GENOME-WIDE CHARACTERIZATION OF THE ROLES OF TRANSCRIPTION FACTORS GAF AND HSF IN THE TRANSCRIPTIONAL HEAT SHOCK RESPONSE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Fabiana de Melo Duarte

May 2017

GENOME-WIDE CHARACTERIZATION OF THE ROLES OF TRANSCRIPTION FACTORS GAF AND HSF IN THE TRANSCRIPTIONAL HEAT SHOCK RESPONSE

Fabiana de Melo Duarte, Ph.D.

Cornell University 2017

In eukaryotes, RNA polymerase II (Pol II) is responsible for the transcription of all protein-coding genes, and regulation of its activity is fundamental for cellular homeostasis and the programmed development of multicellular organisms. The transcription process is regulated by the coordinated action of transcription factors (TFs), which interact with each other, Pol II and specific regulatory sequences to modulate distinct rate-limiting steps in the transcription cycle. Identifying the regulatory TFs and the biochemical processes that are controlled by each factor is therefore critical for understanding how transcription is regulated. When studying mechanisms of transcription regulation, inducible systems are an invaluable resource: regulatory processes can be triggered instantaneously, enabling the tracking of ordered mechanistic events.

We used Precision Run-On sequencing to examine the genome-wide Heat Shock (HS) response in *Drosophila* and the function of two key TFs on the immediate transcription activation or repression of all genes regulated by HS. We identified the primary HS responsive genes and the rate-limiting steps that GAGA-Associated Factor (GAF) and Heat Shock Factor (HSF) regulate. We demonstrated that GAF acts before

promoter-proximally paused Pol II formation, likely at the step of chromatin opening, and that GAF-facilitated Pol II pausing is critical for HS activation. In contrast, HSF is dispensable for establishing or maintaining Pol II pausing, but is critical for the release of paused Pol II into the gene body at a subset of highly-activated genes upon HS induction. Additionally, HSF has no detectable role in the rapid HS-repression of thousands of genes.

As a complementary approach, we have selected RNA aptamers to the human HSF1 and HSF2 with the goal of expressing these aptamers in vivo to disrupt specific macromolecular interactions of these TFs. I describe the results and thorough characterization of a successful RNA aptamer selection to HSF1 and HSF2 using our recently developed SELEX technology. I also describe our development and implementation of a set of SELEX performance metrics to evaluate selection success. Our novel SELEX methodology and analysis tools offer a significant improvement over traditional approaches and provides an efficient platform for the performance and analysis of SELEX experiments.

BIOGRAPHICAL SKETCH

Fabiana de Melo Duarte was born on February 14th 1987 in Lavras, a hilly city in southern Minas Gerais state, Brazil. In Lavras, Fabiana was raised by a family of strong, generous and independent women, including her mother Márcia, grandmother Zenita, her six aunts and her cousin Tati. From her mom – an incredible and exemplary woman who had to overcome many challenges in life from a very young age – Fabiana learned a strong sense of kindness, cheerfulness, fortitude and dedication, which have always guided her life.

Fabiana studied at Colégio Nossa Senhora de Lourdes up till middle school, when she switched to Instituto Presbiteriano Gammon. Halfway through high school, Fabiana moved with her family to the city of Ipatinga, where she finished her high school diploma at Escola Educação Criativa.

Fabiana's interest in the life sciences began when she was just a kid, when she used to spend hours reading her collection of children's books about the human body – a gift from her aunt Merô. At age 13, when she first started taking biology lessons in middle school, Fabiana decided that she wanted to become a biologist to study the fascinating processes that take place inside living cells.

Fabiana has always had ambitious goals for her education, and she worked hard to get into one of the top biology programs in the country, at Universidade Estadual de Campinas (Unicamp) in Campinas, São Paulo. Her undergraduate biology program had a very broad curriculum, including many botany and zoology classes. However, since she

took her first genetics and biochemistry courses in her freshman year, Fabiana knew that it was the molecular aspects of biology that truly captivated her.

In her sophomore year, Fabiana started working with Professors Lucas Argueso and Gonçalo Pereira studying chromosome rearrangements in industrial yeast strains used to produce fuel ethanol. In the summer before her senior year, she did a research internship in the laboratory of Professor Thomas Petes at Duke University, under Professor Argueso's mentorship. After completing her undergraduate degree in 2008, Fabiana stayed at Unicamp to pursue a master's degree with Professors Argueso and Pereira, when she continued her investigations of chromosome instability in yeast. Although the ultimate goal of the project was to generate yeast strains that produce higher amounts of ethanol, Fabiana's main interests were always in the basic and mechanistic aspects of the research.

Professor Argueso, who earned his PhD at Cornell in the Alani laboratory, has always been one of the strongest supporters of Fabiana's goal to pursue a career in academia. He encouraged her to apply to PhD programs at top schools in the United States, where she would have more resources and opportunities to grow as a scientist.

In 2010, Fabiana decided to pursue a PhD degree at the Department of Molecular Biology and Genetics at Cornell University. During her rotation in John Lis' laboratory, she became fascinated with the intricate processes that regulate transcription in eukaryotic cells. Fabiana was also captivated by Professor Lis' brilliance, enthusiasm and passion for science, and she was very fortunate to have the opportunity to join his group. Under Professor Lis' mentorship and through interactions with an incredibly talented

group of colleagues, Fabiana matured as a scientist and developed the required skills to pursue her academic career.

Para Merô.

Por despertar o meu lado cientista,

por me inspirar a conquistar novos desafios,

por seu amor e dedicação incondicionais.

Te amo!

ACKNOWLEDGMENTS

This achievement would not have been possible without the continuous support, guidance and encouragement from an amazing group of individuals.

First and foremost, I would like to thank my PhD advisor, Professor John Lis, for being such a wonderful teacher and mentor. I was extremely lucky to have the opportunity to work with such a brilliant scientist, and his enthusiasm and passion for science have always been truly inspiring. Through his mentorship, John helped me grow as a scientist, and I am coming out of his lab a much better molecular biologist, presenter, writer, mentor and teacher than I was when I started. Furthermore, he creates an amazing lab atmosphere that promotes critical and independent thinking and fruitful scientific discussions, making his lab a very exciting place to work. John has also helped me to delineate a path to achieve my goal of becoming a professor and has been thoroughly guiding me through the transition to the next step in my scientific career. I will be forever grateful to the fundamental role John has had – and will continue to have – in my life.

I would like to thank the other two members of my thesis committee, Professors Jeff Pleiss and Paul Soloway, for their continuous guidance and support during the various stages of my PhD and for providing critical feedback on my thesis projects. I am also grateful to Professor Eric Alani for his guidance, encouragement, insightful suggestions and for always sharing his contagious enthusiasm; and to Professors Charles Danko and Hojoong Kwak, for their critical feedback on data analysis and the interpretation of PRO-seq results.

Besides the professors who played a major role during my PhD years, I would like to express my gratitude to my first scientific advisor, Professor Lucas Argueso, who had a fundamental role in my decision to pursue a PhD. Lucas also encouraged me to apply to PhD programs in the United States, and his guidance and support opened many doors that eventually led to my acceptance to the Genetics and Development program at Cornell University.

Over my years in the Lis lab, I have had the opportunity to interact and learn from an incredibly talented group of scientists. Much more than just colleagues, I have met amazing friends, who have been supportive, generous and encouraging, and contributed to make getting a PhD such a rewarding and fun experience. I am extremely grateful to each and every one of you, and I will always cherish our friendship and the significant role you have all played in my life. In special, I would like to thank Janis Werner, our lab manager, for taking very good care of us and the lab, and for always making sure everything is running smoothly. Abdullah Ozer, my first mentor in the lab, for his patience and guidance, for teaching me a lot of what I know about molecular biology and for always being so ready to share his extensive knowledge. Nick Fuda, for being so generous with his time and for walking me through so many protocols. Iris Jonkers, for being such a dear and close friend and giving critical advice about my projects and career. John Pagano, Leighton Core and Judhajeet Ray, for their guidance and insightful feedback on the various projects I worked on in the lab. Li Yao for her friendship and for bringing so much joy to the early years of my PhD. Lina Bagepalli, for eagerly taking over my aptamer project so we can achieve our ultimate goal of using them in vivo to study HSF function.

I would like to thank my two dear bay mates, Jacob Tome and Jay Mahat, with whom I have established a very close friendship during my time in the lab. They have supported and encouraged me through the most challenging times, shared the most exciting moments and substantially contributed to make the years in the Lis lab some of the best of my life. Jay and I joined the lab at the same time, and we eventually started working on related projects that became the core of both of our theses. His friendship, advice, and fruitful discussions were essential for my success. Jacob and I have shared our trajectories of growth from naïve young students to senior and experienced students that are ready to move on to the next stage. He has participated in every decision I made, from the best way to do an experiment to my ultimate career choices, and his company has brought me a lot of joy and laughter. I am extremely grateful to both of you and our friendship is one of the greatest gifts I have ever received.

Lastly, I would like to express my gratitude to Mike Guertin, whose mentorship was of central importance to the completion of PhD. Before leaving the Lis lab, Mike recruited me to take over his last project, which eventually became the foundation of this dissertation. He not only gave me the opportunity to work on a very exciting project, but he guided me through it and taught me a lot of what I know about computational analysis. Besides being a wonderful mentor and collaborator, we have also established a very close friendship, and Mike, Kris, Evan and Zoey have welcomed me into their lives and became a family away from home.

Besides all the Lis lab members, I would like to thank Kylan Szeto, an honorary member of our lab – from Professor Harold Craighead's lab – that played a major role in the completion of my aptamer project. Kylan's friendship, encouragement and insightful

suggestions were essential during tranquil and challenging times, and his physicist's approach to biology added a new and valuable perspective to how I plan and interpret my experiments. We have also had the chance to work on many experiments together, which were always very fun and productive experiences.

In addition to the Lis lab, I was also very fortunate to be a part of the Department of Molecular Biology and Genetics' community. I would like to thank my classmates, Jae Young Choi, Julia Goodrich, Amanda Yu Guo, Jennifer Apger McGlaughon, Kadeine Campbell Peterson, David Taylor and Kevin Wei, for making the transition into the PhD life such an enjoyable experience and for their friendship and support over the years. In special, I would like to thank Kevin, who has been a close friend since my very first day at Cornell.

During my PhD years, I have established strong friendship bonds with an incredible group of people. These friends became my family in Ithaca, and have been essential to make it feel like home. They were there for me during my happiest and saddest moments, provided support and encouragement and will be forever an integral part of my life. I am extremely grateful to every single one of you and deeply cherish our friendship.

In special, I would like to thank my dear friends Uchita Vaid and Alex Wang, for being the perfect roommates. Sharing a home with them was truly enjoyable, and they have always been full of advice and words of encouragement. Alex was one of my greatest grad school partners. We have shared most of the ups and downs of life as a PhD student, and grown together over the years to become mature scientists. Uchi has always been very inspiring, and has helped me figure out some of my most important decisions. I am also grateful to my dear friend Larissa Di Marzo, for always bringing so

much joy and laughter to even the simplest daily tasks, and for sharing some of my most difficult moments. More than a friend, Lari has become my sister in Ithaca. Bárbara Hufnagel, for being a great friend, always ready to listen and to give precious advice. I would also like to thank Roman Spektor, for playing an essential role during the last year of my PhD. Roman has provided insightful suggestions about my projects and career choices, brought laughter to stressful times and strongly contributed to make writing a thesis a surprisingly fun and tranquil experience. I am also grateful to my dear friends Juliana Magdalon, Jussara Moreira, Renata Polinati and Yin He, for sharing so many fun adventures over my years in Ithaca.

Finally, I would like to thank my wonderful family for making all of this possible. I was immensely lucky to be born into a family of kind, generous and supportive individuals that have always encouraged me to follow my dreams. The role and influence that my family has in my life is so strong that even with the long distance that separated us during my PhD years, it never felt like I was away from home. Words are not enough to express my gratitude to each and every one of you for your love, care, support and encouragement, and for making it possible for me to complete my PhD.

I would like to thank my grandmother Zenita, a strong and amazing woman who has been a source of inspiration and who has always supported me in every possible way. My dear aunt Merô, who played a major role in my education, inspired my love for science, and has always encouraged me to face new challenges. Watching her get sick and leave us over the course of my PhD was the most difficult moment of my life, but it has ultimately been a source of strength and drive and of heightened appreciation for the important things in life. I am also grateful to my aunts Vera, Neuza, Marilda and Gigi, and

to my uncle Dudu, for their unconditional support and dedication. My aunt Marilda has provided for my English education, which made it possible for me to pursue my PhD at Cornell. My aunt Gigi and her husband Brady have been my family in the United States, were extremely helpful during my transition to a foreign country and have always welcomed me with open arms at their home in Connecticut.

I would also like to thank my uncle Lucas. As a successful professor in one of the best universities in Brazil, Lucas was a source of inspiration for my pursuit of a PhD degree, and has been especially invested in my academic career. My cousin and best friend Tati, whom I have deeply admired since I was a kid and who even from a distance is always a constant presence in my life. Tati played a fundamental role during my transition to the United States, and her friendship, support and encouragement were essential during my years in Ithaca. My dear aunt Val, with whom I have a strong connection since the day I was born. Val has always been one of my greatest supporters and advocates, and has always been a reference of love, tenderness and serenity in my life. My little brother André, for being a constant source of joy and laughter and for coming all the way to Ithaca just to attend my thesis defense, bringing our uncle Lucas with him. I am also grateful to my stepfather Wilson, for broadening our horizons and making it possible for me to leave home. And for that I also thank our dear friend Simone, for her love and dedication to my mother. Lastly, I would like to thank the most important and incredible person in my life, my mother Márcia, for being my greatest role model, and for supporting, encouraging and counseling me through every moment of my PhD. Her kindness, wisdom, fortitude and determination inspire everyone around her, and I am exceptionally lucky to have her as a mom.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

CHAPTER 2: TRANSCRIPTION FACTORS GAF AND HSF ACT AT DISTINCT REGULATORY STEPS TO MODULATE STRESS-INDUCED GENE ACTIVATION

CHAPTER 3: SELECTION AND CHARACTERIZATION OF RNA APTAMERS TO STUDY HEAT SHOCK FACTOR FUNCTION AND REGULATION IN VIVO

CHAPTER 4: CONCLUSIONS AND FUTURE DIRECTIONS

LIST OF TABLES

LIST OF ABBREVIATIONS

AD: Activation Domain

CDK9: Cyclin-Dependent Kinase 9

ChIP: Chromatin Immunoprecipitation

CTD: C-Terminal Domain

CYC-T: Cyclin T

DBD: DNA Binding Domain

dHSF: *Drosophila melanogaster* HSF

DSIF: DRB [5,6-dichloro-1-β-D-ribofuranosylbenzimidazole] Sensitivity-Inducing Factor

EDTA: Ethyl-enediaminetetraacetic acid

F-EMSA: Fluorescence Electrophoretic Mobility Shift Assay

FP: Fluorescence Polarization

GAF: GAGA-Associated Factor

GO: Gene Ontology

GRO-seq: Global Run-On sequencing

GTF: General Transcription Factor

HEPES: N-2-hydroxyethylpiperazine-N′-ethanesulfonic acid

HIV: Human Immunodeficiency Virus

HS: Heat Shock

HSE: Heat Shock DNA Element

HSF: Heat Shock transcription Factor

HSP: Heat Shock Protein

K_D: Equilibrium dissociation constant

M1BP: Motif 1 Binding Protein

MEF: Mouse Embryonic Fibroblast

MMLV-RT: Moloney Murine Leukemia Virus Reverse Transcriptase

NELF: Negative Elongation Factor

NHS: Non-Heat Shock

PI: Pausing Index

PIC: Pre-Initiation Complex

Pol II: RNA polymerase II

PRO-seq: Precision nuclear Run-On and sequencing

P-TEFb: Positive Transcription Elongation Factor b

qPCR: quantitative Polymerase Chain Reaction

rNTP: ribonucleoside triphosphate

SDS-PAGE: SDS-Polyacrylamide Gel Electrophoresis

SELEX: Systematic Evolution of Ligands by Exponential Enrichment

Ser2: Serine 2

Ser5: Serine 5

TBE: Tris-Borate-EDTA

TBP: TATA-binding protein

TD: Trimerization Domain

TSS: Transcription Start Site

CHAPTER 1

INTRODUCTION

1.1 Coordinated Regulation of RNA Polymerase II Transcription

The genetic information encoded in segments of an organism's DNA is copied into RNA in a process called transcription, which is executed by large molecular machines called RNA polymerases. In eukaryotes, one of this machines – RNA polymerase II (Pol II) – is responsible for the transcription of all protein-coding genes. Regulation of Pol II activity is fundamental for both cellular homeostasis and for the programmed development of multicellular organisms. The transcription process is regulated by the coordinated action of transcription factors, which can interact with each other, with Pol II and with specific regulatory sequences in the DNA to modulate the distinct steps of the transcription cycle. Initial analyses of the complete human genome estimated the presence of ~2,000 transcription factors (Venter et al. 2001; Lander et al. 2001), and a carefully curated database identified a high-confidence set of 1,391 DNA binding transcription factors (~6% of the total number of protein-coding genes) (Vaquerizas et al. 2009), underscoring the critical role of transcription regulation in human biology.

Only a fraction of the thousands of transcription factors that are involved in the transcription process are true regulatory factors, while the remaining factors are simply 'cogs' in the cycle of transcription (Fuda et al. 2009). A set of sequence elements in the core promoter region directs the binding of General Transcription Factors (GTFs) and the assembly of the Pre-Initiation Complex (PIC) (Juven-Gershon et al. 2008). Specific regulatory transcription factors, which can be activators or repressors, bind to regulatory

sequences in the promoter or at enhancers and ultimately control the status of Pol II. These regulatory factors predominantly execute their functions through interactions with co-activators, which can directly interact with Pol II and GTFs, reorganize nucleosomes, or covalently modify histones and affect the chromatin environment of the gene. Identifying the regulatory transcription factors and the biochemical processes that are controlled by each factor is therefore critical for understanding how transcription is regulated in response to stimuli and during development.

While many of the steps in the transcription cycle can be rate-limiting, one of the challenges in studying transcription regulation is determining the steps that can be regulated by transcription factors in response to signals. The transcription cycle consists of at least eight major steps where transcription can be rate-limiting that can be potentially regulated by transcription factors (Figure 1.1), which include chromatin opening, pre-initiation complex assembly, initiation, promoter-proximal pausing, escape from pausing, productive elongation, termination, and the recycling of the components (Fuda et al. 2009). The transcription cycle starts with the opening of the promoter region, which in some cases can be occluded by nucleosomes (step 1). After nucleosomes are removed, sequence elements in the core promoter region direct the recruitment of GTFs and the assembly of the PIC (step 2). This complex then unwinds the double-stranded DNA around the Transcription Start Site (TSS) so Pol II can engage in active transcription and start synthesizing RNA (step 3). After clearing the core promoter region, Pol II stably transcribes the DNA until it reaches the promoter-proximal pause site, where it pauses while remaining engaged (step 4). Upon receiving the appropriate signals, the paused

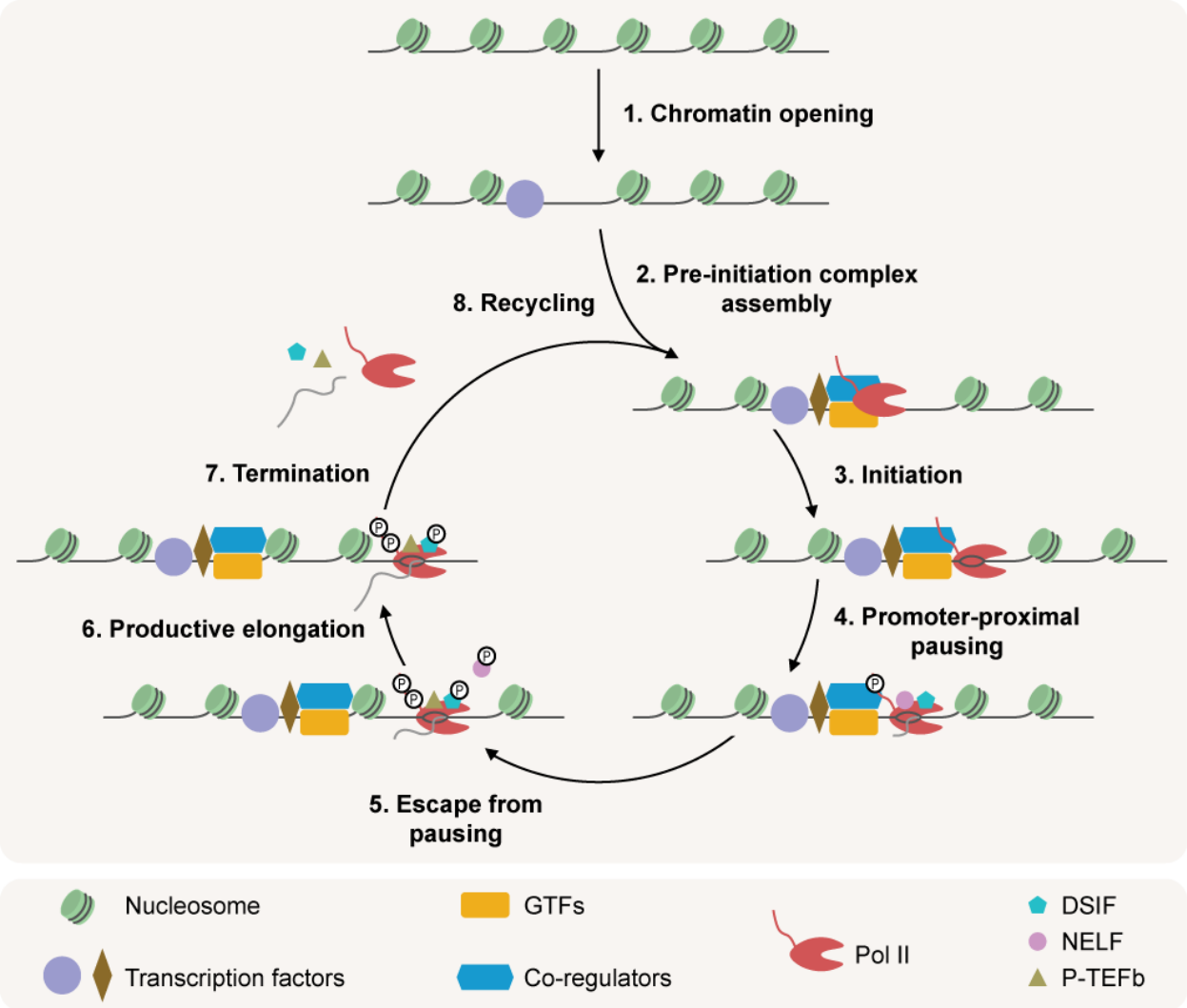


Figure 1.1: Steps in the transcription cycle. Diagram depicting the transcription cycle and its eight major rate-limiting steps. Each step is described in the text. (Adapted from Fuda et al. 2009).

Pol II complex escapes from the pausing region (step 5) and continues to productively elongate through the gene body (step 6). After transcribing the whole gene, Pol II undergoes termination (step 7) and can be recycled to start a new round of transcription (step 8).

1.2 Promoter-Proximal Pausing of Pol II: Major Rate-Limiting Step in the Transcription Cycle

The distribution of Pol II across a gene can serve as an indicator for the steps in the transcription cycle that are rate-limiting for that particular gene. For instance, the absence of Pol II signal across the entire gene body would probably indicate that either chromatin opening (step 1, Figure 1.1) or Pol II recruitment and PIC assembly (step 2, Figure 1.1) is the major rate-limiting step for the regulation of that gene. An even distribution of Pol II signal across the gene would likely suggest that recruitment of Pol II (step 2, Figure 1.1) is the main rate-limiting step. Finally, an accumulation of Pol II signal at the 5' end of the gene, in the promoter-proximal region, would indicate that steps downstream of Pol II recruitment (steps 3-5, Figure 1.1) are rate-limiting.

In the past, transcription regulation studies – performed primarily in bacteria and yeast – have mostly focused on how regulatory signals and transcription factors can control the recruitment and initiation of Pol II, and established these processes as the major rate-limiting steps in the transcription cycle (Ptashne and Gann 1997). However, a myriad of studies have measured Pol II density at individual genes and genome-wide across many species and observed a prominent accumulation of Pol II signal at the 5' end of a significant number of genes (Saunders et al. 2006; Kwak and Lis 2013). These

studies were largely based on Chromatin Immunoprecipitation (ChIP) assays using antibodies against Pol II subunits and nuclear run-on based methods. Pol II ChIP assays, initially focused on individual genes and later performed in a genome-wide manner using either ChIP microarray (ChIP-chip) or ChIP sequencing (ChIP-seq), revealed the enrichment of Pol II in the promoter-proximal region across the genome in *Drosophila melanogaster* and mammalian cells (Muse et al. 2007; Zeitlinger et al. 2007; Guenther et al. 2007). Nevertheless, ChIP using a single Pol II-specific antibody measures all chromatin-associated Pol II, regardless of its transcriptional status, and cannot distinguish between PIC-associated Pol II that has been recruited to the promoter but has not started transcribing yet (step 2, Figure 1.1), Pol II that is in the process of unwinding DNA and initiating transcription (step 3, Figure 1.1), and transcriptionally engaged Pol II that is paused in the promoter-proximal region (step 4, Figure 1.1).

On the other hand, nuclear run-on methods require that the polymerase molecules run-on in vitro in the presence of labeled nucleotides, which are then incorporated into the nascent RNA chain, and therefore only detect polymerases that are transcriptionally engaged. These nuclear run-ons are performed in the presence of high salt or ionic detergent (sarkosyl), which prevents new initiation and removes any chromatin-associated proteins – other than Pol II – that could potentially block the progression of Pol II transcription, enabling the detection of both paused and productively elongating polymerases. By performing nuclear run-ons on a single *Drosophila* gene, Rougvie and Lis (1988) demonstrated that the Pol II molecule at the 5’ end of the gene is transcriptionally engaged, has synthesized a nascent RNA chain of approximately 25 nucleotides in length, but is paused at that position and unable to elongate further into

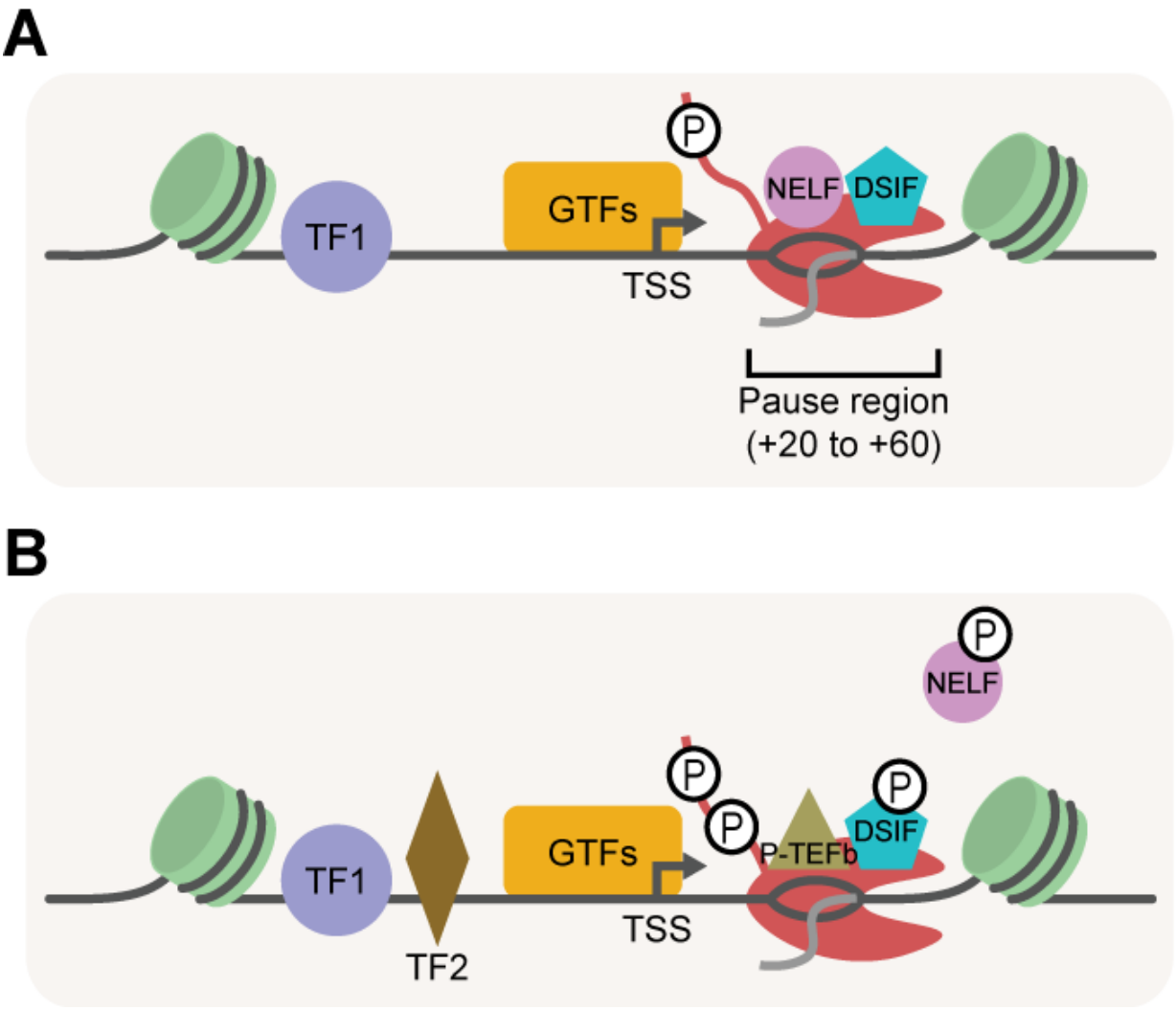
the gene body. Moreover, a more recent study quantitatively compared the outputs of Global Run-On sequencing (GRO-seq) – a nuclear run-on based method that measures the position, levels and orientation of Pol II genome-wide – and ChIP-seq assays in *Drosophila* S2 cells and determined that the vast majority of the Pol II enriched on the promoter-proximal region is transcriptionally engaged (Core et al. 2012). This study also showed that the 5' end accumulated Pol II across the genome can only run on in the presence of sarkosyl, while productively elongating Pol II in the gene bodies do not require the detergent, indicating that promoter-proximal Pol II is indeed tethered in a paused state. Finally, several genome-wide studies have demonstrated that the enrichment of transcriptionally engaged Pol II at the 5' end of genes is widespread in the *Drosophila* and mammalian genomes (Core et al. 2008; Larschan et al. 2011; Min et al. 2011; Core et al. 2012). Taken together, these results revealed the presence of paused, transcriptionally engaged Pol II at the 5' end of the majority of genes in metazoans and established promoter-proximal pausing as a major rate-limiting step in the transcription cycle.

The escape of promoter-proximally paused Pol II into productive elongation is often a highly regulated step. Indeed, genes where paused Pol II was initially discovered, such as heat shock protein genes and c-Myc, are strongly regulated in response to stimuli (heat shock and serum stimulation, respectively), which promote the escape of the paused Pol II into productive elongation (Saunders et al. 2006). In general, promoter-proximal pausing is highly prevalent among genes associated with important signal-responsive pathways, including development, immunological signaling, cell proliferation and

environmental stress responses (Adelman et al. 2009; Levine 2011; Adelman and Lis 2012).

Two main protein complexes act to stabilize paused Pol II in the promoter-proximal region: DSIF (DRB [5,6-dichloro-1-β-D-ribofuranosylbenzimidazole] sensitivity-inducing factor), a two subunit complex composed of Spt4 and Spt5 (Wada et al. 1998), and NELF (Negative Elongation Factor), a multiprotein complex composed of four subunits, A, B, C/D, and E (Yamaguchi et al. 1999; Narita et al. 2003). DSIF and NELF physically interact with each other and with Pol II in the promoter-proximal region and are important to hold the Pol II molecule in the paused state (Yamaguchi et al. 1999) (Figure 1.2A).

P-TEFb (Positive Transcription Elongation Factor b), a two subunit complex composed of Cyclin T (CYC-T) and Cyclin-Dependent Kinase 9 (CDK9), is a protein kinase with a major role in promoter-proximal pause release (Marshall and Price 1995; Price 2000; Ni et al. 2008). P-TEFb overcomes this rate-limiting step by phosphorylating NELF (Fujinaga et al. 2004), DSIF (Kim and Sharp 2001; Yamada et al. 2006), and the Pol II C-Terminal Domain (CTD) at Serine 2 (Ser2) (Ramanathan et al. 2001) (Figure 1.2B), and inhibition of P-TEFb results in a genome-wide decrease in transcription levels (Chao and Price 2001; Jonkers et al. 2014). P-TEFb can be recruited to the promoter-proximal region by regulatory transcription activators to promote the release of paused polymerase and activate the transcription of the gene (TF2 in Figure 1.2B). The transcription activators c-Myc (Eberhardy and Farnham 2002; Kanazawa et al. 2003), NF-κB (Barboric et al. 2001; Nowak et al. 2008), and the Human Immunodeficiency Virus (HIV) TAT transactivator (Price 2000) have all been shown to physically interact with



<u>Figure 1.2:</u> Establishment and release of promoter-proximal paused Pol II. (A) The pausing factors DSIF and NELF act to stabilize the paused Pol II molecule in the pausing region (usually between +20 to +60 nt from the TSS). TF1 represents a class of transcription factors that act upstream of pausing formation. **(B)** The protein kinase P-TEFb is recruited by transcription activators (TF2) and phosphorylates NELF, DSIF and Pol II's CTD at Ser2 to promote the escape of paused Pol II into productive elongation.

P-TEFb, and the heat shock transcription activator HSF has been shown to indirectly recruit P-TEFb upon stress induction (Lis et al. 2000).

While pausing factors such as NELF and DSIF are important to stabilize the paused Pol II complex at the promoter-proximal region, a distinct class of factors, which act upstream of pausing establishment, as well as sequence elements in the promoter region also play an essential role in enabling Pol II pausing (Kwak et al. 2013). Furthermore, chromatin opening and initiation processes have been shown to be spatially and temporally connected with early elongation mechanisms that lead to the formation of paused Pol II (Kwak and Lis 2013). For instance, an open chromatin environment is essential for Pol II recruitment and initiation and the consequent establishment of pausing. Transcription activators and other DNA binding proteins, such as GAGA-Associated Factor (GAF) in *Drosophila*, can bind to the promoter region and promote an open chromatin environment around the TSS (TF1 in Figure 1.2A). GAF can interact with several nucleosome remodeler complexes, including NURF, ISWI, and PBAP, and displace adjacent nucleosomes to generate accessible DNA regions (Tsukiyama et al. 1994; Tsukiyama and Wu 1995; Okada and Hirose 1998; Nakayama et al. 2012; Fuda et al. 2015).

1.3 Regulation of Transcription Activation at Distinct Steps of the Transcription Cycle

As discussed previously, the transcription cycle is composed of multiple rate-limiting steps that can be potentially activated by regulatory transcription factors (Fuda et al. 2009). Certain activators, such as the *Drosophila* GAF described above, act at early stages in

the cycle by facilitating nucleosome displacement (Tsukiyama et al. 1994; Tsukiyama and Wu 1995; Okada and Hirose 1998; Nakayama et al. 2012; Fuda et al. 2015). The mammalian transcription factors Sp1 (Blau et al. 1996) and E2 (Danko et al. 2013) can stimulate transcription initiation, while factors that recruit P-TEFb, such as c-Myc (Eberhardy and Farnham 2002; Kanazawa et al. 2003) and HSF (Lis et al. 2000), act at a later stage (pause escape) to promote the activation of target genes.

Each mode of regulation can serve distinct functions and could have evolved to achieve distinct goals depending on the level of control and speed of response that are required for regulating the transcription of different sets of genes. For instance, regulation at the level of chromatin opening could be important to maintain tissue specific genes in a tight repressive state in the cell types where they are not expressed. On the other hand, regulation at the promoter-proximal pausing step can enable a rapid and synchronous activation of developmentally and environmentally regulated genes (Adelman and Lis 2012).

While many genes display only one major mode of regulation, there are examples of genes whose activation is controlled at more than one step in the transcription cycle. This can be achieved by a sole transcription activator that acts at distinct steps in the cycle of by the action of two or more factors with distinct roles. In these cases, the coordinated action of the key transcription activators (or of the distinct domains of the same activator) is essential for the proper expression of the target gene. The herpes simplex virus activator VP16 has been shown to stimulate both initiation and elongation steps (Blau et al. 1996), and HSF's role in inducing gene body nucleosome loss in

Drosophila is independent from its role in promoting transcription of the *Hsp70* gene (Petesch and Lis 2008).

According to the kinetic synergism model, which was proposed over 20 years ago by Herschlag and Johnson (1993) and recently quantitatively refined by Scholes and colleagues (2017), the combined activation of two or more slow or inefficient steps can lead to a much more rapid and robust increase in gene expression. This synergism can be important to coordinate responses to distinct types of signals or to promote large responses to small changes in the concentrations of the active forms of transcription activators (Herschlag and Johnson 1993). Furthermore, the synergistic effect of two or more transcription factors acting at distinct steps can be especially relevant at genes where promoter-proximal pausing is observed. At these genes, the action of transcription factors that enable the establishment of Pol II pausing is just as necessary as the role of activators that promote the escape from pausing. In other words, a pausing release factor can only execute its function if Pol II pausing had been previously established at the promoter-proximal region; and, transcription factors that act at very early stages require the subsequent release of paused Pol II to fully activate the gene.

1.4 Transcriptional Heat Shock Response as a Model to Study Mechanisms of Transcription Regulation

Inducible systems are an invaluable resource when studying mechanisms of transcription regulation. The regulatory processes can be triggered instantaneously with a specific signal, enabling the tracking of the recruitment kinetics and location of relevant factors and of the ordered mechanistic events that result in the activation or repression of target

genes. As one of the most effective inducible transcription systems, the transcriptional Heat Shock (HS) response in *Drosophila melanogaster* has been widely used to study mechanisms of transcription and its regulation, and many important aspects of transcription regulation were discovered using this model system (Guertin et al. 2010).

The HS response is a highly conserved protective mechanism that responds to elevated temperatures or other forms of stress through the production and accumulation of molecular chaperones, the Heat Shock Proteins (HSPs), which help the cell to cope with stress-induced protein aggregation and misfolding (Lindquist 1986; Lindquist and Craig 1988). This response was first observed by Ritossa in 1962 through the observation of new heat induced puffs on the polytene chromosomes from salivary glands of the fruit fly *Drosophila busckii* (Ritossa 1962). Since this initial observation, a whole new research field has emerged, and a plethora of studies have investigated and characterized many different aspects of the response, providing a very extensive understanding of the mechanisms involved in its regulation.

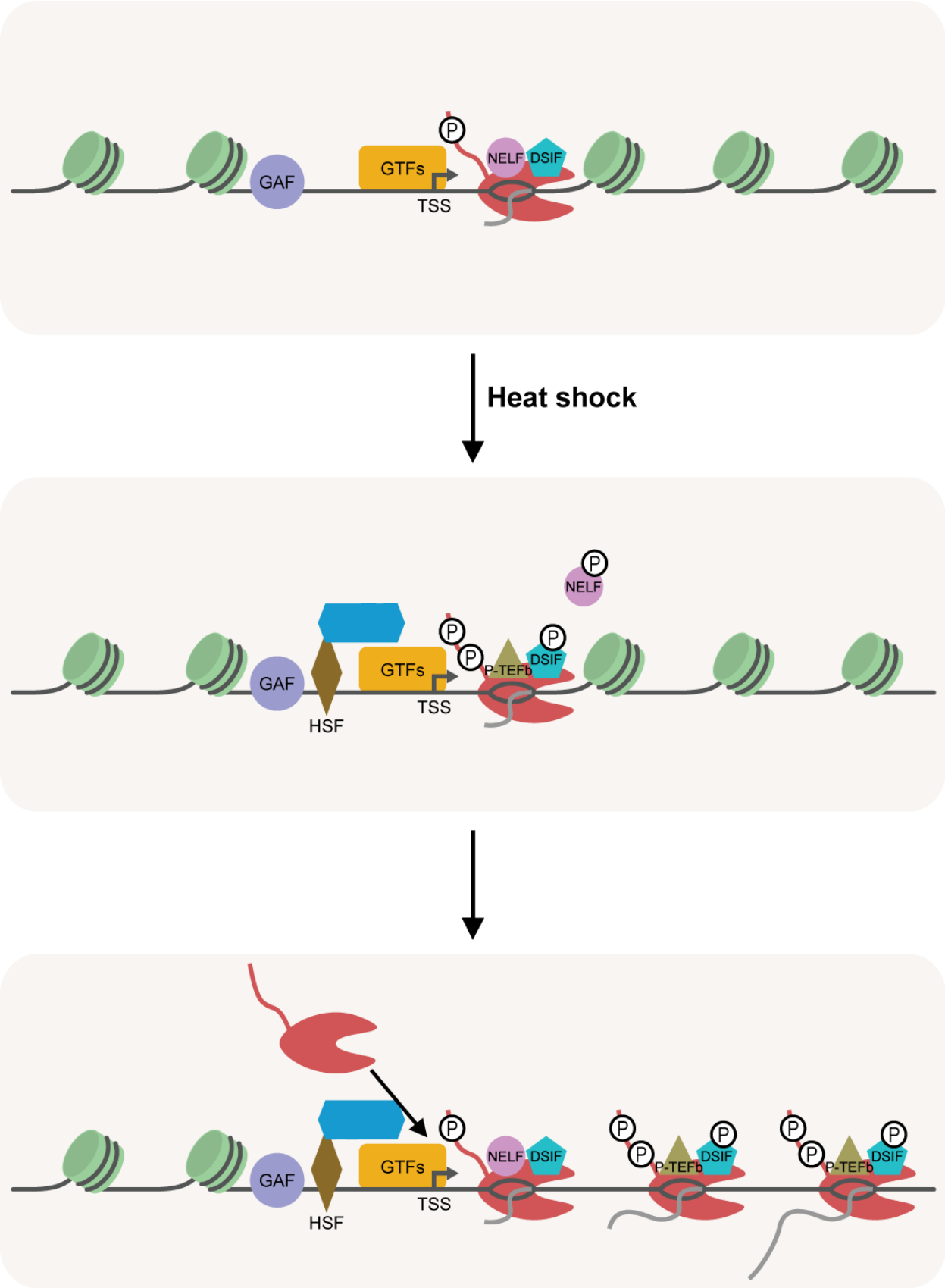
The HS response can be regulated at multiple levels, including transcription, mRNA processing and stability, and translation (Lindquist and Craig 1988; Shalgi et al. 2014). From the very beginning, the transcriptional HS response in *Drosophila melanogaster* has been extensively used as a model to investigate how genes are structured and regulated. For instance, the *Drosophila* HSP genes were among the first eukaryotic genes to be cloned (Schedl et al. 1978; Livak et al. 1978; Craig et al. 1979; Moran et al. 1979; Artavanis-Tsakonas et al. 1979), to have their chromosomal localization and orientation determined (Livak et al. 1978; Moran et al. 1979; Artavanis-Tsakonas et al. 1979; Mirault et al. 1979; Ish-Horowicz et al. 1979; Holmgren et al. 1979;

Wadsworth et al. 1980; Corces et al. 1980; Craig and McCarthy 1980; Voellmy et al. 1981), to have their general chromatin structure characterized before and after heat shock induction (Wu et al. 1979a, 1979b; Wu 1980; Keene et al. 1981), to have their regulatory regions defined (Pelham 1982; Pelham and Bienz 1982), and to have the transcription factor that interacts with these regulatory regions identified (Parker and Topol 1984; Wu 1984, 1985). This factor, which was later named HS transcription Factor (HSF) is considered the master regulator of the HS response from yeast to humans (Wu 1995).

After these initial findings, numerous research groups have been unraveling important details of how the transcriptional HS response is regulated. Among these groups, the Lis laboratory at Cornell University stands out with its major contributions to this field, including the definition of the conserved sequence element that is bound by HSF (Xiao and Lis 1988), the discovery that HSF binds to DNA as a trimer (Perisic et al. 1989), and the first demonstration of promoter-proximal Pol II pausing through studies of the *Drosophila* HS gene *Hsp70* (Gilmour and Lis 1986; Rougvie and Lis 1988). As discussed previously, Pol II pausing later emerged as a major rate-limiting step in the regulation of the majority of genes in metazoans (Adelman and Lis 2012).

Most of our knowledge of HS transcriptional regulation comes from extensively studying the model HS gene *Hsp70* (Guertin et al. 2010) (Figure 1.3). *Hsp70* maintains a promoter-proximally paused Pol II molecule 20-40 bp downstream of the TSS that is released to transcribe the gene at a low level during normal non-stress conditions (Rougvie and Lis 1988; Rasmussen and Lis 1993). The transcription factor GAF (purple circle, Figure 1.3) is bound to the promoter of *Hsp70* prior to HS, and GAF is important

Figure 1.3: Heat shock transcription activation of *Drosophila*'s model HS gene *Hsp70*. Prior to heat shock induction, *Hsp70* maintains a paused, transcriptionally engaged Pol II molecule 20-40 bp downstream of the TSS. The pausing complexes DSIF and NELF are associated with paused Pol II, which is phosphorylated at Ser5 of its CTD. GAF is bound to the promoter at this stage. After heat shock induction, HSF binds to the promoter and recruits P-TEFb, which phosphorylates DSIF, NELF and Pol II's CTD at Ser2 to promote the escape of paused Pol II into productive elongation.


GAF
GTFs
TSS
NELF
DSIF
Heat shock
HSF
P-TEFb

for the establishment and stability of paused Pol II (Lee et al. 1992; Wang et al. 2005; Lee et al. 2008; Kwak et al. 2013). As discussed previously, GAF has a key role in keeping the promoter region open and free of nucleosomes (Tsukiyama et al. 1994; Fuda et al. 2015), which allows the recruitment of general transcription factors and the initiation of transcription by Pol II.

The pausing complexes DSIF (blue pentagon, Figure 1.3) and NELF (pink circle, Figure 1.3) play an important role in mediating Pol II pausing at *Hsp70* (Wu et al. 2003, 2005). Both DSIF and NELF are associated with the promoter region of *Hsp70* before heat shock and colocalize with paused Pol II (Wu et al. 2003, 2005). Individual depletions of DSIF and of one of the NELF subunits reduced the levels of paused Pol II in the promoter-proximal region of *Hsp70* (Wu et al. 2003, 2005). After heat shock induction, NELF dissociates from Pol II, while DSIF remains associated with the elongating polymerase (Wu et al. 2003, 2005).

Paused Pol II is phosphorylated at Serine 5 (Ser5) of Pol II's CTD repeats (black P, Figure 1.3) (Boehm et al. 2003). This phosphorylation is executed by the kinase subunit CDK7 of the TFIIH complex, and a temperature-sensitive mutant of CDK7 decreased *Hsp70*'s pausing levels at non-permissive temperatures (Schwartz et al. 2003).

Upon HS induction, HSF trimerizes, is activated, and is rapidly recruited to the promoter (brown diamond, Figure 1.3), where it binds to its cognate HS DNA Elements (HSEs) (Xiao and Lis 1988). After binding, HSF directly and indirectly recruits co-activators and other factors (blue hexagon, Figure 1.3) (Lis et al. 2000; Saunders et al. 2003; Ardehali et al. 2009) that affect the chromatin structure and composition, and promote the release of Pol II from the paused complex into productive elongation. This

transition from the paused state into productive elongation depends critically on HSF's indirect recruitment of the positive elongation factor P-TEFb (green triangle, Figure 1.3), which phosphorylates DSIF, NELF and Pol II's CTD at Ser2 (black P, Figure 1.3) (Lis et al. 2000; Boehm et al. 2003). Upon heat shock induction, new Pol II molecules are rapidly recruited to *Hsp70* (Zobeck et al. 2010). Although these newly recruited Pol II undergo pausing (Figure 1.3), their rate of release into productive elongation under heat shock conditions is ~100-fold higher than the rate of release under normal uninduced conditions (Lis 1998; Buckley et al. 2014). This molecular cascade of events leads to a massive production of *Hsp70* mRNA.

Although the transcriptional HS response is well characterized for the *Drosophila Hsp70* gene, we lack a comprehensive characterization of the genome-wide changes in transcription. Previous studies have mapped HSF binding sites during normal growth conditions and after HS, and observed that HSF recruitment to a promoter is neither necessary nor sufficient to direct HS gene activation (Trinklein et al. 2004; Guertin and Lis 2010; Gonsalves et al. 2011). Nonetheless, the rules governing the specificity of activation and repression across the *Drosophila* genome remain incomplete. Transcriptional changes after HS have also been measured in *Drosophila* and other organisms (Leemans et al. 2000; Guhathakurta et al. 2002; Murray et al. 2004; Trinklein et al. 2004; Sørensen et al. 2005; Gonsalves et al. 2011; Vihervaara et al. 2013); however, these studies were limited in resolution both temporally and spatially by measuring steady-state levels of mature mRNA. Furthermore, measurement of mRNAs cannot distinguish effects on mRNA stability (Lindquist and Petersen 1990) and pre-mRNA processing (Yost and Lindquist 1986; Shalgi et al. 2014) from transcription, or primary

from secondary effects of the HS response. Therefore, a thorough characterization of the affected genes using a high resolution assay that measures direct changes in transcription is necessary to determine the generality and diversity of the roles of transcription factors such as GAF and HSF in the HS response and define the steps in the transcription cycle that are regulated by these factors.

1.5 Strategy for Dissecting the Function and Mechanisms of Action of Transcription Factors In Vivo

As discussed in the previous sections, major rate-limiting and regulated steps in the transcription cycle have been established. However, our understanding of the direct and indirect interactions between transcription factors, DNA and Pol II that underlie the regulation of these steps remain incomplete. Therefore, our comprehension of transcription regulation would immensely benefit from a comprehensive in vivo investigation of the steps in the transcription cycle that each transcription factor regulates.

As explained in section 1.4, a powerful strategy to study mechanisms of transcription regulation is to observe transcription before and after rapid induction of changes in gene expression by a specific signal, such as heat shock (Figure 1.4). By employing this strategy in normal cells and in cells where the function of a specific transcription factor has been perturbed, one can dissect the roles of this factor in the transcriptional response to the given signal (Figure 1.4). Furthermore, performing these observations in a genome-wide manner provides a comprehensive characterization of the roles of transcription factors and the statistical power to assess their mechanisms of transcription regulation.

