominant bivalent genes, or bivalent genes with no clear dominance (q-value two-tailed Fisher hypergeometric test). (j) Metaprofiles of H3K4me3, H3K27me3, and bivalency in naïve mESCs at developmental and metabolic genes. (k) Gene ontology term enrichment of genes following the classic bivalency model: DEGs that lose bivalency from naïve mESCs (>25% HMD) to NPCs (<10% HMD). Significance computed by two-tailed Fisher hypergeometric test. * $q < 0.05$. ** $q < 0.01$. **** p or $q < 2.2 \times 10^{16}$.

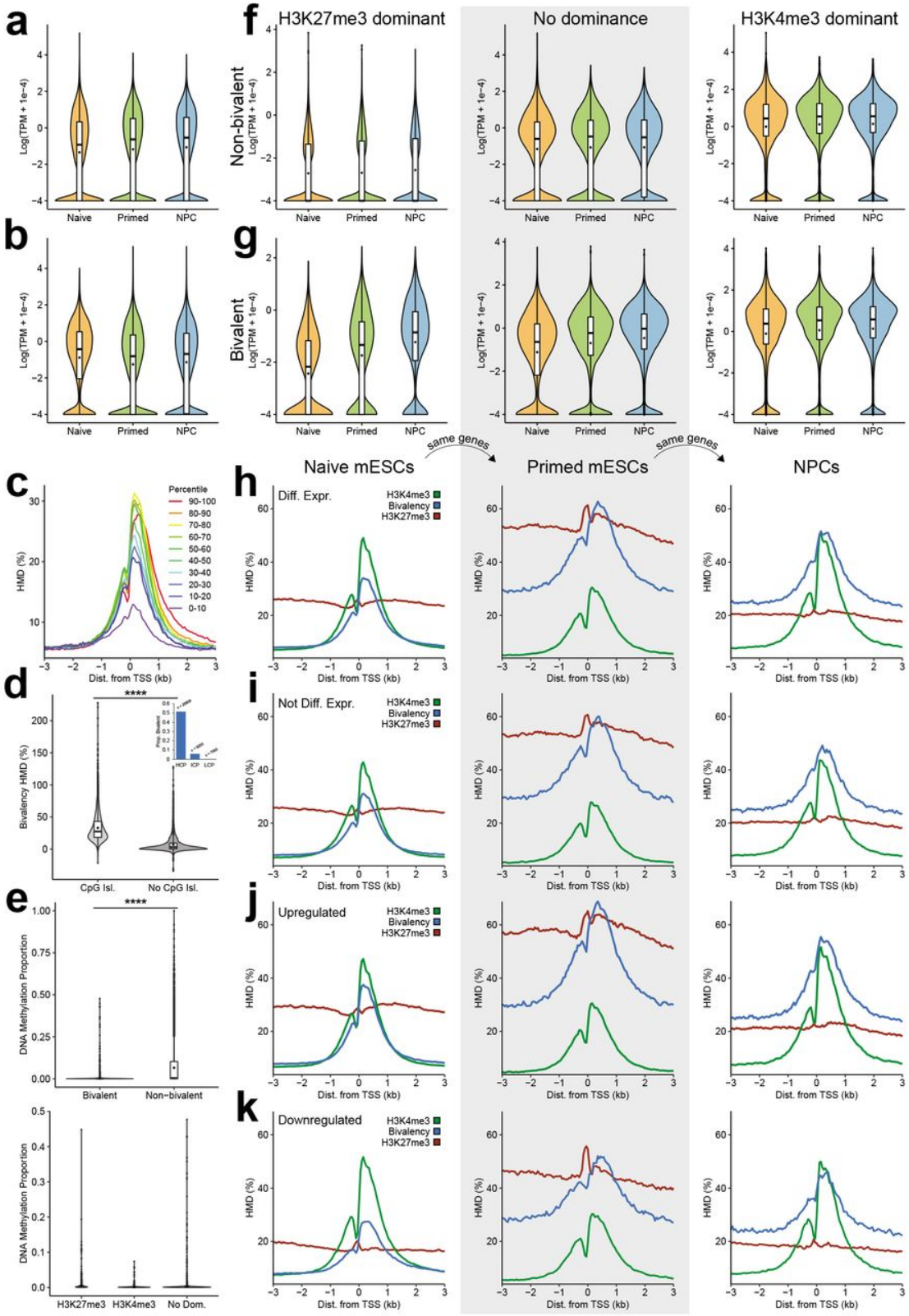


Figure 3.11: Bivalency and differential gene expression.

Figure 3.11 *(previous page)*: (a-b) Violin plots of gene expression for (a) all genes and (b) bivalent (>25% HMD) genes in each cell state. (c) Bivalency metaprofiles in naïve mESCs at promoters binned by gene expression decile. (d) Violin plots of bivalency HMD in naïve mESCs at promoters with and without CpG islands. Inset shows proportion of genes that are bivalent in sets of genes classified by CpG content: high-CpG promoters (HCP), intermediate-CpG promoters (ICP), and low-CpG promoters (LCP), defined as previously described by Mikkelsen et al.18. Total number of genes in each class is provided as n. (e) Violin plots of DNA methylation at bivalent and non-bivalent genes (top), broken by dominance class for bivalent genes (bottom). (f-g) Violin plots of gene expression in (f) non-bivalent (<25% HMD) and (g) bivalent (>25% HMD) genes from naïve mESCs that are H3K27me3 dominant (H3K27me3/H3K4me3 $> e^1$; left), have no clear dominance (centre), or are H3K4me3 dominant (H3K27me3/H3K4me3 $< e^1$; right). (h-k) Metaprofiles of H3K4me3, H3K27me3, and bivalency at genes tracked from naïve mESCs to primed mESCs to NPCs for (h) DEGs, (i) non-DEGs, (j) genes upregulated from naïve mESCs to NPCs, and (k) genes downregulated from naïve mESCs to NPCs. **** p $<10^{16}$ (Welch's two-tailed t-test).

Another pillar of the bivalency model is that bivalent genes are poised to be differentially regulated through differentiation. To test this, we computed the sensitivity and specificity of different bivalency and non-bivalency classes for differentially expressed genes (DEGs; Supplementary Note 5). Counter to the bivalency hypothesis and previous results (92; 91; 156), we found that bivalency was a very poor marker of DEGs; from naïve mESCs to NPCs, bivalency was roughly as sensitive and specific for identifying DEGs as was a *lack* of bivalency (Figure 3.10 d). Though H3K27me3-dominant bivalent genes showed an increase in average gene expression (Figure 3.11 f-g), this class still only had 60% specificity for identifying DEGs, with very low sensitivity (Figure 3.10 d). Promoters of DEGs and non-DEGs from naïve mESCs to NPCs have highly similar histone modification metaprofiles in naïve mESCs (Figure 3.10 e-f) and across differentiation (Figure 3.11 h-k). Comparison of primed mESCs to NPCs displayed similar trends (Figure 3.10 g-h); though sensitivity was higher because most genes are bivalent in primed mESCs, the specificity remained similar between bivalent and non-bivalent genes. Interestingly, whether genes were upregulated, downregulated, or non-DEGs, bivalency still increased over differentiation (Figure 3.11 h-k). Collectively, these analyses show that bivalency is neither sensitively nor specifically associated with poised

DEGs in this system.

We next examined whether bivalency is primarily associated with developmental genes, a central tenet of the original model (91; 92). The first ICeChIP study indirectly hinted that there may be at least two classes of bivalent promoters: an H3K27me3 dominant class associated with developmental genes, and an H3K4me3 dominant class enriched for metabolic genes (96). Direct measurements of bivalency herein unambiguously demonstrate this phenomenon more broadly (Figure 3.4 a, 3.10 i). Overall, bivalent genes are enriched for a broad range of ontology terms, including developmental, metabolic, and immune system process genes (Figure 3.10 i-j), with nearly identical bivalency profiles in naïve mESCs (Figure 3.10 i, Figure 3.12 a). These classes all not only retained, but increased bivalency into NPCs – even immune system process genes, despite being seemingly unrelated to neuronal development. We only found 543 genes that *did* obey the classic bivalency model (Figure 3.10 k), representing less than 5% of the bivalent genes from naïve mESCs, with little difference in bivalency between upregulated and downregulated genes (Figure 3.12 b). Interestingly, these genes were most significantly enriched for metabolic rather than developmental genes (Figure 3.10 k). Taken together, these data suggest that bivalency is neither primarily nor specifically associated with developmental genes in this system.

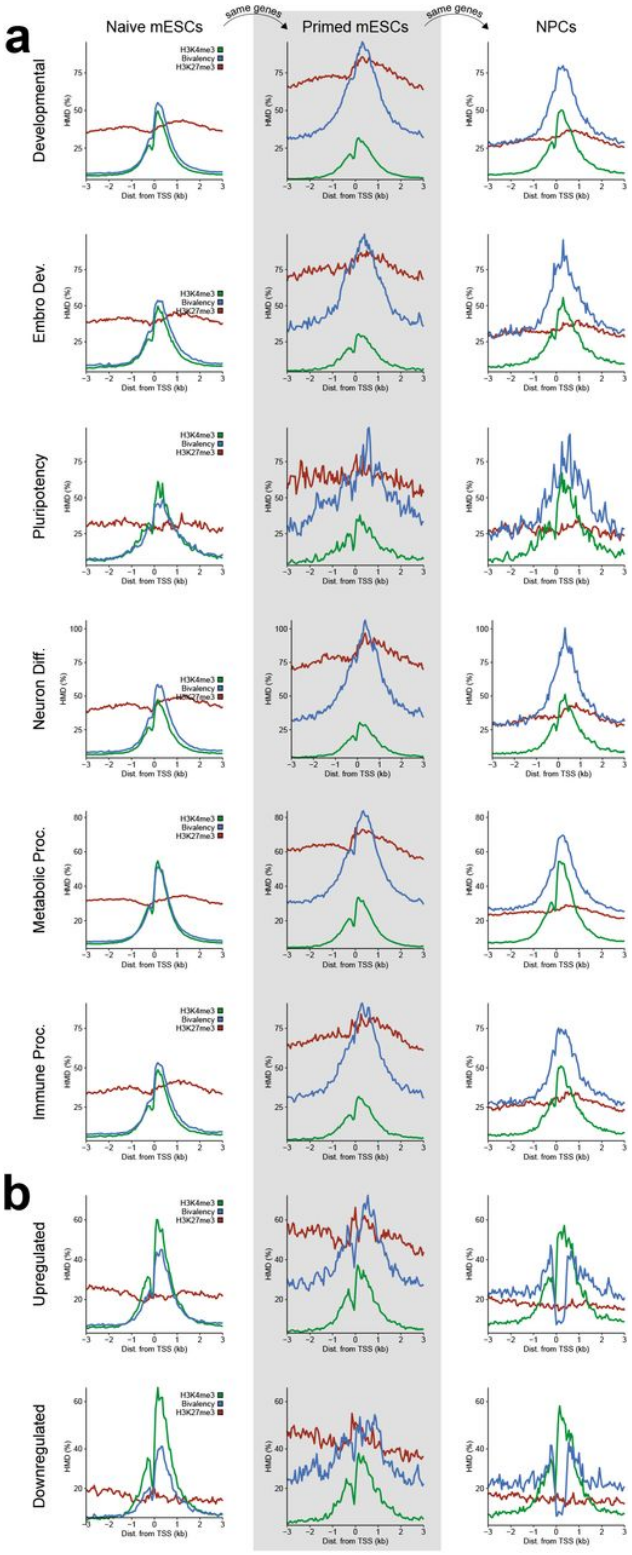


Figure 3.12: Bivalency at different classes of genes.

Figure 3.12 *(previous page)*: (a) Metaprofiles of H3K4me3, H3K27me3, and bivalency at genes tracked from naïve mESCs to primed mESCs to NPCs for bivalent genes of indicated gene ontology terms. (b) Metaprofiles of H3K4me3, H3K27me3, and bivalency at genes tracked across differentiation for genes that lose bivalency at the promoters (0 to +400bp relative to TSS) from naïve mESCs (>25% HMD) to NPCs (<10% HMD) and are upregulated (top) or downregulated (bottom) over differentiation.

3.7 Predicting DEGs with histone PTMs

The premise of the bivalency hypothesis is that the coexistence of H3K4me3 and H3K27me3 synergistically provides additional predictive information about the associated genes upon differentiation beyond that provided by H3K4me3 and H3K27me3 alone. With quantitative measurements of these modifications, this hypothesis can be tested by modelling. We first determined which individual parameters best identified DEGs by measuring the area under the curve (AUC) of receiver operator characteristic (ROC) curves of parameter thresholds. Of the individual histone modifications, H3K4me3 levels were best for identifying DEGs, with the highest AUC of the ROC (Figure 3.13 a; Figure 3.14 a). Bivalency was less predictive of DEGs than were either the log ratio of H3K27me3 and H3K4me3 or DNA methylation (Figure 3.13 a; Figure 3.14 a). And in primed mESCs, far from being predictive of poised genes, bivalency was *inversely* associated with DEGs upon differentiation to NPCs (Figure 3.14 a).

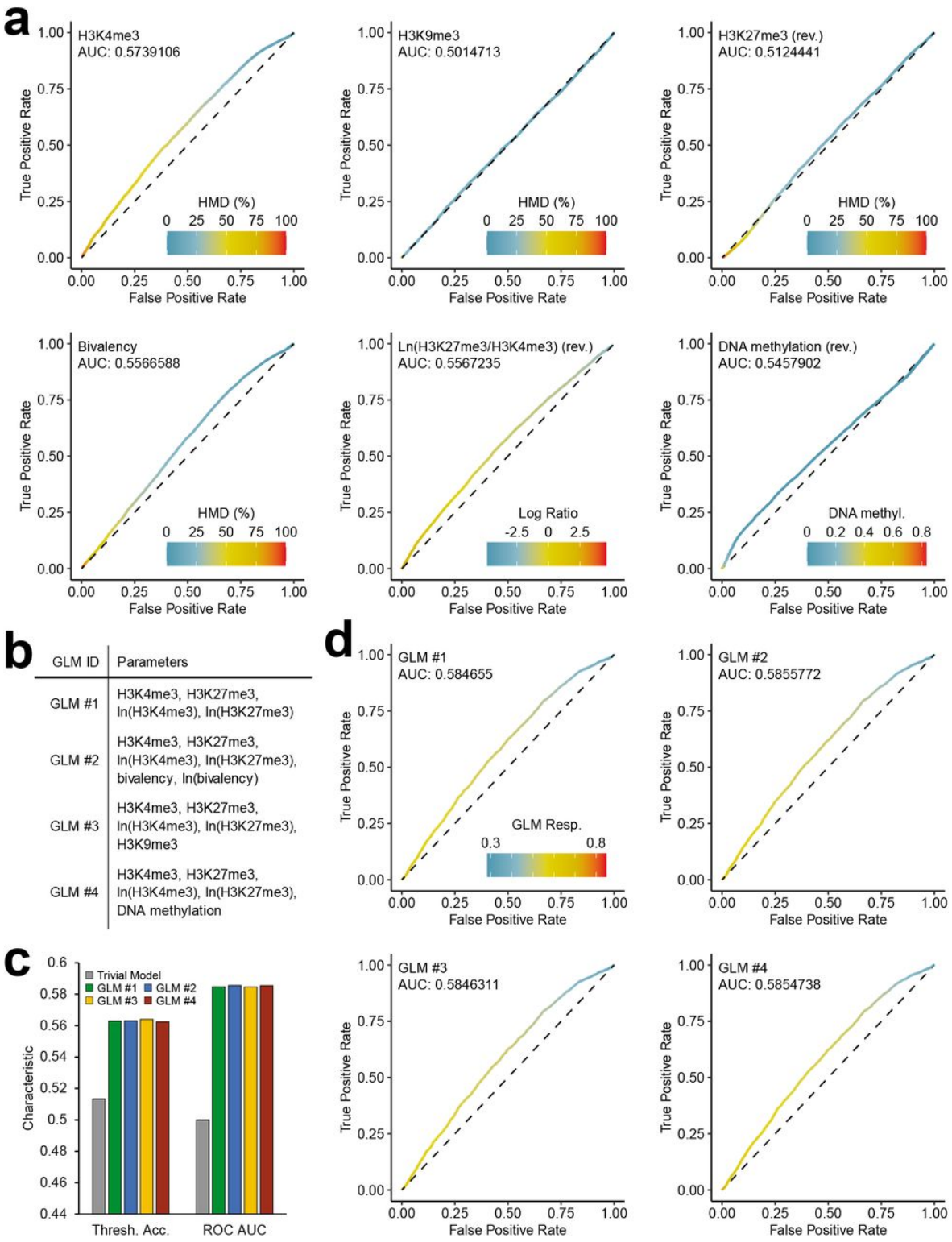
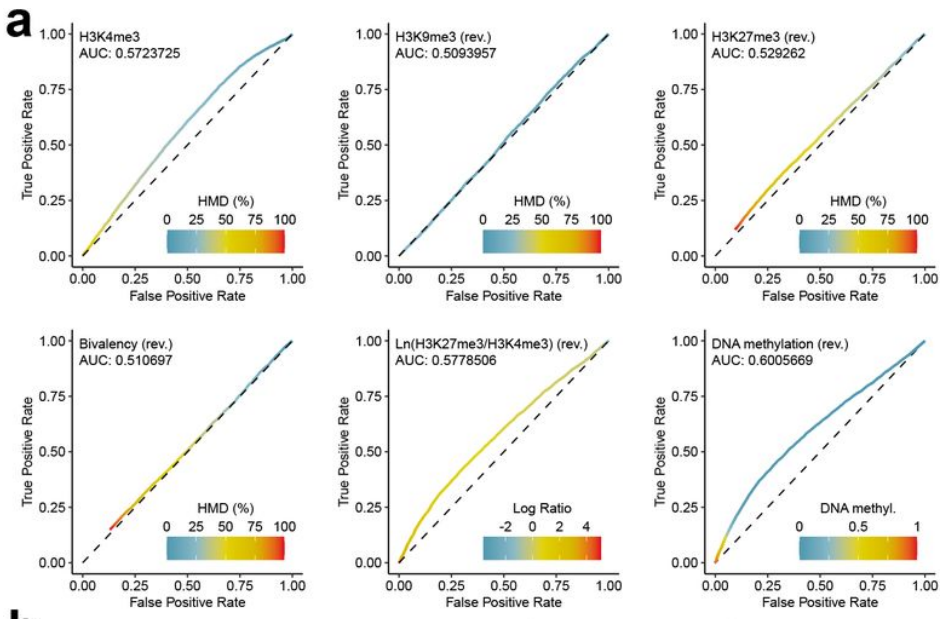


Figure 3.13: Bivalency does not provide appreciably more information than H3K4me3 and H3K27me3 alone for DEG prediciton.

Figure 3.13 *(previous page)*: (a) Receiver operator characteristic (ROC) curves for identifying DEGs from naïve mESCs to NPCs by H3K4me3, H3K9me3, H3K27me3, bivalency, ln(H3K27me3/H3K4me3), or DNA methylation in naïve mESCs. For each point, parameter value threshold used to compute true positive rate (TPR) and false positive rate (FPR) is indicated by the colour. Traits with thresholds identifying non-DEGs rather than DEGs are marked with "rev." (b) Legend for generalized linear models (GLMs) in panels c-d. (c) Accuracy of trivial model and GLMs by threshold accuracy (gene identified as DEG if logistic regression > 0.5; left) and by ROC area under curve (right). (d) ROC curves for identifying DEGs from naïve mESCs to NPCs by different GLMs. For each point, logistic regression threshold value used to compute TPR and FPR is indicated by the colour.

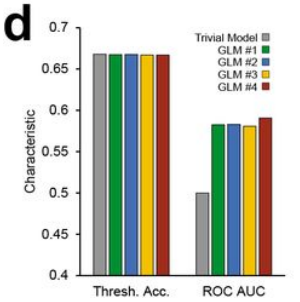


b

Parameters	Naïve to NPC BIC	Naïve to NPC Model Rank	Primed to NPC BIC	Primed to NPC Model Rank
H3K4me3, H3K27me3, log of each, DNA methylation	31678.11	1	28572.21	1
H3K4me3, H3K27me3, log of each	31681.10	2	28690.06	4
H3K4me3, H3K27me3, Bivalency, log of each	31694.95	3	28698.28	5
H3K4me3, H3K27me3, Bivalency, log of each, DNA methylation	31695.11	4	28575.32	2
H3K4me3, H3K27me3	31755.64	5	28743.02	6
H3K4me3, H3K27me3, Bivalency	31765.15	6	28749.58	7
H3K4me3	31799.66	7	28812.13	8
DNA methylation	32013.54	8	28642.26	3
Bivalency	32049.27	9	29061.58	11
Trivial Model	32182.03	10	29060.79	10
H3K27me3	32190.74	11	29019.73	9
H3K9me3	32191.98	12	29070.02	12

c

GLM ID	Parameters
GLM #1	H3K4me3, H3K27me3, ln(H3K4me3), ln(H3K27me3)
GLM #2	H3K4me3, H3K27me3, ln(H3K4me3), ln(H3K27me3), bivalency, ln(bivalency)
GLM #3	H3K4me3, H3K27me3, ln(H3K4me3), ln(H3K27me3), H3K9me3
GLM #4	H3K4me3, H3K27me3, ln(H3K4me3), ln(H3K27me3), DNA methylation



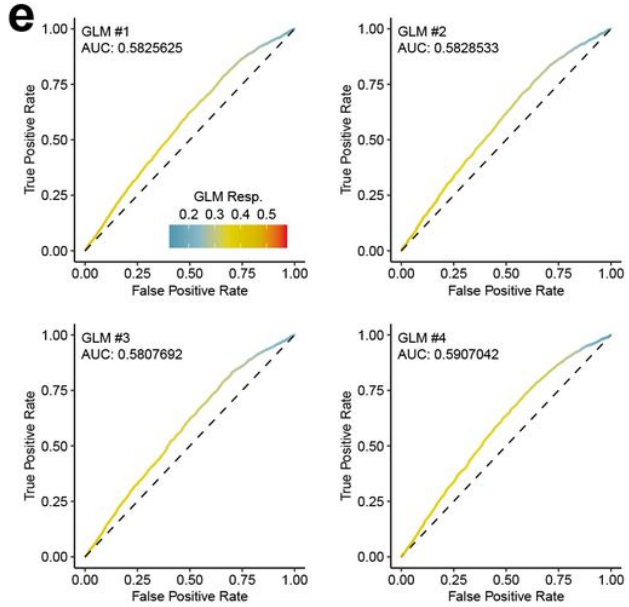


Figure 3.14: Modelling the additional information content provided by bivalency over H3K4me3 and H3K27me3 alone.

