Investigating exogenous stressors impact on transcription in population with Down syndrome

by

Jessica Vy Thùy Huynh-Westfall

B.A., MCD Biology and Psychology, University of California Santa Cruz

M.A., Biological Science, San Jose State University

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 Molecular, Cellular, Developmental Biology

2023

Committee Members:

Dylan Taatjes, Ph.D., Chair

Robin D. Dowell, D.Sc.

Justin Brumbaugh, Ph.D.

Edward Chuong, Ph.D.

Mary A. Allen, Ph.D.

Huynh-Westfall, Jessica Vy Thùy (Ph.D., Molecular, Cellular, Developmental Biology)

Investigating exogenous stressors impact on transcription in population with Down syndrome

Thesis directed by Robin D. Dowell, D.Sc.

The cell's ability to regulate gene transcription in response to external stimuli is crucial for proper cell function. Throughout this thesis, I will delve into the intricacies of cellular responses to external stimuli, specifically focusing on Trisomy 21 (T21) cells. Down syndrome, the most prevalent human autosomal aneuploidy, is caused by triplicate copies of chromosome 21. The first half of this thesis explores the effects of a type I interferon (IFN-β) on T21 and euploid disomic (D21) cells, focusing on both immediate-early transcriptional shifts and subsequent gene expression changes. Though Down syndrome has been linked to heightened interferon activity arising from the extra interferon receptors on chromosome 21, my research suggests that an individual's genetic makeup plays a more decisive role in the earliest responses to IFN-β than the trisomy itself. Next, I explore the heat shock response in T21 cells, a response pathway not explicitly tied to chromosome 21. Given the enhanced inflammatory response typical of Down syndrome, we hypothesized an amplified heat shock response in T21 cells. Our data shows a marginally enhanced heat shock response in Down syndrome, pointing to a broadened stress response mechanism. Through these investigations, we provide a deeper understanding of the impact of T21 on gene regulation. In the appendices, I also describe other efforts undertaken including developing a classifier for offset patterns within motif displacement distributions and efforts on building a reporter construct for transcription factor activity.

Dedication

To all who strive to push science forward with intelligence, passion, and empathy for living beings. To my children, Mai and Violet, though we were miles apart, your love and support illuminate the path of persistence.

Acknowledgements

The research detailed in this thesis would not have been achievable without the invaluable assistance and insight of numerous talented scientists. I am deeply grateful to my mentors, Drs. Robin Dowell and Mary Ann Allen, for their unwavering guidance throughout my academic journey. Robin's profound expertise in genomics and bioinformatics has significantly honed my critical thinking, pushing me to be a meticulous scientist. Her patience and perseverance during the challenges of my projects were instrumental in steering me towards fresh perspectives and areas of exploration. Mary's infectious passion, enthusiasm, and computational prowess were invaluable as we delved deep into data and statistical analysis.

My heartfelt thanks go to the entire DnA lab team, both past and present members, for their consistent assistance and companionship, particularly during trying times. I owe a debt of gratitude to Gilson Sanchez for assisting with my wet-lab experiments, Joe Cardiello for collaborating on the heat shock project, Daniel Ramirez for his meticulous experimental work on the interferon project, and both Jacob Stanley and Rutendo Sigauke for their insights on the computational work.

I also wish to express my appreciation to my thesis committee: Dylan Taatjes, Justin Brumbaugh, and Edward Chuong. Their encouragement to think critically and their consistent support throughout my graduate years were indispensable. Further, I am grateful to the STEM core, BioFrontiers Cell culture facility, FACS facility, Microscope facility, and Sequencing core for their essential support in my research. A special thank you to the administration teams at the MCDB Department and BioFrontiers.

Lastly, my enduring gratitude goes to my partner, Warren, and my children. Even though

our paths often kept us miles apart, they always championed my endeavors, even when they did not quite grasp the intricacies of my work. Knowing that I will soon come back home to you helped me through the most stressful times and brought happiness into my life. I am forever thankful for the unwavering support from my family and friends, who have been my pillars throughout this journey.

Thank you all!

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

General Introduction

Down syndrome (DS), the most prevalent human autosomal anomaly[33], is caused by triplicate copies of chromosome 21. The excess chromosome leads to dysregulation of genes encoded on chromosome 21 - which are over-expressed at or near DNA dosage [80, 162] - as well as numerous genes not located on chromosome 21 also exhibiting altered expression levels[101, 66, 120]. Despite these insights, the precise mechanisms through which a relatively small increase in a limited set of genes gives rise to the multitude of associated co-morbid conditions of Down syndrome remain ambiguous. To identify the molecular basis of DS associated pathologies, we can explore gene expression as it offers a comprehensive perspective on how the genome's information is utilized and manifested[139, 92]. Furthermore it provides an indicator of how the presence of an extra copy of chromosome 21 might contribute to various comorbidities. The study of gene expression in a population of individuals with DS has been instrumental in furthering our understanding of gene dosage effects and identifying dysregulated genes and pathways.

In this thesis, I extend our knowledge of transcriptional dysregulation in Down syndrome by exploring the activation of gene transcription when cells are challenged by external stressors. Whereas prior studies focused on largely unperturbed cells, I sought to ask the question, "How is stress response altered in individuals with Down syndrome?" The work described in this thesis focuses on two different stressors: 1) interferon (IFN) signaling (Chapter 2) in which there are four of the six interferon receptors encoded on chromosome 21 and 2) heat shock (HS) in which there is no known chromosome 21 encoded master regulator (Chapter 3). Leveraging the use of

exogenous stimuli and different time points additionally provides the ability to distinguish between primary and secondary transcription responses. By assessing transcription following the immediate perturbation we can capture the initial transcription that occurs rapidly following a stress. These immediate-early responses subsequently contribute to the expression of secondary genes at a later time point. Therefore, in this introduction, I first discuss Down syndrome, focusing on the genomic features encoded on chromosome 21. I then discuss the process of transcription and how we study it using high-throughput genomics technologies. Finally, I describe the two external stimuli of interest (interferon and heat shock) that have known pathways and response mechanisms, focusing on why these are of interest in Down syndrome specifically.

1.1 The importance of studying the Down syndrome model

Although Langdon Down first described Down syndrome phenotypes in 1866 [52], the molecular cause of Down syndrome was not initially clear. Lejeune, Gautier, and Turpin reported a consistent chromosomal abnormality (trisomy 21) in karyotypes prepared from individuals with Down syndrome [100]. Trisomy 21 is most commonly caused by meiotic nondisjunction during maternal oogenesis (Figure 1.1a). Because the frequency of meiotic nondisjunction increases with age, Down syndrome incidence increases with maternal age[151, 10]. In 5% of cases, trisomy 21 arises from a chromosomal translocation, where a part of chromosome 21 is attached to another chromosome, mostly chromosome 14 t(14;21) or t(21;21) [10] (Figure 1.1b). While DS is typically not inherited, when a parent carries the translocation, they can pass this altered chromosome to their offspring, leading to DS. Mosaic individuals have less severe DS associated phenotypes[58, 132] and will have a mixture of trisomy 21 and typical euploid disomy 21 (D21) cells (Figure 1.1c).

The complexities of Down syndrome arise not just from the genetic impact of an additional chromosome 21 but also from the intricate interplay of various genomic and environmental factors. The triplication of all or part of chromosome 21 leads to overexpression of its genes near the DNA dosage level[80, 162], but it is intriguing to note that several genes not on chromosome 21 also exhibit altered expression[101, 66, 120].

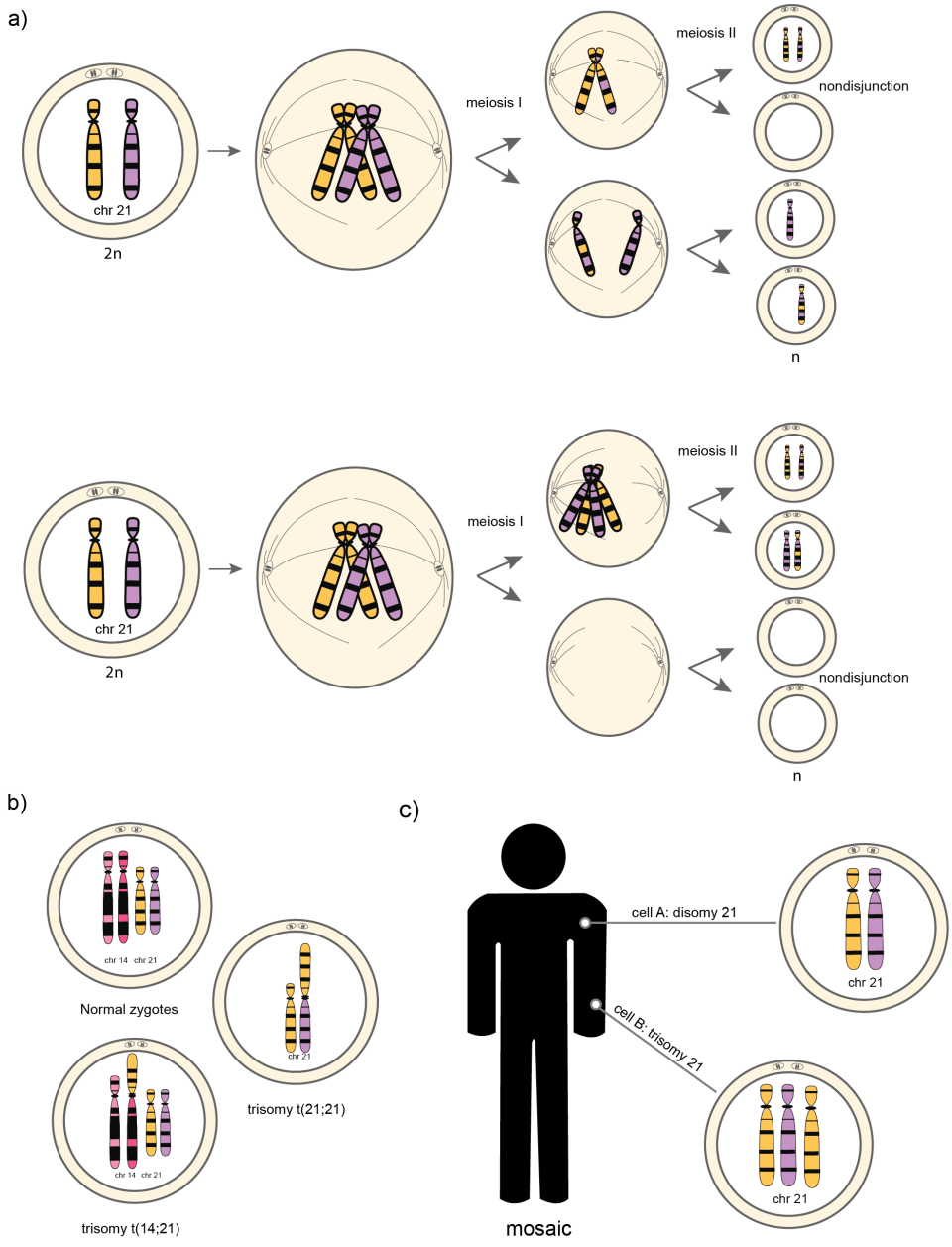


Figure 1.1: **Trisomy 21 causes Down syndrome** a)The most common cause of Down syndrome is nondisjunction during meiosis. In this illustration, we show the meiotic nondisjunction of chromosome 21 in the oocytes, but it can occur in sperm as well. b) Chromosomal translocation of chromosome 21 can also lead to trisomy 21 c) Mosaic individuals have both disomy 21 cells and trisomy 21 cells.

Intriguingly, despite the uniform presence of the additional chromosome, the manifestation of DS varies considerably among individuals. This spectrum of symptoms and severity can be attributed to the fact that, just like all individuals, people with DS possess a high degree of genomic variability. This variability pertains not only to the severity and presence of associated Down syndrome related symptoms but also to the differential expression of genes, both on chromosome 21 and elsewhere in the genome. The variability of Down syndrome phenotypic outcomes is likely modulated by an ensemble of genetic and environmental risk factors[86]. Individuals with Down syndrome are consistently characterized by dysmorphic facial features, mild to moderate intellectual disability, early onset of Alzheimer's disease [176], congenital heart disease [20], autoimmune diseases, mitochondrial dysfunction [8], hematological disorders such as acute lymphoblastic leukemia and acute megakaryocytic leukemia[130], and abnormalities of the immune system characterized by T and B cell lymphopenia, a decrease of naive lymphocytes, and impaired mitogen-induced T cell proliferation [144] (Figure 1.2).

To truly understand and potentially mitigate the effects of Down syndrome we must consider this inherent inter-individual variability. One promising approach is to minimize genomic variability in studies involving individuals with Down syndrome by either utilizing cells from family members of the individual with Down syndrome or sourcing cells from a mosaic individual who possesses some cells with two and others with three copies of chromosome 21. This approach can achieve a more consistent genetic baseline to pave the way for more precise insights.

1.1.1 Genomic origin of Down syndrome

While chromosome 21 is the smallest autosomal chromosome, containing only 1.5% of the genome's DNA, this extra genetic material disrupts typical development, leading to the characteristic physical and cognitive features associated with Down syndrome. Chromosome 21 (HSA21) is $46,709,983$ base pairs long [10, 134] and has annotated 233 protein-coding genes, 423 non-protein-coding genes, and 188 pseudogenes [9].

Chromosome 21 genes are intrinsically linked to many of the distinct features observed in

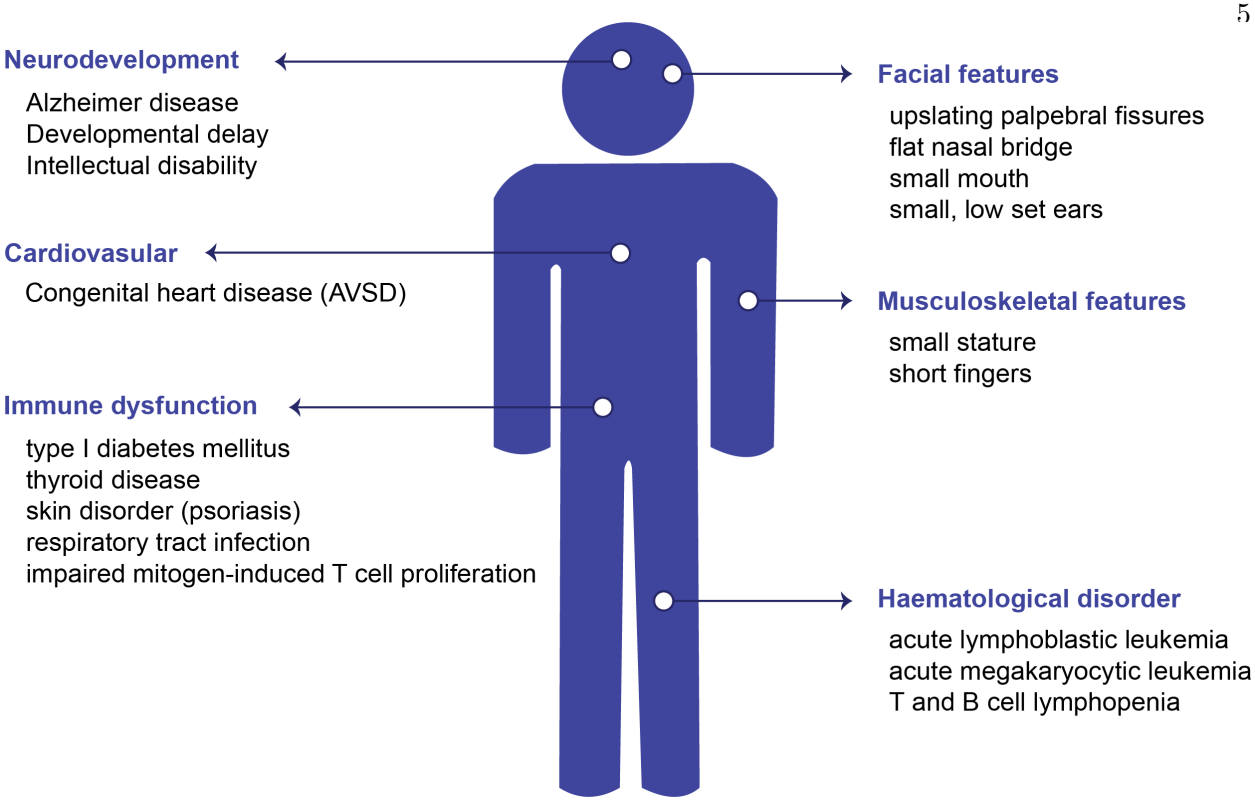


Figure 1.2: **Phenotypes associated in individuals with Down syndrome.** Individuals with Down syndrome exhibit a range of phenotypic variation. While distinct dysmorphic facial features are commonly observed, the syndrome impacts multiple body systems. The figure elucidates some of the diverse pathologies and co-morbid conditions associated with Down syndrome.

individuals with Down syndrome. The region between 21q21 to 21q22.3, known as the Down syndrome critical region, is associated with several Down syndrome hallmarks like distinct facial characteristics, hand formations, and cognitive impairments[124, 121] (Figure 1.2). Interestingly, even a partial trisomy 21 of band 21q22 on chromosome 21 is sufficient to result in phenotypes associated with Down syndrome, suggesting that the gene product of specific genes in this region may contribute to physiological features of Down syndrome[121]. Research indicates that there is not a single chromosomal region that is solely responsible for the DS phenotypes[8, 89, 114, 50]. Two leading hypotheses have been put forward to decipher the correlation between the triplication of chromosome 21 and the array of DS symptoms. The first is the 'dosage imbalance' hypothesis. This suggests that the triplicated chromosome disrupts transcription widespread throughout the

genome. The second, the 'gene dosage' hypothesis, theorizes that the syndrome's features stem from the overexpression of select genes.

An example of 'gene dosage' effect is the overexpression of amyloid-beta precursor protein (APP), a gene linked to early-onset Alzheimer's disease in people with Down syndrome[176]. Other examples encompass genes like HMGN1, tied to an elevated risk for specific types of leukemia and the Runt-related transcription factor 1 (RUNX1) transcription factor associated with certain blood disorders in DS[149, 80]. Contrarily, the 'dosage imbalance' theory proposes that chromosome 21 genes lead to widespread biological dysregulation, indirectly altering specific cellular functions. Some affected processes include chromatin availability (HMGN1, BRWD1), splicing regulation (U2AF1L5, RBM1, U2AF1, DRYK1A), post-transcription regulation (ADARB1), secretory-endosomal functions (DOPEY2, CSTB, SYNJ1), metabolism (SOD1), and protein turnover (USP25) [80, 119]. Furthermore, aneuploidy itself - a condition marked by an atypical chromosome count (and T21 is an aneuploidy) - can disrupt the cellular balance, affecting vital processes like metabolism and DNA repair[80, 153, 159, 139].

1.1.2 Early transcription studies in Down syndrome

Having explored the leading hypotheses regarding the role of genes in Down syndrome, we now focus our attention on transcription and its role in this complex condition. Transcription, the process of converting the genetic code (DNA) into functional RNA molecules, is pivotal in shaping the unique phenotypic expressions of Down syndrome. Emerging techniques in next-generation sequencing have proven instrumental in exploring this critical aspect of cells at a scale and resolution previously unattainable. These high-throughput methodologies, capable of quantifying transcription across the entire genome, have illuminated how the dysregulation of gene expression in DS influences various conditions associated with DS[139, 92]. Thus, we delve into the study of transcription in DS, guided by the power of next-generation sequencing technologies.

Numerous RNA-sequencing studies have sought to understand the impact of triplicated genes on chromosome 21 on overall messenger RNA (mRNA) levels. The consensus among these studies is

that cells from individuals with DS exhibit expression at DNA gene dosage, approximately 1.5-fold a diploid level for chromosome 21 encoded genes, across various cell types and samples. This suggests that there is a lack of dosage compensation for human autosomes[78, 80, 162]. Gene expression profiling techniques, in this context, are indispensable. They enable the influence of T21 on the expression levels of all genes to be examined. While many studies have employed mouse models showcasing partial triplication and Down syndrome-like features[85, 115, 136], limited research has been undertaken in human DS tissues. Investigations spanning various tissues like whole blood[162, 165], fibroblasts[103], T cells[69], placenta[72], and brain[120, 110] have indeed highlighted a primary gene dosage effect. However, the overall transcriptional impact throughout a more diverse set of tissues across the genome remains unrevealed[59]. In addition, the degree of expression level varies between tissues as well as between studies suggesting that the regulation and expression of chromosome 21 genes is likely dynamic and complex.

1.2 Understanding gene expression

Regulating gene expression is crucial for the proper functioning of a cell. Although every cell in an organism has identical DNA, the way genes are turned 'on' or 'off' varies depending on the cell type and its environmental context. This regulation ensures that genes produce proteins only when needed, allowing the cell to adapt to changes and maintain balance. Central to this regulatory process are transcription factors (TFs) and the mechanisms that control their activity. These factors interpret and respond to cellular signals, which often start when a cell detects a chemical stimulus. Generally speaking, gene expression is initiated by cellular signaling; the cell detects a chemical signal through its cell surface receptor leading to activation of its kinase domain and a cascade of downstream kinase phosphorylation of targets. The propagation of a signal is highly coordinated and ultimately results in the expression of specific genes. In multi-cellular organisms, this selective gene regulation results in a diverse array of cell types, each with specialized functions. The Central Dogma of molecular biology articulates this process: genetic information flows from DNA to RNA through transcription, and then from messenger RNA (mRNA) to proteins via translation. This

mechanism enables cells to produce multiple protein molecules from a single gene at a given point in time.

1.2.1 The process of transcription to regulated gene expression

RNA polymerase II is the primary enzyme responsible for transcription as it synthesizes RNA from the DNA template. In eukaryotes, there are three main types of RNA polymerases (I, II, and III), with RNA polymerase II being responsible for transcription of messenger RNA. There are three stages of transcription by RNAP II; initiation, elongation, and termination. At initiation, transcription factor (TF) associates with motif sequences found at regulatory regions (enhancer and promoters) and recruits RNAP II with associated general transcription factor (GTF) aiding in its accurate placement on the DNA sequence. Promoters are cis-acting transcription regulatory sequences located upstream (or $5'$ end) of transcription start sites (TSS) and define the direction of transcription, that is the DNA strand that is 'sense'. Enhancers are sequences that contain recognition sites for multiple TFs, and when bound by specific TFs, enhance transcription of an associated gene.

Initiation of transcription at the TSS of promoters begins with the formation of the open complex where the DNA at the promoter unwinds and opens up after RNAP II binds. After initiation, RNAP II synthesizes a short stretch of RNA approximately 30 to 40 nucleotides long. Two proteins, DRB-sensitivity inducing factor (DSIF) and negative elongation factor (NELF), bind to RNAP II and transcription is paused until P-TEFb phosphorylates Ser2 of the CTD of RNAP II to elongate the transcript. RNAP II continues to synthesize the RNA transcript and the RNAP II CTD gets additionally phosphorylated during elongation, aiding in the recruitment of RNA processing enzymes. Capping enzymes (methyl transferase (MT), guanyl transferase (GT)) associate with the CTD to promote $5'$ cap addition, followed by recruitment of spliceosome by SR-like CTD-associated factors (SCAFs). Elongation continues until after the TES is recognized by RNAP II. In eukaryotes, the newly synthesized pre-mRNA undergoes post-transcription processing that includes $3'$ cleavage and polyadenylation mediated by CTD-associated cleavage-stimulation

factor (CstF) and Cleavage polyadenylation stimulatory factor (CPSF). CstF and CPSF in turn, attract additional cleavage and polyadenylation factors, to add the poly(A) tail to the $3'$ end of the mRNA. Termination occurs when the RNAP II complex dissociates from the DNA[5, 138] (Figure 1.3).

1.2.2 Transcription factors are regulators of gene expression

Transcription factors are the key regulators of cellular processes, both intrinsic (development and differentiation) and extrinsic (response to exogenous signals). In the case of extrinsic, external cues typically initiate cell surface receptor activation and cell signaling, resulting in the modulation of a particular set of TFs. The activity of TFs are tightly controlled to ensure appropriate responses to a myriad of cellular and environmental cues. This control is exerted at multiple levels, from their transcription and translation to post-translational modifications (Figure 1.4), to ensure accurate and timely gene expression. TFs can be further controlled by modulating the accessibility of their binding sites such as cell-type specific chromatin states. Additionally, the binding of a TF to their preferred sequence-specific recognition motifs can alter the chromatin state to either promote or obstruct other factors from binding and thereby regulate gene transcription by altering the activities of nearby RNAP II.

To function effectively, TFs often undergo post-translational modifications within specific signaling pathways. These modifications can activate or inhibit TFs, without affecting their DNA-binding capabilities. This process allows cells to rapidly modulate the activity of existing TF to respond to environmental changes, stress, or signaling cues. One exemplar of this process is Janus kinase/signal transducers and activators of transcription (JAK/STAT) signaling pathway that activates STAT TFs [128]. In this signaling pathway, modifying STAT by phosphorylation allows for rapid cellular response to interferon stimuli. Beyond phosphorylation, other post-translational modifications include ubiquitination, acetylation, glycosylation, methylation, and SUMOylation, which can influence TF localization, stability, activity, and interaction with other proteins[57, 79].

Upon activation, a TF must bind to the chromatin before it can modulate gene expression.

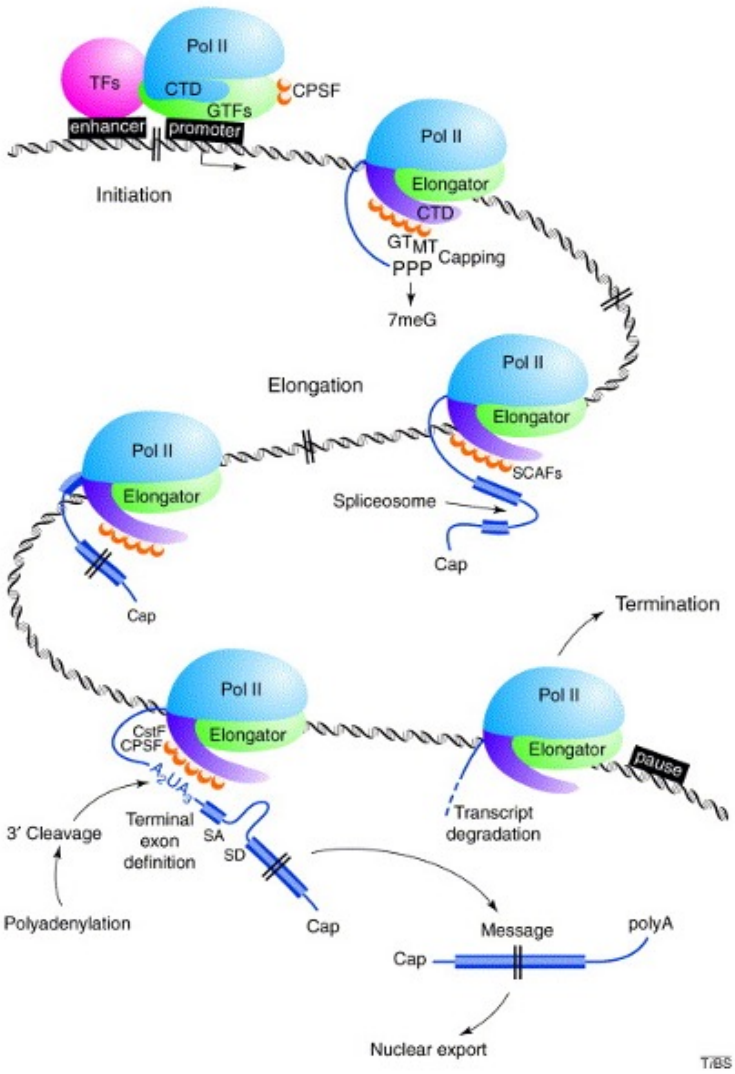


Figure 1.3: **The stages of transcription.** The three stages of transcription by RNAP II are initiation, elongation, and termination. At initiation, TFs (pink) bind at regulatory regions and recruit RNAP II (light blue) and general transcription factors (green). The RNAP II CTD (purple) becomes highly phosphorylated in the elongation stage. Capping enzymes associate with the CTD to promote 5′ cap addition followed by spliceosome recruitment to splice introns from the pre-mRNA. The 3′ cleavage and polyadenylation are mediated by CstF and CPSF. The polyadenylated mRNA is released from the transcription complex. The last step is termination which involves degradation of the remaining nascent transcript (Illustration from Proudfoot 2000)[138].

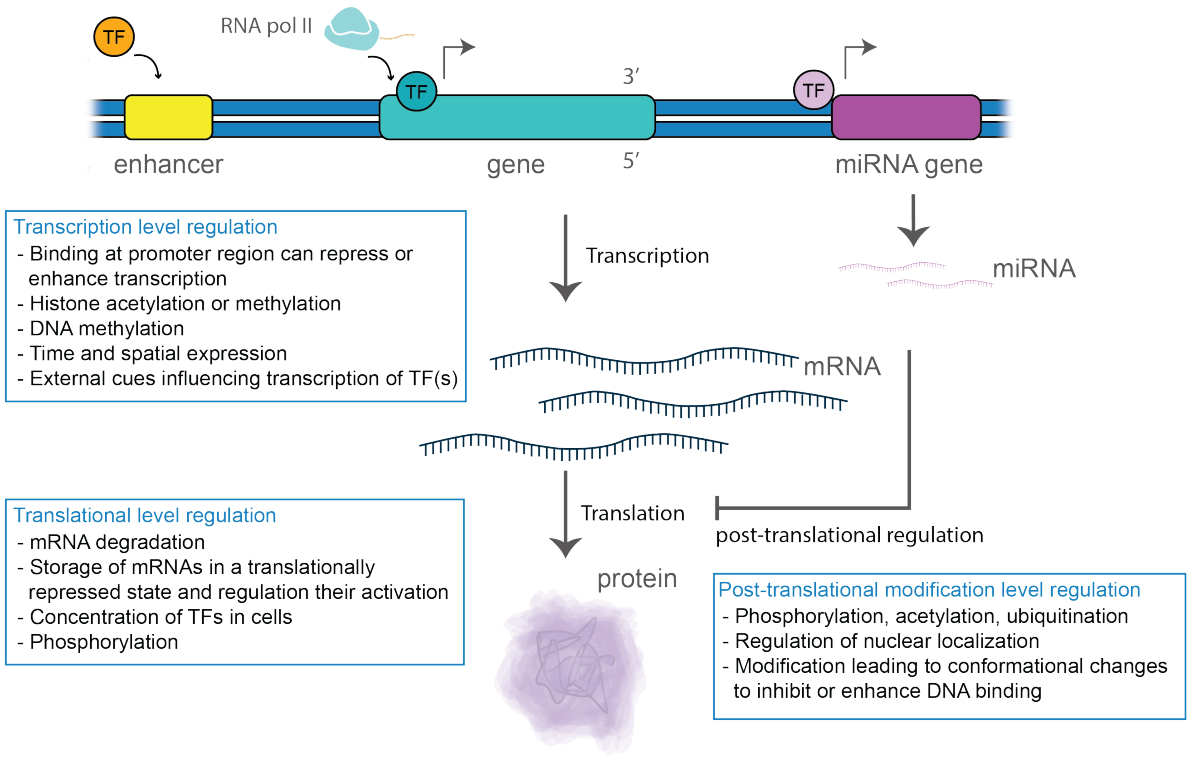


Figure 1.4: **Basic regulation by TFs.** Illustration of the mechanism of regulating TFs, which in turn regulates gene expression. TFs can be regulated through transcription, translation, or post-translation modification. Here, we provide some examples of distinct types of TF regulation.

While some TFs bind less specifically, as in the case of GTFs that are part of the basal transcriptionary machinery, most recognize specific DNA sequences, described by a degenerate sequence recognition motif. These motifs are inferred from TF binding assays *in vitro* such as SELEX [83] or *in vivo* such as ChIP-seq[82, 170]. More recently chromatin profiling technologies such as CUT&RUN[123, 155] have been used to find enrichment of sequences among the TF-bound DNA fragments. The TF motifs have been compiled in motif databases such as HOCOMOCO[91] or JASPAR[61]. A challenge in studying a TF is that knowing the TF binding motif is not sufficient to determine if the site is bound *in vivo*. Furthermore, the DNA sequence alone cannot predict whether TF binding in an *in vivo* setting will alter transcription nearby[15].

Notably, a significant number of TFs are cell-type specific and display expression patterns that are tissue-specific[96]. Additionally, the same TF can bind different loci depending on the context[68] or change their mode of action in different cell types [25]. To study TF binding in specific cell types, large-scale efforts such as ENCODE[46], have profile TFs across thousands of cell types. Although large-scale efforts have profiled TFs across numerous cell types, our knowledge is limited to only those cell types that have been experimentally profiled.

Many TFs, despite being present in low abundance at the protein level[104, 163], play vital roles in cellular processes. Their functional activity does not necessarily correlate with their protein level. TF function can be dependent on post-translation modification or the binding to their interaction partners. ChIP-seq[82] is a commonly used technique to detect and measure the binding of TF to chromatin. However, only a subset of bound sites will alter transcription nearby [15] making it challenging to delineate the functional binding sites from non-specific binding events. Another challenge in TF binding assays is that they are typically limited to one TF at a time, and constrained on having a high quality ChIP-validated antibody.

Understanding gene regulation is intricate, involving multiple TFs. While a plethora of TFs may be involved, they are rarely all assayed in a given cell type. Current databases like HOCOMOCO[91] and JASPAR[28], catalog preferred TF sequence binding motifs but are based on selective experimental methods, which might not capture the complete range of binding contexts. These methods encompass ChIP-seq[82] for identifying DNA sequences bound by a TF *in vivo*, Electrophoretic Mobility Shift Assays (EMSA)[157] to discern DNA-protein binding, DNAse footprinting[62] to detect DNA regions protected by protein binding, and Systematic Evolution of Ligands by EXponential Enrichment (SELEX)[34] which methodically screens amplified DNA or RNA samples to identify molecules with high affinity to a target protein. Despite their reliability, these experimental methods might not capture the complete spectrum of TF binding contexts, potentially missing some motifs or producing false positives. Computationally, in silico methods offer alternative insights by predicting TF binding motifs using statistical models. Notable techniques include the calculating nucleotide probabilities at specific positions using the Position Weight

Matrices (PWM) matrix, leveraging the Multiple EM for Motif Elicitation (MEME)[19] algorithm to identify motifs in a set of unaligned sequences by employing expectation-maximization algorithms, Discriminative Regular Expression Motif Elicitation (DREME)[18] to find core motifs in ChIP-seq data, and Homer[76], which contrasts discovered motifs with known motifs. It is important to consider that binding experiments are inherently enrichment assay that often yield low signals which can produce inconclusive data, making data interpretation challenging.

Gene regulation is fundamentally influenced by noncoding regulatory elements, notably enhancers. These regulatory elements have a role in orchestrating transcription and hence controlling cell-specific gene expression. Enhancers, characterized by dense binding with TFs, are theorized to control the levels of transcription through interaction with target promoters either on the same (cis) chromosome or on a different (trans) chromosome[42]. When a TF binds to an enhancer, it can recruit RNAP II to load subsequently synthesizing short, unstable, bidirectional RNAs near the binding sites[15]. The biogenesis of these transient, bidirectional RNAs, termed enhancer-associated RNAs (eRNAs), are closely associated with the onset processes of transcription initiation. Multiple studies have postulated that eRNAs influence gene expression by facilitating enhancer-promoter looping, modulating the local chromatin landscape, and promoting RNAP II activity at specific gene promoters[87, 105, 177]. However, this perspective is not universally accepted. Some data suggest that eRNAs are simply by-products of active enhancers without any distinctive regulatory role. This view is supported by observations that inhibiting enhancer transcription does not necessarily impact the expression of proximal genes, implying that eRNA production may not be pivotal for gene expression[87]. Others propose that eRNAs' role in enhancer-promoter looping is merely indicative of an already permissive chromatin structure conducive for loop formation, rather than a direct consequence of eRNAs' activity[129]. Moreover, the notably short lifespan of many eRNAs, with many having a half-life of only a few minutes, raises doubts about their ability to exert long-term regulatory function[106].

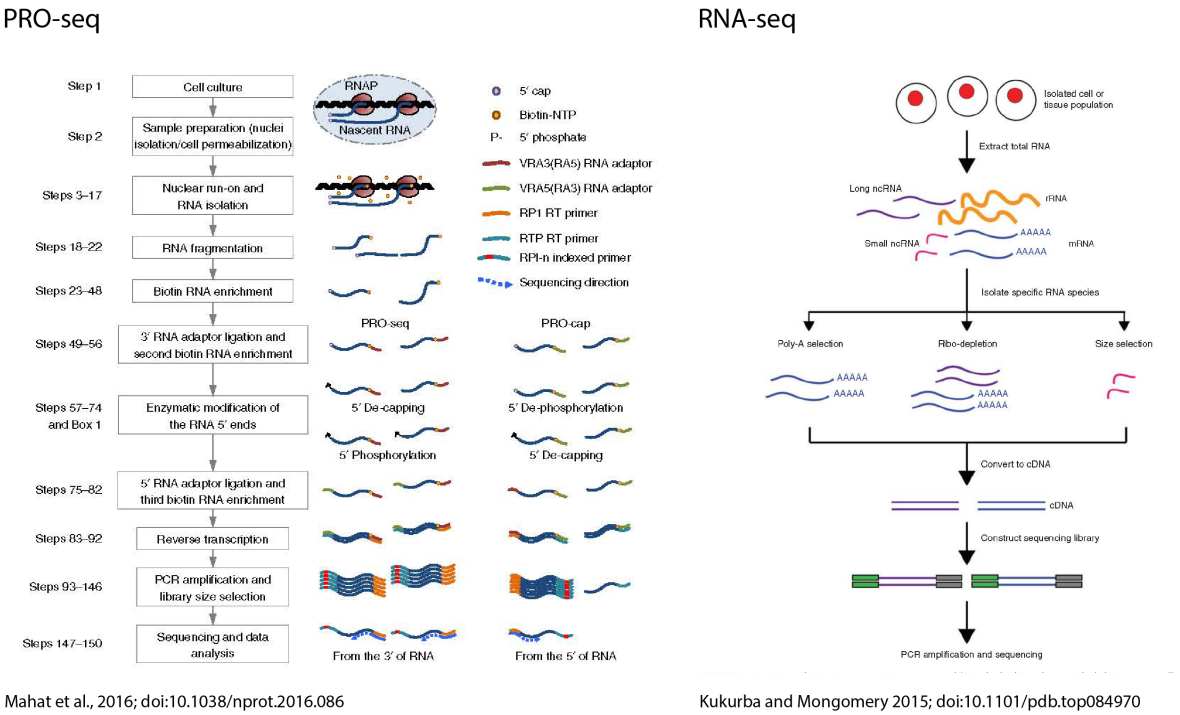
To further our understanding of transcription factor and their function in cells, I have developed tools to help us study the complex relationship between TF and bidirectional transcripts.

In Appendix A, I will discuss tools that I helped develop to look specifically at the activity profile of RUNX1, a TF encoded on chromosome 21. RUNX1 is a master regulator of hematopoiesis and its activity is tightly controlled at the transcriptional and post-transcriptional levels[36]. Prior eRNA profiling of RUNX1 showed it not only had the canonical active pattern (co-localization of the TF motifs with sites of transcription initiation) but also an unusually striking "offset" pattern[15]. I will describe in Appendix A a classifier that I built to categorize TF's eRNA profiles, aiming to identify additional TFs with this novel "offset" pattern. In Appendix B I then introduce a reporter construct that I developed to quantify enhancer influence in transcription. Finally, in Appendix C I describe my contributions to two computational projects aimed at developing improved tools.

1.2.3 Measuring transcription and gene expression

To investigate gene expression alterations during differentiation or due to perturbations, various RNA assays can be employed. These assays can track the generation of transient RNAs or total mRNAs levels in a population of cells that undergo changes due to external influences or natural processes. One effective method to understand these changes in gene expression is by quantifying the levels of specific RNA-seq transcripts. For instance, mRNA abundance in a cell population can be ascertained using techniques like RNA-seq. Alternatively, for a more in-depth view, we can assess nascent transcripts (those that are pre-splice and not yet matured into mRNA) using methods such as global run-on sequencing (GRO-seq)[41] or precision run-on sequencing (PRO-seq)[94].

RNA-sequencing assay detects and quantifies RNAs extracted from biological samples, either cells or tissues, at a given point in time. First, mRNA is enriched through either a polyA selection or ribosomal depletion step. The RNA-seq is then converted to complementary DNA (cDNA) by reverse transcription. Sequencing adapters are ligated to the ends of cDNA fragments and pairs of primers are used to amplify the targeted DNA segment using polymerase chain reaction (PCR) to create short segments called amplicons. These DNA segments are used for library preparation where they are modified to have a sample-specific index to help identify the cell or tissue source. Sequencing



Mahat et al., 2016; doi:10.1038/nprot.2016.086

Kukurba and Mongomery 2015; doi:10.1101/pdb.top084970

Figure 1.5: **Sequencing Protocols.** (left) Nascent RNA-seq captures the newly synthesized RNA transcripts. Illustrated here is PRO-seq. (Illustration from Mahat 2016)[116] (right) RNA-seq measures total mRNA. RNAs are extracted from cells or tissues enriched for mRNA and converted to complementary DNA (cDNA) by reverse transcription. Sequencing adapters are ligated to the ends of the complementary DNA (cDNA) fragments, and amplified by polymerase chain reaction (PCR) to create amplicons that will be sequenced. (Illustration from Kukurba 2015)[90]

adaptors are added to DNA segments to enable parallel sequencing (Figure 1.5). RNA-seq is a valuable tool to study expression as it allows for the detection of lowly expressed genes, alternative splice variants, single nucleotide variants (SNPs), and non-coding RNAs (ncRNAs)[141].

In contrast, nascent RNA-seq sequencing (GRO-seq, PRO-seq), measures the newly produced RNA transcripts prior to maturation (Figure 1.5). These include protein-coding genes, microRNAs (miRNAs), long non-coding RNAs (lncRNAs), and eRNAs. The transcripts captured with nascent sequencing are pre-splicing and therefore need not be stable. Hence eRNAs are captured, and these transcripts provide a readout on the regulatory activity occurring within the enhancer (and promoter) regions. A much larger fraction of the genome is transcribed (signal in nascent) than appears in a steady-state assay such as RNA-seq, suggesting that much transcription is inherently unstable.

With respect to studying interferon stimulus (Chapter 2), I propose that nascent transcription assays also allow us to study the immediate-early response genes that are typically observed rapidly after stimuli and maybe transient.

Regardless of whether one examines mRNA or nascent RNA, after sequencing the resulting data is analyzed by fairly standard bioinformatics pipelines. Reads are first assessed for quality, trimmed (if necessary) to remove any lingering adapter sequence, and then mapped to a reference genome. Reads are then counted at genome features. In RNA-seq the features tend to be annotated genes. However, eRNAs are not annotated. So in nascent sequencing, it is necessary to call transcripts directly from the data before counting. Sites of bidirectional transcription, seen at enhancers and promoters, are typically identified using either a probabilistic model of RNAP II behavior (the Tfit algorithm[17]) or by using a trained classifier (dREG algorithm[45]). Counts are then fed to differential tools such as DESeq2[111] or edgeR[148] for the assessment of statistically significant changes across conditions.

1.2.4 Studying transcription in a trisomy model

While research studies in populations with Down syndrome have provided significant insights into the molecular mechanisms underlying the complex phenotypes associated with the trisomy 21, there remain many questions. Proteomic studies in DS have found that the global protein expression ($\tilde{1}$.4 fold over typical) is slightly less compared to mRNA expression ($\tilde{1}$.5 fold over typical) [80]. Because post-transcriptional regulators such as ADARB2 and miRNAs are encoded on chromosome 21 it has been suggested that the difference in protein versus mRNA expression may be a result of post-transcriptional regulation.

Transcriptional studies in populations with Down syndrome have provided significant insights into the molecular mechanisms underlying the complex phenotypes associated with the trisomy of chromosome 21. Predominantly, these investigations utilize a case-control approach, contrasting gene expression profiles of individuals with trisomy 21 to euploid disomy 21 (often referred to as "typicals"). Key methodologies like microarray, RNA-seq, ATAC-seq (identifies open chromatin), and

4C-seq (examines higher order chromatin structure) have been employed. RNA-seq has been used to capture the differential gene expression pattern between T21 and control euploid (D21) populations [101]. As expected, genes on chromosome 21 manifest altered expression in individuals with DS. Still, this dysregulation extends beyond chromosome 21; pathways impacted include those related to interferon signaling, immune response, and cell cycle regulation, underscoring the widespread impact of the extra chromosome [162, 101]

In what follows, I will discuss two environmental perturbations relevant to Down syndrome. The first, interferons, are proteins that are part of a cell's natural defense against pathogens. As four of the six interferon signaling proteins are encoded on chromosome 21, it is perhaps unsurprising that Down syndrome is considered an interferonopathy. The second, heat shock, has less obvious ties to chromosome 21. Yet individuals with Down syndrome have distinct body temperature regulation compared to typicals. My work focused on these two stressors – one obviously encoded on chromosome 21 and one not so obvious – and how trisomy 21 cells react to these stresses.

1.3 Down syndrome as an interferonopathy

Down syndrome is identified as a Type I interferonopathy due to its pronounced Type I IFN (IFN-I) activity, with associated immune abnormalities evident in these individuals. This heightened IFN-I activity is linked to the presence of four interferon receptors - IFNAR1 and IFNAR2, IFNGR2, IL10RB - encoded on chromosome 21, which are believed to influence the altered IFN-I signaling in DS[80, 162]. The recent Galbraith 2023 paper[65] however found that the IFN-I interferonopathy is more complicated in DS, where the hyperactivity and dysregulation are not attributed only to IFN-I, but what the researcher defined as "mixed-type interferonopathy". They found that the overexpression of all three IFN-Rs contributes to the IFN hyperactivity and dysregulation observed in T21 cells.

Clinically, DS individuals present a dichotomous immune profile. On one side, these individuals often exhibit hyperactive immune responses that predispose them to autoimmune and autoinflammatory diseases such as celiac disease, type I diabetes mellitus, and hypothyroidism [31].

Conversely, epidemiological data suggest that while DS individuals might have reduced susceptibility to certain viral infections due to the antiviral nature of IFN-I, they experience more severe symptoms and higher mortality when they do get infected, especially with respiratory viruses such as influenza and SARS-CoV-2[112, 39, 60, 30]. A case in point would be the recent COVID-19 pandemic, where hospitalized individuals with DS experienced more severe complications and higher rates of sepsis and mechanical ventilation[60, 21, 81]. Delving into IFN signaling in DS provides valuable insights into this population's distinct immune response, enhancing our understanding of this pathway and potentially uncovering immune mechanisms applicable to both individuals with DS and the general population.

1.3.1 Type I receptor and IFN-I signaling in a typical background

There are six interferon receptor subunits: IFNAR1, IFNAR2, IFNGR1, IFNGR2, IFNLR1, and IL10RB. Four of the six IFN receptor subunits - IFNAR1 and IFNAR2, IFNGR2, IL10RB - are encoded on chromosome 21. IFNAR1 and IFNAR2 represent the two subunits of the Type I IFN (IFN-I) receptor. Meanwhile, IFNGR1 and IFNGR2 are the subunits for the Type II IFN receptor. IFNLR1 and IL10RB, while serving as a subunit for the Type III receptor, also function as receptor subunits for three specific interleukins: IL-10, IL-22, and IL-26. These IFN receptors (IFN-Rs) are designed to bind to interferons, cytokines classified into three main types: Type I IFN, Type II, and Type III IFN. Type I IFN encompasses more than ten IFN-α subtypes and a single IFN-β subtype. These IFN-Is play diverse roles in the immune system, including roles in innate and adaptive immune cells during infections caused by viruses, bacteria, parasites, and fungi. When cells encounter pathogens or certain endogenous signals, Type I IFN (IFN-I) responds either directly or through the induction of other proteins to defend against viral threats. The Type II IFN, known as IFN-γ, is mainly produced by natural killers (NKs) and several types of T cells. Type III IFN is comprised of IL-29, IL-28A, and IL-28B (also termed IFN-λ1-3). Their receptors are primarily expressed in specific tissues, predominantly epithelial cells and hepatocytes[145].

Given that the majority of data exists for the triplication of Type I IFN receptors in Down

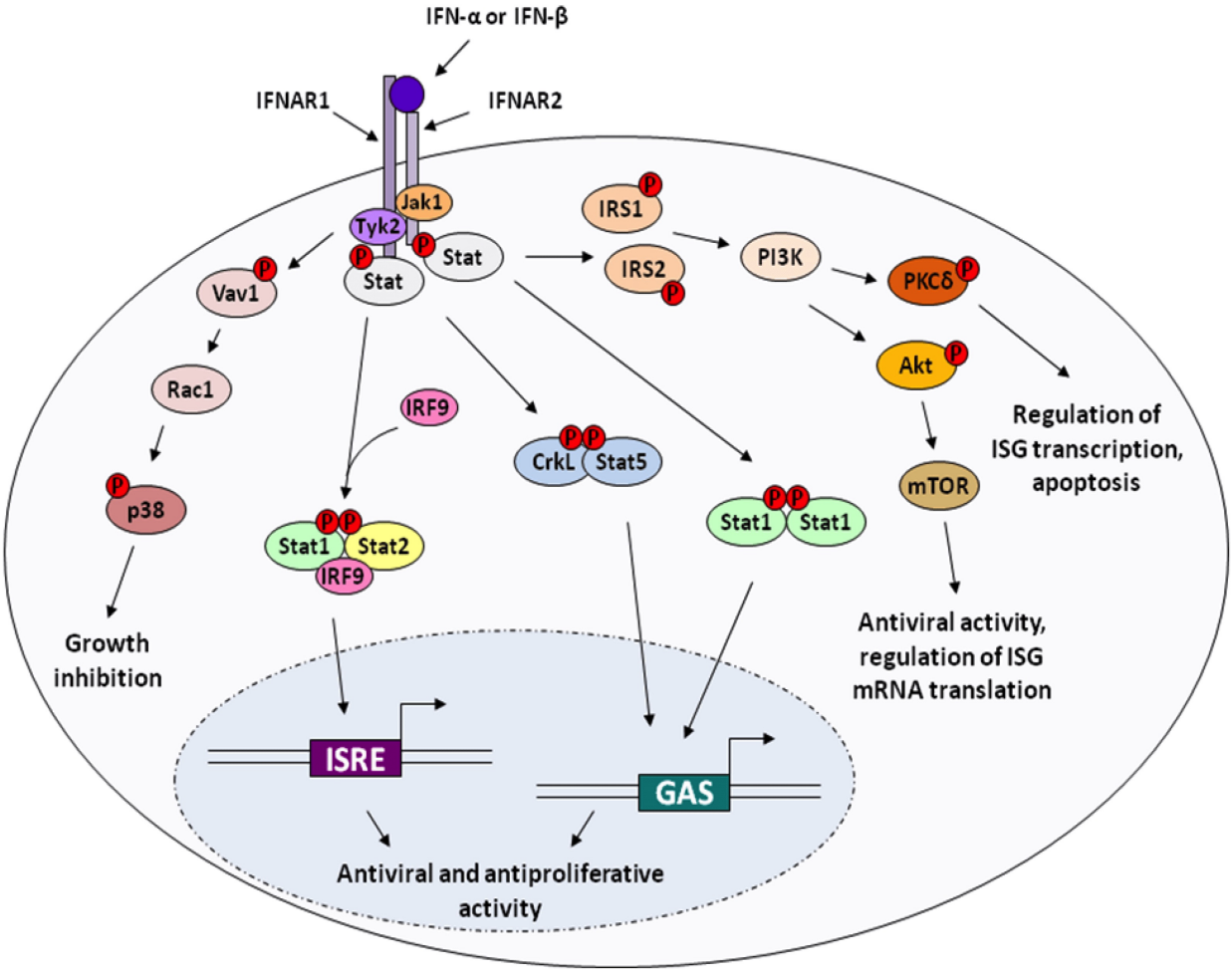


Figure 1.6: **Type I IFN Signaling Pathway.** Type I IFN-R is composed of two transmembrane receptor subunits, IFNAR1 and IFNAR2. The binding of IFN-Is activates the kinases JAK1 and Tyk2. In the canonical signaling pathway, STAT1/2 proteins are phosphorylated and dimerized to ISGF3, which is composed of pSTAT1, pSTAT2, and IRF9, and induces transcription of hundreds of interferon-stimulated genes. Noncanonical signaling pathways include p38 MAPK, pAkt, and pCrk. Phosphorylation of p38 modulates IFN activity and growth inhibition of cells, pCrk (CrkL and CrkII) are tyrosine phosphorylated after IFN treatment, and pArk and mTOR are important in mediating IFN activity. (Illustration from Bekisz 2010)[24].

syndrome, our focus will be on IFN-I. IFN-I cytokines, IFNα and IFN-β , are produced when cells encounter viruses or double-stranded RNA. Upon synthesis, they attach to surface IFN-receptors, triggering a chain of reactions including the phosphorylation of JAK1 and TYK2 kinases. This leads to the activation of pSTAT1 and pSTAT2. Subsequently, the STAT1/STAT2 heterodimer pairs with a third subunit, IRF9, forming the interferon-stimulated gene factor 3 (ISGF3) complex.

This pISGF3 complex initiates the transcription of hundreds of interferon-stimulated genes (ISGs) in the nucleus, establishing an antiviral and antiproliferative state in the infected and neighboring cells [145] (Figure 1.6).

This initiation of the JAK/STAT signaling pathway prompts the expression of immediate-early response genes. These first viral genes are transcribed after infection and three transcription does not require de novo protein synthesis. Their swift activation then paves the way for the expression of secondary response genes, which rely on the newly synthesized regulatory gene product [12, 13]. ISGF3's binding pattern to ISRE is the driver for IFN-I-induced transcription activation and varies based on cell type and the timing and level of IFN stimulation[102]. The subsequent transcription of ISGs mediates restriction of viral replication, hinders cell proliferation, induces apoptosis, and activates subsequent innate and adaptive antiviral immune responses [152].

Responses to viral infections vary widely across populations, even when considering different interferon IFN subtypes[73, 64]. This variability was evident during the recent COVID-19 pandemic, where hospitalized patients showed that IFN-I subtypes have varying potency and are associated with distinct metabolic signatures[64]. Similar patterns were seen in HIV-1 patients, where different IFN-I subtypes had diverse capabilities to inhibit HIV-1 replication ex vivo. Moreover, these subtypes expressed varying levels of core ISGs associated with inflammation and immune system activation [73]. Notably, in HIV-1 patients, the IFN-β subtype exhibited a broader interferome compared to IFN-α [73].

1.3.2 Dysregulation of IFN signaling in Down syndrome

Down syndrome is recognized as having characteristics of a Type I IFN (IFN-I) interferonopathy. This association stems from the presence of four interferon receptors on chromosome 21 and the distinctive IFN-I signaling observed in individuals with Down syndrome[80, 162]. The presence of the triplicate copies of these IFN receptors might be pivotal in disrupting the standard IFN response. This discovery has been an ongoing area of research since the 1970s. Notably, Tan YH's seminal work laid the foundation for this exploration when he found fibroblasts with trisomy 21

to have enhanced protection against vesicular stomatitis virus due to heightened IFN sensitivity compared to typical disomy 21 cells or cells with trisomy 13 or 18[164]. Following this, additional studies revealed that T21 fibroblasts display an increased number of IFN-α receptors, suggesting a molecular basis for the increased IFN sensitivity[54, 67].

Delving deeper into the molecular intricacies, researchers observed an approximately 1.5-fold increase in the protein expression of three interferon receptors from chromosome 21 —IFNAR1, IFNAR2, and IFNGR2 — in T21 B-EBVs and monocytes, compared to typical cells[88]. Despite this general increase, there's considerable variability in protein expression levels, with some overlap evident between T21 and D21 cells[88]. Analysis of cell surface protein expression showed that most white blood cells from individuals with DS exhibit elevated IFNAR1 levels compared to those without DS, though the magnitude of this elevation varies among different cell types[67, 174].

Transitioning from mere protein expression to functional dynamics, it is important to recognize that IFNAR1 and IFNAR2 receptor behavior on the cell surface is dynamic and influenced by IFN-I signaling. Upon cytokine binding, receptors IFNAR1 and IFNAR2 are internalized with aid from the retromer complex, a component of the endosomal protein system. Typically, IFNAR1 is directed towards the lysosome for degradation, while IFNAR2 is recycled back to the plasma membrane. If there is an obstruction to this process, keeping IFNAR subunits at the plasma membrane can amplify IFN-dependent signaling and subsequent gene transcription[38]. Interestingly, individuals with Down syndrome often exhibit abnormalities in their endosomal and retromer complex (Figure 1.7), a trait linked with an elevated Alzheimer's Disease risk[44, 56]. This elevated expression of IFNAR in Down syndrome means more receptor subunits linger on the cell surface after IFN signaling. Yet, the connection between retromer dysregulation in Down syndrome and its impact on IFN signaling has not been explored.

While these protein-level dynamics offer profound insights, the transcriptional landscape in Down syndrome adds another layer of complexity. RNA sequencing data from various cell types indicated that genes on chromosome 21 tend to be over-expressed, approximately at or near a 1.5-fold increase, in cells from individuals with Down syndrome compared to typical individuals.