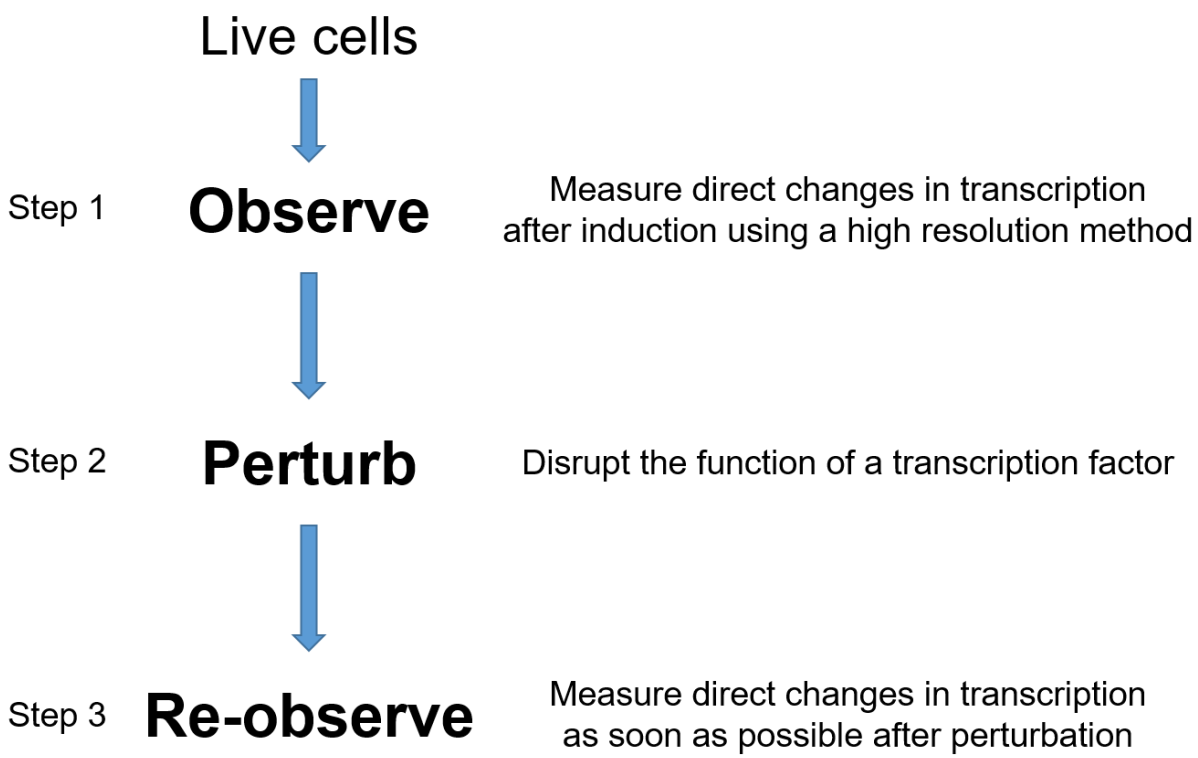


Figure 1.4: Strategy for dissecting the function and mechanisms of action of transcription factors in vivo. Outline of the strategy. A detailed description of the 3 steps is provided in the text.

One of the most critical challenges in the implementation of this strategy is selecting a method that can measure direct changes in transcription genome-wide with high spatial and temporal resolution. Pol II distribution can be tracked using a variety of high resolution methods (Gilmour and Fan 2009; Churchman and Weissman 2011; Rhee and Pugh 2012), including the robust Precision nuclear Run-On and sequencing (PRO-seq) method (Kwak et al. 2013), a base-pair resolution version of the GRO-seq method discussed previously (Core et al. 2008). PRO-seq maps the precise locations of the active sites of all transcriptionally engaged RNA polymerase complexes by affinity-purification and sequencing of nascent RNAs after a terminating biotin-NTP is incorporated to the 3' end of RNA during a nuclear run-on (Kwak et al. 2013). The density of sequencing reads is proportional to the number of transcriptionally-engaged polymerase molecules present at each position when the nuclei were isolated. PRO-seq has base pair resolution, is strand specific, and is not affected by the background levels of accumulated RNAs (Kwak et al. 2013). Therefore, this method can be used to query the genome-wide distribution of transcriptionally-engaged RNA polymerases before and soon after rapid induction with high spatial and temporal resolution (step 1 of strategy, Figure 1.4).

The second major challenge in studying the mechanisms of action of transcription factors is identifying the best suited method to perturb the function of a given factor (step 2 of strategy, Figure 1.4). A variety of methods can be used to deplete or perturb protein function, each one with its advantages and disadvantages (Table 1.1). The vast majority of studies have used traditional methods such as gene knockouts, genetic mutations, and RNAi knockdowns to investigate the roles of protein factors in biological processes. One

Table 1.1: Methods for depleting of perturbing protein function. A list with the most used methods with its main advantages and disadvantages is provided in the table.

Method	Advantages	Disadvantages
Knockouts/Mutations	Many are already available or can be generated by genome editing	Potential for secondary or compensatory effects
RNAi knockdown	Relatively easy and general	Potential for secondary or compensatory effects
Drugs	High specificity, immediate action	Limited availability for inhibiting macromolecular interactions
RNA aptamers	High specificity, can inhibit macromolecular interactions	Somewhat laborious to select and characterize

main disadvantage associated with these methods is the inability to distinguish between primary and secondary or compensatory effects of depleting or modifying the factor. Ideally, one would use a method that can act quickly to inhibit a particular domain of a transcription factor and disrupt specific macromolecular interactions, allowing the dissection of its primary functions on mechanisms of transcription regulation. Inhibitory RNA aptamers (Table 1.1) fit these criteria, and will be discussed in detail in section 1.6.

After disrupting the function of transcription factors using one of the approaches described in Table 1.1, the same genome-wide method used in step 1 can be used to re-observe the direct changes in transcription that happen upon induction when a given factor is perturbed (step 3 of strategy, Figure 1.4). This strategy will reveal the role of transcription factors in the activation or repression of target genes. Furthermore, by using a method that tracks the genome-wide distribution of transcriptionally engaged Pol II at base pair resolution, one can observe the effects of such perturbations on Pol II as it progresses through the steps of the transcription cycle. Although this approach is being

discussed in the context of transcription regulation, the “observe, perturb, re-observe” strategy can be used as a general method to dissect the function of protein factors in biological processes.

1.6 Inhibitory RNA Aptamers as Tools to Dissect the Primary Functions of Transcription Factors in Mechanisms of Transcription Regulation

As discussed in the previous section, the most widely used methods to study protein function (gene knockouts, genetic mutations, and RNAi knockdowns) can result in secondary or compensatory effects that obscure the primary effects of the perturbation and confound the interpretation of results. Furthermore, when using such methods, the distinction between the function of an individual domain rather than the entire protein – which may have multiple domains and functions – is often unclear. As an alternative to overcome these challenges, highly specific inhibitory RNA aptamers that target different surfaces of the protein of interest can be quickly expressed in vivo and used to block macromolecular interactions of a specific protein domain.

Aptamers (Ellington and Szostak 1990) are single-stranded oligonucleotides – DNA, RNA or modified nucleic acids – that can fold into diverse and intricate three-dimensional structures that bind proteins, peptides or small molecules with high affinity. Equilibrium dissociation constants (K_D) typically range from micromolar to picomolar values. Aptamers are selected in vitro using an iterative process called SELEX (Systematic Evolution of Ligands by Exponential Enrichment) (Ellington and Szostak 1990; Tuerk and Gold 1990). In an RNA aptamer SELEX, one starts with an enormously large library of randomized sequences – typically on the order of 10^{14}-10^{15} unique RNA

molecules – that have diverse structures based on sequence variation to identify those that can bind specifically to a target of interest (Figure 1.5). After incubation with the target, the bound species are separated from the unbound ones (partition step), reverse transcribed, PCR amplified and transcribed, and the obtained pools are submitted to further rounds of selection (Figure 1.5). After many rounds, the aptamers that bind with high affinity to the target are usually enriched and dominate the pool of sequences.

Aptamers can be applied in numerous ways to address various biological and technical questions. Of particular relevance, they can act as inhibitors that bind to a protein surface and disrupt specific interactions or functions. Indeed, the Lis laboratory has successfully used inhibitory RNA aptamers in the past to study macromolecular interactions in vitro and in vivo (Shi et al. 1999; Fan et al. 2004, 2005; Zhao et al. 2006; Shi et al. 2007; Sevilimedu et al. 2008; Salamanca et al. 2011, 2014). Shi and colleagues generated transgenic flies that express a multivalent high affinity aptamer to the *Drosophila* SR protein B52 and showed that this aptamer can reverse a variety of phenotypes that are caused by B52 overexpression. Furthermore, this aptamer can inhibit B52-stimulated pre-mRNA splicing in *Drosophila* nuclear extracts (Shi et al. 1999). Fan and colleagues demonstrated that aptamers can act in distinct modes to inhibit the interaction of the yeast TATA-binding protein (TBP) with the TATA DNA element in preformed, higher-order complexes containing the additional general transcription factors TFIIB and TFIIA (Fan et al. 2004). Finally, Salamanca and colleagues generated transgenic flies that express a multivalent aptamer that binds to HSF with high affinity and showed that this aptamer inhibits the expression of HS protein genes and also reverses the phenotypes caused by HSF overexpression (Salamanca et al. 2011). Moreover, when

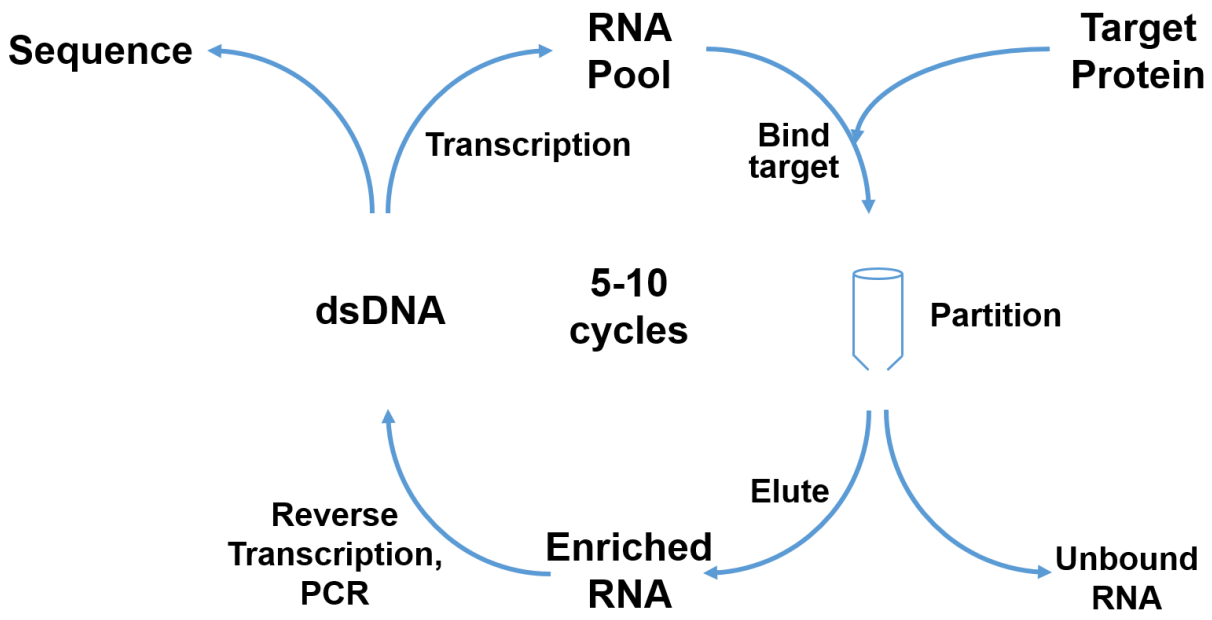


Figure 1.5: In vitro selection of RNA aptamers. Diagram depicting the SELEX procedure and its major steps. Each step is described in the text.

expressed in human cells, this aptamer inhibited binding of HSF to its regulatory DNA elements, induced apoptosis and abolished the colony-forming capability of cancer cells (Salamanca et al. 2014).

Although these studies were highly successful in using RNA aptamers to inhibit protein interactions, there were limitations on the methodology used to select those aptamers, such as the SELEX procedure and the size and quality of the RNA library. Furthermore, multivalent versions of these aptamers had to be constructed to generate aptamers that bind to the target proteins with low nanomolar K_D. More recently, we have significantly improved our methodology by generating a more complex, well-characterized library, developing a more efficient multiplex SELEX procedure that allows the selection of aptamers for many targets at the same time, and using high-throughput sequencing (as opposed to traditional cloning) to analyze the selected pools, which reduces the number of rounds that need to be performed to allow the identification of enriched sequences (Latulippe et al. 2013; Szeto et al. 2014). Furthermore, we have developed a high-throughput assay that can measure the binding affinity of thousands of aptamers in one single experiment, allowing the characterization of the complete set of sequences from a SELEX pool and the identification of the best candidate aptamers for in vivo expression (Tome et al. 2014).

With these improved aptamer selection/characterization technologies, we can generate a collection of inhibitory RNA aptamers that bind with high affinity to distinct surfaces of transcription factors and inhibit their macromolecular interactions. We can then generate transgenic cell lines that express these aptamers under the control of an inducible promoter, enabling the investigation of the primary effects of inhibiting a specific

protein surface shortly after the induction of aptamer expression. This approach can be potentially used as a general, powerful and innovative strategy to study the function of specific protein surfaces/domains in any biological process.

1.7 Research Strategy and Dissertation Outline

The overarching goal of this dissertation was to comprehensively characterize the roles of transcription factors and define the specific steps in the transcription cycle that are regulated by each factor. We focused on the transcription factors HSF and GAF in the context of heat shock induction. As discussed in section 1.4, the transcriptional heat shock response has been used as an effective model to study mechanisms of transcription regulation. Although the roles of HSF and GAF at classical heat shock genes have been characterized, their general roles at distinct steps of the transcription cycle remain unclear, and a genome-wide investigation of these factors would provide the statistical power to assess mechanisms of transcription regulation. To achieve our goals, we implemented the “observe, perturb, re-observe” strategy described in section 1.5, and here we demonstrate the power of this strategy to study the function of specific proteins in a biological process.

For the first step of our strategy (“observe”, Figure 1.6), we used PRO-seq to comprehensively characterize the direct changes in transcription that happen in *Drosophila* S2 cells after a time-course of heat shock induction (Chapter 2). We show that the HS response is rapid and pervasive, with thousands of genes being repressed after 20 minutes of HS and hundreds of genes being activated; moreover, the activated genes are not limited to the classical HSP genes. Promoter-proximal pausing is highly prevalent

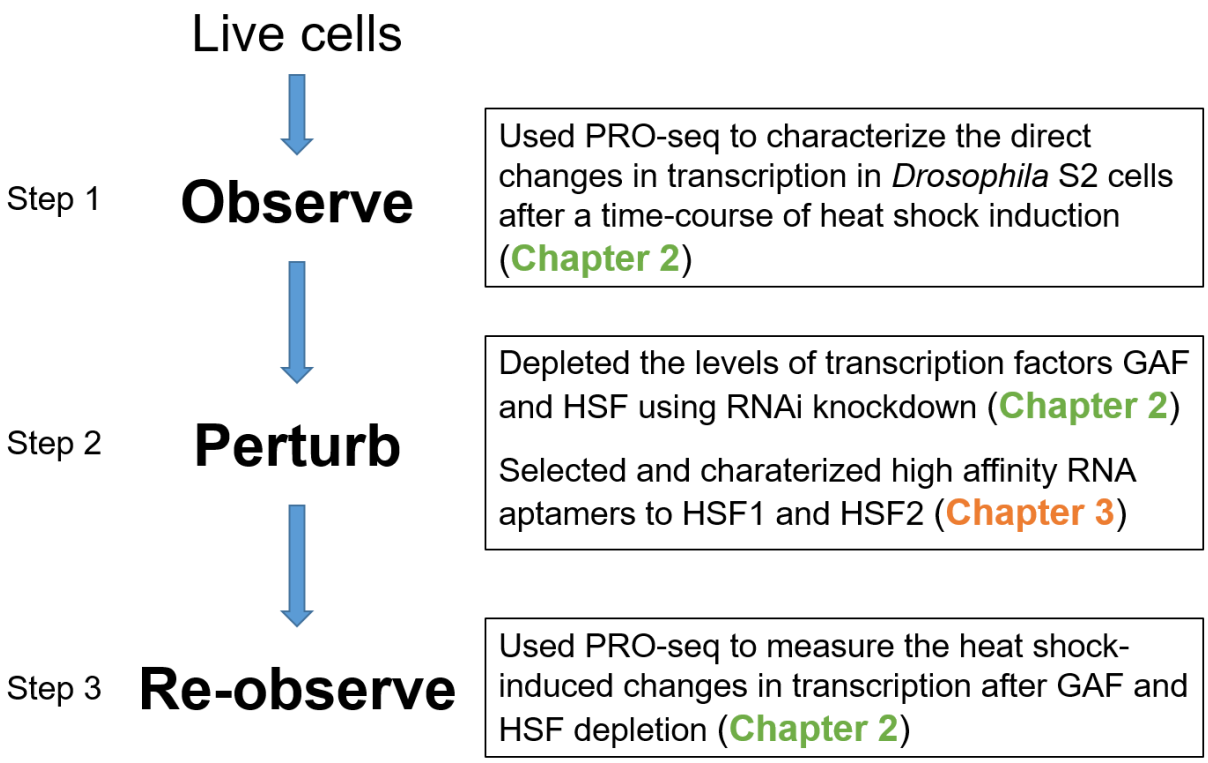


Figure 1.6: Research strategy. Outline of the dissertation's research strategy.

among the activated genes prior to HS, while the repressed class is regulated at the level of transcription initiation.

For the second step of our strategy ("perturb", Figure 1.6), we depleted the levels of HSF and GAF in *Drosophila* S2 cells using RNAi knockdowns (Chapter 2). Furthermore, as a complement to traditional strategies to study protein function, we selected RNA aptamers to the human HSF1 and HSF2 with the goal of expressing these aptamers in vivo to disrupt specific macromolecular interactions of these transcription factors. HSF1 is the human counterpart of the *Drosophila* HSF and HSF2 is a closely related transcription factor with overlapping and distinct functions. In Chapter 3, we describe the results of a successful RNA aptamer selection to HSF1 and HSF2 using our recently developed SELEX reagents and technologies. Moreover, we report a thorough characterization of the sequenced pools from these selections and the development and implementation of a set of SELEX performance metrics that can be used to evaluate the success of a selection. We plan to express the selected aptamers in vivo to dissect the roles of HSF in transcription regulation.

For the last step of our strategy ("re-observe", Figure 1.6), we re-observed the direct changes in transcription that happen upon heat shock in *Drosophila* S2 cells after depleting the levels of GAF and HSF (Chapter 2). We demonstrate that the establishment of promoter-proximal Pol II pausing on a subset of HS activated genes is dependent on GAF binding upstream and proximal to the TSS. Moreover, GAF depletion abrogates pausing and consequently impairs HS activation, indicating that this step in early transcription elongation is essential for gene activation. We also show that the recently identified transcription factor Motif 1 Binding Protein (M1BP) (Li and Gilmour 2013) has a

role in pausing and HS activation of a subset of genes that exhibit GAF-independent pausing. Furthermore, we demonstrate that only a relatively small fraction of HS activated genes are regulated by HSF, and HS activation of these HSF-dependent genes is regulated at the level of pausing release.

CHAPTER 2

TRANSCRIPTION FACTORS GAF AND HSF ACT AT DISTINCT REGULATORY STEPS TO MODULATE STRESS-INDUCED GENE ACTIVATION[1]

2.1 Introduction

The Heat Shock (HS) response in *Drosophila melanogaster* has been an effective model system to discover and study mechanisms of transcription and its regulation (Guertin et al. 2010). This highly conserved protective mechanism (Lindquist and Craig 1988) is regulated at the transcriptional level by the Heat Shock transcription Factor (HSF) (Wu 1995). When activated by stress, HSF potently activates expression of HS genes, resulting in the accumulation of molecular chaperones, the Heat Shock Proteins (HSPs), which helps the cell cope with stress-induced protein aggregation and misfolding (Lindquist and Craig 1988).

The transcriptional HS response has been studied largely using *Hsp70* as a model gene (Guertin et al. 2010). *Hsp70* maintains a promoter-proximally paused RNA Polymerase II (Pol II) molecule 20-40 bp downstream of the Transcription Start Site (TSS) that is released to transcribe the gene at a low level during normal non-stress conditions (Rougvie and Lis 1988; Rasmussen and Lis 1993). The transcription factor GAGA Associated Factor (GAF) is bound to the promoter of *Hsp70* prior to HS, and GAF is important for the establishment and stability of paused Pol II (Lee et al. 1992, 2008; Kwak et al. 2013). GAF has a key role in keeping the promoter region open and free of

---

[1]The contents of this chapter, including all the figures, have been published in Duarte et al. 2016. Genes Dev 30: 1731–46. PMID: 27492368.

nucleosomes (Tsukiyama et al. 1994; Fuda et al. 2015), which allows the recruitment of general transcription factors and the initiation of transcription by Pol II. Upon HS induction, HSF trimerizes and is rapidly recruited to the promoter, where it binds to its cognate HS DNA Elements (HSEs) (Xiao and Lis 1988). After binding, HSF directly and indirectly recruits co-activators and other factors (Lis et al. 2000; Saunders et al. 2003; Ardehali et al. 2009) that affect the chromatin structure and composition, and promotes the release of Pol II from the paused complex into productive elongation. This transition from the paused state into productive elongation depends critically on the positive elongation factor P-TEFb, and has been shown to be a very general step that is essential for the regulation of virtually all genes across different species (Rahl et al. 2010; Jonkers et al. 2014). The net result of this molecular cascade is an increase in transcription levels that can be ~200-fold for some of the HS-regulated genes (Lis et al. 1981).

Although the independent mechanisms of promoter-proximal pausing and escape to productive elongation have been well studied in the context of HS activation of *Hsp70*, we lack a comprehensive characterization of the genome-wide changes in transcription that result from HS in *Drosophila*. A thorough characterization of the affected genes is necessary to determine the generality and diversity of the roles of transcription factors such as GAF and HSF in the HS response and provide the statistical power to assess mechanisms of transcription regulation.

Previous studies have mapped HSF binding sites during normal growth conditions and after HS, and observed that HSF recruitment to a promoter is neither necessary nor sufficient to direct HS gene activation (Trinklein et al. 2004; Guertin and Lis 2010; Gonsalves et al. 2011). Nonetheless, the rules governing the specificity of activation and

repression across the *Drosophila* genome remain incomplete. Transcriptional changes after HS have also been measured in *Drosophila* and other organisms (Leemans et al. 2000; Guhathakurta et al. 2002; Murray et al. 2004; Trinklein et al. 2004; Sørensen et al. 2005; Gonsalves et al. 2011; Vihervaara et al. 2013); however, these studies were limited in resolution both temporally and spatially by measuring steady-state levels of mature mRNA. Furthermore, measurement of mRNAs cannot distinguish effects on mRNA stability (Lindquist and Petersen 1990) and pre-mRNA processing (Yost and Lindquist 1986; Shalgi et al. 2014) from transcription, or primary from secondary effects of the HS response.

To overcome these limitations, we queried the genome-wide distribution of transcriptionally-engaged RNA polymerases before and after HS induction using the Precision nuclear Run-On and sequencing (PRO-seq) assay and quantified differentially expressed genes. PRO-seq has high sensitivity and high spatial and temporal resolution, providing an unprecedented comprehensive view of the transcriptional profiles of cell populations. We show that the HS response is rapid and pervasive, with thousands of genes being repressed after 20 minutes of HS and hundreds of genes being activated; moreover, the activated genes are not limited to the classical HSP genes. Promoter-proximal pausing is highly prevalent among the activated genes prior to HS, and here we demonstrate that its establishment on a subset of genes is dependent on GAF binding upstream and proximal to the TSS. Moreover, GAF depletion abrogates pausing and consequently impairs HS activation, indicating that this step in early transcription elongation is essential for gene activation. We also show that the recently identified transcription factor Motif 1 Binding Protein (M1BP) (Li and Gilmour 2013) has a role in

pausing and HS activation of a subset of genes that exhibit GAF-independent pausing. Furthermore, we demonstrate that only a relatively small fraction of HS activated genes are regulated by HSF, and HS activation of these HSF-dependent genes is regulated at the level of pausing release. This study provides a genome-wide view of HS-induced transcriptional regulation and an understanding of how promoter context affects this process.

2.2 Materials and Methods

GAF, HSF, M1BP and LacZ RNAi treatments

Drosophila S2 cells were grown in M3+BPYE media supplemented with 10% FBS until they reached 3-5 x 10^6 cells/mL. At this point, the cells were split to 1 x 10^6 cells/mL in serum-free M3+BPYE media, and the desired volume of cells was mixed with LacZ, GAF, HSF or M1BP dsRNA to a final concentration of 10 µg/mL. After incubation at 25°C for 45 minutes, an equal volume of M3+BPYE media supplemented with 20% FBS was added to the cells. After 2.5 days, the cells were split 1:2 into two new flasks and more dsRNA was added to keep the final concentration at 10 µg/mL. After 2.5 days the cells were HS treated and harvested for nuclei isolation. The M1BP RNAi treatment was performed separately with its own LacZ control.

The dsRNAs used in these experiments were transcribed from a dsDNA template that had a T7 polymerase promoter at both ends. The DNA templates were generated by PCR using the following primers:

LacZ Forward: GAATTAATACGACTCACTATAGGGAGAGATATCCTGCTGATGAAGC

LacZ Reverse: GAATTAATACGACTCACTATAGGGAGAGCAGGAGCTCGTTATCGC

GAF Forward: GAATTAATACGACTCACTATAGGGATGGTTATGTTGGCTGGCGTCAA

GAF Reverse: GAATTAATACGACTCACTATAGGGATCTTTACGCGTGGTTTGCGT

HSF Forward: GAATTAATACGACTCACTATAGGGAGAGCCTTCCAGGAGAATGCA

HSF Reverse: GAATTAATACGACTCACTATAGGGAGAGCTCGTGGATAACCGGTC

M1BP Forward (from Li and Gilmour 2013):

GAATTAATACGACTCACTATAGGGAGAGCAGCCAAATTGCTTGTCC

M1BP Reverse (from Li and Gilmour 2013):

GAATTAATACGACTCACTATAGGGAGAAGACGGTGAAGACGCCC

Western blot analysis to assess knockdown levels

Western blots were performed using standard conditions, and dilutions of the LacZ-RNAi control samples were used as a quantitative indication of signal linearity. Lab stocks of rabbit anti-HSF and anti-GAF antibodies and guinea pig anti-TFIIS antibody were used at dilutions of 1:2000, 1:500, and 1:3000, respectively. The rabbit anti-M1BP antibody was provided by David Gilmour's lab and was used at a 1:5000 dilution. We used IRDye 800CW donkey anti-rabbit (1 mg/mL) and IRDye 680LT donkey anti-guinea pig (1 mg/mL) as secondary antibodies at a 1:15000 dilution and the membrane was imaged using the LI-COR Odyssey imaging system.

Heat Shock treatments

For the HS treatments, an equal volume of M3+BPYE medium (no serum) at 48°C was added to the cells, and the cultures were incubated at 37°C for the desired time.

Preparation of PRO-seq libraries

Nuclei isolation and PRO-seq library preparation were performed as described previously (Kwak et al. 2013). Approximately 2 x 10^7 nuclei were used for each replicate.

Preparation of RNA-seq libraries

Total RNA from S2 cells was extracted using TRIzol reagent (Thermo Fisher Scientific) and then isolated from the aqueous phase using the E.Z.N.A. Total RNA Kit I (Omega Bio-tek). The following steps were performed by the Cornell RNA Sequencing Core (Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University). PolyA+ RNA was isolated with the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs). TruSeq-barcoded RNA-seq libraries were generated with the NEBNext Ultra Directional RNA Library Prep Kit (New England Biolabs). Each library was quantified with a Qubit 2.0 (dsDNA HS kit; Thermo Fisher Scientific) and the size distribution was determined with a Fragment Analyzer (Advanced Analytical) prior to pooling.

PRO-seq data acquisition

PRO-seq libraries were sequenced in 50 nt runs on the Illumina HiSeq, using standard protocol at the Cornell Biotechnology Resource Center (http://www.BRC.cornell.edu). Raw sequencing reads were processed using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Illumina adapters were removed with the fastx_clipper tool and reads were trimmed to 26-mers using fastx_trimmer. Sequencing reads shorter than 15 nucleotides were discarded.

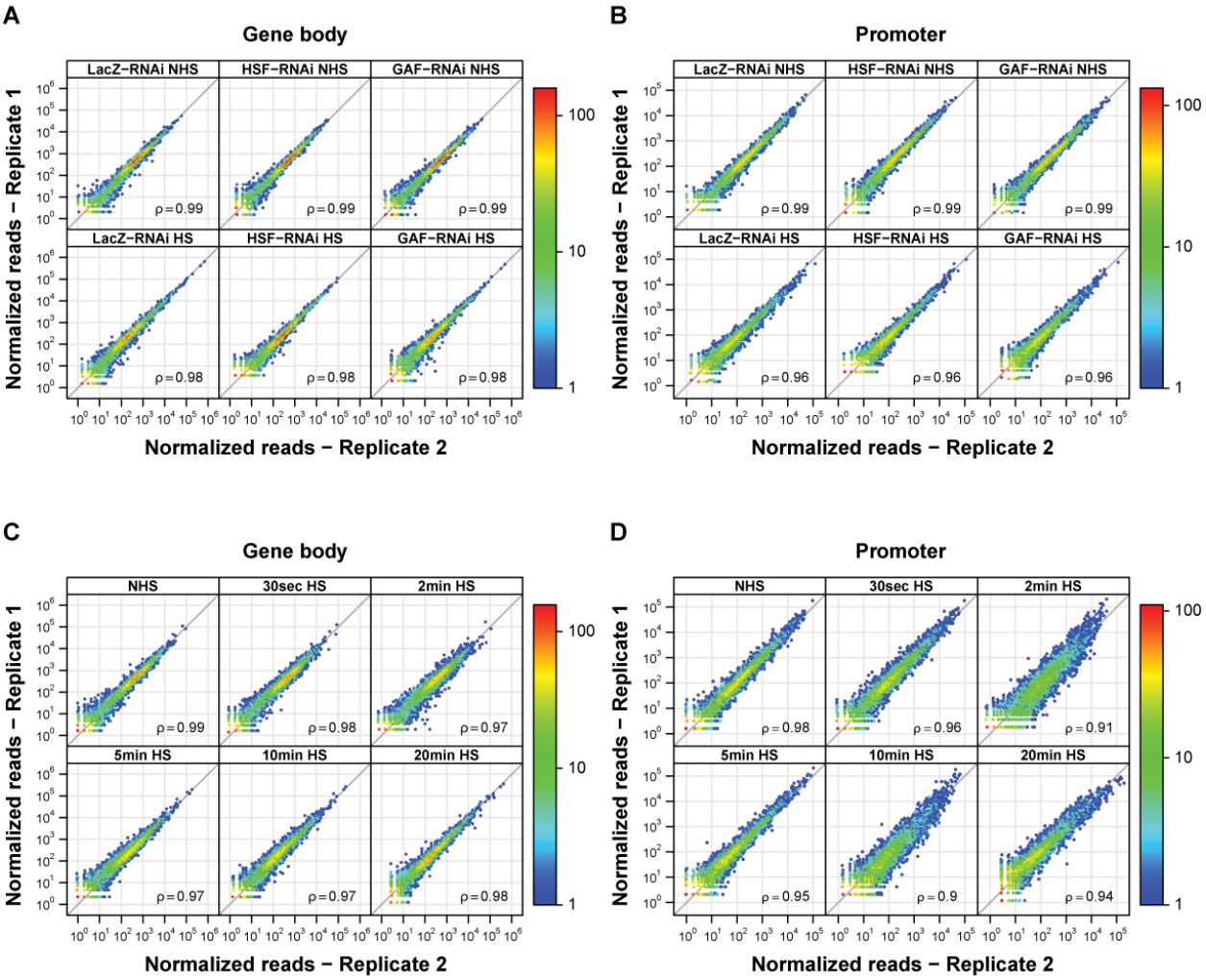
fastx_reverse_complement was then used to generate the reverse complement of the sequencing reads, which correspond to the sense strand of nascent RNA in the nucleus. Reads were aligned uniquely to the *Drosophila melanogaster* dm3 reference genome using Bowtie (Langmead et al. 2009) with up to two mismatches. Histograms of the 3'-end position of each mapped read in base-pair resolution were generated in bedgraph format and used for all subsequent analyses. Table 2.1 contains a summary of sequencing yields and the number of reads that mapped uniquely to the genome or other annotations. Replicates were highly correlated and were pooled for further analyses (Figure 2.1). Sequencing datasets can be found under GEO accession number GSE77607.

PRO-seq normalization method

We used a previously published Pol II ChIP-seq dataset in *Drosophila* S2 cells (Teves and Henikoff 2011) to identify genes whose transcription does not change during HS. Unlike ChIP-seq reads, which can originate from both sense and anti-sense strands, PRO-seq reads are strand specific. Therefore, in order to increase the likelihood of selecting genes that have the majority of their reads originating from the sense strand, we used our PRO-seq LacZ-RNAi control datasets (NHS and 20min HS) to identify and filter out genes that have high levels of transcription in the anti-sense strand. To identify those genes, we calculated the fraction of PRO-seq reads originated from the anti-sense strand for each gene and only kept the ones whose fraction is less or equal than 0.2 for both the NHS and 20min HS conditions. Because of the high background in ChIP-seq

Table 2.1: Sequencing and alignment of PRO-seq libraries. For each replicate, the total number of reads sequenced, number of reads that passed filter, number of reads after clipping, number of reads that did not align to the ribosomal genes, and number of reads that aligned uniquely to the dm3 reference genome are shown in the table.

Library	Total reads	Passed filter	After clipping	Non-ribosomal	Mapped reads
LacZ_RNAi_NHS_rep1	22002207	20498005	18857033	14282348	9443330
LacZ_RNAi_NHS_rep2	45852143	42614140	37948992	29125104	19052661
LacZ_RNAi_20minHS_rep1	38149191	35411265	32429537	26063433	12372995
LacZ_RNAi_20minHS_rep2	48224407	44752452	39159654	30083167	13759500
GAF_RNAi_NHS_rep1	30007019	27839237	25783567	18298791	11898316
GAF_RNAi_NHS_rep2	29223336	27156681	24128417	18720462	12469522
GAF_RNAi_20minHS_rep1	33083184	30743186	27869887	19884430	9288339
GAF_RNAi_20minHS_rep2	26900768	24906543	20804594	15242537	6899682
HSF_RNAi_NHS_rep1	31620891	29340277	26783955	19754947	12848598
HSF_RNAi_NHS_rep2	20661822	19223633	16847824	13244247	8850116
HSF_RNAi_20minHS_rep1	20303622	18904152	17287939	11342865	5292065
HSF_RNAi_20minHS_rep2	44198627	40832316	33567347	22388032	10971965
M1BP_RNAi_NHS_rep1	52108819	52108819	41939060	37382793	22434724
M1BP_RNAi_NHS_rep2	58684062	58684062	46843088	42896942	21490326
M1BP_RNAi_20minHS_rep1	41875931	41875931	30858958	28126011	13899422
M1BP_RNAi_20minHS_rep2	49215679	49215679	31324637	28638156	9749371
LacZ_RNAi_NHS_rep1 (from M1BP-RNAi)	46888842	46888842	38142007	32376920	20653749
LacZ_RNAi_NHS_rep2 (from M1BP-RNAi)	78247125	78247125	66730707	56231838	33911637
LacZ_RNAi_20minHS_rep1 (from M1BP-RNAi)	44850654	44850654	34314687	28456669	13326525
LacZ_RNAi_20minHS_rep2 (from M1BP-RNAi)	39884844	39884844	31949119	26644972	11580641
NHS_rep1	33326090	27513854	24770885	23263961	12081666
NHS_rep2	40386025	34851525	33127425	31427515	17083774
30secHS_rep1	32103863	26577493	24398076	23568732	13946293
30secHS_rep2	43868016	37825590	35476506	34014449	20066701
2minHS_rep1	29483939	24279233	21513613	20261645	12280319
2minHS_rep2	58929589	52883653	49114894	45522711	25046895
5minHS_rep1	26685698	22026373	20040132	18839288	10202616
5minHS_rep2	49420574	42648406	39527809	34244700	18258123
10minHS_rep1	31518051	26025041	23770378	22787845	10255055
10minHS_rep2	36364173	32508148	30318974	28531182	13973044
20minHS_rep1	53294616	47669950	44625103	42605985	19333234
20minHS_rep2	28918287	25102563	23838584	23118980	9322488



<u>Figure 2.1:</u> Biological replicates of PRO-seq libraries were highly correlated for both promoter and gene body regions. (A, B) Correlation plots between PRO-seq reads of biological replicates for the different RNAi treatments (LacZ, HSF and GAF) in **(A)** gene body (200 bp downstream of the TSS to the polyadenylation site) and **(B)** promoter-proximal (150 bp upstream of the TSS to 150 bp downstream of the TSS) regions for 9452 genes. The Spearman's correlation coefficients are shown in the plot. The gray diagonal lines represent a 1:1 fit. **(C, D)** Correlation plots between PRO-seq reads of biological replicates for the different time points after HS treatment in **(C)** gene body and **(D)** promoter-proximal regions for 9452 genes. The Spearman's correlation coefficients are shown in the plot. The gray diagonal lines represent a 1:1 fit.

data, we then focused on genes with highest levels of ser2-P ChIP signal (*Z* score > 3) (Core et al. 2012), assuming these will contain the highest densities of transcribing Pol II over background. In order to obtain a final subset of unaffected genes, we filtered out the ones whose gene body fold-change between NHS and HS conditions is less than 0.85 and greater than 1.15, resulting in 335 genes. The mRNA levels of this subset of genes are also unaffected after HS (Figure 2.2), which is consistent with the transcription levels of these genes not changing after induction.

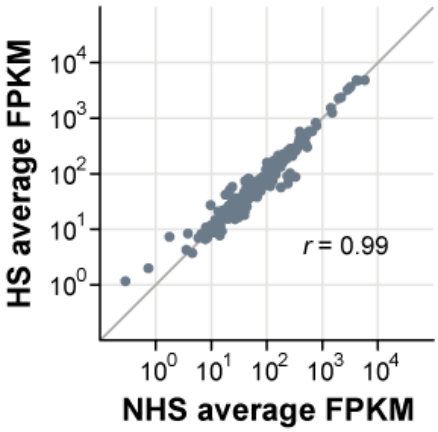


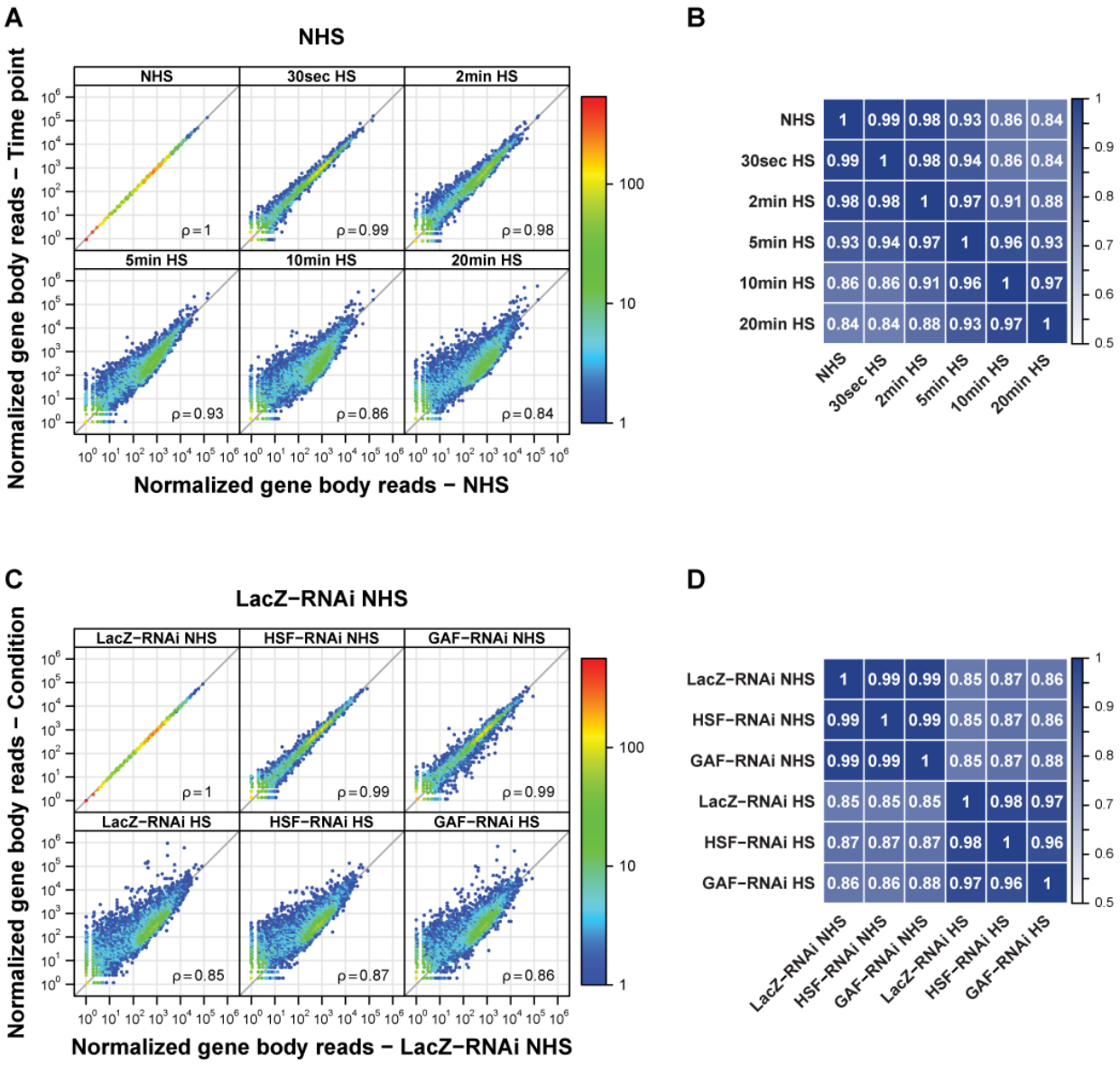
Figure 2.2: mRNA levels of 335 genes used for normalization are not affected by HS. Correlation plot between the RNA-seq FPKM for the NHS and HS conditions for the 335 HS-unaffected genes that were used to normalize our datasets. The FPKM values are the average of two biological replicates. The Pearson's correlation coefficient is shown in the plot. The gray line represents a 1:1 fit.

We then used the sum of the total number of gene body reads for all 335 genes to generate normalization factors in our PRO-seq data to normalize the datasets between replicates and different time points. Since the GAF, HSF and M1BP RNAi treatments did not result in genome-wide changes in transcription in both the NHS and 20min HS time points, we used the same subset of 335 unaffected genes to normalize the datasets between different RNAi treatments (LacZ, GAF, HSF and M1BP).

To access the efficacy of this normalization method, we examined the correlation between gene body and promoter reads for replicates (Figure 2.1) and gene body reads across different time points (Figure 2.3A-B). All replicates show good correlations and time points that are closer to each other have higher correlation coefficients and better fits to the 1:1 diagonal. Moreover, for the RNAi treatments, we examined the correlation between gene body reads across different conditions (NHS and 20min HS) and observed that the NHS datasets have higher correlation coefficients and better fits to the 1:1 diagonal when plotted against each other, and the same was observed for the HS treatments (Figure 2.3C-D). Taken together, these results indicate that the normalization method worked appropriately.

Differential expression analysis using DESeq2

We used DESeq2 (Love et al. 2014) to identify genes whose gene body reads significantly change after HS, starting with a list of 9452 non-overlapping genes described previously (Core et al. 2012). Gene body reads were collected from 200 bp downstream of the TSS, and we used different 3' limits for each time point, assuming a conservative estimate for Pol II transcription elongation rate of 1kb/min. We provided our own normalization factors for the DESeq2 calculations, which were obtained as described above. We used an FDR of 0.001 to identify activated and repressed genes. Unchanged genes were defined as having an adjusted p-value higher than 0.5 and $\log_2$fold-change higher than -0.25 and lower than 0.25.



<u>Figure 2.3:</u> Validation of the PRO-seq normalization method used in our study. (A) Correlation plots between PRO-seq gene body reads of the NHS condition and all other time points after HS treatment for 9452 genes. The Spearman's correlation coefficients are shown in the plot. The gray lines represent a 1:1 fit. **(B)** Correlation matrix showing the Spearman's correlation coefficients for all combinations of time points. The correlation was calculated as in A for each individual sample. **(C)** Correlation plots between PRO-seq gene body reads of the LacZ-RNAi NHS control and all other treatments and conditions for 9452 genes. The Spearman's correlation coefficients are shown in the plot. The gray lines represent a 1:1 fit. **(D)** Correlation matrix showing the Spearman's correlation coefficients for all combinations of RNAi treatments (LacZ, HSF and GAF) and conditions (NHS and HS). The correlation was calculated as in C for each individual sample.

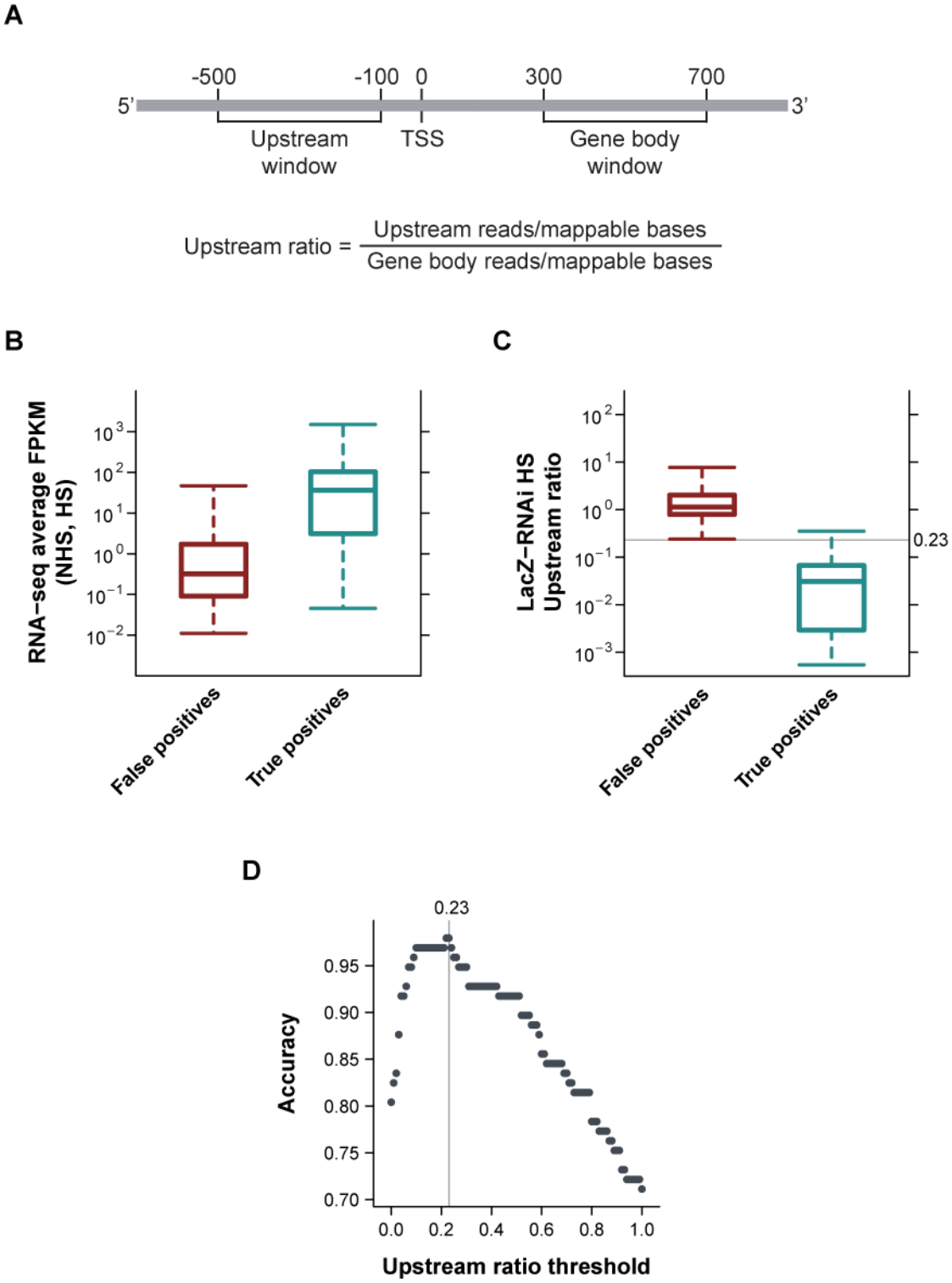
Upstream transcription filter

To minimize the number of false positives caused by changes in run-through transcription originated at the upstream gene, we implemented a filter to exclude from our analyses genes that have high levels of transcription in the region immediately upstream of the TSS. For each gene, we obtained the read counts from a window upstream of the TSS (-500bp to -100bp of the TSS) and a window in the gene body (+300bp to +700bp of the TSS) (Figure 2.4A). The 3' limit of the upstream window was defined as -100bp to the TSS to avoid confounding effects of potentially misannotated TSSs. In the case of the gene body window, the 5' limit was defined as +300bp to the TSS to avoid the region immediately downstream of the TSS, which can contain peaks of promoter-proximal paused Pol II. We then took the ratio of the read counts in these two regions for each gene, taking into consideration the number of mappable positions in the two windows (Figure 2.4A). This ratio was named *upstream ratio* and was later used to exclude false positive genes from our analyses.

In order to verify if we could distinguish true and false positives based on the upstream ratio and define the appropriate cutoff to filter out false positive genes, we visually inspected 100 randomly selected HS activated genes and classified each one as true or false positive based on the presence or absence of run-through transcription. The average mRNA levels (Figure 2.4B) are higher for the genes that were defined as true positives, which provides an independent verification of the criteria that were used to define true and false positives.

The distribution of upstream ratios for the LacZ-RNAi HS condition was very

Figure 2.4: Validation of the upstream transcription filter implemented in our study. **(A)** Diagram of the upstream ratio metric that was used to filter out false positive genes caused by run-through transcription from an upstream gene. **(B)** Box-plot of the average RNA−seq FPKM (NHS and HS) for the true (n=22) and false (n=78) positive subsets classified by visual inspection of 100 randomly selected activated genes. **(C)** Box-plot of the upstream ratio for the LacZ-RNAi HS condition for true and false positive genes. The 0.23 cutoff that was used to separate true from false positives is shown in the plot. **(D)** Accuracy metric ((true positives + true negatives)/total) of upstream ratio filter as a function of tested cutoffs. The cutoff with highest accuracy (0.23) is shown in the plot.