Figure 3.14 *(previous page)*: (a) ROC curves for identifying DEGs from primed mESCs to NPCs by H3K4me3, H3K9me3, H3K27me3, bivalency, ln(H3K27me3/H3K4me3), or DNA methylation in primed mESCs. For each point, parameter value threshold used to compute true positive rate (TPR) and false positive rate (FPR) is indicated by the colour. Traits with thresholds identifying non-DEGs rather than DEGs are marked with "rev." (b) Bayes Information Criterion (BIC) for logistic models identifying DEGs from naïve mESCs or primed mESCs to NPCs with different parameters. (c) Legend for generalized linear models (GLMs). (d) Accuracy of trivial model and GLMs by threshold accuracy (gene identified as DEG if logistic regression > 0.5; left) and by ROC area under curve (right). (e) ROC curves for identifying DEGs from primed mESCs to NPCs by different GLMs. For each point, logistic regression threshold value used to compute TPR and FPR is indicated by the colour.

If bivalency provides additional information over H3K4me3 and H3K27me3, then a model without bivalency will be markedly less explanatory than a model with bivalency. To test this, we conducted logistic regressions with linear models to identify parameters most important for identifying DEGs. Bayes Information Criterion analyses preliminarily hinted that bivalency provided minimal information to this end (Figure 3.14 b; Supplementary Note 6). To more definitively identify whether bivalency provides meaningful predictive information, we conducted hold-out cross-validation on models with H3K4me3, H3K27me3 and either nothing else, bivalency, H3K9me3, or DNA methylation (Figure 3.13 b; Figure 3.14 c; Supplementary Note 6). Parameters other than H3K4me3 and H3K27me3 barely improved model accuracy by two separate metrics (Figure 3.13 c-d; Figure 3.14 d-e; Supplementary Note 4), suggesting that those parameters provide virtually no additional information content to identify DEGs. These data suggest that, in this developmental system, there is little evidence that bivalency has emergent properties in identifying poised genes beyond the combined independent properties of H3K4me3 and H3K27me3.

3.8 Discussion

The bivalency hypothesis is one of the more influential ideas in epigenetics and molecular developmental biology. Persistent interest over the years coupled with widespread deployment and acceptance of sub-optimal bivalency measurement methods has ossified the hypothesis into dogma that extends well beyond any of the experimental data that informed it.

However, this coalescence has not been reached based on functional assays. Indeed, to the extent that functional validation of the bivalency model has been attempted, it has primarily been through deletion of enzymes with pleiotropic effects and functions throughout the genome beyond installation of bivalency (143; 194; 198; 199). Overwhelmingly, the prevailing views on the role of bivalency are derived from ChIP experiments. However, ChIP protocols (200) and antibodies (95; 201; 202; 203; 204) are often highly susceptible to off-target pulldown, and uncalibrated ChIP without exogenous normalization can distort signal and the ability to compare experiments (96; 95; 94), leading to spurious conclusions (95). From the quantitative and specific measurements we made with reICeChIP, we fear that this has been the case with the bivalency hypothesis, at least as far as these analyses in early mESC differentiation permit.

It has been held that bivalency is present at a small, restricted set of promoters early in development; we find that bivalency is widespread, with many thousands of promoters displaying high bivalency levels. It has been held that bivalency primarily exists early in development and resolves upon differentiation; we find that bivalency persists at least through the NPC stage and *increases* over baseline in that span. It has been held that bivalency demarcates poised, developmental genes associated with lineage commitment; we find that bivalency is neither sensitively nor specifically associated with developmental nor differentially expressed genes – and, at worst, may be *inversely* associated with the latter. Moreover, bivalent genes are predominantly not poised in an off state, but are more highly expressed than those that are not bivalent. All told, we find little evidence that bivalency provides

more information in predicting poised gene status than do H3K4me3 and H3K27me3 in an independently additive manner in this system, raising questions as to whether it represents any more than a coincidental overlap of the aforementioned two marks.

Our study is not without caveats. First, we are only able to comment meaningfully on the differentiation paradigm presented here; we cannot definitively infer that these results will hold for the other developmental or clinical contexts. Although the original studies on bivalency indicated that bivalency almost entirely disappeared by the NPC stage (92; 91), this stage is not terminally differentiated, so it is possible that bivalency could resolve in later stages of differentiation. Future studies will be needed to address this possibility in other developmental contexts. Second, though the extant evidence suggests that only trans-bivalency is present at meaningful levels, our method cannot selectively distinguish between *cis*-, *trans*-, and intermediate bivalency conformations (Supplementary Note 1).

The reICeChIP method is not inherently restricted to the study of H3K4me3/H3K27me3 bivalency. With cleavable recombinant affinity reagents targeting other histone modifications (184; 205) it could be used to quantify other combinatorial modification patterns (206; 207; 208; 209) or modification symmetry.

Without serious changes to the standards of ChIP, the limitations of conventional ChIP-seq will continue to pose an existential challenge to the field. Indeed, the divergence between our observations of bivalency and those in the literature can be attributed to the historical lack of tools needed to make quantitative and specific measurements; in that context, the experimental designs and interpretations of the past were reasonable. Fortunately, such tools now exist. And as we have shown in this work, these methods offer a chance for the field to critically evaluate its orthodox models and pave the way for new insights on the chromatin determinants of cell identity and the regulation of development.

3.9 Acknowledgements

We would like to thank Peter Faber, Hannah Whitehurst, and Mikayla Marchuk in the University of Chicago Functional Genomics Facility for Illumina sequencing. We would also like to thank EpiCypher, Inc. for providing some of the histone octamers for this study. A.T.G. was supported by the Harper Dissertation Prize and the Dean's International Student Fellowship of the University of Chicago. R.N.S. was supported by the National Institutes of Health under award number T32-HD007009-45 to the University of Chicago. J. E. was supported by the National Institutes of Health under award number T32-GM007197 and R25-GM109439 to the University of Chicago. This study was supported by the National Institutes of Health, under award numbers R01-GM115945 to A.J.R. and R01-DA036887 to S.K.; and the American Cancer Society, under award number 130230-RSG-16-248-01-DMC to A.J.R.

3.10 Supplementary Notes

Supplementary Note 1

As each nucleosome has two H3 protomers, there are several different configurations of bivalency that a bivalent nucleosome can theoretically adopt, each with a different avidity for ChIP pulldown with immobilized antibody. At one extreme, with the highest avidity, is the symmetric *cis*-bivalency form, where both H3K4 and both H3K27 residues are trimethylated (Figure 3.1e). This nucleosome has the most epitopes for antibody binding and will thus have the highest avidity in pulldown reflected in apical pulldown efficiency (Figure 3.1d). At the other extreme, with the lowest avidity, is the *trans*-bivalency form, where single H3K4me3 and H3K27me3 marks decorate different histone tails (Figure 3.1e). This has the fewest epitopes for antibody binding and will thus have no avidity in pulldown.

This poses a theoretical challenge in normalization and calibration of a ChIP study;

because we cannot separately measure *trans*-bivalency, symmetric *cis*-bivalency, nor any intermediate states, it is impossible for us to definitively state whether a given locus with a given HMD has relatively few nucleosomes that are symmetric *cis*-bivalent or whether it has relatively many nucleosomes that are trans-bivalently modified. To accommodate for this limitation, we include two different bivalent calibrants in our set of nucleosome standards: one that is symmetric *cis*-bivalent and one that is *trans*-bivalent. The bivalency sequential ChIP can then be normalized to either one of these standards, and because these two cases represent the limits of pulldown avidity, normalization to these calibrants will define the theoretical "range" in which true bivalency HMD (i.e. the proportion of nucleosomes with some bivalent configuration) exists (Figure 3.1e). We note that, because the signal from calibration to these standards are scalar multiples of each other, we cannot uniquely distinguish these two configurations in the genome. Absent any prior information about the dominant configuration of bivalency, the proportion of bivalently modified nucleosomes at a given locus will exist in the range defined by calibration to symmetric *cis*- or *trans*-bivalent standards (Figure 3.3a).

In practice, there are a few reasons why this is not a major concern. First, there is no mass spectrometry evidence that H3K4me3 and H3K27me3 exist on the same histone tail, despite specific enrichment for these marks and sensitive detection limits (180; 185), suggesting that configurations other than *trans*-bivalency are at most, extremely minor in abundance. Second, the scarcity of these *cis*-tail modifications is consistent with the bioschemical literature prior to this work that suggests the biogenesis of these *cis*-tail modifications is enzymatically challenging due to antagonistic allosteric effects (see Supplementary Note 4). Third, even if symmetric *cis*- bivalency does exist at some loci, for the purposes of tracking changes in bivalency across differentiation, we can still observe an increase or decrease in bivalency by this calibration method; we simply cannot precisely discern whether the effect is driven by nucleosomes gaining/losing *trans*-bivalency, *cis*-bivalency, or some combination of the two.

The overall amount of bivalency would still increase or decrease in all those scenarios, and so long as our choice of calibrant remains consistent, we can still measure that change regardless of the calibrant that we use for our normalization. Therefore, though we have generated datasets using both calibrants, we present analyses of our reICeChIP bivalency pulldowns calibrated to the trans-bivalent standards.

Supplementary Note 2

Throughout this study, we have defined gene promoters to be the region from 0 to +400bp relative to the TSS, representing the +1 and +2 nucleosomes of each gene. These nucleosomes tend to be well-positioned (210) and, accordingly, are most likely to provide us with adequate read depth to robustly quantify each histone modification. This definition is conservative; we find that H3K4me3 and bivalent domains, which tend to be peak-like, have a median breadth of 550bp at bivalent genes (Figure 3.3d).

The width of these domains raises an important point regarding the measurement of histone modification density as a continuous variable. At a given nucleosome in a single allele of a single cell, there are only three possible states for a histone modification: symmetric, asymmetric or not present. However, nucleosome readers do not typically bind only a single nucleosome at a single position; rather, the local density of the modification across multiple nucleosomes is crucial in localizing these effectors through multivalent avidity-based interactions (211; 212; 213; 214). Indeed, we find that the HMD across sequential nucleosomes relative to the TSS is well autocorrelated (Figure 3.3e). This means that the interpretation of the HMD across a multinucleosomal span becomes more nuanced; a given histone modification may exist at one or more of those nucleosomes. Accordingly, despite the fact that a single nucleosome is essentially ternary in whether it has a given histone modification or not (i.e. HMD of 0% or 100%), a region spanning multiple nucleosomes could have an intermediate HMD; it is this latter quantity that is most relevant for the biological function

imparted to the nearby genomic regions, and this is the quantity we analyse through this work.

Supplementary Note 3

For the datasets presented in this work, the vast majority of promoters have a histone modification density between 0-100%, representing the proportion of nucleosomes at those promoters with the modification of interest (Figure 3.4c; Figure 3.5a). However, at some loci, the measured HMD exceeds 100%. There are several possible reasons for this.

The most important of these possibilities is low input depth. The ICeChIP datasets are normalized to the input read depth at every genomic interval to accommodate for differences in local nucleosome density when computing the HMD. However, this means that at regions that are relatively nucleosome-depleted, there will be few reads in the input, meaning that the denominator of the HMD computation is quite small (Methods). This increased Poisson noise in these regions of low input can result in inflated apparent HMD beyond the physical limit of 100%. To accommodate for this, we can compute 95% confidence intervals for the HMD of each modification at each genomic position, and these confidence intervals virtually always overlap the physically possible range of HMD values (e.g., Figure 3.2c). In naïve mESCs, only 0.5% of the promoters have a bivalency HMD above 100%, and for the vast majority of these promoters (86.1%), the 95% confidence interval error estimate ranges below 100%. The fact the apparent bivalency HMD calibrated by trans-bivalent standards, is broadly constrained to less than 100% further supports the idea that this choice of calibrant is appropriate and not inflationary (Supplementary Note 1).

There are also several other possibilities that are more challenging to accommodate for. First, some regions of the genome are known to be more artefact-prone for sequencing and mapping (215); if the IP sample is enriched for these sequences relative to the input, then that could be disproportionately represented in the IP and have an apparent HMD greater than

100%. Second, the antibodies themselves could skew the apparent HMD. If the antibody is capturing substantial off-target material, then that will result in systematic inflation of the IP, resulting in an inflated HMD. Though ICeChIP barcoded nucleosome standards can help monitor off-target pulldown of some nucleosome species, we can only measure the capture of the standards that we actually have spiked into the experiment. If we do not have nucleosome standards available for a potential off-target modification, then we cannot definitively state that the antibody is not capturing that material. In this context, that is likely most important for H3K27me3 pulldowns; though we cannot state this definitively due to the lack of H3K27me2 standards, it is plausible that we are pulling down some amount of H3K27me2 with these IPs, resulting in slightly inflated apparent H3K27me3 HMD. However, this may not be too problematic; H3K27me2 and H3K27me3 are thought to be recognized by many of the same proteins and to have highly similar functions(197), so the conflation of the two – if present – likely does not pose a significant problem in ascribing biologic function.

On a related note, at some loci, the bivalency HMD goes below 0%. In naïve mESCs, 8.8% of the promoters have a bivalency HMD below 0%, yet for the vast majority of these promoters (90.8%), the 95% confidence interval error estimate ranges above zero. This is because we employ *in silico* signal-correction for the bivalency dataset to remove signal that is attributable to H3K9me3. In essence, we can measure the amount of H3K9me3 pulldown in our bivalency ICeChIP dataset due to nucleosome standards employed, and we can separately measure the H3K9me3 HMD by a highly specific IP. We can then a linear combination correction matrix to remove the signal that is attributable to directly measured H3K9me3 at these loci. This method can effectively reduce the impact of modest off-target binding H3K9me3, but at some loci, will result in a subzero apparent HMD due to random sampling of read depth in the two distinct pulldowns employed.

Finally, at some sets of gene promoters, the trans-bivalency HMD is shown to be greater than the H3K4me3 or H3K27me3 HMD. This apparent discrepancy has a few possible rea-

sons. First, there is some nuance in the interpretation of HMD in the context of single-target ICeChIP and reICeChIP. A nucleosome has two copies of each of its core histone proteins, including histone H3. This means that there are two possible sites of modification on each nucleosome for for each individual modification; if only one of those sites is modified, then that corresponds to an HMD of 50% because only half the possible modification sites are actually modified. However, this is different for the trans-bivalency HMD; by definition, only one trans-bivalency modification pattern can exist on a given nucleosome at any given time. If two "trans-bivalent" modification patterns existed on the same nucleosome simultaneously, then both H3K4 and both H3K27 residues would be trimethylated – which is symmetric *cis*-bivalency. As such, if one H3K4 and one H3K27 residue are trimethylated, then 100% of the possible trans-bivalency configurations for the nucleosome of interest are satisfied, meaning that the trans-bivalency HMD will be 100%. However, in this case, the H3K4me3 and H3K27me3 HMDs will only be 50% because only half the modifiable residues are actually modified.

The other caveat is that symmetrically modified nucleosomes will be pulled down more efficiently than asymmetrically modified nucleosomes due to avidity effects, as can be seen in the pulldown of symmetric vs. asymmetric H3K4me3 and *cis*-bivalency vs. *trans*-bivalency (Figure 3.3), and observed previously (96). This means that calibration to symmetric nucleosome standards will have a larger denominator in computation of HMD and thereby yield lower apparent HMDs; this can also contribute to the lower apparent HMD of H3K4me3 and H3K27me3 relative to trans-bivalency. Accommodating for this phenomenon would require detailed profiling of asymmetric H3K4me3 (which is currently difficult due to the low quality of H3K4me0 antibodies), asymmetric H3K27me3 (which is not currently possible), and distinguishing between trans-bivalency and cis-bivalency (which is also not currently possible). However, as noted in Supplementary Note 1, so long as the method of calibration remains consistent, increases in apparent HMD will still correspond to increases in the modification of

interest. Whether that increase in the target modification is due to asymmetric modification becoming symmetric or due to new gain of the modification at a previously unmodified locus in an instantaneous subpopulation remains unclear, but in both cases, modification density is still being gained at that locus. As such, even with these caveats, we can still quantitatively compare different datasets to each other as we use consistent calibration standards.

Supplementary Note 4

Intriguingly, the catalytic activity of the EZH2-PRC2 core complex on nucleosome substrates is potentiated by pre-existing H3K27me3 (216; 217), yet inhibited by H3K4me3, particularly when symmetric (180; 188; 190). Conversely, symmetric H3K27me3 has been reported to modestly inhibit several of the human COMPASS-family complexes by qualitative assays, although only SET1 complexes were examined at the nucleosome level(191). This presents a potential concern for our data – if the enzyme complexes that install these marks are mutually antagonized by the opposing mark, how might the widespread bivalency we observe arise? As the PRC2 effects are well established with detailed quantitative enzymology(180; 188; 190), which we recapitulate (data not shown), we deployed more quantitative HMTase assays with a larger panel of relevant nucleosomal substrates to evaluate the COMPASS/SET1B/MLL-family core complexes for allosteric modulation by preexisting marks (Figure 3.8).

Supplementary Note 5

In this context, sensitivity refers to the proportion of DEGs that are represented in a specific class of genes (e.g. H3K27me3-dominant bivalent genes), whereas specificity refers to the proportion of that class of genes that are differentially expressed. Under the prevailing bivalency model, bivalency is associated with poised genes that become upregulated or downregulated upon differentiation; as such, it should have high specificity for DEGs.

Supplementary Note 6

The first way we evaluate different models for predicting DEGs is to compute the Bayes Information Criterion (BIC). Though not definitive, this metric estimates whether addition of a parameter to a model improves it more than would be expected from chance alone. When comparing two models, the model with the lower BIC will tend to have more explanatory parameters and/or fewer non-explanatory parameters than the model with the higher BIC. To this end, if BIC increases when a parameter is added, then it can be interpreted that the parameter being added contributes minimal additional explanatory power. Here, we find that adding bivalency to a model increases the BIC, meaning that it is likely (though not definitively) not contributing meaningfully more information in predicting DEG status in this differentiation paradigm.

A more definitive way to evaluate model accuracy is to use hold-out cross-validation. In this method, we split the set of all genes into two groups, one with 80% of the genes (the training set) and one with 20% of the genes (the testing set). We then train our GLMs on the training set and use the derived models to predict DEG status in the testing set. Hold-out cross-validation is a highly effective way of testing whether a model is overfit or underfit upon addition or removal of a parameter. If model accuracy increases substantially, then that would suggest the parameter has explanatory power over that provided by the other parameters. Conversely, if model accuracy decreases substantially, then that suggests that the additional parameter causes overfitting. Minimal changes in model accuracy suggest that the additional parameter contributes little to the model over the existing parameters, positively or negatively.

There are two metrics we use to test the accuracy of the predictions in the testing set. The first is by logistic regression thresholding, in which the gene is predicted to be a DEG if the modelled probability is greater than 0.5. The second is by computing the area under the receiver operator characteristic curve to measure true and false positive rates using different

modelled probabilities as the thresholds.

Overall, we find that the GLM with bivalency barely changes model accuracy by either metric on hold-out cross-validation, with the magnitude of change being similar to that observed by instead adding H3K9me3 or DNA methylation. As such, we can interpret that none of these parameters – including bivalency – meaningfully contributes to the prediction of DEGs beyond what can be achieved with H3K4me3 and H3K27me3 in this system.

3.11 Methods

3.11.1 Cell Culture

Naïve mouse Embryonic Stem Cells (mESCs) were grown from the mESC E14 cell line (129/Ola background)(218) in high glucose DMEM (Invitrogen), supplemented with 15%(v/v) FBS (Gibco), 1%(v/v) non-essential amino acids (Gibco), 1x penicillin/streptomycin (Gibco), 0.1 mM 2-mercaptoethanol (Gibco), 2mM L-glutamine (Gibco), 1000 U/mL LIF (ESG1107 Millipore), 3 μM CHIR99021 (LC Laboratories), 1 μM PD0325901 (LC Laboratories), sterilized using 0.1 μm filter flask (Millipore), stored up to 1 week in 4°C.

Primed mESCs were grown from the mESC E14 cell line (129/Ola background)(218) in high glucose DMEM (Invitrogen), supplemented with 15%(v/v) FBS (Gibco), 1%(v/v) non-essential amino acids (Gibco), 1x penicillin/streptomycin (Gibco), 0.1 mM 2-mercaptoethanol (Gibco), 2mM L-glutamine (Gibco), 1000 U/mL LIF (ESG1107 Millipore), sterilized using 0.1 μm filter flask (Millipore), stored up to 1 week in 4°C.

Naïve and primed mESCs were grown on plates coated with 0.1% bovine gelatin (Sigma), grown to 70-90% confluence and passaged daily at a 1:3 ratio, with a media change 3 hours before passaging, supplemented with 1 vol. of fresh media 8 hours after passaging.

