How transcription factors define cell identity and p53 defends the genome

by

T. Jones

B.S., Wake Forest University, 2018

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Biochemistry

2023

Committee Members:

Dylan J. Taatjes, Chair

Robin D. Dowell

Jones, T. (Ph.D., Biochemistry)

How transcription factors define cell identity and p53 defends the genome

Thesis directed by Dylan J. Taatjes, PhD and Robin D. Dowell, DSc

The regulation of gene expression programs is essential for normal cellular function. These programs are fundamentally regulated at transcription and are orchestrated by sequence-specific transcription factors (TFs). The work presented here focuses on the wide-spread survey of TF activity, as well as an in depth study of a single TF, p53.

TFs function by binding DNA, and alter transcriptional programs through their effector domains. Despite the importance of TFs it remains a challenge to probe when and where TFs are actively regulating transcription as opposed to simply binding DNA. The focus of the first part of this thesis is to define a computational model to assess which TFs are actively regulating from a single nascent RNA-sequencing experiment. Using this model we built a framework in which TFs are categorized into distinct regulatory classes, cell type specific or shared across cell types. From this classification we are able to define distinct TF characteristics that may be related to their regulatory function.

In contrast, the second part of this thesis focuses on a single TF, p53. The p53 trans-activation domain (TAD; an effector domain) plays a major role in p53's ability to alter gene expression programs in response to cellular stressors such as DNA damage. In two separate studies we explore the function of the p53-TAD. First, we explore the inhibition of the p53 response induced by blocking the p53-TAD:Mediator interaction. Next, we explore the activity of the naturally occurring p53 isoform Δ40p53 in conjunction with WTp53 in a stoichiometrically controlled system. The Δ40p53 isoform is missing the first 40 amino acids that make up most of TAD1. In this we find that the Δ40p53:WTp53 isoform is broadly unable to activate transcription, but doesn't display a p53 null phenotype. Taken together, my body of work gives a deeper understanding of TF activity as a whole.

Dedication

To my family and closest friends for always believing in me.

To my mom, Kimberly Maples for always listening with an open heart.

To my dad, Rick Jones for always supporting and encouraging my dreams.

To my partner, Robbie Caron for always standing by me. Without you, this wouldn't have been possible.

I love you all dearly.

Acknowledgements

I've had the unique privilege of having not one, not two, but three advisors for my PhD.

Thank you to Dylan Taatjes for encouraging me to approach this work analytically. You have always pushed me to be my best.

Thank you to Mary Allen for teaching me to approach science rigorously. But also, thank you for inspiring me to approach science with a joyful curiosity and inquisitive spirit.

Thank you to Robin Dowell for being an incredible mentor and role model. You provided me with the support but also the independence (and the touch of snarkiness) I needed to grow into the scientist that I am today. Thank you for believing in me, guiding me and laughing with me.

Thank you to my committee members, Jennifer Kugel, John Rinn and Sabrina Spencer.

Thank you to the members of the Dowell and Taatjes labs. You all have made my experience incredible through the long conversations, the lively science discussions and most importantly, the fun! A big thank you to Dr. Cecilia Levandowski, my mentor upon joining the Dowell and Taatjes labs. She taught me that with finesse, attention and dedication even the most difficult experiments are possible.

Lastly, thank you to core facilities at CU Boulder. Specifically, thank you to Theresa and the cell culture facility, Matt Hynes-Grace and BioFrontiers IT and Amber Scott and Kevyn Jackson in the sequencing facility.

Contents

Chapter

Figures

Figure

Chapter 1

Introduction

It has been over twenty years since the first draft of the human genome was published in 2001 as part of the Human Genome Project. This massive public project incorporated the work of scientists from all over the world to draft the nearly 3 billion base pairs that make up the human genome[1]. About a decade after the conclusion of the Human Genome Project, Next Generation Sequencing (NGS) started to become widely available and with it came the power to address complex biological questions on a genome wide scale[53, 244]. With this technology at their fingertips, scientists began to develop new assays, tools and algorithms to better understand the basis of the regulation of the human genome. From the DNA code[162] to RNA transcription[135] to protein translation[37], scientists developed innovative ways to measure numerous steps of gene regulation.

One step of gene regulation of particular interest is a fundamental step in the central dogma, the process of transcribing DNA into RNA. Transcription in many ways is the basis of all gene regulation. Many cellular signalling cascades initiated by external stimuli end in the activation or repression of transcription [91, 98, 99]. Cell identity is defined by which enhancers and genes are transcribed within the cell[41]. Aberrant transcription often leads to detrimental diseases, such as cancer. In fact, mutations in transcription factors (TFs) are often drivers of oncogenesis[222].

In this introduction, I will discuss the basic regulatory processes surrounding transcription. There will be a particular emphasis on TFs, as they are the proteins that primarily direct transcription. Every aspect of my doctoral work has been touched by TFs in some way. Since much of

this work has revolved around various NGS assays, I will also briefly describe some key assays and analyses that appear throughout this thesis. Finally, an overview of one TF, p53 (gene: TP53) will be discussed as it is the focus of two manuscripts presented in this thesis (Chapters 3, 4).

1.1 Transcription

Fundamentally, transcription is the process of copying double stranded DNA into single stranded RNA. This RNA can then go on to play many distinct cellular roles, with both protein-coding and non-coding functions. Some types of RNA include messenger RNAs (mRNA) that later are translated into protein, enhancer RNAs (eRNA) which are short unstable RNAs associated with active enhancers, transfer RNAs (tRNA) which play a role in peptide synthesis and ribosomal RNAs (rRNA) that catalyzes peptide synthesis.

RNA Polymerase II (PolII) is responsible for the transcription of both mRNAs and eRNAs. PolII dependent transcription relies on a core set of general transcription factors (GTFs). These factors assemble on the DNA template at initiation regions to form the pre-initiation complex (PIC). This is followed by the three main phases of the transcription cycle: initiation, elongation and termination shown in Fig. 1.2. In this section I will provide an overview of the process of transcription, as well as discuss how we can measure transcription output genome wide.

1.1.1 The transcription cycle

1.1.1.1 The pre-initiation complex

There are several GTFs that play a fundamental role in transcription. These PolII associated GTFs include TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH and Mediator. These factors associate the promoter template DNA as shown in Figure 1.1. The PIC dictates the transcription of all coding and most non-coding RNA molecules. Briefly, TFIID binds the DNA upstream and downstream of the transcription start site inducing a bend in the DNA. TFIIA then binds stabilizing TFIID, followed by TFIIB which stabilizes the non-coding strand of the DNA to aid in maintaining the

open DNA confirmation. Next, TFIIF forms a protein bridge with TFIIB stabilizing the open DNA confirmation for PolII to bind. TFIIE interacts with PolII cleft and helps orient TFIIH binding. TFIIH is the ATP-dependent DNA translocase that fully opens the double stranded DNA for transcription to begin. Mediator acts as a scaffold for the other PIC factors and integrates signals from sequence-specific TFs bound at distal regulatory regions. I will focus in more detail on two members of the PIC, Mediator and TFIIH as my research directly pertains to these two GTFs.

Mediator is a 26-subunit protein complex that plays a major role in coordinating the PIC at the site of transcription initiation, such as a gene promoter. Mediator itself associates with DNA in a non-sequence specific manner, yet is capable of regulating transcription at specific initiation sites[202]. One mechanism in which this is possible is through the interaction with sequence-specific TFs[4]. Sequence-specific TFs will be discussed in more detail in section 1.2, but briefly TFs are able to establish favorable conditions for transcription, as well as interact with various Mediator subunits to stabilize the Mediator-PIC assembly at the promoter. This stable assembly of Mediator and the PIC promotes increased transcription output [43, 247]. Additionally, Meditor-TF interactions have been shown to induce conformational changes in Mediator that alter PolII activity[172]. Taken together, Mediator acts as a communication hub between TFs bound at distal regulatory regions (section 1.1.2) and the PIC to attenuate transcription output within the given cellular context[78, 79]. Blocking the interaction between Mediator and a sequence-specific TF, p53, is the focus of Chapter 3.

TFIIH is an essential 10-subunit complex that functions both in the nucleus for DNA repair and transcription activation, but also plays a role in cell cycle regulation[204]. The 7-subunits core functions as a $5'$-$3'$ DNA translocase, unwinding the DNA in an ATP-dependent manner to open the template for active PolII transcription[90]. The dissociable 3-subunit kinase module known as the CAK contains cyclin dependent kinase 7 (CDK7). CDK7 is known to play an important role in cell cycle regulation, management of the transcription cycle through the phosphorylation of the PolII CTD as well as a role in the recruitment of splicing factors for mRNA generation[203].

Assessing the impact of specific CDK7 inhibition on transcription and mRNA biogenesis is the focus of Appendix A.

1.1.1.2 Initiation

Once the PIC is assembled at the promoter, PolII is poised to initiate transcription. At this stage PolII will begin to transcribe either the sense or anti-sense strand of the open DNA template. In terms of genic transcription, only the sense DNA template results in transcription elongation and a viable mRNA molecule. The PIC functions to preferentially orient PolII relative to the sense strand of DNA. The anti-sense transcript will often terminate in a few hundred nucelotides[52, 220]. Both sense and anti-sense transcripts are capped, but the short anti-sense transcript is rapidly degraded. However, the presence of both sense and anti-sense transcription leads to a characteristic bidirectional signal at sites of transcription initiation. The site of transcription initiation will be used in subsequent computational inference of TF activity, or TF inference, discussed in more depth in section 1.2.3.

Once PolII has initiated transcription in the sense direction it quickly faces another regulatory checkpoint, the promoter-proximal pause. Approximately 60 base pairs downstream of initiation PolII pauses[2]. This pause is facilitated by TFIID and negative elongation factors NELF and DSIF[71, 179, 235, 250]. The phosphorylation of NELF and DSIF by PTEF-b within the super elongation complex (SEC) facilitates PolII pause release and permits transcription elongation[164].

1.1.1.3 Elongation

Elongation is the stage of transcription where PolII transcribes the entirety of the gene body. Many processes occur during this stage of transcription, such as co-transcriptional splicing. Splicing is the process of retaining protein coding regions of RNA (exons) and removing the non-coding portions of RNA (introns). Nearly all of the genes in the human genome have multiple splicing variants, leading to alternative isoforms that may be deferentially expressed based on the stage in development, cell type or external stimuli. These isoforms may take on context specific

functions to alter activity depending on the protein in question.

Splicing is a complex process dependent on many cofactors and is impacted by things like PolII transcription rate, however splicing is primarily regulated by the spliceosome[23, 165]. Two major components of the spliceosome SF3B1 and U2AF2 have been shown as direct targets of CDK7 phosphorylation. This implies that CDK7 plays a role in the regulation of spliceosome during elongation[203]. While PolII transcribes approximately 3kb/min, alterations of this rate can cause intron inclusion or exclusion [80].

1.1.1.4 Termination

The final stage of transcription is termination. Termination is where PolII ends transcription and releases the DNA. At the end of the gene, there is a cleavage and poly-adenylation signal (CPS, AAUAA). Once PolII transcribes this signal the cleavage and poly-adenylation complex (CAP) is recruited and the nascent RNA is cleaved and poly-adenylated completing the RNA maturation process. The mature mRNA is subsequently exported from the nucleus. The PolII enzyme however continues to transcribe the DNA template. The nascent RNA associated with PolII is uncapped following the cleavage, which allows the exonuclease Xrn2 to degrade the RNA attached to the elongating polymerase. As PolII slows, Xrn2 reaches PolII and dislodges it from the DNA template[50, 54].

This process results in transcription downstream of all coding genes. One observation in the literature is that some cellular perturbations such as heat shock and viral infection induce increased transcription of the downstream of cleavage RNA (transcriptional run-on) and lack of proper transcription termination[45]. It's unclear why this occurs, and how this may translate to the increased presence of down-stream of gene (DoG) content in the related mRNA molecules. The relationship of increased transcriptional run-on and mRNA containing DoGs is an area of active research.

1.1.2 Enhancers

Promoters are regulatory regions at the 5′ end of genes where the PIC assembles for gene transcription (section 1.1.1.2). Some sequence specific TFs also bind at the promoter, but many bind at intergenic, non-coding regions throughout the genome. TFs will be discussed in more depth in section 1.2, but by binding these regions the TF is able to alter transcription at genes sometimes hundreds of kilobases away. These regions were denoted enhancers for their ability to enhance genic transcription output[40].

The concept of enhancers has been around for over 40 years[21]. The most fundamental definition of an enhancer is that it drive gene expression regardless of position or orientation relative to the target gene. Over time, the definition has expanded to include various histone marks associated with enhancer regions[95]. Canonically, H3K27ac and H3Kme1/2 are associated with enhancer activity. Although, H3Kmethylation may simply be a read out of transcription output at a region. H3Kme3 is often associated with promoter regions, which tend to have a higher transcriptional output than enhancer regions[52]. Generally, multiple enhancers act in conjunction to drive the genic transcriptional response, and the loss or gain of a single enhancer is often not detrimental to gene expression[13, 182]. However, it's shown that a disproportionate number of disease associated single nucleotide polymorphism (SNPs) in intergenic regions are located within enhancer regions[153, 212].

1.1.3 Measuring transcription genome wide

Transcription is a complex, highly regulated process with distinct phases and many interesting features. This lead to the desire for assays that directly measure transcription genome wide[44, 245]. Protocols that achieve this all fundamentally do the same thing, they provide the actively transcribing PolII with a nucleotide variant that is incorporated into the nascent RNA and is used to isolate the nascent RNA molecules. Global Run-On Sequencing (GRO-seq) uses 5-bromouridine 5′-triphosphate[239], Transient Transcription Sequencing (TT-seq) uses 4-thiouridine[215] and Pre-

cision Nuclear Run-On Sequencing (PRO-seq) uses a Biotin-11-NTP[135, 160].

Here I will focus on PRO-seq, as it plays a pivotal role in most chapters in this thesis (Chapters 2, 4 and Appendix A). In PRO-seq, the first step is to isolate nuclei and perform a run-on experiment with a biotinylated-NTP. This biotinylated-NTP is incorporated into the nascent RNA by actively transcribing RNA polymerase and is then used specifically to pull-down nascent RNA through a biotin-streptavadin interaction. This nascent RNA can then be sequenced using NGS and mapped genome wide[135, 160].

PRO-seq yields direct information about transcription within the cellular context of the given experiment. At genic regions we observe evidence of all three main phases of transcription: initiation, elongation and termination. There is a prominent peak at promoter-proximal region at the $5'$ end of all transcribed genes. The promoter-proximal peak could be preset due to early abortive transcription or PolII pausing as previously described. Over the gene body we observe a steady elongation region, followed by a $3'$ termination peak downstream of the CPS (Fig. 1.3A). At the promoter proximal peak we also observe the short, unstable anti-sense transcript, resulting in the bidirectional signal. This bidirectional signal is a fundamental characteristic of polymerase initiation, as it is also observed at intergenic polymerase initiation sites associated with enhancers (Fig. 1.3B, section 1.1.2)[52]. While the function of enhancer associated RNA (eRNA) is incompletely understood, it has been shown that the presence of these transcripts correlates with transcription factor activity[17]. Furthermore, transcriptional output at non-coding enhancer regions is correlated to the transcriptional output at the gene target. A technical benefit of these transcripts is that their distinctive profile can be used to annotate active initiation regions genome wide[18, 240].

1.2 Sequence-specific transcription factors

Transcription is a fundamental process in cellular regulation. Importantly, it's TFs that orchestrate transcriptional programs. There are estimated to be at least 1,600 TFs in the human genome[138]. TFs generally fall into three important categories of gene regulation, 1) they maintain essential gene programs, 2) they drive cell-type specific features defining cell identity, and 3) they

control cellular response to external stimuli[140].

TFs typically consist of at least two key components: a DNA-binding domain (DBD) and an effector domain. The DBD is how the TF directly interacts with DNA in a sequence-specific manner[112]. The effector domain is how the bound TF relays information resulting in altered gene transcription[83]. In Figure 1.4A, I show the effector domains (TAD1 and TAD2) and DNA-binding domain (DBD) for p53. In this section I will discuss how TFs preferentially regulate distinct genomic regions, the role of TFs in gene regulation and how we can infer TF activity from a nascent RNA sequencing assay.

1.2.1 Sequence-specific DNA binding

There are two core classes of TFs, 1) TFs that increase transcription output (activator) or 2) decrease transcription output (repressor). Both classes can be become activated in a myriad of ways, such as post-translational modifications to the TF, induction of transcription of the TF gene or blocked TF protein degradation. Once the TF is activated it binds a specific DNA sequence through the DBD. Many active TF binding sites are at regions of PolII initiation, some preferentially at enhancers and others at promoters (section 1.1.2).

A common NGS assay to survey TF binding genome wide is chromatin-immunoprecipitation sequencing (ChIP-seq). In this assay the DNA is cross-linked to all DNA-associated proteins and is sheared into small fragments. Then the TF of interest is pull-down along with covalently linked DNA using an antibody. The DNA is then isolated, sequenced and mapped genome wide. The resulting data yields low level coverage genome wide with an enrichment of reads (peaks) where that TF was bound. ChIP-seq was performed in Chapters 3 and 4 of this thesis.

Based on ChIP-seq data, as well as complementary binding data such as SELEX and protein binding microarray (PBM) data, researchers are able to determine the "DNA code" that a each TF prefers. This code is commonly referred to as a TF motif. Sequence-specific TFs can bind range of sequences with varying affinities. Therefore, we often represent TF motifs as position specific scoring matrices (PSSMs). These matricies account for the probability of binding a given

nucleotide at a given position within the TF motif. Typically the TF motif will range from 5-30 nucleotides in length and will have a only few positions with a strong nucleotide preference. For example, in the p53 motif shown in Fig. 1.4B there is a strong preference for C at position 4 and G at position 7. Mutations in the p53 DBD that alter binding preferences have been shown to lead to gene dysregulation and often lead to serious disease states in humans, like cancer[22].

TFs are essential for targeted gene regulation, and where they bind determines which genes they regulate. For this reason there is desire in the field to robustly annotate the recognition motif for all TFs in the human genome[138]. One of these databases is HOmo sapiens COmprehensive MOdel COllection (HOCOMOCO)[134]. HOCOMOCO currently has combined 680 annotated TF motifs. One use of PSSMs is to identify sites in the genome that could be bound by the TF. This is usually done by scanning the DNA for matches to the PSSM given a chosen statistical cut-off. Typically, there are far more potential TF sites (sequence matches) than observed binding sites in a ChIP-seq experiment. However, even in a ChIP-seq experiment the presence of a sequence motif can be used to focus researchers on direct versus indirect ChIP peaks. The use of TF sequence motifs in the context of TF inference will be discussed in more depth in section 1.2.3, as it's a key concept in Chapter 2.

1.2.2 TF activation

Once bound and activated the TF goes on to alter gene regulation programs primarily through the effector domain. Effector domains (trans-activation domains in activators, repressive domains in repressors) are often intrinsically disorder regions (IDRs) that interact with transcriptional co-factors to induce transcription changes[74]. Co-factors have many distinct functional classes and activities such as acetyltransferases, H3K4-methyltransferases, bromodomain co-factors and Mediator[92]. The recruitment of co-factors results in many distinct outcomes dependent on the identify of both the TF and the co-factor.

One resulting outcome of TF binding and co-factor recruitment is the reordering of the chromatin landscape. Most of the genome is heterochromatin which is tightly packed and not

transcribed. PIC assembly depends on open and accessible chromatin which can be facilitated by TFs. This aspect of TF activity is commonly referred to as "pioneer" activity[116, 166]. One hypothesis explored in this work is whether pioneer activity is disproportionately observed in cell-type specific TFs (Chapter 2).

1.2.3 Inferring TF activity

The computational inference of active TF regulation (TF inference) integrates the two key pieces of data previously discussed, 1) the characteristic bidirectional signal that enables the precise mapping of PolII initiation genome wide (sections 1.1.1.2,1.1.3) and the use of TF motifs as a proxy for TF binding (section 1.2.1). TFs bind near PolII initiation regions (enhancers and/or promoters) to regulate transcription. By measuring the co-localization of the active regions of PolII initiation and the TF motif instances within these regions we can infer which TFs (as long as we have an annotated PSSM) are active from a single nascent RNA sequencing experiment[17, 207].

Active TFs tend to co-occur with within 100bp of the PolII initiation site (μ) as shown in Fig. 1.3B. One way to score motif displacement (MD-score) from PolII initiation is to simply ask, across all enhancer and promoter regions in a given experiment, does a TF motif fall within one nucleosome distance (150bp) of initiation more than by random chance. This can be calculated with the simple formula[17]:

$$\text{MD-score} = \frac{\text{Hits across} \pm 150\text{bp of } \mu}{\text{Hits across} \pm 1500\text{bp of } \mu} \tag{1.1}$$

Assuming random sequence at the genome composition, an inactive TF would be randomly distributed relative to μ and result in a MD-score of 0.1. A transcription activator currently participating in regulation would have an enrichment of motif hits surrounding μ compared to the flanking region resulting in a score between 0.1-1. An active transcriptional repressor would have a depletion of hits surrounding μ and a score between 0-0.1. These scores are represented in Fig. 1.5A as heat maps, where heat indicates the number of motif hits in that position relative to μ.

The shift in motif instance distribution between two conditions can be used to infer changes in TF activity between conditions[207], such as p53 activation upon treatment with Nutlin-3a (discussed further in section 1.3, Fig. 1.5B). In this work, I use this conceptual framework to determine all TFs that are basally activated and repressed in any cellular context. The focus of Chapter 2 is centered on statistically defining the TFs that drive cell identity and exploring their specific properties and preferences.

1.3 p53: The guardian of the genome

To this point I have discussed transcription and TFs generally, here I will focus in on a single TF, p53. p53 may be the most famous transcription factor due to it's role in cancer biology, cell cycle control, aging and development[121, 122, 129, 132]. Mutations in the TP53 gene are the most common mutations in human cancer[82, 106]. In 1992 p53 was named the guardian of the genome for its ability to mobilize a context appropriate response to DNA damage. When DNA damage is detected, p53 is mobilized to enhancers genome wide to induce cell cycle arrest through activating the gene CDKN1A (p21)[249]. This gives other cellular machinery time to repair the DNA. If the DNA repair fails then p53 can trigger cellular senescence or even apoptosis[6, 9, 139, 208]. Beyond that, p53 is able to activate in response to numerous cellular stressors and subsequently manage the appropriate cellular response (Fig. 1.6). Due to the important role p53 plays in human health, it's an area of great interest to understand it's complex regulatory function. Both Chapters 3 and 4 of this thesis focus on different aspects of p53 function to better understand its role as a transcriptional activator.

1.3.1 Negative regulation of p53 by HDM2

It's important for p53 function that the p53 protein is ready to rapidly respond to DNA damage, but also not perpetually active as that would induce apoptosis. Cells have evolved to constantly express p53, translate the protein and degrade it before it takes action in the nucleus. In a tightly controlled process, HDM2 and HDMX (also denoted MDM2 and MDMX) control

the degradation rate of p53 (Fig. 1.7). HDM2 is an E3 ubiquitin ligase that targets the first trans-activation domain (TAD1) of p53[100, 108, 133] (Fig. 1.4A). This leads to the constant degradation of p53 in the cell until a p53 stimulus occurs, such DNA damage. In response to a myriad of cellular stressors stabilization of p53 via the release of p53 from HDM2. p53 then binds DNA, recruits transcriptional co-factors and activates transcription[111, 236]. A small molecule, Nutlin-3a, was synthesized to disrupt the p53:HDM2 interface with high specificity resulting in p53 activation[229]. This has allowed us to study the p53 response directly without confounding factors of other DNA damage responding TFs[6, 9].

1.3.2 p53 in gene regulation

The p53 TF binds DNA and regulates transcription as a homotetramer consisting of two p53 dimers[186]. The p53 DNA recognition motif is 20 nucleotides long consisting of two 10 nucleotide half sites (Fig. 1.4B), each bound by a homodimer within the functional tetramer. Once bound, the TAD1 of p53 is required for the recruitment of p300 and Mediator[89, 115, 118, 119, 148] (Fig. 1.4A). This recruitment is directly related to p53s abilities as a transcriptional activator[254]. The specific inhibition of TAD1:Mediator is discussed in Chapter 3.

While p53 induces differing cellular signalling cascades in different cell types, the core transcriptional program of approximately 100 direct gene targets remains the same. These genes have diverse biological functions that include inducing cell cycle arrest (CDKN1A, BTG2), DNA-repair (DDB2, XPC), autophagy (DRAM1, SESN1) and apoptosis (BAX, PHLDA3)[6, 9]. Despite regulating the same core set of genes, evidence shows that distinct sets of enhancers control these target genes in different cell types. Current work in progress will provide additional information on how stress-responsive TFs alter their behavior in distinct chromatin context but yield the same genic transcriptional response[218].

1.3.3 p53 alternative isoforms

The primary role of p53 is maintaining the integrity of the genome. However, different cells will have different needs in regard to DNA damage response. One way p53 activity is modulated across cell types and development is through the expression of alternative isoforms[30]. There are twelve distinct p53 isoforms identified in the literature. There are also two additional p53 family members, p63 and p73. These proteins also have many viable isoforms and bind similar DNA recognition motifs to p53, which may alter p53 function[30, 123, 246].

The p53 isoform of particular interest is Δ40p53. In terms of transcription regulation Δ40p53 is missing the first 40 amino acids of p53 which contains most of TAD1. The Δ40p53 however retains the oligomerization domain and the DBD. TAD1 of p53 has been designated the primary transactivation domain in p53 regulation. When Δ40p53 is expressed alone in mice, the mice mimic a p53 null phenotype. However, when this isoform is expressed in conjunction with full length p53, mice undergo a rapid aging phenotype[189]. The Δ40p53 is important to study due to its role in p53 regulation, aging and development. Chapter 4 covers the thorough transcriptional analysis of the Δ40p53 isoform in conjunction with the full length p53 isoform.

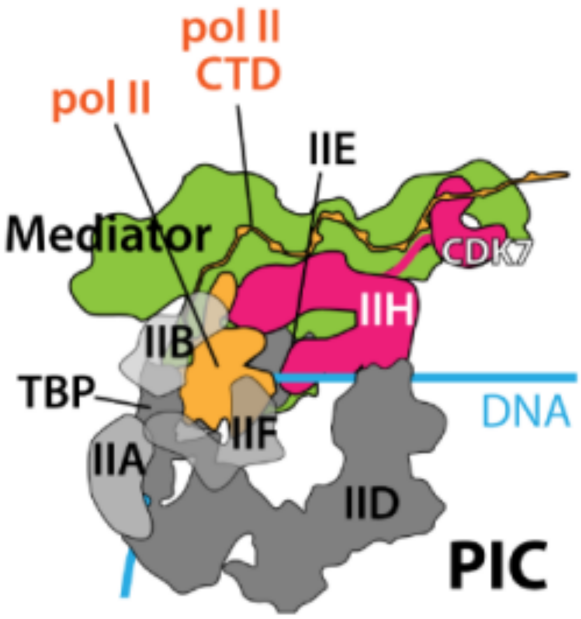


Figure 1.1: **Illustration of components assembled at initiation as part of the pre-initiation complex.** The factors shown assembled for initiation are: TFIID, dark grey; TFIIA,TFIIB, TFIIF and TFIIE, grey; PolII, orange; TFIIH, pink; Mediator, green; DNA, blue. Adapted from Rimel and Taatjes 2018[204].

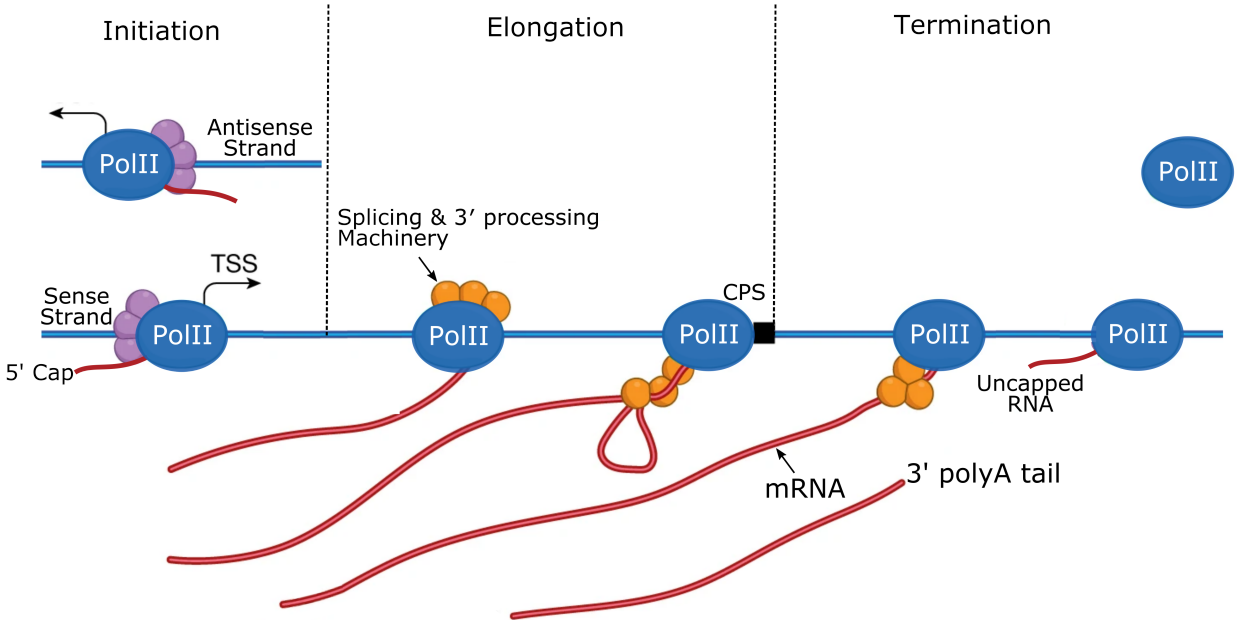


Figure 1.2: **Overview of transcription.** Diagram of the phases of transcription, adapted from Cramer 2019[56]. Initiation (left) leads to bidirectional transcription (RNA in red) as RNA polymerase II (blue circle) can initiate on either strand of DNA (blue line) with the aid of the pre-initiation complex (purple circles). During elongation (middle), RNA polymerase II proceeds through the gene, recruiting splicing and 3′ processing machinery within the gene body (orange circles). After the cleavage site (labeled CPS), RNA polymerase II enters the termination phase (right) where the nascent RNA is cleaved and poly-adenylated. RNA polymerase continues to transcribe past cleavage event before dissociating with the DNA template.

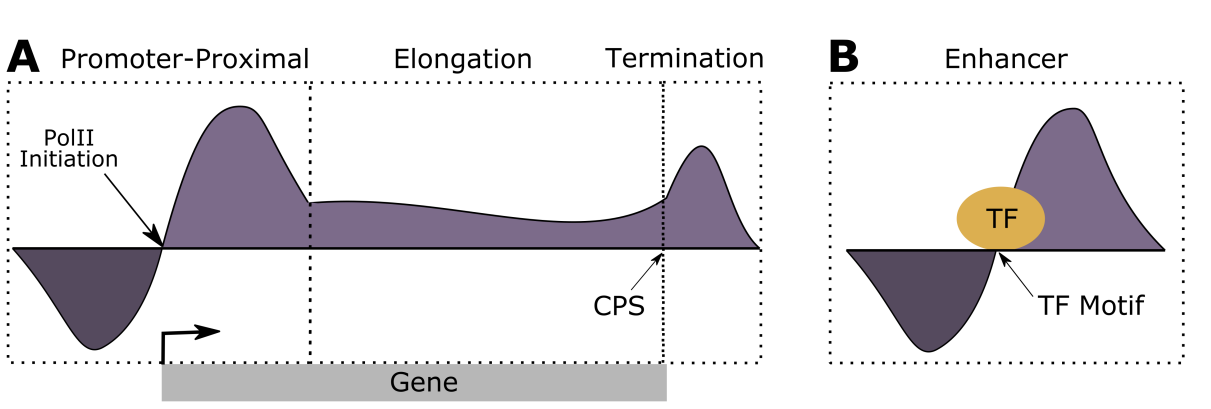


Figure 1.3: **Illustration of PRO-seq data.** A) Illustration of PRO-seq trace at a gene and B) at an enhancer. The positive strand signal is represented in purple, the negative strand signal is represented in dark purple, an example TF protein binding DNA at the region of PolII Initiation is in gold.

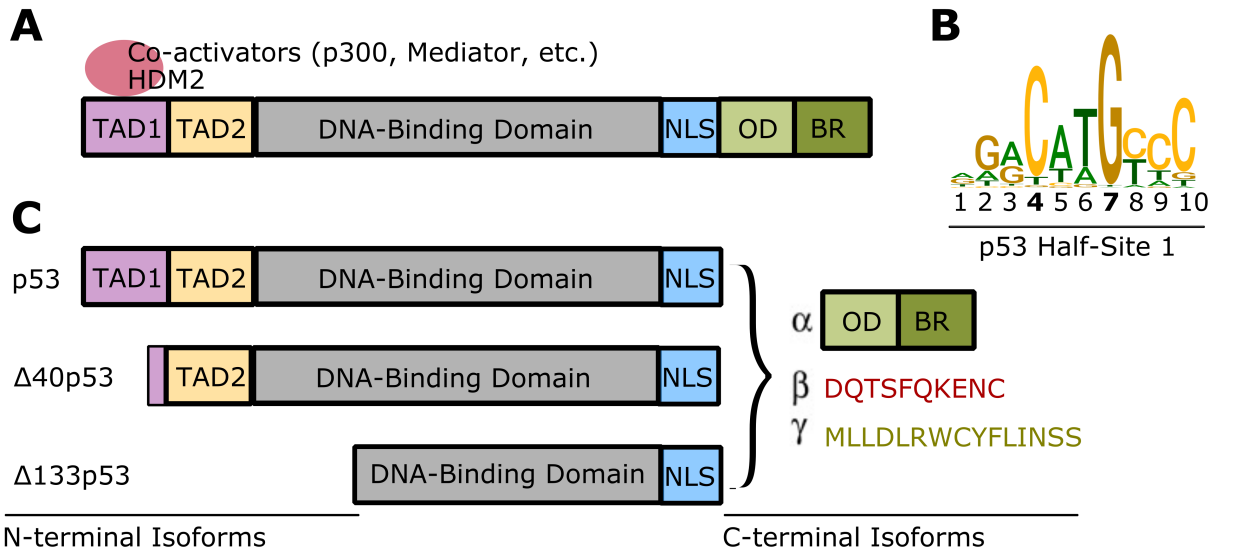


Figure 1.4: **Important TF domains and a DNA recognition motif represented with p53.** A) Example of a TF that contains a trans-activation domain and a DBD. Shown here are the protein domains of p53. p53 has two trans-activation domains (TADs, purple and orange), a DNA-binding domain (grey), a nuclear localization signal (NLS, light blue), a oligomerization domain (OD, light green) and a basic region (BR, dark green). General transcriptional co-factors and HDM2 are represented in pink. B) The first half-site of the p53 DNA recognition motif, shown as a position specific scoring matrix. The p53 motif is composed of two nearly identical 10 nucleotide half-sites. C) Schematic showing the possible isoforms of p53. Adapted from Khoury and Bourdon 2011[123].

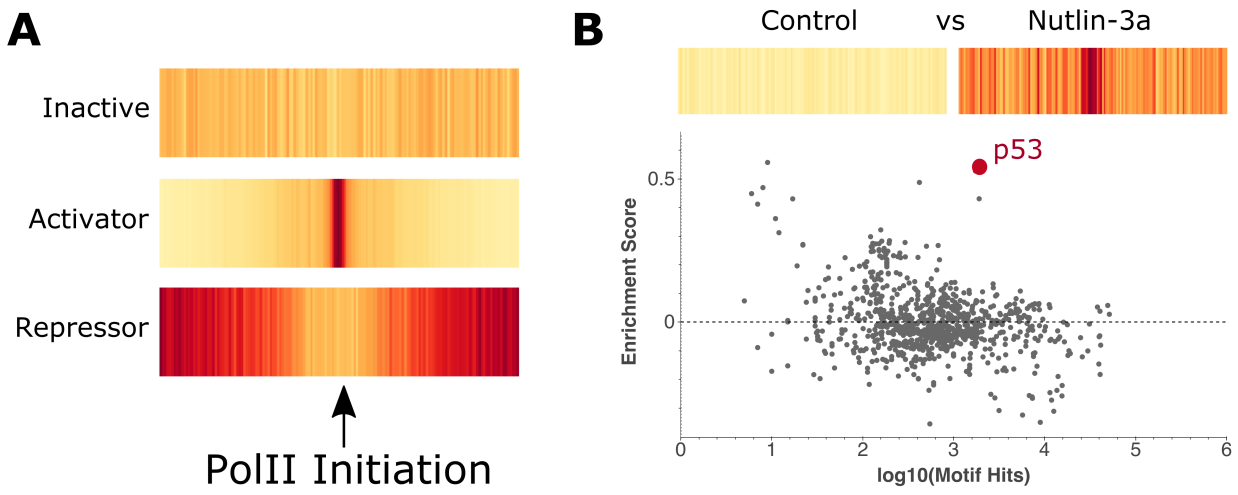


Figure 1.5: **TF Inference.** A) Example heatmaps of an inactive TF, an activator and a repressor. The center of each heatmap is the annotated PolII initiation region (μ) and the heat (red is higher) is the number of motif hits for that TF within 100bp bins across a 3000bp region. B) Additional heatmaps showing significant p53 enrichment after Nutlin-3a activation. TFEA MA plot where each dot represents a single TF with p53 shown in red. The x-axis shows total TF motif hits and the y-axis shows a measure of TF enrichment accounting for the co-localization of a TF motif with PolII initiation regions genome wide demonstrating the ability to infer p53 activity in p53 stimulated versus control conditions[6, 142, 207].

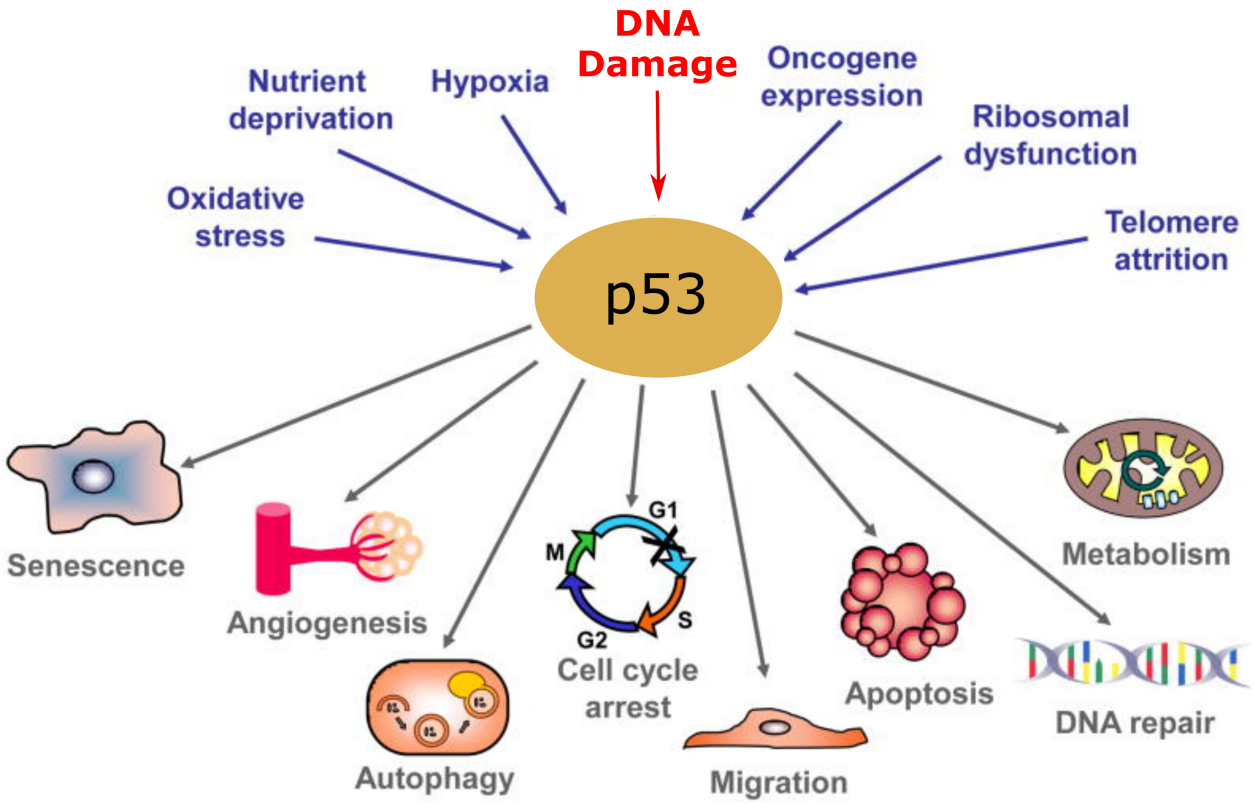


Figure 1.6: **p53 is an important stress response transcription factor.** p53 is able to integrate a number of cellular stresses and orchestrate the appropriate cellular response. Schematic of the roles of p53 adapted from Bieging and Attardi 2011[25].

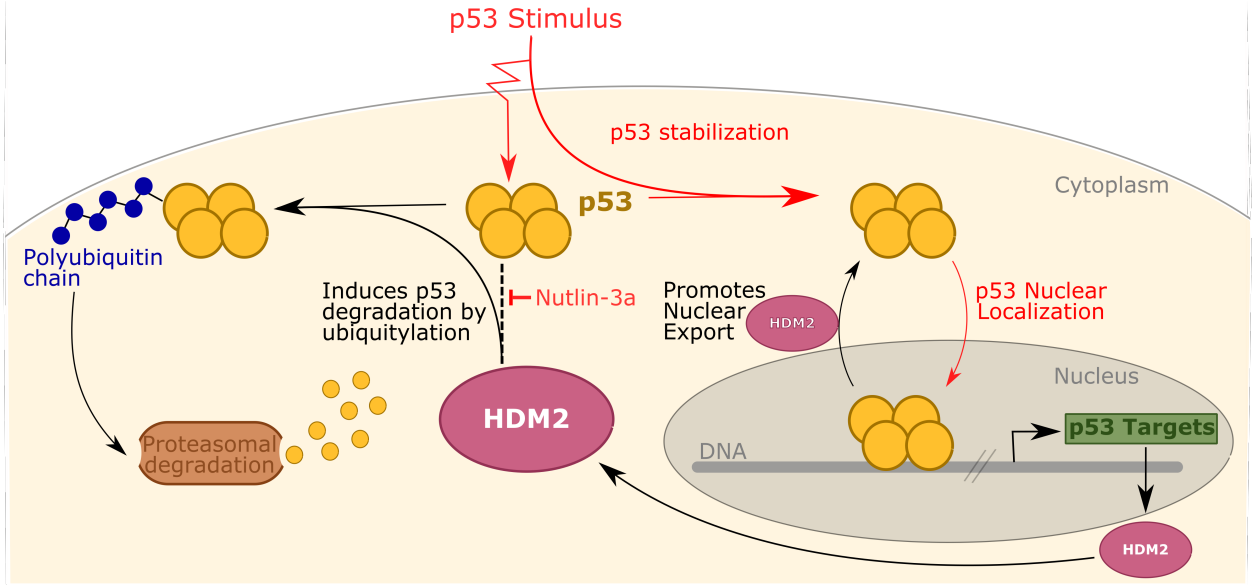


Figure 1.7: **p53 is negatively regulated by HDM2.** Diagram of p53 activation in response to a p53 stimulus. Various p53 stimuli are summarized in Figure 1.6. In unstressed cellular conditions HDM2 (magenta) constantly ubiquinates p53 (gold) and promotes nuclear export. Once polyubiquinated (blue), p53 is shuttled to the proteosome (brown) for degradation. Upon p53 stimulus or treatment with the small molecule, Nutlin-3a (red), the p53:HDM2 interface is blocked allowing p53 to stabilize and localize to the nucleus. In the nucleus p53 binds DNA (dark grey) and promotes the p53 gene regulation program (green). One target for up-regulation is HDM2 itself creating a negative feedback loop. Figure adapted from Chene 2003[47].

Chapter 2

Transcription factors display distinct localization preferences and regulation mechanisms based on function

The work in this chapter is currently ongoing. The code used to generate the TF profiles can be found at `https://github.com/Dowell-Lab/TF_profiler`. Methods are discussed in Appendix B.1.

2.1 Contribution Statement

I preformed the majority of the work in this Chapter. The curation of the nascent sequencing repository (DBNascent) was performed by the Dowell lab members in the summer of 2020. The two primary contributors to subsequent DBNascent data processing which includes mapping, quality control, gene counting and bidirectional calling are Rutendo Sigauke and Lynn Sanford. Work on the DBNascent publication is currently in progress[218].

2.2 Abstract

Transcription factors orchestrate transcription and play a critical role in cellular maintenance, identity and response to external stimuli. Despite the importance of transcription factors it remains a challenge to probe when and where transcription factors are actively regulating transcription. Here we define a computational model to assess which transcription factors are actively regulating from a single nascent RNA-sequencing experiment. We apply our model across 89 published papers to classify TF profiles of 82 distinct cell lines. From this classification we find that transcription

factors that drive cell identity bind enhancer regions with a base composition bias close to genomic background. We postulate that cell type specific transcription factors play an important role not only in driving cell identity but establishing cell type specific enhancers.

2.3 Introduction

Transcription is a fundamental process in defining cellular function, stress response and cell identity[140]. The regulation of gene expression patterns is driven by myriad of sequence-specific transcription factors (TFs) that vary in activity based on both cell type and environmental factors. While there are over 1,600 TFs[138] in the human genome, we don't have a consensus on when or where individual regulatory factors are actively altering gene expression patterns.

Transcription factors orchestrate gene regulation programs by altering the activity of cellular RNA polymerases, primarily RNA polymerase II. Some TFs induce increased transcriptional output (an activator) whereas others decrease transcriptional output (a repressor). Therefore, characterizing when and where TFs are active is necessary to understand the regulatory process. In fact, one of the goals of Encyclopedia of DNA Elements (ENCODE) Consortium was to identify all functional regulatory elements in the human genome[51]. In the ENCODE project, the primary method utilized to assess TF activity was chromatin immunopreciptation assays (ChIP-seq). ChIP informs on the localization of a protein to specific regions within the genome. Functionally, ChIP-seq measures the DNA binding domain of a protein, which is the domain the TF utilizes to directly interact with DNA in a sequence-specific manner[112]. From ChIP-seq studies, it is possible to infer a position specific scoring matrix (PSSM) for a given DNA binding protein. ChIP-seq studies, however, are low throughput in terms of proteins assayed as one sequence-specific protein is evaluated at a time, in one cell type at a time. Furthermore, there is ample evidence of TF binding can occur without altering gene expression[34, 146], as the DNA binding domain is independent of the effector domain. The effector domain is how the TF relays information resulting in altered gene transcription thus is important to study due to it's crucial role in transcriptional regulation[83].

However, measuring the activity of the effector domain – i.e. assaying TF activity has historically been difficult. Part of the difficulty comes from the fact that TF activity can be regulated at multiple stages in the life cycle of a protein. For example, regulation may occur at the level of transcription, translation, post-translational modifications and degradation. Most TFs have their own mechanism of activation, such as MAPK pathway phosphorylation events result in stabilization and activation of MYC [209], or the inhibition of the ubiquitin ligase HDM2 resulting in the stabilization and activation of p53[100, 108, 133]. In these two cases, the MYC and TP53 genes are transcribed and the mRNA and proteins are present in most cellular conditions, despite being inactive until certain stimuli impact the cell. Thus, neither transcription of a TF encoding gene, nor binding of a TF to DNA guarantees it will alter transcription. As the ultimate outcome of the effector domain activity is a change in transcription, nascent transcription assays have the potential to inform on effector domain activity.

Nascent RNA sequencing (such as precision run-on sequencing, PRO-seq)[135, 160]) provides a direct read out of RNA polymerase activity as RNA is captured from the actively catalyzing cellular polymerases. These assays revealed extensive genome-wide transcription, at both genes and enhancers[18, 53, 126, 215]. While the function of bidirectional transcription within enhancers is incompletely understood, a technical benefit of these transcripts is that their distinctive profile can be used to annotate active enhancers genome wide[18, 58].

Studies on individual transcription factors found that activation of a TF resulted in concomitant changes in transcription levels associated with subset of TF binding sites (as measured by ChIP)[6, 94, 97, 156, 195]. Subsequent work generalized these findings, showing a strong co-association of TF binding sites with sites of RNA polymerase II initiation[17, 58]. The model that emerged was that the regulatory activity of the TF (e.g. activity of the effector domain) results in changes to RNA Pol II initiation immediately proximal to the TF sequence motif, within a measurable (by ChIP) TF binding site[17]. Armed with this result, methods were developed to infer changes in TF activity arising in response to a perturbation from differential nascent transcription data and sequence motifs[17, 207]. The effectiveness of these methods indicates that nascent

transcription can serve as a functional readout on the activity of a TF's effector domain.

Here we sought to develop an algorithm for predicting TF activity from a single nascent transcription experiment. To that end, we develop a statistical framework for comparing an individual nascent transcription assay output to a principled, biologically informed statistical expectation. When a TF recognition motif co-localizes with sites of RNA polymerase II more (or less) than expected by chance, we infer that the TF is actively participating in regulation as an activator (or repressor). Our algorithm can be used to identify all the transcription factors currently altering transcription in a single sample, which we call TF profiling. Moreover, we applied our algorithm to 287 high quality nascent RNA sequencing human data sets across 82 unique cell lines. From this compendium we identify three classes of transcription factors: ubiquitous transcription factors, cell type specific transcription factors and environmentally responsive transcription factors. Our method correctly classifies well know TFs, such as Oct4 and Nanog, as only having activity in stem cells. Furthermore, our model identifies unique sequence features inherent to cell type specific TFs, suggesting a model for the establishment of cell type identity.

2.4 Results

2.4.1 An expectation model for TF motif co-occurrences

Our goal is to predict which TFs are actively participating in regulation of nascent transcription. This work builds on the original paper for identifying differential TF activity by Azofeifa et. al. from the Dowell lab[17], so let me first describe a brief description of that effort. For each data set, first sites of Pol II initiation (μ) are identified using the Tfit algorithm[18]. These sites are collectively known as bidirectionals and are located throughout the genome including at gene promoters, enhancers, and within introns. Using the bidirectionals, a simple metric known as the motif displacement (MD) score then calculated. The MD-score quantifies co-localization of TF recognition motifs in DNA sequence (significant hits to the PSSM) with sites of RNA polymerase II initiation. Briefly, the MD-score is the ratio of TF motif hits within $h$bp of μ over the TF

motif hits within Hbps of μ. In the original paper, h was set to 150bps, roughly the size of a single nucleosome. The denominator length, H was set at 1500bps as this length was long enough to reach genome background frequency of nucleotide frequencies. The bulk of the Azofeifa work was on using this between samples to identify TF activity changes in response to a perturbation, a situation where the background motif hit bias (effectively the denominator) was of minimal impact.

However, Azofeifa did briefly examine the MD-score within a single data set as an effort to validate the metric. Azofeifa noted that enhancers have heightened GC-content and therefore developed a simple simulation based method of examining enrichment of a motif hit relative to the collective set of μ (sites of Pol II initiation). For his simulated sequences, Azofeifa used a 4xH matrix of nucleotide frequencies, effectively capturing the positional nucleotide bias of bidirectionals around μ. Simulated sequences were then used to assess the background expectation of a particular motif hit relative to the experimentally observed data. An examination of six cell lines identified known cell type specific TFs active in the appropriate experiments. A larger examination of 491 experiments (34 cell types and 205 treatments) appeared to cluster predominantly by cell type, further validating the MD-score metric as informative.

Notably, the Azofeifa clustering result suggests that a comparison of MD-scores between an experiment and an expected statistical background has the potential to identify which TFs were actively participating in regulation in every sample. Yet Azofeifa used this result only to validate the MD-score metric and not to examine TF activity within individual samples. There were a number of reasons the Azofeifa approach fell short of developing an algorithm for inferring TF regulatory activity. First, the non-stationary background distribution utilized for the simulations fails to capture known dinucleotide biases inherently present in regulatory regions, such as CpG bias. Second, the non-stationary distribution defining background was trained on all bidirectionals, including both enhancers and promoters. While both are biased, we now know that promoters are exceptionally GC rich whereas enhancers show only a mild heightened GC content relative to genomic background (Fig. 2.1C). The bias at promoters and enhancers are fundamentally and significantly (KS Test; p-value=$6.8e^{-4}$, Fig.2.2B) different. Finally, Azofeia used this model on

a large compendia of data to generate the clustering. However, no quality assessment on those data sets was conducted prior to the analysis. Hence while the overall patterns were suggestive, individual TFs did not show high fidelity across experiments. In other words, in the work of Azofeifa, a given TF may appear active in one set of HCT116 control cells but inactive in experiments from a different paper on HCT116 control cells.

Here we seek to develop a robust algorithm for evaluating TF activity from a single nascent transcription experiment. The goal is, for every TF motif within a data base, to ask whether the patterns of co-localization of the motif hits with Pol II initiation deviate from expectation. Deviations from expectation are thus candidate TFs with functionally active effector domains that are actively participating in regulation.

First lets consider the original MD-score metric in a more rigorous mathematical framework. Let $X_i = \mu 1, \mu 2, \ldots$ be the Pol II initiation sites μ for a set of bidirectional locations genome-wide for some experiment i. Let $Y_j = \text{y1,y2,}\ldots$ be the set of all significant motif instances for some TF-DNA binding motif model j genome-wide, which is invariant given the genome of interest. Examples of the distribution of motif hits relative to X_i are given in Figure 2.1B as heatmaps, where heat indicates the number of motif hits in that position relative to μ. We quantify the co-localization of sequence motif hits with Pol II initiation as:

$$\begin{aligned} g(X_i, Y_j; a) &= \sum_{x \in X_i} \sum_{y \in Y_j} \delta(|x - y| < a) \\ md_{i,j} &= g(X_i, Y_j; h) / g(X_i, Y_j; H) \end{aligned} \tag{2.1}$$

Here, $\delta(.)$ is a simple indicator function that returns one if the condition (.) evaluates true and zero if false. The double sum, i.e. $g()$, returns the count of motif hits for a given motif j across the complete set of bidirectional initiation sites X_i in one experiment i that are within a given distance a. Hence the MD-score ($md_{i.j}$) for a given experiment i and TF recognition motif j quantifies co-localization of motif instances near sites of Pol II initiation (h = 150bps) relative to a larger local window (H = 1500bp). Importantly, the use of enhancers defined by nascent RNA-sequencing is essential for this score due to the precision on the position of initiation when compared to enhancers

defined by H3K27ac ChIP-seq (Fig. 2.3A-C).

The motif displacement score (MD-score) gives a starting estimate of TF activity. If the nucleotide distribution within the genome were random, a TF which is not present (off) would score 0.1 whereas an activator would show enrichment around μ and have scores > 0.1 and repressors would show depletion around μ and have scores < 0.1. Genomes, however, how distinct non-stationary patterns including GC-content enrichment at promoters[51, 72] and enhancers[17]. As discussed previously, the simple positional bias model of Azofeifa is also inaccurate as it does not account for the fundamental genomic biases at PolII initiation sites. Gene promoters are highly enriched for CpG islands and are associated with open chromatin[10, 26, 113]. Where the human genome is approximately 60% composed of AT, promoters are approximately 60% GC immediately around the site of Pol II intiation μ (Figure 2.1C). Enhancer initiation regions are also GC rich compared to the genomic background, but more modestly – reaching an equal composition of all four bases (approximately 50% GC, 2.1C) near μ. Because of the fundamental base composition bias at initiation regions, certain motif instances, especially for those motifs that are very AT or very GC rich, will be strongly impacted resulting in many false negatives and positives respectively.

To improve on the sampling model of Azofeifa, we leverage a dinucleotide model of positional nucleotide preference (Figure 2.1D), which better accounts for known biological dinucleotide biases, such as the general preference for CG in CpG islands compared to GC (Fig. 2.2A). To this end, sequences of the length of $2H$ nucleotides were generated accounting for dinucleotide preferences in regions of initiation. Importantly, the positions i are defined relative to μ (e.g. the generated sequence is $\mu\pm$ H). Let $x_n = x_1,x_2,\ldots x_{2H}$ where the probability of a specific nucleotide at each x_i is determined based on the nucleotide x_{n-1}. Thus each position is described by the conditional probability $p(N_{x_i}|N_{x_{i-1}})$, where N represents one of the four standard nucleotides (A, T, C or G). The initial dinucleotide x_1x_2 is calculated as $p(N_1, N_2)$ and all subsequent positions are based on

the conditional probability of the previous position. Therefore, we generate the sequences as:

$$\begin{aligned} x_1 x_2 &= p(N_1, N_2) \\ x_i &= p(N_i | N_{i-1}) \text{ for } i > 2 \end{aligned} \tag{2.2}$$

The model was parameterized using a training data set built from a master list of all control data sets within DBNascent (See Methods B.1) and gives results as shown in Figure 2.2A. Using this model, we simulate promoters and enhancers separately to a depth of 10^6 sequences each using the conditional nucleotide probabilities. Across data sets (SRRs with combined biological replicates) the promoter and enhancer content varies (Fig. 2.7A). To account for the variability in fraction of initiation regions that are promoters (vs enhancers), the simulated initiation regions were subset to an equal proportion of promoter:enhancers of each individual data set. A total of 10^6 sequences with the representative promoter:enhancer ratio were then used as the background model. The 10^6 simulated sequences (with a promoter:enhancer ratio that matches the data set) were then used to calculate the expected MD-score for each TF-motif. The MD-score was calculated based on equation 2.1, assuming for each simulated sequence a μ_e at position x_H.

More rigorously, let $Y_f = y_1, y_2, \ldots$ be the set of all significant motif instances scanned against the given set of simulated sequences. We quantify the co-localization of sequence motif hits within the simulated background model as:

$$\begin{aligned} g(X_e, Y_f; a) &= \sum_{x \in X_e} \sum_{y \in Y_f} \delta(|x - y| < a) \\ md_{e,f} &= g(X_e, Y_f; h) / g(X_e, Y_f; H) \end{aligned} \tag{2.3}$$

A plot of the expected (i.e. model derived, x-axis) to observed (i.e. experimentally observed, y-axis) MD-score for a single experiment is shown in Figure 2.5A. Thus, the expectation model is calculated on a per data set basis to accurately reflect the composition of initiation regions. By taking these steps, we developed a more robust expectation model in terms of relative motif hits, MD-score variance and biological relevance.

2.4.2 Building a TF activity profile

The next step is to assess the statistical significance of TF activity for each TF-motif, i.e. to ask which motifs in Figure 2.5A are significantly more (or less) co-localized than our background model suggests (e.g. than expected). TFs with greater (or less than) expected co-localization are the TFs which we infer as ON (ON-UP and ON-DOWN, respectively) and participating actively in regulation. To this end, the MD-scores for all HOCOMOCO TF motifs (n=388 for TF motifs that have instances near sites of initiation) are compared in two dimensions for all control data sets (n=126), resulting in (n=48,888) points. We fundamentally assume that a portion of TFs will be OFF (not significantly different from expectation) in each data set.

To assess whether the MD-score is higher (ON-UP) or lower (ON-DOWN) than expected, we fit a set proportion of inlier TF MD-scores to a linear equation shown in Figure 2.4A as the purple points. This results in a slope of 1.0 and an intercept of 0.0, with a normal distribution of residuals centered at 0.0 (Fig. 2.4B). The distribution of residuals is used to attribute significance values to all TF motifs within all control data sets (n=48,888). This results in a range of 80-164 TFs that are ON in any given data set (mean=123.5, p-value < 0.05). The ON TFs can be split into two categories, on and up (ON-UP, activators; range of 74-148, mean=109.9) or on and down (ON-DOWN, repressors; range 5-28, mean=13.6) shown in Figure 2.4C. This processes is repeated for all curated perturbation data sets shown in Figure 2.4D-F. There are fewer ON-DOWN TFs compared to ON-UP for two possible reasons. First, there may be fewer functional repressors than activators active within the cell at a given time. However, it may also reflect technical limitations on our ability to call repressors using the MD-score approach (equations 2.1,2.3). The MD-score approach is limited due to the inherently compressed numerical range (0-0.1) of the metric. Ultimately, in a given cell line we define approximately 125 TFs as active that play roles in 1) maintaining cellular homeostasis and 2) aiding in defining cell identity.

Hence, we can thereby classify each TF-motif in each experiment as ON-UP, OFF, or ON-DOWN based on the observed co-localization of the motif with sites of RNA polymearse initiation

and whether it differs from expectation. We refer to the collection of ON TFs for a given cell line as its TF activity profile. We apply this approach to over 70 cell lines with representative nascent RNA sequencing data. In doing this we will define the core set of TFs that drive cell identity in each cell line out of the approximately 400 TFs in the human genome with robustly defined motifs[134]. An example of this classification is shown in Figure 2.5A, where red and blue represent TFs that are classified as ON-UP (red) and ON-DOWN (blue) within the TF profile and grey represents those that are OFF.

Some general observations on the identity of ON and OFF TFs is notable. First, some factors implicated in general cellular processes and proliferation are commonly called ON-UP in the TF profiles including members of the ATF family, ETS family and KLF/SP family. Second, when applied to an ESC data set (n=3 biological replicates) at baseline conditions[219], we call 95 enriched and 9 depleted TFs. Enriched TFs in this profile include the pluripotent factors responsible for ESC self-renewal, Oct4 (pval=$1.3e^{-5}$), Nanog (pval=$2.7e^{-5}$) and SRY-Box Transcription Factors 3 and 4 (pval=0.008, pval=0.03 respectively). Across 3 additional ESC data sets[32, 68, 238] the same pluripotency factors are consistently called as active. Third, the TF MyoD, which is not associated with embryogenesis but is a strong determinate in muscle[59] is not significant in any of the ESC samples. However, when we look at a myoblast data set [141] MyoD is highly enriched (pval= $7.0e^{-9}$). Finally, the blood associated factor, GATA-2, is not significant in the ESC nor the myoblast samples, but when tested against seven separate K562 data sets [29, 52, 64, 184, 231, 240] it is consistently active. Given that the TFs defined by the TF profile within known cellular contexts, such as ESCs, myoblasts and lymphoblasts are meeting expectation given previous work, we have a degree of confidence in the reliability of the TF profiles.

2.4.3 Clustering TF profiles

We next sought a more unbiased examination of the TF activity profiles. Thus we turned to clustering of TF profiles. Here we curated data using DBNascent (See Methods B.1)[218]. After curating the data for sufficient quality and merging biological replicates we ended up with 126

distinct data sets in basal conditions. Basal, or control, conditions are standard cultured cell lines without a treatment condition. These conditions allow us to identify TFs that basally activated in a cell line and drive cellular function. An additional 161 data sets were curated under perturbation or genome modified conditions. These samples have a diverse array of treatments and allow us to identify the TFs that respond to these perturbations or modifications in a cell type specific manner. We used Ward's method to cluster the TF activity profiles (using Euclidean distance) across the DBNascent high-quality samples (287 distinct data sets). We found that the major determinant in clustering was cell identity. This indicates that the TF profiles are successfully capturing both ubiquitously active (TFs that promote cell viability) as well as TFs that drive cell identity (Fig. 2.6).

To develop our most confident baseline TF profiles we took the six cell lines with the most data represented in control conditions and extracted consensus profiles (2.5B). Within these profiles we capture TFs known to function in a given cell line. Some of these have known functions in the given cell line, such as STAT and GATA functions in K562s. Other TFs identified have less well annotated functions and are prime targets for follow-up experiments such as ZNF260 in ESCs and IRF8 in HUVEC cells.

In addition to identifying cell type specific TFs, we also identify ubiquitously enriched TFs. These TFs belong to a handful of well annotated TF families such as ETS, ATF and SP/KLF families. These TFs have highly redundant TF binding motifs known to be generally related to cellular proliferation and are often promoter associated (Fig. 2.7B). Additionally, they may play a key role in permitting accessibility in transcriptionally active regions[257]. Due to the redundant nature of these motifs – as many of them are high in GC content and highly repetitive – our methodology precludes distinguishing between these TFs. There may be subtle differences in which of these TFs is active in a given cell line, but at least a subset of these TFs are always active regardless of cellular condition.

While there are only 4-6 TFs per cell line that are truly cell type specific, they are the major determinant for clustering. Despite data coming from multiple nascent RNA sequencing protocols,

sometimes over a decade apart and from dozens of distinct research groups the clustering is driven by cell line, consistent with the idea that TFs are the major drivers of cell type identity.

Our next question was which PolII initiation regions are driving this signal within the profile. To address this, we assessed the relative number of enhancer and promoter regions unique to a given cell line, or shared across cell lines. We found that enhancers tend to be more unique to the cell line, whereas promoters tend to be shared across cell lines (2.5C).

Since enhancers tend to be more cell line specific and promoters tend to be ubiquitous, we tested which PolII initiation regions were driving the signal for each TF within the TF profiles. In the ESC consensus profile, Nanog is a dominant factor. The vast majority of Nanog motif containing regions in ESC cells (78.9%) are enhancers. Beyond that, when compared to the other five consensus profiles the enhancer regions containing Nanog were disproportionately unique when compared to the promoter distribution. This indicates that Nanog motif containing enhancers in ESC cells are cell type specific. In the ubiquitously shared TF, KLF12, the regions driving the signal tend to be promoter associated and shared between cell lines (Fig. 2.5D). Across TFs within the TF profiles, cell type specific TFs typically regulate at unique enhancers and ubiquitous TFs generally regulate at shared promoters (Fig. 2.7B).

2.4.4 Regulation at Transcription

Given that there is a bias in where classes of TFs regulate, we next tested whether there is also a bias in the TF binding motifs. In this we found that the TFs regulating in more than one cell line (shared TFs) tend to bind more GC rich regions, close to the average GC composition at promoters. Whereas the motifs of cell type specific factors tend have base composition close to genomic background (2.8A).

When compared to all TFs within HOCOMOCO there are a subset of TFs that bind AT rich regions that aren't called as active in any of our data sets. Therefore we searched the cell type expression pattern of these TFs in the GTEX database and the single cell fetal human Atlas. The vast majority of these TFs are expressed in cell types for which no nascent RNA sequencing

has been done. For example, UNCX is a TF implicated in regulation of the cerebellum with an AT-rich binding preference. No nascent sequencing experiment within DBNascent is represented with this tissue type. Although if nascent RNA sequencing was done on that cerebellum, we would anticipate that UNCX would be predicted as active and cell type specific via TF profiling. A more detailed comparison to GTEx and single cell data sets is ongoing.