A
-500
-100
0
300
700
5'
3'
Upstream window
TSS
Gene body window
Upstream ratio = Upstream reads/mappable bases / Gene body reads/mappable bases
B
RNA−seq average FPKM (NHS, HS)
False positives
True positives
C
LacZ−RNAi HS Upstream ratio
0.23
False positives
True positives
D
0.23
Accuracy
Upstream ratio threshold

distinct for true and false positives, with very little overlap (Figure 2.4C), indicating that this metric could be used to identify false positives. In order to define the optimal cutoff, we evaluated the performance of all potential cutoffs from 0 to 1 in 0.01 increments and used the accuracy metric ((true positives + true negatives)/total) to identify the cutoff with the best performance (Figure 2.4D). The horizontal line in Figure 2.4C represents the chosen cutoff (0.23). Filtering out genes with upstream ratios greater than 0.23 eliminates all but one false positive, with only a minor loss of true positive genes.

We then filtered out genes with upstream ratio higher than 0.23 in the HS activated and repressed classes to generate the final subsets of genes that were used in all subsequent analyses. In order to use the condition with highest PRO-seq levels for upstream ratio calculation, we used the NHS upstream ratio to filter the repressed subsets of genes and the HS upstream ratio to filter the activated subsets for every HS and RNAi treatment.

Gene ontology analysis

Gene ontology analysis on HS activated and repressed genes was performed using the Functional Annotation tool from DAVID (Huang et al. 2009), in which 'GOTERM_BP_FAT' was selected.

Composite profiles

The composite profiles in Figures 2.8A, 2.8C, 2.9A, 2.11D, 2.13A-C, 2.14 and 2.17E represent the median from 1000 sub-samplings of 10% of the genes in each class, as

described previously (Core et al. 2012; Danko et al. 2013). The shaded areas in Figures 2.8A, 2.8C and 2.9A represent the 75% confidence intervals.

Promoter-proximal pausing analysis

The “pausing region” was defined as the 50bp interval with highest number of reads within -50 to +150bp of the TSS. This region was defined using the LacZ-RNAi control NHS condition and the same interval was used for all the other treatments and conditions. Pausing index was then calculated as the ratio of the read density in the pausing region (reads/mappable bases) and the read density in the gene body (as defined above). Genes were classified as paused as described previously (Core et al. 2008). We used DESeq2 to identify genes whose pausing levels significantly change after HS, using an FDR of 0.001.

Transcription factor binding analysis

We used *bedtools closest* (Quinlan and Hall 2010) to identify the closest HSF, GAF or M1BP ChIP-seq peak to the TSS of every transcription unit in our list. HSF, GAF or M1BP-bound genes were defined as having a ChIP-seq peak within ±1000 bp of the TSS. For the modENCODE factors, bound genes were defined as having a ‘Regions_of_sig_enrichment’ (from ChIP-chip gff3 file) within ±1000 bp of the TSS.

De novo motif search

De novo motif analysis of the promoter region of HS activated genes (-300 to +50 bp of the TSS) was performed using *MEME* (Bailey and Elkan 1994).

RNA-seq data acquisition and analysis

RNA-seq libraries were sequenced in 100 nt runs on the Illumina HiSeq, using standard protocol at the Cornell Biotechnology Resource Center (http://www.BRC.cornell.edu). Illumina adapters were removed with the fastx_clipper tool (http://hannonlab.cshl.edu/fastx_toolkit/index.html) and sequencing reads shorter than 20 nucleotides were discarded. Reads were aligned to the *Drosophila melanogaster* dm3 reference genome/transcriptome using TopHat2 (Kim et al. 2013), with the following parameters: "--library-type fr-firststrand --no-novel-juncs". Table 2.2 contains a summary of sequencing yields and the number of reads that mapped to the genome/transcriptome or other annotations. Sequencing datasets can be found under GEO accession number GSE77607.

Table 2.2: Sequencing and alignment of RNA-seq libraries. For each replicate, the total number of reads sequenced, number of reads that passed filter, number of reads after clipping, number of reads that did not align to the ribosomal genes, and number of reads that aligned to the dm3 reference genome are shown in the table.

Library	Total reads	Passed filter	After clipping	Non-ribosomal	Mapped reads
NHS_rep1	46630399	33051180	27992112	24951251	21164958
NHS_rep2	29432765	20630727	17824656	17553278	15403098
30minHS_rep1	26881631	19246449	16839161	16345216	14286713
30minHS_rep2	29492622	20610831	18993871	18270231	15925348

FPKM values for each gene were generated with Cuffnorm (Trapnell et al. 2010), using the BAM files generated by TopHat2 as input. Raw counts for each gene were obtained using HTSeq-count (Anders et al. 2014) and used as input for differential expression analysis using DESeq2 (Love et al. 2014). We used an FDR of 0.001 to identify genes whose mRNA levels significantly increase or decrease upon HS.

2.3 Results

Drosophila transcriptional Heat Shock response is rapid and pervasive

We measured nascent transcription levels by PRO-seq in *Drosophila* S2 cells prior to HS (Non-Heat Shock - NHS) and 20 minutes after an instantaneous and continuous HS stress (Figure 2.5A, Table 2.1). PRO-seq maps the active sites of transcriptionally engaged RNA polymerase complexes by affinity-purification and sequencing of nascent RNAs after a terminating biotin-NTP is incorporated during a nuclear run-on experiment (Kwak et al. 2013). The density of sequencing reads is proportional to the number of transcriptionally-engaged polymerase molecules present at each position when the nuclei were isolated. PRO-seq has base pair resolution, is strand specific, and is not affected by the background levels of accumulated RNAs (Kwak et al. 2013). Biological replicates were highly correlated for both promoter and gene body PRO-seq reads (Spearman's coefficient ranged between 0.96-0.99, Figure 2.1A-B, left panels). The expected genome-wide changes in transcription that occur during HS made it unfeasible to use total number of reads to normalize our datasets between conditions. Therefore, we normalized our libraries using a set of genes previously shown to have the same Pol II ChIP-seq signals in NHS and HS *Drosophila* S2 cells (Figure 2.2) where consistent backgrounds of ChIP-seq provide a basis of normalization (Teves and Henikoff 2011) (see Materials and Methods and Figure 2.3 for the normalization method and our validation tests).

We used DESeq2 (Love et al. 2014) to identify genes whose gene body reads significantly change after HS, using an FDR of 0.001. Due to the compactness of the *Drosophila* genome, some of the genes identified as differentially expressed by DESeq2

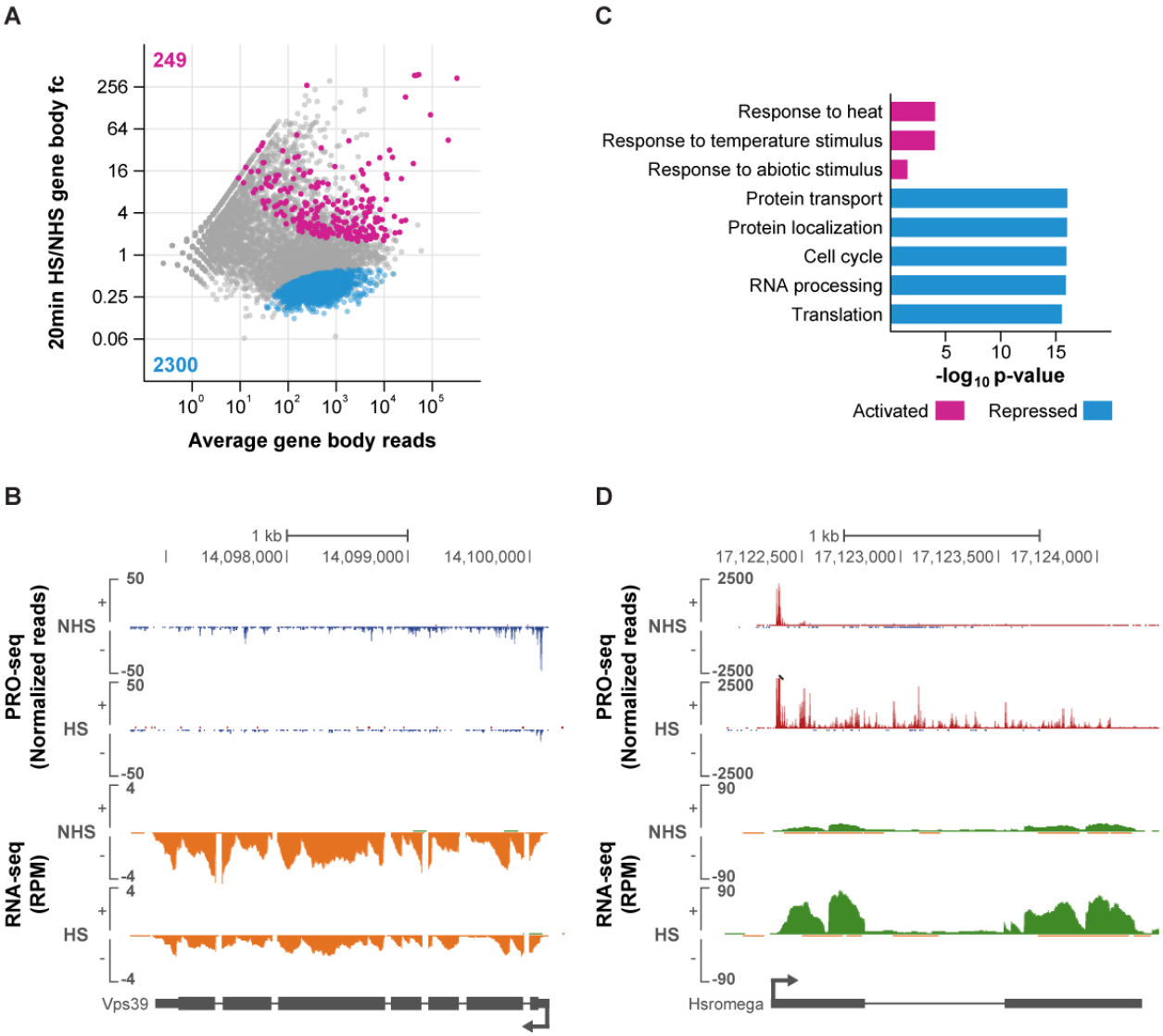


Figure 2.5: ***Drosophila*** **transcriptional Heat Shock response is rapid and pervasive.** **(A)** DESeq2 analysis of PRO-seq gene body reads between 20min HS-treated and NHS cells displayed as an MA plot. Significantly changed genes were defined using an FDR of 0.001. Activated genes that passed our upstream transcription filter (see Materials and Methods) are labeled in magenta and repressed genes in blue. The number of genes in each class is shown in the plot. fc = fold-change. **(B)** Representative view of a HS repressed gene in the UCSC genome browser (Kent et al. 2002). PRO-seq normalized reads for the plus strand are shown in red and for the minus strand in blue. RNA-seq reads for the plus strand are shown in green and for the minus strand in orange. Gene annotations are shown at the bottom. **(C)** Gene ontology terms enriched in the HS activated and repressed classes. **(D)** Representative view of a HS activated gene in the UCSC genome browser (Kent et al. 2002). Axes are the same as in B.

appear to be false positives caused by changes in run-through transcription originating at the upstream gene. To minimize the number of false positives, we implemented a filter to exclude from our analyses genes that have high levels of transcription in the region immediately upstream of the TSS (see Materials and Methods and Figure 2.4 for a description of the implemented filter and validation tests). The genes that passed the filter were classified as activated or repressed. We observed a widespread shutdown of transcription, with 2300 genes being significantly repressed after HS (Figure 2.5A, blue points; Figure 2.5B has an example of a repressed gene – *Vps39*). This finding is in agreement with low resolution studies in *Drosophila* salivary gland polytene chromosomes that have shown that total Pol II levels and transcription decrease in response to HS (Spradling et al. 1975; Jamrich et al. 1977). A previous Pol II ChIP-seq study in *Drosophila* S2 cells has also observed a genome-wide decrease of Pol II levels in gene bodies (Teves and Henikoff 2011). Not surprisingly, measurements of steady-state mRNA levels before and after HS, including micro-array studies and our own RNA-seq data (Figure 2.6, Table 2.2), were unable to detect a genome-wide shutdown of transcription, despite having the sensitivity to detect a decrease in mRNA levels for some genes (Figure 2.5B). Measurements of mRNA do not detect genome-wide transcriptional repression because the reduction of mRNA levels are obscured by steady-state levels of mRNAs already present in the cells; these mRNAs have much longer half-lives than the short HS time points examined herein. Overall, our results greatly expand upon these previous studies, identifying and quantifying the individual genes whose transcription is repressed after HS using a base-pair resolution method that specifically maps transcriptionally-engaged RNA polymerase molecules.

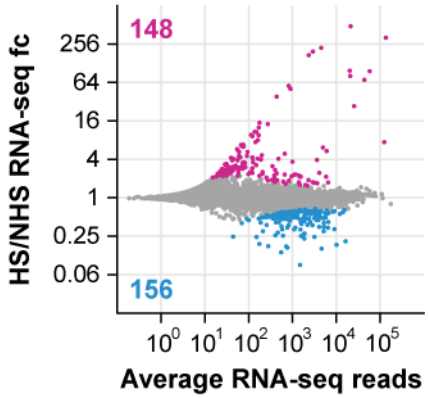


Figure 2.6: Measurement of steady-state mRNA levels by RNA-seq is unable to detect a genome-wide shutdown of transcription after HS. DESeq2 analysis of RNA-seq reads between 30min HS-treated and NHS cells displayed as an MA plot. Significantly changed genes were defined using an FDR of 0.001. Activated genes are labeled in magenta and repressed genes in blue. The number of genes in each class is shown in the plot. fc = fold-change.

Gene Ontology (GO) analysis reveals that the HS repressed class is enriched for genes involved in basic cellular processes, such as cell cycle, RNA processing, protein transport and localization, and translation (Figure 2.5C). This is consistent with previous findings in mammalian cells (Murray et al. 2004; Trinklein et al. 2004), and is expected as cells enter into a defensive non-growth condition triggered by HS stress.

Although not as abundant as the repressed class, hundreds of genes are activated by HS, many very highly (Figure 2.5A, magenta points; Figure 2.5D has an example of an activated gene – *Hsromega*). Notably, we find that all 7 classical HSP genes in our gene list show strong inductions after 20 minutes HS and are among the top 10 genes with highest HS induction, with fold-changes ranging from 44 to 384-fold. Consistent with this result, GO analysis reveals that the HS activated class is enriched for genes involved in the response to temperature and abiotic stimuli (Figure 2.5C). Besides the classical HSP genes, our data reveal the activation of many genes that were not previously

associated with the HS response and provide a comprehensive characterization and quantification of genes whose transcription is directly activated by HS.

We measured nascent transcription levels as a function of time after HS induction to determine how fast activated and repressed genes respond to HS (Figure 2.7, Table 2.1). Biological replicates produced high correlations for PRO-seq reads within either the promoter or gene body regions (Spearman's coefficient ranged between 0.9-0.99, Figure 2.1C-D). The sequential HS time points displayed a progressive increase in the number of genes that were significantly activated (Figure 2.7A, magenta points; Figure 2.7B has an example of an activated gene – *CG13321*) and repressed (Figure 2.7A, blue points; Figure 2.7C has an example of a repressed gene – *CG14005*) by HS. No genes are significantly different after 30 seconds and only a small number of genes are significantly different after 2 minutes of HS. We observe a substantial genome-wide response to HS as early as 5 minutes post-HS; the response is even more pervasive at later time points. Previous studies have shown that classical HSP genes are activated very rapidly (O'Brien and Lis 1993), but herein we demonstrate that many other genes have a rapid response, both for activation and repression. The number of significantly activated and repressed genes further increases after 10 and 20 minutes of HS (Figure 2.7A). Overall, our results demonstrate that the HS response produces an immediate and primary change in the transcription levels of ~27% (~24% repressed, ~3% activated) of the unambiguously mappable mRNA encoding genes (Core et al. 2012), with the repression of thousands of genes and the activation of many hundreds of genes.

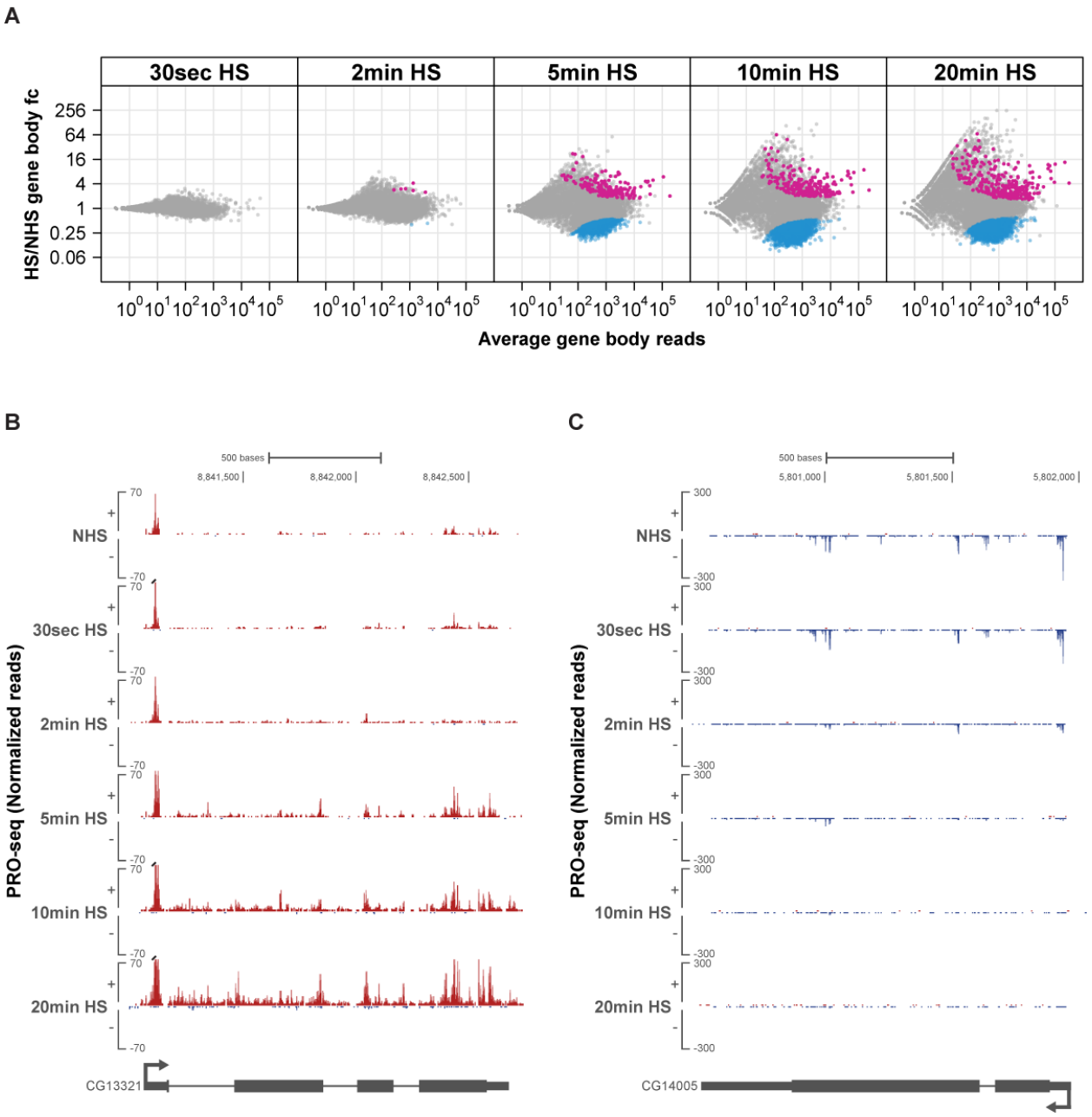


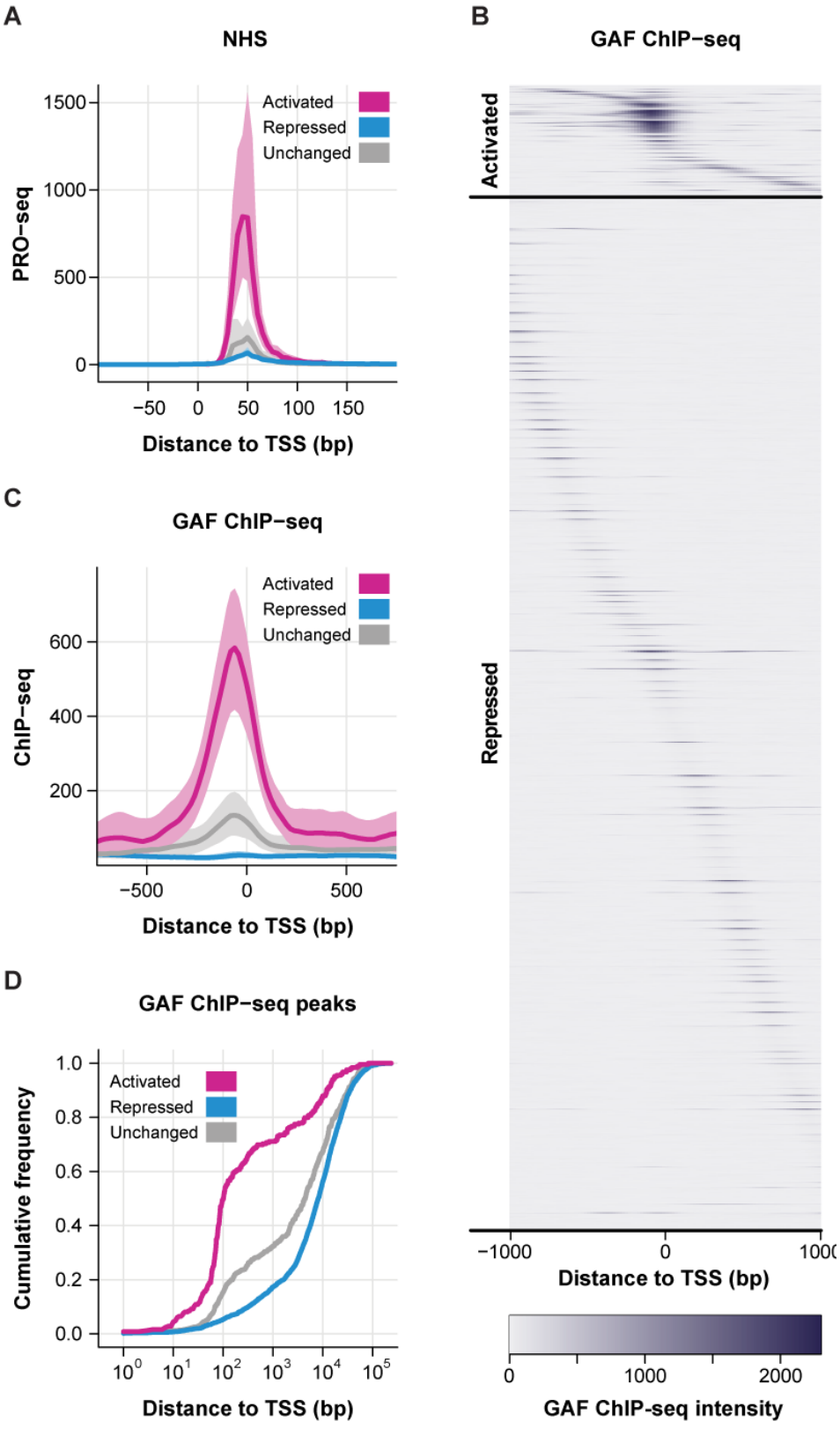
Figure 2.7: Substantial genome-wide response to HS occurs as early as 5 minutes post-HS. (A) DESeq2 analysis of PRO-seq gene body reads between HS-treated and NHS cells displayed as MA plots for the different time points after HS treatment. Significantly changed genes were defined using an FDR of 0.001. Activated genes are labeled in magenta and repressed genes in blue. fc = fold-change. **(B, C)** Representative view of a HS activated **(B)** and repressed **(C)** gene in the UCSC genome browser (Kent et al. 2002). PRO-seq normalized reads for the different time points for the plus strand are shown in red and for the minus strand in blue. Gene annotations are shown at the bottom.

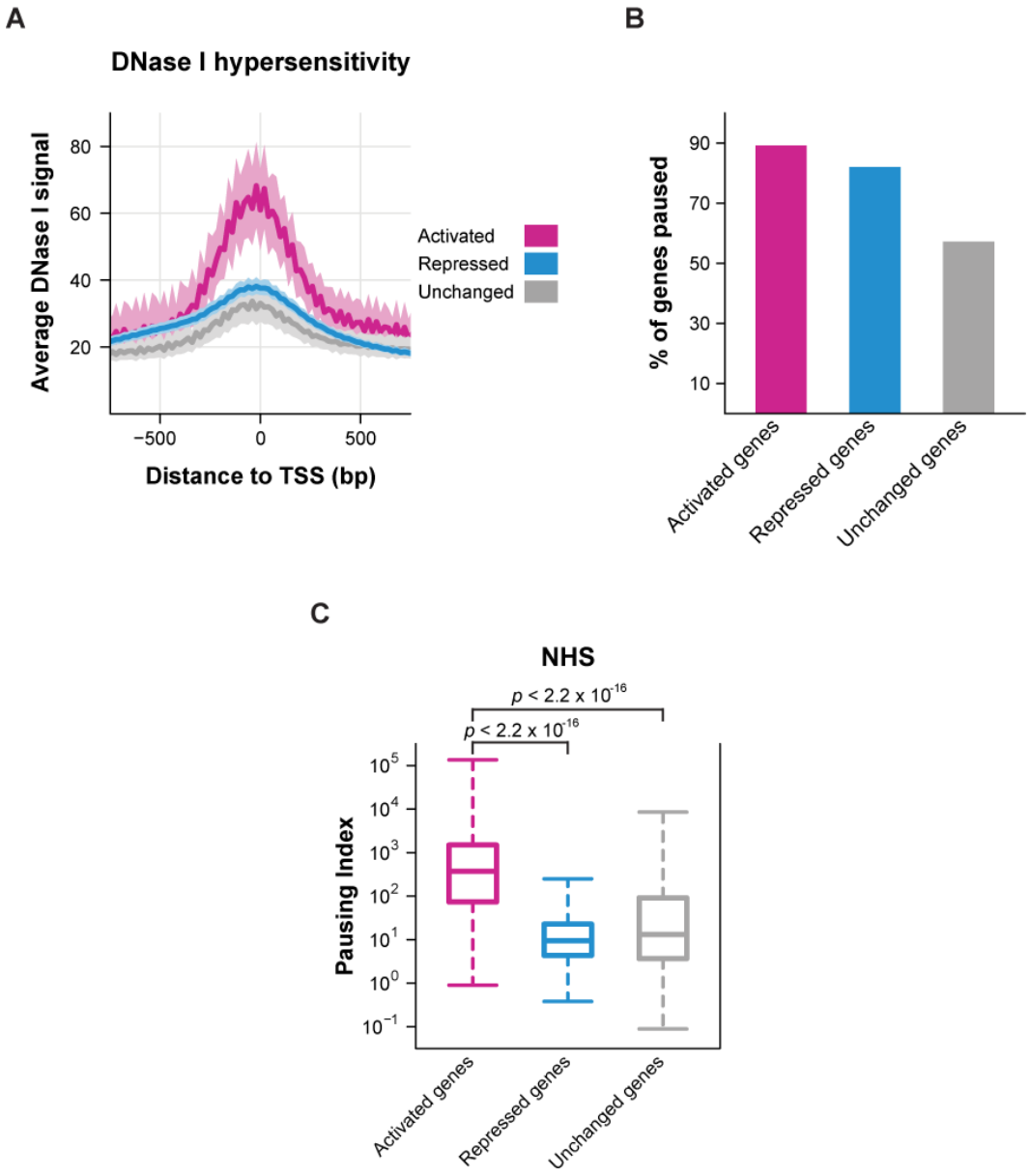
Activated genes are highly paused prior to HS

During normal cell growth, classical HSP genes have a paused, transcriptionally engaged polymerase between 20-50 bp downstream of the TSS (Rougvie and Lis 1988; Rasmussen and Lis 1993). Furthermore, promoter-proximal Pol II pausing is the major regulatory step for the HS activation of the *Hsp70* gene, where it maintains the promoter region open and accessible to transcription factors (Lee et al. 1992; Shopland et al. 1995). We used our PRO-seq data to determine if promoter-proximal pausing is a common feature of HS activated genes. The average PRO-seq read intensity profile across HS activated genes reveals a strong peak in the promoter-proximal region, which is substantially higher than repressed or unchanged gene classes (Figure 2.8A). DNase I hypersensitivity data (Kharchenko et al. 2011) indicates that the promoter region of HS activated genes is more accessible than the other two classes under basal uninduced conditions (Figure 2.9A). This data is consistent with the notion that promoter-proximal pausing is important to maintain an open chromatin environment around the TSS.

We calculated the Pausing Index (PI), which is the ratio of read density in the promoter-proximal region relative to the gene body, for each individual gene (Core et al. 2008). The vast majority of HS activated genes (~90%) were classified as paused (Fisher's exact p-value <= 0.01, Figure 2.9B) (Core et al. 2008). The PI is significantly higher for activated genes compared to the repressed (Mann-Whitney *U* test p-value < 2.2×10^{-16}) and unchanged classes (Mann-Whitney *U* test p-value < 2.2×10^{-16}) (Figure 2.9C). Although the pausing levels of HS repressed genes are not as high as the activated class (Figure 2.8A), a considerable percentage of repressed genes were also classified

Figure 2.8: GAGA factor is highly enriched in the promoter region of HS activated genes prior to HS. (A) PRO-seq read density between -100 to +200 bp to the TSS (in 5 bp bins) for the LacZ-RNAi NHS dataset of HS activated (n=249), repressed (n=2300), and unchanged (n=517) genes. The shaded area represents the 75% confidence interval. **(B)** Heatmap showing the GAF ChIP-seq signal in 20bp windows from ±1 kb to the TSS of HS activated (n=249) and repressed (n=2300) genes. For each class, genes were ordered by the distance between the highest intensity window and the TSS. **(C)** GAF ChIP−seq read density between -750 to +750 bp to the TSS (in 20 bp bins) of HS activated (n=249), repressed (n=2300), and unchanged (n=517) genes. The shaded area represents the 75% confidence interval. **(D)** Cumulative distribution plots of the distance between the closest GAF ChIP-seq peak and the TSS of each gene in the HS activated, repressed, and unchanged classes.


A
NHS
PRO−seq
1500
1000
500
0
−50
0
50
100
150
Distance to TSS (bp)
Activated
Repressed
Unchanged
B
GAF ChIP−seq
Activated
Repressed
−1000
0
1000
Distance to TSS (bp)
0
1000
2000
GAF ChIP-seq intensity
C
GAF ChIP−seq
ChIP−seq
600
400
200
−500
0
500
Distance to TSS (bp)
Activated
Repressed
Unchanged
D
GAF ChIP−seq peaks
Cumulative frequency
1.0
0.8
0.6
0.4
0.2
0.0
10^0
10^1
10^2
10^3
10^4
10^5
Distance to TSS (bp)
Activated
Repressed
Unchanged



<u>Figure 2.9:</u> Promoter region of HS activated genes is more accessible than repressed and unchanged classes prior to HS. (A) DNase I hypersensitivity signal between -750 to +750 bp to the TSS (in 20 bp bins) of HS activated (n=249), repressed (n=2300), and unchanged (n=517) genes. The shaded area represents the 75% confidence interval. **(B)** Percentage of promoter-proximal paused genes in the HS activated, repressed, and unchanged classes. Paused genes were defined as the ones with significantly higher levels of read density in the promoter-proximal region relative to the gene body (Fisher's exact p-value <= 0.01) (Core et al. 2008). **(C)** Box-plot showing the LacZ-RNAi NHS pausing index distribution for HS activated, repressed, and unchanged genes. Mann-Whitney *U* test p-values are shown on the plot.

as paused (~80%, Figure 2.9B). Overall, our results indicate that high levels of promoter-proximal pausing are a general feature of HS-induced genes prior to HS and may play an important role in poising these genes for HS activation by transcription factors.

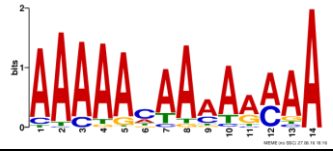
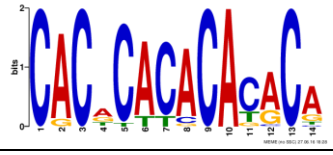
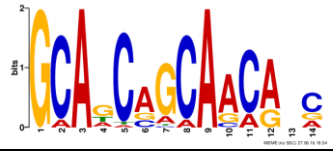
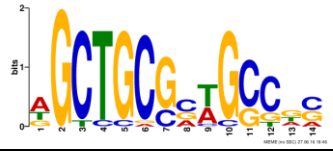
GAGA factor is highly enriched in the promoter region of HS activated genes

To identify candidate factors that play a role in allowing genes to be HS-activated, we screened modENCODE and other publically available genomic transcription factor binding data (Celniker et al. 2009; Li and Gilmour 2013; Fuda et al. 2015) for factors that are differentially enriched in HS activated relative to repressed or unchanged genes prior to HS (Table 2.3). The most significant differential enrichment was observed for GAF (Table 2.3, Figure 2.8B, GAF ChIP-seq data from Fuda et al. 2015). As seen in Figure 2.8B, when compared to the repressed class, HS activated genes show enriched GAF binding immediately upstream of the TSS, which is also evidenced by a peak in the average ChIP-seq intensity profile (Figure 2.8C). Furthermore, de novo motif analysis identified the DNA sequence bound by GAF, the GAGA element (Omichinski et al. 1997; Wilkins 1998), as the most significantly overrepresented motif in the promoter region of HS activated genes (Figure 2.10).

We then identified the closest GAF ChIP-seq peak to the TSS of each gene and plotted the cumulative distribution of these distances for our three gene classes (activated, repressed and unchanged) (Figure 2.8D). GAF binds significantly closer to activated genes than the repressed (Figure 2.8D, Kolmogorov-Smirnov test p-value < 2.2×10^{-16}) and unchanged classes (Figure 2.8D, Kolmogorov-Smirnov test p-value < 2.2 x

Table 2.3: Transcription factor binding data for HS activated, repressed and unchanged genes. Bound genes were defined as having a ChIP-seq peak or ChIP-chip 'Regions_of_sig_enrichment' within ±1000 bp of the TSS. Number: number of genes in each class bound by factor. Fraction: fraction of genes in each class bound by factor (number/total number of genes in each class). The Fisher's exact p-value was calculated for enrichment of bound genes in the activated class relative to the unchanged class. Rows were sorted by the Fisher's exact p-value (Activated x Unchanged).

Factor	Activated genes bound by factor		Repressed genes bound by factor		Unchanged genes bound by factor		Fisher's exact p-value (Activated x Unchanged)
	Number	Fraction	Number	Fraction	Number	Fraction	
GAF	177	0.71	398	0.17	167	0.32	2.40E-24
dMi-2	164	0.66	784	0.34	200	0.39	1.15E-12
LSD1	135	0.54	390	0.17	148	0.29	7.78E-12
SPT16	109	0.44	214	0.09	115	0.22	1.26E-09
Rhino	78	0.31	183	0.08	70	0.14	1.07E-08
ZW5	97	0.39	382	0.17	102	0.20	1.92E-08
dSFMBT	150	0.60	1153	0.50	202	0.39	2.74E-08
RPD3	122	0.49	1099	0.48	149	0.29	4.56E-08
HP1b	118	0.47	687	0.30	145	0.28	1.29E-07
ASH1	59	0.24	117	0.05	52	0.10	8.87E-07
NURF301	179	0.72	2160	0.94	278	0.54	9.32E-07
HP1c	107	0.43	673	0.29	132	0.26	1.06E-06
WDS	172	0.69	2065	0.90	266	0.51	2.29E-06
ISWI	160	0.64	1849	0.80	243	0.47	4.80E-06
MBD-R2	171	0.69	2104	0.91	268	0.52	6.10E-06
PR-Set7	135	0.54	1341	0.58	201	0.39	4.37E-05
dmTopo II	78	0.31	211	0.09	99	0.19	0.00016
dRING	55	0.22	96	0.04	61	0.12	0.00020
PHO	66	0.27	512	0.22	90	0.17	0.00259
Blanks	36	0.14	225	0.10	43	0.08	0.00726
JMJD2A	100	0.40	1503	0.65	160	0.31	0.00758
MLE	57	0.23	666	0.29	88	0.17	0.03366
M1BP	58	0.23	949	0.41	93	0.18	0.05243
Mod(mdg4)-67.2	17	0.07	37	0.02	20	0.04	0.05634
POF	25	0.10	145	0.06	34	0.07	0.06380
Pc	19	0.08	45	0.02	24	0.05	0.06708
BEAF-32 (BEAF-HB antibody)	132	0.53	1995	0.87	244	0.47	0.07617
E(z)	15	0.06	29	0.01	18	0.03	0.07821
Su(var)3-7	29	0.12	151	0.07	43	0.08	0.09039
Chromator (WR antibody)	147	0.59	2253	0.98	279	0.54	0.10635
Chromator (BR antibody)	152	0.61	2261	0.98	290	0.56	0.11085
PCL	15	0.06	13	0.01	21	0.04	0.15395
Su(Hw)	35	0.14	92	0.04	59	0.11	0.17652
BEAF-32 (BEAF70 antibody)	35	0.14	708	0.31	60	0.12	0.19773
CP190	115	0.46	1552	0.67	222	0.43	0.22062
HP1a	11	0.04	119	0.05	17	0.03	0.27801
MRG15	112	0.45	1942	0.84	222	0.43	0.32407
Psc	10	0.04	28	0.01	19	0.04	0.47942
Su(var)3-9	5	0.02	34	0.01	11	0.02	0.63651
CTCF	23	0.09	206	0.09	65	0.13	0.93218
HP2	2	0.01	159	0.07	12	0.02	0.96959
MSL-1	27	0.11	456	0.20	87	0.17	0.99023
JIL-1	89	0.36	2072	0.90	275	0.53	1.00000
HP4	0	0.00	1	0.00	1	0.00	1

Rank	Motif's sequence logo	*MEME E*-value
1		1.6×10^{-90}
2		6.0×10^{-43}
3		2.7×10^{-30}
4		8.5×10^{-20}
5		1.6×10^{-12}
6		1.9×10^{-11}
7		1.2×10^{-5}
8		6.3×10^{-4}

<u>Figure 2.10:</u> De novo motif analysis of the promoter region of HS activated genes. Sequence logos and respective *E*-values generated by *MEME* (Bailey and Elkan 1994) of the motifs that were found enriched in the promoter region (-300 to +50 bp of the TSS) of HS activated genes. Motifs were ranked by *E*-value. The HSE, which varies in the arrangement of its critical 5 bp units, was not identified de novo by *MEME* in the promoter region of HS activated genes; however, individual matches to the HSE's position weight matrix were identified by *FIMO* (Grant et al. 2011) and were significantly enriched in the promoter region of HS activated genes relative to the repressed and unchanged classes.

10^{-16}), and over 70% of activated genes are bound by GAF within ±1kb of the TSS. These results suggest that GAF binding close to the TSS prior to HS is important for the activation of HS-induced genes.

GAF is critical for HS activation when bound immediately upstream of the core promoter

To investigate whether GAF binding is essential for HS activation, we performed PRO-seq in biological replicates of GAF-RNAi treated cells prior to HS and after 20 minutes of HS (Figure 2.11A-B, Table 2.1) (Spearman's coefficient ranged between 0.96-0.99, Figure 2.1A-B, right panels). The decrease in GAF protein levels after the RNAi treatment produced similar numbers of genes that were significantly activated or repressed by HS (Figure 2.12A, compare to Figure 2.5A); however, comparison of the HS gene body reads in the GAF-RNAi and LacZ-RNAi control identified many genes that were significantly affected by the knockdown. The HS PRO-seq levels of 20% of activated genes were affected by GAF-RNAi and nearly all were reduced (Figure 2.11C, left panel), while less than 1% of the repressed class were affected (Figure 2.11C, right panel), demonstrating that GAF is important for HS activation, but not for repression. Greater than 90% of the genes that have reduced HS induction after GAF knockdown have GAF binding within ±1kb of the TSS (Figure 2.11C, left panel). Taken together, these results indicate that promoter-bound GAF is indispensable for the proper activation of many HS activated genes.

GAF is critical for the HS activation of many genes; however, the induction of over 70% of the genes that are bound by GAF prior to HS is not affected by GAF knockdown.

Figure 2.11: GAF's role in HS activation correlates with its function in establishing promoter-proximal pausing prior to HS. (A) Experimental set-up. *Drosophila* S2 cells were treated with either GAF or LacZ RNAi for 5 days. Nuclei were then isolated for PRO-seq after cells were incubated at room temperature (NHS) or heat shocked for 20 minutes (HS). **(B)** Western blot of whole cell extracts from LacZ-RNAi and GAF-RNAi treated cells using antibodies detecting GAF and TFIIS (loading control). 100% is equivalent to 1.5 x 10^6 cells. **(C)** DESeq2 analysis to determine the effect of GAF-RNAi treatment on the PRO-seq gene body reads after HS for the HS activated (n=249) and repressed (n=2300) classes. DESeq2 was used to identify significantly changed genes between GAF-RNAi HS and LacZ-RNAi HS cells and the results are displayed as MA plots. Significantly changed genes were defined using an FDR of 0.001. GAF−bound genes are labeled in purple, significantly changed genes (according to DESeq2) are labeled in green and genes that are both GAF-bound and significantly changed are labeled in orange. fc = fold-change. **(D)** PRO-seq read density between -100 to +200bp to the TSS (in 5 bp bins) for the LacZ-RNAi NHS and GAF-RNAi NHS treatments of genes with GAF-dependent (n=44) or GAF-independent (n=199) HS activation (HS ↑). **(E)** Box-plot showing the GAF-/LacZ-RNAi pausing region fold-change prior to HS (NHS) for genes with GAF-dependent or GAF-independent HS activation. Mann-Whitney *U* test p-value < 2.2 x 10^{-16}. Over 70% of the genes with GAF-dependent HS activation have significantly reduced pausing upon GAF depletion prior to HS, while only 15% of the GAF-independent genes were significantly affected. **(F)** Representative view in the UCSC genome browser (Kent et al. 2002) of a gene with GAF-dependent pausing prior to HS whose activation is inhibited by GAF-RNAi treatment. PRO-seq normalized reads for the different RNAi treatments (LacZ and GAF) before and after HS for the plus strand are shown in red and for the minus strand in blue. Gene annotations are shown at the bottom.

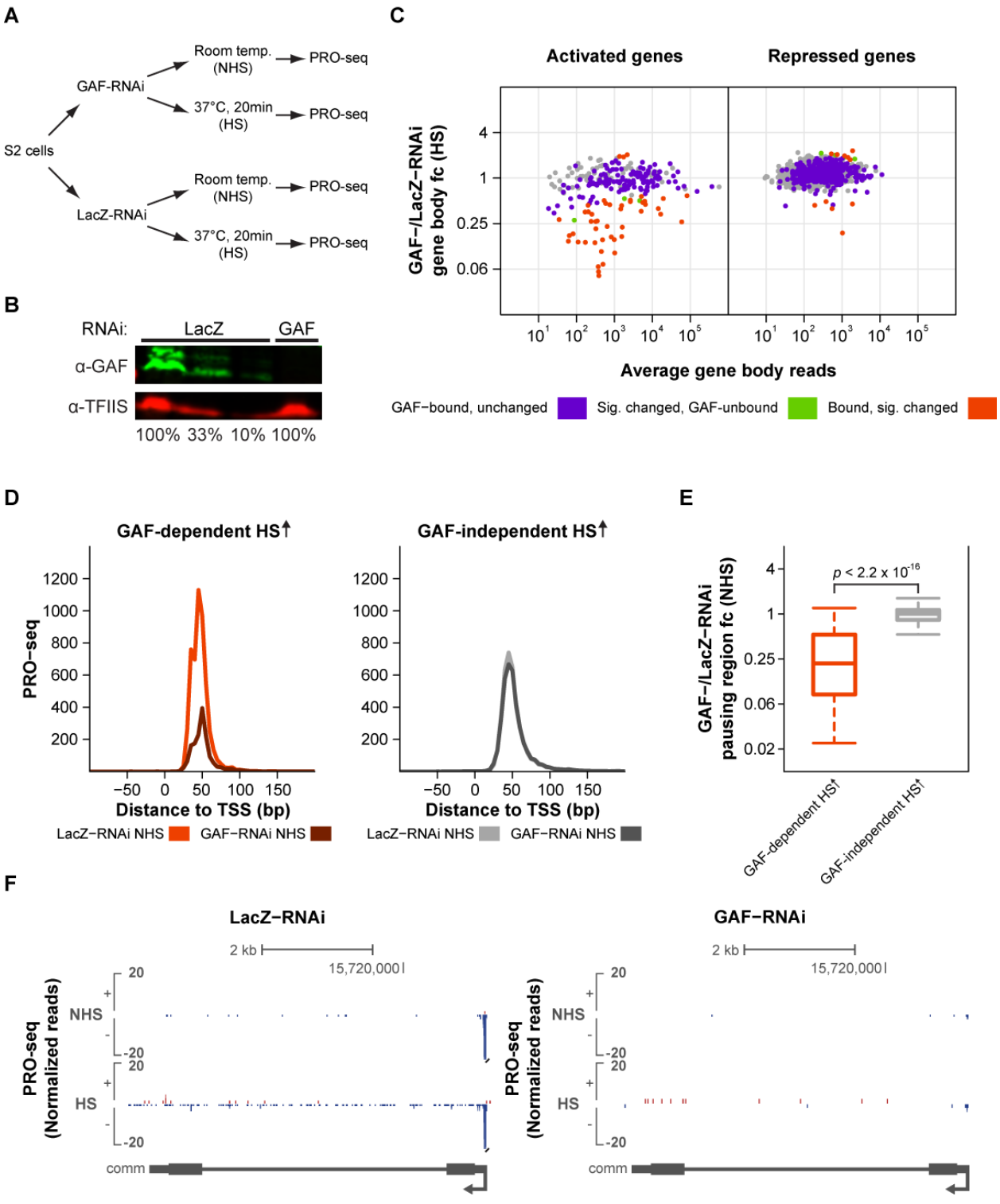
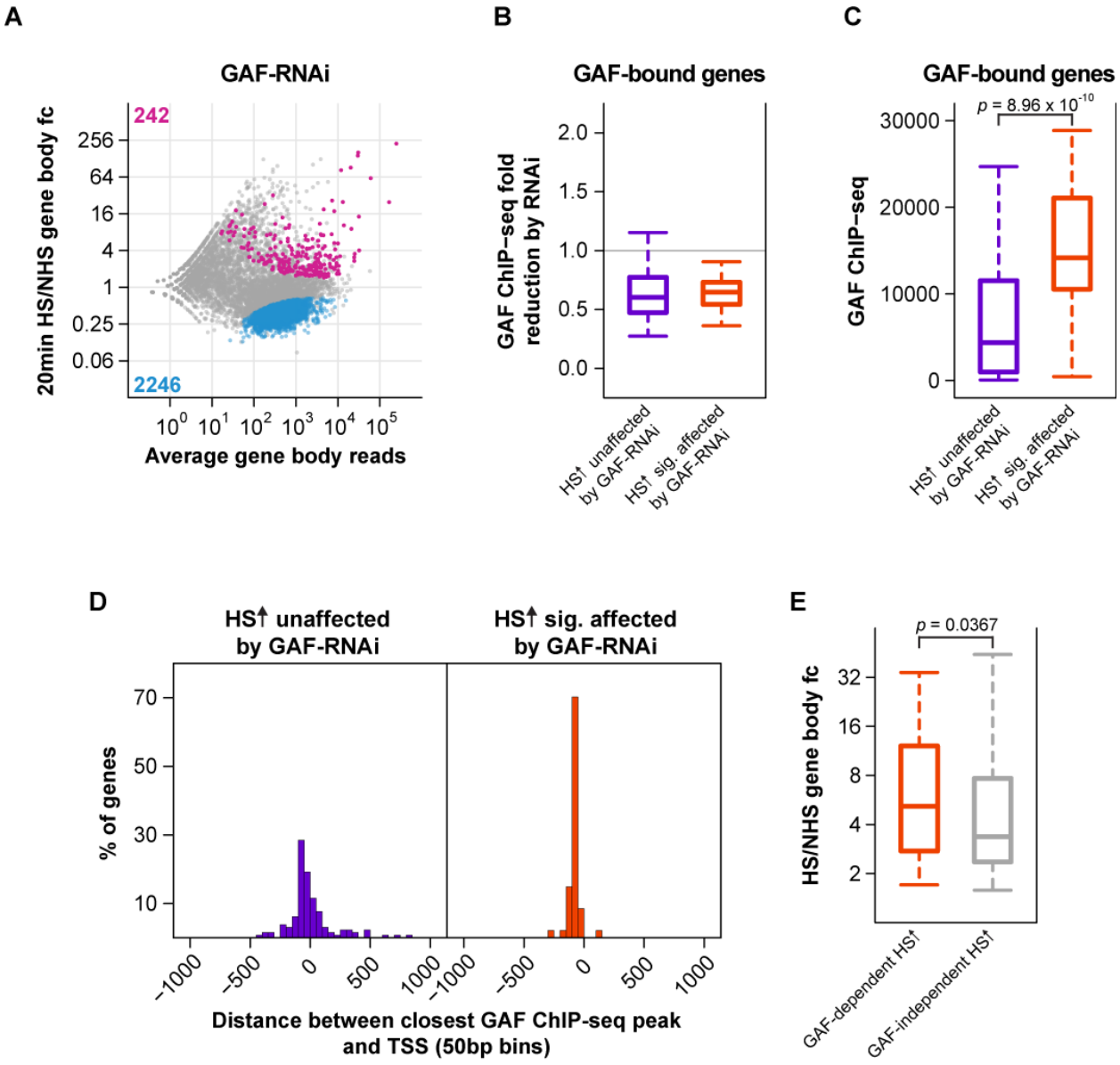

A
S2 cells
GAF-RNAi
LacZ-RNAi
Room temp. (NHS)
37°C, 20min (HS)
PRO-seq
B
RNAi:
LacZ
GAF
α-GAF
α-TFIIS
100% 33% 10% 100%
C
Activated genes
Repressed genes
GAF−/LacZ−RNAi gene body fc (HS)
Average gene body reads
GAF−bound, unchanged
Sig. changed, GAF-unbound
Bound, sig. changed
D
GAF-dependent HS↑
GAF-independent HS↑
PRO−seq
Distance to TSS (bp)
LacZ−RNAi NHS
GAF−RNAi NHS
E
GAF−/LacZ−RNAi pausing region fc (NHS)
p < 2.2 x 10-16
GAF-dependent HS↑
GAF-independent HS↑
F
LacZ−RNAi
GAF−RNAi
2 kb
15,720,000
PRO-seq (Normalized reads)
NHS
HS
comm

Figure 2.12: Higher binding levels and positioning immediately upstream of the core promoter are important for GAF's role in HS activation. (A) DESeq2 analysis of PRO-seq gene body reads of GAF-RNAi treated cells before (NHS) and after 20min HS displayed as an MA plot. Significantly changed genes were defined using an FDR of 0.001. Activated genes that passed our upstream transcription filter (see Materials and Methods) are labeled in magenta and repressed genes in blue. The number of genes in each class is shown in the plot. fc = fold-change. **(B)** Box-plot showing the fold-change of GAF ChIP-seq intensities after treatment with GAF-RNAi for GAF-bound genes whose HS induction is unaffected by GAF knockdown (purple, n=130) and GAF-bound genes whose HS induction is significantly affected by GAF knockdown (orange, n=47). **(C)** Box-plot showing the distribution of GAF ChIP-seq binding intensities for the two classes of genes described in B. Mann-Whitney *U* test p-value = 8.96×10^{-10}. **(D)** Histogram with the distribution of distances between the closest GAF ChIP-seq peak and the TSS of each gene in the two gene classes described in B. The distances are plotted in 50 bp bins from ±1000 bp to the TSS. **(E)** Box-plot showing the distribution of HS/NHS gene body fold-change for genes with GAF-dependent (n=44) or GAF-independent (n=199) HS activation. Mann-Whitney *U* test p-value = 0.0367.