To initiate the adherent monolayer differentiation process to neuronal progenitor cells (NPCs; Day 0)(219; 220), naïve mESCs cells were split onto a gelatinized 6 cm plate at

$1x10^4$ cells/cm2 and allowed to grow for 24 hours. On Day 1, the media was switched to RHB-A (Takara, Y40001) and was subsequently changed every other day. On day 4, cells were split and plated onto Poly-L-Ornithine, laminin-treated 6-cm plates. Prior to cell seeding the plates were treated with 0.01% Poly-L-Ornithine (Millipore, A004C) for at least 20 min, followed by 5 $\mu g/cm^2$ of laminin (Fisher, CB40232) resuspended in basal RHB-A medium (Takara, Y40000). After washing off this treatment, cells were seeded in fresh RHB-A, supplemented with 10 ng/mL of bFGF (PeproTech, 100-18B) and EGF (PeproTech, 315-09). Cells were then split every 4 days at $\geq 20,000 cells/cm^2$ until an appropriate amount of NPCs were cultured for ICeChIP.

3.11.2 Semi-synthetic Histone Preparation

Human histones H3.2(C110A)K4me3, H3.2(C110A)K9me3, H3.2(C110A)K27me3, H3.2(C110A)K4me3-K27me3 were made by semi-synthesis as described previously(96; 221). Asymmetric disulfide linked histone H3K4me3 - H3K27me3 dimers were made by semi-synthesis as described previously(188).

3.11.3 Octamer Reconstitution

Symmetrical H3K4me3, H3K27me3, H3K4me3-K27me3 and H3K9me3 octamers were made as previously described (222; 223). Briefly, equimolar amounts of histone H2A, H2B, H3 and H4 were mixed to the final concentration of 1 mg/ml in unfolding buffer (50 mM Tris-HCl pH 8, 6.3 M Guanidine-HCl, 10 mM 2-mercaptoethanol, 4 mM EDTA), subsequently they were loaded into 3500 M.W.C.O. dialysis tubing (Pierce Snakeskin) and dialyzed in 1000 volumes of refolding buffer (20 mM Tris-HCl pH 7.5, 2 M NaCl, 1mM EDTA, 5 mM DTT), overnight at 4°C. Dialyzed sample was 0.22 μm filtered, and octamers were resolved by S200 gel filtration chromatography (Superdex 200 10/300 GL, GE Healthcare) using refolding buffer as mobile phase. Eluted octamer fractions were pooled and concentrated using centrifugal

filters (Amicon Ultra-4, 10k M.W.C.O., Millipore) to a final concentration of 5-15 μM, diluted with 1 volume of octamer storage buffer (20 mM Tris-HCl pH 7.5, 2 M NaCl, 1 mM EDTA, 5 mM DTT, 55% glycerol), and stored in -20°C. Concentration of octamer was measured spectroscopically using concentrator flow-through as a blank, $\varepsilon_{280nm} = 44700 M^{-1} cm^{-1}$, $M_{oct} \approx 108500 g^1 mol^{-1}$. Octamers were visualized using 18% separating (4% stacking) discontinuous Laemmli SDS-PAGE in Mini-Protean gel running system (Bio-Rad) run for 70 minutes at 22mA, 200V max.

Asymmetrical H3K4me3, H3K27me3 octamers were done as above with the following differences. Equimolar amounts of histone H2A, H2B, H3 and H4 were mixed in unfolding buffer to the total of 1-2 mg, where 90% of histone H3 was trimethylated on Lys 4 or Lys 27 and remaining 10% were unmethylated and had His6-tag at N-terminus with TEV cleavage site. Octamers were reconstituted overnight by dialysis in 1000 volumes of phosphate refolding buffer (50 mM NaxPO4 pH 7.5, 2M NaCl), at 4°C. Octamers were purified by S200 gel filtration chromatography, and his-tagged octamers were isolated using cobalt-based immobilized metal affinity chromatography Dynabeads magnetic particles. Octamers were incubated with magnetic beads for 10 min at 4°C on rotator, followed by two 1 ml washes (50 mM NaxPO4 pH 7.5, 2 M NaCl, 10 mM imidazole), and eluted with 50 ul of elution buffer (50 mM NaxPO4 pH 7.5, 2 M NaCl, 250 mM imidazole, 1 mM EDTA, 1 mM DTT), the elution step was repeated 6 times, fractions were characterized spectroscopically, pooled, diluted with 1 volume of octamer storage buffer, and stored in -20°C.

Asymmetrical *trans*-bivalent H3K4me3-K27me3 octamers were prepared the same way as symmetrical octamers with the following differences. Histones H2A, H2B, H4 and asymmetric disulfide linked histones H3K4me3 - H3K27me3 were mixed 1.2 : 1.2 : 1 : 0.5 ratio. Remaining steps were done as previously described but no reducing agents were used until octamer particles were formed.

All other octamers were obtained from EpiCypher, Inc.

3.11.4 Nucleosome reconstitution

DNA barcodes were constructed based on 601 nucleosome positioning sequence (224). One or both ends of 601 Widom sequence were substituted with 24 bp "barcode" sequence. Each barcode sequence is comprised of two 11 bp sequences absent in human and mouse genome, and constant 2bp linker DNA is added on a free end of the 601 nucleosome positioning sequence.

Nucleosomes were reconstituted as previously described (96). Briefly, 10-100 pmol DNA and histone octamers were mixed in 1 : 1 ratio, at a final concentration >1 μM, and dialyzed in dialysis buttons (Hampton Research) against a non-linear gradient of sodium chloride 2M NaCl $\rightarrow$ 0.2M NaCl in a buffer containing 20 mM Tris-HCl pH 7.5, 1 mM EDTA, 10 mM 2-mercaptoethanol over the course of 12-16 hours (211). Afterwards nucleosomes were recovered, diluted with 1 volume of 2x storage buffer (20 mM Na•Cacodylate pH 7.5, 10% v/v glycerol, 1 mM EDTA, 1x RL Protease Inhibitor Cocktail [1 mM PMSF, 1 mM ABESF, 0.8 μM aprotinin, 20 μM leupeptin, 15 μM pepstatin A, 40 μM bestatin, 15 μM E-64]), and stored at -20°C. Nucleosome concentration was measured by densitometry of 2% agarose gels, 1x TBE (89 mM tris-base, 89 mM boric acid, 2 mM EDTA) run for 30 minutes in 5V/cm electrical field gradient, followed by staining with 1x SYBR Gold (Invitrogen) for >30 minutes. Prior to electrophoresis, nucleosomes were disassembled with 2 M NaCl, roughly 1 pmol of nucleosomes were loaded per well and measured in triplicate against known quantity of free DNA of the same size. For use as ICeChIP standards, the semi-synthetic nucleosomes were diluted to 1 nM concentration using long-term storage buffer (10 mM Na•Cacodylate pH 7.5, 100 mM NaCl, 50% Glycerol, 1 mM EDTA, 1x RL Protease Inhibitor Cocktail, 100 μg/mL BSA(NEB)) and stored at -20°C.

3.11.5 ICeChIP - input preparation

ICeChIP has been done as previously described (96; 186). Briefly, 10^7-10^8 plate adherent cells were released using Accutase (Millipore), quenched with complete medium and collected (500 rcf, 5 min., 4°C). Subsequent steps have been done on ice, cells after each wash were collected by centrifugation (500 rcf, 5 min., 4°C). Cells were washed twice with 10 ml PBS, twice with 5ml buffer N (15 mM Tris pH 7.5, 15 mM NaCl, 60 mM KCl, 8.5% (w/v) Sucrose, 5 mM MgCl2, 1 mM CaCl2 1 mM DTT, 1x RL Protease Inhibitor Cocktail). Cells' membranes was lysed by adding 1 volume of the 2x Lysis Buffer (Buffer N supplemented with 0.6% NP-40 substitute (Sigma)) to the single cell suspension resuspended in 2 PCVs (packed cell volumes) of Buffer N. After 10 minutes incubation on ice, nuclei were collected by centrifugation and resuspended in at least 6 PNV (packed nuclei volumes) of buffer N. Subsequently, nuclei were layered over 7.5 ml Sucrose Cushion N (15 mM Tris pH 7.5, 15 mM NaCl, 60 mM KCl, 30% (w/v) Sucrose, 5 mM MgCl2, 1 mM CaCl2 1 mM DTT, 1x RL Protease Inhibitor Cocktail, 50 μg/mL BSA(NEB)) in a 50 ml centrifuge tube. Nuclei were spun through the sucrose cushion in swinging bucket rotor at 500 rcf for 12 min., 4°C. Nuclei were resuspended in 2 PNVs of buffer N, total nuclei acid content was measured spectroscopically at 260 nm by Nanodrop (Thermo Scientific) (A_{260} = 50 ng/μl), prior to measurement, DNA was stripped from chromatin by adding 18 μl – 98 μl of 2 M NaCl and DNA was fragmented by vortexing and water bath sonication. The quality and quantity of nuclei was measured using a hemocytometer. The apparent concentration of chromatin was adjusted to 1 μg/μl with Buffer N. The following semi-synthetic standards were then spiked in: symmetrical H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K79me2, H3K4me3-K27me3 *cis*-bivalent, asymmetrical H3K4me3, H3K4me3-K27me3 *trans*-bivalent. To fragment the DNA, we aliquoted 100 μg of chromatin and added 1 Worthington unit of Micrococcal nuclease (Worthington) per 4.785 μg of chromatin (measured at 260nm, 1A260 = 50 ng/μl) and incubated at 37°C for 12 minutes. Digestion was stopped by adding 1/10

volume of 11x MNase stop buffer (110 mM EGTA, 110 mM EDTA pH 8.0). Subsequently, nuclei were lysed by slowly adding 5 M NaCl while mixing on a vortex (lowest setting) to the final concentration of 600 mM NaCl. Insoluble material has been spun down at 18000 rcf, 1min., 4°C.

Hydroxyapatite (HAP) chromatography has been done as described previously (96). Briefly, 66 mg of HAP resin (Bio-Rad Macro-Prep® Ceramic Hydroxyapatite Type I 20 μm) was rehydrated with 200 μl of HAP buffer 1 (3.42 mM Na2HPO4 and 1.58 mM NaH2PO4 final pH 7.2, 600 mM NaCl, 1 mM EDTA, 200 μM PMSF). Subsequently, 100 μg of digested soluble chromatin was added to the rehydrated resin and incubated for 10 minutes at 4°C on a rotator. Afterwards, resin slurry was transferred to the centrifugal filter unit (Millipore Ultrafree MC–HV Centrifugal Filter 0.45 μm). Resin was washed 4 times with 200 μl of HAP buffer 1, 4 times with 200 μl of HAP buffer 2 (3.42 mM Na2HPO4 and 1.58 mM NaH2PO4 final pH 7.2, 100 mM NaCl, 1 mM EDTA, 200 μM PMSF), and eluted 3 times with 50 μl HAP elution buffer (342 mM Na2HPO4 and 158 mM NaH2PO4 final pH 7.2, 100 mM NaCl, 1 mM EDTA, 200 μM PMSF), each wash/elution step was accompanied by centrifugation step (600 rcf, 30 sec., 4°C). Concentration of the chromatin was evaluated spectroscopically by Nanodrop (Thermo Scientific) (1A260 = 50 ng/μl). Apparent concentration of the chromatin was adjusted to 20 ng/μl with ChIP buffer 1 with BSA (25 mM Tris pH 7.5, 5 mM MgCl2, 100 mM KCl, 10% (v/v) glycerol, 0.1% (v/v) NP-40 substitute, 100 μg/ml BSA(NEB)).

3.11.6 ICeChIP - immunoprecipitation

ICeChIP was performed as previously described (96) with following modifications: α-H3K4me3 ChIP-5 μg chromatin, 2 μg 304M3B-1xHRV3C (184), 40μl Streptavidin M-280 Dynabeads (Invitrogen). α-H3K9me3 ChIP-3 μg chromatin, 0.5 μg 309M3B (184), 10 μl Streptavidin M-280 Dynabeads (Invitrogen). α-H3K27me3 ChIP - 0.8 μg chromatin, 0.6 μg CST C36B11

lot 8, 5 μl Protein G Dynabeads (Invitrogen). Aforementioned volumes of magnetic beads were washed twice with 200 μl ChIP buffer 1 with BSA. Antibodies were resuspended in 100 μl of ChIP buffer 1 with BSA; subsequently, magnetic beads were collected using magnetic rack, supernatant was removed, and magnetic beads were resuspended with the antibody solution and incubated for at least 1 hr, 4°C, on a rotator. Afterwards unbound antibody was washed away with two 200 μl washes of ChIP buffer 1 with BSA. Streptavidin beads were additionally washed twice with 200 μl of ChIP buffer 1 with BSA supplemented with 5 μM biotin for 10 min, at 4°C, on a rotator for each wash, followed by wash with 200 μl ChIP buffer 1 with BSA. Supernatant was removed on a magnetic rack and specific amount of chromatin, mentioned at the beginning of this chapter, was used to resuspend the magnetic beads. Chromatin was incubated with antibody-beads conjugates for 10-15 minutes, at 4°C, on a rotator.

Subsequently, magnetic beads were washed two times with 200 μl ChIP buffer 2 (25 mM Tris pH 7.5, 5 mM MgCl2, 300 mM KCl, 10% (v/v) glycerol, 0.1% (v/v) NP-40 substitute, 100 μg/ml BSA(NEB)), and one time with 200 μl ChIP buffer 3 (10 mM Tris pH 7.5, 250 mM LiCl, 1 mM EDTA, 0.5% Na•Deoxycholate, 0.5%(v/v) NP-40 substitute, 100 μg/ml BSA(NEB)), 10 minutes, at 4°C, on a rotator with tube change after each wash. These washes were followed by quick 200 μl ChIP buffer 1 (without BSA) wash, and 200μl TE wash (10 mM Tris-HCl pH 8.0, 1 mM EDTA). Chromatin was released from the resin using 50 μl ChIP elution buffer (50 mM Tris pH 7.5, 1 mM EDTA, 1% (w/v) SDS) at 55°C, for 5 minutes. Elution was supplemented with 200 mM NaCl, 10 mM EDTA, 10 μg of Proteinase K (Roche) and incubated at 55°C for 2 hours. DNA was isolated using 3 volumes of Serapure HD (1 mg/ml of 1 μm, hydrophobic, carboxylated, Sera-Mag SpeedBeads (GE 65152105050250), 20% PEG-8000, 2.5 M NaCl, 10 mM Tris pH 7.5, 1 mM EDTA, 0.05% Tween-20, filter sterilized prior to addition of magnetic beads), incubated for 5 minutes at room temperature, collected with magnetic rack and washed twice with >200 μl of 75%

ethanol, on a magnetic rack without disturbing the magnetic beads. Subsequently, ethanol was carefully removed, and DNA was eluted by resuspending beads in 50 μl of TE buffer, magnetic beads were collected using magnetic rack and supernatant was moved to a new tube. In order to limit DNA loss, all operations have been performed using 250 μl siliconized tubes.

3.11.7 reICeChIP

reICeChIP was performed in the same manner as described above with following changes. For the primary IP we have used 5 μg of HAP purified chromatin, 2 μg 304M3B-1xHRV3C – HRV 3C cleavable, biotinylated, αH3K4me3 Fab (PDB:4YHZ) (184), immobilized on 40 μl Streptavidin M-280 Dynabeads (Invitrogen). After 10 minutes incubation of chromatin with antibody-resin conjugate, resin was washed three times with 200 μl ChIP buffer 1 with 100 μg/ml BSA, each wash consisted of 10 minutes incubation at 4°C, on a rotator, followed by tube change. Subsequently, resin was briefly washed with 200 μl ChIP buffer 1 with 100μg/ml BSA, and chromatin was released from the resin with 20 μl of ChIP buffer 1 supplemented with 100 μg/ml BSA and 4 μg HRV3C incubated (GE Healthcare) on ice for 60 minutes, elution step was repeated one more time and both elutions were combined. HRV 3C endoprotease efficiently cleaves at its target sites at 4°C (187) and has a wide range of chemical tolerance (225), but it highly-specific for its cognate cleavage sequence (226) permitting facile use in ChIP under conditions which preserve nucleosomes. Primary elution was added to 0.6 μg CST C36B11, αH3K27me3 mAb, immobilized on 5 μl of Protein G Dynabeads(Invitrogen) for 10 minutes, at 4°C, on a rotator. Subsequently, magnetic beads were washed two times with 200 μl ChIP buffer 2, and one time with 200 μl ChIP buffer 3, 10 minutes, at 4°C, on a rotator, with tube change after each wash. These washes were followed by quick wash with 200 μl ChIP buffer 1 (without BSA). Chromatin was released from the resin by 5 minutes incubation with 50 μl ChIP elution buffer, at 55°C.

Chromatin was Proteinase K digested, and DNA was purified as described in ICeChIP – immunoprecipitation chapter.

3.11.8 Design, expression, and purification of 304M3B-1xHRV3C

304M3B-1xHRV3C Fab is based on previously described Fab 304M3B(PDB:4YHZ)(184). The gene encoding the Fab was modified to contain HRV3C cleavage site at the C-terminus of the heavy chain. To that end, we inserted SSSLEVLFQGP (AGC AGC AGC CTT GAA GTC CTC TTT CAG GGA CCC) sequence just after the position T229 of heavy chain (numbered as in PDB:4YHZ) and before biotinylation acceptor peptide (GLNDIFEAQKIEWHE) (227). The Fab was expressed in the 55244 strain of *E.coli* in the TBG media (Terrific Broth (FisherBrand), 0.8% (v/v) glycerol) with 100 μg/ml carbenicilin, grown for 24 hours, at 30°C, 200 rpm in the Fernbach non-baffled flasks, with constricted airflow. Fab was purified using Protein G-A1 (228) affinity chromatography, followed by cation-exchange chromatography (Resource S, GE Healthcare). Purified Fab was *in vitro* biotinylated using BirA biotin ligase.

3.11.9 Sequencing and Data Analysis

Each sequencing library was made using 10 ng of DNA. Illumina sequencing libraries were made with NEBNext Ultra II DNA Library Prep for Illumina (NEB), according to the manufacturer protocol. DNA libraries were amplified using 8 PCR amplification cycles (C1000, Bio-Rad). Cluster generation and sequencing was performed using the standard Illumina protocols for Illumina HiSeq 4000 by the University of Chicago Functional Genomics Core facility. Data analysis was performed as previously described (96). Briefly, reference genome was modified to contain sequences of semi-synthetic nucleosome barcodes. Reads were mapped to GRCm38/mm10+barcodes reference genome, using Bowtie2 (229), end-to-end, sensitive preset. Subsequently, SAM files were filtered to reject unmapped and

unpaired reads, as well as fragments with length $>$ 200bp and Phred quality score $<$ 20. Paired reads were merged into single interval of the fragment. BEDTools (230) were used to calculate bedgraphs of the genome coverage. IP and input bedgraphs of genome coverage were subsequently merged into multiple entry interval file of genome coverage. Number of DNA fragments coming from semi-synthetic nucleosome standards were counted for each barcode and IP efficiency was calculated for each histone mark.

$$BarcodeIPenrichment = \frac{\sum_1^n IP}{\sum_1^n input}$$

where, n is the identity of the n-th barcode assigned to the specific histone mark.

$$HMD(per/bp) = \frac{\frac{IP}{input} * 100\%}{BarcodeIPenrichment}$$

where, IP and input refer to the depth of the genome coverage at any given position of IP and input for specific histone mark. An IP efficiency can be interpreted as maximal yield of the IP for a given histone mark, or 100% HMD. However, there is a number of factors that can lead to apparent HMD values greater than 100% including: uncertainty due to random sampling of sequenced fragments, in that case standard deviation is approximately equal to square root of sequencing depth, or off-target capture by antibody, where the off-target histone mark or combination of histone marks have greater IP efficiency than the antibody intended target mark.

For all analyses, the HMD averaged over the N+1 and N+2 nucleosomes (taken to be 0 to +400bp into the gene body) was employed as representative of the promoter—this captures the most substantial H3K4me3 and H3K27me3 enrichment.

Genomic browser views were made using IGV. Heatmaps and gene ontology analysis was made using Homer Software (231). Further analysis and sectioning of data was conducted in R using the R code provided in Data Availability. Plots were made using ggplot2 and

Microsoft Excel. For the bivalency tracks, *in silico* antibody off-specificity correction was performed as previously described for H3K9me3 (96). ICeChIP-qPCR was performed with previously described primers (96).

3.11.10 Analysis of External Data

Bisulfite sequencing data was obtained from GEO series accession number GSE41923, dataset accession numbers GSM1027571, GSM1027572, GSM1027573, and GSM1027574. Methylation count files were obtained for each dataset and lifted to mm10. The average methylation for each promoter was then calculated for the 0 to +400bp region relative to the TSS of Refseq promoters using BEDTools.

Bulk RNA-seq data was obtained from GEO series accession numbers GSE108832 and GSE65697, dataset accession numbers GSM2913929, GSM2913930, GSM2913931, GSM1603282, GSM1603283, GSM1603284, GSM1603285, GSM1603286, and GSM1603287. Pseudoalignment was conducted against the Refseq mm10 transcriptome using kallisto (232) with fragment length mean and standard deviation of 200 and 20, respectively, and 100 iterations. Pseudoalignments were then loaded into R for differential expression analysis using sleuth (233), with correction for batch effects between primed mESCs and NPCs due to contribution to principal components of the same. Differentially expressed genes were identified as $q \leq 0.05$. Single-cell RNA-seq data was obtained from GEO series accession number GSE113417 and aligned as above with kallisto.

Suz12 ChIP data to measure PCR2 localization for WT, Ezh2 KO, and Ezh1 KO/Ezh2 KO cells was obtained from GEO series accession number GSE116603, dataset accession numbers GSM3243624, GSM3243625, and GSM3243626. Peak files were obtained for all these datasets lifted to mm10. Ezh2 peaks were identified as peaks lost in Ezh2 KO relative to WT cells. Ezh1 peaks were identified as peaks lost in Ezh1 KO/Ezh2 KO relative to Ezh2 KO cells.