Importantly, TF perform distinct roles between general housekeeping mechanisms to defining cell identity. Cell type specific TFs preferentially bind more GC poor regions that correspond to cell type specific enhancers. These TFs likely establish cell type accessibility patterns. Other TFs are more ubiquitously active and TFs tend to bind at shared promoters. These differences led us to ask whether the TFs themselves are regulated at differently in the cell. To address this question, we took the top significantly enriched TF for each cell type specific cluster, as well as the top significantly enriched ubiquitously shared TF and plotted their cumulative distribution function for the transcription level of the TF itself. We found that the cell type specific TFs fit an exponential distribution, whereas the shared TFs fit a normal distribution (Fig. 2.8B). This indicates that the expression of cell type specific factors is more cell type specific – indicating they are regulated at transcription, as they not transcribed unless they are active within that cell line. Ubiquitously shared TFs activity is not dependent on the transcription level of the TF, as these TFs are always transcribed and at a wide range of levels across samples. When all TFs are assayed for transcription level across samples, there is a bias towards exponential CDFs in cell type specific TFs, and normal distributions for shared TFs. Additionally, when we assay published RNA-seq data (n=1,408 samples), the trends observed here are reproduced, suggesting this is a fundamental feature of the regulation of the TF itself.

For example, when we take the ubiquitously shared TF KLF12 and plot the expression of this TF compared to the enrichment score there is no correlation between transcription level and activity (Fig. 2.8C). While this TF is always active, it's not being regulated at transcription. This is observed across many ubiquitously shared TFs such as SP1 and ETV1 (Fig. 2.9A). When compared to the cell type specific factor Nanog, the transcription level of the TF is positively correlated with

the TF activity. This positive correlation affirms that these TFs are only transcribed in cell lines in which they are actively regulating. The positive correlation between transcription level and activity is observed in many cell line specific factors, including MyoD and GATA-2 (Fig. 2.9B). This is perhaps important because these TFs may be the drivers of cell type specific enhancers, in which opening these heterochromatin regions in inappropriate cellular contexts would lead to aberrant transcription. Thus, the transcription of these TFs is shut down unless necessary within a particular cellular context.

2.5 Discussion

Here we present a model for TF activity inference in which basally ON TFs can be extracted from a single nascent RNA-sequencing experiment. We built a robust expectation model that recapitulates the dinucleotide preferences surrounding PolII initiation sites at gene promoters as well as distal enhancers. Using this model we are able to extract known and novel cell line specific factors from over 70 cell lines, as well as shared regulatory factors that drive basal cellular function and proliferation. Ultimately, this method provides a mechanism to assess ON TFs in any cellular context in a single, high-throughput assay and analysis.

Using the ON TFs across various TF profiles we utilized hierarchical clustering to classify TFs as either shared across cell lines or cell type specific. Using these classes we explored distinct properties of these TFs, such as DNA binding preferences, motif biases, and transcriptional regulation of gene encoding the TF itself. Future work for this study a deeper delve into inducible TFs and how they fit within this TF class framework.

We find that shared transcription factors preferentially bind at promoters and have binding motifs rich in GCs, a similar bias to that of promoters. These TFs tend to have a high degree of redundancy within their motif preference – with many motifs having similar, simple recognition sequences. Often these TFs are not essential, suggesting they may behave cooperatively to retain DNA accessibility[257]. Related is the fact that regions with high GC content exclude nucleosomes enabling higher accessibility for transcription to occur[72]. Shared TFs are enriched for roles in

cellular proliferation such as the ETS, KLF and SP1 families[109, 199, 255]. We also note that these TFs do not appear to be regulated at transcription, as their activity is not correlated with the transcription level of the TF gene itself. So, it may be that shared TFs are constantly transcribed at varying levels, and play more general roles at gene promoters to enhance accessibility and cell viability.

In cell type specific TFs we find that they preferentially bind at enhancers and their binding motifs have a nucleotide bias similar to genomic background, i.e. more AT rich. Additionally, cell type specific TFs are not transcribed unless they are ON within a given cellular context. Taken together we postulate that cell type specific factors not only drive cell identity but establish cell type specific enhancers and open chromatin regions. These TFs are not transcribed unless essential for a specific cellular outcome, otherwise they could possibly establish inappropriate enhancer regions that subsequently induce aberrant transcriptional patterns. In all, our model identifies unique sequence features inherent to cell type specific TFs, suggesting a model for the establishment of cell type identity.

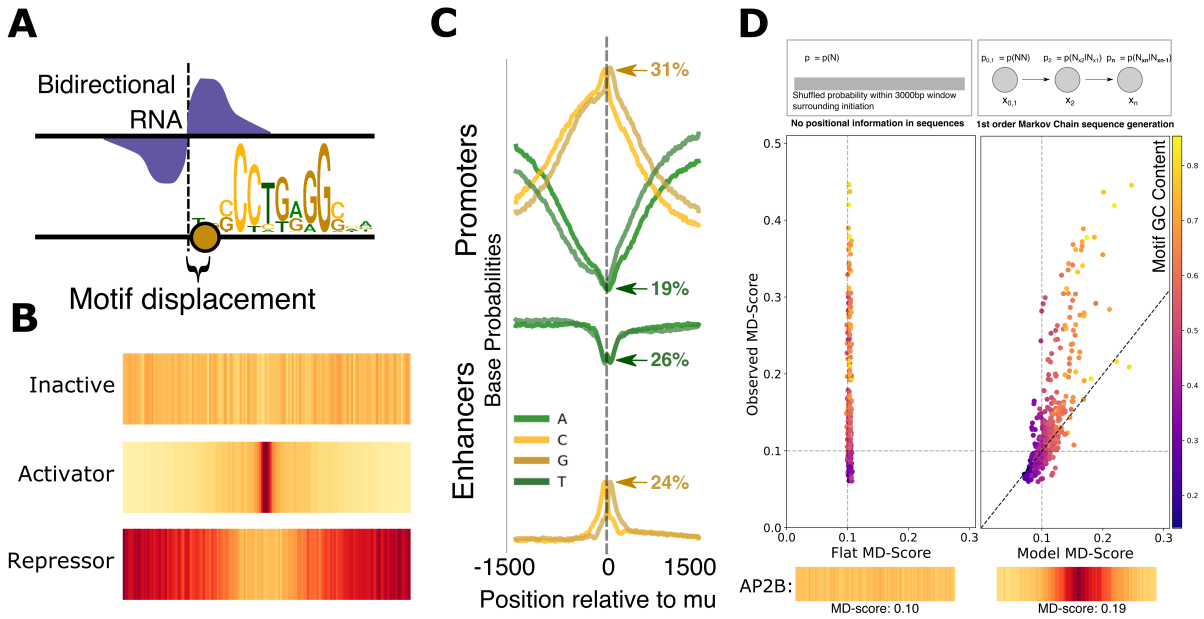


Figure 2.1: **Description of motif displacement scoring and models of background nucleotide distribution.** A) A cartoon representation showing a proximal TF motif (as PSSM) relative to a single bidirectional and its PolII initiation (dashed line). B) Example heat maps showing TF motif co-localization patterns relative to all bidirectional sites in a given data set. More heat (darker red) indicates a higher motif concentration. The three heatmaps here show an inactive TF (top), an activator (middle) and a repressor (bottom). The center of each heatmap is the annotated PolII initiation region (μ). C) Positional base composition relative to sites of Pol II initiation (dashed line) for promoters (top) and enhancers (bottom). D) Top is a graphical representation of the flat, uninformative background model (left) and dinucleotide model (right). Middle shows the observed (y-axis) versus expected (x-axis) MD-score plots for the flat, uninformative background model (left) and the dinucleotide model (right). Bottom are representative heat maps for an individual motif showing the resulting background distribution obtained.

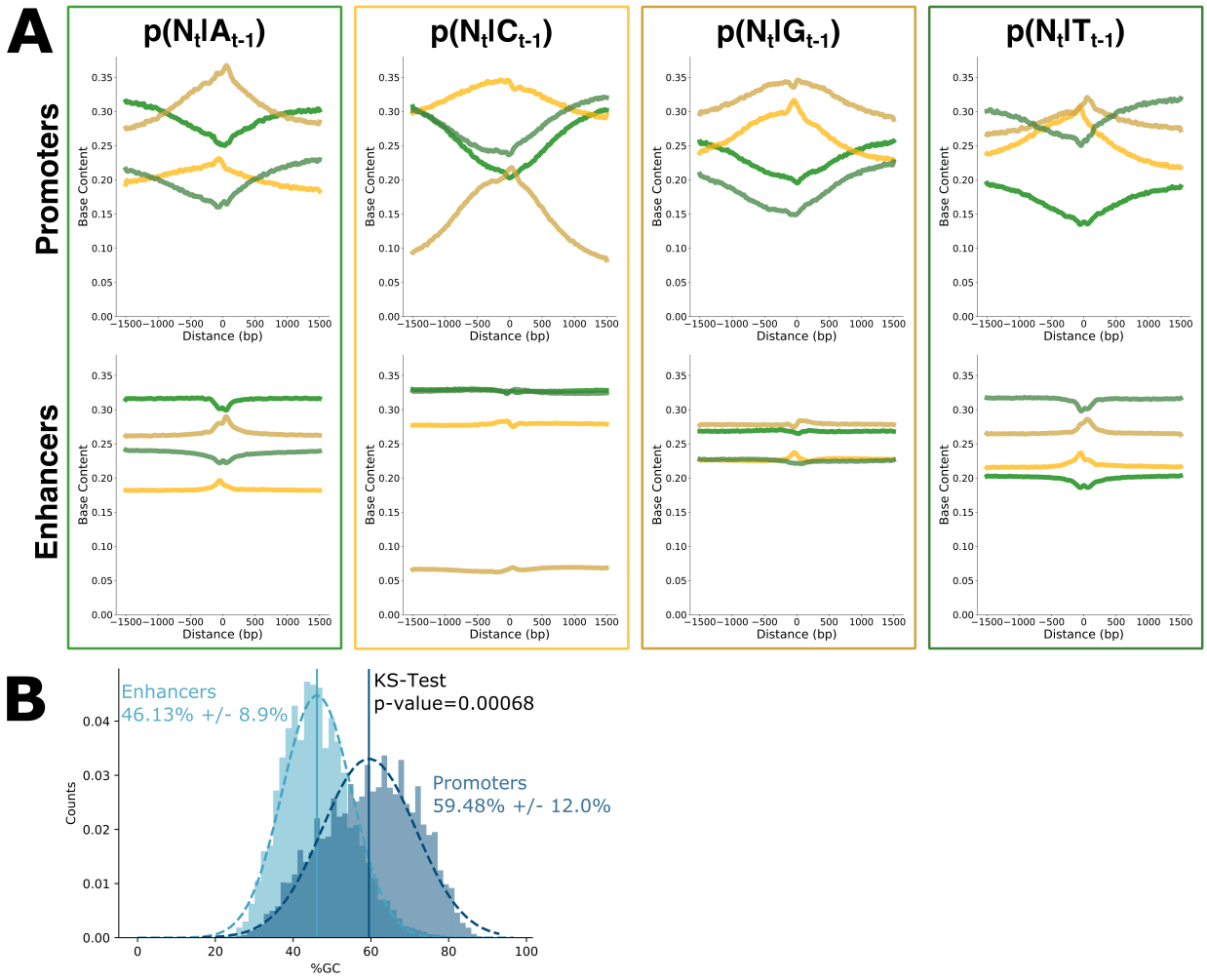


Figure 2.2: **Conditional Probabilities of dinucleotide model.** A) Conditional probabilities used to generate dinucleotide background model. From left to right are the conditional probabilities of each nucleotide given the previous base is A, C, G and T, respectively. Two sets of sequences were simulated within the model, promoter sequences (top) and enhancer sequences (bottom). B) Normal distribution of PolII initiation region's GC content. Promoters (defined by bidirectionals within 1000bp of an annotated TSS in RefSeq) are shown in dark blue with a mean GC content of 59.48%. Enhancers (defined as all bidirectionals that are not at promoters) are shown in light blue with a mean GC content of 48.13%. The normal distributions between enhancers and promoters are statistically different (KS-Test, p-value=0.00068).

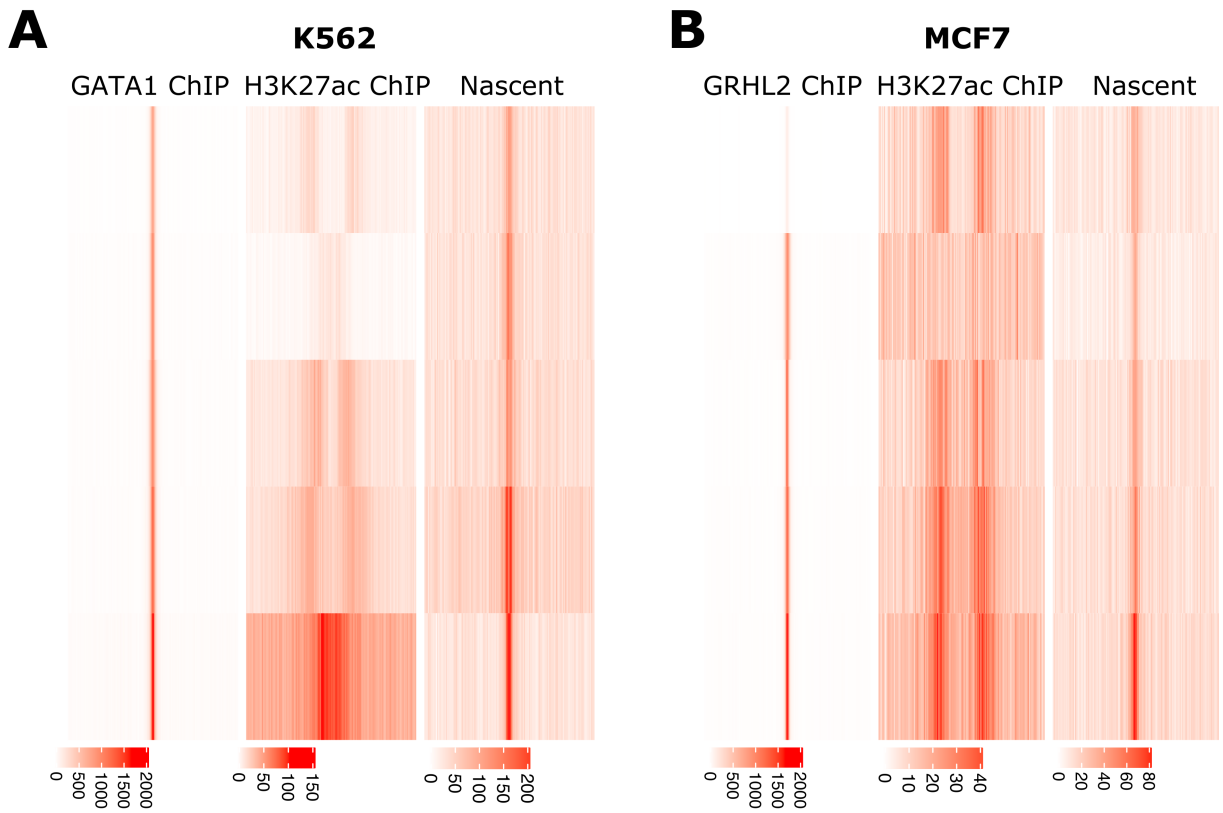
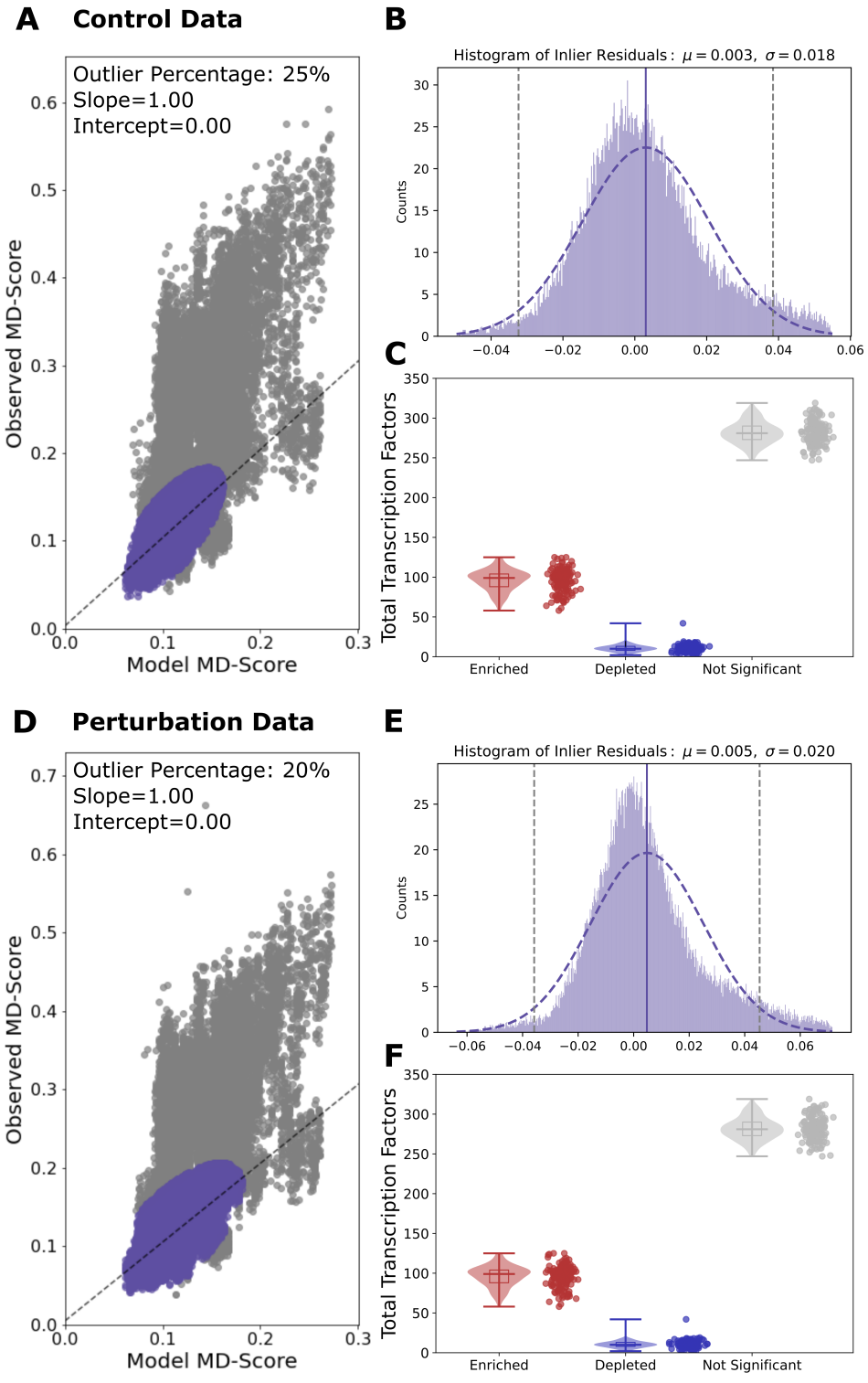
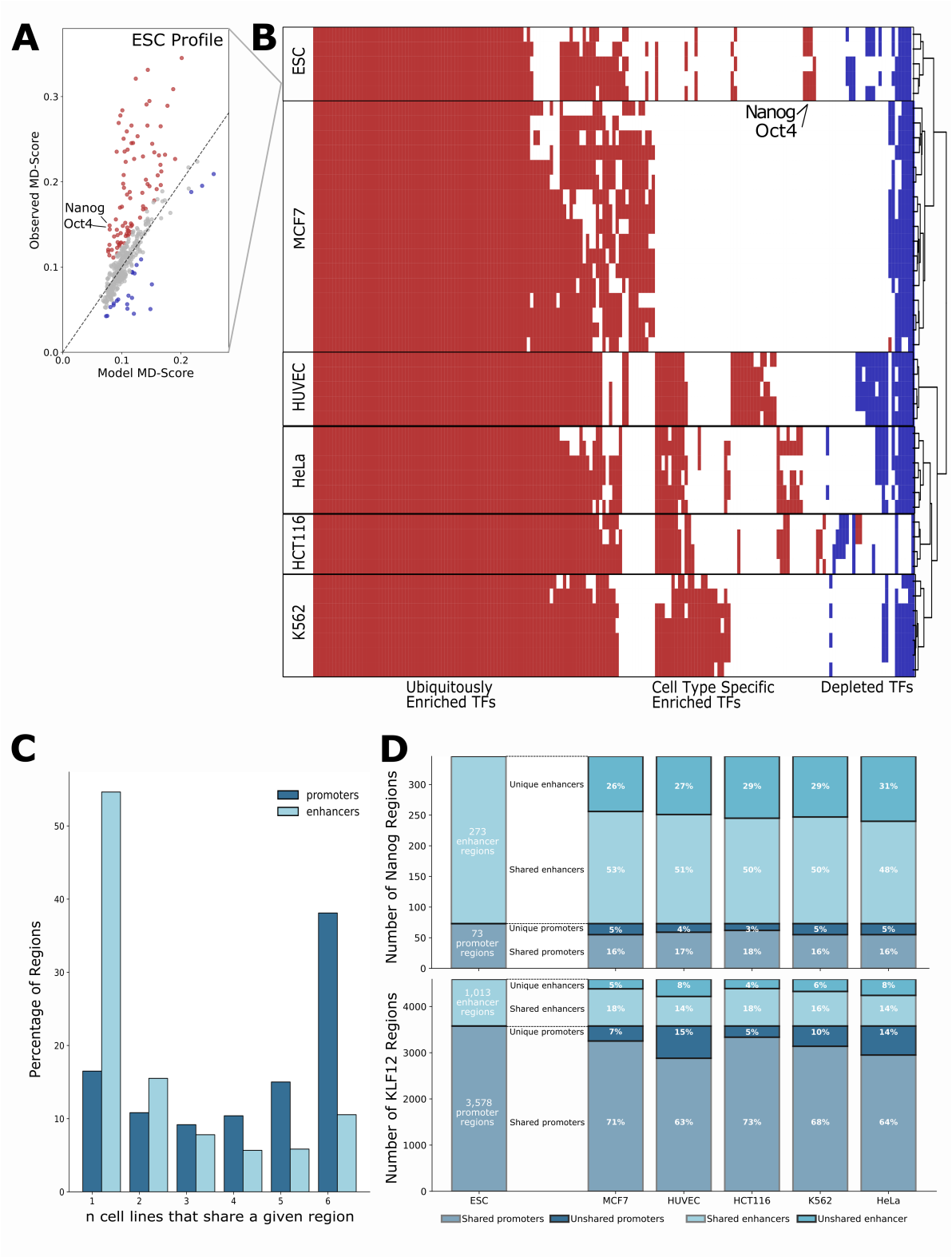


Figure 2.3: **MD-scores calculated from ChIP-seq data compared to nascent PolII initiation data.** These heatmaps show the number of motif hits within a given region. Darker red indicates more motif hits within the 150bp bin across that region. ChIP-seq data was pulled from cistromeBD[167, 258]. Five independent experiments were chosen for each example (top to bottom). A) Shows the cell type specific factor GATA1 in K562 cells. GATA1 ChIP-seq (left) demonstrates high co-localization of GATA1 binding with the GATA1 motif, as expected since ChIP-seq is how the GATA1 motif was defined. This is a low-throughput assay for annotating enhancers, as only one epitope can be pulled down at a time. H3K27ac ChIP-seq (middle) is a high-throughput assay to annotate enhancers. This data demonstrates that the broad histone peaks lose precision on the GATA1 motif:enhancer co-localization. Precision is required for robust MD-Score calculation. Finally, PolII initiation sites annotated from nascent sequencing data (right) demonstrate how a high-throughput assay can be used to precisely capture GATA1 motif:PolII initiation co-localization for robust MD-Score calculation. B) Another example with the cell type specific factor GRHL2 in MCF7 cells. Showing the co-localization of a given region with the GRHL2 motif in GRHL2 ChIP-seq (left), H3K27ac ChIP-seq (middle) and PolII initiation sites annotated from nascent sequencing data (right).



Continued on next page.

Figure 2.4: **Statistical test of ratio of observed to expected MD-score.** A) The dinucleotide model MD-score (expectation) and the experimentally determined MD-score (observation) were calculated for all HOCOMOCO TFs (n=388) from each control data set (n=126). The observed (y-axis) versus expectation (x-axis) for every TF from each control data set (n=48,888) were plotted. The 75% inliers (purple) were fit to a linear equation resulting in a slope of 1.00 and an intercept of 0.00. The 25% outliers (grey) were excluded from the linear fit. B) The residuals of the inliers were fit to a normal distribution (purple) with μ=0.003 (dashed purple line) and σ=0.018 (2σ at dashed grey lines). These values were used to calculate significance values of every TF within the control data sets (n=48,888). C) Violin plots demonstrating the number of TFs that are significantly enriched (ON-UP; red), depleted (ON-DOWn; blue) or off (grey) per data set within control conditions. D) The dinucleotide model MD-score (expectation, x-axis) and the experimentally determined MD-score (observation, y-axis) were calculated for all HOCOMOCO TFs (n=388) from each perturbation or genetically modified data set (n=161). The observed (y-axis) versus expectation (x-axis) for every TF from each perturbation data set (n=62,468) were plotted. The 80% inliers (purple) were fit to a linear equation resulting in a slope of 1.00 and an intercept of 0.00. The 20% outliers (grey) were excluded from the linear fit. E) The residuals of the inliers were fit to a normal distribution (purple) with μ=0.005 (dashed purple line) and σ=0.020 (2σ at dashed grey lines). These values were used to calculate significance values of every TF within the perturbation data sets (n=62,468). F) Violin plots demonstrating the number of TFs that are significantly enriched (red, ON-UP), depleted (blue, ON-DOWN) or off (grey) per data set within perturbation conditions.



Continued on next page.

Figure 2.5: **Cell lines cluster by TF activity profile.** A) Scatterplot of MD-scores expected according to the dinucleotide model (x-axis) compared to those observed (y-axis) in an ESC data set[219]. Each dot is a single TF's PSSM with inferred activators are labeled in red (ON-UP) and inferred repressors in blue (ON-DOWN). B) A two dimensional matrix where rows are control data sets (publications with combined biological replicates of control conditions) and columns are individual TFs. Entries in the matrix are colored red (activator, ON-UP), blue (repressor, ON-DOWN), or white (inactive, OFF) depending on their inferred activity status. Ordering of rows was determined by Ward clustering, which recovers the cell type utilized in the experiment (labeled on left). For succinctness, only cell types with more than four experimental samples are shown. The full data set is shown in Figure 2.6. C) Extent to which various cell types share a given RNA polymerase initiation region. A consensus initiation region profile was generated for each cell lines used to cluster in part B. The percentage of enhancer (light blue) and promoter (dark blue) regions that are only in one cell line (left) are compared to up to shared in all all six (right) cell lines are shown. D) The initiation regions within ESC cells containing either the cell type specific Nanog motif (top) or the ubiquitously shared KLF12 motif (bottom) within 150bp of μ were compared to the other five cell line's initiation regions containing the same TF motif. In ESCs, the Nanog motif predominantly appears in enhancers (light blue) and the KLF12 motif resides predominantly within promoters. When compared to the other consensus regions, the ESC Nanog motif containing enhancers tend to be more unique. The KLF12 promoters tend to be shared across cell lines.

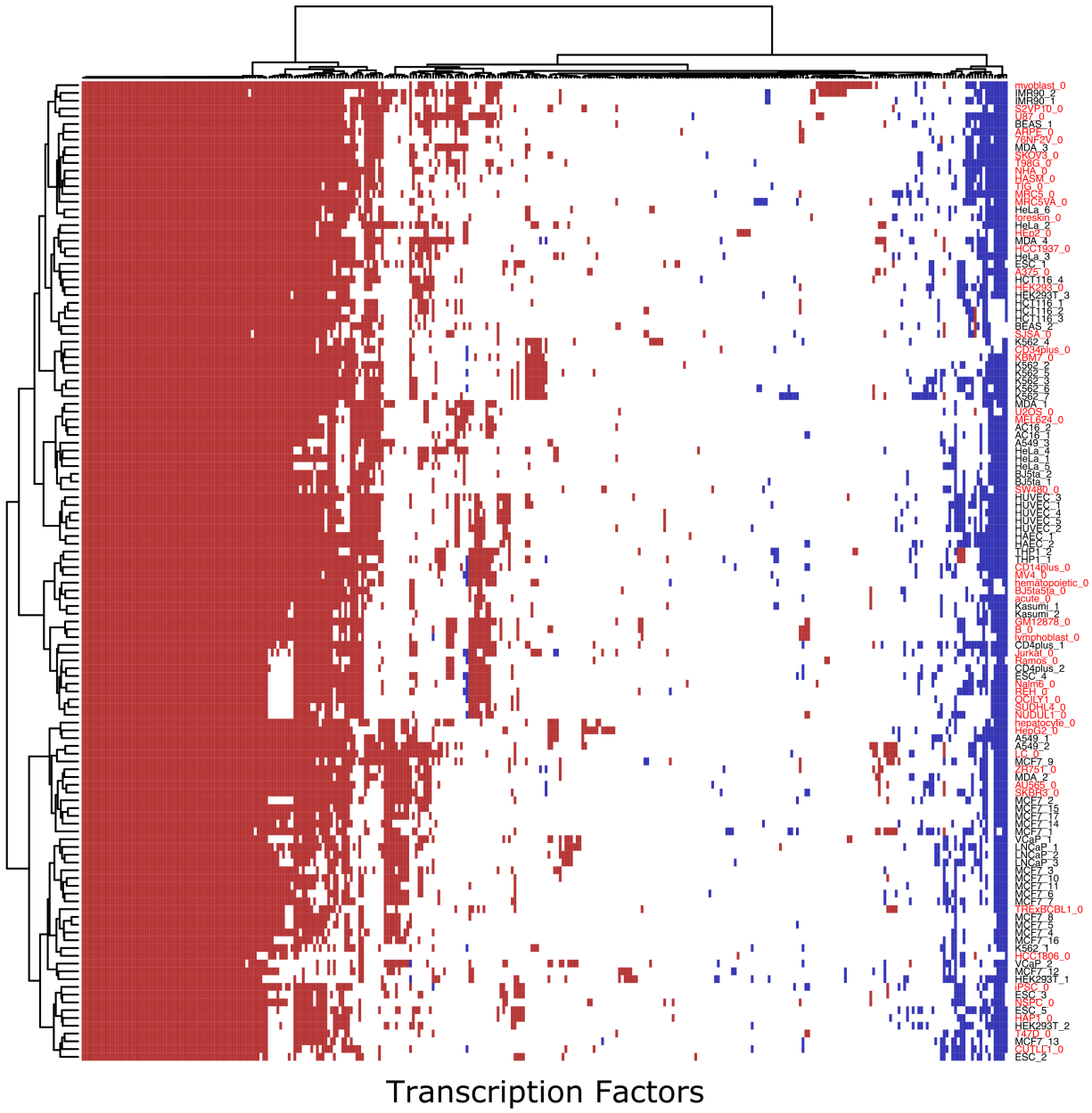
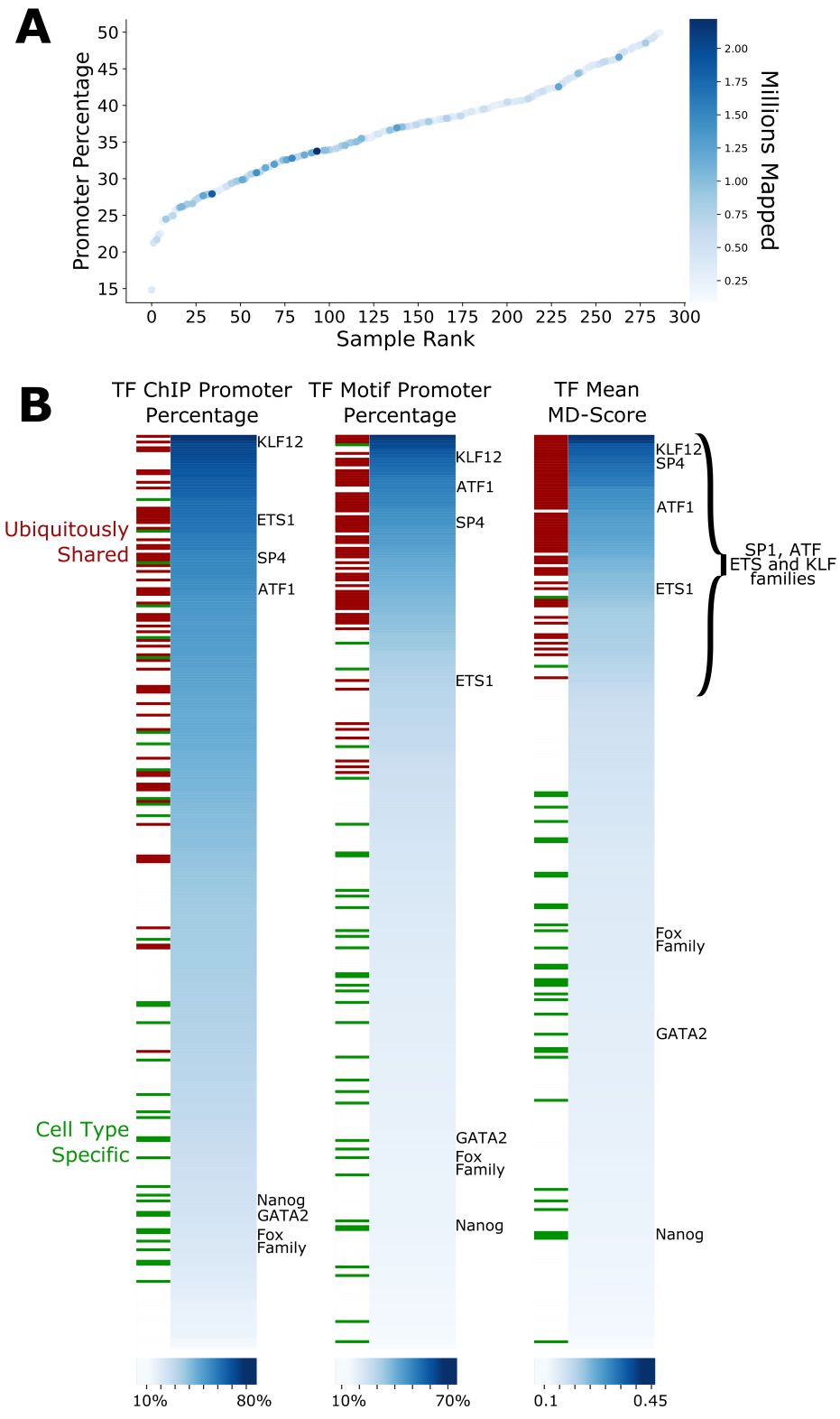


Figure 2.6: **Full matrix of TF activity profiles clustered.** A more complete version of Figure 2.5, showing all control data sets (rows, n=126 data sets). Red text indicates that cell line is only represented once within the curated sample set.



Continued on next page.

Figure 2.7: **Comparison of ChIP, motif and activity profiles.** A) The fraction of all initiation regions that correspond to promoters in all data sets (perturbation and control, n=287) varies. All data sets with greater than 50% promoter fraction of bidirectionals were excluded from all analyses. B) Fraction of events that are promoter associated (dark blue) versus enhancer associated (white) for ChIP (left), sequence motifs (middle) and inferred TF activity (right). In all cases, TFs are colored red if they were found ubiquitously on and green if they were on in a more cell type specific fashion.

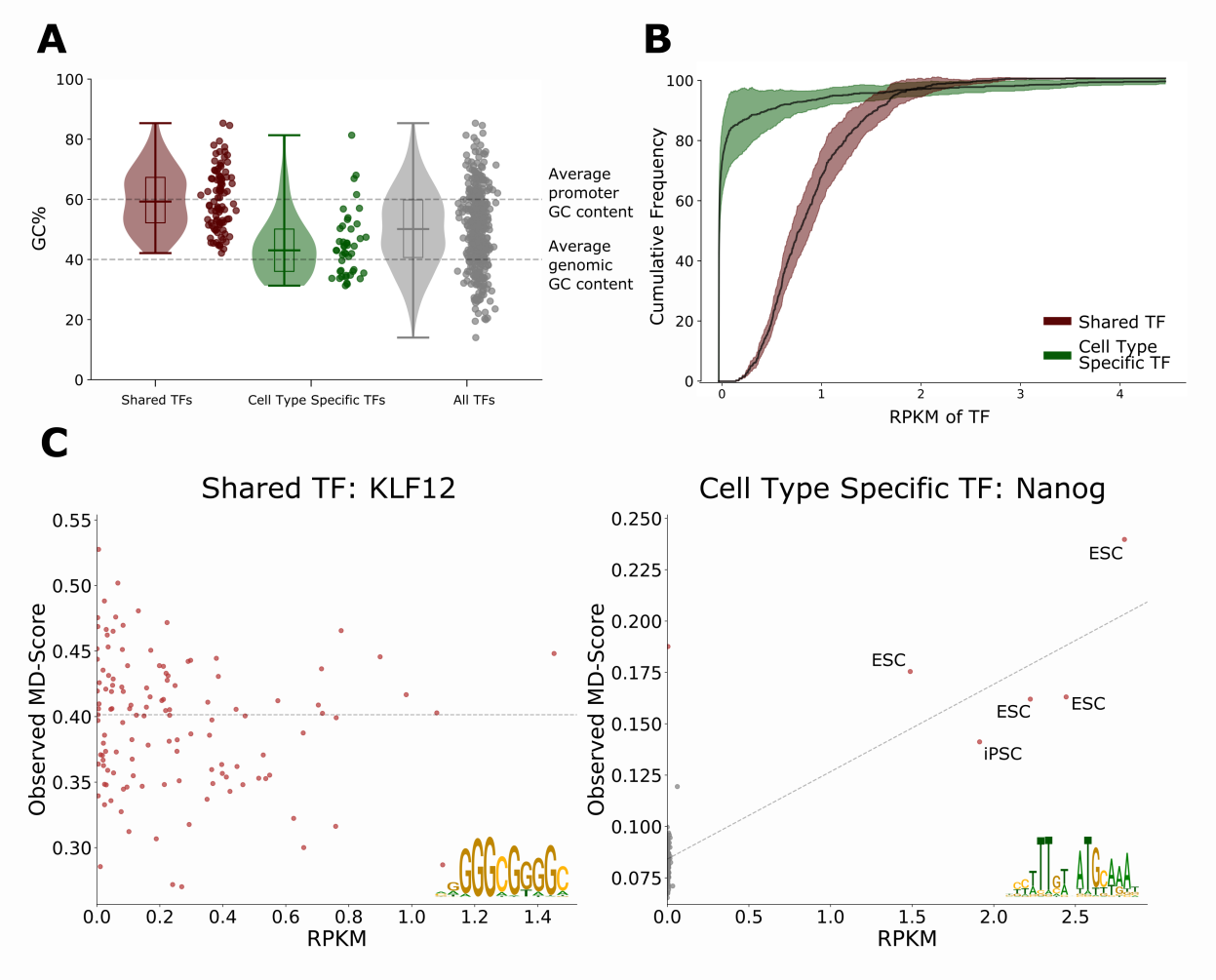


Figure 2.8: **Cell type specific TFs have recognition motifs that are more AT rich.** A) Violin plots of ubiquitously active (red), cell type specific (green) and all motifs (grey). Genome background is ≈ 40% GC (bottom dashed line) but promoters are ≈ 60% GC (top dashed line). B) CDFs of the gene RPKM for the top six most significant ubiquitously shared (red) or cell type specific (green) TFs. Cell type specific TFs were chosen by the top most significant TF per cell line represented in Figure 2.5B. C) Expression patterns of the gene encoding the TF (x-axis) compared to its MD-score (y-axis) for a ubiquitously enriched TF, KLF12, which has a high GC content motif (left) and a TF with a more tissue specific enrichment pattern, Nanog (right). PSSM in lower left corner as a logo.

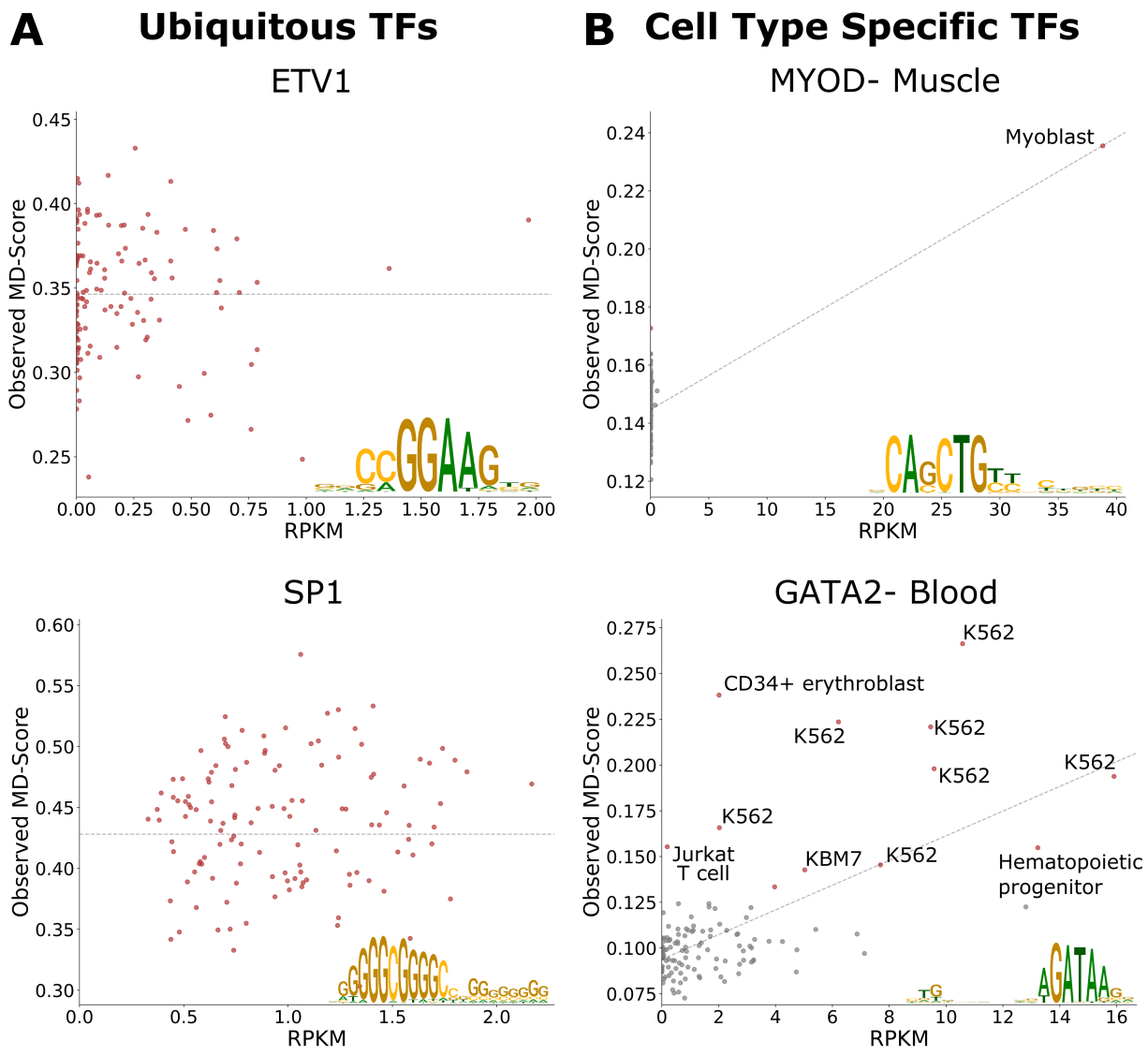


Figure 2.9: **Additional data of TF expression patterns compared to MD-score.** More examples of TF gene expression patterns compared to its MD-score genes with A) ubiquitously enriched and B) cell type specific. Scatter plots as in Figure 2.8.

Chapter 3

Suppression of p53 response by targeting p53-Mediator binding with a stapled peptide

The work in this chapter is as published in: Allen BL, Quach K, Jones T, Levandowski CB, Ebmeier CC, Rubin JD, et al. Suppression of p53 response by targeting p53-Mediator binding with a stapled peptide. Cell Reports 2022;39:110630. https://doi.org/10.1016/j.celrep.2022.110630[5]. All supplemental tables can be found with the published manuscript. Raw sequencing data can be found at GSE135870 and GSE193418. Code associated with this project can be found at `https://zenodo.org/record/6363472#.ZEg2OuzMKso`. Figures 3.5-3.10 were published as supplemental figures in published manuscript. Methods are discussed in Appendix B.2.

3.1 Contribution Statement

My main role in this publication was as the lead bioinformatician. When I joined this project there was already nuclear RNA-seq data produced for four key conditions (DMSO, DMSO+bivalent peptide, Nutlin-3a and Nutlin-3a+bivalent peptide) in two independent experimental designs. For this data I performed mapping, trimming and quality control. While the data was of passing quality the experiments were performed years apart by different scientists. I handled counting, batch correction and differential gene analysis via DESeq2 and edgeR. I also applied EISA analysis and quasR mapping, which utilizes the intronic reads to infer transcriptional impact of the bivalent peptide from the nuclear RNA-sequencing[85]. I performed the pathway analyses, gene snapshots and RNA-seq figure generation. Furthermore, I was responsible for RNA-seq data interpretation.

Upon reviewer request, I produced polII serine5 ChIP-seq in the four conditions previously stated in biological replicate. This experiment was a proxy for Mediator ChIP, which requires far more cell input. The ChIP-seq performed was in approximately one million HCT116 cells per condition, which is approximately 50-fold lower than the field standard. This was necessary as the bivalent peptide was only available in limiting quantities. Regardless, I obtained data of passing quality for the publication, and validated the experiments with ChIP-qPCR. Due to the low cellular input a significant amount of ChIP optimization was required. This included bivalent peptide treatment and timing optimization, chromatin shearing optimization, immunoprecipitation and wash step optimization and finally library preparation and optimization for low input samples. Upon sequencing, I handled mapping, trimming and quality control for the ChIP samples. Since we needed a quantitative comparison I developed new code to normalize the data, ran basic differential analyses in DESeq2 and edgeR as well as generating all ChIP related figures. Finally, I was responsible for ChIP-seq data interpretation.

3.2 Abstract

DNA-binding transcription factors (TFs) remain challenging to target with molecular probes. Many TFs function in part through interaction with Mediator, a 26-subunit complex that controls RNA polymerase II activity genome-wide. We sought to block p53 function by disrupting the p53-Mediator interaction. Through rational design and activity-based screening, we characterized a stapled peptide, with functional mimics of both p53 activation domains, that blocked p53-Mediator binding and selectively inhibited p53-dependent transcription in human cells; importantly, this "bivalent peptide" had negligible impact, genome-wide, on non-p53 target genes. Our proof-of-concept strategy circumvents the TF entirely and targets the TF-Mediator interface instead, with desired functional outcomes (i.e. selective inhibition of p53 activation). Furthermore, these results demonstrate that TF activation domains represent viable starting points for Mediator-targeting molecular probes, as an alternative to large compound libraries. Different TFs bind Mediator through different subunits, suggesting this strategy could be broadly applied to selectively alter

gene expression programs.

3.3 Introduction

Sequence-specific, DNA-binding transcription factors (TFs) drive myriad physiological processes and their mutation or disruption underlies many human diseases[140]. They are unquestionably high-impact targets for molecular therapeutics. Unfortunately, TFs have proven difficult to target with small molecules[35]; their DNA-binding domains are charged and similar to other TFs, and their activation domains are typically unstructured and intrinsically disordered.

Among the estimated ≈1600 TFs in the human genome[138], p53 stands out for its general importance in cancer biology[121, 122, 132]. Across many cell lineages, p53 functions as a tumor suppressor and can paradoxically function as an oncogene if it acquires specific gain-of-function mutations[82] p53 also plays key roles in mammalian development, aging, and stem cell biology. Like many TFs, the p53 protein possesses a DNA-binding domain and an activation domain (AD). The p53AD actually consists of two separate but closely-spaced domains, called AD1 (residues 13 – 29) and AD2 (residues 41 – 60). Whereas most transcriptional activation function can be attributed to p53AD1[118, 119], loss-of-function p53AD1 mutations retain some ability to activate specific subsets of p53 target genes, and mutation of both AD1 and AD2 is required to mimic a p53-null phenotype[36, 117].

The human Mediator complex contains 26 subunits and is generally required for RNA polymerase II (pol II) transcription[137]. A four-subunit kinase module containing CDK8 or CDK19 can reversibly associate with Mediator to control its function[157], but the CDK-Mediator complex was not a focus of this study. Mediator interacts extensively with the pol II enzyme and regulates its function in ways that remain poorly understood; however, a basic aspect of Mediator function is to enable TF-dependent activation of transcription. Mediator was discovered in S. cerevisiae using an in vitro assay to screen for factors required for TF-dependent transcription[78], and similar functions were confirmed for human Mediator complexes[79]. Because TFs do not interact with pol II directly, these and other studies established that TFs regulate pol II function indirectly, through

the Mediator complex. The p53 TF binds Mediator and this interaction has been shown to activate p53 target gene expression in vitro and in cells[115, 172]. Oncogenic mutations in p53AD1 disrupt p53-Mediator interactions[115] and this correlates with loss of p53 function[149]. Whereas specific residues and structural details remain unclear, the p53-Mediator interface appears to involve the MED17 subunit[115, 172]. Interestingly, other TFs (e.g. SREBP or nuclear receptors) activate transcription through interactions with different Mediator subunits[114, 252].

Directly targeting TF activation domains has proven to be a difficult strategy to control TF function. Here we sought to test whether the same outcome could be achieved by targeting Mediator instead. We chose p53 as a test case because it is well-studied, biomedically important, and contains a well-characterized activation domain. An apparent obstacle was that Mediator is large (1.4 MDa, 26 subunits) and its p53 interaction site is not precisely defined. However, we reasoned that the p53 activation domain (residues 13 – 60) evolved to selectively interact with Mediator with high affinity; supporting this concept, the p53AD alone can selectively purify Mediator from human cell extracts[172], and mass spectrometry analysis of p53AD-bound factors revealed Mediator as a top hit (Supplemental Tables 1,2). Consequently, we used the native p53AD structure and sequence as a starting point, rather than screen thousands of drug-like compounds. To directly assess p53-Mediator function, we used a defined in vitro transcription system that recapitulated p53- and Mediator-dependent transcription. Biochemical results were tested further in human cells, using genome-wide approaches. Collectively, these experiments establish that p53 activity can be selectively controlled by targeting its interaction with Mediator.

3.4 Results

3.4.1 An in vitro assay to test p53-activated versus. basal transcription

To screen peptides for the ability to selectively block p53-dependent transcription, we required an assay that enabled p53-dependent activation but that could also support basal (i.e. activator-independent) transcription. We previously established an in vitro transcription assay (Knuesel

et al., 2009) using purified human factors (Figure 3.1A). A key feature of this assay was that both activated and basal transcription could be reconstituted on naked DNA templates (i.e. DNA templates not assembled into chromatin). To adapt this assay for purposes of measuring basal versus. p53-activated transcription, we generated templates with Gal4 DNA binding sites upstream of a TATA-containing promoter sequence (Figure 3.5A, B). Upon titration of a Gal4 DNA Binding Domain-p53 Activation Domain (AD; residues 1-70) fusion protein into this system, we observed pol II-dependent transcription that was dependent on p53AD and Mediator (Figure 3.5C). Reactions containing Gal4-p53AD generally produced about two- to four-fold more transcripts compared to reactions with no activator (Figure 3.5C). Because experiments without Gal4-p53AD produced a low level of basal transcription that could be quantitated, this system allowed assessment of both p53-activated and basal transcription.

3.4.2 Design and synthesis of stapled peptides

We designed hydrocarbon-stapled peptide mimetics of the AD1 and AD2 regions of p53. Hydrocarbon-staples were employed to promote helicity within the peptides. Hydrocarbon-stapled peptides have previously been developed to mimic the α-helical portion of p53AD1 that binds MDM2/MDM4 with the goal of blocking the p53-MDM2/MDM4 interaction and restoring wild-type p53 activity[24, 39]. For the p53AD1 mimetics, we synthesized N-acetylated versions of the penta-arg-containing peptides BP1.2 - BP1.7[196], which are based on residues 14-29 of p53 (Figure 3.1B). This panel of peptides contains an i, i+7 hydrocarbon staple at positions 20 and 27 and five arginine residues grafted into various positions, which were originally introduced to improve peptide cytosolic access and nuclear localization[196]. For the p53AD2 mimetics, we designed a panel of hydrocarbon-stapled peptides that varied the length and position of the hydrocarbon staple and spanned residues 45-57 of p53 (Figure 3.1C). The panel included two peptides with an i, i+7 hydrocarbon staple (AD2-1 and AD2-2), two peptides with an i, i+4 staple (AD2-3 and AD2-4), and one peptide with an i, i+3 staple (AD2-4). Furthermore, both stapled and unstapled variants of the p53AD2 peptides were generated.

3.4.3 Functional screening of stapled peptide mimics of p53AD1 and p53AD2

Starting with the stapled p53AD1 peptides, we tested whether any would block p53-dependent transcription activation without inhibiting basal transcription. Initial screens were completed with 5 μM of each peptide (BP1.2 – BP1.7; Figure 3.1B). At this concentration, all peptides reduced p53 activated transcription, but BP1.4 and BP1.5 did not affect basal transcription (Figure 3.5D). In follow-up experiments, we observed that the BP1.5 peptide negatively affected basal transcription to some degree, in contrast to BP1.4 (Figure 3.5E). We therefore chose the BP1.4 peptide for further testing (also see below). To determine a concentration range in which the BP1.4 peptide selectively blocked p53-activated transcription but not basal transcription, we titrated BP1.4 into transcription reactions at concentrations between 0.9 μM and 9 μM (Figure 3.5F). Interestingly, BP1.4 activated basal transcription at concentrations of 4 μM and above, which could reflect weak binding of BP1.4 to Mediator (i.e. mimicking p53AD) to promote transcription activation. Consistent with this result, promoter-bound pol II complexes are activated upon p53-Mediator binding in vitro[172]. Although basal transcription was inhibited at the 9 μM titration point, the weak BP1.4-dependent activation made the determination of the IC50 for basal transcription impossible using an inhibitor response curve. The IC50 describing the inhibition of p53-activated transcription by BP1.4 was 3.2 $\pm$ 0.2 μM (Figure 3.5G). The concentration window in which BP1.4 selectively blocked activated transcription was therefore relatively narrow, but this issue was circumvented with next-generation peptides (see below). We next tested the p53AD2 peptides (Figure 3.1C) in a similar manner. In contrast to the p53AD1 peptides, the p53AD2 peptides either had no effect on p53-activated transcription or non-specifically inhibited both p53-activated and basal transcription at 5 μM (data not shown). Testing further at different peptide concentrations (i.e. increasing concentration if no activity was observed at 5 μM or decreasing concentration if both activated and basal transcription were inhibited) did not reveal any p53AD2 peptides with specificity for p53-activated transcription. These results were not entirely unexpected, as p53AD2 plays a lesser role (versus. p53AD1) in activation of p53 target genes in vivo[36, 117–119].

3.4.4 A bivalent peptide selectively blocks p53-dependent activation in vitro

We hypothesized that covalently linking two peptides with low-to-moderate affinity could generate a cooperatively binding "bivalent" peptide with improved ability to inhibit p53-dependent activation. Given the inactivity of the p53AD2 peptides tested, we elected to tether the BP1.4 peptide to the wild type p53AD2 sequence. In this way, we hoped to generate a competitive inhibitor of p53AD-Mediator binding by recapitulating the combined landscape of p53AD1/AD2 interactions. And because the p53AD1 portion was stapled (e.g. BP1.4), it would permanently retain the α-helical state, lowering the entropic cost of binding to more effectively compete with WTp53 for Mediator binding. We synthesized and tested three bivalent peptides (BP1.4 + p53AD2 sequence) that contained a 2-, 6- or 10-unit polyethylene glycol (PEG) linker (bivalent peptide 1, 2, or 3; Figure 3.6A). Notably, the bivalent peptides were significantly more potent inhibitors of p53-activated transcription than BP1.4 alone. As shown in Figure 3.6B, bivalent peptides (500 nM) containing either a 6- or 10-unit PEG linker (i.e. bivalent peptide 2 or 3) inhibited p53-activated but not basal transcription. By contrast, the bivalent peptide with a 2-unit PEG linker (i.e. bivalent peptide 1) did not inhibit transcription at 500 nM (Figure 3.6B). We additionally compared bivalent peptides BP1.4 and BP1.5 with the PEG6-linker and found that BP1.4 was slightly more potent (Figure 3.6D, E). Because the bivalent peptide containing a 6-unit PEG linker (i.e. BP1.4PEG6p53AD2; bivalent peptide 2) was easier to synthesize versus. 10-unit PEG, it was used for all future experiments. For simplicity, this molecule (bivalent peptide 2, Figure 3.6A) will be called the "bivalent peptide" throughout this paper. The in vitro transcription assays on naked DNA templates demonstrated improved potency of the bivalent peptide and also confirmed that it inhibited p53-activated transcription but not basal transcription. We next tested its function on more physiologically relevant chromatin templates, in which basal transcription is repressed. In fact, a TF activation domain (such as p53AD) and Mediator are required for transcription on chromatin templates[172, 185], presumably due to the ability of Mediator to relay the activation signal from the TF directly to the pol II enzyme. In vitro transcription assays with chromatin templates revealed

that the bivalent peptide had an IC50 of 85 nM when added to reactions with Gal4-p53AD. By contrast, the BP1.4 peptide alone had an IC50 of 330 nM in these assays (Figure 3.2A, B). Upon introduction of a QS mutation into p53AD2, which blocks its activation function in vivo[117], the IC50 increased to 713 nM, about eight-fold higher than the bivalent peptide and two-fold higher than BP1.4 alone (Figure 3.2A, B). Collectively, these results indicate that both p53AD1 and p53AD2 contribute to Mediator-dependent transcriptional activation in vitro. We next assessed whether the bivalent peptide would selectively block p53-dependent transcription compared with VP16, a viral activation domain. Whereas p53 and VP16 both interact with Mediator[115, 175, 233], they do so through different subunits (MED17 and MED25, respectively). In contrast to Gal4-p53AD (85 nM), the bivalent peptide had an IC50 of 424 nM in the presence of Gal4-VP16 (Figure 3.2C, D). These data, which resulted from experiments in which the only difference was the TF activation domain (i.e. identical DNA templates; identical TF DNA-binding domains), indicated that the bivalent peptide selectively blocked the p53–Mediator interaction versus. the VP16–Mediator interaction; this was further supported by biochemical data (see below). The reduced transcription with Gal4-VP16 at much higher concentration of bivalent peptide likely reflects transcriptional squelching, in which high levels of TF activation domains repress transcription in vitro or in cells, presumably through competition for binding of co-activators such as Mediator[78, 87, 173].

3.4.5 Bivalent peptide directly inhibits p53AD-Mediator interaction

To further test whether the bivalent peptide would selectively block the p53AD–Mediator interaction, we performed a series of biochemical experiments, as outlined in Figure 3.7A. The p53AD can bind Mediator with specificity and apparent high affinity[115]; for example, the p53AD itself is sufficient to selectively isolate Mediator from partially purified cell extracts[172]. As shown in Figure 3.7B, p53AD binding to Mediator was markedly reduced (approximately 60% bound versus. no peptide controls) in the presence of the bivalent peptide; by contrast, the bivalent peptide did not reduce VP16AD binding to Mediator (Figure 3.7C). Furthermore, a QS mutation abolished the ability of the bivalent peptide to block Mediator binding (Figure 3.7B). These data are consistent

with in vitro transcription results (Figure 3.2) and reveal that the bivalent peptide directly blocks p53AD–Mediator interactions. ChIP-seq experiments in HCT116 cells showed general agreement with these in vitro binding results, as described below.

3.4.6 Bivalent peptide suppresses p53 activity in Nutlin-stimulated cells

Prior analysis of the BP1.4 peptide showed that it is not effectively taken up by cells[196], and given its larger size, the bivalent peptide was expected to have poor cellular uptake. To circumvent this issue, we used a well-tested protocol to enhance cell uptake of the bivalent peptide (see Methods). HCT116 cells were evaluated either in the presence of bivalent peptide (Figure 3.3A) or vehicle (water), with or without Nutlin-3a. Nutlin-3a is a small molecule that activates and stabilizes p53 by inhibiting MDM2, a repressor of p53[229]. A 3h treatment time was used based upon experiments that showed the bivalent peptide was biologically active for only a limited time in cells (see Methods). After 3h Nutlin-3a treatment (or DMSO control, bivalent peptide), nuclear RNA was isolated and biological replicate RNA-seq libraries were prepared (Supplemental Table 3). As expected, Nutlin-3a induced expression of p53 target genes (Figure 3.3B; GSEA Figure 3.8A, Supplemental Table 4, RT-qPCR Figure 3.8C), consistent with previous studies in HCT116 cells[6]. Strikingly, however, Nutlin-induced activation of p53 target genes was diminished in cells treated with the bivalent peptide (Figure 3.3C; GSEA Figure 3.8B, RT-qPCR Figure 3.8C). Inhibition by the bivalent peptide was observed across a core set of p53 target genes shown by reduced enrichment in GSEA (Figure 3.3D, Supplemental Table 5). An additional set of control RNA-seq experiments was completed, in biological triplicate, to test whether serum-free media[39, 46] would influence bivalent peptide activity (see Methods). The data (RNA-seq experiment 2; Figure 3.9A, B) were consistent with the first series of biological replicates (RNA-seq experiment 1, Figure 3.3B-D, Figure 3.4A, B, Supplemental Table 4), despite a reduced p53 response to Nutlin. Collectively, these results indicated that the bivalent peptide inhibits activation of p53 target genes in human cells, consistent with the in vitro results.

3.4.7 Bivalent peptide has negligible transcriptional effects in absence of p53 activation

An expectation of our experimental strategy was that the bivalent peptide, which was designed based upon p53AD structure, would selectively block p53 function. HCT116 cells express hundreds of sequence-specific, DNA-binding TFs, including high-level expression of TFs that define the cell lineage[104]. For HCT116 cells, these TFs include SREBF1, ELF3, JUNB, NR2F1, and MYC. To assess the general impact of the bivalent peptide on pol II transcription, we compared RNA-seq data from cells without Nutlin treatment, in the presence or absence of bivalent peptide. The data revealed that the bivalent peptide had no significant impact on pol II transcription, genome-wide, in unstimulated HCT116 cells. For example, only one gene (MT1M) changed significantly, out of 28,260 transcripts analyzed (Figure 3.3E), and similar results were observed in the second set of RNA-seq experiments (Figure 3.9C; RT-qPCR Figure 3.8D). Whereas the volcano plots show fold-change effects with a significance cutoff, GSEA instead reports on trends within a ranked gene list. Interestingly, GSEA results for peptide-treated versus. untreated cells (no Nutlin treatment) showed evidence for modest activation of a small number of pathways, including the p53 pathway (Figure 3.9D). This observation could reflect an ability of the bivalent peptide to mimic p53AD-Mediator binding to activate transcription (as suggested in vitro for the BP1.4 peptide, Figure 3.5F) and/or an ability to reduce MDM2 or MDM4 binding to p53 in HCT116 cells. Taken together, the RNA-seq data from Nutlin-stimulated or uninduced cells (i.e. not treated with Nutlin) indicate that the bivalent peptide is selective for p53 and does not inhibit other TF-Mediator interactions that would otherwise more broadly impact pol II transcription.

3.4.8 Reduced pol II occupancy in peptide-treated cells

ChIP-seq experiments were completed to further probe the effects of the bivalent peptide. We emphasize that ChIP-seq experiments have limitations because data are collected at a single time point, whereas RNA-seq can better represent cumulative effects following a stimulus. The

timing for ChIP-seq was chosen to be 3hr post-Nutlin, based upon ChIP-qPCR experiments at the p21/CDKN1A locus (Figure 3.10A). This time point matched that of the RNA-seq experiments. ChIP-seq analysis of Mediator itself was not feasible based upon the low cell numbers required (ca. 1 million cells/replicate) due to limited quantities of the bivalent peptide. Because the genomic occupancy of Mediator correlates with pol II recruitment and transcription[243] , we completed ChIP-seq experiments for ser5-phosphorylated pol II, whose levels peak at gene 5'-ends (i.e. promoter-proximal regions). The ChIP-seq data showed increased ser5-phosphorylated pol II occupancy at the 5'-ends of p53 target genes in Nutlin-treated cells, as expected. The bivalent peptide reduced pol II levels at gene 5'-ends, in agreement with the RNA-seq data (Figure 3.4A, B; Figure 3.10B, C) and consistent with in vitro data that showed inhibition of p53-Mediator binding by the bivalent peptide. Note that the bivalent peptide decreased pol II occupancy at most but not all p53 target genes (Figure 3.10D), perhaps reflecting differential timing of Nutlin induction. Collectively, the RNA-seq and ChIP-seq data were consistent with the in vitro results and demonstrated that the bivalent peptide 1) blocks transcriptional activation by p53 and 2) has negligible impact on pol II transcription in general; that is, at genes responsive to other TFs (Figure 3.4C).

3.5 Discussion

Whereas few TF-Mediator interactions have been characterized in detail, the importance of bivalent or multi-valent interactions is an emerging theme[57, 102]. Our results demonstrate the importance of both p53 activation domains in Mediator-dependent transcription activation. A bivalent interaction may be required to selectively bind Mediator, as other proteins are bound by p53AD1 or p53AD2 individually[73, 149] or can have their affinity enhanced by p53AD phosphorylation[131]. For instance, p53AD phosphorylation will increase its binding affinity for CBP/p300[73]. The stapled, bivalent peptide is likely a poor substrate for site-specific phosphorylation, which may contribute to its effectiveness in cells. However, we cannot exclude the possibility that other functionally relevant interactions are influenced by the bivalent peptide, which may contribute to its cellular activity.

Although p53 normally functions as a tumor suppressor, gain-of-function p53 mutations are common and can be oncogenic[82, 121]; thus, blocking the activity of such p53 mutants is a viable therapeutic strategy[38]. Recent work has also shown that suppression of p53 function may have applications for tissue regeneration[217] or to prevent drug resistance[241]. Stapled peptides have shown promise as molecular therapeutics[178], and peptide drugs represented approximately $50 billion in U.S. sales in 2019, with more than 50 new drug approvals over the past 20 years[180]. The strategy outlined here demonstrates that stapled peptides derived from TF activation domains can be effective molecular probes; however, further optimization is required for potential clinical use. Our staple design for p53AD1 was based upon previous studies[24] and enforces an α-helical conformation. Future experiments are needed to structurally define the p53-Mediator interface targeted by the bivalent peptide. Indeed, whereas biochemical data suggest that p53 interacts with the MED17 subunit[115], precise details about the molecular interface are lacking. Among Mediator subunits, MED17 is noteworthy because it represents a core structural subunit, along with MED14[256]. Data from yeast suggest that MED17 may be more important for pol II transcription, genome-wide, compared with MED14[107, 225]. Given the central role for p53 in diverse physiological functions, its interaction with MED17 may ensure robust p53 responses independent of MED14 status or cell type.