A
GAF-RNAi
242
2246
20min HS/NHS gene body fc
256
64
16
4
1
0.25
0.06
10^0 10^1 10^2 10^3 10^4 10^5
Average gene body reads
B
GAF-bound genes
GAF ChIP−seq fold reduction by RNAi
2.0
1.5
1.0
0.5
0.0
HS↑ unaffected by GAF-RNAi
HS↑ sig. affected by GAF-RNAi
C
GAF-bound genes
$p = 8.96 \times 10^{-10}$
GAF ChIP−seq
30000
20000
10000
0
HS↑ unaffected by GAF-RNAi
HS↑ sig. affected by GAF-RNAi
D
HS↑ unaffected by GAF-RNAi
HS↑ sig. affected by GAF-RNAi
% of genes
70
50
30
10
−1000
−500
0
500
1000
Distance between closest GAF ChIP-seq peak and TSS (50bp bins)
E
$p = 0.0367$
HS/NHS gene body fc
32
16
8
4
2
GAF-dependent HS↑
GAF-independent HS↑

These two classes of GAF-bound genes, which respond differentially to GAF knockdown, cannot simply be explained by differences in the response to the RNAi treatment, since the GAF ChIP-seq signal for both classes is similarly reduced by the knockdown (Figure 2.12B). However, GAF-bound genes with GAF-dependent HS activation have significantly higher GAF ChIP-seq intensities when compared to the GAF-bound genes with GAF-independent HS activation (Figure 2.12C, Mann-Whitney *U* test p-value = 8.96×10^{-10}). The class of GAF-bound, HS-activated genes whose induction is dependent on GAF has a strong preference for GAF binding immediately upstream of the TSS, between -100 to -50 bp (Figure 2.12D, right panel). Taken together, these results suggest that higher binding levels and positioning upstream and proximal to the TSS are essential for GAF's role in HS activation.

GAF's role in HS activation correlates with its function in establishing promoter-proximal pausing prior to HS

GAF has been shown to have a role in the establishment of promoter-proximal pausing and consequent HS activation of two classical HSP genes (Glaser et al. 1990; Lee et al. 1992; Lu et al. 1993; O'Brien et al. 1995), and a recent study has demonstrated that pausing was significantly reduced on a large subset of GAF-bound genes upon GAF depletion (Fuda et al. 2015). However, the role of GAF-mediated pausing in gene activation has not yet been studied in a comprehensive genome-wide manner. We hypothesize that GAF's role in HS activation is connected to its ability to create promoter-proximal pausing prior to HS.

To test this hypothesis, we compared the NHS promoter-proximal PRO-seq reads for the LacZ-RNAi control and GAF-RNAi treatment between the subset of GAF-bound genes whose HS induction is dependent on GAF (GAF-dependent HS activation) and the HS activated genes whose induction is unaffected by GAF depletion (GAF-independent HS activation). As observed in Figure 2.11D, there is a substantial reduction in the NHS pausing levels after GAF knockdown for genes with GAF-dependent HS activation, while the NHS pausing levels of the GAF-independent class are largely unaffected. To quantify this effect, we compared the LacZ-RNAi and GAF-RNAi NHS reads in the pausing region for genes with GAF-dependent or GAF-independent HS activation. As observed in Figure 2.11E, most genes with GAF-dependent HS activation have reduced number of reads (fold-change < 1) in the pausing region upon GAF knockdown prior to HS. In contrast, the distribution of fold-changes for the GAF-independent class is centered around 1, indicating that GAF binding prior to HS is not essential to establish pausing at these genes (Mann-Whitney *U* test p-value < 2.2×10^{-16}). Figure 2.11F has an example of a HS-activated gene that displays GAF-dependent pausing prior to HS whose induction is inhibited by GAF knockdown. Taken together, these results indicate that GAF's role in HS activation strongly correlates with its function in establishing promoter-proximal pausing prior to HS.

Figure 2.11D also shows that GAF-dependent genes have higher levels of promoter-proximal pausing prior to HS than the GAF-independent ones. Interestingly, they also have higher average HS/NHS induction (Figure 2.12E, Mann-Whitney *U* test p-value = 0.0367); however, the distribution of HS/NHS fold-changes for these two classes

mostly overlap. Additionally, there was no preferential enrichment for classical HSP genes in either class.

Insulator proteins and M1BP are enriched in the promoter region of HS activated genes with GAF-independent induction

While nearly all activated genes display promoter-proximal pausing prior to HS, we have shown that GAF is essential for pausing establishment and HS activation on a subset of these genes. To identify factors that can contribute to the establishment of pausing on GAF-independent genes, we screened modENCODE and other publically available chromatin factor ChIP-seq or ChIP-chip datasets for factors that are differentially enriched in the promoter region of GAF-independent relative to GAF-dependent genes prior to HS (Table 2.4). Among the factors with the most significant differential enrichment were the transcription factor M1BP (ChIP-seq data from Li and Gilmour 2013), the insulator protein BEAF-32 (ChIP-chip data from Schwartz et al. 2012), and the chromodomain containing protein Chromator (ChIP-chip data from Kharchenko et al. 2011) (Figure 2.13A-C). M1BP is a recently discovered zinc-finger transcription factor that has been shown to orchestrate promoter-proximal pausing in a GAF-independent manner (Li and Gilmour 2013). BEAF-32 is one of the insulator associated proteins identified in *Drosophila* (Zhao et al. 1995), and Chromator was initially identified as a mitotic spindle protein and later implicated in the regulation of chromosome structure through partial cooperation with BEAF-32 (Rath et al. 2006; Gan et al. 2011), and both of these proteins are enriched at the boundaries of physical chromosomal domains (Hou et al. 2012; Sexton et al. 2012). The factor with

Table 2.4: Transcription factor binding data for genes with GAF-dependent or GAF-independent HS activation. Bound genes were defined as having a ChIP-seq peak or ChIP-chip 'Regions_of_sig_enrichment' within ±1000 bp of the TSS. Number: number of genes in each class bound by factor. Fraction: fraction of genes in each class bound by factor (number/total number of genes in each class). The Fisher's exact p-value was calculated for enrichment of bound genes in the GAF-independent HS activation class relative to the GAF-dependent HS activation class. Rows were sorted by the Fisher's exact p-value.

Factor	GAF-dependent HS activation genes bound by factor		GAF-independent HS activation genes bound by factor		Fisher's exact p-value
	Number	Fraction	Number	Fraction	
JIL-1	4	0.09	81	0.41	2.06E-05
M1BP	1	0.02	55	0.28	4.61E-05
Chromator (WR antibody)	14	0.32	129	0.65	6.20E-05
Chromator (BR antibody)	16	0.36	132	0.66	0.000252
BEAF-32 (BEAF-HB antibody)	17	0.39	113	0.57	0.021831
BEAF-32 (BEAF70 antibody)	2	0.05	32	0.16	0.030694
MSL-1	2	0.05	25	0.13	0.096098
MLE	7	0.16	50	0.25	0.132302
MRG15	16	0.36	93	0.47	0.138967
CP190	17	0.39	95	0.48	0.176596
Su(var)3-9	0	0.00	5	0.03	0.364945
WDS	29	0.66	139	0.70	0.365259
RPD3	21	0.48	98	0.49	0.494011
Blanks	6	0.14	30	0.15	0.510872
Pc	3	0.07	16	0.08	0.537647
CTCF	4	0.09	19	0.10	0.594259
dSFMBT	27	0.61	120	0.60	0.615388
Mod(mdg4)-67.2	3	0.07	14	0.07	0.629614
ZW5	18	0.41	78	0.39	0.650570
PHO	12	0.27	51	0.26	0.666779
HP2	0	0.00	2	0.01	0.670034
Psc	2	0.05	8	0.04	0.736402
POF	5	0.11	19	0.10	0.749252
ISWI	30	0.68	127	0.64	0.763005
MBD-R2	32	0.73	136	0.68	0.771283
JMJD2A	20	0.45	78	0.39	0.825469
dRING	12	0.27	43	0.22	0.844243
NURF301	34	0.77	141	0.71	0.852175
PR-Set7	27	0.61	106	0.53	0.873958
Su(var)3-7	7	0.16	22	0.11	0.874285
Rhino	17	0.39	60	0.30	0.897410
HP1b	25	0.57	91	0.46	0.933170
HP1c	23	0.52	82	0.41	0.933902
ASH1	14	0.32	43	0.22	0.946667
E(z)	5	0.11	10	0.05	0.965234
PCL	5	0.11	10	0.05	0.965234
Su(Hw)	10	0.23	25	0.13	0.971201
LSD1	29	0.66	102	0.51	0.974208
SPT16	25	0.57	81	0.41	0.982664
HP1a	5	0.11	6	0.03	0.994206
dmTopo II	21	0.48	56	0.28	0.995928
dMi-2	36	0.82	125	0.63	0.996479
HP4	0	0.00	0	0.00	1

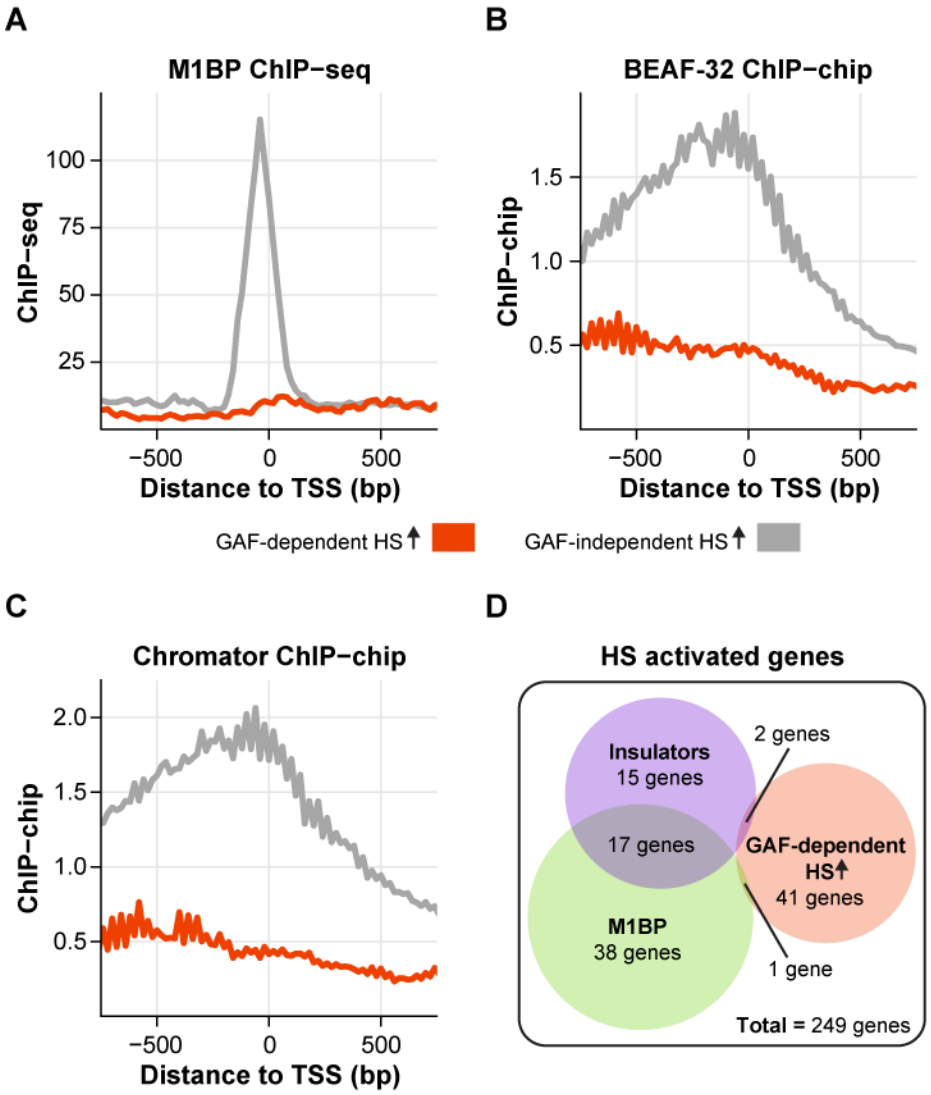


Figure 2.13: Insulator proteins and M1BP are enriched in the promoter region of HS activated genes with GAF-independent induction. (A-C) M1BP ChIP−seq **(A)**, BEAF-32 ChIP-chip (BEAF-HB antibody) **(B)**, and Chromator ChIP-chip (BR antibody) **(C)** signal between -750 to +750 bp to the TSS (in 20 bp bins) of genes with GAF-dependent (n=44) or GAF-independent (n=199) HS activation (HS ↑). **(D)** Venn diagram showing the overlap between HS activated genes with GAF-dependent activation and genes bound by M1BP or both insulator proteins (BEAF-32 and Chromator – only genes with insulator binding detected by both antibodies for these two proteins were considered) within ±1kb of the TSS.

the highest differential enrichment for promoter-bound genes in our screen was the tandem kinase JIL-1 (Table 2.4), which has been previously shown to interact with Chromator (Rath et al. 2006). However, a comparison between the JIL-1 ChIP-chip intensities of genes with GAF-dependent and GAF-independent HS activation did not show the same striking differences that were observed for M1BP, BEAF-32 and Chromator (Figure 2.14). Remarkably, almost no overlap exists between genes with GAF-dependent HS activation and genes bound by M1BP or insulator proteins within ±1kb of the TSS (Figure 2.13D). The mutually exclusive distributions of GAF and M1BP in promoter-proximal pausing has been previously reported (Li and Gilmour 2013), and our results suggest a possible role for M1BP in pausing and HS activation. Similarly to M1BP, the mutually exclusive distribution of insulator proteins and the GAF-dependent subset suggests that BEAF-32 and/or Chromator may have a role in generating promoter-proximal pausing when bound proximally to the TSS. GAF has also been classified as an insulator protein with enhancer-blocking activity (Ohtsuki and Levine 1998; Schweinsberg et al. 2004), which suggests a possible overlap between insulator function and a role in maintaining an open chromatin environment that enables promoter-proximal pausing, and opens the possibility for a novel role of BEAF-32 and Chromator as pausing factors. Another possible explanation is that these insulator proteins reside between GAF and the TSS, therefore blocking any activity of GAF on the promoter, which could explain why pausing is not affected by GAF depletion at insulator-bound promoters.

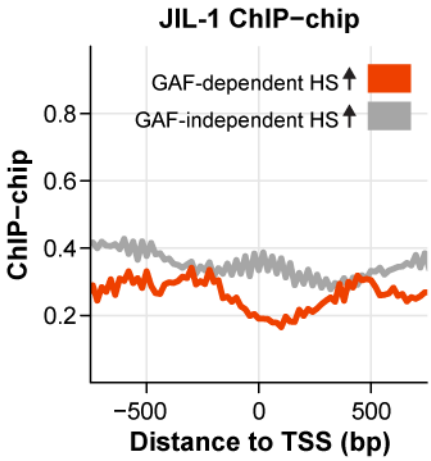
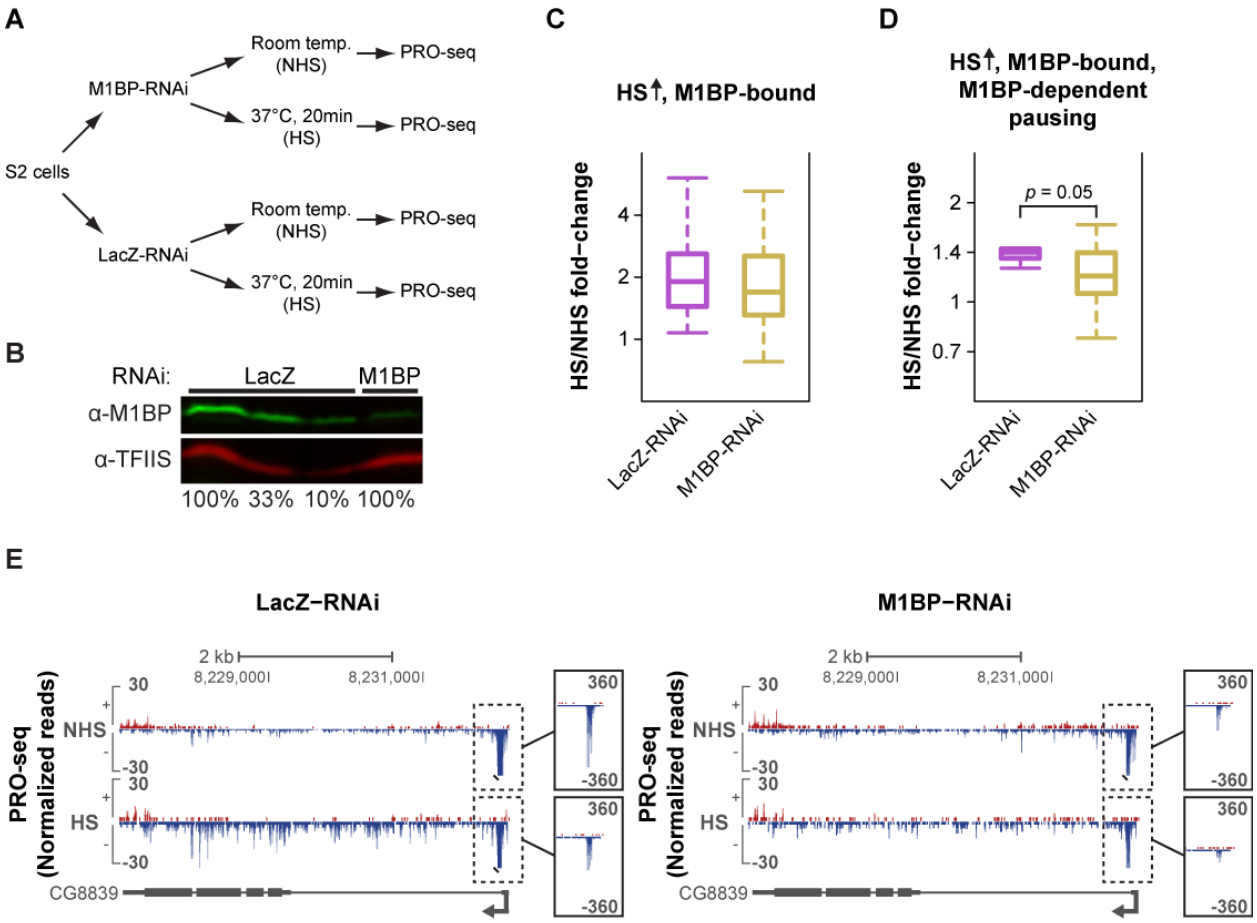


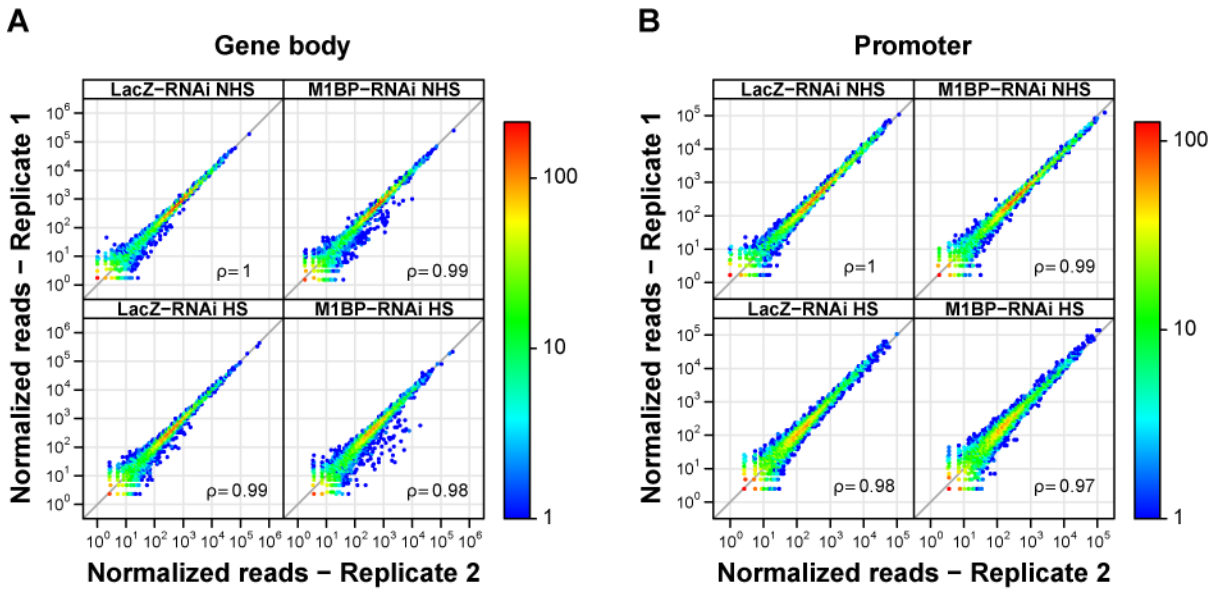
Figure 2.14: Genes with GAF-dependent or GAF-independent HS activation have similar JIL-1 ChIP-chip profiles. JIL-1 ChIP-chip signal between -750 to +750 bp to the TSS (in 20 bp bins) of genes with GAF-dependent (n=44) or GAF-independent (n=199) HS activation.

M1BP is important for promoter-proximal pausing and HS activation of a subset of M1BP-bound HS activated genes

To investigate whether M1BP indeed has a role in pausing and HS activation of M1BP-bound genes with GAF-independent HS activation, we performed PRO-seq in biological replicates of M1BP-RNAi treated cells prior to HS and after 20 minutes of HS (Figure 2.15A-B, Table 2.1) (Spearman's coefficient ranged between 0.97-0.99, Figure 2.16A-B, right panels). As seen in Figure 2.15C, M1BP depletion by RNAi has a small effect in the HS activation of M1BP-bound, HS activated genes when compared to the LacZ-RNAi control; however, this effect was not statistically significant. To assess if like GAF, M1BP's role in HS activation is associated with its role in establishing promoter-proximal pausing, we then focused on the subset of HS activated, M1BP-bound genes that display M1BP-dependent pausing prior to HS. As seen in Figure 2.15D, M1BP knockdown has a significant effect on the HS activation of this subset of genes (Mann-Whitney *U* test p-



<u>Figure 2.15:</u> M1BP is important for pausing and HS activation of a subset of M1BP-bound genes with GAF-independent induction. (A) Experimental set-up. *Drosophila* S2 cells were treated with either M1BP or LacZ RNAi for 5 days. Nuclei were then isolated for PRO-seq after cells were incubated at room temperature (NHS) or heat shocked for 20 minutes (HS). **(B)** Western blot of whole cell extracts from LacZ-RNAi and M1BP-RNAi treated cells using antibodies detecting M1BP and TFIIS (loading control). 100% is equivalent to 1.5 x 10^6 cells. **(C)** Box-plot showing the HS/NHS fold-change of M1BP-RNAi or LacZ-RNAi control cells for all M1BP-bound, HS activated genes (HS ↑). **(D)** Box-plot showing the HS/NHS fold-change of M1BP-RNAi or LacZ-RNAi control cells for M1BP-bound, HS activated genes with M1BP-dependent pausing. Mann-Whitney *U* test p-value = 0.05. **(E)** Representative view in the UCSC genome browser (Kent et al. 2002) of a gene with M1BP-dependent pausing prior to HS whose activation is decreased by M1BP-RNAi treatment. PRO-seq normalized reads for the different RNAi treatments (LacZ and M1BP) before and after HS for the plus strand are shown in red and for the minus strand in blue. Gene annotations are shown at the bottom.



<u>Figure 2.16:</u> Biological replicates of M1BP-RNAi and LacZ-RNAi control PRO-seq libraries were highly correlated for both promoter and gene body regions. (A, B) Correlation plots between PRO-seq reads of biological replicates for the different RNAi treatments (LacZ and M1BP) in **(A)** gene body (200 bp downstream of the TSS to the polyadenylation site) and **(B)** promoter-proximal (150 bp upstream of the TSS to 150 bp downstream of the TSS) regions for 9452 genes. The Spearman's correlation coefficients are shown in the plot. The gray diagonal lines represent a 1:1 fit.

value = 0.05), indicating that M1BP has a role in pausing establishment and HS activation of at least a subset of M1BP-bound genes. Figure 2.15E has an example of a gene that displays M1BP-dependent pausing prior to HS whose HS induction is affected by M1BP knockdown. Thus M1BP, like GAF, is important for pausing and HS activation of a subset of genes, supporting the hypothesis that pausing is a pre-requisite for HS activation.

<u>HSF is essential for the induction of only a small minority of HS activated genes</u>

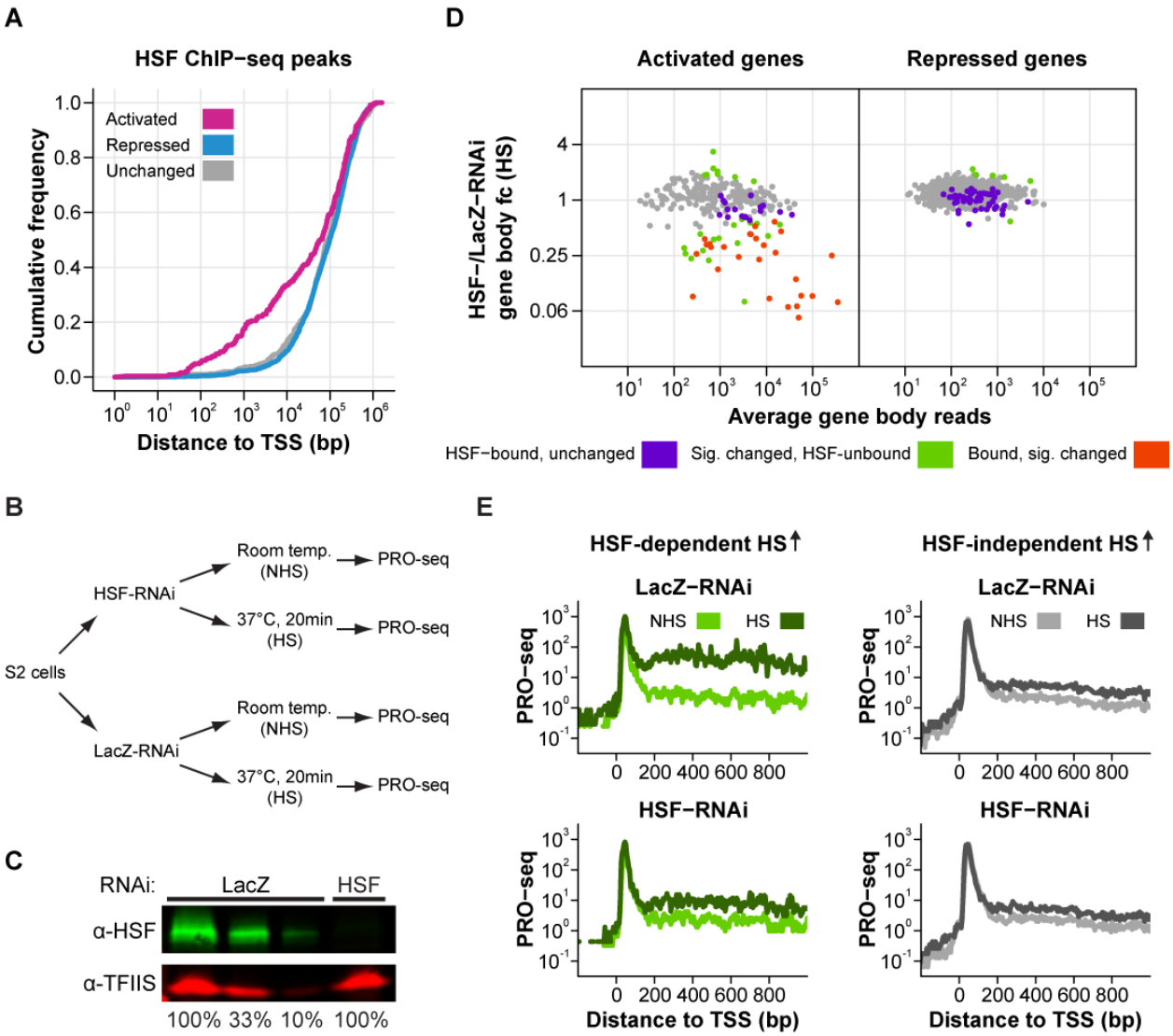
HSF is the evolutionarily conserved master regulator of the HS response and is essential for the activation of classical HSP genes (Wu 1995). Inducible HSF binding at those genes is critical for the recruitment of the positive elongation factor P-TEFb (Lis et al. 2000),

which modulates the release of Pol II into productive elongation. We used our previously published HSF ChIP-seq datasets, performed before and after 20min of HS induction (Guertin and Lis 2010), to determine if HSF also preferentially binds to non-canonical HS activated genes. HSF ChIP-seq peaks are closer to the TSS in the HS activated class when compared to the repressed (Figure 2.17A, Kolmogorov-Smirnov test p-value = 4.04×10^{-12}) and unchanged gene classes (Figure 2.17A, Kolmogorov-Smirnov test p-value = 1.6×10^{-7}). Surprisingly, even though HSF is enriched in the proximity of activated genes, less than 20% of those genes have an HSF ChIP-seq peak within ±1kb of the TSS. The existence of HSF-independent genes has been previously demonstrated in *Drosophila* (Gonsalves et al. 2011). However, our study substantially expands the number of identified genes and offers a more comprehensive view of HSF-independent regulation due to the considerably higher resolution and sensitivity afforded by our binding and nascent transcription assays. These results indicate that HSF can activate genes when bound to distal enhancer sites or that there are other factors dictating the induction of HS activated genes.

HSF activates genes by stimulating the release of paused Pol II

To investigate HSF's roles during the HS-induced transcriptional response, we performed PRO-seq in biological replicates of HSF-RNAi treated cells prior to HS and after 20 minutes of HS (Figure 2.17B-C, Table 2.1) (Spearman's coefficient ranged between 0.96-0.99, Figure 2.1A-B, middle panels). Comparison of the HS gene body reads in the LacZ-RNAi control and HSF-RNAi for activated and repressed genes shows that the HS PRO-

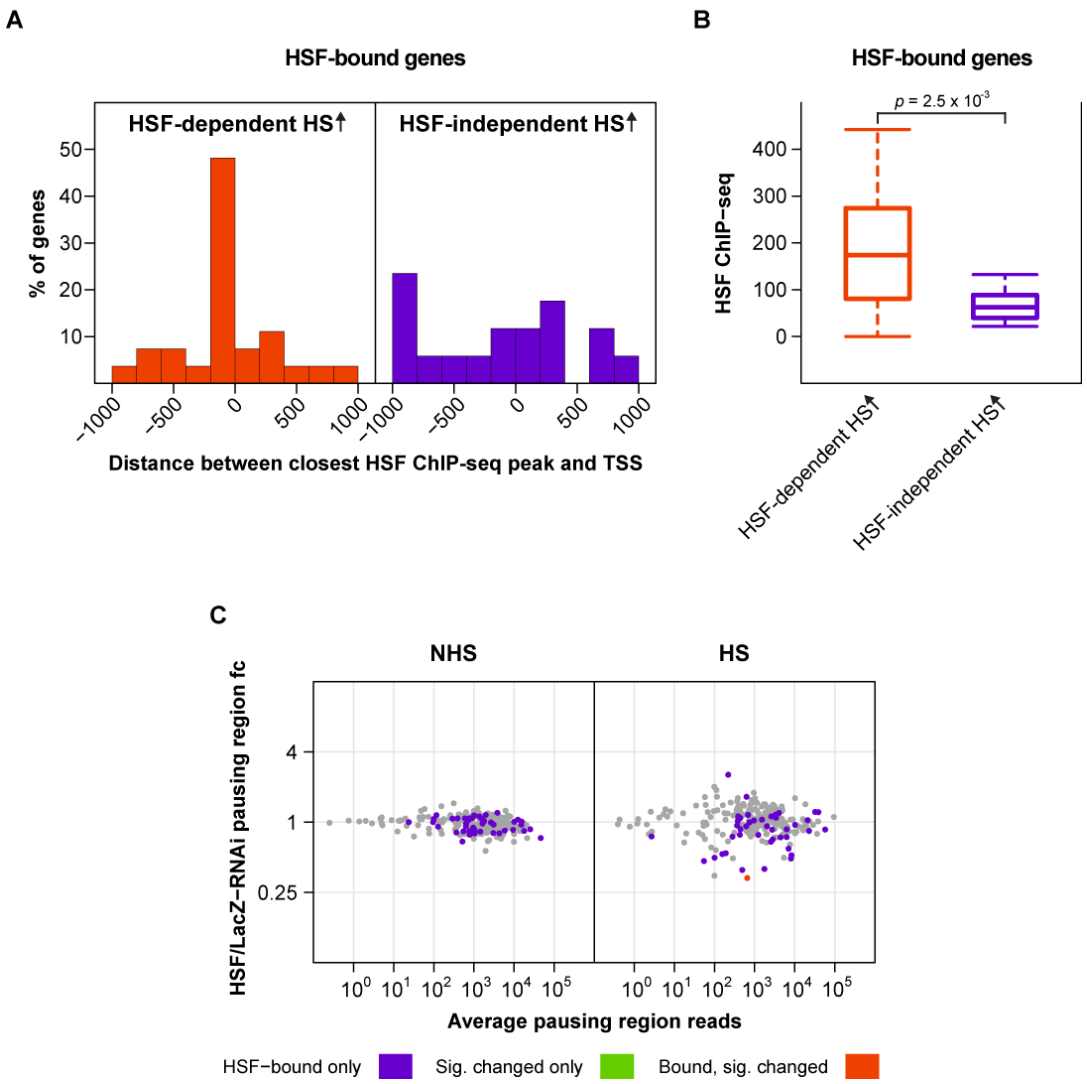
<u>Figure 2.17:</u> HSF is essential for the induction of only a small minority of HS activated genes and activates genes by stimulating the release of paused Pol II. (A) Cumulative distribution plots of the distance between the closest HSF ChIP-seq peak and the TSS of each gene in the HS activated (n=249), repressed (n=2300), and unchanged (n=517) classes. **(B)** Experimental set-up. *Drosophila* S2 cells were treated with either HSF or LacZ RNAi for 5 days. Nuclei were then isolated for PRO-seq after cells were incubated at room temperature (NHS) or heat shocked for 20 minutes (HS). **(C)** Western blot of whole cell extracts from LacZ-RNAi and HSF-RNAi treated cells using antibodies detecting HSF and TFIIS (loading control). 100% is equivalent to 1.5 x 10^6 cells. **(D)** DESeq2 analysis to determine the effect of HSF-RNAi treatment on the PRO-seq gene body reads after HS for the HS activated and repressed classes. We used DESeq2 to identify significantly changed genes between HSF-RNAi HS and LacZ-RNAi HS cells and the results are displayed as MA plots. Significantly changed genes were defined using an FDR of 0.001. HSF−bound genes are labeled in purple, significantly changed genes (according to DESeq2) are labeled in green and genes that are both HSF-bound and significantly changed are labeled in orange. fc = fold-change. **(E)** PRO-seq read density between -200 to +1000 bp to the TSS (in 5 bp bins) of genes with HSF-dependent (n=44) or independent (n=197) HS activation (HS ↑) for the LacZ-RNAi and HSF-RNAi datasets before (NHS) and after HS.


A
HSF ChIP−seq peaks
Activated
Repressed
Unchanged
Cumulative frequency
Distance to TSS (bp)
D
Activated genes
Repressed genes
HSF−/LacZ−RNAi gene body fc (HS)
Average gene body reads
HSF−bound, unchanged
Sig. changed, HSF−unbound
Bound, sig. changed
B
S2 cells
HSF-RNAi
LacZ-RNAi
Room temp. (NHS)
37°C, 20min (HS)
PRO-seq
C
RNAi:
LacZ
HSF
α-HSF
α-TFIIS
100% 33% 10% 100%
E
HSF-dependent HS↑
HSF-independent HS↑
LacZ−RNAi
HSF−RNAi
NHS
HS
PRO−seq
Distance to TSS (bp)

seq levels of 20% of activated genes were affected by HSF-RNAi, while a significant change was only observed for <1% of the repressed class, demonstrating that HSF is important for HS activation, but not for repression (Figure 2.17D).

Most of the activated genes that have HSF binding within ±1kb of the TSS have reduced HS induction after HSF knockdown (Figure 2.17D, left panel, orange points). HSF-bound genes with compromised induction upon HSF knockdown are enriched for HSF binding immediately upstream (within 200 bp) of the TSS, while the unaffected class displays a random distribution of distances (Figure 2.18A). Furthermore, HSF-bound genes with reduced HS induction have significantly higher HSF ChIP-seq binding intensity when compared to the unaffected class (Figure 2.18B, Mann-Whitney *U* test p-value = 2.5×10^{-3}), indicating that higher HSF binding levels and positioning upstream and proximal to the TSS are important for the induction of HSF's target genes. Additionally, comparison of all induced, HSF-dependent genes to the remainder of HS-activated genes (HSF-independent HS activation) showed that genes depending on HSF have stronger HS induction (Figure 2.17E). As expected, HSF is essential for the HS activation of all 7 classical HSP genes in our gene list, which strongly contributes to the higher HS induction of HSF-dependent genes relative to genes with HSF-independent HS activation (Figure 2.17E).

HSF depletion does not affect the induction of most genes that are not bound by HSF within ±1kb of the TSS (Figure 2.17D, left panel, gray points). Nonetheless, the presence of significantly changed genes with no proximal HSF binding (Figure 2.17D, left panel, green points) indicates that HSF may be able to mediate activation at distal



<u>Figure 2.18:</u> Higher HSF binding levels and positioning upstream and proximal to the TSS are important for the induction of HSF's target genes. (A) Histogram with the distribution of distances between the closest HSF ChIP-seq peak and the TSS of HSF-bound genes with HSF-dependent (left panel, n=27) or independent (right panel, n=17) HS activation. The distances are plotted in 200 bp bins from ±1000 bp to the TSS. **(B)** Box-plot showing the distribution of HSF ChIP-seq binding intensities for the two classes of genes described in A. Mann-Whitney *U* test p-value = 2.5 x 10^{-3}. **(C)** DESeq2 analysis to determine the effect of HSF-RNAi treatment on the PRO-seq pausing region reads before (NHS) and after 20min HS (HS). DESeq2 was used to identify significantly changed genes between HSF-RNAi and LacZ-RNAi cells and the results are displayed as MA plots. Significantly changed genes were defined using an FDR of 0.001. HSF−bound genes are labeled in purple, significantly changed genes (according to DESeq2) are labeled in green and genes that are both HSF-bound and significantly changed are labeled in orange.

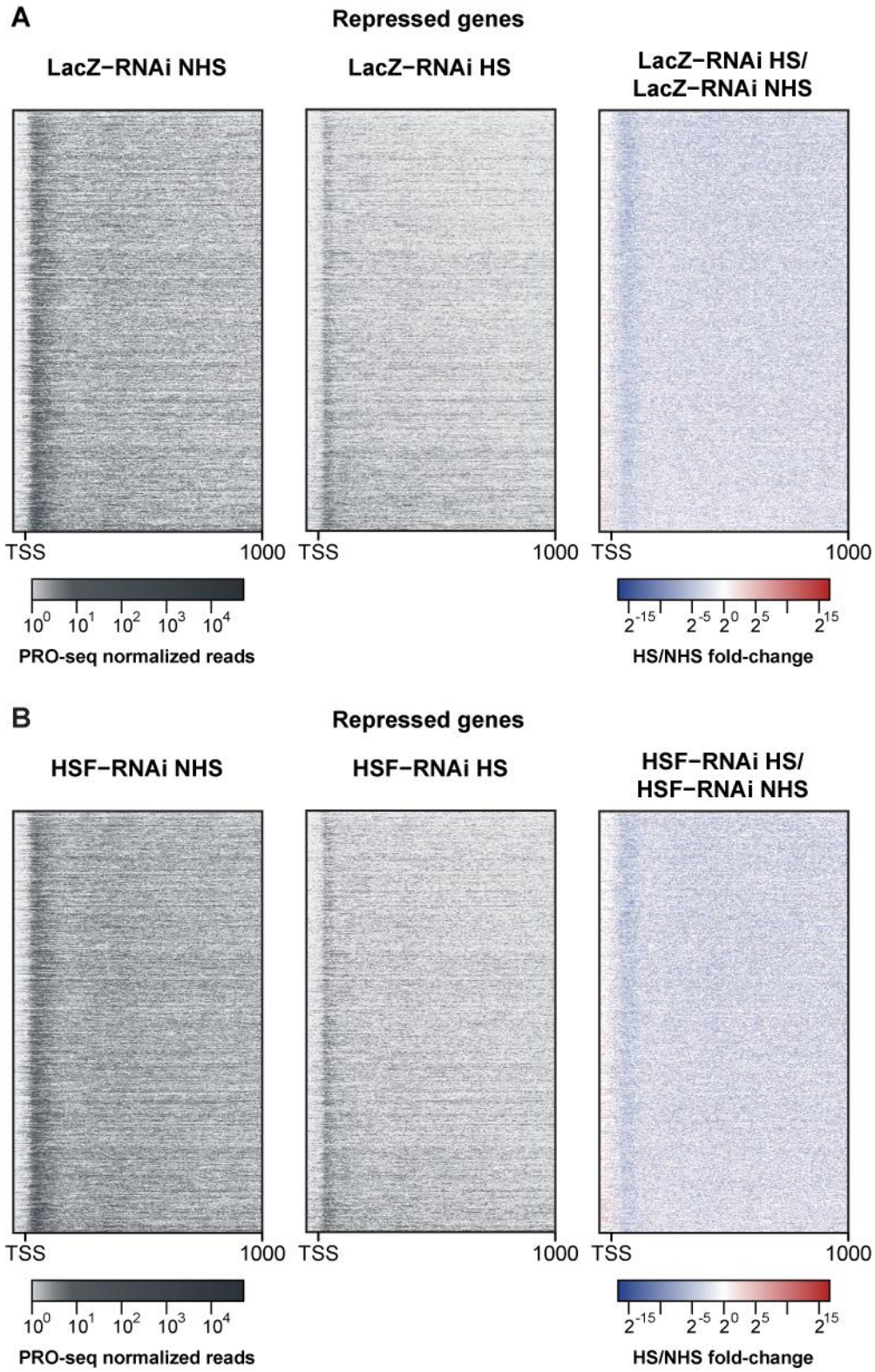
enhancer sites on a small subset of genes. The enhancer activity of HSF had been previously shown in a focused study of *Hsp70* to be weak and require large arrays of HSF binding sites (Bienz and Pelham 1986). This early study and the rarity with which we find HSF acting at a distance might be explained if such long-range interactions required specialized binding sites and chromatin architecture. Clearly, the preferred mode of HSF action is close to promoters.

Composite profiles show that the average pausing levels of genes with HSF-dependent and HSF-independent HS activation are not affected by HS in both LacZ-RNAi and HSF-RNAi conditions (Figure 2.17E), indicating that neither HSF depletion nor HS have much of an effect on pausing. Quantification of the effect of HSF knockdown on pausing levels in both NHS and HS conditions for all HS activated genes revealed that the pausing level of only one gene was significantly affected by the knockdown (Figure 2.18C). Taken together, these results suggest that HSF acts mainly at the release of paused Pol II into productive elongation, which is consistent with the critical role of HSF in the recruitment of the pause release factor P-TEFb to *Hsp70* (Lis et al. 2000).

HS transcriptional repression results in a decrease of promoter-proximally paused Pol II

Our data revealed that HS induction causes a vast transcriptional shutdown, with thousands of genes being repressed after 20min of increased temperatures (Figure 2.5A). To elucidate the mechanisms involved in this repression, we observed the PRO-seq profile for all repressed genes plotted as heatmaps before and after HS (Figure 2.19A). This analysis indicates the presence of enriched PRO-seq reads in the region

Figure 2.19: HS transcriptional repression is HSF-independent and results in a decrease of promoter-proximally paused Pol II. Heatmaps showing the NHS PRO-seq levels (left panel), HS PRO-seq levels (middle panel) and the fold-change between the two conditions (right panel) between -50 to +1000 bp to the TSS (in 5 bp bins) of HS repressed genes (n=2300) for the LacZ-RNAi **(A)** and HSF-RNAi **(B)** treatments. Genes in both A and B are sorted by the HS/NHS PRO-seq fold-change in the LacZ-RNAi condition (highest to lowest).


A
Repressed genes
LacZ−RNAi NHS
LacZ−RNAi HS
LacZ−RNAi HS/
LacZ−RNAi NHS
TSS
1000
TSS
1000
TSS
1000
10^0 10^1 10^2 10^3 10^4
PRO-seq normalized reads
2^{-15} 2^{-5} 2^0 2^5 2^{15}
HS/NHS fold-change
B
Repressed genes
HSF−RNAi NHS
HSF−RNAi HS
HSF−RNAi HS/
HSF−RNAi NHS
TSS
1000
TSS
1000
TSS
1000
10^0 10^1 10^2 10^3 10^4
PRO-seq normalized reads
2^{-15} 2^{-5} 2^0 2^5 2^{15}
HS/NHS fold-change

immediately downstream of the TSS, representing promoter-proximally paused polymerases. Both gene body reads and reads in this promoter-proximal region are reduced after HS, which is also evidenced by the overall blue color of the fold-change heatmap (Figure 2.19A, right panel). The overall NHS and HS distributions and the HS/NHS fold-change are very similar after HSF depletion by knockdown (Figure 2.19B), indicating that HSF does not play a role in gene repression by HS.

2.4 Discussion

In this study, we used PRO-seq to comprehensively characterize the direct changes in Pol II distribution that occur in *Drosophila* S2 cells in the minutes following HS. We show that the HS response is more general than previously appreciated, with thousands of genes being repressed and hundreds activated by heat. This latter class is not limited to the group of cellular chaperones that are known to be activated by stress (Lindquist and Craig 1988), and includes hundreds of other genes with various cellular functions. Surprisingly, only 20% of the activated genes are regulated by HSF, which was previously believed to be the major orchestrator of the response. Moreover, we show that promoter-proximal pausing is highly pronounced and prevalent among activated genes prior to HS. GAF, which has been shown to be important for establishing pausing, is highly enriched at the promoter of HS activated genes, and our results suggest that GAF-mediated pausing in a subset of these genes is essential for HS activation. Furthermore, our results indicate that HS activation of HSF-dependent genes is regulated at the level of pausing release, whereas HS repression of thousands of genes is regulated at the step of transcription initiation in *Drosophila*, and this process is independent of HSF. Very

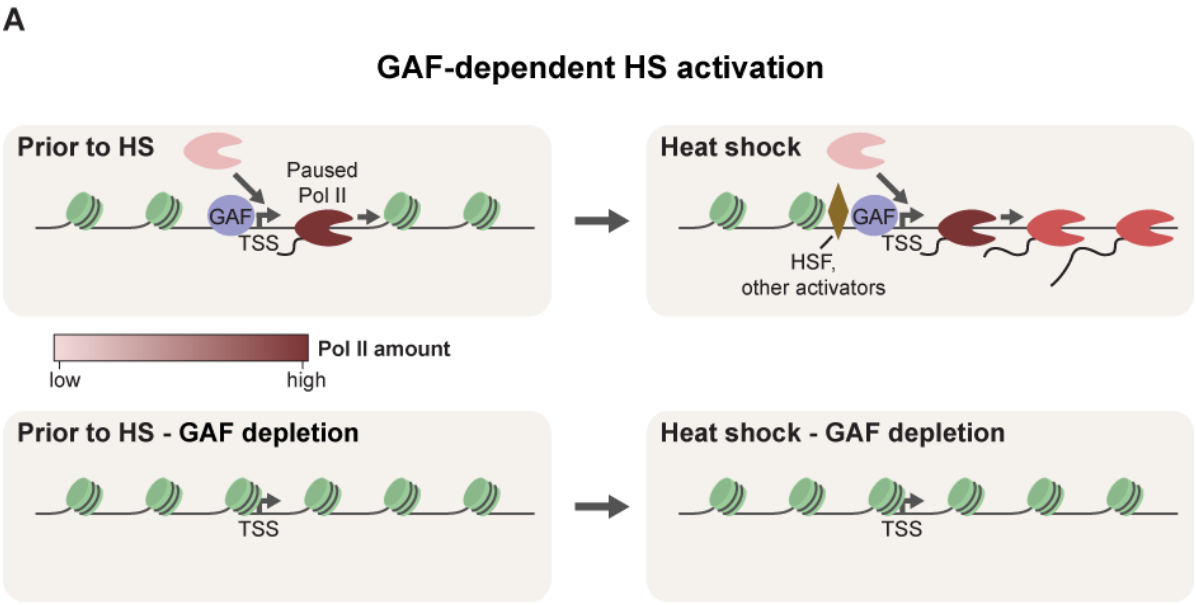
recently, we have shown in mouse that HS activation is similarly regulated by HSF at the level of pause release; however, in contrast to *Drosophila*, HS repression of genes is also mediated at pause release (Mahat et al. 2016). In both mammals and *Drosophila*, the widespread transcriptional repression is independent of HSF. Overall, by measuring how transcription changes after HS, our results provide insights into mechanisms of transcription activation and repression, the key regulating factors, and the steps in the transcription cycle that are modulated.