Set1A ChIP data was obtained from GEO series accession number GSE98988, dataset accession numbers GSM2629676, GSM2629677, GSM2629678, and GSM2629691. FastQ files were downloaded for the input and ChIP datasets for each replicate, then aligned to mm10 using Bowtie2 in end-to-end mode with the sensitive preset. Peak calling was then conducted on the alignments with MACS2 (234), and consensus peaks for each replicate were identified.

Mll2 ChIP data was obtained from GEO series accession number GSE78708, dataset accession number GSM2073022. Peaks were obtained and lifted to mm10.

3.11.11 Methyltransferase assays

Enzymatic complexes were procured from Reaction Biology Corporation. Methyltransferase reactions were done using following concentrations of the enzymatic complexes: 200 nM hsMLL1(3745-3969), 200 nM hsMLL2(5319-5537), 400 nM hsMLL3(4689-4911), 200 nM hsMLL4(2490-2715), 800 nM hsSet1A(1418-1707), 800 nM hsSet1B(1629-1923), in a complex with hsWDR5(22-334), haRbBP5(1-538), hsAsh2L(2-534), 2x(hsDPY-30(1-99)), supplemented with 4%(v/v) RBC MLL enhancer (Reaction Biology Corp); 800 nM hsEzh1 (2-747), 120 nM hsEzh2 (2-746), in a complex with hsAEBP2 (2-517), hsEED (2-441), hsRbAp48 (2-425) and hsSUZ12 (2-739) supplemented with 3.6mM hsJarid2 (119-574) provided by Dr. Peter Lewis's laboratory. 30 ng/μl of semi-synthetic nucleosome substrate, 10 μM [^{3}H]-SAM (50-80 Ci/mmol, Perkin Elmer Health Sciences), and enzymatic complexes were mixed in the Reaction Buffer (50 mM Tris pH 8.0, 91 mM NaCl, 5 mM MgCl2, 1 mM DTT, 10% glycerol, 1 mM PMSF) and incubated at 30°C. At designated time points, 4 μl of reactions were spotted on P81 Ion Exchange Cellulose Chromatography Paper (Reaction Biology Corp). Spotted paper was washed 4 times with 250 ml of 50 mM NaHCO3 pH 9.0, for 5 minutes on a platform shaker, briefly washed with acetone, air-dried and immersed in scintillation fluid. 3H decay rate was measured by scintillation counter (LS 6000IC, Beckman).

3.11.12 Data and Software Availability

ICeChIP-seq data generated for this study has been deposited at the Gene Expression Omnibus (GEO) under accession numbers GSE108747 and GSE183155. R markdown file for analysis and sectioning of datasets is provided at
https://www.github.com/shah-rohan/bivalency/.

CHAPTER 4

CONCLUSIONS

In this work, I have discussed multiple mechanisms governing gene expression and ultimately, cellular identity. In Chapter 2, I discuss my work identifying an unannotated cryptic nucleic acid binding domain of YY1, as well as a previously uncharacterized autoregulatory feature of the transcription factor. These characteristics could describe features applicable to more TFs within the proteome. In Chapter 3, I have investigated the coexistence of contrasting chromatin marks, H3K27me3 and H3K4me3, throughout the differentiation scheme of mESCs to NPCs and provided evidence countering the currently accepted bivalency model. With this final chapter I will discuss how my work fits into the field of gene regulation and propose possible routes for further investigation.

4.1 Cryptic nucleic acid binding domains and their possible regulation of nuclear molecular machinery

Essential components within gene regulatory networks are transcription factors, and within this thesis I have investigated the TF, Yin Yang 1 (YY1). Given the highly pleiotropic nature of YY1, it has been difficult to understand the contextual cues that define YY1's function. Previous biochemical dissections of YY1's activation and repression domains have provided a foundational basis of which regions of the protein are responsible for these contrasting transcriptional activities (Figure 2.1). Among the many functionalities I have discussed throughout this thesis, the ability to bind RNA is a relatively new capability attributed to this TF. In the current state of the literature, there is conflict ascribing YY1's RNA binding capabilities to a specific interface (113; 39). One group states that the N-terminal portion of YY1 (AA 1-297) confers this functionality (39), while another study states that the more thoroughly conserved C-terminal ZnF module (AA 293-414) is responsible for this function

(113). My work reconciles this dispute by elucidating that both domains of the protein are capable of binding RNA (Figure 2.5, Figure 2.8). Now, a finer dissection of these RNA binding regions, coupled to *in vivo* experiments, can delineate how RNA binding influences YY1 functionality.

ZnF binding of both DNA and RNA is a characteristic observed for other essential TFs such as TFIIIA (235) and CTCF (76; 77; 236; 237). A proposed function for the dual-binding nature of these transcription factors is the transcription factor trapping model (39). Due to transient non-specific interactions between TFs and RNA molecules, RNA serves as a trapping molecule that can retain a high local concentration of the protein factors necessary for transcriptional firing after a primary act of transcription. Our data not only supports this model for YY1 but also posits the idea of RNA binding acting as a guide for proper genome localization and association due to the REPO-NAB's newly elucidated capability of binding both types of nucleic acids (98; 238). The fact that there are now two possible binding sites for RNA prompts future investigation.

My work posits that there is potential overlap of the DNA- and RNA-binding interfaces within the ZnF module. Previous work used NMR spectroscopy to identify shifted residues in an RNA bound conformation of the ZnF module (113). This group observed the largest chemical shifts within the first two ZnFs of YY1, which coincides with the ZnF knuckle RNA binding domain resident within YY1 (141). However, a notable aspect for the staging of these experiments is that the RNA sequence utilized in these NMR experiments was identified through systematic evolution of ligands by exponential enrichment (SELEX) using a YY1 mutant that compromised the folding of the fourth zinc finger. This could explain their observation of the largest chemical shifts residing within the first two zinc fingers of YY1. Regardless, point mutations of these shifted residues did not ablate RNA binding; therefore, the group mutated 5 residues that constitute a contiguous interface separate from the amino acids involved in DNA binding, hypothesizing that the RNA and DNA binding interfaces

could be mutually exclusive (113). Although mutating this contiguous interface caused an 8-fold drop in RNA affinity, my work counters the notion that these binding interfaces are mutually exclusive. The mutations I instituted in my ZnF3 pentamutant target amino acids engaged in specific base and nonspecific backbone interactions within the central CATT motif of YY1's consensus sequence. We observed diminished binding of both DNA and RNA substrates across equivalent concentration ranges for the WT ZnF module and the ZnF3 mutant (Figure 2.5) indicating that the DNA and RNA binding interfaces probably share a similar interface for binding or that my mutations destablize the proper folding of the zinc fingers. Delineation of the these nucleic acid binding interfaces are critical for further dissection of the contribution that RNA binding has on YY1's overall functionality.

A caveat that requires further investigation is the structural fidelity of my ZnF3 mutant. Circular dichroism and fluorescent thermal shift assays can determine whether I have actually disrupted amino acids that participate in RNA binding or instituted mutations that unfold the ZnF module. Structural biology represents a much more robust and precise avenue for delineation of the RNA binding interface within YY1's ZnFs. Given the high affinity the ZnFs display for a variety of RNA substrates (Figure 2.6), devoid of sequence- and length-specificity, a high resolution structure of this binding interaction would define both the overlapping and exclusive interfaces for dsDNA and ssRNA binding. I hypothesize that shorter substrates (12-22 nt) would primarily interact with the first two ZnFs, as previously observed (113). An intriguing aspect of such a structure will be the extent that ZnFs 3 and 4 play in binding longer ssRNA molecules. We observed tighter associations with longer substrates when we performed our length truncations of the ARID1A RNA (Figure 2.6). Moreover, obtaining a structure of ssRNA and YY1's ZnFs will delineate the role that structural motifs play in ZnF RNA binding. We, and others, (113; 125; 74) have described the lack of sequence specificity for RNA binding but, these observations do not preclude the possibility that a common structural motif resides within these RNA molecules and is responsible for recogni-

tion by the ZnF module. Given the spectrum of RNA folds within our panel of nucleic acids (Figure 2.5, 2.6), and the abundance of CU nucleotides across our panel, I would hypothesize that these RNAs adopt some hairpin structure and that these interactions span from ZnF1 to ZnF3.

A novel insight of my work is the elucidation of the nucleic acid binding capabilities of the REPO-NAB domain. Initially named for the ability to interact with the EED subunit present in both Polycomb Repressive complexes (42; 235; 43; 44), the REPO domain is a prime example of a protein domain that is predicted to be intrinsically disordered yet, adopts a stable conformation when in complex with a protein interaction partner (129). The possibility of nucleic acid binding for this domain could add another layer of regulation dictating YY1 functionality. For insight into how these interactions are modulated in an *in vivo* context and the phenotypic consequences for such perturbations, the mESC degron line described in Appendix A could be a very useful tool. Transient transfection of epitope-tagged mutant versions of YY1 in an environment of WT degradation could answer many inquiries regarding YY1's localization and function. Here are a couple paradigms I would investigate if given the time:

1. Although we initially set out to uncover a discrete RNA binding interface of YY1, we have demonstrated that there are multiple RNA binding regions within the protein. Furthermore, we have discovered that the REPO-NAB domain can bind both types of nucleic acids from our panel, dsDNA and ssRNA (Figure 2.8). The question of: does the REPO-NAB display sequence specificity for its interactions with nucleic acids can now be addressed using our FKBP-degron system. The current state of the transcription factor field posits that the intrinsically ordered domains of TFs aid in genome scanning and orientation of their structured DNA binding domains (97; 101; 238). Now we can ask: how much of WT YY1's localization is influenced by the REPO-NAB's ability to bind nucleic acids?

ChIP and PARCLIP of an epitope-tagged REPO-NAB would provide information on the

sequence specificity and localization for this domain of YY1. Given the similar affinities across our panel of nucleic acids (Figure 2.8 B), I expect a mixture of on-target and off-target localization throughout the genome. Intersection with a variety of datasets (chromatin accessiblity, histone modifying complexes such as PRC1 and PRC2, cell-type specific RNA-seq) would give insight into the governing principles that the REPO-NAB imparts onto WT YY1. I imagine that a large subset of binding sites observed within FLAG-REPO-NAB ChIP or CLIP datasets, would coincide with members of the Polycomb Repressive Complexes. In a scenario in which we see co-occupancy of FLAG-REPO-NAB and EED, or EZH2 subunits, of the PRC2 complex, a possible mechanism that would be supported by my work is the active tethering of chromatin modifying complexes via the REPO-NAB domain. Glutamines 207, 209, and 211 are freely available for nucleic acid binding (Figure 2.8 C), and mutating these residues weakens both dsDNA and ssRNA binding (Figure 2.8 E), though notably, attenuates dsDNA binding more than ssRNA binding. With this in mind, I would be very intrigued in comparing CHIP and PARCLIP datasets between the WT REPO-NAB construct and the glutamine mutant we have designed in my work. I hypothesize that the glutamine mutant would exhibit a higher degree of aberrant localization throughout the genome, or have much less occupancy on chromatin, resulting in lower CHIP signals. Comparison of CLIP data could elucidate RNA molecules of higher affinity for these constructs. I would be most interested in observing whether there is a decrease in PRC2 occupancy when we transiently express the glutamine mutant. Due to its decreased ability to bind dsDNA, would the glutamine mutant be able to tether chromatin modifying complexes to chromatin? This is a functionality typically endowed upon TF domains when in complex with larger molecular machinery and such an defect in PRC2 chromatin occupancy would highlight the importance of the REPO-NAB's dsDNA binding capabilities.

Another aspect that can be further dissected are the key residues responsible for dsDNA and ssRNA binding within the REPO-NAB. I mutated three glutamines to alanines for

our mutant, but combinatorial mutations of these residues could uncover a separation of function mutant in which solely dsDNA is disrupted, and ssRNA binding is maintained. Comparison of this separation of function mutant to the WT REPO-NAB could demonstrate how differences in the ability to bind these nucleic acids affect localization and genomic "scanning" by the REPO-NAB domain. It would also begin to provide insight into the inherent competition that exists for this domain to bind both nucleic acids.

Finally, instituting separation of function mutations of the REPO-NAB into full-length YY1 could demonstate how essential the nucleic acid binding interactions of the REPO-NAB are for the overall localization of the full-length protein. Currently, IDRs are posited to aid in the search for genomic binding sites and proper orientation of the DNA binding domains of TFs. Intrinsic to this hypothesis, when IDRs are disrupted/swapped (238; 97; 101), or the RNA binding capabilities of TFs are compromised (74), there is an increase off-target binding events. When observed with single molecule tracking there is an increase in the population of TFs that adhere to free diffusion kinetics instead of a more confined and directed search for binding sites. Therefore, if a REPO-NAB separation of function mutant ablated its dsDNA capabilities yet maintained its RNA binding capabilities was uncovered, I'd hypothesize that this mutant would maintain subdiffusive properties and generally be tethered to chromatin, scanning genomic loci for on-target or slightly off-target dsDNA binding motifs of YY1. However, since the dsDNA binding capabilities of this mutant would be thoroughly attenuated, I imagine that the frequency of binding non-cognate dsDNA sequences for this mutant would be much higher in comparison to a WT REPO-NAB.

Critical to the context-dependent nature of YY1, protein-protein interactions thoroughly influence how YY1 functions. In regards to the REPO-NAB, an investigation into whether nucleic acid binding induces conformational changes or competes with protein-protein interaction interfaces is addressable in an *in vivo* context as well. *In vitro* pulldown assays between members of the polycomb complex and the REPONAB could be assessed with or

without initial preincubation of the REPONAB with specific nucleic acids. Incorporating insights gained from experiments outlined previously, utilizing a panel of nucleic acids, perhaps those that were enriched within ChIP and PARCLIP datasets, would prove useful in these experiments. Preincubation with these types of nucleic acids could render this binding interface ineligible for protein-protein interactions. Timing plays a critical role in the staging of these experiments. Furthermore, experiments could be performed to query whether nucleic acid binding disrupts Polycomb complex stability by preassembling the REPONAB with Polycomb complex components. Determining whether the complex remains intact in the presence of escalating amounts of nucleic acid could provide insight into feedback loops governing acts of transcription and the recycling of chromatin remodeling complexes.

2. Through our work, we also identified an autoinhibitory mechanism of the N-terminus (AA 1 - 297) on the nucleic acid binding activity of the ZnF module of YY1 (AA 293-414). Due to the nature of our *in vitro* approaches, the magnitude of this autoinhibition in an *in vivo* setting is a query open for investigation. ChIP for a FLAG tagged version of the ZnFs could demonstrate a spectrum of activities due to the spurious binding activities of the ZnF module. Since the ZnF module exhibits a higher affinity for ssRNA molecules than dsDNA molecules, I would expect widespread association throughout the genome and an enrichment of the ZnFs at highly transcribed genes. Coupled to PARCLIP, we could identify the RNA molecules associated with the ZnF module, though I suspect that there would be little to no sequence specificity as our results and others have demonstrated. To test the non-specific nature of RNA binding interactions, a FLP-FRT site could be integrated into the YY1-FKBP tagged mESC cell line which would allow for insertion of a transcriptional unit that is under the control of an inducible dosage-responsive promoter which would allow for varying levels of RNA production at this engineered site. If we observed a positive correlation between ZnF occupancy and RNA production at our transcriptional unit, we could then posit that the N-terminus of YY1 plays a major role in guiding the ZnF module for proper localization and

association with chromatin. This would drastically change the mechanistic understanding of YY1's activity, as this autoinhibitory layer of regulation has not been documented before.

Agnostic of the inhibitory mechanisms we see in an *in vivo* context, determining the interfaces responsible for the *in vitro* inhibition that we observe is another inquiry to pursue. For this investigation, we could implement biophysical approaches such as Nuclear Magnetic Resonance (NMR). Inhibition/modulation of nucleic acid binding activities by disordered regions is a phenomenon that's been previously described in other TFs (239; 240). NMR spectra were used in both of these studies to determine the inter/intra-molecular interactions that regulated key functionalities for master regulator TFs, Pu.1 and p53. For Pu.1, its PEST domain (AA 117-165), which resides within its N-terminal IDR, attenuates the ability for homodimers to bind cognate dsDNA binding sites when a high concentration of Pu.1 is recruited to a cognate binding site. This inhibition results in a *trans* negative feedback loop. For p53, the N-terminal Transactivation Domain (NTAD, AA 1-61) competes with dsDNA substrates for interaction with its DBD. Cognate dsDNA molecules outcompete the NTAD, displacing this autoinhibitory mechanism and inducing a conformational shift in the TF via the displacement of the NTAD. However, when p53 is in an environment with non-cognate dsDNA substrates, the NTAD maintains its inhibitory regulation and attenuates these binding interactions. A critical element of both of these autoregulatory mechanisms is that they are charge-mediated. Both groups varied the electrostatic nature of their *in vitro* experiments, either via phosphomimetic mutations or varying salt concentrations, and observed dampened effects of the regulatory properties they discovered.

It should be noted that these TFs belong to separate TF families than YY1 but, the regulatory implications observed within their IDRs, which are enriched within eukaryotic TFs, provide justification for further interrogation of these regulatory properties. In contrast to the electrostatic-driven interactions of Pu.1 and p53, I mutated a stretch of 11 negatively charged amino acids within YY1's disordered N-terminus to alanines (AA 43-53, Fig 2.11)

as we posited that this negative stretch could interact with positively charged amino acids within the ZnF module (Figure 2.12 A). From my binding experiments, I conclude that this negative stretch of amino acids is dispensable for the autoinhibitory regulation that the N-terminus can impose upon the ZnF module, implying that this inhibitory capability is embedded within the rest of the IDR. Furthermore, by expressing and purifying a truncated N-terminus, (N-terminal IDR, AA 1-163, Figure 2.12) we can conclude that the REPO-NAB domain plays some role in orienting the inhibitory interfaces of the N-terminus for the ZnF module. Or, that the REPO-NAB itself plays a role in this inhibition. Determining what residues of YY1's N-terminus confers this autoinhibitory mechanism would be pivotal to understanding the mechanism of this inter/intra-molecular interaction and could provide insights into generalizable regulatory mechanisms within TFs.

3. Finally, it is clear that a major factor dictating YY1's functionality is its protein interaction partners. Immunoprecipitation followed by mass spectometry could begin to identify the variance in associating protein factors for each of the epitope tagged mutants. Given that intrinsically disordered regions have been attributed as licensing regions responsible for association with other protein complexes which can lead to the coacervation of transcriptional condensates, I believe this system could provide insight into what contextual factors dictate YY1's occupancy and therefore functionality.

These types of experiments can be extended to the wide spectrum of YY1 protein interaction partners, many of which interact with the ZnF module of this protein (241; 242). Delineation of how RNA or DNA binding of both nucleic acid binding domains (the ZnF module and the REPO-NAB) impacts YY1's localization within cells and its protein interaction partners could provide more concrete rules dictating the contextual nature of this pleiotropic and constitutively expressed transcription factor.

4.2 Interpreting the coexistence of H3K4me3 and H3K27me3 chromatin marks throughout the genome

Given the accepted methods for ChIP-seq analyses, the bivalency model provides an intuitive and elegant explanation for the overlap of activating and repressive chromatin marks within undifferentiated cell types. Via the identification and engineering of highly specific antibodies for our chromatin marks of interest, we have developed a quantitative assay capable of assessing the merits of the bivalency model. From the work described in Chapter 4, we obtain results that directly contradict the bivalency model and should thus warrant, not only a re-evaluation of this biological dogma, but also, hopefully, a reassessment of how the field of molecular biology approaches ChIP.

The first notion of the bivalency model states that cell-type specific genes are decorated with activating and repressive marks in order to establish "poised" states of chromatin that will undergo "resolution" through differentiation towards a final transcriptional program. Our results demonstrate that, while lineage-commitment genes do possess both H3K4me3 and H3K27me3 marks, bivalent chromatin is a widespread characteristic, not a phenomenon that solely exists at these deterministic genes. Our results showcase that H3K27me3 is pervasively found throughout the genome, albeit at lower HMD values than canonically accepted enriched loci. This quantitative observation was only made possible due to the rigorous testing of specific antibodies for our chromatin marks and the utilization of our semi-synthetic nucleosome standards. In characteristic ChIP-seq analyses, true-postivie loci containing low levels of H3K27me3 would be disregarded due to peak calling algorithms which solely look for relatively high levels of the modification. Our approach is able to specifically capture lower levels of the mark and provide confidence that this modification is installed at the loci we are identifying. Given this data, we have provided evidence questioning the first aspect of the bivalency model.

However, the pervasive nature of H3K27me3 deposition raises a separate question. Why

is most of the genome covered with this repressive mark? Due to the widespread incorporation of exogenous transposable elements within the human genome (3; 4), I interpret this deposition as a buffering mechanism to prevent spurious transcriptional activation. While previous studies have utilized knockout systems to disrupt key chromatin modifying enzymes, these approaches induce pleiotropic effects that extend beyond their roles in bivalency (143; 194; 199; 198). To address how inhibition of the PRC2 complex, the sole installer of H3K27me3, would affect bivalent chromatin dynamics, I would use a specific inhibitor towards the EED componenet of PRC2 at different timepoints throughout our differentiation process (243). This inhibitor was observed to bind to the same binding pocket as an H3K27me peptide, the activating substrate that promotes PRC2's catalytic activity (244). This drug also demonstrated high selectivity for PRC2, exhibiting a >2000 fold increase in IC50 concentration for 21 other methyltransferases. When immobilized on streptavidin beads and exposed to Karpas422 cell lysate, the only proteins that bound to this drug were EZH2, SUZ12, EED, and a known PRC2 partner, MTF2. This selectivity reduces the impact of pleiotropic effects that previous KO experiments have utilized and allows precise temporal control of PRC2 inhibition. Indeed, upon inhibition of H3K27me3 deposition, global changes can be seen in both gene expression and H3K27me3 occupancy (243). My only addendum to this study would be to also look at the production of ncRNA to gain insight into the spurious transcriptional activity that could emerge in the noncoding genome.