Historically, TFs have been intractable as therapeutic targets, although progress has been made[42]. This proof-of-concept study suggests that desired transcriptional outcomes can be achieved by avoiding the TF entirely and targeting the human Mediator complex instead. Additional support for this concept was provided by Arthanari et al.[183], in which the Med15-Pdr1 interaction was blocked with a small molecule in yeast (C. glabrata). Analogous to our results with p53, they showed that disruption of the Mediator-Pdr1 interaction prevented activation of Pdr1 target genes in yeast cells. Whereas many, if not most, TFs target the Med15 subunit in yeast[210], TFs bind many different sites on human Mediator[194]. An implication is that blocking a single Mediator-TF interaction will not affect other signal-responsive or lineage-specific TFs, thus providing a means to selectively alter gene expression patterns. Our results suggest this strategy

could be applied toward p53 and other human TFs that target distinct Mediator subunits.

3.6 Limitations of this study

The RNA-seq experiments were completed with nuclear RNA after only 3hr Nutlin treatment, to better capture direct versus. indirect effects. Analysis across longer time points could reveal more gene expression changes from peptide treatment, but this would also increase the contribution from indirect effects due to p53 activity changes. Whereas the in vitro experiments can reliably assess direct p53-Mediator effects and mechanism, the complexity of factors (i.e. proteins, nucleic acids, metabolites) that converge on active genes in human cells prevents a complete understanding of cellular mechanisms. We could not obtain Mediator ChIP-seq data to assess its recruitment in cells; our ChIP-seq experiments were limited to about 1 million cells/replicate due to limited amounts of the bivalent peptide. This is about 50-fold less than typical for pol II ChIP-seq[11] and over 100-fold less than published Mediator ChIP-seq experiments[197].

There are some caveats regarding the data within this manuscript. Regarding the RNA-seq, the main source of variance within the RNA-seq data is between the biological replicates. Less than 5% of the variance could be attributed to the p53 response to Nutlin, likely due to the experiments being done at times years apart by different scientists. After batch correction, there is a mild p53 response observed when expected. p53 is the strongest perturbation response, but we likely lost a lot of the more subtle responses due to this replicate problem. This issue is more exacerbated in RNA-seq experiment set 2. Regarding the ChIP-seq, while the data was passing quality it was on the lower end in regards to polII Ser5 enrichment. There were two samples failed qc (DMSO+bivalent peptide and Nutlin+bivalent peptide in biological replicate one) and I was unable to reproduce due to limiting amounts of bivalent peptide. Since they did not pass quality metrics they were excluded from subsequent analyses. Normalization was performed but due to low complexity of the samples, and two samples failing, in depth statistical analysis were not practical. For this reason, the ChIP-seq data was mainly used to validate trends observed in the RNA-seq.

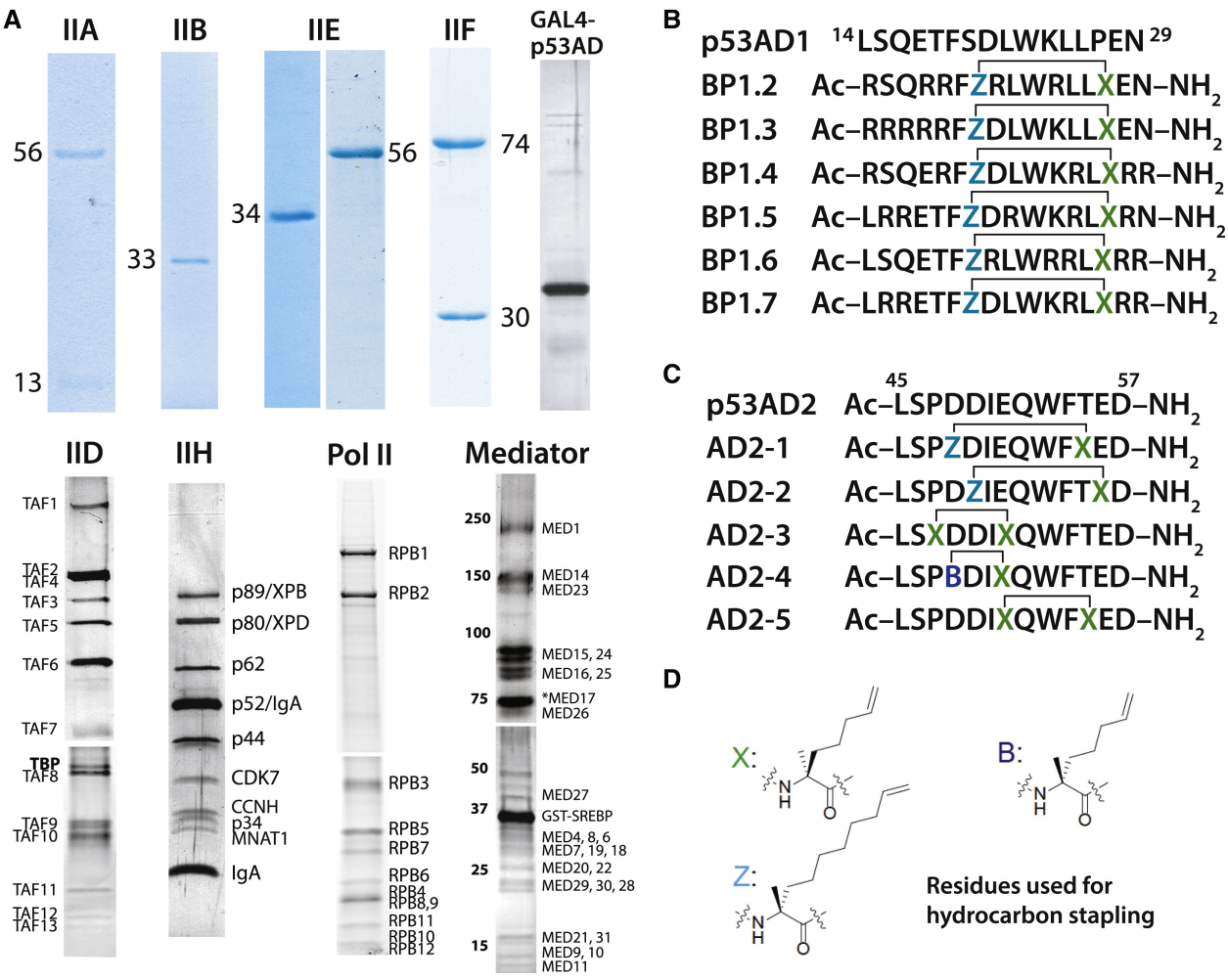


Figure 3.1: **Human factors and peptides used for the in vitro transcription assays.** (A) Purified PIC factors. (B) WT p53AD1 sequence and sequences of p53AD1 peptides containing diverse penta-arg motifs. (C) WT p53AD2 sequence and sequences of p53AD2 peptides. (D) Residues Z, X, and B represent α,α-disubstituted amino acids with olefin tethers for hydrocarbon-stapling. Unstapled structures shown.

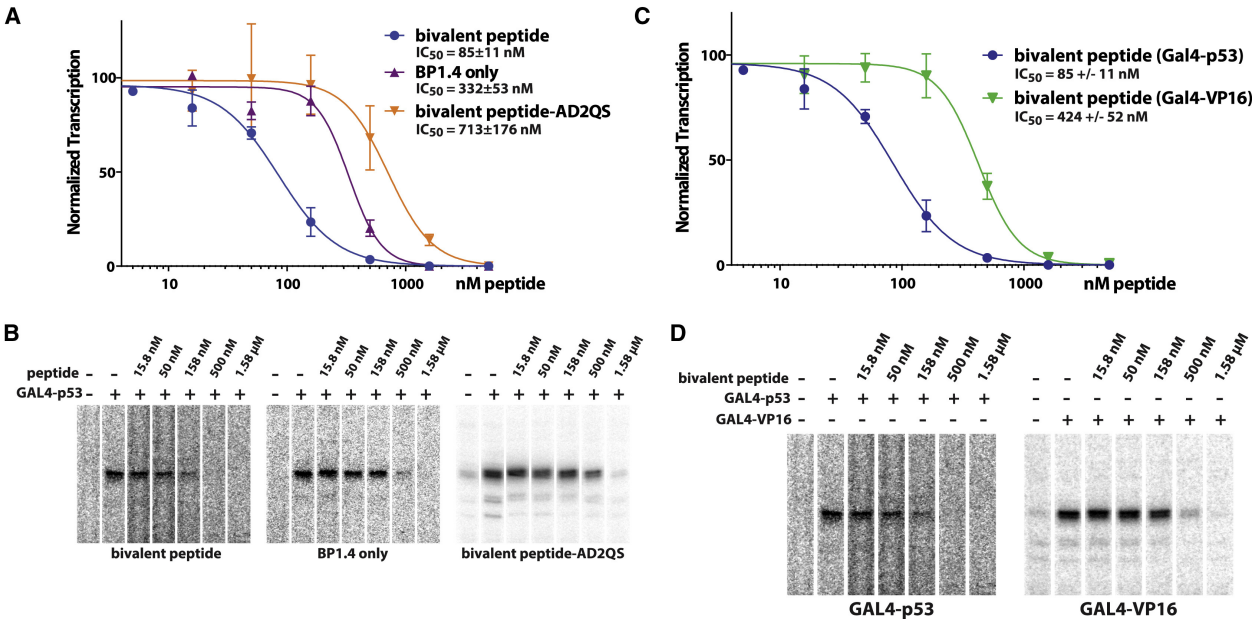
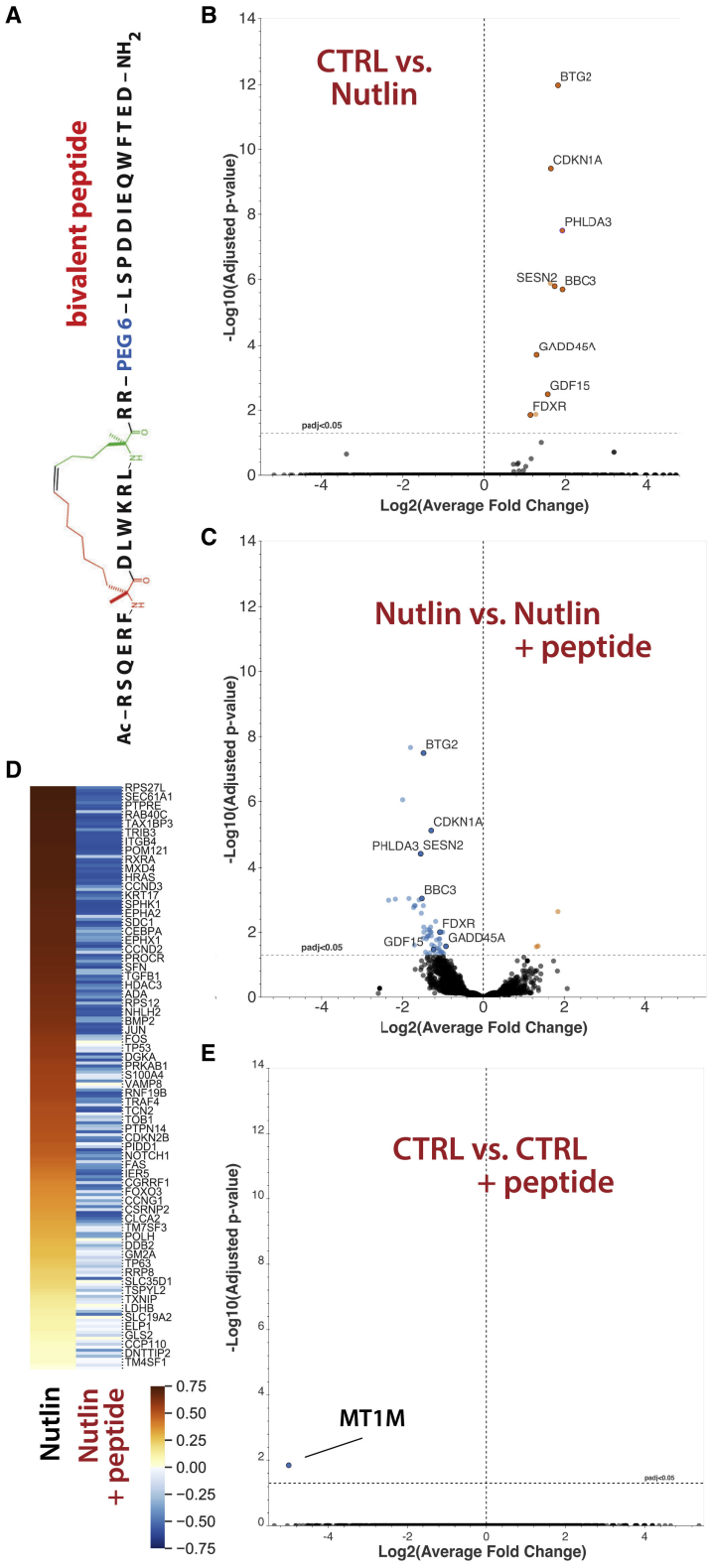


Figure 3.2: **In vitro transcription on chromatin templates reveals bivalent peptide is a potent and selective inhibitor of p53-dependent transcription.** (A) IC50 plot showing activity of bivalent peptide (n = 3 to 8 biological replicates) versus. BP1.4 (stapled p53AD1 mimic; n = 2 to 6 biological replicates) or a bivalent peptide with a mutated p53AD2 region (n = 3 to 9 biological replicates). (B) Representative data from experiments plotted in A. (C) IC50 plot showing that bivalent peptide is selective for p53; repressive activity is reduced with GAL4-VP16 (n = 2 to 6 biological replicates), which binds a different Mediator subunit compared with p53. (D) Representative data from experiments plotted in C. Vertical lines in plots represent standard error of the mean (panel A, C).



Continued on next page.

Figure 3.3: **The bivalent peptide blocks activation of p53 target genes but has negligible effect on pol II transcription in the absence of p53 activation.** (A) Schematic of the stapled, bivalent peptide. (B) Volcano plot showing significant induction of p53 target genes upon Nutlin treatment. (C) Volcano plot showing that the bivalent peptide reduces expression of p53 target genes in Nutlin-treated cells. (D) Enrichment score (GSEA) heatmap for Nutlin and Nutlin+peptide at a core set of p53 target genes. The p53 pathway is the most significantly altered pathway in both contexts: most upregulated in Nutlin and most downregulated in Nutlin+peptide (Figure 3.8A, B). (E) Volcano plot showing that the bivalent peptide causes virtually no significant changes in pol II transcription in the absence of p53 activation. Whereas MT1M met the significance cutoff, the data showed an outlier in one DMSO replicate, suggesting it does not reflect a true biological difference (see Methods). RNA-seq data were obtained in biological replicate.

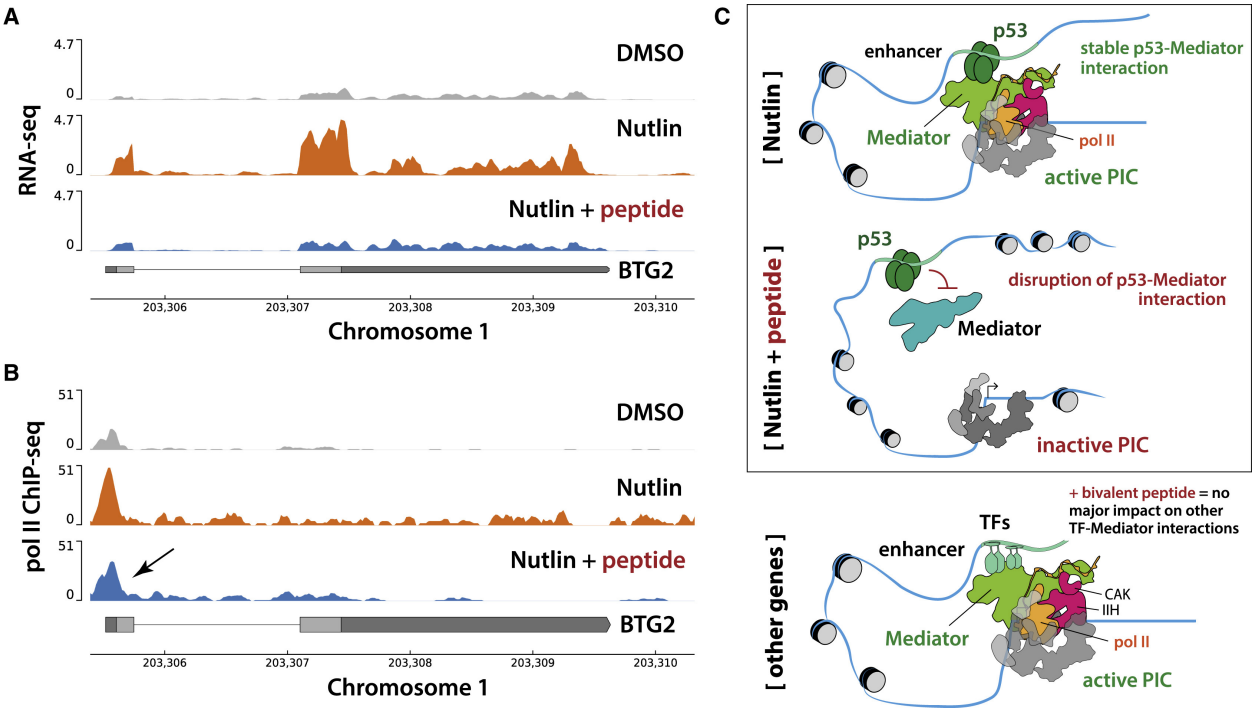


Figure 3.4: **Representative RNA-seq and ChIP-seq data; model.** (A) RNA-seq traces at the p53 target gene BTG2, showing bivalent peptide blocks Nutlin-dependent activation. The y-axis is read-depth normalized read density. RNA-seq data were obtained in biological replicate. (B) ChIP-seq traces showing decreased pol II CTD Ser5P occupancy at the BTG2 promoter region in bivalent peptide-treated cells. The y-axis is normalized read density (see Methods); ChIP-seq data were obtained in single or biological replicates (see Methods). (C) Model. The bivalent peptide effectively competes with p53 to reduce its binding to Mediator, which reduces activation of p53 target genes. At non-p53 target genes, which are activated through other TF-Mediator interactions (via different Mediator subunits), the bivalent peptide has minimal impact on pol II transcription.

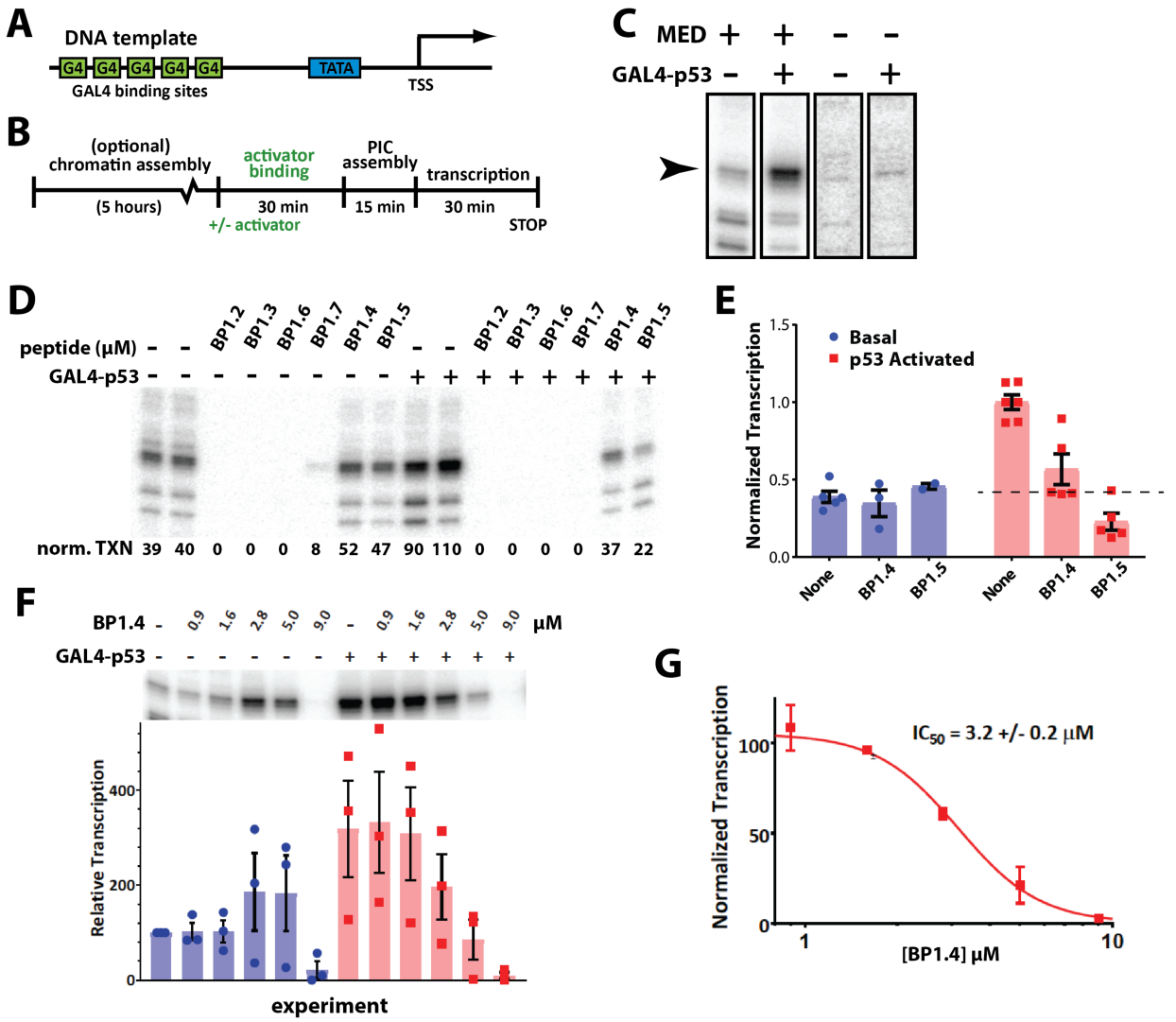


Figure 3.5: **In vitro screening protocol; functional screening of stapled and unstapled p53AD1 mimics.** (A) Promoter DNA template scheme. (B) Overview of reconstituted in vitro transcription assay (chromatin assembly optional; PIC = Pre-Initiation Complex: TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, Mediator, and pol II). (C) Representative in vitro transcription data from naked DNA templates, showing p53- and Mediator-dependence. (D) Representative data (in vitro transcription) using each peptide at 5 μM concentration, in presence (+) or absence (-) of GAL4-p53. Note that only BP1.4 and BP1.5 show ability to inhibit p53-activated transcription while not markedly affecting basal transcription. (E) Scatter plot summarizing in vitro transcription data for BP1.4 and BP1.5 peptides in absence (basal) or presence (activated) of GAL4-p53. Dashed line represents basal transcription level. (F) Representative data (top) and scatter plot (bottom) summarizing results from titration experiments with BP1.4 peptide under basal (- GAL4-p53) or activated (+ GAL4-p53) conditions. (G) IC50 plot summarizing inhibitory activity of BP1.4 peptide. For data panels (D-G), transcription was normalized to GAL4-p53 in absence of added peptide.

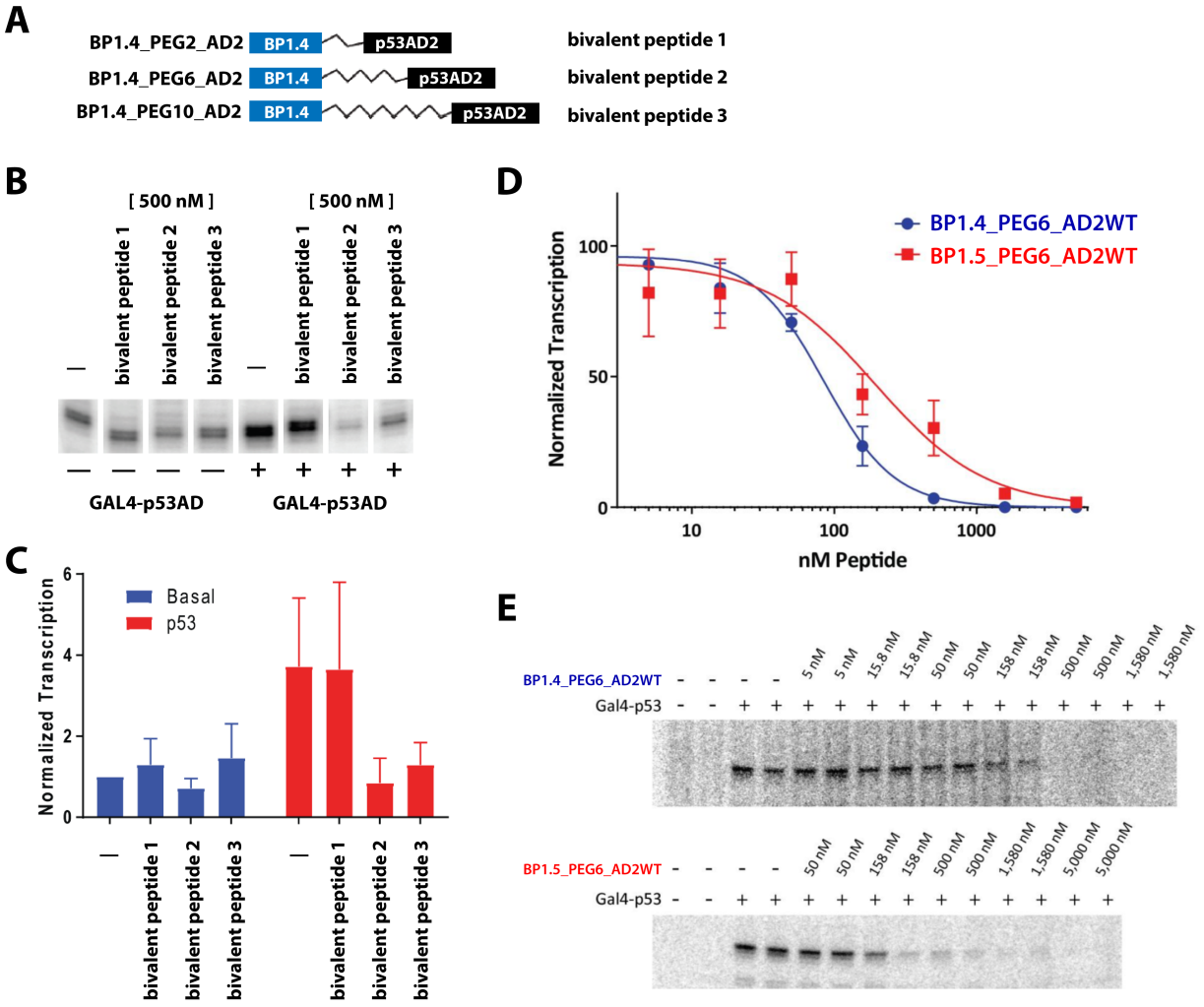


Figure 3.6: **Testing different PEG linker lengths to tether BP1.4 (stapled p53AD1 mimic) to p53AD2 sequence.** (A) Schematic of 3 different PEG linkers to generate bivalent peptide 1, 2, or 3. (B) Representative in vitro transcription data and (C) bar plot summary (n = 4; bars = s.e.m.) of bivalent peptides with different PEG linker lengths. Note that a PEG6 or PEG10 linker showed enhanced ability to block p53-activated transcription, whereas PEG2 linker (i.e. bivalent peptide 1) did not. (D) IC50 plot showing activity of bivalent peptide (i.e. BP1.4PEG6AD2WT) versus. an AD1 derivative (BP1.5). Points with error bars represent standard error of the mean, with n = 3 to 8 (blue dots) or n = 3 to 4 (red squares). The IC50 value for BP1.5PEG6AD2WT was 200 ± 88nM whereas the IC50 for the bivalent peptide BP1.4PEG6AD2WT was 85 ± 11nM (see Figure 3.2). (E) Representative in vitro transcription data used for the plot in panel D.

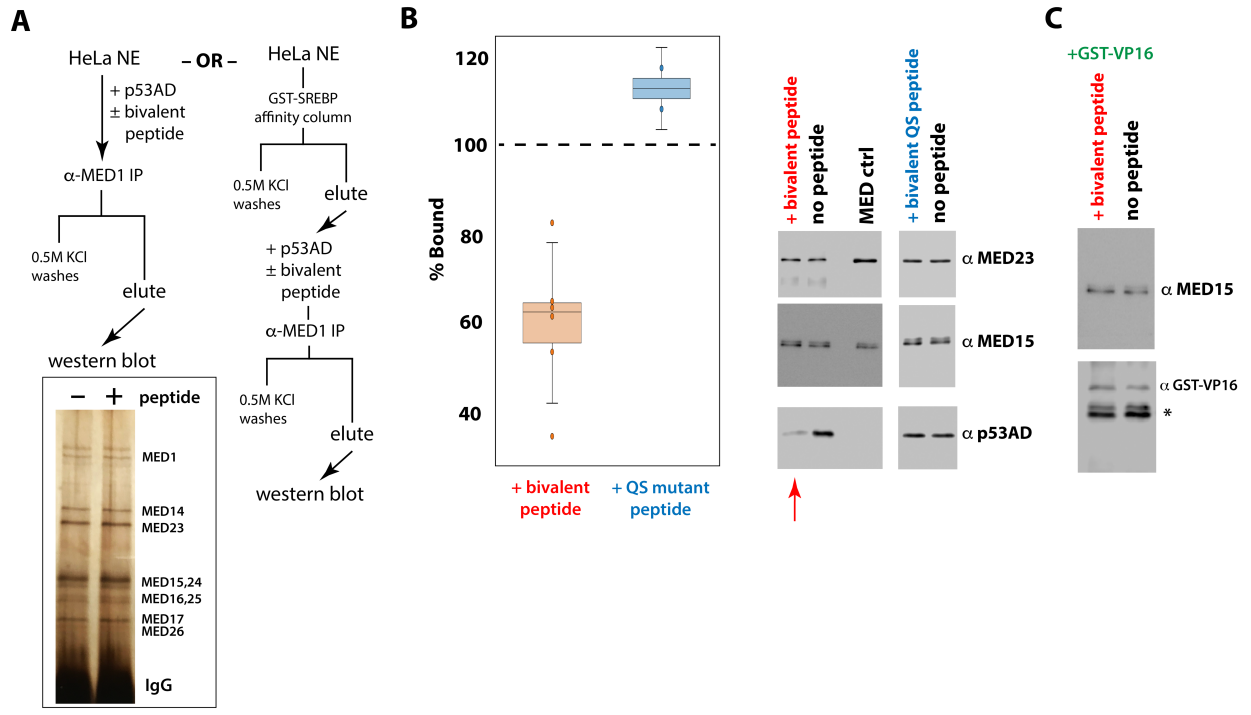


Figure 3.7: **Bivalent peptide blocks p53AD-Mediator binding.** (A) Overview of binding assays used. A crude Mediator sample was isolated from HeLa NE with a GST-SREBP affinity column [181] prior to incubation with p53AD (2 μM) ± bivalent peptide (5 μM) in one of the protocols (right). Inset: silver-stained gel of the MED1 IP, showing a relatively pure Mediator sample. (B) Scatterplot (left) summarizing p53AD binding to Mediator in presence of bivalent peptide or a QS mutation in p53AD2. The percent binding is relative to p53AD bound to Mediator in absence of added peptide (dashed line); p53AD quantitation was normalized to total Mediator, as assessed by quantitation of MED15 signal (and/or MED1 in some cases; n = 4 biological and 6 total replicates). Representative data (western blot) shown at right. Similar results were obtained with either protocol shown in panel A. (C) Binding assay shown at right in panel A was used to probe VP16-Mediator binding (n = 2), which revealed that VP16-Mediator binding is not inhibited by the bivalent peptide. Normalization of bound GST-VP16 to MED15 in fact showed a 1.67-fold increase in Mediator-bound VP16 in the +peptide experiments. Representative western blot shown. Asterisk: free GST.

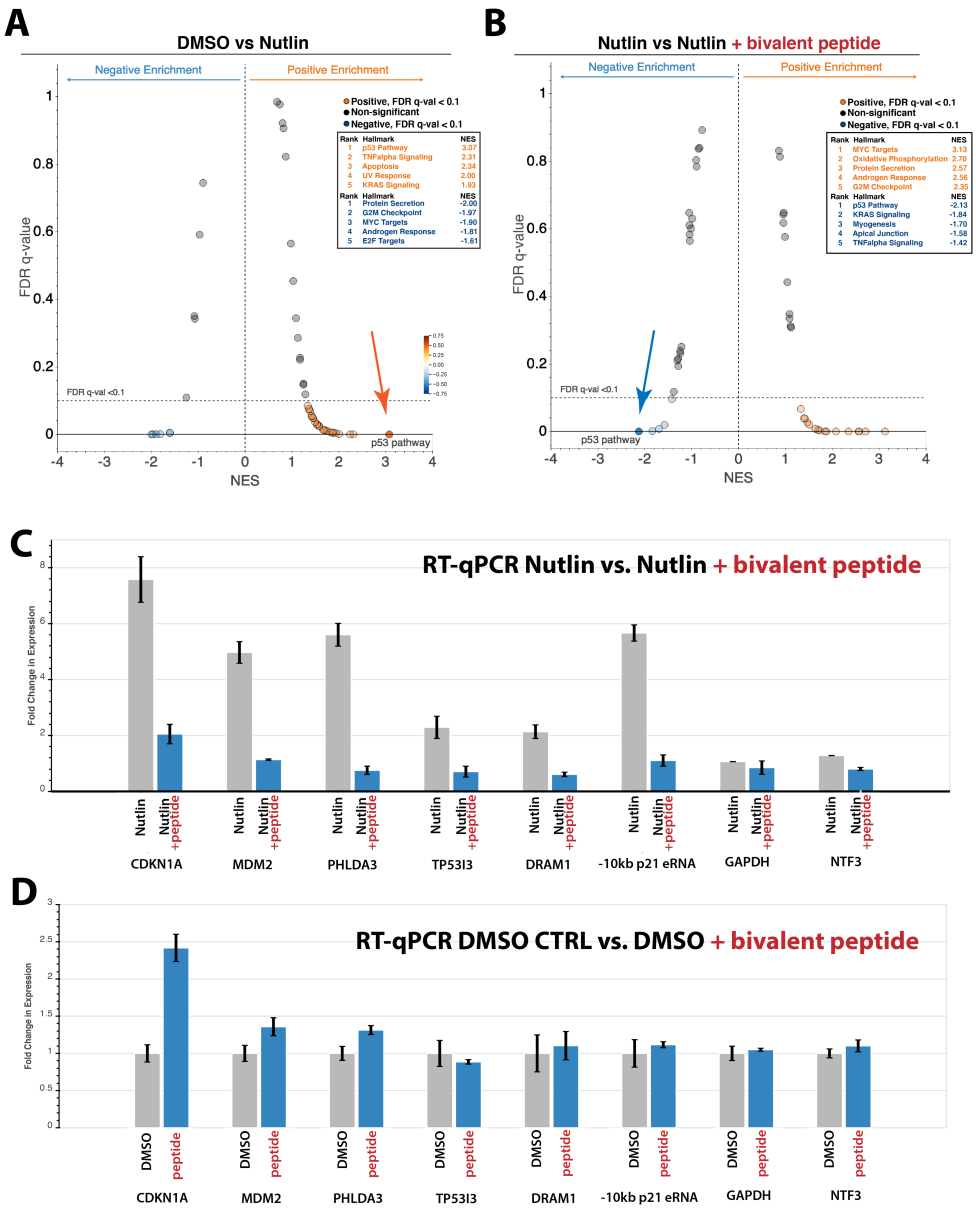
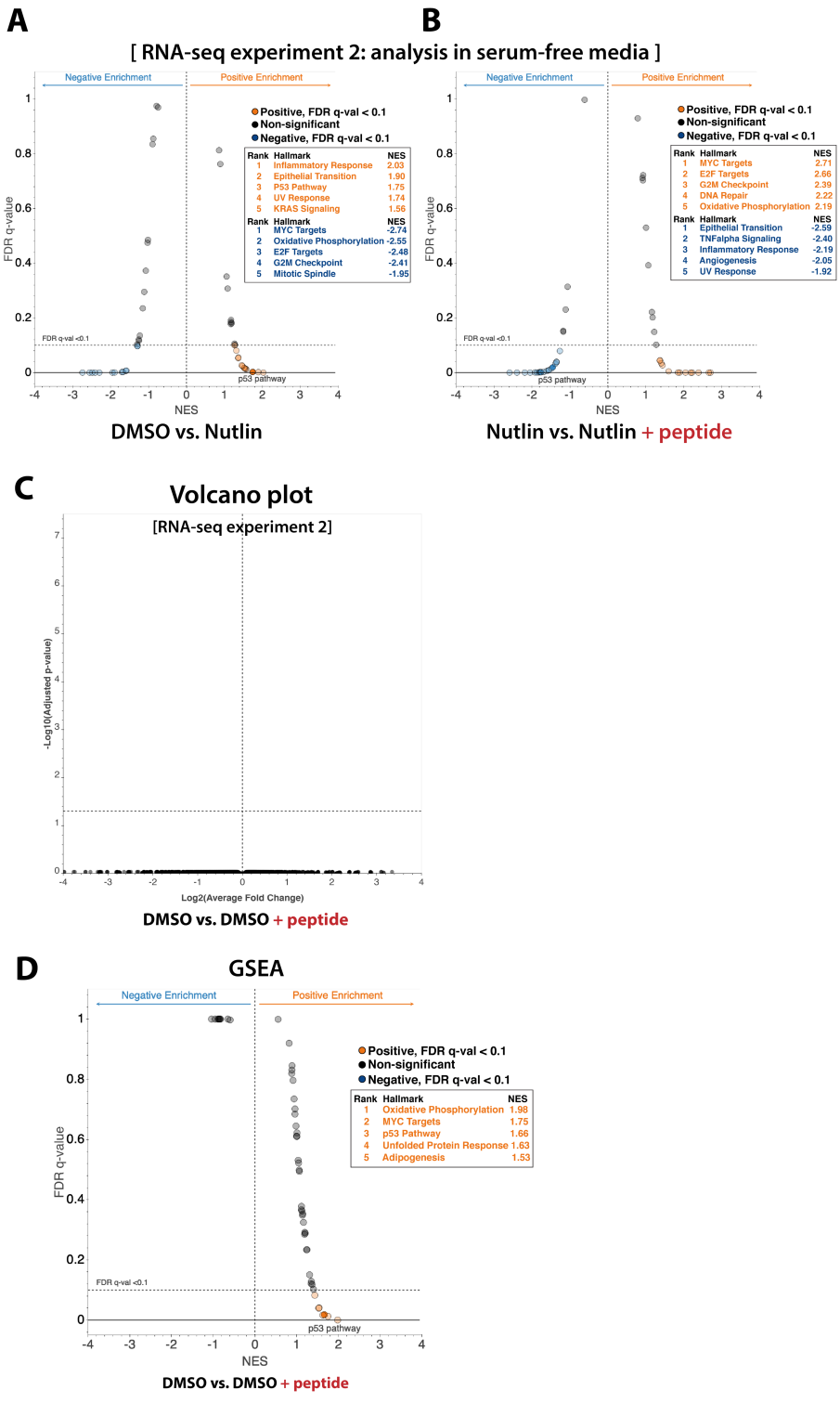


Figure 3.8: **Additional RNA-seq results and supporting RT-qPCR data.** (A, B) GSEA moustache plots (RNA-seq experiment 1) that show (A) Nutlin treatment strongly induces p53 pathway activation (NES: 3.07), as expected, and that (B) in Nutlin-treated cells, the bivalent peptide inhibits p53 pathway activation (NES: -2.13). (C, D) RT-qPCR results are consistent with RNA-seq data and show suppression of p53 target gene expression by the bivalent peptide (C) or negligible effects in the absence of Nutlin-dependent p53 activation (D). Bars represent standard error of the mean (n = 2).



Continued on next page.

Figure 3.9: **Bivalent peptide blocks p53 response in Nutlin-treated HCT116 cells, but causes no significant changes in pol II transcription in absence of p53 activation (RNA-Seq experiment 2); evidence for weak p53 activation in DMSO control + peptide experiments.** (A, B) GSEA moustache plots (RNA-seq experiment 2) show (A) Nutlin treatment induces p53 pathway activation (NES: 1.75), and that (B) in Nutlin-treated cells, the bivalent peptide inhibits p53 pathway activation (NES: -1.45). Note that RNA-seq experiment 2 (biological triplicate samples) differed from RNA-seq experiment 1 (biological replicate samples) in that serum-free media was used, based upon reports that it could enhance peptide uptake by cells [39, 46]. In retrospect, this likely ensured a weak Nutlin response due to p53 activation triggered by serum removal [28, 216]. Consistent with this notion, the p53 pathway activation was markedly reduced in Nutlin-treated cells (NES: 1.75) compared with experiment 1, with GSEA NES = 3.07 for p53 pathway in Nutlin-treated cells (Figure S3.8A). Despite these limitations, GSEA results from RNA-seq experiment 2 show suppression of p53 activation by the bivalent peptide. (C) Volcano plot showing that the bivalent peptide causes no significant changes in pol II transcription in the absence of p53 activation (RNA-seq experiment 2). These results are similar to RNA-seq experiment 1 (Figure 3.3E). (D) GSEA moustache plot (RNA-seq experiment 1) comparing DMSO control conditions ± bivalent peptide. Note a weak p53 pathway activation (NES=1.66) that is consistent with in vitro data (Figure S3.5F); in cells, the weak activation could reflect a modest disruption of p53-MDM2 interactions, but this remains to be rigorously tested.

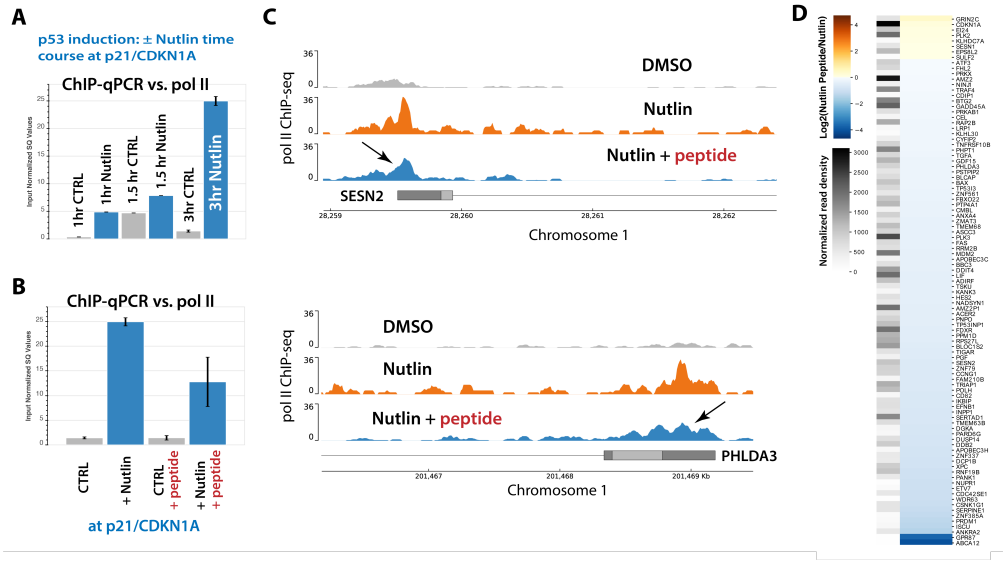


Figure 3.10: **Summary of pol II CTD Ser5P ChIP-seq data.** (A) Time-course ChIP- qPCR that provided the basis for a 3hr post-Nutlin time point for ChIP-seq experiments. This time point is identical to that used for RNA-seq. (B) ChIP-qPCR at the p21/CDKN1A locus, showing i) increased pol II occupancy in Nutlin-stimulated cells but ii) reduced pol II occupancy in cells treated with the bivalent peptide. (C) Example pol II CTD Ser5P ChIP-seq traces at p53 target gene promoters. Consistent with the RNA-seq data, pol II occupancy is increased in Nutlin-treated cells, but pol II occupancy is reduced in cells treated with the bivalent peptide. (D) Summary heatmap at core p53 target genes [9] showing overall reduced pol II occupancy at promoter regions (see Methods). The heatmap represents the log2(Nutlin+Peptide/Nutlin adjusted read counts) in which blue indicates a reduction of pol II occupancy in the Nutlin+peptide versus. Nutlin samples. The separate black- and-white heatmap represents the average Signal-to-Noise Ratio (SNR) normalized read counts, derived from the CHIPIN method [193]. Promoters with higher read density will have darker shading. (E) Ser5P pol II ChIP-seq metagene at the TSSs of a subset of p53 target genes. Note that ChIP-seq provides data at a snap-shot in time, whereas RNA-seq represents cumulative effects over time. Thus, we observe two populations of responses at p53 target genes [9] as seen by ChIP-seq (t = 3hr) in which the adjusted read density in Nutlin+Peptide sample compared to Nutlin is either unchanged or higher (27 TSSs) or lower (70 TSSs). This probably reflects differential induction of p53 target genes at the t = 3hr time point and is also reflected in the heatmap in panel D. The ChIP-seq metagene shown is derived from the set of 70 TSSs and shows that Nutlin + bivalent peptide has lower Ser5P pol II occupancy compared with Nutlin, which is overall consistent with the RNA-seq data.

Chapter 4

The naturally occurring Δ40p53 isoform inhibits eRNA transcription and enables regulation by signal-specific transcription factors during p53 activation

The work in this chapter is as published in: Levandowski CB, Jones T, Gruca M, Ramamoorthy S, Dowell RD, Taatjes DJ. The Δ40p53 isoform inhibits p53-dependent eRNA transcription and enables regulation by signal-specific transcription factors during p53 activation. PLOS Biology 2021;19:e3001364. https://doi.org/10.1371/journal.pbio.3001364[142]. All supplemental tables can be found with the published manuscript. Raw sequencing data can be found at GSE147703 and GSE227931. Figures 4.5-4.23 were published as supplemental figures in published manuscript. Methods are discussed in Appendix B.3.

4.1 Contribution Statement

The lead scientist for this project was Dr. Cecilia Levandowski, and it embodied the bulk of her graduate work. I joined this project as a junior graduate student to learn from Dr. Levandowski. On this project from the wet-lab perspective I took over cell culture work, aided in growth rate, flow-cytometry and metabolimics experiments and aided in preparing PRO-seq, RNA-seq and ChIP-seq libraries. I aided in the TP63 knockdown experiments and individually completed all of the follow up qPCR. I also completed all of the validation qPCR of two additional CRISPR clones per genome edited MCF10A cell line. From the dry-lab perspective I analyzed the bulk of the data in the publication with the help of Margaret Gruca.

In terms of NGS library preparation I produced two additional PRO-seq libraries in MCF10A

cells. There are a total of four conditions: WT MCF10A +/- Nutlin-3a and p53 null MCF10 +/- Nutlin-3a. One biological replicate was produced with the help of Dr. Levandowski, the other I produced independently. With these samples I also ran qPCR experiments to validate that the WT sample showed an expected p53 response to Nutlin-3a but the p53 null sample did not. I helped produce an additional RNA-seq data set in the three genome edited MCF10A cells +/- 5-fluorouracil (5FU). Finally, I optimized and helped perform the p53 ChIP-seq experiments in the three genome edited MCF10A cells +/- Nutlin-3a. This optimization entailed a complete ChIP protocol optimization that integrated techniques used by four separate academic labs.

In terms of data analysis, I aided or completed analysis for all three NGS library types. For PRO-seq, I performed the bidirectional calling and analysis[18]. Using the curated bidirectionals I ran transcription factor enrichment analyses, including that done by TFEA[207]. I ran the differential expression analysis by DESeq2 and pathway analysis by GSEA in conjunction with Ms. Gruca. I performed the in depth analyses comparing the DESeq2 and GSEA results across conditions, as well as generating all finalized figures related to PRO-seq for this manuscript. Within the PRO-seq samples I validated that each sample had the expected coverage of the p53 trans-activation-domain 1, that is missing in the Δ40p53 isoform. In RNA-seq, I performed mapping, trimming, quality control, differential analysis and pathway analysis in conjunction with Ms. Gruca. I fully analyzed an additional, publicly available data set [251] with 5-FU treatment to corroborate our findings within this manuscript. I performed the in-depth comparisons between the RNA-seq and PRO-seq results. Finally, I generated all finalized figures related to RNA-seq for this manuscript. In regards to ChIP-seq, I performed all mapping, trimming, quality control and analysis. I generated all finalized figures related to ChIP-seq for this manuscript. Beyond that, I played a major role in data interpretation along with Dr. Levandowski to produce the finalized manuscript.

4.2 Abstract

The Δ40p53 isoform heterotetramerizes with WTp53 to regulate development, aging, and stress responses. How Δ40p53 alters WTp53 function remains enigmatic because their co-expression causes tetramer heterogeneity. We circumvented this issue with a well-tested strategy that expressed Δ40p53:WTp53 as a single transcript, ensuring a 2:2 tetramer stoichiometry. Human MCF10A cell lines expressing Δ40p53:WTp53, WTp53, or WTp53:WTp53 (as controls) from the native TP53 locus were examined with transcriptomics (PRO-seq, RNA-seq), metabolomics, and other methods. Δ40p53:WTp53 was transcriptionally active and although phenotypically similar to WTp53 under normal conditions, it failed to induce growth arrest upon Nutlin-induced p53 activation. This occurred via Δ40p53:WTp53-dependent inhibition of eRNA transcription and subsequent failure to induce mRNA biogenesis, despite similar genomic occupancy to WTp53. A different stimulus (5-fluorouracil) also showed Δ40p53:WTp53-specific changes in mRNA induction; however, other transcription factors (e.g. E2F2) could then drive the response, yielding similar outcomes versus WTp53. Our results establish that Δ40p53 tempers WTp53 function to enable compensatory responses by other stimulus-specific transcription factors. Such modulation of WTp53 activity may be an essential physiological function for Δ40p53. Moreover, Δ40p53:WTp53 functional distinctions uncovered herein suggest an eRNA requirement for mRNA biogenesis and that human p53 evolved as a tetramer to support eRNA transcription.

4.3 Introduction

Transcription factors (TFs) are the primary drivers of cell state and cell physiology[140]. As a testament to their biological importance, an entire population of fibroblasts will form myotubes upon expression of a single TF, MyoD[223]. As a TF, p53 coordinates cellular stress responses and also plays key roles in cancer, aging, and stem cell biology[121, 129, 132]. Regulation of p53 function across these diverse biological circumstances involves an array of mechanisms, with well-studied examples that include changes in post-translational modification[93] or protein stability[100, 132, 133].

Less well understood are p53 isoforms[30], which represent naturally occurring truncated products that result from alternative splicing from the native TP53 locus, or from altered translation of its mRNA. Among the p53 isoforms that have been identified, the Δ40p53 isoform is arguably the most biologically relevant, yet also remains one of the most enigmatic[7]. Isoform Δ40p53 has many aliases such as p44, p53/p47, and ΔNp53; Δ40p53 lacks only the N-terminal 39 amino acids, which encompass the first p53 activation domain (AD1). All other p53 domains, including the oligomerization and DNA-binding domains, are retained. AD1 is a key p53 domain that drives most p53 transcriptional responses in vivo[118, 119] and is required for stable recruitment of co-activators such as Mediator and CBP/p300. AD1 is also recognized by the MDM2 protein, an E3 ubiquitin ligase that negatively regulates p53 function through proteasomal degradation[100, 133]. In addition to AD1, the p53 N-terminus contains a second activation domain, AD2. Loss of both AD1 and AD2 or mutation of key residues in each domain (L22, W23 in AD1; L53, W54 in AD2) results in complete loss of p53 TF function and resembles a p53-null phenotype[36]. Notably, p53AD2 is capable of activating a subset of p53 transcripts in the presence of L22, W23 AD1 mutations, to support induction of senescence and tumor suppression[36, 117]. In the absence of WTp53, the Δ40p53 isoform is transcriptionally inactive in vitro, despite forming stable tetramers[151]. Moreover, Δ40p53 expression in a p53-null background does not alter the p53-null phenotype[161]. Thus, Δ40p53 requires WTp53 to impact p53 transcriptional responses and to cause phenotypic changes. This was best exemplified in mouse studies in which truncated or naturally occurring Δ40p53 isoforms were expressed together with WTp53 in roughly a 1:1 ratio[161, 226]. These "Δ40p53 + WTp53" mice adopted an accelerated aging phenotype, which, in addition to premature death, involved physiological changes observed with normal aging, such as increased senescence, early-onset osteoporosis, and memory loss[103, 161, 188]. Whereas the molecular and cellular mechanisms by which Δ40p53 acts to alter WTp53 function remain unclear, it involves formation of hetero-dimers with WTp53. The Δ40p53 isoform forms mixed Δ40p53:WTp53 tetramers when expressed together with WTp53[55, 65]. Generation of Δ40p53 can occur by alternate splicing; however, the primary mechanism appears to be through alternate translation via an internal ribosomal entry

site (IRES)[242, 253]. Cellular levels of Δ40p53 have been shown to increase in response to diverse types of stress[31, 86, 177, 190], suggesting that Δ40p53 levels naturally fluctuate throughout the lifespan of an organism. Whereas enforced co-expression of Δ40p53 with WTp53 causes accelerated aging in mice[161], the direct correlation between cellular stress and Δ40p53 levels suggests a mechanism whereby chronic stress may cause transcriptional changes that contribute to mammalian aging[75]. A major barrier to understanding how Δ40p53 affects WTp53 function is that co-expression (i.e. Δ40p53 + WTp53) will confound analyses due to tetramer heterogeneity, including formation of "contaminating" WTp53 tetramers. Under such circumstances, the activity of Δ40p53:WTp53 tetramers cannot be de-coupled from WTp53 function. The cryo-EM structure of the WTp53 tetramer[186] revealed a straightforward means to link Δ40p53 and WTp53 as a single transcript while preserving p53 tetramer structure (Fig. 4.1A). As a proof-of-principle, we tested the function of tethered Δ40p53:WTp53 tetramers (versus standard WTp53 tetramers) in a series of biochemical and cell-based experiments[151]. As expected, the flexible sequence linking Δ40p53 with WTp53—which was longer than necessary to enable conformational flexibility—did not affect p53 activity. For example, tethered p53 tetramers purified exactly as WTp53 tetramers (e.g. identical over a size-exclusion column) and tethered versions of WTp53 (i.e. WTp53:WTp53) mimicked phenotypic and gene expression changes induced by WTp53 in H1299 cells[151]. In fact, the global gene expression changes (mRNA) induced by WTp53 versus WTp53:WTp53 were essentially identical[151]. These results affirmed the tethering strategy and served as a "proof of concept" for the more rigorous analysis described here, in which we used CRISPR-Cas9 to generate homozygous knock-in cell lines that expressed WTp53, Δ40p53:WTp53 or WTp53:WTp53 from the native TP53 locus. Using PRO-seq, we measured rapid transcriptional responses following p53 activation, whereas RNA-seq probed subsequent changes in steady-state mRNA levels, and these transcriptional responses were linked to p53 occupancy using ChIP-seq. Combined with metabolomics and phenotypic assays, we have better defined how Δ40p53 alters WTp53 function in human cells. Notably, the Δ40p53 isoform tempers WTp53 activity, which allows other TFs to drive stimulus-specific responses. We also uncovered unexpected aspects of Δ40p53 function that

link eRNA transcription and mRNA biogenesis, and that suggest that four complete p53 activation domains (i.e. AD1 + AD2) must occupy a p53 binding site to induce eRNA transcription.

4.4 Results

4.4.1 Generation of genome-edited cell lines

Three distinct genome-edited (CRISPR-Cas9) MCF10A cell lines were generated: a WTp53 control, in which WTp53 was simply inserted back into the native TP53 locus, a WTp53:WTp53 control, to probe for potential tether-specific effects, and Δ40p53:WTp53 (Fig. 4.1A). In each case, the p53 cDNA sequence was inserted at the first translational start site in exon 2 of the TP53 gene, to ensure expression would be controlled through the native p53 promoter (Fig. 4.5A,B). We chose MCF10A cells because they endogenously express WTp53 and are derived from non-tumorigenic mammary tissue. MCF10A cells therefore have a stable genome that is not prone to mutations or polyploidy. Edited cells were sorted based on mCherry selection, and single cell clones were expanded and verified homozygous using PCR (Fig. 4.5C), western blot (Fig. 4.5D), and sequencing (Fig. 4.6). The endogenous regulation of p53 expression in genome-edited cells was tested by treating cells with the small molecule Nutlin-3a, which disrupts the p53 interaction with MDM2[229]. Consequently, Nutlin-3a activates p53 and increases its protein levels. As shown in Figure 4.1B, Nutlin-3a increased p53 protein levels in all three genome-edited cell lines, confirming each was regulated similar to WTp53 in non-edited MCF10A cells. Additional verification results are shown in Figure 4.7. Together, these three MCF10A cell lines provided a means to evaluate Δ40p53 function under physiologically relevant conditions. Because Δ40p53:WTp53 was expressed as a single transcript, a fixed 2:2 tetramer stoichiometry was assured, avoiding tetramer heterogeneity that results from Δ40p53 + WTp53 co-expression.

4.4.2 WTp53 and Δ40p53:WTp53 cells are phenotypically similar under normal growth conditions

In non-stressed conditions, cellular p53 activity is typically very low. All three cell lines (WTp53, WTp53:WTp53, and Δ40p53:WTp53) were derived from the same parental MCF10A line and were therefore isogenic except at the TP53 locus. Under normal growth conditions, no significant change in cell cycle was observed between WTp53, WTp53:WTp53, or Δ40p53:WTp53 cells (Fig. 4.8A). Furthermore, the growth rate of each cell line was similar (Fig. 4.1C); whereas growth was slightly enhanced in Δ40p53:WTp53 cells, the increase was not statistically significant. As a tumor suppressor, p53 significantly impacts cell metabolism[132]. Therefore, we compared the metabolomes of WTp53, WTp53:WTp53, and Δ40p53:WTp53 cells. Consistent with the cell cycle and cell growth assays, untargeted metabolomics experiments confirmed that each p53 knock-in cell line had similar (but not identical) basal levels of metabolites (Supplemental Table 1). Metabolites relevant to cell cycle progression are shown in Figure 4.8B.

4.4.3 Nutlin-3a exposes phenotypic changes in Δ40p53:WTp53 cells (versus WTp53)

We next examined how each cell line (WTp53, WTp53:WTp53, or Δ40p53:WTp53) would respond to p53 activation by Nutlin-3a. Notably, Nutlin-3a is non-genotoxic and highly specific for p53[9]; thus, unlike a typical physiological stimulus (e.g. DNA damage), auxiliary pathways are not activated by Nutlin-3a. As shown in Figure 4.1D, p53 activation by Nutlin-3a triggered a marked increase in G1 and a decrease in S and G2 phases in WTp53 cells. These results are characteristic of G1 arrest and represent a typical p53 response[121]. Also as expected, WTp53:WTp53 cells were similar to WTp53 cells (Fig. 4.9A). By contrast, cell cycle data for Nutlin-treated Δ40p53:WTp53 cells resembled control-treated (DMSO) cells, with only a modest Nutlin-dependent shift in G1 that did not reach statistical confidence of $p \leq 0.01$ (Fig. 4.1D). A p53 target gene that drives cell cycle arrest is CDKN1A (a.k.a. p21), and p21 can be induced by mixed Δ40p53:WTp53 tetramers, as shown by us[151] and others[161]. We measured p21 protein levels in Nutlin-treated WTp53

versus Δ40p53:WTp53 cells and observed increases in each line (Fig. 4.1E; WTp53:WTp53 in Fig. 4.9B). Although p21 protein levels were greater in WTp53 cells, it was evident that Nutlin-3a activated Δ40p53:WTp53 tetramers, consistent with increased Δ40p53:WTp53 protein levels upon Nutlin-3a treatment (Fig. 4.1B). Parallel experiments in p53-null MCF10A cells confirmed that the Nutlin-dependent increase in p21 protein was p53-dependent (Fig. 4.1E). Also as expected, cell cycle changes were not observed in p53-null cells in response to Nutlin-3a treatment, nor was there a change in p21 mRNA levels (Fig. 4.9C). Others have shown that cell stress transiently increases Δ40p53 protein levels [31, 55, 86, 169, 177] and this has implications for aging, as chronic induction of Δ40p53 over time may contribute to physiological aging. To simulate transient periods of cell stress (i.e. p53 pathway activation), we subjected each cell line (WTp53, WTp53:WTp53, or Δ40p53:WTp53) to repeated cycles of Nutlin-3a treatment (20 hr), followed by 48 hr recovery. As shown in 4.9D Figure, WTp53:WTp53 cells responded similarly to WTp53, as expected. By contrast, a stark difference was observed between WTp53 and Δ40p53:WTp53 cells. Growth was arrested in WTp53 cells, but proliferation continued in Δ40p53:WTp53 cells, even after multiple rounds of Nutlin treatment (Fig. 4.9D). This result contrasts with untreated (DMSO) WTp53 versus Δ40p53:WTp53 cells, which showed no significant difference in growth rate (Fig. 4.1C). Because the metabolic needs for proliferating versus arrested cells will be distinct, we also compared the metabolomes of WTp53 versus Δ40p53:WTp53 cells following Nutlin-3a treatment (Supplemental Table 2). As expected, Nutlin-treated WTp53 cells showed metabolic changes that reflected their reduced proliferation (Supplemental Table 2; similar results in WTp53:WTp53 cells). By contrast, metabolic differences evident in Δ40p53:WTp53 cells were consistent with their maintenance of proliferation and were observed even prior to Nutlin stimulation (Supplemental Table 1). For instance, decreased levels of sphingosine or increased levels of sphingomyelin metabolites in Δ40p53:WTp53 cells (Fig. 4.10) is each independently consistent with maintenance of cell cycle and proliferation[96]. Conversely, WTp53 and WTp53:WTp53 cells showed the opposite trends in these metabolites (versus Δ40p53:WTp53), consistent with the induction of cell cycle arrest upon Nutlin treatment. Although these results could reflect a complete lack of p53 pathway activation

by Nutlin-3a in Δ40p53:WTp53 cells, this did not appear to be the case based upon data shown in Figure 4.1B (Nutlin-dependent increase in Δ40p53:WTp53 levels) or Figure 4.1E (p21 protein induction). To further probe the underlying mechanisms, we next assessed how Nutlin-dependent p53 activation affected the transcriptomes of Δ40p53:WTp53 versus WTp53 cells.

4.4.4 Δ40p53:WTp53 differentially affects the pol II transcriptome upon Nutlin-3a treatment

To compare and contrast transcriptional changes in Δ40p53:WTp53 cells (versus WTp53), we used Precision nuclear Run-On sequencing (PRO-seq), which measures nascent transcription genome-wide[135]. PRO-seq detects transcripts from all three RNA polymerases and measures all types of pol II transcripts, including non-coding RNAs and non-annotated regions. Cells (Δ40p53:WTp53, WTp53, and WTp53:WTp53) were treated with Nutlin-3a for three hours, and differential transcription was quantified using DEseq2[8]. The 3 hour time point was determined empirically (Fig. 4.11), with reference to Nutlin-treated HCT116 cells that we evaluated previously[6]. This early time point following Nutlin stimulation favored identification of direct (i.e. primary) p53 transcriptional targets. In WTp53 cells, 4607 annotated regions were differentially transcribed (p-value $\leq$ 0.01) after Nutlin-3a treatment (Fig. 4.2A), with similar transcriptional changes in WTp53:WTp53 cells (Fig. 4.12). By contrast, only 315 annotated regions were differentially transcribed (p-value $\leq$ 0.01) in Δ40p53:WTp53 cells after Nutlin-3a treatment (Fig. 4.2A). Of these 315 Nutlin-induced transcripts in Δ40p53:WTp53 cells, 298 were also observed in WTp53 cells, and only 17 annotated genes were differentially transcribed in Δ40p53:WTp53 versus WTp53 cells (Fig. 4.13A), with most linked to cell growth. We confirmed that a significant p53 response was occurring in Δ40p53:WTp53 cells by completing parallel PRO-seq experiments in p53-null MCF10A cells (Fig. 4.13B-D). Thus, the magnitude of Nutlin-3a induction was simply greater in WTp53 cells versus Δ40p53:WTp53, with examples shown in Figure 4.13E. Gene Set Enrichment Analysis (GSEA) of the Nutlin-induced genes revealed substantial differences, with pathways associated with growth and cell cycle progression increased in Δ40p53:WTp53 cells (Fig. 4.2B). Enrichment of

the E2F pathway in Nutlin-treated Δ40p53:WTp53 cells (versus WTp53) was notable because the E2F TF family has overlapping function with p53[192]. The reduced p53 pathway activity (GSEA, Fig. 4.2B) was consistent with dampened response in Δ40p53:WTp53 cells compared with WTp53. An Integrated Pathway Analysis (IPA) of the Nutlin-induced genes showed results consistent with GSEA, with growth and cell cycle pathways enhanced in Δ40p53:WTp53 cells (Fig. 4.13F). In particular, mTOR signaling and IGF-1 signaling were increased in Δ40p53:WTp53 cells compared to WTp53; each of these pathways has been linked to Δ40p53 expression in human cells[151] or mouse models[161], as potential contributors to the Δ40p53 accelerated aging phenotype. Although the transcriptional changes summarized in Figure 4.13A may contribute to the increased proliferation of Δ40p53:WTp53 cells (versus WTp53) after Nutlin-3a treatment (Fig. 4.9D), we emphasize that the PRO-seq data were obtained after only 3 hr treatment; changes in steady-state mRNA levels (i.e. RNA-seq) after a longer period of Nutlin-3a treatment are described later.