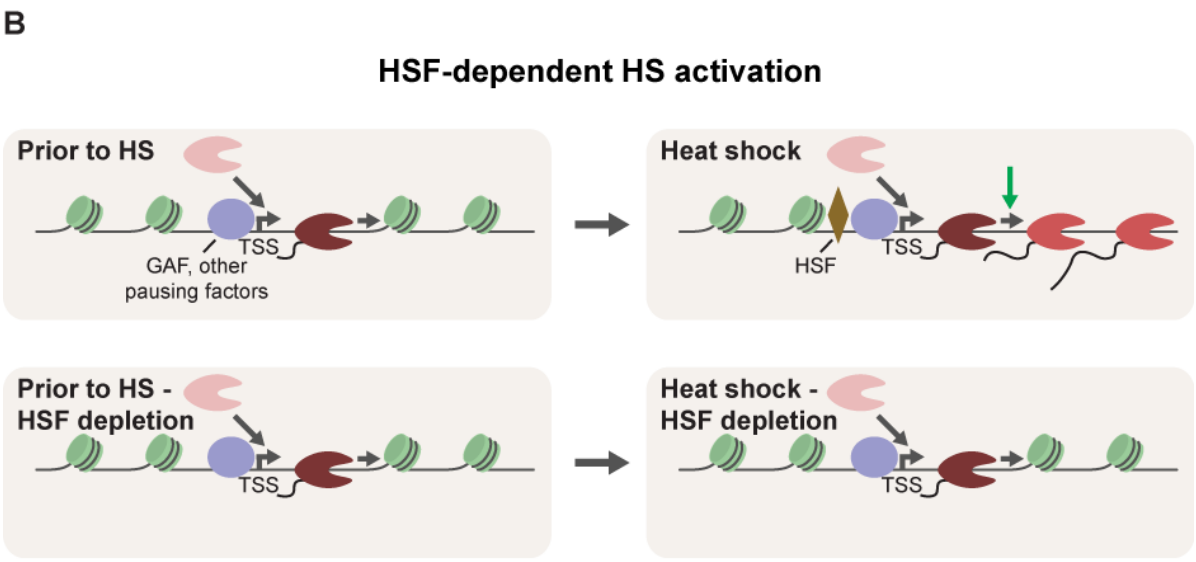
GAF-mediated promoter-proximal pausing is essential for the HS activation of a subset of genes

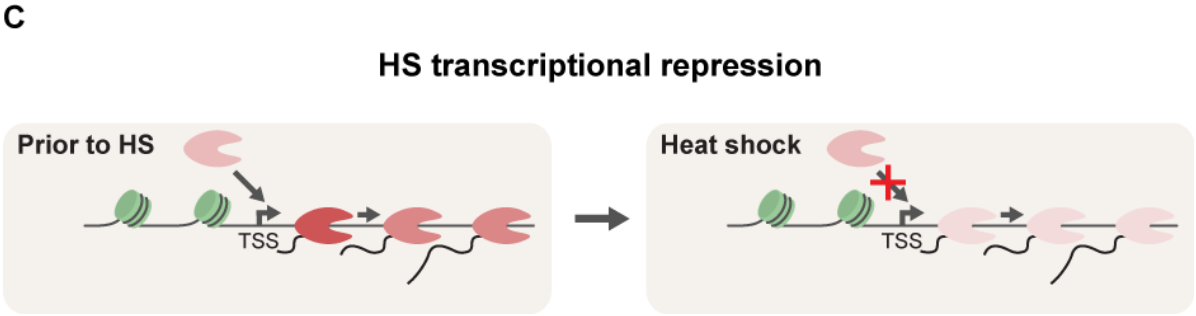
Classical HSP genes accumulate paused Pol II molecules between 20-50 bp downstream of the TSS prior to HS (Rougvie and Lis 1988; Rasmussen and Lis 1993). Our results and analyses greatly expand upon these previous findings and indicate that pausing is a common feature among HS activated genes and is not specific to the highly induced class of molecular chaperones. Previous studies have shown that the paused Pol II complex on *Hsp70* and many other genes is remarkably stable (Henriques et al. 2013; Buckley et al. 2014; Jonkers et al. 2014), and this stably paused molecule can help to maintain an open chromatin environment that is accessible to transcription factors that will promote the release of Pol II into productive elongation, mediating a rapid response to HS. The open chromatin state of our newly identified HS activated genes is confirmed with the higher DNase I hypersensitivity signal observed in the promoter region relative to repressed and unchanged genes (Figure 2.9A).

GAF has been previously shown to be important for establishing pausing and is highly enriched in the promoter region of activated genes prior to HS. GAF is essential for HS activation of a subset of GAF-bound genes that have high levels of GAF binding in the region immediately upstream of the core promoter, indicating that GAF's positioning and levels are important for its role in the HS response. We also observed that the pausing levels of genes with GAF-dependent HS activation are dramatically reduced upon GAF depletion prior to HS. Transgenic studies of the model *Hsp70* gene have demonstrated that presence of the GAF binding element is essential for generating pausing at this gene and that pausing level changes created by mutating the core promoter strongly correlate with the promoter's potential to induce transcription upon HS induction (Lee et al. 1992). Our results expand upon these studies and demonstrate that GAF depletion prior to HS in the native chromatin environment of a subset of HS activated genes abrogates Pol II pausing levels and the consequent induction of these genes by HS. Importantly, other GAF-bound genes that maintain pausing upon GAF knockdown, due to the activity of other pausing factors like M1BP and possibly the insulator proteins BEAF-32 and Chromator, remain fully HS inducible. We propose a model where GAF-mediated pausing is essential to maintain an open chromatin environment at the promoter region prior to HS (Figure 2.20A). When GAF is depleted by knockdown and pausing is not properly established, then the promoter loses its potential to induce transcription after HS (Figure 2.20A).

Figure 2.20: Summary of proposed mechanisms of HS transcriptional regulation. Model depicting the mechanisms of transcriptional regulation proposed in our study for **(A)** GAF-dependent HS activation, **(B)** HSF-dependent HS activation and **(C)** HS transcriptional repression. Red X represents a step that is being inhibited, and green arrow represents a step induced by HS. Nucleosomes are shown in green.


A
GAF-dependent HS activation
Prior to HS
Paused
Pol II
GAF
TSS
Heat shock
GAF
TSS
HSF,
other activators
Pol II amount
low
high
Prior to HS - GAF depletion
TSS
Heat shock - GAF depletion
TSS


B
HSF-dependent HS activation
Prior to HS
TSS
GAF, other
pausing factors
Heat shock
HSF
TSS
Prior to HS -
HSF depletion
TSS
Heat shock -
HSF depletion
TSS


C
HS transcriptional repression
Prior to HS
TSS
Heat shock
TSS

HSF acts at the step of promoter-proximal pausing release

HSF depletion has almost no effect on pausing reads both before and after HS (Figure 2.18C), and the average pausing levels are largely unaffected by HS and HSF knockdown (Figure 2.17E). The amount of pausing is determined by the transcription recruitment/initiation rate and the rate of escape into productive elongation. If HSF was acting at the step of Pol II initiation, we would expect the pausing levels to be reduced upon HSF depletion, which is not observed. Therefore, we propose a model where after being recruited to the promoter region upon HS, HSF promotes the release of Pol II into the gene body (Figure 2.20B), likely through the indirect recruitment of P-TEFb, which has been shown to be the case for classical HSP genes (Lis et al. 2000). Pausing also maintains an open chromatin environment that is accessible to transcription activators such as HSF. In this model, the activity of factors that are important for establishing pausing prior to HS, such as GAF, is crucial for HSF-dependent HS activation (Figure 2.20B), and failure to generate pausing prevents the induction of HSF target genes. Less than 20% of the genes with HSF-dependent HS activation are also dependent on GAF for activation, indicating that the action of other factors such as M1BP and possibly BEAF-32 and Chromator is important for pausing and consequent HS activation of these HSF-dependent genes.

HS causes a rapid and broad reduction in transcription, which is regulated at the transcription initiation step and independent of HSF

Early low resolution studies in *Drosophila* polytene chromosomes have shown that HS causes a genome wide downregulation of transcription (Spradling et al. 1975; Jamrich et

al. 1977), presumably to reduce the accumulation of misfolded protein aggregates. Although this has been a paradigm in the HS field, higher resolution genome-wide studies have failed to identify all the primary genes that are repressed by heat, mostly due to the limitations of measuring steady-state levels of mRNA, which requires that the mRNAs already present in the cells have shorter half-lives than the HS time points used in the experiment. Our results provide definitive evidence to support the widespread shutdown of transcription caused by HS. We identify and quantify the genes with significantly reduced transcription and demonstrate that the HS repressive response is very rapid, with over a thousand genes being repressed after only 5 minutes of HS (Figure 2.7A). Furthermore, the Pol II density in the promoter-proximal region, which represents the paused Poll molecules, is also significantly reduced across all HS repressed genes (Figure 2.19). The accumulation of Pol II in the pausing region depends on both the transcription initiation rate and the rate of escape into productive elongation. The reduction in pausing levels thus indicates that the recruitment and initiation of Pol II is affected by HS (Figure 2.20C).

The HS-induced binding of HSF is not essential for the genome-wide transcriptional repression (Figure 2.19), and given the magnitude of this repressive response, we believe that it is unlikely that one single transcription repressor is responsible for inhibiting transcription initiation in all HS repressed genes. We consider three possible mechanisms, which are not mutually exclusive, that could be responsible for HS-mediated repression. (1) The activity of a general transcription factor that is involved in recruitment of Pol II to the promoter could be modulated by heat stimulus. (2) Changes in nucleosomal composition or positioning induced by heat could generate an

unfavorable chromatin environment that would prevent transcription initiation and elongation. A previous study has demonstrated that HS results in decreased nucleosome turnover genome-wide within gene bodies; however, the same pattern was observed after drug inhibition of Pol II elongation, arguing that reduced nucleosome turnover may be a consequence rather than the cause of the genome-wide transcriptional repression (Teves and Henikoff 2011). (3) A genome-wide rearrangement of the 3D chromatin structure could either disrupt long-range interactions that are needed for transcription or allow new long-range interactions that repress transcription initiation, which is supported by a recent study in a different *Drosophila* cell line that demonstrated that HS induces a genome-wide rearrangement in the 3D nuclear architecture (Li et al. 2015). Any model must accommodate our new observations that 1) recruitment of Pol II is the step in the transcription cycle that is regulated, 2) HSF is not involved in the repression; 3) the specifically repressed genes identified here and their level of down-regulation must be accommodated by any proposed regulatory factor interactions.

CHAPTER 3

SELECTION AND CHARACTERIZATION OF RNA APTAMERS TO STUDY HEAT SHOCK FACTOR FUNCTION AND REGULATION IN VIVO[1]

3.1 Introduction

The Heat Shock (HS) response in *Drosophila melanogaster* has been used for many decades as a model system to study transcription regulation (reviewed in Guertin et al. 2010). This highly conserved protective mechanism (Lindquist and Craig 1988) is regulated at the transcriptional level by the transcription activator Heat Shock Factor (HSF). When activated by stress, HSF strongly induces the expression of HS genes, which results in the accumulation of molecular chaperones – the Heat Shock Proteins – that help the cell to cope with stressful conditions.

While HSF activity in *Drosophila* is encoded by a single gene, mammals have evolved multiple HSFs with overlapping and distinct functions, with HSF1 serving as the major regulator of the HS response (Rabindran et al. 1991; Schuetz et al. 1991; Sarge et al. 1991; Xiao et al. 1999). Its activity is not required for viability and normal cell growth, but it is essential for survival in stress conditions (Xiao et al. 1999). Besides the induction of HS Proteins, HSF1 plays important regulatory roles in other processes, such as development, postnatal growth and protection during inflammatory responses (Xiao et al. 1999). Furthermore, Lindquist and colleagues have demonstrated that HSF1 is critical for tumor formation and maintenance in response to carcinogens and oncogenes

---

[1]Portions of this chapter have been published in Latulippe et al. 2013. Anal Chem 85: 3417–24. PMID: 23398198.

(Dai et al. 2007); however, the roles of HSF1 in this process are not completely understood.

We were specifically interested in understanding in more detail the molecular interactions of HSF and the specific mechanisms that are used to execute its functions. This factor has 3 major domains: 1) trimerization domain (TD) (Rabindran et al. 1993; Wu 1995), which is important for trimer formation and activation, since only the trimeric form of HSF is able to bind DNA; 2) DNA binding domain (DBD) (Wu 1995), which is responsible for recognizing HS DNA Elements (HSEs) and recruiting HSF to the promoter of target genes; 3) activation domain (AD) (Wisniewski et al. 1996), which is responsible for recruiting co-activators and other factors that will promote the transcription of HS genes. Over the years, our research group and others have used different types of biochemical and imaging assays to investigate the molecular interactions of HSF, and we already know many of the factors with which HSF interacts (Guertin et al. 2010). However, important questions remain unanswered. For instance, we know that HSF is important for the recruitment of the kinase P-TEFb, which promotes the escape of Pol II from promoter-proximal pausing into productive elongation (Boehm et al. 2003; Peterlin and Price 2006). However, this interaction is not direct, and we still do not know the specific molecular mechanism that is involved in this recruitment (Lis et al. 2000; Boehm et al. 2003). Therefore, although we have learned important aspects of HSF interactions, we are still limited by the available methods and lack approaches that allow us to perturb the activity of specific factors to tease apart molecular interactions.

Macromolecular interactions are generally studied using strategies for perturbing protein function, such as RNAi, knockouts and genetic mutations. When using such

methods, the distinction between the function of an individual domain rather than the entire protein – which may have multiple domains and functions – as well as the primary and secondary effects of disrupting the interaction of a protein with its partners are often unclear, making it difficult to determine the primary functions of each protein (or domain). To overcome this limitation, ligands are needed that can be quickly produced in vivo, bind with high affinity and specificity to a particular region of a protein and thus block its interactions with one or more partners. To address this need, we endeavored to generate highly specific inhibitory RNA aptamers to target different surfaces of the proteins of interest.

Aptamers (Ellington and Szostak 1990) are single-stranded oligonucleotides – DNA, RNA or modified nucleic acids – that can fold into diverse and intricate three-dimensional structures that bind proteins, peptides or small molecules with high affinity. Equilibrium dissociation constants (K_D) typically range from micromolar to picomolar values. These molecules are selected in vitro using an iterative process called SELEX (Systematic Evolution of Ligands by Exponential Enrichment) (Ellington and Szostak 1990; Tuerk and Gold 1990). In this technique, one starts with an enormously large library of randomized sequences – typically on the order of 10^{14}-10^{15} unique molecules – that have diverse structures based on sequence variation to identify those that can bind specifically to a target of interest. After incubation with the target, the bound species are separated from the unbound ones (partition step), reverse transcribed (when using RNA libraries), PCR amplified and transcribed, and the obtained pools are submitted to further rounds of selection. After many rounds, the aptamers that bind with high affinity to the target are usually enriched and dominate the pool of sequences.

Aptamers can be applied in a number of ways to address various biological and technical questions. In particular, they can act as inhibitors that bind to a protein surface and disrupt specific interactions or functions. When expressed in vivo in a temporally and spatially controlled manner, these aptamers provide a way to rapidly disrupt targeted domains of proteins and efficiently assess their primary functions and mechanisms of actions.

Our group has been successfully using inhibitory RNA aptamers for many years to study macromolecular interactions in vivo (Shi et al. 1999; Fan et al. 2004, 2005; Zhao et al. 2006; Shi et al. 2007; Salamanca et al. 2011). However, there were some limitations on the methodology used to select those aptamers, such as the efficiency of the SELEX procedure and the size and quality of the RNA library. More recently we have significantly improved our methodology by generating a more complex, well-characterized library, developing a more efficient multiplex SELEX procedure that allows the selection of aptamers for many targets at the same time, and using high-throughput sequencing (as opposed to traditional cloning) to analyze the selected pools, which reduces the number of rounds that need to be performed to allow the identification of enriched sequences (Latulippe et al. 2013; Szeto et al. 2014).

The first step in the application of this inhibitory aptamer technology to study HSF's primary functions and mechanisms of action was the selection of RNA aptamers to individual domains of HSF. One of the major challenges in the selection of aptamers that bind to the target protein with high affinity is the identification of the best candidate sequences from a high-throughput sequenced pool that contains thousands of distinct sequences. Here, we describe the results of a successful RNA aptamer selection to HSF1

and HSF2 using the new SELEX reagents and technologies developed by our group in collaboration with Harold Craighead's group (School of Applied and Engineering Physics, Cornell University). Furthermore, we report a thorough characterization of the sequenced pools from these selections and the development and implementation of a set of SELEX performance metrics that can be used to evaluate the success of a selection. We plan to express the selected aptamers in vivo to dissect the roles of HSF in transcription regulation.

3.2 Materials and Methods

Purification of recombinant protein targets

Recombinant proteins were expressed in BL21 (DE3) *E. coli* transformed with plasmids that encode for GST-tagged human HSF1, HSF2, HSF1-DBD, HSF2-DBD, HSF1-TD-AD and HSF2-TD-AD. One liter LB cultures supplemented with 100 µg/mL ampicillin were inoculated with 10 mL from a 30 mL starter LB culture derived from a single colony and grown at 37°C until the OD600 reached approximately 0.6. Protein expression was induced by the addition of IPTG to a final concentration of 0.2 mM. After an overnight incubation at 18°C, bacteria were collected by centrifugation and the resulting pellet was processed according to the manufacturer's instructions for glutathione-agarose (Thermo Scientific) resin. SDS-polyacrylamide gel electrophoresis (SDS-PAGE) was used to verify the purity and quality of the final protein product. Resulting protein preps were dialyzed against 1× PBS (supplemented with 5 mM 2-mercaptoethanol and 0.01% Triton X-100) and stored in small aliquots after addition of glycerol to a final concentration of 20%.

Preparation of protein-immobilized resin

For each selection round, a fresh batch of protein-bound resin was prepared. Glutathione-agarose resin was extensively washed (4 times) with SELEX binding buffer [10 mM N-2-hydroxyethylpiperazine-N′-ethanesulfonic acid (HEPES)-KOH pH 7.6, 125 mM NaCl, 25 mM KCl, 1 mM $MgCl_2$, and 0.02% Tween-20] to remove any residual storage components. GST-tagged proteins were prepared as described above and immobilized onto the washed resin at 4°C for 2 hours with constant mixing. The protein-bound resin was then degassed in a vacuum desiccator for approximately 20 min and carefully pipetted into the microcolumn device.

Microcolumn SELEX protocol

All of the solutions were degassed prior to use and introduced into the microcolumns via a standard syringe pump (Harvard Apparatus). First, yeast tRNA at a final concentration of 500 µg/mL in SELEX binding buffer was introduced to block any possible nonspecific RNA binding sites. For each loading step, the RNA library was diluted in 1 mL of SELEX binding buffer, heat-denatured at 65°C for 5 min, renatured by cooling down to room temperature for 10 min while degassing, and then spiked with 200 units of SUPERase-In RNase inhibitor. A 10 µL aliquot was collected and used as a standard for the quantitative polymerase chain reaction (qPCR) analysis. The RNA library was injected into the microcolumns at a rate of 1 µL/min (approximately 14 hours). Each device was then washed with 3 mL of SELEX binding buffer at a rate of 100 µL/min to remove unbound RNA. Finally, the RNA was eluted from individual microcolumns by flowing elution buffer [SELEX binding buffer + 50 mM ethyl-enediaminetetraacetic acid (EDTA)] at a rate of 50

μL/min for 6 min. Eluted RNA and the input samples were phenol/chloroform and chloroform extracted and isopropanol precipitated together with 1 μL of GlycoBlue and 50 μg of yeast tRNA, and the resulting pellet was resuspended in RNase-free water. Both the resuspended pools and standards were reverse transcribed with Moloney murine leukemia virus reverse transcriptase (MMLV-RT) using the “Apt Lib Const REV” oligo (5’-AAGCTTCGTCAAGTCTGCAGTGAA-3’). Residual RNA was eliminated by treating the samples with RNase H and RNaseA/T1 cocktail. A small amount (less than 5%) of the cDNA product was analyzed on a LightCycler 480 qPCR instrument using “T7 pro-Apt Lib Const FOR” (5’-GATAATACGACTCACTATAGGGAATGGATCCACATCTACGA-3’) and “Apt Lib Const REV” to determine the amount of RNA library that was retained on each device. The cDNA samples from each round were PCR amplified using the same oligos (“T7 pro-Apt Lib Const FOR” and “Apt Lib Const REV”), gel purified from an 8% native PAGE and then subjected to phenol/chloroform and chloroform extractions and ethanol precipitation. A fraction of the purified PCR product was used to make the RNA pool for the next round of SELEX. A typical 60 μL transcription reaction consisted of ~100 ng of template DNA, 455 nmol of each ribonucleoside triphosphate (rNTP), T7 RNA polymerase and 60 units of SUPERase-In RNase inhibitor. The reactions were incubated at 37°C overnight and the resulting RNA pool was treated with DNase I to remove the template DNA, phenol/chloroform and chloroform extracted and isopropanol precipitated. The purified pools were then verified by denaturing PAGE for length and purity and quantified by Qubit BR RNA assay before being used for the next round of SELEX.

High-throughput sequencing of selected pools

The SELEX DNA pools were PCR amplified with primers containing a unique 6 nt barcode and the adapters necessary for Illumina sequencing. Sequenced pools and barcodes used are described in Table 3.1. After PCR, the libraries were gel purified from an 8% PAGE, phenol/chloroform and chloroform extracted and ethanol precipitated. The 10 libraries were then mixed together and sequenced in a 100 nt single-end run on the Illumina HiSeq 2000, using standard protocol at the Cornell Biotechnology Resource Center (http://www.BRC.cornell.edu). The reads generated by this 100 nt sequencing run were used for all the processing and clustering steps described below. However, 100 nt reads could only yield the first 68 nt of the 70 nt random region (26 nt forward constant region + 6 nt barcode + 68 nt random region = 100 nt). Therefore, we later submitted the library to a 113 nt single-end run on the Illumina HiSeq to obtain the complete sequence of the aptamers.

Table 3.1: High-throughput sequencing of SELEX libraries. The 10 SELEX DNA libraries that were sequenced on the Illumina HiSeq 2000 and the IDs and sequences of the barcodes used are shown in the table. The final DNA pools for all target proteins (after 5 rounds of SELEX) were sequenced and we also sequenced the DNA pools obtained after 3 rounds of SELEX for HSF1 TD-AD, HSF1 and HSF2.

Library	Barcode ID	Barcode sequence
GST - Round 5	3	CGAGAT
HSF1 TD-AD - Round 5	4	ACACTG
HSF2 TD-AD - Round 5	5	CATTCG
HSF1 - Round 5	6	GCATAG
HSF2 - Round 5	7	ACTAGC
HSF1 DBD - Round 5	8	CAGTAC
HSF2 DBD - Round 5	9	TGCAAC
HSF1 TD-AD - Round 3	10	TTGCGA
HSF1 - Round 3	11	GCTACA
HSF2 - Round 3	12	GAGCAA

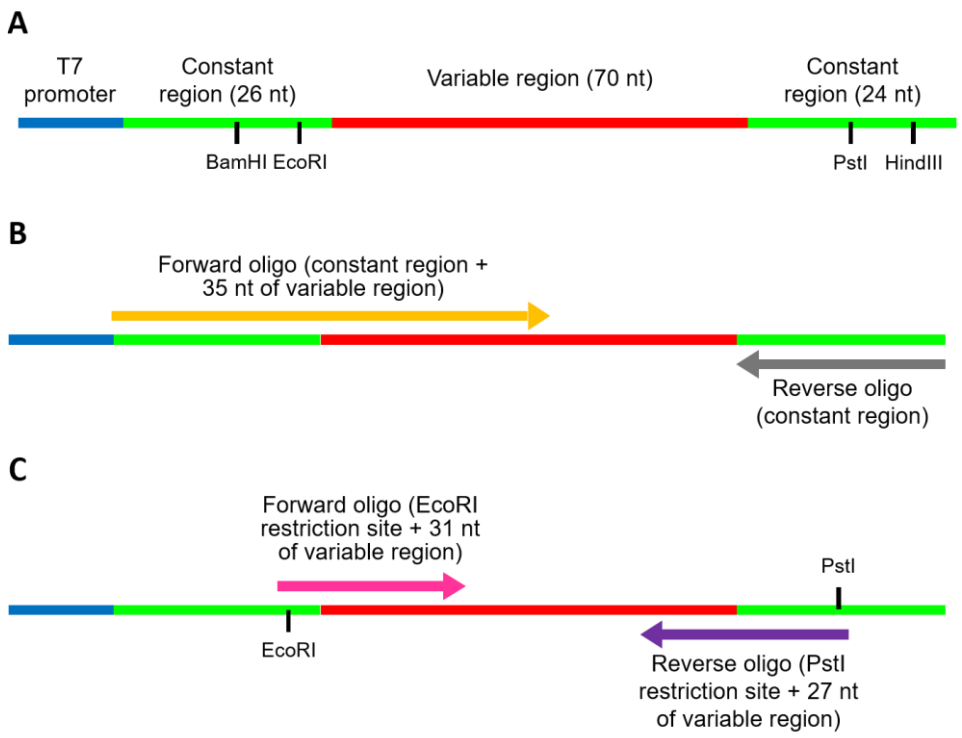
The raw sequencing reads were processed as described previously (Latulippe et al. 2013). Briefly, reads with low quality scores were filtered out and the remaining reads were separated based on the barcodes. The forward constant regions were removed; however, given the length of the reads, the reverse constant region and the Illumina sequencing adapters were not present and did not need to be removed. Remaining reads that were between 64 and 72 nt in length and identical in sequence were collapsed and reads with 80% or higher sequence identity were clustered together. This clustering was performed to account for sequencing errors that may induce apparently distinct reads from a single aptamer. The cluster representative, which is the read with the highest multiplicity within each cluster, was identified as the true aptamer sequence. The multiplicity of each cluster was defined as the sum of multiplicities within the cluster and was normalized by the total number of reads in each individual library. Representative sequences and their normalized multiplicities were tabulated and sorted based on the multiplicity across different pools. For the selections for which we sequenced the DNA pools from both round 3 and 5 (HSF1 TD-AD, HSF1 and HSF2) we also calculated enrichment factors (ratio of the normalized multiplicity of a given cluster in round 5's pool/normalized multiplicity of the same cluster in round 3's pool) for each cluster.

Cloning of candidate aptamer sequences from SELEX DNA pools

Candidate aptamer sequences were PCR amplified from the final DNA pool (round 5) of each protein target with Phusion DNA Polymerase using two different cloning strategies. These two strategies took advantage of the four restriction sites that are present in the two constant regions (Figure 3.1A, BamHI and EcoRI in the forward constant region and

PstI and HindIII in the reverse constant region). The first strategy used aptamer specific forward oligos and “Apt Lib Const REV” (Figure 3.1B, Table 3.2). Aptamer specific oligos span the entire forward constant region and 35 nt of the variable region specific to each aptamer. The BamHI restriction site in the forward constant region and the PstI site in the reverse constant region were used for cloning the aptamers. The second strategy used aptamer specific forward and reverse oligos and only the minimal amount of the constant regions that was necessary to reach the constant regions’ restriction sites that were more proximal to the variable region (Figure 3.1C, EcoRI in the forward region and PstI in the reverse region). A few aptamers had either an EcoRI or a PstI site in their variable region, so the second restriction site had to be used (Table 3.3).

The resulting PCR products were double digested with the restriction enzymes listed in Tables 3.2 and 3.3 and ligated into the plasmid pGEM3Z-N70Apt, which had been cut with the same enzymes. This plasmid was obtained by cloning a random aptamer sequence together with the T7 promoter into the pGEM3Z vector (Promega) between NarI and HindIII sites (Latulippe et al. 2013). The obtained clones were sequence verified and all of the candidate aptamers that were used in subsequent assays were prepared from the sequence verified constructs to ensure the purity of the intended aptamer in each preparation.



<u>Figure 3.1:</u> Strategies used to clone candidate aptamer sequences from SELEX DNA pools. (A) Each aptamer in the library consists of a 70 nt variable region flanked by two constant regions. The restriction sites in the constant regions are shown in the figure. **(B)** The first cloning strategy used an aptamer specific forward oligo and "Apt Lib Const REV", which spans the entire reverse constant region. The forward oligos span the entire forward constant region and 35 nt of the variable region. **(C)** The second cloning strategy used aptamer specific forward and reverse oligos. These oligos span only the minimal amount of both constant regions that was necessary to reach their most proximal restriction sites.

Table 3.2: Oligos and restriction sites used for cloning candidate aptamers. The sequences of the forward oligos used are shown, with the forward constant region underlined. The same reverse oligo ("Apt Lib Const REV") and restriction sites (BamHI and PstI) were used for all the aptamers listed.

Aptamer ID	Forward oligo	Reverse oligo	Restriction sites
HSF1_TD_AD_R5_1	GGAATGGATCCACATCTACGAATTCACCGTTGACCCGTCAACCTATGCATCCCAAACCCA	Apt Lib Const REV	BamHI, PstI
HSF1_TD_AD_R5_10	GGAATGGATCCACATCTACGAATTCCTACCCGTTTGTGACACTACCGATAGCGACCAGGT	Apt Lib Const REV	BamHI, PstI
HSF1_TD_AD_R5_22	GGAATGGATCCACATCTACGAATTCTACAACCCGACATGTCAAGTAACGTTACTTCTCCC	Apt Lib Const REV	BamHI, PstI
HSF1_R5_1	GGAATGGATCCACATCTACGAATTCCACTCCATAGTATCTAGAAGCCCTGCCGAAAACGA	Apt Lib Const REV	BamHI, PstI
HSF1_R5_6	GGAATGGATCCACATCTACGAATTCATGCCAAACCCGGACTATAGTGATACGGACGGAGA	Apt Lib Const REV	BamHI, PstI
HSF2_R5_2	GGAATGGATCCACATCTACGAATTCAATCAAGTCCCCAGACTCAGCAACACTGGACAGCG	Apt Lib Const REV	BamHI, PstI
HSF2_R5_5	GGAATGGATCCACATCTACGAATTCAGAACTGGGCGGACAATTAATATAACGGCACACTG	Apt Lib Const REV	BamHI, PstI

Table 3.3: Oligos and restriction sites used for cloning candidate aptamers. The sequences of the forward and reverse oligos used are shown, with the respective constant regions underlined.

Aptamer ID	Forward oligo	Reverse oligo	Restriction sites
HSF1_R5_2	TACGAATTCCGGGCCTCACGCAGCA CGCCCCTGCGCGTAG	AGTCTGCAGTGAAGAGCTTGGCGTC TGTGGGTGCGGGTCC	EcoRI, PstI
HSF1_R5_3	TACGAATTCGCCACGGAAGACACCT TAGTATCCATACCCT	AGTCTGCAGTGAAGTCAGGAGGAGA TTGCGAATCTCTGGC	EcoRI, PstI
HSF1_R5_4	TACGAATTCCGCCCTCTGGCTTCCA GGTGCTTCAGTAAAT	AGTCTGCAGTGAACGCCTGATCGCT GGGTGTCAGGTAAGC	EcoRI, PstI
HSF1_R5_5	TACGAATTCCCGGCCTTTCCGCTTC CACGTTCCTGAGCTG	AGTCTGCAGTGAAATCGCCAGGGCA GCGTGCTGGCGTCGG	EcoRI, PstI
HSF1_R5_7	TACGAATTCGCGGGCGAAACCGGA CTAGCCTCCCGGCGTA	AGTCTGCAGTGAATATATGGTGGAT GGATTCGAACTCTTA	EcoRI, PstI
HSF1_R5_8	TACGAATTCCCTGCTCGACCCGAAT GTCCACGTAGTCGAG	AGTCTGCAGTGAAACTAAGAAGTAC TTCTTATCTTAGTCT	EcoRI, PstI
HSF1_R5_9	TACGAATTCCCTCGCTTACTACAGC AGCTAGCCCACCCGG	AGTCTGCAGTGAAATGTGTGCCGGT CACTTTAGATCACTG	EcoRI, PstI
HSF1_R5_10	TACGAATTCCCCGTCCTAAAGCACC CTCGACTATGGAAAG	AGTCTGCAGTGAATGCCCTCATCTG TGTTTCATGGTTCCC	EcoRI, PstI
HSF1_R5_11	TACGAATTCCCGGCTCCTCTTAGTC GCAAATAGTCTGGCG	AGTCTGCAGTGAATCCCTTGCGGGC ACTACGTCGGGTTCG	EcoRI, PstI
HSF1_R5_19	TACGAATTCCCCAACTACCCACGAA ACCAATTGGGTTGCA	AGTCTGCAGTGAATATCGCTTCACT GAGGGCGGCATGTTT	EcoRI, PstI
HSF1_R5_31	TACGAATTCCACCACAAGCCTGACG CTTGGGAAACTTACG	AGTCTGCAGTGAAAGCGTACGTCGG GTTCGGCGCTTTCGT	EcoRI, PstI
HSF1_R5_62	TACGAATTCCCCAGTGCACGAGGCA CCACAAAGACAAATC	AGTCTGCAGTGAAGGCTAAGTGTAT CCCATATGGGCCCTT	EcoRI, PstI
HSF1_R5_97	TACGAATTCCCACGCTGCTGTGTTC CGATCCATGTGAAAA	AGTCTGCAGTGAACTGAACTATGTT ACCCGGGGCCTGCCG	EcoRI, PstI
HSF1_R5_154	TACGAATTCGATTCGAAATCTAGAG CAACTCAGGATAGAC	AGTCTGCAGTGAATCGATTCTACTT GTACAGGTCTAGTGT	EcoRI, PstI
HSF1_R5_248	TACGAATTCCTATCCGCCACTGCCC TTTCACTAGACCCGG	AGTCTGCAGTGAAGCTGCGTGTGGC CCTTACCCTAGTGTG	EcoRI, PstI
HSF1_R5_368	TACGAATTCCGCGGCCAAGTCATTC TGATTGCCCAGTGAC	AGTCTGCAGTGAAATGCTAGTAGTT CCCCAAGAATCGGAC	EcoRI, PstI
HSF1_R5_560	TACGAATTCCGGTCCCCCTCCCCCC TCCGCACCACCGCAA	AGTCTGCAGTGAAGACTCCTAAGCG TTCCCGGCGCATAAC	EcoRI, PstI
HSF2_R5_1	TACGAATTCGGAAAGGCGTACAAGA CGTACCTGGACCCAG	AGTCTGCAGTGAATCGCCGAATCGG GCGCGTGTATCTACA	EcoRI, PstI
HSF2_R5_3	AATGGATCCACATCTACGAATCCTC CTCACTAACGCCACGTAAGCACCCG GATGATATGG	AGTCTGCAGTGAAACGTTAATCGAG TCGGCTTTACAGTGTACCCTGGCCA TATCATCCGG	BamHI, PstI
HSF2_R5_4	TACGAATTCGCCCTGCAGACATATC CACGGCAGCCACTAGAAATATAGTA CCCGGTAAGA	GCTAAGCTTCGTCAAGTCTGCAGTG AACCATATCCGCCATCTACCTCTTAC CGGGTACTA	EcoRI, HindIII
HSF2_R5_6	TACGAATTCAACAAACCTGCGATGA CCAAGCACCGACGGA	AGTCTGCAGTGAACGGCATATACGG CCTTACCCGGTACTA	EcoRI, PstI
HSF2_R5_7	TACGAATTCGAGCGCAAGAACCTAC CAGGTAATCCAGAAC	AGTCTGCAGTGAACTTAATCCACAC CGTTGTTCGGTTATT	EcoRI, PstI

Fluorescence electrophoretic mobility shift assay (F-EMSA)

Candidate aptamers were PCR amplified from the sequence verified constructs described above using “T7 pro-Apt Lib Const FOR” and “Apt Lib Const REV”. The PCR products were purified with DNA Clean & Concentrator spin columns (Zymo Research) and used

as templates for in vitro transcription reactions with T7 RNA polymerase. The reactions were incubated at 37°C overnight and the resulting RNA products were treated with DNase I to remove the template DNA, phenol/chloroform and chloroform extracted, isopropanol precipitated and gel purified from a denaturing 8% PAGE. The purified products were then verified by denaturing PAGE for length and purity and 3′-end labeled with fluorescein 5-thiosemicarbazide as described previously (Pagano et al. 2007). Binding reactions (50 µL) were prepared by mixing 2 nM of 3'-end labeled RNA aptamer with protein concentrations that ranged from 0.2 to 2000 nM in 1.5-fold increments in SELEX binding buffer containing 0.01% IGEPAL CA-630, 10 µg/mL yeast tRNA and 3U of SUPERase-In RNase inhibitor. Reactions were incubated at room temperature for 2 h, mixed with loading dye, and then loaded into the wells of a refrigerated 1.5% agarose gel prepared with 0.5× Tris-Borate-EDTA (TBE) buffer with 1 mM $MgCl_2$. The gel was run for 80 min at 100 V in refrigerated 0.5× TBE. Images were acquired with the fluorescein scan settings on a Typhoon 9400 imager (GE Healthcare Life Sciences). The resulting bands were quantified with the ImageQuant software and the data were fit to the Hill equation using Igor (Wavemetrics) to estimate the equilibrium dissociation constants (K_D).

Fluorescence polarization (FP) assay

Binding reactions were prepared by mixing 2 nM of 3'-end labeled RNA aptamer with protein concentrations that ranged from 0.2 to 2000 nM in 1.5-fold increments in SELEX binding buffer containing 0.01% IGEPAL CA630, 10 µg/mL yeast tRNA and 3U of SUPERase-In RNase inhibitor. The reactions were prepared in black 96-well half area microplates (Corning) and then incubated at room temperature for 2 hours. The plates

were read on a Synergy H1 Microplate Reader (BioTek); fluorescence polarization was measured as $(F_{\parallel}-F_{\perp})/(F_{\parallel}+F_{\perp})$ using the Ex: 485/20 Em: 528/20 filter set.

3.3 Results

Selecting RNA aptamers to different HSF domains

The first step in the development of our strategy to use inhibitory RNA aptamers to dissect the specific functions and interactions of HSF was the selection of RNA aptamers that can bind with high affinity and specificity to the three major domains – activation, trimerization and DNA binding – of the human HSF1 and HSF2.

These aptamers were selected using a new SELEX approach developed in collaboration with Harold Craighead's group (Latulippe et al. 2013). Every step and component of the procedure was thoroughly characterized and optimized and the HSF aptamer selection was used as a proof-of-principle to demonstrate the efficiency of our new approach (Latulippe et al. 2013).

The single stranded RNA library used in the SELEX contains 5×10^{15} unique sequences and was in vitro transcribed from a DNA library that was chemically synthesized by GenScript (Latulippe et al. 2013). Each RNA molecule contains a core 70 nucleotide random region flanked by two constant regions (120 nt total length, 5′-GGGAAUGGAUCCACAUCUACGAAUUC-N70-UUCACUGCAGACUUGACGAAGCUU-3′) as described previously (Cox et al. 1998; Hall et al. 2009). The starting pool of the SELEX experiment contained approximately 5 copies of each unique sequence in the library.

As protein targets for the SELEX, we used the human full-length HSF1 and HSF2,

variants that contain only the DNA-binding domain (HSF1-DBD and HSF2-DBD) and variants lacking the DBD (HSF1-TD-AD and HSF2-TD-AD), with the goal of isolating aptamers that can bind to all of HSF1 and HSF2 domains. GST-tagged versions of these proteins were cloned and expressed in *E. coli* and purified through affinity chromatography using glutathione agarose beads (Figure 3.2).

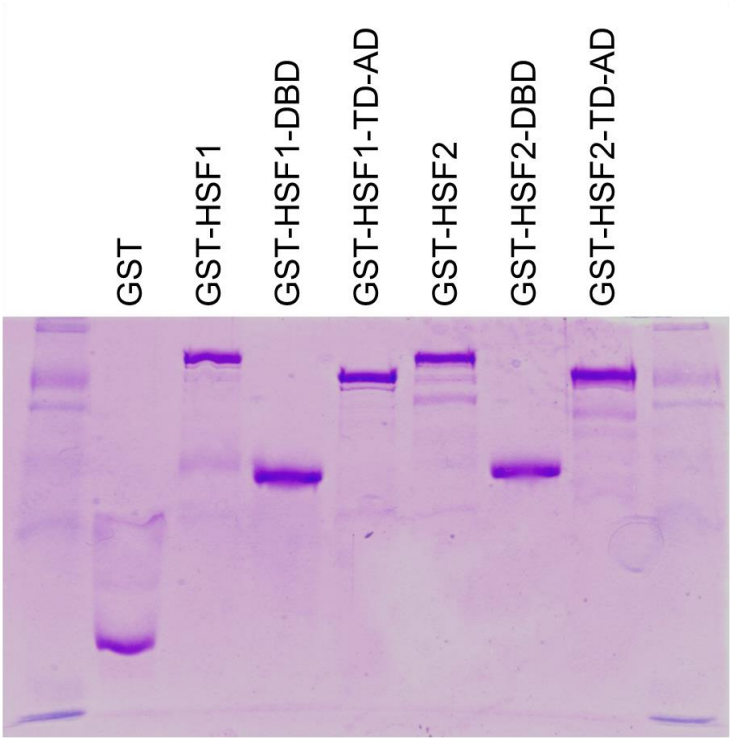


Figure 3.2: Protein targets used in the SELEX experiment. Coomassie blue stained SDS-PAGE gel showing the GST-tagged proteins that were used as targets in the SELEX experiment.

The SELEX was performed using a new device that was designed and fabricated by our collaborators in the Craighead lab: miniaturized affinity chromatography columns, which have an internal chamber that can be loaded with protein-bound resin. The RNA library, washing and elution buffers were injected into these microcolumns and flowed

through the resin at a controlled flow-rate using standard syringes connected to a pump. This device presents many advantages over traditional strategies. Since it is affinity chromatography-based, the same affinity tags and resins that were used to purify the proteins can be used to immobilize the targets inside the microcolumns. Moreover, they can directly interface with each other and with most standard fluid handling systems, making it easier to multiplex and automate the method and to interchange between serial and parallel conditions, allowing the selection of aptamers for many targets at once (Latulippe et al. 2013).

Each protein/domain was incubated with glutathione-agarose resin (~1µg/µL final concentration) and the mixtures were loaded into a 20µL microcolumn. Before injecting the RNA library into the protein-loaded microcolumns, we incubated the library with just the resin to reduce the likelihood of selecting aptamers that bind to the resin. We performed a total of 5 selection rounds. For the first round, the targets were connected serially (Figure 3.3A) in the following order: 1) GST, 2) HSF1-TD-AD, 3) HSF2-TD-AD, 4) HSF1, 5) HSF2, 6) HSF1-DBD, 7) HSF2-DBD. The microcolumn loaded with GST-immobilized resin was added as an in-line negative selection to enhance the specificity to the target proteins. After the RNA library injection and washing steps the microcolumns were disconnected and the elution step was performed with the microcolumns arranged in a parallel configuration (Figure 3.3B). The eluates were reverse transcribed, PCR amplified and transcribed using T7 RNA polymerase and the pools obtained were used as the starting material for round 2. For rounds 2 through 5, the microcolumns were arranged in a parallel configuration with a 10 µL microcolumn filled with GST-immobilized resin serially connected to the inlet of each of the six microcolumns with the target proteins

(in-line negative selection, Figure 3.3C). The GST microcolumns were disconnected after the RNA injection step for the subsequent washing and elution of the target protein-bound aptamers from each microcolumn (Figure 3.3B).

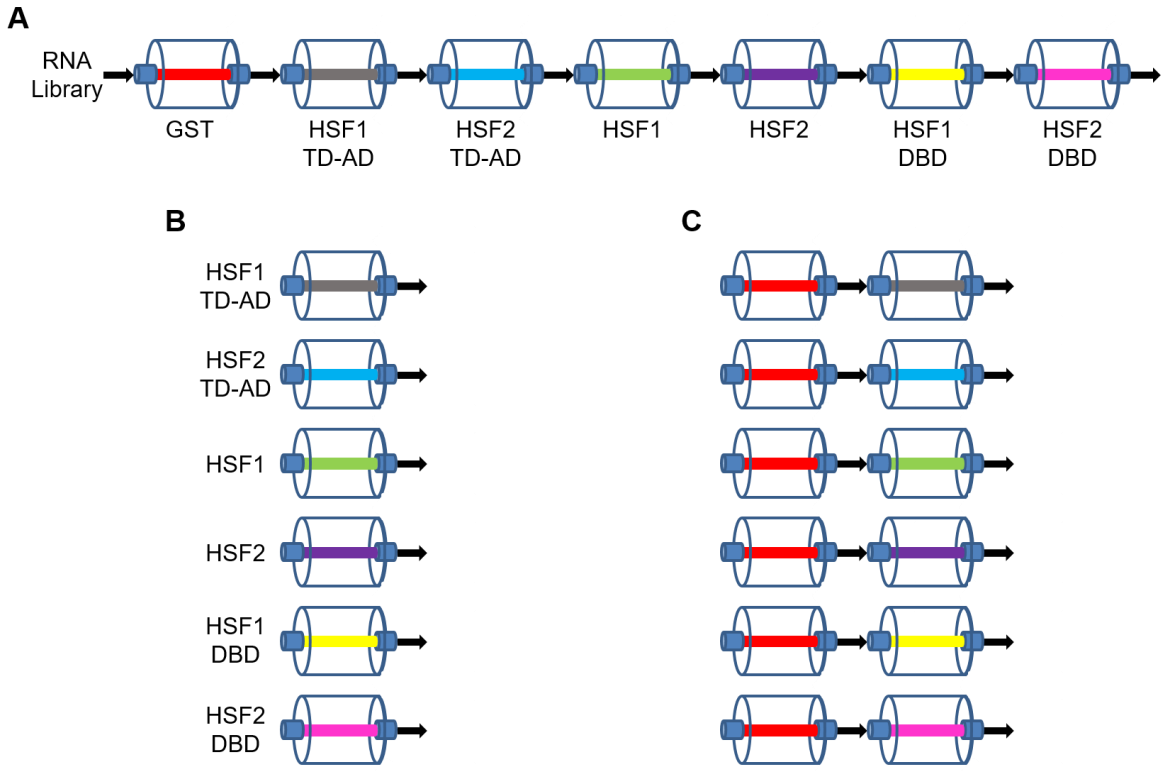


Figure 3.3: Microcolumn configurations used in the SELEX experiment. (A) For the first round of SELEX, the microcolumns filled with protein-immobilized resin were connected in a serial configuration for the RNA library injection and washing steps. **(B)** The microcolumns were arranged in a parallel configuration for the elution step of the first round of SELEX and for the washing and elution steps of rounds 2 through 5. **(C)** For the RNA injection step of rounds 2 through 5, the microcolumns were arranged in a parallel configuration with a GST microcolumn serially connected to the inlet of each of the six microcolumns with the target proteins.

Developing SELEX performance metrics to evaluate the success of aptamer selections

We sequenced the pools obtained after 5 rounds of selection for each protein/domain using the Illumina HiSeq platform to identify the sequences that were selected within each

pool. In order to assess how each sequence was enriched during the SELEX process, we also sequenced the pools obtained after 3 rounds of selection for HSF1, HSF2 and HSF1-TD-AD. Each pool was uniquely tagged with a sequence barcode, which allowed the sequencing of all the pools on the same lane. The raw sequencing data was processed and clustered as described in the Materials and Methods and the total number of processed reads per pool ranged from ~6 million to ~9 million.

We thoroughly characterized the selected pools for HSF1 and HSF2 and implemented a set of SELEX performance metrics to evaluate if our selections were successful. These analyses were based on two values: (1) multiplicity, which corresponds to the normalized copy number of each sequence in a pool; (2) enrichment, which corresponds to the ratio of the multiplicity of a given sequence in round 5's pool/multiplicity of the same sequence in round 3's pool (see Materials and Methods for a detailed explanation of how multiplicity and enrichment were calculated). For both HSF1 and HSF2, there was a noticeable shift toward higher multiplicity values from round 3 to round 5. In round 3, the top 20 highest multiplicity sequences for each protein represented only ~0.04% of the total pool. However, in round 5, the top 20 sequences represented 85% and 76.5% of the HSF1 and HSF2 selected pools, respectively (Tables 3.4 and 3.5 and Figure 3.4). In addition, of the top 20 highest multiplicity sequences in round 3, between a quarter and a half of them were also among the top 20 highest multiplicity sequences in round 5 for HSF2 and HSF1.

Table 3.4: HSF1 high-throughput sequencing results. Top 20 highest multiplicity sequences from the round 5 pool of the HSF1 SELEX and the corresponding round 3 ranks and multiplicities. The multiplicity values are presented as multiplicity per 10 million reads (normalized by total number of sequencing reads, which was 5,930,382 for round 5 and 8,072,400 for round 3). Only the random region of each aptamer sequence is shown.