While ICeChIP addresses many of the drawbacks inherent to ChIP, there are still caveats to our population-wide methodology which deserve further dissection. Although ICeChIP is quantitative in its nature, HMD cannot distinguish between allelic or population-wide distribution of our chromatin marks of interest. When we make the claim that there is 50% HMD at a specific locus for H3K4me3, we could be describing that 1 of 2 alleles contain this mark within every single cell of a given population, or that 50% of cells within a population have this modification installed at both of their alleles. While ICeChIP allows us to make

claims on the quantitative levels of histone modification within a population, these two models represent fundamentally different versions of reality. Distinguishing between these two possibilities would provide insightful biological implications, as both models suggest differing scales of cell or tissue regulation. Single-cell ChIP-seq would begin to answer such a question and protocols have been published (245; 246). In general, single-cell protocols have gained traction as biologists are publishing single-cell atlases across a spectrum of biological systems (247). Be that as it may, conflicts reside within the protocols utilized to analyze such massive datasets, and therefore the reproducibility and interpretation of such endeavors are called into question (248). In regards to single-cell ChIP, the genomic coverage and read depth is insufficient to achieve the mononucleosome resolution represented in bulk ChIP-seq protcols, and with the lack of internal standards the drawbacks of conventional ChIP resurface. Perhaps in the future a technology will be developed to satisfy this type of investigation but, until then, we shall remain in the dark as to which model represents truth.

Finally, our work in Chapter 4 can only comment on the bivalency model in the context of mESC differentiation to NPCs. Although this was the seminal differentiation pathway that established the bivalency model (92; 91), other differentiation paths or differentiation into a more terminal cell state could possibly follow the mechanisms proposed by the bivalency model. In a similar regard, we also only investigated the coexistence of H3K27me3 and H3k4me3 within our work. A wonderful aspect of reICeChIP-seq is the adaptability of the protocol. As long as specific antibodies for the moieties of interest are utilized, introduction and engineering of the HRV 3C protease cleavage site can facilitate investigation into whichever marks of interest or differentiation programs investigators choose to look into.

4.3 Significance

In this dissertation I have delved into two specific niches of regulatory components that maintain the fidelity of transcriptional activation and repression. I have uncovered novel

regulatory features for the "housekeeping" TF Yin Yang 1, which shall hopefully provide a basis for further investigation into its context-dependent functionality and add to the lexicon of regulatory properties for TFs as a whole. I have also discussed the pitfalls of conventional ChIP-seq, and have presented work using a methodology that we have developed (ICeChIP and ReICeChIP) to re-evaluate the strength of the bivalency model. I believe this work provides outlines for how to mechanistically approach investigations of basic biology and highlights how, even in this era of massive dataset production, a rigorous look at the finer details of molecular machinery is essential for the formation of proper biological conclusions.

APPENDIX A

GENERATION OF AN E14 MESC YY1 DEGRON LINE

In Chapter 2 of this thesis, I describe my work in generating YY1 truncation and substitution mutants to delineate its RNA-binding interface. Defining such an interface would allow us to generate an RNA-binding deficient mutant of YY1 which we could then utilize within a mammalian cell culture system. Expression of this RNA binding deficient mutant (or of any of the mutants I've generated) in cells would allow us to start assessing how RNA binding influences YY1's genomic occupancy and modulate YY1's protein interactome. In order to produce data that are not confounded by the presence of WT endogenous YY1, I generated an E14 mESC YY1-degron line, in which both copies of YY1 are tagged with an FKBP tag for rapid and temporally controlled protein degradation using the CRISPR/Cas9 genome editing system. This system was previously utilized to generate a V6.5 mESC line (37), and our protocol closely follows theirs.

Briefly, five hundred thousand E14 mESCs were transfected with 2.5 μg of Cas9-GFP plasmid and 1.25 μg of repair plasmid 1 (pAW62.YY1.FKBP.knock-in.mCherry) and 1.25 μg of repair plasmid 2 (pAW63.YY1.FKBP.knock-in.BFP). Both plasmids were ordered from Addgene. Cells were allowed to expand for 48 hours, and were then sorted for GFP expression on an Invitrogen Bigfoot Cell Sorter at the University of Chicago Cytometry and Antibody Technology Core Facility (Figure A.1).

Collected cells were expanded for 5 days and then sorted for BFP and mCherry expression, indicating that both alleles of YY1 had been properly tagged. Cells were sorted into single cell populations within 96-well plates and allowed to expand until mESC colonies began to form (Figure A.2).

Upon stable expansion into 6-well plates, genomic DNA and protein lysates were extracted from multiple monoclonal populations and assayed for genomic integration of the FKBP tag and targeted degradation of endogenous YY1 after cell populations were treated

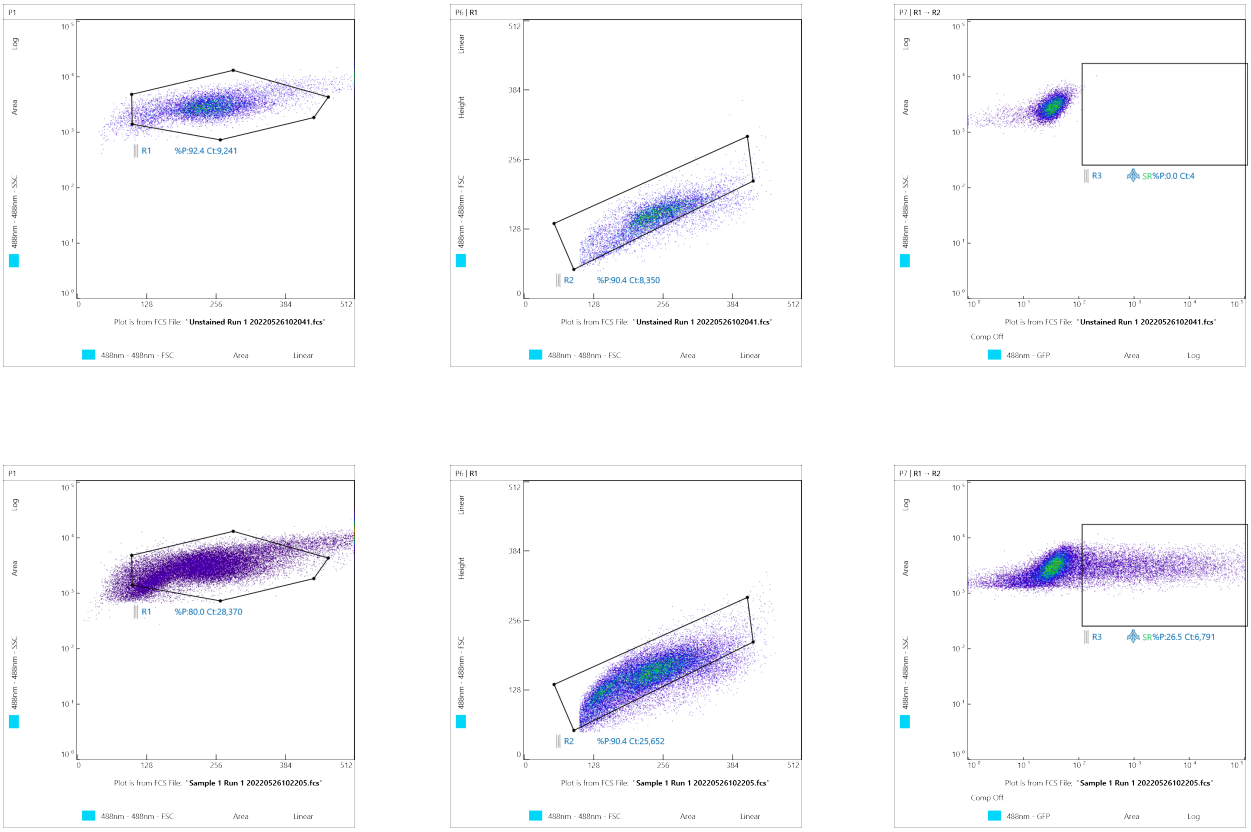


Figure A.1: FACS sorting for Cas9-eGFP expression. (Top) WT mESCs that have not undergone CRISPR-Cas9 genome editing. (Bottom) Population of cells sorted for expansion due to expression of Cas9-eGFP

with 500 nM of dTAG-47 for 24 hours (Figure A.3).

From this protocol, clone M1 showed the most promise in being a monoclonal mESC line that has both alleles of YY1 tagged for controlled degradation. I hope that this cell line can be used as a tool to investigate the scenarios I outlined previously.

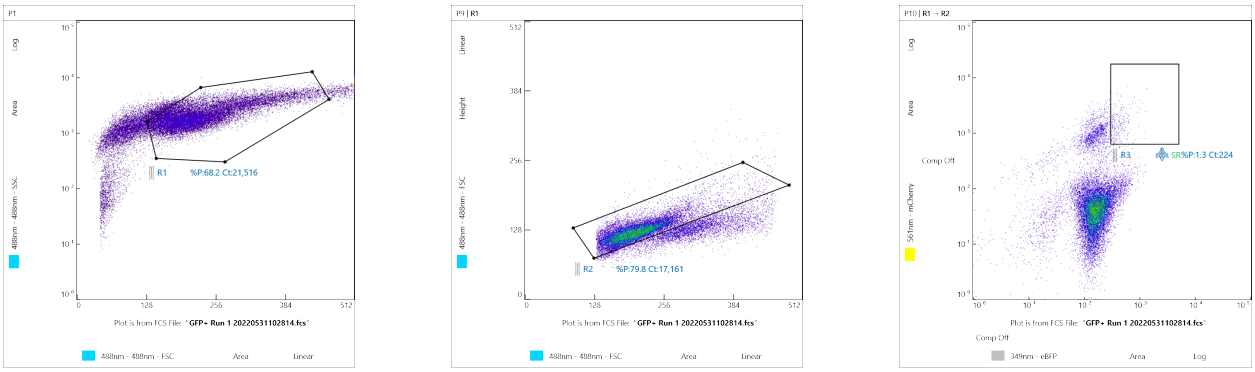


Figure A.2: FACS sorting for mCherry+ and BFP+ positive cells.

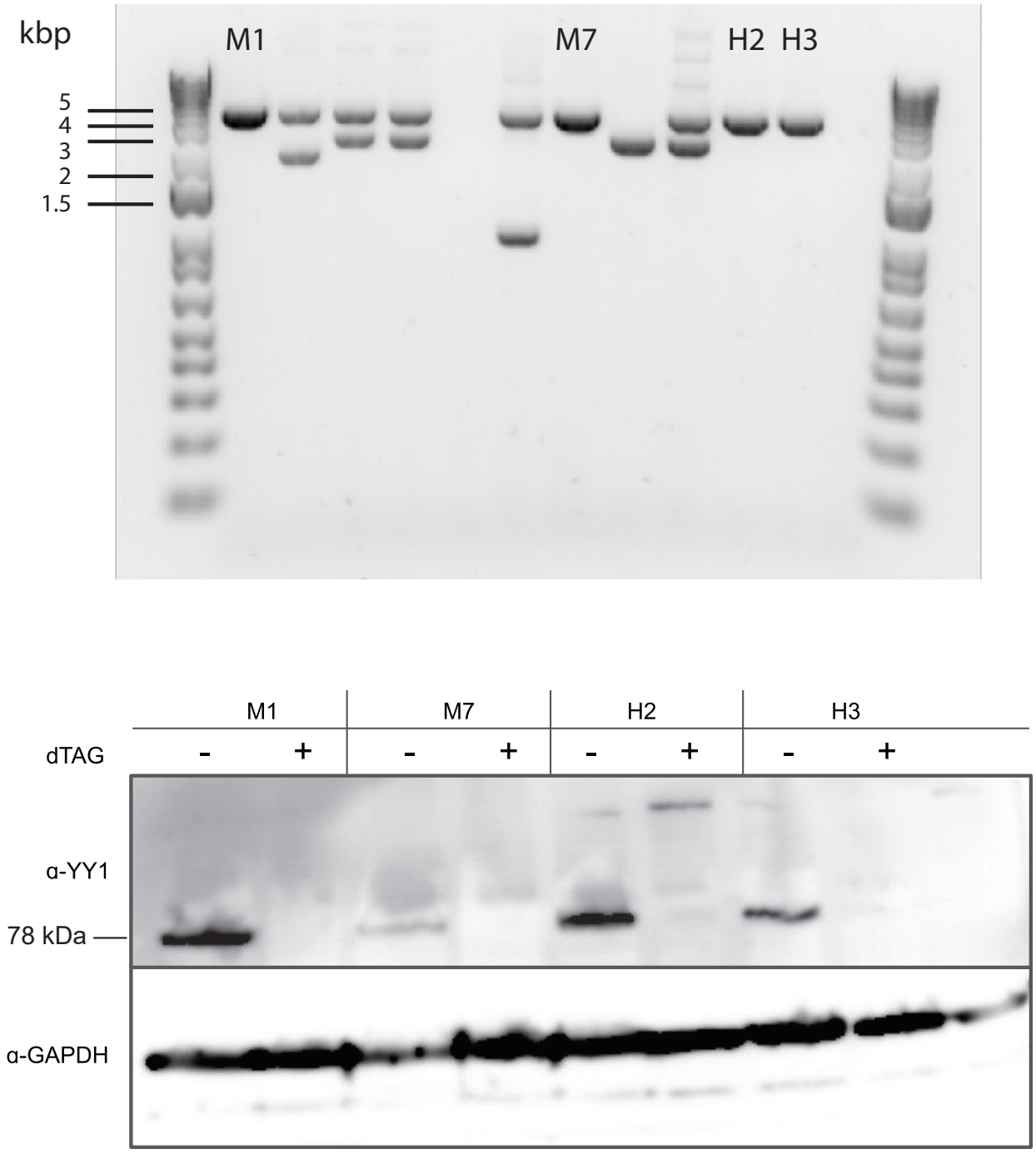


Figure A.3: PCR and Western blot indicating homozygous tagging of endogenous YY1 with the FKBP tag for inducible protein degradation. (Top) Genomic DNA amplification of C-terminal region of YY1. WT band size $\tilde{2}$508 bp. FKBP Knock-in band size $\tilde{4}$kbp

REFERENCES

[1] A. Uzman, "Molecular biology of the cell (4th ed.): Alberts, b., johnson, a., lewis, j., raff, m., roberts, k., and walter, p.," *Biochemistry and Molecular Biology Education*, vol. 31, p. 212–214, July 2003.

[2] A. Bolzer, G. Kreth, I. Solovei, D. Koehler, K. Saracoglu, C. Fauth, S. Müller, R. Eils, C. Cremer, M. R. Speicher, and T. Cremer, "Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes," *PLoS Biology*, vol. 3, p. e157, Apr. 2005.

[3] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.-F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H.-C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. R. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. A. Smit, E. Stupka,

J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S.-P. Yang, R.-F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, A. Patrinos, M. J. Morgan, International Human Genome Sequencing Consortium, C. f. G. R. Whitehead Institute for Biomedical Research, The Sanger Centre:, Washington University Genome Sequencing Center, US DOE Joint Genome Institute:, Baylor College of Medicine Human Genome Sequencing Center:, RIKEN Genomic Sciences Center:, Genoscope and CNRS UMR-8030:, I. o. M. B. Department of Genome Analysis, GTC Sequencing Center:, Beijing Genomics Institute/Human Genome Center:, T. I. f. S. B. Multimegabase Sequencing Center, Stanford Genome Technology Center:, University of Oklahoma's Advanced Center for Genome Technology:, Max Planck Institute for Molecular Genetics:, L. A. H. G. C. Cold Spring Harbor Laboratory, GBF—German Research Centre for Biotechnology:, a. i. i. l. u. o. h. *Genome Analysis Group (listed in alphabetical order, U. N. I. o. H. Scientific management: National Human Genome Research Institute, Stanford Human Genome Center:, University of Washington Genome Center:, K. U. S. o. M. Department of Molecular Biology, University of Texas Southwestern Medical Center at Dallas:, U. D. o. E. Office of Science, and The Wellcome Trust:, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, feb 2001. Publisher: Nature Publishing Group.

[4] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M.-H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy,

L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.-H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.-H. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu, "The Sequence of the Human Genome," *Science*, vol. 291, pp. 1304–1351, Feb. 2001. Publisher: American Association for the Advancement of Science.

[5] S. Boyle, "The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells," *Human Molecular Genetics*, vol. 10, p. 211–219, Feb. 2001.

[6] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science*, vol. 326, p. 289–293, Oct. 2009.

[7] F. Jacob and J. Monod, "Genetic regulatory mechanisms in the synthesis of proteins," *Journal of Molecular Biology*, 1961.

[8] D. Hnisz, B. J. Abraham, T. I. Lee, A. Lau, V. Saint-André, A. A. Sigova, H. A. Hoke, and R. A. Young, "Super-enhancers in the control of cell identity and disease," *Cell*, vol. 155, no. 4, p. 934, 2013. ISBN: 0092-8674.

[9] S. Rao, M. Huntley, N. Durand, E. Stamenova, I. Bochkov, J. Robinson, A. Sanborn, I. Machol, A. Omer, E. Lander, and E. Aiden, "A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, vol. 159, p. 1665–1680, Dec. 2014.

[10] A. L. Sanborn, B. T. Yeh, J. T. Feigerle, C. V. Hao, R. J. Townshend, E. L. Aiden, R. O. Dror, and R. D. Kornberg, "Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to mediator," *eLife*, vol. 10, pp. 1–42, 2021.

[11] S. Lindquist, "Regulation of protein synthesis during heat shock," *Nature*, vol. 293, p. 311–314, Sept. 1981.

[12] A. Tissiéres, H. K. Mitchell, and U. M. Tracy, "Protein synthesis in salivary glands of drosophila melanogaster: Relation to chromosome puffs," *Journal of Molecular Biology*, vol. 84, p. 389–398, Apr. 1974.

[13] A. Ali, S. Bharadwaj, R. O'Carroll, and N. Ovsenek, "Hsp90 interacts with and regulates the activity of heat shock factor 1 in xenopus oocytes," *Molecular and Cellular Biology*, vol. 18, p. 4949–4960, Sept. 1998.

[14] S. Chowdhary, A. S. Kainth, and D. S. Gross, "Heat shock protein genes undergo dynamic alteration in their three-dimensional structure and genome organization in response to thermal stress," *Molecular and Cellular Biology*, vol. 37, Dec. 2017.

[15] M. Du, S. H. Stitzinger, J.-H. Spille, W.-K. Cho, C. Lee, M. Hijaz, A. Quintana, and I. I. Cissé, "Direct observation of a condensate effect on super-enhancer controlled gene bursting," *Cell*, vol. 187, pp. 331–344.e17, Jan. 2024.

[16] W.-K. Cho, J.-H. Spille, M. Hecht, C. Lee, C. Li, V. Grube, and I. I. Cisse, "Mediator and rna polymerase ii clusters associate in transcription-dependent condensates," *Science*, vol. 361, p. 412–415, July 2018.

[17] W. A. Whyte, D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, and R. A. Young, "Master transcription factors and mediator establish super-enhancers at key cell identity genes," *Cell*, vol. 153, no. 2, pp. 307–319, 2013. Publisher: Elsevier Inc.

[18] K. Takahashi and S. Yamanaka, "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors," *Cell*, vol. 126, pp. 663–676, Aug. 2006.

[19] S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch, "The Human Transcription Factors," *Cell*, vol. 172, no. 4, pp. 650–665, 2018. Publisher: Elsevier Inc.

[20] P. Cramer, "Organization and regulation of gene transcription," *Nature*, 2019. Publisher: Springer US.

[21] T. Matsui, J. Segall, P. Weil, and R. Roeder, "Multiple factors required for accurate initiation of transcription by purified rna polymerase ii.," *Journal of Biological Chemistry*, vol. 255, p. 11992–11996, Dec. 1980.

[22] J. D. Parvin, H. Timmers, and P. A. Sharp, "Promoter specificity of basal transcription factors," *Cell*, vol. 68, pp. 1135–1144, Mar. 1992.

[23] J. Tycko, N. DelRosso, G. T. Hess, Aradhana, A. Banerjee, A. Mukund, M. V. Van, B. K. Ego, D. Yao, K. Spees, P. Suzuki, G. K. Marinov, A. Kundaje, M. C. Bassik, and L. Bintu, "High-throughput discovery and characterization of human transcriptional effectors," *Cell*, vol. 183, pp. 2020–2035.e16, Dec. 2020.

[24] J. Tycko, M. V. Van, Aradhana, N. DelRosso, H. Ye, D. Yao, R. Valbuena, A. Vaughan-Jackson, X. Xu, C. Ludwig, K. Spees, K. Liu, M. Gu, V. Khare, A. X. Mukund, P. H. Suzuki, S. Arana, C. Zhang, P. P. Du, T. S. Ornstein, G. T. Hess, R. A. Kamber, L. S. Qi, A. S. Khalil, L. Bintu, and M. C. Bassik, "Development of compact transcriptional effectors using high-throughput measurements in diverse contexts," *Nature Biotechnology*, Nov. 2024.

[25] N. DelRosso, J. Tycko, P. Suzuki, C. Andrews, Aradhana, A. Mukund, I. Liongson, C. Ludwig, K. Spees, P. Fordyce, M. C. Bassik, and L. Bintu, "Large-scale mapping and mutagenesis of human transcriptional effector domains," *Nature*, vol. 616, p. 365–372, Apr. 2023.

[26] N. D. Tippens, J. Liang, A. K.-Y. Leung, S. D. Wierbowski, A. Ozer, J. G. Booth, J. T. Lis, and H. Yu, "Transcription imparts architecture, function and logic to enhancer units," *Nature Genetics*, vol. 52, p. 1067–1075, Sept. 2020.