4.4.5 Δ40p53:WTp53 tetramers fail to induce eRNA transcription

In addition to gene expression changes in annotated regions, the PRO-seq data revealed stark differences in transcription of eRNAs. Whereas WTp53 cells had 510 differentially transcribed (p-value ≤ 0.01) eRNAs after 3 hr Nutlin-3a treatment, only 118 eRNAs changed in Δ40p53:WTp53 cells. Among the 510 WTp53 eRNAs (Fig. 4.2C), 109 mapped to p53 binding sites identified by ChIP-seq (see below; similar results in WTp53:WTp53 cells, Fig. 4.14). These 109 p53-associated eRNAs are indicative of direct eRNA activation by p53[6, 170]. By contrast, only 6 eRNAs mapped to p53 binding sites in Δ40p53:WTp53 cells, revealing a defect in its ability to induce eRNA transcription. The examples shown in Figure 4.2D and Figure 4.15 further highlight the contrast in eRNA induction between WTp53 and Δ40p53:WTp53 tetramers. We note that Nutlin-induced eRNAs that mapped to sites not bound by p53 likely represent weak p53 binding (i.e. below cutoff) or secondary effects from p53 activation. Whereas the biological roles for eRNA transcription remain unclear, they display cell-type specific expression patterns[13] and rapidly respond to external stimuli. Most relevant to this study, eRNAs are transcribed in response to p53 binding

events[6] and correlate with expression of p53 target genes[170]. To further probe eRNA transcriptional changes, we applied Transcription Factor Enrichment Analysis (TFEA), an improved computational method[207] that can detect changes in bidirectional eRNA transcription. TFEA then maps these eRNAs to the genome and identifies any underlying TF binding motifs. In this way, TFEA can accurately identify which TFs are being activated or repressed during a stimulus. As shown in Figure 4.16A, TFEA revealed a robust p53 activation in Nutlin-treated WTp53 cells (and WTp53:WTp53 cells, Fig. 4.16B), as expected. In Δ40p53:WTp53 cells, however, no evidence of p53 activation was apparent (Fig. 4.16C); moreover, a TFEA comparison of WTp53 versus Δ40p53:WTp53 cells indicated a defect in p53 activation by Δ40p53:WTp53 tetramers (Fig. 4.16D). Metagene analyses further illustrated the contrast between eRNA transcription in Nutlin-treated WTp53 versus Δ40p53:WTp53 cells (Fig. 4.2E; WTp53:WTp53 data in Fig. 4.16E). In agreement with the TFEA results, ChIP-seq analyses confirmed that Δ40p53:WTp53 binding fails to induce eRNA transcription, in direct contrast to WTp53 (see below). Collectively, these results establish that Δ40p53:WTp53 tetramers fail to induce eRNA transcription upon p53 activation.

4.4.6 Genomic occupancy of Δ40p53:WTp53 is identical to WTp53

Because eRNA induction is dependent upon TF binding[6, 17, 81], it was possible that eRNAs were not impacted in Δ40p53:WTp53 cells because Δ40p53:WTp53 did not bind p53 sequences, even after Nutlin treatment. However, Δ40p53 contains the entire DNA-binding domain (residues 101-300); therefore, we expected that the genomic occupancy of Δ40p53:WTp53 tetramers would resemble that of WTp53. As shown in Figure 4.2F and Figure 4.17A,B, ChIP-seq data revealed that Δ40p53:WTp53 bound the same sites as WTp53, genome-wide, as expected given their identical DNA-binding domains. The corresponding eRNAs are shown in Fig. 4.2D. Consistent with TFEA results (Fig. 4.16C) these data show that Δ40p53:WTp53 binding fails to induce eRNA transcription, in contrast to WTp53. Metagene analyses showed that occupancy of both WTp53 and Δ40p53:WTp53 increased after Nutlin-3a treatment, as expected (Fig. 4.2G; WTp53:WTp53 shown in Fig. 4.17C). No Δ40p53:WTp53-specific binding sites were detected. Note that the

antibody used for ChIP-seq recognizes the p53 N-terminus, which is absent in Δ40p53; thus, Δ40p53:WTp53 possesses 50% fewer epitopes per tetramer, and this will reduce its overall signal. Nevertheless, occupancies of WTp53:WTp53 and Δ40p53:WTp53 appeared nearly identical under basal or Nutlin-treated conditions (Fig. 4.17B). Because ChIP-seq is qualitative, we cannot draw conclusions about the relative levels of p53 binding across the cell lines. Moreover, proteins with increased molecular weight (e.g. WTp53:WTp53) are more susceptible to degradation during sonication[187], which may have decreased ChIP efficiency in tethered p53 cell lines. Despite these caveats, it was evident that 1) Δ40p53:WTp53 was appropriately mobilized in response to Nutlin-3a treatment, similar to WTp53, that 2) Δ40p53:WTp53 occupied the same p53 binding sites (versus WTp53), genome-wide, under basal and Nutlin-induced conditions, and that 3) unlike WTp53, Δ40p53:WTp53 binding fails to induce eRNA transcription.

4.4.7 RNA-seq data suggest defective mRNA biogenesis for transcripts induced by Δ40p53:WTp53

We next completed biological replicate RNA-seq experiments, to compare cellular mRNA levels with the nascent RNA changes identified by PRO-seq. We treated each cell line (WTp53, Δ40p53:WTp53, and WTp53:WTp53) with Nutlin-3a for 20 hr. This time point was chosen to allow time for Nutlin-induced changes to manifest in the steady-state mRNA transcriptome. As shown in Figure 4.3A, 1132 genes were differentially expressed (p-value ≤ 0.01) in WTp53 cells, with similar results in WTp53:WTp53 cells (Fig. 4.18A). In stark contrast, Δ40p53:WTp53 cells showed essentially no Nutlin response at the mRNA level; CDKN1A/p21 was the only gene that was induced (Fig. 4.3A). The lack of p53 activation in Δ40p53:WTp53 cells is further highlighted by the heat map of p53 pathway genes, shown in Figure 4.18B. As with the PRO-seq data (Fig. 4.2B), GSEA identified the same pathways upon comparison of RNA-seq data from Nutlin-treated Δ40p53:WTp53 versus WTp53 cells (Fig. 4.3B). For instance, the inhibited (p53 pathway) and activated (G2M checkpoint, E2F targets) pathways were consistent, again supporting the enhanced proliferation of Δ40p53:WTp53 cells during Nutlin treatment (Fig. 4.9D). As shown in Figure

4.3C, approximately 10% of genes (439/4338) were identified as differentially expressed (p-value ≤ 0.01) in both the PRO-seq (nascent transcription, 3 hr) and RNA-seq (steady-state mRNA, 20 hr) experiments in Nutlin-treated WTp53 cells (WTp53:WTp53 data shown in S14C Fig). This percentage is roughly consistent with previous studies that compared GRO-seq and RNA-seq data in Nutlin-treated cells[9], and provides a benchmark for the efficiency of mRNA biogenesis during a Nutlin-induced p53 response. Although the number of differentially expressed nascent transcripts (n = 315) was reduced in Nutlin-treated Δ40p53:WTp53 cells at 3 hr (Fig. 4.3C), over 30 genes would be expected to be induced at the mRNA level, but this was not observed. These results suggest that mRNA biogenesis is defective for transcripts induced by Δ40p53:WTp53. Moreover, the mRNA data show striking parallels with bidirectional eRNA transcription. Like the mRNA transcriptome, changes in p53-dependent eRNA transcription were largely absent in Δ40p53:WTp53 cells.

4.4.8 The p53 paralogs p63 or p73 do not impact Δ40p53:WTp53 response

Analysis of PRO-seq and RNA-seq data revealed that whereas p63 was expressed in MCF10A cells, p73 was not (Fig. 4.19). Stable knockdown of p63 (Fig. 4.20A) in each cell line (WTp53, WTp53:WTp53, or Δ40p53:WTp53) did not impact proliferation or the cell cycle compared with controls (Fig. 4.20B,C); moreover, p63 was not induced by Nutlin-3a in any cell line (Fig. 4.19). Finally, RT-qPCR experiments showed that loss of p63 did not significantly impact CDKN1A/p21 or PUMA gene induction in Nutlin-treated WTp53, WTp53:WTp53, or Δ40p53:WTp53 cells (Fig. 4.20D). Collectively, these results suggest that p63 and p73 do not contribute to the differential phenotypic or transcriptional responses in Δ40p53:WTp53 cells.

4.4.9 WTp53 and Δ40p53:WTp53 support similar cellular responses to 5-fluorouracil, via distinct TFs

Finally, we asked whether a different p53 stimulus would cause the same transcriptional defects in Δ40p53:WTp53 cells; that is, a lack of response at the mRNA level. Because Nutlin-3a is highly selective for p53 (i.e. other TFs are not affected), it provided an efficient means to interrogate

p53-specific transcriptional effects. However, Nutlin-3a is not physiologically relevant, and typical stress responses activate multiple pathways and their respective signal-specific TFs. We therefore selected 5-fluorouracil (5FU), a well-studied, clinically relevant chemotherapeutic that not only activates the p53 pathway but also other signaling cascades associated with DNA damage[248]. WTp53, WTp53:WTp53, and Δ40p53:WTp53 cells were treated with 375 μM 5FU for 20 hr. Cell cycle analysis showed that 5FU caused similar changes in WTp53 versus Δ40p53:WTp53 cells (Fig. 4.4A), with a significant increase in S-phase and a significant decrease in G2 phase. As expected, WTp53 and WTp53:WTp53 cells also showed similar results upon treatment with 5FU (Fig. 4.21A). These results were not entirely surprising, since each cell line (WTp53, WTp53:WTp53, and Δ40p53:WTp53) was isogenic apart from the TP53 locus; thus, all auxiliary pathways were expected to respond similarly. To further compare and contrast the cellular responses to 5FU, we conducted biological replicate RNA-seq experiments. As shown in Figure 4.4B, 5FU caused differential expression of 926 genes in WTp53 cells and 823 genes in Δ40p53:WTp53 cells (p-value ≤ 0.01; WTp53:WTp53 data shown in Fig. 4.21B). Of these, 415 were shared among WTp53 and Δ40p53:WTp53 and many of these genes were p53 targets (Fig. 4.4B), as revealed by IPA. About half (408 out of 823) of the differentially expressed genes in Δ40p53:WTp53 cells were regulated by other TFs such as E2F1 (Fig. 4.4B). This result indicated that, in contrast to the targeted, p53-specific Nutlin response, the complex stress response induced by 5FU enabled other signal-specific TFs to augment the attenuated p53 response in Δ40p53:WTp53 cells. As shown in Figure 4.4C, mRNA levels of the TF E2F2 were selectively up-regulated in Δ40p53:WTp53 cells, suggesting that E2F2 helps compensate for the reduced p53 response in cells expressing Δ40p53:WTp53. In WTp53 cells, E2F2 mRNA levels decreased by 57% (versus DMSO, with a 10% decrease in WTp53:WTp53 cells) upon 5FU treatment. By contrast, E2F2 mRNA levels increased by 398% in Δ40p53:WTp53 cells (fold-change of 14.4), and E2F2 pathway genes were selectively up-regulated in Δ40p53:WTp53 cells (Fig. 4.4D, E;WTp53:WTp53 Fig. 4.22A). Interestingly, the E2F family of TFs have overlapping functions with p53, but p53 activation typically downregulates E2F gene expression[192]. E2F2 induction in 5FU-treated Δ40p53:WTp53 cells suggested that, instead of

p53, E2F and other TFs drive the transcriptional response toward cellular outcomes similar to WTp53. This concept was further supported by pathway analyses, which revealed that despite a different set of active and repressed TFs in Δ40p53:WTp53 cells versus WTp53 (Fig. 4.22B), the upstream regulators identified by IPA were similar (Fig. 4.22C). Consistent with these results, 5FU treatment caused similar cell cycle changes in WTp53 versus Δ40p53:WTp53 cells (Fig. 4.4A). To further probe p53-dependent effects on the 5FU response, and whether E2F and other TFs compensate for reduced p53 activity in Δ40p53:WTp53 cells, we completed cell cycle analyses in p53-null MCF10A cells following 20hr treatment with 375 μM 5FU (conditions identical to prior experiments). The results showed that cellular responses were similar to WTp53 (Fig. 4.22D), suggesting that, despite strong p53 activation by 5FU in WT cells, p53 is not required for cell cycle changes in this context. We next analyzed published RNA-seq data that compared 5FU-treated WTp53 and p53-null HCT116 cells[251]. Notably, GSEA identified similar pathways and TFs (e.g. E2F) activated by 5FU in p53-null cells (Fig. 4.22E), compared with Δ40p53:WTp53 cells (Fig. 4.4D), providing further support for a network of TFs that can compensate for p53 if its activity is reduced (e.g. in Δ40p53:WTp53 cells) or even lost (as in p53-null cells). Taken together, the results from 5FU-treated cells revealed that other signal-specific TFs can drive transcriptional programs under conditions with tempered p53 activity, such as in Δ40p53:WTp53 cells (Fig. 4.4F).

4.5 Discussion

The Δ40p53 isoform is naturally occurring, and its levels appear to increase during cell stress[31, 55, 86, 177, 190] or during specific developmental stages[227]. Because spontaneous TP53 mutations can yield a proliferative advantage in cultured cells[171], we chose the MCF10A cell line for this study because it is genetically stable (Fig. 4.6), unlike many cancer-derived cell lines. MCF10A cells also endogenously express WTp53, and we confirmed that WTp53 or WTp53:WTp53 cells were phenotypically similar to unedited MCF10A cells under normal or p53-stimulated conditions (Fig. 4.7). We also tested additional genome-edited cell clones to rule out potential effects from clonal expansion (Fig. 4.23). Whereas more cell lines could be tested with

identical sets of experiments, others have shown that basic p53 transcriptional responses are similar across cell types[9, 76], and p53 transcriptional response was the primary focus of this work. Based upon the transcriptomics data, several themes have emerged about how Δ40p53 alters WTp53 function (Fig. 4.4F). Expression of Δ40p53 1) modulates the transcriptional activity of WTp53, such that mixed Δ40p53:WTp53 tetramers suppress typical p53 transcriptional responses. Despite this, 2) Δ40p53:WTp53 tetramers retain the ability to activate p53 target genes, as seen most clearly in 5FU-treated cells (Fig. 4.4B) or upon comparison with p53-null cells (Fig. 4.13B-D). That is, Δ40p53:WTp53 tetramers do not functionally mimic p53-null conditions. This ability of Δ40p53 to dampen the p53 response 3) allows other sequence-specific, DNA-binding TFs to tune cellular stress responses, because they are not dominated by p53. For example, the gene expression signatures for the signal-specific TF E2F2 substantially contributed to the 5FU response in Δ40p53:WTp53 cells (Fig. 4.20, Fig. 4.4C-E). In this way, Δ40p53 allows augmentation of the p53 response by other stress- or signal-responsive TFs. This functional versatility may be especially important in specific contexts, cell types, or developmental stages that require a modified p53 response[33, 228]; for instance, to allow cell cooperation[61] and proliferation during embryonic development[227] or wound healing[234]. Furthermore, 4) the continual, enforced expression of Δ40p53 in the context of WTp53 (i.e. Δ40p53 + WTp53) causes accelerated aging in mice[161]. Whereas the biology of aging is complex[154], we observed that numerous pathways implicated in organismal aging were impacted in predictable ways by Δ40p53:WTp53. For example, the mTOR and IGF signaling pathways were up-regulated in Δ40p53:WTp53 cells (versus WTp53; Fig. 4.13F), and each pathway is broadly implicated in aging[27, 213]. Concomitant activation of p53 and the mTOR pathway was also noted upon transfection of Δ40p53:WTp53 in p53-null H1299 cells[151]; notably, simultaneous activation of p53 and mTOR can trigger senescence[62, 128]. Senescent cells accumulate during physiological aging[20] and, consistent with the aging phenotype, mice that co-express N-terminal p53 truncations with WTp53 exhibit early senescence-associated phenotypes[226].

4.5.1 Δ40p53 tempers WTp53 function, enabling other TFs to drive cellular processes

Related to item 3) and 4), above, the contrast between Nutlin-3a and 5FU is especially informative. Because Nutlin-3a is exquisitely specific for p53[9, 229], it was a valuable tool to selectively interrogate the p53 response. However, Nutlin stimulation is not physiologically relevant and does not activate other signal-specific TFs (i.e. only p53). In contrast, 5FU induces a DNA damage response that activates multiple signaling cascades. A clear distinction in mRNA biogenesis was observed in cells treated with 5FU. Hundreds of mRNAs were induced in Δ40p53:WTp53 cells after 20 hr treatment with 5FU, including many canonical p53 target genes (Fig. 4.4B). These results suggest that functional coordination among other signal-specific TFs helps drive mRNA biogenesis in Δ40p53:WTp53 cells. Whereas 5FU-treated Δ40p53:WTp53 cells showed diminished p53 activation (versus WTp53 cells), E2F2 expression levels (and its downstream target genes; Fig. 4.4D, E) increased significantly, revealing that other signal-responsive TFs augment cellular responses in the presence of Δ40p53:WTp53. Thus, the dampened p53 response in Δ40p53:WTp53 cells allows other TFs to drive the cellular response. These results are consistent with the concept of collaborative TF networks that enable compensatory responses if the function of typical "driver" TFs is compromised[152]. Such compensatory mechanisms appear to hard-wire cellular stress responses to ensure robust and consistent outcomes; this may be especially important in vivo, to coordinate transcriptional responses across organs and tissues. Notably, increased levels of Δ40p53 occur during ER stress[31, 177], serum stimulation after starvation[55], and oxidative stress[86]. Multiple TFs and signaling cascades are activated under these circumstances, and the presence of Δ40p53:WTp53 tetramers may be important to allow integration of other pathway- and signal-specific TFs to modulate the cellular response in cell-type or context-specific ways.

4.5.2 eRNA transcription and mRNA biogenesis

Many studies have linked eRNA transcription with mRNA expression[14], including in response to p53 activation[143, 170]. A general theme has been that increased eRNA transcription correlates with increased expression of protein-coding genes; moreover, the timing of eRNA induction implies a regulatory role, as eRNAs are rapidly induced after a stimulus, followed by increased mRNA levels of protein-coding genes. Here, we inadvertently established a system that, for the first time, allowed analysis of the transcriptional response to p53 activation in the presence or absence of p53-dependent eRNA induction, in largely isogenic cell lines. Our results implicate eRNA transcription as a pre-requisite for mRNA biogenesis. In support of this conclusion, eRNA transcription at p53 binding sites was blocked in Δ40p53:WTp53 cells (e.g. Fig. 4.16C)—despite similar genomic occupancy versus WTp53—and this tracked with mRNA levels. For instance, PRO-seq data showed that several hundred transcripts increased after 3 hour Nutlin treatment in Δ40p53:WTp53 cells (Fig. 4.2A), but only one gene (0.3%), CDKN1A/p21, increased at the mRNA level (Fig. 4.3A). By contrast, PRO-seq data from Nutlin-induced WTp53 cells showed an increase in 4607 transcripts, with 1132 increased at the mRNA level (25%; Fig. 4.3C). These results establish a direct connection between eRNA transcription and mRNA biogenesis (Fig. 4.4F). Indeed, the experiments described herein were i) completed in virtually isogenic cell lines, with ii) virtually the same TF (only lacking two of four AD1 domains for Δ40p53:WTp53), with iii) the same genomic occupancy and iv) under the same stimulus. The mechanistic basis remains to be determined, but eRNA-dependent regulation of RNA processing (e.g. splicing, cleavage, polyadenylation), or nuclear export could contribute.

4.5.3 Four complete p53 activation domains are required for eRNA transcription

Because Δ40p53:WTp53 was expressed from the native TP53 locus, the timing and the levels of induction matched WTp53; moreover, ChIP-seq data indicated that occupancy of Δ40p53:WTp53 versus WTp53 on genomic DNA was similar in basal and Nutlin-induced conditions. The lack of

p53-dependent eRNA induction by Δ40p53:WTp53 is therefore attributed to loss of only two of the four p53AD1 regions in the tetramer. That is, two activation domains are not sufficient for normal p53 tetramer function in cells. These findings suggest that p53 may have evolved to function as a tetramer (i.e. delivering four ADs to genomic DNA), at least in part, to induce eRNA transcription. Although additional experiments are needed to better define the molecular mechanisms that underlie this unexpected result, it is notable that p53 binding sites are more isolated in the human genome[230], whereas other TFs function within enhancers with clustered binding sites for many factors[105]. Potentially, four complete p53 activation domains (i.e. AD1 + AD2) are needed to stably recruit cofactors, such as Mediator, chromatin remodelers, and CBP/p300, to induce eRNA transcription. Alternately, four complete p53 activation domains might promote formation of molecular condensates that could help drive eRNA transcription by RNA polymerase II[84]. In support of this hypothesis, the activation domains of p53 are intrinsically disordered and p53 has been shown to form phase separated condensates in vitro[120] . These mechanisms (cofactor recruitment or condensate formation) are not mutually exclusive and may act synergistically in cells.

4.5.4 Implications for p53 tetramer structure

To date, no structural data exist for the entire p53 tetramer at atomic resolution, and it is evident that the p53 tetramer is structurally dynamic[110, 130]. This presents challenges for structural analysis, even with cryo-EM. The flexibly tethered p53 tetramers described herein (WTp53:WTp53 or Δ40p53:WTp53) were designed based upon the cryo-EM structure of the native p53 tetramer at intermediate resolution[186]. As far as we are aware, this structure by Orlova et al. is the only instance in which wild-type, full-length p53 was used, and it remains the highest resolution p53 tetramer structure to date (13.7 Å). Whereas alternate structural models of the p53 tetramer have been proposed[168, 224], these have used mutant versions of p53 (truncations or with four mutations in each core DNA-binding domain; thus, 16 mutations per tetramer) that showed evidence for structural instability and heterogeneity, and yielded lower resolution information. Previous proof-

of-concept biochemical and transcriptomics experiments demonstrated that the flexible tethers linking p53 monomers did not disrupt normal p53 tetramer function. For instance, WTp53:WTp53 tetramers matched gene expression changes induced by WTp53 in H1299 cells[151]. Similar results were obtained in genome-edited MCF10A cells here (Fig. 4.12), and ChIP-seq data showed similar genomic occupancy for WTp53 versus WTp53:WTp53 (Fig. 4.17). Likewise, MCF10A cells expressing WTp53 or WTp53:WTp53 were phenotypically indistinguishable under normal growth conditions or in response to Nutlin-3a or the genotoxic agent 5FU. These results best support the structural model of Orlova et al.[12, 186]. Nevertheless, we emphasize that our flexible tether strategy would enable formation of the "alternate" p53 tetramer structures as well[168, 224]. Because high-resolution data are lacking for WTp53 tetramers, its structural organization has remained controversial, and it remains plausible that full-length p53 may adopt multiple structural states that are functionally relevant.

4.5.5 CDKN1A/p21 induction and Δ40p53:WTp53 biological functions

Despite the inability of Δ40p53:WTp53 tetramers to induce eRNA transcription or mRNA production in Nutlin-treated cells, an exception was CDKN1A (a.k.a. p21), which is a well-established p53 target gene. CDKN1A inhibits the cyclin-dependent kinases (CDKs) CDK2 and CDK4/6, which phosphorylate RB-related proteins to promote G1/S cell cycle arrest. Whereas Nutlin-dependent activation of p53 triggered cell cycle arrest in WTp53 cells, this was not observed with Δ40p53:WTp53 (Fig. 4.1D). This phenotypic difference (p53 activation without cell cycle arrest) suggests Δ40p53 expression may be oncogenic; however, this was not observed in mouse models[161]. Notably, p21 is still induced in Nutlin-treated Δ40p53:WTp53 cells (although less compared with WTp53), at the protein and mRNA level (Fig. 4.1E, Fig. 4.3A). This low-level p21 expression is likely an important aspect of Δ40p53:WTp53 function in biology. Loss of p21 results in polyploidy, due to defects in DNA damage response and cycling in the absence of mitosis[237]. Low-level p21 expression by Δ40p53:WTp53 may help maintain genomic stability by retaining the mitotic checkpoint to prevent polyploidy. As a naturally occurring isoform, this basic function may

be essential while Δ40p53 regulates stress responses throughout the mammalian lifespan.

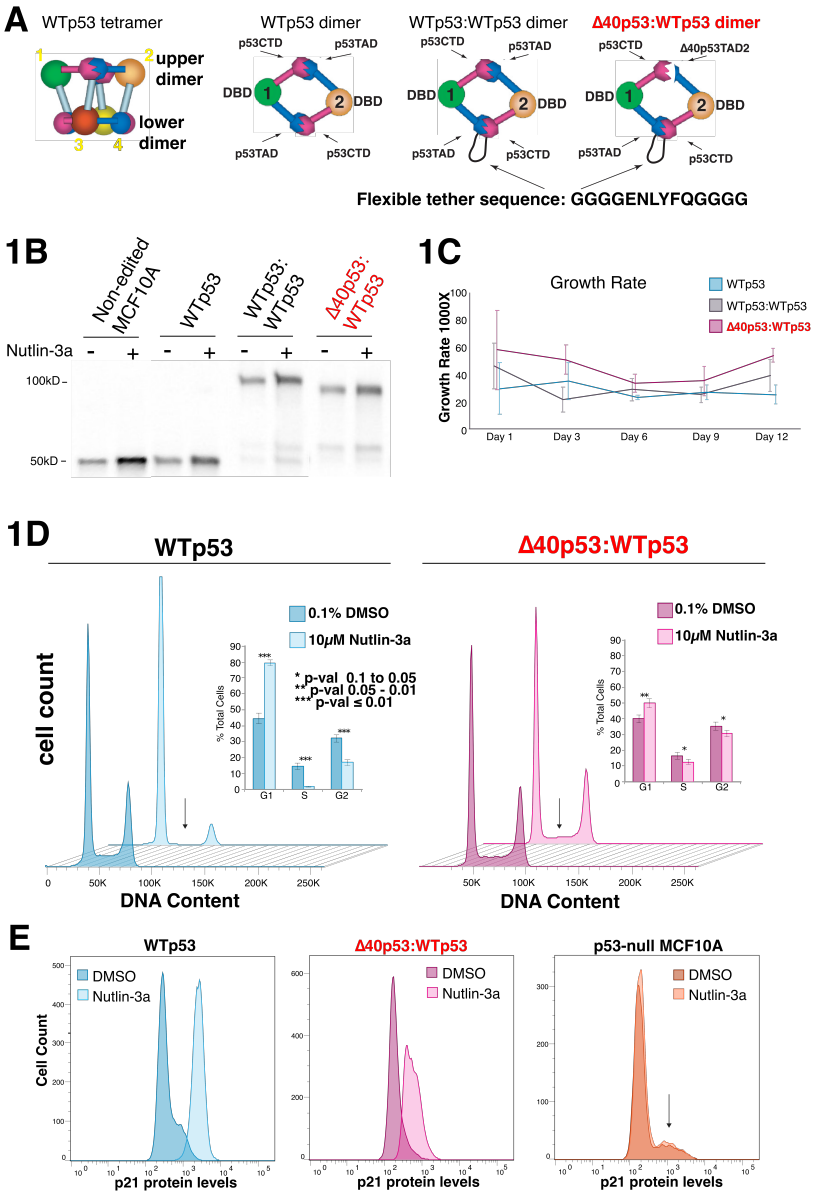


Figure 4.1: **Analysis of Δ40p53:WTp53 tetramers as a single entity; phenotypic comparisons under normal growth and Nutlin-induced conditions** (A) Schematic of WTp53 tetramer [186] and strategy to generate Δ40p53:WTp53 tetramers with a fixed 2:2 stoichiometry. Note the flexible tether is longer than necessary to allow conformational flexibility [151]. (B) Western blot to probe p53 levels before (-) or 6 hr after Nutlin-3a treatment. Non-edited MCF10A cells express WTp53 and are shown as a control. 20 μg total protein was loaded in each lane. (C) Growth rate measured over 5 treatment cycles; each cycle encompassed 20 hours growth under basal (0.1% DMSO) conditions, splitting cells 1:10, then growth for another 48 hours (3 biological replicates; bars = s.e.m.). (D) Cell cycle analysis (propidium iodide); chart (inset) represents the average of 6 biological replicates (bars = s.e.m.). Arrow highlights loss of S-phase in WTp53 cells, in contrast with Δ40p53:WTp53 cells. (E) Measurement of p21 protein levels by FACS.

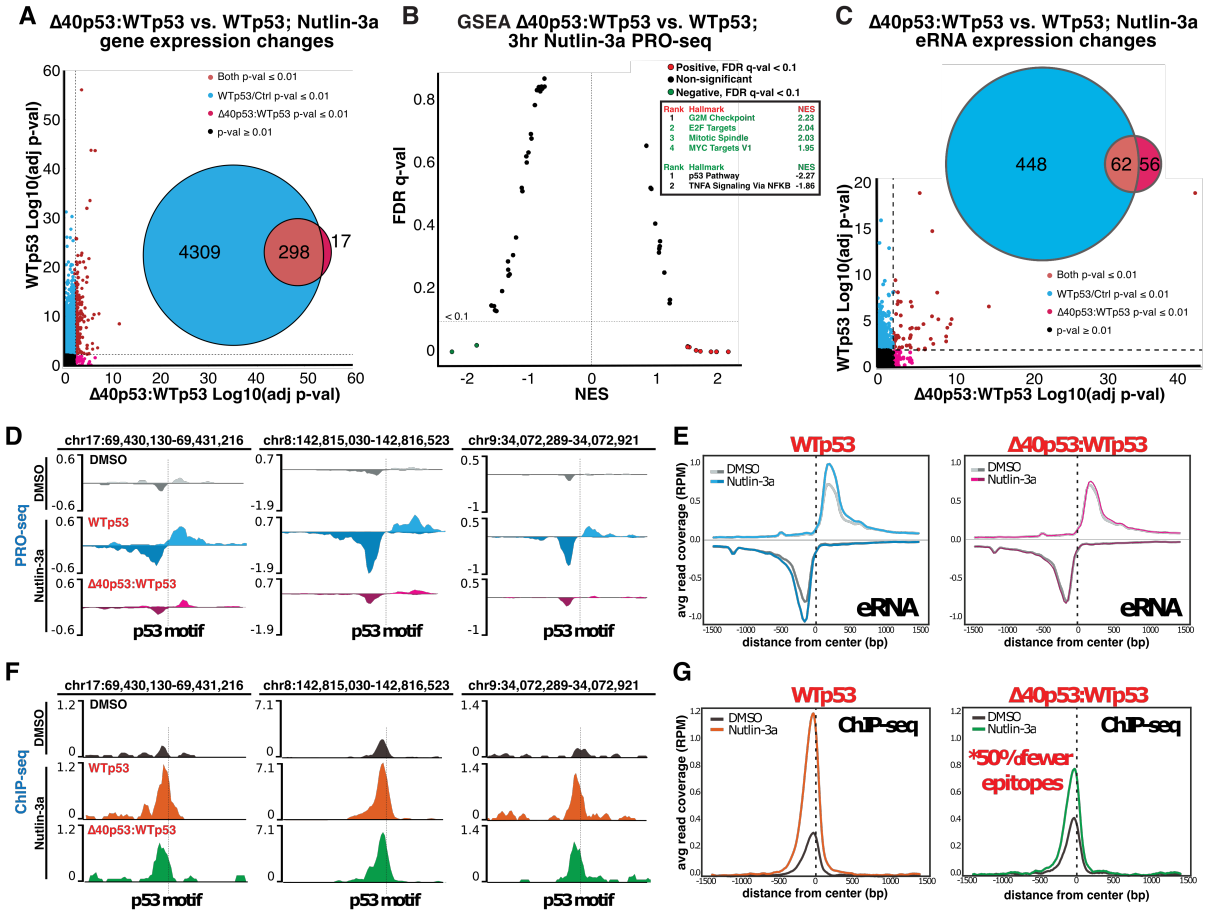


Figure 4.2: **Δ40p53 alters WTp53 function; Δ40p53:WTp53 fails to induce eRNA transcription despite similar genomic occupancy versus WTp53.** (A) Summary of PRO-seq data for WTp53 (y-axis) versus Δ40p53:WTp53 (x-axis). Dashed line represents p-value 0.01. Venn diagram shows overlap between WTp53 and Δ40p53:WTp53 cells. (B) Gene Set Enrichment Analysis (GSEA) based upon PRO-seq data (3 hr Nutlin-3a) comparing Δ40p53:WTp53 versus WTp53. Pathways with false discovery rate (q-val <0.1) are colored dots. Red represents increased in Δ40p53:WTp53 compared to WTp53 and green is decreased. X-axis is normalized enrichment score (NES). Top significant pathways are in the ranked list. (C) Summary of PRO-seq data of eRNA transcription for WTp53 (y-axis) versus Δ40p53:WTp53 (x-axis). Dashed line represents p-value 0.01. Venn diagram shows overlap between WTp53 and Δ40p53:WTp53 cells. (D) Examples of PRO-seq data, showing Nutlin-induced eRNA transcription in WTp53 cells, but not Δ40p53:WTp53. Location of p53 binding motif (p-val < 1x10-5) indicated with dashed line. (E) Metagene analysis showing average eRNA peak height, genome-wide, at p53-responsive eRNAs (p-val < 0.25) in WTp53 or Δ40p53:WTp53 cells. (F) Examples of ChIP-seq data in Nutlin-treated (6 hr) WTp53 or Δ40p53:WTp53 cells, compared with DMSO controls; aligns with PRO-seq locations from panel D. (G) Metagene analyses showing average ChIP-seq signal, genome-wide, at p53 binding sites in control versus Nutlin-treated WTp53 or Δ40p53:WTp53 cells.

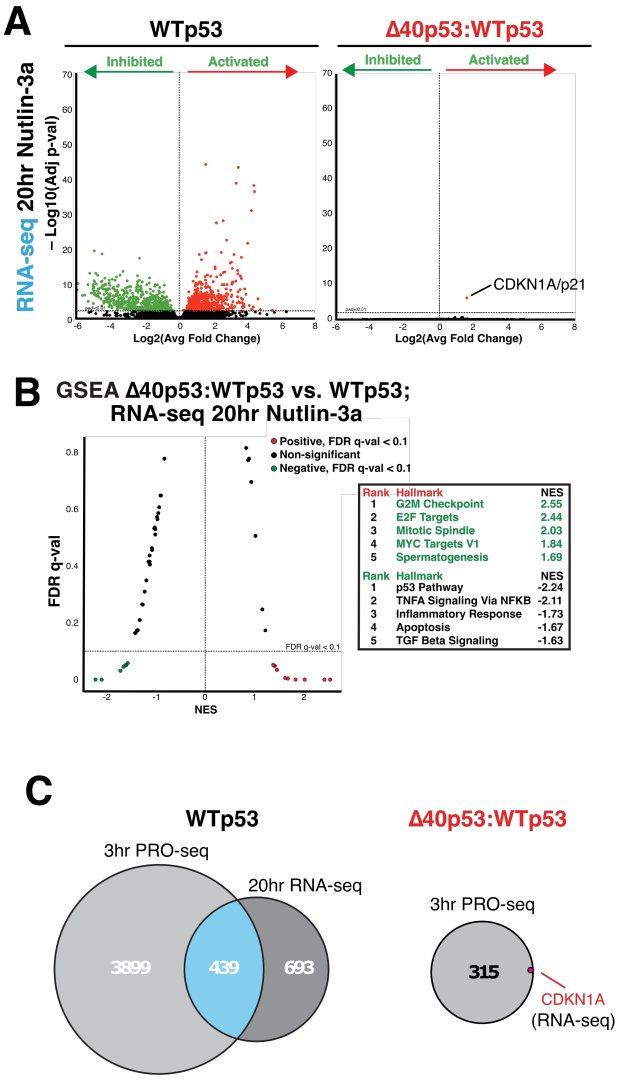
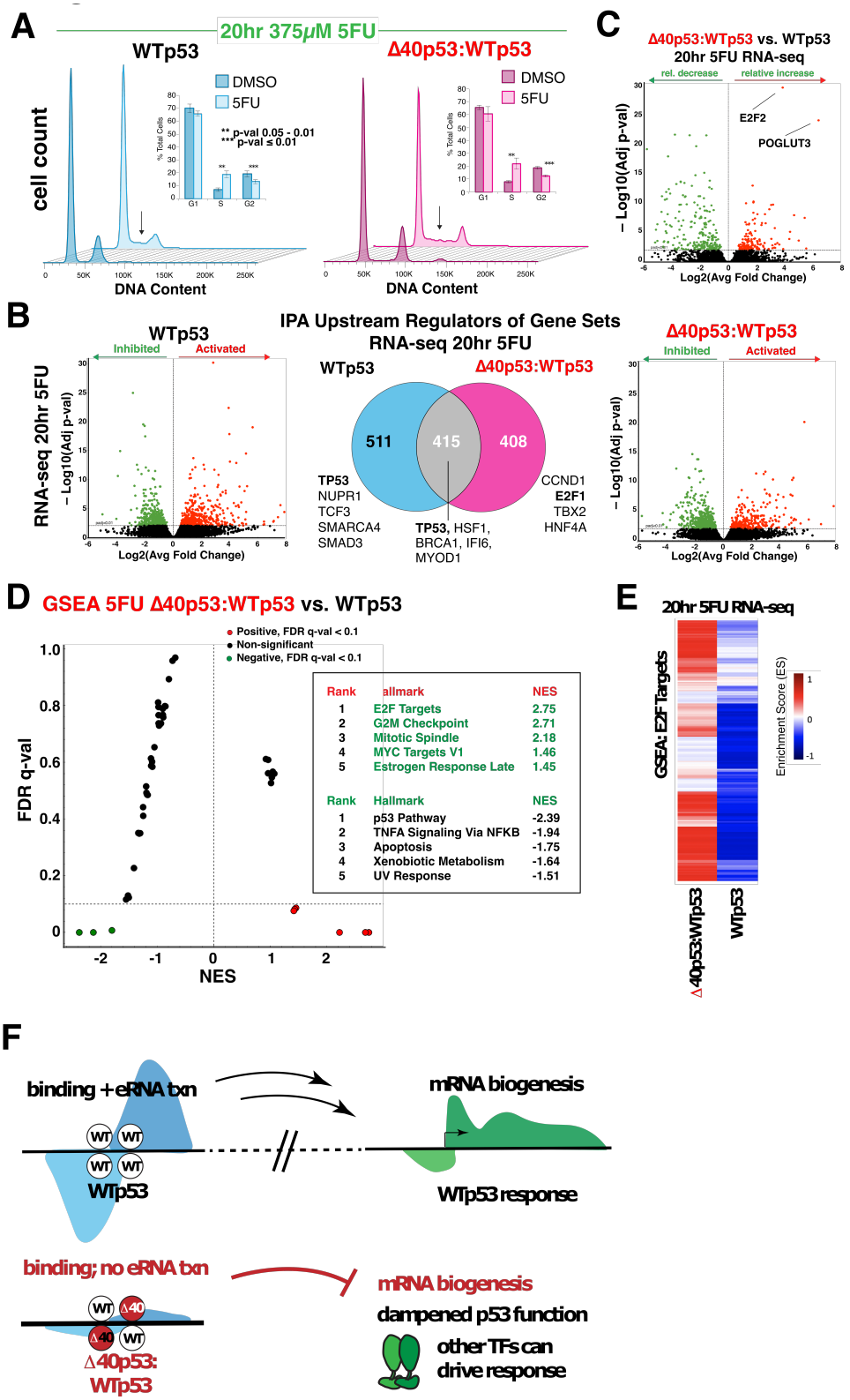


Figure 4.3: **p53 activation fails to increase mRNA levels in Δ40p53:WTp53 cells.** (A) Volcano plots showing differentially expressed mRNAs after 20 hr Nutlin treatment (versus DMSO controls) in WTp53 or Δ40p53:WTp53 cells. Green dots represent down-regulated and red dots up-regulated transcripts (p-val < 0.01). (B) Gene Set Enrichment Analysis (GSEA) based upon RNA-seq data (20hr Nutlin-3a) comparing Δ40p53:WTp53 versus WTp53. Pathways with false discovery rate q-val <0.1 are colored dots. Red represents increased in Δ40p53:WTp53 compared to WTp53 and green is decreased. X-axis is normalized enrichment score (NES). Top significant pathways are shown in ranked list. (C) Venn diagrams showing overlap among significantly induced transcripts from PRO-seq data (3hr Nutlin) and RNA-seq data (20hr Nutlin). Whereas about 10% of nascent transcripts show corresponding increases at the mRNA level in WTp53 cells, only CDKN1A/p21 shows this behavior in Δ40p53:WTp53 cells (1/316; 0.3%).



Continued on next page.

Figure 4.4: **Cellular response to 5FU similar in WTp53 versus Δ40p53:WTp53 cells, but driven by distinct TFs.** (A) Cell cycle analysis (propidium iodide); chart (inset) represents the average of 3 experiments (bars = s.e.m.). Arrows highlights increased S-phase in both WTp53 and Δ40p53:WTp53 cells after 5FU treatment. (B) Volcano plots showing differentially expressed mRNAs after 20hr 5FU treatment (versus DMSO controls) in WTp53 or Δ40p53:WTp53 cells. Green dots represent down-regulated and red dots up-regulated transcripts (p-val < 0.01). Venn diagram shows the overlap among Ingenuity Pathway Analysis (IPA) upstream regulators in 5FU-treated WTp53 versus Δ40p53:WTp53 cells. For each subset of genes (WTp53-specific, shared, or Δ40p53:WTp53-specific), the top IPA transcription regulators associated with those genes are listed (minimum Z-score cutoff 2.0). (C) Volcano plot showing relative mRNA differences in 5FU-treated Δ40p53:WTp53 versus WTp53 cells. (D) Moustache plot showing Gene Set Enrichment Analysis (GSEA) based upon RNA-seq data (20hr 5FU treatment), comparing Δ40p53:WTp53 versus WTp53. Consistent with Fig. 4.2B and 4.3B. Pathways with false discovery rate q-val < 0.1 are colored dots. Red represents increased in Δ40p53:WTp53 compared to WTp53 and green is decreased. X-axis is normalized enrichment score (NES). Top significant pathways are shown in ranked list. (E) RNA-seq data (GSEA) shows that differential E2F2 TF activity characterizes the Δ40p53:WTp53 transcriptional response to 5FU. (F) Model. WTp53 induces eRNA transcription and drives cellular stress responses, whereas Δ40p53:WTp53 tetramers enable a tunable p53 response that allows other signal specific TFs to govern transcriptional outcomes. Despite Nutlin-dependent activation of p53 target gene nascent transcription (PRO-seq, 3hr) in Δ40p53:WTp53 cells, subsequent increases in mRNA levels (RNA-seq, 20hr) were not observed; moreover, p53-dependent induction of eRNAs was absent in Δ40p53:WTp53 cells. This disconnect between nascent transcription of p53 target genes and mRNA induction implicates eRNA transcription as a regulator of mRNA biogenesis.

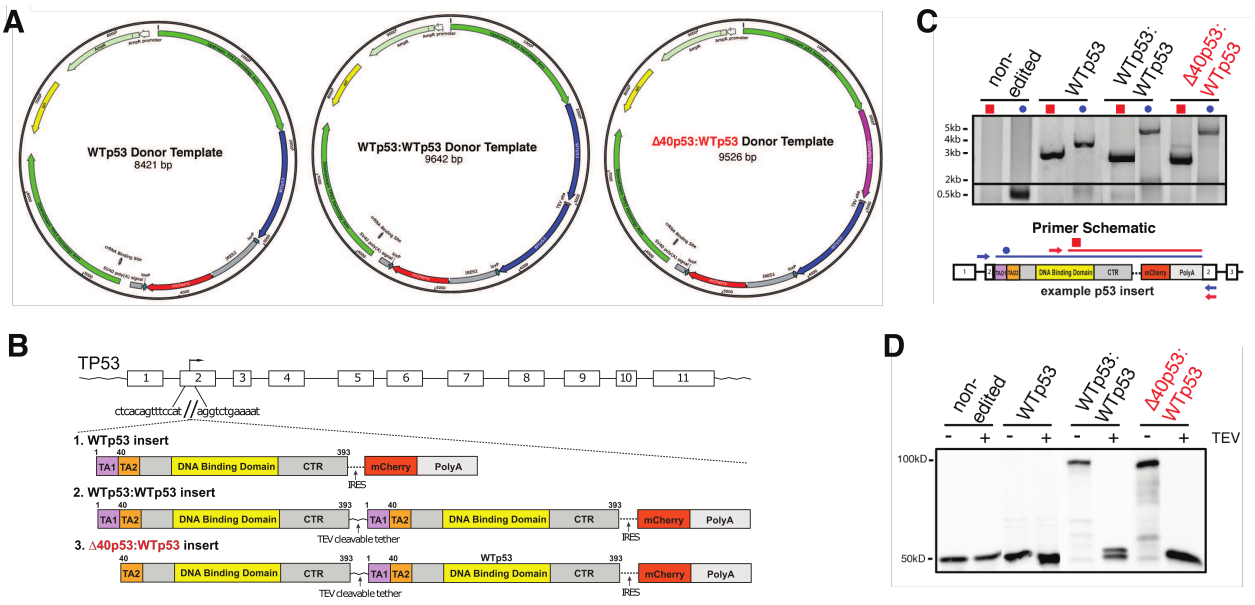
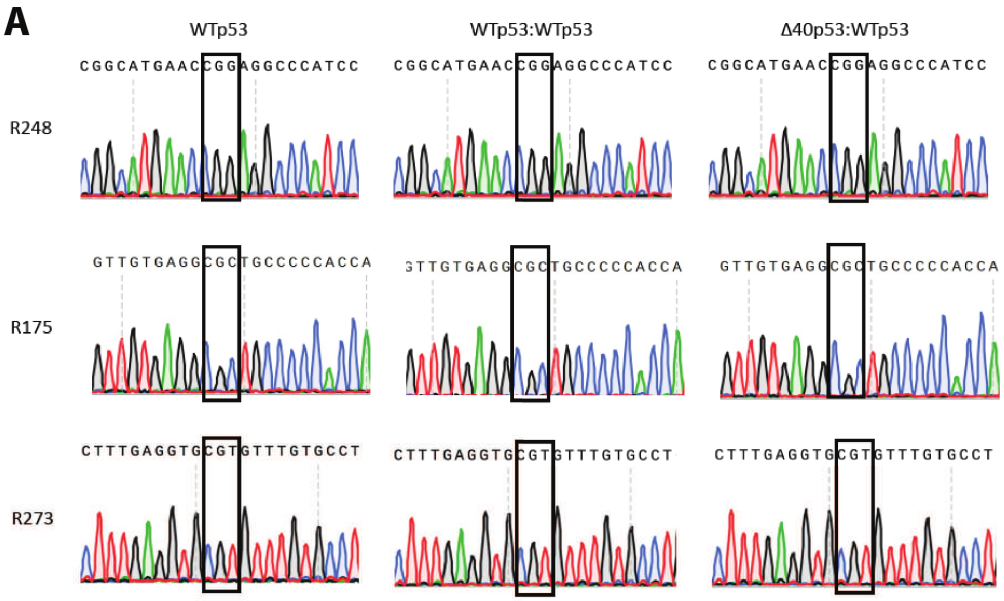


Figure 4.5: **Additional information about genome-edited cell lines.** (A) Overview of donor repair plasmids co-transfected with Cas9 RNPs. (B) Scheme for CRISPR-Cas9 insertions at the native TP53 locus: WTp53 (2820bp), WTp53:WTp53 (4041bp), and Δ40p53:WTp53 (3924bp). TA1: transactivation domain 1; TA2: transactivation domain 2; CTR: C-terminal region; IRES: internal ribosomal entry site; PolyA: SV40 polyA sequence with terminator to prevent downstream transcription. (C) PCR validation. Blue primer sets span the insertion and red primer sets have forward primer within the insertion and the reverse primer outside the insertion. Lack of 500bp band with blue primers indicated homozygous insertion. PCR experiments were completed on genomic DNA isolated from each of the indicated cell lines. Non-edited MCF10A cells were tested as an additional control. (D) Western blot validation of each genome-edited cell line, using a p53 antibody (DO-1). As designed, each tethered construct (Δ40p53:WTp53 or WTp53:WTp53) migrated at around 100kDa, indicative of dimer expression as a single transcript. The flexible tether for Δ40p53:WTp53 or WTp53:WTp53 contained a TEV cleavage site, and treatment with TEV protease (+, as shown) cleaved the dimers to p53 monomers, as expected. Note that the Δ40p53 monomer lacks the epitope detected by the p53 DO-1 antibody.



	Average Percentage of Reads in TA1 (Stderr, n=4)	Average Percentage of Reads in Central Region (Stderr, n=4)	Ratio of Reads in TA1/Central Region	Reads in TA1 compared to WTp53	Expectation
WTp53	2.48 +/- 0.30	12.29 +/- 0.20	0.20	1.00	1
WTp53:WTp53	2.84 +/- 0.16	14.79 +/- 0.20	0.19	0.95	1
Δ40p53:WTp53	1.42 +/- 0.13	12.36 +/- 0.47	0.12	0.57	0.5

Figure 4.6: **Additional validation of CRISPR-Cas9 knock-in cell lines.** (A) Sequencing results from the p53 DNA-binding domain show no mutations at common hotspots in any of the cell lines. Because p53 mutations can yield proliferative advantages in culture [171], DNA-binding domain sequencing was performed several times throughout the project to ensure no mutations occurred during the course of our experiments. (B) Internal validation of the edited p53 cell lines, using PRO-seq data. The PRO-seq data was mapped to the inserted p53 sequences at the native TP53 locus. RNA sequence corresponding to the first 39 amino acids of p53 was reduced by half in the Δ40p53:WTp53 cell line versus WTp53, as expected. This method also allowed verification of p53 copy number at 2 per cell line.

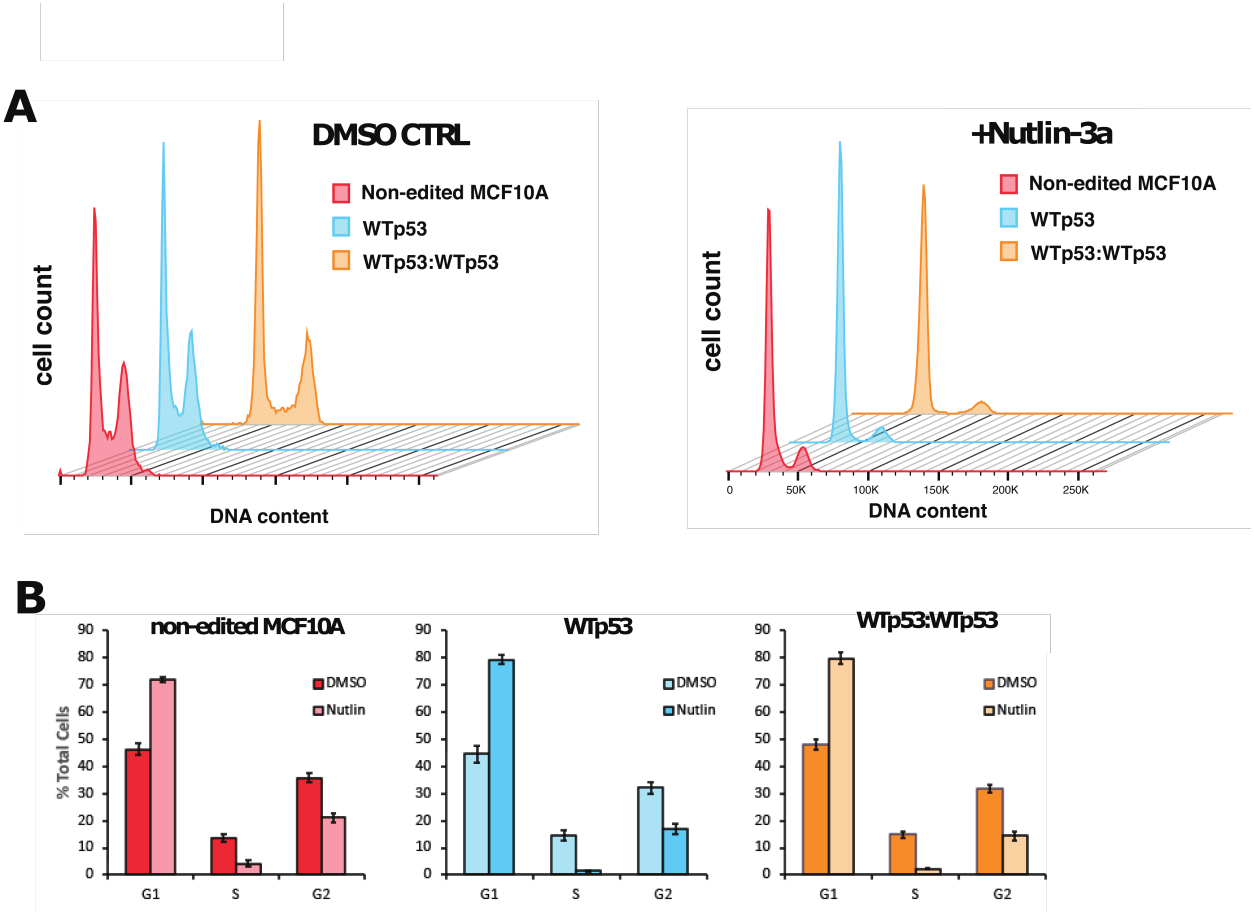


Figure 4.7: **Endogenous MCF10A cells phenotypically match CRISPR-Cas9 edited WTp53 and WTp53:WTp53 cell lines.** Unedited "off-the-shelf" MCF10A cells, edited WTp53, and WTp53:WTp53 cells each display the same cell cycle phenotypes. (A) Cell cycle analysis (propidium iodide) and bar plots (B) represent the average of 6 biological replicates (bars = standard error of mean). Note that MCF10A cells endogenously express WTp53.

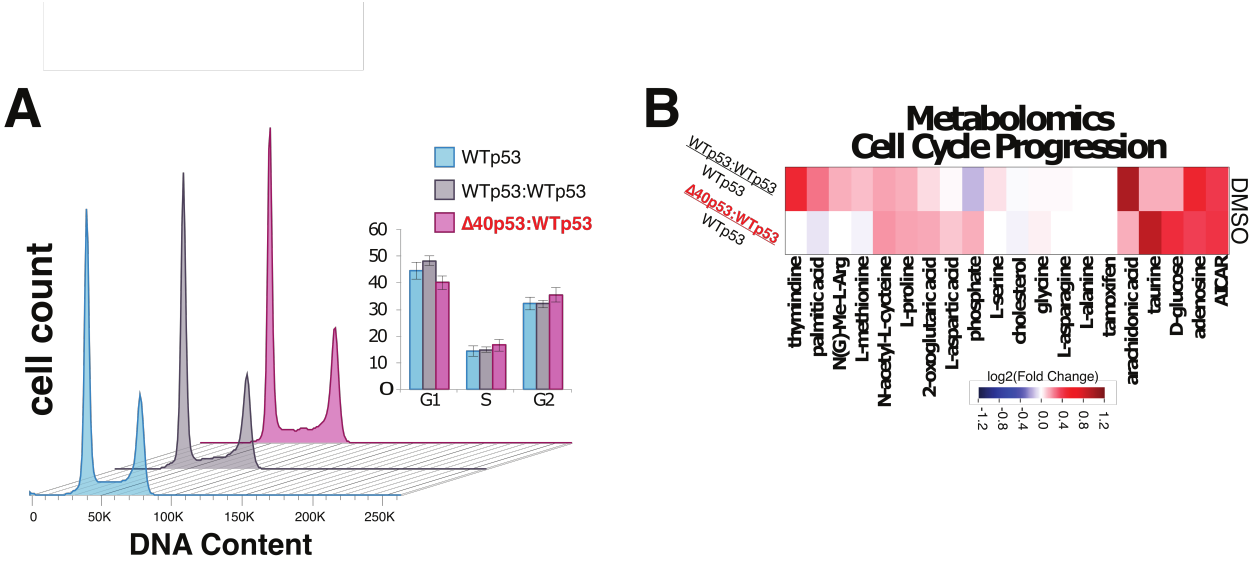


Figure 4.8: **WTp53 versus Δ40p53:WTp53 phenotypic and metabolic similarity under normal growth conditions but differences upon p53 activation with Nutlin-3a.** (A) Cell cycle analysis (propidium iodide) of each genome-edited cell line. Chart (inset) represents the average of 6 experiments (bars = standard error of mean). (B) Sample of metabolomics data comparing IPA identified cell cycle metabolites under normal growth conditions in each cell line (6 biological replicates each). Log2(Fold Change) was normalized to WTp53.

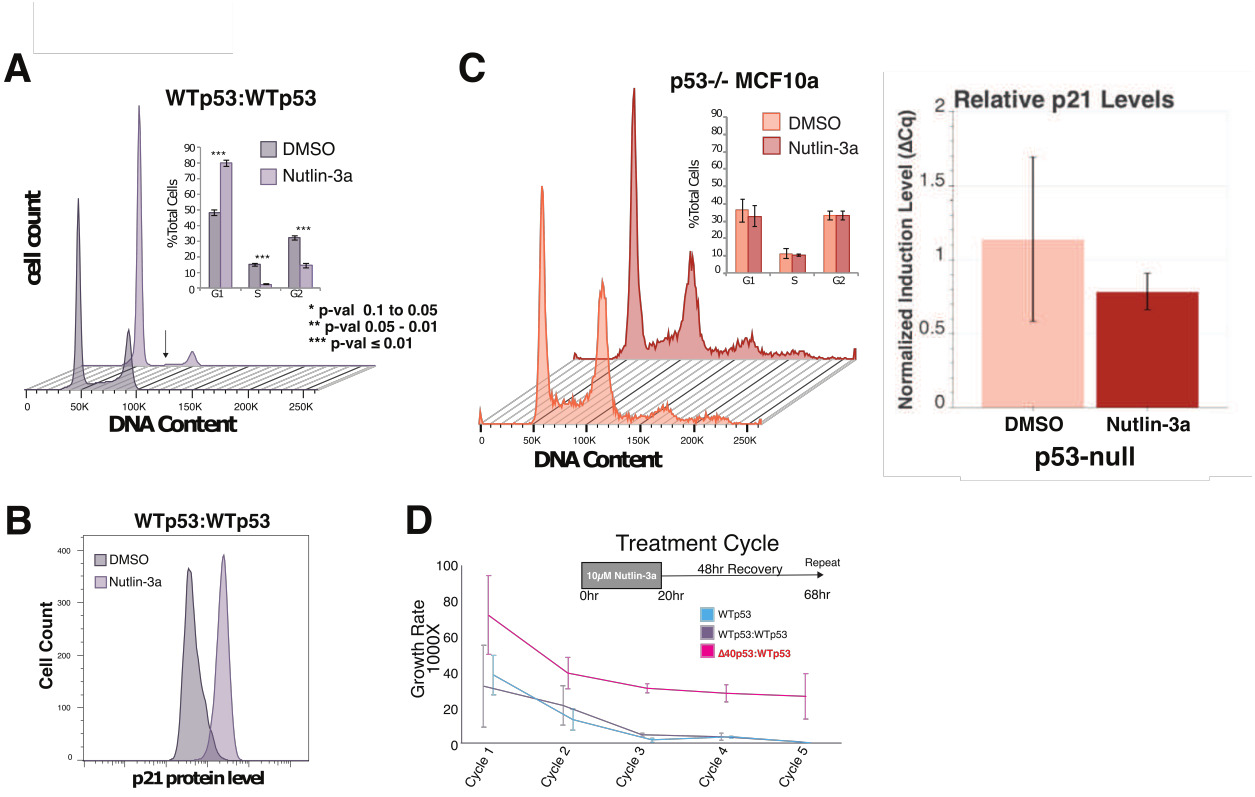


Figure 4.9: **WTp53:WTp53 cells are metabolically and phenotypically similar to WTp53 cells.** (A) Cell cycle analysis (propidium iodide)in WTp53:WTp53 cells; chart (inset) represents the average of 6 biological replicate experiments (bars = standard error of mean). Arrow highlights the loss of S-phase cells, similar to WTp53 cells (and in contrast with Δ40p53:WTp53 cells). (B) Measurement of p21 protein levels by FACS; similar to WTp53, p21 protein levels increase in WTp53:WTp53 cells after Nutlin-3a treatment. (C) Cell cycle analysis (propidium iodide) in p53-null MCF10a cells; chart (inset) represents the average of 3 biological replicate experiments (bars = standard error of mean). Bar plot shows time correlated p21 RT-qPCR (not statistically significant pval > 0.05). (D) Growth rate measured over 5 treatment cycles; each cycle encompassed Nutlin treatment for 20 hours, followed by splitting cells 1:10, then growth under normal conditions for 48 hours (3 biological replicates; bars = s.e.m.). Note the similar growth characteristics of WTp53 versus WTp53:WTp53 cells.

Sphingolipid Metabolites

		Δ40p53:WTp53 / WTp53:WTp53	Δ40p53:WTp53 / WTp53
Sphingolipid Synthesis	sphinganine	0.97	0.71
	sphingadienine	0.65	0.34
	phytosphingosine	0.97	0.51
Dihydroceramides	N-palmitoyl-sphinganine (d18:0/16:0)	1.57	1.52
Ceramides	N-palmitoyl-sphingosine (d18:1/16:0)	0.94	0.73
	N-stearoyl-sphingosine (d18:1/18:0)*	1.17	1.21
	N-palmitoyl-sphingadienine (d18:2/16:0)*	0.45	0.40
	N-palmitoyl-heptadecasphingosine (d17:1/16:0)*	0.82	0.67
	ceramide (d18:1/14:0, d16:1/16:0)*	0.86	0.68
	ceramide (d18:1/17:0, d17:1/18:0)*	0.84	1.01
Hexosylceramides (HCER)	glycosyl-N-palmitoyl-sphingosine (d18:1/16:0)	0.55	0.54
	glycosyl-N-stearoyl-sphingosine (d18:1/18:0)	0.75	0.66
	glycosyl ceramide (d18:1/20:0, d16:1/22:0)*	0.96	1.00
	glycosyl ceramide (d18:2/24:1, d18:1/24:2)*	0.36	0.38
Lactosylceramides (LCER)	lactosyl-N-palmitoyl-sphingosine (d18:1/16:0)	0.72	0.63
	lactosyl-N-nervonoyl-sphingosine (d18:1/24:1)*	0.96	0.86
Dihydrosphingomyelins	myristoyl dihydrosphingomyelin (d18:0/14:0)*	1.77	1.67
	palmitoyl dihydrosphingomyelin (d18:0/16:0)*	1.58	1.66
	behenoyl dihydrosphingomyelin (d18:0/22:0)*	2.07	2.49
	sphingomyelin (d18:0/18:0, d19:0/17:0)*	3.25	3.52
	sphingomyelin (d18:0/20:0, d16:0/22:0)*	3.05	3.66
Sphingosines	sphingosine	0.89	0.45
	hexadecasphingosine (d16:1)*	0.69	0.40
	heptadecasphingosine (d17:1)	0.78	0.50
	eicosanoylsphingosine (d20:1)*	0.83	0.63

down — p-val ≤ 0.05, ratio < 1.00

up — p-val ≤ 0.05, ratio ≥ 1.00

1.21 — non-significant

0.84 — 0.05 < p < 0.10, ratio < 1.00

Figure 4.10: **Metabolomics show increased sphingolipid metabolites in Δ40p53:WTp53 cells.** Metabolic changes in Δ40p53:WTp53 cells are similar versus WTp53:WTp53 cells and versus WTp53 cells after Nutlin-3a treatment for 20hr.

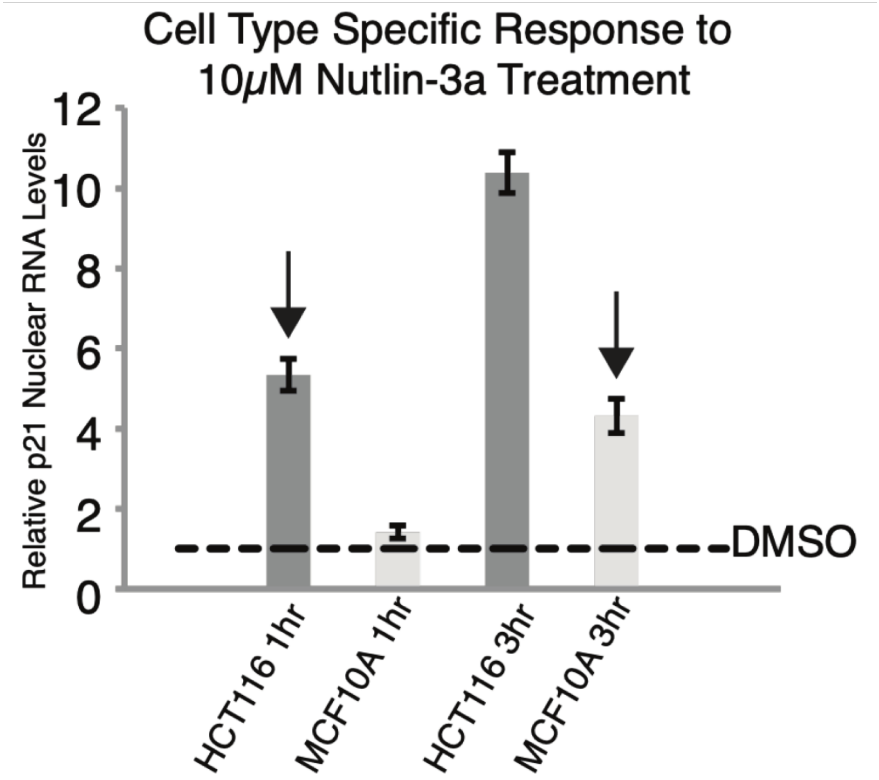


Figure 4.11: **Justification for PRO-seq analysis after 3-hour Nutlin-3a in MCF10A cells.** RT-qPCR of nuclear RNA levels for CDKN1A/p21 in either HCT116 or MCF10A cells at 1hr and 3hr. These data show a delayed p53 response in MCF10A cells versus HCT116, with the 3hr induction roughly matching the 1hr p21 nuclear RNA levels from HCT116 cells (arrows). Past GRO-seq results in 1hr Nutlin-treated HCT116 cells [6] were used to guide the 3hr time point chosen for MCF10A.

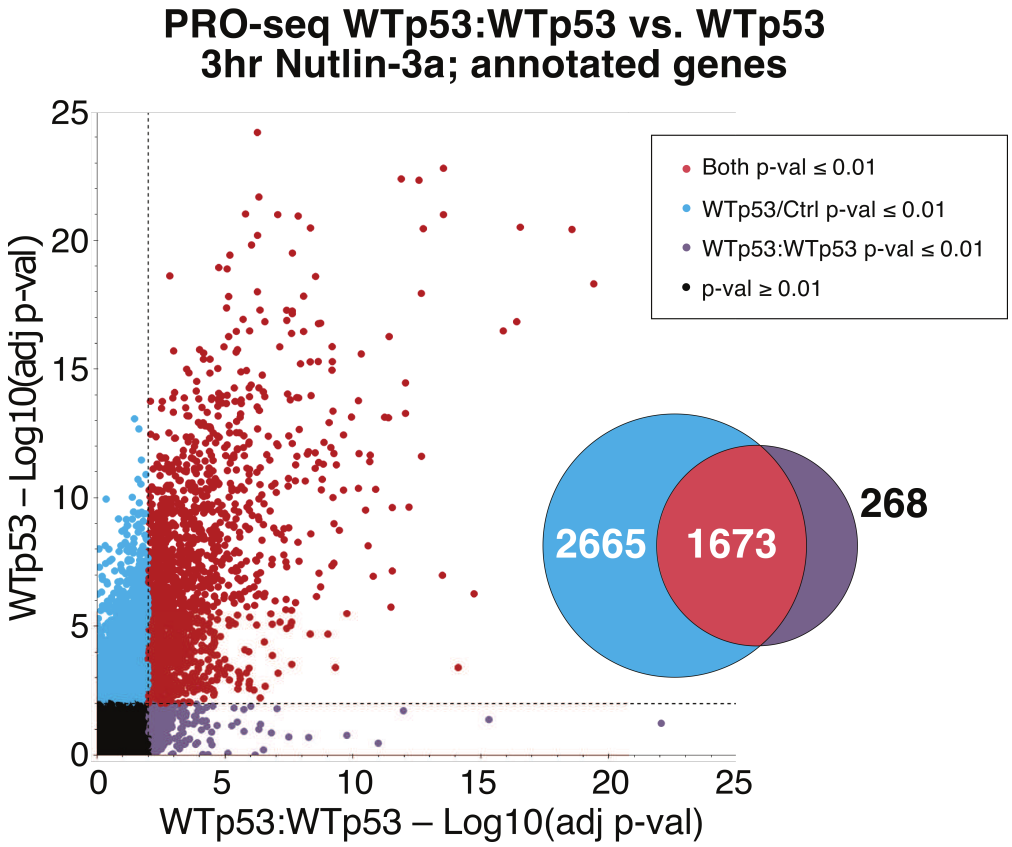


Figure 4.12: **Nutlin induces similar transcriptional changes in WTp53:WTp53 and WTp53 cells.** Summary of PRO-seq data for annotated genes differentially transcribed in WTp53 (y-axis) versus WTp53:WTp53 cells (x-axis). Dashed line represents p-value 0.01. Venn diagram shows overlap between WTp53 and WTp53:WTp53 cells.