Aptamer ID	Sequence	Round 5 Multiplicity	Round 3 Rank	Round 3 Multiplicity
HSF1_R5_1	CACTCCATAGTATCTAGAAGCCCTGCCGAAAACGAGAGCGCTTACGCTTATAAATGATCAACCTACCTCG	1340479	11	133
HSF1_R5_2	CGGGCCTCACGCAGCACGCCCCTGCGCGTAGATAGACTGATCGGGACCCGCACCCACAGACGCCAAGCTC	1274151	9	152
HSF1_R5_3	GCCACGGAAGACACCTTAGTATCCATACCCTCTCTAGTAAGAAGCCAGAGATTCGCAATCTCCTCCTGAC	1261696	6	190
HSF1_R5_4	CGCCCTCTGGCTTCCAGGTGCTTCAGTAAATCAACGACTACCGGCTTACCTGACACCCAGCGATCAGGCG	621764	22	98
HSF1_R5_5	CCGGCCTTTCCGCTTCCACGTTCCTGAGCTGAAATAAAAACGGCCGACGCCAGCACGCTGCCCTGGCGAT	550234	54	55
HSF1_R5_6	ATGCCAAACCCGGACTATAGTGATACGGACGGAGAGGTGGTTCCACTTTTTCCCCTTGCATAGGCAATG	530607	24	93
HSF1_R5_7	GCGGGCGAAACCGGACTAGCCTCCCGGCGTAAGAGTTCGAATCCATCCACCATATA	469695	3	264
HSF1_R5_8	CCTGCTCGACCCGAATGTCCACGTAGTCGAGAGGTACTGCTACAGACTAAGATAAGAAGTACTTCTTAGT	405537	30	79
HSF1_R5_9	CCTCGCTTACTACAGCAGCTAGCCCACCCGGAACTTGGGGCGACAGTGATCTAAAGTGACCGGCACACAT	358828	7	165
HSF1_R5_10	CCCGTCCTAAAGCACCCTCGACTATGGAAAGAGAGGACAAAGGGGAACCATGAAACACAGATGAGGGCA	272611	27	84
HSF1_R5_11	CCGGCTCCTCTTAGTCGCAAATAGTCTGGCGTGATTAAGGCACCCGAACCCGACGTAGTGCCCGCAAGGGA	271480	10	147
HSF1_R5_12	CCTAGGTCCATACTACAATAGCTCGATCTTGTGCCAGTGAAGGGGTACATCCGGGGACAGGGGGAAAGGGG	191659	1	500
HSF1_R5_13	TCCTCACTAACGCCACGTAAGCACCCGGATGATATGGCCAGGGTACACTGTAAAGCCGACTCGATTAACGT	170782	5	190
HSF1_R5_14	ACGTCTCAGTAAAGGCCGCGACGGCCATATGGCTGTTTGACACGGCAACCTCGCTAACTCGGGAGCAACC	156627	41	66
HSF1_R5_15	CGAACTTTAGCCCGACGCACGCGTAGCTAACCGACCCACGTCCTATAAACGGACGGAAGCCGGGTGCGAGC	143621	26	89
HSF1_R5_16	AACCCTACCTACATGCAAATAGTCAAGCCAAGTCCGGAGAACGTTACCAATGTGGACAACCTTCGACCGG	127020	2	347
HSF1_R5_17	AATGGCAACACGAAATTCGGAATCCAACAACTAGAGAATAGACTCCCAGTTCGTCACCGTTAGCCCCCAA	98130	40	68
HSF1_R5_18	ACCGGGCTACAGCCAGCACGCCCCTGGCCTAGATGACACTCATTCCTGGTCGATGTCAAAACGTGAACCC	97392	8	154
HSF1_R5_19	CCCAACTACCCACGAAACCAATTGGGTTGCAACAAGACCATAAACATGCCGCCCTCAGTGAAGCGATA	84796	35	73
HSF1_R5_20	AGGCCCTCAGCAACCGAAGCCTACTAAACCCACACGTAATGACAAGCGGCTCCGAGAGGTAAACGGAGTAT	76493	43	64

Table 3.5: HSF2 high-throughput sequencing results. Top 20 highest multiplicity sequences from the round 5 pool of the HSF2 SELEX and the corresponding round 3 ranks and multiplicities. The multiplicity values are presented as multiplicity per 10 million reads (normalized by total number of sequencing reads, which was 8,644,268 for round 5 and 9,350,602 for round 3). Only the random region of each aptamer sequence is shown.

Aptamer ID	Sequence	Round 5 Multiplicity	Round 3 Rank	Round 3 Multiplicity
HSF2_R5_1	GGAAAGGCGTACAAGACGTACCTGGACCC AGACCAAAGACACCTGTAGATACACGCGCC CGATTCGGCGA	1627399	51	87
HSF2_R5_2	AATCAAGTCCCCAGACTCAGCAACACTGGA CAGCGATATGCAGATAACCAAGACCAA	682671	6	205
HSF2_R5_3	TCCTCACTAACGCCACGTAAGCACCCGGAT GATATGGCCAGGGTACACTGTAAAGCCGAC TCGATTAACGT	660708	47	93
HSF2_R5_4	GCCCTGCAGACATATCCACGGCAGCCACTA GAAATATAGTACCCGGTAAGAGGTAGATGG CGGATATGG	569105	33	103
HSF2_R5_5	AGAACTGGGCGGACAATTAATATAACGGCA CACTGATAGGTGCTAGGCCGCCACGTCGA CGACGATCCAT	567516	511	27
HSF2_R5_6	AACAAACCTGCGATGACCAAGCACCGACG GACATAGGAATCTAGTACCGGGTAAGGCCG TATATGCCG	558955	49	89
HSF2_R5_7	GAGCGCAAGAACCTACCAGGTAATCCAGAA CTGCTTGCGAGCCAATAACCGAACAACGGT GTGGATTAAG	485019	342	33
HSF2_R5_8	CAGCCGCTTCCACTGACCTTGAGTGCGCC GCGATAATGTTCACGGACAAAGGGCTGCAG GCAGCCTGTATG	384008	9	197
HSF2_R5_9	CCCGTCCTAAAGCACCCTCGACTATGGAAA GAGAGGACAAAGGGGAACCATGAAACACA GATGAGGGCA	370668	69	73
HSF2_R5_10	TACCCCTATTACTGCTCGCGCTCGGAACCC ATTGAACTCTAGGATACACGGGCCTGACAA ACCGATGCGCC	304918	93	64
HSF2_R5_11	CAAAATGACCCAGATAAGCCCCCGAATAGA CCCTGGTCGGGGACCTCACACTCGTATGCT GTTAGTGAC	225959	295	36
HSF2_R5_12	CACACCGCCTGAAGCCCCCGGATAAAGAG CGCCGGGAAAACCTTACCACTACCAGCCCC CGTCATATGCC	212557	116	59
HSF2_R5_13	CCCTGACTGTGAAATGAATTATACTCGGATA AGACACGGATACCTAGAATAAGGGATGTTG ATCCAACTCC	176490	162	50
HSF2_R5_14	CCCTCCTCACCTTGCGTACGAAGATCCGCA GGACAAGGAAATCATCCTGAGCAACGATAG GTTAGCTCCC	146975	41	98
HSF2_R5_15	TCTAAACGGACTAGTGAGAGATACTAGCGC ATCGAACCGGACAACTTTAGTATAGTTAGA GCGTAGGCCA	139564	650	22
HSF2_R5_16	CCGGACCGAACCCGAAATTGTAACTATGCA TAGTGTGATTATCATACGGGGCATTCTAGA GGAAAGGGAA	123854	45	97
HSF2_R5_17	AACCCTAAACGCGGTACCCACGTAAGCCCC CGGATAGAGAGAACAGTGGCCGGGGAATT CACCTCAAGCC	121634	108	61
HSF2_R5_18	GCCGCGTACGTAGACGTACATCACGCAAAA GCGAGGGAAACCGATCCAAGTGAACCCGA ACCCTACGTAG	117221	10	189
HSF2_R5_19	CCTAGGTCCATACTACAATAGCTCGATCTT GTGCCAGTGAAGGGGTACATCCGGGGACA GGGGGAAAGGGG	87884	1	242
HSF2_R5_20	ACAGTAGCAAGACACCTGGAGACCTCCCTA ACCTGCAAATGGTGAGAGCAGGAACCATCA GATACCATAC	86977	11	161

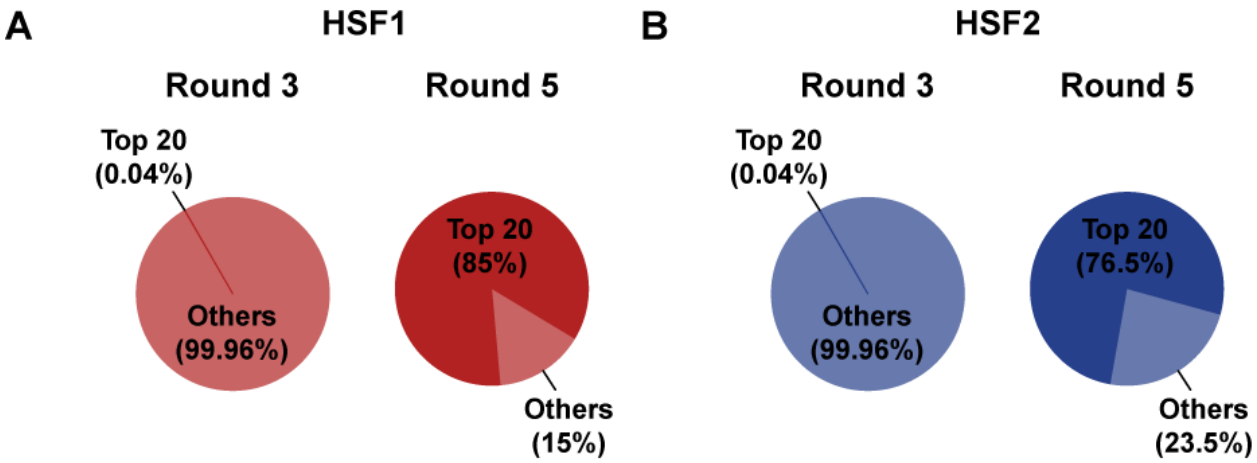


Figure 3.4: Top 20 highest multiplicity sequences in rounds 3 and 5 of HSF1 and HSF2 selections. The fraction of the total number of sequencing reads in rounds 3 and 5 that is represented by the top 20 sequences with highest multiplicities for the HSF1 **(A)** and HSF2 **(B)** selections.

The trend toward higher multiplicity values from round 3 to round 5 can also be observed when analyzing the multiplicity distributions for the top 3000 sequences within each pool. As seen in Figure 3.5, the round 5 histograms for both HSF1 and HSF2 are shifted toward higher multiplicity values when compared to the round 3 histograms and there is an evident decrease in the number of sequences with low multiplicity values. For example, the bin with the highest number of sequences in the HSF1 round 3 histogram contains the sequences with the lowest multiplicity values (multiplicity values higher than 1 and lower than 2.08). Although in the round 5 histogram this same bin contains the highest number of sequences, the total number is considerably reduced from round 3 to round 5 (Figure 3.5A, reduced from 2057 to 289). This same trend is observed for the HSF2 histograms. Taken together, these results indicate that 5 rounds of SELEX for HSF1 and HSF2 were sufficient for markedly enriching the pools with sequences with high multiplicity values from a starting RNA library that contained only 5 copies of each

sequence. These high multiplicity sequences likely correspond to aptamers that can bind to the target proteins with high affinity.

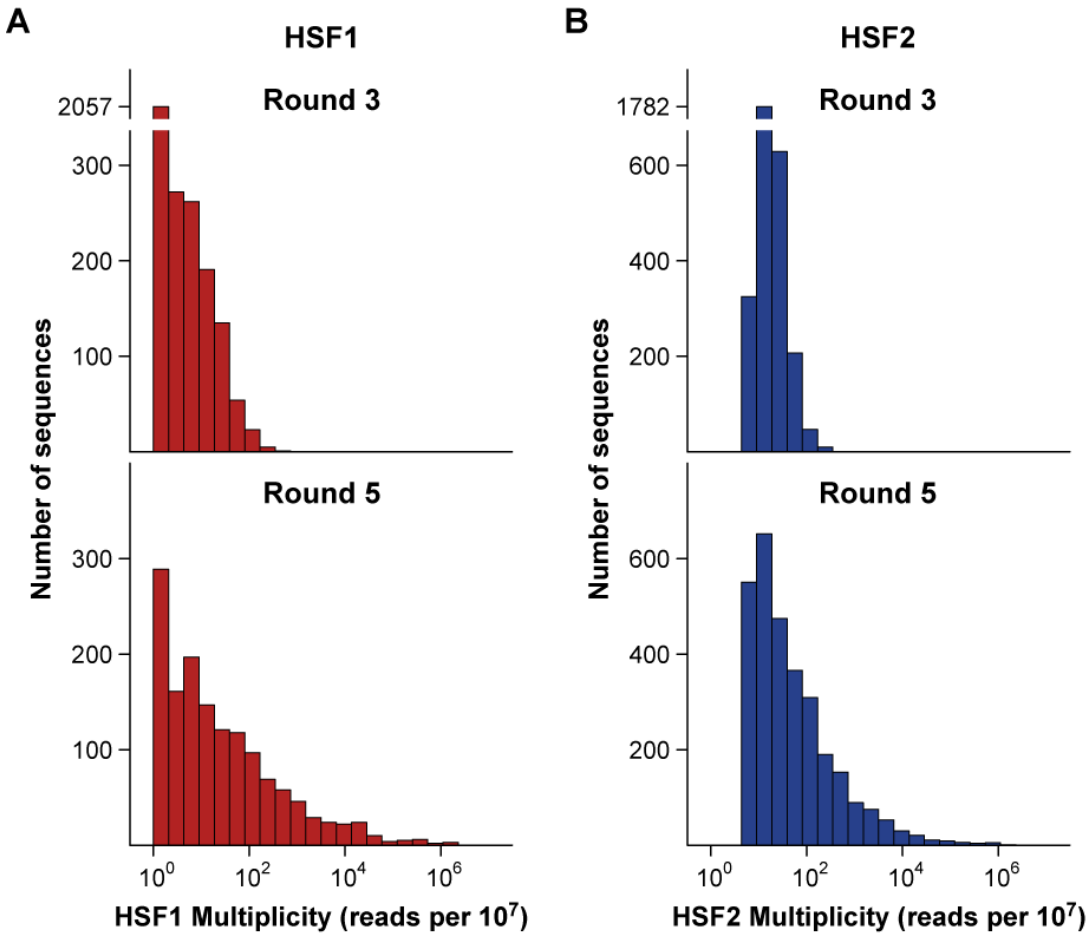


Figure 3.5: Multiplicity distributions for rounds 3 and 5 of HSF1 and HSF2 selections. Multiplicity histograms for the top 3000 sequences in rounds 3 and 5 of the HSF1 **(A)** and HSF2 **(B)** selections. For the HSF1 round 5 histogram, only 1432 sequences were detected in the pool, therefore the total number of sequences in the histogram is 1452. The total number of sequences in all the other histograms is 3000. The multiplicity values are presented as multiplicity per 10 million reads (normalized by total number of sequencing reads).

To further investigate the evolution of the RNA sequences during the SELEX experiment, we compared the enrichment of each sequence from round 3 to round 5 with their multiplicities in round 5. As seen in Figure 3.6, there is a strong correlation between

round5/round3 enrichment and round 5 multiplicity for both HSF1 and HSF2 selections, indicating that sequences with higher multiplicities enrich faster and more efficiently during the process. Furthermore, the majority of sequences in both selections have enrichment values higher than 1. A poor correlation between these two metrics would indicate that factors other than target binding were playing a major role at enriching for sequences with high multiplicities. For example, PCR amplification biases could enrich for sequences with high multiplicities very early on in the process and generate sequences that have high multiplicities and low round5/round 3 enrichments. Therefore, good correlation between enrichment and multiplicity is a strong indication that the SELEX was successful in enriching for sequences with high multiplicities that likely bind to the target proteins with high affinities.

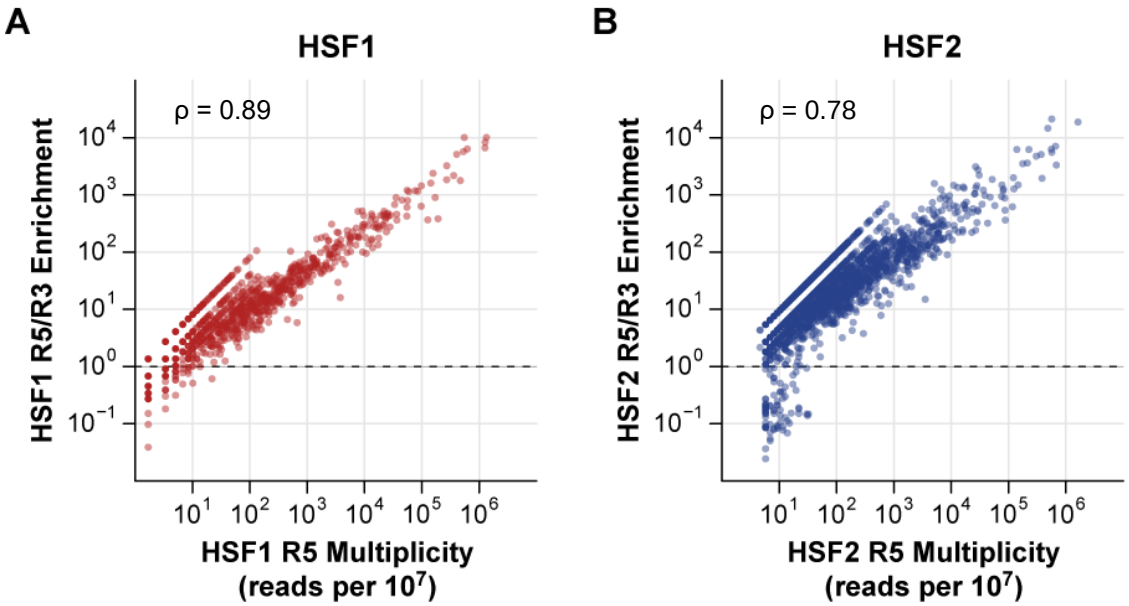


Figure 3.6: Relationship between enrichment and multiplicity values for HSF1 and HSF2 selections. Scatter plots of round 5/round 3 enrichment values versus round 5 multiplicity values for the top 3000 sequences in the HSF1 **(A)** and HSF2 **(B)** selections. The total number of sequences in the top 3000 of round 5 that were also present in the top 3000 of round 3 was 777 for HSF1 and 541 for HSF2. The multiplicity values are presented as multiplicity per 10 million reads (normalized by total number of sequencing reads). The Spearman's correlation coefficients are shown in the plot.

Enrichment and multiplicity metrics weakly correlate with binding affinities to target proteins

Based on the strong correlation between round5/round3 enrichment and round 5 multiplicity that was observed for both HSF1 and HSF2 selections (Figure 3.6), we hypothesized that the sequences with highest enrichment and multiplicity values were likely to bind the target proteins with the highest affinities. To further investigate the relationship between enrichment, multiplicity and binding affinity, we characterized the binding of a subset of sequences identified in the round 5 pool of the HSF1 SELEX. For this analysis, we chose the top 11 sequences with the highest multiplicity values in round 5 and seven other sequences that encompass the full range of enrichment and multiplicity values that were observed for the HSF1 selection (Table 3.6 and Figure 3.7). These full-length sequences were cloned and sequence verified as described in the Materials and Methods.

The 18 selected aptamers were fluorescently end-labeled and then tested for binding to HSF1 by Fluorescence Electrophoretic Mobility Shift Assay (F-EMSA) and Fluorescence Polarization (FP) assays (Pagano et al. 2007). The same equilibration reactions that were prepared for the FP measurements can be subsequently ran on a gel for the F-EMSA, which provides an efficient and fast way to perform two complimentary binding assays that rely on different physical properties of the aptamer/protein complexes (Pagano et al. 2011). The binding curves from the F-EMSA and FP assays for each of the 18 sequences are shown in Figures 3.8 and 3.9, respectively, and the estimated K_Ds from both experiments are shown in Table 3.6.

Table 3.6: Relationship between round5/round3 enrichment, round 5 multiplicity and binding affinity for the HSF1 SELEX. 18 characterized sequences from the round 5 pool of the HSF1 selection. Multiplicity and enrichment values were calculated as described in the Materials and Methods. Equilibrium dissociation constants (K_D) were estimated from both Fluorescence Electrophoretic Mobility Shift Assays (F-EMSA) and Fluorescence Polarization (FP) assays. Only the random region of each aptamer sequence is shown.

Aptamer ID	Sequence	Round 5 Multiplicity	Round5/ Round 3 Enrichment	K_D F-EMSA (nM)	K_D FP (nM)
HSF1_R5_1	CACTCCATAGTATCTAGAAGCCCTGCCGAAAACGAGAGCGCTTACGCTTATAAATGATCAACCTACCTCG	1340479	10113	27.3	37.4
HSF1_R5_2	CGGGCCTCACGCAGCACGCCCCTGCGCGTAGATAGACTGATCGGGACCCGCACCCACAGACGCCAAGCTC	1274151	8362	40.5	70.6
HSF1_R5_3	GCCACGGAAGACACCTTAGTATCCATACCCTCTCTAGTAAGAAGCCAGAGATTCGCAATCTCCTCCTGAC	1261696	6657	87.7	26.8
HSF1_R5_4	CGCCCTCTGGCTTCCAGGTGCTTCAGTAAATCAACGACTACCGGCTTACCTGACACCCAGCGATCAGGCG	621764	6353	33.1	22.7
HSF1_R5_5	CCGGCCTTTCCGCTTCCACGTTCCTGAGCTGAAATAAAAACGGCCGACGCCAGCACGCTGCCCTGGCGAT	550234	10095	44.1	58.4
HSF1_R5_6	ATGCCAAACCCGGACTATAGTGATACGGACGGAGAGGTGGTTCCACTTTTTCCCCTTGCATAGGCAATG	530607	5711	54.2	96
HSF1_R5_7	GCGGGCGAAACCGGACTAGCCTCCCGGCGTAAGAGTTCGAATCCATCCACCATATA	469695	1780	37.8	125.3
HSF1_R5_8	CCTGCTCGACCCGAATGTCCACGTAGTCGAGAGGTACTGCTACAGACTAAGATAAGAAGTACTTCTTAGT	405537	5115	77.1	19.1
HSF1_R5_9	CCTCGCTTACTACAGCAGCTAGCCCACCCGGAACTTGGGGCGACAGTGATCTAAAGTGACCGGCACACAT	358828	2178	37.3	42.1
HSF1_R5_10	CCCGTCCTAAAGCACCCTCGACTATGGAAAGAGAGGACAAAGGGGAACCATGAAACACAGATGAGGGCA	272611	3236	104.6	42.5
HSF1_R5_11	CCGGCTCCTCTTAGTCGCAAATAGTCTGGCGTGATTAAGGCACCCGAACCCGACGTAGTGCCCGCAAGGGA	271480	1842	170.3	55.6
HSF1_R5_19	CCCAACTACCCACGAAACCAATTGGGTTGCAACAAGACCATAAACATGCCGCCCTCAGTGAAGCGATA	84796	1160	65.2	41.3
HSF1_R5_31	CACCACAAGCCTGACGCTTGGGAAACTTACGTAGTGCGTGTCCACGAAAGCGCCGAACCCGACGTACGCT	28415	441	36.9	34.5

Table 3.6 (Continued)

Aptamer ID	Sequence	Round 5 Multiplicity	Round5/ Round 3 Enrichment	K_D F-EMSA (nM)	K_D FP (nM)
HSF1_R5_62	CCCAGTGCACGAGGCAC CACAAAGACAAATCATAC CCCTATAAAAGGGCCCA TATGGGATACACTTAGCC	10313	181	44.7	37.2
HSF1_R5_97	CCACGCTGCTGTGTTCC GATCCATGTGAAAAGAG TGGCATACCCGGCAGGC CCCGGGTAACATAGTTC AG	3379	83	53.6	58.2
HSF1_R5_154	GATTCGAAATCTAGAGCA ACTCAGGATAGACCGAA CACACTATACACTAGACC TGTACAAGTAGAATCGA	1013	34	139.4	45.7
HSF1_R5_368	CGCGGCCAAGTCATTCT GATTGCCCAGTGACTCC TGTGTCCGATTCTTGGG GAACTACTAGCAT	105	6	189.2	-
HSF1_R5_560	CGGTCCCCCTCCCCCCT CCGCACCACCGCAAAAG ATCCCAGATGTTATGCG CCGGGAACGCTTAGGAG TC	30	2	61.8	81.9

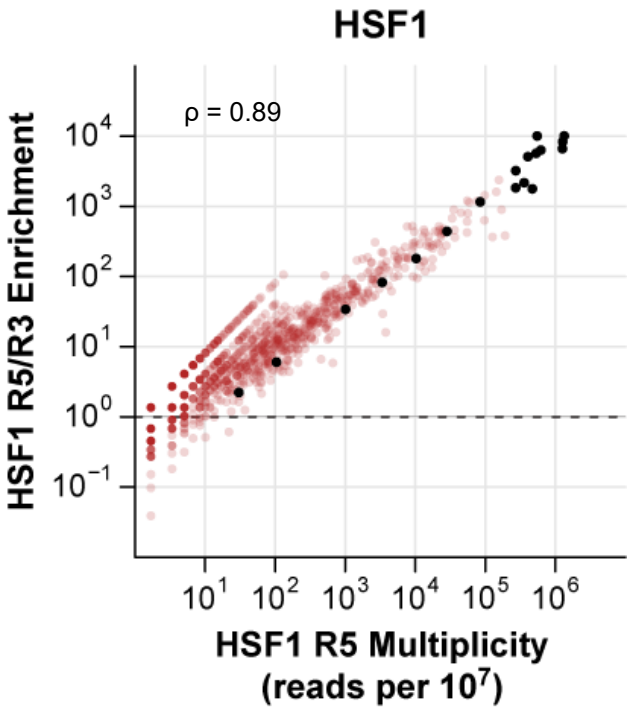


Figure 3.7: 18 characterized sequences from the round 5 pool of the HSF1 selection. Scatter plot of round 5/round 3 enrichment versus round 5 multiplicity for the top 3000 sequences in the HSF1 selection from Figure 3.6A with the 18 characterized sequences highlighted in black. The Spearman's correlation coefficient is shown in the plot.

Figure 3.8: Evaluation of 18 selected sequences binding to HSF1 using F-EMSA. Binding curves measured by F-EMSA for 18 selected sequences binding to a two-thirds dilution series (from 2000 nM to 0.2 nM) of HSF1 protein. The solid lines are the best fits of the Hill equation to the experimental data for each sequence with the corresponding K_D values given in the figure legends.

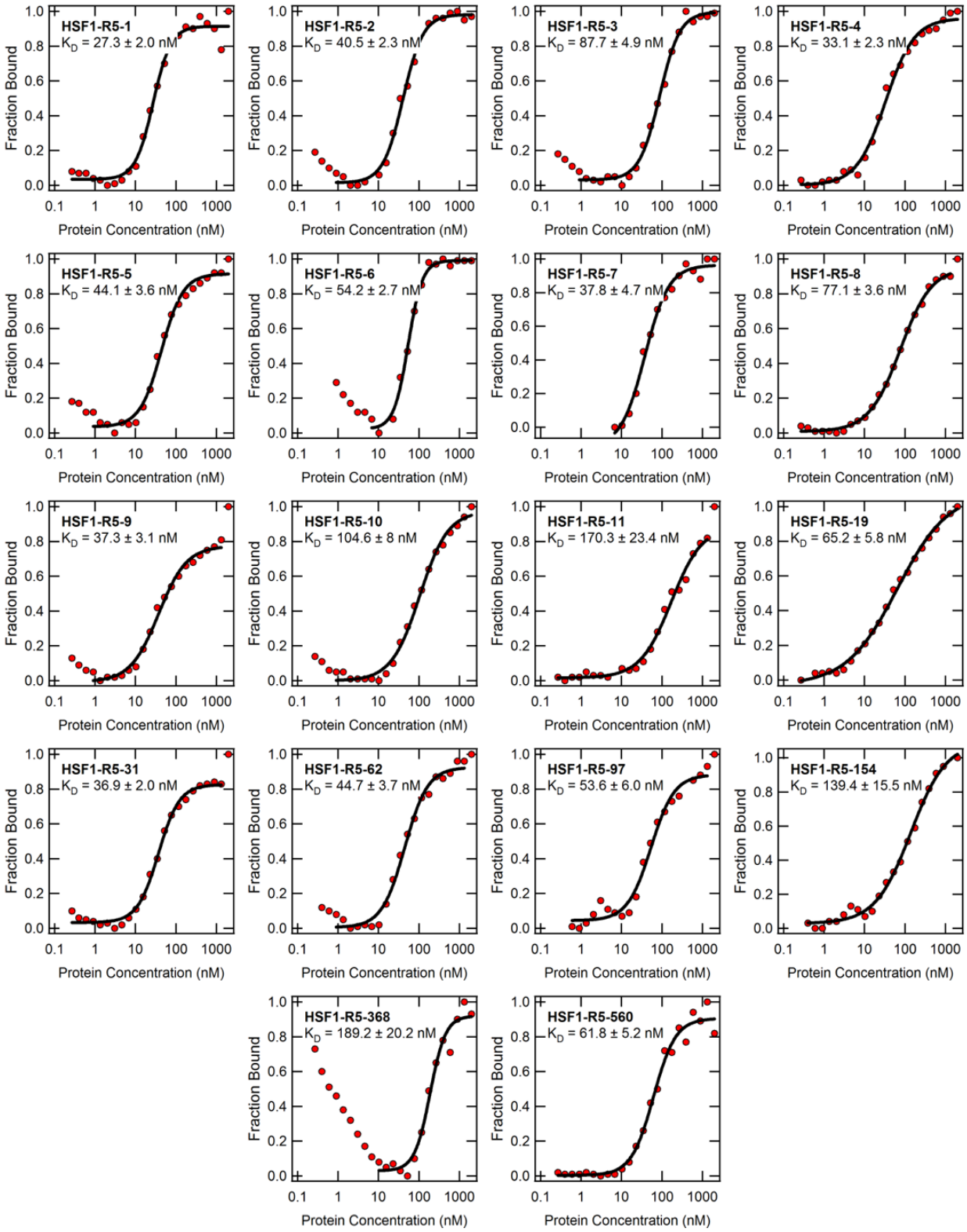
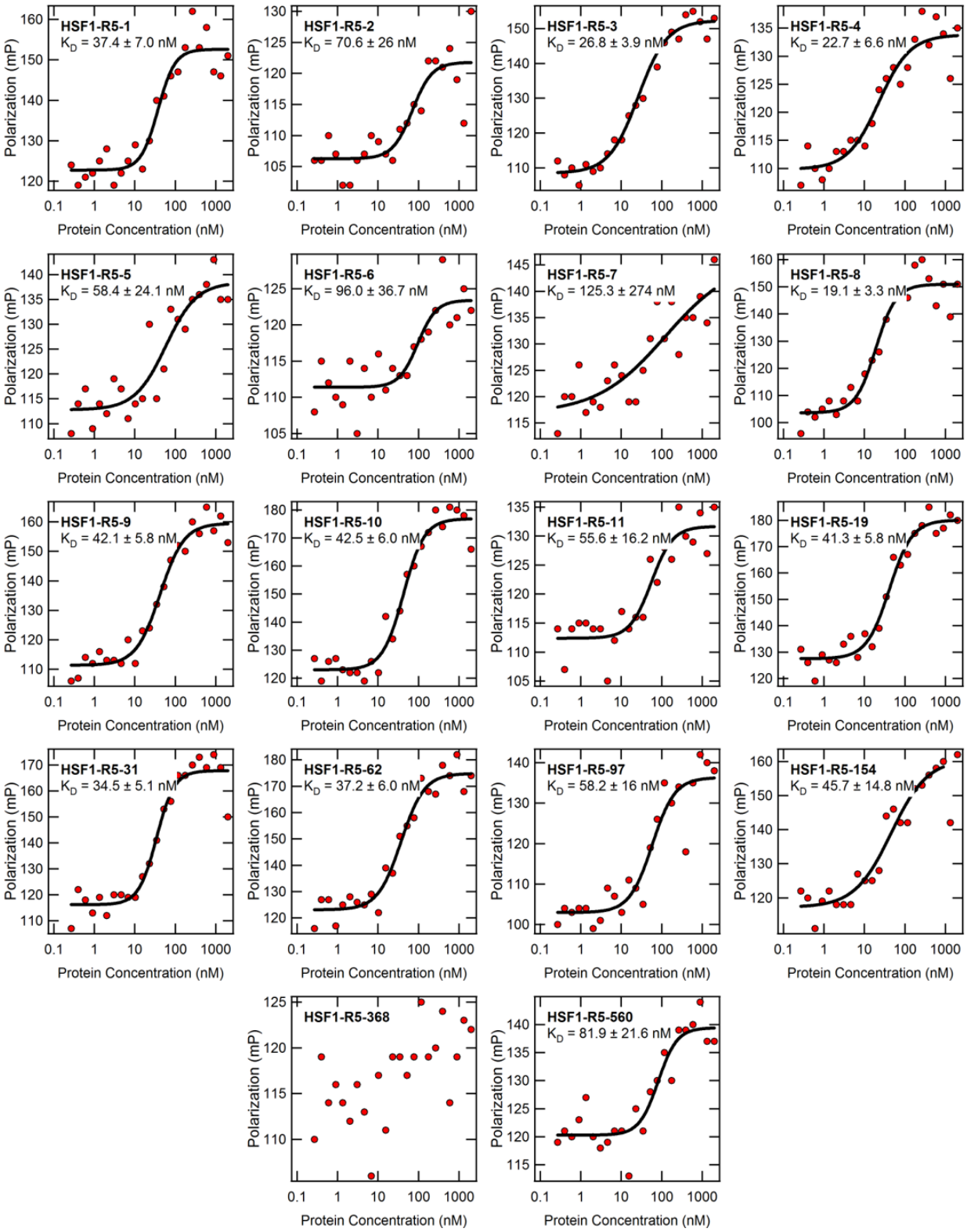

HSF1-R5-1
K_D = 27.3 ± 2.0 nM
HSF1-R5-2
K_D = 40.5 ± 2.3 nM
HSF1-R5-3
K_D = 87.7 ± 4.9 nM
HSF1-R5-4
K_D = 33.1 ± 2.3 nM
HSF1-R5-5
K_D = 44.1 ± 3.6 nM
HSF1-R5-6
K_D = 54.2 ± 2.7 nM
HSF1-R5-7
K_D = 37.8 ± 4.7 nM
HSF1-R5-8
K_D = 77.1 ± 3.6 nM
HSF1-R5-9
K_D = 37.3 ± 3.1 nM
HSF1-R5-10
K_D = 104.6 ± 8 nM
HSF1-R5-11
K_D = 170.3 ± 23.4 nM
HSF1-R5-19
K_D = 65.2 ± 5.8 nM
HSF1-R5-31
K_D = 36.9 ± 2.0 nM
HSF1-R5-62
K_D = 44.7 ± 3.7 nM
HSF1-R5-97
K_D = 53.6 ± 6.0 nM
HSF1-R5-154
K_D = 139.4 ± 15.5 nM
HSF1-R5-368
K_D = 189.2 ± 20.2 nM
HSF1-R5-560
K_D = 61.8 ± 5.2 nM
Fraction Bound
Protein Concentration (nM)
0.0
0.2
0.4
0.6
0.8
1.0
0.1
1
10
100
1000

Figure 3.9: Evaluation of 18 selected sequences binding to HSF1 using FP. Binding curves measured by FP for 18 selected sequences binding to a two-thirds dilution series (from 2000 nM to 0.2 nM) of HSF1 protein. The solid lines are the best fits of the Hill equation to the experimental data for each sequence with the corresponding K_D values given in the figure legends.


HSF1-R5-1
K_D = 37.4 ± 7.0 nM
HSF1-R5-2
K_D = 70.6 ± 26 nM
HSF1-R5-3
K_D = 26.8 ± 3.9 nM
HSF1-R5-4
K_D = 22.7 ± 6.6 nM
HSF1-R5-5
K_D = 58.4 ± 24.1 nM
HSF1-R5-6
K_D = 96.0 ± 36.7 nM
HSF1-R5-7
K_D = 125.3 ± 274 nM
HSF1-R5-8
K_D = 19.1 ± 3.3 nM
HSF1-R5-9
K_D = 42.1 ± 5.8 nM
HSF1-R5-10
K_D = 42.5 ± 6.0 nM
HSF1-R5-11
K_D = 55.6 ± 16.2 nM
HSF1-R5-19
K_D = 41.3 ± 5.8 nM
HSF1-R5-31
K_D = 34.5 ± 5.1 nM
HSF1-R5-62
K_D = 37.2 ± 6.0 nM
HSF1-R5-97
K_D = 58.2 ± 16 nM
HSF1-R5-154
K_D = 45.7 ± 14.8 nM
HSF1-R5-368
HSF1-R5-560
K_D = 81.9 ± 21.6 nM
Polarization (mP)
Protein Concentration (nM)

As seen in Figure 3.10, there is a weak correlation between binding affinity – as measured by the K_D values – and the enrichment and multiplicity metrics. The Spearman's rank correlation coefficients of -0.49 and -0.41 indicate that there is a trend toward lower K_D values (higher affinity) with increasing multiplicity or enrichment values. However, this correlation is weak, and the aptamer with the lowest multiplicity and enrichment values among the ones we tested binds to HSF1 with relatively high affinity (HSF1-R5-560, Table 3.6). Conversely, HSF1-R5-11, which is among the top 11 sequences with the highest multiplicity values, binds to HSF1 with relatively low affinity. Furthermore, all 18 tested sequences bind to HSF1 with K_Ds lower than 190 nM, which suggests that most sequences in the final pool of the HSF1 SELEX are able to bind to the target protein, irrespective of each sequence's copy number and of how efficiently it enriched during the SELEX process. These results indicate that binding affinity alone is not the main factor governing the enrichment of a particular sequence during the SELEX process. Although we did not observe a strong correlation between binding affinity and the enrichment and multiplicity metrics, our analyses demonstrate that the SELEX process efficiently selected for sequences that are able to bind to the target and suggest that a tight correlation between multiplicity and enrichment can serve as an indicator of the success of a selection.

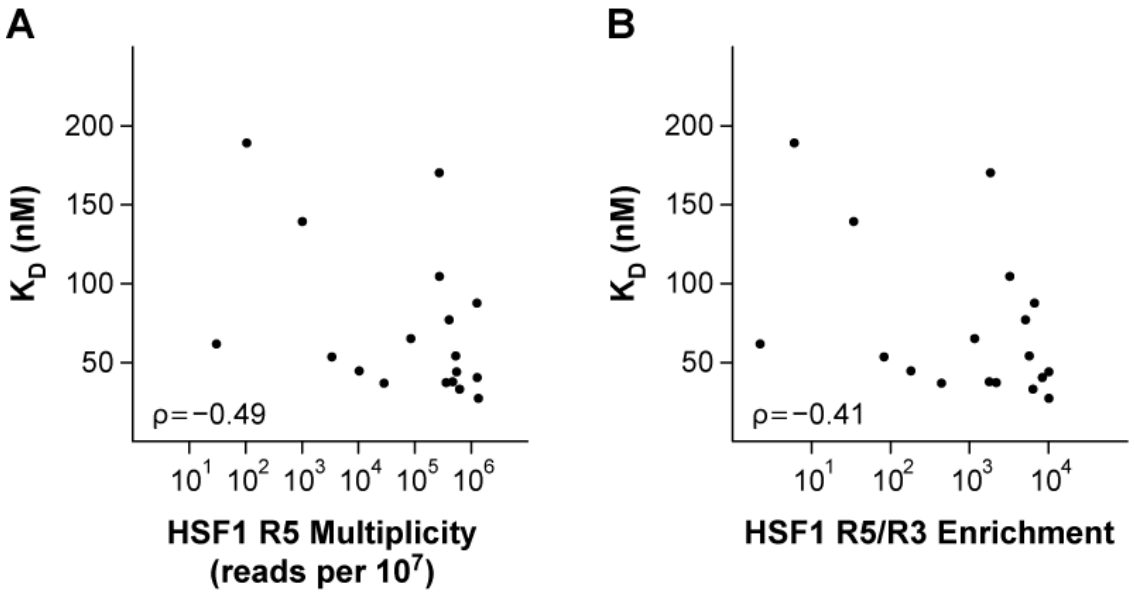


Figure 3.10: Relationship between multiplicity, enrichment and binding affinity. Scatter plots of estimated K_Ds from F-EMSA versus round 5 multiplicity **(A)** and round 5/round 3 enrichment **(B)** for the 18 selected sequences from the round 5 pool of the HSF1 selection. The Spearman's correlation coefficients are shown in the plot.

Strong correlation between multiplicity and enrichment metrics can function as an indicator of the success of a selection

According to the results described in the previous sections, the SELEX process for HSF1 was successful in selecting for sequences that bind to the target protein with high affinity. For this selection, the sequence multiplicity in the final pool (round 5) correlated well with round5/round3 enrichment, which suggested that a strong correlation between these two metrics could serve as an indicator of a successful selection. To further investigate if the correlation between multiplicity and enrichment is associated with selection success, we characterized the binding affinity of a subset of sequences from the other two selections for which we sequenced both round 3 and round 5 pools: HSF1-TD-AD and HSF2. Similarly to the HSF1 SELEX, the multiplicity and enrichment metrics for the HSF2

selection were strongly correlated (Figure 3.6B); however, the same trend was not observed for the HSF1-TD-AD selection, where these two metrics were poorly correlated (Figure 3.11B). Therefore, we hypothesized that the HSF2 selection was more likely to have yielded aptamers that bind to the target protein with high affinity when compared to the HSF1-TD-AD selection.

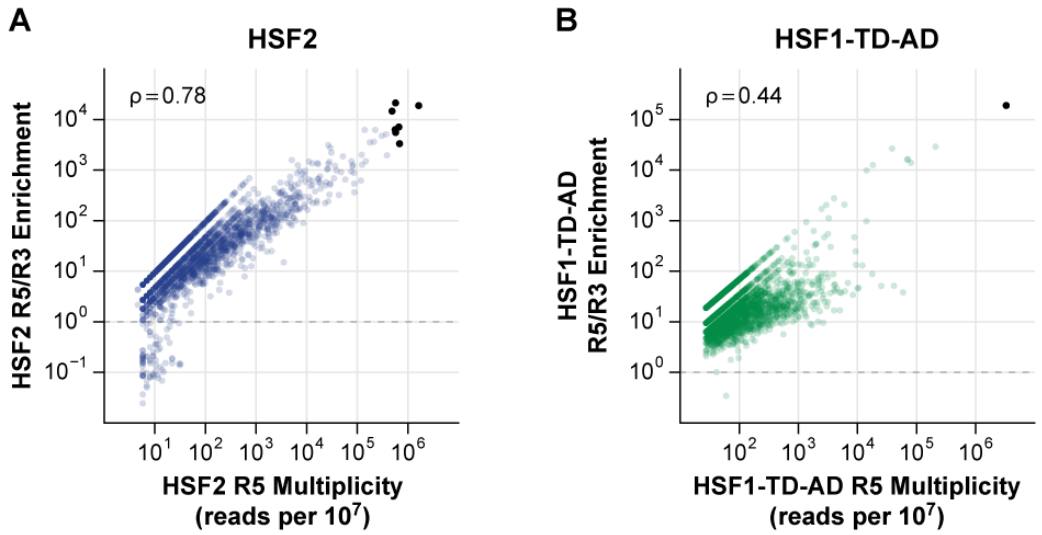
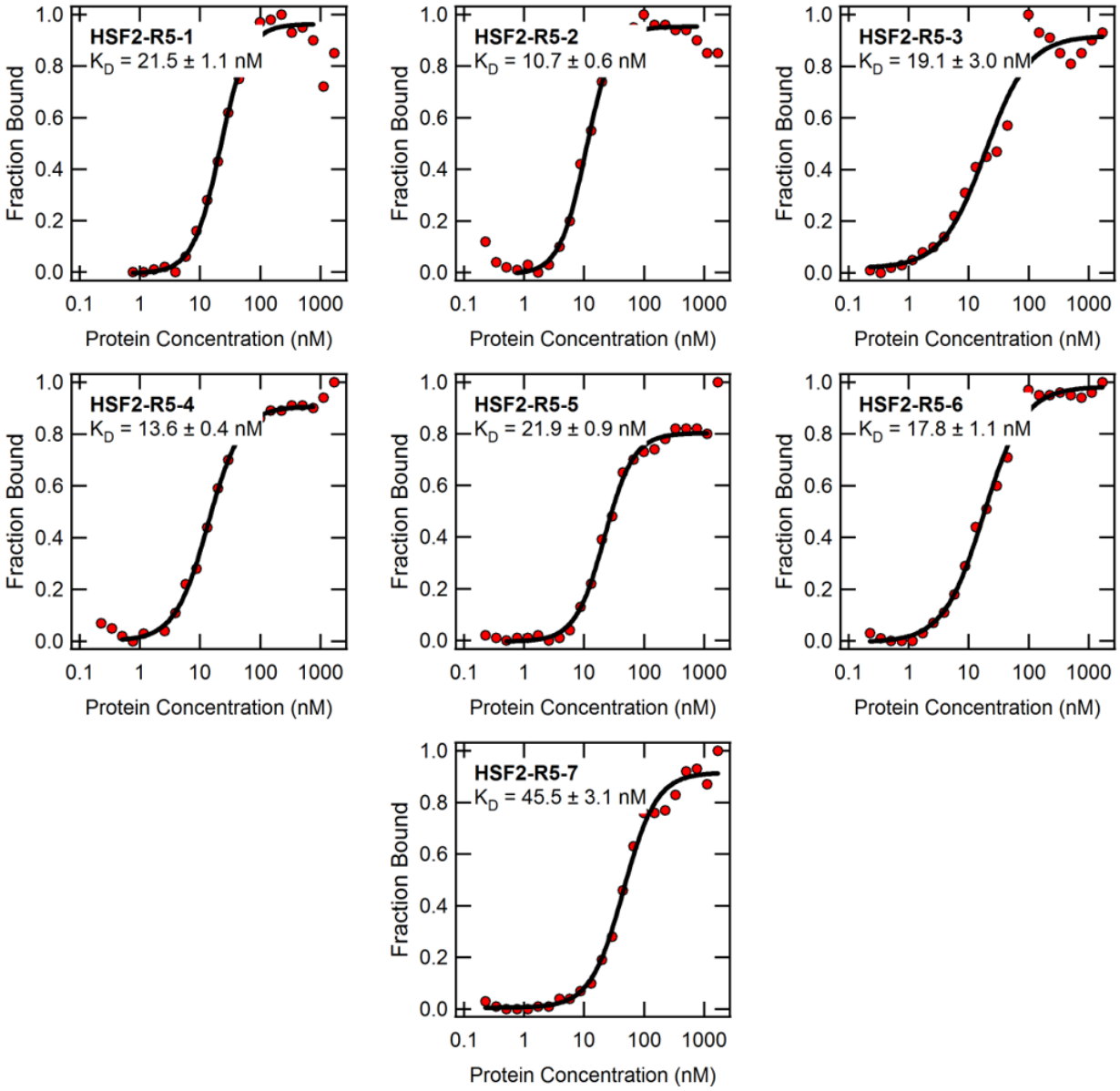


Figure 3.11: Characterized sequences from the round 5 pools of the HSF2 and HSF1-TD-AD selections. Scatter plots of round 5/round 3 enrichment versus round 5 multiplicity for the top 3000 sequences in the HSF2 **(A)** and HSF1-TD-AD **(B)** selections. Panel **(A)** is the same as Figure 3.6B, with the 7 characterized sequences highlighted in black. In panel **(B)**, one of the characterized sequences from the round 5 pool of the HSF1-TD-AD selection is labeled in black (HSF1-TD-AD-R5-1). However, the other two tested sequences – HSF1-TD-AD-R5-10 and HSF1-TD-AD-R5-22 – are not shown in the plot because their multiplicity value in round 3 was 0. The Spearman's correlation coefficient is shown in the plot.

To test our hypothesis, we used F-EMSA to measure the binding of seven sequences from the round 5 pool of the HSF2 selection and three sequences from the round 5 pool of the HSF1-TD-AD selection (Figure 3.11). As seen in Figure 3.12 and Table 3.7, all tested sequences from the HSF2 selection bind to HSF2 with high affinity.

However, no shift was observed for all three tested sequences from the HSF1-TD-AD selection, indicating that these sequences cannot bind to the target protein (Figure 3.13). This includes HSF1-TD-AD-R5-1, which alone represented 33% of the round 5 sequenced pool, HSF1-TD-AD-R5-10 and HSF1-TD-AD-R5-22 (Table 3.7). This result indicates that the multiplicity metric alone cannot determine if a sequence will bind to the target, since sequences with extremely high copy number in the final pool did not show any shift in the F-EMSA (Figure 3.13). The multiplicity versus enrichment scatter plot for HSF1-TD-AD shows that the majority of sequences in the round 5 pool have low enrichment values (Figure 3.11B). Together, these results strongly suggest that a tight correlation between enrichment and multiplicity is an indication that a particular SELEX was successful in yielding sequences that can bind to the target protein with high affinity.



<u>Figure 3.12:</u> Evaluation of seven selected sequences binding to HSF2 using F-EMSA. Binding curves measured by F-EMSA for seven selected sequences binding to a two-thirds dilution series (from 2000 nM to 0.2 nM) of HSF2 protein. The solid lines are the best fits of the Hill equation to the experimental data for each sequence with the corresponding K_D values given in the figure legends.

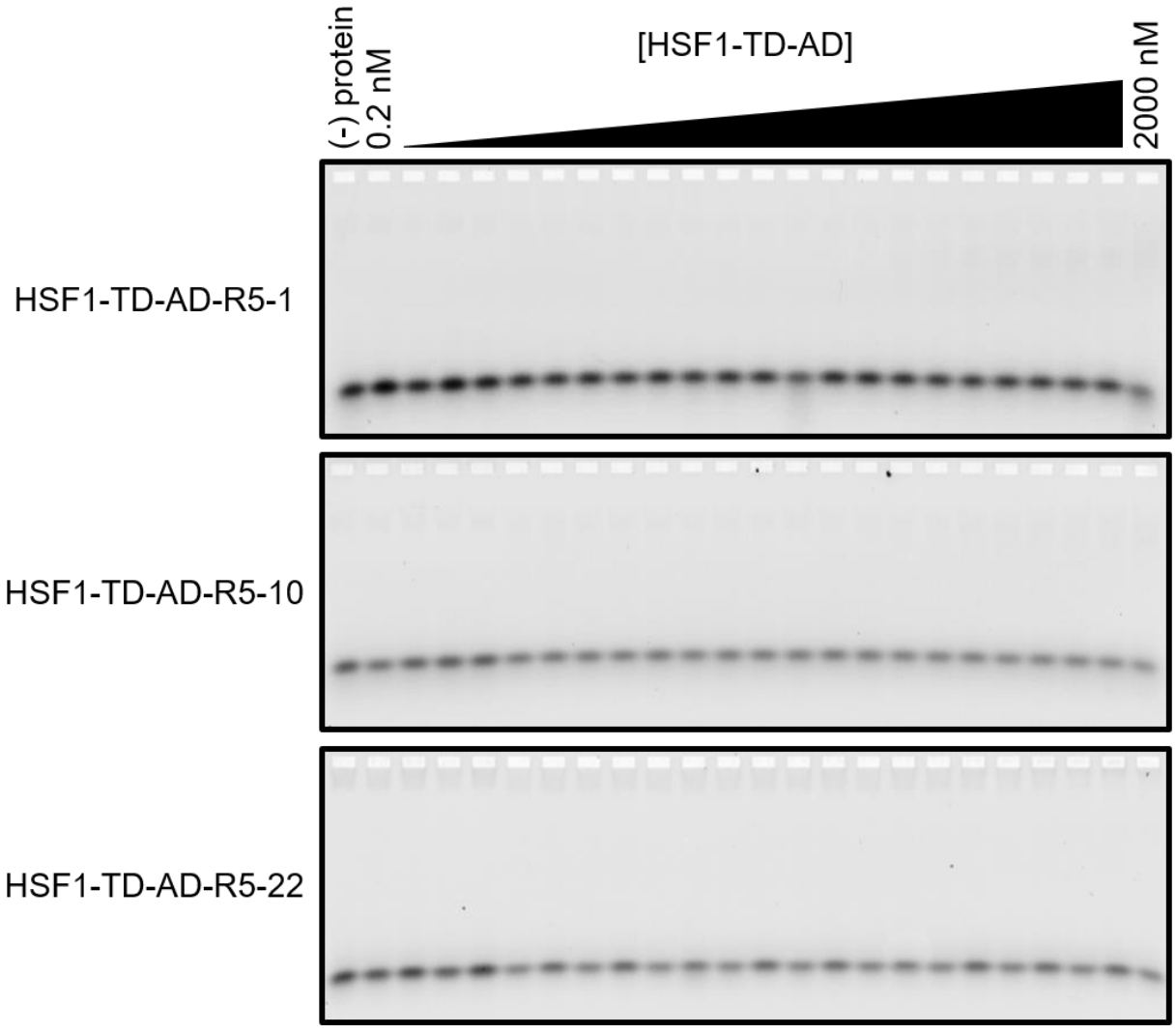


Figure 3.13: Evaluation of three selected sequences binding to HSF1-TD-AD using F-EMSA. F-EMSA results for HSF1-TD-AD-R5-1, HSF1-TD-AD-R5-10 and HSF1-TD-AD-R5-22 binding to a two-thirds dilution series (from 2000 nM to 0.2 nM) of HSF1-TD-AD protein.