[27] J. Zuin, G. Roth, Y. Zhan, J. Cramard, J. Redolfi, E. Piskadlo, P. Mach, M. Kryzhanovska, G. Tihanyi, H. Kohler, M. Eder, C. Leemans, B. van Steensel, P. Meister, S. Smallwood, and L. Giorgetti, "Nonlinear control of transcription through enhancer–promoter interactions," *Nature*, vol. 604, p. 571–577, Apr. 2022.

[28] R. Daber, S. Stayrook, A. Rosenberg, and M. Lewis, "Structural Analysis of Lac Repressor Bound to Allosteric Effectors," *Journal of Molecular Biology*, vol. 370, pp. 609–619, July 2007.

[29] J. L. Brown, D. Mucci, M. Whiteley, M.-L. Dirksen, and J. A. Kassis, "The Drosophila Polycomb Group Gene pleiohomeotic Encodes a DNA Binding Protein with Homology to the Transcription Factor YY1," *Molecular Cell*, vol. 1, pp. 1057–1064, June 1998. Publisher: Elsevier.

[30] Y. Shi, E. Seto, L. S. Chang, and T. Shenk, "Transcriptional repression by YY1, a human GLI-Krüppel-related protein, and relief of repression by adenovirus E1A protein," *Cell*, vol. 67, no. 2, pp. 377–388, 1991.

[31] N. Hariharan, D. E. Kelley, and R. P. Perry, "Delta, a transcription factor that binds to downstream elements in several polymerase II promoters, is a functionally versatile zinc finger protein.," *Proceedings of the National Academy of Sciences*, vol. 88, pp. 9799–9803, Nov. 1991. Publisher: Proceedings of the National Academy of Sciences.

[32] J. R. Flanagan, K. G. Becker, D. L. Ennist, S. L. Gleason, P. H. Driggers, B. Z. Levi, E. Appella, and K. Ozato, "Cloning of a negative transcription factor that binds to the

upstream conserved region of Moloney murine leukemia virus.," *Mol Cell Biol*, vol. 12, pp. 38–44, Jan. 1992.

[33] M. E. Donohoe, X. Zhang, L. McGinnis, J. Biggers, E. Li, and Y. Shi, "Targeted Disruption of Mouse Yin Yang 1 Transcription Factor Results in Peri-Implantation Lethality," *Molecular and Cellular Biology*, vol. 19, no. 10, pp. 7237–7244, 1999.

[34] E. B. Affar, F. Gay, Y. Shi, H. Liu, M. Huarte, S. Wu, T. Collins, E. Li, and Y. Shi, "Essential Dosage-Dependent Functions of the Transcription Factor Yin Yang 1 in Late Embryonic Development and Cell Cycle Progression," *Molecular and Cellular Biology*, vol. 26, no. 9, pp. 3565–3581, 2006.

[35] M. Gabriele, A. T. V.-v. Silfhout, P.-L. Germain, A. Vitriolo, R. Kumar, E. Douglas, E. Haan, K. Kosaki, T. Takenouchi, A. Rauch, K. Steindl, E. Frengen, D. Misceo, C. R. J. Pedurupillay, P. Stromme, J. A. Rosenfeld, Y. Shao, W. J. Craigen, C. P. Schaaf, D. Rodriguez-Buritica, L. Farach, J. Friedman, P. Thulin, S. D. McLean, K. M. Nugent, J. Morton, J. Nicholl, J. Andrieux, A. Stray-Pedersen, P. Chambon, S. Patrier, S. A. Lynch, S. Kjaergaard, P. M. Tørring, C. Brasch-Andersen, A. Ronan, A. v. Haeringen, P. J. Anderson, Z. Powis, H. G. Brunner, R. Pfundt, J. H. M. Schuurs-Hoeijmakers, B. W. M. v. Bon, S. Lelieveld, C. Gilissen, W. M. Nillesen, L. E. L. M. Vissers, J. Gecz, D. A. Koolen, G. Testa, and B. B. A. d. Vries, "YY1 Haploinsufficiency Causes an Intellectual Disability Syndrome Featuring Transcriptional and Chromatin Dysfunction," *The American Journal of Human Genetics*, vol. 100, pp. 907–925, June 2017. Publisher: Elsevier.

[36] A. Usheva and T. Shenk, "Yy1 transcriptional initiator: Protein interactions and association with a dna site containing unpaired strands," *Proceedings of the National Academy of Sciences*, vol. 93, p. 13571–13576, Nov. 1996.

[37] A. S. Weintraub, C. H. Li, A. V. Zamudio, A. A. Sigova, N. M. Hannett, D. S. Day, B. J. Abraham, M. A. Cohen, B. Nabet, D. L. Buckley, Y. E. Guo, D. Hnisz, R. Jaenisch, J. E. Bradner, N. S. Gray, and R. A. Young, "YY1 Is a Structural Regulator of Enhancer-Promoter Loops," *Cell*, vol. 171, pp. 1573–1588.e28, Dec. 2017. Publisher: Elsevier.

[38] J. Wang, X. Wu, C. Wei, X. Huang, Q. Ma, X. Huang, F. Faiola, D. Guallar, M. Fidalgo, T. Huang, D. Peng, L. Chen, H. Yu, X. Li, J. Sun, X. Liu, X. Cai, X. Chen, L. Wang, J. Ren, J. Wang, and J. Ding, "YY1 Positively Regulates Transcription by Targeting Promoters and Super-Enhancers through the BAF Complex in Embryonic Stem Cells," *Stem Cell Reports*, vol. 10, no. 4, pp. 1324–1339, 2018. Publisher: ElsevierCompany.

[39] A. A. Sigova, B. J. Abraham, X. Ji, B. Mollinie, N. M. Hannett, Y. E. Guo, M. Jangi, C. C. Giallourakis, P. A. Sharp, and R. A. Young, "Transcription factor trapping by RNA in gene regulatory elements," *Science*, vol. 350, no. 6263, pp. 978–982, 2015.

[40] Q. Zhou, R. W. Gedrich, and D. A. Engel, "Transcriptional repression of the c-fos gene by yy1 is mediated by a direct interaction with atf/creb," *Journal of Virology*, vol. 69, p. 4323–4330, July 1995.

[41] W. MacLellan, T. Lee, R. Schwartz, and M. Schneider, "Transforming growth factor-beta response elements of the skeletal alpha-actin gene. combinatorial action of serum response factor, yy1, and the sv40 enhancer-binding protein, tef-1," *Journal of Biological Chemistry*, vol. 269, p. 16754–16760, June 1994.

[42] F. H. Wilkinson, K. Park, and M. L. Atchison, "Polycomb recruitment to DNA in vivo by the YY1 REPO domain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 51, pp. 19296–19301, 2006.

[43] L. Atchison, A. Ghias, F. Wilkinson, N. Bonini, and M. L. Atchison, "Transcription factor YY1 functions as a PcG protein in vivo," *EMBO J*, vol. 22, pp. 1347–1358, Mar. 2003.

[44] D. P. E. Satijn, K. M. Hamer, J. den Blaauwen, and A. P. Otte, "The Polycomb Group Protein EED Interacts with YY1, and Both Proteins Induce Neural Tissue in XenopusEmbryos," *Molecular and Cellular Biology*, vol. 21, pp. 1360–1369, Feb. 2001. Publisher: Taylor & Francis _eprint: https://doi.org/10.1128/MCB.21.4.1360-1369.2001.

[45] M. J. Solomon, P. L. Larsen, and A. Varshavsky, "Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene," *Cell*, vol. 53, pp. 937–947, June 1988. Publisher: Elsevier.

[46] H. S. Rhee and B. F. Pugh, "ChIP-exo Method for Identifying Genomic Location of DNA-Binding Proteins with Near-Single-Nucleotide Accuracy," *Current Protocols in Molecular Biology*, vol. 100, no. 1, pp. 21.24.1–21.24.14, 2012. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471142727.mb2124s100.

[47] P. J. Skene and S. Henikoff, "An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites," *eLife*, vol. 6, p. e21856, Jan. 2017. Publisher: eLife Sciences Publications, Ltd.

[48] A. Yang, Z. Zhu, P. Kapranov, F. McKeon, G. M. Church, T. R. Gingeras, and K. Struhl, "Relationships between p63 binding, dna sequence, transcription activity, and biological function in human cells," *Molecular Cell*, vol. 24, p. 593–602, Nov. 2006.

[49] J. Wang, J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder, and Z. Weng, "Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors," *Genome Research*, vol. 22, p. 1798–1812, Sept. 2012.

[50] M. J. Hangauer, I. W. Vaughn, and M. T. McManus, "Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs," *PLoS Genetics*, vol. 9, no. 6, 2013. ISBN: 1553-7404 (Electronic)\r1553-7390 (Linking).

[51] J. Brosius, T. J. Dull, and H. F. Noller, "Complete nucleotide sequence of a 23s ribosomal rna gene from escherichia coli.," *Proceedings of the National Academy of Sciences*, vol. 77, p. 201–204, Jan. 1980.

[52] J. Brosius, M. L. Palmer, P. J. Kennedy, and H. F. Noller, "Complete nucleotide sequence of a 16s ribosomal rna gene from escherichia coli.," *Proceedings of the National Academy of Sciences*, vol. 75, p. 4801–4805, Oct. 1978.

[53] J. A. STEITZ, "Polypeptide chain initiation: Nucleotide sequences of the three ribosomal binding sites in bacteriophage r17 rna," *Nature*, vol. 224, p. 957–964, Dec. 1969.

[54] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl, "Identification of novel genes coding for small expressed rnas," *Science*, vol. 294, p. 853–858, Oct. 2001.

[55] R. C. Lee and V. Ambros, "An extensive class of small rnas in caenorhabditis elegans," *Science*, vol. 294, p. 862–864, Oct. 2001.

[56] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14," *Cell*, vol. 75, p. 843–854, Dec. 1993.

[57] H.-C. Lee, W. Gu, M. Shirayama, E. Youngman, D. Conte, and C. Mello, "C.elegans pirnas mediate the genome-wide surveillance of germline transcripts," *Cell*, vol. 150, p. 78–87, July 2012.

[58] A. G. Matera and Z. Wang, "A day in the life of the spliceosome," *Nature Reviews Molecular Cell Biology*, vol. 15, p. 108–121, Jan. 2014.

[59] M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn, and E. S. Lander, "Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals," *Nature*, vol. 458, p. 223–227, Feb. 2009.

[60] M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn, "Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses," *Genes amp; Development*, vol. 25, p. 1915–1927, Sept. 2011.

[61] Y. Jeon and J. T. Lee, "YY1 Tethers Xist RNA to the inactive X nucleation center," *Cell*, vol. 146, no. 1, pp. 119–133, 2011. Publisher: Elsevier Inc.

[62] M. S. Werner and A. J. Ruthenburg, "Nuclear Fractionation Reveals Thousands of Chromatin-Tethered Noncoding RNAs Adjacent to Active Genes," *Cell Reports*, vol. 12, no. 7, pp. 1089–1098, 2015. Publisher: The Authors.

[63] M. S. Werner, M. A. Sullivan, R. N. Shah, R. D. Nadadur, A. T. Grzybowski, V. Galat, I. P. Moskowitz, and A. J. Ruthenburg, "Chromatin-enriched lncRNAs can act as cell-type specific activators of proximal gene transcription," *Nature Structural and Molecular Biology*, vol. 24, no. 7, pp. 596–603, 2017. arXiv: 15334406 Publisher: Nature Publishing Group ISBN: 1545-9985 (Electronic) 1545-9985 (Linking).

[64] X. H. Yang, R. D. Nadadur, C. R. Hilvering, V. Bianchi, M. Werner, S. R. Mazurek, M. Gadek, K. M. Shen, J. A. Goldman, L. Tyan, J. Bekeny, J. M. Hall, N. Lee, C. Perez-Cervantes, O. Burnicka-Turek, K. D. Poss, C. R. Weber, W. de Laat, A. J. Ruthenburg, and I. P. Moskowitz, "Transcription-factor-dependent enhancer transcription defines a gene regulatory network for cardiac rhythm," *eLife*, vol. 6, pp. 1–21, 2017. ISBN: 2158-1592.

[65] K. C. Wang, Y. W. Yang, B. Liu, A. Sanyal, R. Corces-Zimmerman, Y. Chen, B. R. Lajoie, A. Protacio, R. A. Flynn, R. A. Gupta, J. Wysocka, M. Lei, J. Dekker, J. A. Helms, and H. Y. Chang, "A long noncoding rna maintains active chromatin to coordinate homeotic gene expression," *Nature*, vol. 472, p. 120–124, Mar. 2011.

[66] J. Kung, B. Kesner, J. An, J. Ahn, C. Cifuentes-Rojas, D. Colognori, Y. Jeon, A. Szanto, B. del Rosario, S. Pinter, J. Erwin, and J. Lee, "Locus-specific targeting to the x chromosome revealed by the rna interactome of ctcf," *Molecular Cell*, vol. 57, p. 361–375, Jan. 2015.

[67] C.-K. Chen, M. Blanco, C. Jackson, E. Aznauryan, N. Ollikainen, C. Surka, A. Chow, A. Cerase, P. McDonel, and M. Guttman, "Xist recruits the x chromosome to the nuclear lamina to enable chromosome-wide silencing," *Science*, vol. 354, p. 468–472, Oct. 2016.

[68] J. Engreitz, E. S. Lander, and M. Guttman, *RNA Antisense Purification (RAP) for Mapping RNA Interactions with Chromatin*, p. 183–197. Springer New York, Dec. 2014.

[69] M. D. Simon, "Capture hybridization analysis of rna targets (chart)," *Current Protocols in Molecular Biology*, vol. 101, Jan. 2013.

[70] B. Sridhar, M. Rivas-Astroza, T. C. Nguyen, W. Chen, Z. Yan, X. Cao, L. Hebert, and S. Zhong, "Systematic mapping of rna-chromatin interactions in vivo," *Current Biology*, vol. 27, p. 602–609, Feb. 2017.

[71] X. Li, B. Zhou, L. Chen, L.-T. Gou, H. Li, and X.-D. Fu, "Grid-seq reveals the global rna–chromatin interactome," *Nature Biotechnology*, vol. 35, p. 940–950, Sept. 2017.

[72] S. A. Quinodoz, P. Bhat, P. Chovanec, J. W. Jachowicz, N. Ollikainen, E. Detmar, E. Soehalim, and M. Guttman, "Sprite: a genome-wide method for mapping higher-order 3d interactions in the nucleus using combinatorial split-and-pool barcoding," *Nature Protocols*, vol. 17, p. 36–75, Jan. 2022.

[73] C. He, S. Sidoli, R. Warneford-Thomson, D. C. Tatomer, J. E. Wilusz, B. A. Garcia, and R. Bonasio, "High-Resolution Mapping of RNA-Binding Regions in the Nuclear Proteome of Embryonic Stem Cells," *Molecular Cell*, vol. 64, no. 2, pp. 416–430, 2016. Publisher: Elsevier Inc.

[74] O. Oksuz, J. E. Henninger, R. Warneford-Thomson, M. M. Zheng, H. Erb, A. Vancura, K. J. Overholt, S. W. Hawken, S. F. Banani, R. Lauman, L. N. Reich, A. L. Robertson, N. M. Hannett, T. I. Lee, L. I. Zon, R. Bonasio, and R. A. Young, "Transcription factors interact with RNA to regulate genes," *Molecular Cell*, vol. 83, pp. 2449–2463.e13, July 2023. Publisher: Elsevier.

[75] A. S. Hansen, A. Amitai, C. Cattoglio, R. Tjian, and X. Darzacq, "Guided nuclear exploration increases CTCF target search efficiency," *Nature Chemical Biology*, vol. 16, no. 3, pp. 257–266, 2020. Publisher: Springer US.

[76] A. S. Hansen, T.-H. S. Hsieh, C. Cattoglio, I. Pustova, R. Saldaña-Meyer, D. Reinberg, X. Darzacq, and R. Tjian, "Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region in CTCF," *Molecular Cell*, vol. 76, pp. 395–411.e13, Nov. 2019. Publisher: Elsevier.

[77] R. Saldaña-Meyer, J. Rodriguez-Hernaez, T. Escobar, M. Nishana, K. Jácome-López, E. P. Nora, B. G. Bruneau, A. Tsirigos, M. Furlan-Magaril, J. Skok, and D. Reinberg, "RNA Interactions Are Essential for CTCF-Mediated Genome Organization," *Molecular Cell*, vol. 76, pp. 412–422.e5, Nov. 2019. Publisher: Elsevier.

[78] Z. E. Holmes, D. J. Hamilton, T. Hwang, N. V. Parsonnet, J. L. Rinn, D. S. Wuttke, and R. T. Batey, "The Sox2 transcription factor binds RNA," *Nature Communications*, vol. 11, no. 1, 2020. Publisher: Springer US.

[79] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, "Crystal structure of the nucleosome core particle at 2.8 Å resolution," *Nature*, vol. 389, pp. 251–260, Sept. 1997. Publisher: Nature Publishing Group.

[80] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond, "Solvent mediated interactions in the structure of the nucleosome core particle at 1.9Å resolution," *Journal of Molecular Biology*, vol. 319, p. 1097–1113, June 2002.

[81] G. Millán-Zambrano, A. Burton, A. J. Bannister, and R. Schneider, "Histone post-translational modifications — cause and consequence of genome function," *Nat Rev Genet*, vol. 23, pp. 563–580, Sept. 2022. Publisher: Nature Publishing Group.

[82] R. Liu, J. Wu, H. Guo, W. Yao, S. Li, Y. Lu, Y. Jia, X. Liang, J. Tang, and H. Zhang, "Post-translational modifications of histones: Mechanisms, biological functions, and therapeutic targets," *MedComm*, vol. 4, May 2023.

[83] S. Chen, L. Jiao, M. Shubbar, X. Yang, and X. Liu, "Unique Structural Platforms of Suz12 Dictate Distinct Classes of PRC2 for Chromatin Binding," *Molecular Cell*, vol. 69, pp. 840–852.e5, Mar. 2018.

[84] H. Santos-Rosa *et al.*, "Active genes are tri-methylated at k4 of histone h3," *Nature*, vol. 419, pp. 407–411, 2002.

[85] J. Wysocka *et al.*, "A phd finger of nurf couples histone h3 lysine 4 trimethylation with chromatin remodelling," *Nature*, vol. 442, pp. 86–90, 2006.

[86] N. D. Heintzman *et al.*, "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome," *Nat. Genet.*, vol. 39, pp. 311–318, 2007.

[87] M. G. Guenther, S. S. Levine, L. A. Boyer, R. Jaenisch, and R. A. Young, "A chromatin landmark and transcription initiation at most promoters in human cells," *Cell*, vol. 130, pp. 77–88, 2007.

[88] M. Vermeulen *et al.*, "Selective anchoring of tfiid to nucleosomes by trimethylation of histone h3 lysine 4," *Cell*, vol. 131, pp. 58–69, 2007.

[89] S. M. Lauberth *et al.*, "H3k4me3 interactions with taf3 regulate preinitiation complex assembly and selective gene activation," *Cell*, vol. 152, pp. 1021–1036, 2013.

[90] E. Blanco, M. González-Ramírez, A. Alcaine-Colet, S. Aranda, and L. Di Croce, "The Bivalent Genome: Characterization, Structure, and Regulation," *Trends in Genetics*, vol. 36, no. 2, pp. 118–131, 2020.

[91] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.-K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein, "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells," *Nature*, vol. 448, pp. 553–560, Aug. 2007.

[92] B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber, and E. S. Lander, "A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells," *Cell*, vol. 125, no. 2, pp. 315–326, 2006.

[93] D. S. Gilmour and J. T. Lis, "Detecting protein-dna interactions in vivo: distribution of rna polymerase on specific bacterial genes.," *Proceedings of the National Academy of Sciences*, vol. 81, p. 4275–4279, July 1984.

[94] D. A. Orlando, M. W. Chen, V. E. Brown, S. Solanki, Y. J. Choi, E. R. Olson, C. C. Fritz, J. E. Bradner, and M. G. Guenther, "Quantitative ChIP-Seq Normalization Reveals Global Modulation of the Epigenome," *Cell Reports*, vol. 9, pp. 1163–1170, Nov. 2014. Publisher: Elsevier.

[95] R. N. Shah, A. T. Grzybowski, E. M. Cornett, A. L. Johnstone, B. M. Dickson, B. A. Boone, M. A. Cheek, M. W. Cowles, D. Maryanski, M. J. Meiners, R. L. Tiedemann, R. M. Vaughan, N. Arora, Z.-W. Sun, S. B. Rothbart, M.-C. Keogh, and A. J. Ruthenburg, "Examining the Roles of H3K4 Methylation States with Systematically Characterized Antibodies," *Molecular Cell*, vol. 72, pp. 162–177.e7, Oct. 2018. Publisher: Elsevier.

[96] A. T. Grzybowski, Z. Chen, and A. J. Ruthenburg, "Calibrating ChIP-Seq with Nucleosomal Internal Standards to Measure Histone Modification Density Genome Wide," *Molecular Cell*, vol. 58, no. 5, pp. 886–899, 2015. arXiv: 15334406 Publisher: Elsevier Inc. ISBN: 4546474849.

[97] T. Jana, S. Brodsky, and N. Barkai, "Speed–Specificity Trade-Offs in the Transcription Factors Search for Their Genomic Binding Sites," *Trends in Genetics*, pp. 1–12, 2021. Publisher: Elsevier Ltd.

[98] S. Brodsky, T. Jana, K. Mittelman, M. Chapal, D. K. Kumar, M. Carmi, S. Brodsky, T. Jana, K. Mittelman, M. Chapal, D. K. Kumar, and M. Carmi, "Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity Article Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity," *Molecular Cell*, pp. 1–13, 2020. Publisher: Elsevier Inc.

[99] M. Mazzocca, A. Loffreda, E. Colombo, T. Fillot, D. Gnani, P. Falletta, E. Monteleone, S. Capozi, E. Bertrand, G. Legube, Z. Lavagnino, C. Tacchetti, and D. Mazza, "Chromatin organization drives the search mechanism of nuclear factors," *Nat Commun*, vol. 14, p. 6433, Oct. 2023. Number: 1 Publisher: Nature Publishing Group.