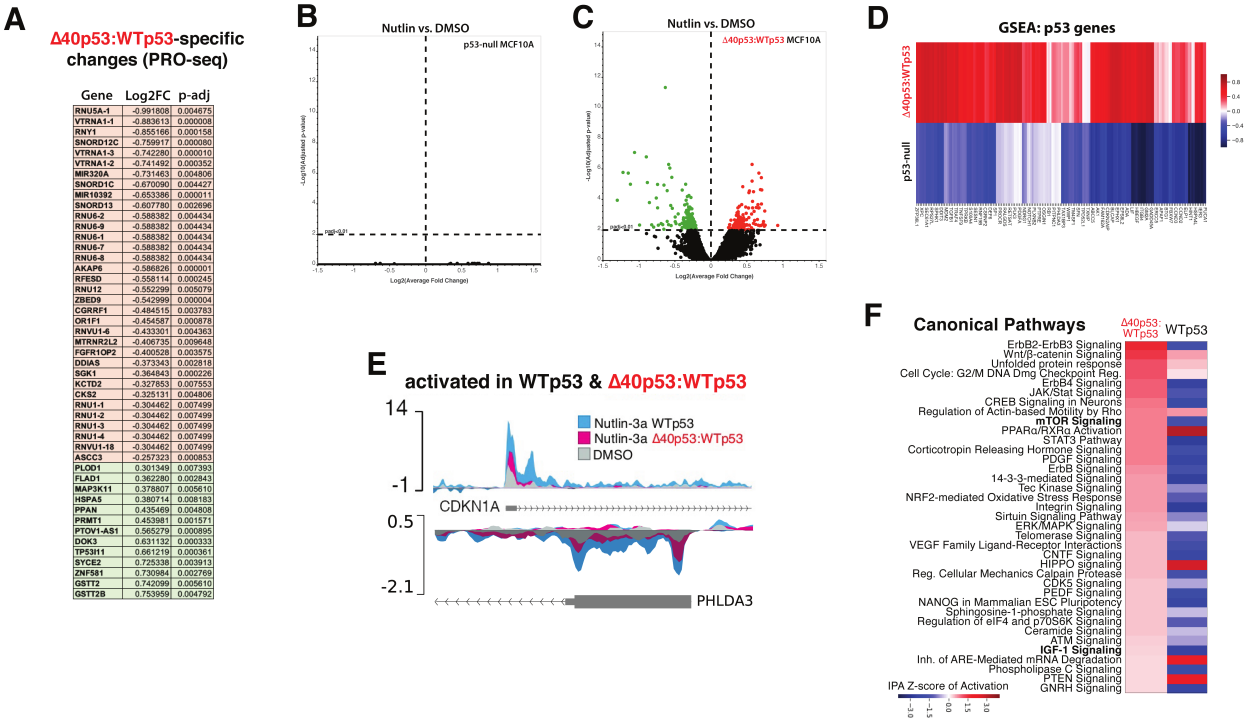


Figure 4.13: **PRO-seq data in p53-null MCF10A cells confirms robust p53 response in Δ40p53:WTp53 cells.** (A) List of the 17 genes that are selectively transcribed in Δ40p53:WTp53 cells (PRO-seq data, 3 hr Nutlin). (B-C) Volcano plots that show differentially expressed transcripts (PRO-seq) at gene bodies after 3hr Nutlin-3a treatment (versus DMSO controls) in (B) p53-null or (C) Δ40p53:WTp53 MCF10A cells. Green dots represent down-regulated and red dots up-regulated transcripts (p-val < 0.01). (D) Ward cluster maps of enrichment of canonical p53 target genes in either Δ40p53:WTp53 (left) or p53-null MCF10A cells (right) after 3hr Nutlin-3a treatment (PRO-seq). (E) Representative PRO-seq data from Nutlin-treated WTp53 or Δ40p53:WTp53 cells, and DMSO controls. Although Nutlin induction is observed in each cell line, activation is reduced in Δ40p53:WTp53 cells. (F) Ingenuity Pathway Analysis (IPA) based upon PRO-seq data (3 hr Nutlin treatment). Pathways highlighted in bold font have been linked to progeroid phenotypes in mice expressing Δ40p53 + WTp53[161, 226].

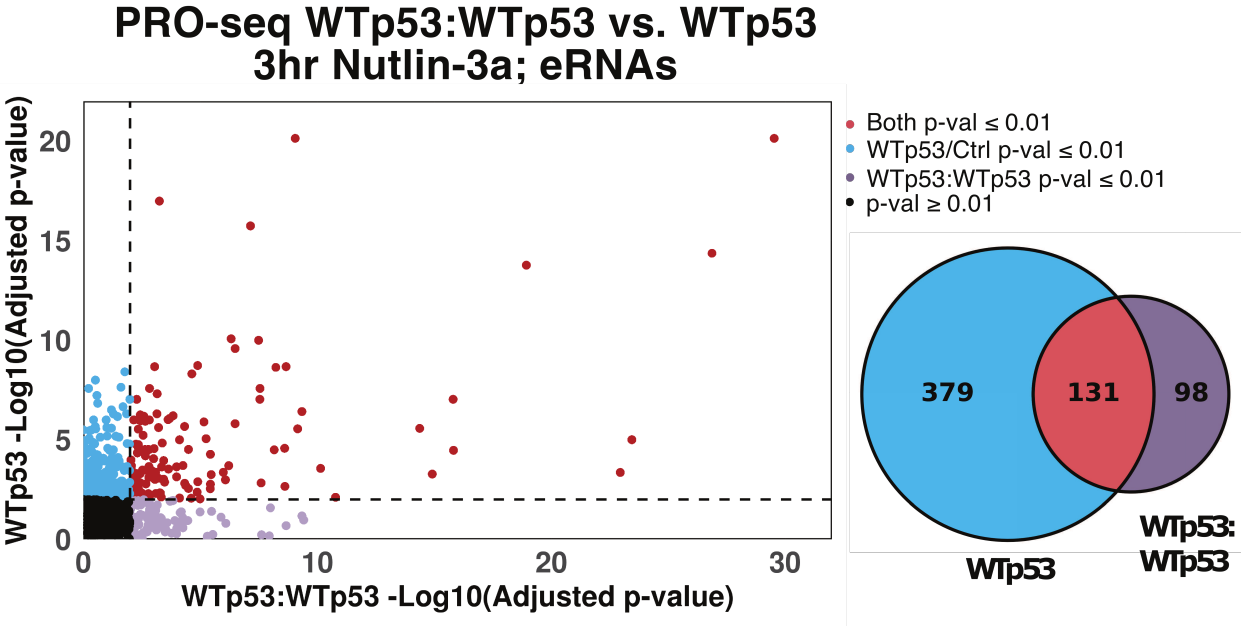


Figure 4.14: **Nutlin induces similar transcriptional changes (eRNA) in WTp53:WTp53 and WTp53 cells.** Summary of PRO-seq data of eRNA transcription for WTp53 (y-axis) versus WTp53:WTp53 (x-axis). Dashed line represents p-value 0.01. Venn diagram shows overlap between WTp53 and WTp53:WTp53 cells.

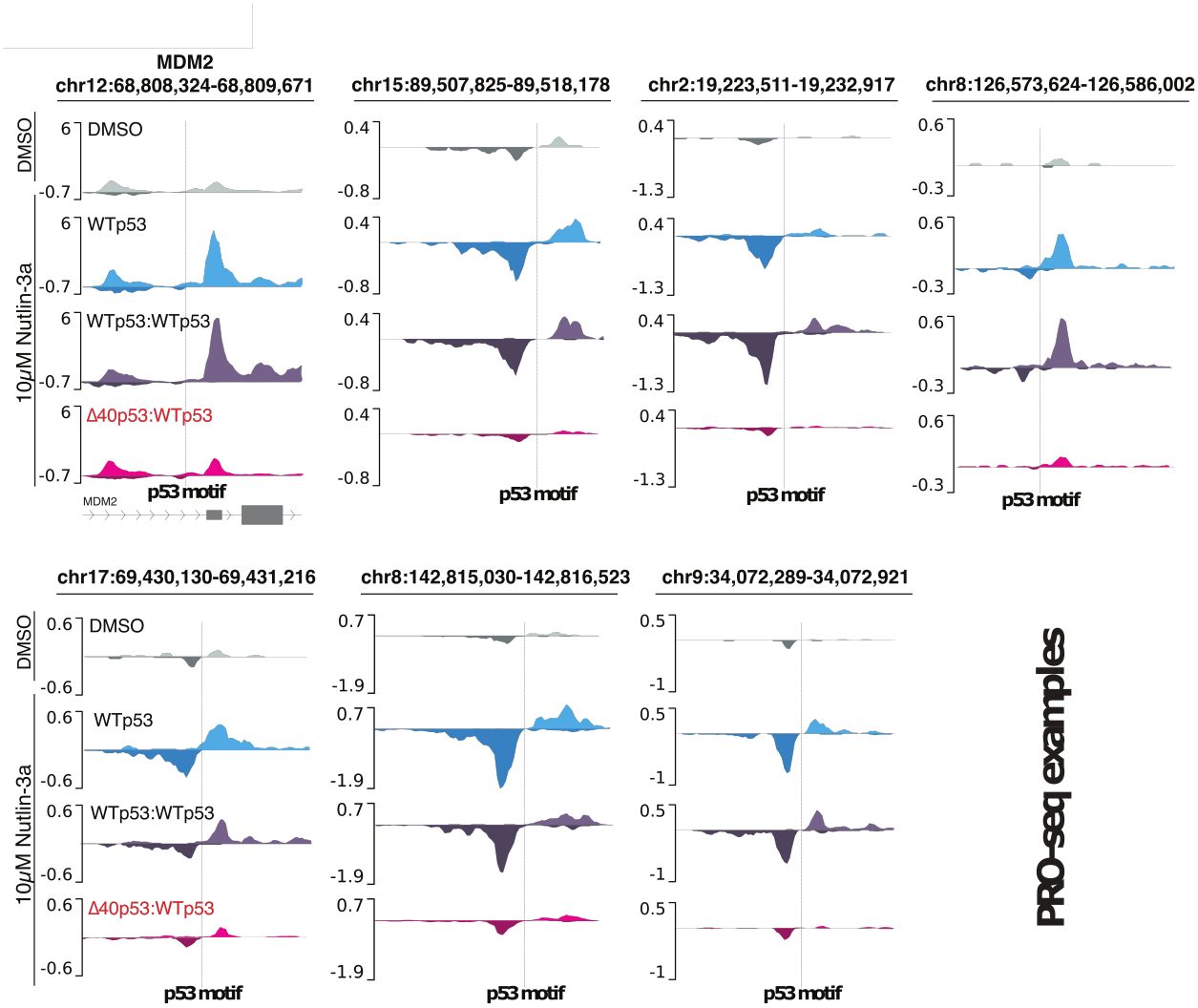


Figure 4.15: **PRO-seq data from all three cell lines (3 hr Nutlin-treated) and DMSO control.** Note the similar responses in WTp53 and WTp53:WTp53 cells, whereas Δ40p53:WTp53 cells lack the ability to induce bidirectional eRNA transcription. The p53 binding motif (p-value < 1x10-5) is shown with a dashed line, and the peaks correspond directly with ChIP-seq peaks shown in Fig. 4.17.

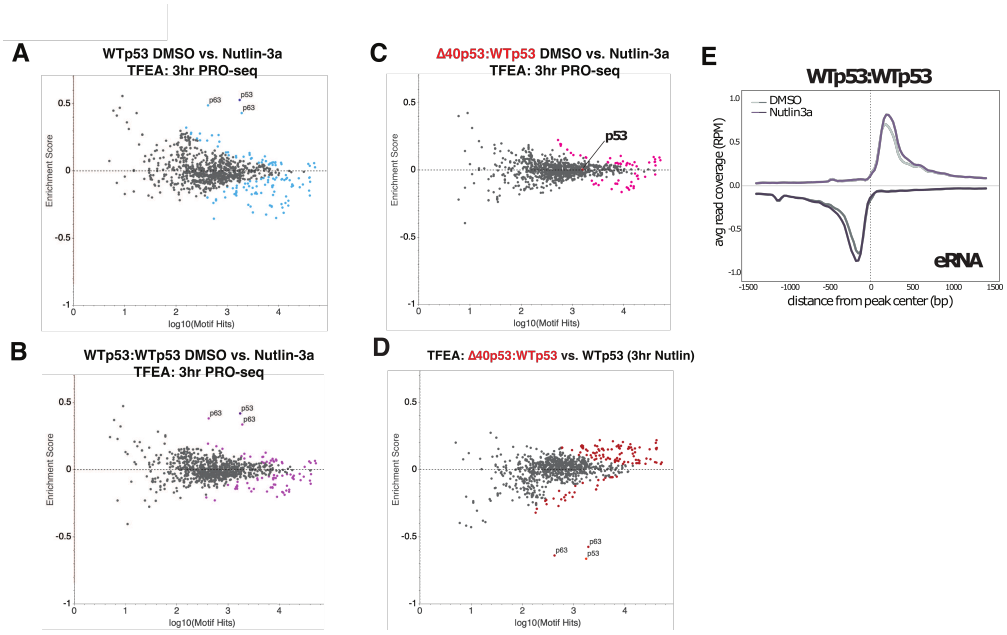


Figure 4.16: **TFEA reveals differential TF activation in Nutlin-treated cells.** (A-C) Transcription Factor Enrichment Analysis (TFEA) [207] from PRO-seq data (3 hr Nutlin-3a) in WTp53, WTp53:WTp53, and Δ40p53:WTp53 cells. Whereas p53 induction is robust in WTp53 and WTp53:WTp53 cells, this is not observed in Nutlin-treated Δ40p53:WTp53 cells (p-val < 1x10-6). This reflects a lack of eRNA transcription at p53 binding sites in Δ40p53:WTp53 cells. Note that TP53 and TP63 have almost identical binding motifs. (D) Transcription Factor Enrichment Analysis (TFEA) [207] from PRO-seq data (3 hr Nutlin-3a) in WTp53 versus Δ40p53:WTp53 cells, indicating reduced p53 (and p63) activity in Δ40p53:WTp53 cells (p-val < 1x10-6). This reflects a lack of eRNA transcription at p53 binding sites in Δ40p53:WTp53 cells. Note that TP53 and TP63 have almost identical binding motifs. (E) Metagene analysis showing average eRNA peak height, genome-wide, at p53 responsive eRNAs (p-val < 0.25) in WTp53:WTp53 cells.

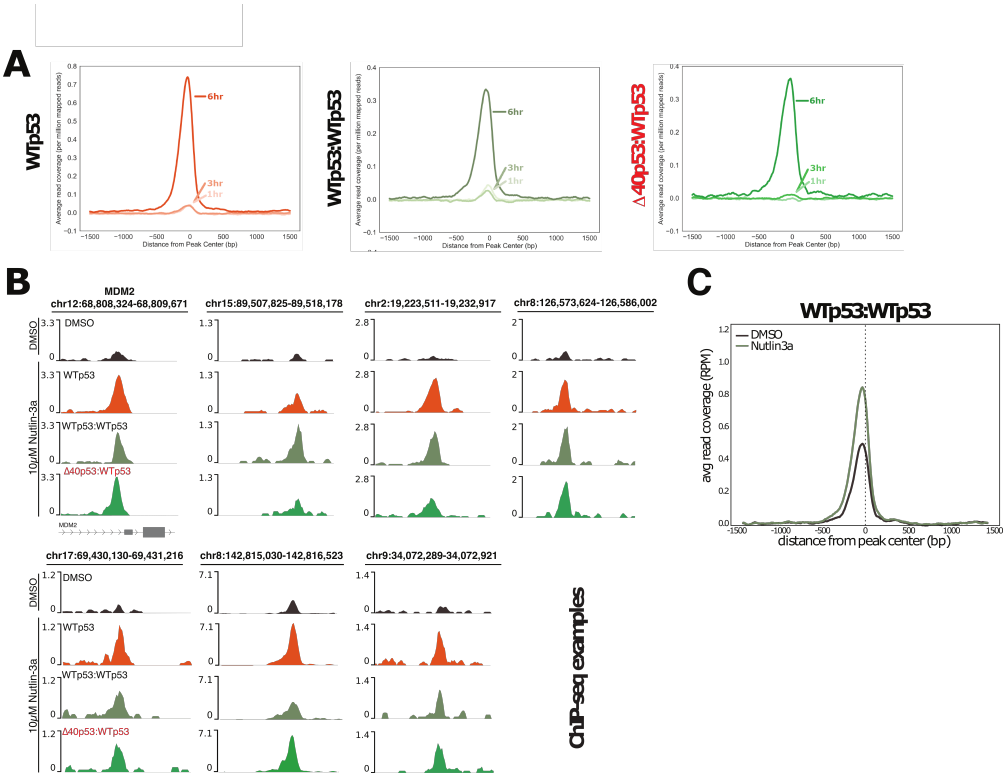


Figure 4.17: **Additional ChIP-seq data across all 3 cell lines.** (A) A series of ChIP-seq experiments were completed 1hr, 3hr, or 6hr after Nutlin-3a treatment, to assess the time-dependence of p53 occupancy changes. Metagene analyses are shown (average ChIP-seq signal) at p53 binding sites in Nutlin-3a-treated WTp53, WTp53:WTp53 or Δ40p53:WTp53 cells at 1hr, 3hr and 6hr. To our knowledge, following Nutlin treatment, all published p53 ChIP-seq data sets used 6hr or longer time points. (B) Examples of ChIP-seq data in Nutlin-treated (6 hr) cells, compared with DMSO controls. ChIP-seq peaks correspond directly with PRO-seq eRNA peaks shown in Fig. 4.16 (C) Metagene analyses (filtered for peaks containing p53 motif) showing average ChIP-seq signal, genome-wide, at p53 binding sites in DMSO control versus Nutlin-treated WTp53:WTp53 cells.

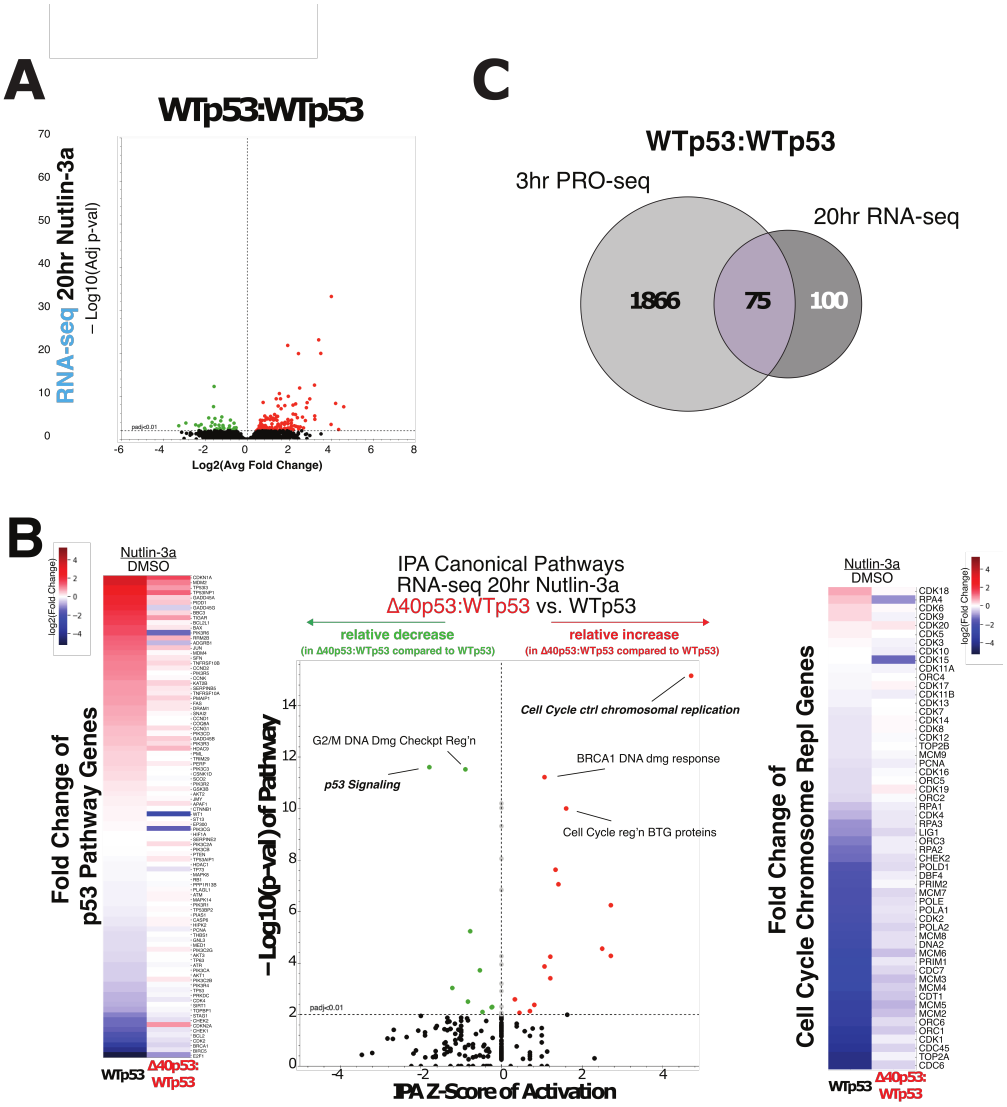


Figure 4.18: **Transcriptional changes in WTp53:WTp53 cells; IPA comparisons for Δ40p53:WTp53 and WTp53.** (A) Volcano plot showing differentially expressed mRNAs after 20 hr Nutlin treatment (versus DMSO controls) in WTp53:WTp53 cells. Green dots represent down-regulated and red dots up-regulated transcripts (p-val < 0.01). (B) Venn diagrams showing overlap among significantly induced transcripts from PRO-seq data (3 hr Nutlin) and RNA-seq data (20 hr Nutlin) in WTp53:WTp53 cells. (C) Ingenuity Pathway Analysis (IPA) of canonical pathways from RNA-seq data after 20 hr Nutlin-3a treatment in Δ40p53:WTp53 versus WTp53 cells flanked by heat maps of log2(fold change) of genes from p53 on the left, and cell cycle chromosome replication pathway on the right. P53 and cell cycle chromosome replication represent the most significantly down-regulated or up-regulated pathways, respectively, in Δ40p53:WTp53 cells (versus WTp53).

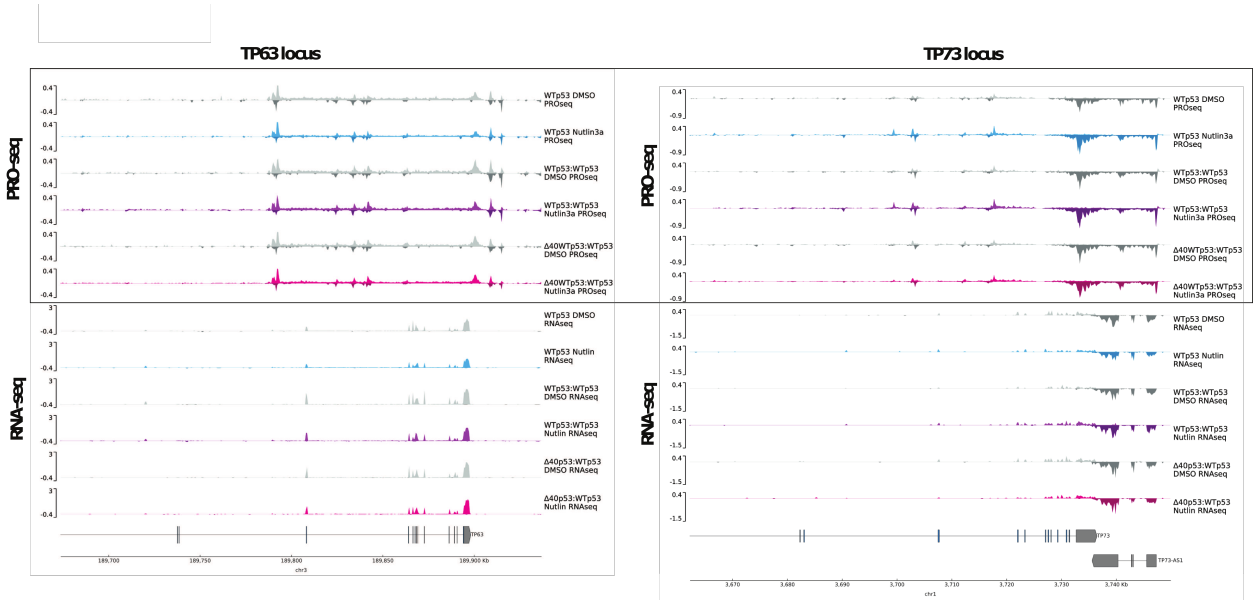


Figure 4.19: **PRO-seq and RNA-seq data at TP63 and TP73 loci in WTp53, WTp53:WTp53, and Δ40p53:WTp53 cells.** Genome browser views of p63 and p73, showing that p63 is transcribed (PRO-seq) and expressed (RNA-seq) in the MCF10A cell lines listed at right. The p73 locus, by contrast, shows transcription of p73-AS1 (p73 antisense 1), a mature antisense transcript. Neither p63 or p73 was induced by Nutlin-3a treatment. Because p73 was not transcribed in MCF10A cells, we focused on p63, but we note some transcription is observed from the adjacent TP73-AS1 locus. Top: DMSO and Nutlin-3a tracks for PRO-seq in WTp53, WTp53:WTp53, and Δ40p53:WTp53 cells; bottom: DMSO and Nutlin-3a tracks for RNA-seq in WTp53, WTp53:WTp53, and Δ40p53:WTp53 cells.

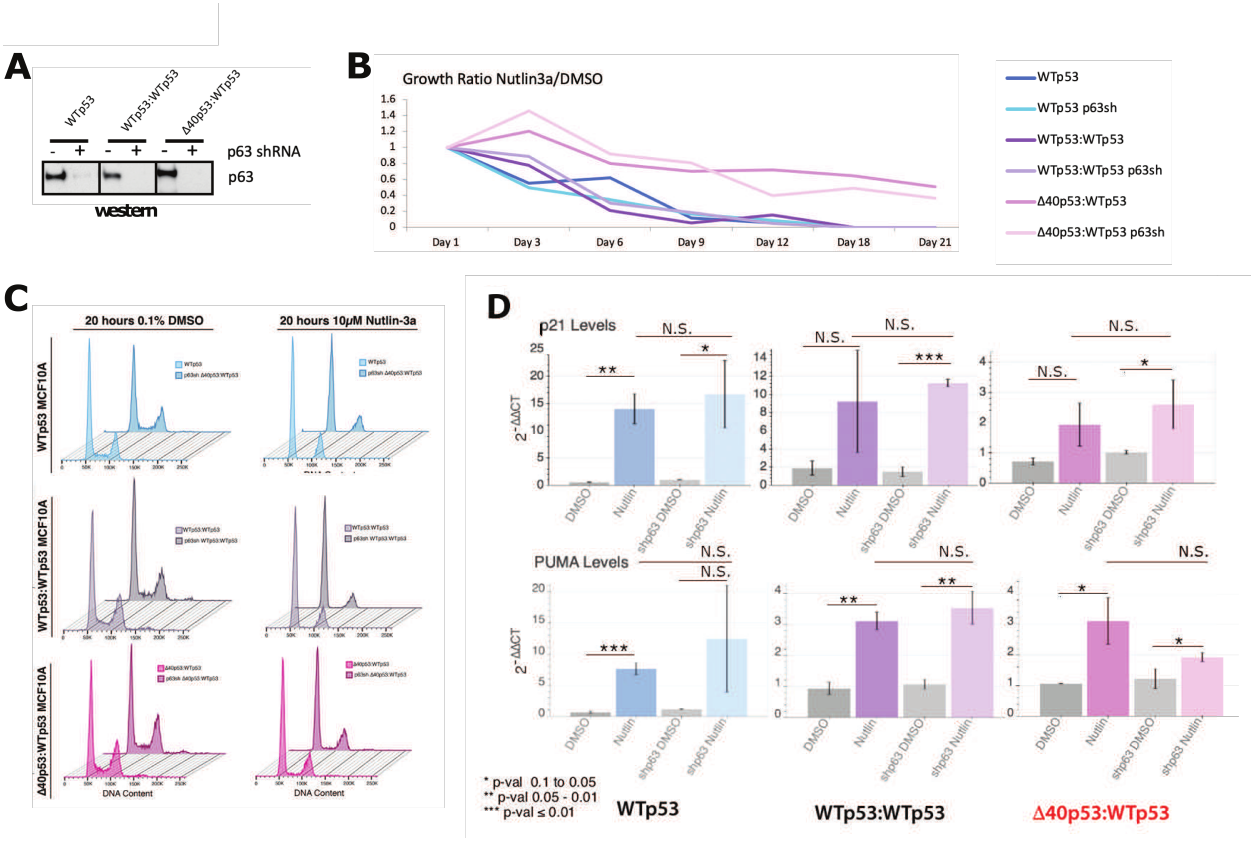


Figure 4.20: **The p53 paralog p63 does not impact Δ40p53:WTp53 phenotype or function.** (A) Western blot to probe p63 levels before or after lentiviral p63 knock down. (B) Growth rate measured over 5 Nutlin treatment cycles; each cycle encompassed 20 hours under basal (0.1% DMSO) versus Nutlin-treated conditions, splitting cells 1:10, then growth for another 48 hours. (C) Cell cycle analysis (propidium iodide) of each genome-edited cell line in either control or p63 knockdown. Left shows the cell cycle after 20hr of 0.1% DMSO treatment (control), right shows the cell cycle after 20hr of Nutlin3a treatment. (D) Knockdown of p63 shows no impact on p53 target gene induction in WTp53, WTp53:WTp53, or Δ40p53:WTp53 cell lines. RT-qPCR data are shown for p21 and PUMA (2 biological replicates; bars = s.e.m.). N.S. designates non-significant comparisons.

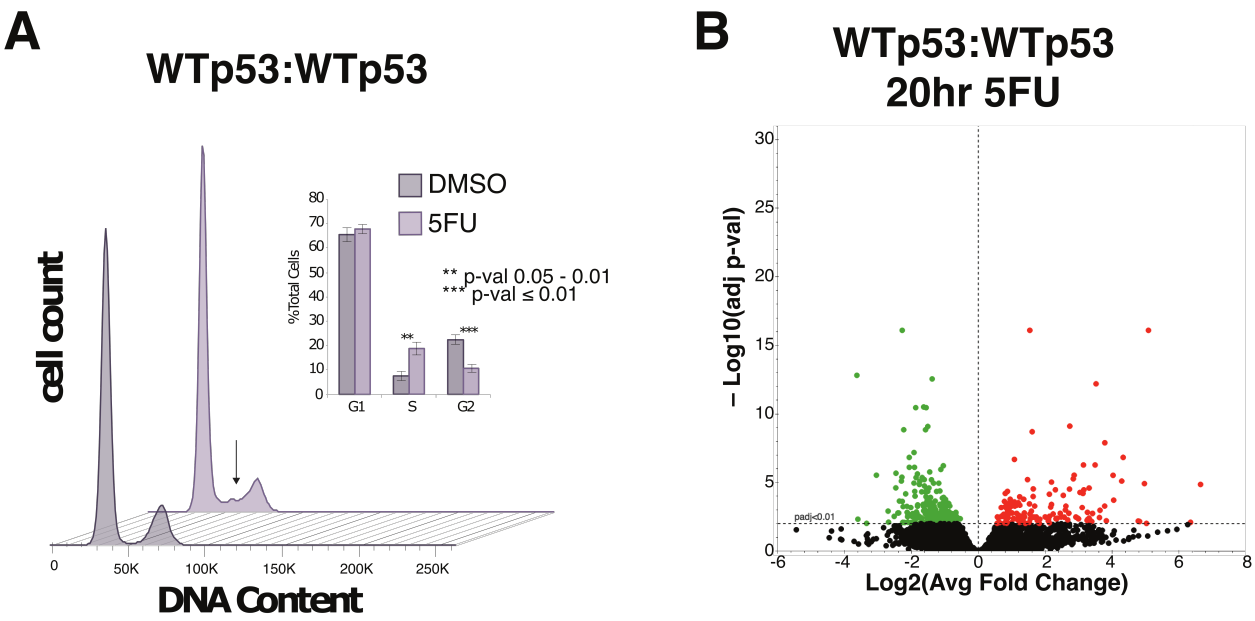


Figure 4.21: **Additional data on cellular responses to 5FU.** (A) Cell cycle analysis (propidium iodide); chart (inset) represents the average of 3 experiments (bars = standard error of mean). Arrow highlights increased S-phase in 5FU-treated WTp53:WTp53 cells, similar to both WTp53 and Δ40p53:WTp53 cells (Fig. 4.4A). (B) Volcano plot showing differentially expressed mRNAs after 20 hr 5FU treatment (versus DMSO controls) in WTp53:WTp53 cells. Green dots represent down-regulated and red dots up-regulated transcripts (p-val < 0.01).

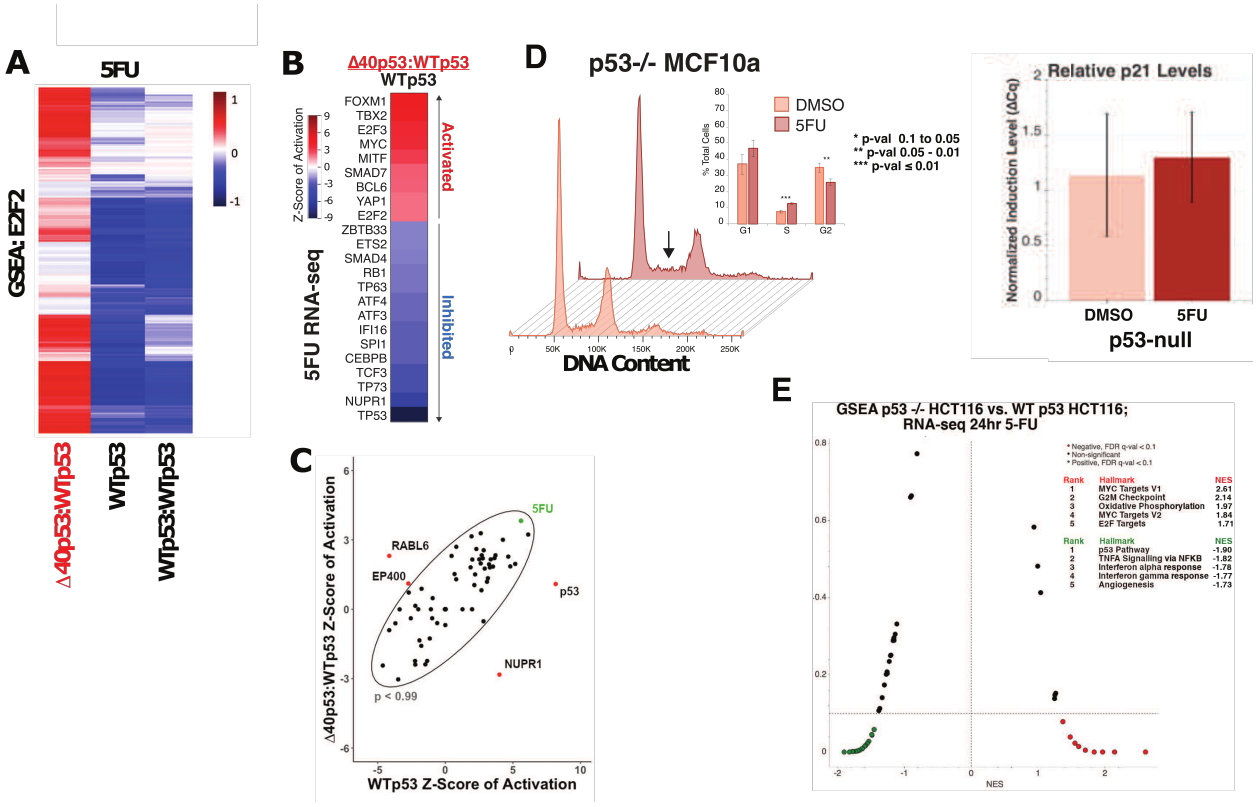


Figure 4.22: **The E2F pathway is activated selectively in 5FU-treated Δ40p53:WTp53 cells.** (A) RNA-seq data (GSEA) shows that differential E2F2 TF activity characterizes the Δ40p53:WTp53 transcriptional response to 5FU in comparison to WTp53 and WTp53:WTp53. (B) Ingenuity Pathway Analysis (IPA) showing upstream regulatory transcription factors, inferred from RNA-seq data in 5FU-treated Δ40p53:WTp53 versus WTp53 cells. (C) The general upstream regulators (IPA) controlling the cellular response to 5FU are similar (shared inside oval) between Δ40p53:WTp53 and WTp53 cells, with a few exceptions ($p < 0.01$). (D) Cell cycle analysis (propidium iodide); chart (inset) represents the average of 3 experiments (bars = standard error of mean). Arrow highlights increased S-phase in 5FU-treated p53-null MCF10a cells (E) Gene Set Enrichment Analysis (GSEA) based upon RNA-seq data from Yang et. al. comparing p53 -/- HCT116 cells versus WT HCT116 cells after a 24hr 5FU treatment. Pathways with false discovery rate q-val <0.1 are colored dots. Red represents increased in p53 -/- cells compared to WT p53 cells and green is decreased. X-axis is normalized enrichment score. Top significant pathways are in the ranked list.

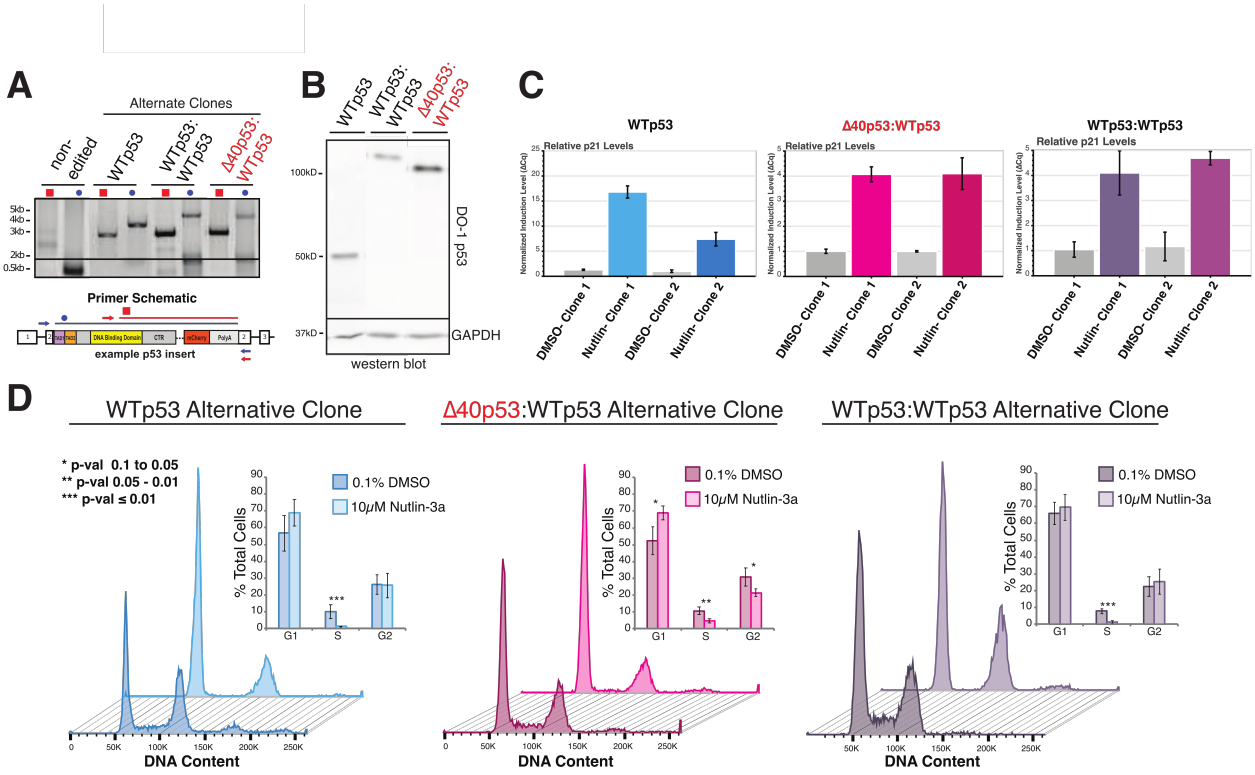


Figure 4.23: **Alternative p53 clones show similar phenotypic patterns as original clones selected.** (A) PCR validation of additional clones. Blue primer sets span the insertion and red primer sets have forward primer within the insertion and the reverse primer outside the insertion. Lack of 500bp band with blue primers indicated homozygous insertion. PCR experiments were completed on genomic DNA isolated from each of the indicated cell lines. Non-edited MCF10A cells were tested as an additional control. (B) Western blot validation of each genome-edited cell line, using a p53 antibody (DO-1). As designed, each tethered construct (Δ40p53:WTp53 or WTp53:WTp53) migrated at around 100kDa, indicative of dimer expression as a single transcript. (C) Clonal differences between the WTp53, WTp53:WTp53, or Δ40p53:WTp53 clones are minimal as shown by the relative induction read by RT-qPCR data are shown for p21 (2 biological replicates; bars = s.e.m.). (D) Cell cycle analysis (propidium iodide) of the second clone for each genome-edited cell line. Cell cycle is performed at 20hr of 0.1% DMSO treatment (control) and 20hr of Nutlin3a (treatment).

Chapter 5

Conclusion

Throughout this thesis, I have focused on the regulation and impact of transcription factor (TF) activity. First, I present a computational model that identifies all the TFs currently altering transcription from a single nascent RNA-sequencing assay, which is called TF profiling. TF profiling was used to define TF classes, such as cell type specific or shared across cell types. These TF classes were then used to elucidate distinct characteristics in terms of TF binding preferences and regulation. In p53-centric studies, I focused on the function of the p53 trans-activation domains (TAD) and their role in p53 transcriptional regulation. We show that the p53-TAD:Mediator interface can be blocked to inhibit p53 transcriptional activation. Additionally, we show that the naturally occurring p53 isoform, Δ40p53 (missing most of TAD1), in conjunction with WTp53, blunts the cellular p53 response, but does not have a p53 null phenotype. Overall, this work delves into broadly surveying TF activity and specifically modifying TF function.

5.1 Transcription factor profiling enables novel exploration of TF activity.

In Chapter 2, I present a robust statistical model, called TF profiling, that enables the identification of actively regulating TFs from a single nascent RNA-sequencing assay. Using this model we are able to extract known and novel cell type specific TFs, as well as TFs are actively regulating in many cell types.

This model can be applied to any cell type or tissue with nascent RNA-sequencing data to robustly determine all actively regulating TFs. This is an effective hypothesis generation tool that

is easily applied using the publicly available TF Profiler (`https://github.com/Dowell-Lab/TF_profiler`). Additionally, we plan to build a website associated with the data in the this thesis with searchable TF profiles based on cell type, tissue type or any TF of interest. Using the data produced in this study, we provide a general resource assessing TF activity in numerous cellular contexts which can be utilized for countless follow-up experiments.

Future work could explore how distinct TF classes (cell type specific or shared) may interact or alter chromatin in a way to impact each others behavior. One hypothesis is that cell type specific TFs establish open chromatin regions that are then maintained by shared TFs[257]. Taking this a step farther, it's possible that these cell type specific open regions are now targets for the binding of inducible TFs. That is to say, cell type specific TFs establish cell type specific enhancers and thus determine which inducible TF binding sites will be utilized in a given cellular context. Alterations of these binding sites may alter inducible TF activity and modulate their response. If cell type specific TFs and regions are known, we may be able to predict differing stimulus responses in different cellular contexts. Consistent with this, Sigauke et. al. shows that while p53 has a highly conserved network of gene targets[6, 9], the enhancers utilized to drive those targets differ on a cell type specific basis[218]. It's possible that the determinant for alternative enhancer usage in distinct cellular contexts is, in fact, the cell type specific TFs present.

Other future work involving the TF profiles includes integrating them with transcriptional regulatory networks (TRNs). In Sigauke et. al. we show that the transcription level of enhancer associated bidirectionals is correlated with the transcription level of the gene in a cell type specific manner[14, 218]. This correlation can be used to assess which enhancers are driving which genes, effectively building TRNs through nascent RNA-sequencing data. A natural follow-up to the TF profiling study in conjunction with the TRN work in Sigauke et. al. is attributing specific enhancers to TFs within the TF profiles, then linking those enhancers to the gene(s) they regulate. This would enable one to pair every TF within a profile to every gene that it potentially regulates.

5.2 The modulation of TF trans-activation domains and its impact on transcription.

My work establishes the importance of the p53-TADs in the p53 transcriptional response. I show that the p53-TAD:Mediator interface is a targetable region for specific transcription inhibition. I also show how the naturally occurring isoform Δ40p53 harnesses the loss of p53-TAD1 to blunt the p53 response.

In Chapter 3, we demonstrate the importance of both p53 activation domains in Mediator-dependent transcription activation. In this study we demonstrate that we can effectively and specifically inhibit the p53 transcriptional response by targeting Mediator where p53-TAD would normally bind using a bivalent peptide. An implication is that blocking a single Mediator-TF interaction will not affect other stimulus-responsive or lineage-specific TFs, thus providing a means to selectively alter gene expression patterns. As TFs have been historically difficult to therapeutically target[42], this study demonstrates that targeting Mediator-TF interfaces may be a viable alternative.

In future studies, it would be interesting to scale the blocking p53-TAD:Mediator for specific p53 inhibition to other TF:Mediator interactions. The first objective would be to build a compendium of known TF:Mediator interaction sites. Additional TFs of biomedical interest with unknown interaction sites could also be probed for these interfaces. The study here uses a bivalent peptide to block the Mediator-p53 interaction. While effective for a proof-of-concept study, the switch to small molecule inhibitors would be essential for viable therapeutics. This is primarily due to the ease of which peptides can be metabolized by humans. These additional studies would be of interest for the development of additional therapeutics, specifically in chemotherapy as many mutated TFs gain oncogenic function[139, 209].

In Chapter 4, we probe the function of naturally occurring Δ40p53 isoform in conjunction with the WTp53 isoform in a stoichiometrically controlled system (2Δ40p53:2WTp53). We show that the Δ40p53 isoform modulates the WTp53 function to suppress the typical p53 response.

However, the Δ40p53:WTp53 does not induce a p53 null phenotype. Instead, some p53 target genes are still activated upon p53 related stress, indicating that Δ40p53:WTp53 dampens the p53 response without abolishing it. Related to this finding is that the Δ40p53:WTp53 doesn't generate p53 induced eRNAs. Many studies have linked eRNA transcription with gene expression[14]. As the Δ40p53:WTp53 causes a dampened p53 response at genes, there is also dampened eRNA production.

Δ40p53:WTp53 can be used as an exemplar to better understand how TFs may use isoforms to regulate their function. Additionally, it's a model for how TFs with TAD mutations have altered, but not absent, transcriptional responses. One possible mode of these attenuated transcriptional responses could be permitting other sequence-specific, DNA-binding TFs to drive the cellular response. Future experiments include probing other TF-TAD isoforms or mutations and their direct impact on transcriptional output. The focus of these studies would be to explore the way that TFs utilize their TADs in different contexts to modulate function. In all, the work presented in this thesis makes significant contributions towards understand TF activity.

Bibliography

[1] US DOE Joint Genome Institute: Hawkins Trevor 4 Branscomb Elbert 4 Predki Paul 4 Richardson Paul 4 Wenning Sarah 4 Slezak Tom 4 Doggett Norman 4 Cheng Jan-Fang 4 Olsen Anne 4 Lucas Susan 4 Elkin Christopher 4 Uberbacher Edward 4 Frazier Marvin 4, RIKEN Genomic Sciences Center: Sakaki Yoshiyuki 9 Fujiyama Asao 9 Hattori Masahira 9 Yada Tetsushi 9 Toyoda Atsushi 9 Itoh Takehiko 9 Kawagoe Chiharu 9 Watanabe Hidemi 9 Totoki Yasushi 9 Taylor Todd 9, Genoscope, CNRS UMR-8030: Weissenbach Jean 10 Heilig Roland 10 Saurin William 10 Artiguenave Francois 10 Brottier Philippe 10 Bruls Thomas 10 Pelletier Eric 10 Robert Catherine 10 Wincker Patrick 10, Institute of Molecular Biotechnology: Rosenthal André 12 Platzer Matthias 12 Nyakatura Gerald 12 Taudien Stefan 12 Rump Andreas 12 Department of Genome Analysis, GTC Sequencing Center: Smith Douglas R. 11 Doucette-Stamm Lynn 11 Rubenfield Marc 11 Weinstock Keith 11 Lee Hong Mei 11 Dubois JoAnn 11, Beijing Genomics Institute/Human Genome Center: Yang Huanming 13 Yu Jun 13 Wang Jian 13 Huang Guyang 14 Gu Jun 15, et al. Initial sequencing and analysis of the human genome. nature, 409(6822):860–921, 2001.

[2] K. Adelman and J. T. Lis. Promoter-proximal pausing of rna polymerase ii: emerging roles in metazoans. Nat Rev Genet, 13(10):720–31, 2012.

[3] M. S. Akhtar, M. Heidemann, J. R. Tietjen, D. W. Zhang, R. D. Chapman, D. Eick, and A. Z. Ansari. Tfiih kinase places bivalent marks on the carboxy-terminal domain of rna polymerase ii. Mol Cell, 34(3):387–93, 2009.

[4] Nader Alerasool, He Leng, Zhen-Yuan Lin, Anne-Claude Gingras, and Mikko Taipale. Identification and functional characterization of transcriptional activators in human cells. Molecular cell, 82(3):677–695, 2022.

[5] B. L. Allen, K. Quach, T. Jones, C. B. Levandowski, C. C. Ebmeier, J. D. Rubin, T. Read, R. D. Dowell, A. Schepartz, and D. J. Taatjes. Suppression of p53 response by targeting p53-mediator binding with a stapled peptide. Cell Rep, 39(1):110630, 2022.

[6] M. A. Allen, Z. Andrysik, V. L. Dengler, H. S. Mellert, A. Guarnieri, J. A. Freeman, K. D. Sullivan, M. D. Galbraith, X. Luo, W. L. Kraus, R. D. Dowell, and J. M. Espinosa. Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. Elife (Cambridge), 3:e02200, 2014.

[7] T. Anbarasan and J. C. Bourdon. The emerging landscape of p53 isoforms in physiology, cancer and degenerative diseases. Int J Mol Sci, 20(24):6257, 2019.

[8] S. Anders and W. Huber. Differential expression analysis for sequence count data. Genome Biol, 11(10):R106, 2010.

[9] Z. Andrysik, M. D. Galbraith, A. L. Guarnieri, S. Zaccara, K. D. Sullivan, A. Pandey, M. MacBeth, A. Inga, and J. M. Espinosa. Identification of a core tp53 transcriptional program with highly distributed tumor suppressive activity. Genome Res, 27(10):1645–1657, 2017.

[10] Francisco Antequera and Adrian Bird. Number of cpg islands and genes in human and mouse. Proceedings of the National Academy of Sciences, 90(24):11995–11999, 1993.

[11] Y. Aoi, E. R. Smith, A. P. Shah, E. J. Rendleman, S. A. Marshall, A. R. Woodfin, F. X. Chen, R. Shiekhattar, and A. Shilatifard. Nelf regulates a promoter-proximal step distinct from rna pol ii pause-release. Mol Cell, 78(2):261–274 e5, 2020.

[12] R. Aramayo, M.B. Sherman, K. Brownless, R. Lurz, A.L. Okorokov, and E.V. Orlova. Quaternary structure of the specific p53-dna complex reveals the mechanism of p53 mutant dominance. Nucleic Acids Research, 39:8960–8971, 2011.

[13] E. Arner, C. O. Daub, K. Vitting-Seerup, R. Andersson, B. Lilje, F. Drablos, A. Lennartsson, M. Ronnerblad, O. Hrydziuszko, M. Vitezic, T. C. Freeman, A. M. Alhendi, P. Arner, R. Axton, J. K. Baillie, A. Beckhouse, B. Bodega, J. Briggs, F. Brombacher, M. Davis, M. Detmar, A. Ehrlund, M. Endoh, A. Eslami, M. Fagiolini, L. Fairbairn, G. J. Faulkner, C. Ferrai, M. E. Fisher, L. Forrester, D. Goldowitz, R. Guler, T. Ha, M. Hara, M. Herlyn, T. Ikawa, C. Kai, H. Kawamoto, L. M. Khachigian, S. P. Klinken, S. Kojima, H. Koseki, S. Klein, N. Mejhert, K. Miyaguchi, Y. Mizuno, M. Morimoto, K. J. Morris, C. Mummery, Y. Nakachi, S. Ogishima, M. Okada-Hatakeyama, Y. Okazaki, V. Orlando, D. Ovchinnikov, R. Passier, M. Patrikakis, A. Pombo, X. Y. Qin, S. Roy, H. Sato, S. Savvi, A. Saxena, A. Schwegmann, D. Sugiyama, R. Swoboda, H. Tanaka, A. Tomoiu, L. N. Winteringham, E. Wolvetang, C. Yanagi-Mizuochi, M. Yoneda, S. Zabierowski, P. Zhang, I. Abugessaisa, N. Bertin, A. D. Diehl, S. Fukuda, M. Furuno, J. Harshbarger, A. Hasegawa, F. Hori, S. Ishikawa-Kato, Y. Ishizu, M. Itoh, T. Kawashima, M. Kojima, N. Kondo, M. Lizio, T. F. Meehan, C. J. Mungall, M. Murata, H. Nishiyori-Sueki, S. Sahin, S. Nagao-Sato, J. Severin, M. J. de Hoon, J. Kawai, T. Kasukawa, T. Lassmann, et al. Gene regulation. transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science, 347(6225):1010–4, 2015.

[14] P. R. Arnold, A. D. Wells, and X. C. Li. Diversity and emerging roles of enhancer rna in regulation of gene expression and cell fate. Front Cell Dev Biol, 7:377, 2019.

[15] E. Aronesty. Comparison of sequencing utility programs. The Open Bioinformatics Journal, 7:1–8, 2013.

[16] K. A. Audetat, M. D. Galbraith, A. T. Odell, T. Lee, A. Pandey, J. M. Espinosa, R. D. Dowell, and D. J. Taatjes. A kinase-independent role for cyclin-dependent kinase 19 in p53 response. Mol Cell Biol, 37(13):e00626–16, 2017.

[17] J. G. Azofeifa, M. A. Allen, J. R. Hendrix, T. Read, J. D. Rubin, and R. D. Dowell. Enhancer rna profiling predicts transcription factor activity. Genome Res, 28:334–344, 2018.

[18] Joseph G. Azofeifa and Robin D. Dowell. A generative model for the behavior of rna polymerase. Bioinformatics, 33(2):227–234, Jan 2017.

[19] Timothy L Bailey, James Johnson, Charles E Grant, and William S Noble. The meme suite. Nucleic acids research, 43(W1):W39–W49, 2015.

[20] D. J. Baker, B. G. Childs, M. Durik, M. E. Wijers, C. J. Sieben, J. Zhong, R. A. Saltness, K. B. Jeganathan, G. C. Verzosa, A. Pezeshki, K. Khazaie, J. D. Miller, and J. M. van Deursen. Naturally occurring p16(ink4a)-positive cells shorten healthy lifespan. Nature, 530(7589):184–9, 2016.

[21] Julian Banerji, Sandro Rusconi, and Walter Schaffner. Expression of a β-globin gene is enhanced by remote sv40 dna sequences. Cell, 27(2):299–308, 1981.

[22] Evan H Baugh, Hua Ke, Arnold J Levine, Richard A Bonneau, and Chang S Chan. Why are there hotspot mutations in the tp53 gene in human cancers? Cell Death & Differentiation, 25(1):154–160, 2018.

[23] Susan M Berget, Claire Moore, and Phillip A Sharp. Spliced segments at the $5'$ terminus of adenovirus 2 late mrna. Proceedings of the National Academy of Sciences, 74(8):3171–3175, 1977.

[24] F. Bernal, A.F. Tyler, S.J. Korsmeyer, L.D. Walensky, and G.L. Verdine. Reactivation of the p53 tumor suppressor pathway by a stapled p53 peptide. J Am Chem Soc, 129:2456–2457, 2007.

[25] Kathryn T Bieging and Laura D Attardi. Deconstructing p53 transcriptional networks in tumor suppression. Trends in cell biology, 22(2):97–106, 2012.

[26] Adrian P Bird. Dna methylation and the frequency of cpg in animal dna. Nucleic acids research, 8(7):1499–1504, 1980.

[27] A. Bitto, A. M. Wang, C. F. Bennett, and M. Kaeberlein. Biochemical genetic pathways that modulate aging in multiple species. Cold Spring Harb Perspect Med, 5(11):a025114, 2015.

[28] M. Blagosklonny, G. Wu, K. Somasundaram, and W. Eldeiry. Wild-type p53 is not sufficient for serum starvation-induced apoptosis in cancer cells but accelerates apoptosis in sensitive cells. Int J Oncol, 11(6):1165–70, 1997.

[29] Amit Blumberg, Yixin Zhao, Yi-Fei Huang, Noah Dukler, Edward J. Rice, Alexandra G. Chivu, Katie Krumholz, Charles G. Danko, and Adam Siepel. Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data. BMC Biology, 19(1):30, February 2021.

[30] J. Bourdon, K. Fernandes, F. Murray-Zmijewski, F. Liu, A. Diot, D.P. Xirodimas, M.K. Saville, and D.P. Lane. p53 isoforms can regulate p53 transcriptional activity. Genes & Development, 19:2122–2137, 2005.

[31] K. Bourougaa, N. Naski, C. Boularan, C. Mlynarczyk, M. M. Candeias, S. Marullo, and R. Fahraeus. Endoplasmic reticulum stress induces g2 cell-cycle arrest via mrna translation of the p53 isoform p53/47. Mol Cell, 38(1):78–88, 2010.

[32] Maria Bouvy-Liivrand, Ana Hernández de Sande, Petri Pölönen, Juha Mehtonen, Tapio Vuorenmaa, Henri Niskanen, Lasse Sinkkonen, Minna Unelma Kaikkonen, and Merja

Heinäniemi. Analysis of primary microrna loci from nascent transcriptomes reveals regulatory domains governed by chromatin architecture. Nucleic Acids Research, 45(17):9837–9849, September 2017.

[33] M. E. Bowen, J. McClendon, H. K. Long, A. Sorayya, J. L. Van Nostrand, J. Wysocka, and L. D. Attardi. The spatiotemporal pattern and intensity of p53 activation dictates phenotypic diversity in p53-driven developmental syndromes. Dev Cell, 50(2):212–228 e6, 2019.

[34] Chris A Brackley, James Johnson, Steven Kelly, Peter R Cook, and Davide Marenduzzo. Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains. Nucleic acids research, 44(8):3503–3512, 2016.

[35] J. E. Bradner, D. Hnisz, and R. A. Young. Transcriptional addiction in cancer. Cell, 168(4):629–643, 2017.

[36] C.A. Brady, D. Jiang, S.S. Mello, T.M. Johnson, L.A. Jarvis, M.M. Kozak, D.K. Broz, S. Basak, E.J. Park, M.E. McLaughlin, A.N. Karnezis, and L.D. Attardi. Distinct p53 transcriptional programs dictate acute dna-damage responses and tumor suppression. Cell, 145:571–583, 2011.

[37] Gloria A. Brar and Jonathan S. Weissman. Ribosome profiling reveals the what, when, where and how of protein synthesis. Nature Reviews Molecular Cell Biology, 16(1111):651–664, Nov 2015.

[38] R. Brosh and V. Rotter. When mutants gain new powers: news from the mutant p53 field. Nat Rev Cancer, 9(10):701–13, 2009.

[39] C. J. Brown, S. T. Quah, J. Jong, A. M. Goh, P. C. Chiam, K. H. Khoo, M. L. Choong, M. A. Lee, L. Yurlova, K. Zolghadr, T. L. Joseph, C. S. Verma, and D. P. Lane. Stapled peptides with improved potency and specificity that activate p53. ACS Chem Biol, 8(3):506–12, 2013.

[40] AD Buffry, CC Mendes, and AP McGregor. The functionality and evolution of eukaryotic transcriptional enhancers, 2016.

[41] M. Bulger and M. Groudine. Functional and mechanistic diversity of distal transcription enhancers. Cell, 144:327–339, 2011.

[42] J. H. Bushweller. Targeting transcription factors in cancer - from undruggable to reality. Nat Rev Cancer, 19(11):611–624, 2019.

[43] G.T. Cantin, J.L. Stevens, and A. J. Berk. Activation domain-mediator interactions promote transcription preinitiation complex assembly on promoter dna. Proc Natl Acad Sci U S A, 100(21):12003–12008, 2003.

[44] J. F. Cardiello, G. J. Sanchez, M. A. Allen, and R. D. Dowell. Lessons from ernas: understanding transcriptional regulation through the lens of nascent rnas. Transcription, 11(1):3–18, 2020.

[45] Joseph F. Cardiello, James A. Goodrich, and Jennifer F. Kugel. Heat shock causes a reversible increase in rna polymerase ii occupancy downstream of mrna genes, consistent with a global loss in transcriptional termination. Molecular and Cellular Biology, 38(18):e00181–18, Aug 2018.

[46] Y. S. Chang, B. Graves, V. Guerlavais, C. Tovar, K. Packman, K. H. To, K. A. Olson, K. Kesavan, P. Gangurde, A. Mukherjee, T. Baker, K. Darlak, C. Elkin, Z. Filipovic, F. Z. Qureshi, H. Cai, P. Berry, E. Feyfant, X. E. Shi, J. Horstick, D. A. Annis, A. M. Manning, N. Fotouhi, H. Nash, L. T. Vassilev, and T. K. Sawyer. Stapled alpha-helical peptide drug development: a potent dual inhibitor of mdm2 and mdmx for p53-dependent cancer therapy. Proc Natl Acad Sci U S A, 110(36):E3445–54, 2013.

[47] Patrick Chène. Inhibiting the p53–mdm2 interaction: an important target for cancer therapy. Nature reviews cancer, 3(2):102–109, 2003.

[48] K. C. Clopper and D. J. Taatjes. Chemical inhibitors of transcription-associated kinases. Curr Opin Chem Biol, 70:102186, 2022.

[49] T. H. Collet, T. Sonoyama, E. Henning, J. M. Keogh, B. Ingram, S. Kelway, L. Guo, and I. S. Farooqi. A metabolomic signature of acute caloric restriction. J Clin Endocrinol Metab, 102(12):4486–4495, 2017.

[50] Sheila Connelly and James L Manley. A functional mrna polyadenylation signal is required for transcription termination by rna polymerase ii. Genes & development, 2(4):440–452, 1988.

[51] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. Nature, 489(7414):57, 2012.

[52] Leighton J. Core, André L. Martins, Charles G. Danko, Colin T. Waters, Adam Siepel, and John T. Lis. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature Genetics, 46(12):1311–1320, Dec 2014.

[53] Leighton J Core, Joshua J Waterfall, and John T Lis. Nascent rna sequencing reveals widespread pausing and divergent initiation at human promoters. Science, 322(5909):1845–1848, 2008.

[54] M. A. Cortazar, R. M. Sheridan, B. Erickson, N. Fong, K. Glover-Cutter, K. Brannan, and D. L. Bentley. Control of rna pol ii speed by pnuts-pp1 and spt5 dephosphorylation facilitates termination by a sitting duck torpedo mechanism. Mol Cell, 76(6):896–908 e4, 2019.

[55] S. Courtois, G. Verhaegh, S. North, M. G. Luciani, P. Lassus, U. Hibner, M. Oren, and P. Hainaut. Deltan-p53, a natural isoform of p53 lacking the first transactivation domain, counteracts growth suppression by wild-type p53. Oncogene, 21(44):6722–8, 2002.

[56] P. Cramer. Organization and regulation of gene transcription. Nature, 573(7772):45–54, 2019.

[57] Simon L. Currie, Jedediah J. Doane, Kathryn S. Evans, Niraja Bhachech, Bethany J. Madison, Desmond K. W. Lau, Lawrence P. McIntosh, Jack J. Skalicky, Kathleen A. Clark, and Barbara J. Graves. Etv4 and ap1 transcription factors form multivalent interactions with three sites on the med25 activator-interacting domain. Journal of Molecular Biology, 429(20):2975–2995, Oct 2017.

[58] Charles G Danko, Stephanie L Hyland, Leighton J Core, Andre L Martins, Colin T Waters, Hyung Won Lee, Vivian G Cheung, W Lee Kraus, John T Lis, and Adam Siepel. Identification of active transcriptional regulatory elements from GRO-seq data. Nat Meth, 12(5):433–438, 05 2015.

[59] Robert L. Davis, Harold Weintraub, and Andrew B. Lassar. Expression of a single transfected cdna converts fibroblasts to myoblasts. Cell, 51(6):987–1000, Dec 1987.

[60] C. D. Dehaven, A. M. Evans, H. Dai, and K. A. Lawton. Organization of gc/ms and lc/ms metabolomics data into chemical libraries. J Cheminform, 2(1):9, 2010.

[61] M. Dejosez, H. Ura, V. L. Brandt, and T. P. Zwaka. Safeguards for cell cooperation in mouse embryogenesis shown by genome-wide cheater screen. Science, 341(6153):1511–4, 2013.