Table 3.7: Characterized sequences from the round 5 pool of the HSF2 and HSF1-TD-AD selections. Multiplicity and enrichment values were calculated as described in the Materials and Methods. Equilibrium dissociation constants (K_D) were estimated from Fluorescence Electrophoretic Mobility Shift Assays (F-EMSA). Only the random region of each aptamer sequence is shown.

Aptamer ID	Sequence	Round 5 Multiplicity	Round5/ Round 3 Enrichment	K_D F-EMSA (nM)
HSF2_R5_1	GGAAAGGCGTACAAGA CGTACCTGGACCCAGA CCAAAGACACCTGTAGA TACACGCGCCCGATTC GGCGA	1627399	18787	21.5
HSF2_R5_2	AATCAAGTCCCCAGACT CAGCAACACTGGACAG CGATATGCAGATAACCA AGACCAA	682671	3325	10.7
HSF2_R5_3	TCCTCACTAACGCCACG TAAGCACCCGGATGATA TGGCCAGGGTACACTG TAAAGCCGACTCGATTA ACGT	660708	7101	19.1
HSF2_R5_4	GCCCTGCAGACATATC CACGGCAGCCACTAGA AATATAGTACCCGGTAA GAGGTAGATGGCGGAT ATGG	569105	5543	13.6
HSF2_R5_5	AGAACTGGGCGGACAA TTAATATAACGGCACAC TGATAGGTGCTAGGCC GCCACGTCGACGACGA TCCAT	567516	21226	21.9
HSF2_R5_6	AACAAACCTGCGATGAC CAAGCACCGACGGACA TAGGAATCTAGTACCGG GTAAGGCCGTATATGC CG	558955	6297	17.8
HSF2_R5_7	GAGCGCAAGAACCTAC CAGGTAATCCAGAACTG CTTGCGAGCCAATAACC GAACAACGGTGTGGAT TAAG	485019	14630	45.5
HSF1_TD_AD_R5_1	ACCGTTGACCCGTCAA CCTATGCATCCCAAACC CATCGATCCAGTATCAG GCCGGCGATCCGACCA CTGG	3291408	189219	-
HSF1_TD_AD_R5_10	CTACCCGTTTGTGACAC TACCGATAGCGACCAG GTGCTGCGACCAGGCC GGCAATCCGTCCCGAA GCACCTTGG	19273	NA	-
HSF1_TD_AD_R5_22	TACAACCCGACATGTCA AGTAACGTTACTTCTCC CTAGGACCAGGCCGGC AATCCGACCCCTCCGA GTGG	7201	NA	-

Top candidate aptamer in HSF2 selection binds to the DNA binding domain of HSF2

We decided to comprehensively characterize the binding properties of one of the top candidate sequences for HSF1 and HSF2. We chose the highest ranked sequences from the round 5 pools that were also highly ranked in round 3. For HSF1, this was the first-ranked sequence in round 5, hereafter referred to as HSF1-R5-1, which was the 11th-ranked sequence in round 3 (Table 3.4). For HSF2, this was the second-ranked sequence in round 5, hereafter referred to as HSF2-R5-2, which was the sixth-ranked sequence in round 3 (Table 3.5). The mfold (Zuker 2003) predicted secondary structures of these two candidate aptamers are shown in Figure 3.14. HSF1-R5-1 and HSF2-R5-2 alone represented 13.4% and 6.8% of the corresponding round 5 pools, respectively.

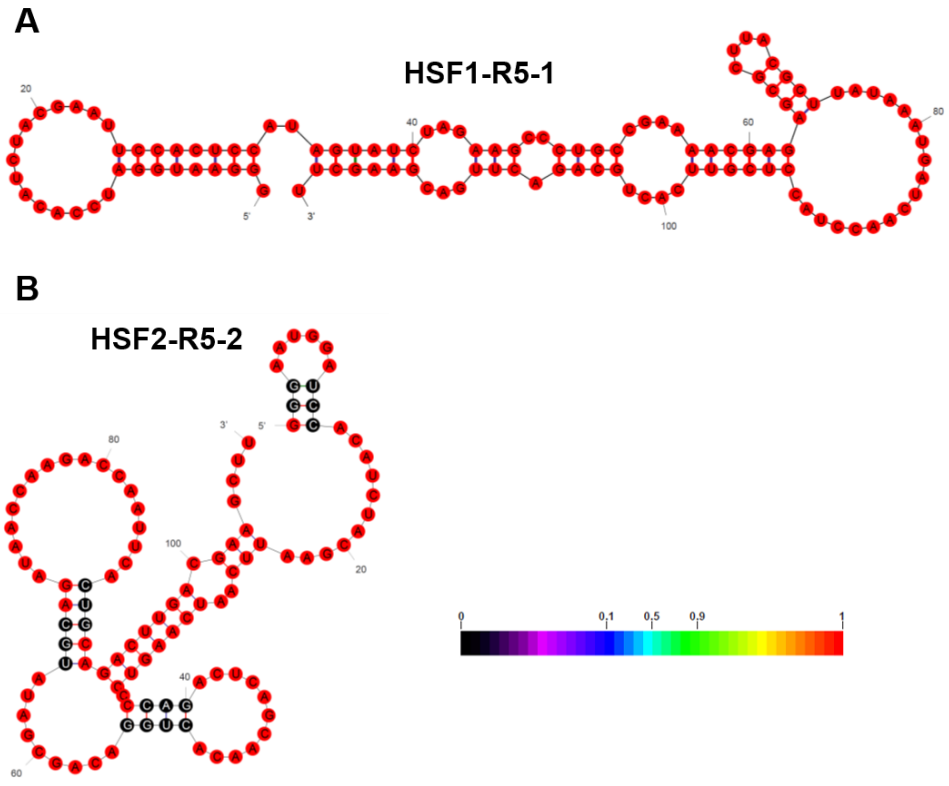


Figure 3.14: Predicted secondary structures of candidate aptamers. The mfold predicted secondary structures of HSF1-R5-1 **(A)** and HSF2-R5-2 **(B)** are shown. The structures are annotated using p-num with the color representing the probability according to the color key.

The candidate RNA aptamers were fluorescently end-labeled and then tested for binding to their HSF targets by F-EMSA and FP assays (Pagano et al. 2007). An image of a typical F-EMSA result is shown in Figure 3.15A for HSF1-R5-1 aptamer binding to HSF1 protein. The fraction of bound aptamer was calculated as a function of protein concentration and then plotted as shown in Figures 3.15B and 3.15C for various aptamer-protein pairings; K_D values were determined by fitting each data set to the Hill equation. HSF1-R5-1 and HSF2-R5-2 showed high-affinity binding to both HSF1 and HSF2 (K_D < 20 nM). Interestingly, both aptamers also bound to hexahistidine-tagged *Drosophila melanogaster* HSF (dHSF), although slightly more weakly ($K_D \sim 70$ nM), and no binding was observed to the GST-tag alone, which was used to purify the target proteins used in the SELEX. The F-EMSA results were confirmed by the FP assays (Figures 3.16 and 3.17). Thus, the observed binding is not due to the affinity tags on the protein targets but rather to specific domains of the targets themselves. Contrary to their functional similarity, these two aptamers did not show any similarity in primary sequence or in secondary structure, as predicted by mfold (Zuker 2003) (Figure 3.14). Given that the highest degree of sequence similarity between HSF1, HSF2, and dHSF is in the DNA binding and trimerization domains (DBD-TD) (Anckar and Sistonen 2007) and that the previously selected dHSF aptamer was found to bind the DBD-TD of dHSF (Zhao et al. 2006), we predicted these novel HSF aptamers were likely to bind the HSF proteins in a similar fashion.

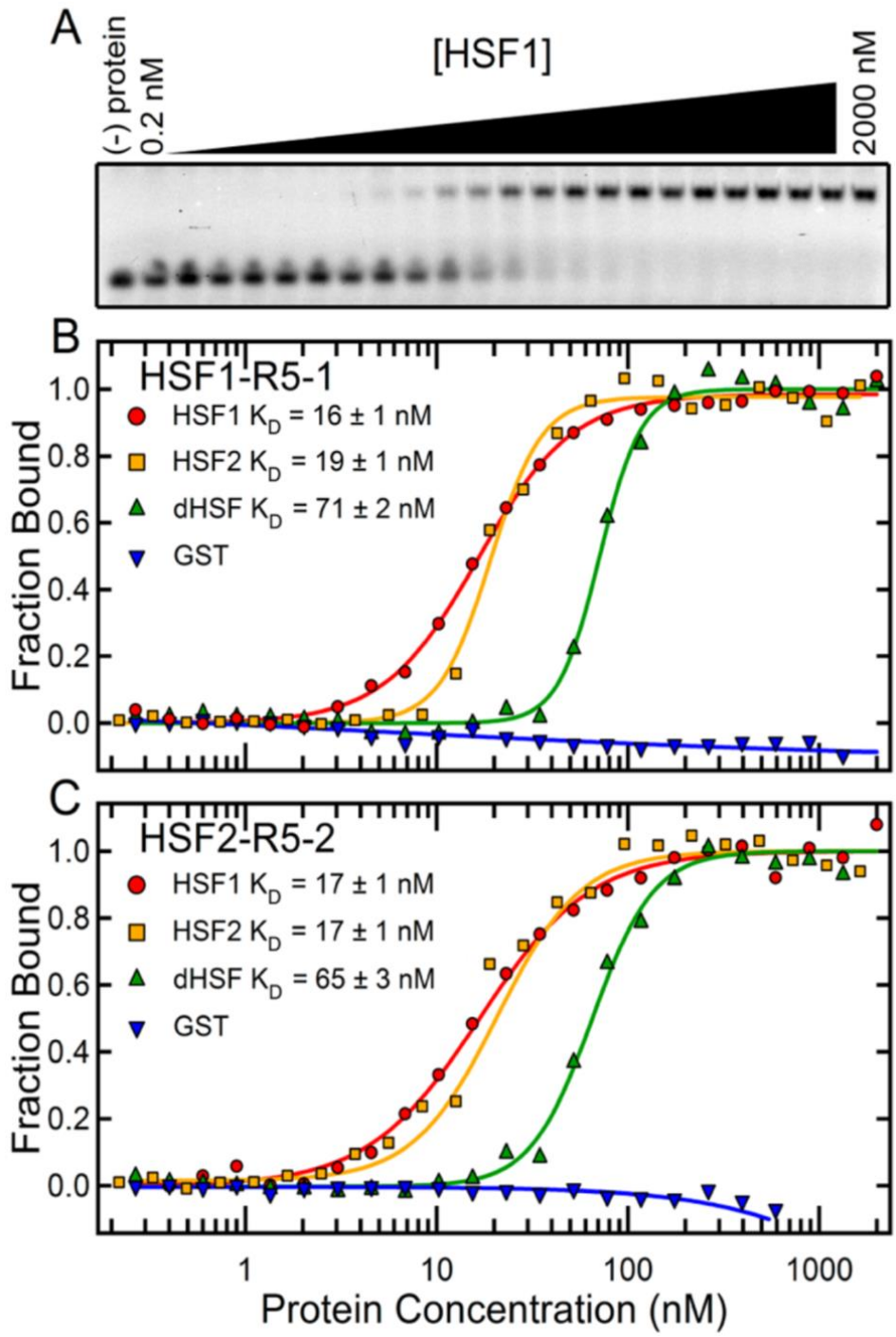
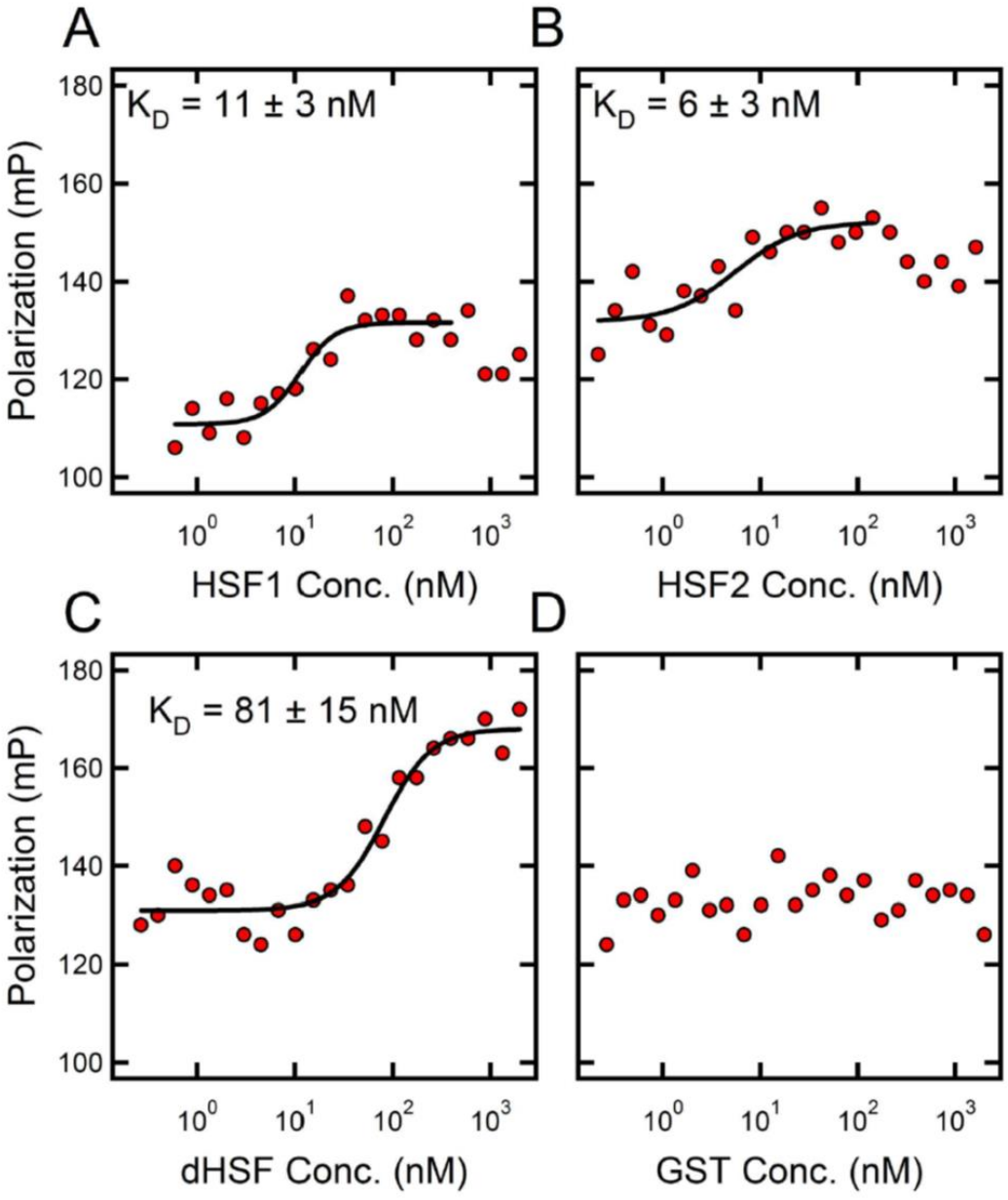
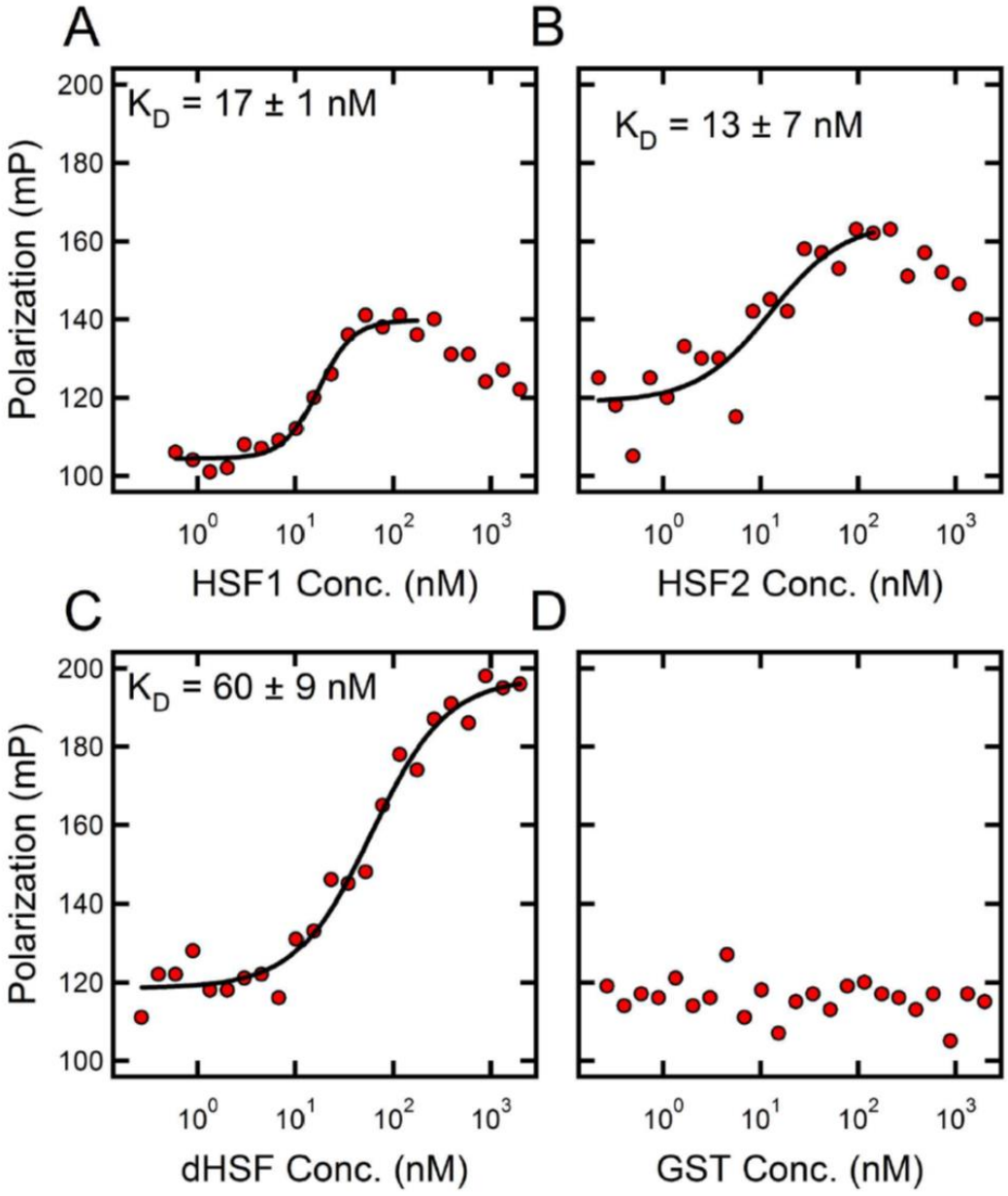


Figure 3.15: Evaluation of candidate aptamers binding to target proteins using F-EMSA. (A) Typical F-EMSA results for HSF1-R5-1 aptamer binding to a two-thirds dilution series (from 2000 nM to 0.2 nM) of HSF1 protein. **(B, C)** Binding curves measured by F-EMSA for HSF1-R5-1 and HSF2-R5-2 aptamers to HSF1, HSF2, dHSF, and GST tag. The same dilution series in panel A was used in panels B and C. The solid lines are the best fits of the Hill equation to the experimental data for each aptamer-target pair with the appropriate K_D values given in the figure legends.



<u>Figure 3.16:</u> Evaluation of HSF1-R5-1 binding to target proteins using Fluorescence polarization (FP). FP assay results for binding of HSF1-R5-1 aptamer to HSF1 **(A)**, HSF2 **(B)**, dHSF **(C)**, and GST tag **(D)**. The solid line in each panel (except D) is the best fit of the Hill equation to the experimental data and the corresponding K_D value is shown inside each plot.



<u>Figure 3.17:</u> Evaluation of HSF2-R5-2 binding to target proteins using Fluorescence polarization (FP). FP assay results for binding of HSF2-R5-2 aptamer to HSF1 **(A)**, HSF2 **(B)**, dHSF **(C)**, and GST tag **(D)**. The solid line in each panel (except D) is the best fit of the Hill equation to the experimental data and the corresponding K_D value is shown inside each plot.

In order to determine the specific regions of HSF1 and HSF2 that are bound by HSF1-R5-1 and HSF2-R5-2, we used F-EMSA to measure the binding of these two aptamers to truncated versions of the proteins that contain only the DNA binding domain (HSF1-DBD and HSF2-DBD) or that lack the DBD (HSF1-TD-AD and HSF2-TD-AD). As seen in Figure 3.18, HSF1-R5-1 did not show any binding to HSF1-DBD and HSF1-TD-AD, which suggests that this aptamer could potentially bind to the linker region between the DNA binding domain and the trimerization domain. Another possibility is that HSF1-R5-1 binds to the trimeric version of the HSF1 DNA binding domain. Full-length HSF produced from *E. coli* cells is present in solution in its oligomeric version (Clos et al. 1990; Rabindran et al. 1991); however, the HSF1-DBD truncation is unable to form trimers due to the lack of the trimerization domain. Therefore, it is possible that an aptamer that binds to the trimeric version of the DNA binding domain would bind to the full-length HSF1 but not to the HSF1-DBD truncation. Further experiments are necessary to map the specific region of HSF1 that is bound by HSF1-R5-1. Unlike HSF1-R5-1, HSF2-R5-2 binds to HSF2-DBD with slightly higher affinity than the full-length protein (Figure 3.19), which suggests that this aptamer binds to the DNA binding domain of HSF2. Given the high degree of sequence similarity between the DNA binding domain of HSF2 and HSF1 (Anckar and Sistonen 2007), this finding explains why HSF2-R5-2 is able to bind both HSF1 and HSF2 with similar affinities.

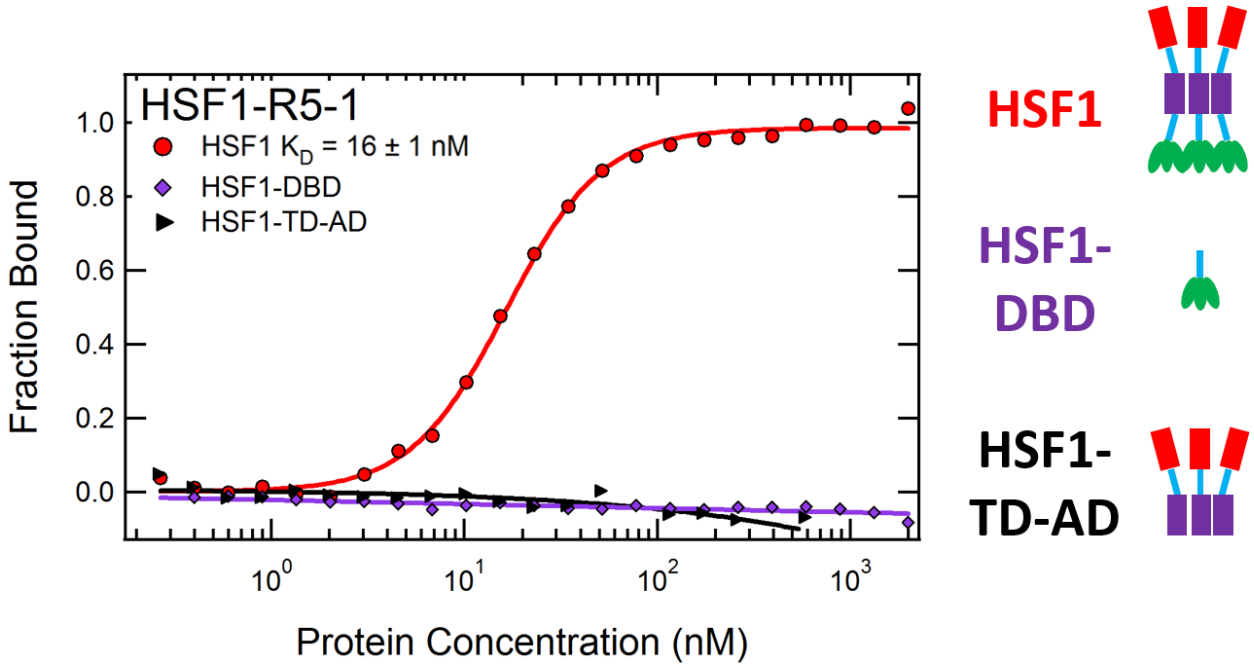
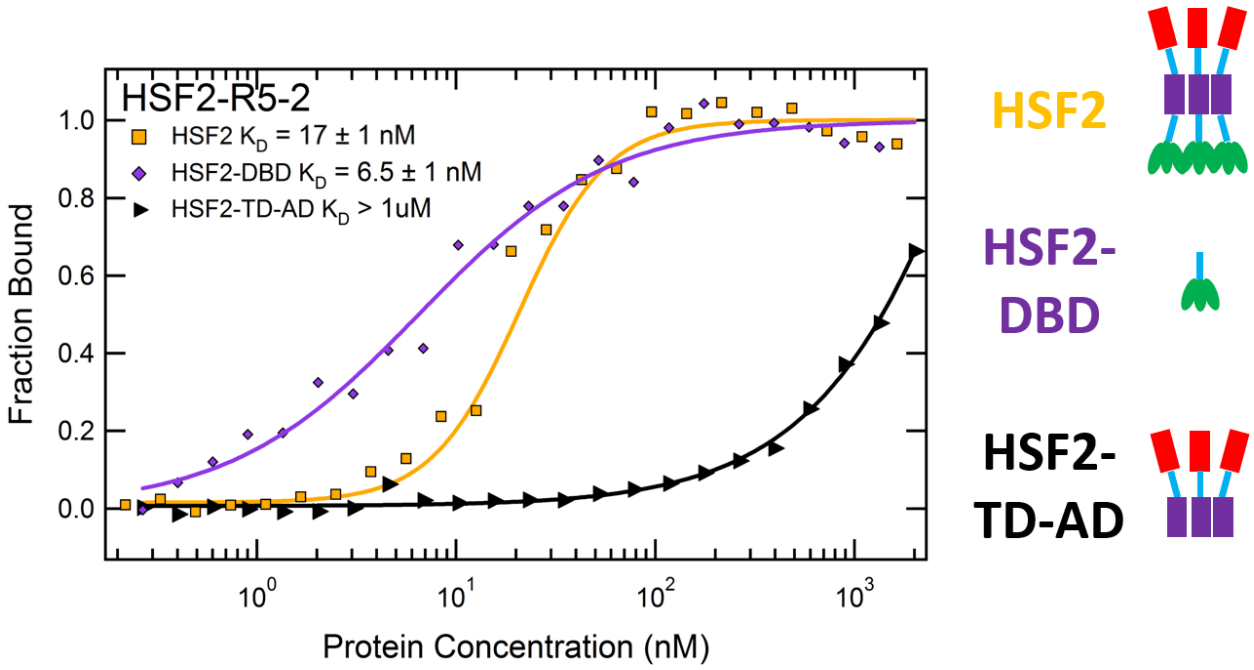


Figure 3.18: Evaluation of HSF1-R5-1 binding to HSF1 truncations using F-EMSA. Binding curves measured by F-EMSA for HSF1-R5-1 aptamer binding to a two-thirds dilution series (from 2000 nM to 0.2 nM) of HSF1, HSF1-DBD and HSF1-TD-AD. The solid red line is the best fit of the Hill equation to the experimental data for HSF1-R5-1 binding to HSF1 with the corresponding K_D value given in the figure legend. Cartoons representing the full-length protein and the truncations used in the experiment are shown on the right.



<u>Figure 3.19:</u> Evaluation of HSF2-R5-2 binding to HSF2 truncations using F-EMSA. Binding curves measured by F-EMSA for HSF2-R5-2 aptamer binding to a two-thirds dilution series (from 2000 nM to 0.2 nM) of HSF2, HSF2-DBD and HSF2-TD-AD. The solid lines are the best fits of the Hill equation to the experimental data for HSF2-R5-2 binding to the target proteins with the corresponding K_D values given in the figure legend. Cartoons representing the full-length protein and the truncations used in the experiment are shown on the right.

3.4 Discussion

We have described the results of a successful RNA aptamer selection for the human transcription factors HSF1 and HSF2. This experiment served as a proof-of-principle for our new RNA aptamer library and microcolumn-based SELEX technology, which were developed in collaboration with Harold Craighead's group (Latulippe et al. 2013; Szeto et al. 2014). The final enriched pools for both HSF1 and HSF2 contain thousands of potential binders, and we have identified dozens of aptamers within these pools that bind to HSF1 and HSF2 with high affinity (low nanomolar K_D). These aptamers will ultimately be

expressed in vivo in an induced manner and used as high specificity inhibitors to study the functions of these transcription factors.

To increase the likelihood of selecting aptamers that bind to distinct domains of HSF, we performed SELEX to the full-length HSF1 and HSF2 and to truncated versions of these proteins that contained only the DNA binding domain (HSF1-DBD and HSF2-DBD) or that lacked the DBD (HSF1-TD-AD and HSF2-TD-AD). The multiplex capability of our microcolumns allowed the concomitant RNA aptamer selection for all our targets in a highly efficient manner (Latulippe et al. 2013). Furthermore, the microcolumns' design enabled the addition of in-line negative selections against the GST tag to all our targets in all 5 rounds of SELEX. This negative selection appears to have been successful in decreasing the chance of enriching for aptamers that bind to the GST tag, since the candidate aptamers tested showed no shift to the tag in F-EMSA assays.

One of the major improvements in the SELEX methodology from previous experiments performed in our lab was the use of high-throughput sequencing to assess the selection results. Besides a significant reduction in the number of SELEX cycles that need to be performed, high-throughput sequencing of the selected pools allows a more comprehensive view of the types and numbers of sequences that are present in the final pools. Furthermore, by also sequencing an earlier pool (round 3), we were able to assess how each sequence enriched during the process.

Even though the use of high-throughput sequencing provides extensive information about the selected pools, one of the main challenges when performing SELEX is the identification of the best candidate aptamers for follow-up experiments. In our study, we have implemented a set of SELEX performance metrics to evaluate the success of a

selection based on the high-throughput sequencing results, and we have found that sequencing an earlier round was fundamental for our analyses. In a successful SELEX, three rounds of selection were sufficient to enrich for sequences with high multiplicity values from a starting pool that contained only 5 copies of each unique sequence (Figure 3.5). An additional two rounds of selection – for a total of 5 rounds – considerably enriched the round 3 pools, with the top 20 sequences representing ~80% of the final (round 5) pool for both HSF1 and HSF2 (Figure 3.4). These results indicate that our SELEX methodology was highly successful in enriching for sequences with high multiplicity values in the final pool.

Although the main goal of a SELEX experiment is to exponentially enrich for sequences that bind to the target (Ellington and Szostak 1990; Tuerk and Gold 1990), we have observed that the presence of sequences with high multiplicity values in the final pool is not a sufficient indicator of the success of a selection. For instance, the final pool of our HSF1-TD-AD SELEX contains multiple sequences with extremely high multiplicity values, with the top sequence alone representing 33% of the pool; however, none of the tested sequences are able to bind to the target (Figure 3.13). Although multiplicity alone cannot predict selection success, a strong correlation between each sequence's multiplicity value with its enrichment from round 3 to round 5 is highly associated with a successful SELEX. The HSF1 and HSF2 selections, where round 5 multiplicity and round5/round3 enrichment had a strong correlation (Figure 3.6), generated aptamers with very high affinity to the targets, while the HSF1-TD-AD SELEX, where this correlation was weak (Figure 3.11B), had a poor SELEX outcome. A tight correlation between these two metrics is associated with a SELEX process in which sequences gradually enriched from

round to round, and is therefore more likely to have selected for aptamers that bind to the target. For example, weak correlation between these two metrics due to the presence of multiple sequences with high multiplicity and low enrichment values could indicate that sequence evolution was governed by factors other than target-affinity, such as PCR biases. Indeed, Cho and colleagues observed a large population of sequences with high copy numbers but with minimal enrichment values in their SELEX, and attributed this finding to biases during library synthesis or PCR (Cho et al. 2010). We conclude that a tight correlation between enrichment and multiplicity is a strong indicator of the success of an aptamer selection.

Although in a successful selection there is a strong correlation between multiplicity and enrichment, these two metrics showed a very weak correlation with binding affinity (Figure 3.10). In our HSF1 SELEX, even sequences with some of the lowest enrichment and multiplicity values were still able to bind to the target, and all the tested sequences had K_D values below 190 nM, which suggested that most sequences in the final pool can bind to HSF1. This result indicated that binding affinity alone was not the main factor governing the enrichment of sequences during the selection process, and that our SELEX conditions were not only selecting for binding, but for additional properties of the aptamer-target interaction.

One possibility is that the enrichment of sequences in the final pool is correlated with the rate constants of the aptamer-protein interactions. For instance, slower or faster dissociation rate constants could potentially explain the differential enrichment and copy number of two aptamers with similar K_D values. In our SELEX, we used very slow RNA loading steps (1 µL/min flow rate, approximately 14 hours) and somewhat slow washing

steps (100 µL/min flow rate, approximately 30 minutes), which conceivably could have preferentially enriched for sequences with slower off rates. Further experiments are necessary to investigate the correlation between rate constants and aptamer enrichment and multiplicity. Furthermore, it will be extremely valuable to determine if our selection enriched for sequences with slower off rates, given that it can be a desirable property. Indeed, previous studies have specifically adjusted their SELEX conditions to enrich for aptamers with slow off rates (Geiger et al. 1996; Davis and Szostak 2002; Gold et al. 2010). Moreover, according to the “drug-target residence time model”, the off rate of a drug-target complex, and not the binding affinity per se, dictates much of the in vivo efficacy of the drug (Copeland 2016), which indicates that slow off rates can be extremely advantageous for aptamers that will be used as inhibitory drugs in vivo.

In sum, we have presented the results of a successful RNA aptamer selection to the human transcription factors HSF1 and HSF2. The identified high affinity aptamers will be expressed in vivo in an inducible manner and used to dissect the molecular mechanisms of these factors. Furthermore, our novel SELEX methodology and implemented set of analysis tools offer a significant improvement over traditional approaches and provides an efficient platform for the performance and analysis of SELEX experiments.

CHAPTER 4

CONCLUSIONS AND FUTURE DIRECTIONS

4.1 General Discussion

The overarching goal of this dissertation was to investigate the roles of transcription factors in the regulation of rate-limiting steps in the transcription cycle. We focused on the transcription factors GAF and HSF in the context of the transcriptional response to heat shock induction. However, the lessons learned here about the coordinated regulation of a transcriptional response by transcription factors acting at distinct steps can serve as a general model for our understanding of the role played by transcription factors in the response to environmental or developmental signals. Furthermore, we generated and comprehensively characterized a collection of high affinity RNA aptamers to HSF that will be used as powerful tools to further our understanding of the mechanistic roles of this transcription factor. While focusing on the selection of aptamers to HSF, we implemented new SELEX technologies and analysis tools that can be easily applied to the selection of aptamers to any protein.

Lessons learned from “observing, perturbing, re-observing” the transcriptional heat shock response in *Drosophila* S2 cells

To achieve our main goal, we employed the “observe, perturb, re-observe” strategy to investigate the roles of GAF and HSF in the transcriptional response to heat shock induction. For the first step (“observe”, Figure 1.4), we used PRO-seq to comprehensively characterize the direct changes in transcription that happen in *Drosophila* S2 cells after a

time-course of heat shock induction. Here, the use of a high spatial and temporal resolution method such as PRO-seq was essential for determining the identity, levels and speed of response of all the genes that were either activated or repressed by heat shock induction. The comprehensive characterization of a global transcriptional response such as heat shock provides valuable clues to the diverse mechanisms that are used to regulate transcription. We have learned that the HS activated class is not limited to the classical molecular chaperones and that promoter-proximal Pol II pausing is a prevalent feature at this class of genes prior to heat shock induction. We have also identified the genes whose transcription is directly repressed by heat shock induction and shown that this is a regulated response rather than an indiscriminate global shut-down of transcription as had been previously proposed. Furthermore, we demonstrated that HS-induced transcriptional repression is mediated at the level of transcription initiation.

For the second step ("perturb", Figure 1.4), we depleted the levels of GAF and HSF in *Drosophila* S2 cells using standard RNAi knockdown methods. We have also generated high affinity RNA aptamers to the human HSF1 and HSF2 to be used as powerful inhibitors of macromolecular interactions of these transcription factors. The potential application of these tools to study the primary functions of HSF1 and HSF2 and their mechanisms of action in vivo will be discussed in the next section (Section 4.2).

For the last step ("re-observe", Figure 1.4), we used PRO-seq to measure the direct changes in transcription upon heat shock after depleting the levels of GAF and HSF. Depletion of these transcription factors coupled with the high-resolution of the PRO-seq method elucidated the general and distinct roles of GAF and HSF in the regulation of the transcriptional heat shock response. GAF-mediated Pol II pausing prior to heat shock

is essential for the activation of a subset of GAF-bound genes, while HSF activates its target genes by promoting the release of paused Pol II into productive elongation. Interestingly, we have shown that HSF, which was considered the ‘master’ regulator of the heat shock response, is essential for the activation of only 20% of the HS activated genes. Furthermore, neither GAF nor HSF have a role in HS-induced transcriptional repression.

An invaluable advantage of using a model cell line such as *Drosophila* S2 is the availability of a wealth of chromatin factor ChIP-seq or ChIP-chip datasets from modENCODE and other sources. We took advantage of these datasets to show that the transcription factor M1BP and the insulator proteins BEAF-32 and Chromator are enriched in the promoter region of HS activated genes with GAF-independent pausing prior to heat shock. We used PRO-seq to measure the direct changes in transcription upon heat shock after M1BP depletion by RNAi knockdown and showed that similarly to GAF, M1BP-mediated pausing prior to heat shock is important for the activation of a subset of M1BP-bound genes.

Overall, we demonstrated the “observe, perturb, re-observe” strategy as a powerful approach to investigate the roles of protein factors in biological processes. The application of this strategy to the transcriptional heat shock response revealed the steps in the transcription cycle that are under regulation for distinct classes of genes, provided the statistical power to evaluate mechanisms of regulation of the transcription factors GAF and HSF and identified new players with important roles in the response. This thorough characterization also revealed important aspects of transcription regulation that require further investigation, which will be discussed in the next section.

4.2 Future Directions

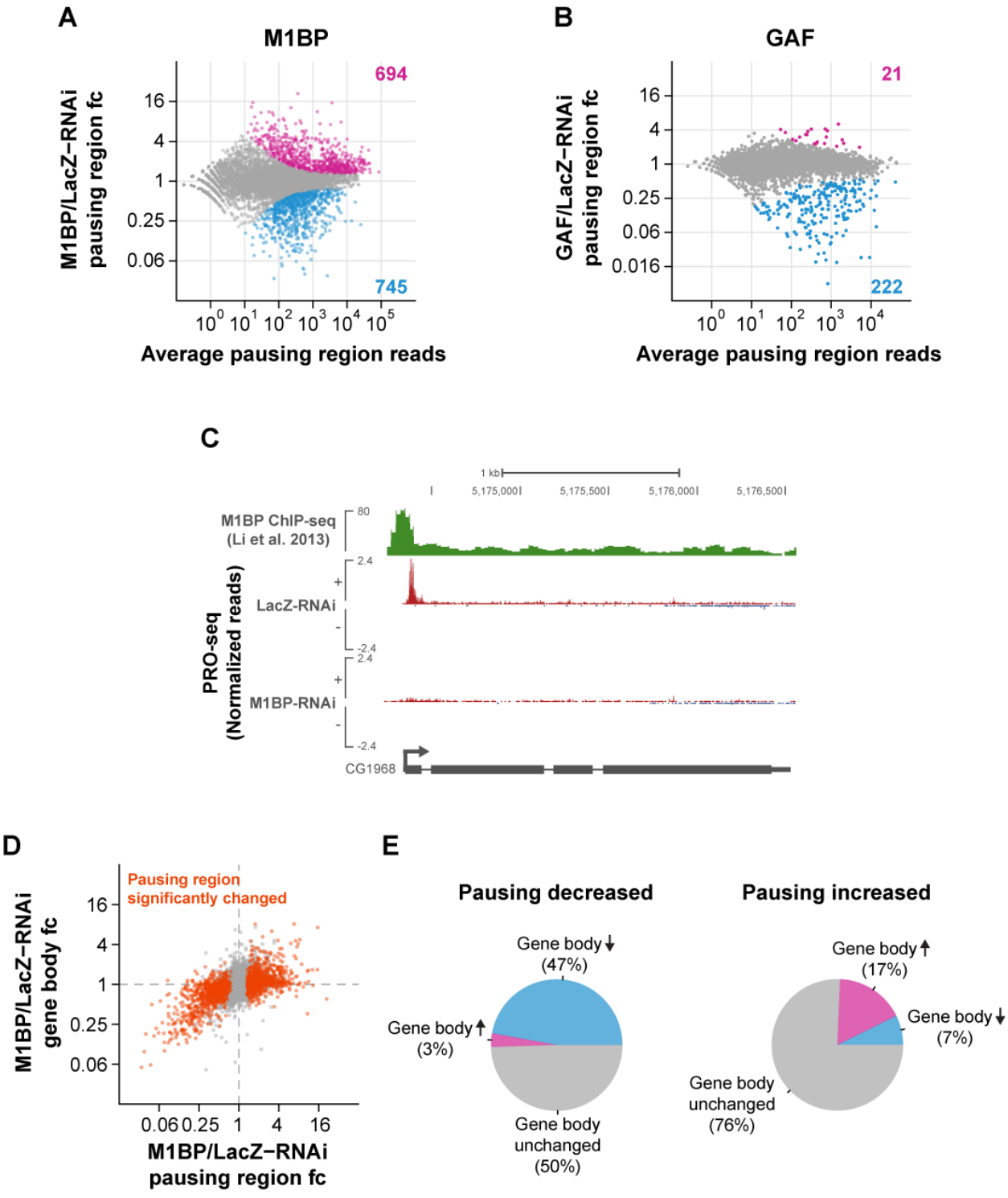
The work presented in this dissertation provided significant contributions to our understanding of the general roles of transcription factors in the regulation of distinct steps in the transcription cycle. It also laid the foundation for numerous follow-up studies, some of which will be discussed in this section.

Characterizing the roles of GAF and M1BP in the establishment of promoter-proximal Pol II pausing

As described in Chapter 2, we have shown that the recently discovered transcription factor M1BP (Li and Gilmour 2013) is enriched in the promoter region of HS activated genes with GAF-independent pausing prior to heat shock. To further investigate M1BP's role in heat shock activation, we performed PRO-seq in biological replicates of LacZ and M1BP-RNAi treated cells before and after 20min of HS. Analysis of these datasets revealed that M1BP is important for pausing and HS activation of a subset of M1BP-bound genes that are GAF-independent (Chapter 2).

Although the initial goal of this experiment was to understand M1BP's function in heat shock activation, a thorough characterization of this dataset will be invaluable for investigating M1BP's role in pausing, and the similarities and differences between GAF- and M1BP-facilitated pausing. Preliminary analyses revealed that M1BP-RNAi has a substantial effect on pausing, which is more extensive than the effect of GAF-RNAi (Figure 4.1A, B. M1BP-RNAi: 694 genes have increased pausing and 745 genes have decreased pausing. GAF-RNAi: 21 genes have increased pausing and 222 genes have

Figure 4.1: M1BP knockdown has a substantial effect on promoter-proximal Pol II pausing, which is mostly independent of gene body changes. (A, B) DESeq2 analysis of PRO-seq pausing region reads after M1BP-RNAi treatment **(A)** and GAF-RNAi treatment **(B)**. Results are displayed as MA plots. Significantly changed genes were defined using an FDR of 0.001. Genes with significantly increased pausing levels are labeled in magenta and genes with significantly decreased pausing levels are labeled in blue. The number of genes in each class is shown in the plot. fc = fold-change. **(C)** Representative view in the UCSC genome browser (Kent et al. 2002) of an M1BP-bound gene whose pausing levels are dramatically decreased by M1BP-RNAi treatment. ChIP-seq reads for M1BP (from Li and Gilmour 2013) are shown in green. PRO-seq normalized reads for the different RNAi treatments (LacZ and M1BP) for the plus strand are shown in red and for the minus strand in blue. Gene annotations are shown at the bottom. **(D)** Scatter plot of M1BP/LacZ-RNAi gene body fold-changes versus M1BP/LacZ-RNAi pausing region fold-changes. Genes with significantly changed pausing region reads are labeled in orange. fc = fold-change. **(E)** Fraction of the total number of genes whose pausing region reads significantly decreased (left) or increased (right) after M1BP-RNAi treatment that is represented by genes with significantly increased, decreased or unchanged gene body reads.


A
M1BP
694
745
M1BP/LacZ−RNAi pausing region fc
Average pausing region reads
B
GAF
21
222
GAF/LacZ−RNAi pausing region fc
Average pausing region reads
C
1 kb
5,175,000
5,175,500
5,176,000
5,176,500
M1BP ChIP-seq (Li et al. 2013)
PRO-seq (Normalized reads)
LacZ-RNAi
M1BP-RNAi
CG1968
D
Pausing region significantly changed
M1BP/LacZ−RNAi gene body fc
M1BP/LacZ−RNAi pausing region fc
E
Pausing decreased
Gene body ↓ (47%)
Gene body ↑ (3%)
Gene body unchanged (50%)
Pausing increased
Gene body ↑ (17%)
Gene body ↓ (7%)
Gene body unchanged (76%)

decreased pausing). Figure 4.1C has an example of an M1BP-bound gene whose pausing levels are dramatically reduced by M1BP knockdown. Moreover, we have shown that M1BP-RNAi's effects on pausing are mostly independent of gene body changes (Figure 4.1D, E).

Interestingly, we observed a substantial enrichment of GAF binding in the promoter region of genes whose pausing increases upon M1BP-RNAi treatment (Figure 4.2). We hypothesize that the positioning of GAF and M1BP relative to the TSS can determine the role of each factor in establishing/maintaining promoter-proximal Pol II pausing. To test this hypothesis, we propose to precisely map the genome-wide positioning of GAF and M1BP using a high resolution method such as ChIP-exo (Rhee and Pugh 2011) or ChIP-nexus (He et al. 2015). ChIP-exo, which was initially developed by the Pugh laboratory, uses exonuclease treatments to map the genome-wide position of chromatin-bound proteins with near base pair resolution (Rhee and Pugh 2011). More recently, the Zeitlinger laboratory described and improved version of the ChIP-exo protocol – called ChIP-nexus – that uses molecular barcoding to reduce PCR biases and a DNA self-circularization step to improve ligation efficiency (He et al. 2015). We propose to perform ChIP-nexus for GAF and M1BP in LacZ, GAF and M1BP-RNAi treated cells. The analysis of these datasets, coupled with the PRO-seq datasets presented in this dissertation, will significantly advance our understanding of the functions of GAF and M1BP in establishing and maintaining promoter-proximal Pol II pausing. Furthermore, the rules learned from these experiments will be essential for guiding our search for the mammalian counterparts of GAF and M1BP.

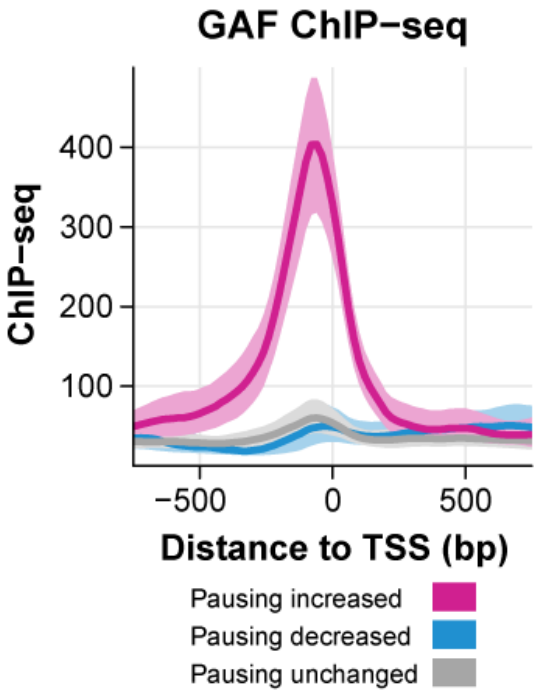


Figure 4.2: GAF is substantially enriched in the promoter region of genes whose pausing levels increase upon M1BP-RNAi treatment. GAF ChIP−seq read density between -750 to +750 bp to the TSS (in 20 bp bins) of genes with increased, decreased and unchanged pausing upon M1BP-RNAi treatment. The shaded area represents the 75% confidence interval.