[100] Y. Chen, C. Cattoglio, G. M. Dailey, Q. Zhu, R. Tjian, and X. Darzacq, "Mechanisms governing target search and binding dynamics of hypoxia-inducible factors," *eLife*, vol. 11, p. e75064, Nov. 2022. Publisher: eLife Sciences Publications, Ltd.

[101] D. K. Kumar, F. Jonas, T. Jana, S. Brodsky, M. Carmi, and N. Barkai, "Complementary strategies for directing in vivo transcription factor binding through DNA binding domains and intrinsically disordered regions," *Molecular Cell*, vol. 83, pp. 1462–1473.e5, May 2023. Publisher: Elsevier.

[102] D. Panne, T. Maniatis, and S. C. Harrison, "An atomic model of the interferon-beta enhanceosome," *Cell*, vol. 129, pp. 1111–1123, June 2007.

[103] H. Göös, M. Kinnunen, K. Salokas, Z. Tan, X. Liu, L. Yadav, Q. Zhang, G.-H. Wei, and M. Varjosalo, "Human transcription factor protein interaction networks," *Nat Commun*, vol. 13, p. 766, Feb. 2022.

[104] J. Tycko, N. DelRosso, G. T. Hess, Aradhana, A. Banerjee, A. Mukund, M. V. Van, B. K. Ego, D. Yao, K. Spees, P. Suzuki, G. K. Marinov, A. Kundaje, M. C. Bassik, and L. Bintu, "High-Throughput Discovery and Characterization of Human Transcriptional Effectors," *Cell*, vol. 183, pp. 2020–2035.e16, Dec. 2020.

[105] L. F. Soto, Z. Li, C. S. Santoso, A. Berenson, I. Ho, V. X. Shen, S. Yuan, and J. I. Fuxman Bass, "Compendium of human transcription factor effector domains," *Molecular Cell*, vol. 82, pp. 514–526, Feb. 2022.

[106] J.-S. Lee, R. H. See, K. M. Galvin, J. Wang, and Y. Shi, "Functional interactions between YY1 and adenovirus E1A," *Nucleic Acids Research*, vol. 23, pp. 925–931, Mar. 1995.

[107] Y. Shi, J. S. Lee, and K. M. Galvin, "Everything you have ever wanted to know about Yin Yang 1...," *Biochimica et Biophysica Acta - Reviews on Cancer*, vol. 1332, no. 2, 1997.

[108] E. Seto, Y. Shi, and T. Shenk, "YY1 is an initiator sequence-binding protein that directs and activates transcription in vitro," *Nature*, vol. 354, pp. 241–245, Nov. 1991. Number: 6350 Publisher: Nature Publishing Group.

[109] A. López-Perrote, H. E. Alatwi, E. Torreira, A. Ismail, S. Ayora, J. A. Downs, and O. Llorca, "Structure of Yin Yang 1 oligomers that cooperate with RuvBL1-RuvBL2 ATPases," *Journal of Biological Chemistry*, vol. 289, no. 33, pp. 22614–22629, 2014.

[110] S. Wu, Y. Shi, P. Mulligan, F. Gay, J. Landry, H. Liu, J. Lu, H. H. Qi, W. Wang, J. A. Nickoloff, C. Wu, and Y. Shi, "A YY1-INO80 complex regulates genomic stability through homologous recombination-based repair," *Nature Structural and Molecular Biology*, vol. 14, no. 12, pp. 1165–1172, 2007.

[111] X. Zhang, R. M. Blumenthal, and X. Cheng, "Updated understanding of the protein–DNA recognition code used by C2H2 zinc finger proteins," *Current Opinion in Structural Biology*, vol. 87, p. 102836, Aug. 2024.

[112] H. B. Houbaviy, A. Usheva, T. Shenk, and S. K. Burley, "Cocrystal structure of YY1 bound to the adeno-associated virus P5 initiator," *Proceedings of the National Academy of Sciences*, vol. 93, no. 24, pp. 13577–13582, 2002.

[113] D. C. Wai, M. Shihab, J. K. Low, and J. P. Mackay, "The zinc fingers of YY1 bind single-stranded RNA with low sequence specificity," *Nucleic Acids Research*, vol. 44, no. 19, pp. 9153–9165, 2016.

[114] Z. R. Belak and N. Ovsenek, "Assembly of the Yin Yang 1 Transcription Factor into Messenger Ribonucleoprotein Particles Requires Direct RNA Binding Activity*," *Journal of Biological Chemistry*, vol. 282, pp. 37913–37920, Dec. 2007.

[115] Z. S. Chen, M. Ou, S. Taylor, R. Dafinca, S. I. Peng, K. Talbot, and H. Y. E. Chan, "Mutant GGGGCC RNA prevents YY1 from binding to Fuzzy promoter which stimulates Wnt/β-catenin pathway in C9ALS/FTD," *Nat Commun*, vol. 14, p. 8420, Dec. 2023. Publisher: Nature Publishing Group.

[116] S. Nabeel-Shah, S. Pu, J. D. Burns, U. Braunschweig, N. Ahmed, G. L. Burke, H. Lee, E. Radovani, G. Zhong, H. Tang, E. Marcon, Z. Zhang, T. R. Hughes, B. J. Blencowe, and J. F. Greenblatt, "C2H2-zinc-finger transcription factors bind RNA and function in diverse post-transcriptional regulatory processes," *Molecular Cell*, vol. 0, Sept. 2024. Publisher: Elsevier.

[117] D. U. Gorkin, I. Barozzi, Y. Zhao, Y. Zhang, H. Huang, A. Y. Lee, B. Li, J. Chiou, A. Wildberg, B. Ding, B. Zhang, M. Wang, J. S. Strattan, J. M. Davidson, Y. Qiu, V. Afzal, J. A. Akiyama, I. Plajzer-Frick, C. S. Novak, M. Kato, T. H. Garvin, Q. T. Pham, A. N. Harrington, B. J. Mannion, E. A. Lee, Y. Fukuda-Yuzawa, Y. He, S. Preissl, S. Chee, J. Y. Han, B. A. Williams, D. Trout, H. Amrhein, H. Yang, J. M. Cherry, W. Wang, K. Gaulton, J. R. Ecker, Y. Shen, D. E. Dickel, A. Visel, L. A. Pennacchio, and B. Ren, "An atlas of dynamic chromatin landscapes in mouse fetal development," *Nature*, vol. 583, pp. 744–751, July 2020. Number: 7818 Publisher: Nature Publishing Group.

[118] T.-H. S. Hsieh, C. Cattoglio, E. Slobodyanyuk, A. S. Hansen, X. Darzacq, and R. Tjian, "Enhancer–promoter interactions and transcription are largely maintained upon acute loss of CTCF, cohesin, WAPL or YY1," *Nat Genet*, vol. 54, pp. 1919–1932, Dec. 2022. Number: 12 Publisher: Nature Publishing Group.

[119] M. Figiel, F. Szubert, E. Luchinat, P. Bonarek, A. Baranowska, K. Wajda-Nikiel, M. Wilamowski, P. Miłek, M. Dziedzicka-Wasylewska, L. Banci, and A. Górecki, "Zinc controls operator affinity of human transcription factor YY1 by mediating dimerization via its N-terminal region," *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, vol. 1866, p. 194905, Mar. 2023.

[120] H. B. Houbaviy and S. K. Burley, "Thermodynamic analysis of the interaction between YY1 and the AAV P5 promoter initiator element," *Chemistry & Biology*, vol. 8, pp. 179–187, Feb. 2001.

[121] K. Chen, Y. Lu, K. Shi, D. B. Stovall, D. Li, and G. Sui, "Functional analysis of YY1 zinc fingers through cysteine mutagenesis," *FEBS Letters*, vol. 593, no. 12, pp. 1392–1402, 2019.

[122] S. R. Yant, W. Zhu, D. Millinoff, J. L. Slightom, M. Goodman, and D. L. Gumucio, "High affinity YY1 binding motifs: identification of two core types (ACAT and CCAT) and distribution of potential binding sites within the human β globin cluster," *Nucleic Acids Research*, vol. 23, pp. 4353–4362, Nov. 1995.

[123] A. Shrivastava and K. Calame, "An analysis of genes regulated by the multi-functional transcriptional regulator Yin Yang-1," *Nucl Acids Res*, vol. 22, no. 24, pp. 5151–5155, 1994.

[124] J. K. Guo, M. R. Blanco, W. G. Walkup, G. Bonesteele, C. R. Urbinati, A. K. Banerjee, A. Chow, O. Ettlin, M. Strehle, P. Peyda, E. Amaya, V. Trinh, and M. Guttman, "Denaturing purifications demonstrate that PRC2 and other widely reported chromatin proteins do not appear to bind directly to RNA in vivo," *Molecular Cell*, vol. 0, Feb. 2024. Publisher: Elsevier.

[125] D. Heller, R. Krestel, U. Ohler, M. Vingron, and A. Marsico, "ssHMM: extracting intuitive sequence-structure motifs from high-throughput RNA-binding protein data," *Nucleic Acids Research*, vol. 45, pp. 11004–11018, Nov. 2017.

[126] R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, "ViennaRNA Package 2.0," *Algorithms for Molecular Biology*, vol. 6, p. 26, Nov. 2011.

[127] F. M. Golebiowski, A. Górecki, P. Bonarek, M. Rapala-Kozik, A. Kozik, and M. Dziedzicka-Wasylewska, "An investigation of the affinities, specificity and kinetics involved in the interaction between the Yin Yang 1 transcription factor and DNA," *The FEBS Journal*, vol. 279, no. 17, pp. 3147–3158, 2012. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1742-4658.2012.08693.x.

[128] K. Shrinivas, B. R. Sabari, E. L. Coffey, I. A. Klein, A. Boija, A. V. Zamudio, J. Schuijers, N. M. Hannett, P. A. Sharp, R. A. Young, and A. K. Chakraborty, "Enhancer Features that Drive Formation of Transcriptional Condensates," *Molecular Cell*, vol. 75, pp. 549–561.e7, Aug. 2019.

[129] J. Habchi, P. Tompa, S. Longhi, and V. N. Uversky, "Introducing protein intrinsic disorder," *Chemical Reviews*, vol. 114, no. 13, pp. 6561–6588, 2014.

[130] C. Alfieri, M. C. Gambetta, R. Matos, S. Glatt, P. Sehr, S. Fraterman, M. Wilm, J. Müller, and C. W. Müller, "Structural basis for targeting the chromatin repressor Sfmbt to Polycomb response elements," *Genes and Development*, vol. 27, no. 21, pp. 2367–2379, 2013.

[131] G. Erdős, M. Pajkos, and Z. Dosztányi, "IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation," *Nucleic Acids Research*, vol. 49, pp. W297–W303, July 2021.

[132] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J.

Read, and D. Baker, "Accurate prediction of protein structures and interactions using a three-track neural network," *Science*, vol. 373, pp. 871–876, Aug. 2021. Publisher: American Association for the Advancement of Science.

[133] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J. M. Jumper, "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature*, pp. 1–3, May 2024. Publisher: Nature Publishing Group.

[134] A. Górecki, P. Bonarek, A. K. Górka, M. Figiel, M. Wilamowski, and M. Dziedzicka-Wasylewska, "Intrinsic disorder of human Yin Yang 1 protein," *Proteins: Structure, Function, and Bioinformatics*, vol. 83, no. 7, pp. 1284–1296, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.24822.

[135] S. Ahmad, M. M. Gromiha, and A. Sarai, "Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information," *Bioinformatics*, vol. 20, pp. 477–486, Mar. 2004.

[136] M. Corley, M. C. Burns, and G. W. Yeo, "How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms," *Molecular Cell*, vol. 78, pp. 9–29, Apr. 2020.

[137] J. J. Ellis, M. Broom, and S. Jones, "Protein–RNA interactions: Structural analysis and functional classes," *Proteins: Structure, Function, and Bioinformatics*, vol. 66, no. 4, pp. 903–911, 2007. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.21211.

[138] M. M. Hoffman, M. A. Khrapov, J. C. Cox, J. Yao, L. Tong, and A. D. Ellington, "AANT: the Amino Acid–Nucleotide Interaction Database," *Nucleic Acids Research*, vol. 32, pp. D174–D181, Jan. 2004.

[139] W. Wang, S. Qiao, G. Li, J. Cheng, C. Yang, C. Zhong, D. B. Stovall, J. Shi, C. Teng, D. Li, and G. Sui, "A histidine cluster determines YY1-compartmentalized coactivators and chromatin elements in phase-separated enhancer clusters," *Nucleic Acids Research*, vol. 50, no. 9, pp. 4917–4937, 2022. Publisher: Oxford University Press.

[140] X. Wang, L. S. Bigman, H. M. Greenblatt, B. Yu, Y. Levy, and J. Iwahara, "Negatively charged, intrinsically disordered regions can accelerate target search by DNA-binding proteins," *Nucleic Acids Research*, p. gkad045, Feb. 2023.

[141] A. Ficzycz and N. Ovsenek, "The Yin Yang 1 Transcription Factor Associates with Ribonucleoprotein (mRNP) Complexes in the Cytoplasm of Xenopus Oocytes *," *Journal of Biological Chemistry*, vol. 277, no. 10, pp. 8382–8387, 2002.

[142] R. Xiao, J. Y. Chen, Z. Liang, D. Luo, G. Chen, Z. J. Lu, Y. Chen, B. Zhou, H. Li, X. Du, Y. Yang, M. San, X. Wei, W. Liu, E. Lécuyer, B. R. Graveley, G. W. Yeo, C. B. Burge, M. Q. Zhang, Y. Zhou, and X. D. Fu, "Pervasive Chromatin-RNA Binding Protein Interactions Enable RNA-Based Regulation of Transcription," *Cell*, vol. 178, no. 1, pp. 107–121.e18, 2019.

[143] G. Mas, E. Blanco, C. Ballaré, M. Sansó, Y. G. Spill, D. Hu, Y. Aoi, F. Le Dily, A. Shilatifard, M. A. Marti-Renom, and L. Di Croce, "Promoter bivalency favors an open chromatin architecture in embryonic stem cells," *Nature Genetics*, vol. 50, p. 1452–1462, Sept. 2018.

[144] S. Xiao *et al.*, "Comparative epigenomic annotation of regulatory dna," *Cell*, vol. 149, pp. 1381–1392, 2012.

[145] B. D. Strahl and C. D. Allis, "The language of covalent histone modifications," *Nature*, vol. 403, pp. 41–45, 2000.

[146] T. Kouzarides, "Chromatin modifications and their function," *Cell*, vol. 128, pp. 693–705, 2007.

[147] A. J. Bannister and T. Kouzarides, "Regulation of chromatin by histone modifications," *Cell Res.*, vol. 21, pp. 381–395, 2011.

[148] L. Ringrose, H. Ehret, and R. Paro, "Distinct contributions of histone h3 lysine 9 and 27 methylation to locus-specific stability of polycomb complexes," *Mol. Cell*, vol. 16, pp. 641–653, 2004.

[149] L. A. Boyer *et al.*, "Polycomb complexes repress developmental regulators in murine embryonic stem cells," *Nature*, vol. 441, pp. 349–353, 2006.

[150] D. O'Carroll *et al.*, "The polycomb-group geneezh2 is required for early mouse development," *Mol. Cell. Biol.*, vol. 21, pp. 4330–4336, 2001.

[151] D. Pasini, A. P. Bracken, J. B. Hansen, M. Capillo, and K. Helin, "The polycomb group protein suz12 is required for embryonic stem cell differentiation," *Mol. Cell. Biol.*, vol. 27, pp. 3769–3779, 2007.

[152] D. Pasini, A. P. Bracken, M. R. Jensen, E. Lazzerini Denchi, and K. Helin, "Suz12 is essential for mouse development and for ezh2 histone methyltransferase activity," *EMBO J.*, vol. 23, pp. 4061–4071, 2004.

[153] V. Azuara *et al.*, "Chromatin signatures of pluripotent cell lines," *Nat. Cell Biol.*, vol. 8, pp. 532–538, 2006.

[154] C. A. Gifford *et al.*, "Transcriptional and epigenetic dynamics during specification of human embryonic stem cells," *Cell*, vol. 153, pp. 1149–1163, 2013.

[155] W. Xie *et al.*, "Epigenomic analysis of multilineage differentiation of human embryonic stem cells," *Cell*, vol. 153, pp. 1134–1148, 2013.

[156] P. Voigt, W.-W. Tee, and D. Reinberg, "A double take on bivalent promoters," *Genes Dev.*, vol. 27, pp. 1318–1338, 2013.

[157] K. Kanayama *et al.*, "Genome-wide mapping of bivalent histone modifications in hepatic stem/progenitor cells," *Stem Cells Int.*, vol. 2019, 2019. Article e9789240.

[158] Y. Zhou *et al.*, "Bivalent histone codes on wnt5a during odontogenic differentiation," *J. Dent. Res.*, vol. 97, pp. 99–107, 2018.

[159] M. J. Burney *et al.*, "An epigenetic signature of developmental potential in neural stem cells and early neurons," *STEM CELLS*, vol. 31, pp. 1868–1880, 2013.

[160] K. Tanimura, T. Suzuki, D. Vargas, H. Shibata, and T. Inagaki, "Epigenetic regulation of beige adipocyte fate by histone methylation," *Endocr. J.*, vol. 66, pp. 115–125, 2019.

[161] S. S. Dhar *et al.*, "An essential role for utx in resolution and activation of bivalent promoters," *Nucleic Acids Res.*, vol. 44, pp. 3659–3674, 2016.

[162] B. J. Abraham, K. Cui, Q. Tang, and K. Zhao, "Dynamic regulation of epigenomic landscapes during hematopoiesis," *BMC Genomics*, vol. 14, p. 193, 2013.

[163] S. Seenundun *et al.*, "Utx mediates demethylation of h3k27me3 at muscle-specific genes during myogenesis," *EMBO J.*, vol. 29, pp. 1401–1411, 2010.

[164] K. Cui *et al.*, "Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation," *Cell Stem Cell*, vol. 4, pp. 80–93, 2009.

[165] J. A. Dahl, A. H. Reiner, A. Klungland, T. Wakayama, and P. Collas, "Histone h3 lysine 27 methylation asymmetry on developmentally-regulated promoters distinguish the first two lineages in mouse preimplantation embryos," *PLOS ONE*, vol. 5, p. e9150, 2010.

[166] N. L. Vastenhouw *et al.*, "Chromatin signature of embryonic pluripotency is established during genome activation," *Nature*, vol. 464, pp. 922–926, 2010.

[167] N. L. Vastenhouw and A. F. Schier, "Bivalent histone modifications in early embryogenesis," *Curr. Opin. Cell Biol.*, vol. 24, pp. 374–386, 2012.

[168] O. Alder *et al.*, "Ring1b and suv39h1 delineate distinct chromatin states at bivalent genes during early mouse lineage commitment," *Development*, vol. 137, pp. 2483–2492, 2010.

[169] M. Denholtz *et al.*, "Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization," *Cell Stem Cell*, vol. 13, pp. 602–616, 2013.

[170] M. Vieux-Rochas, P. J. Fabre, M. Leleu, D. Duboule, and D. Noordermeer, "Clustering of mammalian hox genes with other h3k27me3 targets within an active nuclear domain," *Proc. Natl. Acad. Sci.*, vol. 112, pp. 4672–4677, 2015.

[171] S. Kundu *et al.*, "Polycomb repressive complex 1 generates discrete compacted domains that change during differentiation," *Mol. Cell*, vol. 65, pp. 432–446.e5, 2017.

[172] T. L. Messier *et al.*, "Oncofetal epigenetic bivalency in breast cancer cells: H3k4 and h3k27 tri-methylation as a biomarker for phenotypic plasticity," *J. Cell. Physiol.*, vol. 231, pp. 2474–2481, 2016.

[173] D. Kaukonen *et al.*, "Analysis of h3k4me3 and h3k27me3 bivalent promotors in her2+ breast cancer cell lines reveals variations depending on estrogen receptor status and significantly correlates with gene expression," *BMC Med. Genomics*, vol. 13, p. 92, 2020.

[174] D. S. Dunican *et al.*, "Bivalent promoter hypermethylation in cancer is linked to the h327me3/h3k4me3 ratio in embryonic stem cells," *BMC Biol.*, vol. 18, p. 25, 2020.

[175] M. L. Burr *et al.*, "An evolutionarily conserved function of polycomb silences the mhc class i antigen presentation pathway and enables immune evasion in cancer," *Cancer Cell*, vol. 36, pp. 385–401.e8, 2019.

[176] H. Patani *et al.*, "Transition to naïve human pluripotency mirrors pan-cancer dna hypermethylation," *Nat. Commun.*, vol. 11, p. 3671, 2020.

[177] D. Hu *et al.*, "The mll2 branch of the compass family regulates bivalent promoters in mouse embryonic stem cells," *Nat. Struct. Mol. Biol.*, vol. 20, pp. 1093–1097, 2013.

[178] D. Hu *et al.*, "Not all h3k4 methylations are created equal: Mll2/compass dependency in primordial germ cell specification," *Mol. Cell*, vol. 65, pp. 460–475.e6, 2017.

[179] A. T. Grzybowski, R. N. Shah, W. F. Richter, and A. J. Ruthenburg, "Native internally calibrated chromatin immunoprecipitation for quantitative studies of histone post-translational modifications," *Nat. Protoc.*, vol. 14, pp. 3275–3302, 2019.

[180] P. Voigt *et al.*, "Asymmetrically modified nucleosomes," *Cell*, vol. 151, pp. 181–193, 2012.

[181] A. Weiner *et al.*, "Co-chip enables genome-wide mapping of histone mark co-occurrence at single-molecule resolution," *Nat. Biotechnol.*, vol. 34, pp. 953–961, 2016.