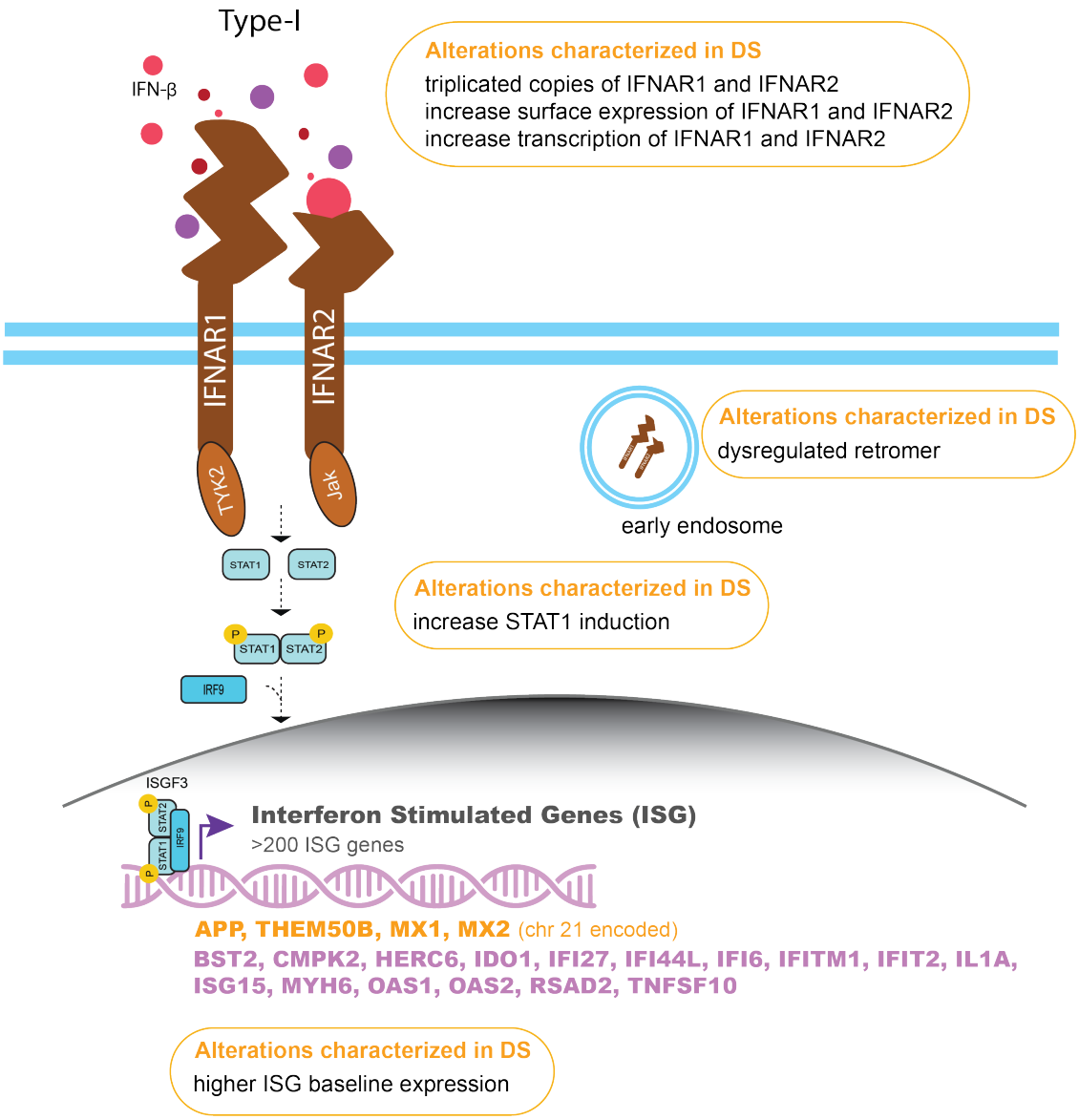


Figure 1.7: **Type I Interferon dysregulation in Down syndrome.** Illustration of IFN binding to Type I IFN leads to activation of JAK/STAT signaling followed by transcription of IFN-stimulated genes. Alterations characterized in individuals with Down syndrome are highlighted in yellow.

Intriguingly, this overexpression is not universally true for all genes, with a few exceptions — most notably, the interferon receptors[80, 162]. Furthermore, downstream ISGs consistently exhibit higher expression levels in DS immune cells than in typical cells[162, 11]. These initial insights hint at the non-uniform elevation of IFN-I receptor expression in Down syndrome (Figure 1.7), sparking ongoing investigations into the underlying reasons for this variability.

Taking a step back, the broader implications of this dysregulated interferon response in Down syndrome cannot be ignored. While an initial surge in IFN response can be protective, its chronic elevation can become detrimental. This pattern could explain observations in the population with DS population, where sustained high levels of circulating IFN could induce the transcription of ISGs creating a positive feedback loop, amplifying IFN signaling and consequently impacting the body system of these individuals[162, 11]. The altered interferon response in DS can influence the outcome of viral infections. On the one hand, a heightened interferon response can provide better initial protection against viral infections. On the other hand, an overactive immune response can lead to immune-related complications, as seen in severe cases of COVID-19 where hospitalized patients with DS faced more severe complications than the general population[60, 21, 81]. The heightened severity in patients with DS might be tied to elevated IFN levels fostering autoantibodies, given their known propensity for autoimmunity[21]. Epidemiology studies have found that individuals with DS often fare better initially during viral infections compared to typical individuals, but unfortunately, experience more severe symptoms and higher mortality rates upon prolonged infection[112, 39, 60, 30].

While the molecular mechanisms underlying IFN dysregulation in DS remains a subject of active research, it is evident that the distinct genomic and cellular profiles of T21 significantly shape immune responses and overall health.

1.3.3 Looking forward to Chapter 2

In this thesis, I will investigate the response in a human cohort population comprising both typical individuals and those with trisomy 21 when exposed to IFN-β (detailed in Chapter 2). This research primarily offers two novel insights into the field: the employment of a direct IFN perturbation and the exploration of the immediate-early response genes captured by nascent transcription. For the direct IFN perturbation, IFN-β was chosen over its counterparts, IFN-alpha (IFN-α) and IFN-gamma (IFN-γ). Notably, IFN-β is a cytokine associated with type I receptors, leading to the activation of the downstream transcription factor complex, interferon-stimulated gene factor 3 (ISGF3)[1]. Although IFN-α also targets Type I receptors, its 13 subtypes present

nomenclature inconsistencies across vendors, potentially hindering reproducibility. Therefore, to maintain clarity, our study centers on IFN-β. It should be noted that IFN-γ activates Type II receptors, but there is crosstalk between Type I and II making it difficult to distinguish between Type II and I responses.

1.4 Heat shock response in Down syndrome

Clinical findings suggest that individuals with Down syndrome have a different body temperature regulation compared to typical individuals. Individuals with DS tend to overheat more readily, and their reduced sweating contributes to overheating [48]. Interestingly, although Heat Shock Factor (HSF) genes are not encoded on chromosome 21, individuals with DS often display an amplified Heat Shock Response (HSR)[3, 48].

Exposure to external stresses, such as temperature fluctuations, can harm proteins, prompting cells to initiate a HSR. The HSR is an adaptive cellular response to various stressors, activating thermotolerance, a mechanism that deactivates Heat Shock Proteins (HSPs) to combat stress. Induced by stress, HSPs counteract protein denaturation, aggregation, and subsequent cell death, thereby playing a vital role in thermotolerance and preventing stress-induced cellular demise. This understanding is crucial for therapeutic strategies related to hyperthermia. The primary regulatory element of the HSR is heat shock transcription factor 1 (HSF1)[127, 99], a transcription factor encoded on chromosome 8.

The amplified response in Down syndrome may be attributed to the chronic activation of the interferon response, closely tied to the DS condition (See Chapter 2 for further details). Genes for interferon receptors reside on chromosome 21, and the resultant gene dosage effect in DS elicits chronic stress. This might initiate the HSR even in the absence of HSF1 gene overexpression. Furthermore, there's a noted overexpression of Hsp70, a HSR chaperone, in individuals with DS[179]. Though Hsp70 is not encoded on chromosome 21, its heightened presence might be a compensatory action to the persistent cellular stress in DS, thereby strengthening the HSR.

1.4.1 Classical Heat Shock Response Pathway

The HSR is a cellular defense mechanism activated in response to various stressors, ensuring survival by maintaining homeostasis. Typically, a temperature increase of 5 to 10°C above the optimal triggers a pronounced HSR in most organisms. While elevated temperatures are a common trigger, the HSR can also be activated by other stimuli such as protein misfolding, oxidative stress, and viral or bacterial infections. These stressors can lead to a temperature increase leading to a higher concentration of unfolded proteins. When these unfolded proteins are detected, molecular chaperones release previously bound and inactive Hsf1 monomers. These monomers then travel to the cell's nucleus, forming Hsf1 trimers. The trimers are activated by phosphorylation and activate transcription of HSP[147, 168] (Figure 1.8). These HSPs are molecular chaperones that assist in protein folding, assembly, and translocation, as well as controlling protein secretion. The prominent chaperones include HSP 70 family, HSP 40 and HSP 90 families, the smallHSPs, and the chaperonins[2].

1.4.2 Dysregulation of HSR in Down syndrome

The relationship between Down syndrome and the Heat Shock Response is primarily influenced by the cellular stress that individuals with DS undergo. This stress emerges from the overproduction of proteins and the subsequent misfolding of these proteins[110]. Additionally, the elevated IFN signaling in DS further exacerbates cellular stress, potentially activating the HSR[162]. Although direct connections between the DS pathology and a heightened HSR remain elusive, there is evidence suggesting that the HSR might indirectly influence DS pathology.

The linkage of HSR and DS pathology is observed between the deficient level of Heat Shock Protein Hsp70 and its link to increased risk of Alzheimer's disease pathology in individuals with Down syndrome. These individuals exhibit a deficiency in the molecular chaperone protein, Hsp70, which is believed to be correlated with increased neuronal death. Hsp70 serves a protective role, guarding cells against death by inhibiting specific cell-death pathways like the caspase cascade and

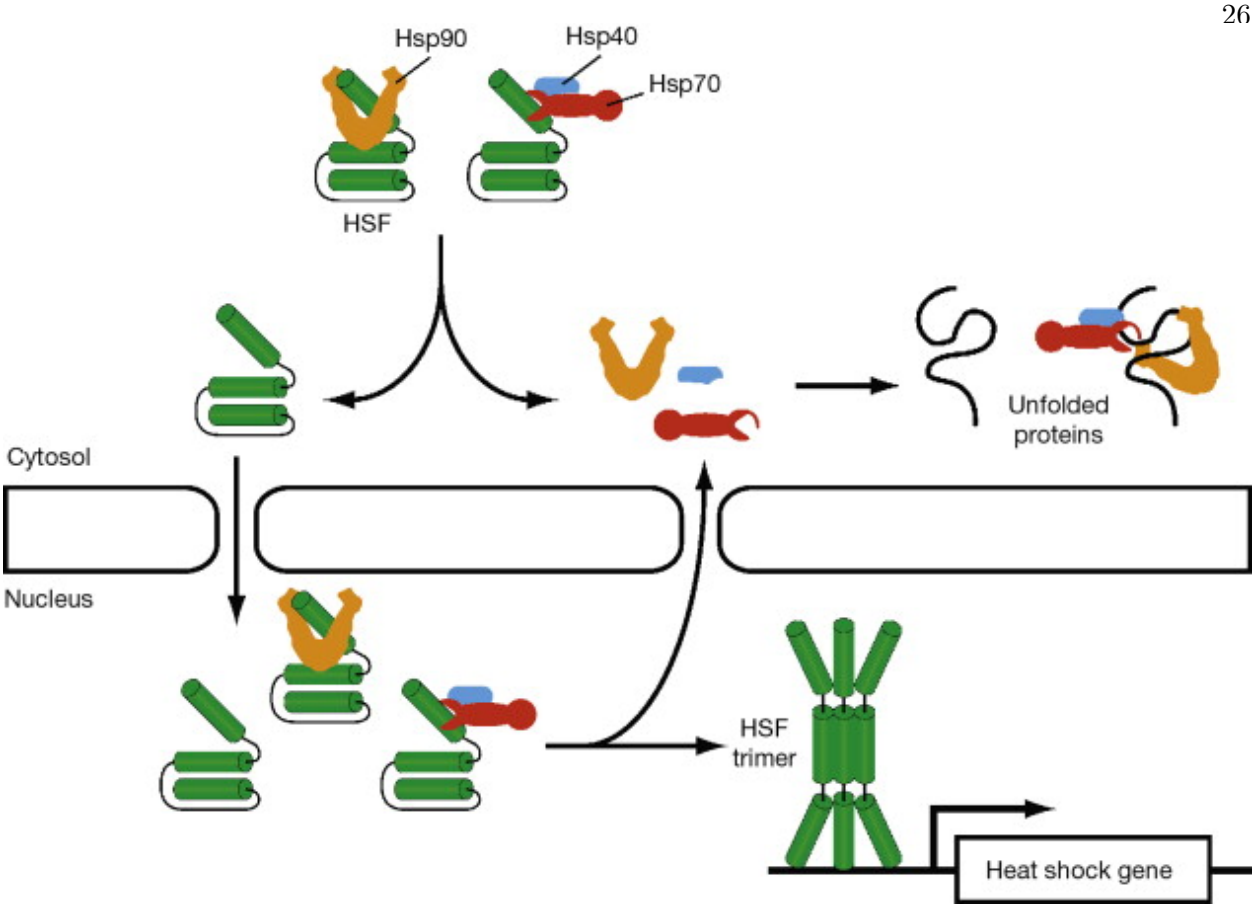


Figure 1.8: **Regulation of heat shock response.** HSR is activated when molecular chaperone proteins detect stress. The increased protein detected leads to monomeric Hsf1 being released from chaperones and transported to the nucleus. The trimeric Hsf1 binds to the heat shock gene promoter in an inert state. The phosphorylation of trimeric Hsf1 activates the transcription of heat shock genes. (Illustration from Thibault and Ng 2013)[168]

Stress-activated protein kinases/Jun amino-terminal kinases (SAPK/JNK) signaling. Consequently, a deficiency in Hsp70 is linked to increased neuronal death in DS individuals[180]. Furthermore, the brains of these individuals show abnormal patterns in the expression of molecular chaperones, possibly explaining the development of late onset pathologies such as misfolded proteins resembling Alzheimer's disease characteristics (AD-like tangles and plagues) found in many brain regions of adult individuals with DS[180].

HSR is triggered at temperatures exceeding 4°C. The lower thermal threshold for heat shock reflects how inflammatory mediators such as IFN-β, can reduce the induction of the stress response

during viral infections[75]. While our exploration in Chapter 3 does not directly study the interplay between IFN and HSR, we are interested in how exposing cells to a low-grade heat shock (42°C, approximately 5°C above optimal) may differ for a trisomy 21 genotype. In this context, cells may be under chronic stress, but the regulator for HSR is not evidently located on chromosome 21.

1.4.3 Looking forward to Chapter 3

In this research, our objective was to delve deeper into the intricate genetic and cellular dynamics associated with the HSR in DS. By subjecting both trisomy 21 and typical disomy 21 cell lines to an acute heat shock perturbation, we aimed to understand the impact of external stress on DS in the absence of a primary regulator on chromosome 21. While Chapter 2 delved into interferon stress, which is explicitly encoded on chromosome 21, our goal here was to elucidate DS-associated phenotypes with less direct genetic ties, leading us to focus on the Heat Shock Response in Chapter 3.

1.5 Summary

DS arises due to the additional copy of chromosome 21, although the exact mechanism through which the slight increase in the number of certain genes leads to the various associated conditions of DS remains not fully understood. A key approach to understanding the implications of this additional chromosome is to study transcription, the first step in gene expression. This process is profoundly impacted by the trisomy, leading to varied gene expression profiles that can give rise to diverse DS-related conditions. Current studies on populations with DS use various genomic techniques to understand how this additional chromosome affects gene expression. Insights from studying DS also offer perspectives on other conditions, like cancers, that feature aneuploidy.

The present thesis aims to further explore transcriptional dysregulation in DS, with an emphasis on the effects of external stressors. While past research primarily analyzed cells in their typical states, the current investigation delves into how stress responses differ in individuals with DS. Two main stressors are explored: interferon signaling, with four out of six receptors encoded on

chromosome 21, and heat shock, which lacks a chromosome 21-encoded master regulator.

In subsequent chapters, this research will delve into interferon responses, with a focus on IFN-β and its signaling pathways (Chapter 2). The study will then shift to the heat shock response, aiming to understand how a low-grade heat shock might differentially impact cells with trisomy 21, given the backdrop of chronic stress in these cells but with no HSR master regulator on chromosome 21 (Chapter 3).

By combining insights from these chapters, the hope is to paint a clearer picture of how trisomy 21 affects cellular responses to stressors, with potential implications for therapeutic strategies. Lastly, tools developed to classify transcription factors based on their role and how they interact with regulatory regions such as enhancers are discussed in the Appendix (A, B and C).

Chapter 2

Distinguishing the primary and secondary transcriptional response across the population to IFN-beta

This chapter is in preparation for publication, currently as:

J. Westfall, D. Rameriz, R.D. Dowell, and MA Allen. Transcriptional response to interferon and its impact in an interferonopathy model. (in preparation)

2.1 Contributions

The following chapter describes the collective work in the Dowell aNd Allen (DNA) laboratory. The main wet lab experiments were generated by Dr. Daniel Ramirez as part of his doctoral work. I conducted the quality control and subsequent analysis of the sequencing data. Dr. Mary Ann Allen provided significant intellectual and computational guidance on this process. I wrote the majority of the manuscript. Drs. Mary Ann Allen and Robin Dowell reviewed the manuscript for preparation for submission.

2.2 Introduction

The innate immune system plays a pivotal role in defending our bodies against pathogens, particularly viruses. Central to this defense mechanism is the interferon response. When a cell detects a virus, it releases interferons (IFNs), inflammatory cytokines that act as a signal to neighboring cells to activate the interferon response. These IFNs are categorized into three distinct types of interferon: Type I, Type II, and Type III, each with its distinct roles and receptors. Each

has its specific receptors and distinct roles in the immune response. Type I IFN (IFN-I), such as IFN-α and IFN-β, bind to the receptors of IFN-I, IFNAR1 and IFNAR2, triggering a signaling cascade via the JAK/STAT pathway. This leads to downstream immune defense and gene expression of interferon-stimulated genes (ISGs)[158, 1].

Interestingly, not everyone responds to viral infections in the same way. This variability was evident during the recent COVID-19 pandemic, where hospitalized patients showed that IFN-I subtypes have varying potency and are associated with distinct metabolic signatures[64]. Moreover, COVID-19 patients can have heterogeneous responses in their gene expression level of ISG that is cell-type dependent[175]. Having an enhanced IFN-I signaling, as found in individuals with DS[162, 43], further complicates the interferon response. While individuals with Down syndrome generally exhibit fewer infections than typical euploid population[60, 118, 112, 39, 30], when they do contract infections like COVID-19, the consequences are often more severe leading to hospitalization and higher incidence of mortality[21, 81, 37]. A probable reason for the immune dysregulation lies in their genetics; the triplication of chromosome 21 in Down syndrome leads to an overproduction of various IFN receptors. This overexpression can tilt the immune balance, making Down syndrome a kind of interferonopathy. Individuals with DS show signs of immune dysregulation that includes a higher incidence of autoimmune and autoinflammatory disease[31, 47, 27, 93] in addition to elevated inflammatory markers[162, 88]. This dysregulated immune response in the DS population has been attributed to the IFN receptors coded on chromosome 21.

Studying IFN-I response in the context of Down syndrome provides us with a unique perspective on how the interferon response is modulated when there are three copies of chromosome 21. This elevated IFN-I activity has been associated with the presence of four (out of six) interferon receptors, IFNAR1, IFNAR2, IFNGR2, and IL10RB, being encoded on chromosome 21. The extra copies of the four receptors are believed to contribute to the enhanced IFN-I signaling in DS[80, 11, 137, 162]. While Down syndrome was initially identified as a Type I interferonopathy because of the increased IFN-I activity[162], recent research has shown a broader and more complex picture. A recent study comparing variable IFN signaling across multiple cell types derived from individuals with

DS found that the interferonopathy associated with this population is more complicated than simply hyperactivity in IFN-I. Rather, there is a mixture of overexpression of all three types of IFN receptors with a major contribution of type II and type III to the IFN hyperactivity than previous studies showed[65]. The triplication of several IFN receptors on chromosome 21 might skew the interferon response in an unexpected way shedding light on why DS have a unique immune profile. Delving into IFN signaling in DS provides valuable insights into this population's distinct immune response, enhancing our understanding of this pathway and potentially uncovering immune mechanisms applicable to both DS individuals and the general population.

Most studies, to date, have examined the transcriptional profiles of tissues and cells derived from individuals with DS under baseline conditions[162, 165, 103, 69, 72, 120, 110]. To further our understanding of the complexity of IFN response, we are adopting a novel approach. Instead of examining just the baseline immune response, we will be introducing an external perturbation: treating cells with IFN-I cytokine IFN-β. Although IFN-β is known to induce a cellular inflammatory response, distinguishing the primary transcriptional response from the secondary response across the human population remains unclear [122, 156, 37, 70]. Thus, we will leverage nascent RNA assays (as a primary response) in conjunction with RNA-seq (to measure secondary response) to provide a comprehensive view of the transcriptional landscape under IFN-β treatment including investigating the immediate-early response. Through this innovative approach, we aim to unravel the complexities of interferon responses across a diverse population that includes individuals with DS, thereby furthering our understanding of this genetic condition.

2.3 Result

2.3.1 Measuring immediate-early and subsequent response to IFN-beta

In order to gain insight into the transcriptional and gene expression response of human cells to interferon stimuli, our study aims to characterize the transcriptional response in human lymphoblastoid cells upon exposure to type I interferon. Specifically, we will incubate lymphoblastoid

cells with interferon beta (IFN-β), a cytokine renowned for its immunomodulatory, antiviral, antitumor, and anti-inflammatory effects. To this end, we used lymphoblastoid cell lines derived from a cohort of ten diverse individuals varying in gender, ethnicity, genotype, and age (Table 2.1).

Table 2.1: **Lymphoblastoid Cell line Information.** Ten lymphoblastoid cell lines (LCLs) were used to generate the intrahuman datasets. Seven of the LCLs were from Coriell Institute for Medical Research (NIGMS Repository) and three LCLs from the Linda Crnic Institute for Down syndrome (AB Nexus program). The table contains information on the LCLs including the internal ID used in the lab (these names are not the real name of the origin), the Short Read Archive (SRA) ID, the country of origin, the ethnicity, the biological sex, and the cell line source. Additionally, the table is color coded so that typical individuals are colored violet and individuals with Down syndrome are colored orange. F: Female, M: Male; D21: Disomy 21, T21: Trisomy 21

Internal ID	SRA	Genotype	Sex	Source	Country	Ethnicity
Khaondo	GM19024	D21	F	Coriell / NIGMS	Kenya	Luhya
Niyilolawa	GM18489	D21	F	Coriell / NIGMS	Nigeria	Yoruba
Srivathani	HG03645	D21	F	Coriell / NIGMS	Sri Lanka	Tamil
Ursula	GM12878	D21	F	Coriell / NIGMS	United States	Caucasian
ChenChao	GM18530	D21	M	Coriell / NIGMS	China	Han
Dave	TIC0001672	T21	M	Nexus Biobank	United States	NA
Ethan	172	T21	M	Nexus Biobank	United States	NA
Eric	259	D21	M	Nexus Biobank	United States	NA
Pedro	HG02150	D21	M	Coriell / NIGMS	Peru	Peruvian
Sengbe	HG03077	D21	M	Coriell / NIGMS	Sierra Leones	Mende

Interferons are known to stimulate an immediate early response followed by a delayed induction of additional genes. To capture both of these processes, we leverage two types of RNA sequencing protocols: nascent RNA sequencing at 60 minutes after IFN-β (to capture the immediate early response) and steady-state RNA-seq (to capture the subsequent induction) at 180 minutes after IFN-β. (Figure 2.1). Hereafter, we refer to the changes observed in PRO-seq as the primary response whereas subsequent changes detectable in RNA-seq are secondary, as they are induced more slowly, at time scales that suggest they require cellular protein synthesis. See Methods (Section 2.5) for complete details of the experiment.

All cell lines were incubated with interferon beta (IFN-β) or BSA as a negative control for

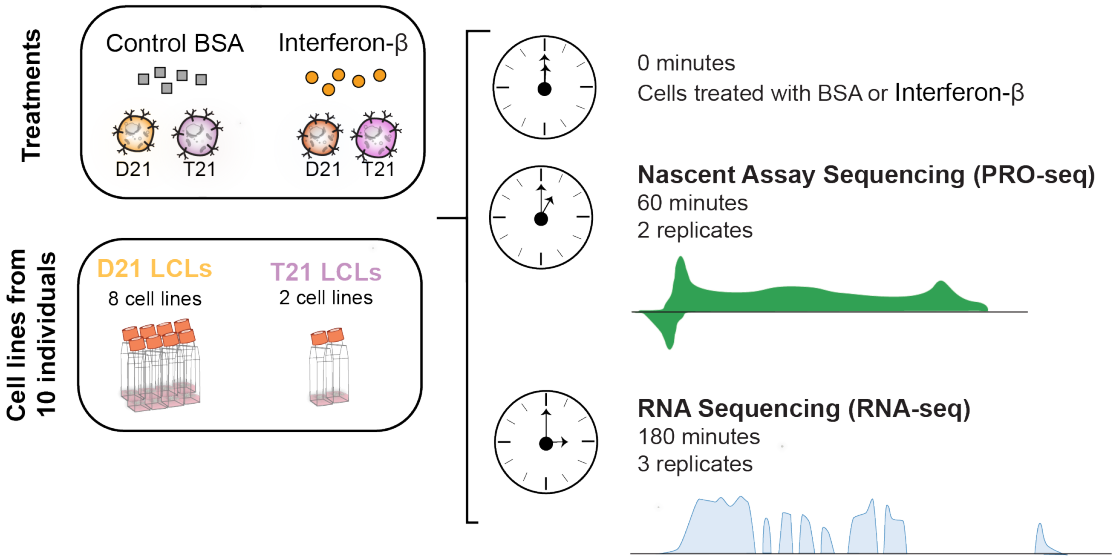


Figure 2.1: **Experimental design.** Cartoon depicting the experimental design. Lymphoblastoid cell lines were exposed to IFN-β or negative control BSA for 60 minutes or 180 minutes for PRO-seq (green) and RNA-seq (blue), respectively.

both RNA sequencing assays. PRO-seq was obtained in duplicate, whereas RNA-seq was assayed in triplicate. The PRO-seq samples were sequenced to an average total depth of 41.7 M reads per replicate for the PRO-seq assay (Table S2.4). Because two of the PRO-seq libraries had low complexity libraries (HISAT2 total deduplicated reads $< 15\%$), those two individual samples were dropped from all downstream analyses. The RNA-seq samples were sequenced to a total depth of an average 34.7 M reads per replicate and were of uniformly good quality (Table S2.5).

Reads were mapped to GRCh38/hg38 reference genome. The reads mapped to genes using the annotation 'exon' for RNA-seq samples and 'gene length' with the $5'$ of genes truncated by 750bp for PRO-seq samples were counted using Rsubread featureCounts[108]. As expected for these two assays, the PRO-seq mapped to a larger fraction of the genome, with many intron and intergenic genome regions. This is expected as the assay measures transcription pre-splicing (hence introns are included) and recovers all transcripts, including highly unstable intergenic enhancer associated

transcripts. In contrast to PRO-seq, RNA-seq reads mapped mainly to the exonic genome regions and therefore cover less of the genome (Table S2.2).

As an example, we plot two individuals worth of data for ISG20 (Figure 2.2), an interferon-inducible $3'$-$5'$ exonuclease that inhibits replication of several human RNA viruses. ISG20 is modulated by type I and type II IFNs and under the control of the transcription factors IRF-1[71, 49]. We observed that IFN-β perturbation lead to an increase in transcription for both RNA assays. The level of transcription varies between the two individuals, but they showed similar responses. The remaining cell lines also had a similar response to IFN-β (Figures S2.17, S2.18).

2.3.2 No major changes to nascent transcription profiles in Down syndrome

Widespread gene dysregulation has been observed in multiple studies of Down syndrome[162, 88, 67, 174, 80]. The altered transcription profile includes not only dosage induced chromosome 21 differences[78], but also changes in the expression of thousands of genes across all chromosomes. It has been speculated that this dysregulation may arise from a fundamental shift in RNA polymerase II activity. The proposed shift is thought to be mediated through DYRK1A, a kinase encoded on chromosome 21 that is known to phosphorylate RNA polymerase II. Thus we first sought to ascertain whether the two samples from individuals with Down syndrome displayed any categorical shifts in their nascent transcription profiles.

To address this question, we first sought to determine whether the overall activity profile of RNA polymerase II has shifted. In nascent, genes have a fairly generic profile of reads that corresponds to three phases of RNA polymerase II activity. RNA polymerase II initiation leads to bidirectional transcription at the $5'$ end of the gene, corresponding to the transcription start site. Within the body of the gene, RNA polymerase II has fairly consistent processivity leading to a nearly uniform signal. After the polyadenylation site, RNA polymerase II is thought to slow as part of the termination process, leading to a rise in nascent signal followed by a steady reduction. By calculating metagenes (the average profile over a collection of genes) we can recapitulate these signatures at genes within our nascent data (Figure 2.3). Importantly, our metagene show no

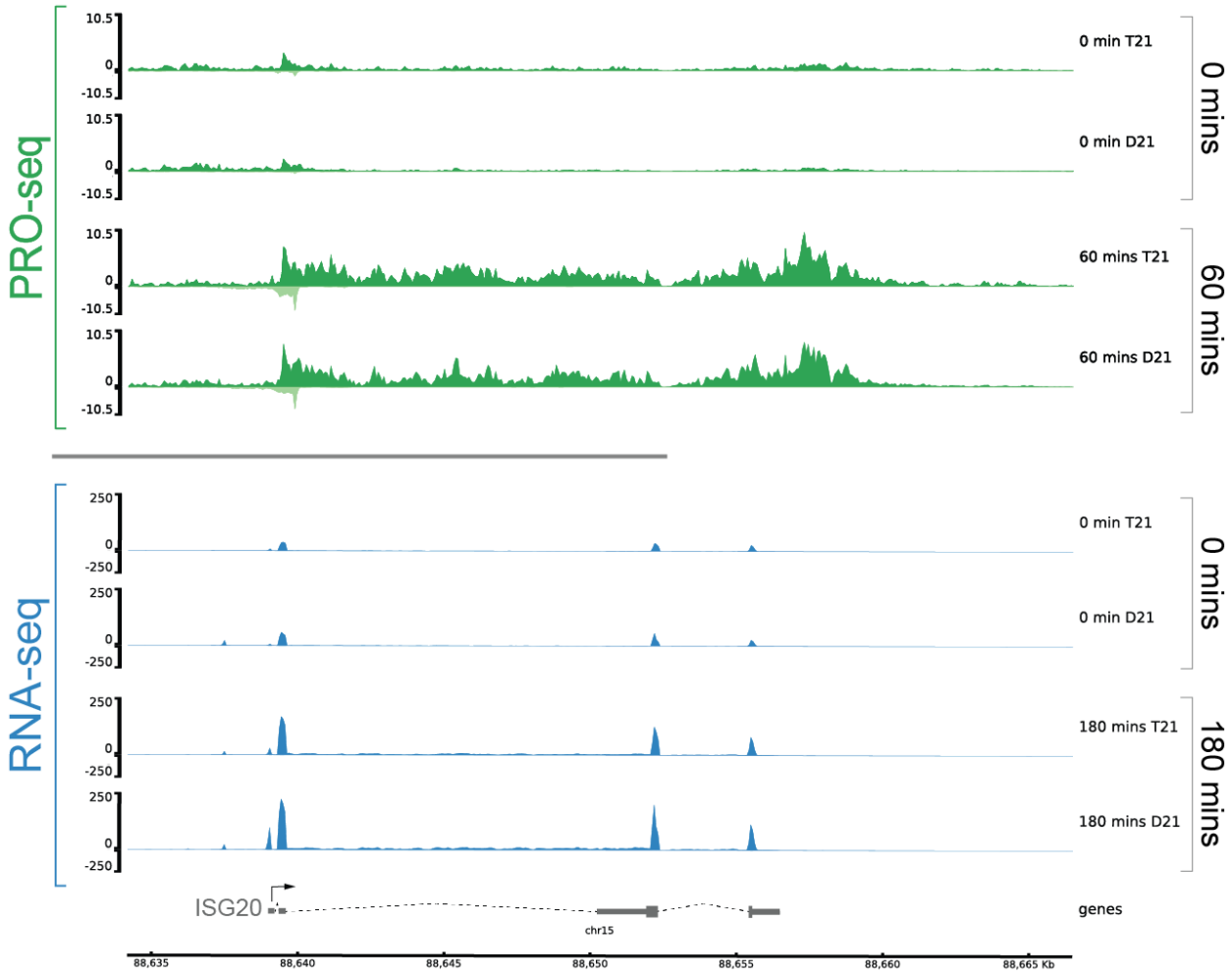


Figure 2.2: **ISG20 response to IFN-β.** Two individuals show response to IFN-β at the ISG20 gene (chr15: 88,635,670-88,656,483). The top four tracks (green) show two cell lines (Internal IDs: Dave and Khaondo) at 0 and 60 minutes post IFN-β stimulation. The next four tracks (blue) are the same two cell lines at 0 and 180 minutes post IFN-β stimulation.

discernible difference between T21 and D21 in either the elongation or termination region. We do see a modest difference in the initiation region, where the shape of the bidirectional is identical but with slightly elevated transcription in the T21 samples. However, this difference is driven almost exclusively by the 'Dave' sample and dissipates when compared to a larger collection of D21 samples (not shown).

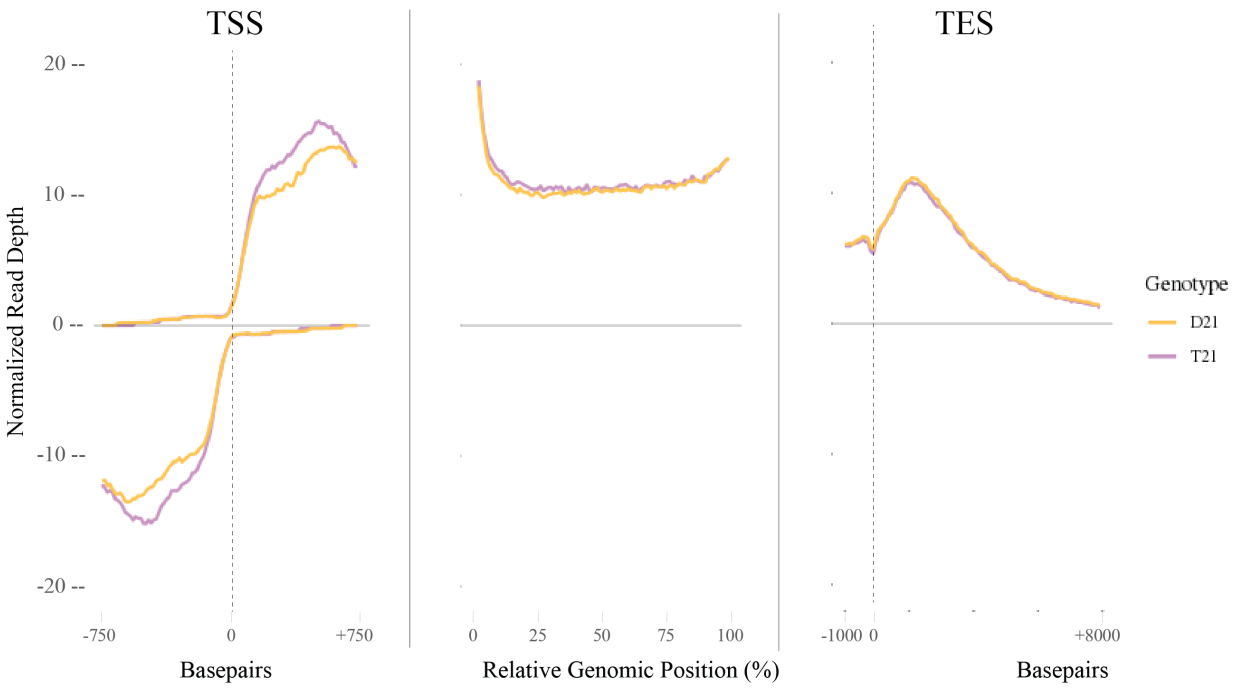


Figure 2.3: **Nascent RNA metagene profile under baseline condition.** A metagene summarizing the typical nascent transcription profile of trisomy 21 (Dave and Ethan) compared to euploid disomy 21 (Khaondo and Eric). The metagene summarizes signals over approximately 5500 genes (of which 50 are on chromosome 21) in the baseline conditions (BSA), selected for adequate transcription across all four individuals. Left: Initiation region profile ± 750 bps of the annotated transcription start site (TSS). Center: Elongation region profile (+750 to -1000 of the annotated end), normalized to percentage of the region. Right: Termination profile surrounding the annotated end of the gene (labeled TES; -1000 of the TES to +8000 bp).

We next sought to determine whether the T21 samples had more sites of RNA polymerase II (RNAP II) initiation genome-wide. Transient, short, non-coding RNAs are produced at sites of

transcription initiation, an inherent aspect of the bidirectional nature of RNAP II activity[32, 87, 140]. The bidirectionals RNAs associated with initiation are seen at both enhancers and promoters and can be identified by Tfit[17] in these samples (Figure 2.4). Importantly, Tfit will identify sites of RNA polymerase II initiation even when they are unidirectional (e.g. the 'bidirectional' has zero reads in one of the two directions). Generally, the term 'enhancer RNA' (eRNA) is used to refer to bidirectionals not at the 5′ end of annotated genes. In some nascent sequencing papers, bidirectional regions are also referred to as transcribed regulatory elements (TREs). To access the number of initiation sites across samples, Tfit was used to call bidirectionals which were then merged across samples using muMerge[150]. PRO-seq reads were then counted over all bidirectionals detected in the collection of samples. While the precise number of detected bidirectionals varies across the individuals, there does not appear to be an increase in the number of bidirectionals identified in the T21 samples(Figure 2.5).

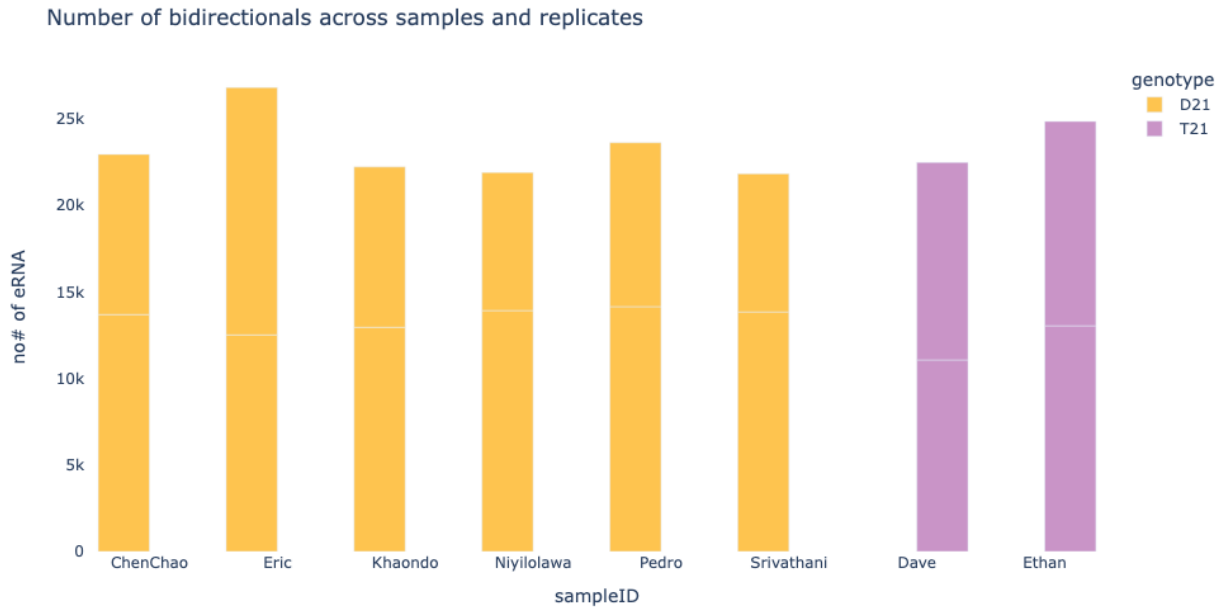


Figure 2.4: **Bidirectionals detected in each individual.** The total number of bidirectionals captured in each individual sample, as defined by Tfit, muMerged across replicates.

The current gene-dosage hypothesis of trisomy 21 argues that all genes encoded on chromosome

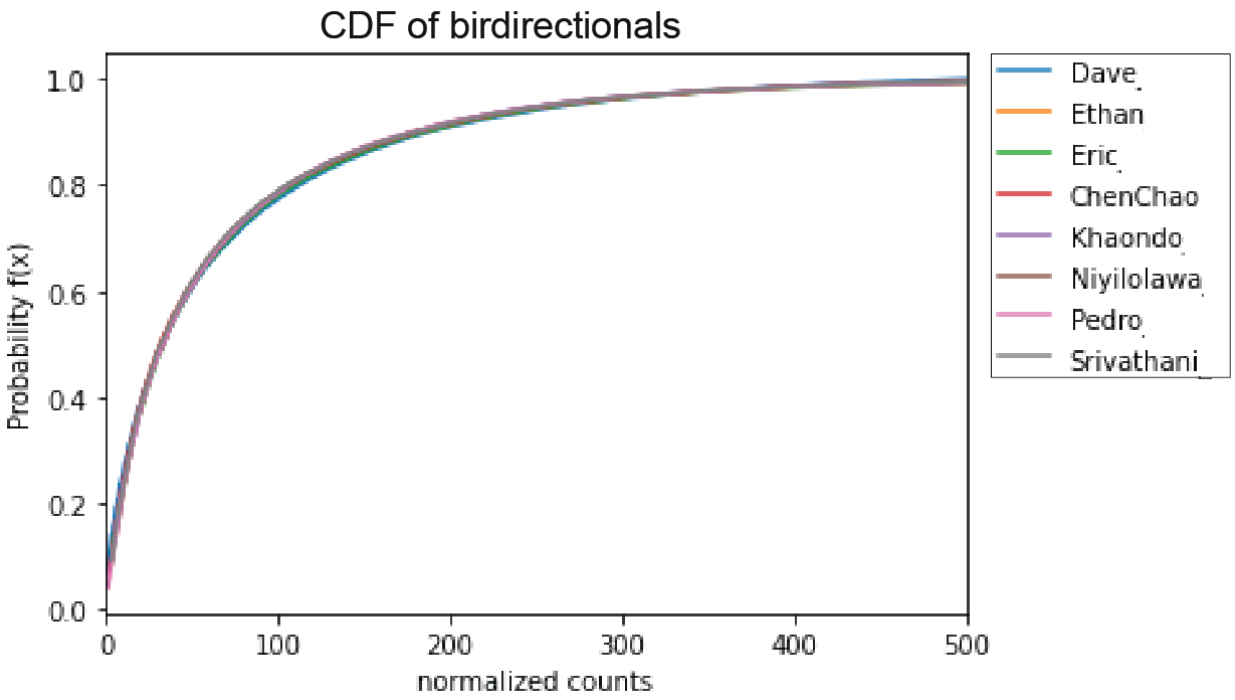


Figure 2.5: **CDF of bidirectionals in each individual.** CDF plot of bidirectionals across individuals in baseline condition (BSA) demonstrate that the individual samples are calling roughly the same bidirectionals. No individual sample has altered calls of bidirectionals.

21 are transcribed at DNA dosage levels[78]. Many of these increases persist at the protein level[80]. Therefore, the altered transcription profiles observed in T21 cells arise from changes in the dosage of a small number of transcriptional regulators encoded on chromosome 21.

2.3.3 An interferonopathy model for Down syndrome

Previous studies have found that the four chromosome 21 encoded IFN-Rs are expressed at higher levels in primary cell lines from individuals with DS[162]. Consistent with previous studies, we find all four of these genes are expressed at higher levels in the T21 samples (Figure 2.6). These genes were also transcribed at higher levels in nascent transcription (Figure S2.20). We notice however that the two interferon receptor genes not encoded on chromosome 21, IFNGR1 and IFNLR1, had similar transcription levels (at 60 minutes) and expression levels (at 180 minutes)

across all individuals in the basal state (Figures S2.20, S2.19).

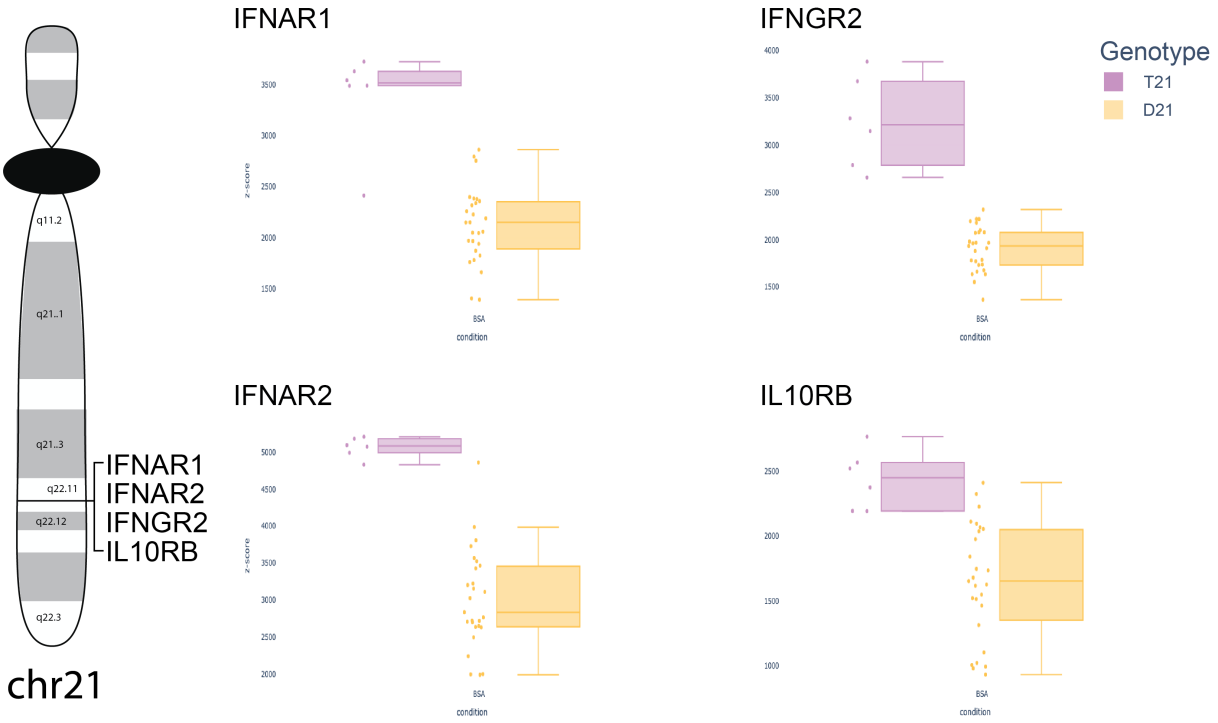


Figure 2.6: **Chromosome 21 with IFN receptor genes labeled.** Chromosome 21 encodes four of the six interferon receptors. T21 individuals have a higher mean level of expression (RNA-seq) in control conditions (BSA) compared to D21 for these receptors, consistent with the interferonopathy model.

Given that the six receptors did not show consistent elevated levels, we next turned our attention to the IFN-score, a metric that measures the overall IFN response of a cell line, a method described in Galbraith 2022 [64]. The metric summarizes the transcription status of a set of annotated IFN response genes for each cell line. We observed that the IFN score for one of trisomy 21 cell lines is much higher relative to all the disomy 21 cell lines. The other trisomy 21 cell line had a comparable IFN score suggesting that trisomy 21 cell lines have variable transcription levels. However, it is worth noting that the trisomy 21 cell line with a lower IFN score is still higher than its age and gender matched sibling. This finding suggests that trisomy 21 genotypes have a variable base IFN transcription state that depends on their underlying genotype (Figure 2.7).

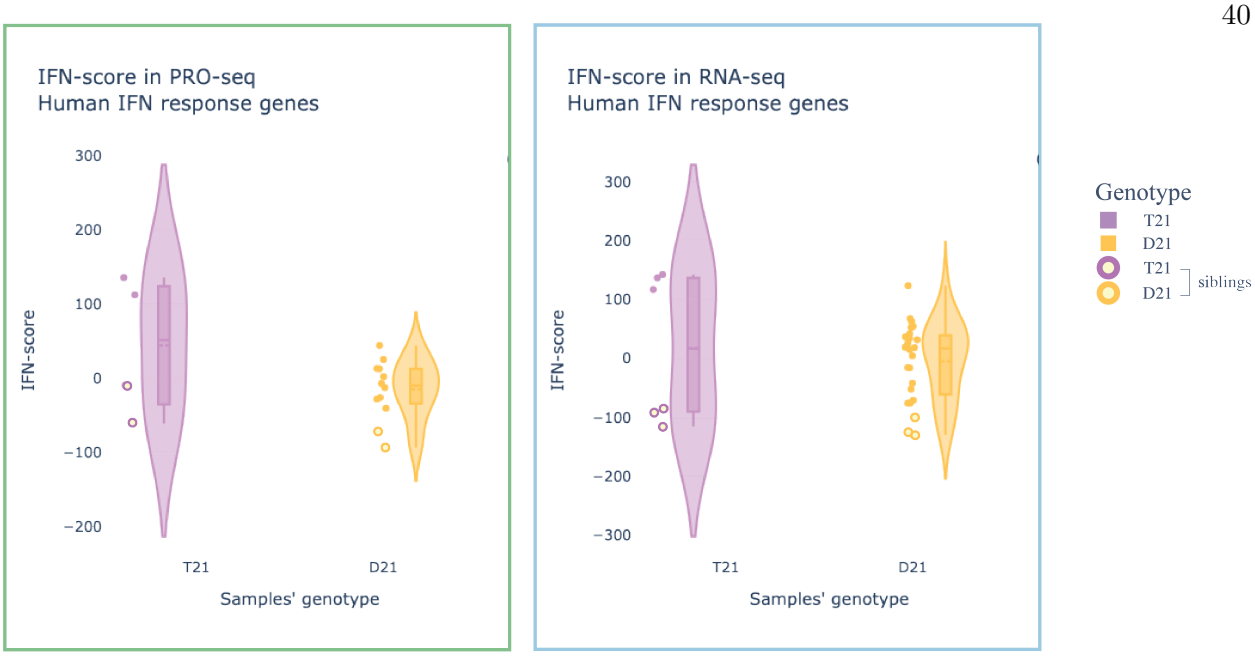


Figure 2.7: **IFN-score for samples.** Calculated IFN-score[64] for all samples between the two assays. T21 samples showed a higher variance in their scores compared to D21 samples.

Interestingly, we wondered whether the increase in IFN-score was more than what would be expected given the DNA dosage hypothesis. To address this question, we did a DNA copy number normalization[78] and re-evaluated the IFN-score. We found that once the DNA copy number was accounted for the IFN-score in trisomy 21 was lower than for disomy 21. (Figure S2.24). This suggests that the DNA count skews the normalized read counts and fold change estimates calculated in DESeq2, which null hypothesis expects a fold change between two samples is 1.0 versus 1.5 in T21.

Thus our data supports the interferonopathy model of Down syndrome and extends the observed dysregulation to nascent transcription. While extensive work has been conducted examining the levels of IFN genes in populations of individuals with Down syndrome[64] and their response to infection[60, 21, 81], very little is know about the immediate transcriptional response to interferon directly. In our case, we use IFN-β to stimulate a population of cells, including two cell lines with T21. Thus we next consider whether the T21 cell lines respond to IFN-β in a manner that is

different from the euploid disomy 21 population.

In our analysis, we wanted to understand the impact of chromosome copy number (T21 versus D21) on gene expression after 180 minutes of IFN-β treatment. To do this, we utilized the DESeq2 likelihood ratio test (LRT) to identify genes that react differently to IFN-β treatment in T21 cells compared to D21 cells. Before applying the LRT, we excluded genes with no reads (TPM = 0) in any of the samples. This step helps to reduce noise and provides a more accurate estimation of our model parameters. From the initial 14,907 genes, we compared a comprehensive model (considering genotype and treatment as potential factors affecting gene expression) to a reduced model (considering only treatment). Our analysis identified 3,857 genes (significant threshold padj < 0.01 and an absolute log2FoldChange > 1) where the genotype has a significant effect on the gene expression in the context of treatment. Interestingly, while T21 cells showed a different baseline gene expression compared to D21 cells under control conditions, both DNA copy numbers exhibited a similar response to treatment (see Figure S2.25). Despite T21 starting at a different baseline expression, the magnitude of change after treatment was largely consistent with D21 (Figures 2.8, 2.9). This could be due to the extra chromosome in T21 contributing to baseline levels. Alternatively, certain regulatory mechanisms or intrinsic 'ceiling effects' might standardize responses regardless of starting levels.

Given that only 13% of the total genes showed differential expression in our LRT analysis, we expanded our focus. We wanted to explore genes influenced by chromosome copy number, treatment (0 and 180 minutes IFN-β), or the interaction of chromosome copy number and treatment. For this, we created a subset of genes showing significant expression differences at 0 and 180 minutes post-IFN-β treatment. We excluded genes already identified in our LRT analysis. A heatmap of these gene expression changes across individuals revealed a consistent pattern for the majority of individuals (Figure 2.10). Notably, any variation observed was predominantly on individual genetic variations rather than the presence or absence of an extra chromosome 21.

The genes were separated by the direction of regulation, and then further separated by the genotypes and conditions. We asked if the transcription levels after cell lines are incubated with

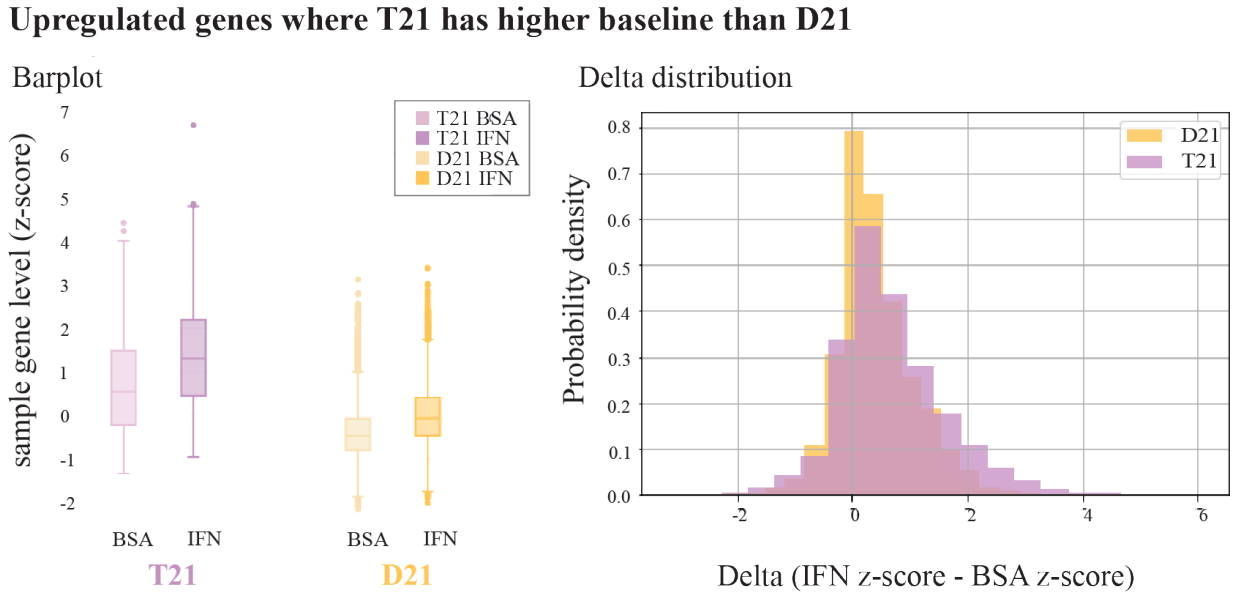


Figure 2.8: **Upregulated genes with higher baseline in T21 defined by LRT.** The likelihood ratio model identifies a group of genes (n=862) that are upregulated when responding to IFN-β in a genotype dependent manner. The T21 has an elevated baseline (BSA) and increases in level after IFN-β treatment. Trisomy 21 in purple, disomy cells in orange.

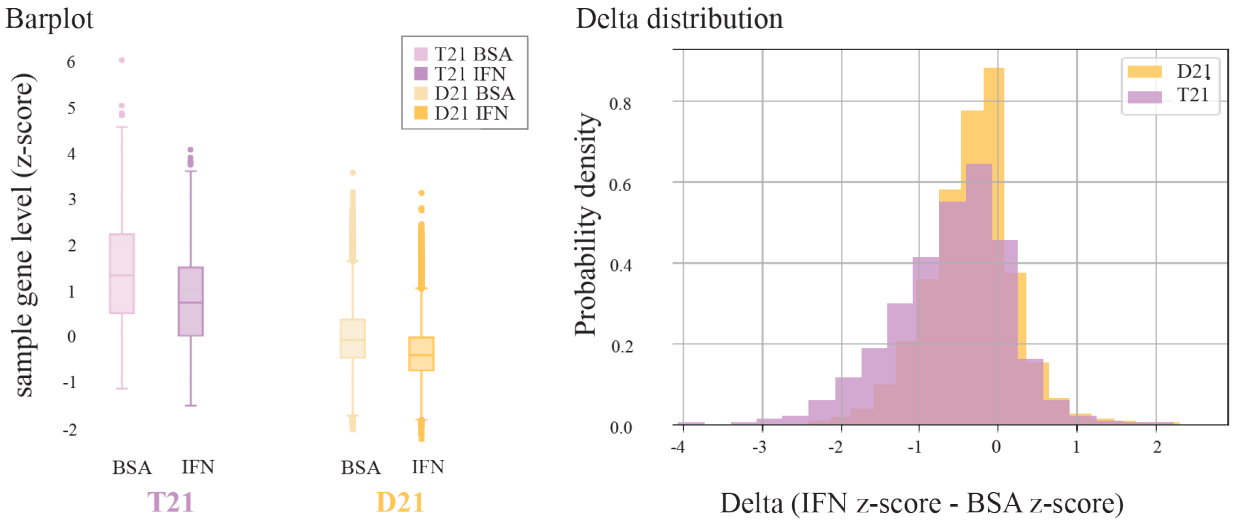


Figure 2.9: **Downregulated genes with higher baseline in T21 defined by LRT.** The likelihood ratio model identifies a group of genes (n=1183) that starts at a higher baseline level and decreases in level when responding to IFN-β in a genotype dependent manner. Trisomy 21 in purple, disomy cells in orange.