Investigating the roles of insulator proteins in establishing promoter-proximal Pol II pausing

Similarly to M1BP, the insulator proteins BEAF-32 and Chromator were also found to be enriched in the promoter region of HS activated genes with GAF-independent pausing prior to heat shock (Figure 2.13, Chapter 2). GAF has also been classified as an insulator protein with enhancer-blocking activity (Ohtsuki and Levine 1998; Schweinsberg et al. 2004), which suggests a possible overlap between insulator function and a role in maintaining an open chromatin environment that enables promoter-proximal pausing. Therefore, we hypothesized that BEAF-32 and/or Chromator might have a role in generating pausing when bound proximally to the TSS. Another possible explanation is

that these insulator proteins reside between GAF and the TSS, therefore blocking any activity of GAF on the promoter, which could explain why pausing is not affected by GAF depletion at insulator-bound promoters. To test these hypotheses, we propose to perform PRO-seq after knocking down BEAF-32 and Chromator – to assess how their depletion affects pausing levels – and ChIP-nexus for both factors – to precisely map their positioning relative to the TSS and to other factors such as GAF and M1BP. We have already performed PRO-seq after BEAF-32-RNAi treatment in collaboration with Professor Craig Hart (Louisiana State University). Further analysis of this dataset, coupled with the additional experiments we have proposed, will advance our understanding of the function of insulator proteins and their potential novel role in establishing promoter-proximal pausing.

Characterizing the roles of the nucleosome remodeler NURF in the establishment of GAF-mediated pausing

In Chapter 2, we have demonstrated that GAF has an essential role in the establishment of promoter-proximal pausing and the consequent induction of a subset of HS activated genes. Furthermore, previous studies have shown that GAF is critical for moving nucleosomes at both promoter and intergenic regions (Tsukiyama et al. 1994; Fuda et al. 2015), so we proposed a model in which GAF's role in establishing pausing is connected to its ability to create an open chromatin environment around the TSS. However, it is still unclear whether GAF can act as a pioneer factor and remove nucleosomes by itself or if it requires ATP-dependent cofactors to execute this function.

Tsukiyama and Wu have demonstrated in vitro that GAF's ability to remove nucleosomes from the *Drosophila Hsp70* promoter is greatly facilitated by the ATP-dependent chromatin remodeler NURF (Tsukiyama et al. 1994; Tsukiyama and Wu 1995). Moreover, a recent genome-wide study has shown that NURF interacts physically and functionally with transcription factors to organize nucleosomes downstream of active promoters (Kwon et al. 2016).

We propose to investigate the general role of NURF in the establishment of GAF-dependent promoter-proximal pausing. Initially, PRO-seq after knocking down NURF will reveal if this factor is important for pausing and if there is an overlap between the genes that are affected by GAF and NURF RNAi treatments. A broader effect of NURF knockdown will indicate that this nucleosome remodeler has a role in establishing pausing through interactions with transcription factors other than GAF.

We also propose to perform MNase-seq after knocking down NURF to assess its effects on nucleosome levels and positioning. We hypothesize that there will be an overlap between the genes whose pausing and nucleosome levels are affected by NURF knockdown. ChIP-seq or the high resolution ChIP-nexus can also be performed to enable the identification of NURF's direct targets in S2 cells.

Identifying transcription factors that promote the activation of HSF-independent genes

For many years, the transcription factor HSF has been considered the 'master' regulator of the transcriptional heat shock response. However, we have shown that HSF is important for the activation of only a small minority of HS activated genes (Chapter 2), and a related study in MEFs has observed similar results for the mouse HSF1 (Mahat et

al. 2016). In both studies, thorough motif searches at HS activated genes have failed to identify all the additional transcription factors that promote the activation of HSF-independent genes (Chapter 2, Mahat et al. 2016).

We propose to perform 'factor-independent' chromatin accessibility assays such as DNase-seq (Song and Crawford 2010) or ATAC-seq (Buenrostro et al. 2015) before and after heat shock induction to search for transcription factors with potential novel roles in heat shock activation. Both DNase-seq and ATAC-seq use enzyme accessibility as a proxy for open chromatin regions across the genome. In both methods, DNA regions that are bound by DNA binding proteins are protected from enzyme cleavage, generating 'footprints' that can be used to infer transcription factor binding sites. Therefore, we propose to use one of these assays to identify new transcription factor footprints generated after heat shock induction. Although both methods have comparable sensitivities, ATAC-seq provides a simpler and more efficient experimental workflow (Buenrostro et al. 2013), so we recommend using ATAC-seq for this follow-up experiment.

After determining the set of significantly changed HS-induced footprints, a motif search should be performed in these regions to identify significantly overrepresented motifs. The motif search can be performed de novo using *MEME* (Bailey et al. 2009) or through integrated frameworks that search databases of known motifs, such as *rtfbsdb* (Wang et al. 2016). *rtfbsdb* can also integrate RNA-seq or PRO-seq data to restrict the analyses to motifs associated with transcription factors that are expressed in the cell type of interest (Wang et al. 2016). The transcription factors that recognize the significantly overrepresented motifs will correspond to potential novel regulators of the transcriptional

heat shock response, and their role can be validated through follow-up studies. If an identified significantly overrepresented motif cannot be associated with a known transcription factor, the motif's DNA sequence can be used in affinity chromatography assays to identify novel transcription factors.

Investigating the mechanisms underlying HS-induced transcriptional repression

In Chapter 2, we have demonstrated that HS induction causes a rapid reduction in the transcription levels of thousands of genes, which is regulated at the initiation step and independent of HSF (Figure 2.19). Interestingly, we have recently shown in related studies in MEFs (Mahat et al. 2016) and in human cells (Vihervaara et al., *under review*) that HS-induced transcriptional repression is regulated at the level of promoter-proximal pausing release, indicating that *Drosophila* and mammals have evolved different mechanisms to repress transcription upon heat shock. However, the mechanisms underlying this transcriptional repression are not completely understood in both mammals and *Drosophila*.

In Section 2.4, we pointed out that given the magnitude of this repressive response, it is unlikely that one single transcription repressor is responsible for inhibiting transcription initiation in all HS repressed genes. We have also presented three possible mechanisms, which are not mutually exclusive, that could be responsible for HS-mediated repression. (1) The activity of a general transcription factor that is involved in recruitment of Pol II to the promoter could be modulated by the heat shock signal. (2) Changes in nucleosomal composition or positioning induced by the heat stress could generate an unfavorable chromatin environment that would prevent transcription initiation and

elongation. A previous study has demonstrated that HS results in decreased nucleosome turnover genome-wide within gene bodies; however, a decrease in nucleosome turnover was also observed after drug inhibition of Pol II elongation, arguing that reduced nucleosome turnover may be a consequence rather than the cause of the genome-wide transcriptional repression (Teves and Henikoff 2011). (3) A genome-wide rearrangement of the 3D chromatin structure could either disrupt long-range interactions that are needed for transcription or allow new long-range interactions that repress transcription initiation, which is supported by a recent study in a different *Drosophila* cell line that demonstrated that HS induces a genome-wide rearrangement in the 3D nuclear architecture (Li et al. 2015). Further investigation is necessary to evaluate the role of each of these three alternatives in modulating HS-induced transcriptional repression.

Regarding the third possibility, a thorough characterization of the changes in 3D chromatin structure upon heat shock would substantially advance our understanding of the transcriptional heat shock response. Although Li and colleagues have reported that HS induces a genome-wide rearrangement in the 3D nuclear architecture in a different *Drosophila* cell line (Li et al. 2015), preliminary studies from our laboratory using the in situ Hi-C method (Rao et al. 2014) in *Drosophila* S2 cells have failed to observe genome-wide changes in nuclear architecture, and indicate that HS does not induce general conformational changes. Therefore, further studies are necessary to evaluate the role of 3D rearrangements in HS-induced transcriptional repression.

Investigating the role of long-range interactions in HS-induced transcriptional activation

Besides validating a potential role in repression, investigation of the HS-induced changes in 3D nuclear architecture will likely provide important clues to our understanding of mechanisms of transcription activation. For instance, we have shown in Chapter 2 that HSF may be able to mediate activation at distal enhancer sites on a small subset of HS-activated genes; however, HSF's general role in activating transcription from distant enhancers or in mediating long-range chromatin interactions has not been characterized.

Analyses of the in situ Hi-C dataset described above will reveal the pre-existing and HS-induced long-range interactions and indicate if the HS-activated genes interact with each other and with other genomic regions. Furthermore, methods that specifically enrich for chromatin interactions mediated by a protein of interest, such as ChIA-PET (Fullwood et al. 2009) and the recently developed HiChIP (Mumbach et al. 2016) can be used to identify the specific long-range interactions mediated by HSF. Although groundbreaking advances have been made in the chromosome conformation field, current methods still lack the resolution and the signal-to-noise ratio that are needed to comprehensively characterize enhancer-promoter interactions, and the development of new technologies is required to further advance our understanding of how these interactions are regulated.

Inhibitory RNA aptamers to different HSF domains and their use in dissecting HSF mechanisms of transcription activation

In Chapter 3, we presented the results of a successful RNA aptamer selection to the human transcription factors HSF1 and HSF2. Sequencing analyses of the SELEX pools

identified thousands of enriched sequences, and affinity assays confirmed that many candidates bind to HSF1 and HSF2 with high affinity. Moreover, further analyses suggested that most of the sequences in the final pools can bind to the targets (Chapter 3).

These aptamers were selected with the final goal of being used as tools to dissect HSF's primary functions and mechanisms of action during heat shock activation. To that end, we would ideally generate a collection of high affinity inhibitory RNA aptamers to all of HSF's specific domains. Given the large number of HSF-binding aptamers identified by our SELEX, further characterization of these pools will likely identify aptamers that bind to HSF's three major domains (DNA binding, trimerization and activation domains).

To identify aptamers that bind to each domain, we initially attempted to perform affinity assays with numerous truncations of HSF1 containing individual domains or combinations of domains. However, this approach has proven to be challenging and has failed to identify aptamers that bind to distinct domains. Therefore, we propose to use protease footprinting assays (such as in Sevilimedu et al. 2008) as an alternative approach to map the specific regions of HSF1 that are bound by each aptamer. Further optimization will be necessary to identify the ideal combination of proteases and mass spectrometry can be used to characterize the composition of specific cleavage bands.

After identifying the best candidates, we will express these aptamers in vivo in mammalian cells using Pol III expression systems. The human U6 promoter has been successfully used to drive small RNA expression in human cell nuclei through the pAV U6+27 vector (Good et al. 1997). Inserts driven by this promoter are efficiently transcribed and stay in the nucleus. At their 5' end, the transcribed products contain the first 27

nucleotides of the U6 RNA, which includes the full sequence required for 5' γ-phosphomethyl 'capping' and stabilization. Furthermore, stable stems at the 3' end of the insert immediately upstream of the Pol III polyU terminator stabilize the transcript from 3'-5' exonuclease degradation (Good et al. 1997). Preliminary data indicate that the pAV U6+27 vector produced large amounts of candidate HSF1 aptamers when transiently transfected into HEK-293T cells (Figure 4.3), with an estimated nuclear concentration of approximately 25μM for all tested aptamers.

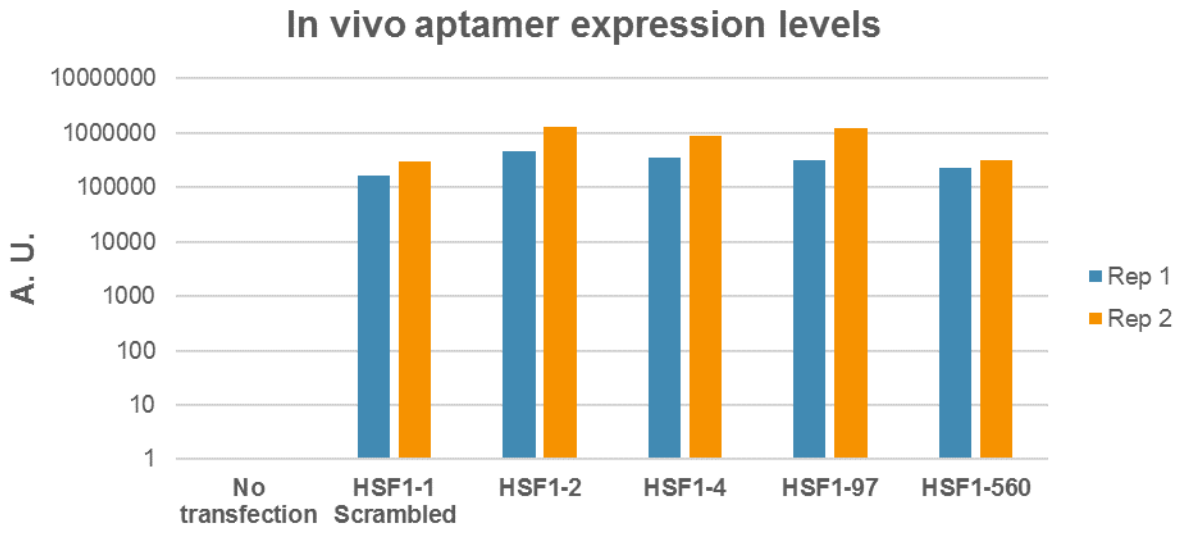


Figure 4.3: Aptamer expression driven by pAV U6+27 vector in HEK-293T cells. Candidate HSF1 aptamers were cloned into the pAV U6+27 vector and transiently transfected into HEK-293T cells. After 2 days, total RNA was extracted and the aptamer levels were measured using RT-qPCR. A.U. = arbitrary units, expression levels were normalized to the no transfection control. Data is shown for two biological replicates for each aptamer.

These results indicate that high in vivo expression levels of candidate HSF1 aptamers can be easily achieved by transiently transfecting cells with the pAV U6+27 vector. We propose to use this simple and effective system to screen for aptamers that

inhibit HSF1's function in vivo. This inhibition can be assessed by measuring the effect of aptamer expression on heat shock transcriptional induction of classical heat shock genes using RT-qPCR. ChIP-qPCR can also be used to assess if the expressed aptamers affect the binding of HSF1 to its known regulatory sequence elements.

After identifying the best candidate aptamers, which will ideally bind and inhibit different domains of HSF1, we will generate stable cells lines that inducibly express these aptamers. Inducible expression can be achieved by modifying our Pol III expression systems using available methods (Meissner et al. 2001; Kappel et al. 2007; Zhou et al. 2008).

These aptamers will then be used as tools to dissect the precise mechanistic roles and macromolecular interactions of each domain of HSF1 during transcription regulation using genome-wide assays. We will perform ChIP-seq and PRO-seq before and after heat shock induction to examine the effects of aptamer expression on the recruitment of HSF1 and its interacting proteins to target genes and on the expression of these genes, respectively. We expect that the various aptamers will have distinct effects. For instance, aptamers to the DNA binding domain will likely reduce the levels of HSF1 binding to its target DNA elements (measured by ChIP-seq) and can be used to study the primary effects of the inhibition of HSF1 recruitment and to determine if the decrease in HSF1 occupancy is directly correlated to a decrease in gene expression (measured by PRO-seq). Since HSF1 trimerization is important for DNA binding, we expect to observe a similar effect when using aptamers that block the trimerization domain. Besides the identification of aptamers that bind and inhibit the trimerization domain, it would be valuable to identify an aptamer that binds to the trimeric form of HSF1 and is capable of

holding it together at this state. We can then investigate how this aptamer affects the duration of the HS response and the mechanisms that control its recovery. Moreover, one could also expect that this aptamer can induce some level of HS response during NHS conditions.

In contrast, we expect that the activation domain aptamers will not affect the recruitment of HSF1 to its target genes. Nevertheless, they can be used to investigate the recruitment levels and dynamics of factors that are known to be recruited after HSF1 binding – such as P-TEFb. Furthermore, the identification of aptamers that bind different regions of the activation domain will be invaluable to tease apart the molecular interactions of HSF1. For instance, if an aptamer for a specific region of the activation domain inhibits the activation of only a subset of genes, this may indicate that HSF1 interacts with different co-activators and uses different pathways to activate its target genes. Moreover, our lab has previously demonstrated in *Drosophila* that there is a rapid, transcription-independent nucleosome loss at *Hsp70* upon HS and that this phenomenon is dependent on HSF and on the HSF-dependent recruitment of the histone acetyl-transferase Tip60 (Petesch and Lis 2008, 2012). We propose to use aptamers to distinct regions of the activation domain – coupled with MNase-seq and PRO-seq – to tease apart HSF1 interactions and decipher how it is involved with transcription-uncoupled nucleosome loss and how this differs from its role in the dramatic transcription activation of hundreds of genes. We hope to identify aptamers that can affect the dramatic transcription activation but do not interfere with the nucleosome loss and vice-versa, which will enable the identification of specific surfaces of this factor that are involved in each phenomenon.

Finally, we will evaluate the effectiveness of our inhibitory RNA aptamer strategy by comparing these datasets with those generated after depleting HSF levels using traditional methods (Mahat et al. 2016; Duarte et al. 2016; Vihervaara et al., *under review*). We expect that aptamer expression will be effective in overcoming the main limitations imposed by traditional approaches – such as confounding secondary or compensatory effects and the inability to inhibit specific domains and macromolecular interactions of a protein. These studies will hopefully serve as a proof-of-principle and establish our inhibitory RNA aptamer strategy as a general method to study macromolecular interactions in vivo.

REFERENCES

Adelman K, Kennedy MA, Nechaev S, Gilchrist DA, Muse GW, Chinenov Y, Rogatsky I. 2009. Immediate mediators of the inflammatory response are poised for gene activation through RNA polymerase II stalling. *Proc Natl Acad Sci* **106**: 18207–18212.

Adelman K, Lis JT. 2012. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* **13**: 720–31.

Anckar J, Sistonen L. 2007. Heat shock factor 1 as a coordinator of stress and developmental pathways. *Adv Exp Med Biol* **594**: 78–88.

Anders S, Pyl PT, Huber W. 2014. HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–9.

Ardehali MB, Yao J, Adelman K, Fuda NJ, Petesch SJ, Webb WW, Lis JT. 2009. Spt6 enhances the elongation rate of RNA polymerase II in vivo. *EMBO J* **28**: 1067–77.

Artavanis-Tsakonas S, Schedl P, Mirault ME, Moran L, Lis J, Leder A, Enquist LW, Norman B, Leder P. 1979. Genes for the 70,000 dalton heat shock protein in two cloned D. melanogaster DNA segments. *Cell* **17**: 9–18.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202-8.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**: 28–36.

Barboric M, Nissen RM, Kanazawa S, Jabrane-Ferrat N, Peterlin BM. 2001. NF-kappaB binds P-TEFb to stimulate transcriptional elongation by RNA polymerase II. *Mol Cell* **8**: 327–37.

Bienz M, Pelham HRB. 1986. Heat shock regulatory elements function as an inducible enhancer in the xenopus hsp70 gene and when linked to a heterologous promoter.

Cell **45**: 753–760.

Blau J, Xiao H, McCracken S, O'Hare P, Greenblatt J, Bentley D. 1996. Three functional classes of transcriptional activation domain. *Mol Cell Biol* **16**: 2044–55.

Boehm AK, Saunders A, Werner J, Lis JT. 2003. Transcription factor and polymerase recruitment, modification, and movement on dhsp70 in vivo in the minutes following heat shock. *Mol Cell Biol* **23**: 7628–37.

Buckley MS, Kwak H, Zipfel WR, Lis JT. 2014. Kinetics of promoter Pol II on Hsp70 reveal stable pausing and key insights into its regulation. *Genes Dev* **28**: 14–9.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**: 1213–1218.

Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. In *Current Protocols in Molecular Biology*, Vol. 109 of, p. 21.29.1-21.29.9, John Wiley & Sons, Inc., Hoboken, NJ, USA.

Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, et al. 2009. Unlocking the secrets of the genome. *Nature* **459**: 927–30.

Chao S-H, Price DH. 2001. Flavopiridol Inactivates P-TEFb and Blocks Most RNA Polymerase II Transcription in Vivo. *J Biol Chem* **276**: 31793–31799.

Cho M, Xiao Y, Nie J, Stewart R, Csordas AT, Oh SS, Thomson JA, Soh HT. 2010. Quantitative selection of DNA aptamers through microfluidic selection and high-throughput sequencing. *Proc Natl Acad Sci* **107**: 15373–15378.

Churchman LS, Weissman JS. 2011. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**: 368–73.

Clos J, Westwood JT, Becker PB, Wilson S, Lambert K, Wu C. 1990. Molecular cloning and expression of a hexameric Drosophila heat shock factor subject to negative regulation. *Cell* **63**: 1085–97.

Copeland RA. 2016. The drug-target residence time model: a 10-year retrospective. *Nat Rev Drug Discov* **15**: 87–95.

Corces V, Holmgren R, Freund R, Morimoto R, Meselson M. 1980. Four heat shock proteins of Drosophila melanogaster coded within a 12-kilobase region in chromosome subdivision 67B. *Proc Natl Acad Sci U S A* **77**: 5390–3.

Core LJ, Waterfall JJ, Gilchrist DA, Fargo DC, Kwak H, Adelman K, Lis JT. 2012. Defining the status of RNA polymerase at promoters. *Cell Rep* **2**: 1025–35.

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**: 1845–8.

Cox JC, Rudolph P, Ellington AD. 1998. Automated RNA Selection. *Biotechnol Prog* **14**: 845–850.

Craig EA, McCarthy BJ. 1980. Four Drosophila heat shock genes at 67B: characterization of recombinant plasmids. *Nucleic Acids Res* **8**: 4441–57.

Craig EA, McCarthy BJ, Wadsworth SC, Baxter JD, Goodman HM, Goldschmidt-Clermont M, Moran L, Tissières A. 1979. Sequence organization of two recombinant plasmids containing genes for the major heat shock-induced protein of D. melanogaster. *Cell* **16**: 575–88.

Dai C, Whitesell L, Rogers AB, Lindquist S. 2007. Heat shock factor 1 is a powerful multifaceted modifier of carcinogenesis. *Cell* **130**: 1005–18.

Danko CG, Hah N, Luo X, Martins AL, Core L, Lis JT, Siepel A, Kraus WL. 2013. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol Cell* **50**: 212–222.

Davis JH, Szostak JW. 2002. Isolation of high-affinity GTP aptamers from partially structured RNA libraries. *Proc Natl Acad Sci U S A* **99**: 11616–21.

Duarte FM, Fuda NJ, Mahat DB, Core LJ, Guertin MJ, Lis JT. 2016. Transcription factors GAF and HSF act at distinct regulatory steps to modulate stress-induced gene activation. *Genes Dev* **30**: 1731–46.

Eberhardy SR, Farnham PJ. 2002. Myc Recruits P-TEFb to Mediate the Final Step in the Transcriptional Activation of the cad Promoter. *J Biol Chem* **277**: 40156–40162.

Ellington AD, Szostak JW. 1990. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**: 818–22.

Fan X, Shi H, Adelman K, Lis JT. 2004. Probing TBP interactions in transcription initiation and reinitiation with RNA aptamers that act in distinct modes. *Proc Natl Acad Sci U S A* **101**: 6934–9.

Fan X, Shi H, Lis JT. 2005. Distinct transcriptional responses of RNA polymerases I, II and III to aptamers that bind TBP. *Nucleic Acids Res* **33**: 838–45.

Fuda NJ, Ardehali MB, Lis JT. 2009. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461**: 186–92.

Fuda NJ, Guertin MJ, Sharma S, Danko CG, Martins AL, Siepel A, Lis JT. 2015. GAGA Factor Maintains Nucleosome-Free Regions and Has a Role in RNA Polymerase II Recruitment to Promoters. *PLoS Genet* **11**: e1005108.

Fujinaga K, Irwin D, Huang Y, Taube R, Kurosu T, Peterlin BM. 2004. Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. *Mol Cell Biol* **24**: 787–95.

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed Y Bin, Orlov YL, Velkov S, Ho A, Mei PH, et al. 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**: 58–64.

Gan M, Moebus S, Eggert H, Saumweber H. 2011. The Chriz–Z4 complex recruits JIL-1 to polytene chromosomes,a requirement for interband-specific phosphorylation of H3S10. *J Biosci* **36**: 425–438.

Geiger A, Burgstaller P, von der Eltz H, Roeder A, Famulok M. 1996. RNA aptamers that bind L-arginine with sub-micromolar dissociation constants and high enantioselectivity. *Nucleic Acids Res* **24**: 1029–36.

Gilmour DS, Fan R. 2009. Detecting transcriptionally engaged RNA polymerase in eukaryotic cells with permanganate genomic footprinting. *Methods* **48**: 368–374.

Gilmour DS, Lis JT. 1986. RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. *Mol Cell Biol* **6**: 3984–9.

Glaser RL, Thomas GH, Siegfried E, Elgin SC, Lis JT. 1990. Optimal heat-induced expression of the Drosophila hsp26 gene requires a promoter sequence containing (CT)n.(GA)n repeats. *J Mol Biol* **211**: 751–61.

Gold L, Ayers D, Bertino J, Bock C, Bock A, Brody EN, Carter J, Dalby AB, Eaton BE, Fitzwater T, et al. 2010. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* **5**: e15004.

Gonsalves SE, Moses AM, Razak Z, Robert F, Westwood JT. 2011. Whole-genome analysis reveals that active heat shock factor binding sites are mostly associated with non-heat shock genes in Drosophila melanogaster. *PLoS One* **6**: e15934.

Good PD, Krikos AJ, Li SX, Bertrand E, Lee NS, Giver L, Ellington A, Zaia JA, Rossi JJ, Engelke DR. 1997. Expression of small, therapeutic RNAs in human cell nuclei. *Gene Ther* **4**: 45–54.

Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. 2007. A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. *Cell* **130**: 77–88.

Guertin MJ, Lis JT. 2010. Chromatin landscape dictates HSF binding to target DNA elements. *PLoS Genet* **6**.

Guertin MJ, Petesch SJ, Zobeck KL, Min IM, Lis JT. 2010. Drosophila heat shock system as a general model to investigate transcriptional regulation. *Cold Spring Harb Symp Quant Biol* **75**: 1–9.

Guhathakurta D, Palomar L, Stormo GD, Tedesco P, Johnson TE, Walker DW, Lithgow G, Kim S, Link CD. 2002. Identification of a novel cis-regulatory element involved in the heat shock response in Caenorhabditis elegans using microarray gene expression and computational methods. *Genome Res* **712**: 701–712.

Hall B, Micheletti JM, Satya P, Ogle K, Pollard J, Ellington AD. 2009. Design, synthesis, and amplification of DNA pools for in vitro selection. *Curr Protoc Mol Biol* **Chapter 24**: Unit 24.2.

He Q, Johnston J, Zeitlinger J. 2015. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat Biotechnol* **33**: 395–401.

Henriques T, Gilchrist DA, Nechaev S, Bern M, Muse GW, Burkholder A, Fargo DC, Adelman K. 2013. Stable Pausing by RNA Polymerase II Provides an Opportunity to Target and Integrate Regulatory Signals. *Mol Cell* **52**: 517–528.

Herschlag D, Johnson FB. 1993. Synergism in transcriptional activation: a kinetic view. *Genes Dev* **7**: 173–9.

Holmgren R, Livak K, Morimoto R, Freund R, Meselson M, Goldschmidt-Clermont M, Moran L, Tissières A, Smith M. 1979. Studies of cloned sequences from four drosophila heat shock loci. *Cell* **18**: 1359–1370.

Hou C, Li L, Qin ZS, Corces VG. 2012. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol Cell* **48**: 471–84.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**: 44–57.

Ish-Horowicz D, Pinchin SM, Schedl P, Artavanis-Tsakonas S, Mirault ME, Goldschmidt-Clermont M, Moran L, Tissières A, Georgiev GP. 1979. Genetic and molecular analysis of the 87A7 and 87C1 heat-inducible loci of D. melanogaster. *Cell* **18**: 1351–8.

Jamrich M, Greenleaf AL, Bautz EK. 1977. Localization of RNA polymerase in polytene chromosomes of Drosophila melanogaster. *Proc Natl Acad Sci* **74**: 2079–2083.

Jonkers I, Kwak H, Lis JT. 2014. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife* **3**: e02407.

Juven-Gershon T, Hsu J-Y, Theisen JW, Kadonaga JT. 2008. The RNA polymerase II core promoter — the gateway to transcription. *Curr Opin Cell Biol* **20**: 253–259.

Kanazawa S, Soucek L, Evan G, Okamoto T, Peterlin BM. 2003. c-Myc recruits P-TEFb for transcription, cellular proliferation and apoptosis. *Oncogene* **22**: 5707–5711.

Kappel S, Matthess Y, Kaufmann M, Strebhardt K. 2007. Silencing of mammalian genes by tetracycline-inducible shRNA expression. *Nat Protoc* **2**: 3257–3269.

Keene MA, Corces V, Lowenhaupt K, Elgin SC. 1981. DNase I hypersensitive sites in Drosophila chromatin occur at the 5' ends of regions of transcription. *Proc Natl Acad Sci U S A* **78**: 143–6.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996–1006.

Kharchenko P V, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, et al. 2011. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471**: 480–5.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36.

Kim JB, Sharp PA. 2001. Positive Transcription Elongation Factor b Phosphorylates hSPT5 and RNA Polymerase II Carboxyl-terminal Domain Independently of Cyclin-dependent Kinase-activating Kinase. *J Biol Chem* **276**: 12317–12323.

Kwak H, Fuda NJ, Core LJ, Lis JT. 2013. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**: 950–3.

Kwak H, Lis JT. 2013. Control of transcriptional elongation. *Annu Rev Genet* **47**: 483–508.

Kwon SY, Grisan V, Jang B, Herbert J, Badenhorst P. 2016. Genome-Wide Mapping Targets of the Metazoan Chromatin Remodeling Factor NURF Reveals Nucleosome Remodeling at Enhancers, Core Promoters and Gene Insulators ed. B. Hendrich. *PLOS Genet* **12**: e1005969.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Larschan E, Bishop EP, Kharchenko P V, Core LJ, Lis JT, Park PJ, Kuroda MI. 2011. X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. *Nature* **471**: 115–8.

Latulippe DR, Szeto K, Ozer A, Duarte FM, Kelly C V, Pagano JM, White BS, Shalloway D, Lis JT, Craighead HG. 2013. Multiplexed microcolumn-based process for efficient selection of RNA aptamers. *Anal Chem* **85**: 3417–24.

Lee C, Li X, Hechmer A, Eisen M, Biggin MD, Venters BJ, Jiang C, Li J, Pugh BF, Gilmour DS. 2008. NELF and GAGA factor are linked to promoter-proximal pausing at many genes in Drosophila. *Mol Cell Biol* **28**: 3290–300.

Lee H, Kraus KW, Wolfner MF, Lis JT. 1992. DNA sequence requirements for generating paused polymerase at the start of hsp70. *Genes Dev* **6**: 284–295.

Leemans R, Egger B, Loop T, Kammermeier L, He H, Hartmann B, Certa U, Hirth F, Reichert H. 2000. Quantitative transcript imaging in normal and heat-shocked Drosophila embryos by using high-density oligonucleotide arrays. *Proc Natl Acad Sci U S A* **97**: 12138–43.

Levine M. 2011. Paused RNA Polymerase II as a Developmental Checkpoint. *Cell* **145**: 502–511.

Li J, Gilmour DS. 2013. Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor. *EMBO J* **32**: 1829–41.

Li L, Lyu X, Hou C, Takenaka N, Nguyen HQ, Ong C-T, Cubeñas-Potts C, Hu M, Lei EP, Bosco G, et al. 2015. Widespread Rearrangement of 3D Chromatin Organization Underlies Polycomb-Mediated Stress-Induced Silencing. *Mol Cell* **58**: 216–231.

Lindquist S. 1986. The Heat-Shock Response. *Annu Rev Biochem* **55**: 1151–1191.

Lindquist S, Craig EA. 1988. The heat-shock proteins. *Annu Rev Genet* **22**: 631–77.

Lindquist S, Petersen R. 1990. Selective translation and degradation of heat-shock messenger RNAs in Drosophila. *Enzyme* **44**: 147–66.

Lis J. 1998. Promoter-associated pausing in promoter architecture and postinitiation transcriptional regulation. *Cold Spring Harb Symp Quant Biol* **63**: 347–56.

Lis JT, Mason P, Peng J, Price DH, Werner J. 2000. P-TEFb kinase recruitment and function at heat shock loci. *Genes Dev* **14**: 792–803.

Lis JT, Neckameyer W, Dubensky R, Costlow N. 1981. Cloning and characterization of nine heat-shock-induced mRNAs of Drosophila melanogaster. *Gene* **15**: 67–80.

Livak KJ, Freund R, Schweber M, Wensink PC, Meselson M. 1978. Sequence organization and transcription at two heat shock loci in Drosophila. *Proc Natl Acad Sci U S A* **75**: 5613–7.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Lu Q, Wallrath LL, Granok H, Elgin SC. 1993. (CT)n (GA)n repeats and heat shock elements have distinct roles in chromatin structure and transcriptional activation of the Drosophila hsp26 gene. *Mol Cell Biol* **13**: 2802–2814.

Mahat DB, Salamanca HH, Duarte FM, Danko CG, Lis JT. 2016. Mammalian heat shock response and mechanisms underlying its genome-wide transcriptional regulation. *Mol Cell* **62**: 63–78.

Marshall NF, Price DH. 1995. Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *J Biol Chem* **270**: 12335–8.

Meissner W, Rothfels H, Schäfer B, Seifart K. 2001. Development of an inducible pol III transcription system essentially requiring a mutated form of the TATA-binding protein. *Nucleic Acids Res* **29**: 1672–82.

Min IM, Waterfall JJ, Core LJ, Munroe RJ, Schimenti J, Lis JT. 2011. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev* **25**: 742–54.

Mirault ME, Goldschmidt-Clermont M, Artavanis-Tsakonas S, Schedl P. 1979. Organization of the multiple genes for the 70,000-dalton heat-shock protein in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **76**: 5254–8.

Moran L, Mirault ME, Tissières A, Lis J, Schedl P, Artavanis-Tsakonas S, Gehring WJ. 1979. Physical map of two D. melanogaster DNA segments containing sequences coding for the 70,000 dalton heat shock protein. *Cell* **17**: 1–8.

Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, Chang HY. 2016. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13**: 919–922.

Murray JI, Whitfield ML, Trinklein ND, Myers RM, Brown PO, Botstein D. 2004. Diverse and specific gene expression responses to stresses in cultured human cells. *Mol Biol Cell* **15**: 2361–74.

Muse GW, Gilchrist DA, Nechaev S, Shah R, Parker JS, Grissom SF, Zeitlinger J, Adelman K. 2007. RNA polymerase is poised for activation across the genome. *Nat Genet* **39**: 1507–1511.

Nakayama T, Shimojima T, Hirose S. 2012. The PBAP remodeling complex is required for histone H3.3 replacement at chromatin boundaries and for boundary functions. *Development* **139**: 4582–90.

Narita T, Yamaguchi Y, Yano K, Sugimoto S, Chanarat S, Wada T, Kim D, Hasegawa J, Omori M, Inukai N, et al. 2003. Human transcription elongation factor NELF: identification of novel subunits and reconstitution of the functionally active complex. *Mol Cell Biol* **23**: 1863–73.

Ni Z, Saunders A, Fuda NJ, Yao J, Suarez J-R, Webb WW, Lis JT. 2008. P-TEFb is critical for the maturation of RNA polymerase II into productive elongation in vivo. *Mol Cell Biol* **28**: 1161–70.

Nowak DE, Tian B, Jamaluddin M, Boldogh I, Vergara LA, Choudhary S, Brasier AR. 2008. RelA Ser276 Phosphorylation Is Required for Activation of a Subset of NF- B-Dependent Genes by Recruiting Cyclin-Dependent Kinase 9/Cyclin T1 Complexes. *Mol Cell Biol* **28**: 3623–3638.

O'Brien T, Lis JT. 1993. Rapid changes in Drosophila transcription after an instantaneous heat shock. *Mol Cell Biol* **13**: 3456–3463.

O'Brien T, Wilkins RC, Giardina C, Lis JT. 1995. Distribution of GAGA protein on Drosophila genes in vivo. *Genes Dev* **9**: 1098–1110.

Ohtsuki S, Levine M. 1998. GAGA mediates the enhancer blocking activity of the eve promoter in the Drosophila embryo. *Genes Dev* **12**: 3325–3330.

Okada M, Hirose S. 1998. Chromatin Remodeling Mediated by Drosophila GAGA Factor and ISWI Activates fushi tarazu Gene Transcription In Vitro. *Mol Cell Biol* **18**: 2455–2461.

Omichinski JG, Pedone P V, Felsenfeld G, Gronenborn AM, Clore GM. 1997. The solution structure of a specific GAGA factor–DNA complex reveals a modular binding mode. *Nat Struct Biol* **4**: 122–132.

Pagano JM, Clingman CC, Ryder SP. 2011. Quantitative approaches to monitor protein-nucleic acid interactions using fluorescent probes. *RNA* **17**: 14–20.

Pagano JM, Farley BM, McCoig LM, Ryder SP. 2007. Molecular basis of RNA recognition by the embryonic polarity determinant MEX-5. *J Biol Chem* **282**: 8883–94.

Parker CS, Topol J. 1984. A Drosophila RNA polymerase II transcription factor binds to the regulatory site of an hsp 70 gene. *Cell* **37**: 273–83.

Pelham HR, Bienz M. 1982. A synthetic heat-shock promoter element confers heat-inducibility on the herpes simplex virus thymidine kinase gene. *EMBO J* **1**: 1473–7.

Pelham HRB. 1982. A regulatory upstream promoter element in the Drosophila Hsp 70 heat-shock gene. *Cell* **30**: 517–528.

Perisic O, Xiao H, Lis JT. 1989. Stable binding of Drosophila heat shock factor to head-to-head and tail-to-tail repeats of a conserved 5 bp recognition unit. *Cell* **59**: 797–806.

Peterlin BM, Price DH. 2006. Controlling the elongation phase of transcription with P-TEFb. *Mol Cell* **23**: 297–305.

Petesch SJ, Lis JT. 2012. Activator-induced spread of poly(ADP-ribose) polymerase promotes nucleosome loss at Hsp70. *Mol Cell* **45**: 64–74.

Petesch SJ, Lis JT. 2008. Rapid, transcription-independent loss of nucleosomes over a

large chromatin domain at Hsp70 loci. *Cell* **134**: 74–84.

Price DH. 2000. P-TEFb, a cyclin-dependent kinase controlling elongation by RNA polymerase II. *Mol Cell Biol* **20**: 2629–34.

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

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–2.

Rabindran SK, Giorgi G, Clos J, Wu C. 1991. Molecular cloning and expression of a human heat shock factor, HSF1. *Proc Natl Acad Sci U S A* **88**: 6906–10.

Rabindran SK, Haroun RI, Clos J, Wisniewski J, Wu C. 1993. Regulation of heat shock factor trimer formation: role of a conserved leucine zipper. *Science* **259**: 230–4.

Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA, Adelman K, Wei W, et al. 2010. c-Myc regulates transcriptional pause release. *Cell* **141**: 432–45.

Ramanathan Y, Rajpara SM, Reza SM, Lees E, Shuman S, Mathews MB, Pe'ery T. 2001. Three RNA Polymerase II Carboxyl-terminal Domain Kinases Display Distinct Substrate Preferences. *J Biol Chem* **276**: 10913–10920.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**: 1665–1680.

Rasmussen EB, Lis JT. 1993. In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proc Natl Acad Sci* **90**: 7923–7927.

Rath U, Ding Y, Deng H, Qi H, Bao X, Zhang W, Girton J, Johansen J, Johansen KM. 2006. The chromodomain protein, Chromator, interacts with JIL-1 kinase and regulates the structure of Drosophila polytene chromosomes. *J Cell Sci* **119**: 2332–41.

Rhee HS, Pugh BF. 2011. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell* **147**: 1408–1419.

Rhee HS, Pugh BF. 2012. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**: 295–301.

Ritossa F. 1962. A new puffing pattern induced by temperature shock and DNP in Drosophila. *Experientia* **18**: 571–573.

Rougvie AE, Lis JT. 1988. The RNA polymerase II molecule at the 5′ end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell* **54**: 795–804.

Salamanca HH, Antonyak MA, Cerione RA, Shi H, Lis JT. 2014. Inhibiting Heat Shock Factor 1 in Human Cancer Cells with a Potent RNA Aptamer ed. S.D. Westerheide. *PLoS One* **9**: e96330.

Salamanca HH, Fuda N, Shi H, Lis JT. 2011. An RNA aptamer perturbs heat shock transcription factor activity in Drosophila melanogaster. *Nucleic Acids Res* **39**: 6729–40.

Sarge KD, Zimarino V, Holm K, Wu C, Morimoto RI. 1991. Cloning and characterization of two mouse heat shock factors with distinct inducible and constitutive DNA-binding ability. *Genes Dev* **5**: 1902–11.

Saunders A, Core LJ, Lis JT. 2006. Breaking barriers to transcription elongation. *Nat Rev Mol Cell Biol* **7**: 557–67.

Saunders A, Werner J, Andrulis ED, Nakayama T, Hirose S, Reinberg D, Lis JT. 2003. Tracking FACT and the RNA polymerase II elongation complex through chromatin in vivo. *Science* **301**: 1094–6.

Schedl P, Artavanis-Tsakonas S, Steward R, Gehring WJ, Mirault ME, Goldschmidt-Clermont M, Moran L, Tissières A. 1978. Two hybrid plasmids with D. melanogaster DNA sequences complementary to mRNA coding for the major heat shock protein. *Cell* **14**: 921–9.

Scholes C, DePace AH, Sánchez Á. 2017. Combinatorial Gene Regulation through Kinetic Control of the Transcription Cycle. *Cell Syst* **4**: 97–108.e9.

Schuetz TJ, Gallo GJ, Sheldon L, Tempst P, Kingston RE. 1991. Isolation of a cDNA for HSF2: evidence for two heat shock factor genes in humans. *Proc Natl Acad Sci U S A* **88**: 6911–5.

Schwartz BE, Larochelle S, Suter B, Lis JT. 2003. Cdk7 is required for full activation of Drosophila heat shock genes and RNA polymerase II phosphorylation in vivo. *Mol Cell Biol* **23**: 6876–86.

Schwartz YB, Linder-Basso D, Kharchenko P V, Tolstorukov MY, Kim M, Li H-B, Gorchakov AA, Minoda A, Shanower G, Alekseyenko AA, et al. 2012. Nature and function of insulator protein binding sites in the Drosophila genome. *Genome Res* **22**: 2188–98.

Schweinsberg S, Hagstrom K, Gohl D, Schedl P, Kumar RP, Mishra R, Karch F. 2004. The enhancer-blocking activity of the Fab-7 boundary from the Drosophila bithorax complex requires GAGA-factor-binding sites. *Genetics* **168**: 1371–84.

Sevilimedu A, Shi H, Lis JT. 2008. TFIIB aptamers inhibit transcription by perturbing PIC formation at distinct stages. *Nucleic Acids Res* **36**: 3118–27.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**: 458–72.

Shalgi R, Hurt JA, Lindquist S, Burge CB. 2014. Widespread inhibition of posttranscriptional splicing shapes the cellular transcriptome following heat shock. *Cell Rep* **7**: 1362–70.

Shi H, Fan X, Sevilimedu A, Lis JT. 2007. RNA aptamers directed to discrete functional sites on a single protein structural domain. *Proc Natl Acad Sci U S A* **104**: 3742–6.

Shi H, Hoffman BE, Lis JT. 1999. RNA aptamers as effective protein antagonists in a multicellular organism. *Proc Natl Acad Sci U S A* **96**: 10033–8.

Shopland LS, Hirayoshi K, Fernandes M, Lis JT. 1995. HSF access to heat shock elements in vivo depends critically on promoter architecture defined by GAGA factor, TFIID, and RNA polymerase II binding sites. *Genes Dev* **9**: 2756–2769.

Song L, Crawford GE. 2010. DNase-seq: A High-Resolution Technique for Mapping

Active Gene Regulatory Elements across the Genome from Mammalian Cells. *Cold Spring Harb Protoc* **2010**: pdb.prot5384-prot5384.

Sørensen JG, Nielsen MM, Kruhøffer M, Justesen J, Loeschcke V. 2005. Full genome gene expression analysis of the heat stress response in Drosophila melanogaster. *Cell Stress Chaperones* **10**: 312.

Spradling A, Penman S, Pardue ML. 1975. Analysis of drosophila mRNA by in situ hybridization: Sequences transcribed in normal and heat shocked cultured cells. *Cell* **4**: 395–404.

Szeto K, Reinholt SJ, Duarte FM, Pagano JM, Ozer A, Yao L, Lis JT, Craighead HG. 2014. High-throughput binding characterization of RNA aptamer selections using a microplate-based multiplex microcolumn device. *Anal Bioanal Chem* **406**: 2727–2732.

Teves SS, Henikoff S. 2011. Heat shock reduces stalled RNA polymerase II and nucleosome turnover genome-wide. *Genes Dev* **25**: 2387–97.

Tome JM, Ozer A, Pagano JM, Gheba D, Schroth GP, Lis JT. 2014. Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. *Nat Methods* **11**: 683–8.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–5.

Trinklein ND, Murray JI, Hartman SJ, Botstein D, Myers RM. 2004. The role of heat shock transcription factor 1 in the genome-wide regulation of the mammalian heat shock response. *Mol Biol Cell* **15**: 1254–61.

Tsukiyama T, Becker PB, Wu C. 1994. ATP-dependent nucleosome disruption at a heat-shock promoter mediated by binding of GAGA transcription factor. *Nature* **367**: 525–32.

Tsukiyama T, Wu C. 1995. Purification and properties of an ATP-dependent nucleosome remodeling factor. *Cell* **83**: 1011–20.

Tuerk C, Gold L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**: 505–10.

Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. 2009. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**: 252–63.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The Sequence of the Human Genome. *Science (80-)* **291**: 1304–1351.

Vihervaara A, Sergelius C, Vasara J, Blom MAH, Elsing AN, Roos-Mattjus P, Sistonen L. 2013. Transcriptional response to stress in the dynamic chromatin environment of cycling and mitotic cells. *Proc Natl Acad Sci U S A* **110**: E3388-97.

Voellmy R, Goldschmidt-Clermont M, Southgate R, Tissières A, Levis R, Gehring W. 1981. A DNA segment isolated from chromosomal site 67B in D. melanogaster contains four closely linked heat-shock genes. *Cell* **23**: 261–70.

Wada T, Takagi T, Yamaguchi Y, Ferdous A, Imai T, Hirose S, Sugimoto S, Yano K, Hartzog GA, Winston F, et al. 1998. DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes Dev* **12**: 343–56.

Wadsworth SC, Craig EA, McCarthy BJ. 1980. Genes for three Drosophila heat-shock-induced proteins at a single locus. *Proc Natl Acad Sci U S A* **77**: 2134–7.

Wang YV, Tang H, Gilmour DS. 2005. Identification in vivo of different rate-limiting steps associated with transcriptional activators in the presence and absence of a GAGA element. *Mol Cell Biol* **25**: 3543–52.

Wang Z, Martins AL, Danko CG. 2016. RTFBSDB: an integrated framework for transcription factor binding site analysis. *Bioinformatics* **32**: 3024–3026.

Wilkins R. 1998. GAGA factor binding to DNA via a single trinucleotide sequence element. *Nucleic Acids Res* **26**: 2672–2678.

Wisniewski J, Orosz A, Allada R, Wu C. 1996. The C-terminal region of Drosophila heat shock factor (HSF) contains a constitutively functional transactivation domain.

Nucleic Acids Res **24**: 367–74.

Wu C. 1985. An exonuclease protection assay reveals heat-shock element and TATA box DNA-binding proteins in crude nuclear extracts. *Nature* **317**: 84–7.

Wu C. 1995. Heat shock transcription factors: structure and regulation. *Annu Rev Cell Dev Biol* **11**: 441–69.

Wu C. 1980. The 5’ ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature* **286**: 854–60.

Wu C. 1984. Two protein-binding sites in chromatin implicated in the activation of heat-shock genes. *Nature* **309**: 229–34.

Wu C-H, Lee C, Fan R, Smith MJ, Yamaguchi Y, Handa H, Gilmour DS. 2005. Molecular characterization of Drosophila NELF. *Nucleic Acids Res* **33**: 1269–1279.

Wu C-H, Yamaguchi Y, Benjamin LR, Horvat-Gordon M, Washinsky J, Enerly E, Larsson J, Lambertsson A, Handa H, Gilmour D. 2003. NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila. *Genes Dev* **17**: 1402–1414.

Wu C, Bingham PM, Livak KJ, Holmgren R, Elgin SC. 1979a. The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. *Cell* **16**: 797–806.

Wu C, Wong YC, Elgin SC, Kretschmer P, Nienhuis AW, Moran L, Tissières A. 1979b. The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. *Cell* **16**: 807–14.

Xiao H, Lis J. 1988. Germline transformation used to define key features of heat-shock response elements. *Science (80-)* **239**: 1139–1142.

Xiao X, Zuo X, Davis AA, McMillan DR, Curry BB, Richardson JA, Benjamin IJ. 1999. HSF1 is required for extra-embryonic development, postnatal growth and protection during inflammatory responses in mice. *EMBO J* **18**: 5943–52.

Yamada T, Yamaguchi Y, Inukai N, Okamoto S, Mura T, Handa H. 2006. P-TEFb-

Mediated Phosphorylation of hSpt5 C-Terminal Repeats Is Critical for Processive Transcription Elongation. *Mol Cell* **21**: 227–237.

Yamaguchi Y, Takagi T, Wada T, Yano K, Furuya A, Sugimoto S, Hasegawa J, Handa H. 1999. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell* **97**: 41–51.

Yost HJ, Lindquist S. 1986. RNA splicing is interrupted by heat shock and is rescued by heat shock protein synthesis. *Cell* **45**: 185–193.

Zeitlinger J, Stark A, Kellis M, Hong J-W, Nechaev S, Adelman K, Levine M, Young RA. 2007. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet* **39**: 1512–1516.

Zhao K, Hart CM, Laemmli UK. 1995. Visualization of chromosomal domains with boundary element-associated factor BEAF-32. *Cell* **81**: 879–889.

Zhao X, Shi H, Sevilimedu A, Liachko N, Nelson HCM, Lis JT. 2006. An RNA aptamer that interferes with the DNA binding of the HSF transcription activator. *Nucleic Acids Res* **34**: 3755–61.

Zhou H, Huang C, Xia XG. 2008. A tightly regulated Pol III promoter for synthesis of miRNA genes in tandem. *Biochim Biophys Acta* **1779**: 773–9.

Zobeck KL, Buckley MS, Zipfel WR, Lis JT. 2010. Recruitment timing and dynamics of transcription factors at the Hsp70 loci in living cells. *Mol Cell* **40**: 965–75.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**: 3406–15.