[182] S. Kinkley *et al.*, "rechip-seq reveals widespread bivalency of h3k4me3 and h3k27me3 in cd4+ memory t cells," *Nat. Commun.*, vol. 7, p. 12514, 2016.

[183] E. Shema *et al.*, "Single-molecule decoding of combinatorially modified nucleosomes," *Science*, vol. 352, pp. 717–721, 2016.

[184] T. Hattori *et al.*, "Antigen clasping by two antigen-binding sites of an exceptionally specific antibody for histone methylation," *Proc. Natl. Acad. Sci.*, vol. 113, pp. 2092–2097, 2016.

[185] N. L. Young *et al.*, "High throughput characterization of combinatorial histone codes," *Mol. Cell. Proteomics MCP*, vol. 8, pp. 2266–2284, 2009.

[186] M. Brand, S. Rampalli, C.-P. Chaturvedi, and F. J. Dilworth, "Analysis of epigenetic modifications of chromatin at specific gene loci by native chromatin immunoprecipitation of nucleosomes isolated using hydroxyapatite chromatography," *Nat. Protoc.*, vol. 3, pp. 398–409, 2008.

[187] S. Raran-Kurussi, J. Tözsér, S. Cherry, J. E. Tropea, and D. S. Waugh, "Differential temperature dependence of tobacco etch virus and rhinovirus 3c proteases," *Anal. Biochem.*, vol. 436, pp. 142–144, 2013.

[188] C. C. Lechner, N. D. Agashe, and B. Fierz, "Traceless synthesis of asymmetrically modified bivalent nucleosomes," *Angew. Chem. Int. Ed.*, vol. 55, pp. 2903–2906, 2016.

[189] H. Marks *et al.*, "The transcriptional and epigenomic foundations of ground state pluripotency," *Cell*, vol. 149, pp. 590–604, 2012.

[190] F. W. Schmitges *et al.*, "Histone methylation by prc2 is inhibited by active chromatin marks," *Mol. Cell*, vol. 42, pp. 330–341, 2011.

[191] D.-H. Kim *et al.*, "Histone h3k27 trimethylation inhibits h3 binding and function of set1-like h3k4 methyltransferase complexes," *Mol. Cell. Biol.*, vol. 33, pp. 4936–4946, 2013.

[192] A. Shilatifard, "The compass family of histone h3k4 methylases: Mechanisms of regulation in development and disease pathogenesis," *Annu. Rev. Biochem.*, vol. 81, pp. 65–95, 2012.

[193] C. C. Sze *et al.*, "Histone h3k4 methylation-dependent and -independent functions of set1a/compass in embryonic stem cell self-renewal and differentiation," *Genes Dev.*, vol. 31, pp. 1732–1737, 2017.

[194] S. Denissov, H. Hofemeister, H. Marks, A. Kranz, G. Ciotta, S. Singh, K. Anastassiadis, H. G. Stunnenberg, and A. F. Stewart, "Mll2 is required for h3k4 trimethylation on bivalent promoters in embryonic stem cells, whereas mll1 is redundant," *Development*, vol. 141, p. 526–537, Feb. 2014.

[195] K. Aoyama *et al.*, "Ezh1 targets bivalent genes to maintain self-renewing stem cells in ezh2-insufficient myelodysplastic syndrome," *iScience*, vol. 9, pp. 161–174, 2018.

[196] W. Béguelin *et al.*, "Ezh2 and bcl6 cooperate to assemble cbx8-bcor complex to repress bivalent promoters, mediate germinal center formation and lymphomagenesis," *Cancer Cell*, vol. 30, pp. 197–213, 2016.

[197] E. Lavarone, C. M. Barbieri, and D. Pasini, "Dissecting the role of h3k27 acetylation and methylation in prc2 mediated control of cellular identity," *Nat. Commun.*, vol. 10, p. 1679, 2019.

[198] H. K. Tan, C.-S. Wu, J. Li, Z. H. Tan, J. R. Hoffman, C. J. Fry, H. Yang, A. Di Ruscio, and D. G. Tenen, "Dnmt3b shapes the mca landscape and regulates mcg for promoter bivalency in human embryonic stem cells," *Nucleic Acids Research*, vol. 47, p. 7460–7475, June 2019.

[199] M. A. Eckersley-Maslin, A. Parry, M. Blotenburg, C. Krueger, Y. Ito, V. N. R. Franklin, M. Narita, C. S. D'Santos, and W. Reik, "Epigenetic priming by dppa2 and 4 in pluripotency facilitates multi-lineage commitment," *Nature Structural & Molecular Biology*, vol. 27, p. 696–705, June 2020.

[200] S. Kasinathan, G. A. Orsi, G. E. Zentner, K. Ahmad, and S. Henikoff, "High-resolution mapping of transcription factor binding sites on native chromatin," *Nat. Methods*, vol. 11, pp. 203–209, 2014.

[201] I. Bock *et al.*, "Detailed specificity analysis of antibodies binding to modified histone tails with peptide arrays," *Epigenetics*, vol. 6, pp. 256–263, 2011.

[202] T. A. Egelhofer *et al.*, "An assessment of histone-modification antibody quality," *Nat. Struct. Mol. Biol.*, vol. 18, pp. 91–93, 2011.

[203] S. Nishikori *et al.*, "Broad ranges of affinity and specificity of anti-histone antibodies revealed by a quantitative peptide immunoprecipitation assay," *J. Mol. Biol.*, vol. 424, pp. 391–399, 2012.

[204] S. M. Fuchs, K. Krajewski, R. W. Baker, V. L. Miller, and B. D. Strahl, "Influence of combinatorial histone modifications on antibody and effector protein recognition," *Curr. Biol.*, vol. 21, pp. 53–58, 2011.

[205] T. Hattori *et al.*, "Recombinant antibodies to histone post-translational modifications," *Nat. Methods*, vol. 10, pp. 992–995, 2013.

[206] C. D. Allis and T. Jenuwein, "The molecular hallmarks of epigenetic control," *Nat. Rev. Genet.*, vol. 17, pp. 487–500, 2016.

[207] C. Lu, M. Coradin, K. A. Janssen, S. Sidoli, and B. A. Garcia, "Combinatorial histone h3 modifications are dynamically altered in distinct cell cycle phases," *J. Am. Soc. Mass Spectrom*, vol. 32, pp. 1300–1311, 2021.

[208] T. Suganuma and J. L. Workman, "Signals and combinatorial functions of histone modifications," *Annu. Rev. Biochem.*, vol. 80, pp. 473–499, 2011.

[209] S. Henikoff, "Histone modifications: Combinatorial complexity or cumulative simplicity?," *Proc. Natl. Acad. Sci.*, vol. 102, pp. 5308–5309, 2005.

[210] M. Radman-Livaja and O. J. Rando, "Nucleosome positioning: how is it established, and why does it matter?," *Dev. Biol.*, vol. 339, pp. 258–266, 2010.

[211] A. J. Ruthenburg *et al.*, "Recognition of a mononucleosomal histone modification pattern by bptf via multivalent interactions," *Cell*, vol. 145, pp. 692–706, 2011.

[212] S. B. Rothbart *et al.*, "Multivalent histone engagement by the linked tandem tudor and phd domains of uhrf1 is required for the epigenetic inheritance of dna methylation," *Genes Dev.*, vol. 27, pp. 1288–1298, 2013.

[213] P. Savitsky *et al.*, "Multivalent histone and dna engagement by a phd/brd/pwwp triple reader cassette recruits zmynd8 to k14ac-rich chromatin," *Cell Rep.*, vol. 17, pp. 2724–2737, 2016.

[214] M. Yun, J. Wu, J. L. Workman, and B. Li, "Readers of histone modifications," *Cell Res.*, vol. 21, pp. 564–578, 2011.

[215] H. M. Amemiya, A. Kundaje, and A. P. Boyle, "The encode blacklist: Identification of problematic regions of the genome," *Sci. Rep.*, vol. 9, p. 9354, 2019.

[216] R. Margueron *et al.*, "Role of the polycomb protein eed in the propagation of repressive histone marks," *Nature*, vol. 461, pp. 762–767, 2009.

[217] W. Yuan *et al.*, "Dense chromatin activates polycomb repressive complex 2 to regulate h3 lysine 27 methylation," *Science*, vol. 337, pp. 971–975, 2012.

[218] M. Hooper, K. Hardy, A. Handyside, S. Hunter, and M. Monk, "Hprt-deficient (lesch–nyhan) mouse embryos derived from germline colonization by cultured cells," *Nature*, vol. 326, pp. 292–295, 1987.

[219] L. Conti *et al.*, "Niche-independent symmetrical self-renewal of a mammalian tissue stem cell," *PLOS Biol.*, vol. 3, p. e283, 2005.

[220] E. Abranches *et al.*, "Neural differentiation of embryonic stem cells in vitro: A road map to neurogenesis in the embryo," *PLOS ONE*, vol. 4, p. e6286, 2009.

[221] Z. Chen, A. T. Grzybowski, and A. J. Ruthenburg, "Traceless semisynthesis of a set of histone 3 species bearing specific lysine methylation marks," *Chembiochem Eur. J. Chem. Biol.*, vol. 15, pp. 2071–2075, 2014.

[222] K. Luger, T. J. Rechsteiner, and T. J. Richmond, *Preparation of nucleosome core particle from recombinant histones. in Methods in Enzymology vol. 304 3–19*. Academic Press, 1999.

[223] P. N. Dyer *et al.*, "Reconstitution of nucleosome core particles from recombinant histones and dna," *Methods Enzymol*, vol. 375, pp. 23–44, 2004.

[224] P. T. Lowary and J. Widom, "New dna sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning," *J. Mol. Biol.*, vol. 276, pp. 19–42, 1998.

[225] R. Ullah *et al.*, "Activity of the human rhinovirus 3c protease studied in various buffers, additives and detergents solutions for recombinant protein production," *PLOS ONE*, vol. 11, 2016. Article e0153436.

[226] M. G. Cordingley, P. L. Callahan, V. V. Sardana, V. M. Garsky, and R. J. Colonno, "Substrate requirements of human rhinovirus 3c protease for peptide cleavage in vitro," *J. Biol. Chem.*, vol. 265, pp. 9062–9065, 1990.

[227] D. Beckett, E. Kovaleva, and P. J. Schatz, "A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation," *Protein Sci. Publ. Protein Soc.*, vol. 8, pp. 921–929, 1999.

[228] L. J. Bailey *et al.*, "Applications for an engineered protein-g variant with a ph controllable affinity to antibody fragments," *J. Immunol. Methods*, vol. 415, pp. 24–30, 2014.

[229] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short dna sequences to the human genome," *Genome Biol.*, vol. 10, p. R25, 2009.

[230] A. R. Quinlan and I. M. Hall, "Bedtools: a flexible suite of utilities for comparing genomic features. bioinforma," *Oxf. Engl.*, vol. 26, pp. 841–842, 2010.

[231] S. Heinz *et al.*, "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities," *Mol. Cell*, vol. 38, pp. 576–589, 2010.

[232] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic rna-seq quantification," *Nat. Biotechnol.*, vol. 34, pp. 525–527, 2016.

[233] H. Pimentel, N. L. Bray, S. Puente, P. Melsted, and L. Pachter, "Differential analysis of rna-seq incorporating quantification uncertainty," *Nat. Methods*, vol. 14, pp. 687–690, 2017.

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

[235] D. Lu, M. Alexandra Searles, and A. Klug, "Crystal structure of a zinc-finger–rna complex reveals two modes of molecular recognition," *Nature*, vol. 426, p. 96–100, Nov. 2003.

[236] S. Sun, B. Del Rosario, A. Szanto, Y. Ogawa, Y. Jeon, and J. Lee, "Jpx rna activates xist by evicting ctcf," *Cell*, vol. 153, p. 1537–1551, June 2013.

[237] H. J. Oh, R. Aguilar, B. Kesner, H.-G. Lee, A. J. Kriz, H.-P. Chu, and J. T. Lee, "Jpx rna regulates ctcf anchor site selection and formation of chromosome loops," *Cell*, vol. 184, pp. 6157–6173.e24, Dec. 2021.

[238] D. A. Garcia, T. A. Johnson, D. M. Presman, G. Fettweis, K. Wagh, L. Rinaldi, D. A. Stavreva, V. Paakinaho, R. A. Jensen, S. Mandrup, A. Upadhyaya, and G. L. Hager, "An intrinsically disordered region-mediated confinement state contributes to the dynamics and function of transcription factors," *Molecular Cell*, vol. 81, pp. 1484–1498.e6, Apr. 2021.

[239] S. Xhani, S. Lee, H. M. Kim, S. Wang, S. Esaki, V. L. T. Ha, M. Khanezarrin, G. L. Fernandez, A. V. Albrecht, J. M. Aramini, M. W. Germann, and G. M. K. Poon, "Intrinsic disorder controls two functionally distinct dimers of the master transcription factor pu.1," *Science Advances*, vol. 6, Feb. 2020.

[240] A. S. Krois, H. J. Dyson, and P. E. Wright, "Long-range regulation of p53 dna binding by its intrinsically disordered n-terminal transactivation domain," *Proceedings of the National Academy of Sciences*, vol. 115, Nov. 2018.

[241] J. S. Lee, K. M. Galvin, and Y. Shi, "Evidence for physical interaction between the zinc-finger transcription factors YY1 and Sp1," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 90, no. 13, pp. 6145–6149, 1993.

[242] J. S. Lee, K. M. Galvin, R. H. See, R. Eckner, D. Livingston, E. Moran, and Y. Shi, "Relief of yy1 transcriptional repression by adenovirus e1a is mediated by e1a-associated protein p300.," *Genes & Development*, vol. 9, p. 1188–1198, May 1995.

[243] W. Qi, K. Zhao, J. Gu, Y. Huang, Y. Wang, H. Zhang, M. Zhang, J. Zhang, Z. Yu, L. Li, L. Teng, S. Chuai, C. Zhang, M. Zhao, H. Chan, Z. Chen, D. Fang, Q. Fei, L. Feng, L. Feng, Y. Gao, H. Ge, X. Ge, G. Li, A. Lingel, Y. Lin, Y. Liu, F. Luo, M. Shi, L. Wang, Z. Wang, Y. Yu, J. Zeng, C. Zeng, L. Zhang, Q. Zhang, S. Zhou, C. Oyang, P. Atadja, and E. Li, "An allosteric prc2 inhibitor targeting the h3k27me3 binding pocket of eed," *Nature Chemical Biology*, vol. 13, p. 381–388, Jan. 2017.

[244] A. Laugesen, J. W. Højfeldt, and K. Helin, "Molecular Mechanisms Directing PRC2 Recruitment and H3K27 Methylation," *Molecular Cell*, vol. 74, pp. 8–18, Apr. 2019. Publisher: Elsevier.

[245] K. Grosselin, A. Durand, J. Marsolier, A. Poitou, E. Marangoni, F. Nemati, A. Dahmani, S. Lameiras, F. Reyal, O. Frenoy, Y. Pousse, M. Reichen, A. Woolfe, C. Brenan, A. D. Griffiths, C. Vallot, and A. Gérard, "High-throughput single-cell chip-seq identifies heterogeneity of chromatin states in breast cancer," *Nature Genetics*, vol. 51, p. 1060–1066, May 2019.

[246] S. J. Wu, S. N. Furlan, A. B. Mihalas, H. S. Kaya-Okur, A. H. Feroze, S. N. Emerson, Y. Zheng, K. Carson, P. J. Cimino, C. D. Keene, J. F. Sarthy, R. Gottardo, K. Ahmad, S. Henikoff, and A. P. Patel, "Single-cell cut&tag analysis of chromatin modifications in differentiation and tumor progression," *Nature Biotechnology*, vol. 39, p. 819–824, Apr. 2021.

[247] L. Pan, P. Parini, R. Tremmel, J. Loscalzo, V. M. Lauschke, B. A. Maron, P. Paci, I. Ernberg, N. S. Tan, Z. Liao, W. Yin, S. Rengarajan, and X. Li, "Single cell atlas: a single-cell multi-omics human cell encyclopedia," *Genome Biology*, vol. 25, Apr. 2024.

[248] T. Chari and L. Pachter, "The specious art of single-cell genomics," *PLOS Computational Biology*, vol. 19, p. e1011288, Aug. 2023.

[249] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass, "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities," *Mol Cell*, vol. 38, pp. 576–589, May 2010.

[250] T. H. S. Hsieh, C. Cattoglio, E. Slobodyanyuk, A. S. Hansen, O. J. Rando, R. Tjian, and X. Darzacq, "Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding," *Molecular Cell*, pp. 1–15, 2020. Publisher: Elsevier Inc.

[251] W. A. Bickmore, "The Spatial Organization of the Human Genome," *Annual Review of Genomics and Human Genetics*, 2013.

[252] I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B.-K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, I. Dunham, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kellis, J. Khatun, P. Kheradpour, A. Kundaje, T. Lassmann, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. J. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, M. Snyder, M. J. Pazin, R. F. Lowdon, L. A. L. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, E. D. Green, P. J. Good, E. A. Feingold, B. E. Bernstein, E. Birney, G. E. Crawford, J. Dekker, L. Elnitski, P. J. Farnham, M. Gerstein, M. C. Giddings, T. R. Gingeras, E. D. Green, R. Guigó, R. C. Hardison, T. J. Hubbard, M. Kellis, W. J. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, M. Snyder, J. A. Stamatoyannopoulos, S. A. Tenenbaum, Z. Weng, K. P. White, B. Wold, J. Khatun, Y. Yu, J. Wrobel, B. A. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, M. C. Giddings, B. E. Bernstein, C. B. Epstein, N. Shoresh, J. Ernst, P. Kheradpour, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X. Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, L. D. Ward, R. C. Altshuler, M. L. Eaton, M. Kellis, S. Djebali,

C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Röder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, P. Batut, I. Bell, K. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. P. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, G. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca, B. A. Risk, D. Robyr, X. Ruan, M. Sammeth, K. S. Sandhu, L. Schaeffer, L.-H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, Y. Hayashizaki, J. Harrow, M. Gerstein, T. J. Hubbard, A. Reymond, S. E. Antonarakis, G. J. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigó, T. R. Gingeras, K. R. Rosenbloom, C. A. Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. P. Barber, M. S. Cline, T. R. Dreszer, S. G. Heitner, D. Karolchik, W. J. Kent, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney, T. S. Furey, L. Song, L. L. Grasfeder, P. G. Giresi, B.-K. Lee, A. Battenhouse, N. C. Sheffield, J. M. Simon, K. A. Showers, A. Safi, D. London, A. A. Bhinge, C. Shestak, M. R. Schaner, S. Ki Kim, Z. Z. Zhang, P. A. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Ni, N. U. Rashid, M. J. Kim, S. Adar, Z. Zhang, T. Wang, D. Winter, D. Keefe, E. Birney, V. R. Iyer, J. D. Lieb, G. E. Crawford, G. Li, K. S. Sandhu, M. Zheng, P. Wang, O. J. Luo, A. Shahab, M. J. Fullwood, X. Ruan, Y. Ruan, R. M. Myers, F. Pauli, B. A. Williams, J. Gertz, G. K. Marinov, T. E. Reddy, J. Vielmetter, E. Partridge, D. Trout, K. E. Varley, C. Gasper, and The ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, pp. 57–74, Sept. 2012. Publisher: Nature Publishing Group.

[253] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position," *Nat Methods*, vol. 10, pp. 1213–1218, Dec. 2013. Publisher: Nature Publishing Group.

[254] M. F. Pereira, V. Finazzi, L. Rizzuti, D. Aprile, V. Aiello, L. Mollica, M. Riva, C. Soriani, F. Dossena, R. Shyti, D. Castaldi, E. Tenderini, M. T. Carminho-Rodrigues, J. F. Bally, B. B. A. de Vries, M. Gabriele, A. Vitriolo, and G. Testa, "YY1 mutations disrupt corticogenesis through a cell type specific rewiring of cell-autonomous and non-cell-autonomous transcriptional programs," *Mol Psychiatry*, pp. 1–17, Feb. 2025. Publisher: Nature Publishing Group.

[255] B. Bintu, L. J. Mateo, J.-H. Su, N. A. Sinnott-Armstrong, M. Parker, S. Kinrot, K. Yamaya, A. N. Boettiger, and X. Zhuang, "Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells," *Science*, vol. 362, Oct. 2018.

[256] A. N. Boettiger, B. Bintu, J. R. Moffitt, S. Wang, B. J. Beliveau, G. Fudenberg, M. Imakaev, L. A. Mirny, C.-t. Wu, and X. Zhuang, "Super-resolution imaging reveals

distinct chromatin folding for different epigenetic states," *Nature*, vol. 529, p. 418–422, Jan. 2016.

[257] M. Gabriele, H. B. Brandão, S. Grosse-Holz, A. Jha, G. M. Dailey, C. Cattoglio, T.-H. S. Hsieh, L. Mirny, C. Zechner, and A. S. Hansen, "Dynamics of ctcf- and cohesin-mediated chromatin looping revealed by live-cell imaging," *Science*, vol. 376, p. 496–501, Apr. 2022.

[258] H. K. Tan *et al.*, "Dnmt3b shapes the mca landscape and regulates mcg for promoter bivalency in human embryonic stem cells," *Nucleic Acids Res.*, vol. 47, pp. 7460–7475, 2019.

[259] C. Terranova *et al.*, "Global developmental gene programing involves a nuclear form of fibroblast growth factor receptor-1 (fgfr1)," *PLOS ONE*, vol. 10, 2015. Article e0123380.

[260] Q. R. Xing *et al.*, "Parallel bimodal single-cell sequencing of transcriptome and chromatin accessibility," *Genome Res.*, vol. 30, pp. 1027–1039, 2020.

[261] Y. Zhang *et al.*, "Model-based analysis of chip-seq (macs)," *Genome Biol.*, vol. 9, p. R137, 2008.