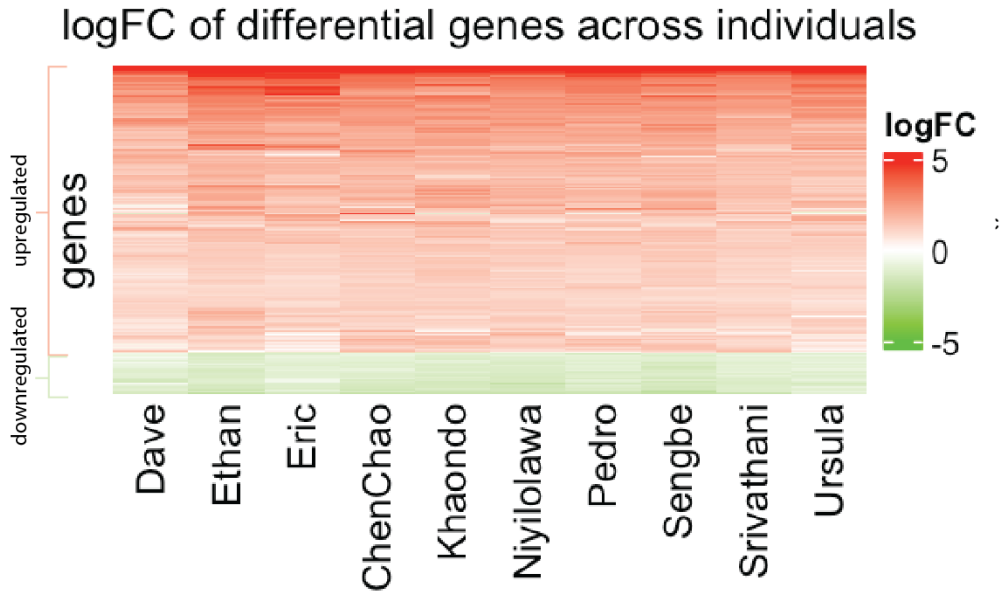


Figure 2.10: **Heatmap of significantly differential expressed genes in individuals.** Heatmap displaying the log2 fold change of genes found significant in the differential expression analysis. Each column represents an individual. While the majority of individuals exhibit consistent patterns in gene expression changes, variations are observed predominantly on an individual basis, rather than being associated with the presence or absence of an extra chromosome 21.

IFN-β are higher in the trisomy 21 genotype. The delta in both genotypes was similar when looking at IFN-β perturbation and negative BSA conditions. When we compared transcription levels of genes we see that although the trisomy 21 cell lines have a higher transcription level in BSA, after the cells were perturbed the trisomy 21 did not have a larger magnitude shift in transcription level (Figure 2.25).

2.3.4 A population response to IFN-beta

We next sought to examine the population response to IFN-β, using DESeq2 to identify genes with significant changes in response to IFN-β stimulation. The analysis was executed completely independently for each individual, focusing solely on the effect of treatment (comparing IFN-β to negative control). Genes were deemed significantly differentially expressed at an significant threshold of adjusted p-value < 0.01. For each individual, we identify a large number of genes as upregulated

in PRO-seq at 60 minutes which gives rise to a larger response at 180 minutes in RNA-seq (Figure 2.11. All the individuals' responses following perturbation showed a similar magnitude of response; With 298-576 upregulated genes in PRO-seq and 458-699 upregulated genes in RNA-seq (Figures S2.21, S2.22).

Figure 2.11: **Immediate-early and subsequent response to IFN-β in Dave** An example of the MA plots (top) for PRO-seq (left, green) and RNA-seq (right, blue) for the Dave individual (Trisomy 21) comparing IFN-β perturbation to control. Recall that PRO-seq is a comparison between 0 and 60 minutes; RNA-seq between 0 and 180 minutes. Genes are colored blue if they meet the significance threshold of adjusted p-value < 0.01 but are not annotated as an ISG. Genes are colored orange if they are annotated as an ISG. Below are example volcano plots for the same comparisons.

To validate the IFN-β exposure activates a consistent set of transcription factors across the individuals, we used TFEA[150] on the PRO-seq data. TFEA looks for over-enrichment of TF motifs with sites of differential transcription. We found that the upstream transcription regulators are those expected for IFN-β exposure, namely TFs involved in interleukin pathways activated during inflammatory and immune responses as well as the growth hormone receptor signaling pathway via

JAK-STAT (Figure 2.12). The regulators are consistent across all the cell lines, with the exception of the Khaondo cell line samples, which could be either biological or technical. The MA plot of the enrichment score for this individual (Figure S2.23) tightens on the x-axis as the mean expression increases suggesting a cleaner dispersion pattern.

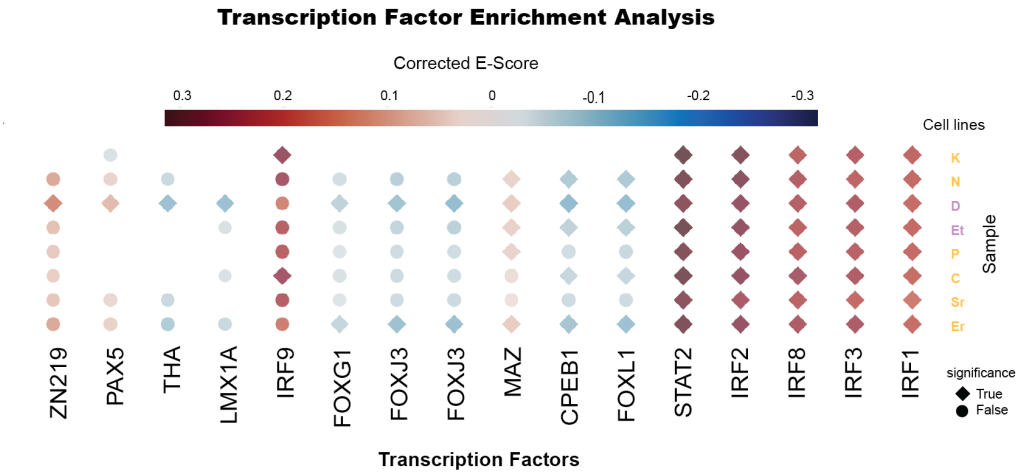


Figure 2.12: **TF enrichment analysis (TFEA) across cell lines.** Transcription factor enrichment analysis (TFEA)[150] identifies a consistent set of transcription factors (TFs) present across most cell lines. Notably, STAT2, a TF regulated by IFN-β and IFN-α that is activated within the JAK-STAT signaling pathway, exhibited the highest enrichment score in all the cell lines. interferon regulatory factors, a family of TFs essential for regulating IFN and associated immune response to viral infection, cell growth, and differentiation, consistently showed strong enrichment scores across all cell lines. Other TFs with significant scores for most cell lines are mainly associated with interleukin pathways that are activated during inflammatory and immune responses, as well as the growth hormone receptor signaling pathway through JAK/STAT. Khaondo, a cell line was not significant for TFs with negative enrichment-score may be biological or technical, as this cell line MA plot suggests higher quality data relative to the other cell lines.

We wanted to compare the signals from both sequencing assays across individuals, focusing initially on those genes with some signal in every individual. Thus we first remove any genes in which there was an individual sample with a low mean normalized count indicated by the NA value in the adjusted p-value column which reduced the number of total genes analyzed from 28,266 to 19,678 genes. To compare the signal of a gene transcribed at 60 minutes to its corresponding expression at 180 minutes, we excluded genes that were found in a single sequencing assay which

took our total gene analyzed to 12,242 genes. This filtering only removed 101-262 statistically significant genes per individual (Figures S2.26, S2.27).

We recognize that for each individual we were constrained to two replicates per condition and that this would reduce our statistical power to estimates of gene expression changes. To account for this we leverage the use of all the individuals as replicates to get a population-wide differential expression analysis. We anticipated that for low expressed genes, the higher variability may be harder to decipher if a gene is significantly differential expressed with a stringent cutoff hence we used a looser significant threshold (adjusted p-value < 0.1) in the population-wide analysis. We used the population to remove false positives from our individual differential expression analysis, reasoning that genes that respond in only one person and only one assay were likely to be false positive with regards to their response to IFN-β. Therefore these genes that are found only in one person but not in the population were removed from all subsequent analyses.

Interestingly, the vast majority of the genes removed for this criteria were present in only one comparison and were typically down-regulated. This suggests that the genes that are down-regulated by IFN-β are either very individual specific (and of low effect size and washed out in the population comparison) or the determination of IFN-β down-regulated genes after short-term IFN-β exposure are more likely to be false positives (Figure S2.28).

When we considered the number of genes that are statistically significant in either sequencing assays in at least a single individual, we found a total of 1,146 upregulated genes and 2,189 down-regulated genes (Figure 2.13). We note that the majority of the upregulated genes showed a consistent response across all individuals.

2.3.5 Temporal dynamics of IFN-beta stimulation

By leveraging two RNA sequencing protocols, we categorized genes based on their response timing, differentiating between immediate-early response genes and those activated later following the synthesis of early gene proteins. The three categories of responsive genes are transient, direct, and secondary. Genes that respond at 60 minutes (PRO-seq only), was termed transient (average

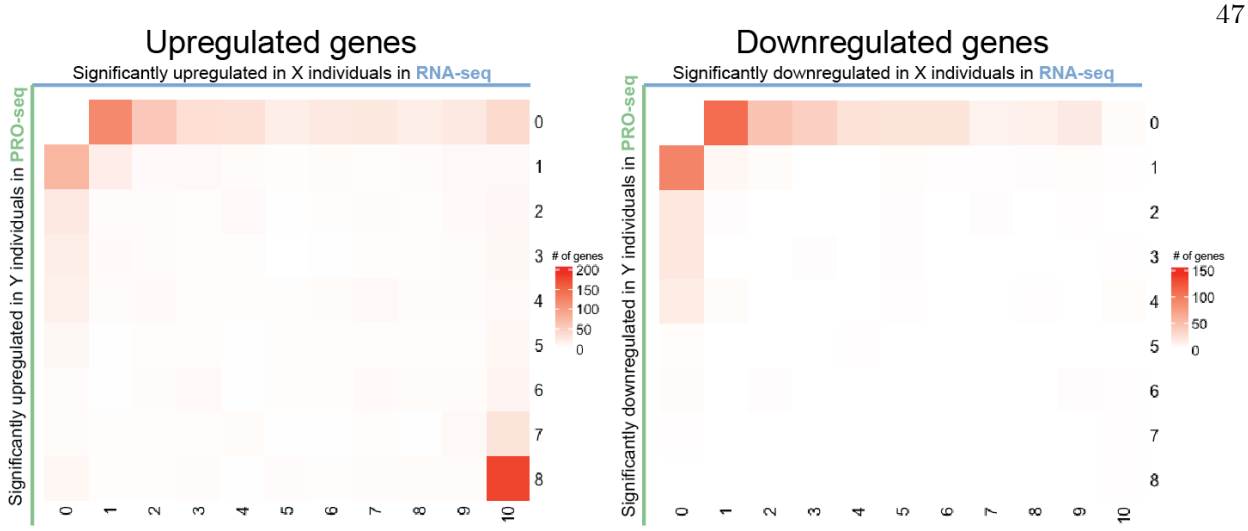
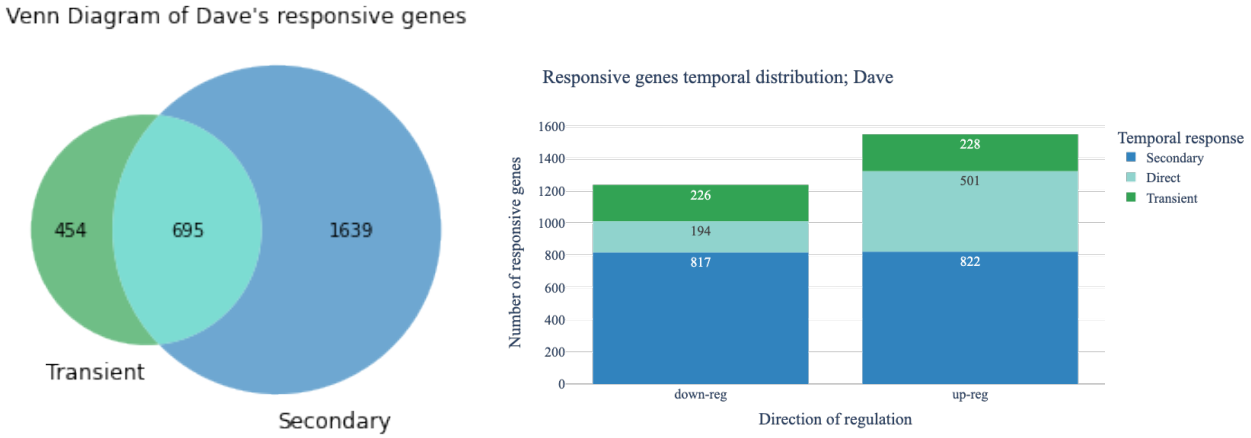


Figure 2.13: **Population response to IFN-β.** Across the population, the majority of genes are upregulated (left heatmap) in all individuals and both assays. The downregulated genes (right heatmap) are more variable, with the majority showing individual responses.

358 upregulated, 215 down-regulated), genes that respond at both 60 minutes (PRO-seq) and 180 minutes (RNA-seq) were termed direct (average 379 upregulated, 135 down-regulated), and genes that only respond at 180 minutes only (RNA-seq) were called secondary (1038 upregulated, 970 down-regulated) (Figure 2.14, Table S2.3). We next investigated whether the genes in each category were organized into known gene sets. Since the upregulation of genes often has different functional implications than down-regulation, we argue that separating the genes on the direction of transcription can provide clearer insights into which pathways are being activated or suppressed. This, however, leads to too small of a gene set per individual to gain statistical power. To account for this we considered genes that were common in the majority of individuals (> 5 individuals) (Figure 2.15, S2.29). To evaluate consistency across cell lines, we use GO enrichment in each individual cell line to ask what terms were enriched for. For the transient, the gene set was too small and many of the terms are not significant (Figure S2.30). Our GO term enrichment for direct and secondary temporal response identifies processes associated with interferon response and downstream transcription regulation respectively (Figure S2.31, S2.32).

To find those genes that are differential in a subset of individuals, we employed DESeq2



(a) **Venn Diagram of Temporal Distribution**

(b) **Temporal distribution of responsive genes**

Figure 2.14: **Temporal distribution of responsive genes.** Distribution of temporal gene response in a single individual; Dave. 2.14a Venn diagram shows the total genes found to be transient (60 mins), direct (60 mins, 180 mins), or secondary (180 mins). 2.14b Responsive genes are further divided based on their direction of regulation showing that there are more secondary response genes.

altHypothesis at a lfcThreshold $< log2(1.5)$ to find genes that are not changing. This approach is extremely conservative and enables us to find those gene sets for which we have the highest confidence. To visualize the intersection of differentially expressed gene sets across cell lines, we plotted a heatmap of differentially expressed genes and the associated number of individuals where the gene DE is significant 2.13. We ran the althypothesis analysis for both the 60 minutes time point (PRO-seq) and the 180 minutes time point (RNA-seq). For each gene, we asked who the gene was differential in and at what time point. We found that the highest heat was found in the gene set where the gene is differentially expressed across all individuals in the upregulated genes, suggesting a similarity in response to interferon activation among individuals.

Most of the significantly upregulated genes in response to IFN-β perturbation were captured at both 60 minutes and 180 minutes. However, a small subset of upregulated genes was found to be significant only in a subpopulation of individuals (Figure S2.26). Conversely, only two genes were found to be significant in all individuals for both RNA sequencing assays in the case of significantly down-regulated genes(Figure S2.27). The number of down-regulated genes was relatively smaller after IFN-β perturbation. Among the down-regulated differential genes, the largest intersection and

Upregulated genes across population and the associated temporal response classification

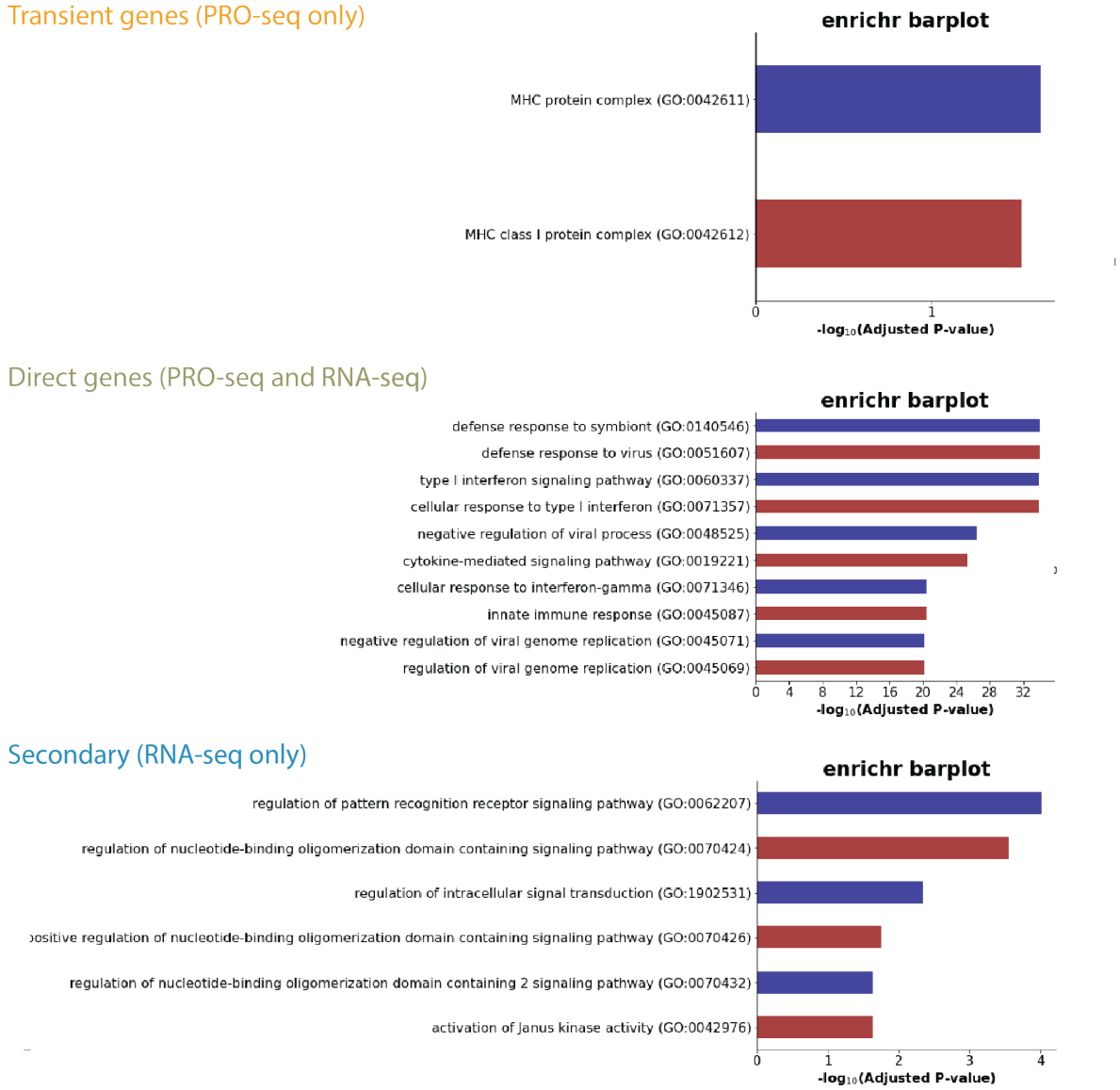


Figure 2.15: **Gene ontology enrichment for upregulated genes in population.** Enrichr tool use GO terms to identify biological pathway and molecular function associated with genes statistically significant in population and classified in temporal response. Upregulated genes classified as transient are related to immediate early response. Direct genes that are upregulated responses to virus and type I interferon signaling. Secondary upregulated genes are enriched for downstream signaling pathways.

the majority of the smaller intersections were true for the steady-state RNA assay (Figure 2.13).

This suggests that IFN-β acts as predominantly an activator, leading to an upregulation of gene transcription.

2.3.6 Individual variation in regulatory regions can influence transcription levels

Overall, the transcription (PRO-seq) and expression (RNA-seq) changes in response to IFN-β were remarkably consistent across the population. However, there were a small number of genes that respond in an individual or subpopulation specific fashion (Figure 2.16). Therefore, we next asked whether these subpopulation differences could be linked to known sequence variations.

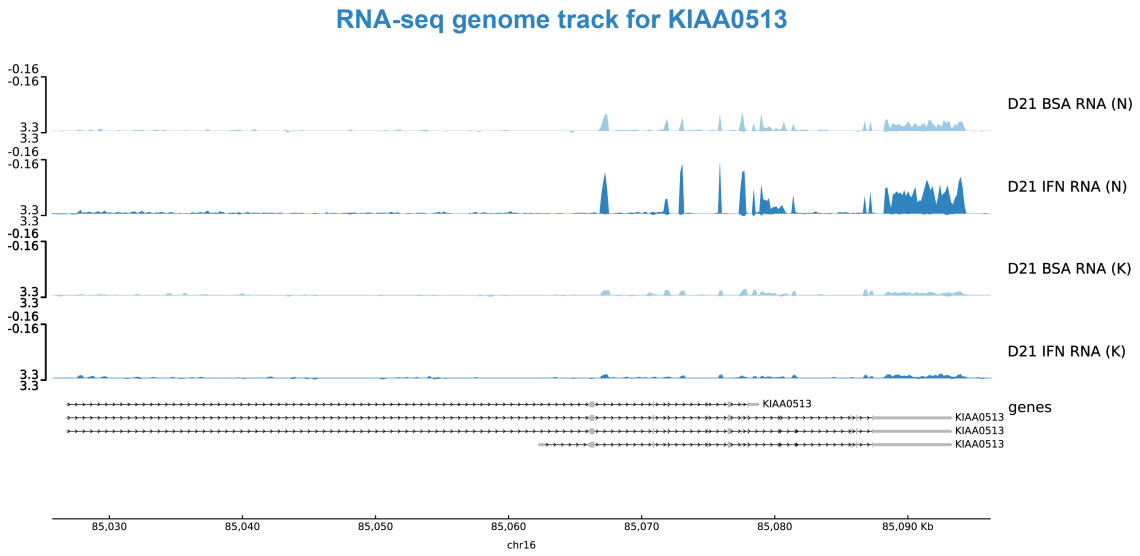


Figure 2.16: **Variable response observed at KIAA0513 in RNA-seq.** The gene KIAA0513 (chr16:85,025,709-85,096,230) is found to be expressed in distinct patterns active after IFN-β perturbation in a subpopulation. Two cell lines are shown: Khaondo and Niyilolawa. See Figure S2.33 for all cell lines.

All ten individuals have existing genomic sequencing data and variant calls. However, for many of the individuals, the depth of the genome sequence is low and therefore the variant calls are quite noisy. We started off with over 74 million SNPs, thus we sought to account for this noise by excluding genetic variations not captured in the dbSNPv138 database[154] in addition to removing any SNPs that did not have annotated genotypes for all ten individuals which reduced the number of SNP to approximately 28.5 million. Additionally, we focused on variants that do not overlap

annotated hg38 genes (approximately 15 million SNPs) and reside near enhancer regions. Our reasoning for this is that SNP in enhancer regions are linked to disease[42] which we reason is their influence in driving transcription of genes. Manually inspecting SNPs that are within ± 50 nts of the center of bidirectional loci from our muMerge master list (approximately 1200), we have visually observed some candidates where there are variances in transcription in a subpopulation of individuals. More work is necessary to further refine these and identify those variants that we believe are responsible for differences in transcription response.

2.4 Discussion

Interferon has been extensively studied in the context of an immune response, primarily utilizing microarrays and steady-state RNA-sequencing in blood cell types [162, 165, 103, 69, 72, 110, 120]. However, it was previously unknown how T21 cells respond to IFN-β transcriptionally, e.g. the immediate-early response. We find that the population, which included two individuals with T21, responds remarkably similar to IFN-β. The majority of the immediate response is to upregulate a large number of genes associated with major histocompatibility (MHC) which are antigens that play a role in responding to viral infection (Figure 2.15). This immediate response is amplified by 180 minutes in steady-state RNA-seq. While there were a limited number of genes that were immediately down-regulated (PRO-seq) and a few in the secondary response (RNA-seq), most of these down-regulated genes exhibited individual specific patterns. These findings suggest that interferon beta stimulation predominantly acts as a gene transcription activator.

Of particular note in our population were two individuals with trisomy 21. Initial expectations were influenced by existing literature indicating that trisomy 21 is an interferonopathy and exhibits increased interferon activation even at basal level [162]. Whether this increased basal IFN activation would translate into an exaggerated or ablated response to subsequent IFN-β stimulation was unknown. Surprisingly, we found that the T21 cells responded remarkably similar to euploid disomic cells when stimulated with IFN-β. We did not identify a significant increase in bidirectionals within individuals with trisomy 21, nor did we observe a higher number of differentially expressed genes.

Overall genes responded with a similar magnitude of change, however many in the IFN pathway did start at a higher level, especially those encoded on chromosome 21.

Given the additional gene dosage resulting from the trisomy 21 genotype, our initial assumption was that there would be significant differences in transcriptional regulation. However, our findings challenge this preconceived notion and suggest that individual differences play a more prominent role in the variation of transcriptional regulation than the trisomy 21 genotype itself. Contrary to our expectations, our findings suggest that individual variations contribute significantly to the diversity of transcriptional regulation in response to interferon stimulation. The alternations seen in T21 cell lines are not solely dependent on the presence of additional gene dosage on chromosome 21. These individual differences may arise from a multitude of factors, including genetic variations, epigenetic modifications, and environmental influences. This realization highlights the importance of considering individual variations in future studies of interferon response and gene regulation. By shifting our focus towards understanding the unique characteristics and contributions of each individual, we can uncover valuable insights into the complex mechanisms governing transcriptional regulation. By embracing the inherent variability within individuals, we can advance our knowledge and develop more targeted interventions and treatments for individuals with immune dysregulation.

Interestingly, only one out of the two trisomy 21 individuals displayed a higher (basal) IFN-score compared to the disomy 21 individuals. The lower IFN score T21 cell line, however, was elevated relative to his age and gender matched brother. Together these two cell lines suggest that while an extra copy of chromosome 21 does lead to an interferonopathy, the extent of the elevated IFN score within the T21 population is likely to be dependent on the overall underlying genotype of the individual. In other words, there are many modifiers of the extent of interferonopathy within the genetic background.

Remarkably, despite the interferonopathy, the T21 cell line showed no greater variability in response than we observed from any two random cell lines. How do we reconcile this finding with the obvious clinical differences in response to infection? In our work, we examined a single cell type (lymphoblastoid cells) to IFN-β for a relatively short time frame (180 min). Clinical responses to

infection involve multiple cell types and prolonged exposure to a variety of signals beyond just IFN-β. The distinct IFN scores of the two T21 cell lines further suggest that the genomic background likely also plays a strong part in influencing how interferonopathy leads to clinical differences. More work, including a larger population, more cell types, and longer time points is necessary to begin to disentangle how interferonopathy leads to an altered clinical response.

2.5 Methods

2.5.1 Lymphoblastoid cell culture conditions

Ten human lymphoblastoid cell lines (LCLs) were obtained; three cell lines from Nexus Biobank (COMIRB 08-1276) at the Linda Crnic Institute and seven cell lines from NHGRI Repository at Coriell Institute. Table 2.1 provides information about the cell lines. The human LCLs were cultured upright in vent-cap T-25 suspension flasks (Corning 430639) in 10 mL RPMI-1640 media (Gibco 72400-047) plus 15% FBS (Gibco 10437-028) and 100 units/mL Penicillin-Streptomycin (Gibco 15140-122). The cells were cultured at 37°C with 5% CO_2. The cells were passaged approximately every two to three days by pelleting the cells via centrifugation (300xg, 5 minutes) and re-suspended in culture media until confluency between 400,000 cells/mL to 800,000 cells/mL.

2.5.2 Interferon perturbation for sequencing assays

2.5.2.1 PRO-seq culture conditions

Each LCLs was cultured in three T-25 flasks for technical replicates. IFN-β (Kingfisher Biotech Ref. RP1788H-100 Lot. KU4428KU) was reconstituted to 200 µg/ml in sterile water. Prior to the cell collection, each cell line was treated with 100 ng/mL IFN-β or with 0.00004% BSA, as the negative control. The nuclei were isolated 60 minutes after incubation with IFN-β. Two biological replicates were processed on two separate dates. Per condition, IFN-β or BSA, the LCL cultures range from 3 to 25 million cells. All cultures and treatments were processed in parallel.

2.5.2.2 RNA-seq culture conditions

For the RNA-seq assays, each LCL cell line was moved onto a separate 48-well plate. The cells were plated at a concentration of 125,000 cells/well and left incubating in a total volume of 250 µL after the 100 ng/mL IFN-β or 0.00004% BSA addition. After the 3 hours IFN-β treatment incubations, 1 mL of RNA lysis buffer was added to the 48-well plate wells for a total volume of 1250 µL, and the plates were stored at -70°C until all three replicates were ready to be processed together. The biological replicates were processed on three different dates. All cultures and treatments were processed in parallel.

2.5.3 Sequencing library preparation

2.5.3.1 PRO-seq nuclei extraction

The LCLs nuclei were isolated as described in [41] with some modifications. The Lymphoblastoid cell lines (LCLs) were collected after 60 minutes IFN-β treatment incubation and washed twice with ice-cold PBS. The cell pellets were resuspended in 6 mL of lysis buffer (0.1% DEPC-DI water with 10 mM Tris-HCl pH 7.4, 2 mM $MgCl_2$, 3 mM $CaCl_2$, 0.5% IGEPAL, 10% Glycerol, 1 mM DTT, SUPERase-IN RNAse inhibitor (Invitrogen Ref. AM2696), and protease inhibitor cocktail (Roche Ref. 11836170001)) and centrifuged for 15 minutes at 4°C at 1000 x g. The pellets were resuspended in 1 mL lysis buffer using Finntip wide orifice pipette tips (Thermo Scientific Ref. 9405163). An additional 4 mL more of lysis buffer was added to the cell suspension and the solution was centrifuged for 15 minutes at 4°C at 1000 x g. The pellets were resuspended a second time in 1 mL lysis buffer using Finntip wide orifice pipette tips, transferred to low binding 1.7 mL eppendorf tubes (Costar Ref. 3207), and centrifuged for 5 minutes at 4°C at 1000 x g. The pellets were carefully resuspended using Finntip wide orifice pipette tips in 500 µL freezing buffer (0.1% DEPC-DI water with 50 mM Tris-HCl pH 8.0, 5 mM $MgCl_2$, 40% Glycerol, 0.1 mM EDTA pH 8.0, and SUPERase-IN RNAse inhibitor), and centrifuged for 2 minutes at 4°C at 2000 x g. The resulting nuclei pellets were resuspended a final time in 110 µL of freezing buffer (0.1% DEPC-DI

water with 50 mM Tris-HCl pH 8.0, 5 mM $MgCl_2$, 40% Glycerol, 0.1 mM EDTA pH 8.0, and SUPERase-IN RNAse inhibitor), using Finntip wide orifice pipette tips. To count the nuclei yield, 10 µL of the resuspended nuclei was added to 990 µL of PBS. The remaining 100 µL resuspended nuclei were snap-frozen in liquid nitrogen and stored at -70°C before being used for the PRO-seq nuclear run-on reactions.

2.5.3.2 PRO-seq library preparation and sequencing

PRO-seq datasets were prepared as described in [55], which in turn is a modified protocol from [116]. Briefly, between 2 to 18 million nuclei per dataset were used for the PRO-seq transcription run-on using a mixture of rNTPs and Biotin-11-CTP (0.025 mM Biotin-11-CTP (PerkinElmer Ref. NEL542001EA), 0.025 mM rCTP (Promega Ref. E604B), 0.125 mM rATP (Promega Ref. E601B), 0.125 mM rGTP at 0.125 mM (Promega Ref. E603B), and 0.125 mM rUTP (Promega Ref. E6021)). 1% of S2 Drosophila melanogaster nuclei relative to the number of the sample nuclei were added during the run-on reaction as a normalization spike-in. Total RNA was extracted using phenol/chloroform precipitation. Isolated RNA was fragmented using a base hydrolysis with NaOH. Biotinylated fragmented nascent transcripts were isolated using a first streptavidin Dynabeads M-280 (Invitrogen Ref. 11206D) pull down, and 10 µL VRA3 3' RNA adaptor[84] (/5Phos/rUrNrNrNrNrNNGATCGTCGGACTGTAGAACTCTGAAC/3InvdT/) was ligated at their 3' end. A second streptavidin bead pull-down was performed, followed by the enzymatic modifications of the RNA fragment 5' ends with a pyrophosphohydrolase and a polynucleotide kinase (PNK)(NEB, Ref. M0201), and the 10 µL VRA5 RNA adaptor[84] (/5InvddT/CCTTGGCACCCGAGAATTCCANrNrNrNrNrNrC) was ligated at their fixed 5' ends. A third streptavidin bead pull-down was performed, followed by the reverse transcription of the resulting adaptor-ligated libraries. The libraries were cleaned up with AMPure XP beads (Beckman Coulter Ref. A63881). The libraries were amplified using 13 PCR cycles and cleaned up again with another round of AMPure XP beads. The resulting library concentrations were measured with the Qubit dsDNA high-sensitivity assay (Invitrogen Ref. Q32851), and their size distributions

were assessed using the Agilent High Sensitivity D1000 ScreenTape. The biological replicates were processed independently.

The first two PRO-seq biological replicates were sequenced using a NextSeq 500. Base calls and demultiplexing were done using Bcl2Fastq2 (v2.2.0). The FASTQ files sequenced on the sequential dates were concatenated. A third biological replicate was done for three of the cell lines due to the low complexity of the second replicate and was sequenced using a NextSeq 2000. All datasets were sequenced as single-end 76 base pair long reads.

Table S2.4 describes the number of reads per PRO-seq library. The samples were sequenced to a depth range between 2,854,374 and 53,450,928 reads with an average of 41.6 million reads.

2.5.3.3 RNA-seq library preparation and sequencing for intrahuman datase

The LCLs were collected after 180 minutes of IFN-β treatment incubation. Total RNA was extracted using the Quick-RNA MiniPrep Plus (Zymo Research Ref. R1058) and the RNA concentrations were measured using a Qubit HS RNA kit, yielding concentrations ranging from 2 ng/µL to 12 ng/µL. The RNA-seq libraries were prepared following the KAPA mRNA HyperPrep Kit instruction (KR1352 – v7.21) using the KAPA mRNA HyperPrep Kit (Roche Ref. KK8581), KAPA mRNA Capture Kit (Roche Ref. KK8441), and KAPA Pure Beads (Roche Ref. KK8545). For most samples, 250 ng of total RNA was used as input with an RNA fragmentation step of 6 minutes at 94°C, and using 11 cycles in the amplification step. A lower concentration of 150-100 ng of total RNA was used and during the amplification step, 12-14 cycles were used. The finalized library concentrations were obtained using the Qubit dsDNA high-sensitivity assay kit (Invitrogen Ref. Q32851), with final concentrations ranging from 2 ng/µL to 21 ng/µL.

The three biological replicates were pooled together and sequenced on a NovaSeq 6000 as paired-end 150 base pair long reads. Table 2.5 describes the number of reads per RNA-seq library. The samples were sequenced to a depth range between 23750002 and 49718713 reads with an average of 34.7 million reads. Mapping statistics regarding reads aligned against the GRCh38 reference genome with HISAT2 2.1.0 aligner had an average 99% overall alignment rate and average 80%

uniquely mapped.

2.5.4 Sequence library processing

2.5.4.1 PRO-seq datasets processing

PRO-seq fastq were processed through our in-house Nextflow (v21.10.6) pipeline, Nascent-Flow (v1.3), available on Dowell-Lab Github at `https://github.com/Dowell-Lab/Nascent-Flow/`. The fastq read quality was assessed using FastQC (v0.11.8). The reads were trimmed using BBDuk (v38.05) with the parameters ktrim=r, qtrim=10, k=23, mink=11, hdist=1, maq=10, minlen=25, and literal=AAAAAAAAAAAAAAAAAAAAAA. The trimmed reads were aligned to the GRCh38/hg38 reference genome using HISAT v2.1.0 with parameters –very-sensitive –no-spliced-alignment.

The reads distribution is captured by RSeQC (v3.0.0) read_distribution.py program. This module takes the BAM input to calculate how mapped reads are distributed over genome features CDS exon, 5′UTR exon, 3′ UTR exon, intron, and intergenic regions.

To obtain a PRO-seq count table reads were counted over hg38 reference genes using Rsubread featureCounts with parameters isGTFAnnotationFile=TRUE, GTF.featureType="gene_length", GTF.attrType="transcript_id", useMetaFeatures=TRUE, allowMultiOverlap=FALSE, isPairedEnd=FALSE, and strandSpecific=1. In the PRO-seq, reads were counted over the gene +750 nucleotides from the annotated transcription start site and used parameter

2.5.4.2 RNA-seq datasets processing

RNA-seq fastq files were processed through our in-house Nextflow (v21.10.6) pipeline, RNAseq-Flow (v1.1), available on Github at `https://github.com/Dowell-Lab/RNAseq-Flow+6`. The fastq read quality was assessed using FastQC (v0.11.8). The reads were trimmed using bbduk (v38.05) to remove adapter sequences as well as short or low quality reads. The trimmed reads were aligned to GRCh38/hg38 reference genome (release number 109, downloaded August 2019) using HISAT (v2.1.0). The resulting SAM files were converted to BAM files using Samtools v1.8. For visualization,

BedGraph files were generated using Bedtools v2.28.0 and converted to TDF files using IGVtools v2.3.75. Quality metrics were assessed with FastQC v0.11.8. To generate RNA-seq count table reads were counted over genes using R Rsubread featureCounts with parameters isGTFAnnotationFile=TRUE, GTF.featureType="exon", GTF.attrType="gene_id", useMetaFeatures=TRUE, allowMultiOverlap=FALSE, largestOverlap=TRUE, and isPairedEnd=TRUE.

2.5.5 Differential Expression Analysis

Distinct count tables were used for differential expression analysis. The PRO-seq counts were pre-processed before running analysis to select the longest annotated gene length. To ensure the results between the RNA-seq and PRO-seq assays were comparable, only genes with reads in both RNA-seq and PRO-seq were considered.

Differential transcription analysis was done with DESeq2 (v1.36.0) for R (v4.2.1). The DESeq2 was run separately for each of the ten individuals contrasting IFN-β and control (BSA) treatment with default alpha 0.1. To further filter for genes that respond to IFN-β and are significant in each cell line between treatment, a significance cutoff alpha of 0.01 was used.

2.5.6 Bidirectional processing and analysis

2.5.6.1 Bidirectional loci calls

Bidirectional loci were called using Tfit(v1.2) using our in-house Nextflow (v20.07.1) pipeline, Bidirectional-Flow (v0.3), available on Dowell-Lab Github at `ttps://github.com/Dowell-Lab/Bidirectional-Flow`.

2.5.7 Building bidirectional annotation list

To generate a consensus list from multiple replicates and conditions we use TFEA (v1.1.1) muMerge[150] module in two parts. First, we muMerge (parameters –verbose, –remove_singletons) each cell line's control and treatment replicates without singletons. Second, we merge (parameters –verbose, –save_sampids) all the individual files treating each as a replicate to get the master list

of all bidirectionals excluding those found only in a single sample. We will be referring to this annotation as "muMerge master list".

2.5.8 Enrichment of regulatory factors in PRO-seq via TFEA

To detect positional motif enrichment, we fed a ranked bidirectional list to Transcription factor enrichment analysis (TFEA)[150]. To generate the ranked list, we first padded bidirectionals in the muMerge master list by 500 nts on either side before generating a count table (over the padded muMerge master list). A ranked list was generated in DESeq2 using the formula -log(p-value) * sign(log2FoldChange) on differential expression analysis using the bidirectional count table.

2.5.9 IFN-score

IFN-score calculation was adapted from [64]. To capture interferon signaling in each sample as a single value, we calculate from the RNA-seq data the type I (Interferon Alpha) and type II (Interferon Gamma) scores. The z-score was first calculated for each gene in the sample based on the mean and standard deviation of the negative BSA condition. Per sample, the sum of the z-score for genes in the GSEA Hallmark Interferon Alpha and Hallmark Interferon Gamma gene sets which consists of 224 annotated genes. The IFN score per sample was plotted and grouped by the genotype T21 and D21.

2.5.10 Likelihood Ratio Test

We implement DESeq2's likelihood ratio test (LRT) to identify any genes that show changes in expression that are different in T21 cells compared to D21. LRT compares the full model (genotype and treatment) to the reduced model (treatment only) to identify significant genes (significant cutoff adjusted p-value < 0.01). The significant genes were clustered using the clustering tool, degPatterns (DEGreport v1.32.0). degPatterns uses a hierarchical clustering approach based on pair-wise correlations and clusters groups of genes with similar expression profiles.

To compare significantly differential genes expressed similarly in the full model design, genes

that were found significant in the LRT were removed from the full model. To reduce noise, a cutoff in log2FoldChange of ± 1.0 was applied to filter out genes that may have a small log2FoldChange because the gene had low read counts, or there may be high or unequal variance in the data.

2.5.11 Metagene plot

To plot the metagene, we compile the chr, start, stop, and strand information for a subset of genes using the GRCh38 reference genome RefSeq. Each gene is split into 3 components; TSS (-750bps/+750bps from the annotated start), TES (-1000bps/+8000bps from the annotated stop), and genebody (position between +750bp TSS and TES-8000bps). Next, the regions were split into 100 bins and the samples PRO-seq reads were counted over the bins using Rsubread featureCounts. The counts were normalized based on sizeFactor obtained from DESeq2. The metagene plots are composed of the average median coverage over the region at single base pair resolution.

To select a subset of genes for the baseline comparison, genes with adequate transcription were filtered to not include any annotated ISGs and were not found significantly differential expressed (significant cutoff adjusted p-value < 0.1) when perturbed with IFN-β.

2.5.12 SNP identification filtered by logFoldChange threshold

To identify SNPs, we seek to filter out genes that have cell line variation in response. To find unchanging genes we use the altHypothesis="lessAbs" function in DESeq2 and set a lfcThreshold of 0.58, which is equivalent to 1.5. fold. This defines any gene < 1.5 as not changing

This method is more conservative than DESeq2's default DEG method as it targets genes with a fold change of less than 1.5, setting a higher threshold to identify genes with more substantial changes This analysis is more stringent and reduces the likelihood of false positives.

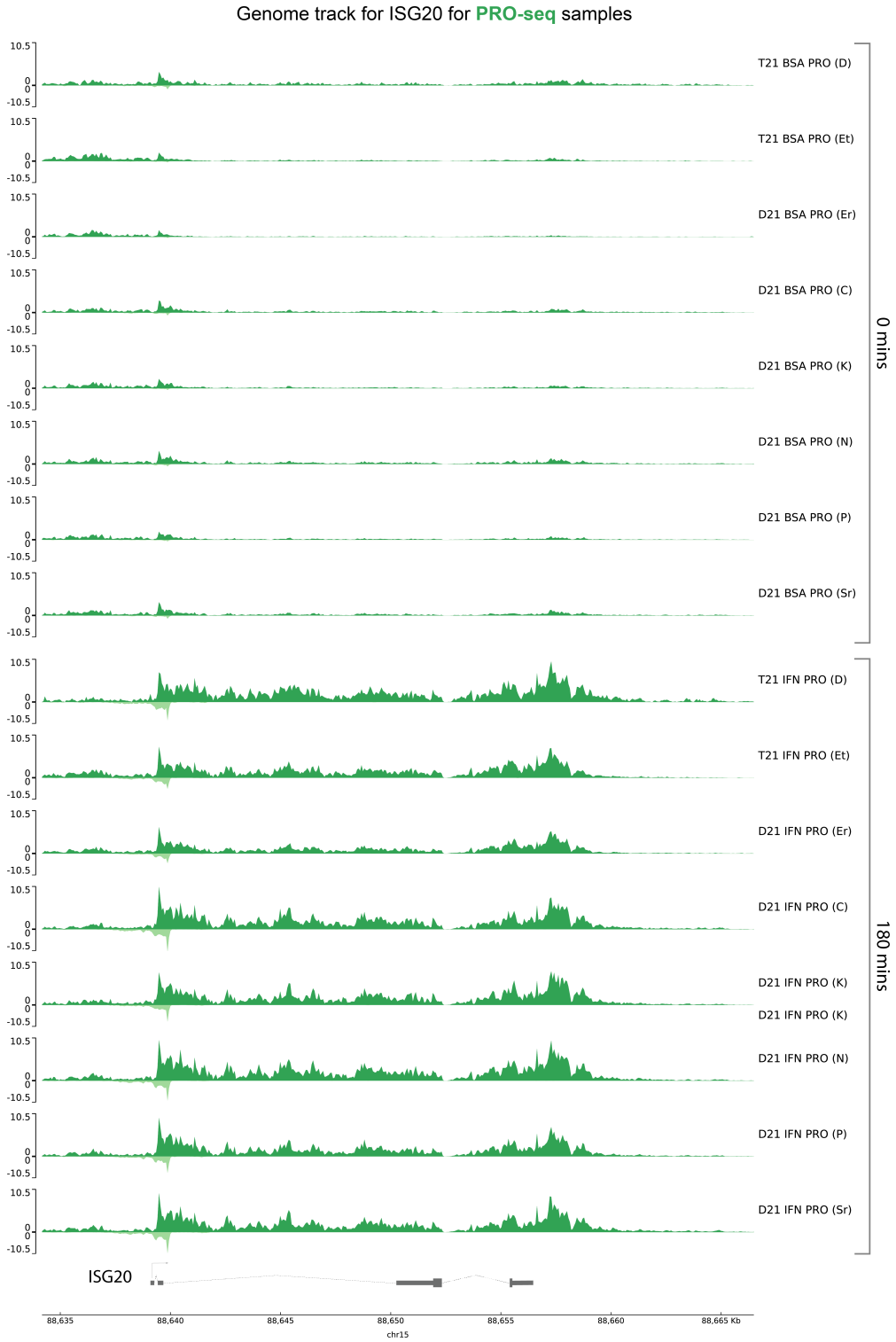
2.5.13 GitHub

All code is available at https://github.com/Dowell-Lab/IFN_population

2.6 Supplemental Tables and Figures

Table 2.2: **Sequencing reads distribution.** Reads map to distinctly different locations between the transcription (PRO-seq) and steady state (RNA-seq) assays.

Region	PRO-seq read distribution %	RNA-seq read distribution %
TSS 1kb	2%	0.2%
Exon	19%	83%
TES	23%	7%
Intron	51%	8%
Intergenic	7%	2%

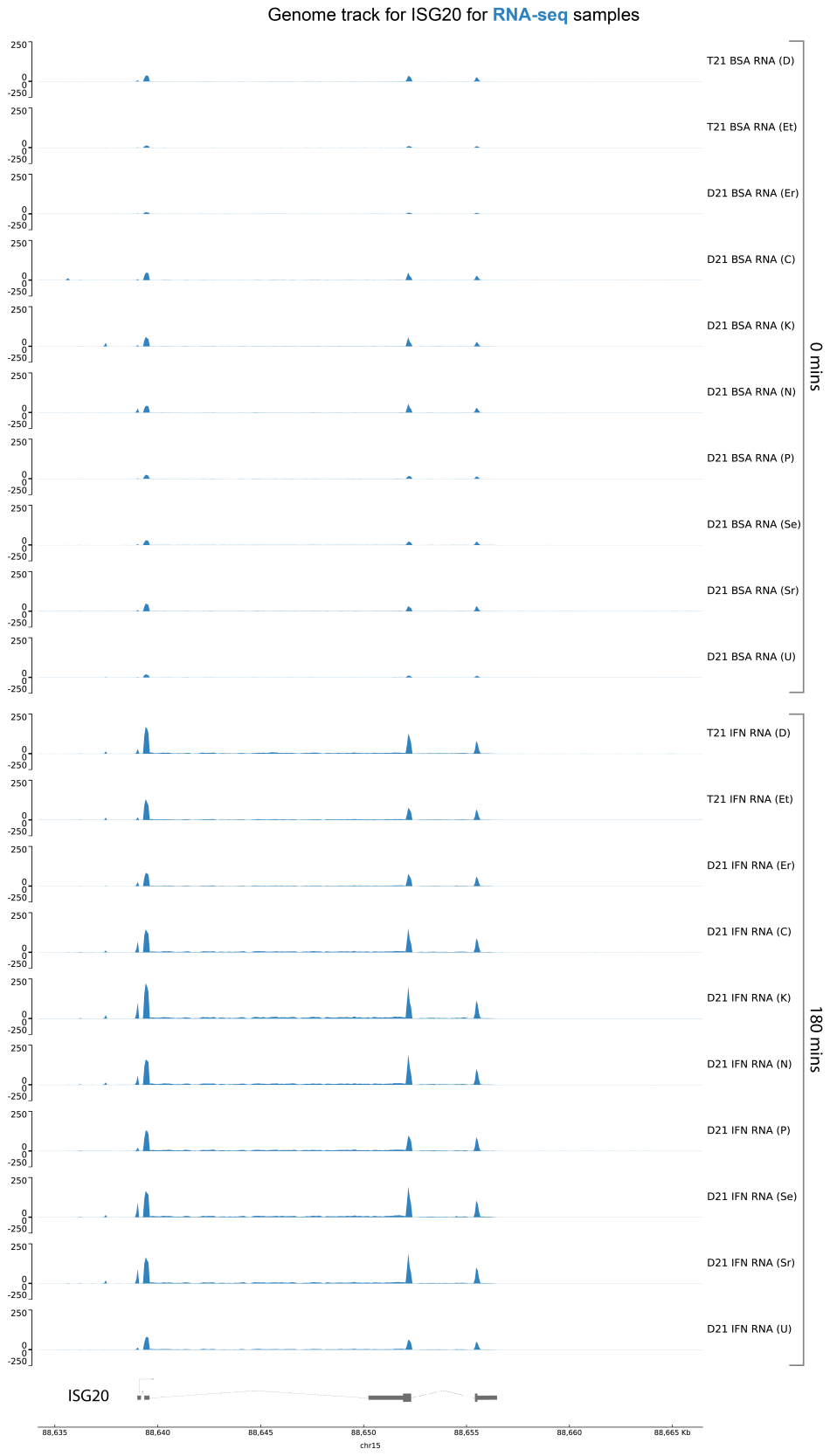


Continued on next page.

Figure 2.17: **PRO-seq genome track for ISG20** Genome track for ISG20 (chr15: 88,635,670-88,656,483), in all eight cell lines (2 individuals dropped for low quality libraries; Sengbe and Ursula) for PRO-seq at 0 (first 8 tracks) and 60 minutes (last 8 tracks). The cell lines have a similar immediate-early transcription response to the IFN-β stimulation.

Table 2.3: **Temporal distribution of responsive genes.** Temporal distribution of responsive genes in all samples. Transient captured at 60 minutes, Secondary captured at 180 minutes, and Direct are genes found at both 60 minutes and 180 minutes. "NA" for Sengbe and Ursula PRO-seq samples that were dropped for low-quality libraries.

Sample	Transient (up-reg, down-reg)	Direct (up-reg, down-reg)	Secondary (up-reg, down-reg)
ChenChao	623, 191	266, 186	951, 972
Dave	501, 817	228, 194	822, 226
Eric	398, 43	93, 21	965, 928
Ethan	131, 100	562, 137	1063, 1099
Khaondo	681, 284	308, 220	1035, 1260
Niyilolawa	251, 174	652, 186	960, 1062
Pedro	164, 61	494, 91	891, 1043
Sengbe	NA	NA	1604, 1444
Srivathani	115, 51	429, 45	880, 804
Ursula	NA	NA	1217, 859



Continued on next page.

Figure 2.18: **RNA-seq genome track for ISG20** Genome track for ISG20 (chr15: 88,635,670-88,656,483), in all ten cell lines for RNA-seq at 0 (first 10 tracks) and 180 minutes (bottom 10 tracks). All the cell lines display an increase in transcription in response to the IFN-β stimulation compared to the negative BSA control track.

Figure 2.19: **RNA-seq BSA transcription of IFN receptors** Type I, II, and III IFN receptors in BSA between D21 and T21 cell lines. IFN receptors found on chromosome 21 have a higher basal expression in the T21 cell line compared to D21, but those off chromosome 21 do not.

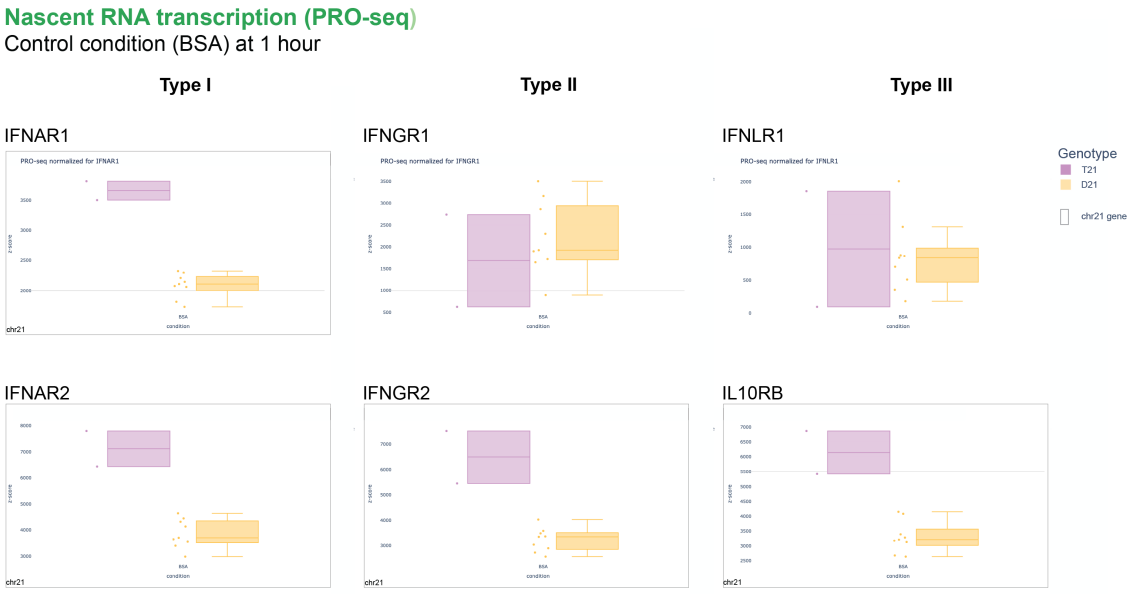


Figure 2.20: **PRO-seq BSA transcription of IFN receptors** Type I, II, and III IFN receptors in BSA between D21 and T21 cell lines. IFN receptors found on chromosome 21 have a higher basal transcription in the T21 cell line compared to D21, but those off chromosome 21 do not.

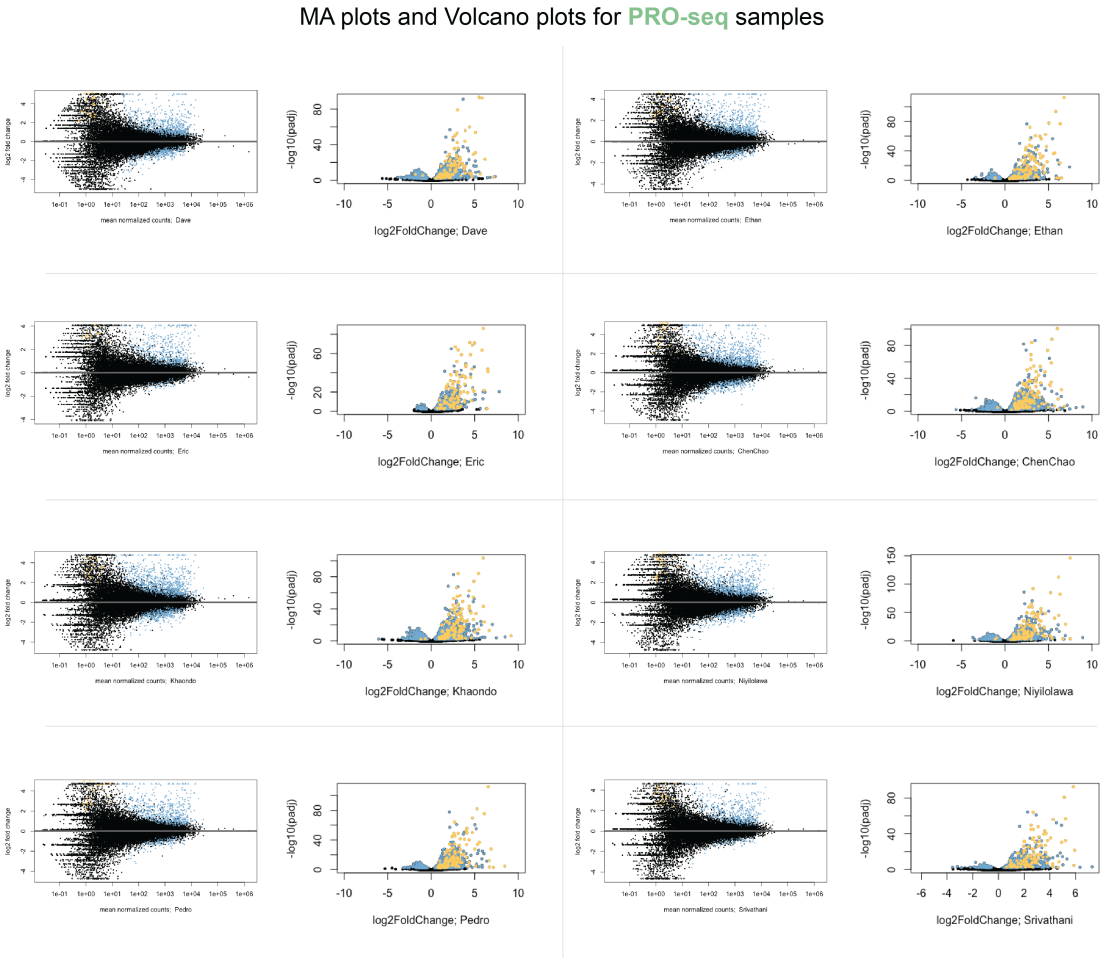


Figure 2.21: **PRO-seq MA and volcano plot.** MA and volcano plot for eight (out of ten) cell lines after 60 minutes of incubation with IFN-β. The Type I and Type II interferon response genes are labeled on the volcano plot in yellow. Similar to RNA-seq most of the differential genes are up-regulated including the interferon response genes.