[62] Z. N. Demidenko, L. G. Korotchkina, A. V. Gudkov, and M. V. Blagosklonny. Paradoxical suppression of cellular senescence by p53. Proc Natl Acad Sci U S A, 107(21):9660–4, 2010.

[63] Sarah Diab, Mingfeng Yu, and Shudong Wang. Cdk7 inhibitors in cancer therapy: the sweet smell of success? Journal of medicinal chemistry, 63(14):7458–7474, 2020.

[64] Noah Dukler, Gregory T. Booth, Yi-Fei Huang, Nathaniel Tippens, Colin T. Waters, Charles G. Danko, John T. Lis, and Adam Siepel. Nascent RNA sequencing reveals a dynamic global transcriptional response at genes and enhancers to the natural medicinal compound celastrol. Genome Research, 27(11):1816–1829, 2017.

[65] M. Dumble, L. Moore, S. M. Chambers, H. Geiger, G. Van Zant, M. A. Goodell, and L. A. Donehower. The impact of altered p53 dosage on hematopoietic stem cell dynamics during aging. Blood, 109(4):1736–42, 2007.

[66] C.C. Ebmeier and D.J. Taatjes. Activator-mediator binding regulates mediator-cofactor interactions. Proc Natl Acad Sci U S A, 107(25):11283–11288, 2010.

[67] Sylvain Egloff and Shona Murphy. Cracking the rna polymerase ii ctd code. Trends in genetics, 24(6):280–288, 2008.

[68] Conchi Estarás, Chris Benner, and Katherine A Jones. SMADs and YAP compete to control elongation of β-catenin:LEF-1-recruited RNAPII during hESC differentiation. Mol Cell, 58(5):780–93, Jun 2015.

[69] Charles R. Evans, Alla Karnovsky, Melissa A. Kovach, Theodore J. Standiford, Charles F. Burant, and Kathleen A. Stringer. Untargeted lc–ms metabolomics of bronchoalveolar lavage fluid differentiates acute respiratory distress syndrome from health. Journal of Proteome Research, 13(2):640–649, Feb 2014.

[70] C. B. Fant, C. B. Levandowski, K. Gupta, Z. L. Maas, J. Moir, J. D. Rubin, A. Sawyer, M. N. Esbin, J. K. Rimel, O. Luyties, M. T. Marr, I. Berger, R. D. Dowell, and D. J. Taatjes. Tfiid enables rna polymerase ii promoter-proximal pausing. Mol Cell, 78(4):785–793, 2020.

[71] Charli B. Fant, Cecilia B. Levandowski, Kapil Gupta, Zachary L. Maas, John Moir, Jonathan D. Rubin, Andrew Sawyer, Meagan N. Esbin, Jenna K. Rimel, Olivia Luyties, Michael T. Marr, Imre Berger, Robin D. Dowell, and Dylan J. Taatjes. Tfiid enables rna polymerase ii promoter-proximal pausing. Molecular Cell, 78(4):785 – 793.e8, 2020.

[72] Romain Fenouil, Pierre Cauchy, Frederic Koch, Nicolas Descostes, Joaquin Zacarias Cabeza, Charlène Innocenti, Pierre Ferrier, Salvatore Spicuglia, Marta Gut, Ivo Gut, et al. Cpg islands and gc content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome research, 22(12):2399–2408, 2012.

[73] J.C. Ferreon, C.W. Lee, M. Arai, M.A. Martinez-Yamout, H.J. Dyson, and P.E. Wright. Cooperative regulation of p53 by modulation of ternary complex formation with cbp/p300 and hdm2. Proc Natl Acad Sci U S A, 106(16):6591–6596, 2009.

[74] John J Ferrie, Jonathan P Karr, Robert Tjian, and Xavier Darzacq. "structure"-function relationships in eukaryotic transcription factors: The role of intrinsically disordered regions in gene regulation. Molecular Cell, 2022.

[75] T. Finkel and N. J. Holbrook. Oxidants, oxidative stress and the biology of ageing. Nature, 408(6809):239–47, 2000.

[76] M. Fischer. Census and evaluation of p53 target genes. Oncogene, 36(28):3943–3956, 2017.

[77] R. P. Fisher. Secrets of a double agent: Cdk7 in cell-cycle control and transcription. J Cell Sci, 118(Pt 22):5171–80, 2005.

[78] P.M. Flanagan, R.J. Kelleher-III., M.H. Sayre, H. Tschochner, and R. Kornberg. A mediator required for activation of rna polymerase ii transcription in vitro. Nature, 350:436–438, 1991.

[79] J D Fondell, H Ge, and R G Roeder. Ligand induction of a transcriptionally active thyroid hormone receptor coactivator complex. Proceedings of the National Academy of Sciences, 93(16):8329–8333, Aug 1996.

[80] N. Fong, H. Kim, Y. Zhou, X. Ji, J. Qiu, T. Saldi, K. Diener, K. Jones, X. D. Fu, and D. L. Bentley. Pre-mrna splicing is facilitated by an optimal rna polymerase ii elongation rate. Genes Dev, 28(23):2663–76, 2014.

[81] H. L. Franco, A. Nagari, V. S. Malladi, W. Li, Y. Xi, D. Richardson, K. L. Allton, K. Tanaka, J. Li, S. Murakami, K. Keyomarsi, M. T. Bedford, X. Shi, W. Li, M. C. Barton, S. Y. R. Dent, and W. L. Kraus. Enhancer transcription reveals subtype-specific gene expression programs controlling breast cancer pathogenesis. Genome Res, 28(2):159–170, 2018.

[82] William A. Freed-Pastor and Carol Prives. Mutant p53: one name, many proteins. Genes & Development, 26(12):1268–1286, Jun 2012. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press publisher: Cold Spring Harbor Lab PMID: 22713868.

[83] Seth Frietze and Peggy J Farnham. Transcription factor effector domains. A handbook of transcription factors, pages 261–277, 2011.

[84] E. E. M. Furlong and M. Levine. Developmental enhancers and chromosome topology. Science, 361(6409):1341–1345, 2018.

[85] Dimos Gaidatzis, Lukas Burger, Maria Florescu, and Michael B Stadler. Analysis of intronic and exonic reads in rna-seq data characterizes transcriptional and post-transcriptional regulation. Nature biotechnology, 33(7):722–729, 2015.

[86] V. Gambino, G. De Michele, O. Venezia, P. Migliaccio, V. Dall'Olio, L. Bernard, S. P. Minardi, M. A. Della Fazia, D. Bartoli, G. Servillo, M. Alcalay, L. Luzi, M. Giorgio, H. Scrable, P. G. Pelicci, and E. Migliaccio. Oxidative stress activates a specific p53 transcriptional response that regulates cellular senescence and aging. Aging Cell, 12(3):435–45, 2013.

[87] G. Gill and M. Ptashne. Negative effect of the transcriptional activator gal4. Nature, 334(6184):721–4, 1988.

[88] K. Glover-Cutter, S. Larochelle, B. Erickson, C. Zhang, K. Shokat, R. P. Fisher, and D. L. Bentley. Tfiih-associated cdk7 kinase functions in phosphorylation of c-terminal domain ser7 residues, promoter-proximal pausing, and termination by rna polymerase ii. Mol Cell Biol, 29(20):5455–64, 2009.

[89] Steven R Grossman. p300/cbp/p53 interaction and regulation of the p53 response. European journal of biochemistry, 268(10):2773–2778, 2001.

[90] S. Grunberg, L. Warfield, and S. Hahn. Architecture of the rna polymerase ii preinitiation complex and mechanism of atp-dependent promoter opening. Nat Struct Mol Biol, 19(8):788–96, 2012.

[91] Yan-Jun Guo, Wei-Wei Pan, Sheng-Bing Liu, Zhong-Fei Shen, Ying Xu, and Ling-Ling Hu. Erk/mapk signalling pathway and tumorigenesis. Experimental and Therapeutic Medicine, 19(3):1997–2007, Mar 2020.

[92] V. Haberle and A. Stark. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol, 19(10):621–637, 2018.

[93] A. Hafner, M. L. Bulyk, A. Jambhekar, and G. Lahav. The multiple mechanisms that regulate p53 activity and cell fate. Nat Rev Mol Cell Biol, 20(4):199–210, 2019.

[94] Nasun Hah, Shino Murakami, Anusha Nagari, Charles G. Danko, and W. Lee Kraus. Enhancer transcripts mark active estrogen receptor binding sites. Genome Research, 23(8):1210–1223, 2013.

[95] Marc S Halfon. Studying transcriptional enhancers: the founder fallacy, validation creep, and other biases. Trends in Genetics, 35(2):93–103, 2019.

[96] Y. A. Hannun and L. M. Obeid. Sphingolipids and their metabolism in physiology and disease. Nat Rev Mol Cell Biol, 19(3):175–191, 2018.

[97] Paul E Hardin and Satchidananda Panda. Circadian timekeeping and output mechanisms in animals. Current opinion in neurobiology, 23(5):724–731, 2013.

[98] S. Harris and A.J. Levine. The p53 pathway: positive and negative feedback loops. Oncogene, 24:2899–2908, 2005.

[99] Douglas A. Harrison. The jak/stat pathway. Cold Spring Harbor Perspectives in Biology, 4(3):a011205, Mar 2012.

[100] Y. Haupt, R. Maya, A. Kazaz, and M. Oren. Mdm2 promotes the rapid degradation of p53. Nature, 387(6630):296–9, 1997.

[101] Oliver Hendy, Leonid Serebreni, Katharina Bergauer, Felix Muerdter, Lukas Huber, Filip Nemčko, and Alexander Stark. Developmental and housekeeping transcriptional programs in drosophila require distinct chromatin remodelers. Molecular Cell, 82(19):3598–3612, 2022.

[102] E. Herbig, L. Warfield, L. Fish, J. Fishburn, B. A. Knutson, B. Moorefield, D. Pacheco, and S. Hahn. Mechanism of mediator recruitment by tandem gcn4 activation domains and three gal11 activator-binding domains. Mol Cell Biol, 30(10):2376–90, 2010.

[103] G. W. Hinkal, C. E. Gatza, N. Parikh, and L. A. Donehower. Altered senescence, apoptosis, and dna damage response in a mutant p53 model of accelerated aging. Mech Ageing Dev, 130(4):262–71, 2009.

[104] D. Hnisz, B. J. Abraham, T. I. Lee, A. Lau, V. Saint-Andre, A. A. Sigova, H. A. Hoke, and R. A. Young. Super-enhancers in the control of cell identity and disease. Cell, 155(4):934–47, 2013.

[105] D. Hnisz, K. Shrinivas, R. A. Young, A. K. Chakraborty, and P. A. Sharp. A phase separation model for transcriptional control. Cell, 169(1):13–23, 2017.

[106] Monica Hollstein, David Sidransky, Bert Vogelstein, and Curtis C Harris. p53 mutations in human cancers. Science, 253(5015):49–53, 1991.

[107] F. C. Holstege, E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, M. R. Green, T. R. Golub, E. S. Lander, and R. A. Young. Dissecting the regulatory circuitry of a eukaryotic genome. Cell, 95(5):717–28, 1998.

[108] R. Honda, H. Tanaka, and H. Yasuda. Oncoprotein mdm2 is a ubiquitin ligase e3 for tumor suppressor p53. FEBS letters, 420(1):25–27, Dec 1997.

[109] Tien Hsu, Maria Trojanowska, and Dennis K Watson. Ets proteins in biological control and cancer. Journal of cellular biochemistry, 91(5):896–903, 2004.

[110] Fang Huang, Sridharan Rajagopalan, Giovanni Settanni, Richard J. Marsh, Daven A. Armoogum, Nick Nicolaou, Angus J. Bain, Eitan Lerner, Elisha Haas, Liming Ying, and Alan R. Fersht. Multiple conformations of full-length p53 detected with single-molecule fluorescence resonance energy transfer. Proceedings of the National Academy of Sciences, 106(49):20758–20763, Dec 2009.

[111] Lei Huang, Zheng Yan, Xiaodong Liao, Yuan Li, Jie Yang, Zhu-Gang Wang, Yong Zuo, Hidehiko Kawai, Miriam Shadfan, Suthakar Ganapathy, and Zhi-Min Yuan. The p53 inhibitors mdm2/mdmx complex is required for control of p53 activity in vivo. Proceedings of the National Academy of Sciences, 108(29):12001–12006, Jul 2011.

[112] Sachi Inukai, Kian Hong Kock, and Martha L Bulyk. Transcription factor–dna binding: beyond binding site motifs. Current opinion in genetics & development, 43:110–119, 2017.

[113] Ilya P Ioshikhes and Michael Q Zhang. Large-scale human promoter mapping using cpg islands. Nature genetics, 26(1):61–63, 2000.

[114] M. Ito, C. X. Yuan, H. J. Okano, R. B. Darnell, and R. G. Roeder. Involvement of the trap220 component of the trap/smcc coactivator complex in embryonic development and thyroid hormone action. Mol. Cell, 5(4):683–93, 2000.

[115] Mitsuhiro Ito, Chao-Xing Yuan, Sohail Malik, Wei Gu, Joseph D Fondell, Soichiro Yamamura, Zheng-Yuan Fu, Xiaolong Zhang, Jun Qin, and Robert G Roeder. Identity between trap and smcc complexes indicates novel pathways for the function of nuclear receptors and diverse mammalian activators. Molecular Cell, 3(3):361–370, Mar 1999.

[116] Makiko Iwafuchi-Doi and Kenneth S Zaret. Cell fate control by pioneer transcription factors. Development, 143(11):1833–1837, 2016.

[117] Dadi Jiang, Colleen A. Brady, Thomas M. Johnson, Eunice Y. Lee, Eunice J. Park, Matthew P. Scott, and Laura D. Attardi. Full p53 transcriptional activation potential is dispensable for tumor suppression in diverse lineages. Proceedings of the National Academy of Sciences, 108(41):17123–17128, Oct 2011.

[118] G.S. Jimenez, M. Nister, J.M. Stommel, M. Beeche, E.A. Barcarse, X.Q. Zhang, S. O'Gorman, and G.M. Wahl. A transactivation-deficient mouse model provides insights into trp53 regulation and function. Nat Genet., 26(1):37–43, 2000.

[119] T.M. Johnson, E.M. Hammond, A. Giaccia, and L.D. Attardi. The p53qs transactivation-deficient mutant shows stress-specific apoptotic activity and induces embryonic lethality. Nat Genet., 37(2):145–152, 2005.

[120] K. Kamagata, S. Kanbayashi, M. Honda, Y. Itoh, H. Takahashi, T. Kameda, F. Nagatsugi, and S. Takahashi. Liquid-like droplet formation by tumor suppressor p53 induced by multivalent electrostatic interactions between two disordered domains. Sci Rep, 10(1):580, 2020.

[121] E. R. Kastenhuber and S. W. Lowe. Putting p53 in context. Cell, 170(6):1062–1078, 2017.

[122] K. H. Khoo, C. S. Verma, and D. P. Lane. Drugging the p53 pathway: understanding the route to clinical efficacy. Nat Rev Drug Discov, 13(3):217–36, 2014.

[123] Marie P Khoury and Jean-Christophe Bourdon. p53 isoforms: an intracellular microprocessor? Genes & cancer, 2(4):453–465, 2011.

[124] D. Kim, B. Langmead, and S. L. Salzberg. Hisat: a fast spliced aligner with low memory requirements. Nat Methods, 12(4):357–60, 2015.

[125] M. Kim, H. Suh, E. J. Cho, and S. Buratowski. Phosphorylation of the yeast rpb1 c-terminal domain at serines 2, 5, and 7. J Biol Chem, 284(39):26421–6, 2009.

[126] Tae-Kyung Kim, Martin Hemberg, Jesse M Gray, Allen M Costa, Daniel M Bear, Jing Wu, David A Harmin, Mike Laptewicz, Kellie Barbara-Haley, Scott Kuersten, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature, 465(7295):182–187, 2010.

[127] Matthew T. Knuesel, Krista D. Meyer, Carrie Bernecky, and Dylan J. Taatjes. The human cdk8 subcomplex is a molecular switch that controls mediator coactivator function. Genes & Development, 23(4):439–451, Feb 2009. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press publisher: Cold Spring Harbor Lab PMID: 19240132.

[128] L. G. Korotchkina, O. V. Leontieva, E. I. Bukreeva, Z. N. Demidenko, A. V. Gudkov, and M. V. Blagosklonny. The choice between p53-induced senescence and quiescence is determined in part by the mtor pathway. Aging (Albany NY), 2(6):344–52, 2010.

[129] V. Krizhanovsky and S.W. Lowe. Stem cells: The promises and perils of p53. Nature, 460:1085–1086, 2009.

[130] A. S. Krois, H. J. Dyson, and P. E. Wright. Long-range regulation of p53 dna binding by its intrinsically disordered n-terminal transactivation domain. Proc Natl Acad Sci U S A, 115(48):E11302–E11310, 2018.

[131] A. S. Krois, J. C. Ferreon, M. A. Martinez-Yamout, H. J. Dyson, and P. E. Wright. Recognition of the disordered p53 transactivation domain by the transcriptional adapter zinc finger domains of creb-binding protein. Proc Natl Acad Sci U S A, 113(13):E1853–62, 2016.

[132] F. Kruiswijk, C. F. Labuschagne, and K. H. Vousden. p53 in survival, death and metabolic health: a lifeguard with a licence to kill. Nat Rev Mol Cell Biol, 16(7):393–405, 2015.

[133] Michael H. G. Kubbutat, Stephen N. Jones, and Karen H. Vousden. Regulation of p53 stability by mdm2. Nature, 387(66306630):299–303, May 1997.

[134] Ivan V Kulakovskiy, Ilya E Vorontsov, Ivan S Yevshin, Ruslan N Sharipov, Alla D Fedorova, Eugene I Rumynskiy, Yulia A Medvedeva, Arturo Magana-Mora, Vladimir B Bajic, Dmitry A Papatsenko, et al. Hocomoco: towards a complete collection of transcription factor binding models for human and mouse via large-scale chip-seq analysis. Nucleic acids research, 46(D1):D252–D259, 2018.

[135] Hojoong Kwak, Nicholas J. Fuda, Leighton J. Core, and John T. Lis. Precise maps of rna polymerase reveal how promoters direct initiation and pausing. Science, 339(6122):950–953, Feb 2013.

[136] N. Kwiatkowski, T. Zhang, P. B. Rahl, B. J. Abraham, J. Reddy, S. B. Ficarro, A. Dastur, A. Amzallag, S. Ramaswamy, B. Tesar, C. E. Jenkins, N. M. Hannett, D. McMillin, T. Sanda, T. Sim, N. D. Kim, T. Look, C. S. Mitsiades, A. P. Weng, J. R. Brown, C. H. Benes, J. A. Marto, R. A. Young, and N. S. Gray. Targeting transcription regulation in cancer with a covalent cdk7 inhibitor. Nature, 511(7511):616–20, 2014.

[137] E. Lambert, K. Puwakdandawa, Y. F. Tao, and F. Robert. From structure to molecular condensates: emerging mechanisms for mediator function. FEBS J, 2021.

[138] S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch. The human transcription factors. Cell, 172(4):650–665, 2018.

[139] D. P. Lane. Cancer. p53, guardian of the genome. Nature, 358(6381):15–16, Jul 1992.

[140] Tong Ihn Lee and Richard A. Young. Transcriptional regulation and its misregulation in disease. Cell, 152(6):1237–1251, Mar 2013.

[141] Gary LeRoy, Ozgur Oksuz, Nicolas Descostes, Yuki Aoi, Rais A Ganai, Havva Ortabozkoyun Kara, Jia-Ray Yu, Chul-Hwan Lee, James Stafford, Ali Shilatifard, et al. LEDGF and HDGF2 relieve the nucleosome-induced barrier to transcription in differentiated cells. Science Advances, 5(10):eaay3068, 2019.

[142] C. B. Levandowski, T. Jones, M. Gruca, S. Ramamoorthy, R. D. Dowell, and D. J. Taatjes. The delta40p53 isoform inhibits p53-dependent erna transcription and enables regulation by signal-specific transcription factors during p53 activation. PLoS Biol, 19(8):e3001364, 2021.

[143] N. Leveille, C. A. Melo, K. Rooijers, A. Diaz-Lagares, S. A. Melo, G. Korkmaz, R. Lopes, F. A. Moqadam, A. R. Maia, P. J. Wijchers, G. Geeven, M. L. den Boer, R. Kalluri, W. de Laat, M. Esteller, and R. Agami. Genome-wide profiling of p53-regulated enhancer rnas uncovers a subset of enhancers controlled by a lncrna. Nat Commun, 6:6520, 2015.

[144] Bo Li, Triona Ni Chonghaile, Yue Fan, Stephen F Madden, Rut Klinger, Aisling E O'Connor, Louise Walsh, Gillian O'Hurley, Girish Mallya Udupi, Jesuchristopher Joseph, et al. Therapeutic rationale to target highly expressed cdk7 conferring poor outcomes in triple-negative breast cancercdk7 expression and function in triple-negative breast cancer. Cancer research, 77(14):3834–3845, 2017.

[145] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. Bioinformatics, 25(16):2078–2079, Aug 2009.

[146] Xiao-yong Li, Stewart MacArthur, Richard Bourgon, David Nix, Daniel A Pollard, Venky N Iyer, Aaron Hechmer, Lisa Simirenko, Mark Stapleton, Cris L Luengo Hendriks, et al. Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm. PLoS biology, 6(2):e27, 2008.

[147] Y. Liao, G. K. Smyth, and W. Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7):923–30, 2014.

[148] N.L. Lill, S.R. Grossman, D. Ginsberg, J. DeCaprio, and D.M. Livingston. Binding and modulation of p53 by p300/cbp coactivators. Nature, 387:823–827, 1997.

[149] J. Lin, J. Chen, B. Elenbaas, and A.J. Levine. Several hydrophobic amino acids in the p53 amino-terminal domain are required for transcriptional activation, binding to mdm-2 and the adenovirus 5 e1b 55-kd protein. Genes & Development, 8:1235–1246, 1994.

[150] S. Lin, B. T. Staahl, R. K. Alla, and J. A. Doudna. Enhanced homology-directed human genome engineering by controlled timing of crispr/cas9 delivery. Elife, 3:e04766, 2014.

[151] Shih-Chieh Lin, Edward D. Karoly, and Dylan J. Taatjes. The human δnp53 isoform triggers metabolic and gene expression changes that activate mtor and alter mitochondrial function. Aging Cell, 12(5):863–872, 2013.

[152] V. M. Link, S. H. Duttke, H. B. Chun, I. R. Holtman, E. Westin, M. A. Hoeksema, Y. Abe, D. Skola, C. E. Romanoski, J. Tao, G. J. Fonseca, T. D. Troutman, N. J. Spann, T. Strid, M. Sakai, M. Yu, R. Hu, R. Fang, D. Metzler, B. Ren, and C. K. Glass. Analysis of genetically diverse macrophages reveals local and domain-wide mechanisms that control transcription factor binding and function. Cell, 173(7):1796–1809 e17, 2018.

[153] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. Nature genetics, 45(6):580–585, 2013.

[154] C. Lopez-Otin, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer. The hallmarks of aging. Cell, 153(6):1194–217, 2013.

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

[156] Xin Luo, Minho Chae, Raga Krishnakumar, Charles G Danko, and W Lee Kraus. Dynamic reorganization of the AC16 cardiomyocyte transcriptome in response to TNFα signaling revealed by integrated genomic analyses. BMC Genomics, 15:155–155, 2014.

[157] O. Luyties and D. J. Taatjes. The mediator kinase module: an interface between cell signaling and transcription. Trends Biochem Sci, 47:314–327, 2022.

[158] Zachary L Maas and Robin D Dowell. Modeling variability and efficiency of spike-ins in nascent sequencing experiments. In review, 2023.

[159] Dig B Mahat, H. Hans Salamanca, Fabiana M Duarte, Charles G Danko, and John T Lis. Mammalian heat shock response and mechanisms underlying its genome-wide transcriptional regulation. Molecular Cell, 62(1):63–78, Apr 2016.

[160] Dig Bijay Mahat, Hojoong Kwak, Gregory T Booth, Iris H Jonkers, Charles G Danko, Ravi K Patel, Colin T Waters, Katie Munson, Leighton J Core, and John T Lis. Base-pair-resolution genome-wide mapping of active rna polymerases using precision nuclear run-on (pro-seq). Nature protocols, 11(8):1455–1476, 2016.

[161] B. Maier, W. Gluba, B. Bernier, T. Turner, K. Mohammad, T. Guise, A. Sutherland, M. Thorner, and H. Scrable. Modulation of mammalian life span by the short isoform of p53. Genes & Development, 18:306–319, 2004.

[162] Elaine R. Mardis. Dna sequencing technologies: 2006–2016. Nature Protocols, 12(22):213–218, Feb 2017.

[163] Jason J Marineau, Kristin B Hamman, Shanhu Hu, Sydney Alnemy, Janessa Mihalich, Anzhelika Kabro, Kenneth Matthew Whitmore, Dana K Winter, Stephanie Roy, Stephane Ciblat, et al. Discovery of sy-5609: a selective, noncovalent inhibitor of cdk7. Journal of Medicinal Chemistry, 65(2):1458–1480, 2021.

[164] N. F. Marshall and D. H. Price. Purification of p-tefb, a transcription factor required for the transition into productive elongation. Journal of Biological Chemistry, 270(21):12335–8, 1995.

[165] A. Gregory Matera and Zefeng Wang. A day in the life of the spliceosome. Nature Reviews Molecular Cell Biology, 15(22):108–121, Feb 2014.

[166] Alexandre Mayran and Jacques Drouin. Pioneer transcription factors shape the epigenetic landscape. Journal of Biological Chemistry, 293(36):13795–13804, 2018.

[167] Shenglin Mei, Qian Qin, Qiu Wu, Hanfei Sun, Rongbin Zheng, Chongzhi Zang, Muyuan Zhu, Jiaxin Wu, Xiaohui Shi, Len Taing, et al. Cistrome data browser: a data portal for chip-seq and chromatin accessibility data in human and mouse. Nucleic acids research, page gkw983, 2016.

[168] R. Melero, S. Rajagopalan, M. Lazaro, A. C. Joerger, T. Brandt, D. B. Veprintsev, G. Lasso, D. Gil, S. H. Scheres, J. M. Carazo, A. R. Fersht, and M. Valle. Electron microscopy studies on the quaternary structure of p53 reveal different binding modes for p53 tetramers in complex with dna. Proc Natl Acad Sci U S A, 108(2):557–62, 2011.

[169] J. P. Melis, E. M. Hoogervorst, C. T. van Oostrom, E. Zwart, T. M. Breit, J. L. Pennings, A. de Vries, and H. van Steeg. Genotoxic exposure: novel cause of selection for a functional deltan-p53 isoform. Oncogene, 30(15):1764–72, 2011.

[170] C. A. Melo, J. Drost, P. J. Wijchers, H. van de Werken, E. de Wit, J. A. Oude Vrielink, R. Elkon, S. A. Melo, N. Leveille, R. Kalluri, W. de Laat, and R. Agami. ernas are required for p53-dependent enhancer activity and gene transcription. Mol Cell, 49(3):524–35, 2013.

[171] F. T. Merkle, S. Ghosh, N. Kamitaki, J. Mitchell, Y. Avior, C. Mello, S. Kashin, S. Mekhoubad, D. Ilic, M. Charlton, G. Saphier, R. E. Handsaker, G. Genovese, S. Bar, N. Benvenisty, S. A. McCarroll, and K. Eggan. Human pluripotent stem cells recurrently acquire and expand dominant negative p53 mutations. Nature, 545(7653):229–233, 2017.

[172] K.D. Meyer, S. Lin, C. Bernecky, Y. Gao, and D.J. Taatjes. p53 activates transcription by directing structural shifts in mediator. Nat Struct Mol Biol, 17(6):753–760, 2010.

[173] M. E. Meyer, H. Gronemeyer, B. Turcotte, M. T. Bocquel, D. Tasset, and P. Chambon. Steroid hormone receptors compete for factors that mediate their enhancer function. Cell, 57(3):433–42, 1989.

[174] K. Meyer-Arendt, W. M. Old, S. Houel, K. Renganathan, B. Eichelberger, K. A. Resing, and N. G. Ahn. Isoformresolver: A peptide-centric algorithm for protein inference. J Proteome Res, 10(7):3060–75, 2011.

[175] A.G. Milbradt, M. Kulkarni, T. Yi, K. Takeuchi, Z.Y. Sun, R.E. Luna, P. Selenko, A.M. Naar, and G. Wagner. Structure of the vp16 transactivator target in the mediator. Nat Struct Mol Biol, 18:410–415, 2011.

[176] Chitvan Mittal, Olivia Lang, William KM Lai, and B Franklin Pugh. An integrated saga and tfiid pic assembly pathway selective for poised and induced promoters. Genes & Development, 36(17-18):985–1001, 2022.

[177] C. Mlynarczyk and R. Fahraeus. Endoplasmic reticulum stress sensitizes cells to dna damage-induced apoptosis through p53-dependent suppression of p21(cdkn1a). Nat Commun, 5:5067, 2014.

[178] C. Morrison. Constrained peptides' time to shine? Nat Rev Drug Discov, 17(8):531–533, 2018.

[179] G.W. Muse, D.A. Gilchrist, S. Nechaev, R. Shah, J.S. Parker, S.F. Grissom, J. Zeitlinger, and K. Adelman. Rna polymerase is poised for activation across the genome. Nat Genet., 39(12):1507–1511, 2007.

[180] M. Muttenthaler, G. F. King, D. J. Adams, and P. F. Alewood. Trends in peptide drug discovery. Nat Rev Drug Discov, 20(4):309–325, 2021.

[181] A.M. Naar, P. A. Beaurang, S. Zhou, S. Abraham, W. Solomon, and R. Tjian. Composite co-activator arc mediates chromatin-directed transcriptional activation. Nature, 398:828–832, 1999.

[182] Alice Neal, Svanhild Nornes, Pakavarin Louphrasitthiphol, Natalia Sacilotto, Mark D Preston, Lucija Fleisinger, Sophie Payne, and Sarah De Val. Ets factors are required but not sufficient for specific patterns of enhancer activity in different endothelial subtypes. Developmental Biology, 473:1–14, 2021.

[183] J. L. Nishikawa, A. Boeszoermenyi, L. A. Vale-Silva, R. Torelli, B. Posteraro, Y. J. Sohn, F. Ji, V. Gelev, D. Sanglard, M. Sanguinetti, R. I. Sadreyev, G. Mukherjee, J. Bhyravabhotla, S. J. Buhrlage, N. S. Gray, G. Wagner, A. M. Naar, and H. Arthanari. Inhibiting fungal multidrug resistance by disrupting an activator-mediator interaction. Nature, 530(7591):485–9, 2016.

[184] Einari A Niskanen, Marjo Malinen, Päivi Sutinen, Sari Toropainen, Ville Paakinaho, Anniina Vihervaara, Jenny Joutsen, Minna U Kaikkonen, Lea Sistonen, and Jorma J Palvimo. Global SUMOylation on active chromatin is an acute heat stress response restricting transcription. Genome Biol, 16:153, Jul 2015.

[185] Anders M. Näär, Pierre A. Beaurang, Karen M. Robinson, Jon D. Oliner, Daina Avizonis, Sigrid Scheek, Jörk Zwicker, James T. Kadonaga, and Robert Tjian. Chromatin, tafs, and a novel multiprotein coactivator are required for synergistic activation by sp1 and srebp-1a in vitro. Genes & Development, 12(19):3020–3031, Oct 1998. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press publisher: Cold Spring Harbor Lab.

[186] A.L. Okorokov, M.B. Sherman, C. Plisson, V. Grinkevich, K. Sigmundsson, G. Selivanova, J. Milner, and E.V. Orlova. The structure of p53 tumor suppressor protein reveals the basis for its functional plasticity. EMBO J, 25:5191–5200, 2006.

[187] N. A. Pchelintsev, P. D. Adams, and D. M. Nelson. Critical parameters for efficient sonication and improved chromatin immunoprecipitation of high molecular weight proteins. PLoS One, 11(1):e0148023, 2016.

[188] M. Pehar, K. J. O'Riordan, M. Burns-Cusato, M. E. Andrzejewski, C. G. del Alcazar, C. Burger, H. Scrable, and L. Puglielli. Altered longevity-assurance activity of p53:p44 in the mouse causes memory loss, neurodegeneration and premature death. Aging Cell, 9(2):174–90, 2010.

[189] Mariana Pehar, Mi Hee Ko, Mi Li, Heidi Scrable, and Luigi Puglielli. P44, the 'longevity-assurance'isoform of p53, regulates tau phosphorylation and is activated in an age-dependent fashion. Aging Cell, 13(3):449–456, 2014.

[190] B. H. Phang, R. Othman, G. Bougeard, R. H. Chia, T. Frebourg, C. L. Tang, P. Y. Cheah, and K. Sabapathy. Amino-terminal p53 mutations lead to expression of apoptosis proficient p47 and prognosticate better survival, but predispose to tumorigenesis. Proc Natl Acad Sci U S A, 112(46):E6349–58, 2015.

[191] Hemali P Phatnani and Arno L Greenleaf. Phosphorylation and functions of the rna polymerase ii ctd. Genes & development, 20(21):2922–2936, 2006.

[192] S. Polager and D. Ginsberg. p53 and e2f: partners in life and death. Nat Rev Cancer, 9(10):738–48, 2009.

[193] L. Polit, G. Kerdivel, S. Gregoricchio, M. Esposito, C. Guillouf, and V. Boeva. Chipin: Chip-seq inter-sample normalization based on signal invariance across transcriptionally constant genes. BMC Bioinformatics, 22(1):407, 2021.

[194] Z. C. Poss, C. C. Ebmeier, and D. J. Taatjes. The mediator complex and transcription regulation. Crit Rev Biochem Mol Biol, 48(6):575–608, 2013.

[195] Janusz Puc, Piotr Kozbial, Wenbo Li, Yuliang Tan, Zhijie Liu, Tom Suter, Kenneth A. Ohgi, Jie Zhang, Aneel K. Aggarwal, and Michael G. Rosenfeld. Ligand-dependent enhancer activation regulated by topoisomerase-I activity. Cell, 160(3):367 – 380, 2015.

[196] K. Quach, J. LaRochelle, X. H. Li, E. Rhoades, and A. Schepartz. Unique arginine array improves cytosolic localization of hydrocarbon-stapled peptides. Bioorg Med Chem, 26(6):1197–1202, 2018.

[197] M. Quevedo, L. Meert, M. R. Dekker, D. H. W. Dekkers, J. H. Brandsma, D. L. C. van den Berg, Z. Ozgur, IJcken W. F. J. van, J. Demmers, M. Fornerod, and R. A. Poot. Mediator complex interaction partners organize the transcriptional network that defines neural stem cells. Nat Commun, 10(1):2669, 2019.

[198] A. R. Quinlan and I. M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6):841–2, 2010.

[199] Whitney Rabacal, Sudheer K Pabbisetty, Kristen L Hoek, Delphine Cendron, Yin Guo, Damian Maseda, and Eric Sebzda. Transcription factor klf2 regulates homeostatic nk cell proliferation and survival. Proceedings of the National Academy of Sciences, 113(19):5370–5375, 2016.

[200] Fidel Ramírez, Vivek Bhardwaj, Laura Arrigoni, Kin Chung Lam, Björn A. Grüning, José Villaveces, Bianca Habermann, Asifa Akhtar, and Thomas Manke. High-resolution tads reveal dna sequences underlying genome organization in flies. Nature Communications, 9(11):189, Jan 2018.

[201] K.A. Resing, K. Meyer-Arendt, A.M. Mendoza, L.D. Aveline-Wolf, J.R. Jonscher, K.G. Pierce, W.M. Old, H.T. Cheung, S. Russell, J.L. Wattawa, G.R. Goehle, R.D. Knight, and N.G. Ahn. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. Anal. Chem, 76:3556–3568, 2004.

[202] William F Richter, Shraddha Nayak, Janet Iwasa, and Dylan J Taatjes. The mediator complex as a master regulator of transcription by rna polymerase ii. Nature Reviews Molecular Cell Biology, 23(11):732–749, 2022.

[203] J. K. Rimel, Z. C. Poss, B. Erickson, K.B. Hamman, S. Hu, C.C. Ebmeier, J.L. Johnson, J.J. Marineau, J.P. Carulli, M. Geyer, P. White, M. Breault, L. Tao, P. DeRoy, C. Clavet, S. Nayak, J. Iwasa, D.L. Bentley, R.D. Dowell, W. M. Old, and D. J. Taatjes. Selective inhibition of cdk7 reveals high-confidence targets and new models for tfiih function in transcription. Genes Dev, 34:1452–1473, 2020.

[204] J.K. Rimel and D. J. Taatjes. The essential and multi-functional tfiih complex. Protein Sci., 27:1018–1037, 2018.

[205] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. Nucleic Acids Res, 43(7):e47, 2015.

[206] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. bioinformatics, 26(1):139–140, 2010.

[207] J. D. Rubin, J. T. Stanley, R. F. Sigauke, C. B. Levandowski, Z. L. Maas, J. Westfall, D. J. Taatjes, and R. D. Dowell. Transcription factor enrichment analysis (tfea) quantifies the activity of multiple transcription factors from a single experiment. Commun Biol, 4(1):661, 2021.

[208] K. M. Ryan, A. C. Phillips, and K. H. Vousden. Regulation and function of the p53 tumor suppressor protein. Current Opinion in Cell Biology, 13(3):332–7, 2001.

[209] Simone E Salghetti, So Young Kim, and William P Tansey. Destruction of myc by ubiquitin-mediated proteolysis: cancer-associated and transforming mutations stabilize myc. The EMBO journal, 18(3):717–726, 1999.

[210] A. L. Sanborn, B. T. Yeh, J. T. Feigerle, C. V. Hao, R. J. Townshend, E. Lieberman Aiden, R. O. Dror, and R. D. Kornberg. Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to mediator. Elife, 10, 2021.

[211] Ioannis Sanidas, Hanjun Lee, Purva H Rumde, Gaylor Boulay, Robert Morris, Gabriel Golczer, Marcelo Stanzione, Soroush Hajizadeh, Jun Zhong, Meagan B Ryan, et al. Chromatin-bound rb targets promoters, enhancers, and ctcf-bound loci and is redistributed by cell-cycle progression. Molecular cell, 82(18):3333–3349, 2022.

[212] Sarah K Sasse, Amber Dahlin, Lynn Sanford, Margaret A Gruca, Arnav Gupta, Ann Chen Wu, Carlos Ibarren, Robin D Dowell, Scott T Weiss, and Anthony N Gerber. Glucocorticoid-regulated bidirectional enhancer rna transcription pinpoints functional genetic variants linked to asthma. medRxiv, pages 2022–11, 2022.

[213] R. A. Saxton and D. M. Sabatini. mtor signaling in growth, metabolism, and disease. Cell, 168(6):960–976, 2017.

[214] Miriam Merzel Schachter and Robert P Fisher. The cdk-activating kinase cdk7: taking yes for an answer. Cell Cycle, 12(20):3239–3240, 2013.

[215] Björn Schwalb, Margaux Michel, Benedikt Zacher, Katja Frühauf, Carina Demel, Achim Tresch, Julien Gagneur, and Patrick Cramer. Tt-seq maps the human transient transcriptome. Science, 352(6290):1225–1228, 2016.

[216] Y. Shi, E. Felley-Bosco, T. M. Marti, K. Orlowski, M. Pruschy, and R. A. Stahel. Starvation-induced activation of atm/chk2/p53 signaling sensitizes cancer cells to cisplatin. BMC Cancer, 12:571, 2012.

[217] A. Shoffner, V. Cigliola, N. Lee, J. Ou, and K. D. Poss. Tp53 suppression promotes cardiomyocyte proliferation during zebrafish heart regeneration. Cell Rep, 32(9):108089, 2020.

[218] Rutendo F Sigauke, Lynn Sanford, Taylor Jones, Zachary L Maas, Mary A Allen, and Robin D Dowell. Atlas of nascent rna transcripts reveals high confidence enhancer associated bidirectionals linked with genes across different tissue types. In progress, 2023.

[219] A. A. Sigova, A. C. Mullen, B. Molinie, S. Gupta, D. A. Orlando, M. G. Guenther, A. E. Almada, C. Lin, P. A. Sharp, C. C. Giallourakis, and R. A. Young. Divergent transcription of long noncoding rna/mrna gene pairs in embryonic stem cells. Proc Natl Acad Sci U S A, 110(8):2876–81, 2013.

[220] Robert J. Sims, Rimma Belotserkovskaya, and Danny Reinberg. Elongation by rna polymerase ii: the short and long of it. Genes & Development, 18(20):2437–2468, Oct 2004.

[221] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 102(43):15545–50, 2005.

[222] Ayako Suzuki, Hideki Makinoshima, Hiroyuki Wakaguri, Hiroyasu Esumi, Sumio Sugano, Takashi Kohno, Katsuya Tsuchihara, and Yutaka Suzuki. Aberrant transcriptional regulations in cancers: genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines. Nucleic Acids Research, 42(22):13557–13572, Dec 2014.

[223] S.J. Tapscott, R.L. Davis, M.J. Thayer, P.F. Cheng, H. Weintraub, and A.B. Lassar. Myod1: a nuclear phosphoprotein requiring a myc homology region to convert fibroblasts to myoblasts. Science, 242:405–411, 1988.

[224] H. Tidow, R. Melero, E. Mylonas, S. M. Freund, J. G. Grossmann, J. M. Carazo, D. I. Svergun, M. Valle, and A. R. Fersht. Quaternary structures of tumor suppressor p53 and a specific p53 dna complex. Proc Natl Acad Sci U S A, 104(30):12324–9, 2007.

[225] J. P. Tourigny, K. Schumacher, M. M. Saleh, D. Devys, and G. E. Zentner. Architectural mediator subunits are differentially essential for global transcription in saccharomyces cerevisiae. Genetics, 217(3), 2021.

[226] S.D. Tyner, S. Venkatachalam, J. Choi, S. Jones, N. Ghebranious, H. Igelmann, X. Lu, G. Soron, B. Cooper, C. Brayton, S.H. Park, T. Thompson, G. Karsenty, A. Bradley, and L.A. Donehower. p53 mutant mice that display early ageing-associated phenotypes. Nature, 415:45–53, 2002.

[227] E. Ungewitter and H. Scrable. Delta40p53 controls the switch from pluripotency to differentiation by regulating igf signaling in escs. Genes Dev, 24(21):2408–19, 2010.

[228] J. L. Van Nostrand, C. A. Brady, H. Jung, D. R. Fuentes, M. M. Kozak, T. M. Johnson, C. Y. Lin, C. J. Lin, D. L. Swiderski, H. Vogel, J. A. Bernstein, T. Attie-Bitach, C. P. Chang, J. Wysocka, D. M. Martin, and L. D. Attardi. Inappropriate p53 activation during development induces features of charge syndrome. Nature, 514(7521):228–32, 2014.

[229] L.T. Vassilev, B.T. Vu, B. Graves, D. Carvajal, F. Podlaski, Z. Filipovic, N. Kong, U. Kammlott, C. Lukacs, C. Klein, N. Fotouhi, and E.A. Liu. In vivo activation of the p53 pathway by small-molecule antagonists of mdm2. Science, 303(5659):844–848, 2004.

[230] A. Verfaillie, D. Svetlichnyy, H. Imrichova, K. Davie, M. Fiers, Z. Kalender Atak, G. Hulselmans, V. Christiaens, and S. Aerts. Multiplex enhancer-reporter assays uncover unsophisticated tp53 enhancer logic. Genome Res, 26(7):882–95, 2016.

[231] A. Vihervaara, D. B. Mahat, M. J. Guertin, T. Chu, C. G. Danko, J. T. Lis, and L. Sistonen. Transcriptional response to stress is pre-wired by promoter and enhancer architecture. Nat Commun, 8(1):255, 2017.

[232] L. Viladevall, C.V. St. Amour, A. Rosebrock, S. Schneider, C. Zhang, J.J. Allen, K.M. Shokat, B. Schwer, J.K. Leatherwood, and R.P. Fisher. Tfiih and p-tefb coordinate transcription with capping enzyme recruitment at specific genes in fission yeast. Mol Cell, 33:738–751, 2009.

[233] E. Vojnic, A. Mourao, M. Seizl, B. Simon, L. Wenzeck, L. Lariviere, S. Baumli, K. Baumgart, M. Meisterernst, M. Sattler, and P. Cramer. Structure and vp16 binding of the mediator med25 activator interaction domain. Nat Struct Mol Biol, 18:404–409, 2011.

[234] B. Vollmar, A. M. El-Gibaly, C. Scheuer, M. W. Strik, H. P. Bruch, and M. D. Menger. Acceleration of cutaneous wound healing by transient p53 inhibition. Lab Invest, 82(8):1063–71, 2002.

[235] T. Wada, T. Takagi, Y. Yamaguchi, A. Ferdous, T. Imai, S. Hirose, S. Sugimoto, K. Yano, G. A. Hartzog, F. Winston, S. Buratowski, and H. Handa. Dsif, a novel transcription elongation factor that regulates rna polymerase ii processivity, is composed of human spt4 and spt5 homologs. Genes Dev, 12(3):343–56, 1998.

[236] Mark Wade, Ee Tsin Wong, Mengjia Tang, Jayne M Stommel, and Geoffrey M Wahl. Hdmx modulates the outcome of p53 activation in human tumor cells. Journal of Biological Chemistry, 281(44):33036–33044, 2006.

[237] T. Waldman, C. Lengauer, K. W. Kinzler, and B. Vogelstein. Uncoupling of s phase and mitosis induced by anticancer agents in cells lacking p21. Nature, 381(6584):713–6, 1996.

[238] Allen Wang, Feng Yue, Yan Li, Ruiyu Xie, Thomas Harper, Nisha A Patel, Kayla Muth, Jeffrey Palmer, Yunjiang Qiu, Jinzhao Wang, Dieter K Lam, Jeffrey C Raum, Doris A Stoffers, Bing Ren, and Maike Sander. Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. Cell Stem Cell, 16(4):386–99, Apr 2015.

[239] Isabel X. Wang, Leighton J. Core, Hojoong Kwak, Lauren Brady, Alan Bruzel, Lee McDaniel, Allison L. Richards, Ming Wu, Christopher Grunseich, John T. Lis, and Vivian G. Cheung. RNA-DNA differences are generated in human cells within seconds after RNA exits polymerase II. Cell Reports, 6(5):906 – 915, 2014.

[240] Zhong Wang, Tinyi Chu, Lauren A. Choate, and Charles G. Danko. Identification of regulatory elements from nascent transcription using dreg. Genome Research, 29(2):293–303, Feb 2019.

[241] M. R. Webster, M. E. Fane, G. M. Alicea, S. Basu, A. V. Kossenkov, G. E. Marino, S. M. Douglass, A. Kaur, B. L. Ecker, K. Gnanapradeepan, A. Ndoye, C. Kugel, A. Valiga, J. Palmer, Q. Liu, X. Xu, J. Morris, X. Yin, H. Wu, W. Xu, C. Zheng, G. C. Karakousis, R. K. Amaravadi, T. C. Mitchell, F. V. Almeida, M. Xiao, V. W. Rebecca, Y. J. Wang, L. M. Schuchter,

M. Herlyn, M. E. Murphy, and A. T. Weeraratna. Paradoxical role for wild-type p53 in driving therapy resistance in melanoma. Mol Cell, 77(3):633–644 e5, 2020.

[242] S. Weingarten-Gabbay, D. Khan, N. Liberman, Y. Yoffe, S. Bialik, S. Das, M. Oren, and A. Kimchi. The translation initiation factor dap5 promotes ires-driven translation of p53 mrna. Oncogene, 33(5):611–8, 2014.

[243] W. A. Whyte, D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, and R. A. Young. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell, 153(2):307–19, 2013.

[244] Brian T Wilhelm, Samuel Marguerat, Stephen Watt, Falk Schubert, Valerie Wood, Ian Goodhead, Christopher J Penkett, Jane Rogers, and Jürg Bähler. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature, 453(7199):1239–1243, 2008.

[245] Erin M Wissink, Anniina Vihervaara, Nathaniel D Tippens, and John T Lis. Nascent rna analyses: tracking transcription and its regulation. Nature Reviews Genetics, 20(12):705–723, 2019.

[246] David Wolf, Nicholas Harris, Naomi Goldfinger, and Varda Rotter. Isolation of a full-length mouse cdna clone coding for an immunologically distinct p53 molecule. Molecular and cellular biology, 5(1):127–132, 1985.

[247] Shwu-Yuan Wu, Tianyuan Zhou, and Cheng-Ming Chiang. Human mediator enhances activator-facilitated recruitment of rna polymerase ii and promoter recognition by tata-binding protein (tbp) independently of tbp-associated factors. Molecular and cellular biology, 23(17):6229–6242, 2003.

[248] M. D. Wyatt and 3rd Wilson, D. M. Participation of dna repair in the response to 5-fluorouracil. Cell Mol Life Sci, 66(5):788–99, 2009.

[249] Yue Xiong, Gregory J Hannon, Hui Zhang, David Casso, Ryuji Kobayashi, and David Beach. p21 is a universal inhibitor of cyclin kinases. Nature, 366(6456):701–704, 1993.

[250] Y. Yamaguchi, T. Takagi, T. Wada, K. Yano, A. Furuya, S. Sugimoto, J. Hasegawa, and H. Handa. Nelf, a multisubunit complex containing rd, cooperates with dsif to repress rna polymerase ii elongation. Cell, 97(1):41–51, 1999.

[251] C. M. Yang, M. K. Kang, W. J. Jung, J. S. Joo, Y. J. Kim, Y. Choi, and H. P. Kim. p53 expression confers sensitivity to 5-fluorouracil via distinct chromatin accessibility dynamics in human colorectal cancer. Oncol Lett, 21(3):226, 2021.

[252] F. Yang, B.W. Vought, J.S. Satterlee, A.K. Walker, Z. Jim Sun, J.L. Watts, R. DeBeaumont, R. Mako Saito, S.G. Hyberts, S. Yang, C. Macol, L. Iyer, R. Tjian, S. van den Heuvel, A.C. Hart, G. Wagner, and A. M. Naar. An arc/mediator subunit required for srebp control of cholesterol and lipid homeostasis. Nature, 442:700–704, 2006.

[253] Yili Yin, Charles W. Stephen, M. Gloria Luciani, and Robin Fåhraeus. p53 stability and activity is regulated by mdm2-mediated induction of alternative p53 translation products. Nature Cell Biology, 4(66):462–467, Jun 2002.

[254] Xinyang Yu and Michael J Buck. Defining tp53 pioneering capabilities with competitive nucleosome binding assays. Genome research, 29(1):107–115, 2019.

[255] Jing-Ping Zhang, Hua Zhang, Hong-Bo Wang, Yan-Xian Li, Gui-Hong Liu, Shan Xing, Man-Zhi Li, and Mu-Sheng Zeng. Down-regulation of sp1 suppresses cell proliferation, clonogenicity and the expressions of stem cell markers in nasopharyngeal carcinoma. Journal of translational medicine, 12(1):1–12, 2014.

[256] H. Zhao, N. Young, J. Kalchschmidt, J. Lieberman, L. El Khattabi, R. Casellas, and F. J. Asturias. Structure of mammalian mediator complex reveals tail module architecture and interaction with a conserved core. Nat Commun, 12(1):1355, 2021.

[257] Yongbing Zhao, Supriya V. Vartak, Andrea Conte, Xiang Wang, David A. Garcia, Evan Stevens, Seol Kyoung Jung, Kyong-Rim Kieffer-Kwon, Laura Vian, Timothy Stodola, Francisco Moris, Laura Chopp, Silvia Preite, Pamela L. Schwartzberg, Joseph M. Kulinski, Ana Olivera, Christelle Harly, Avinash Bhandoola, Elisabeth F. Heuston, David M. Bodine, Raul Urrutia, Arpita Upadhyaya, Matthew T. Weirauch, Gordon Hager, and Rafael Casellas. "stripe" transcription factors provide accessibility to co-binding partners in mammalian genomes. Molecular Cell, 82(18):3398–3411.e11, Sep 2022.

[258] Rongbin Zheng, Changxin Wan, Shenglin Mei, Qian Qin, Qiu Wu, Hanfei Sun, Chen-Hao Chen, Myles Brown, Xiaoyan Zhang, Clifford A Meyer, et al. Cistrome data browser: expanded datasets and new tools for gene regulatory analysis. Nucleic acids research, 47(D1):D729–D735, 2019.

Appendix A

The TFIIH kinase CDK7 governs RNA polymerase II function, RNA processing and cellular proliferation networks in the nucleus

The work in this section is currently ongoing. Contributing authors from the Taatjes Lab include Taylor Jones, Jenna Rimel, Olivia Luyties, Jessica Rodino and Grace Shelby.

Additional contributing authors include Jason Marineau, Junjie Feng, Koh, Kotecha, Carulli, Cooper, Das, Hamman, Hu, Madduri, Rajagopal. Principal investigators include Dylan Taatjes, Robin Dowell, and Basil Gerber. The focus of this section will be in data produced from the Taatjes lab, with a focus on NGS data. All raw data in this section can be found in GEO at GSE228346. This data is private until publication or March 2026. Methods are discussed in Appendix B.4.

A.1 Contribution Statement

My role in this project is centered around NGS data production and analysis. There are two main types of sequencing data in this project, RNA-seq and PRO-seq. The RNA-seq data was produced by Dr. Jenna Rimel and analyzed by me. The PRO-seq data was both produced and analyzed by me.

In terms of RNA-seq data analysis I performed mapping, trimming and quality control. I also have performed DESeq2 and GSEA analyses. In the future I plan to also perform splicing and downstream of gene analyses (DoG). Additional RNA-seq data sets will also be processed that were produced by Syros. For PRO-seq: performed mapping, trimming, qc, DESeq2 and GSEA analyses. I ran numerous analyses in regards to pause-indices and elongation rates. More complex analyses

were performed in terms of TF enrichment. For this I ran bidirectional calling and analysis[18], complex TFEA analyses[207] as well as comparative analyses between TFEA and the OV90 TF profile (discussed in Chapter 2). Additional analyses will be performed for a deeper dive into TF function, exploring the activation of inducible genes in the context of CDK7 inhibition, and rigorously comparing RNA-seq and PRO-seq data.

A.2 Abstract

CDK7 is an important regulatory kinase with defined roles as a transcriptional regulator in the nucleus and as the master cell cycle regulator in the cytoplasm. Here we use the potent and selective CDK7 inhibitor, SY5609, to characterize both the transcriptional response and impact of RNA maturation induced by CDK7 inhibition. We find that CDK7 inhibition does not prevent inducible transcriptional factor activity, but instead dramatically reduces the function of basally active transcription factors involved in cellular maintenance and proliferation. These "basally active" transcription factors tend to be promoter associated imply a promoter specific role for CDK7 activity. Additionally, CDK7 inhibition reduces steady-state RNA production for genes affiliated with cellular proliferation; for example, genes required for mitosis. Collectively, these results reveal a novel mechanism by which CDK7 coordinates its nuclear and cytoplasmic functions to control cellular proliferation. In the nucleus, CDK7 kinase activity maintains the activity of basal TFs that drive proliferative gene expression programs; in the cytoplasm, the CDK7 kinase activates other CDKs that drive the cell cycle. The novel nuclear function for CDK7 identified here establishes a link between its nuclear and cytoplasmic functions, which ensures that gene expression programs in the nucleus match the cell cycle requirements, driven in part by CDK7-regulated CDKs in the cytoplasm. Collectively, these results further establish CDK7 as a master regulator of cell homeostasis and proliferation.

A.3 Introduction

Cyclin-dependent kinase 7 (CDK7) plays an important role in both cell cycle regulation and transcription. CDK7 operates in a 3-subunit complex containing cyclin H and Mat1 commonly called the cyclin-activating kinase (CAK). The CAK has well defined cytosolic roles as the "master cell cycle regulator"[77]. The CAK catalyzes the phosphorylation of the "activation loop" (a.k.a. T-loop) of other cell cycle kinases, subsequently activating their kinase abilities[214]. In the nucleus, the CAK associates with the general transcription factor (GTF) TFIIH. TFIIH is a member of the PolII pre-initation complex, thus is poised at the promoter of actively transcribed genes. It's known that the CAK is responsible for the phosphorylation of Ser5 and Ser7 of the PolII C-terminal domain (CTD)[3, 88, 125], which is associated with the transition from initiating to elongating PolII[204]. The PolII CTD provides a docking site for transcriptional co-factors and RNA processing factors. The phosphorylation of the PolII CTD coordinates the timing of co-factor recruitment[67, 191]. Additionally, the CAK has been shown to phosphorylate factors such as NELF, DSIF and P-TEFb, which regulate pol II pausing and elongation, splicing factors such as U2AF and SF3B1 and RNA processing facors such as DDX21[203]. In all, inhibition of CDK7 leads to major defects in $5'$ capping[232], successful elongation, splicing and termination [63, 88, 204, 232].

Due to CDK7's role in both cell cycle control and transcription, numerous cancers depend on elevated CDK7 activity to drive oncogenesis[144]. Thus, inhibitors of CDK7 have been studied as potential anticancer therapeutics. Inhibition of CDK7 has anti-proliferative effects across a range of cancer cell lines as well as in mouse models[136]. Over the years, CDK7 inhibitors have vastly improved in terms of specificity and potency, with several inhibitors currently in clinical trials as cancer therapeutics[48].

The development of potent and specific CDK7 inhibitors has enabled us to probe the direct transcriptional impact of CDK7 inhibition (CDK7i). In a single nascent sequencing assay (PRO-seq) assay, we are able to address the direct transcriptional responses to CDK7i. Upon CDK7i we see influences on every stage of the transcription cycle, as well as altered sequence-specific TF activity

within promoter regions. Contrasting this data with steady-state RNA-sequencing (RNA-seq), we helped determine the overall cellular impact of CDK7i. Our results show widespread effects of CDK7 inhibition, including defects in RNA processing and unanticipated repression of promoter-associated TFs. Together, our results reveal novel mechanisms by which CDK7 coordinates cytoplasmic and nuclear functions to control of cell proliferation.

A.4 Results

In this section I will cover NGS data from two main assays, the first is PRO-seq which is described in section 1.1.3. For the PRO-seq experiments, cells were pre-treated for 30min with either DMSO (control, 0.005%) or 50nM SY5609, a potent and specific CDK7 inhibitor[163]. The cells were then split into three groups, 1) the nuclei were immediately harvested, 2) the cells were heat shocked (42°C) for 20min then the nuclei were harvested or 3) the cells were heat shocked for 45min then the nuclei were harvested. All PRO-seq samples were produced in biological replicate and sequenced to approximately 100 million reads per replicate. A graphic of the PRO-seq treatment scheme is shown in Figure A.1A.

The next assay is RNA-seq, which measures steady-state RNA. In RNA-seq, total RNA is isolated from cells. Then there is an enrichment step selecting for the poly-adenylation of mature mRNA. Finally, there is a library preparation step and sequencing. Here, all samples were pre-treated for 30min with either DMSO (control) or a CDK7 inhibitor, which was SY5609 unless otherwise stated. One set of cells were immediately harvested after the 30min pre-treatment with SY5609 or control. The next set was heat shocked (42°C) for 30min followed by a 60min recovery. In this set the cells were treated with one of three possible CDK7 inhibitors, 1) SY5609, 2) SY5102 or 3) 3MB-PP1 in a CDK7 analog sensitive OV90 cell line. The last set of RNA-seq data was continuously heat shocked for 120min after the pre-treatment with SY5609 or control. All RNA-seq samples were produced in biological triplicate and sequenced to approximately 220 million reads per replicate. A graphic of the RNA-seq treatment scheme is shown in Figure A.1B.

A.4.1 CDK7 influences all stages of PolII transcription

CDK7 has known roles in the transcription cycle related to 5′ capping[232], the transition between initiation and elongation[88], co-transcriptional splicing and PolII termination[88, 203]. From PRO-seq, an example trace of HSP90 (Fig. A.2) exemplifies the impact of CDK7i on the transcription cycle. In the absence of heat shock, HSP90 is not transcribed. However, after a 45min heat shock there is a shift in the promoter proximal peak, where it is greater upon CDK7i. This implies a defect in the transition from promoter-proximal pausing into successful elongation, which occurs at gene 5′ ends. Additionally, with a 45min heat shock there is clearly additional transcriptional run-on in the CDK7i condition, suggesting a termination defect at gene 3′ ends. Splicing information is not captured by PRO-seq, but is by RNA-seq. Preliminary RNA-seq splicing analyses (not shown here) imply that CDK7i induces widespread splicing defects consistent with work published by Rimel et. al.[203].

We observe that CDK7i induces a general $\approx 20\%$ reduction in overall transcriptional output across gene bodies based upon PRO-seq data (Fig. A.3A). While there were few significantly changing genes in the condition SY5609 versus DMSO (no heat shock), there was a general, global reduction of gene body transcription. When total normalized gene counts for DMSO (x-axis) and SY5609 (y-axis) were plotted, we observed that the greater gene mass fell below the 1:1 line (slope of best fit 0.9, Fig. A.3B). Since this was a global trend, proper normalization is essential. Many differential expression algorithms normalize to the inner-quartile of unchanging genes[155, 206]. If all genes are subtly changing, the internal normalization of these programs will wash out this signal. For this reason, we systematically normalized the data to the 3′ end of transcribing genes. We selected genes that are long enough that, assuming a polymerase rate of 3kb/min, the 3′ end couldn't be reached within our maximum treatment time (75min, 225kb). The precise normalization methodology is described in Maas and Dowell 2023[158], and normalization factors are shown in Figure A.3C.

As previously stated, CDK7 plays an important role in the transition between transcription

initiation and elongation[88]. Approximately 60 base pairs downstream of initiation PolII pauses before productive elongation[2]. In PRO-seq data, the ratio of reads within the promoter proximal peak to the reads within the elongation region is commonly used to quantify the changes in pausing between conditions. This metric is called the pause-index (PI)[2, 53, 71, 159]. Genes that are down-regulated by heat shock have increased PIs (relatively more reads in the promoter proximal peak:gene body) compared to non-heat shocked samples in both control and CDK7i conditions (Fig. A.4A). This implies relatively less productive elongation for this subset of heat shock genes. Genes that are up-regulated by heat shock have decreased PIs (relatively fewer reads in the promoter proximal peak:gene body) compared to non-heat shocked samples in both control and CDK7i conditions (Fig. A.4B). This implies increased productive elongation for this set of genes. These results are consistent with those published in Mahat et. al.[159]. When assessing how CDK7i impacts PI compared to control conditions in the context of heat shock, we see a general increase in PI at all genes significantly impacted by heat shock (Fig. A.5). This illustrates the importance of CDK7 in the transition into productive elongation, regardless of up-regulation or down-regulation of the gene target.

Next, we asked what specific gene expression programs were altered due to 1) heat shock, 2) heat shock in the context of CDK7i and 3) CDK7i alone. In both heat shock and heat shock + CDK7i conditions, the predominant response was similar (Fig. A.6A-B, A.7A-B). Surprisingly, heat shocked cells pre-treated with CDK7i were able to mount a robust heat shock response equivalent to control heat shock cells. This implies that the inhibition of CDK7 has negligible effects on inducible PolII transcription. When we contrast CDK7i versus control (at 0min, 20min and 45min heat shock) we find a relatively mild change in gene transcription programs (Fig. A.8A-B). One observation is that almost all genes up-regulated by CDK7i (versus DMSO) are extreme run-on from up-stream genes in the CDK7i condition. Heat shock also increases transcriptional run-on[45]. In SY5609 versus DMSO with 45min heat shock we see compounding effects of transcriptional run-on both heat shock and CDK7i. In fact, the most significantly up-regulated gene in SY5609 versus DMSO 45min heat shock is DYNC1H1 (p-value=$8.53e^{-7}$), the gene down-stream of HSP90 shown

in Figure A.2.

One gene that is truly significantly up-regulated by 45min heat shock in SY5609 versus DMSO is the RNA PolII subunit POLR2A (p-value=0.005, log2(Fold Change)=1.53). RNA PolII is trending up by 20min heat shock in SY5609 versus DMSO (p-value=0.02, log2(Fold Change)=1.29). One possible explanation is that to compensate for the generally decreased transcriptional output in CDK7i conditions (Fig. A.3A-B), cellular systems up-regulate RNA PolII to handle the stress response induced by heat shock.

A.4.2 CDK7 inhibition reduces mRNA associated with proliferative gene programs

We observed a robust heat shock response in gene transcription in both control and CDK7i conditions (Fig. A.6A-B, A.7A-B). This manifested at the steady-state RNA level as we observe a strong heat shock response in both control and CDK7i conditions. The heat shock response (20min and 45min heat shock versus no heat shock) in CDK7i conditions is functionally equivalent to the response in control conditions (Fig. A.9A-B, A.10A-B). This affirms that CDK7i has no impact on inducible transcriptional and RNA maturation programs.

Despite a robust heat shock response in both control and CDK7i conditions, when contrasting CDK7i versus control in the context of heat shock, we see alterations in steady-state RNA. At 0min heat shock there was little significant impact on steady-state RNA (n=6 significant genes, all down-regulated). This is likely because a 30min treatment time is too short for steady-state changes to manifest. Even so, by 30min the transcript encoding for the transcription factor KLF2 is significantly reduced almost two fold in CDK7i conditions (p-value=$2.64e^{-5}$, log2(Fold Change)=0.63, Fig. A.11A). KLF2 is part of the Kruppel family of transcription factors (TFs) which bind the CACCC box within promoters. Generally, this family of TFs are associated with proliferative programs[199].

Both RNA-seq experiments contrasting SY5609 versus DMSO in the context of heat shock have far more significantly changing gene signatures (Fig. A.11A). This is likely partially due to the heat shock stimulus, and partially due to the increased experiment time permitting steady-state

RNA changes to manifest (120min and 150min experiment length before RNA harvest; see Fig. A.1 for details). In both heat shock conditions SY5609 versus DMSO causes mostly down-regulation of gene targets. These gene targets are not related to the heat shock response, but rather many are generally related to the reduction of cellular proliferation (Fig. A.11B).

In analyzing the RNA-seq data with alternative CDK7 inhibitors SY5102 and 3MB-PP1 versus DMSO we find that the CDK7 inhibition response in the context of heat shock is consistent to that of SY5609 (Fig. A.12A-B). This indicates that we have a reliable set of genes that are impacted by CDK7i, which are enriched for the negative regulation of cellular proliferation (Fig. A.11B). Taken together, despite the modest effect of CDK7i on gene transcription, there is a general reduction of steady-state RNA related to cellular proliferation.