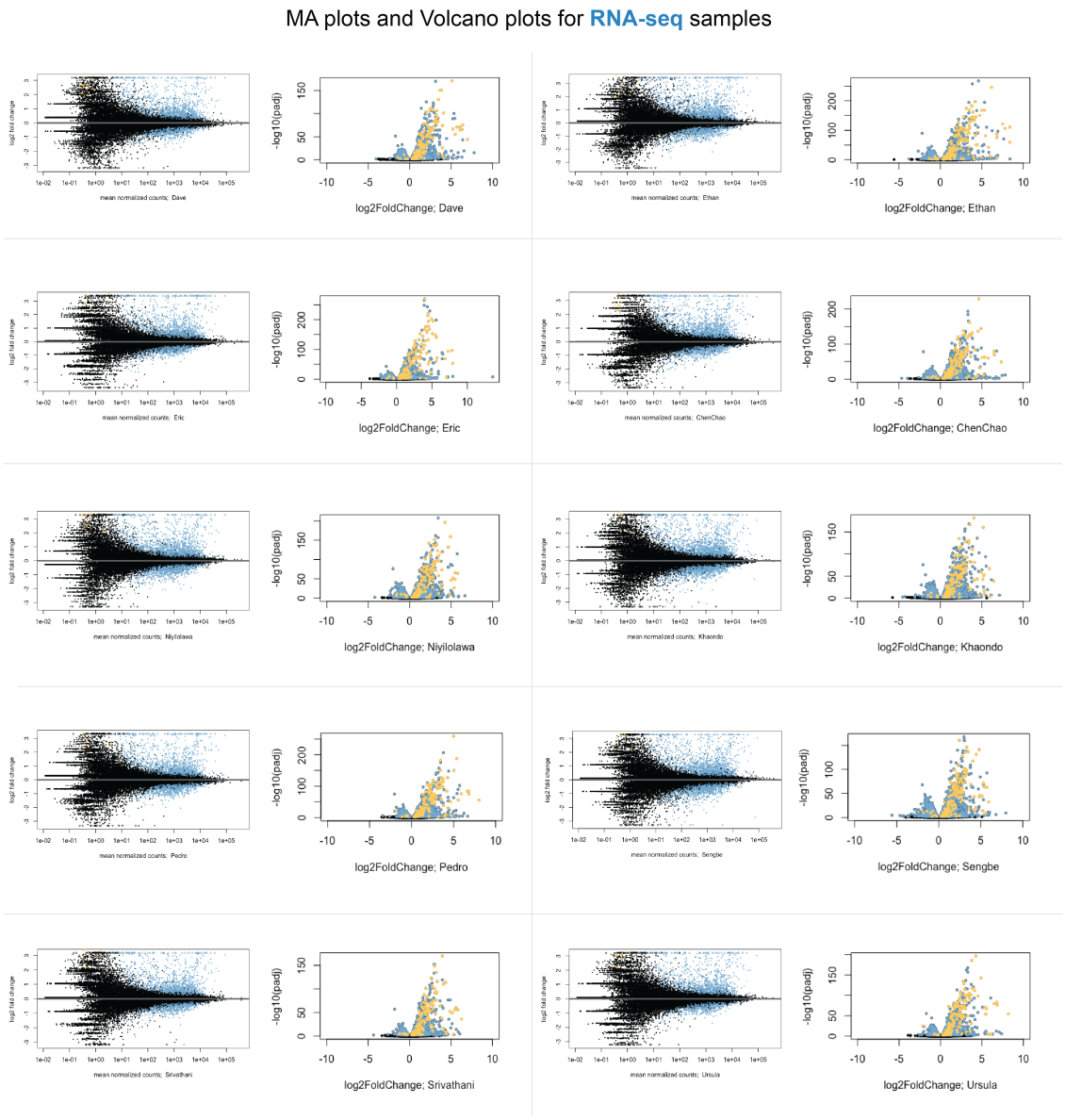
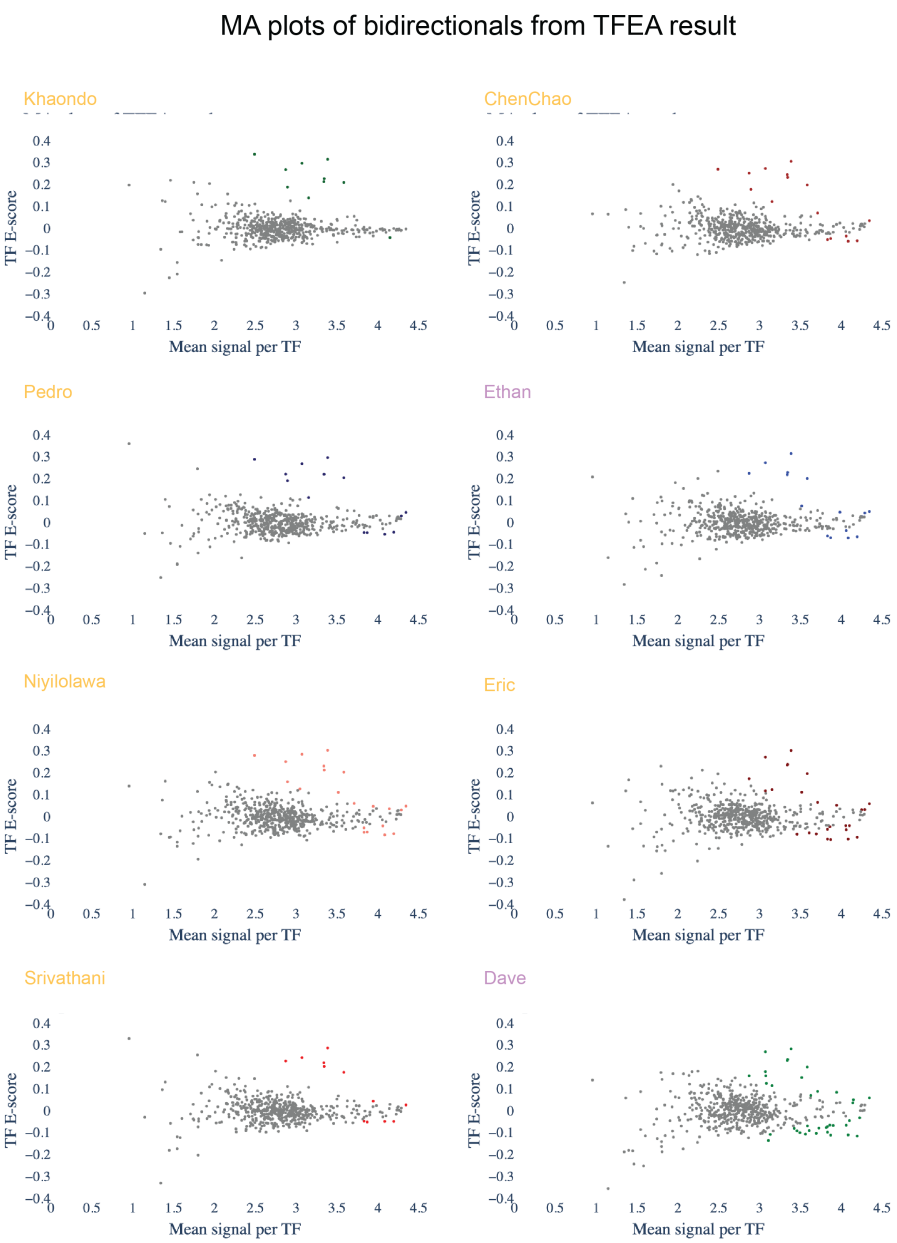


Figure 2.22: **RNA-seq MA and volcano plot** MA and volcano plot for all ten cell lines after 180 minutes incubation with IFN-β show an increase in gene expression. Volcano plots with Type I and Type II interferon response genes labeled in yellow show that most IFN response genes are upregulated after IFN-β perturbation. There are some downregulated interferon response genes in the steady-state RNA-seq assay.



Continued on next page.

Figure 2.23: **TFEA MA plots of each individual.** Each dot is a single transcription factor motif. X-axis is proportional to the number of bidirectionals that contain at least one motif instance within 3kb of the identified μ (center of the bidirectional). Y-axis is the enrichment score, which quantifies co-localization of motif hits with μ relative to a larger local background.

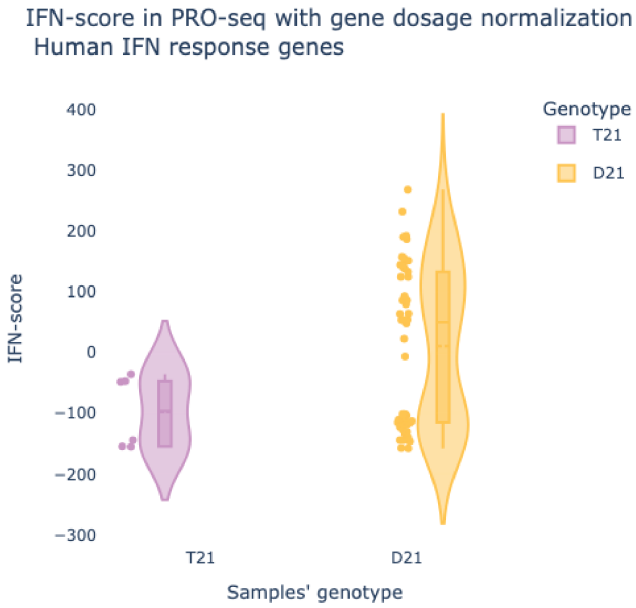


Figure 2.24: **IFN-score after gene dosage normalization.** Gene dosage normalized chromosome 21 prior to calculating IFN-score tightened the distribution in T21 samples.

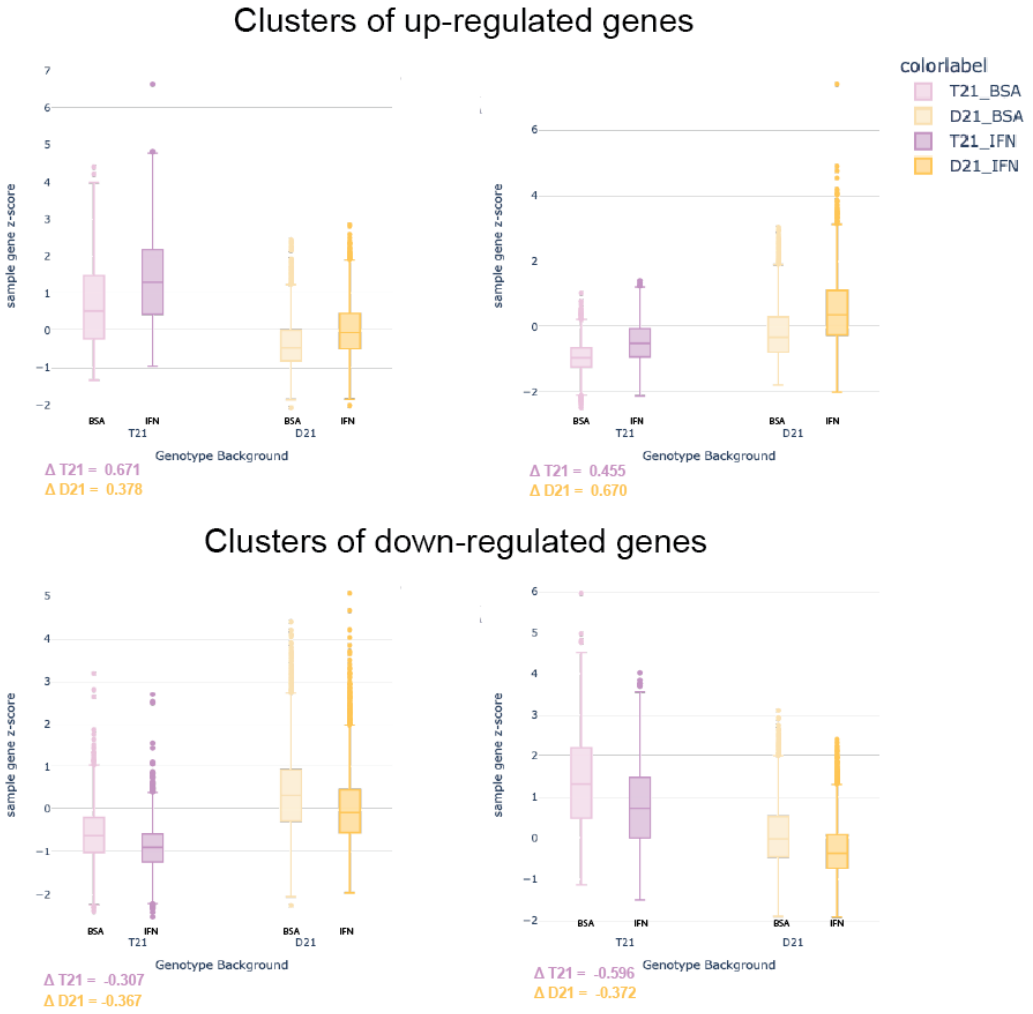


Figure 2.25: **Differential gene expression clustered based on patterns in T21 versus D21 considering IFN-β treatment.** Four clusters defined by the LRT model comparing a comprehensive and reduced model are plotted as bar plots representing samples' gene expression profiles. While the baseline gene expression for T21 varies from D21 in control conditions, both genotypes respond similarly to the treatment. The change (delta) in gene expression post-IFN-β remains consistent across genotypes, despite T21 baseline expression level.

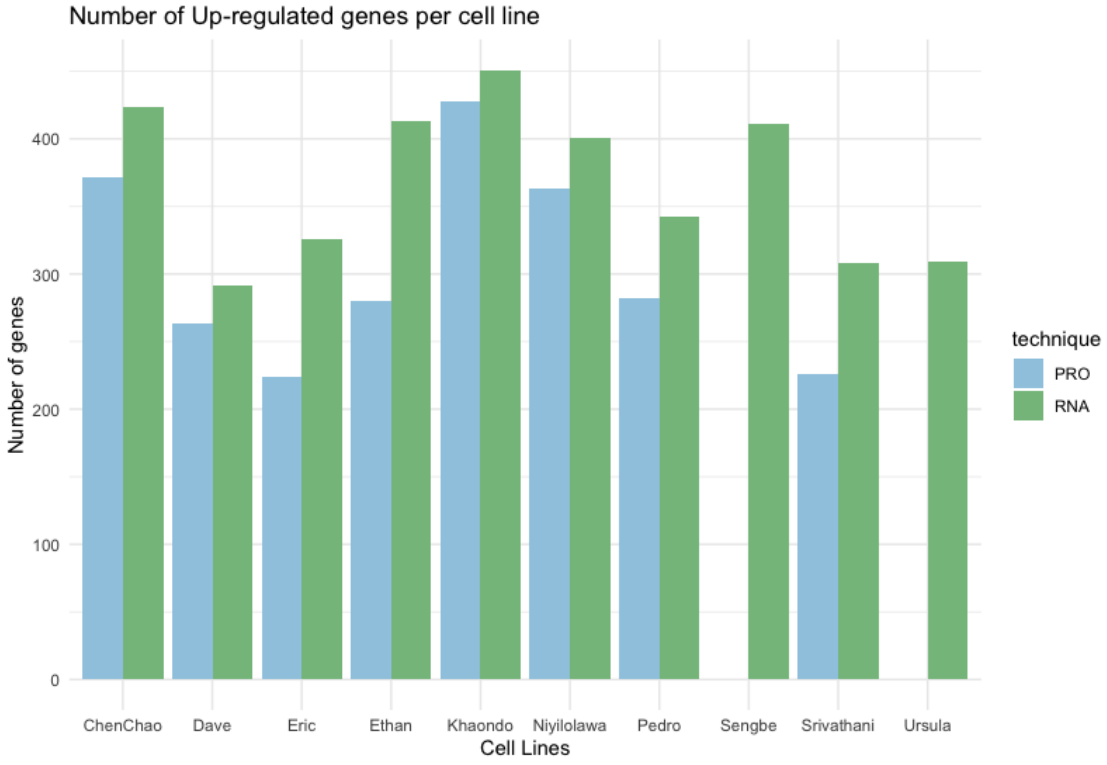


Figure 2.26: **Numbers of up-regulated genes per individual.** Numbers of up-regulated genes in each assay per individual. Blue bars are the number of genes that were transcribed at 60 minutes (2 individuals dropped for low quality libraries; Sengbe and Ursula). Green bars are the number of genes expressed at 180 minutes.

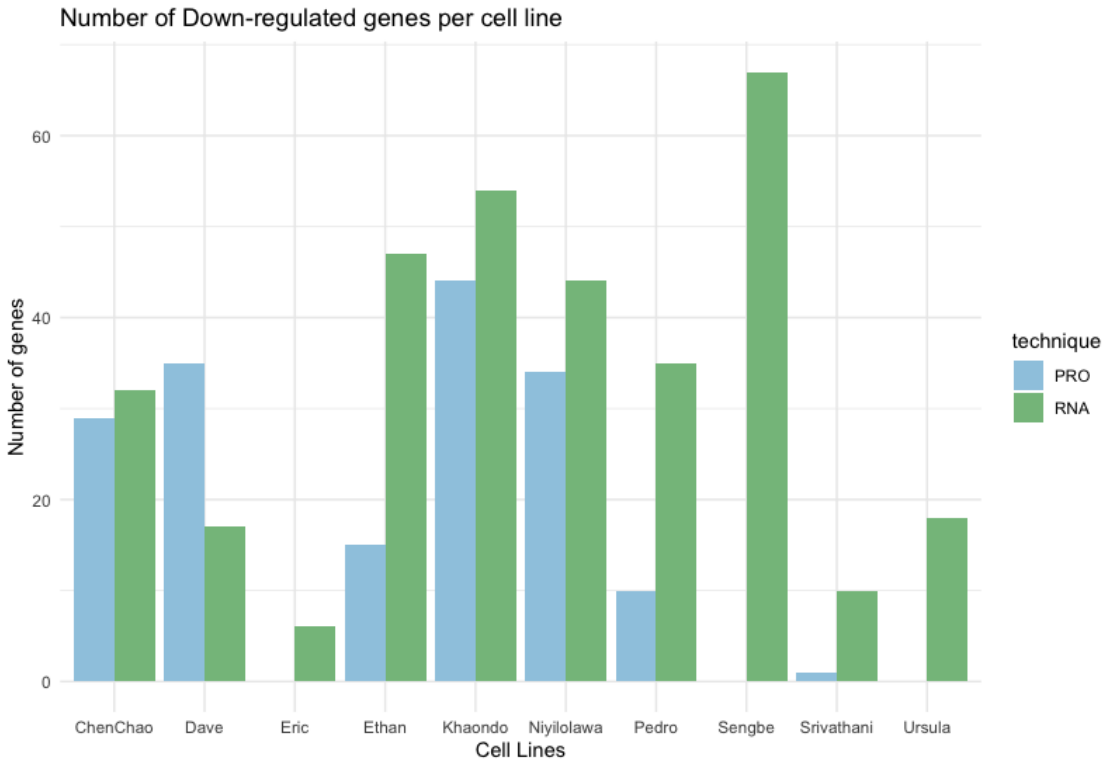


Figure 2.27: **Numbers of down-regulated genes per individual.** Numbers of down-regulated genes in each assay per individual. Blue bars are the number of genes that were transcribed at 60 minutes (2 individuals dropped for low quality libraries; Sengbe and Ursula). Green bars are the number of genes expressed at 180 minutes.

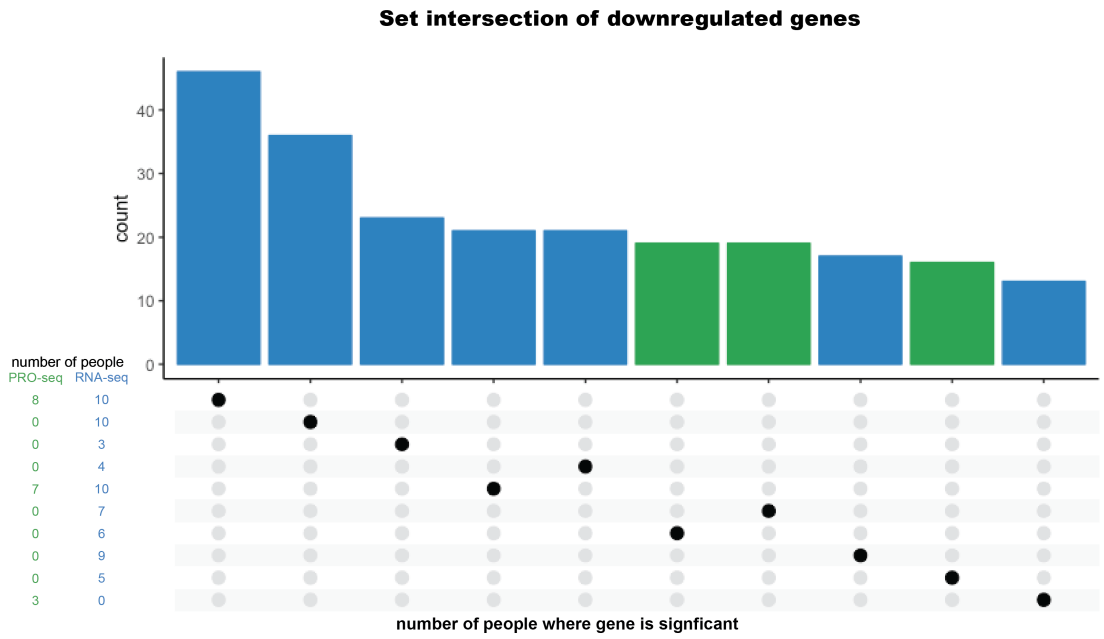


Figure 2.28: **UpSet plot of Down-regulated genes.** UpSet plot of down-regulated genes showing the set intersection of genes based on set of number of people the gene is significantly differential transcribed/expressed.

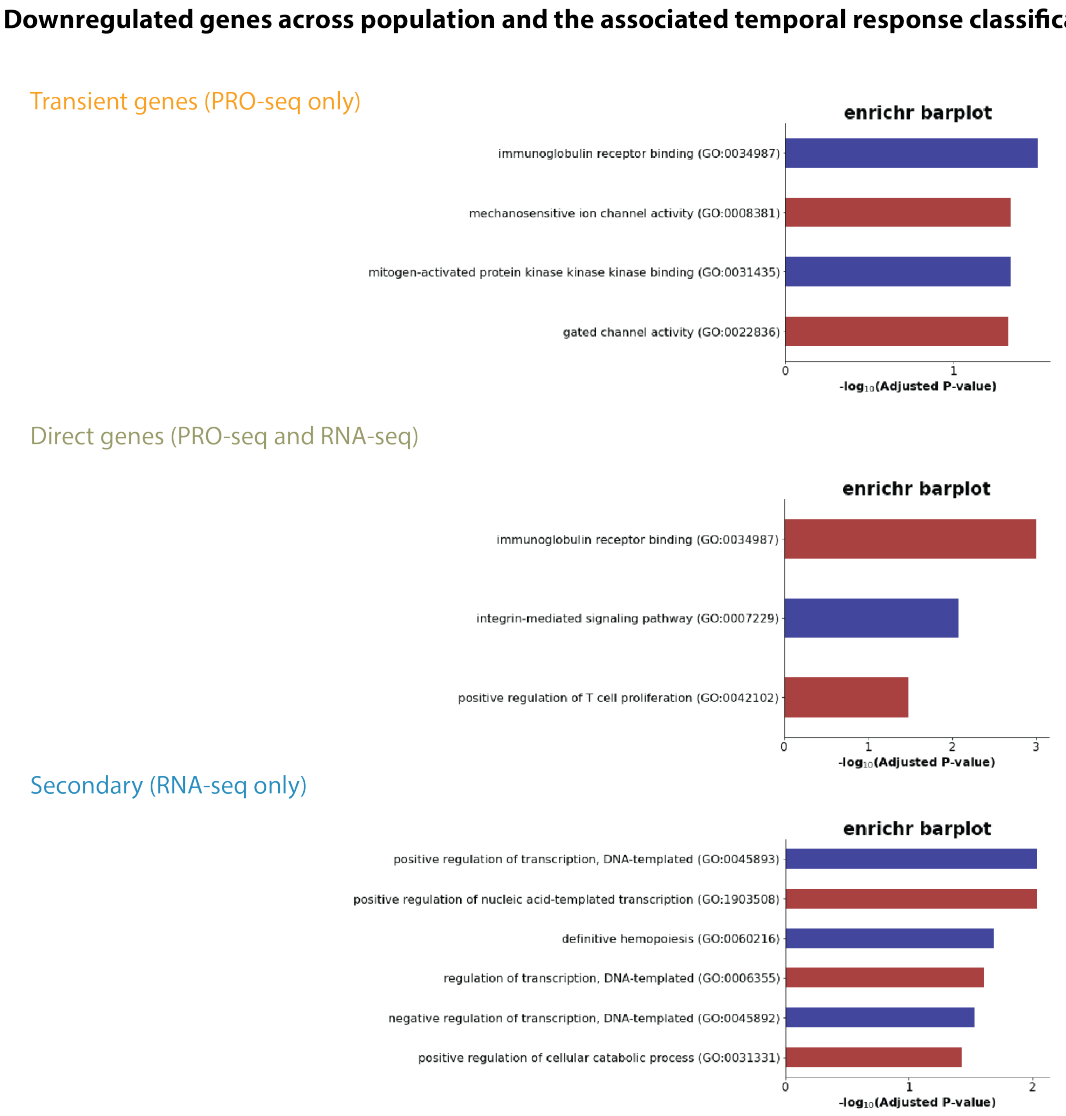


Figure 2.29: **Gene ontology enrichment for downregulated genes in population.** Enrichr tool use GO terms to identify biological pathway and molecular function associated with genes statistically significant in population and classified in temporal response. Downregulated transient genes are enriched for receptors and channel activity. Direct downregulated genes are related to immune response. Secondary downregulated genes are enriched for transcription.

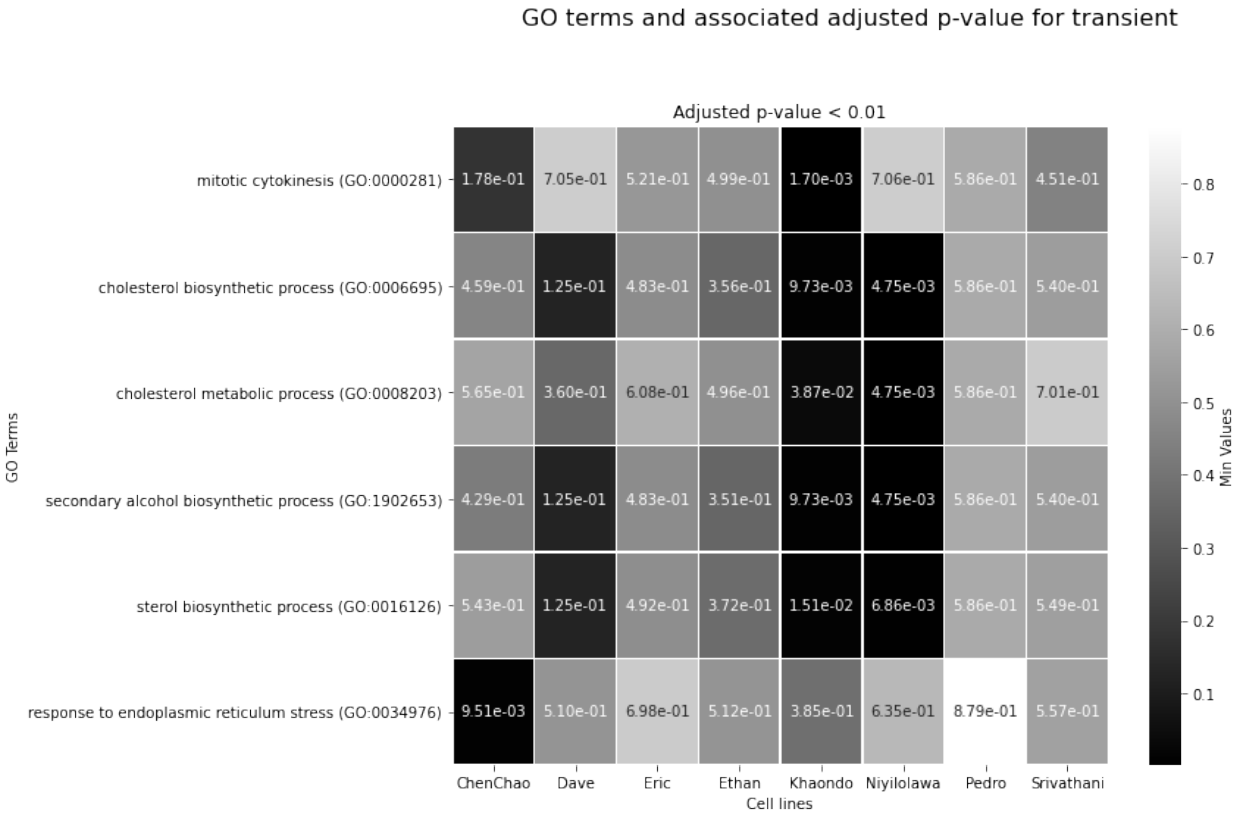


Figure 2.30: **Transient genes GO terms.** Heatmap of top significant genes categorized by the temporal response and the associated adjusted p-value per cell lines. Heatmap scale darker heat equivalent to the more significant p-value. Transient GO term enrichment includes metabolic processes. The adjusted p-value for many of the terms are not significant.

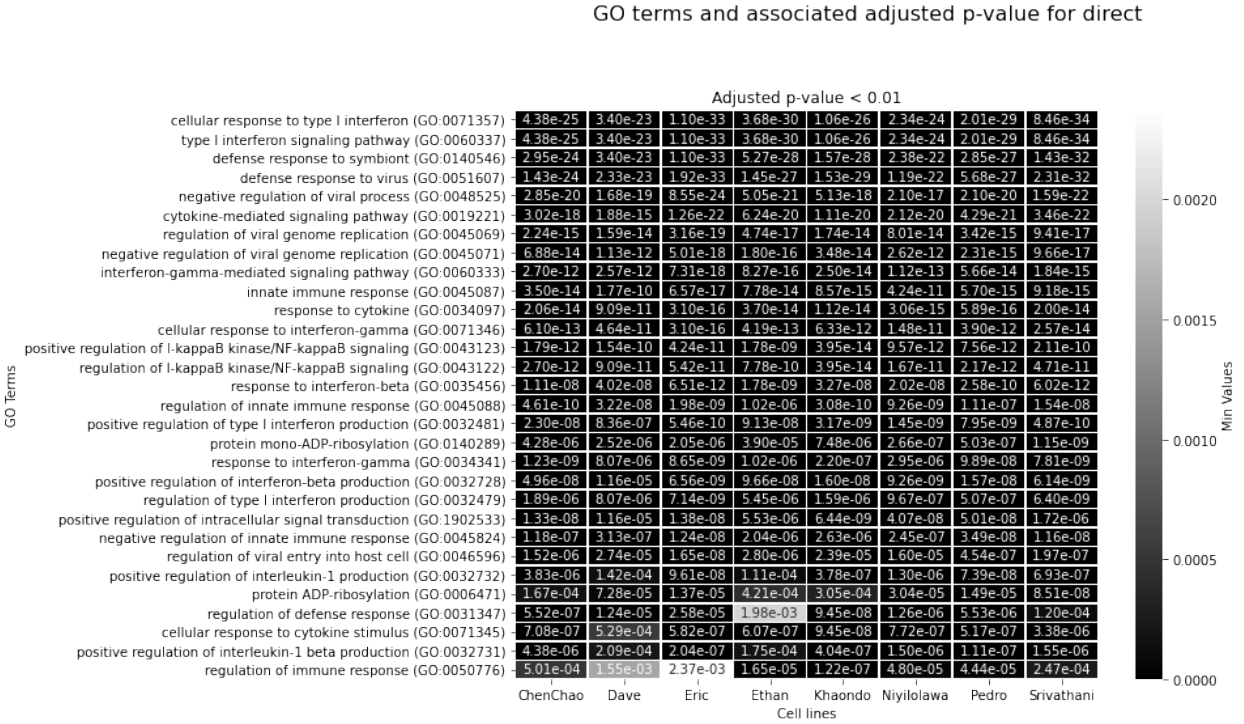


Figure 2.31: **Direct genes GO terms.** Heatmap of top significant genes for direct temporal response and the associated adjusted p-value per cell lines. Heatmap scale darker heat equivalent to the more significant p-value. Direct GO terms are predominantly related to interferon response in addition to response to cytokine and virus.

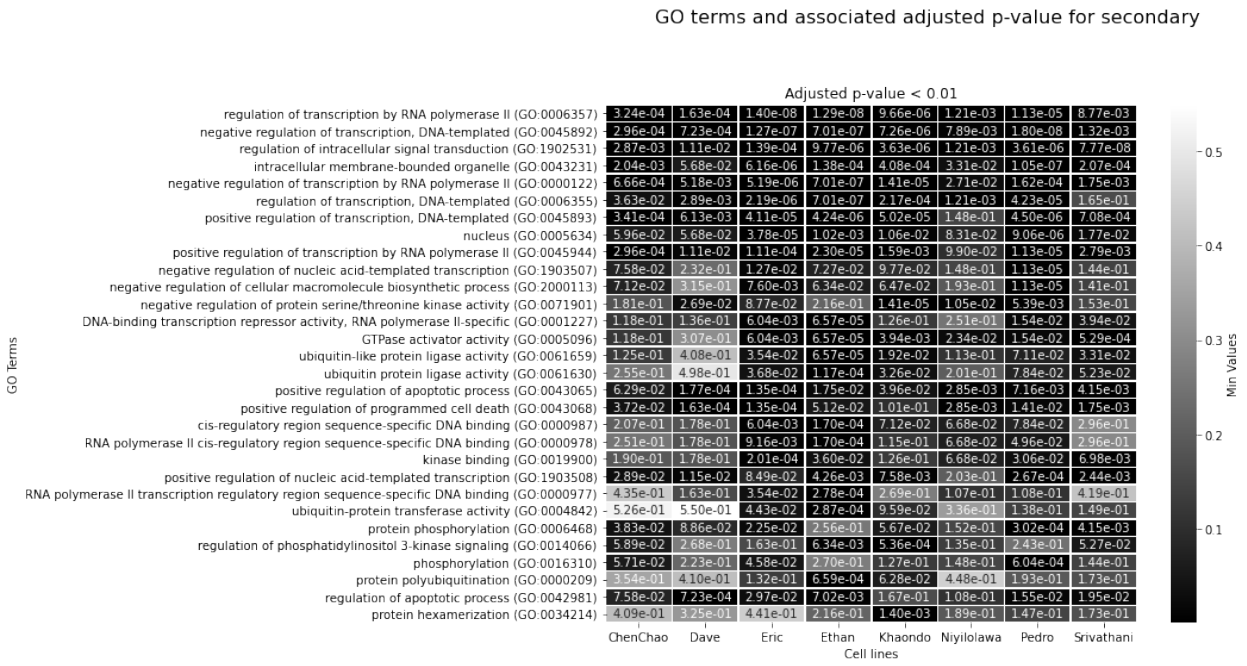
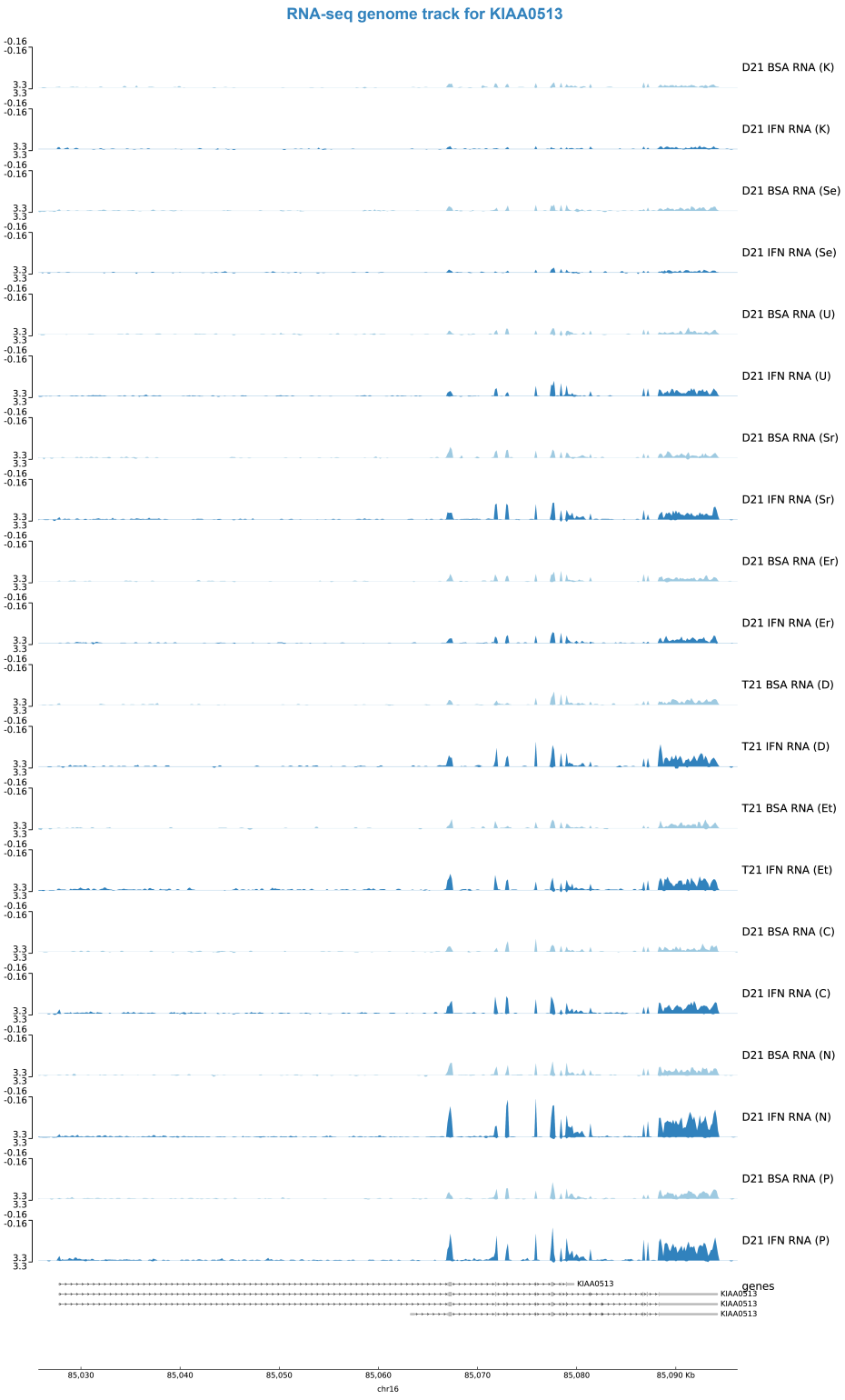


Figure 2.32: **Secondary genes GO terms.** Heatmap of top significant genes for secondary temporal response and the associated adjusted p-value per cell lines. Heatmap scale darker heat equivalent to the more significant p-value. Secondary GO terms are predominantly related to cellular regulation of transcription and secondary response to viral infection such as ubiquitin ligase activity and apoptotic processes.

Table 2.4: **Sequencing depth of the IFN PRO-seq intrahuman datasets.** Not shown Sengbe and Ursula; samples dropped for low quality libraries

Dataset	Read number	Dataset	Read number
PRO-BSA-ChenChao-1	42,002,990	PRO-IFNB-ChenChao-1	42,058,600
PRO-BSA-ChenChao-2	46,360,142	PRO-IFNB-ChenChao-2	46,200,586
PRO-BSA-Dave-1	41,194,471	PRO-IFNB-Dave-1	42,453,894
PRO-BSA-Dave-2	44,249,637	PRO-IFNB-Dave-2	40,934,032
PRO-BSA-Dave-3	35,464,522	PRO-IFN-Dave-3	31,737,744
PRO-BSA-Eric-1	39,029,604	PRO-IFNB-Eric-1	42,101,766
PRO-BSA-Eric-2	40,696,392	PRO-IFNB-Eric-2	46,652,910
PRO-BSA-Eric-3	28,547,374	PRO-IFN-Eric-3	28,787,496
PRO-BSA-Ethan-1	41,596,634	PRO-IFNB-Ethan-1	41,620,402
PRO-BSA-Ethan-2	40,623,819	PRO-IFNB-Ethan-2	43,416,781
PRO-BSA-Ethan-3	28,639,578	PRO-IFN-Ethan-3	33,508,006
PRO-BSA-Khaondo-1	42,423,753	PRO-IFNB-Khaondo-1	41,437,714
PRO-BSA-Khaondo-2	42,875,523	PRO-IFNB-Khaondo-2	46,542,139
PRO-BSA-Niyilolawa-1	41,631,139	PRO-IFNB-Niyilolawa-1	41,501,879
PRO-BSA-Niyilolawa-2	46,611,255	PRO-IFNB-Niyilolawa-2	40,392,956
PRO-BSA-Pedro-1	41,821,224	PRO-IFNB-Pedro-1	42,071,170
PRO-BSA-Pedro-2	42,975,706	PRO-IFNB-Pedro-2	45,989,703
PRO-BSA-Srivathani-1	41,795,630	PRO-IFNB-Srivathani-1	42,054,804
PRO-BSA-Srivathani-2	47,045,002	PRO-IFNB-Srivathani-2	41,652,259



Continued on next page.

Figure 2.33: **RNA-seq genome track for KIAA0513.** Genome track for KIAA0513 (chr16:85,025,709-85,096,230) in all ten cell lines. Most of the cell lines show up-regulation in response to the IFN-β stimulation, but two cell lines (Khaondo and Sengbe; top) show little to no response.

Table 2.5: **Sequencing depth of the IFN RNA-seq intrahuman datasets.**

Dataset	Read number	Dataset	Read number
RNA-BSA-Ursula-1	31,909,634	RNA-IFN-Ursula-1	40,162,379
RNA-BSA-Ursula-2	36,552,450	RNA-IFN-Ursula-2	30,464,008
RNA-BSA-Ursula-3	35,459,008	RNA-IFN-Ursula-3	32,786,917
RNA-BSA-DR-1	35,955,569	RNA-IFN-DR-1	31,248,938
RNA-BSA-DR-2	36,194,499	RNA-IFN-DR-2	32,219,991
RNA-BSA-DR-3	29,918,724	RNA-IFN-DR-3	29,311,192
RNA-BSA-Sengbe-1	34,534,303	RNA-IFN-Sengbe-1	36,101,570
RNA-BSA-Sengbe-2	36,240,125	RNA-IFN-Sengbe-2	39,571,943
RNA-BSA-Sengbe-3	26,787,072	RNA-IFN-Sengbe-3	25,737,414
RNA-BSA-Khaondo-1	38,882,432	RNA-IFN-Khaondo-1	33,931,715
RNA-BSA-Khaondo-2	42,343,160	RNA-IFN-Khaondo-2	48,990,204
RNA-BSA-Khaondo-3	25,064,983	RNA-IFN-Khaondo-3	32,175,085
RNA-BSA-Niyilolawa-1	36,792,381	RNA-IFN-Niyilolawa-1	49,718,713
RNA-BSA-Niyilolawa-2	36,034,857	RNA-IFN-Niyilolawa-2	46,966,788
RNA-BSA-Niyilolawa-3	25,716,202	RNA-IFN-Niyilolawa-3	33,621,764
RNA-BSA-Pedro-1	39,598,373	RNA-IFN-Pedro-1	33,245,218
RNA-BSA-Pedro-2	34,300,383	RNA-IFN-Pedro-2	34,119,312
RNA-BSA-Pedro-3	30,474,367	RNA-IFN-Pedro-3	24,924,245
RNA-BSA-Srivathani-1	31,412,960	RNA-IFN-Srivathani-1	42,189,675
RNA-BSA-Srivathani-2	34,607,358	RNA-IFN-Srivathani-2	28,661,148
RNA-BSA-Srivathani-3	33,511,863	RNA-IFN-Srivathani-3	31,859,262
RNA-BSA-ChenChao-1	26,505,929	RNA-IFN-ChenChao-1	37,207,581
RNA-BSA-ChenChao-2	33,966,765	RNA-IFN-ChenChao-2	31,794,771
RNA-BSA-ChenChao-3	30,666,149	RNA-IFN-ChenChao-3	33,286,304
RNA-BSA-Dave-1	36,687,953	RNA-IFN-Dave-1	23,750,002
RNA-BSA-Dave-2	49,676,332	RNA-IFN-Dave-2	34,934,265
RNA-BSA-Dave-3	31,907,175	RNA-IFN-Dave-3	39,546,353
RNA-BSA-Eric-1	34,468,542	RNA-IFN-Eric-1	38,176,754
RNA-BSA-Eric-2	38,251,062	RNA-IFN-Eric-2	39,738,019
RNA-BSA-Eric-3	34,477,884	RNA-IFN-Eric-3	32,169,657
RNA-BSA-Ethan-1	26,222,595	RNA-IFN-Ethan-1	27,211,017
RNA-BSA-Ethan-2	47,751,448	RNA-IFN-Ethan-2	41,846,890
RNA-BSA-Ethan-3	36,767,133	RNA-IFN-Ethan-3	43,914,436

Chapter 3

Characterizing Primary transcriptional responses to short term heat shock in paired fraternal lymphoblastoid lines with and without Down syndrome

Portions of this chapter are currently under review. Adapted from:
Cardiello JF; **Westfall J**; Dowell RD; Allen MA. Characterizing primary transcriptional responses to short-term heat shock in paired fraternal lymphoblastoid lines with and without Down syndrome. bioRxiv 2023 Feb 2;2023.01.17.524431. doi: 10.1101/2023.01.17.524431

3.1 Introduction

The Heat Shock Response (HSR) is a highly conserved cellular defense mechanism that gets activated in response to various stressors, predominantly elevated temperatures. This mechanism triggers the synthesis of a unique group of proteins termed Heat Shock Protein (HSP). These HSPs serve vital protective functions, acting as molecular chaperones that aid in protein folding, inhibit protein aggregation, and assist in the removal of malfunctioning proteins. Previous studies have shown that HSR and HSPs are implicated in neurodegenerative diseases where protein misfolding is a hallmark, in cancer where they help tumor cells cope with the tumor microenvironment, and in infectious diseases where certain pathogens can manipulate the host's HSR[127]. Previously, studies have identified elevated baseline HSP levels in the cells of individuals with Down syndrome[7]. Using publicly available transcriptomic data sets, we observed that there is upregulation of numerous heat shock-related genes in the blood samples of these individuals when compared to those without trisomy 21. Intriguingly, despite the presence of an extra copy in Down syndrome, chromosome 21

does not contain any recognized major regulators of the heat shock response. This study ventures to decipher the cause behind this amplified HSR in the context of trisomy 21.

3.2 Significance

The cellular and molecular mechanisms that contribute to the phenotype observed in individuals with Down syndrome are multifaceted. A prominent anomaly among these is the dysregulated heat shock response. While individuals with Down syndrome inherently exhibit elevated levels of HSPs, their cellular machinery might not amplify these proteins' synthesis as distinctly under stress when compared to typical individuals. This study probed the transcriptional and gene expression effects of acute heat shock perturbation on cells derived from individuals with and without trisomy 21. Our findings indicated that cells with trisomy 21 mounted a more pronounced heat shock response than disomy 21 cells. This highlights potential mechanisms behind this heightened response, possibly including compensatory reactions to chronic stress introduced by the extra chromosome 21.

3.3 Contributions

This chapter describes the collective work in the Dowell and Allen (DnA) laboratory. Dr. Joseph Cardiello, a post-doctoral in Dr. Mary Ann Allen's laboratory, spearheaded this project, generating most of the sequencing libraries and conducting the majority of the data analysis. My role in this endeavor was to process the Lymphoblastoid cell line (LCL) for ATAC-seq and prepare the sample sequencing libraries, with assistance from Dr. Mary Ann Allen. Dr. Cardiello wrote the initial manuscript and all authors contributed to subsequent refining, revisions, and reviewer responses.

More specifically, my contribution to this publication involved the cell culture work including growing up the Induced Pluripotent Stem Cell (iPSC) mosaic cell lines, perturbing the LCLs to heat-shock conditions, collecting the cells after treatment, and processing the cells for ATAC-seq. I contributed to the preparation of the manuscript by illustrating the experimental design, drafting the z-score normalization calculation in the method section, providing edits, and reviewing the

manuscript in preparation for publication submission.

This section includes most of the cited manuscript, encompassing both my contributions and those of others to maintain a coherent narrative flow. This work was supported by RUNX1 RO1 grant and SIE fellowship funding.

3.4 Paper Contents

Characterizing Primary transcriptional responses to short term heat shock in paired fraternal lymphoblastoid lines with and without Down syndrome.

Cardiello Joseph F., Westfall Jessica, Dowell Robin, Allen Mary Ann

Abstract

Heat shock stress induces genome wide changes in transcription regulation, activating a coordinated cellular response to enable survival. Using publicly available transcriptomic and proteomic data sets comparing individuals with and without trisomy 21, we noticed many heat shock genes are up-regulated in blood samples from individuals with trisomy 21. Yet no major heat shock response regulating transcription factor is encoded on chromosome 21, leaving it unclear why trisomy 21 itself would cause a heat shock response, or how it would impact the ability of blood cells to subsequently respond when faced with heat shock stress. To explore these issues in a context independent of any trisomy 21 associated co-morbidities or developmental differences, we characterized the response to heat shock of two lymphoblastoid cell lines derived from brothers with and without trisomy 21. To carefully compare the chromatin state and the transcription status of these cell lines, we measured nascent transcription, chromatin accessibility, and single cell transcript levels in the lymphoblastoid cell lines before and after acute heat shock treatment. The trisomy 21 cells displayed a more robust heat shock response after just one hour at 42°C than the matched disomic cells. We suggest multiple potential mechanisms for this increased heat shock response in lymphoblastoid cells with trisomy 21 including the possibility that cells with trisomy 21 may exist in a hyper-reactive state due to chronic stresses. Whatever the mechanism, abnormal heat

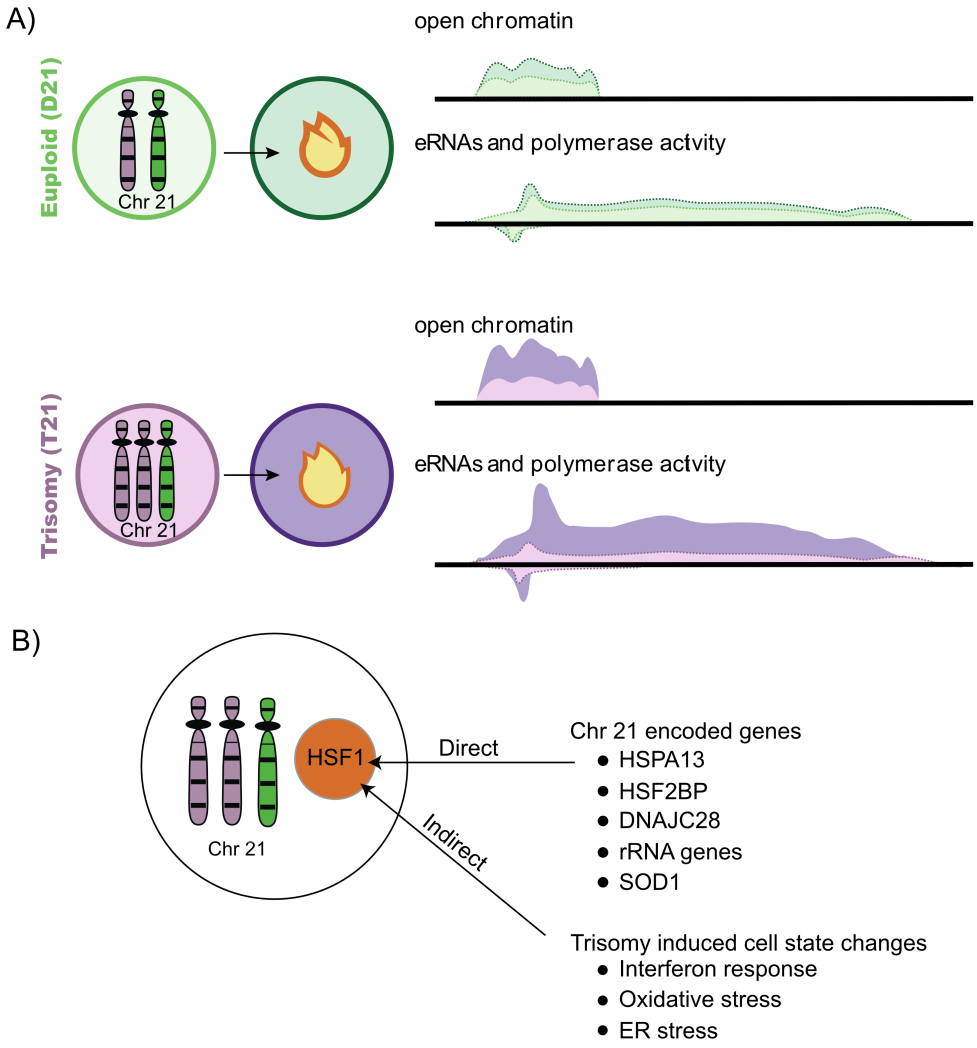


Figure 3.1: **Graphical abstract.** A: Cells with trisomy 21 have increased acute heat shock activated HSF1 transcription factor function as determined by both PRO-seq and ATAC-seq analysis. B: Several trisomy 21 cellular changes may contribute to an increased response to heat shock.

shock response in individuals with Down syndrome may hobble immune responses during fever and contribute to health problems in these individuals.

Introduction

Trisomy 21 is caused by an extra copy of chromosome 21. For the most part, it is unclear how this extra copy of 1% of the genome leads to phenotypes associated with trisomy 21. Trisomy 21 cells have been demonstrated to show an increased reaction to key cellular perturbations. For instance, trisomy 21 cells show an increased interferon response relative to typical cells which is likely driven by overexpression of four interferon receptor genes encoded on chromosome 21([162, 11, 137, 174]). Similarly, trisomic cells show signs of an elevated oxidative stress response, which may tie into the chromosome 21 encoded, oxidative stress responsive NRF2 or SOD1 genes ([97, 182, 131]). Despite the clear ties between these chromosome 21 located genes and the unusual cellular responses, it is unclear if the increased response to these two perturbations is driven solely by regulatory genes encoded on chromosome 21 or if other factors contribute such as a general trisomic stress response, or a trisomy 21 derived chronic stress. Moreover, it is unclear how trisomy 21 cells respond to perturbations that do not have primary regulators on chromosome 21.

Heat shock is a potentially lethal stress and therefore, it activates various cellular response processes, including the unfolded protein response. While three genes encoded on chromosome 21 are heat shock activated (HSF2BP, DNAJC28, HSPA13), none of these heat shock genes are known to be upstream regulators of the heat shock response. The major regulator of heat shock, HSF1, is a transcription factor located on chromosome 8. HSF1 is ubiquitously expressed, but its transcription factor activity is highly regulated through post-translational modifications, nuclear import, and protein interactions ([7, 26, 173, 171]). Following heat shock there is an increase in HSF1 DNA binding and HSF1 activity (reviewed in [7, 77]). Activation of HSF1 results in genome wide transcription changes including activation of the production of heat shock proteins, and repression of thousands of genes ([146, 117]).

We used heat shock to investigate whether trisomy 21 impacts the ability of cells to mount a

robust stress response in which the master regulators do not reside on chromosome 21. Previous studies provide mixed messages about how aneuploid cells respond to heat shock. A study by Beach et al found that aneuploidy in yeast leads to increased cell to cell variation in response to heat shock ([22]). Another study found that human fibroblasts with trisomy 21 failed to properly activate the expression of a couple of key heat shock proteins after heat shock ([3]). In published untreated clinical blood samples, we found an irregular elevation of heat shock target genes in individuals with trisomy 21 (Figure 3.2). Therefore, to more directly assess whether trisomy 21 blood cells properly respond to heat shock, we examined the primary effects of heat shock on lymphoblastoid cell lines from two brothers. We found by multiple omics assays that after a short, mild heat shock stress, the trisomy 21 lymphoblastoid cell line activates primary HSF1 regulated transcriptional responses more robustly than the diploid cells.

3.5 Results

3.5.1 Individuals with trisomy 21 have elevated levels of genes related to heat shock in some blood cell lineages.

To determine whether heat shock response is influenced by trisomy 21, we first examined publicly available clinical data. Several projects have collected unperturbed RNA-seq data from clinical samples of multiple blood and skin cell types in individuals with and without trisomy 21, including the Human Trisome Project [109]. In the Human Trisome Project samples ([109, 162, 11, 137, 174], data accessed May 23rd, 2022), we noted that the RNA levels for several heat shock regulated genes were higher in individuals with trisomy 21, particularly in T cells and monocytes (see Figure 3.2A). Specifically, transcript levels for the heat shock regulated genes HSPA8, DNAJA1, HSPH1, HSPA1A, and SERPINH1 are elevated in trisomy 21 groups compared to clinical controls. Furthermore, the transcript levels for HSF1, the master heat shock regulating transcription factor, appears changed in some trisomy 21 cell types (see Figure 3.2B). Published proteomic data further confirmed that HSPA1A, HSPA8, and DNAJB1 protein levels are elevated in blood samples from

individuals with trisomy 21 ([161]). Importantly, transcript and proteomic levels were not increased consistently for all HSF1 regulated genes in all blood cell types (Figure S3.6). Thus the clinical data present a perplexing picture of how trisomy 21 influences the expression of heat shock related genes.

3.5.2 Greater heat shock induced increase in chromatin accessibility at HSF1 sites in trisomic cells

To investigate the trisomic heat shock response outside of clinical complexities, we set out to characterize the acute heat shock response in paired cell lines with and without trisomy 21 (Figure 3.3A). To determine how chromatin accessibility changes as blood cells respond to heat shock, age and gender matched lymphoblastoid cells derived from two brothers were assayed for transposase-accessible chromatin (ATAC-seq) under control conditions 37°, and mild heat shock treatment, e.g. 1 hour at 42° (See Figure 3.3A for design). We used lymphoblastoid cells because they are a readily available blood like cell type. A short time point (1 hour) was chosen to focus on the primary response to heat shock, as this avoids most secondary or downstream effects that arise from cellular feedback mechanisms.

We first confirmed the heat shock response was observable under these conditions in both cell lines by manual inspection of several well-known heat shock genes including HSPB1 and SERPINH1 (Figure 3.3B). These genes showed the expected response, i.e. opening of the chromatin at the promoter region in heat shock compared to control. We called peaks in this data using HMMRATAC and determined a list of peaks that were differentially expressed after heat shock in the trisomic cells and the disomic cells. All genes within 25 kilobases of a differential ATAC-seq peak were used for GO analysis. GO analysis showed the heatshock pathway was strongly activated in both the trisomy 21 and disomic samples. Interestingly, more peaks were called as differently expressed between the control and heat shocked conditions in the trisomic samples than the disomic samples.

We next sought to infer changes in transcription factor activity in response to heat shock in an unbiased fashion for both cell lines. To this end, we employed transcription factor enrichment analysis (TFEA)[150] on HMMRATAC[166] called peaks to determine which transcription factor

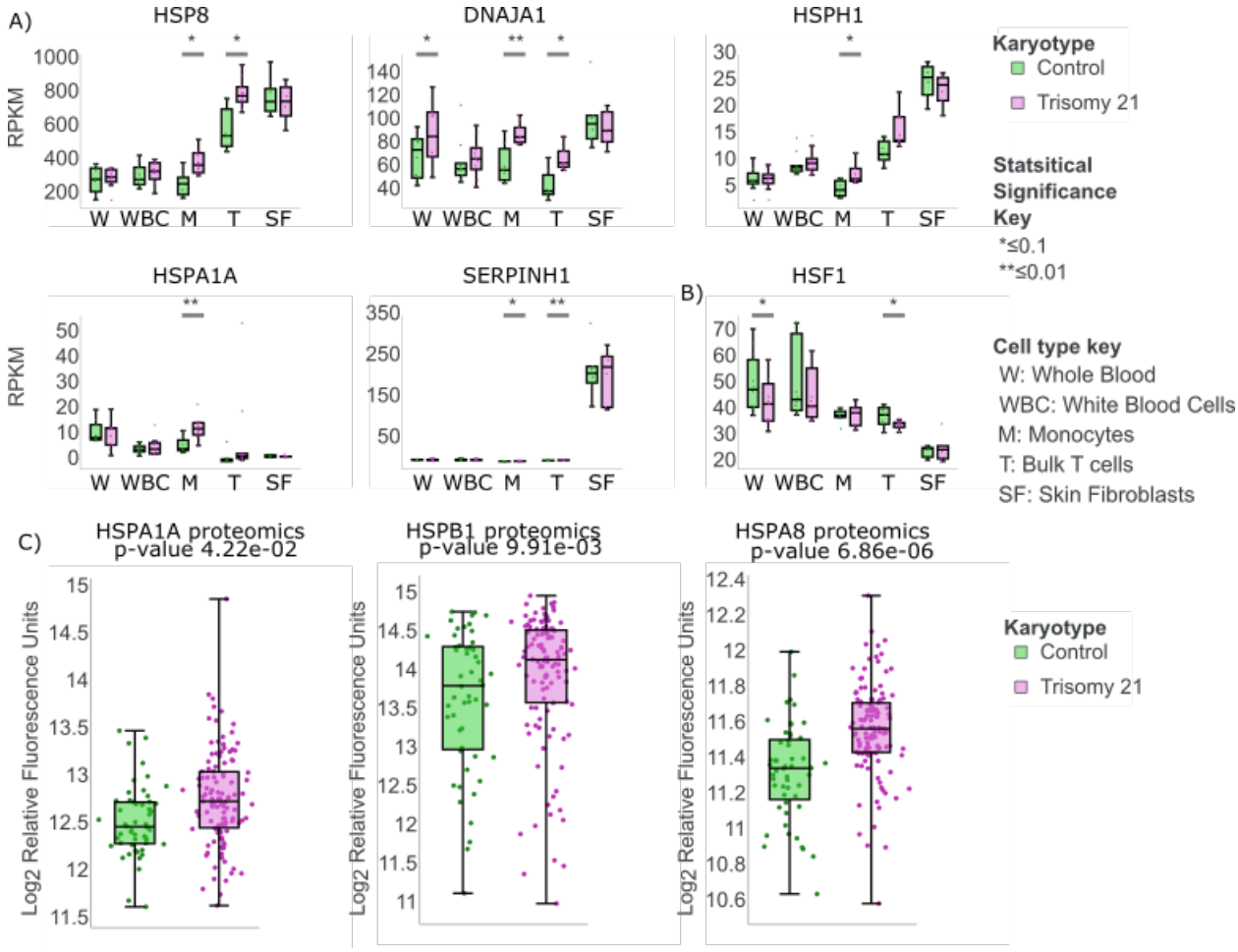


Figure 3.2: **Individuals with trisomy 21 have elevated levels of some heat shock regulated genes under normal conditions.** All Data from the Human Trisome project [109]. A: Several heat shock genes (HSPA8, DNAJA1, HSPH1, HSPA1A, SERPINH1) are differentially expressed (RNA-seq) in multiple blood cell lineages in individuals with trisomy 21 (purple) compared to disomic controls (green). Multiple clinical samples shown: whole blood (W), white blood cells (WBC), monocytes (M), bulk T cells (T), and skin fibroblasts (SF). Significance key: * $\leq$ 0.1, ** $\leq$ 0.01. B: HSF1 transcript levels in the same blood cell lineages. C: Clinical blood sample proteomics data ([161]) shows elevated levels for some heat shock induced genes in plasma from people with trisomy 21.

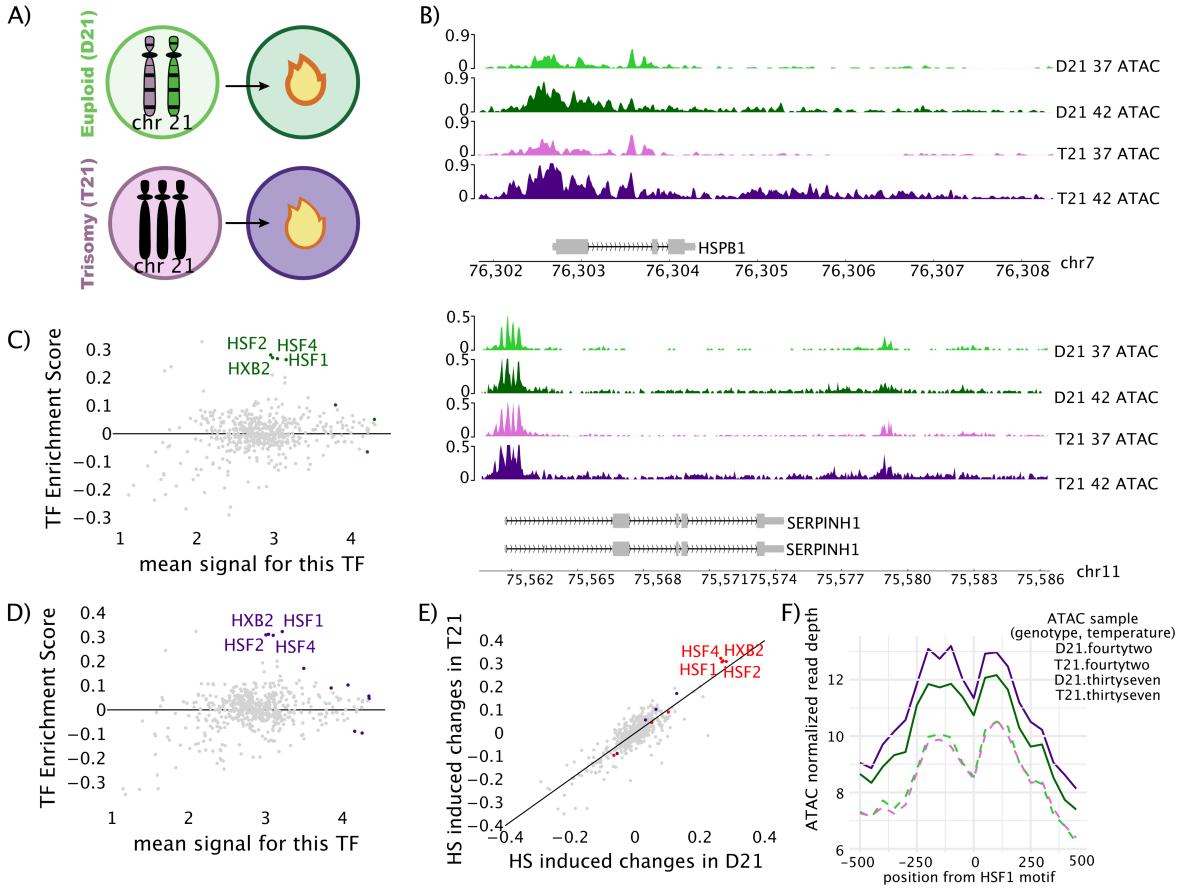


Figure 3.3: **After acute heat shock, cells with trisomy 21 have increased chromatin accessibility near heat shock response elements compared to disomic controls.** A: Conceptual diagram of the conditions analyzed. Disomic cells (green) and trisomic cells (purple) at control (37°, light color) or mild heat shock (42°, dark color). B: ATAC-seq data traces of two heat shock regulated genes (HSPB1, SERPHIN1). All traces for an individual gene have the same y-axis scale. C: MA plot of heat shock induced changes in transcription factor activity in the disomic cell line based on TFEA (Transcription Factor Enrichment Analysis) analysis [150] of ATAC-seq data. Grey: non-siginficant TFs, colored: GC-corrected P-adjusted value of $p<1x10^{-10}$. D: MA plot of heat shock induced changes in transcription factor activity in the trisomic cell line based on TFEA analysis [150] of ATAC-seq data. Grey: non-siginficant TFs, colored: GC-corrected P-adjusted value of $p<1x10^{-10}$. E: Scatter plot comparing heat shock induced changes in transcription factor activity (E-values from TFEA) between disomic cells (X-axis) and trisomic cells (y-axis). Red: significant in both comparisons, Purple: Significant in trisomy only, Green: Significant in disomy only. F: Averaged signal of ATAC-seq data for 1 kilobase region around active HSF1 motifs.

motifs co-associate with observed changes in chromatin accessibility genome wide. TFEA calculates a E-score, or enrichment score, which measures the motif occurrence in regions of interest that have altered accessibility/transcription. For example, if heat shock causes a transcription factor to bind its motif, open chromatin and activates transcription nearby (like HSF1), the E-score for that TF will be positive and high. On the other hand, if a transcription factor is binding to its motif, opening chromatin and activating transcription at 37° and after heat shock that TF leaves DNA then the E-score for that will be negative. Note, transcription factors that repress transcription will show the opposite Escore pattern in PRO-seq. In our data, in both the disomic and trisomic cell lines, TFEA infers that the transcription factors HSF1, HSF2, HSF4, and HXB2 were robustly induced by heat shock (Figure 3.3C, 3.3D, S3.8B). In addition, we directly compared the ATAC-seq TFEA inferred heat shock induced TF changes between the two cell lines and found that the same TFs were significantly upregulated in activity, though the activation was slightly more robust in the trisomic cells after this short, mild heat shock treatment (red TFs in Figure 3.3E).

Therefore, we next sought confirm the above result by characterizing changes in accessibility at known bound HSF1 motifs. To this end, we downloaded HSF1 ChIP-seq data from lymphoblastoid lines and filtered the data for binding sites with the HSF1 motifs ([181], [125], [40]). We then graphed the ATAC-seq signal over the known HSF1 bound sites genome wide (Figure S3.8A). Upon heat shock, both cell lines show increases in accessibility at HSF1 sites, but the trisomic cell line has a more open ATAC-seq signal post heat shock (Figures 3.3F, S3.8A). Overall, our ATAC-seq data suggests that trisomic cells display a slightly elevated chromatin accessibility at HSF1 bound sites after heat shock, compared to disomic cells. This lead us to question whether the difference in chromatin was leading to concomitant changes in gene transcription in the trisomic cells upon heat shock.

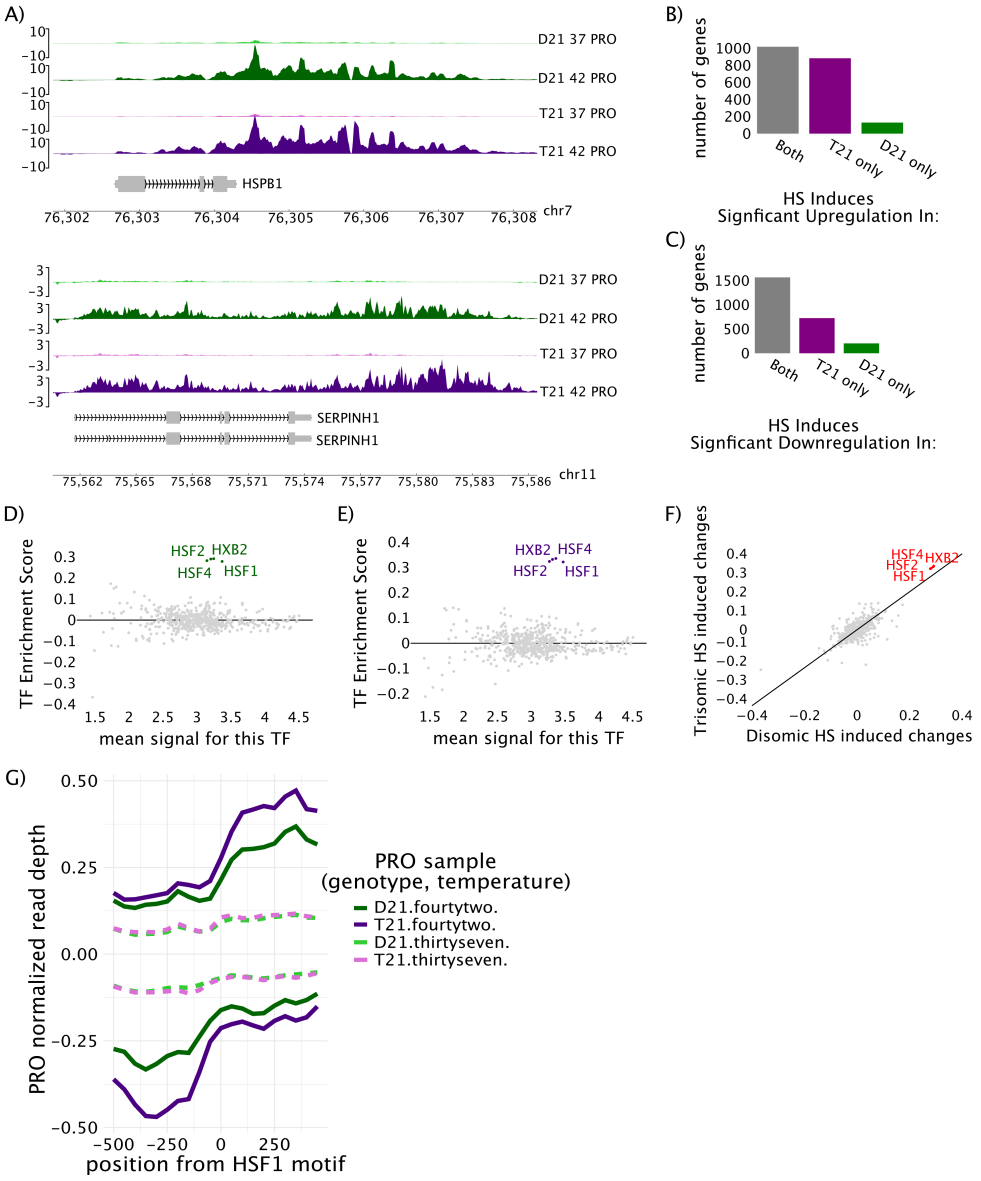


Figure 3.4: **A mild heat shock treatment induces more robust transcriptional changes in the trisomic cell line compared to disomic control.** A: Total read count corrected PRO-seq gene traces at two heat shock regulated genes (Hspb1, Serpinh1) in the four cell types/conditions. B: Number of genes with heat shock induced increases (by DESeq2) in gene transcription (PRO-seq) in one or both cell lines. C: Number of genes with heat shock induced repression (by DESeq2) of gene transcription (PRO-seq) identified in one or both cell lines. D: MA plot of heat shock induced changes in transcription factor activity in the disomic cell line, via TFEA analysis [150]. Grey: non-significant TFs, colored: significant at GC corrected p-adjusted value of $p<1x10^{-10}$. E: MA plot of heat shock induced changes in transcription factor activity in the trisomic cell line, via TFEA analysis [150]. F: Scatter plot comparing TFEA derived GC-corrected E-score values for PRO-seq differences between disomic heat shock (X-axis) and trisomic heat shock (y-axis). Red: significant ($p<1x10^{-10}$) in both comparisons G: Average metaplot of PRO-seq data surrounding (± 500 bp) lymphoblastoid-active HSF1 motifs (at zero).

3.5.3 The trisomic cell line displays larger heat shock induced increases in transcription at HSF1 motifs.

To compare observed changes in chromatin accessibility to changes in transcription, we performed precision run-on sequencing (PRO-seq) in the trisomic and disomic cells at the same time points (before and 1 hr HS) used for the ATAC-seq (Figure 3.3A for design). In both cell lines, heat shock genes such as HSPB1 and SERPINH1 were transcribed at higher levels after heat shock (Figures 3.4A, S3.7A). We used DESeq2 to assess differential gene transcription after heat shock in both samples (Figures 3.4B, 3.4C). Many genes showed a reduction in transcription in response to heat shock in both cell lines (Figures 3.4C, S3.7B). The trisomic sample revealed more genes with significant changes in transcription in response to heat shock than the disomic sample (Figures 3.4B, 3.4C). Moreover, genes that were differentially transcribed in both samples showed a general trend of being induced to a greater extent in the trisomic cell line (Figures S3.7B, S3.7C).

To determine if HSF1 was the only TF with increased transcription associated with its motifs, we used TFEA to infer transcription factor activity changes based the PRO-seq data (independent of the changes in ATAC-seq). To this end, we used Tfit (Transcription fit) to identify all sites of bidirectional transcription within each PRO-seq data set ([17]). Regions of transcription were combined across conditions and replicates using muMerge. Consistent with the ATAC results, TFEA results on the PRO-seq data show a robust activation of HSF1, 2, and 4 in response to heat shock in both the disomic and trisomic cell line (Figures 3.4D, 3.4E). Additionally, many TFs showed a subtle but non-significant reduction in activity in response to heat shock in both cell lines, consistent with widespread transcription repression in response to heat shock treatment (Figures 3.4D, 3.4E). A direct comparison of the heat shock induced changes to TF activity inferred by PRO-seq signal revealed a higher relative activation of HSF TFs in the trisomic cell line compared to the disomic cell line (Figure 3.4F).

Since transcription factor binding sites co-occur with enhancer RNAs which are readily detected by the PRO-seq assay, we next examined nascent transcription at HSF1 binding sites

in response to heat shock. We hypothesized that a more sensitive or robust HSF1 activation in the trisomic cells might explain the increased genome wide changes in chromatin accessibility and transcription in the trisomic cell line compared to the disomic line. Genome wide, heat shock led to an increase in PRO-seq signal at bound HSF1 motifs in both trisomic and disomic cell lines, confirming the activation of HSF1 motif adjacent eRNAs in both cell lines (Figures S3.8C, S3.4G). Though PRO-seq levels began at similar levels in the two cell lines under control conditions, after just the one hour of mild heat shock treatment we noted a more robust transcriptional response in the trisomic cell line compared to the disomic cell line (Figure 3.4G). Collectively, both the PRO-seq and the ATAC-seq suggest that though the transcriptional response to heat shock is similar between the two cell lines, it is more robust in the trisomic cells after just one hour of mild heat shock.

3.5.4 Single cell RNA sequencing confirms the increased heat shock response in trisomic cells is population wide rather than the result of outlier hyper-stressed or dying cells.

The increase in heat shock response observed in trisomic cells could arise from either a small number of hyper responsive cells or a more consistent population wide effect. To address this question, we applied single cell RNA sequencing (scRNA-seq) to the same two cell lines at the same time points, namely before and 1 hr after heat shock. As a control, we first confirmed that the scRNA-seq protocol detected the expected higher quantity of chromosome 21 transcripts in the trisomy 21 sample. To this end, we plotted the depth normalized counts per cell for chromosome 21 encoded genes and found that these transcripts are present in higher quantities in the trisomic cell line than the disomic cell line (Figures 3.5A, 3.5B). Additionally, we examined the transcript levels for known heat shock responsive genes and confirmed heat shock induced increases/decreases in the transcript level for these genes in both cell lines (Figures 3.5C, 3.5D).

To address whether the observed increase in trisomy heat shock response was driven by a small number of outlier cells, we summed all Z-scores for all heat shock genes across individual cells. If a small population of cells had an usually strong heat shock response, those cells should be

expected to have a sum that was an outlier, exceedingly higher than the other cells. Furthermore, the non-outlier cells would be expected to have sums roughly equal to that of the disomic cells. On the other hand, if the whole population of trisomic cells was showing a slightly more robust heat shock then the entire population of disomic cells, then the distribution of cell sums would be shifted. The results show that there is not a small population of super responsive heat shock cells in the trisomy sample. Instead the heat shock difference appears to be population wide: essentially all trisomic cells appear to have an increase in heat shock transcripts relative to the disomic cells (Figure 3.5E). The same result is obtained when using the median of Z-scores rather than the mean (Figure 3.5F). Although some cells respond more strongly than others to heat shock, the T21 cells do not show a different overall pattern than the D21 cells. This suggests that the increased heat shock response in trisomic cell lines is a population wide phenomenon, not the result of a small number of hyper responding cells.

3.6 Discussion

In this study we found that the presence of a third copy of chromosome 21 did not disrupt the cellular ability to mount a heat shock response. Rather, we observed that the trisomic cells were surprisingly agile at changing gene expression in response to this mild perturbation and appeared to increase chromatin accessibility and transcription at HSF1 motifs more readily at this early time point, than the disomic control. Our global analysis of changes in chromatin accessibility and nascent transcription nearby annotated human TF motifs, found that changes in response to heat shock were highly correlated between the two cell lines, but that the degree to which some transcription factors were modulated in response to heat shock differed between the two cell lines. These results indicate that the presence of an extra chromosome 21 copy does not disrupt any major signaling events required for the appropriate transcription response immediately following exposure to heat shock stress.

In this study the trisomic cells responded more aggressively to the short heat shock treatment. After heat shocking the lymphoblastoid cell lines for just one hour at 42 °C, we found a more robust

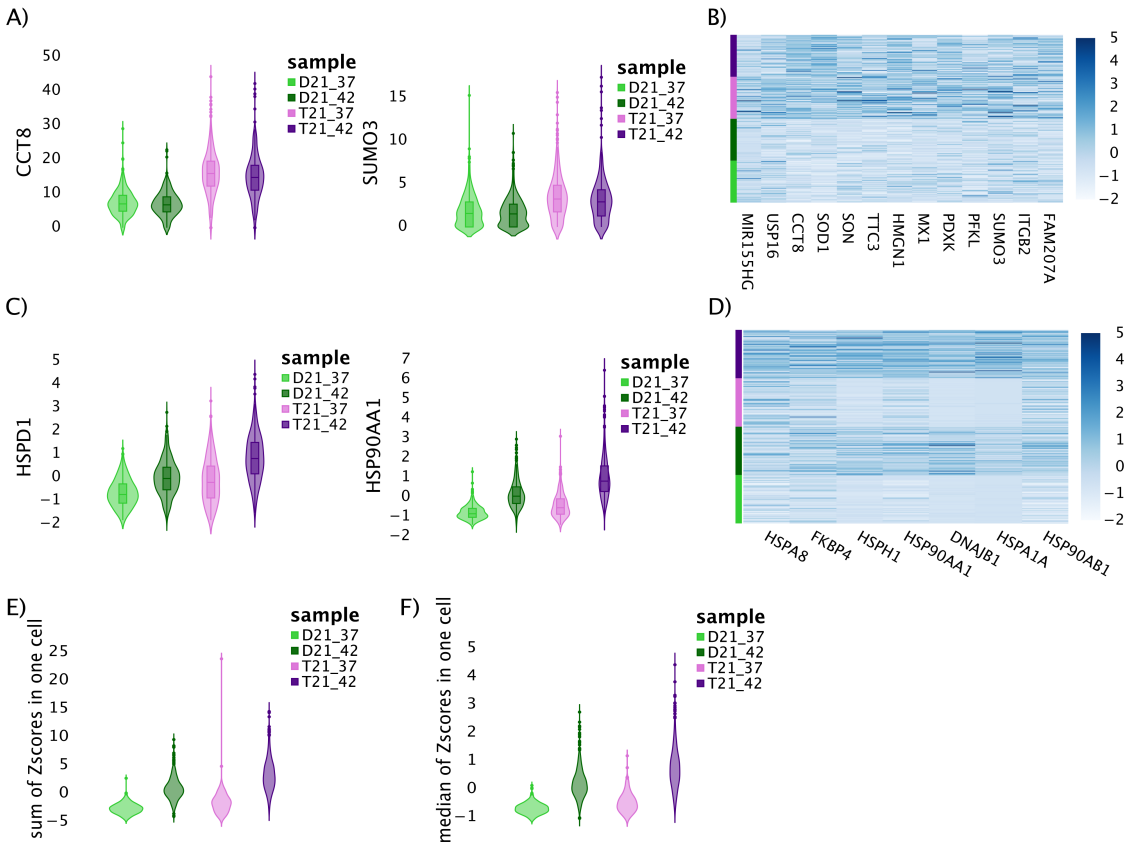


Figure 3.5: **Single Cell RNA-seq indicates that the change in heat shock induced gene expression in trisomy 21 cells is population wide.** A: Violin plots of percent of total normalized scRNA-seq gene counts per cell of two genes (CCT8, SUMO3) present on chromosome 21 in the control (light colors) and HS conditions (dark colors) of the disomic (green) and trisomic (purple) lymphoblastoid cell lines. B: Heatmap of the Z-scores of chromosome 21 genes showing a general up-regulation in expression of genes on chromosome 21. C: Violin plots of two heat shock responsive genes (HSPD1, HSP90AA1) not encoded on chromosome 21, the y-axis is the levels across more than 500 cells in each sample via single cell RNA-seq. D: Heatmap of the Z-scores of heat shock genes show a general up-regulation in expression of heat shock genes, rather than a few cells with extreme heat shock phenotypes. E: Violin plots showing of the sum of Z-scores of all heat shock genes for more than 500 cells in each sample. The genes showed must be present in at least 75% of cells. F: Violin plots showing the median Z-score of all heat shock genes for more than 500 cells in each sample.

activation of HSF1 activity in trisomic cells as inferred from changes in ATAC-seq data, PRO-seq data, and HSF1 regulated steady state transcript levels in scRNA-seq data. The lack of any major heat shock response machinery on chromosome 21 is one reason why heat shock was chosen for this study: we wanted to understand how an extra chromosome might impact global gene regulation as cells mount a transcription program primarily involving genes not present on the aneuploid chromosome. Three genes with links to the heat shock response are present on chromosome 21: HSPA13, DNAJC28 and HSF2BP. It is possible that one of these genes may be impacting the heat shock response directly(Figure 3.1B). However, the low level expression of these genes in lymphoblastoids and lack of known signaling or transcription factor activities would make a causal role of these genes in increased heat response a surprising mechanism.

HSF1 activation is generally thought of as a response to heat shock but can also occur as a result of heat shock independent stresses such as proteotoxic stress from ribosomal gene imbalances ([4]). Crosstalk from other cell stress pathways like those activated in response to oxidative stress, or ER stress, could also play a part in priming cells to over-respond to heat shock. Chromosome 21 encoded genes such as SOD1, DYRK1A, and four interferon response receptors are directly involved in other cellular stress response pathways. Previous studies have suggested that trisomy caused increased dosage of these genes may lead to the over-activation of stress response pathways in cell and tissue samples from individuals with trisomy 21 ([162, 11, 137, 174, 97, 182, 131]. We do see some evidence, that the interferon response may be overactivated in the trisomic lymphoblastoid cell line in our data (Figures 3.3, S3.8C). If any of these other stress response pathways are chronically overactivated in the trisomic cells, crosstalk between these stress responses may cause the overactivated heat shock response observed in this study, and might impact other cellular responses to perturbations.

Despite observing increased heat shock induced HSF1 activity in the trisomic cells, we did not detect a difference between the two cell lines in the transcription levels of the HSF1 transcript itself by PRO-seq or the level of HSF1 transcripts by scRNA-seq (Figures 3.5 S3.9). However, HSF1 is a highly regulated TF [7, 26, 173, 171], so there are many possible mechanisms that could lead to

an elevated HSF1 activation in trisomic cells downstream of HSF1 transcript level changes. A few genes of many with known abilities to regulate HSF1 activity include DAXX, TPR, Mediator, and ribosomal components [26, 173, 4]. The transcript levels for some of these heat shock regulating factors are elevated in the trisomic cells in our scRNA-seq data (Figures 3.5, S3.9). Furthermore, we found that many of these genes were detected in higher quantities in various blood cell types from individuals with trisomy compared to controls, though the results were often inconsistent between blood cell types (Figure 3.2). Note, none of these known HSF1 regulating factors is encoded on chromosome 21, so any trisomy caused misregulation of these genes is likely to be more complex than a trisomy caused gene dosage imbalance. We hypothesize that one of the direct or indirect trisomy caused gene expression changes described above may lead to the observed changes in the regulation of one or more HSF1 regulating genes resulting in an overactivated heat shock response in trisomic blood cells.

Previous studies on the effect of trisomy 21 on stress responses have often focused on stresses known to impact cellular networks regulated by genes on chromosome 21, in which case the increased gene dosage of regulatory genes is expected to lead to a misregulated cellular response. Additionally, measuring the levels of stress response relevant genes in clinical samples can be complicated by the presence of a host of co-morbidity conditions associated with increased occurrence in patients with Down syndrome, but which might be expected to lead to elevated bodily stress. The observation here that a lymphoblastoid cell line with trisomy 21 over-responded to a stress not known to be regulated by a chromosome 21 gene suggests that trisomy 21 may be causing more widespread primary effects on basic gene expression regulation in blood cells than expected.

Because this study began with only a single set of disomic/trisomic cell lines, the next steps for this work need to include more extensive clinical analyses including studying whether this misregulation of HSF1 is a hallmark of Down syndrome and identifying the cell types affected and whether this leads to the depletion or death of any specific immune cell subtype. There are a number of potential consequences that might unfold as a result of misregulating the transcriptional response immediately after heat shock. The irregular blood cell response to heat shock might hamper the

ability of the immune system to respond to infection in the presence of a fever in individuals with Down syndrome. Further, if blood cells with Trisomy 21 over-respond to many common cellular stresses, that might mean that a common trisomy associated mechanism may be hampering the ability of patient's immune systems to respond typically and appropriately to perturbations, and could prove treatable if identified. Future studies would need to investigate whether trisomic blood cells reveal abnormalities in cell survival, or immune cell activation during heat shock stress.

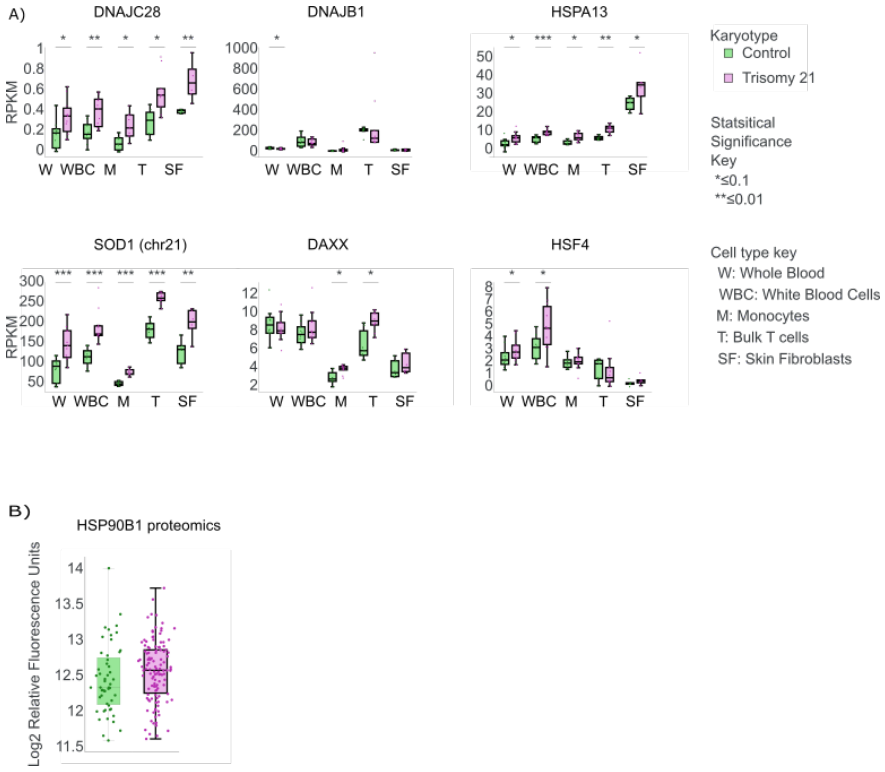


Figure 3.6: **Heat shock genes altered in transcript or protein levels.** All Data from the Human Trisome Project. A: Heat shock relevant genes (DNAJB1, DNAJC28, HSPA13, SOD1, DAXX, HSF4), are elevated (RNA-seq) in individuals with trisomy 21 (green) relative to disomic controls (green). B: Proteomic data for HSPB1 which shows an increased protein levels in individuals with trisomy 21.

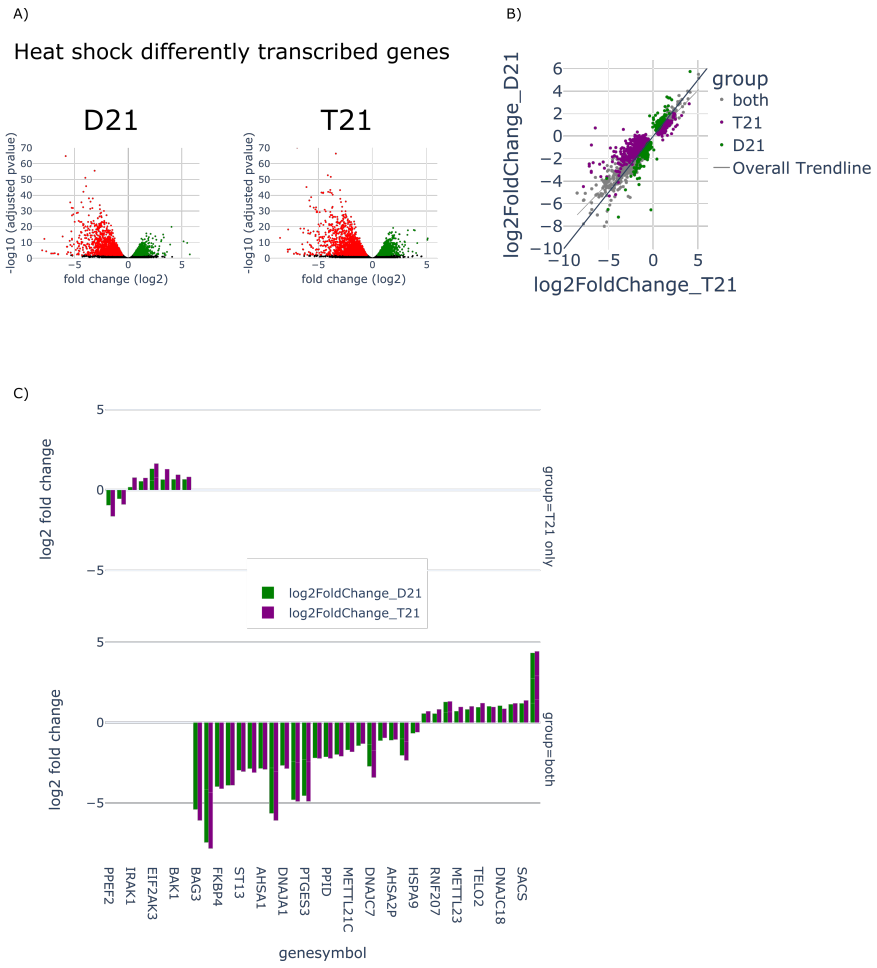


Figure 3.7: **Extended plots of PRO-seq gene transcription** A: Volcano plots of differentially transcribed genes after heat shock. Left: Disomy, Right: Trisomy. Green: up regulated after heat shock, Red: down regulated. B: Scatter plot of the log fold change of the genes changed via heat shock in either trisomy 21 (purple), disomy 21 (green), or both (grey). Best fit line is drawn relative to grey dots only. C: A bar graph of the log fold change of Heat shock genes (as defined by the GO term HEAT_SHOCK_PROTEIN_BINDING.The top plot contains the heat shock genes changed only in the trisomic sample. The bottom plot contains heat shock genes that are differently expressed in both cell types.

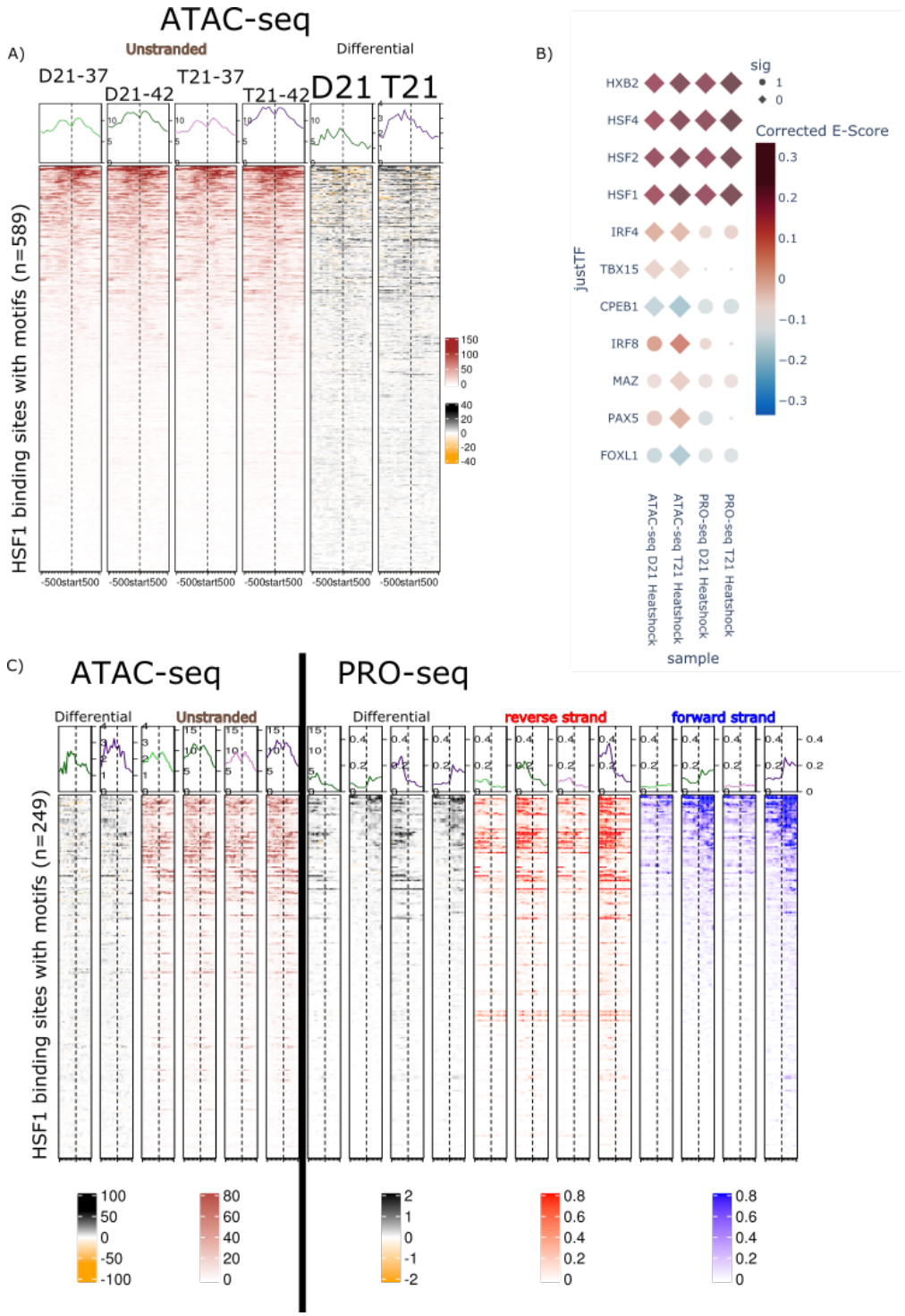

A)
ATAC-seq
Unstranded
Differential
D21-37
D21-42
T21-37
T21-42
D21
T21
HSF1 binding sites with motifs (n=589)
-500start500
B)
sig
1
0
Corrected E-Score
HXB2
HSF4
HSF2
HSF1
IRF4
TBX15
CPEB1
IRF8
MAZ
PAX5
FOXL1
justTF
ATAC-seq D21 Heatshock
ATAC-seq T21 Heatshock
PRO-seq D21 Heatshock
PRO-seq T21 Heatshock
sample
C)
ATAC-seq
PRO-seq
Differential
Unstranded
Differential
reverse strand
forward strand
HSF1 binding sites with motifs (n=249)

Figure 3.8: **(previous page) Extended heatmaps of ATAC-seq and PRO-seq signal over HSF1 sites.** A: Heatmap of ATAC-seq data surrounding HSF1 motifs in HSF1 ChIPseq peaks. First four columns (in red) show DESeq2 size factor normalized ATAC-seq data surrounding (± 500 nts) a HSF1 motif (dashed center line) at all 589 HSF1 ChIP-seq peaks. Top: line graph of median ATAC-seq depth per position. Fifth and Sixth columns (in black) show the difference (control vs heat shock) in ATAC-seq signal after heat shock. B: A plot showing TFs reported as changed by TFEA in at least one comparison. Diamonds: statically significant (p-value $1x10^{-10}$), Circles: not significant. Left two columns show the TFEA Escore for ATAC-seq after heat shock in Disomic and Trisomic cell line, whereas the right two columns show TFEA Escore for PRO-seq. Escore, or enrichment score, measures the motif occurrence with open chromatin (ATAC) or transcription (PRO). Hence red Escores indicate enrichment of the motif within regions of increased transcription/accessibility after heatshock whereas blue indicate enrichment of motif within regions of reduced transcription/accessibility. C: Heatmaps of inter-genic HSF1 bound regions for ATAC-seq and PRO-seq data. Columns correspond to: differential ATAC-seq (black, columns 1-2), ATAC-seq signal (red, columns 3-6), differential PRO-seq signal (black, columns 7-10), reverse strand PRO-seq (red, columns 11-14), forward strand PRO-seq (blue, columns 16-18). Top: line graph of median depth per position, dashed line is site of HSF1 motif.

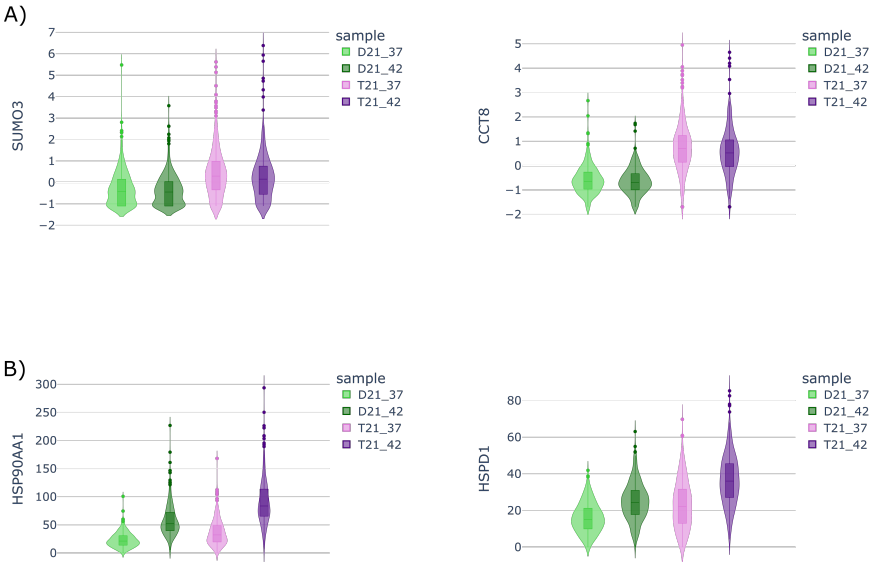


Figure 3.9: **Extended plots of scRNAseq.** A: The same genes as in Figure 4A but with Z scores instead of raw counts. B: The same genes as in Figure 4C but with raw counts instead of Z scores.

Chapter 4

Conclusions

In this dissertation, I sought to understand how Trisomy 21 cells respond to different external stimuli. Stress, whether environmental, molecular, or physiological, often triggers a cascade of cellular responses, changing gene transcription. Essentially, transcription adjusts RNA synthesis based on both external and internal signals. For instance, when cells detect foreign viral RNA or DNA, they produce interferon to initiate an anti-viral defense.

In Chapter 2, I focused on IFN-β on a population of distinct individuals, including two with trisomy 21 DNA copy number. Down syndrome (DS) has been described as an interferonopathy owing to the presence of multiple interferon receptors - IFNAR1, IFNAR2, IFNGR2, and IL10RB - on chromosome 21. Despite the elevated IFN transcription in baseline conditions, the T21 cells showed a rather remarkably consistent response to IFN-β compared to a population of euploid disomic 21 individuals. It appears that genetic background plays a more significant role in the variability of IFN-β responses than the additional chromosome 21 characteristic of DS.

Since individuals with DS have a known elevated interferon response - due to the receptors of type I interferon being encoded on chromosome 21 — I was prompted to investigate if these individuals might respond when an exogenous stressor that is not obviously encoded on chromosome 21. In Chapter 3, we thus examined the Heat Shock Response (HSR), a cellular defense mechanism triggered by various stressors, such as high temperatures, oxidative stress, and immune signaling. Given the pronounced inflammatory response in those with Down syndrome, we hypothesized that T21 cells might exhibit an intensified HSR, perhaps as a compensatory action. To validate this, we

analyzed the HSR in two siblings, one with DS, ensuring a closely matched genetic background. Our results revealed a marginally enhanced response in the DS sibling, suggesting a heightened stress response in them, even for pathways not directly regulated by chromosome 21.

My doctoral work began with a keen interest in the intricacies of blood development, especially in relation to Down syndrome and transcription. A notable irregularity in blood development among individuals with DS sparked my interest in Runt-related transcription factor 1 (RUNX1), a vital regulator of hematopoiesis encoded on chromosome 21. As a multifaceted transcription factor, RUNX1 exhibits various isoforms that seemingly serve distinct roles depending on the cell type. In my research (see Appendix A), I aimed to elucidate the "Offset MD" pattern observed in RUNX1 within multipotent hematopoietic cells. This pattern is not exclusive to RUNX1 but extended to other transcription factors (TFs). Consequently, I developed the Motif Enrichment classifier to differentiate and classify TF based on these Motif Distribution patterns across different cell types. Through classifying TF distribution patterns, I hoped to identify additional TFs with offset patterns that might provide insight into their cellular roles.

Furthermore, to comprehend RUNX1's role more holistically, it was essential to locate genome regions displaying the "Offset" signature. Doing so could provide a clearer picture of their regulatory influence on transcription. Notably, RUNX1's unique enhancer-associated RNA (eRNA) profile varies across cell types, implying these profiles are likely cell-type-specific, stemming from their individualized roles within cells. This insight motivated the development of the Regulatory Activity Decoder (RAD) construct (outlined in Appendix B). In addition to exploring RUNX1, the RAD construct can also be leveraged to study transcription factor activity patterns across cell types and transcription regulation shifts based on DNA copy numbers.

At its core, this dissertation seeks to illuminate the intricacies of gene regulation. Below, I'll outline the limitations and future work related to each of these projects.

4.1 Transcription regulation under interferon perturbation

My study depended on the limited individual cell lines available to me. Specifically, I had data from only two individuals with Down syndrome – who were dramatically distinct with respect to their IFN score in baseline (BSA) conditions. This presented a challenge when comparing their interferon responses to typical individuals. Ideally, a much larger cohort is needed for a more comprehensive view of the range of responses within a population of individuals with Trisomy 21. Another limitation was our time points which focus on the the early response to IFN-β to which we did not detect distinct responses between cell lines. Given that the secondary response captured with RNA-seq had more variance between cell lines, a longer time point may provide more insights into the contribution of copy number to transcription alteration. Likewise, it would be interesting to examine other cell types and potentially even alternative IFNs (IFN-α or IFN-γ).

When analyzing a population of data, one must decide whether to analyze the samples independently and then merge the results or to evaluate all the samples concurrently. I tried both strategies to variable success. Using all the data in a single DESeq2 design table gained considerable power to detect small changes but at the expense of washing out individual specific changes that were at the heart of our interest. Thus most of my work focused on independent analysis first. However, this strategy likely has higher noise, as evidenced by the inconsistency in down-regulated genes. Perhaps better tools for population level analysis will be available in the future that better balance these trade-offs.

On a related note, one particularly challenging aspect of this work was that it is difficult to discern changes in transcription levels when a gene is not transcribed or lowly transcribed prior to the perturbation. To address this, I made analytical adjustments to filter out potential noise from low-expressed genes, which gave me cleaner results but has the risk of potentially omitting biologically relevant small changes that are inherently difficult to detect.

Additionally, my research did not delve into gene isoforms, variants of genes that can differ in function. A prime example is IFNAR2, a type I receptor that is encoded on chromosome 21,

which exists in three isoforms, each with varying roles in interferon signaling; IFNAR2a (soluble truncated form), IFNAR2b (truncated transmembrane form missing the intracellular domain), and IFNAR2c (long transmembrane form that facilitates IFN-I signaling). Notably, several crucial splicing regulators (U2AF1L5, RBM1, U2AF1, DRYK1A) are located on chromosome 21[80, 119]. The presence of certain genes on chromosome 21 might influence IFNAR2 isoform and IFN signaling, yet this connection remains unstudied. To adequately tackle splicing would require significantly deeper RNA-seq data.

Lastly, while elevated baseline interferon-stimulated gene (ISG) levels were observed in T21 cell lines, it is worth considering how this sustained stress might alter the chromatin structure both before and after perturbation. Future research could deploy techniques like ATAC-seq to examine chromatin accessibility and its potential alterations under chronic stress conditions.

4.2 Transcription regulation under low grade heat shock

In the heat shock work, the primary strength of our study was the minimized genetic variance by studying siblings. However, this approach simultaneously poses a limitation: relying solely on a single pair of disomic and trisomic cell lines. To genuinely understand if the dysregulation of heat shock factor genes is emblematic of Down syndrome, more expansive clinical research involving a larger participant pool is essential. While we began our investigation with Lymphoblastoid cell line (LCL) — given their pronounced responsiveness to cellular stress — it is possible that other cell subtypes, like cancer cells (known for their active Heat Shock Response (HSR), could be more fitting for this study. Notably, as blood cells exhibit a robust IFN response, our study might inadvertently focus more on the mechanism of IFN pathways rather than HSR.

The DnA lab now has Induced Pluripotent Stem Cell (iPSC) lines derived from an individual mosaic for T21. Thus the two iPSC cell lines are isogenic, differing only in the copy number of chromosome 21. These cells would, ideally, further minimize genetic background variation allowing for a clearer picture of T21 induced versus individual variation based changes. While this strategy has its merits, it is worth noting that iPSCs are generally less reactive than other cell types,

and recalcitrant to IFN stimulation in general. To address this, future studies should consider differentiating iPSCs into cell types more pertinent for such research.

4.3 Transcription activity in different cell types

In collaboration with The Gates Institute at the University of Colorado Anschutz, we have embarked on developing iPSC featuring a Runt-related transcription factor 1 (RUNX1) dosage normalized knockout. Originating from an individual with mosaic T21, these cell lines underwent normalization via a CRISPR knockout system. A T21 RUNX1$^{+/+/-}$ cell line was produced which allows us to study DNA copy number variation for a single TF. Our ongoing research with these cells aims to determine the influence of RUNX1 on blood cell typology. Once established, introducing the Regulatory Activity Decoder (RAD) construct could shed light on transcription modifications in relation to varying DNA copy numbers.

My offset study primarily focused on RUNX1, potentially overlooking other transcription factors with significant roles. Addressing this limitation would entail broadening our analysis to other relevant TFs. In either case, substantial experimental work is needed to further support the biological relevance of the offset pattern. For example, ChIP-seq or Cut-and-Run for TFs in offset-inducing cell lines would confirm physical occupancy of these sites by this TF. Genetic manipulation, via a CRISPR strategy, could then begin to explore the impact of these offset sites on cellular activity and differentiation.

I've also developed two promising tools that, with refinement, can further the understanding of TF activity. The ME Classifier could be instrumental in identifying active TFs in cells. Categorizing TFs based on their MD signature not only infers TF activity but also explores novel signatures of potential biological importance, such as the offset pattern. The RAD construct, on the other hand, holds promise in evaluating regulatory regions, especially enhancers. Although I highlighted potential issues with the RAD construct's readout, I have also suggested avenues for optimization. For instance, integrating the construct into the genome could address the limitations introduced by transfection-related stress. Moreover, there is observable interference between the two dual-reporter

regions, but this can likely be mitigated by altering the number of transfer DNA (tDNA) in the insulator region, reducing the chances of cross-talk.

4.4 Concluding Remarks

This dissertation offers valuable insights into the influence of stressors on transcription and TF activity. TF play a pivotal role in controlling transcription, which is fundamental to regulating gene expression. Contrary to the prevailing belief that transcription dysregulation in DS will lead to dramatic T21-specific transcription profiles, my research suggests that individual genetic variations have a more profound impact on the cell's response to environmental stimuli.

The findings from this research could reshape how population studies are approached. Acknowledging the variability in transcription regulation across populations might be crucial when modeling diseases.

Bibliography

[1] M. H. Abdolvahab, M. Mofrad, and H. Schellekens. Interferon beta: from molecular level to therapeutic effects. International review of cell and molecular biology, 326:343–372, 2016.

[2] V. R. Agashe and F.-U. Hartl. Roles of molecular chaperones in cytoplasmic protein folding. In Seminars in cell & developmental biology, volume 11, pages 15–25. Elsevier, 2000.

[3] S. Aivazidis, C. M. Coughlan, A. K. Rauniyar, H. Jiang, L. A. Liggett, K. N. Maclean, and J. R. Roede. The burden of trisomy 21 disrupts the proteostasis network in down syndrome. PloS one, 12(4):e0176307, 2017.

[4] B. Albert, I. C. Kos-Braun, A. K. Henras, C. Dez, M. P. Rueda, X. Zhang, O. Gadal, M. Kos, and D. Shore. A ribosome assembly stress response regulates transcription to maintain proteome homeostasis. Elife, 8:e45002, 2019.

[5] B. Alberts. Molecular biology of the cell. Garland science, 2017.

[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, et al. Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. Elife, 3:e02200, 2014.

[7] J. Anckar and L. Sistonen. Regulation of hsf1 function in the heat stress response: implications in aging and disease. Annual review of biochemistry, 80:1089–1115, 2011.

[8] S. E. Antonarakis, R. Lyle, E. T. Dermitzakis, A. Reymond, and S. Deutsch. Chromosome 21 and down syndrome: from genomics to pathophysiology. Nature reviews genetics, 5(10):725, 2004.

[9] S. E. Antonarakis, B. G. Skotko, M. S. Rafii, A. Strydom, S. E. Pape, D. W. Bianchi, S. L. Sherman, and R. H. Reeves. Down syndrome. Nature Reviews Disease Primers, 6(1):9, 2020.

[10] F. Antonaros, R. Zenatelli, G. Guerri, M. Bertelli, C. Locatelli, B. Vione, F. Catapano, A. Gori, L. Vitale, M. C. Pelleri, et al. The transcriptome profile of human trisomy 21 blood cells. Human genomics, 15(1):1–14, 2021.

[11] P. Araya, K. A. Waugh, K. D. Sullivan, N. G. Núñez, E. Roselli, K. P. Smith, R. E. Granrath, A. L. Rachubinski, B. E. Estrada, E. T. Butcher, et al. Trisomy 21 dysregulates t cell lineages toward an autoimmunity-prone state associated with interferon hyperactivity. Proceedings of the National Academy of Sciences, 116(48):24231–24241, 2019.

[12] A. Arvin, G. Campadelli-Fiume, E. Mocarski, P. S. Moore, B. Roizman, R. Whitley, and K. Yamanishi. Human herpesviruses: biology, therapy, and immunoprophylaxis. Cambridge University Press, 2007.

[13] N. Au-Yeung, R. Mandhana, and C. M. Horvath. Transcriptional regulation by stat1 and stat2 in the interferon jak-stat pathway. Jak-stat, 2(3):e23931, 2013.

[14] J. Azofeifa, M. A. Allen, M. Lladser, and R. Dowell. Fstitch: A fast and simple algorithm for detecting nascent rna transcripts. In Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, pages 174–183, 2014.

[15] J. G. Azofeifa, M. A. Allen, J. R. Hendrix, J. D. Rubin, and R. D. Dowell. Enhancer rna profiling predicts transcription factor activity. Genome research, 28(3):334–344, 2018.

[16] J. G. Azofeifa, M. A. Allen, M. E. Lladser, and R. D. Dowell. An annotation agnostic algorithm for detecting nascent rna transcripts in gro-seq. IEEE/ACM transactions on computational biology and bioinformatics, 14(5):1070–1081, 2016.

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

[18] T. L. Bailey. Dreme: motif discovery in transcription factor chip-seq data. Bioinformatics, 27(12):1653–1659, 2011.

[19] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine learning, 21:51–80, 1995.

[20] G. M. Barlow, X.-N. Chen, Z. Y. Shi, G. E. Lyons, D. M. Kurnit, L. Celle, N. B. Spinner, E. Zackai, M. J. Pettenati, A. J. Van Riper, et al. Down syndrome congenital heart disease: a narrowed region and a candidate gene. Genetics in Medicine, 3(2):91, 2001.

[21] P. Bastard, L. B. Rosen, Q. Zhang, E. Michailidis, H.-H. Hoffmann, Y. Zhang, K. Dorgham, Q. Philippot, J. Rosain, V. Béziat, et al. Autoantibodies against type i ifns in patients with life-threatening covid-19. Science, 370(6515):eabd4585, 2020.

[22] R. R. Beach, C. Ricci-Tam, C. M. Brennan, C. A. Moomau, P.-h. Hsu, B. Hua, R. E. Silberman, M. Springer, and A. Amon. Aneuploidy causes non-genetic individuality. Cell, 169(2):229–242, 2017.

[23] S. Behjati and P. S. Tarpey. What is next generation sequencing? Archives of Disease in Childhood-Education and Practice, 98(6):236–238, 2013.

[24] J. Bekisz, S. Baron, C. Balinsky, A. Morrow, and K. C. Zoon. Antiproliferative properties of type i and type ii interferon. Pharmaceuticals, 3(4):994–1015, 2010.

[25] I. Berest, C. Arnold, A. Reyes-Palomares, G. Palla, K. D. Rasmussen, H. Giles, P.-M. Bruch, W. Huber, S. Dietrich, K. Helin, et al. Quantification of differential transcription factor activity and multiomics-based classification into activators and repressors: difftf. Cell reports, 29(10):3147–3159, 2019.

[26] J. K. Björk and L. Sistonen. Regulation of the members of the mammalian heat shock factor family. The FEBS journal, 277(20):4126–4139, 2010.

[27] B. L. Bloemers, C. J. Broers, L. Bont, M. E. Weijerman, R. J. Gemke, and A. M. van Furth. Increased risk of respiratory tract infections in children with down syndrome: the consequence of an altered immune system. Microbes and Infection, 12(11):799–808, 2010.

[28] J. C. Bryne, E. Valen, M.-H. E. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, and A. Sandelin. Jaspar, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic acids research, 36(suppl_1):D102–D106, 2007.

[29] J. D. Buenrostro, B. Wu, H. Y. Chang, and W. J. Greenleaf. Atac-seq: a method for assaying chromatin accessibility genome-wide. Current protocols in molecular biology, 109(1):21–29, 2015.

[30] M. J. Bull. Down syndrome. New England Journal of Medicine, 382(24):2344–2352, 2020.