A.4.3 CDK7 activates a common core TF network that drives proliferation across cell types

Since CDK7i has an anti-proliferative role at the level of steady-state RNA, we asked whether TFs associated with proliferation were also impacted by CDK7i. To address this we performed TF inference analyses. TF inference is discussed in section 1.2.3, but briefly TF inference measures the co-localization of TF motifs with PolII initiation sites genome-wide[17, 207]. These methods depend on reliable identification of PolII initiation regions, which are annotated based on their canonical bidirectional signal[18, 240]. A summary of the PolII initiation regions identified in this study are summarized in Figure A.13A-C.

We first ran transcription factor enrichment analysis (TFEA) assessing the heat shock response (20min/45min vs 0min) in control and CDK7i conditions. We found that the main responsive TFs were heat shock factors (HSF1, HSF2, SRF1, etc.; Fig. A.14, A.15). There was no major difference in heat shock factor activity in response to CDK7i compared to control, indicating that CDK7i is not strongly impacting stimulus-responding TFs. The heat shock factors are able to mobilize and alter heat shock gene targets in transcription (Fig. A.6A-B, A.7A-B) which manifest into steady-state RNA (Fig. A.9A-B, A.10A-B).

When we look at SY5609 versus DMSO (no heat shock; Fig. A.16A) we find a set of TFs that have a high frequency of motif hits, and tend to be subtly activated or repressed. Many of these significantly changing TFs are associated with proliferation, such as KLF and ETS factors. This led us to ask whether the TFs that are impacted by CDK7i are disproportionately enriched for factors that are active in baseline cellular conditions (ie no heat shock, no CDK7i). To test this we used another TF inference method called TF profiling. This method is discussed in depth in Chapter 2, but briefly we infer TFs that are activated in baseline conditions by comparing experimentally calculated motif displacement scores (observed, y-axis) to motif displacement scores calculated from a statistically relevant model (expectation, x-axis; Fig. A.16B). When we compare the TFs that are ON in baseline conditions, to TFs that are changing upon CDK7i, we find that, generally, all bassally active TFs have altered activity upon CDK7i. Generally, the TFs called as ON-UP in baseline conditions are repressed by CDK7i (Fig. A.16C). Next, we asked if the regions driving this effect were predominantly promoters (ie at the transcription start site of a gene) or at enhancers. We found that the TFs repressed by CDK7i were disproportionately promoter associated (Fig. A.16D).

This observation is retained through the heat shock time course when comparing CDK7i to control (Fig. A.17), but stronger enhancer driven signals begin to appear over time. This is possibly due to the mobilized heat shock response at enhancers being equal in control and CDK7i, but the secondary effects may manifest less robustly in CDK7i conditions. For example, TFs downstream of heat shock factors (ie secondary TFs) may not be as robustly transcribed (Fig. A.3A-B), leading them to mount a less effective secondary transcriptional response. One general concern with this methodology is that the highest impacted TFs by CDK7i tend to have very GC rich motifs which can lead to biases in TF inference methods due to the GC rich nature of PolII initiation sites (Fig. A.13B). This was accounted for by correcting for all TFEA results with a single linear fit, shown in Figure A.18 and discussed further in section B.4.

A.5 Discussion and future directions

In this work we show that CDK7 has a general regulatory role in the transition from initiation to elongation, and have preliminary evidence that it's important for PolII termination. The inhibition of CDK7 causes a general ≈20% decrease in global gene transcription. After sufficient time (30min treatment with SY5609, followed by a 45min heat shock), it seems that cellular programs begin to compensate for the general reduction of transcription by up-regulating RNA PolII itself.

We also show that CDK7 has negligible effects on inducible transcription (heat shock). In CDK7i conditions, cells are able to mobilize heat shock factors in response to heat shock. These factors still impact their gene targets, which go on to become viable, steady-state transcripts. This implies that inducible gene programs (such as heat shock) are regulated by distinct mechanisms and/or factors. This is consistent with work by others, indicating that the requirements at inducible gene promoters differ from those that are basally operational[101, 176]. Furthermore, stimulus responding TFs tend to be associated with intergenic regions, rather than promoters (work in progress, see Chapter 2). It could be that CDK7 plays a more promoter specific role with TFs, as well as in productive elongation in genic regions. One potential caveat is that TF inference is picking up large changes in promoters, and attributing the behavior to sequence specific TFs, when in reality it could be that CDK7 is simply modulating the behavior of RNA PolII, as indicated by the broad increase in promoter-proximal pausing[63, 67, 88, 191, 204, 232]. Regardless, this alludes to a promoter specific role for CDK7 in transcriptional regulation.

We additionally show that CDK7 has a pro-proliferative role in the nucleus through two aspects. The first is that, CDK7 is required to maintain the function of core TFs that drive proliferation in most cell lines. It is possible that CDK7 exerts control over these core TFs through the TF RB1. RB1 is a high confidence target for CDK7 inhibition[203]. Preliminary data in the lab also shows that RB1 is depleted within 30min of SY5609 treatment, and remains reduced for up to 4hrs by western blot (work by Grace Shelby). It has been recently shown that RB1 plays an important role coordinating core TFs to control the cell cycle and MAPK-responsive genes[211].

The second mechanism by which CDK7 appears to regulate cell proliferation is through regulation of genes that control cell proliferation. Proliferative genes are not dramatically impacted at the transcriptional level, leading to two possible explanations for the negative regulation of these gene programs. The first could be that the general reduction of gene transcription caused by CDK7i ($\approx$ 20%) reduces the accumulated steady-state RNA enough to be observable by our detection methods. Alternatively, CDK7i may induce detrimental splicing or termination defects (in the form of down-stream of gene expression- DoGs) disproportionately at proliferative genes. We plan to do in-depth splicing and termination analyses for this study. We also plan to do a long-read sequencing experiment (0min heat shock and 120min heat shock +/- SY5609) to validate results regarding our splicing and termination analyses. Regardless, the current results taken together reveal a mechanism by which CDK7 coordinates cytoplasmic and nuclear control of cell proliferation.

Future work, beyond the scope of this study, could explore the relationship between transcriptional run-on and DoG expression associated with steady-state RNA. We observe here that heat shock[45] and CDK7 inhibition both increase transcriptional run-on in a compounding effect. We observe (but have not quantified) DoGs in the RNA-seq data for both heat shock and for CDK7 inhibition. How and why certain genes have excessive transcriptional run-on is unknown. How that run-on correlates with DoGs, and whether this process is functionally relevant is also unknown. This could serve as a basis for an entirely independent project using the wealth of data generated for this study.

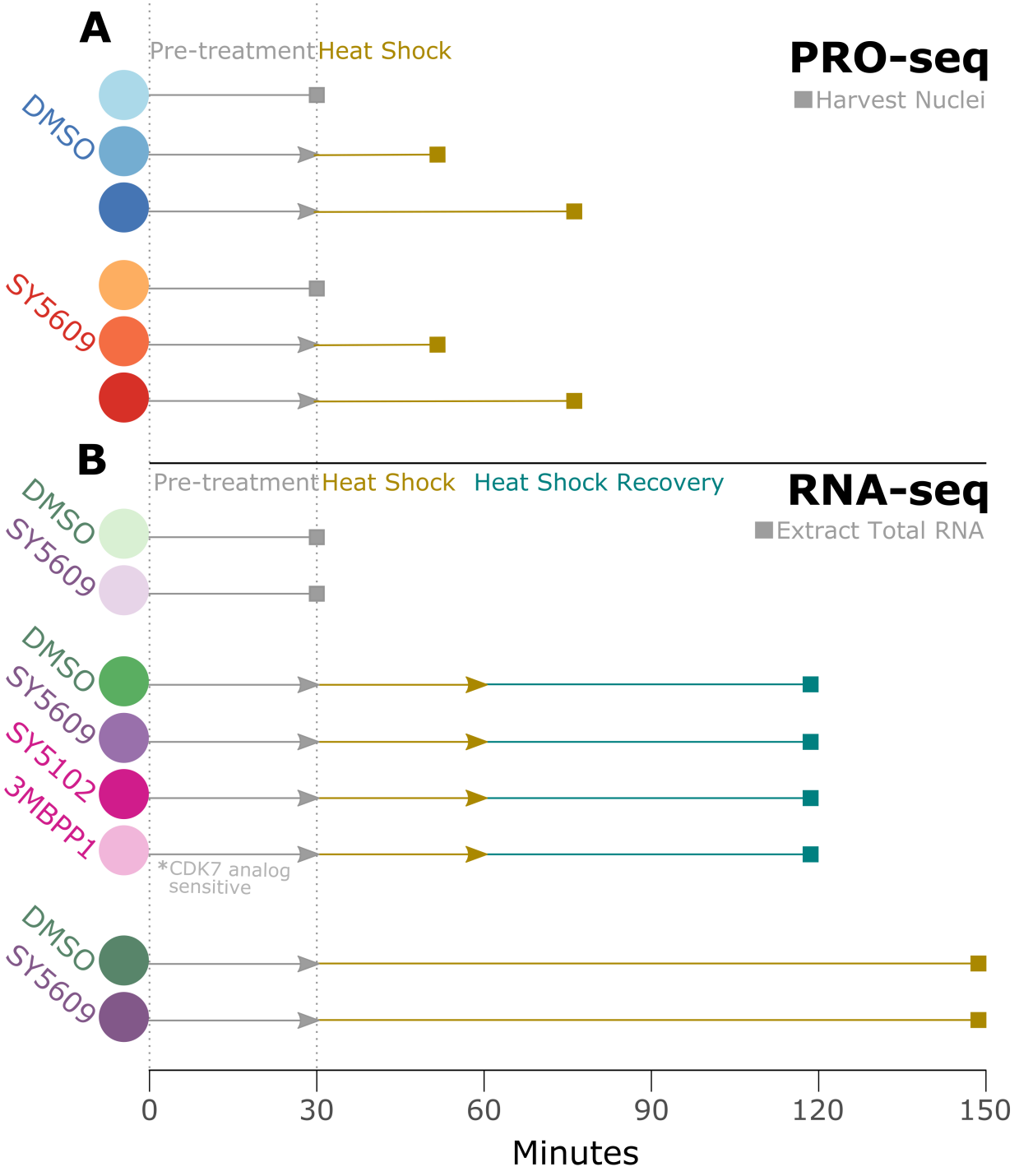


Figure A.1: **Experimental design for OV90 CDK7 inhibition in the context of heat shock** Schematic for the experimental design. A) Cells for PRO-seq were pre-treated (grey) for 30min with either DMSO (blue circles) or SY5609 (orange circles). Cells were heat shocked at 42°C (gold) for 0min, 20min or 45min before the nuclei were harvested for library preparation. B) Cells for RNA-seq were pre-treated (grey) for 30min with either DMSO (green circles) or a CDK7 inhibitor (purple/pink circles). There are three sets of RNA-seq experiments. In the first set, the RNA was extracted immediately after the pre-treatment (top). The second set (middle) was pre-treated with DMSO or one of three distinct CDK7 inhibitors, SY5609, SY5102 or 3MB-PP1 (analog sensitive cell line). This set was heat shocked at 42°C for 30min (gold) before undergoing a 60min recovery at 37°C (teal). The final set (bottom) was pre-treated with DMSO or SY5609 before undergoing a 120min heat shock at 42°C with no recovery.

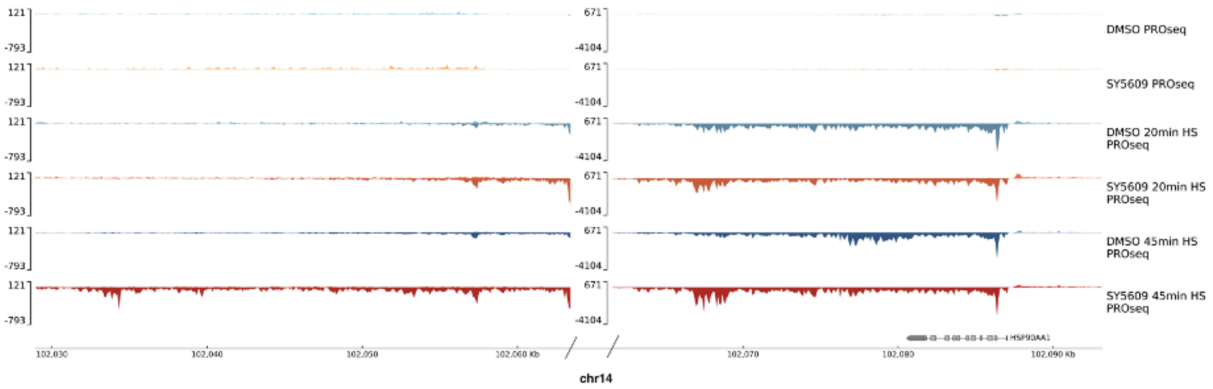


Figure A.2: **Example trace at HSP90.** This PRO-seq trace at a canonical heat shock gene. With no heat shock (top two traces) this gene is not transcribed. At 20min heat shock (middle two traces) the gene is activated. At this point it already seems like there is more run-on in the SY5609 sample compared to DMSO. By 45min heat shock (bottom two traces) there is clearly additonal run on in the SY5609 sample compared to DMSO, indicating that there may be a splicing/termination defect.

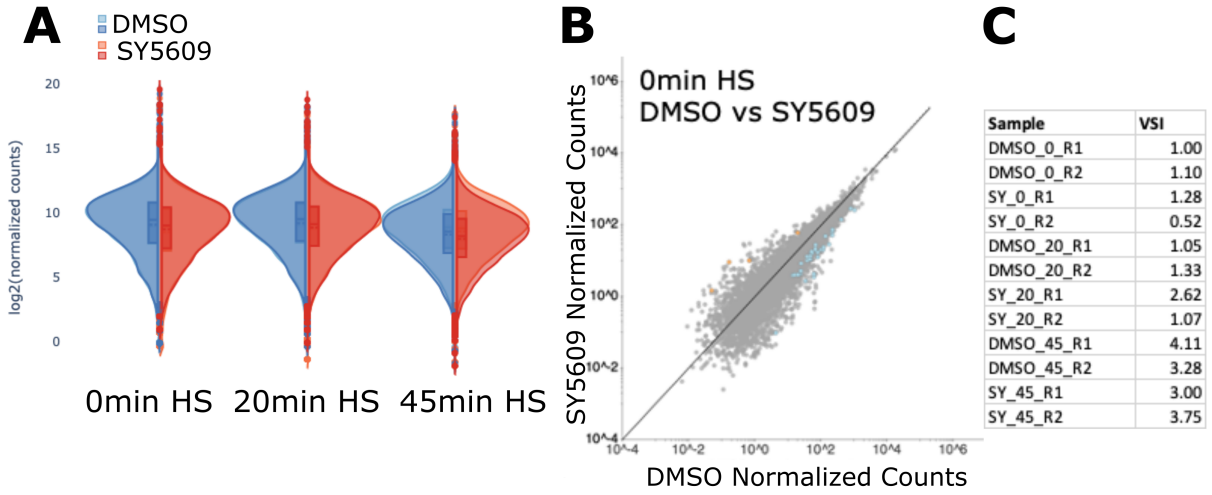


Figure A.3: **SY5609 treatment in OV90 cells causes a mild global reduction of gene body transcription in normalized PRO-seq data.** A) Violin plots demonstrating the average 20% decrease in gene body transcription upon SY5609 treatment (blue-DMSO replicate 1 and replicate 2, orange- SY5609 replicate 1 and replicate 2). 2) Scatter plot of normalized count values for DMSO (x-axis) versus SY5609 (y-axis). The expectation of no reduction in gene body transcription is a slope 1.0 (plotted here). The best fit slope is 0.9. The central mass of normalized gene counts falls below the 1:1 line. This also indicates a general, mild reduction in gene body transcription upon treatment with SY5609. C) Sample normalized values calculated using the virtual spike-in values. These values are relative normalization factors (relative to DMSO 0 R1).

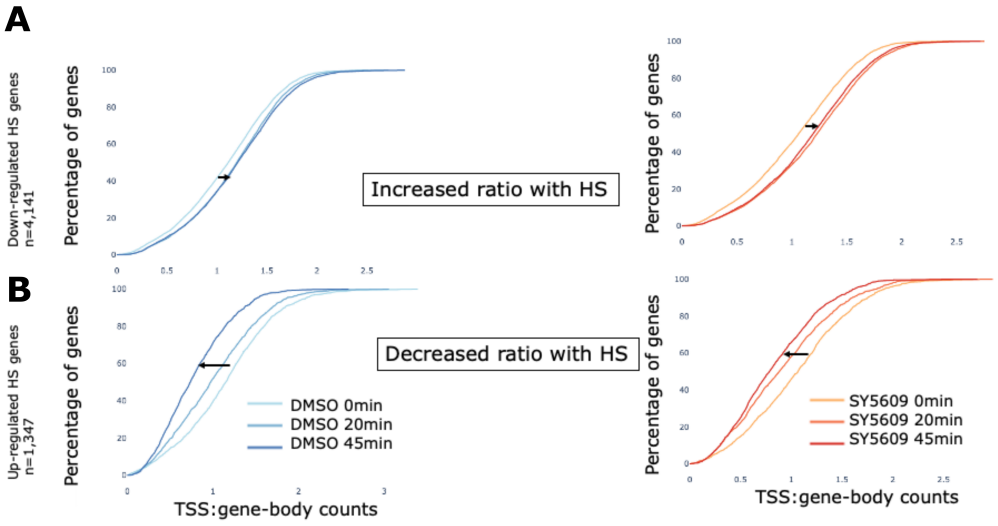


Figure A.4: **Heat shock treatment induces increased pause-index at down-regulated genes, and increased pause release (decreased pause index).** Pause-index (PI; calculated as the ratio of promoter-proximal signal (TSS):gene body signal) plotted as a cumulative distribution. The heat shock (HS) treatment gives the expected result of A) an increase in PI at genes down-regulated with heat shock in both DMSO (darker blue indicates longer heat shock) and SY5609 (darker orange indicates longer heat shock) conditions, and B) a decrease in PI at genes up-regulated with heat shock[159].

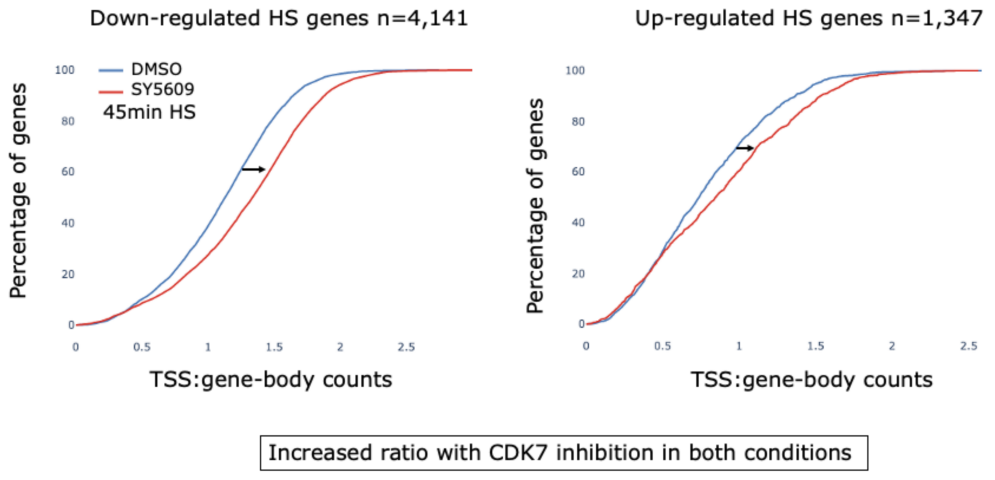


Figure A.5: **SY5609 treatment increases pause-index relative to control.** In 45min heat shock (HS) conditions, SY5609 (orange) causes increased pause-index (PI) regardless of up-regulation or down-regulation compared to the DMSO control (blue). This indicates a lack of productive elongation upon CDK7 inhibition.

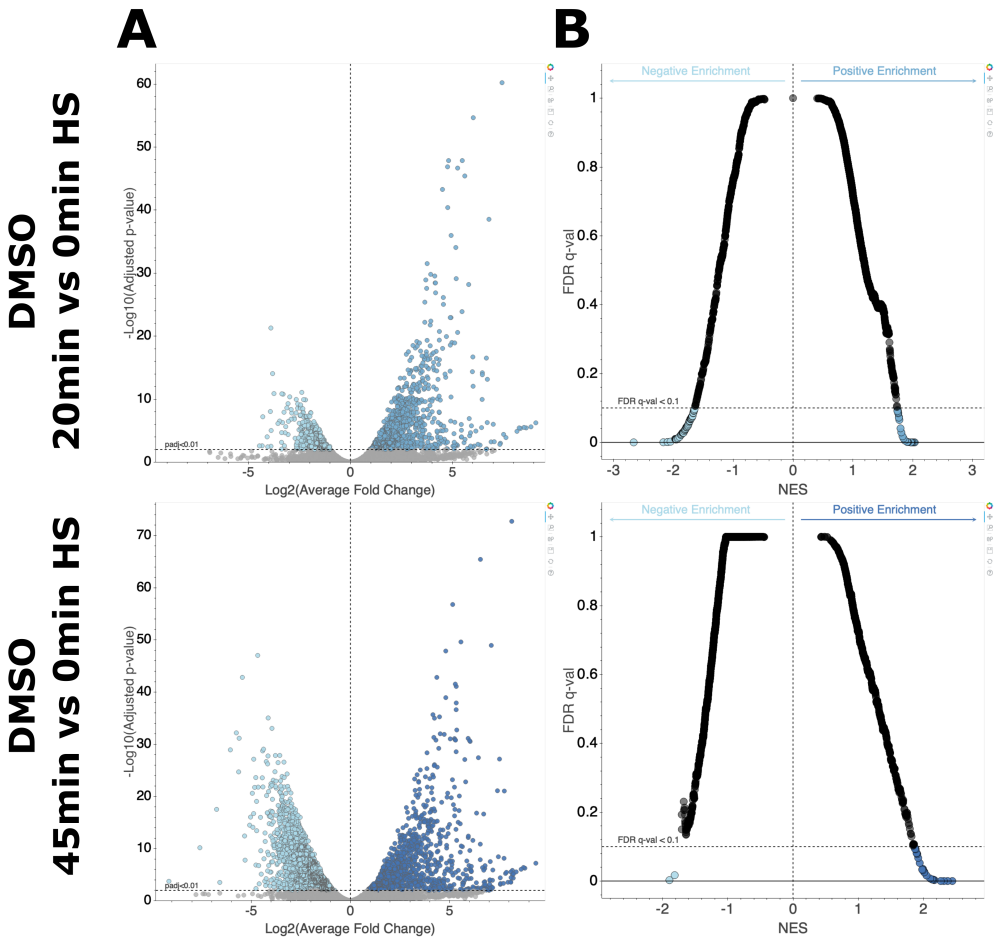


Figure A.6: **PRO-seq differential expression and gene enrichment analyses upon heat shock.** A) PRO-seq gene body results of the differential gene profile at 20min (top) and 45min (bottom) heat shock versus 0min control. Each point represents a coding gene in the genome. The x-axis shows the log2 fold change of the gene counts between the two conditions, where genes on the left (light blue) are significantly down-regulated by HS and on the right (dark blue) are significantly up-regulated by HS (p-value < 0.01). The y-axis is the -log10 of the adjusted p-value, where a higher value indicates greater confidence that a given gene is differentially expressed. B) The gene enrichment analysis results based on the gene set in A. The x-axis is the normalized enrichment score (NES) of a GO biological processes pathway. The y-axis is the FDR corrected q-value, where points at the bottom are the highest confidence pathways. One of the highest confidence up-regulated pathways at both 20min and 45min HS is cellular heat shock response (20min: NES=2.03, q-val=0; 45min: NES=2.15, q-val=0.004).

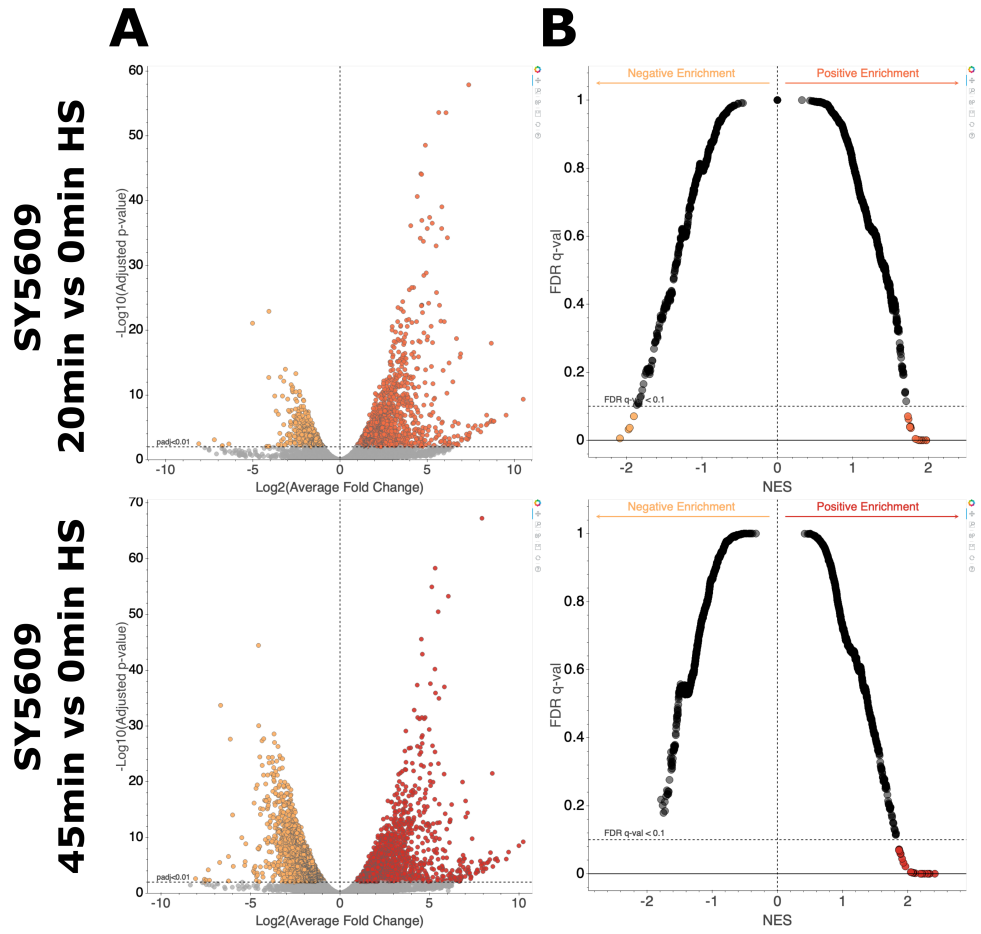
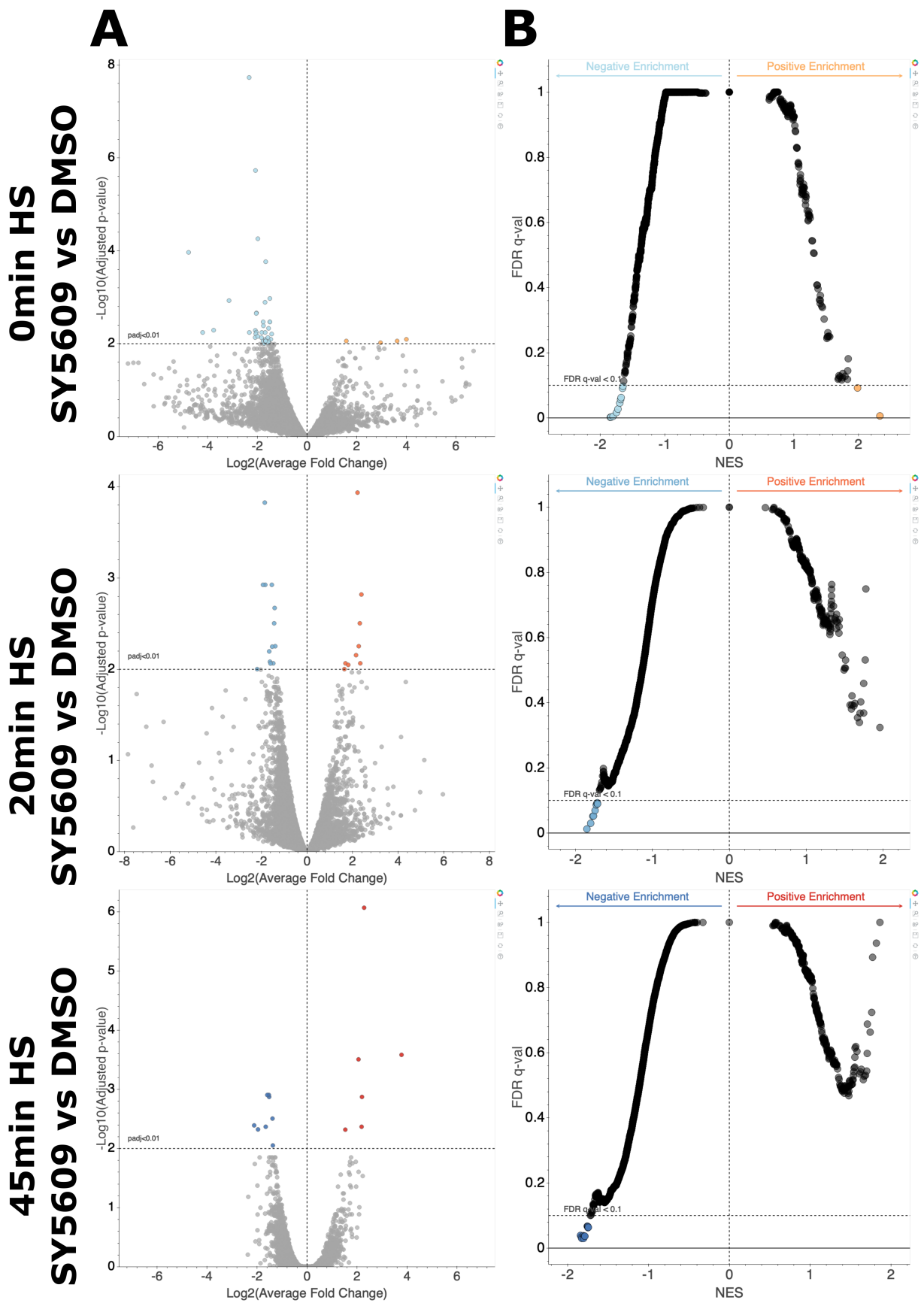


Figure A.7: **PRO-seq differential expression and gene enrichment analyses upon heat shock in the context of CDK7 inhibition.** A) PRO-seq gene body results of the differential gene profile at 20min (top) and 45min (bottom) heat shock versus 0min control after a 30min pre-treatment of SY5609. Each point represents a coding gene in the genome. The x-axis shows the log2 fold change of the gene counts between the two conditions, where genes on the left (light orange) are significantly down-regulated by HS and on the right (dark orange) are significantly up-regulated by HS (p-value < 0.01). The y-axis is the -log10 of the adjusted p-value, where a higher value indicates greater confidence that a given gene is differentially expressed. B) The gene enrichment analysis results based on the gene set in A. The x-axis is the normalized enrichment score (NES) of a GO biological processes pathway. The y-axis is the FDR corrected q-value, where points at the bottom are the highest confidence pathways. One of the highest confidence up-regulated pathways at both 20min and 45min HS is cellular heat shock response after SY5609 treatment (20min: NES=1.89, q-val=0; 45min: NES=2.26, q-val=0).



Continued on next page.

Figure A.8: **PRO-seq differential expression and gene enrichment analyses contrasting CDK7 inhibition versus control across heat shock time points.** A) PRO-seq gene body results of the differential gene profile +/- SY5609 at 0min, 20min and 45min heat shock (HS). Each point represents a coding gene in the genome. The x-axis shows the log2 fold change of the gene counts between the two conditions, where genes on the left (blue) are significantly down-regulated by CDK7 inhibition and on the right (orange) are significantly up-regulated by CDK7 inhibition (p-value < 0.01). The y-axis is the -log10 of the adjusted p-value, where a higher value indicates greater confidence that a given gene is differentially expressed. B) The gene enrichment analysis results based on the gene set in A. The x-axis is the normalized enrichment score (NES) of a GO biological processes pathway. The y-axis is the FDR corrected q-value, where points at the bottom are the highest confidence pathways. One of the highest confidence down-regulated pathways across all time points is cell cycle DNA replication (0min: NS=-1.68, q-val=0.05, 20min: NES=-1.73, q-val=0.08; 45min: NES=-1.74, q-val=0.06).

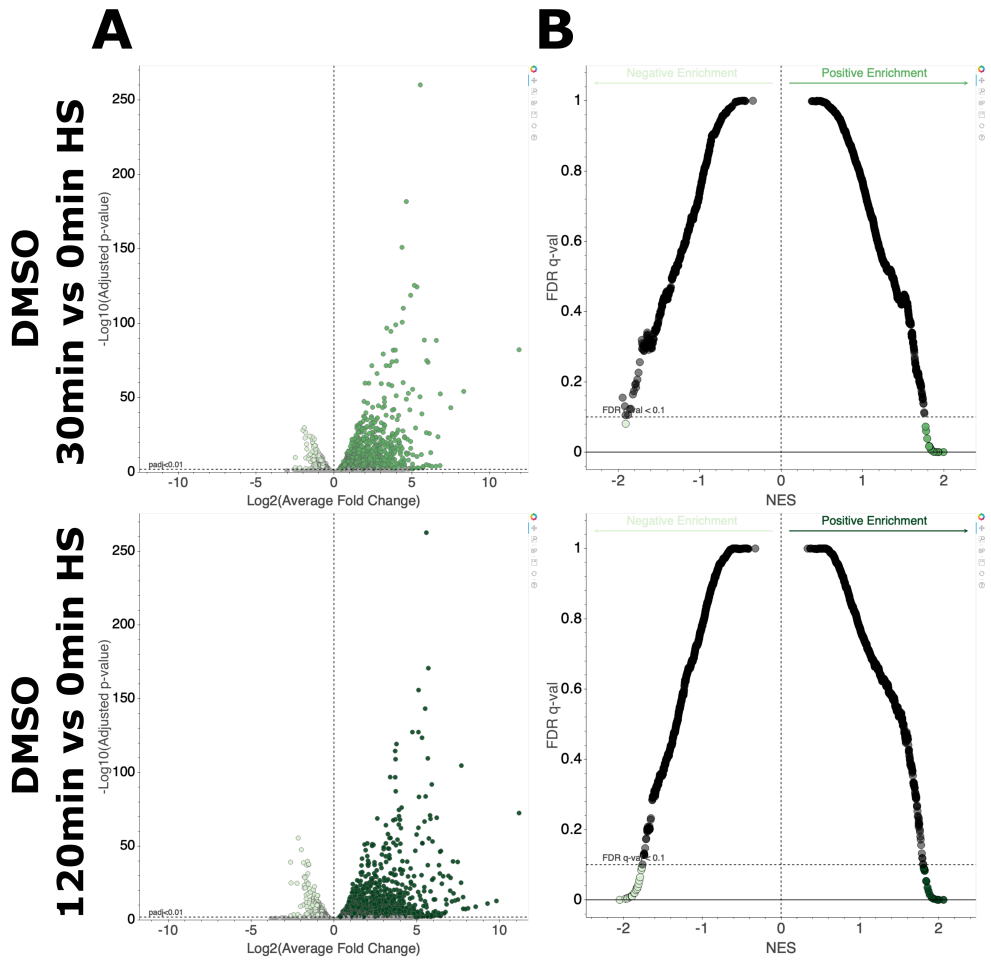


Figure A.9: **RNA-seq differential expression and gene enrichment analyses upon heat shock.** A) RNA-seq results of the differential gene profile at 30min (top) and 120min (bottom) heat shock versus 0min control. Each point represents a coding gene in the genome. The x-axis shows the log2 fold change of the gene counts between the two conditions, where genes on the left (light green) are significantly down-regulated by HS and on the right (dark green) are significantly up-regulated by HS (p-value < 0.01). The y-axis is the -log10 of the adjusted p-value, where a higher value indicates greater confidence that a given gene is differentially expressed. B) The gene enrichment analysis results based on the gene set in A. The x-axis is the normalized enrichment score (NES) of a GO biological processes pathway. The y-axis is the FDR corrected q-value, where points at the bottom are the highest confidence pathways. One of the highest confidence up-regulated pathways at both 30 and 120min HS is cellular heat shock response (30min: NES=1.90, q-val=0.0002; 120min: NES=1.91, q-val=0.006).

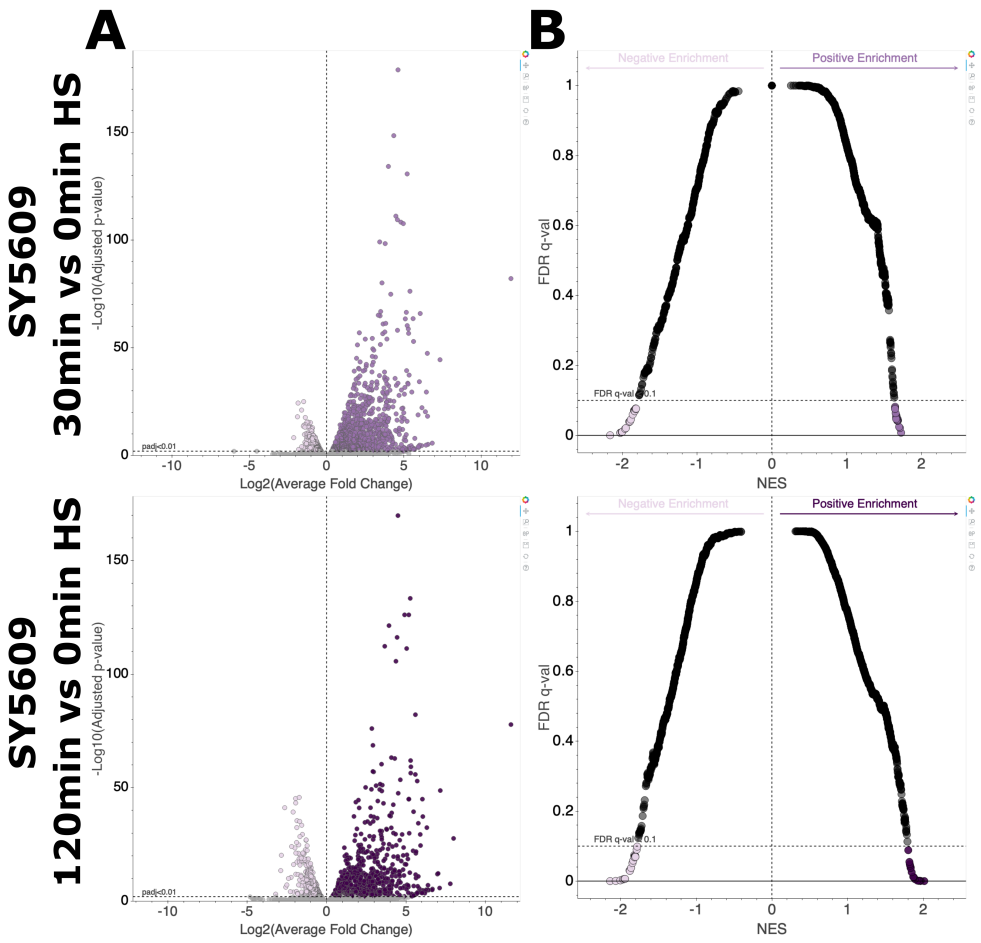
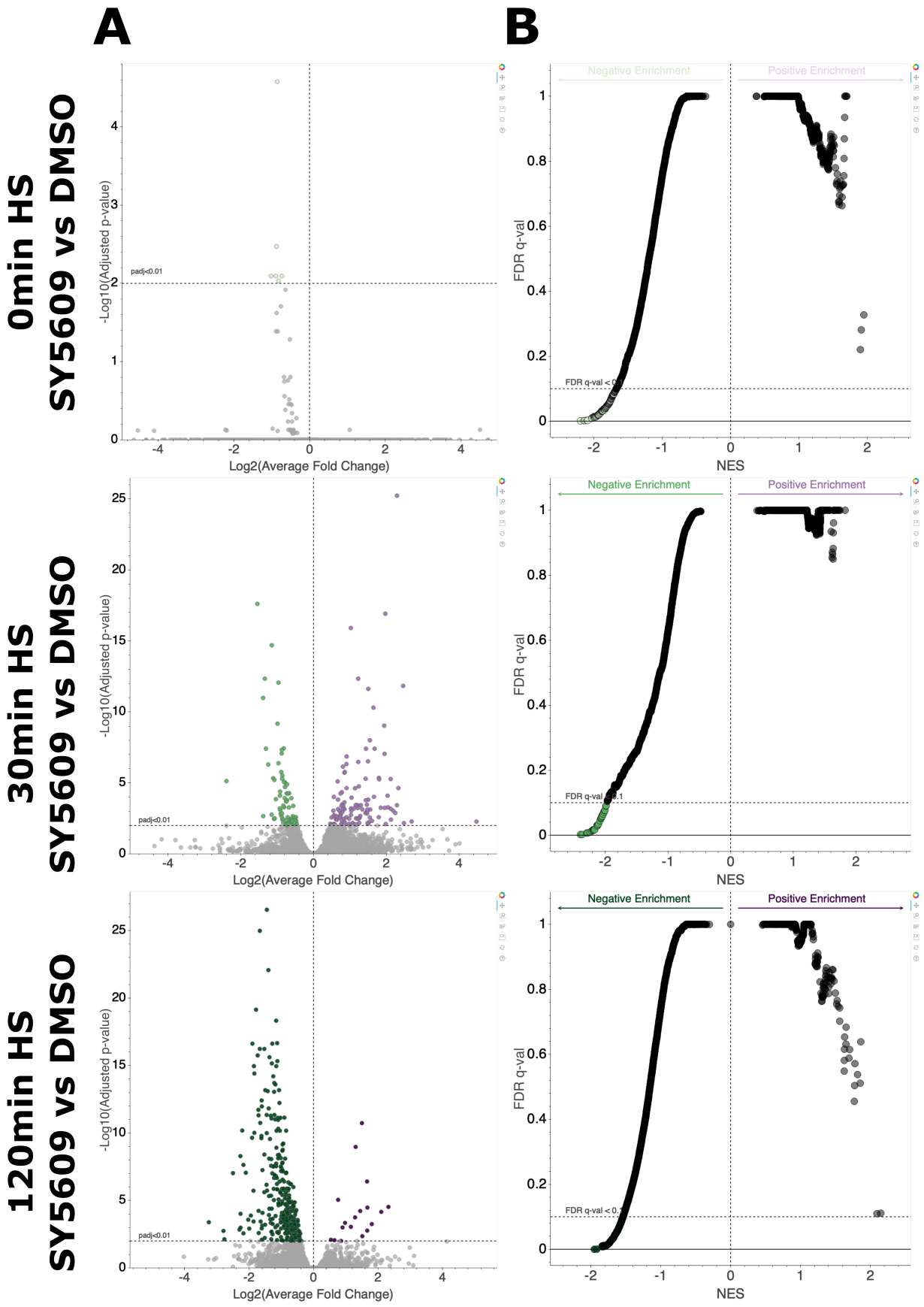


Figure A.10: **RNA-seq differential expression and gene enrichment analyses upon heat shock in the context of CDK7 inhibition.** A) RNA-seq results of the differential gene profile at 30min (top) and 120min (bottom) heat shock versus 0min control. Each point represents a coding gene in the genome. The x-axis shows the log2 fold change of the gene counts between the two conditions, where genes on the left (light purple) are significantly down-regulated by HS and on the right (dark purple) are significantly up-regulated by HS (p-value < 0.01). The y-axis is the -log10 of the adjusted p-value, where a higher value indicates greater confidence that a given gene is differentially expressed. B) The gene enrichment analysis results based on the gene set in A. The x-axis is the normalized enrichment score (NES) of a GO biological processes pathway. The y-axis is the FDR corrected q-value, where points at the bottom are the highest confidence pathways. One of the highest confidence up-regulated pathways at both 30 and 120min HS is cellular heat shock response (30min: NES=1.68, q-val=0.04; 120min: NES=1.92, q-val=0.001).



Continued on next page.

Figure A.11: **RNA-seq differential expression and gene enrichment analyses contrasting CDK7 inhibition versus control across heat shock time points.** A) RNA-seq differential gene results +/- SY5609 at 0min, 30min and 120min heat shock (HS). Each point represents a coding gene in the genome. The x-axis shows the log2 fold change of the gene counts between the two conditions, where genes on the left (green) are significantly down-regulated by CDK7 inhibition and on the right (purple) are significantly up-regulated by CDK7 inhibition (p-value < 0.01). The y-axis is the -log10 of the adjusted p-value, where a higher value indicates greater confidence that a given gene is differentially expressed. B) The gene enrichment analysis results based on the gene set in A. The x-axis is the normalized enrichment score (NES) of a GO biological processes pathway. The y-axis is the FDR corrected q-value, where points at the bottom are the highest confidence pathways. Many down-regulated pathways center around the reduction of cellular proliferation.

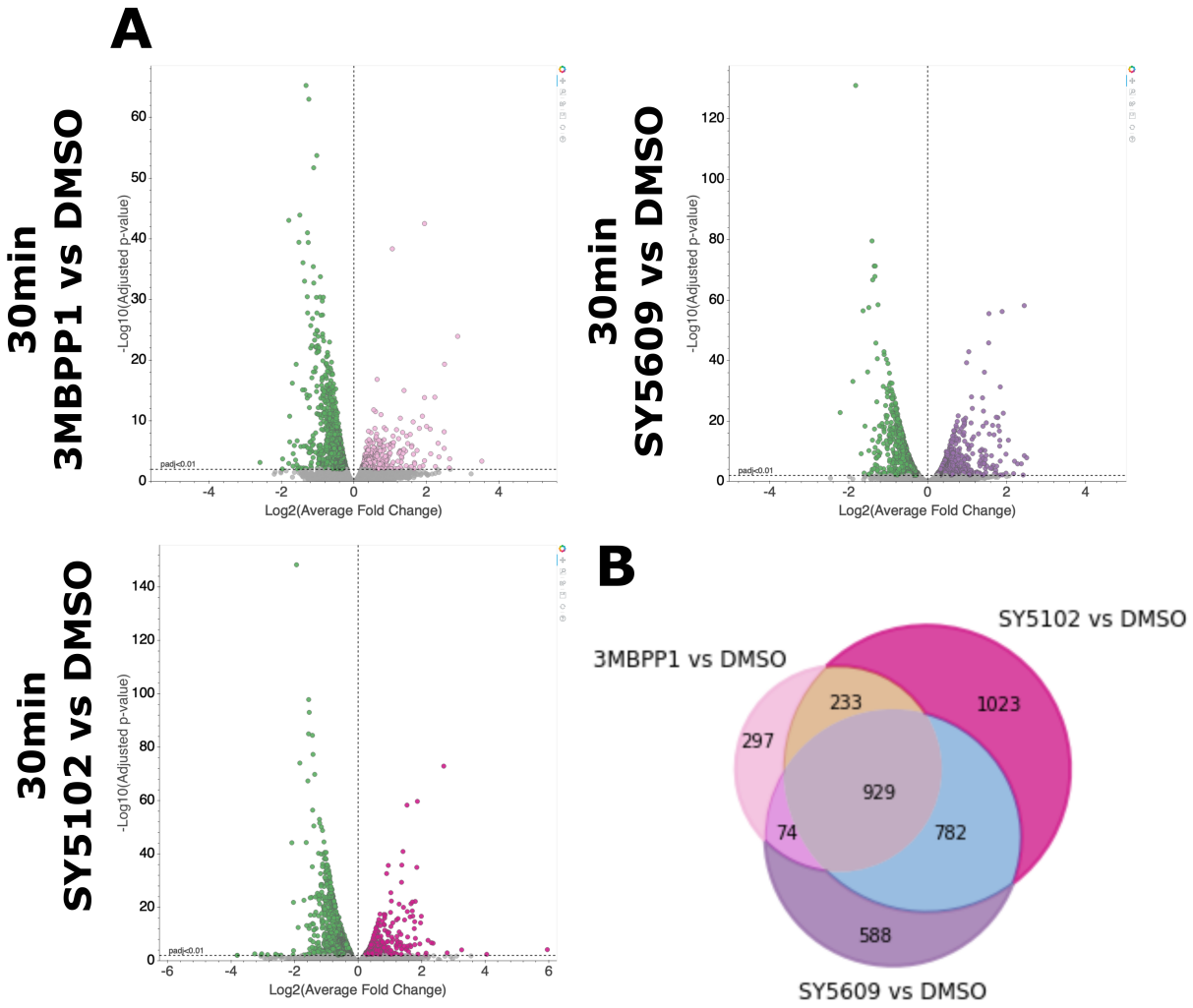


Figure A.12: **RNA-seq differential expression analysis in three CDK7 inhibition conditions at 30min heat shock.** A) RNA-seq differential gene results +/- 3MBPP1 in OV90 CDK7 analog sensitive cells (top left), +/- non-covalent inhibitor SY5102 (bottom left) and +/- SY5609 (top right). Each point represents a coding gene in the genome. The x-axis shows the log2 fold change of the gene counts between the two conditions, where genes on the left (green) are significantly down-regulated by CDK7 inhibition and on the right (purple) are significantly up-regulated by CDK7 inhibition (p-value < 0.01). The y-axis is the -log10 of the adjusted p-value, where a higher value indicates greater confidence that a given gene is differentially expressed. B) The overlap of significant genes (p-value < 0.01) upon CDK7 inhibition showing that the CDK7 gene targets are largely shared across different inhibition conditions.

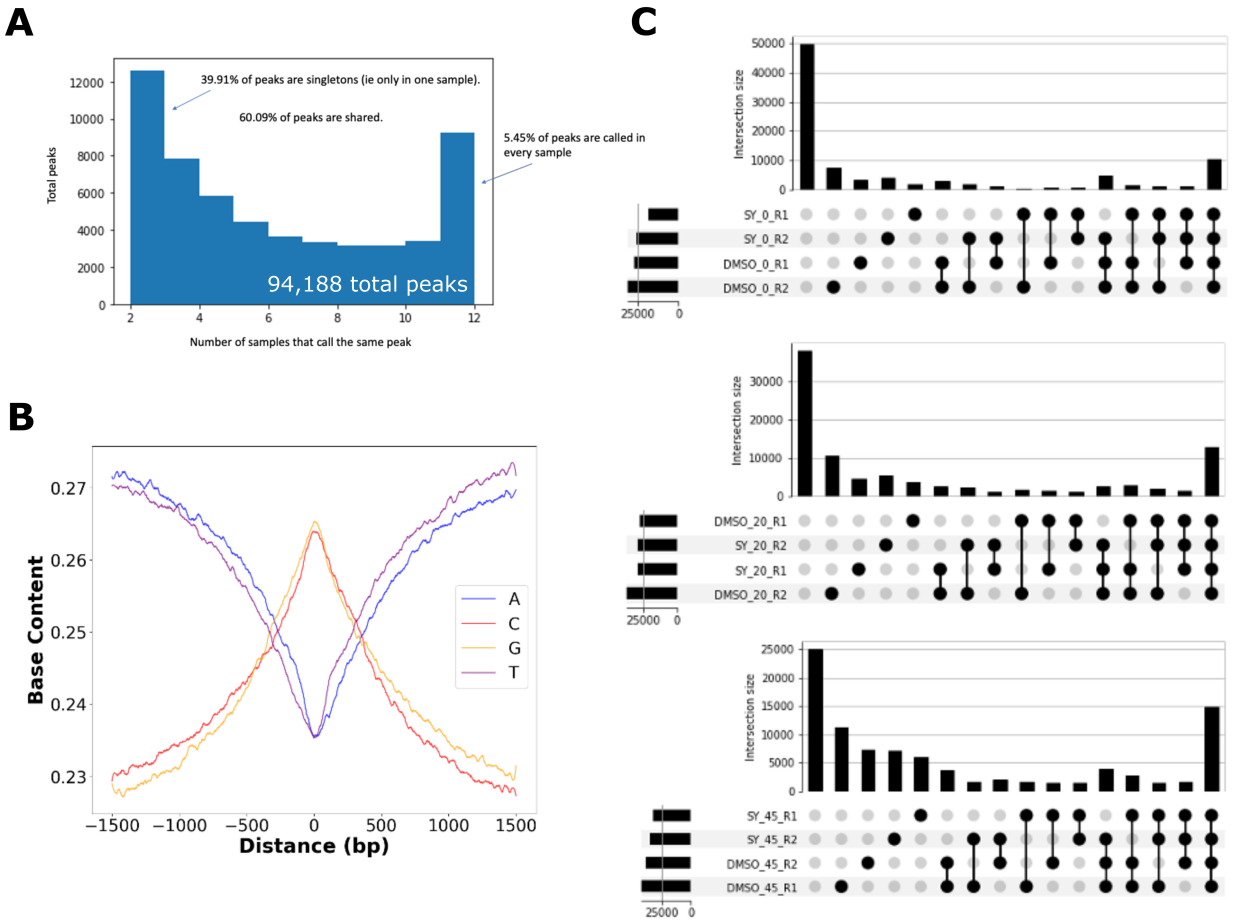


Figure A.13: **Description of bidirectional transcripts identified in OV90 PRO-seq experiments.** Tfit was used to identify bidirectional transcripts across the twelve PRO-seq samples (+/- SY5609 at 0min, 20min and 45min heat shock in biological replicate)[18]. These samples were merged into a master file by muMerge[207] and were used for all subsequent bidirectional analysis, which includes TFEA results discussed in Figures A.14,A.15,A.16,A.17 and A.18, as well as TF profile results in Figure A.16. A) Histogram of 94,188 tfit regions that are in one to twelve samples. This shows roughly 40% of bidirectionals are identified in only one sample, approximately 60% of regions are shared across samples. Only 5.5% of regions are in all samples, these regions are enriched for promoters. B) Base composition profile for the 94,188 bidirectional regions in OV90 cells. The y-axis shows the probability of a base (A/T/C/G) occurring at a given position. On the x-axis 0bp represents the location of PolII initiation (μ). This shows the sharp GC preference at initiation that declines as distance from μ increases. C) Upset plots showing region intersections at 0min heat shock (HS; top), 20min HS (middle) and 45min HS (bottom). The bars on the left represent total regions in the condition indicated. The dots represent the conditions that a given region is represented in from none of the conditions in that HS time point (far left) to all conditions for that HS time point (far right). The top bars show how many regions meet the condition intersections defined by the dots below. The upset plots show no evidence of novel bidirectionals specific to SY5609. The highest fraction of region representation is at 45min HS, possibly due to new bidirectional regions forming in response to the external stress. A large factor in region representation on a per sample basis is read depth, where deeper samples capture more bidirectional regions. The lowest read depth sample is SY5609 0min HS Rep 2 (SY 0 R2, 59.6 million reads) and has the fewest bidirectionals captured across all conditions. Given this data, it is recommended to have at least 100 million reads for robust bidirectional identification.

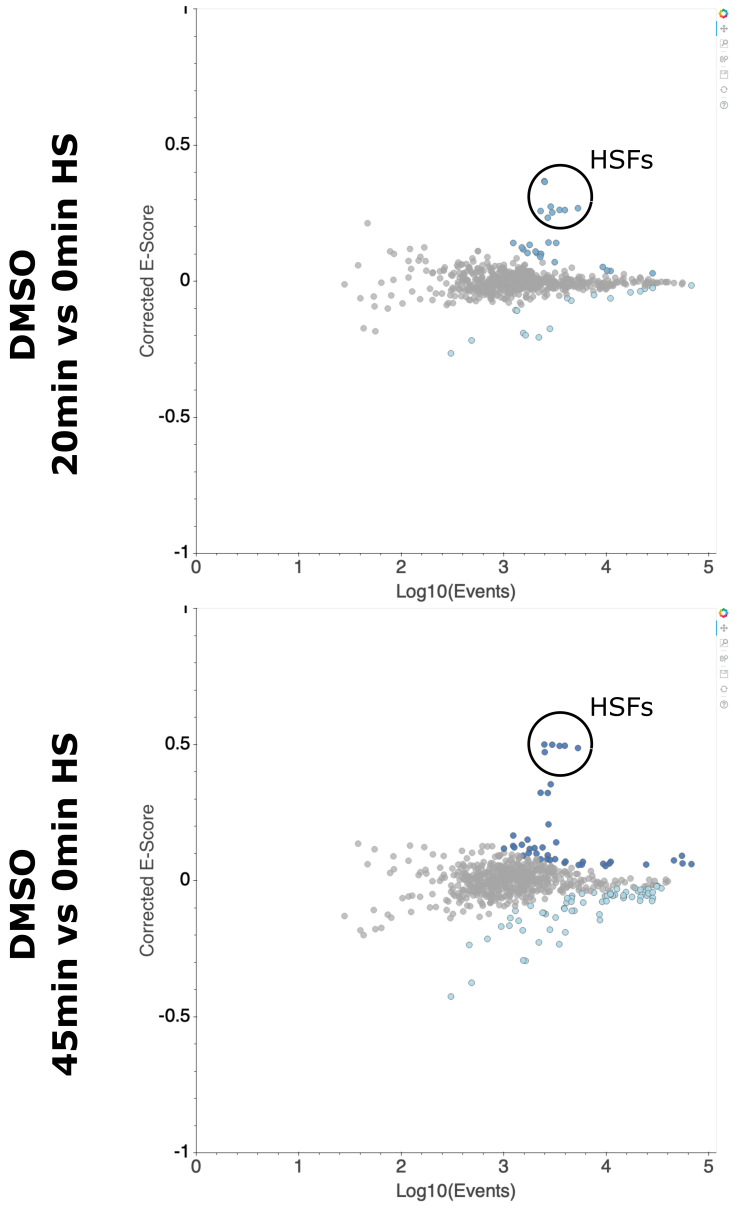


Figure A.14: **PRO-seq transcription factor enrichment results upon heat shock.** Transcription factor enrichment analysis (TFEA) MA plot for 20min (top) and 45min (bottom) heat shock (HS) where each dot represents a single TF with significantly activated in dark blue and significantly repressed in light blue (p-value < 0.01). The x-axis shows total TF motif hits and the y-axis shows a measure of TF enrichment accounting for the co-localization of a TF motif with PolII initiation regions genome wide[207]. Heat shock factors (HSFs) are the most highly enriched TFs in these condition.

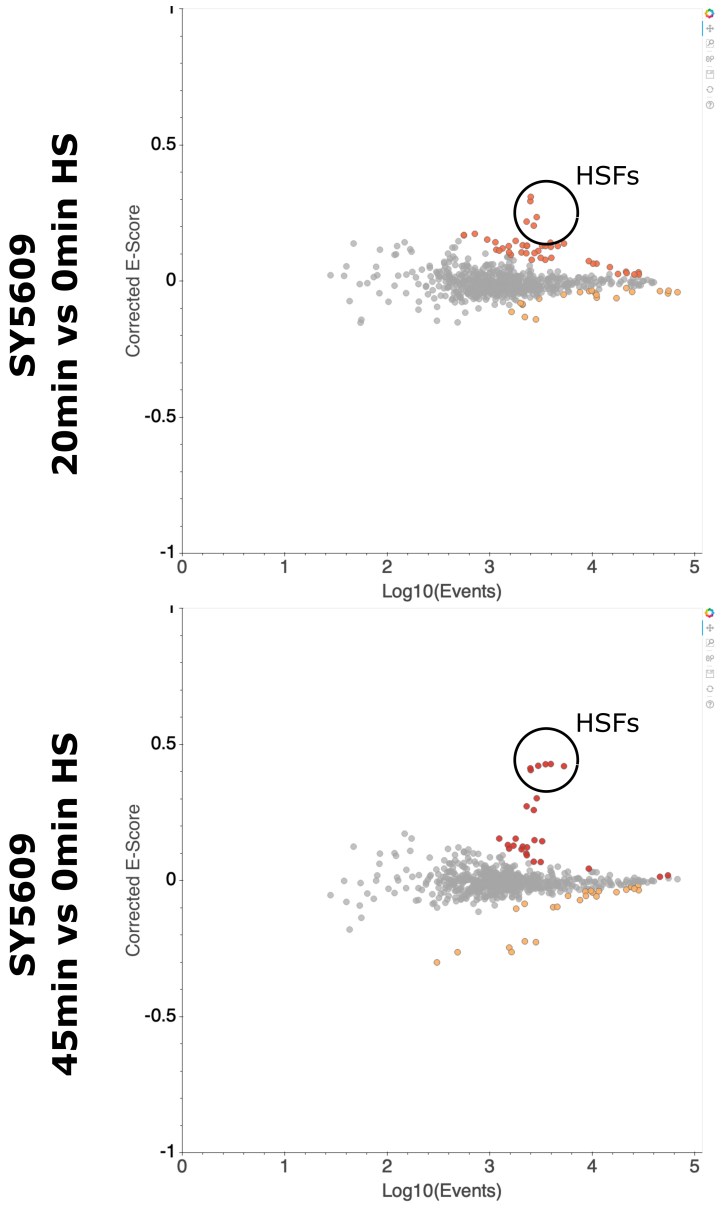


Figure A.15: **PRO-seq transcription factor enrichment results upon heat shock in the context of CDK7 inhibition.** Transcription factor enrichment analysis (TFEA) MA plot for 20min (top) and 45min (bottom) heat shock (HS) where each dot represents a single TF with significantly activated in dark orange and significantly repressed in light orange (p-value < 0.01). The x-axis shows total TF motif hits and the y-axis shows a measure of TF enrichment accounting for the co-localization of a TF motif with PolII initiation regions genome wide[207]. Heat shock factors (HSFs) are the most highly enriched TFs in these condition.

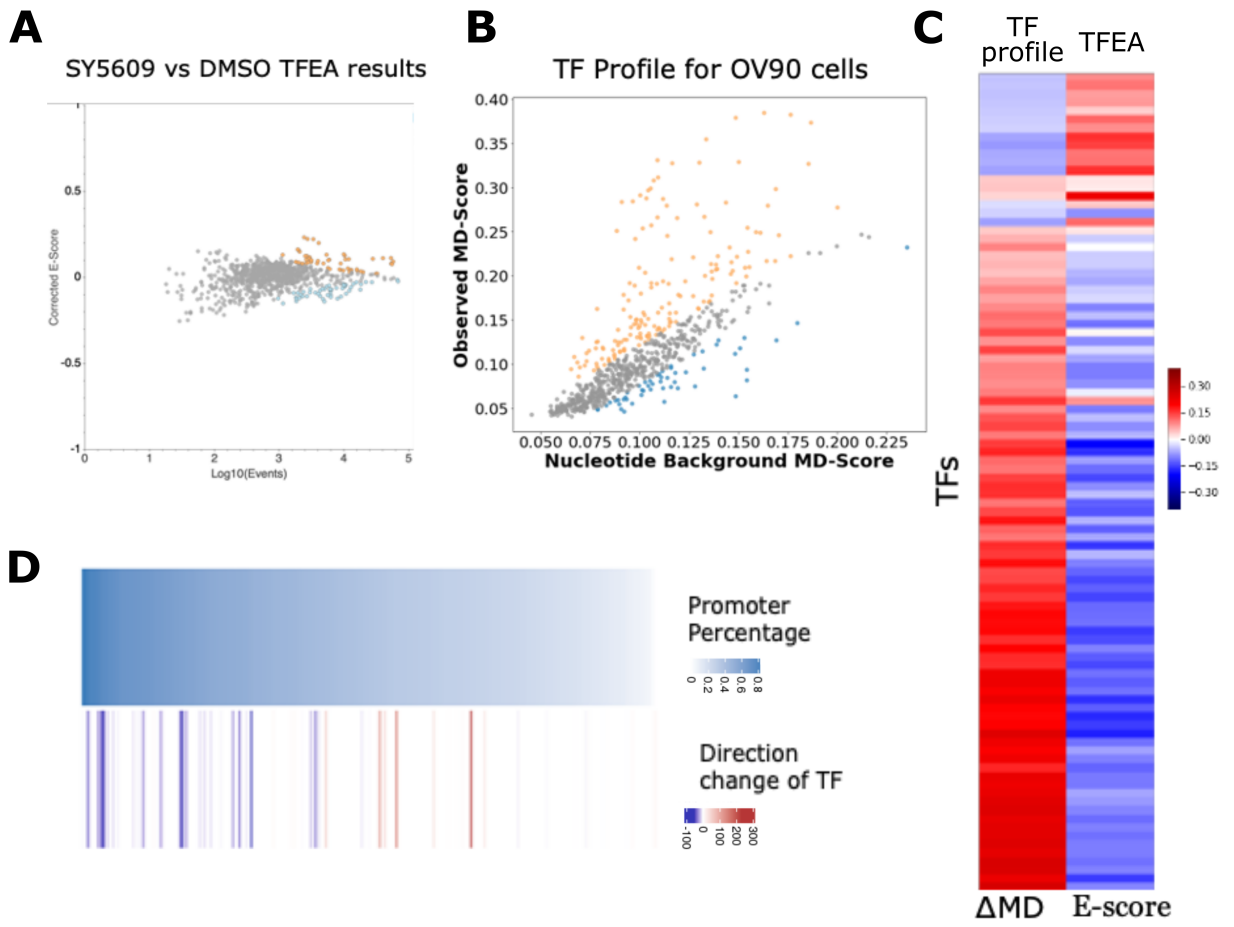


Figure A.16: **Treatment with SY5609 reduces activity of promoter associated TFs.** A) TFEA MA plot for SY5609 versus DMSO where each dot represents a single TF with significantly activated in orange and significantly repressed in blue (p-value < 0.01). The x-axis shows total TF motif hits and the y-axis shows a measure of TF enrichment accounting for the co-localization of a TF motif with PolII initiation regions genome wide[207]. B) TF profile for OV90 cells showing basally activated TFs. Discussed in depth in Chapter 2, each point represents an individual TF. The TF motif-displacement score is assessed in experimental conditions (y-axis) or against a position specific background model (x-axis). The contrast between the experimental data and the model provides the means to attribute significance to basally activated TFs, where orange is enriched and blue is depleted. C) Heatmap contrasting the TF profile identified factors (left, from analysis in panel B) with the TFEA enrichment results (right, from panel A). Red indicates an enriched TF, blue indicates a repressed or depleted TF. TFs that are enriched/active in the TF profile tend to have negative enrichment scores (repressed/inactive) in TFEA. This result suggests that CDK7 inhibition by SY5609 causes widespread repression of basally active TFs in OV90 cells. D) Heatmap showing the percent promoter content for all TFs identified by ChIP-seq from cistromeDB[167, 258](top) where darker blue indicates a higher promoter percentage, compared to the TFEA results (bottom: blue=repressed; red=enriched, heat shows -log10(p-adj) multiplied by the direction of enrichment change) showing that SY5609 treatment disproportionately impacts TFs with motifs that are more promoter associated.