[31] M. J. Bull and C. on Genetics. Health supervision for children with down syndrome, 2011.

[32] 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.

[33] CDC. Data and statistics on down syndrome, 06 2017.

[34] C. Chai, Z. Xie, and E. Grotewold. Selex (systematic evolution of ligands by exponential enrichment), as a powerful tool for deciphering the protein–dna interaction space. Plant Transcription Factors: Methods and Protocols, pages 249–258, 2011.

[35] I. Chambers and S. R. Tomlinson. The transcriptional foundation of pluripotency. Development, 2009.

[36] C.-K. Cheng, T. H. Wong, T. S. Wan, A. Z. Wang, N. P. Chan, N. C. Chan, C.-K. Li, and M. H. Ng. Runx1 upregulation via disruption of long-range transcriptional control by a novel t (5; 21)(q13; q22) translocation in acute myeloid leukemia. Molecular Cancer, 17(1):1–6, 2018.

[37] S. Chevrier, Y. Zurbuchen, C. Cervia, S. Adamo, M. E. Raeber, N. de Souza, S. Sivapatham, A. Jacobs, E. Bachli, A. Rudiger, et al. A distinct innate immune signature marks progression from mild to severe covid-19. Cell Reports Medicine, 2(1), 2021.

[38] D. Chmiest, N. Sharma, N. Zanin, C. Viaris de Lesegno, M. Shafaq-Zadah, V. Sibut, F. Dingli, P. Hupé, S. Wilmes, J. Piehler, et al. Spatiotemporal control of interferon-induced jak/stat signalling and gene transcription by the retromer complex. Nature communications, 7(1):13476, 2016.

[39] A. K. Clift, C. A. Coupland, R. H. Keogh, H. Hemingway, and J. Hippisley-Cox. Covid-19 mortality risk in down syndrome: results from a cohort study of 8 million adults. Annals of internal medicine, 174(4):572–576, 2021.

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

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

[42] O. Corradin and P. C. Scacheri. Enhancer variants: evaluating functions in common disease. Genome medicine, 6(10):1–14, 2014.

[43] Y. J. Crow and D. B. Stetson. The type i interferonopathies: 10 years on. Nature Reviews Immunology, 22(8):471–483, 2022.

[44] M. E. Curtis, D. Yu, and D. Praticò. Dysregulation of the retromer complex system in down syndrome. Annals of neurology, 88(1):137–147, 2020.

[45] C. G. Danko, S. L. Hyland, L. J. Core, A. L. Martins, C. T. Waters, H. W. Lee, V. G. Cheung, W. L. Kraus, J. T. Lis, and A. Siepel. Identification of active transcriptional regulatory elements from gro-seq data. Nature methods, 12(5):433–438, 2015.

[46] C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, et al. The encyclopedia of dna elements (encode): data portal update. Nucleic acids research, 46(D1):D794–D801, 2018.

[47] Y. C. de Hingh, P. W. van der Vossen, E. F. Gemen, A. B. Mulder, W. C. Hop, F. Brus, and E. de Vries. Intrinsic abnormalities of lymphocyte counts in children with down syndrome. The Journal of pediatrics, 147(6):744–747, 2005.

[48] A. Debiec-Bak, D. Wojtowicz, L. Pawik, A. Ptak, and A. Skrzek. Analysis of body surface temperatures in people with down syndrome after general rehabilitation exercise. Journal of Thermal Analysis and Calorimetry, 135:2399–2410, 2019.

[49] G. Degols, P. Eldin, and N. Mechti. Isg20, an actor of the innate immune response. Biochimie, 89(6-7):831–835, 2007.

[50] J.-M. Delabar, D. Theophile, Z. Rahmani, Z. Chettouh, J.-L. Blouin, M. Prieur, B. Noel, and P.-M. Sinet. Molecular mapping of twenty-four features of down syndrome on chromosome 21. European Journal of Human Genetics, 1(2):114–124, 1993.

[51] B. Dey, S. Thukral, S. Krishnan, M. Chakrobarty, S. Gupta, C. Manghani, and V. Rani. Dna–protein interactions: methods for detection and analysis. Molecular and cellular biochemistry, 365(1-2):279–299, 2012.

[52] J. L. Down. On some of the mental affections of childhood and youth. J. & A. Churchill, 1887.

[53] K. E. Elagib, F. K. Racke, M. Mogass, R. Khetawat, L. L. Delehanty, and A. N. Goldfarb. Runx1 and gata-1 coexpression and cooperation in megakaryocytic differentiation. blood, 101(11):4333–4341, 2003.

[54] C. J. Epstein, N. H. McManus, L. B. Epstein, A. A. Branca, S. B. D'Alessandro, and C. Baglioni. Direct evidence that the gene product of the human chromosome 21 locus, ifrc, is the interferon-α receptor. Biochemical and biophysical research communications, 107(3):1060–1066, 1982.

[55] 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, et al. Tfiid enables rna polymerase ii promoter-proximal pausing. Molecular cell, 78(4):785–793, 2020.

[56] A. Filippone and D. Praticò. Endosome dysregulation in down syndrome: a potential contributor to alzheimer disease pathology. Annals of neurology, 90(1):4–14, 2021.

[57] T. M. Filtz, W. K. Vogel, and M. Leid. Regulation of transcription factor activity by interconnected post-translational modifications. Trends in pharmacological sciences, 35(2):76–85, 2014.

[58] K. Fishler and R. Koch. Mental development in down syndrome mosaicism. American Journal of Mental Retardation: AJMR, 96(3):345–351, 1991.

[59] D. R. FitzPatrick. Transcriptional consequences of autosomal trisomy: primary gene dosage with complex downstream effects. Trends in Genetics, 21(5):249–253, 2005.

[60] V. Fitzpatrick, A. Rivelli, S. Chaudhari, L. Chicoine, G. Jia, A. Rzhetsky, and B. Chicoine. Prevalence of infectious diseases among 6078 individuals with down syndrome in the united states. Journal of Patient-Centered Research and Reviews, 9(1):64, 2022.

[61] O. Fornes, J. A. Castro-Mondragon, A. Khan, R. Van der Lee, X. Zhang, P. A. Richmond, B. P. Modi, S. Correard, M. Gheorghe, D. Baranašić, et al. Jaspar 2020: update of the open-access database of transcription factor binding profiles. Nucleic acids research, 48(D1):D87–D92, 2020.

[62] K. R. Fox. Dnase i footprinting. Drug-DNA Interaction Protocols, pages 1–22, 1997.

[63] M. Fu, H. Chen, Z. Cai, Y. Yang, Z. Feng, M. Zeng, L. Chen, Y. Qin, B. Cai, P. Zhu, et al. Forkhead box family transcription factors as versatile regulators for cellular reprogramming to pluripotency. Cell Regeneration, 10:1–11, 2021.

[64] M. D. Galbraith, K. T. Kinning, K. D. Sullivan, P. Araya, K. P. Smith, R. E. Granrath, J. R. Shaw, R. Baxter, K. R. Jordan, S. Russell, et al. Specialized interferon action in covid-19. Proceedings of the National Academy of Sciences, 119(11):e2116730119, 2022.

[65] M. D. Galbraith, A. L. Rachubinski, K. P. Smith, P. Araya, K. A. Waugh, B. Enriquez-Estrada, K. Worek, R. E. Granrath, K. T. Kinning, N. Paul Eduthan, et al. Multidimensional definition of the interferonopathy of down syndrome and its response to jak inhibition. Science Advances, 9(26):eadg6218, 2023.

[66] K. Gardiner and A. C. Costa. The proteins of human chromosome 21. American Journal of Medical Genetics Part C: Seminars in Medical Genetics, 142(3):196–205, 2006.

[67] A.-M. Gerdes, M. Hørder, P. H. Petersen, and V. Bonnevie-Nielsen. Effect of increased gene dosage expression on the α-interferon receptors in down's syndrome. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1181(2):135–140, 1993.

[68] J. Gertz, T. E. Reddy, K. E. Varley, M. J. Garabedian, and R. M. Myers. Genistein and bisphenol a exposure cause estrogen receptor 1 to bind thousands of sites in a cell type-specific manner. Genome research, 22(11):2153–2162, 2012.

[69] S. Giannone, P. Strippoli, L. Vitale, R. Casadei, S. Canaider, L. Lenzi, P. D'Addabbo, F. Frabetti, F. Facchin, A. Farina, et al. Gene expression profile analysis in human t lymphocytes from patients with down syndrome. Annals of Human Genetics, 68(6):546–554, 2004.

[70] A. B. Glick, A. Wodzinski, P. Fu, A. D. Levine, and D. N. Wald. Impairment of regulatory t-cell function in autoimmune thyroid disease. Thyroid, 23(7):871–878, 2013.

[71] C. Gongora, G. Degols, L. Espert, T. D. Hua, and N. Mechti. A unique isre, in the tata-less human isg20 promoter, confers irf-1-mediated responsiveness to both interferon type i and type ii. Nucleic acids research, 28(12):2333–2341, 2000.

[72] S. J. Gross, J. C. Ferreira, B. Morrow, P. Dar, B. Funke, D. Khabele, and I. Merkatz. Gene expression profile of trisomy 21 placentas: a potential approach for designing noninvasive techniques of prenatal diagnosis. American journal of obstetrics and gynecology, 187(2):457–462, 2002.

[73] K. Guo, G. Shen, J. Kibbie, T. Gonzalez, S. M. Dillon, H. A. Smith, E. H. Cooper, K. Lavender, K. J. Hasenkrug, K. Sutter, et al. Qualitative differences between the ifnα subtypes and ifnβ influence chronic mucosal hiv-1 pathogenesis. PLoS pathogens, 16(10):e1008986, 2020.

[74] Z. Guo, L. Wang, R. Eisensmith, and S. Woo. Evaluation of promoter strength for hepatic gene expression in vivo following adenovirus-mediated gene transfer. Gene therapy, 3(9):802–810, 1996.

[75] J. D. Hasday and I. S. Singh. Fever and the heat shock response: distinct, partially overlapping processes. Cell stress & chaperones, 5(5):471, 2000.

[76] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. Molecular cell, 38(4):576–589, 2010.

[77] C. Huang, J. Wu, L. Xu, J. Wang, Z. Chen, and R. Yang. Regulation of hsf1 protein stabilization: An updated review. European journal of pharmacology, 822:69–77, 2018.

[78] S. Hunter, R. D. Dowell, J. Hendrix, J. Freeman, and M. A. Allen. Transcription dosage compensation does not occur in down syndrome. bioRxiv, pages 2023–06, 2023.

[79] T. Hunter and M. Karin. The regulation of transcription by phosphorylation. Cell, 70(3):375–387, 1992.

[80] S. Hwang, P. Cavaliere, R. Li, L. J. Zhu, N. Dephoure, and E. M. Torres. Consequences of aneuploidy in human fibroblasts with trisomy 21. Proceedings of the National Academy of Sciences, 118(6):e2014723118, 2021.

[81] T. Illouz, A. Biragyn, M. Frenkel-Morgenstern, O. Weissberg, A. Gorohovski, E. Merzon, I. Green, F. Iulita, L. Flores-Aguilar, M. Dierssen, et al. Specific susceptibility to covid-19 in adults with down syndrome. Neuromolecular medicine, 23:561–571, 2021.

[82] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-dna interactions. Science, 316(5830):1497–1502, 2007.

[83] A. Jolma, T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J. M. Vaquerizas, J. Yan, M. J. Sillanpää, et al. Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. Genome research, 20(6):861–873, 2010.

[84] J. Judd, L. A. Wojenski, L. M. Wainman, N. D. Tippens, E. J. Rice, A. Dziubek, G. J. Villafano, E. M. Wissink, P. Versluis, L. Bagepalli, et al. A rapid, sensitive, scalable method for precision run-on sequencing (pro-seq). BioRxiv, pages 2020–05, 2020.

[85] P. Kahlem, M. Sultan, R. Herwig, M. Steinfath, D. Balzereit, B. Eppens, N. G. Saran, M. T. Pletcher, S. T. South, G. Stetten, et al. Transcript level alterations reflect gene dosage effects across multiple tissues in a mouse model of down syndrome. Genome research, 14(7):1258–1267, 2004.

[86] A. Karmiloff-Smith, T. Al-Janabi, H. D'Souza, J. Groet, E. Massand, K. Mok, C. Startin, E. Fisher, J. Hardy, D. Nizetic, et al. The importance of understanding individual differences in down syndrome. F1000Research, 5, 2016.

[87] T.-K. Kim, M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature, 465(7295):182–187, 2010.

[88] X.-F. Kong, L. Worley, D. Rinchai, V. Bondet, P. V. Jithesh, M. Goulet, E. Nonnotte, A. S. Rebillat, M. Conte, C. Mircher, et al. Three copies of four interferon receptor genes underlie a mild type i interferonopathy in down syndrome. Journal of clinical immunology, 40:807–819, 2020.

[89] J. R. Korenberg, X. Chen, R. Schipper, Z. Sun, R. Gonsky, S. Gerwehr, N. Carpenter, C. Daumer, P. Dignan, and C. Disteche. Down syndrome phenotypes: the consequences of chromosomal imbalance. Proceedings of the National Academy of Sciences, 91(11):4997–5001, 1994.

[90] K. R. Kukurba and S. B. Montgomery. Rna sequencing and analysis. Cold Spring Harbor Protocols, 2015(11):pdb–top084970, 2015.

[91] I. V. Kulakovskiy, I. E. Vorontsov, I. S. Yevshin, R. N. Sharipov, A. D. Fedorova, E. I. Rumynskiy, Y. A. Medvedeva, A. Magana-Mora, V. B. Bajic, D. 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.

[92] L. Kuras and K. Struhl. Binding of tbp to promoters in vivo is stimulated by activators and requires pol ii holoenzyme. Nature, 399(6736):609–613, 1999.

[93] M. Kusters, R. Verstegen, E. Gemen, and E. De Vries. Intrinsic defect of the immune system in children with down syndrome: a review. Clinical & Experimental Immunology, 156(2):189–193, 2009.

[94] H. Kwak, N. J. Fuda, L. J. Core, and J. T. Lis. Precise maps of rna polymerase reveal how promoters direct initiation and pausing. Science, 339(6122):950–953, 2013.

[95] M. T. Lam, H. Cho, H. P. Lesch, D. Gosselin, S. Heinz, Y. Tanaka-Oishi, C. Benner, M. U. Kaikkonen, A. S. Kim, M. Kosaka, et al. Rev-erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature, 498(7455):511–515, 2013.

[96] 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.

[97] C. Lanzillotta and F. Di Domenico. Stress responses in down syndrome neurodegeneration: State of the art and therapeutic molecules. Biomolecules, 11(2):266, 2021.

[98] D. S. Latchman. Transcription factors: an overview. The international journal of biochemistry & cell biology, 29(12):1305–1312, 1997.

[99] L. Le Breton and M. P. Mayer. A model for handling cell stress. Elife, 5:e22850, 2016.

[100] J. Lejenu. Les chromosomes humains en culture de tissues. Com. Rend. Acad. Sci., 248:602–603, 1959.

[101] A. Letourneau, F. A. Santoni, X. Bonilla, M. R. Sailani, D. Gonzalez, J. Kind, C. Chevalier, R. Thurman, R. S. Sandstrom, Y. Hibaoui, et al. Domains of genome-wide gene expression dysregulation in down's syndrome. Nature, 508(7496):345–350, 2014.

[102] S. Leviyang. Interferon stimulated binding of isre is cell type specific and is predicted by homeostatic chromatin state. Cytokine: X, 3(4):100056, 2021.

[103] C.-M. Li, M. Guo, M. Salas, N. Schupf, W. Silverman, W. B. Zigman, S. Husain, D. Warburton, H. Thaker, and B. Tycko. Cell type-specific over-expression of chromosome 21 genes in fibroblasts and fetal hearts with trisomy 21. BMC medical genetics, 7(1):1–15, 2006.

[104] J. J. Li, P. J. Bickel, and M. D. Biggin. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ, 2:e270, 2014.

[105] W. Li, D. Notani, Q. Ma, B. Tanasa, E. Nunez, A. Y. Chen, D. Merkurjev, J. Zhang, K. Ohgi, X. Song, et al. Functional roles of enhancer rnas for oestrogen-dependent transcriptional activation. Nature, 498(7455):516–520, 2013.

[106] W. Li, D. Notani, and M. G. Rosenfeld. Enhancers as non-coding rna transcription units: recent insights and future perspectives. Nature Reviews Genetics, 17(4):207–223, 2016.

[107] X.-y. Li, S. MacArthur, R. Bourgon, D. Nix, D. A. Pollard, V. N. Iyer, A. Hechmer, L. Simirenko, M. Stapleton, C. L. L. Hendriks, et al. Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm. PLoS Biol, 6(2):e27, 2008.

[108] Y. Liao, G. K. Smyth, and W. Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7):923–930, 2014.

[109] Linda Crnic Institute. Human Trisome Project, 2021.

[110] H. Lockstone, L. Harris, J. Swatton, M. Wayland, A. Holland, and S. Bahn. Gene expression profiling in the adult down syndrome brain. Genomics, 90(6):647–660, 2007.

[111] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology, 15(12):1–21, 2014.

[112] Y. N. Löwensteyn, E. W. Phijffer, J. V. Simons, N. M. Scheltema, N. I. Mazur, H. Nair, and L. J. Bont. Respiratory syncytial virus-related death in children with down syndrome: the rsv gold study. The Pediatric infectious disease journal, 39(8):665, 2020.

[113] M.-C. Luo, S.-Y. Zhou, D.-Y. Feng, J. Xiao, W.-Y. Li, C.-D. Xu, H.-Y. Wang, and T. Zhou. Runt-related transcription factor 1 (runx1) binds to p50 in macrophages and enhances tlr4-triggered inflammation and septic shock. Journal of Biological Chemistry, 291(42):22011–22020, 2016.

[114] R. Lyle, F. Béna, S. Gagos, C. Gehrig, G. Lopez, A. Schinzel, J. Lespinasse, A. Bottani, S. Dahoun, L. Taine, et al. Genotype–phenotype correlations in down syndrome identified by array cgh in 30 cases of partial trisomy and partial monosomy chromosome 21. European Journal of Human Genetics, 17(4):454–466, 2009.

[115] R. Lyle, C. Gehrig, C. Neergaard-Henrichsen, S. Deutsch, and S. E. Antonarakis. Gene expression from the aneuploid chromosome in a trisomy mouse model of down syndrome. Genome research, 14(7):1268–1274, 2004.

[116] D. B. Mahat, H. Kwak, G. T. Booth, I. H. Jonkers, C. G. Danko, R. K. Patel, C. T. Waters, K. Munson, L. J. Core, and J. T. Lis. Base-pair-resolution genome-wide mapping of active rna polymerases using precision nuclear run-on (pro-seq). Nature protocols, 11(8):1455, 2016.

[117] D. B. Mahat, H. H. Salamanca, F. M. Duarte, C. G. Danko, and J. T. Lis. Mammalian heat shock response and mechanisms underlying its genome-wide transcriptional regulation. Molecular cell, 62(1):63–78, 2016.

[118] L. Malle, P. Bastard, A. Martin-Nalda, T. Carpenter, D. Bush, R. Patel, R. Colobran, P. Soler-Palacin, J.-L. Casanova, M. Gans, et al. Atypical inflammatory syndrome triggered by sars-cov-2 in infants with down syndrome. Journal of Clinical Immunology, 41(7):1457–1462, 2021.

[119] L. Malle and D. Bogunovic. Down syndrome and type i interferon: not so simple. Current Opinion in Immunology, 72:196–205, 2021.

[120] R. Mao, X. Wang, E. L. Spitznagel, L. P. Frelin, J. C. Ting, H. Ding, J.-w. Kim, I. Ruczinski, T. J. Downey, and J. Pevsner. Primary and secondary transcriptional effects in the developing human down syndrome brain and heart. Genome biology, 6(13):1–20, 2005.

[121] M. K. McCormick, A. Schinzel, M. B. Petersen, G. Stetten, D. J. Driscoll, E. S. Cantu, L. Tranebjaerg, M. Mikkelsen, P. C. Watkins, and S. E. Antonarakis. Molecular genetic approach to the characterization of the "down syndrome region" of chromosome 21. Genomics, 5(2):325–331, 1989.

[122] R. Medzhitov and T. Horng. Transcriptional control of the inflammatory response. Nature Reviews Immunology, 9(10):692–703, 2009.

[123] M. P. Meers, T. D. Bryson, J. G. Henikoff, and S. Henikoff. Improved cut&run chromatin profiling tools. elife, 8:e46314, 2019.

[124] A. Mégarbané, A. Ravel, C. Mircher, F. Sturtz, Y. Grattau, M.-O. Rethoré, J.-M. Delabar, and W. C. Mobley. The 50th anniversary of the discovery of trisomy 21: the past, present, and future of research and treatment of down syndrome. Genetics in Medicine, 11(9):611, 2009.

[125] S. Mei, Q. Qin, Q. Wu, H. Sun, R. Zheng, C. Zang, M. Zhu, J. Wu, X. Shi, L. 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.

[126] S. Moncini, A. Bevilacqua, M. Venturin, C. Fallini, A. Ratti, A. Nicolin, and P. Riva. The 3'untranslated region of human cyclin-dependent kinase 5 regulatory subunit 1 contains regulatory elements affecting transcript stability. BMC Molecular Biology, 8:1–14, 2007.

[127] R. I. Morimoto. Regulation of the heat shock transcriptional response: cross talk between a family of heat shock factors, molecular chaperones, and negative regulators. Genes & development, 12(24):3788–3796, 1998.

[128] R. Morris, N. J. Kershaw, and J. J. Babon. The molecular details of cytokine signaling via the jak/stat pathway. Protein Science, 27(12):1984–2009, 2018.

[129] K. Mousavi, H. Zare, S. Dell'Orso, L. Grontved, G. Gutierrez-Cruz, A. Derfoul, G. L. Hager, and V. Sartorelli. ernas promote transcription by establishing chromatin accessibility at defined genomic loci. Molecular cell, 51(5):606–617, 2013.

[130] C. G. Mullighan, J. R. Collins-Underwood, L. A. Phillips, M. G. Loudin, W. Liu, J. Zhang, J. Ma, E. Coustan-Smith, R. C. Harvey, C. L. Willman, et al. Rearrangement of crlf2 in b-progenitor–and down syndrome–associated acute lymphoblastic leukemia. Nature genetics, 41(11):1243, 2009.

[131] G. Pagano and G. Castello. Oxidative stress and mitochondrial dysfunction in down syndrome. Neurodegenerative Diseases, pages 291–299, 2012.

[132] P. Papavassiliou, C. Charalsawadi, K. Rafferty, and C. Jackson-Cook. Mosaicism for trisomy 21: a review. American Journal of Medical Genetics Part A, 167(1):26–39, 2015.

[133] N. Pencovich, R. Jaschek, A. Tanay, and Y. Groner. Dynamic combinatorial interactions of runx1 and cooperating partners regulates megakaryocytic differentiation in cell line models. Blood, The Journal of the American Society of Hematology, 117(1):e1–e14, 2011.

[134] A. Piovesan, M. C. Pelleri, F. Antonaros, P. Strippoli, M. Caracausi, and L. Vitale. On the length, weight and gc content of the human genome. BMC research notes, 12(1):1–7, 2019.

[135] R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard. Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. Genome research, 21(3):447–455, 2011.

[136] M.-C. Potier, I. Rivals, G. Mercier, L. Ettwiller, R. Moldrich, J. Laffaire, L. Personnaz, J. Rossier, and L. Dauphinot. Transcriptional disruptions in down syndrome: a case study in the ts1cje mouse cerebellum during post-natal development. Journal of neurochemistry, 97:104–109, 2006.

[137] R. K. Powers, R. Culp-Hill, M. P. Ludwig, K. P. Smith, K. A. Waugh, R. Minter, K. D. Tuttle, H. C. Lewis, A. L. Rachubinski, R. E. Granrath, et al. Trisomy 21 activates the kynurenine pathway via increased dosage of interferon receptors. Nature communications, 10(1):1–11, 2019.

[138] N. Proudfoot. Connecting transcription to messenger rna processing. Trends in biochemical sciences, 25(6):290–293, 2000.

[139] M. Ptashne and A. Gann. Transcriptional activation by recruitment. Nature, 386(6625):569–577, 1997.

[140] K. Pulakanti, L. Pinello, C. Stelloh, S. Blinka, J. Allred, S. Milanovich, S. Kiblawi, J. Peterson, A. Wang, G.-C. Yuan, et al. Enhancer transcribed rnas arise from hypomethylated, tet-occupied genomic regions. Epigenetics, 8(12):1303–1320, 2013.

[141] D. Qin. Next-generation sequencing and its clinical application. Cancer biology & medicine, 16(1):4, 2019.

[142] J. Y. Qin, L. Zhang, K. L. Clift, I. Hulur, A. P. Xiang, B.-Z. Ren, and B. T. Lahn. Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. PloS one, 5(5):e10611, 2010.

[143] J. R. Raab, J. Chiu, J. Zhu, S. Katzman, S. Kurukuti, P. A. Wade, D. Haussler, and R. T. Kamakaka. Human trna genes function as chromatin insulators. The EMBO journal, 31(2):330–350, 2012.

[144] G. Ram and J. Chinen. Infections and immunodeficiency in down syndrome. Clinical & Experimental Immunology, 164(1):9–16, 2011.

[145] I. Rauch, M. Müller, and T. Decker. The regulation of inflammation by interferons and their stats. Jak-Stat, 2(1):e23820, 2013.

[146] J. Ray, P. R. Munn, A. Vihervaara, J. J. Lewis, A. Ozer, C. G. Danko, and J. T. Lis. Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. Proceedings of the National Academy of Sciences, 116(39):19431–19439, 2019.

[147] K. Richter, M. Haslbeck, and J. Buchner. The heat shock response: life on the verge of death. Molecular cell, 40(2):253–266, 2010.

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

[149] E. J. Rozen, C. D. Ozeroff, and M. A. Allen. Run (x) out of blood: emerging runx1 functions beyond hematopoiesis and links to down syndrome. Human Genomics, 17(1):83, 2023.

[150] 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): Quantifying the activity of hundreds of transcription factors from a single experiment. Nature Communications Biology, 2021.

[151] G. M. Savva, J. K. Morris, D. E. Mutton, and E. Alberman. Maternal age-specific fetal loss rates in down syndrome pregnancies. Prenatal Diagnosis: Published in Affiliation With the International Society for Prenatal Diagnosis, 26(6):499–504, 2006.

[152] J. W. Schoggins. Interferon-stimulated genes: what do they all do? Annual review of virology, 6:567–584, 2019.

[153] J. M. Sheltzer, E. M. Torres, M. J. Dunham, and A. Amon. Transcriptional consequences of aneuploidy. Proceedings of the National Academy of Sciences, 109(31):12644–12649, 2012.

[154] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbsnp: the ncbi database of genetic variation. Nucleic acids research, 29(1):308–311, 2001.

[155] P. J. Skene and S. Henikoff. An efficient targeted nuclease strategy for high-resolution mapping of dna binding sites. elife, 6:e21856, 2017.

[156] S. T. Smale and G. Natoli. Transcriptional control of inflammatory responses. Cold Spring Harbor perspectives in biology, 6(11):a016261, 2014.

[157] M. F. Smith and S. Delbary-Gossart. Electrophoretic mobility shift assay (emsa). Colorectal Cancer: methods and protocols, pages 249–257, 2001.

[158] G. Stark, I. Kerr, B. Williams, R. Silverman, and R. Schreiber. How cells respond to interferons annu rev biochem. Annual Review of Biochemistry, 1998.

[159] S. Stingele, G. Stoehr, K. Peplowska, J. Cox, M. Mann, and Z. Storchova. Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. Molecular systems biology, 8(1):608, 2012.

[160] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences, 100(16):9440–9445, 2003.

[161] K. D. Sullivan, D. Evans, A. Pandey, T. H. Hraha, K. P. Smith, N. Markham, A. L. Rachubinski, K. Wolter-Warmerdam, F. Hickey, J. M. Espinosa, and T. Blumenthal. Trisomy 21 causes changes in the circulating proteome indicative of chronic autoinflammation. Scientific Reports, 7(1):14818, 2017.

[162] K. D. Sullivan, H. C. Lewis, A. A. Hill, A. Pandey, L. P. Jackson, J. M. Cabral, K. P. Smith, L. A. Liggett, E. B. Gomez, M. D. Galbraith, J. DeGregori, and J. M. Espinosa. Trisomy 21 consistently activates the interferon response. eLife, 5:e16220, jul 2016.

[163] A. Tacheny, M. Dieu, T. Arnould, and P. Renard. Mass spectrometry-based identification of proteins interacting with nucleic acids. Journal of proteomics, 94:89–109, 2013.

[164] Y. Tan, E. Schneider, J. Tischfield, C. Epstein, and F. Ruddle. Human chromosome 21 dosage: effect on the expression of the interferon induced antiviral state. Science, 186(4158):61–63, 1974.

[165] Y. Tang, M. B. Schapiro, D. N. Franz, B. J. Patterson, F. J. Hickey, E. K. Schorry, R. J. Hopkin, M. Wylie, T. Narayan, T. A. Glauser, et al. Blood expression profiles for tuberous sclerosis complex 2, neurofibromatosis type 1, and down's syndrome. Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society, 56(6):808–814, 2004.

[166] E. D. Tarbell and T. Liu. Hmmratac: a hidden markov modeler for atac-seq. Nucleic acids research, 47(16):e91–e91, 2019.

[167] C. Teschendorf, K. H. Warrington Jr, D. W. Siemann, and N. Muzyczka. Comparison of the ef-1 alpha and the cmv promoter for engineering stable tumor cell lines using recombinant adeno-associated virus. Anticancer research, 22(6A):3325–3330, 2002.

[168] G. Thibault and D. Ng. Encyclopedia of Biological Chemistry II. Elsevier, 2013. Heat/Stress Responses.

[169] I. J. Tripodi, T. J. Callahan, J. T. Westfall, N. S. Meitzer, R. D. Dowell, and L. E. Hunter. Applying knowledge-driven mechanistic inference to toxicogenomics. Toxicology in Vitro, 66:104877, 2020.

[170] B. Turner. Chip with native chromatin: advantages and problems relative to methods using cross-linked material. Mapping Protein/DNA Interactions by Cross-Linking, 2001.

[171] A. Vihervaara and L. Sistonen. Hsf1 at a glance. Journal of cell science, 127(2):261–266, 2014.

[172] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. Nature methods, 17(3):261–272, 2020.

[173] R. Voellmy. On mechanisms that control heat shock transcription factor activity in metazoan cells. Cell stress & chaperones, 9(2):122, 2004.

[174] K. A. Waugh, P. Araya, A. Pandey, K. R. Jordan, K. P. Smith, R. E. Granrath, S. Khanal, E. T. Butcher, B. E. Estrada, A. L. Rachubinski, et al. Mass cytometry reveals global immune remodeling with multi-lineage hypersensitivity to type i interferon in down syndrome. Cell reports, 29(7):1893–1908, 2019.

[175] A. J. Wilk, A. Rustagi, N. Q. Zhao, J. Roque, G. J. Martínez-Colón, J. L. McKechnie, G. T. Ivison, T. Ranganath, R. Vergara, T. Hollis, et al. A single-cell atlas of the peripheral immune response in patients with severe covid-19. Nature medicine, 26(7):1070–1076, 2020.

[176] F. K. Wiseman, T. Al-Janabi, J. Hardy, A. Karmiloff-Smith, D. Nizetic, V. L. Tybulewicz, E. M. Fisher, and A. Strydom. A genetic cause of alzheimer disease: mechanistic insights from down syndrome. Nature Reviews Neuroscience, 16(9):564–574, 2015.

[177] E. M. Wissink, A. Vihervaara, N. D. Tippens, and J. T. Lis. Nascent rna analyses: tracking transcription and its regulation. Nature Reviews Genetics, 20(12):705–723, 2019.

[178] L. Yao, J. Liang, A. Ozer, A. K.-Y. Leung, J. T. Lis, and H. Yu. A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers. Nature biotechnology, 40(7):1056–1065, 2022.

[179] B. Yoo, R. Seidl, N. Cairns, and G. Lubec. Heat-shock protein 70 levels in brain of patients with Down syndrome and Alzheimer's disease. Springer, 1999.

[180] B. C. Yoo, R. Vlkolinsky, E. Engidawork, N. Cairns, M. Fountoulakis, and G. Lubec. Differential expression of molecular chaperones in brain of patients with down syndrome. Electrophoresis, 22(6):1233–1241, 2001.

[181] R. Zheng, C. Wan, S. Mei, Q. Qin, Q. Wu, H. Sun, C.-H. Chen, M. Brown, X. Zhang, C. A. Meyer, and X. S. Liu. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Research, 47(D1):D729–D735, 11 2018.

[182] P. J. Zhu, S. Khatiwada, Y. Cui, L. C. Reineke, S. W. Dooling, J. J. Kim, W. Li, P. Walter, and M. Costa-Mattioli. Activation of the isr mediates the behavioral and neurophysiological abnormalities in down syndrome. Science, 366(6467):843–849, 2019.

Appendix A

Predicting RUNX1 transcription factor activity through the use of a Motif Enrichment Classifier

In this Appendix, I outline work conducted on clustering transcription factor (TF) motif distribution patterns. This work builds on the work of former graduate students (Joey Azofeifa and Jonathan Rubin, [15, 150]) and leverages the "barcodes" that the Azofeifa et al 2018[15] created for examining the distribution of TF motif instances (genome mapping of the motifs) relative to sites of RNA polymerase II initiation (inferred using the Tfit[17] algorithm).

A.1 Background

Transcription factors (TFs) are proteins that orchestrate the transcription of DNA into RNA. This regulation allows for altered gene expression as necessary for cell development and environmental responses. TFs have a DNA-binding domain that gives them the ability to recognize and bind to specific DNA sequences. Binding is typically measured by ChIP-seq, DNA footprinting, and southwestern blotting[51]. The set of sequences recognized by a particular TF is represented as the DNA binding motif (also known as a position specific scoring matrix (PSSM)). It should be noted that not all sites of binding lead to RNA polymerase II (RNAP II) activity changes nearby [107, 96, 98]. Furthermore, not all motif sequence occurrences in the DNA are bound by TFs.

At a subset of binding sites, a TF alters RNA polymerase II activity nearby. TF proteins are able to interact with multiple other proteins and protein complexes to regulate transcription through a variety of mechanisms that include priming the regulatory regions and altering DNA

chromatin folding[98, 96]. TF regulatory activity can be detected by the proximity of short, unstable, bidirectional RNA transcripts proximal to the TF binding site[116, 23, 15], effectively providing a readout on nearby TF activity. However, the presence of transcription does not tell you which TF(s) are active. Luckily, this can be inferred in perturbation experiments based on the relative proximity of a TF motif to the patterns of altered transcription[15, 150].

Two tools, Tfit[17, 15] and dREG[45] have been developed to identify sites of bidirectional transcription directly from nascent sequencing data. This pattern of bidirectional transcription is seen at nearly all sites of RNA polymerase II initiation genome-wide, including both promoters and enhancers[41, 94, 23].

The Dowell lab has developed a suite of tools[15, 150] to combine these sites of bidirectional transcription with sequence information to infer TF activity changes in response to perturbation. Briefly, a score is calculated that measures motif proximity to the precise location of RNA polymerase II initiation at sites of differential transcription. This co-localization score is diagnostic for which TFs are driving observed changes. A canonical active transcription factor protein will have significant occurrences of bidirectionals whose centers are co-localized to the TF motifs and an inactive transcription factor will have a uniform distribution of TF motifs relative to RNA centers[15]. This pattern can be visualized in a heat map when the heat is the abundance of motif counts at that relative position (relative to the site of RNA polymerase II initiation), see Figure A.1.

The goal of this work was to cluster and classify the patterns present within these heat maps.

A.2 Computational characterization of motif displacement distributions

A.2.1 Motif Displacement and Enrichment Data

Initially, the bidirectional calls and motif scans were obtained from the Azofeifa paper[15]. Briefly, FStitch[14] was used to identify regions of the genome that are transcribed within nascent transcription data sets. These regions were then fed to Tfit[17], a probabilistic model that identifies the signature of RNA polymerase II activity through the identification of bidirectional transcripts

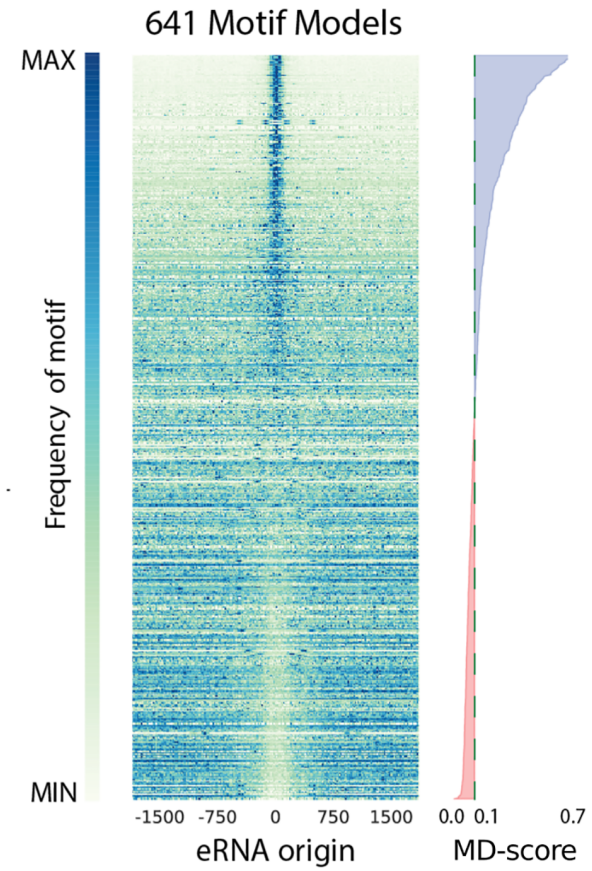
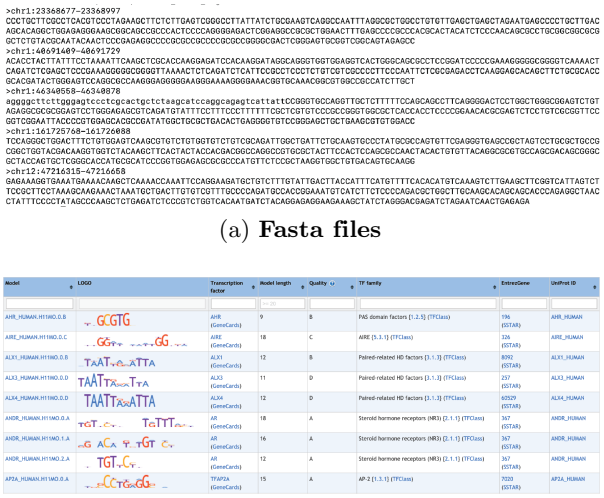


Figure A.1: **Motif co-localization with eRNA origins varies by cell type.** From Azofeifa et al., 2016[16], the heatmap shows the motif distribution patterns in a single cell type. Each row is a single TF motif distribution and each column is the binning of motif count where the heat is proportional to the frequency of a motif instance at that distance from an RNA center.

in nascent transcription data. A database of TF motifs, HOCOMOCO[91], was used to scan the sequence (Figure A.2a) surrounding each site of RNA polymerase II initiation which is defined as the center of the bidirectional for motif hits (Figure A.2c). The nascent transcription dataset includes details on 641 human TF motifs (Figure A.2b). The dataset presents counts for each position in relation to the RNA polymerase II initiation sites, which is defined as zero. The data spans from −1500 bp to +1500 bp (Figure A.2d). This process is illustrated in the schematic Figure A.3.

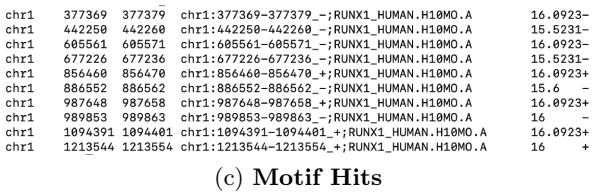
The Motif Distribution (MD) score was calculated for each of the 641 proteins (transcription factors). The MD score counts the number of motifs across the 3000 base pair window to identify Active and Inactive calls. A 300 base pair window was based on biological reasoning that there is 150 bp on either DNA strand wraps around a histone protein in a cell. To capture background noise, we found that 3000 base pair is a sufficient window size.

$$\text{MD score} = \frac{\text{Motif counts in 300 bp}}{\text{Motif counts in 3000 bp}}$$

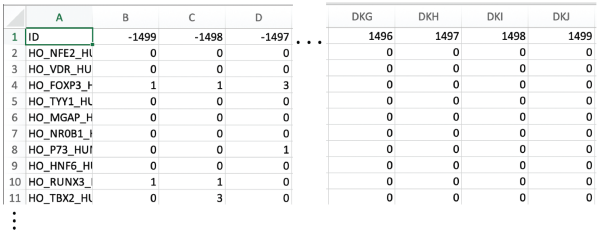


(a) **Fasta files**

(b) **HOCOMOCO Human motifs**



(c) **Motif Hits**



(d) **Alignment of bidirectionals and TF motif hits**

Figure A.2: **Input data for TFPeakDetect.** A.2a The input data originally is downloaded as a fasta file. A.2b The HOCOMOCO database with human TF motifs A.2c Each motif is mapped to positions within the reference human genome (eg., hg38) A.2d Aligning the bidirectional against the mapped motif sequences, we can calculate the positional distribution of motif instances relative to sites of RNA polymerase II initiation (the data underlying the heat maps).

To understand our data we first started by building summary plots of the entire data set –

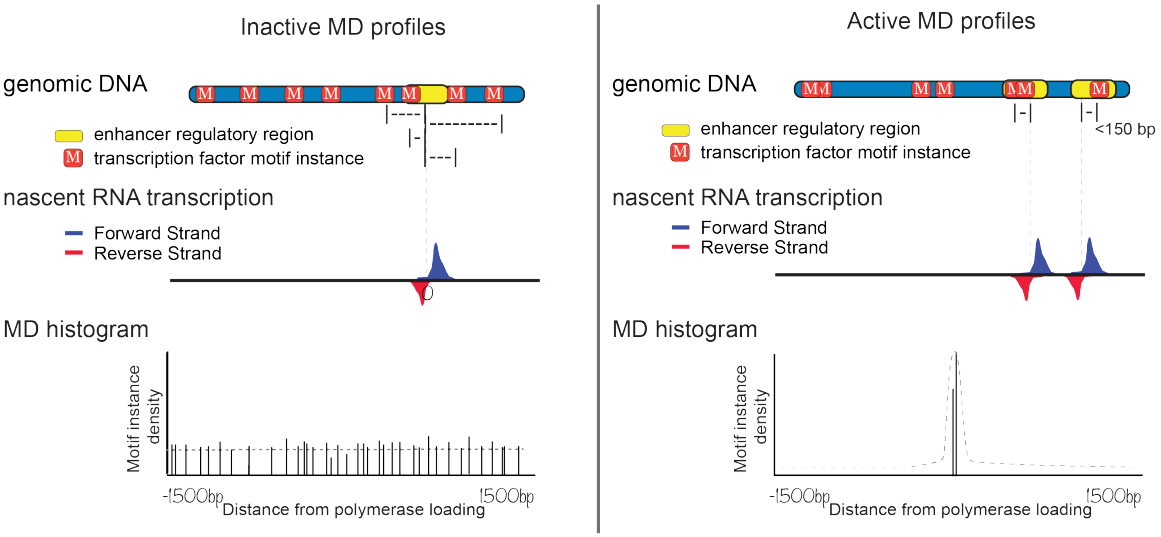
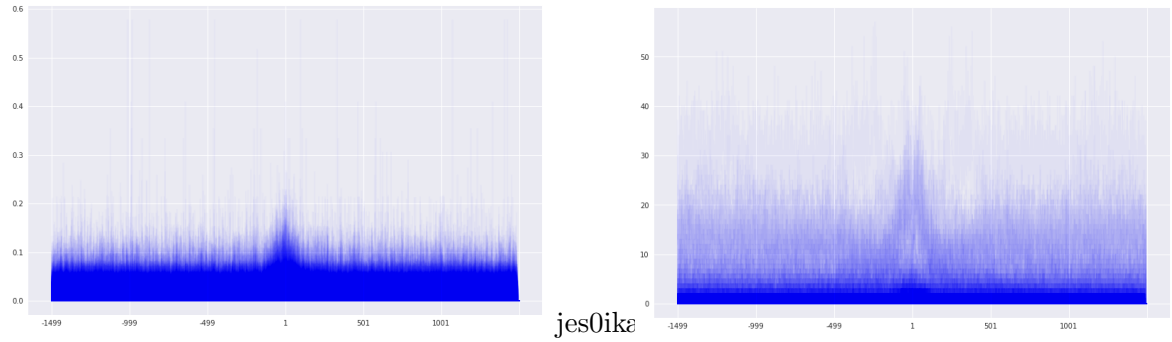


Figure A.3: **Schematic of eRNA profiling to calculate MD score.** Illustration of how MD score is calculated to distinguish between "Inactive" MD (left) and "Active" MD (right). Top bar is the genomic DNA with the static motif mapping, the middle bar is the nascent RNA with bidirectional defined by Tfit, and the bottom plot is the resulting MD histogram that counts motif relative to sites of RNA polymerase II initiation (bidirectional's center). This is the data underlying the heat map.

all 641 proteins overlaid. To this end, we generate a metadata plot to observe any distinguishing patterns (Figure A.4a). Two patterns are strongly observed in this plot; the uniform signal across the 3000 base pair reflects the "Inactive" (e.g. background) pattern and a strong peak at position 0 reflects the "Active" motifs. The Azofeifa paper[15] examined the most common distances between the center and the position of max motif hits, finding that most sites were within 150 base pairs and no maxima exceeded 500 base pairs. For this reason, we use the edges (1000 to 1500 bp on both the positive and negative sides) to estimate the background and normalize accordingly (Figure A.4b). The normalized metaplot further accentuates the centered "active" peak but highlights the noisy nature of the data.



(a) Raw metaplot (b) Normalized metaplot

Figure A.4: **Metaplots of data set** Overlapping of all the 641 TF data points in a single cell sample. **(a)** Overlay of 641 TF MD for specific cell sample **(b)** Normalized metaplot plot of 641 TF for specific cell sample

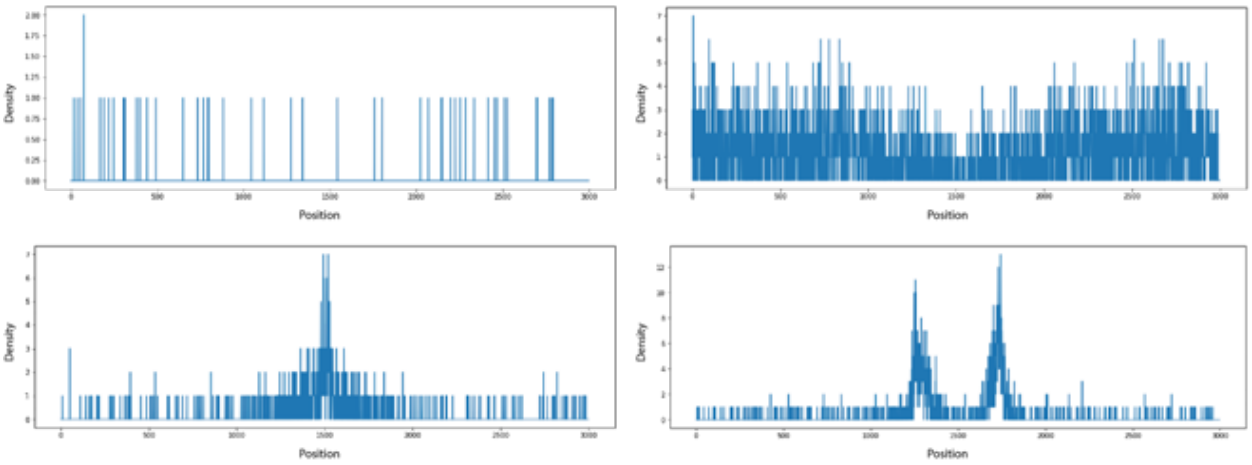
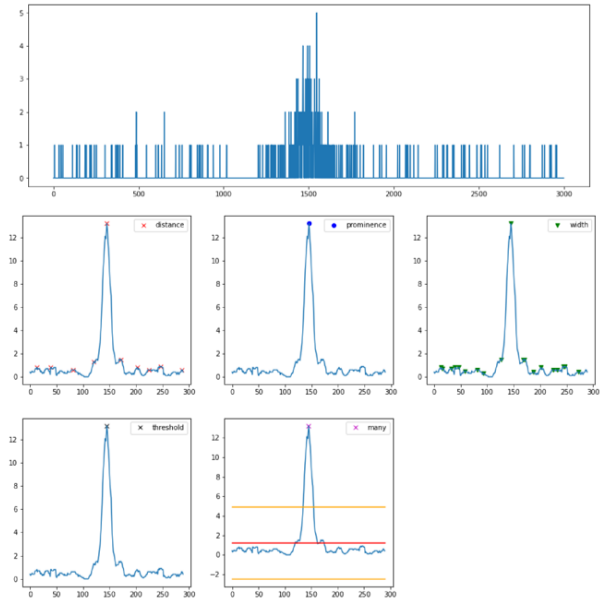


Figure A.5: **Identifying Patterns in Sample Dataset.** Sample of the four main patterns found in dataset; Inactive (top left), Valley (top right), Single Peak (bottom left), and Offset Peaks (bottom right).

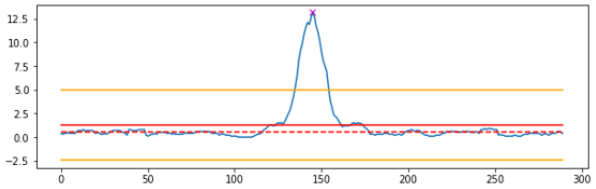
A.2.2 Classifying motif distributions patterns

Next, we return to the individual motif displacement distributions per sample. After visual analysis of the data, I decided to base our initial clustering on a peak detection scheme. Peaks and valleys represent some of the most significant pattern attributes in the data. A small sample includes single peaks, multiple peaks, valleys, and unclassifiable noise (Figure A.5). By programmatically identifying these peaks and valleys, I can quickly and easily apply labels to a training dataset. This data can then be fed into a learning model.

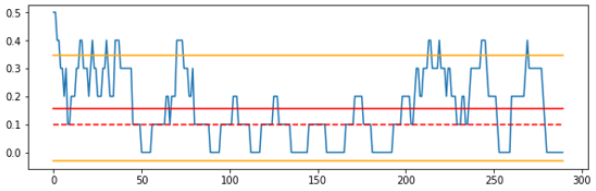
I tested several peak detection and noise filtering algorithms, and found SciPy find_peaks[172] to be well suited for the task. The SciPy find_peaks library uses multiple cooperative metrics to detect peaks inside a signal while filtering background noise. The function takes an array and finds all the local maxima by comparing the neighboring values. A subset of these peaks can be selected by specifying conditions that account for neighboring values with properties that account for the entire data block as an aggregate. I use a combination of peak height, the distance between peaks, peak prominence, and peak width to identify the most relevant peaks (Figure A.6a). In the TFPeakDetect algorithm, I defined the find_peaks' parameters prominence=4, width=3, and distance=20.



(a) **SciPy find_peaks peak detection metrics**



(b) **Smoothing data to detect relevant peaks**



(c) **Noise exclusion example**

Figure A.6: **TFPeakDetect algorithm detects signal enrichment.** The SciPy find_peak library has different optional parameters that can be used to identify relevant peaks A.6a Example of the parameters in SciPy find_peak library that were used to identify relevant peaks A.6b Pre-processing data with sliding window smooths peak width. A single peak was detected in this dataset with a standard deviation of 3.817. A.6c Implementing a sliding window excludes noise in data to optimize peak (signal) detection.

However, even this tool is not powerful enough to filter through all of the noise on its own. Before feeding our data into SciPy find_peaks, I first preprocessed the data using a simple 2-phase

filter. The first phase is a windowed sum function, it splits the data into blocks of 10 units each and sums the elements in the blocks. This reduces the total number of data points in our processed sample and accentuates peaks. Synthetically increasing the size of the peaks helped SciPy find_peaks work most effectively. From there, I used a simple sliding window mean function to help smooth anomalies and increase the width of real peaks. The smoothing made it easier to smooth out the false positive peaks and improve the detection rate of real peaks. The preprocessing is necessary to clean the data to allow me to identify data sets with peaks while ignoring those that only present noise (Figures A.6b, A.6c).

Another pattern that I observed in our data was an inverse peak at the RNA polymerase II initiation (Figure A.5. To identify this pattern termed valley, I had to add an extra step to the processing chain. As the name implies, find_peaks only identifies peaks in datasets. I needed to provide a dynamic method to invert the dataset before peak detection, that way the valleys appear as peaks when evaluated by find_peaks. If the extrema solver detects a sustained region of data points that are all under a specific threshold, then it will mirror the dataset over the x-axis before solving for peaks. With this, I am able to detect both peaks and valleys with a high enough degree of relevance to use this information as labels for our training data (Figure A.7).

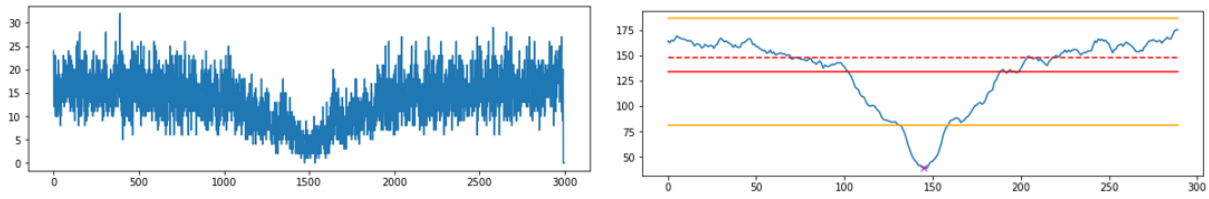
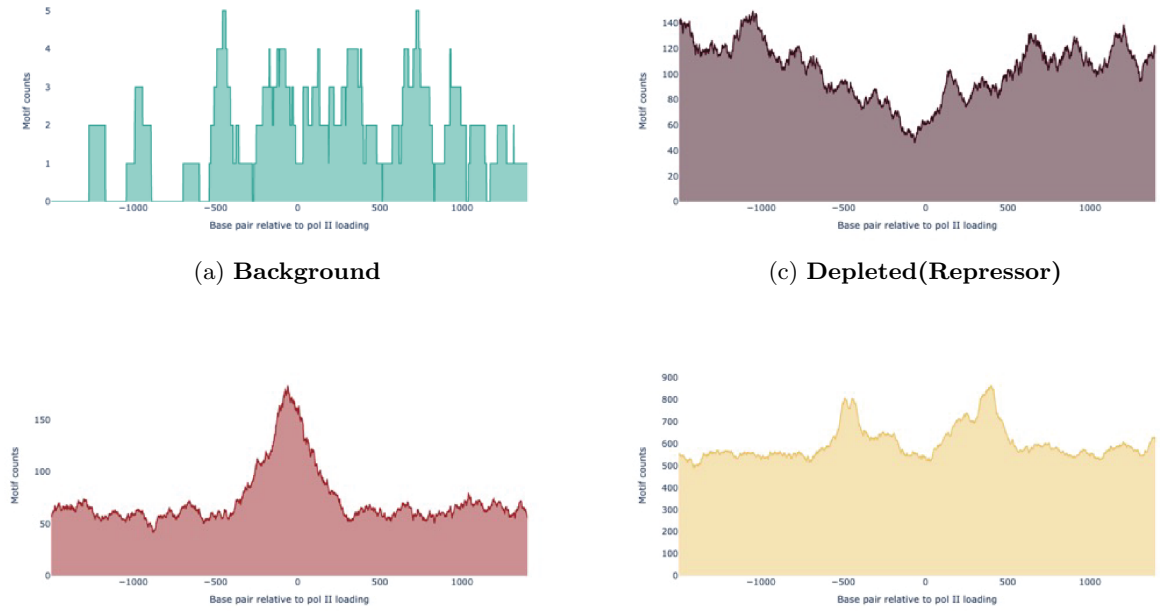


Figure A.7: **Algorithm Detects Valley Pattern.** Inverting the dataset before peak detection allows inverted peaks to be detected as signals. The inverted peaks are termed "valleys" to distinguish them from "peaks".

Using this algorithm, I was able to classify each motif displacement distribution by the number of peaks and valleys therein. I then clustered the motif displacement distributions based on these summaries and identified four prominent patterns (Figure A.8); background (inactive), depleted, centered (active), and offset. These patterns were present in every dataset regardless of cell type



(a) **Background**

(c) **Depleted(Repressor)**

(b) **Centered(Active)**

(d) **Offset**

Figure A.8: **Motif Enrichment Classifier Patterns in A Single Dataset.** The pre-processed data for a single dataset was run through the ME Classifier and four distinct patterns correlated to TF activity were found. The four patterns are A.8a Background, A.8b Active, A.8c Depleted patterns, also termed as Repressor, and A.8d Offset.

(Figure A.9).

A.2.3 Validating Patterns Based on Known TF Activity

The first three patterns correspond to known TF activities: the background (Figure A.8a) implies the TF is off, the depleted pattern is seen with repressors[95] (Figure A.8c), and the centered pattern is seen with activators[15] (Figure A.8b). To validate the patterns, I sought to confirm TFs that are active or inactive in the appropriate cell types. FoxD1 and FoxO6 are examples of TF that have active or inactive activity in specific cell types. The FoxD TF subfamily is found to accelerate induced pluripotent stem cells (iPSCs) generation from mouse fibroblasts as early as day 4, while FoxO subfamily impede this process[63] (Figure A.10).

To further validate that the Motif Enrichment Classifier algorithm can predict a transcription factors activity across multiple cell types. For this task I considered Nanog, a key TF known to

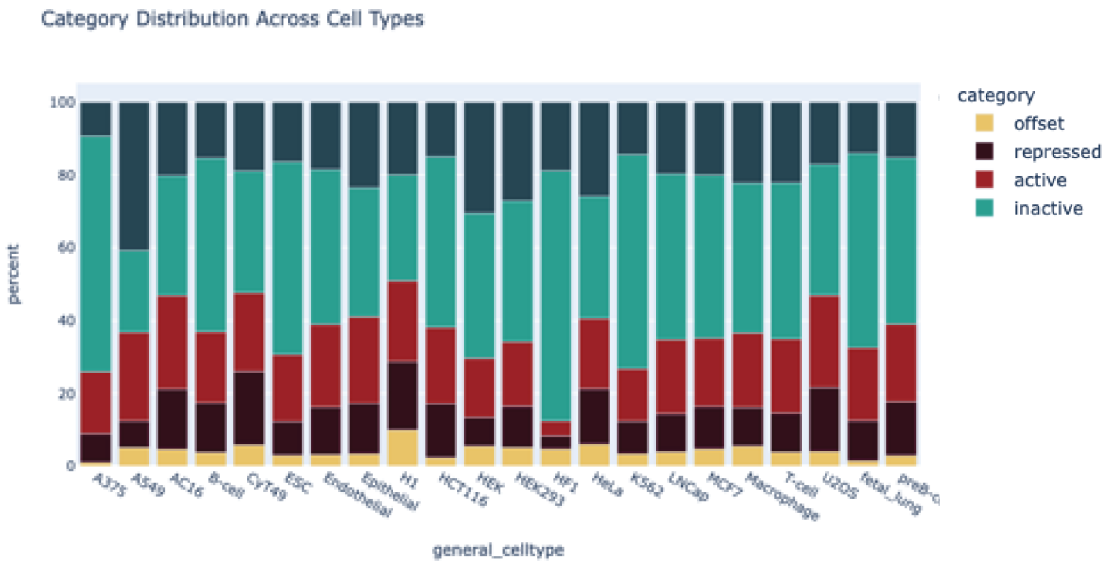
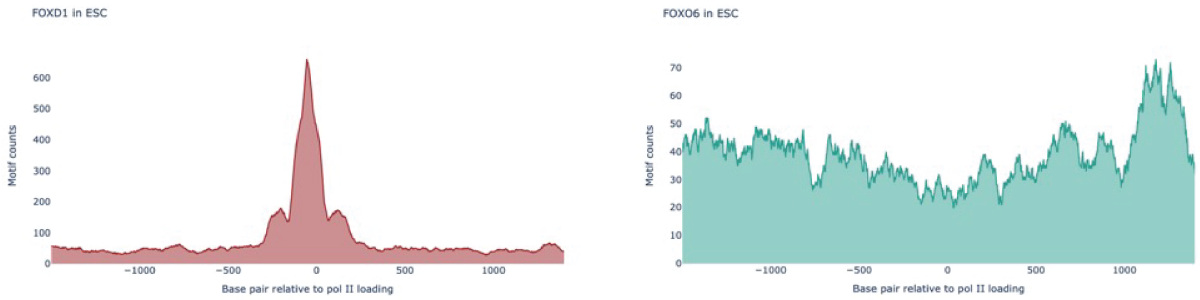


Figure A.9: **Categorical distribution of patterns across cell types.** In any given cell type, the four prominent patterns are observed in similar proportionate across cell types. Approximately 65% background, 20% centered, 10% depleted, and 5% offset. Patterns that appear to be most likely noise and I am not confident to categorize or have an explanation are labeled "uncategorized". n=105 datasets from the year 2018 and earlier for 641 TF Hocomoco motifs.

be active in pluripotent cell types[35]. The Motif Classifier run across 105 datasets for the Nanog motif found that for the majority of cell types, Nanog displayed the Background pattern suggesting it is not active in those cell types. Nanog has the centered pattern in pluripotent cell types such as embryonic stem cells (ESC) suggesting they are active. Strangely the U20S cell type also has a centered pattern, but at closer examination of this cell type did not have transcription over the Nanog gene region. If the gene is not transcribed the TF probably does not function as a repressor because it is not transcribed, hence not present in the cell. Additionally, Nanog had the depleted classification which is unexpected. This classification either meant that Nanog has repressor activity that has not been investigated or poor quality of dataset leads to a false signal. This type of results led to the motivation of the Nascent Database which addresses the quality of data.



(a) **FoxD1 Active Pattern** (b) **FoxO6 Inactive Pattern**

Figure A.10: **TF-Cell Type Dependent Patterns.** Verifying TF patterns in a cell-type dependent manner A.10a FoxD1 has active MD pattern in ESC and is found in studies to accelerate iPSCs generation A.10b FoxO6 has inactive MD pattern in ESC and is found in studies to impede iPSCs generation

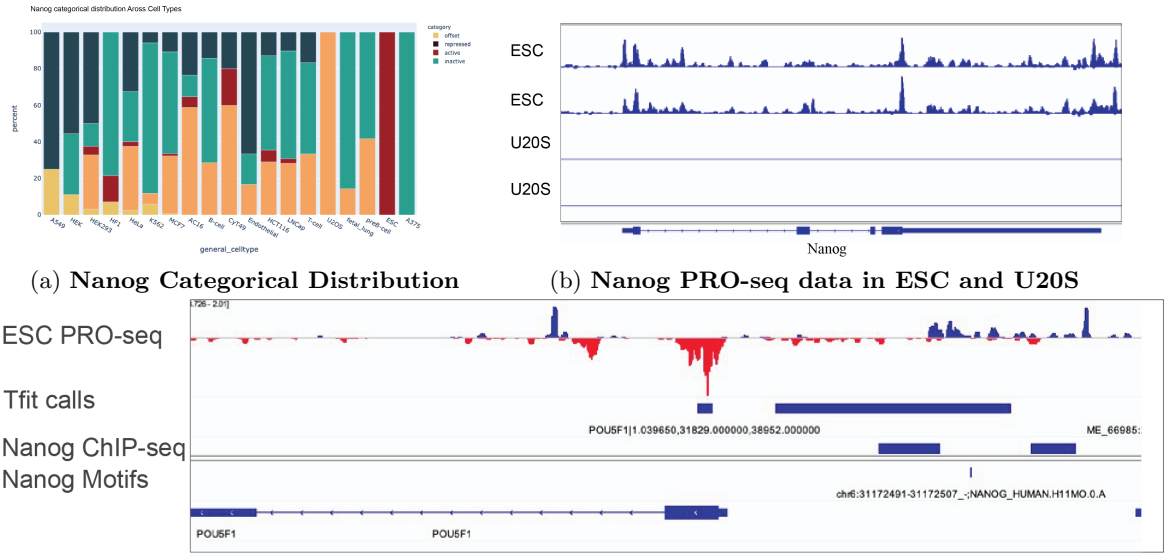
A.2.4 Filtering for Quality Data

The TFPeakDetect initially relied on data from bidirectional calls and motif scans sourced from the Azofeifa paper[15]. The 2018 Tfit was specifically designed to model RNA polymerase II loading. While it excels at pinpointing the center of bidirectionals, it's less adept at identifying enhancer peaks when compared to other peak identification tools[178]. Furthermore, the 2018 dataset did not adequately address data quality or the accuracy of Tfit calls. This oversight in filtering out lower-quality datasets led to inaccuracies, including false peak calls and subsequent misclassification.

A.2.4.1 Nascent Repository Has Consistent Data Quality

To enhance quality control, the Dowell lab built a curated Nascent Repository (available at `http://nascent-dev.int.colorado.edu`). This compared various nascent transcription assay protocols across cell types, consolidating data from approximately 290 studies. They meticulously curated nascent RNA sequencing experiments from the SRAs of over 2800 experiments. This rigorous approach ensured streamlined data processing, uniform transcription level summaries, and consistent quality control. A quality metric was devised, categorizing datasets into tiers based on

their quality. To address the 2018 Tfit susceptibility of multi-mapping of reads, Dr. Robin Dowell made revisions to Tfit in 2022 to tackle these issues, such as refining the identification of the RNA polymerase loading site and enhancing the sensitivity to repeated regions. For consistency, every dataset in the repository was reprocessed using the 2022 Tfit version.



(a) **Nanog Categorical Distribution**

(b) **Nanog PRO-seq data in ESC and U20S**

(c) **Tracks overlapping ESC PRO-seq data with Tfit and Nanog motif.**

Figure A.11: **Nanog Categorial Distribution Across Cell Types.** A.11a Nanog Active Pattern found in ES cells A.11b Motif Classifier confirms that Nanog is active in pluripotent cell types. A.11c Track with PRO-seq in ESC, Tfit calls, Nanog ChIP-seq, and Nanog motifs.

Working with the improved dataset, I focused on Tier 1 and 2 datasets, which had the highest quality scores, and limited my analysis to baseline datasets to avoid external perturbation influences. Subsequent motif scans with the updated Tfit bidirectional calls were conducted on 1200 human TF motifs. Although the cleaner data and the revised 2022 Tfit bidirectional calls were seemingly more precise at calling eRNAs, the reduced motif hits made it challenging for the TFPeakDetect algorithm to discern true peaks in low-transcription scenarios. For TFs with distinct activities like Nanog, the TFPeakDetect algorithm called clear peaks allows the Motif Enrichment classifier to categorize activity more effectively (Figure A.12). In contrast, for TFs with less discernible activities, like RUNX1 (Figure A.15), or universally observed patterns with unclear functions, such as offset, questions arose regarding Tfit's capability in accurately identifying less pronounced bidirectionals

and whether these observed inconsistencies are biologically genuine or data artifacts.

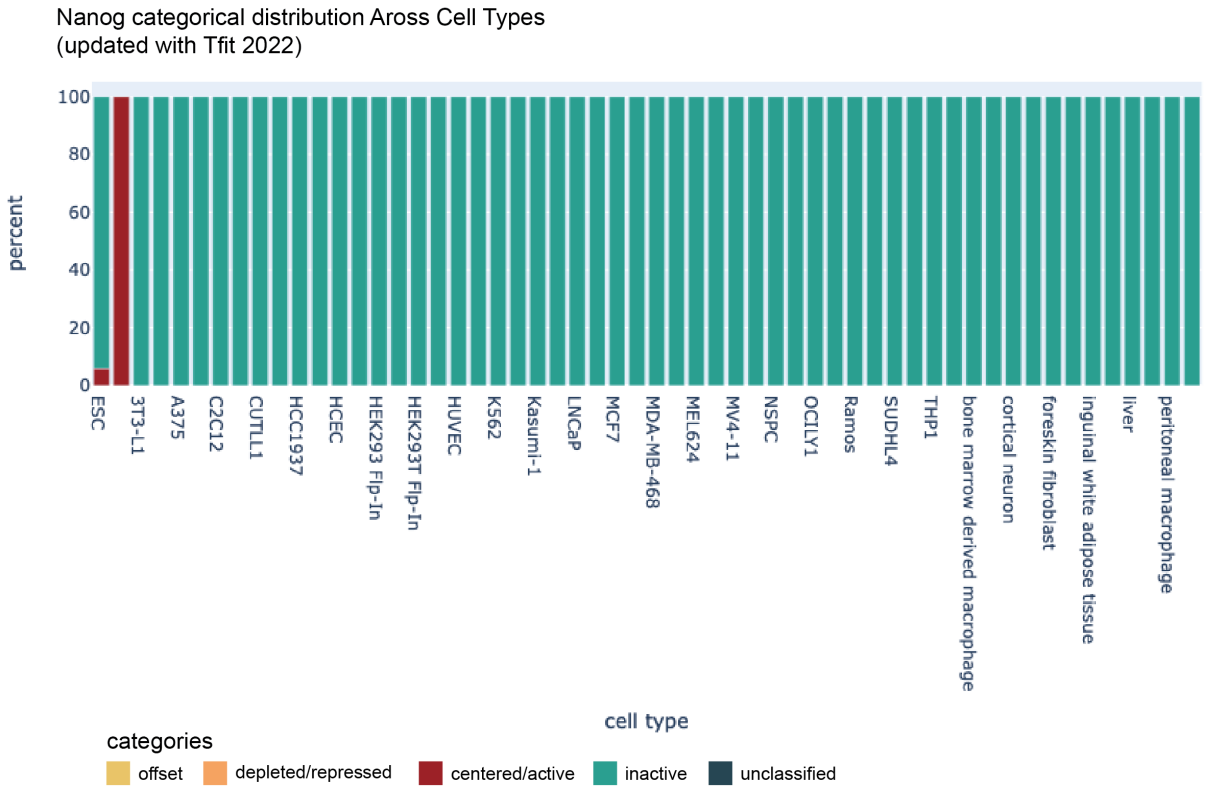
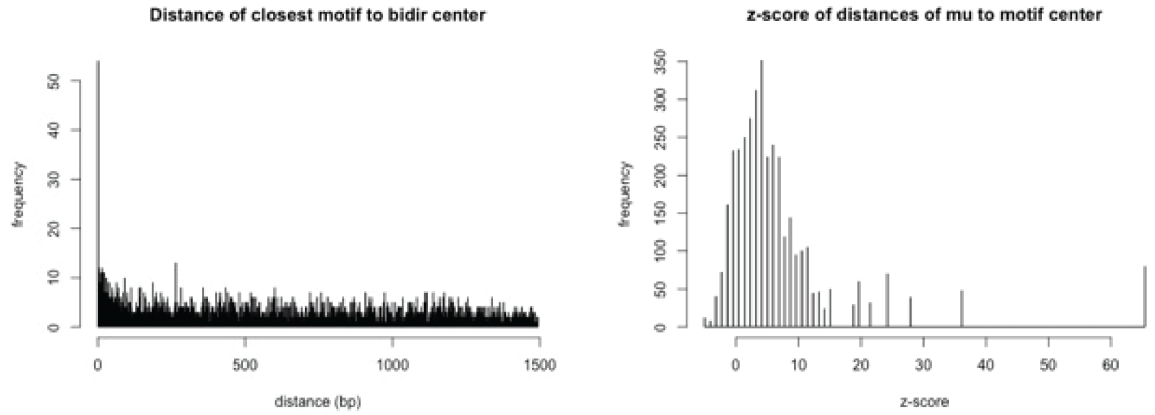


Figure A.12: **Nanog Categorical Distribution With Filter Quality Data.** TFPeakDetect and ME Classifier on filtered 2022 data with Nanog TF motifs fit more with our expectations. Nanog was found inactive in the majority of the cell types except ESC in which it is found active.

A.2.4.2 Normalizing Motif Scans to Enrich for Signal

In an effort to enhance the peak-finding algorithm's precision in identifying peaks in a dataset, I focused on amplifying the signal from the peak relative to the expected background distribution of motifs. The motif scan identifies motifs distributed within a ± 1500 range of the RNA polymerase II initiation site. Given that RNA polymerase II does not initially differentiate between the two strands of DNA during transcription (until it detects an open reading frame on the sense strand), I anticipate equal motif occurrences both upstream and downstream of the RNAP II initiation site. To handle datasets with small sample sizes, I treated distances as absolute values from the RNAP II initiation point, making -1500bp and 1500bp equivalent (Figure A.13a). Drawing inspiration from

Storey's FDR correction method[160], I ranked motifs based on their count frequencies, estimated the background using the interquartile range (IQR), and employed a z-score to assess the signal strength. I found this approach reduced background noise and enhanced the detection of the centered signal (Figure A.13b), facilitating its interpretation. However, for certain patterns, such as the offset, the technique required further refinement to accurately position the motifs. The primary goal was to address patterns that are inherently challenging to interpret, but a complete solution remains elusive.



(a) **Active Pattern** (b) **Signal Enrichment of Active**

Figure A.13: **Signal Enrichment Method.** RUNX1 has an Active pattern in Jurkat T cells. A.13a Signal enrichment of of RUNX1 Active pattern A.13b After normalization of Active pattern has wider peak

A.3 Exploring Offset Pattern

A.3.1 Background

RUNX1 has all three motif distribution pattern that is found in a cell-type dependent manner. RUNX1 drives transcription in T cells and B cells and has the "Active MD" confirming that the TF is functional in lymphocytes. In cell types where the RUNX1 is not known to have transcription activity, it has the "Inactive MD". The last motif distribution pattern, "Offset MD" was observed in multipotent cells such as K562 and pluripotent cells such as ESC. The role of this pattern is

Gene Ontology terms	p-value	GO category
chromatin binding	1.08E-02	Molecular function
RNAP II activating transcription factor binding	2.34E-02	Molecular function
DNA-binding transcription activator activity	3.33E-08	Molecular function
RNAP II proximal promoter sequence-specific DNA binding	2.20E-10	Molecular function
cell fate commitment	5.82E-30	Biological function
positive regulation of cell differentiation	4.62E-03	Biological function
positive regulation of transcription by RNAP II	4.19E-09	Biological function
negative regulation of transcription by RNAP II	1.17E-07	Biological function
positive regulation of multicellular organismal process	1.54E-04	Biological function
animal organ development	1.82E-06	Biological function

Table A.1: **Offset MD Gene Ontology terms. The subset of transcription factors with an Offset MD pattern in any cell type was input into GO to identify shared terms. The GO terms for Offset MD pattern suggest that the TFs function as chromatin modifiers.**

unknown. Using the Motif Classifier, I found the set of TFs that has the offset pattern in any cell type. The set was used as input into the gene ontology (GO) system to identify shared terms to better understand their collective functional roles. The GO terms suggest that this pattern is associated with TFs functioning as a chromatin modifier (Table A.1).

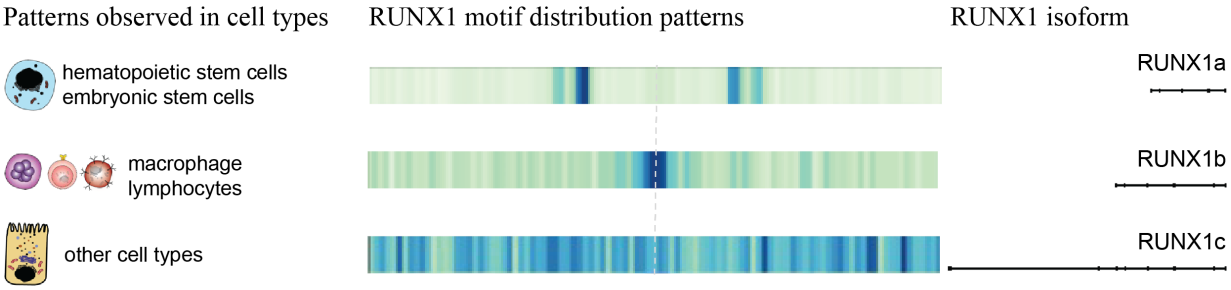


Figure A.14: **RUNX1 Motif Distribution Patterns is Cell-type Dependent.** RUNX1 MD patterns are cell-type specific and correlated to the dominant isoform present in the cell. The "Active MD" is observed in lymphocytes and "Offset MD" is observed in multipotent and pluripotent cell types. The "Inactive MD" is observed in cells where RUNX1 protein is not present.

A.4 Validating the offset: RUNX1 ChIP

To evaluate RUNX1's "Offset" pattern, I wanted to ask if RUNX1 TF was bound at the regions that have the "Offset" patterns. To do that I would use ChIP-seq to enrich for DNA-RUNX1

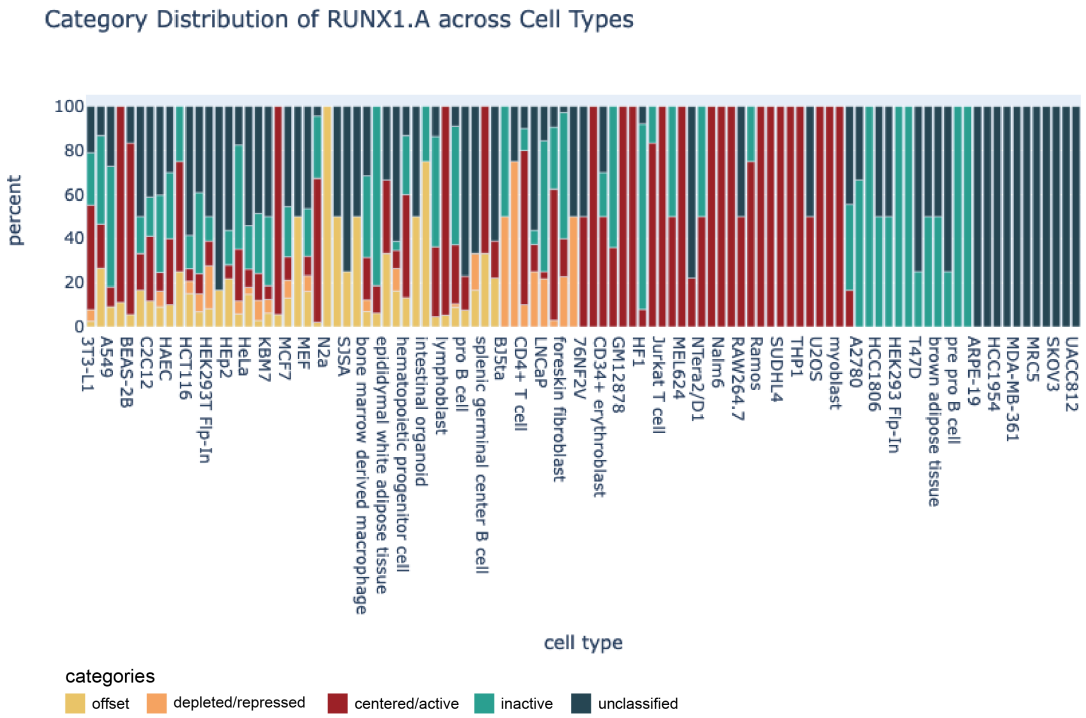


Figure A.15: **RUNX1 Categorical Distribution With Filter Quality Data.** Using data filtered for quality data (Tier 1 and 2) to categorize RUNX1 MD patterns by cell type.

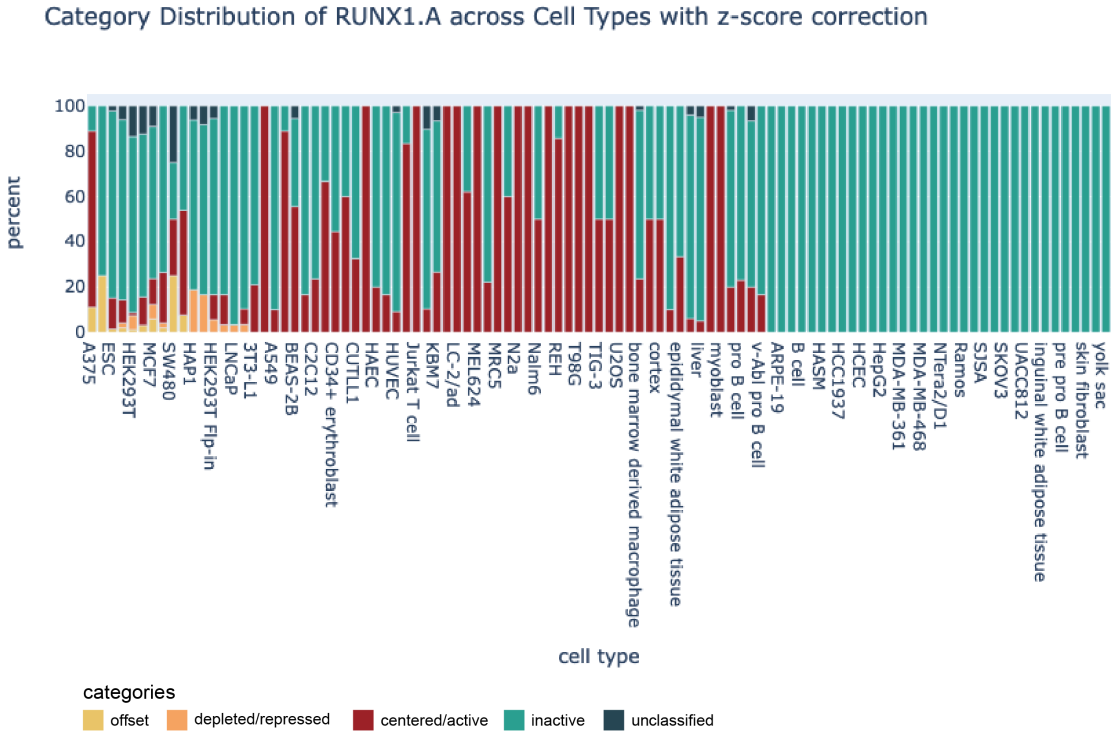


Figure A.16: **RUNX1 Categorical Distribution after normalization.** Using a z-score correction in addition to filtering for high quality data we found that the data fit the expectations more. ESC had the offset and inactive patterns while Jurkat T cells and pro B cells had the active pattern.

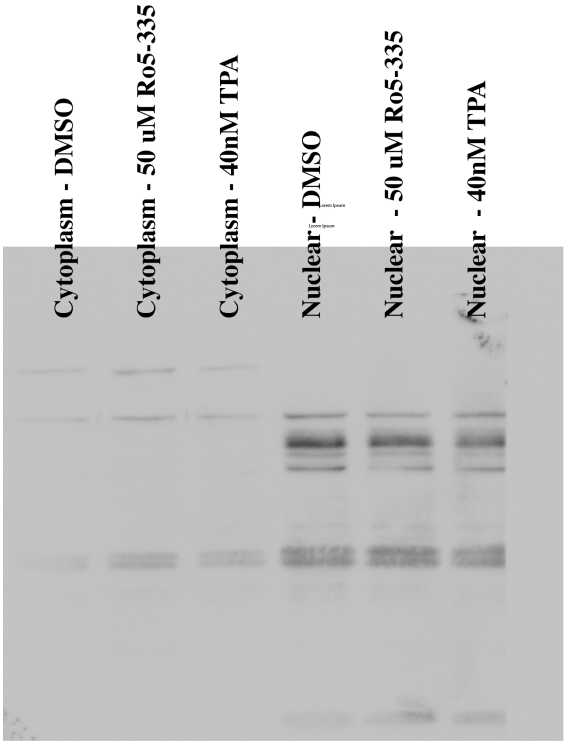


Figure A.17: **RUNX1 Western Blot in K562 cells.** Anti-RUNX1 antibody targeting RUNX1 in K562 cells. 30 µg of protein loaded on 12% gel. Exposure was for 120 seconds. A band was observed in the nuclear extract for RUNX1 at 48 kDa. Activator TPA and repressor Ro5-335 do not appear to change the intensity of the band.

interaction in K562 cells.

To confirm the presence of RUNX1 TF in K562 cells, a western blot was done using an anti-RUNX1 antibody (Abcam Ref. AB23980). Additionally, I tested two compounds that were found to activate (40 nM 12-O-Tetradecanoylphorbol 13-acetate (TPA); Sigma-Aldrich Ref. P1585)[113] or repressed (10 µM Ro5-3355; Tocris Ref. 4694)[53]. Although RUNX1 was found expressed in the nuclear extract, I notice that 30 µg of protein required a long exposure to observe the band on the gelA.17 suggesting that while RUNX1 is present, it is either lowly expressed in K562 or the antibody is not specific for RUNX1.

Since ChIP-seq is an enrichment assay, the quality of the antibody is an important factor. Since antibodies can vary from target to target, I requested a non-commercial antibody from Groner's laboratory[133]. The western blot for RUNX1 with this antibody required longer exposure

to be able to detect the protein. As a comparison of anti-RUNX1 antibody targets for different isoforms and MD patterns, I harvested cells from lymphoblastoid (LCL) and K562. I lost more DNA during the immunoprecipitation in the K562 compared to the LCL sample as measured by Qubit and the tapestation suggested that the DNA samples were of low complexity. One possible explanation would be the sheering of the samples affected the epitope and hence the antibody binding. Another possible optimization would be to add additional bead clean-up to remove adapter dimers that may have interfered with the outcome. Due to the quality issue, I did not move forward with the ChIP-seq assays since it is highly dependent on the quality of the pulldown.

A.5 Github Repository

The Python and R scripts developed for this body of work can be found under the Github Repository `https://github.com/jessicatwes/ME-classifier`.

A.6 Conclusion and Future Work

Transcription factors are vital for cell function and understanding them can help us understand how genes are regulated. The goal of this project was to be able to use high-throughout assays and from a single experiment be able to infer which TFs are functional. In addition to just inferring if a TF is on or off, I seek to add additional knowledge by investing in other TF motif distribution patterns. TFs can provide us insight into which pathways are crucial for the cells.

Appendix B

Development of Regulatory Activity Decoder Construct (RAD Construct) to evaluate enhancer activity

B.1 Background

While techniques like ChIP-seq[40] and open chromatin assays[29] have advanced our study of transcription factor (TF), it remains challenging to ascertain when TF exist as proteins in the cell and are actively influencing transcription[135]. The regulation of TFs can occur at various stages, including transcription, translation, localization, or through post-translation modifications (Figure 1.4). Hence, understanding when a specific TF is present and active is pivotal for discerning regulation, building computational models, and designing appropriate experiments.

In Appendix A, we detail our lab's method of using enhancer-associated RNA (eRNA) profiling, also called Motif Distribution (MD) patterns, to predict the presence and activity of a TF in altering transcription. Our findings indicate that when a TF is active, short bidirectional RNAs can be observed near the TF motif[15]. An "Active" TF has a centered MD pattern, suggesting that the TF motif co-localizes with sites of RNA polymerase II (RNAP II) more so than expected by chance. We can quantify this association by calculating the Motif Distribution (MD) score, which considers the number of motifs within a specified window (See Appendix A for details on the MD score).

It is important to note that just because a TF protein is physically present does not mean it actively participates in transcription regulation. To address this, I aimed to develop a reporter assay that can assess the activity of individual TF in a quantifiable manner. Not only could this tool help us validate our enhancer-associated RNA (eRNA) profiling predictions both in terms of

quality (i.e., the accuracy of active predictions) and quantity (e.g., assessing if a higher MD-score[15] implies that a TF is more active in one cell type than another), but it could, in turn, help adjust the TFPeakDetect and ME classifier algorithm (see Appendix A for details on the algorithms) to maximize the sensitivity of our peak calls.

I anticipated that the reporter could be used to inform me if the MD-score metric is truly a readout on TF activity. For example, if we wanted to ask what context surrounding a TF motif is sufficient to elicit an appropriate TF response, we could place the TF motif in multiple reporter enhancer segments to quantify its activity based on the responding enhancer. Alternatively, if we want to ask if a TF targets a specific enhancer, we could use RNA interference (RNAi) to target the TF and ask if there is an associated loss of enhancer activity. Moreover, using a suitably designed reporter, we can examine various patterns seen in the ME Classifier classification (as elaborated in Appendix A).

Runt-related transcription factor 1 (RUNX1), a TF essential for blood development, exhibits not only the "Active" and "Inactive" MD patterns but also a third "Offset" pattern depending on the cell type, as depicted in Figure A.14. Through the reporter, I suggest that we analyze RUNX1 activity based on its MD pattern in relation to the cell type. Additionally, by utilizing a RUNX1 knockdown cell line, we can delve into the role RUNX1 plays in enhancer and transcription regulation. By knocking down RUNX1, my aim is to study its influence on enhancer activity, the transcription of bidirectional genes, or the transcription of specific target genes.

B.2 Design of the Regulatory Activity Decoder (RAD) Construct

The Regulatory Activity Decoder (RAD) is a dual reporter construct designed to enable the accurate prediction of enhancer activity. It makes use of the unique property of two distinct fluorescent reporters. The mScarlet fluorescent protein (RFP) is ubiquitously expressed and serves as a normalization control, ensuring uniformity in data interpretation across different cell types and plasmid copies. On the other hand, the green fluorescent protein (GFP) provides a dynamic readout of enhancer activity, offering real-time monitoring of enhancer activity in living cells (Figure B.1b,

B.1c).

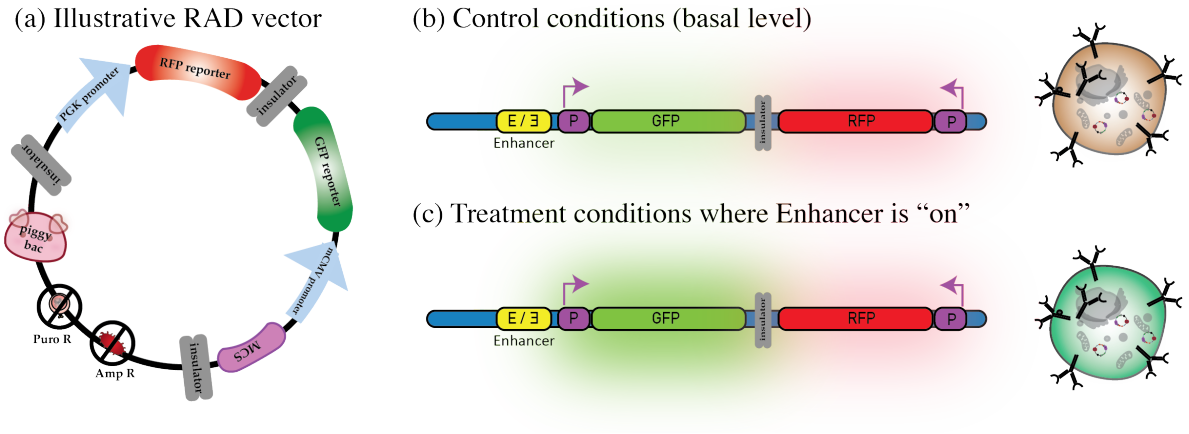


Figure B.1: **Regulatory Activity Decoder (RAD) Construct and schematic.** B.1a RAD Construct Illustrative. Illustration of the main component of the dual fluorophore reporter construct. B.1b Control conditions (basal level). GFP and RFP are expected to be lowly expressed at basal level with the GFP/RFP ratio is near 1. B.1c Treatment conditions where Enhancer is "on". GFP fluorescent signal should be greater than RFP with GFP/RFP ratio > 1.

The primary objective in designing this plasmid is to enable the accurate prediction of enhancer activity. To achieve this goal, the plasmid construct incorporates a multiple cloning site (MCS), where the enhancer of interest can be cloned. This MCS serves as a versatile and flexible platform for the insertion of enhancer sequences, allowing for their subsequent analysis and evaluation of their regulatory potential. The design ensures that the enhancer can be easily integrated into the plasmid, facilitating comprehensive assessments of its activity and function. The plasmid contained several other design features illustrated in Figure B.1a. Two antibiotic-resistance genes were used to screen for positively transformed bacteria (ampicillin) and to eventually select for positively transfected mammalian host cells (puromycin). Finally, we include a PiggyBac transposon, which is a genetic element that enables random integration of the reporter cassette into the host genome.

B.2.1 Main Components of RAD construct

B.2.1.1 Flurorphore Reporters

Briefly, the fluorophore genes are downstream of a minimal promoter; murine cytomegalovirus immediate-early promoter (mCMV) promoter[142] is upstream of the MCS and eGFP fluorophore and mouse phosphoglycerate kinase 1 promoter (PGK) promoter[142] upstream of mScarlet fluorophore. The mCMV promoter was selected because its expression is constant across most cells used in cell culture at a medium expression level. The role of this reporter is to measure changes in transcription activation, hence we needed a reporter gene that has a low basal level. PGK was selected because it is a consistently weak promoter, thus the rate of transcription of the mScarlet gene is low and a good option for internal control.

B.2.1.2 Multiple Cloning System (MCS)

The MCS segment of the construct contains multiple recognition sites for restrictive enzymes, facilitation the insertion of target Enhancer region for study. This MCS has thirteen restriction enzyme (RE) sites that include both sticky end and blunt ends (see Figure B.2). Sticky ends have overhanging bases resulting from enzyme's staggered cut. These overhangs can base-pair with complementary DNA sequences, ensuring more precise ligation. In contrast, blunt ends are straight cuts with no overhangs making them less efficient for cloning because of the lack of specific base-pairing. However, blunt ends can ligate to any other blunt-ended vector, thus are more versatile. The design of the MCS provides a range of cloning options to allow for customization of cloning approaches.

B.2.1.3 Insulator

Insulators are DNA sequences that acts as boundaries and were used to separate the two reporter domains. This ensure that the genetic regulatory elements influence the activity of only their target genes and do not 'cross-talk' between genes located in adjacent domains. Insulators

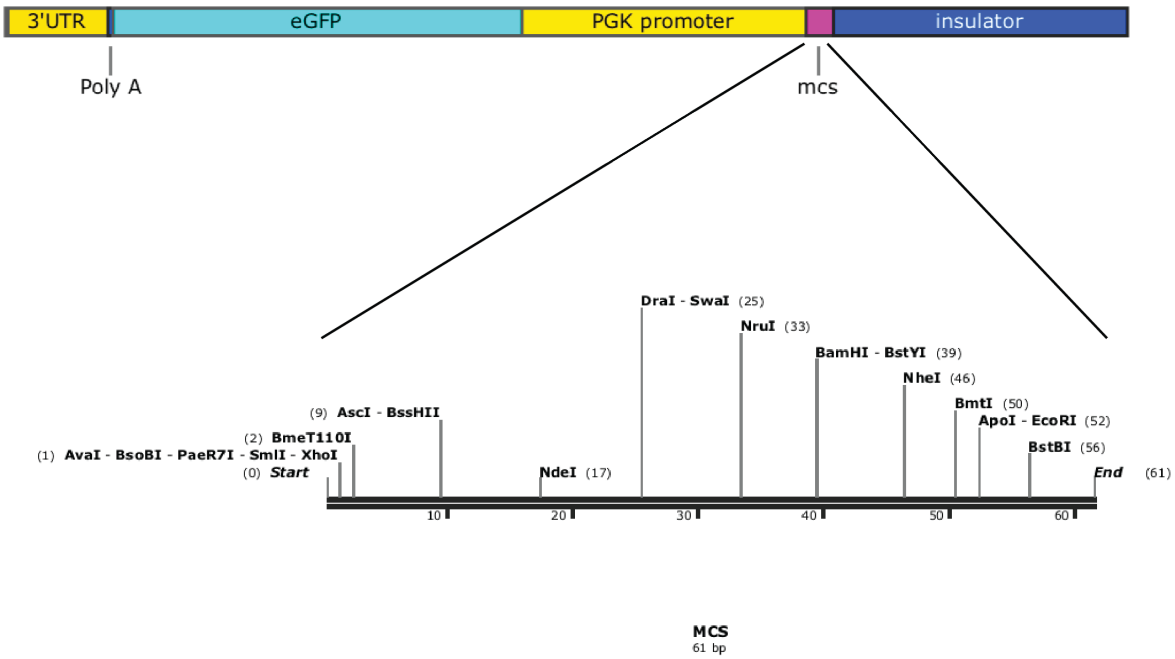


Figure B.2: **Multiple Cloning System (MCS).** MCS has thirteen restriction enzyme sites that include both sticky and blunt ends for cloning target enhancer region into the construct.

are elements that can block enhancers from activating RNA pol II transcribed promoters. Transfer DNAs (tDNAs) are repetitive sequences derived from the integration of mobile genetic elements, specifically from transfer RNA genes. These sequences are dispersed throughout the human genome and some tDNA have been observed to possess insulator-like properties. Three tDNA clusters were used for the insulator based on their ability to bind specific transcription factor TFIIIC and has binding sites for CCCTC-binding factor (CTCF) [143]. The tDNA clusters used were TMEM tRNA cluster (hg38 chr17:8030134-8031757), Chr19tDNA cluster (hg38 chr19:1334277-1334919) and Aloxe3 tRNA cluster (hg38 chr17:8064306-8067241).

B.2.1.4 piggyBac Transposon

The piggyBac transposon is a genetic element that can efficiently transpose, or 'jump', between the vector and insert itself into mammalian cell chromosomes. This transposition mechanism is

facilitated by the recognition of specific inverted terminal repeat (ITR) sequences located on both ends of the transposon vector by the piggyBac transposase enzyme. Once recognized, the content between the ITR is moved from its original position and integrated into the preferred TTAA target sequences on the host cell's chromosome. Additionally, the advantage is that the integration is reversible, meaning you can express the piggyBac transposase in cells where the transposon has integrated and excise the element from the host genome. The reversible integration is of value when validating the phenotypic effects of the transgene.

B.2.1.5 Antibiotic Resistance

The vector has two antibiotic resistance genes. The puromycin resistance is used to select transducing cells. The ampicillin resistance gene allows the plasmid to be maintained by ampicillin selection in E. coli competent cells.

B.2.1.6 3' untranslated region

A 3' untranslated region(3'-UTR) has regulatory mechanisms that include modulation of mRNA stability and degradation, translation efficiency, and transport of mRNA out of the nucleus. We used the 3' UTR sequence from Cyclin-dependent kinase 5 regulatory subunit (CDK5R1) that showed low gene expression activity[126].

B.2.1.7 Kozak sequence

Facilitates translation initiation of ATG start codon downstream of the Kozak sequence.

B.2.1.8 Polyadenylation signal

Polyadenylation signal (rGB pA, BGH pA, and poly A signal) placed at the end of the gene and MCS allows for transcription termination and polyadenylation of mRNA transcribed by RNA polymerase II.

B.2.2 VectorBuilder Summary

We outsourced our custom designed plasmid vector to VectorBuilder. VectorBuilder specializes in providing vectors for viral and non-viral gene delivery. I selected the Mammalian Gene Expression PiggyBac Vector as the backbone. From VectorBuilder's database, I selected the sequences for the fluorophore, antibiotic resistance, and promoter genes. The insulators and MCS sequences were designed in-house and provided to VectorBuilder. VectorBuilder provided me with the full sequence for the construct. We confirmed the constructed MCS sequence using Sanger sequencing with our primer RAD_MCS_R (Sequence 5'-GCTAAGGAGAACGGACCTCAG-3'). This primer will be used downstream to verify the correct Enhancer sequence ligated into the MCS. For more detail on the construct, see the Vector Summary by the VectorBuilder B.6.

B.3 Validating the RAD construct with p53 enhancer regions

To ensure the functionality of the RAD construct, we created plasmids incorporating p53 target enhancers that, as previously identified by our lab, showed transcriptional changes upon Nutlin-3 treatment [6]. The 2014 study by Allen et al. leveraged nascent transcription to observe immediate cellular transcriptional responses post-Nutlin-3 treatment. The tumor protein p53 functions as a transcription factor and tumor suppressor, inhibiting cell growth and slowing cell division. Nutlin-3 activates p53 by binding to MDM2, an oncoprotein responsible for repressing p53 (Figure B.3a). This study confirmed two regions, DRAM1 (Figures B.4, B.5) and CDKN1A, which had eRNAs arise in proximity to the p53 motif after Nutlin-3 exposure, suggesting that they are direct p53 targets [6].

B.3.1 Experimental Methods

B.3.1.1 Cloning the p53 Enhancer Region into the RAD construct

To assemble the RAD plasmid with an embedded enhancer sequence, I used PCR cloning to amplify the DNA fragment that will ligate into the MCS region of the vector. The target DNA

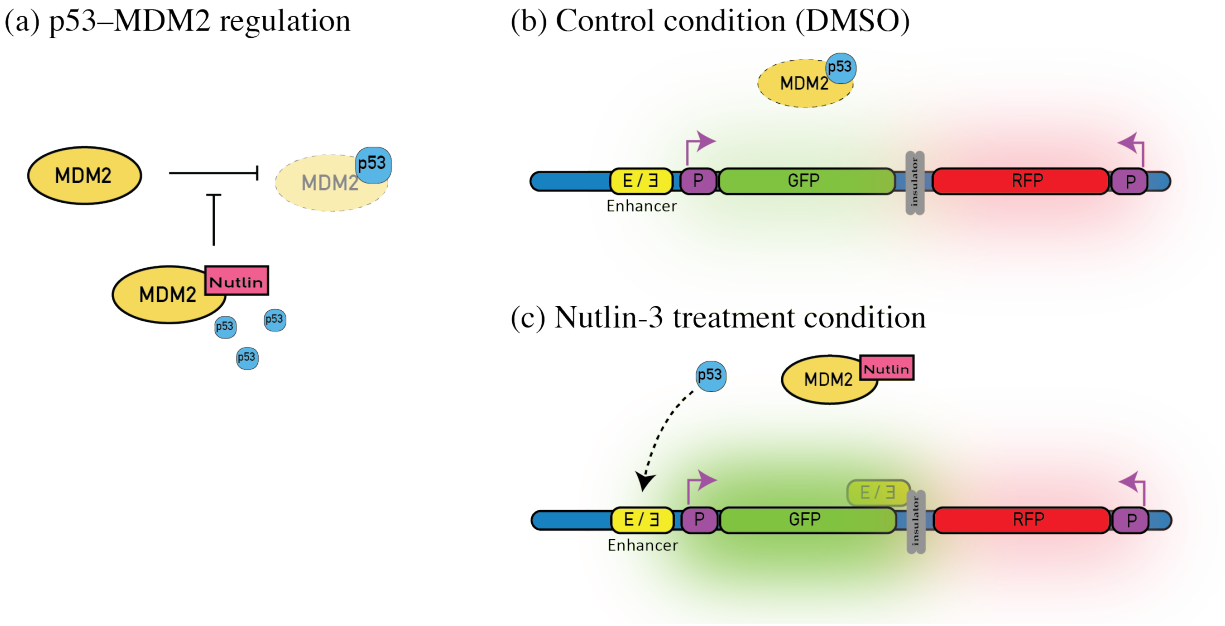


Figure B.3: **Schematic of DRAM1 RAD plasmid with Nutlin-3 treatment.** **(a)** MDM2 regulatory role on p53. Nutlin-3 activates p53 by binding to MDM2, **(b)** Control condition p53 is bound to MDM2. **(c)** Nutlin-3 treatment condition activates p53.

template was isolated from purified human derived cell lines (HEK293 and HeLa) genomic DNA using OneTaq DNA Polymerase (NEB Ref. M0480), 200 nM Forward and Reverse PCR primers, and 100 ng genomic DNA in final volume of 25 µL/reaction using an Annealing gradient $57 - 62°C$. Multiple primers were designed for the DRAM1 region because this region was found to be very repetitive and not well conserved when the sequence was input into Basic Local Alignment Search Tool (BLAST). The DRAM1 sequence mapped to 200+ sites in the human genome, thus multiple primers were designed before a successful one was able to isolate the region. The PCR primers were designed to have RE overhang and As. The sequence for the final PCR primers that were used to extract the template is available in Table B.1. The PCR amplification mixture was run on a 1.5% agarose gel and the extracted DNA band of expected size was purified using QIAquick Gel Extraction (Qiagen Ref. 28704).

The primers included restriction enzyme sequences to allow for ligation of the DNA amplicon into the MCS of the RAD construct. Two separate double digestion was done on the PCR amplicon

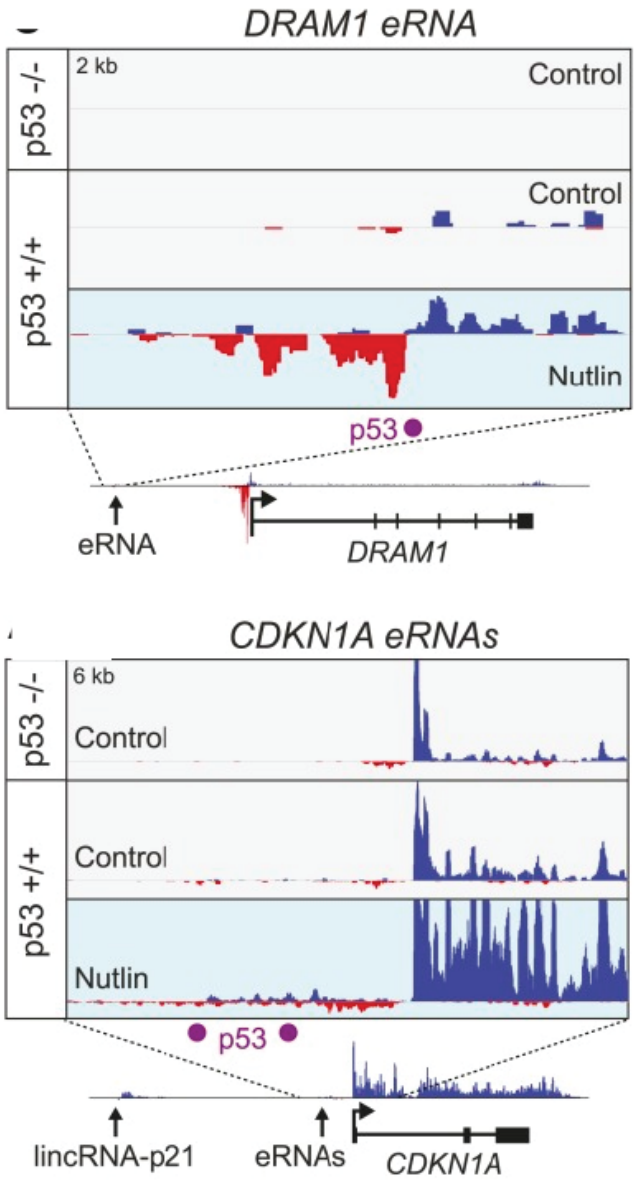


Figure B.4: **p53 targets from Allen 2014.** Allen et al., 2014 paper show that GRO-seq analysis reveal DRAM1 and CDKN1A to be immediate p53 targets[6].

and the RAD construct with restrictive enzymes EcoRI (NEB Ref. R0101) and AscI (NEB R0558). A ligation of 7 parts amplicon: 1 part linearized RAD construct was performed overnight at 18°C using Anza T4 DNA ligase (Invitrogen Ref. IVGN2104) and heat-inactivated at 65°C for 10 minutes. The ligated product was transformed into TOP10 competent cells and the transformation mixture was

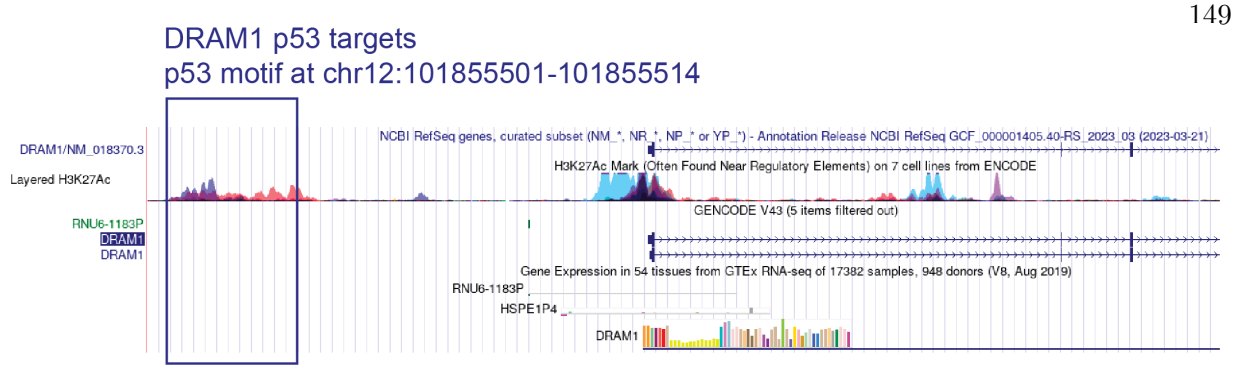


Figure B.5: **DRAM1 Enhancer Region.** UCSC genome browser shows that an enhancer located in proximity to DRAM1 gene has a p53 motif and serves as a p53 target.

grown on Luria-Bertani (LB) broth + 100 µg/mL Ampicillin plates. To determine transformation efficiency, 0.1 µg/mL pUC19 control was transformed. Additional negative controls were linearized RAD with no template and an empty RAD vector.

Although all PCR were successful, only the DRAM1 PCR products were cloned into the backbone. We referred to these plasmid as DRAM1_707 and DRAM1_911 referencing the length of the DRAM1 amplicon (Table B.1). The two DRAM1 plasmids (DRAM1_707 and DRAM1_911) that were constructed were sent out for Sanger sequencing to confirm that the DNA insert sequence was correct. The primer used for sequencing was RAD_MCS_RC (Sequence 5'-CTGAGGTCCGTTCTCCTTAGC-3').

The ligated plasmid was transformed into DH5-α competent cells with 0.1 µg/mL pUC19 control to determine transformation efficiency and plated on LB plates with 100 µg/mL Ampicillin. Additionally, a negative control of linearized plasmid + ligase was used. A plasmid prep was prepared and a plasmid sample was outsourced for Sanger sequencing to validate the sequence of the DNA template.

B.3.1.2 Transfection of RAD plasmids into mammalian cells

The two DRAM1 plasmids, DRAM1_707 and DRAM1_911, were transfected into mammalian cells separately. We proceeded with both plasmids to address the question of whether length of

Primer	Sequence (5'-3')	Tm
DRAM1-707bp-AscI-Fwd	5'-AAA AGG CGC GCC CAT ATA TTT TCT CTC TTC AAC ATA AAC TGG-3'	°C
DRAM1-707bp-EcoRI-Rev	5'-GGC CGG GAA TTC GGA AAA GAA AGA GAA GAA CTA GCT TTA G-3'	°C
DRAM1-911bp-AscI-Fwd	5'-AAA AGG CGC GCC TTT ACA TAT ATT TTC TCT CTT CAA CAT AAA C-3'	°C
DRAM1-911bp-EcoRI-Rev	5'-GGC CGG GAA TTC TAT CTT AGC TAT GTA AAA ACA TGT ACT CTT G-3'	°C
CDKN1A-790bp-AscI-Fwd	5'-AAA AGG CGC GCCC AAT TAC TAA CCA CTT GTC AGA AAC AAT AAA TC-3'	61.5°C
CDKN1A-790bp-EcoRI-Rev	5'-GGC CGG GAA TTC CTG TTC AGA GTA ACA GGC TAA GGT TTA C-3'	61.2°C
CDKN1A-777bp-AscI-Fwd	5'-AAA AGG CGC GCC CTC TGC TCA ATA ATG TTC TAT CTT TGT TCC-3'	60.87°C
CDKN1A-777bp-EcoRI-Rev	5'-GGC CGG GAA TTC TAC TAA GTG TCT AGT ACT ATT CAG TGC TTT-3'	59.91°C

Table B.1: **Primers for p53 enhancer templates.** : Two set of primers for p53 enhancer targets; DRAM1 and CDKN1A. The blue color indicates the EcooRI (G'AATTC) or AscI (GG'CGCGCC) RE site.

the amplicon had an impact on efficiency of enhancer activity. The human colorectal carcinoma (HCT116) cell line was chosen for the transfection because it showed a strong response to Nutlin-3[6]. Additionally, I transfected the plasmids into the Human Embryonic Kidney 293 (HEK293FT) cell line. The cell lines were cultured in complete Dulbecco's Modified Eagle Medium (DMEM) media consisting of DMEM media, 2 mM L-glutamine, and 10% fetal bovine serum (FBS).

Initially, I attempted to introduce the DRAM1 plasmids into the cells using a nucleofector, specifically the Lonza 4D-nucleofector (Lonza Ref. VACA-1003). For the HCT116 cells, I used the program D-032 and for the HEK293FT, program FS-100 was used to transfect 2 µg of plasmid into $1 * 10^6$ cells. Unfortunately, the cells did not survived and died seven hours post-nucleofection. Multiple repeats of the nucleofection yield the same results.

To address this, I transitioned to using Lipofectamine 3000 transfection reagent (Invitrogen Ref. L3000001) for the plasmid transfection, which yielded a better cell survival rate post-transfection.

The mammalian cells were grown to 70-90% confluency prior to the transfection. Following the manufacturer protocol for Lipofectamine 3000 transfection reagent, 2µg of DRAM1 plasmid was transfected into approximately $7 * 10^5$ cells in a 6-well plate. The cells were incubated for 2 days at 37°C, 5% CO_2 in media with 10 µg/mL puromyocin before treatment with Nutlin-3. Two controls were applied; the negative control had only Lipofectamine reagent and the positive control had the empty RAD construct only (no DNA template). The same protocol was applied to HCT116 cells.

B.3.1.3 Nutlin-3a treatment and live imaging

The transfected cells were seeded into 96-well plate at a concentration of $1 * 10^4$ cells/well six hours after transfection. The cells were treated with 10 µM Nutlin-3a[6] or DMSO (control), and moved to the imaging facility for live imaging on the Opera Phenix confocal microscope. The microscope was set to capture an image per well every 60 minutes for 12 cycles on two channels for eGFP and mScarlet flurophores.

B.3.2 Live imaging analysis

The images were imported into ImageJ software. The background signal was captured by selecting 10 small areas with no fluorescence. A selected number of cells were selected per replicate and were tracked across the twelve time points. The cells' measurements were captured for each channel including the area, integrated density and mean gray value. The total cell fluorescence was corrected by subtracting the mean background signal expanded by the area of the selected cell from the integrated density. This correction was done for each cell in both channels and helps compare the fluorescence intensity between cells. To normalize the fluorescent to account for plasmid copy number, the ratio of GFP/RFP per cell was then calculated.

B.3.3 Evaluation of Enhancer activity

To evaluate the function of the RAD construct we selected an enhancer known to respond to Nutlin-3 treatment and is a validated p53 target[6]. We transfected the two DRAM1 plasmids,

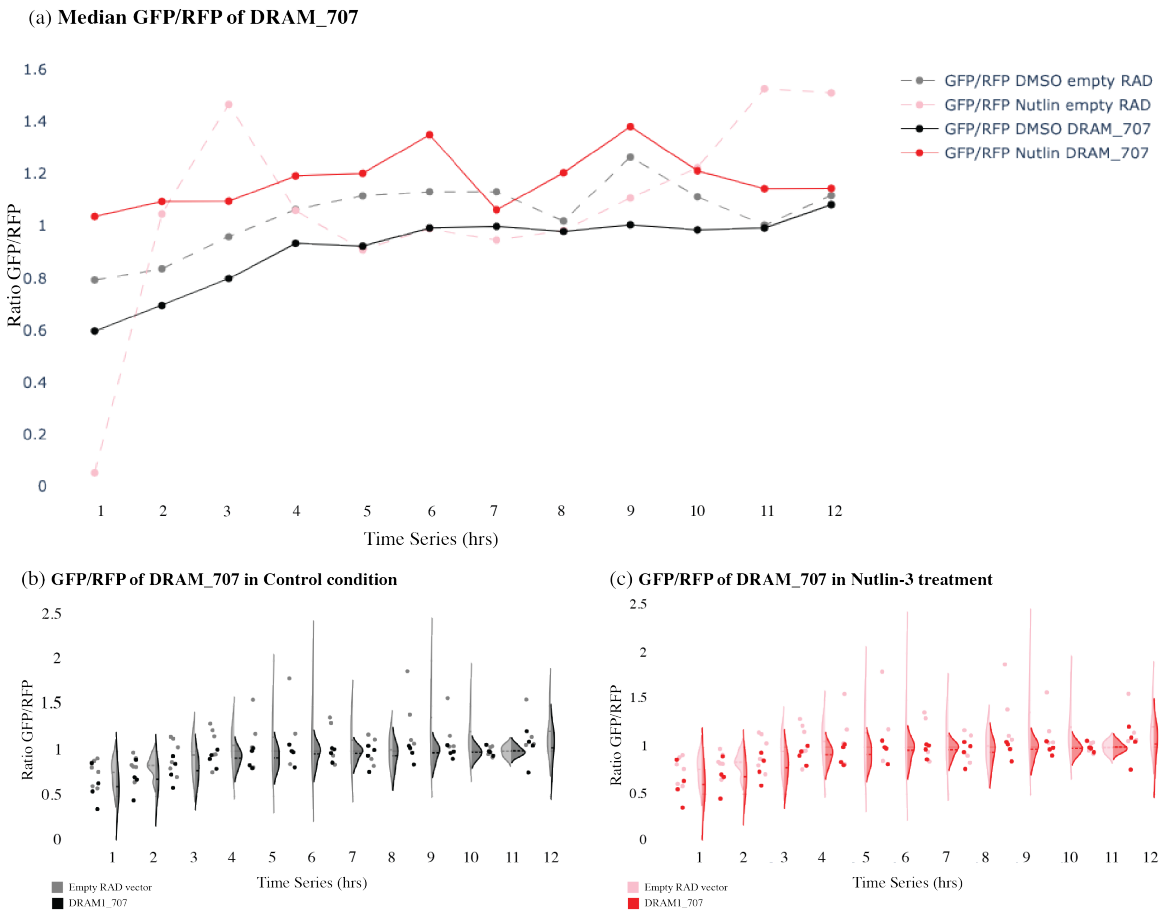


Figure B.6: **Ratio of GRF/RFP for cells with DRAM_707 plasmid.** Plot of the ratio of GRF/RFP to compare the two conditions; Control (DMSO) and Nutlin-3 perturbation. B.6a Over the twelve hours time course, the median GFP/RFP ratio was plotted for both cells transfected with empty RAD vectors (Control) and DRAM1_707 plasmids. The cells with DRAM1_707 plasmids had an increase in fluorescence for up to six hours then dropped off at the seventh hour. There is a slight increase of fluorescence GFP/RFP ratio in the post-Nutlin-3 treated DRAM_707 plasmids compared to the empty RAD construct (Control and Nutlin-3 treatment). B.6b GFP/RFP of DRAM1_707 in Control condition and B.6c GFP/RFP of DRAM1_707 in Nutlin-3 treatment tracks individual cells over the twelve hours time course.