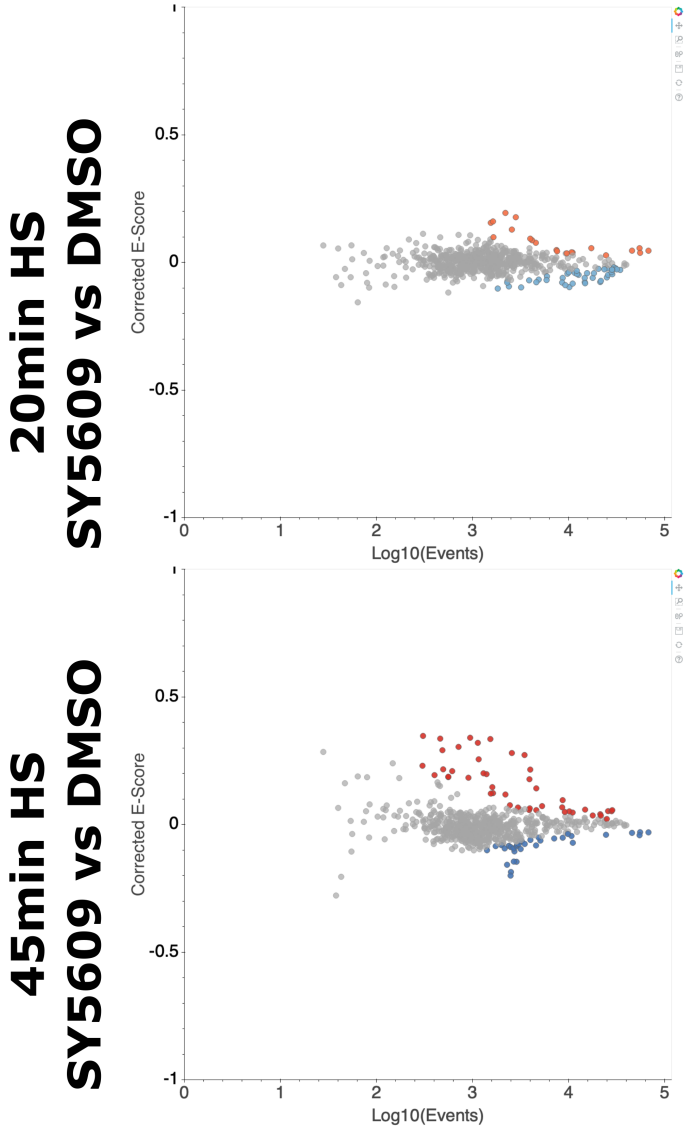


Figure A.17: **PRO-seq transcription factor enrichment results contrasting CDK7 inhibition versus control across heat shock time points.** Transcription factor enrichment analysis (TFEA) MA plot for +/- SY5609 at 20min (top) and 45min (bottom) heat shock (HS). Each dot represents a single TF with significantly activated in orange and significantly repressed in blue (p-value < 0.01). The x-axis shows total TF motif hits and the y-axis shows a measure of TF enrichment accounting for the co-localization of a TF motif with PolII initiation regions genome wide[207]. Promoter associated TFs are the most highly repressed TFs in these conditions. SY5609 versus DMSO 0min HS is shown in Figure A.16.

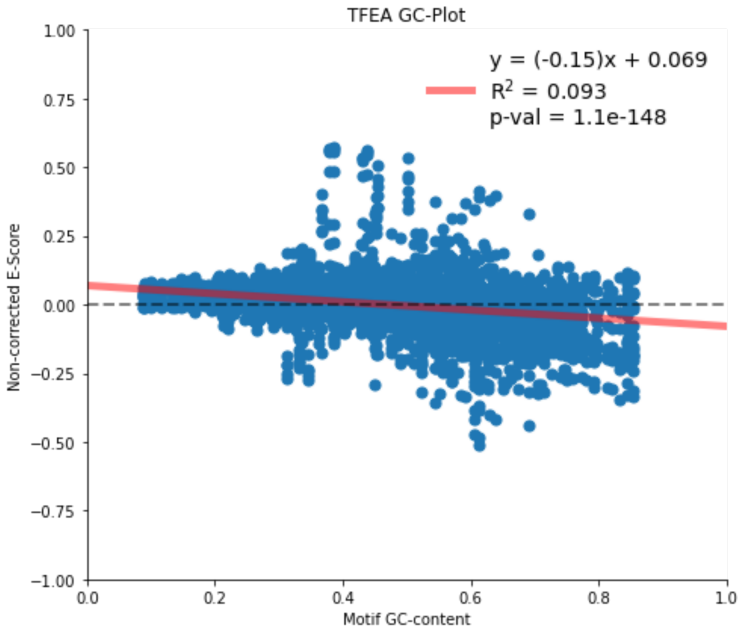


Figure A.18: **Correction for GC-bias in TFEA results across all conditions.** TFEA has an internal feature which corrects from the over-representation of GC rich motifs at initiation regions (Fig. A.13B)[207]. The non-corrected TF enrichment score (E-score) on the y-axis is a measure of TF activation or repression in a given comparison, 0min heat shock SY5609 versus DMSO for example. The x-axis shows the GC content of each TF motif used in the experiment. The points represent the non-corrected E-score for a given TF in a single comparison. A linear regression of the points is fit to correct for GC biases and corrected E-Scores are the basis of all subsequent analyses. In solving the linear equation the most GC rich motifs will have the largest correction value (ie the larger the x-value, the more the y-value will change). Across individual comparisons the motif GC content will be invariable but the non-corrected E-score will vary. One observation was that CDK7 inhibition seemed preferentially impact TFs associated with promoters (Fig. A.16D). Promoter associated factors tend to have more GC rich motif preferences (discussed in detail in Chapter 2). For this reason, one concern was that the observation that CDK7 inhibition impacts promoter-associated TFs was simply an artifact of the GC correction. For this reason, all relevant comparisons in this study (DMSO heat shock time course, SY5609 heat shock time course and SY5609 versus DMSO across all heat shock, n=9 comparisons) were plotted for a single GC correction all together, shown here. This means that the GC correction was equivalent across all conditions, and the observation that CDK7 inhibition preferentially impacts promoter associated TFs was still observed.

Appendix B

Methods

B.1 TF Inference Methods

Curating data from DBNascent: All nascent RNA-seq data was curated from NCBI GEO by the Dowell Lab. Data processing was primarily performed by Rutendo Sigauke and Lynn Sanford[218]. To prepare the data for TF profiling all tfit[18] calls in PRO-seq and GRO-seq human experiments (mapped to hg38) were examined. A minimum quality score of "4" was required for all samples. This score indicates a minimum of 5 million reads, over 50% of reads map to the reference genome and less than 95% duplication. As many biological replicates were kept as possible, but if one replicate passed this quality metric it was still used for subsequent analysis. All tfit calls between biological replicates within a given paper were merged using muMerge[207]. This was done to maximize number of reliable calls per experimental data set. After merging biological replicates, we required that the region within $2h$ of μ had a base composition of at least 50% GC content. The reason for this requirement some data sets have prohibitively low complexity– likely due to failed pull-down of nascent RNA. This results in tfit calling noise as PolII initiation regions and causes the base composition surrounding "μ" to be close to genomic background (ie $\approx$ 40% GC). The final requirement is that at least 50% regions must not be at an annotated promoter. If promoter regions are over-represented then we lose the ability to call cell type specific TFs, and only call shared TFs leaving the data set overall uninformative.

After this curation process, we ended up with 126 distinct data sets from 88 papers that represent 79 unique cell lines under basal conditions (n=299 unique biological samples). An additional

161 data sets were curated from 65 papers that represent 46 unique cell lines under perturbation or genome modified conditions (n=411 unique biological samples).

Building a training data set: The training data was based on the 126 distinct control data sets. First, we used muMerge[207] to merge all 126 data sets (ie tfit calls merged by biological replicate) into a single master file. Due to technical limitations, the merge had to be performed step-wise rather than all data at once. Mainly, if a single region was called in many data sets, the confidence of μ would converge to 0 nucleotides causing the program to fail. For this reason, all samples of a given tissue type were merged into a tissue specific region file. Any region less than 20nt were windowed to be at least 20nt in length. The tissue specific region files were then merged into the master file.

This resulted in a master file that represented all PolII initiation regions captured in 79 unique cell lines across 88 papers. These regions were then divided into two populations, enhancers and promoters. All regions (windowed by h=150) within 1000bp (300bp upstream, 700bp downstream) of refseq annotated transcription start site (TSS) were defined as promoters. Any region outside of this parameter was defined as an enhancer. This resulted in a total of 53,244 promoter regions, and 611,963 enhancer regions within the master annotation file.

Both regions were treated separately, but identically to extract two sets of probabilities to build 10^6 simulated sequences. First all regions were windowed by H=1500. Sequences were extracted using bedtools getfasta. Using all of the extracted regions, conditional probabilities were calculated in a position specific manner for the standard nucleotides (A,T,C,G), shown in Figure 2.2A. Code used to calculate the probabilities and generate the sequences can be found at `https://github.com/Dowell-Lab/TF_profiler`, within the sequence_generator module.

Quantification of TF gene transcription: Refseq gene counts for every SRR within DBNascent were counted over hg38 using Rsubread, featurecounts[218]. For all SRRs within a biological replicate for a given data set (both control, n=126 and perturbation, n=161), the mean RPKM was calculated for every gene. Only the highest mean RPKM isoform for every gene was retained for additional analyses.

Motif scanning: Motif scanning was performed using the MEME suite function FIMO scan[19]. This scan was performed using a flat background model (equal distribution assumed of the four canonical nucleotides). The threshold was set to 1e-5. The motif files used were curated from HOCOMOCO version 11[134]. The scan was performed across the human genome (hg38) and these motif instances were used for subsequent analysis. Internal to the TF profiler program the motif scan can also be performed de novo across only the bidirectional regions provided, or take in pre-scanned regions genome wide. Motif scanning was performed on simulated sequences using the same parameters. TF profiler will generate sequences based on the provided annotation file and scan those regions for motif hits, or will take pre-calculated distance tables for the background model.

TF profile distance calculations: To measure TF co-localization the relative distance between a motif hit and the center of the bidirectional transcript must be assessed. The distance for all motif hits within the large window (H=1500) of a given region were calculated. This distance is calculated as the distance from the center of the motif to the distance of the center of the bidirectional. For motifs and regions of odd length the center is rounded to the nearest even integer per the native python rounding function.

Each motif instance is associated with the distance to the center of the bidirectional as well as two ranking metrics for the occasion that there is more than one motif instance in a given bidirectional region. The distance rank assesses which motif instance is closest to the center of the bidirectional, where 1 is the closest. The quality rank defined by the fimo score where 1 is the highest quality motif hit. All motif instances within the large window are stored within the distance tables. Code used to generate these tables can be found at `https://github.com/Dowell-Lab/TF_profiler`, within the distance module. For scoring purposes, only one motif hit per region is counted. If there are multiple motif instances within a given region then 1) the highest quality instance is kept based on fimos quality score. If there are multiple hits of equivalent highest quality then the hit closest to the center is used.

TF profile scoring: The motif-displacement score (MD-score) and associated statistics

are described in detail in section 2.4.1. Briefly, the MD-score is defined by equation 2.1. This metric measures motif co-localization with the center of the bidirectional transcripts within a given data set. This score is compared to the model calculated score defined by equation 2.3.

To statistically define whether the MD-score is higher (ON-UP) or lower (ON-DOWN) than expected, we fit a set proportion of inlier TF MD-scores to a linear equation. For the meta-analysis in this study all TFs across all data sets were set in two main linear fits, control conditions (Fig. 2.4A, n=48,888 points, 126 data sets with 388 TFs) and perturbation conditions (Fig. 2.4D, n=62,468 points, 161 data sets with 388 TFs). The outlier percentage was tuned to obtain a slope of 1.0 and an intercept of 0.0. The normal distribution of the residuals of the inliers was used to attribute a p-value for each TF across all data sets (Fig. 2.4B,E). The data was fit all together to get a better estimate the distribution of the residuals. This was attempted using single SRRs (ie biological replicates were not merged prior to the analysis) and resulted in abnormally high slopes ($m \approx 1.5$), negative intercepts and very large variance resulting in very few TFs being called significant. The TF profiler program fits the residuals of the inliers for a single data set at a time by default. This can be found at `https://github.com/Dowell-Lab/TF_profiler`, within the scoring and statistics modules.

Clustering and TF classes: Clustering TF profiles was performed using the R package ComplexHeatmaps which utilizes the native R function hclust. Profiles were defined using numerical representation of ON-UP as 1, ON-DOWN as -1 and not significant as 0. To clean noise within the data, if a TF was not represented in at least 50% of the data sets in a given cell line it was treated as not significant.

The Eucledian distances were used followed by wards method to cluster the profiles. Numerous distance metrics and clustering metrics were tested, including pearson and kendall distances as well as k-means and complete clustering. Additionally, seriation was attempted via the seriation R package. While most methods reliably recapitulated the cell specific clusters, eucledian/ward did so most cleanly and consistently. Seriation methods most robustly captured cell type of all other methods, but the cell specific TF clusters were not visually clear, thus were not used for central

figures. Clustering in a cut-off independent fashion was performed by clustering the -log(p-value) for each TF across profiles. This methodology also robustly recapitulated cell line specific clusters. Manual reordering of the dendrogram nodes for clarity was performed using the R package dendextend.

TF classes were extracted from this data in four main categories, unique, shared, ubiquitously shared and environmentally responsive. Each cell line was assigned a consensus TF profile, in which the TFs are represented in at least 50% of the data sets of a given cell line. This resulted in one consensus profile per cell line that were then systematically compared. If a TF was only in a specific cell type, it was identified as unique. If a TF was in some but not cell types, it was identified as shared. If a TF was in all profiles it was identified as ubiquitously shared. Finally, if a TF was called in a perturbation profile, but not in the basal profile for the same cell type it was identified as environmentally responsive. The core classes in this study were defined from the six most highly represented cell lines within DBNascent (HCT116, MCF7, ESC, HUVEC, K562, HeLa).

CistromeDB data curation: Both TF and histone ChIP-seq region data was obtained from CistromeDB[167, 258]. Within this database there are six total quality parameters assessed for every ChIP-seq experiment. These can be broken into two main categories, mapping and peak quality. The mapping scores account for sequence quality, number of unique sequences and unique molecule representation after sub-sampling the data. The peak scores account for the number of peaks, the signal to noise ratio and the overlap of peaks with accessible regions. In order for the ChIP-seq sample to be used here we required the sample to pass at least one parameter within both mapping and peak scores.

CistromeDB contained TF ChIP-seq data for 316 TFs within HOCOMOCO v11 core set (n=388 total) that passed the defined qc standards. The percentage of promoter associated regions was calculated by the total number of regions within 1000bp of the refseq annotated transcription start site (TSS) over the total number of regions within that sample. In many cases there were multiple ChIP-seq samples for a single TF. In this case the mean promoter percentage was used. Heatmaps in figure 2.3 use cistromeDB regions from five independent samples per condition. Dis-

tance tables were generated using the TF profiler program. The R package ComplexHeatmaps was used to plot the motif localization using the generated distance tables.

B.2 Peptide Methods

As published in Allen et. al. [5].

Affinity purification-mass spectrometry: Affinity purification was completed from HeLa cell nuclear extract with a GST fusion of the p53AD (residues 1–70) immobilized on Glutathione-Sepharose beads (GE Life Sciences). After binding, the resin was washed five times with 10 column volumes (CV) 0.5 M KCl HEGN (20 mM Hepes pH 7.6; 0.1 mM EDTA; 10% Glycerol; 0.1% NP-40 alternative) and once with 10 CV 0.15 M KCl HEGN (0.02% NP-40 alternative). Bound proteins were eluted with 30mM GSH in elution buffer (80 mM Tris, 0.1 mM EDTA, 10% Glycerol, 0.02% NP-40, 100 mM KCl) and applied to a 15% to 40% linear glycerol gradient (in 0.15 M KCl HEG) and centrifuged for 6 h at 55,000 rpm. Twenty-two 100μL fractions were removed and Mediator-containing fractions (¿1.0 MDa) were combined for proteomics analysis. GST-p53AD (residues 1-70) affinity purified Mediator complex-containing fractions were precipitated on ice by adding 20% (w/v) TCA, 0.067mg/mL insulin and 0.067%(w/v) sodium deoxycholate. Precipitated protein pellets were washed twice with -20°C acetone and air dried. Proteins were trypsin digested using a slightly modified filter-aided sample prep (FASP) protocol (Wisniewski et al., 2009). Briefly, protein pellets were suspended with 4% (v/v) SDS, 0.1M Tris pH 8.5, 10mM TCEP, incubated 30min at ambient temperature to reduce disulfides. Reduced proteins were diluted with 8M Urea, 0.1M Tris pH 8.5 and iodoacetamide was added to 10mM and incubated 30 minutes in total darkness. Reduced and alkylated proteins were then transferred to a Microcon YM-30 (Millipore) spin concentrator and washed twice with 8M Urea, 0.1M Tris pH 8.5 to remove SDS. Three washes were performed with 2M urea, 0.1M Tris pH 8.5, then trypsin and 2mM CaCl2 were added and incubated approximately 2 hours in a 37°C water bath. Digested peptides were eluted and acidified with 5% (v/v) formic acid. Peptides were then desalted online and fractionated with a Phenomenex Jupiter C18 (5μM 300Å, 0.25 x 150mm) custom fabricated column using a two dimensional LC/MS/MS

method (Agilent 1100). Seven steps of increasing acetonitrile (6, 8, 10, 12, 14, 17, 65% acetonitrile with 10mM ammonium formate) at a flowrate of 5μL/minute was used to elute peptides for second dimension analyses with an Acclaim PepMap C18 (3μm 100Å, 0.075 x 150mm) column (Dionex). Peptides were gradient eluted at 0.2μL/minute from 5 to 25% acetonitrile, 0.1% formic acid in 100 minutes and detected with an Agilent MSD Trap XCT (3D ion trap) mass spectrometer. All spectra were searched using Mascot v2.2 (Matrix Sciences) against the International Protein Index (IPI) human database version 3.65 with a maximum of two missed cleavages and a mass tolerance of 2.0 daltons for MS1 and 0.8 daltons for MS2 spectra. Peptides were accepted above a Mascot ion score corresponding to a 1% false discovery rate (1% FDR) determined by a separate search of a reversed IPI v3.65 human database. Peptides were then filtered and protein identifications were assembled using in-house software as described[174, 201].

Purification of PIC factors for in vitro transcription: TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, Mediator, and pol II were purified as described[70].

In vitro transcription: Chromatinized templates and in vitro transcription assays were generated and completed as described[127]. Briefly, each activator (GAL4-p53AD or GAL4-VP16AD) was titrated to yield maximum transcription. While the activator bound the template, the general transcription factors (GTFs) were mixed in 0.1 M HEMG (10 mM HEPES pH 7.6, 100 mM KCl, 0.1 mM EDTA, 10% glycerol, 5.5 mM MgCl2) to give approximate final concentrations of 40 nM TFIIA, 10 nM IIB, 0.8 nM TFIID, 10 nM TFIIE, 10 nM TFIIF, 0.5 nM TFIIH and 2 nM pol II. A non-limiting amount of Mediator was then diluted in a separate salts mix (10 mM HEPES pH 7.6, 100 mM KCl, 2.5% PVA, 2.5% PEG, 7.5 mM MgCl2), along with 400 U of RNAseOUT, about 300 ng PC4 and about 300 ng HMGB1. On ice, the desired concentration of peptide was then added to the Mediator mix, followed by the GTF mix at a 5:11 ratio. The GTFs, Mediator and peptide were then incubated at least 5 minutes at 30 °C. Then, 15 μL of the mixture was added to each reaction. PIC assembly proceeded for 15 minutes, then transcription was initiated by adding 5 μL of a solution containing 5 mM of each NTP. After thirty minutes, reactions were stopped with the addition of 150 μL Stop Buffer (20 mM EDTA, 200 mM NaCl, 1% SDS, 100 μg/mL Proteinase

K, 100 μg/mL glycogen) and incubating at 37 °C for 15 minutes. RNA was isolated with 100 μL phenol/chloroform/isoamyl alcohol (pH 7.7-8.3); 140 μL of the aqueous phase was mixed with 5 μL, 7.5 M ammonium acetate and 5 μL of twenty-fold diluted, radiolabeled (32P) Reverse Transcriptase (RT) probe and transferred to a 500 μL microfuge tube. The RNA was then precipitated by adding 375 μL, 100% cold ethanol and placing at -20 °C for at least an hour.

Radiolabeling of reverse transcription primer: A reverse transcriptase (RT) primer was synthesized to complement the RNA transcript 85 bases downstream of the transcription start site. The RT primer was radiolabeled in polynucleotide kinase (PNK) buffer (70 mM Tris-HCl pH 7.6, 10 mM MgCl2, 5 mM DTT) with the addition of about 150 μCi [γ-32P]ATP, 6 U of T4 PNK and 48 ng of the RT primer in a final volume of 10 μL. The reactions were then incubated at 37 °C for 45 minutes. A glycogen mixture (10 mM Tris pH 7.5, 34 mM EDTA, 1.33 mg/mL glycogen) was then added to bring the volume to 25 μL, and the reaction was passed through a G-25 column to remove excess free [γ-32P]ATP. An additional 25 μL of TE buffer (10 mM Tris pH 7.5, 1 mM EDTA) was added. The radiolabeled primer was then stored at 4 °C until needed (up to 1 week).

Primer extension: Reactions were spun down at 14K RPM (Eppendorf 5415) for 20 minutes and the ethanol was removed. Pellets were then briefly dried (speedvac) and resuspended in 10 μL Annealing Buffer (10 mM Tris-HCl pH 7.8, 1 mM EDTA, 250 mM KCl). The resuspended RNA was then incubated in a thermocycler as follows: 85 °C for 2 minutes, cool to 58 °C at 30 sec/degree, 58 °C for 10 minutes, 57 °C for 20 minutes, 56 °C for 20 minutes, 55 °C for 10 minutes, and cool to 25 °C at 30 seconds/degree. 38 μL of RT mix (20 mM Tris-HCl pH 8.7, 10 mM MgCl2, 0.1 mg/mL actinomycin D, 330 μM of each dNTP, 5 mM DTT, 0.33 U/μL Moloney Murine Leukemia Virus (MMLV) reverse transcriptase) was then added to the annealing reactions and allowed to extend for forty-five minutes at 37 °C. Reactions were then stopped and precipitated by adding 300 μL cold ethanol and placed at -20 °C for at least an hour.

In vitro transcription analysis by denaturing polyacrylamide gel electrophoresis: The cDNA reactions were spun down at 14K RPM (Eppendorf 5415) for 25 minutes and the ethanol was removed from the pellets. After briefly drying pellets (speedvac), cDNA was resuspended

in 6 μL formamide loading buffer (75% formamide, 4 mM EDTA, 0.1 mg/mL xylene cyanol, 0.1 mg/mL bromophenol blue, 33 mM NaOH), heated for 3 minutes at 90 °C and loaded onto a denaturing polyacrylamide gel (89 mM Tris base, 89 mM boric acid, 2 mM EDTA, 7 M Urea, 6% acrylamide/bisacrylamide [19:1]). Gels were run at 35 W for about 1.5 hours, then transferred to filter paper and dried for 1 hour at 80 °C. Gels were then exposed on a phosphorimager screen.

Peptide synthesis reagents: All purchased reagents were used without further purification. Standard Fmoc-protected amino acids were purchased from Novabiochem (San Diego, CA). Fmoc-protected olefinic amino acids, (S)-N-Fmoc-2-(4'- pentenyl)alanine and (R)-N-Fmoc-2-(7'-octenyl)alanine, were purchased from Okeanos Tech Jiangsu Co., Ltd (Jiangsu, P.R. China). Rink amide resin, N,N-dimethylformamide (DMF), N-hydroxybenzotriazole (HOBt), and Grubbs Catalyst^{TM} 1st Generation were purchased from Sigma-Aldrich (St. Louis, MO). Trifluoroacetic acid (TFA) and dichloroethane (DCE) were purchased from Acros Organics (Fair Lawn, NJ). N,N,N',N'- tetramethyl-uronium-hexafluoro-phosphate (HBTU) and diisopropylethylamine (DIEA) were purchased from AmericanBio (Natick, MA). Anhydrous piperazine and 6-chlorobenzotriazole-1-yloxy-tris-pyrrolidinophosphonium hexafluorophosphate (PyClocK) was purchased from EMD Millipore (Billerica, MA). Acetic anhydride was purchased from ThermoScientific, Pierce Biotechnology (Rockford, IL).

Solid phase peptide synthesis: Peptides were synthesized using standard Fmoc chemistry with Rink amide resin on Biotage Initiatior+ Alstra from Biotage (Charlotte, NC) using microwave acceleration. Fmoc deprotections were performed using 5% piperazine with 0.1 M HOBt to reduce aspartimide formation in DMF. Coupling reactions were performed using 5 equivalents of amino acid, 4.9 equivalents of HBTU, 5 equivalents of HOBt, and 10 equivalents of DIEA in DMF at 75°C for 5 min. Fmoc-NH-(PEGn)-COOH linkers were coupled as amino acids were. All arginine residues were double coupled at 50°C. Olefinic 55 side-chain bearing residues were coupled using 3 equivalents of amino acid, 3 equivalents of PyClocK, and 6 equivalents of DIEA and stapled for 2 hours at room temperature PyClocK. Residues following olefinic residues were double coupled using standard coupling procedures. N-terminally capped peptides were generated by treating

Fmoc-deprotected resin with 100 equivalents acetic anhydride and 100 equivalents DIEA for 10 minutes at room temperature. Following synthesis, resin was washed thoroughly with alternating DMF (5 mL) and DCM (10 mL) washes before subsequent cyclizing, labeling, and cleavage.

Ring closing olefin metathesis: Peptides containing olefinic amino acids were washed with DCM (3 x 1 min) and DCE (3 x 1 min) prior to cyclizing on resin using Grubbs Catalyst I (20 mol % compared to peptide, or 1 equivalent compared to resin) in DCE (4 mL) for 2 h under N2. The cyclization step was performed twice (Kim et al., 2011). The resin was then washed three times with DCM (5 mL) before washing with MeOH (5 mL x 5 min) twice to shrink the resin. The resin was dried under a stream of nitrogen overnight.

Peptide cleavage: After shrinking and drying overnight, the peptide was cleaved from the resin using a 3 mL solution of trifluoroacetic acid (TFA) (81.5%), thioanisole (5%), phenol (5%), water (5%), ethanedithiol (EDT) (2.5%) and triisopropylsilane (TIPS) (%) for 2 hours at RT on an orbital shaker. Cleaved peptides were precipitated in diethyl ether (40 mL, chilled to -80°C), pelleted by centrifugation, washed with additional diethyl ether (40 mL, -80°C), pelleted, redissolved in a solution of acetonitrile (ACN) and water (15% CAN), frozen, lyophilized to dryness, and reconstituted in 1 mL dimethyl sulfoxide (DMSO) prior to purification by high-performance liquid chromatography (HPLC).

Peptide purification by HPLC: Peptide solutions were filtered through nylon syringe filters (0.45 μm pore size, 4 mm diameter, Thermo Fisher Scientific) prior to HPLC purification. Peptides were purified using an Agilent 1260 Infinity HPLC system on a reverse phase Triaryl-C18 column (YMC-Triaryl-C18, 150 mm x 10 mm, 5 μm, 12 nm) (YMC America, Inc.) over H2O/ACN gradients containing 0.1% TFA. Peptides were detected at 214 nm and 280 nm. Peptide purity was verified using a Shimadzu Analytical ultra-performance liquid chromatography (UPLC) system (ES Industries, West Berlin; Shimadzu Corporation, Kyoto, Japan) and a C8 reverse phase (Sonoma C8(2), 3 μm, 100 Å, 2.1 x 100 mm) analytical column. Analytical samples were eluted over a gradient of 15- 57 60% ACN in water containing 0.1% TFA over 15 min with detection at 214 and 280 nm.

In vitro binding assays: Starting from 180 μL HeLa nuclear extract (which contains Mediator), bivalent peptide was added (to 5μM concentration) followed by addition of purified p53AD (residues 1-70; to 2 μM concentration). A parallel experiment lacked added bivalent peptide. Each sample was allowed to incubate, with mixing, for 2 hours at 4°C. Each sample was then incubated, with mixing, over an anti-MED1 affinity resin (to immunoprecipitate Mediator from the sample) for 90 minutes at 4°C. The resin was then washed 4 times with 20 resin volumes with 0.5M KCl HEGN (20 mM HEPES, pH 7.9; 0.1 mM EDTA, 10% glycerol, 0.1% NP-40) and once with 0.15M KCl HEGN (0.02% NP-40). Material that remained bound to the resin (i.e. Mediator) was eluted with 1M glycine, pH 2.2 and subsequently probed by western. As an alternate protocol, HeLa nuclear extract (1 mL) was first incubated over a GST-SREBP affinity column, washed 5 times with 0.5M HEGN, once with 0.15M HEGN, and eluted with 30 mM glutathione buffer, as described[66]. This material (160 μL), which is enriched in Mediator, was then incubated with p53AD (residues 1-70; to 2 μM concentration) or GST-VP16AD (residues 411-490) in the presence or absence of bivalent peptide or its p53AD2 QS mutant (5 μM) at 4°C for 1 hour. Then each sample was incubated, with mixing, over an anti-MED1 affinity resin, washed, eluted, and probed by western as described above.

Western blotting: Protein samples were run on 7 or 9% acrylamide gels and transferred onto a nitrocellulose membrane for western blotting. Westerns were scanned on an ImageQuant LAS 4000 series imager, and ImageJ software was used to measure band intensity, which was normalized to MED15 for quantitation.

Experimental time frame for RNA-seq experiments: In a series of experiments in SJSA cells, we initially tested whether the bivalent peptide could cause a phenotypic change. SJSA cells are unusually sensitive to Nutlin-3a[229] and therefore if the p53 response could be persistently blocked by the bivalent peptide, peptide-treated cells would show enhanced survival following Nutlin treatment. Starting with a 24-hour Nutlin treatment (10 μM), we observed no significant effect of the bivalent peptide: similar percentages of cell death were observed in control vs. peptide-treated populations as analyzed by CellTiter-Glo assay (Promega). Although these

results could be attributed to poor cellular uptake of the bivalent peptide (see below), we also suspected that the peptide was active in cells for only a limited time (e.g. before being secreted or degraded). We next determined that a 6-hour Nutlin treatment time was the shortest that would still trigger significant SJSA cell death within 24-48 hours. However, we obtained similar results with 6-hour Nutlin treatment times (bivalent peptide). Note these experiments did not implement electroporation to increase peptide uptake. It remains plausible that phenotypic effects would result from methods that ensured increased bivalent peptide uptake in SJSA cells. We next tested the prospect of RNA-seq experiments, in hopes that gene expression changes and shorter time frames would allow an assessment of bivalent peptide effects. Here, we used HCT116 cells, which show strong transcriptional response to Nutlin[6]. For RNA-Seq experiments, we needed a time frame long enough to allow accumulation of p53 target gene mRNAs but short enough to enable maximum activity of the bivalent peptide (e.g. prior to its secretion, export, and/or degradation). Using RT-qPCR assays, we confirmed that a 3-hour Nutlin treatment was a minimum amount of time to reliably detect induction of p53 target genes. Parallel RT-qPCR assays confirmed that the bivalent peptide was blocking activation of p53-target genes in HCT116 cells during this time frame (e.g. Figure S4C, D3.8).

Electroporation of bivalent peptide into HCT116 cells (RNA-Seq experiment 1): Two 6-well plates (HCT116 cells) were grown to about 80% confluency. Cells were then trypsinized, washed with PBS, and resuspended in 150 μL Neon Buffer R. The cells were then split into two groups: No peptide and 10 μM peptide. The cells were then drawn into a 10 μL Neon pipet tip, electroporated and ejected into 2 mL of McCoy's 5A media without antibiotic. For each experiment, two cell electroporation aliquots were added to media containing either 0.1% DMSO (control) or 10 μM Nutlin-3a (in DMSO, to a final concentration of 0.1%). For wells containing cells electroporated with peptide, an additional 200 nM peptide was added to the well to allow for peptide uptake during the experiment. The 6-well plate was then placed back at 37 °C for 3 hours. After 3 hours, cells were scraped from the plates, transferred to a 15 mL conical vial, pelleted at 1,000 x g, and washed in 10 mL phosphate buffered saline (PBS) solution. To isolate the nuclei,

cells were resuspended in 10 mL lysis buffer (10 mM Tris-HCl, pH 7.4, 2 mM MgCl2, 3 mM CaCl2, 0.5% NP-40, 10% glycerol) and thoroughly mixed. The nuclei were spun down at 1,000xg for 10 minutes, the lysis buffer was removed, and 1 mL of TRIzol was added. The nuclear RNA was isolated as described in the TRIzol instructions, except an additional phenol/chloroform extraction and chloroform-only extraction were performed to reduce contaminants. RNA was precipitated and washed twice with 75% ethanol to further remove contaminants. The RNA was then converted to cDNA using the High Capacity cDNA kit from Thermo Fisher Scientific.

Electroporation of bivalent peptide into HCT116 cells (RNA-Seq experiment 2): One 15cm plate (HCT116 cells) was grown to about 70% confluency. Cells were then trypsinized, washed with PBS, and resuspended in 40 μL Neon Buffer R. The cells were then split into two groups: No peptide and 10 μM peptide. The cells were then drawn into a 10 μL Neon pipet tip, electroporated and ejected into 2 mL of serum-free McCoy's 5A media without antibiotic. Serum-free media was tested based upon prior reports that it may enhance cellular peptide uptake[39, 46]. For each experiment, two cell electroporation aliquots were added to media containing 0.1% DMSO (control) or 10 μM Nutlin-3a (in DMSO, to a final concentration of 0.1%). For wells containing cells electroporated with peptide, an additional 200 nM peptide was added to the well to allow for peptide uptake during the experiment. The 6-well plate was then placed back at 37 °C for 3 hours. After 3 hours, cells were scraped from the plate, transferred to 2 mL eppendorf tubes, pelleted at 1,000 x g, and washed in 1 mL cold phosphate buffered saline (PBS) solution. To isolate the nuclei, cells were resuspended in 0.5 mL lysis buffer (10 mM Tris-HCl, pH 7.4, 2 mM MgCl2, 3 mM CaCl2, 0.5% NP-40, 10% glycerol) and mixed by pipetting up and down 20 times. The nuclei were spun down at 1,000 x g for 10 minutes, the lysis buffer was removed, and 200 μL of TRIzol was added. The RNA was isolated as described in the TRIzol instructions, except an additional phenol/chloroform extraction and chloroform-only extraction were performed. RNA was precipitated and washed twice with 75% ethanol. The RNA was then converted to cDNA using the High Capacity cDNA kit from Thermo Fisher Scientific.

RNA-seq data analysis: The RNA-seq data were mapped to hg38 and processed using a Nextflow pipeline v1.1 (https://github.com/Dowell-Lab/RNAseq-Flow). Batch correction for experiment 2 was performed in R using the removeBatchEffect function provided by the limma package[205] from the R programming language. Gene counts were generated using featureCounts[147] and differential gene expression analysis was performed using DESeq2 [8]. Duplicate transcripts were filtered for those with the highest RPKM, leaving 28,260 transcripts. RNA-seq was performed on nuclear RNA instead of total RNA to better assess acute transcriptional changes; furthermore, analyses were completed at 3h post-Nutlin instead of longer time points, to reduce indirect/secondary effects from p53 activation. Because of this strategy, fewer differential mRNA products were expected (e.g. vs. total RNA analysis at 12h post-Nutlin). Nevertheless, numerous p53 target genes showed differential expression with statistical confidence. A prerank file was generated in R using the results from the differential analysis results and used in the Broad Institute's Gene Set Enrichment Analysis (GSEA 4.0.3) software, using hallmark pathways gene sets (hall.v7.4)[221]. Heatmaps were generated in Python using seaborn 0.9.0. Other plots (bar plots, box plots, volcano, moustache) were made with the python package bokeh 1.4.0. Gene traces were made using pyGenomeTracks 3.5, part of the deeptools suite. Note that for comparisons between peptide-treated vs. untreated cells in the absence of Nutlin stimulation (e.g. Figure 3E3.3), MTM1 met our significance threshold; however, MT1M was highly expressed in only one replicate of the DMSO control (R1 RPKM = 8.61; R2 RPKM = 0.63) and MT1M expression in other conditions (Nutlin, DMSO + peptide, or Nutlin + peptide) was less than 1.0 across all replicates, suggesting a sampling artifact rather than a true biological difference.

RT-qPCR: Experiments were performed as described[16]; primers used are shown in Table S6.

ChIP-seq: HCT116 cells were grown to about 80% confluency in one 15cm plate. Cells were then trypsinized, washed with PBS, and resuspended in 250 μL Neon Buffer R to a concentration of 2.2 million cells per 100 μL. The cells were then split into either a no peptide or a peptide group. The peptide group had 10 μM peptide during electroporation and 200 nM peptide on the

plate during the 3hr incubation. The cells were drawn into a 100 μL Neon pipet tip, electroporated and ejected into 700 μL of McCoy's 5A media without antibiotic. This sample was then split, 400 μL (containing 1 million cells) each into 5mL of antibiotic-free media with 0.1% DMSO or with 10 μM Nutlin-3a. The plates containing the four conditions were placed at 37°C for 3 hours. After this incubation, the cells were fixed with 1% formaldehyde for 10 min and quenched with 125mM glycine for 5 min before nuclei isolation in NRO buffer. All buffers were as described[142] unless otherwise noted. Shearing was performed based on the Covaris truChIP Chromatin Shearing Kit (Covaris: PN 520237) and sheared in 1 mL of Covaris D3 buffer for 8 min, with a duty factor 5% and peak power of 75,200 cycles/burst with a Covaris M220 Focused-ultrasonicator. Protein G Dynabeads (6 μg antibody 25 μL beads per condition. Beads: Invitrogen, #10003D. Antibody: pol II Ser5P clone 3E8 Millipore Sigma, 04-1572) were loaded with the crosslinked chromatin and allowed to incubate with shaking at 4°C overnight (18hr) in a final buffer composition of 15 mM Tris-HCl pH 8, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 5% Glycerol, 0.1% Sodium Deoxycholate, 0.1% SDS and the protease inhibitor cocktail. Note that the buffer composition and antibody:bead ratio was optimized empirically for this assay, to maximize sample recovery. The beads were then washed at 4°C with 2x IP Buffer (15 mM Tris-HCl pH 8, 150 mM NaCl, 1 mM EDTA, 5% Glycerol, 0.1% Sodium Deoxycholate, 1% Triton X-100 and the protease inhibitor cocktail), 2x RIPA Buffer, 2x LiCl Wash Buffer and 2x TE Salt Buffer. Samples were kept on ice throughout the wash steps, and all washes were completed in less than 30 minutes; each wash step was 1 min to reduce sample loss. After washes, the sample was eluted from beads with PK buffer containing 100 μg of Proteinase K (New England BioLabs: P81075) and incubated at 65°for 2 hr with periodic vortexing. The eluted sample was removed from the magnetic beads and reverse-crosslinked for 16 hr at 65°C. DNA was purified by phenol chloroform extraction using Light-5PRIME Phase Lock Tubes (Quanta Bio: 2302820) based on the manufacturer's instructions. To increase yield, samples were ethanol precipitated from glycogen (20μg/sample), sodium acetate (pH 5.2) and 10mM MgCl2 and kept at -20°overnight. Isolated DNA was tested by qPCR at the CDKN1A promoter (Forward 5'-CCAGGAAGGGCGAGGAAA, Reverse 5'- GGGACCGATCCTAGACGAACTT) and the BTG2

promoter (Forward 5'- AGGGTAACGCTGTCTTGTGG, Reverse 5'- CAGGAGGCTGGAGAG-GAAG). Subsequently, libraries with approximately 1 ng of input DNA were prepared using the KAPA Hyper Prep Kit (KAPA Biosystems: KK8502) and sequenced on Illumina NextSeq V2 high output 75-cycle. Adapters were diluted 1:10 before use and amplified for 12-14 cycles before sequencing.

ChIP-seq data analysis: The ChIP-seq data were mapped to hg38 and processed using Nextflow pipeline v1.0 (https://github.com/Dowell-Lab/ChIP-Flow). QC metrics showed that biological replicate 2 was higher quality. The Nutlin + peptide replicate 1 showed low complexity, high background and high duplication; despite this, we still observed similar trends in Nutlin response consistent with replicate 2. Due to these complexity and background issues, replicate 1 in the DMSO + peptide and Nutlin + peptide were not used in subsequent analyses. Due to the nature of pol II Ser5P ChIP-seq (i.e. Ser5P is enriched at gene 5'-ends), we focused on transcriptional start sites (TSS) and counted over a 1000bp window at all RefSeq TSSs (NCBI Reference Sequences for hg38 downloaded from the UCSC track browser on May 18, 2018; n=82,500 annotated transcripts). All counting was performed with bedtools multicov (v 2.28.0) and only the highest RPKM annotated start site per transcript was retained (n= 28,260). The data were then normalized using a Signal to Noise ratio (SNR) equivalent to the CHIPIN quantile method as described[193]. To perform this method, we estimated signal per condition by summing the top 75% expressing TSSs, excluding 1) p53 target genes and 2) any differentially expressed genes (padj ¡ 0.05) in any of the 6 pairwise comparisons in RNA-seq experiment 1 and RNA-seq experiment 2 (excluded n= 2,820 transcripts, approximately n 18,000 transcripts remained for normalization). To estimate noise, we randomly shuffled 82,500 1000bp regions with bedtools shuffle and counted using bedtools multicov (v 2.28.0). These counts were then filtered twice by dropping all zero-read regions and by keeping only the 10th-90th quantile of reads. We did this to remove 1) any lowly expressed and highly variable regions and 2) to remove any regions that may overlap with a gene body. This left approximately 50,000 regions to estimate the noise per condition. The average signal per condition (reads per region assessed) was then divided by the average noise per condition (reads per region

assessed). This gave us a SNR factor that was applied to bedgraph files for 1) bigwig conversion using UCSC kentUtils for visualization or 2) count tables for quantification. ChIP figures were generated with the same software outlined in the RNA-seq analysis section, seaborn 0.9.0, bokeh 1.4.0 and pyGenomeTracks 3.5.

Quantification and statistical analysis: RNA-seq experiments were completed in biological replicate (experiment 1) or biological triplicate (experiment 2). Statistical analysis of RNA-seq data is described in the Method Details. ChIP-seq experiments were completed in biological replicate; however, single replicates for DMSO + peptide and Nutlin + peptide did not pass quality control (see Methods). The number of replicates for each in vitro transcription experiment is indicated in the Figure Legends. Statistical analysis of in vitro transcription data is provided in Figure Legends and Method Details.

B.3 Δ40p53 Methods

As published in Levandowski et. al. [142].

Cas9 protein purification: Cas9 purification was completed as described[150]

Cas9-RNP formation: sgRNA was formed by adding tracrRNA (IDT cat# 1072533) and crRNA (TP53 exon 2, positive strand, AGG PAM site, sequence: GATCCACTCACAGTTTCCAT) in a 1:1 ratio and heated to 95°then slowly cooled to room temperature over 1 hour. Purified Cas9 was added to sgRNA at a ratio of 1:1.2 and incubated at 37°C for 15 min, forming Cas-9 RNP. Cas9 RNP was used at 10 μM concentration within an hour.

Donor plasmid construction: Vector Builder was used to construct the plasmid. The insert (see 4.5A Fig) was flanked by 1.5kb homology arms, and mCherry was inserted as a selection marker. Insertion sizes were as follows: WTp53: 2820bp, WTp53:WTp53: 4041bp, and Δ40p53:WTp53: 3924bp.

CRISPR-Cas9 genome editing: MCF10A cells were split 24 hours prior to each experiment and grown to approximately 70% confluence on a 15cm plate. Cells were washed with PBS, followed by trypsinization (4ml per plate) and resuspended in re-suspension media (8 mL

DMEM/F12 containing 20% horse serum and 1x pen/strep). 5x105 cells were placed in 1.5mL eppendorf tubes for transfection with neon transfection system (Invitrogen, #MPK5000). Cells were re-suspended in resuspension Buffer R (Invitrogen, #MPK1025) with 10 μM Cas9-RNP and 1 μg donor plasmid (WTp53, WTp53:WTp53, or Δ40p53:WTp53). 10 μL Neon pipet tip (Invitrogen, #MPK1025), electroporated using the Neon Transfection Kit (1400V, 20 ms width, 2 pulses). Transfected cells were transferred to 2 mL antibiotic free media. Cut location: hg38 chr17:7,676,510. Insertion location: hg38 chr17:7,676,591; after the first ATG for TP53 in exon 2. Cells were single cell sorted into 96 well plates based on mCherry expression. Clones were then verified with sequencing, PCR, and western blot.

MCF10A Cell Culture: MCF10a cells cultured in DMEM/F12 (Invitrogen #11330-032) media containing 5% horse serum (LifeTech #16050-122), 20ng/mL epidermal growth factor EGF (Peprotech #AF-100-15), 0.5ug/mL Hydrocortisone (Sigma #H0888-1g), 100ng/mL Cholera toxin (Sigma #C8052-2mg), 10ug/mL insulin (Sigma #I1882-200mg), and 1x Gibco 100x Antibiotic-Antimycotic (Fisher Sci, 15240062) penicillin-streptomycin.

PCR Verification of insertions at TP53 locus: NEB Phusion polymerase (NEB, #M0530S) was used to the manufacturer's specifications. SeqTP53exon2Forward and SeqTP53exon2Reverse primers were used at 65-68°C; homozygous knock-in clones had no 500bp product. Product was sequenced using TP53exon2-Forward primer. Using DBDForward and SeqTP53exon2Reverse primers at 67.8°C, a 2542bp band indicated knock-in for at least one allele (either homozygous or heterozygous clone). A homozygous clone was then verified if SeqTP53exon2Forward and SeqTP53exon2Reverse primers at 65-68°C had no 500bp product. Because mutations in p53 can yield a survival advantage[171], we periodically re-sequenced the TP53 locus. Primers used were CL802 & CL803 at 65°C, and DBDForward and DBDReverse were used for sequencing. For sequencing downstream of insert, we used DBDForward and SeqTP53exon2Reverse at 70°C. For amplifying upstream of insert, we used CRISPR3 seq primer and CL803 at 70°C, and DBDReverse and SeqTP53exon2Forward primers were used for sequencing. Primer sequences are listed in Supplement Table 3.

Compound Treatments: Nutlin-3a (Selleck, #S8059) stock concentration 10mM, and treatment concentration 10μM). 5FU (Selleck, #S1209) stock concentration 100mM, and treatment concentration 375μM. Equivalent volumes of DMSO vehicle were used for controls.

Metabolomics: Cells were harvested after 20 hr treatment with Nutlin-3a (10 μM) or 0.1% DMSO controls, with six biological replicates for each cell line and each condition. Sample preparation was carried out at Metabolon Inc. (Durham, NC), in a manner similar to a previous study[49]. Briefly, individual samples were subjected to methanol extraction then split into aliquots for analysis by ultrahigh-performance liquid chromatography/mass spectrometry (UHPLC/MS). The global biochemical profiling analysis comprised four unique arms consisting of reverse phase chromatography, positive ionization methods optimized for hydrophilic compounds (LC/MS Pos Polar) and hydrophobic compounds (LC/MS Pos Lipid), reverse phase chromatography with negative ionization conditions (LC/MS Neg), as well as a HILIC chromatography method coupled to negative ionization (LC/MS Polar)[69]. All of the methods alternated between full scan MS and data dependent MSn scans. The scan range varied slightly between methods but generally covered 70–1000 m/z. Metabolites were identified by automated comparison of the ion features in the experimental samples to a reference library of chemical standard entries that included retention time, molecular weight (m/z), preferred adducts, and in-source fragments as well as associated MS spectra, and curated by visual inspection for quality control using software developed at Metabolon, Inc. Identification of known chemical entities was based on comparison to metabolomic library entries of purified standards[60]. A summary of all metabolomics data is shown in Supplemental Table 4.

Statistical Analysis of metabolomics data: Two types of statistical analyses were performed: (1) significance tests and (2) classification analysis. Standard statistical analyses were performed in ArrayStudio on log-transformed data. For analyses not standard in ArrayStudio, the R program (http://cran.r-project.org/) was used. Following log transformation and imputation of missing values, if any, with the minimum observed value for each compound, Welch's two sample t-Test was used as a significance test to identify biochemicals that differed significantly (p ¡ 0.05)

between experimental groups. An estimate of the false discovery rate (q-value) was calculated to take into account the multiple comparisons that normally occur in metabolomic-based studies. Classification analyses included principal components analysis (PCA), hierarchical clustering, and random forest. For the scaled intensity graphics, each biochemical in original scale (raw area count) was rescaled to set the median across all samples and time-points equal to 1.

Cell cycle analysis: WTp53, WTp53:WTp53, or Δ40p53:WTp53 cells were treated with 10μM Nutlin-3a or 375μM 5FU for 20 hours in parallel with DMSO controls, 0.1% and 0.375% respectively. Propidium Iodide (PI) Flow Cytometry Kit (abcam, ab139418) was used as specified by manufacturer. Samples were then placed on ice and analyzed with a BS FACSCellesta cell analyzer. FLOWJO was used to analyze FACS data. All experiments were performed in biological triplicate.

Immunoflourescence: WTp53, WTp53:WTp53, and Δ40p53:WTp53 were treated with 10μM Nutlin-3a or 0.1% DMSO for 20 hours, collected, and fixed in 66% ethanol for at least 24 hr. Cells were then incubated with a blocking/permeabilization buffer (3% BSA, 0.1% Triton X-100) for 1 hr at room temperature. Primary antibody, anti-p21 (CST, 2947) at 1:250 dilution, staining was carried out overnight at 4 °C in the blocking buffer and visualized using secondary antibodies conjugated to Alexa Fluor 488.

Growth Rate Calculations: To examine how growth rate change over time, cells were treated with 0.1% DMSO or 10μM Nutlin-3a for 20 hours, passaged 1:10, and given 48 hr to recover (treatment cycle; 4.8A Fig). Growth rate was calculated by dividing the total cell growth by time of growth (68 hours). Total cell growth was calculated by counting with Nexcelom Bioscience Cellometter Auto T4 Bright Field Cell Counter. All experiments were performed in biological triplicate.

p63 lentiviral knockdown: WTp53, WTp53:WTp53, and Δ40p53:WTp53 cells were transduced with viral construct targeting p63 (shRNA TRCN0000006506; obtained from University of Colorado Functional Genomics Core) under a constitutive hU6 promoter with puromycin resistance. Knockdown cells were selected with 48hr treatment of 2μg/mL puromycin followed by

recovery in standard growth media.

RT-qPCR: WTp53, WTp53:WTp53, Δ40p53:WTp53 cells p63 knockdown, or alternate p53 cell clones were treated with 10μM Nutlin-3a or DMSO control for 20hr before RNA isolation using TRizol (Invitrogen, #15596026) as specified by manufacturer. Total RNA was converted to cDNA using High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems, 4368813) as specified by manufacturer, followed by a clean-up performed using 0.8X AMPure XP beads (Beckman). Sybr Select Master Mix (Thermo, 4472908) was added to cDNA at 0.1ng/μL and amplified per manufacturer instructions. ΔΔCT values were background normalized using a gene desert. PCR primers are listed in Supplemental Table 3.

RNA-seq: Total RNA was isolated from WTp53, WTp53:WTp53, and Δ40p53:WTp53 cells treated with 10μM Nutlin-3a, 375μM 5FU, or DMSO controls (0.1% and 0.375% respectively) using TRizol (Invitrogen, #15596026) as specified by manufacturer. 1μg of total RNA with a RIN number of ≥ 8 was used for RNA-seq library prep. Sample was enriched for mRNA using NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490) and library was prepared using NEBNext Ultra II Directional RNA Library Prep Kit from Illumina (E7765).

ChIP-seq: MCF10A cell lines (WTp53, WTp53:WTp53, and Δ40p53:WTp53; 60 x 106 cells per experiment) were treated with 10 μM Nutlin-3a or 0.1% DMSO for 6 hours then crosslinked with 1% formaldehyde for 10 min at 25°followed by glycine (0.125 M) quenching for 5 min. Nuclei were isolated by resuspending cells in NRO buffer (80 μl/million cells; 10 mM Tris-HCl [pH 8], 4 mM MgCl2, 10 mM NaCl, 0.5% [vol/vol] NP-40, 1mM DTT, and the protease inhibitor cocktail (1mM Benzamidine (Sigma, #B6506-100G), 1mM Sodium Metabisulfite (Sigma, #255556-100G), 0.25mM Phenylmethylsulfonyl Fluoride (American Bioanalytical, #AB01620) 0.012 TIU/mL aprotinin (Sigma, #A6106), followed by a 5 min incubation on ice, a low speed spin, and then a final wash with NRO buffer. The isolated nuclei were prepared for shearing based on the Covaris truChIP Chromatin Shearing Kit (Covaris: PN 520237) and sheared for 11 min with Covaris M220 Focused-ultrasonicator. 25μL of Protein G Dynabeads beads (Invitrogen, #10003D) was used per 100μg chromatin (15 million nuclei). Beads were incubated in blocking solution (PBS, 0.5% BSA) then 6

μg of DO1 p53 antibody (BD Biosciences: BD554293) was conjugated to beads in blocking solution, nutating at 4°for 4 hours. Conjugated beads were washed 1x block solution and 1x IP buffer (15 mM Tris-HCl [pH 8], 150 mM NaCl, 1 mM EDTA, 1% Triton X-100 and the protease inhibitor cocktail). Sheared chromatin was added (100μg chromatin (15 million nuclei)) to conjugated beads in IP buffer and then nutating at 4°for at least 12 hours. Bound chromatin was washed 3x with IP buffer, 3x with RIPA buffer (20 mM Tris-HCl [pH 8], 500 mM NaCl, 1 mM EDTA, 1% Triton X-100 and 0.1% SDS), 2x with LiCl buffer (20 mM Tris-HCl [pH 8], 500 mM LiCl, 1 mM EDTA, 1% sodium deoxycholate and 1% NP-40) and 2x with TE Salt buffer (10 mM Tris-HCl [pH 8], 50 mM NaCl, 1 mM EDTA). Sample was eluted from beads with PK buffer (10 mM Tris-HCl [pH 8], 1 mM EDTA and 1% SDS) containing 100 μg of Proteinase K (New England BioLabs: P81075) incubated at 50oC, with shaking, for 1-hour then at 65°for 1-hour. Eluted DNA was transferred to a new tube and incubated at 65°for 12 hours to reverse crosslinking. DNA was purified by phenol chloroform extraction using Light-5PRIME Phase Lock Tubes (Quanta Bio: 2302820) based on the manufacturer's instructions. Libraries were prepared using the KAPA Hyper Prep Kit (KAPA Biosystems: KK8502) and sequenced on Illumina NextSeq V2 high output 75-cycle.

PRO-seq nuclei preparation: WTp53, WTp53:WTp53, Δ40p53:WTp53, or p53-null MCF10A cells were seeded on three 15cm dishes (1x107 cells per dish) for each treatment 24 hr prior to the experiments (70% confluency at the time of the experiment). Cells were treated simultaneously with 10μM Nutlin-3a or 0.1% DMSO for 3 hours. After treatment, cells were washed 3x with ice cold PBS, and then treated with 10 ml (per 15 cm plate) ice-cold lysis buffer (10 mM Tris–HCl pH 7.4, 2 mM MgCl2, 3 mM CaCl2, 0.5% NP-40, 10% glycerol, 1 mM DTT, 1x Protease Inhibitors (1mM Benzamidine (Sigma B6506), 1mM Sodium Metabisulfite (Sigma 255556), 0.25mM Phenylmethylsulfonyl Fluoride (American Bioanalytical AB01620), and 4U/mL SUPERase-In). Cells were centrifuged with a fixed-angle rotor at 1000g for 15 min at 4°C. Supernatant was removed and pellet was resuspended in 1.5 mL lysis buffer to a homogenous mixture by pipetting 20-30X before adding another 8.5 mL lysis buffer. Suspension was centrifuged with a fixed-angle rotor at 1000g for 15 min at 4°C. Supernatant was removed and pellet was resuspended in 1 mL of lysis

buffer and transferred to a 1.7 mL pre-lubricated tube (Costar cat. No. 3207). Suspensions were then pelleted in a microcentrifuge at 1000g for 5 min at 4°C. Next, supernatant was removed and pellets were resuspended in 500 μL of freezing buffer (50 mM Tris pH 8.3, 40% glycerol, 5 mM MgCl2, 0.1 mM EDTA, 4U/ml SUPERase-In). Nuclei were centrifuged 2000g for 2 min at 4°C. Pellets were resuspended in 100 μL freezing buffer. To determine concentration, nuclei were counted from 1 μL of suspension and freezing buffer was added to generate 100 μL aliquots of $1e^7$ nuclei. Aliquots were flash frozen in liquid nitrogen and stored at -80°C.

PRO-seq Nuclear run-on and RNA preparation: Nuclear run-on experiments were performed with biological duplicates as described[160] with the following modifications: the final concentration of non-biotinylated CTP was raised from 0.25 μM to 25 μM, a clean-up and size selection was performed using 1X AMPure XP beads (1:1 ratio) (Beckman) prior to test PCR and final PCR amplification, and the final library clean-up and size selection was accomplished using 1X AMPure XP beads (1:1 ratio) (Beckman).

Sequencing: Sequencing of RNA-seq, ChIP-seq, and PRO-Seq libraries was performed at the BioFrontiers Sequencing Facility (UC-Boulder). Single-end fragment libraries (75 bp) were sequenced on the Illumina NextSeq 500 platform (RTA version: 2.4.11, Instrument ID: NB501447), demultiplexed and converted BCL to fastq format using bcl2fastq (bcl2fastq v2.20.0.422); sequencing data quality was assessed using FASTQC (v0.11.5) (`https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`) and FastQ Screen (v0.11.0, `https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/`). Trimming and filtering of low-quality reads was performed using BBDUK from BBTools (v37.99) (https://www.osti.gov/servlets/purl/1241166) and FASTQ-MCF from EAUtils (v1.05)[15]. Alignment to the human reference genome (hg38) was carried out using Hisat2 (v2.1.0)[124] in unpaired, no-spliced-alignment mode with an hg38 index, and alignments were sorted and filtered for mapping quality (MAPQ¿10) using Samtools (v1.5)[145].

RNA-seq Computational Analysis: The RNA-seq data was processed using a Nextflow pipeline v1.1 (https://github.com/Dowell-Lab/RNAseq-Flow). A full pipeline report of the run as well as a quality control report generated by MultiQC (v. 1.7), including trimming, mapping,

coverage, splicing, and complexity metrics are included in Supplemental Table 5. Gene counts were generated using featureCounts[147] and differential gene expression analysis was performed using DESeq2[8]. Duplicate genes were filtered and those with the highest FPKM were kept for analysis. Qiagen Ingenuity Pathway Analysis (IPA) and Gene Set Enrichment Analysis (GSEA 4.03)[221] were used for identification of activated and inhibited pathways, and upstream regulators based on expression changes.

ChIP-seq Computational Analysis: All ChIP-seq data was processed (mapped and quality checked) using a Nextflow pipeline, ChIP-Flow v1.0 (https://github.com/Dowell-Lab/ChIP-Flow). A full pipeline report of the run as well as a quality control report generated by MultiQC (v. 1.7), including trimming, mapping, coverage, and complexity metrics are included in Supplemental Table 6. Peak calls were generated using MACS2 narrowPeak. The q-value default cutoff was also decreased from the default of 0.05 to 1e-5. Blacklisted regions (those having artificially high signal and read mapping, http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/hg38-human) were removed using BEDTools intersect[198]. PyGenomeTracks [200] was used to generate track images.

Data processing, visualization and identification of eRNAs: Two biological replicates for each treatment (DMSO/Nutlin-3a) were generated for each of the three cell lines were processed using a Nextflow pipeline for nascent data (https://doi.org/10.5281/zenodo.2641755). A full pipeline report of the run as well as a quality control report generated by MultiQC (v. 1.7), including trimming, mapping, coverage, and complexity metrics, is included in Supplemental Table 7. PyGenomeTracks [200] was used to generate track images. Tfit was used to identify regions with bidirectional transcription[18]. Note that in Nutlin-treated Δ40p53:WTp53 cells, 86 eRNAs were identified with a p-value ≤ 0.01; after removal of all repetitive regions, 25 eRNAs remained. Transcription Factor Enrichment Analysis (TFEA) was used to identify changes in bidirectional transcription and map to underlying TF sequence motifs to infer changes in TF activity[207].

Differential transcription analysis of genes and bidirectionals/enhancers (PRO-seq): Using the RefSeq: NCBI Reference Sequences for hg38, including both NM and NR acces-

sion types (downloaded from the UCSC track browser on May 18, 2018), counts were calculated for each sorted BAM file using multiBamCov in the BEDTools suite (v. 2.25.0). Genes (NM accession type) and lncRNAs (NR accession type) were then filtered such that only the isoform with the highest number of reads per annotated length was kept in order to minimize duplicate samples being included in differential transcription analysis. DESeq2 (v. 1.20.0, Bioconductor release v. 3.7) was then used to determine differentially transcribed genes between the different treatments both within and between timepoints. A prerank file was generated using the results from the differential analysis from each pairwise comparison and used in differential pathway analysis using GSEA 4.0.3, using hallmark pathways gene sets. Qiagen Ingenuity Pathway Analysis (IPA, v7.2) was used for identification of activated and inhibited pathways based on transcriptional changes. For bidirectional/enhancer comparisons, all bidirectional prediction Tfit calls were merged using mumerge software (merge component of TFEA) to generate an annotation file. Counts were then calculated for each sample using multicov from the BEDTools suite (v. 2.28.0)[198] and DESeq2 [8] was used to calculate differentially transcribed bidirectionals. Transcription factor enrichment analysis (TFEA) was used to assess p53 activation based on eRNA expression[207].

Additional cell line validation: Cell lines were internally validated by mapping PRO-seq reads to the p53 construct as a mini "genome" using Hisat2 (v2.1.0)[124] and alignments were sorted using Samtools (v1.5)[145]. Counts were then calculated for each sample using multicov from the BEDTools suite (v. 2.28.0)[198] and regions were compared for their read density over TAD1 against to the central region (TAD2, DBD and C-terminal region) of p53.

Quantification and statistical analysis: PRO-seq and RNA-seq experiments were completed in biological replicate. ChIP-seq experiments were completed with biological triplicates. Metabolomics experiments were completed with six biological replicates. Statistical analysis of sequencing data and metabolomics data is provided in Method Details.

B.4 OV90 CDK7i Methods

OV90 cell culture and treatment details: Roughly 7 million early-passage wild-type OV90 cells were grown in MCDB 105 media 1:1 with Medium 199 and supplemented with 15% FBS and 1% Pen-strep per ATCC guidelines. Cells were then treated in T-75 flasks with either 50nM SY5609, 50nM SY5102, 2.5uM 3MB-PP1 (OV90 CDK7as cells) or equivalent DMSO (0.005%) for 30min at 37°C and were either harvested or heat shocked at 42°C for a set time described in Figure A.1. Some experiments involved a recovery for 1hr at 37°C[45].

RNA-seq library preparation and sequencing: After conclusion of the treatment scheme, cells were harvested by washing with cold D-PBS, scraped in D-PBS into 15mL conical tubes, and centrifuged at 1000rpm for 5min 4°C. RNA was extracted with TRIzol according to the manufacturer's instructions. DNaseI digestion was then performed with the RNeasy kit (Qiagen) using the on-column protocol[203]. RNA was eluted in 50L RNase-free water and samples analyzed by Qubit and TapeStation. Samples were concentrated, 300-400ng/μL, and of high quality, RIN values ¿8.6. For deep mRNA sequencing to asses splicing alterations, poly(A) selection was performed with the KAPA mRNA HyperPrep Kit with 1μg of RNA containing 1:100 diluted ERCC RNA spike-in mix (Invitrogen Thermo Fisher). Samples were then sequenced on the NovaSEQ6000, using a paired-end 150bp cycle (2x150)[203].

RNA-seq computational analysis: The RNA-seq data was processed using a Nextflow pipeline (https://github.com/Dowell-Lab/RNAseq-Flow). Gene counts were generated using featureCounts[147] and differential gene expression analysis was performed using DESeq2[8, 155]. Duplicate coding genes were filtered first by transcriptional start site isoforms identified by PRO-seq then by the highest FPKM of remaining possible isoforms. A prerank file was generated using the results from the differential analysis from each pairwise comparison and used in differential pathway analysis using GSEA 4.0.3[221], using go biological processes pathways gene set.

PRO-seq nuclei preparation and library preparation: PRO-seq libraries were prepared as described in Levandowski et. al.[142]. Briefly, after treatment, cells were washed with ice

cold PBS, and then treated with ice-cold lysis buffer. The nuclei were isolated and resuspended in 500 μL of freezing buffer. Nuclei were centrifuged 2000g for 2 min at 4°C. Pellets were resuspended in 100 μL freezing buffer. To determine concentration, nuclei were counted from 1 μL of suspension and freezing buffer was added to generate 100 μL aliquots of $1e^7$ nuclei. A single aliquot of nuclei per condition was used to perform a nuclear run-on as described in Mahat et. al. 2016, with the modifications discussed in Levandowski et. al. 2021[142, 160]. Sequencing was performed at University of Nevada Reno sequencing core on the Illumina NextSeq 2000, using a single-end 80bp cycle.

PRO-seq computational analysis: The PRO-seq data was processed using a Nextflow pipeline (`https://doi.org/10.17605/OSF.IO/NDHJ2`). Gene-centric analyses were performed by counting over refseq h38 regions using bedtools multicov in the BEDTools suite (v. 2.25.0). All count data was normalized using the virtual spike-in method[158]. A single isoform for each coding transcript was selected by first the highest transcriptional start site signal, then by highest FPKM over the gene body. Differential expression results were generated by DESeq2[8, 155]. A prerank file was generated using the DESeq2 results from each pairwise comparison and used in GSEA 4.0.3[221], using go biological processes pathways gene set.

Tfit was used to identify regions with bidirectional transcription[18]. These regions were then merged using muMerge to generate a single bidirectional master file (Fig. A.13)[207]. The master file was used to count coverage over all bidirectionals using bedtools multicov and was normalized by the virtual spike-in[158]. These regions were then assessed for differential expression using DESeq2[8]. The differential expression results were used to build the ranked lists for TFEA input, allowing for the virtual spike-in to be accounted for in transcription factor enrichment analyses. Initial runs for TFEA were performed to extract non-corrected E-scores. All non-corrected E-scores for all relevant comparisons in this study (n=9, DMSO heat shock time course, SY5609 heat shock time course and SY5609 versus DMSO across all heat shock) were corrected in one linear fit, equally across all conditions (Fig. A.18). TF Profiling was performed using annotated peaks defined by the muMerge master file, filtered for only regions that contained signal in DMSO 0min heat shock

replicate 1 or replicate 2. PyGenomeTracks [200] was used to generate track images.