DRAM1_707 and DRAM1_911, into HCT116 cells, which have a robust response to Nutlin-3[6]. Using fluorescence microscopy on live cells, we monitored RFP fluorophores to gauge plasmid copy number, and GFP fluorophores to measure enhancer activity. Ideally, an empty RAD vector, serving as a negative control, would exhibit minimal GFP-positive cells, helping to establish baseline RFP activity. Meanwhile, the RAD+DNA template plasmid under identical conditions should display a

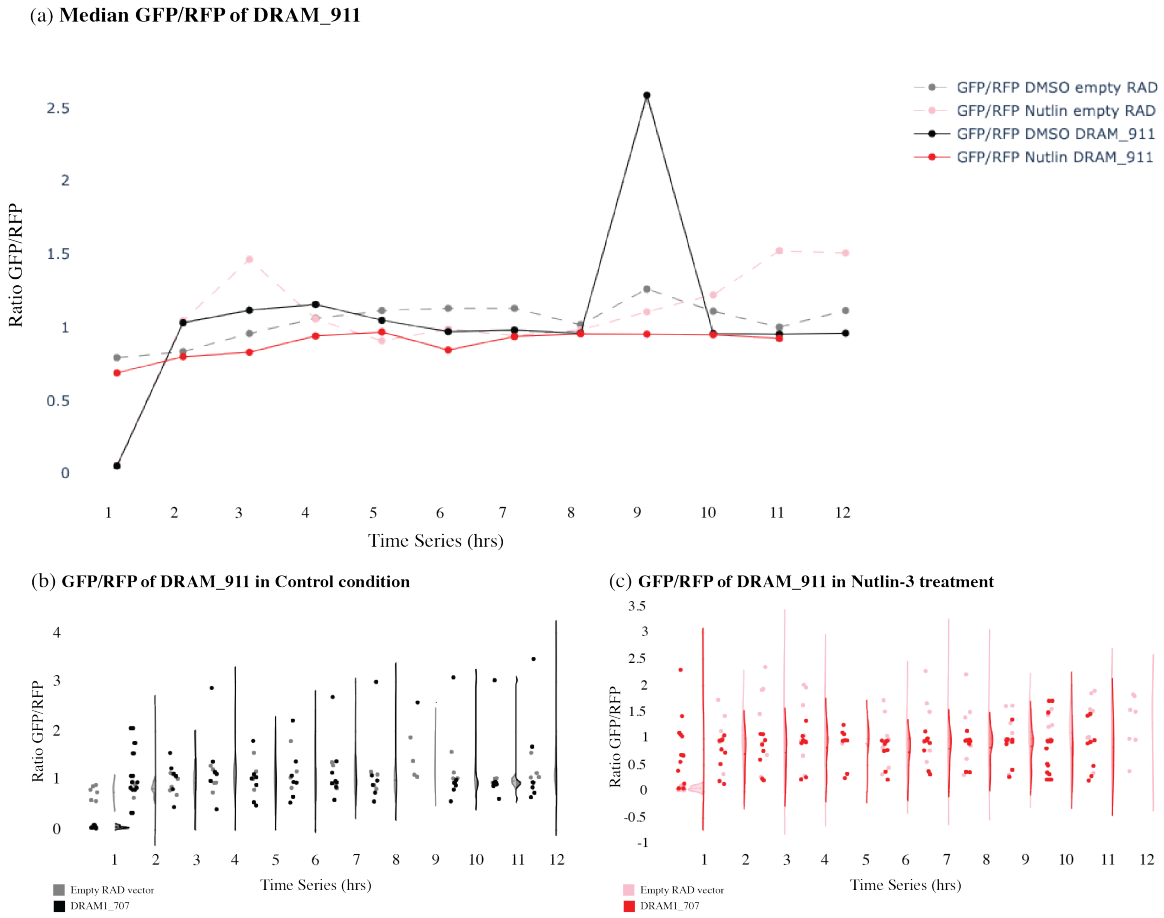


Figure B.7: **Ratio of GRF/RFP for cells with DRAM_911 plasmid.** Plot of the ratio of GRF/RFP to compare the two conditions; Control (DMSO) and Nutlin-3 perturbation. B.7a Over the twelve hours time course, the GFP/RFP ratio was plotted for both cells transfected with empty RAD vectors (Control) and DRAM1_911 plasmids. The GFP/RFP ratio for the control condition and Nutlin-3 treatment is similar for both the cells with empty RAD vectors and DRAM1_911 plasmids with a spike at nine hours for the DRAM1_911 plasmid. The GFP/RFP ratio distribution of individual cells B.7b GFP/RFP of DRAM1_911 in Control condition and B.7c GFP/RFP of DRAM1_911 in Nutlin-3 treatment suggest that at nine hours for the DMSO, there was a technical error in capturing the fluorescent read out for DRAM1_911 plasmid. The overall observation suggests that the cells with DRAM1_911 plasmids did not have an increase in the GFP fluorescence upon perturbation greater than cells with empty RAD vectors.

GFP signal with intensity increasing with time due to the enhancer activity leading to transcription and subsequent protein accumulation stemming from Enhancer-Promoter interactions (Figure B.1b, c and B.3b, c).

Yet, my observations deviated from these expectations. The negative control, anticipated to show low GFP expression, exhibited a surprisingly elevated GFP signal nearly equivalent to the cells with DRAM1_911 plasmid (Figure B.7). The cells with DRAM1_707 plasmids had an increase in in fluorescence up to six hours then had a dropped off at the seventh hour. There is slight increased of fluorescence GFP/RFP ratio in the post-Nutlin-3 treated cells with DRAM_707 plasmids compared to the control condition as well as cells with empty RAD vectors in both control and post-Nutlin-3 treatment (Figure B.6). However, the signal of GFP is lower in the post-Nutlin-3 treatment of cells with DRAM1_707 plasmids compared to the negative control (empty RAD vectors), implying that the dual reporter system was not functioning as expected.

To ensure that the observations made with the negative preliminary result were not due to the region we targeted, we selected two more potential enhancer regions. Gilson Sanchez, a postdoctoral researcher in the Dowell lab, designed two new RAD plasmids targeting the SBE4 and VDRE regions — TGF-β targets responsive to Ghrelin. After verifying these targets using quantitative PCR, Gilson transfected the RAD+insert vectors into SH-SY5Y and HEPG2 cell lines. Seven hours post-transfection, he subjected them to varied experimental conditions to assess dose response, using concentrations ranging from 75 to 300nM Ghrelin or 2 to 10 ng/mL TGF-β. Visually, Gilson observed a weak GFP signal in cells containing the RAD + DNA template plasmid. When I analyzed the live imaging data he gathered, the findings mirrored those of the DRAM1 plasmids experiment. Irrespective of the treatment dose, the GFP signal remained consistent between the RAD+DNA template plasmid and the negative control (empty RAD vector), indicating an absence of a dose-dependent RFP activity increase. These results suggest that the dual reporter system requires further refinement to enhance its sensitivity and specificity.

B.4 Conclusion and future direction

The dual reporter exhibited low background noise for both GFP and RFP fluorescence. If the reporter functions as expected, only the GFP fluorescence should have increased after perturbation, indicating higher enhancer activity and subsequent transcription of the GFP gene. A potential

reason for this discrepancy could be the low transcription rate of the mCMV promoter placed upstream of the GFP gene. Delving further into the literature it was found that researchers found that the activity of mCMV varied across different cell types, suggesting that the mammalian cell used might not have been the best choice for testing the RAD construct[142, 167]. A more consistently expressing promoter in vivo, such as the human elongation factor 1α promoter (EF1-α), might have been more suitable [74, 167].

Another concern involves the insulator in the RAD construct. It seems it failed to prevent cross-talk between the enhancer and the mPGK promoter upstream of the mScarlet gene. Referring to the design in Raab et al., 2012, it was observed that a single tDNA was insufficient to block the enhancer. While two tDNAs offered moderate blocking, it took four tDNAs to achieve robust blocking[143]. This could suggest a weak insulator design in our construct.

Furthermore, observations from Nina Ripin, a post-doctorate in Roy Parker's lab, highlighted an accumulation of stress granules in cells approximately six hours post-plasmid transfection. By the 24-hour mark, a minority (1%) of cells still had these granules, and under external stress, only half-formed granules. This indicated significant cell stress. A proper cell response was not observed until 48 hours post-transfection. Factors like plasmid, siRNA, and lipofectamine concentrations influenced the post-transfection cellular response. Considering Parker's lab recent findings, an optimization strategy for the RAD+insert reporter construct would be to leverage the PiggyBac transposase encoded on the RAD plasmid backbone.

To optimize the project further, my first step would involve integrating the current RAD construct design prior to cell perturbation. The stress from the transfection might have reduced the detected fluorescence signal. If redesigning the construct, I would enhance the insulator by adding more tDNA units to prevent cross-talk between reporter blocks. Additionally, replacing the mCMV promoter with a more robust one would be beneficial.

B.5 Contribution of other lab members

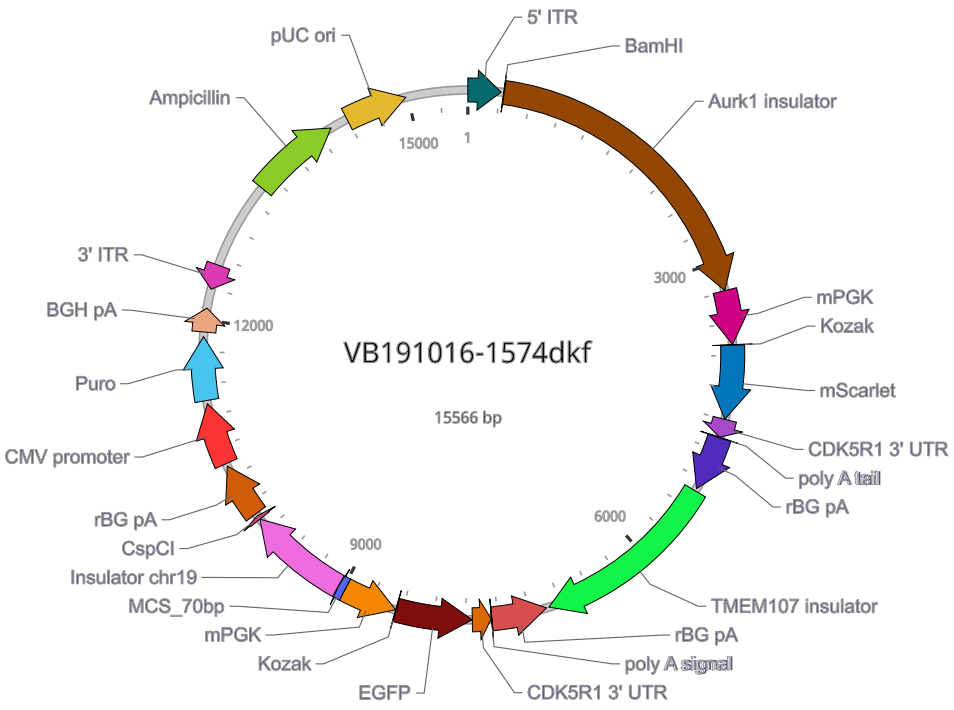
Mary Ann Allen and Gilson Sanchez were involved in the initial design of the construct. Gilson Sanchez led the creation of the TGF-β enhancer for the preliminary validation of the RAD construct.

B.6 Vector Builder Information

Vector Summary

Vector ID	VB191016-1574dkf
Vector Name	pPB[Exp]-Puro-Aurk1 insulator:mPGK>mScarlet:CDK5R1 3' UTR:rBG pA:TMEM107 insulator:rev(mPGK>EGFP:CDK5R1 3' UTR:rBG pA):chr19tDNA insulator:CspCI
Date Created (Pacific Time)	2019-10-16
Vector Size	15566 bp
Vector Type	Mammalian Gene Expression PiggyBac Vector
Plasmid Copy Number	High
Antibiotic Resistance	Ampicillin
Cloning Host	VB UltraStable (or alternative strain)

Vector Map



Vector Components

Name	Position	Size (bp)	Type	Description	Application notes
5' ITR	1-313	313	ITR	piggyBac 5' inverted terminal repeat	note=Unknown feature type:ITR color: #a949ca; direction: RIGHT
BamHI	337-342	6	misc_feature	*None*	note=BamHI
Aurk1 insulator	343-3284	2942	misc_feature	*None*	note=Aurk1 insulator
mPGK	3285-3795	511	misc_feature	*None*	note=mPGK
Kozak	3796-3801	6	misc_feature	*None*	note=Kozak
mScarlet	3802-4500	699	misc_feature	*None*	note=mScarlet
CDK5R1 3' UTR	4501-4674	174	misc_feature	*None*	note=CDK5R1 3' UTR
poly A tail	4675-4680	6	misc_feature	*None*	note=poly A tail
rBG pA	4681-5202	522	misc_feature	*None*	note=rBG pA
TMEM107 insulator	5227-7012	1786	ORF	*None*	note=Unknown feature type:ORF color: #0ed8aa; direction: RIGHT
rBG pA	complement (7037-7558)	522	misc_feature	*None*	note=rBG pA
poly A signal	7559-7564	6	misc_feature	*None*	note=poly A signal
CDK5R1 3' UTR	complement (7565-7741)	177	misc_feature	*None*	note=CDK5R1 3' UTR
EGFP	complement (7742-8461)	720	misc_feature	*None*	note=EGFP
Kozak	complement (8462-8467)	6	misc_feature	*None*	note=Kozak
mPGK	complement (8468-8978)	511	misc_feature	*None*	note=mPGK

Name	Position	Size (bp)	Type	Description	Application notes
MCS_70bp	8979-9048	70	misc_feature	*None*	note=MCS_70bp
Insulator chr19	9049-10018	970	misc_feature	*None*	note=Insulator chr19
CspCI	10019-10058	40	misc_feature	*None*	note=CspCI
rBG pA	10081-10602	522	PolyA_signal	Rabbit beta-globin polyadenylation signal	full_name=Rabbit beta-globin polyadenylation signal note=Unknown feature type:PolyA_signal color: #1f36a9; direction: RIGHT
CMV promoter	10628-11215	588	Promoter	Human cytomegalovirus immediate early enhancer/promoter	full_name=Human cytomegalovirus immediate early promoter note=Unknown feature type:Promoter color: #db6901; direction: RIGHT
Puro	11247-11846	600	ORF	Puromycin resistance gene	full_name=Puromycin resistance gene note=Unknown feature type:Marker color: #800c0c; direction: RIGHT
BGH pA	11890-12114	225	PolyA_signal	Bovine growth hormone polyadenylation signal	full_name=Bovine growth hormone polyadenylation note=Unknown feature type:PolyA_signal color: #0c5d90; direction: RIGHT
3' ITR	complement (12296-12530)	235	ITR	piggyBac 3' inverted terminal repeat	note=Unknown feature type:ITR color: #f58600; direction: LEFT
Ampicillin	13362-14222	861	ORF	Ampicillin resistance gene	full_name=Ampicillin resistance gene note=Unknown feature type:ORF color: #5566f5; direction: RIGHT
pUC ori	14393-14981	589	Rep_origin	pUC origin of replication	full_name=pUC origin of replication note=Unknown feature type:Rep_origin color: #ef6cdf; direction: RIGHT

Note: Components added by user are listed in bold red text.

Vector Sequence

```
1    TTAACCCTAG AAAGATAGTC TGCGTAAAAT TGACGCATGC ATTCTTGAAA TATTGCTCTC TCTTTCTAAA TAGCGCGAAT
81   CCGTCGCTGT GCATTTAGGA CATCTCAGTC GCCGCTTGGA GCTCCCGTGA GGCGTGCTTG TCAATGCGGT AAGTGTCACT
161  GATTTTGAAC TATAACGACC GCGTGAGTCA AAATGACGCA TGATTATCTT TTACGTGACT TTTAAGATTT AACTCATACG
241  ATAATTATAT TGTTATTTCA TGTTCTACTT ACGTGATAAC TTATTATATA TATATTTTCT TGTTATAGAT ATCATCAACT
321  TTGTATAGAA AAGTTGGGAT CCCCTGGGCT TCTAGTGACT GCTGGGAAGT CTGGAGAAAG GTAATATTTG TATTTAATAT
401  CCTTGTGTTC AGTGTTAAGT AAACATACCT AGAACAGCAG AGATTATTAA AAACATTTAA AAAGGCCGGG CACAGTGGCT
481  CACGCCACTA AAATACAAAA TTATCCGGGC GTGGTGGTGC ATGCCTGTAA TCCCAGCTAC TCGGGACGCT GAGGCAGGAG
561  AATCCCTTTG AACCCGTGGG CGGAGGTTGC GGTGAGCCGC GATAGTGCCA TTGCATTCCA GCCTGGGCAA CAAGAGAGAA
641  ACTCCGTCTC AAGGGAAAAA AAAATACAAA CAAAAAATAA TAATTTAAAA AATTATTTAA CAATTTCCAT ACAACAGGGA
721  ATTTTTTTAT TTTTAAATTT ATTTTTACTT TTATAGACCT TAGGGATACA TGCGCAGTTT TGCCACAGGG ATATATCACA
801  TAATGTTGAA AGTCTGCACC CAACTTCTGA ATAACGTTCA TATTTGTCTC ATTCCCCTCC TATCCTTCTA CTGAGGAGGC
881  TTTTTTTTTT TCCTATCCTG AAAAATGTCA CCGTAGGTCT ATTTGATTAG GGAAAACAAT GTTTTGTTTG ACCCCGACGT
961  GATTTGAACA CGCAACCTTC TGATCTGGAG TCAGACGCGC TACCGTTGCG CCACGAGGCC TGTCCAACAG CGTGACTTGA
1041 TGGGTCTGTT GAGCAGCAAT AACGGGCGGA CTTTTGTCAT GAAGCCCCAC CTTTAAGATT TGTCAGCGCA TTTCTCTGAG
1121 GCACCAGCTG ACGAGATGTG GGCGCATGTA CGTAGGAATT AAATTCATCC TCTGGATAAA TAATAGAGAG GCAAGTTGCG
1201 TTTAGGCGCA TTTGGCTTAA GGCCCAATTC CTTAACTGGA ATCACATTGA GCTCGTTGAG CTCACTCTAC AGAGAAGCCA
1281 TTTTTATGTG TCTAGGCCTT TTTTTTTGTT TTCTGGAGAT GATAAATCAT GGTGTCGCTG CAGTTCTCGC CCCCTGGACA
1361 AATCTTAGAC GTAATCCCAG CTCTGCCGGT GACTCAGCTT CCCTGGATCT TGGTTTCCTC GTCTGTAGAC TGGGCATAAC
1441 CCTACAGGTT CATGTGGGGT GGGTGGCGCG CGCTAGCGGT GAAGGTCACT CACAATTGCG CGCTGGGCAG ACGACGGCAG
1521 CCATTACTTT TACCTCGATC GCTGTTTTCC TGGGACCGCA CGGGTCCAAC CCGACTCATC CCAACCAACC TGAGGTATGA
1601 AAACCAGGAA AGAGAGCTAG CACCGGAGCG TTGGTGGTAT AGTGGTAAGC ATAGCTGCCT TCCAAGCAGT TGACCCGGGT
1681 TCGATTCCCG GCCAACGCAA GTCGTTTTGG GTGTTTTTTC CCCCCCCCGC CTTTTCCTTT TCGTGTTTTC TGGGCCCCAG
1761 CATCGTTGAG GGTTTTCGTG AGGTTTTCCT GAGGAAACTT CCGCTCCGAA AGGACCCACT TTCCGCTACA CCCGCGACCA
1841 CGGCTGGACC ACCGCGCTCC TGACGGATGC GCCCTGCAAG CCCCTCCAGG CGAGAGCAGC GCGTCGGAAT CTCGGCCCGG
1921 GGTTGCCTAC TGTTCCCGGA TCGCAGTGCG AATCCACAAA AAGCAGGGAG GCGAAAGGAA ACAAAAGCCC GGGCTCTGCA
2001 CCCTGGACGG CTGGGGGCCA AAGCTACTTG CCCAGGTTCA AGCAGAGACT GGAAATGCGC ACTTCCGTCA AGGAGGGAGC
2081 AAGCTAGGCC AAGTCTGAGC GCTAATGAGC ATACTGAGCA GAGCAGATAA TGGCCACTTT AGGTACACGT TTGAACTCTA
2161 AGTTGCTTGA AGTAAAATGG AAACTTAGAA TGCGAGCTGG AGAGGAAGGC CAAACTCTCC TCCCCTCCTC CTTTCCCTCT
2241 TTTTTATAGC TCCATCCCAG GGACCAGGAC GGATTTAATG AAAGCAAAAA GATAGAAGTT TCAAAACTGC TAGTACTCTC
2321 CCCGTCGGGG AATCGAACCC CGGTCTCCCG CGTGACAGGC GGGGATACTC ACCACTATAC TAACGAGGAG GGACCTGTAG
2401 AGACCTTTCT CCGAATCCTT GATCTGAGAC TTTACACTCT GCGTGCTGAC AGCAGGGTTC CTAGGGGTTC TGTTCTTCCA
2481 ACCACTGATA AACGCTGACC AGAAACCGTA GGTGTATTTC AGGCACCAAT CGTCAGATGG AGTGACCGAA AGAGACAGAG
2561 TGCCTTCAGG ACTGCAGAGA ACCAAGTTTT GACAGTTCCT GGGCTCTCCC GAACACTGTC ATCCACAACT ATAGATTGAC
2641 GCGTGTCAAT CAAAGCGAAA CCAGGCTGTG GTGGCCGAGT TGAGACACCA GGAAGGAAGT CAGGAAGCAG AAAGGGAGGG
2721 ACCTGCCCAA GTCTGGGCGT GCCTCGTCTC TTCCTGACAT GCCAGGCGGG TCTGGGGATT GAGATGCTTT CTGCTACCGC
2801 GGTGCCACAA AAGAGCTTAC TGTACATTGA TGACTTACAG TCAAGCCTCC TGCAGCGCTG GTCAGGTAAC CTATTCCTTG
2881 AAGACACTCT TGTCTCAGTA AAAGGTAAAG GAGGGCTCGT CCGGGATTTG AACCCGGGAC CTCTCGCACC CGAAGCGAGA
2961 ATCATACCCC TAGACCAACG AGCCGACGTG CGGACGTTGC CGCGAACCGC CTTAGAGGTC GTGCCAGGCT TGCTGTAGTG
3041 CTGGGTCCAC TATGCATGGC GGAACGGTCC GGGCGCACGC TCACGGACCA GCCTCCCCCA GGCCGAGTAT TTTGGAGCAC
3121 TGGGCTGGGA ATCTCTTGGC TCCGGGCCGC GAGCTCCGGC TCCTCCCAGG AAATAGCGTC AAGGAAGTGG GAGGGAGTGG
3201 CCTCGGCCTC GCCCCCGCGC CGCCTTGCAC GACTGCCTTG AACACCCGGG TTGTTGCTTC CTTTGGTTAC CGACTTGGGG
3281 GACCTTCTAC CGGGTAGGGG AGGCGCTTTT CCCAAGGCAG TCTGGAGCAT GCGCTTTAGC AGCCCCGCTG GGCACTTGGC
3361 GCTACACAAG TGGCCTCTGG CCTCGCACAC ATTCCACATC CACCGGTAGG CGCCAACCGG CTCCGTTCTT TGGTGGCCCC
3441 TTCGCGCCAC CTTCTACTCC TCCCCTAGTC AGGAAGTTCC CCCCCGCCCC GCAGCTCGCG TCGTGCAGGA CGTGACAAAT
3521 GGAAGTAGCA CGTCTCACTA GTCTCGTGCA GATGGACAGC ACCGCTGAGC AATGGAAGCG GGTAGGCCTT TGGGGCAGCG
```

3601 GCCAATAGCA GCTTTGCTCC TTCGCTTTCT GGGCTCAGAG GCTGGGAAGG GGTGGGTCCG GGGGCGGGCT CAGGGGCGGG
3681 CTCAGGGGCG GGGCGGGCGC CCGAAGGTCC TCCGGAGGCC CGGCATTCTG CACGCTTCAA AAGCGCACGT CTGCCGCGCT
3761 GTTCTCCTCT TCCTCATCTC CGGGCCTTTC GACCTGCCAC CATGGTGAGC AAGGGCGAGG CAGTGATCAA GGAGTTCATG
3841 CGGTTCAAGG TGCACATGGA GGGCTCCATG AACGGCCACG AGTTCGAGAT CGAGGGCGAG GGCGAGGGCC GCCCCTACGA
3921 GGGCACCCAG ACCGCCAAGC TGAAGGTGAC CAAGGGTGGC CCCCTGCCCT TCTCCTGGGA CATCCTGTCC CCTCAGTTCA
4001 TGTACGGCTC CAGGGCCTTC ACCAAGCACC CCGCCGACAT CCCCGACTAC TATAAGCAGT CCTTCCCCGA GGGCTTCAAG
4081 TGGGAGCGCG TGATGAACTT CGAGGACGGC GGCGCCGTGA CCGTGACCCA GGACACCTCC CTGGAGGACG GCACCCTGAT
4161 CTACAAGGTG AAGCTCCGCG GCACCAACTT CCCTCCTGAC GGCCCCGTAA TGCAGAAGAA GACAATGGGC TGGGAAGCGT
4241 CCACCGAGCG GTTGTACCCC GAGGACGGCG TGCTGAAGGG CGACATTAAG ATGGCCCTGC GCCTGAAGGA CGGCGGCCGC
4321 TACCTGGCGG ACTTCAAGAC CACCTACAAG GCCAAGAAGC CCGTGCAGAT GCCCGGCGCC TACAACGTCG ACCGCAAGTT
4401 GGACATCACC TCCCACAACG AGGACTACAC CGTGGTGGAA CAGTACGAAC GCTCCGAGGG CCGCCACTCC ACCGGCGGCA
4481 TGGACGAGCT GTACAAGTGA GCGGGTCTAG TGGAAAGAGC AGCAGACAAG GGGGTAGTGA GCGGGTCTAG TGGAAAGATT
4561 GTGTCTGGTC GTTTGACCAC ACACCGCCCT GATTTGCTGT TTTCTTTTTT TAGGGAGAAG GGCTTTTCTT TAGTGGAGAA
4641 ATGGAACTCG CCCCCCTACC CCCTTGTCTG CTGCAATAAA TCCTCAGGTG CAGGCTGCCT ATCAGAAGGT GGTGGCTGGT
4721 GTGGCCAATG CCCTGGCTCA CAAATACCAC TGAGATCTTT TTCCCTCTGC CAAAAATTAT GGGGACATCA TGAAGCCCCT
4801 TGAGCATCTG ACTTCTGGCT AATAAAGGAA ATTTATTTTC ATTGCAATAG TGTGTTGGAA TTTTTTGTGT CTCTCACTCG
4881 GAAGGACATA TGGGAGGGCA AATCATTTAA AACATCAGAA TGAGTATTTG GTTTAGAGTT TGGCAACATA TGCCCATATG
4961 CTGGCTGCCA TGAACAAAGG TTGGCTATAA AGAGGTCATC AGTATATGAA ACAGCCCCCT GCTGTCCATT CCTTATTCCA
5041 TAGAAAAGCC TTGACTTGAG GTTAGATTTT TTTTATATTT TGTTTTGTGT TATTTTTTTC TTTAACATCC CTAAAATTTT
5121 CCTTACATGT TTTACTAGCC AGATTTTTCC TCCTCTCCTG ACTACTCCCA GTCATAGCTG TCCCTCTTCT CTTATGGAGA
5201 TCCAAGTTTG TACAAAAAAG CAGGCTGTTC CGTCTGGTTC CCATCCTTCC ATCCGTTTCT AGGCAGGTCC GTCCCCACTG
5281 GGCTGTAAGC TTTGGCATGC CCTGGCGATC AGCTCGGGAC CCTCTATCTG GGCGTTGGCA GGACGCCGGG GGCCGGGAGG
5361 GACAGACCGC TAAGCCTGCA TGCCATAGTC ACTGCCCTGG GGTGCCACTC GCCCGGCTCG TCCTACAGGG CTGGCTCGGC
5441 GAGCGCAGAT ACGACCCCGC AGCTGTTCAG AGGGGCAGAA ATGCCCTAGG TGCCCATCCA TGCCTCGATT CATGACCCTG
5521 GCCTCCAGGG CGCACAGTGG TCATGGGGAG ACCTGAGCTG CCGAGTGGCC GGCCGACCTC GTGGCGCAAC GGTAGCGCGT
5601 CTGACTCCAG ATCAGAAGGT TGCGTGTTCA AATCACGTCG GGGTCAGCGG CTATTTTTCT TCGGTTTTTA TTAACCCCCT
5681 TTATTTTAAA CTACGGTCGA GCTTCAGCGT TCAGGTCATT GAAGAAGCAA TATCTCCTTG GGGCCTGAAG GAGAGGGGTT
5761 TCTGGAAGTT CCAAGGCCGC CCCGTCTGGA CAGCCCAACC ATCGCGCGGG GATTTTTGCG ATGCATGCGG GTACCGTAAT
5841 TCTGGGCGGA TAACGCGGGT CCTAAGACAG GAGCAGTTCT AGACCTCTCA GCAGAGGGAC GAGGGTCGGG CCATCACGCA
5921 TGGAAGAAGT CGGTCTCTGA TCTACGAGTT CTCTTTCCAG TGCCGAGCGG ATTCCTTCCA AATGTGCAGC CTTCACGACG
6001 TAGGCGAGCC CCACCTGCAG GAAGTTCAGG TTCCAGAGAA GTGAGATGCG GAGGGCAGTC TGAACAGCGA GGGCTGTCTG
6081 CAGACGAGGT GGCCGAGTGG TTAAGGCGAT GGACTGCTAA TCCATTGTGC TCTGCACGCG TGGGTTCGAA TCCCATCCTC
6161 GTCGGCTAAG GAAGTCCTGT GCTCAGTTTT GTAGCATCAA AACTAGGATT TCTCTTGTTA CCCCCAGTCA CTCCATTCAG
6241 TTTTCGTGTC TTTCCCAGCT GCATCCATCC TTTCCTCATT TTCGTATGCA GCCGACTTTT TGTGACATCT TTGTATTCAT
6321 TCTCTGCAAT TCAGCTGACC TGGCCAAGGA AACAAGATCC TAAGCGTCTT TCCGGCGGCG CCGTGGCTTA GTTGGTTAAA
6401 GCGCCTGTCT AGTAAACAGG AGATCCTGGG TTCGAATCCC AGCGGTGCCT CCGTGTTTCC CCCACGCTTT TGCCAACATT
6481 AAACATTGTG AGGACAGTTG CAGAAACTCA TAACTTCCAT CCTACATGGT TTACTCACGT ACCCATCTAT CCTCTCCCGG
6561 TGCATCTGCC ACACGCTGTT GGGTTTTTGC TCTTCGTGCA CATGGTACTT GCGCCTCGAC CTGCAGTTAC ACCAGTCGCA
6641 TCATCTGTAC TTGCCAGTAC TGTCTGTTAC TCGTCTAGGT GCTCAGTTAG TCAGTGTTTG TGTTCTGTTT TCATTTCTCA
6721 TAAAGCAGTC CCACAACTGA AGGTTTTTCC CGAGAGAACT GAACAGTATT GTAACTAAAT AGATTTATTT CAAGCTTTTC
6801 GTAGCAACTG GCCGGTTAGC TCAGTTGGTT AGAGCGTGGT GCTAATAACG CCAAGGTCGC GGGTTCGATC CCCGTACGGG
6881 CCAGGATTGA AACTTTTCGA AAGTACGATT ACTGCACTCC GTTTTAGAGC CAAGTAACGT CTCTGGGGAA AAACAGCGCC
6961 ACATTTCCAA TCCCAGAACA GGGAGCGTAT TGGAGCGCAT TCTAAAGTGG GCACCCAGCT TTCTTGTACA AAGTGGGATC
7041 TCCATAAGAG AAGAGGGACA GCTATGACTG GGAGTAGTCA GGAGAGGAGG AAAAATCTGG CTAGTAAAAC ATGTAAGGAA
7121 AATTTTAGGG ATGTTAAAGA AAAAAATAAC ACAAAACAAA ATATAAAAAA AATCTAACCT CAAGTCAAGG CTTTTCTATG
7201 GAATAAGGAA TGGACAGCAG GGGGCTGTTT CATATACTGA TGACCTCTTT ATAGCCAACC TTTGTTCATG GCAGCCAGCA
7281 TATGGGCATA TGTTGCCAAA CTCTAAACCA AATACTCATT CTGATGTTTT AAATGATTTG CCCTCCCATA TGTCCTTCCG
7361 AGTGAGAGAC ACAAAAAATT CCAACACACT ATTGCAATGA AAATAAATTT CCTTTATTAG CCAGAAGTCA GATGCTCAAG

```
7441  GGGCTTCATG ATGTCCCCAT AATTTTTGGC AGAGGGAAAA AGATCTCAGT GGTATTTGTG AGCCAGGGCA TTGGCCACAC
7521  CAGCCACCAC CTTCTGATAG GCAGCCTGCA CCTGAGGATT TATTGCAGCA GACAAGGGGG TAGGGGGGCG AGTTCCATTT
7601  CTCCACTAAA GAAAAGCCCT TCTCCCTAAA AAAAGAAAAC AGCAAATCAG GGCGGTGTGT GGTCAAACGA CCAGACACAA
7681  TCTTTCCACT AGACCCGCTC ACTACCCCCT TGTCTGCTGC TCTTTCCACT AGACCCGCTC ATTACTTGTA CAGCTCGTCC
7761  ATGCCGAGAG TGATCCCGGC GGCGGTCACG AACTCCAGCA GGACCATGTG ATCGCGCTTC TCGTTGGGGT CTTTGCTCAG
7841  GGCGGACTGG GTGCTCAGGT AGTGGTTGTC GGGCAGCAGC ACGGGGCCGT CGCCGATGGG GGTGTTCTGC TGGTAGTGGT
7921  CGGCGAGCTG CACGCTGCCG TCCTCGATGT TGTGGCGGAT CTTGAAGTTC ACCTTGATGC CGTTCTTCTG CTTGTCGGCC
8001  ATGATATAGA CGTTGTGGCT GTTGTAGTTG TACTCCAGCT TGTGCCCCAG GATGTTGCCG TCCTCCTTGA AGTCGATGCC
8081  CTTCAGCTCG ATGCGGTTCA CCAGGGTGTC GCCCTCGAAC TTCACCTCGG CGCGGGTCTT GTAGTTGCCG TCGTCCTTGA
8161  AGAAGATGGT GCGCTCCTGG ACGTAGCCTT CGGGCATGGC GGACTTGAAG AAGTCGTGCT GCTTCATGTG GTCGGGGTAG
8241  CGGCTGAAGC ACTGCACGCC GTAGGTCAGG GTGGTCACGA GGGTGGGCCA GGGCACGGGC AGCTTGCCGG TGGTGCAGAT
8321  GAACTTCAGG GTCAGCTTGC CGTAGGTGGC ATCGCCCTCG CCCTCGCCGG ACACGCTGAA CTTGTGGCCG TTTACGTCGC
8401  CGTCCAGCTC GACCAGGATG GGCACCACCC CGGTGAACAG CTCCTCGCCC TTGCTCACCA TGGTGGCAGG TCGAAAGGCC
8481  CGGAGATGAG GAAGAGGAGA ACAGCGCGGC AGACGTGCGC TTTTGAAGCG TGCAGAATGC CGGGCCTCCG GAGGACCTTC
8561  GGGCGCCCGC CCCGCCCCTG AGCCCGCCCC TGAGCCCGCC CCCGGACCCA CCCCTTCCCA GCCTCTGAGC CCAGAAAGCG
8641  AAGGAGCAAA GCTGCTATTG GCCGCTGCCC CAAAGGCCTA CCCGCTTCCA TTGCTCAGCG GTGCTGTCCA TCTGCACGAG
8721  ACTAGTGAGA CGTGCTACTT CCATTTGTCA CGTCCTGCAC GACGCGAGCT GCGGGGCGGG GGGGAACTTC CTGACTAGGG
8801  GAGGAGTAGA AGGTGGCGCG AAGGGGCCAC CAAAGAACGG AGCCGGTTGG CGCCTACCGG TGGATGTGGA ATGTGTGCGA
8881  GGCCAGAGGC CACTTGTGTA GCGCCAAGTG CCCAGCGGGG CTGCTAAAGC GCATGCTCCA GACTGCCTTG GGAAAAGCGC
8961  CTCCCCTACC CGGTAGAACT CGAGAGGCGC GCCCATATGA TTTAAATATC GCGAAGATCG ATAGCGATCG CAGCTAGCGA
9041  ATTCGAATGG CTCGGAGAAG CCCGGAGAGG ACCGCGGCCA CGACGCCGGC ACCGACCCCC GACGCCACCA CGACACCCGA
9121  GGCTCCCACG GCCCCAGACC CGCGCGGCCC CACACCCGCC GCCGGTGCCG CGCTCACCTG ACAGCACCGC CATCTTGGCC
9201  TCGGCCTTCA GACAACCTCT GAGGTCCGTT CTCCTTAGCG CTCCGGGTTC CCGGGGCCGC CGCCCAAAGC GGCTCGGAGC
9281  GCATGCGGAA ACCGGAACCT GGAGCCGGGA GGTTCCGCTG CGCTTTCGGC CTCCGTGGTC GTGTTGCGCC GTGCACCCCC
9361  AGATCACCCG AGGCGGAAAG GTCATCGGAG TGCGCACGTG GGGGCCGCGG GTCTGCGGCG GACCGACAAA AACAGAGGCG
9441  ACACTCATTT TCTTGTTATT CCAGGAGGGC AAAAGTCACC ACGCGTCCCT GGGTGGGCTC GAACCACCAA CCTTTCGGTT
9521  AACAGCCGAA CGCGCTAACC GATTGCGCCA CAGAGACGAC ACTGCGCTCC GGCGCCCTGG GAGACCTTAC ATAGGATGAC
9601  GGGCCTGCTG ACGCGAAACG ACTGCGCCTG CGCAGGGAAG GCGCAGCGCG GGACGGAGTC CCGGCGAGAG CGACTCTGCC
9681  GCCCAGCCGA AATAGCTCAG TTGGGAGAGC GTTAGACTGA AGATCTAAAG GTCCCTGGTT CGATCCCGGG TTTCGGCAGA
9761  AGTTTTAGCG CCTCTTGCGA TCACTGATGT CTTTCGTTCA GGATTGGTCC AGCTGCCTCG CAGCTATCCC GGACCCCTTA
9841  CCGCAACATC ACCGCATTCT GCAGAGGAGT CGCCGCGATG TGGACTCATC CCAAAGTAAG GGGGCTCAGA CCTCACAACC
9921  CCAAAGCCTC CTCTTCCCAC CCCTAAGAGG TCATAATTTC AGCTCAAACT TCCCCCTTCC CTGACACACC CTCGATTTTC
10001 CATCTTGCTT AAATTGATAG TCAAACCGCT ATCCACGCCC ATTGATGTAC TGCCAAAACA ACTTTATTAT ACATAGTTGA
10081 TCCTCAGGTG CAGGCTGCCT ATCAGAAGGT GGTGGCTGGT GTGGCCAATG CCCTGGCTCA CAAATACCAC TGAGATCTTT
10161 TTCCCTCTGC CAAAAATTAT GGGGACATCA TGAAGCCCCT TGAGCATCTG ACTTCTGGCT AATAAAGGAA ATTTATTTTC
10241 ATTGCAATAG TGTGTTGGAA TTTTTTGTGT CTCTCACTCG GAAGGACATA TGGGAGGGCA AATCATTTAA AACATCAGAA
10321 TGAGTATTTG GTTTAGAGTT TGGCAACATA TGCCCATATG CTGGCTGCCA TGAACAAAGG TTGGCTATAA AGAGGTCATC
10401 AGTATATGAA ACAGCCCCCT GCTGTCCATT CCTTATTCCA TAGAAAAGCC TTGACTTGAG GTTAGATTTT TTTTATATTT
10481 TGTTTTGTGT TATTTTTTTC TTTAACATCC CTAAAATTTT CCTTACATGT TTTACTAGCC AGATTTTTCC TCCTCTCCTG
10561 ACTACTCCCA GTCATAGCTG TCCCTCTTCT CTTATGGAGA TCCCTCGACC TGCAGCCCAA GCTTCGCGTT GACATTGATT
10641 ATTGACTAGT TATTAATAGT AATCAATTAC GGGGTCATTA GTTCATAGCC CATATATGGA GTTCCGCGTT ACATAACTTA
10721 CGGTAAATGG CCCGCCTGGC TGACCGCCCA ACGACCCCCG CCCATTGACG TCAATAATGA CGTATGTTCC CATAGTAACG
10801 CCAATAGGGA CTTTCCATTG ACGTCAATGG GTGGAGTATT TACGGTAAAC TGCCCACTTG GCAGTACATC AAGTGTATCA
10881 TATGCCAAGT ACGCCCCCTA TTGACGTCAA TGACGGTAAA TGGCCCGCCT GGCATTATGC CCAGTACATG ACCTTATGGG
10961 ACTTTCCTAC TTGGCAGTAC ATCTACGTAT TAGTCATCGC TATTACCATG GTGATGCGGT TTTGGCAGTA CATCAATGGG
11041 CGTGGATAGC GGTTTGACTC ACGGGGATTT CCAAGTCTCC ACCCCATTGA CGTCAATGGG AGTTTGTTTT GGCACCAAAA
11121 TCAACGGGAC TTTCCAAAAT GTCGTAACAA CTCCGCCCCA TTGACGCAAA TGGGCGGTAG GCGTGTACGG TGGGAGGTCT
    ATATAAGCAG AGCTCTCTGG CTAACTAGAG AACCCACTGC GCCACCATGA CCGAGTACAA GCCCACGGTG CGCCTCGCCA
```

```
11201 CCCGCGACGA CGTCCCCAGG GCCGTACGCA CCCTCGCCGC CGCGTTCGCC GACTACCCCG CCACGCGCCA CACCGTCGAT
11281 CCGGACCGCC ACATCGAGCG GGTCACCGAG CTGCAAGAAC TCTTCCTCAC GCGCGTCGGG CTCGACATCG GCAAGGTGTG
11361 GGTCGCGGAC GACGGCGCCG CGGTGGCGGT CTGGACCACG CCGGAGAGCG TCGAAGCGGG GGCGGTGTTC GCCGAGATCG
11441 GCCCGCGCAT GGCCGAGTTG AGCGGTTCCC GGCTGGCCGC GCAGCAACAG ATGGAAGGCC TCCTGGCGCC GCACCGGCCC
11521 AAGGAGCCCG CGTGGTTCCT GGCCACCGTC GGCGTCTCGC CCGACCACCA GGGCAAGGGT CTGGGCAGCG CCGTCGTGCT
11601 CCCCGGAGTG GAGGCGGCCG AGCGCGCCGG GGTGCCCGCC TTCCTGGAGA CCTCCGCGCC CCGCAACCTC CCCTTCTACG
11681 AGCGGCTCGG CTTCACCGTC ACCGCCGACG TCGAGGTGCC CGAAGGACCG CGCACCTGGT GCATGACCCG CAAGCCCGGT
11761 GCCTGACTCG AGTCTAGAGG GCCCGTTTAA ACCCGCTGAT CAGCCTCGAC TGTGCCTTCT AGTTGCCAGC CATCTGTTGT
11841 TTGCCCCTCC CCCGTGCCTT CCTTGACCCT GGAAGGTGCC ACTCCCACTG TCCTTTCCTA ATAAAATGAG GAAATTGCAT
11921 CGCATTGTCT GAGTAGGTGT CATTCTATTC TGGGGGGTGG GGTGGGGCAG GACAGCAAGG GGGAGGATTG GGAAGACAAT
12001 AGCAGGCATG CTGGGGATGC GGTGGGCTCT ATGGCTCGAG TTAATTAACG AGAGCATAAT ATTGATATGT GCCAAAGTTG
12081 TTTCTGACTG ACTAATAAGT ATAATTTGTT TCTATTATGT ATAGGTTAAG CTAATTACTT ATTTTATAAT ACAACATGAC
12161 TGTTTTTAAA GTACAAAATA AGTTTATTTT TGTAAAAGAG AGAATGTTTA AAAGTTTTGT TACTTTATAG AAGAAATTTT
12241 GAGTTTTTGT TTTTTTTTAA TAAATAAATA AACATAAATA AATTGTTTGT TGAATTTATT ATTAGTATGT AAGTGTAAAT
12321 ATAATAAAAC TTAATATCTA TTCAAATTAA TAAATAAACC TCGATATACA GACCGATAAA ACACATGCGT CAATTTTACG
12401 CATGATTATC TTTAACGTAC GTCACAATAT GATTATCTTT CTAGGGTTAA ATAATAGTTT CTAATTTTTT TATTATTCAG
12481 CCTGCTGTCG TGAATACCGA GCTCCAATTC GCCCTATAGT GAGTCGTATT ACAATTCACT GGCCGTCGTT TTACAACGTC
12561 GTGACTGGGA AAACCCTGGC GTTACCCAAC TTAATCGCCT TGCAGCACAT CCCCCTTTCG CCAGCTGGCG TAATAGCGAA
12641 GAGGCCCGCA CCGATCGCCC TTCCCAACAG TTGCGCAGCC TGAATGGCGA ATGGGACGCG CCCTGTAGCG GCGCATTAAG
12721 CGCGGCGGGT GTGGTGGTTA CGCGCAGCGT GACCGCTACA CTTGCCAGCG CCCTAGCGCC CGCTCCTTTC GCTTTCTTCC
12801 CTTCCTTTCT CGCCACGTTC GCCGGCTTTC CCCGTCAAGC TCTAAATCGG GGGCTCCCTT TAGGGTTCCG ATTTAGTGCT
12881 TTACGGCACC TCGACCCCAA AAAACTTGAT TAGGGTGATG GTTCACGTAG TGGGCCATCG CCCTGATAGA CGGTTTTTCG
12961 CCCTTTGACG TTGGAGTCCA CGTTCTTTAA TAGTGGACTC TTGTTCCAAA CTGGAACAAC ACTCAACCCT ATCTCGGTCT
13041 ATTCTTTTGA TTTATAAGGG ATTTTGCCGA TTTCGGCCTA TTGGTTAAAA AATGAGCTGA TTTAACAAAA ATTTAACGCG
13121 AATTTTAACA AAATATTAAC GCTTACAATT TAGGTGGCAC TTTTCGGGGA AATGTGCGCG GAACCCCTAT TTGTTTATTT
13201 TTCTAAATAC ATTCAAATAT GTATCCGCTC ATGAGACAAT AACCCTGATA AATGCTTCAA TAATATTGAA AAAGGAAGAG
13281 TATGAGTATT CAACATTTCC GTGTCGCCCT TATTCCCTTT TTTGCGGCAT TTTGCCTTCC TGTTTTTGCT CACCCAGAAA
13361 CGCTGGTGAA AGTAAAAGAT GCTGAAGATC AGTTGGGTGC ACGAGTGGGT TACATCGAAC TGGATCTCAA CAGCGGTAAG
13441 ATCCTTGAGA GTTTTCGCCC CGAAGAACGT TTTCCAATGA TGAGCACTTT TAAAGTTCTG CTATGTGGCG CGGTATTATC
13521 CCGTATTGAC GCCGGGCAAG AGCAACTCGG TCGCCGCATA CACTATTCTC AGAATGACTT GGTTGAGTAC TCACCAGTCA
13601 CAGAAAAGCA TCTTACGGAT GGCATGACAG TAAGAGAATT ATGCAGTGCT GCCATAACCA TGAGTGATAA CACTGCGGCC
13681 AACTTACTTC TGACAACGAT CGGAGGACCG AAGGAGCTAA CCGCTTTTTT GCACAACATG GGGGATCATG TAACTCGCCT
13761 TGATCGTTGG GAACCGGAGC TGAATGAAGC CATACCAAAC GACGAGCGTG ACACCACGAT GCCTGTAGCA ATGGCAACAA
13841 CGTTGCGCAA ACTATTAACT GGCGAACTAC TTACTCTAGC TTCCCGGCAA CAATTAATAG ACTGGATGGA GGCGGATAAA
13921 GTTGCAGGAC CACTTCTGCG CTCGGCCCTT CCGGCTGGCT GGTTTATTGC TGATAAATCT GGAGCCGGTG AGCGTGGGTC
14001 TCGCGGTATC ATTGCAGCAC TGGGGCCAGA TGGTAAGCCC TCCCGTATCG TAGTTATCTA CACGACGGGG AGTCAGGCAA
14081 CTATGGATGA ACGAAATAGA CAGATCGCTG AGATAGGTGC CTCACTGATT AAGCATTGGT AACTGTCAGA CCAAGTTTAC
14161 TCATATATAC TTTAGATTGA TTTAAAACTT CATTTTTAAT TTAAAAGGAT CTAGGTGAAG ATCCTTTTTG ATAATCTCAT
14241 GACCAAAATC CCTTAACGTG AGTTTTCGTT CCACTGAGCG TCAGACCCCG TAGAAAAGAT CAAAGGATCT TCTTGAGATC
14321 CTTTTTTTCT GCGCGTAATC TGCTGCTTGC AAACAAAAAA ACCACCGCTA CCAGCGGTGG TTTGTTTGCC GGATCAAGAG
14401 CTACCAACTC TTTTTCCGAA GGTAACTGGC TTCAGCAGAG CGCAGATACC AAATACTGTT CTTCTAGTGT AGCCGTAGTT
14481 AGGCCACCAC TTCAAGAACT CTGTAGCACC GCCTACATAC CTCGCTCTGC TAATCCTGTT ACCAGTGGCT GCTGCCAGTG
14561 GCGATAAGTC GTGTCTTACC GGGTTGGACT CAAGACGATA GTTACCGGAT AAGGCGCAGC GGTCGGGCTG AACGGGGGGT
14641 TCGTGCACAC AGCCCAGCTT GGAGCGAACG ACCTACACCG AACTGAGATA CCTACAGCGT GAGCTATGAG AAAGCGCCAC
14721 GCTTCCCGAA GGGAGAAAGG CGGACAGGTA TCCGGTAAGC GGCAGGGTCG GAACAGGAGA GCGCACGAGG GAGCTTCCAG
14801 GGGGAAACGC CTGGTATCTT TATAGTCCTG TCGGGTTTCG CCACCTCTGA CTTGAGCGTC GATTTTTGTG ATGCTCGTCA
14881 GGGGGGCGGA GCCTATGGAA AAACGCCAGC AACGCGGCCT TTTTACGGTT CCTGGCCTTT TGCTGGCCTT TTGCTCACAT
14961
```

```
15041 GTTCTTTCCT GCGTTATCCC CTGATTCTGT GGATAACCGT ATTACCGCCT TTGAGTGAGC TGATACCGCT CGCCGCAGCC
15121 GAACGACCGA GCGCAGCGAG TCAGTGAGCG AGGAAGCGGA AGAGCGCCCA ATACGCAAAC CGCCTCTCCC CGCGCGTTGG
15201 CCGATTCATT AATGCAGCTG GCACGACAGG TTTCCCGACT GGAAAGCGGG CAGTGAGCGC AACGCAATTA ATGTGAGTTA
15281 GCTCACTCAT TAGGCACCCC AGGCTTTACA CTTTATGCTT CCGGCTCGTA TGTTGTGTGG AATTGTGAGC GGATAACAAT
15361 TTCACACAGG AAACAGCTAT GACCATGATT ACGCCAAGCT CGAAATTAAC CCTCACTAAA GGGAACAAAA GCTGGTACCT
15441 CGCGCGACTT GGTTTGCCAT TCTTTAGCGC GCGTCGCGTC ACACAGCTTG GCCACAATGT GGTTTTTGTC AAACGAAGAT
15521 TCTATGACGT GTTTAAAGTT TAGGTCGAGT AAAGCGCAAA TCTTTT
```

Validation by Restriction Enzyme Digestion

Restriction Enzymes	Cutting Sites	DNA Fragments (bp)
AscI	8988	15566
NaeI	5571, 9088, 12904	3517, 3816, 8233
ApaLI	3851, 6597, 9352, 13478, 14724	2746, 2755, 4126, 1246, 4693
AsiSI	9029	15566
EcoRI	9040	15566
ApaLI+AscI	3851, 6597, 8988, 9352, 13478, 14724	2746, 2391, 364, 4126, 1246, 4693
ApaLI+NaeI	3851, 5571, 6597, 9088, 9352, 12904, 13478, 14724	1720, 1026, 2491, 264, 3552, 574, 1246, 4693
ApaLI+AsiSI	3851, 6597, 9029, 9352, 13478, 14724	2746, 2432, 323, 4126, 1246, 4693
ApaLI+EcoRI	3851, 6597, 9040, 9352, 13478, 14724	2746, 2443, 312, 4126, 1246, 4693

Appendix C

Other contributions

In this Appendix, I outline additional work that I completed during my Ph.D. candidacy. The first paper, Tripodi IJ. et al 2019 [169], implements a machine learning-based prediction algorithm that combines knowledge from human biology, biochemistry, and toxicology data to predict the canonical mechanism of toxicity of certain chemicals and suggests possible mechanisms for those less studied. The second paper, Rubin J. et al 2021 [150], introduces Transcription Factor Enrichment Analysis (TFEA). TFEA is a computational method that detects positional motif enrichment associated with changes in transcription between two different conditions. Both papers were led by other, now graduated, graduate students in the Dowell laboratory.

C.1 Applying knowledge-driven mechanistic inference to toxicogenomics

The work described in the section is described in detail in the publication:
Tripodi IJ; Callahan TJ; **Westfall JT**; Meitzer NS; Dowell RD; Hunter LE. Applying knowledge-driven mechanistic inference to toxicogenomics. Toxicology in Vitro 2020; Volume 66, 2020, 104877. doi: 10.1016/j.tiv.2020.104877

C.1.1 Contributions

The Tripodi paper describes a collaboration between the Dowell laboratory at the University of Colorado Boulder and the Lawrence Hunter laboratory at the University of Colorado School of Medicine (Anschutz). Ignacio Tripodi, a graduate student that was co-mentor in the Dowell and

Hunter laboratories, was the lead for this project. He developed the mechanistic inference framework (MechSpy) and wrote the software and the majority of the manuscript. Ignacio identified some compounds that the method had predicted canonical toxicity mechanisms in specific tissue types and at different doses. I was tasked with verifying these predictions.

I designed three independent experiments that targeted mitochondrial-mediated toxicity, oxidative stress, and caspases homeostasis disruption. Each experiment selected cell types that closely resemble the target tissues such as the use of HUH-7 or Hep-G2 to explore compounds that act on hepatocytes found in liver tissue or A549 cells to capitulate lung tissues. The experimental designs also took into consideration doses and duration of exposure. To investigate mitochondrial-mediated toxicity, I used the MITO-ID Membrane Potential Detection Kit (Enzo Life Sciences Ref. ENZ-51018) which measures the mitochondrial membrane potential (MMP) in live cells. I used CellROX (Thermo Fisher Ref. C10444) to measure oxidation by reactive oxygen species (ROS) in live cells after exposure to a compound. Lastly, for caspase-mediated apoptosis, I used the Caspase 3 (Cleaved) Human ELISA Kit (Thermo Fisher Ref. KHO1091) to quantify the level of Caspase 3 in live cells after exposure to a compound. See the published manuscript for complete details.

C.2 Transcription factor enrichment analysis (TFEA) quantifies the activity of multiple transcription factors from a single experiment

The work described in the section is described in detail in the publication:
Rubin JD; Stanley, JT; Sigauke, RF; Levandowski CB; Maas ZL; **Westfall JT**; Taatjes DJ; Dowell RD. Transcription factor enrichment analysis (TFEA) quantifies the activity of multiple transcription factors from a single experiment. Communications Biology volume 4, Article number: 661 (2021) doi: 10.1038/s42003-021-02153-7

C.2.1 Contributions

The work described in this section was a collective work in the Dowell and Taatjes laboratories. Jonathon Rubin, a graduate student that was co-mentor in the Dowell and Taatjes laboratories,

spearheaded this project, developing the TFEA computation method and writing the majority of the manuscript. Additionally, Dr. Jacob Stanley developed muMerge, a statistically principled method of generating a consensus list of ranked regions of interest (ROIs) from multiple replicates and conditions. The ROIs were fed into the TFEA method to return the positional motif enrichment. Zach Maas contributed to the TFEA code throughout to improve the stability of the mathematical computation. Cecilia Levandowski generated PRO-seq for the paper. Rutendo Sigauke tested the beta version of TFEA on perturbation datasets and performed an analysis to determine how the muMerge regions overlap TF ChIP-seq data. Rutendo additionally implemented time-series analysis for TF enrichment over time.

My role in this endeavor was to assist in communicating the methodology. I generated the graphics for the publication for both TFEA and muMerge. I also alpha-tested both methods and provided user feedback. See the published manuscript for complete details of the method.