RNA Turnover and Trisomy: Inferring RNA Decay Rates Throughout the Interferon Response in Down Syndrome

by

Samuel Hunter

B.A., University of Utah, 2016

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, and Developmental Biology

2023

This thesis entitled:
RNA Turnover and Trisomy: Inferring RNA Decay Rates Throughout the Interferon Response in Down Syndrome
written by Samuel Hunter
has been approved for the Department of Molecular, Cellular, and Developmental Biology

Robin Dowell

Prof. Justin Brumbaugh

Prof. Dylan Taatjes

Date ____________________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Hunter, Samuel (Ph.D., Molecular Biology)

RNA Turnover and Trisomy: Inferring RNA Decay Rates Throughout the Interferon Response in Down Syndrome

Thesis directed by Prof. Robin Dowell

Trisomy 21 is a genetic abnormality referring to the presence of an extra copy of chromosome 21. This increase in DNA copy number has drastic consequences on the production and regulation of RNA. These consequences ripple throughout the genome, leading to an aberrant interferon response. These "interferonopathies" have been well characterized in steady state assays; however, it is unknown whether differences in RNA production and turnover dynamics exist in the trisomy 21 interferon response. Difficulties in answering this question are threefold. First, in comparison with their steady state counterparts, nascent RNA production assays (such as GRO-seq and PRO-seq) are challenging and prone to technical noise. Second, differential analysis of these assays in trisomic backgrounds requires additional statistical considerations. Lastly, high throughput RNA turnover assays utilize modified RNA-seq protocols with toxic metabolic labels, which can confound results. Thus, in this study, I first benchmarked and demonstrated technical signatures left by protocol variations in nascent sequencing library generation. Next, I developed a modified pipeline for differential analysis when using trisomic libraries, which properly account for the trisomic condition and normalizes read counts accordingly. Finally, I integrated this knowledge into an analysis of PRO- and RNA-seq libraries generated from a time series of interferon treatment in trisomic and disomic lymphoblastoid cell lines. From this study, I developed a computational method to infer RNA degradation rates across time.

Dedication

To my family. Thank you for your unending support in every one of my endeavors.

Acknowledgements

Special thanks to Dr. Mary A. Allen for the initial design, generation, and analysis of the nascent sequencing experiments used throughout my studies. Thanks to Dr. Chi Zhang for his contributions to the time series analysis and initial RNA degradation experiments. Thanks to Dr. Jacob Stanley for his work and instruction on degradation rate modeling and estimation. Further thanks to Paul Hartley at the University of Reno Nevada Genomics Center for providing counsel on the generation of nascent sequencing libraries, and the Anschutz Genomics Center for sequencing RNA-seq libraries. Thanks to all members of the Dowell Lab for their endless support in conducting experiments, proofreading manuscripts, and providing extensive feedback throughout the course of my education.

Contents

Chapter

Chapter 1

Introduction

Transcriptional regulation in the human genome is a balancing act between gene template availability, RNA production rates, and the controlled destruction of existing RNA. Disruptions in these systems are increasingly linked to a variety of human disorders[63, 22, 82]. Aneuploidy represents an extreme case of this genomic dysregulation. In individuals with Down syndrome (Trisomy 21), the presence of an extra copy of chromosome 21 can lead to severe autoimmune disorders and other aberrant immune responses. These aberrations manifest on a molecular level, with observable transcriptional dysregulation genome-wide[118].

With the advent of next-generation sequencing technology, researchers have generated dozens of profiles of trisomic cells and their response to various stimuli. However, interpreting these results and thus characterizing the transcriptome in aneuploidy poses a unique set of challenges. One would expect the presence of an extra chromosome to upregulate the level of transcription proportional to DNA dosage, although there are varying accounts on whether this trend is ubiquitous[50, 56]. The increase in RNA over chromosome 21 genes also shifts expected yields in sequenced reads over these regions, and thus how these counts should be treated in modeling and statistical analysis, leading to disagreements about expression levels across the field[41].

Gene dosage dysregulation causes widespread shifts in the RNA levels of immune-related genes. Indeed, tremendous resources have been directed towards researching Down syndrome as an immune disorder, due to the increased levels of interferon-relevant gene transcription. However, it is unclear whether T21 also affects immune-related RNA degradation rates and the global transcription equilibrium. As such, special consideration must be taken for each step of the experimental pipeline: What biases in sequencing library preparation could influence the observed signal? How does aneuploidy affect differential expression algorithms? And finally, after these technical aspects have been accounted for, how does aneuploidy disrupt RNA regulatory pathways of trisomic samples in response to interferon?

In particular, this final question is multifaceted; the steady-state level of RNA for any gene is influenced by two contributing factors, the rate of transcription and the rate of degradation. In steady-state conditions, previous research has established that transcription rates in conjunction with total RNA levels can be used to infer per-molecule degradation rates[15]. However, this model is not necessarily true in non-steady-state conditions; by definition, total RNA levels have shifted across a period of time when transcription rates and degradation rates have fallen out of equilibrium, necessitating new measurements of both. Measuring transcription and degradation directly can be difficult, especially in a dynamic-state cell. In particular, previous research relied upon nucleotide analogs to measure degradation, some of which trigger a cellular response themselves. Such protocols also fail to capture RNA species with low transcription rates[102].

Here, I aim to show that the steady-state solution could be expanded to a dynamic setting: knowing total RNA levels for each gene and their transcription rates across time allows one to infer the per-molecule degradation rate per unit time. If an accurate rate is to be calculated, this experiment necessitates a dense time-series analysis, where matched RNA- and transcription-measuring libraries are generated. Additionally, for each of these datasets, protocol-specific biases need to be accounted for, to avoid interpreting systemic noise as part of a change in biological signal. To allow comparisons between disomic and trisomic cells, the analysis needs to be aware of the ploidy number of the gene, as the presence of aneuploidy can result in technical artifacts in differential expression algorithms[41].

I thus took the following approach: First, run-on protocols required a rigorous comparison between different isolation steps and library preparations to determine which parts of the signal were influenced by systemic noise. As such, I analyzed potential differences inherent to these protocols.

Next, differential expression algorithms are not suited for aneuploidy by default. As such, I worked to develop a “best practices” pipeline for differential analysis. Finally, once these technical factors were accounted for, I sought to develop a differential equation, utilizing transcription rates and total RNA levels in interferon-treated cells across time to infer changes in per molecule degradation rates. Thus, I generated RNA-seq and PRO-seq data in a time series for both disomic and trisomic cells. For the remainder of this chapter, I will provide background information for the interferon response pathway, RNA degradation, and Down syndrome, and will summarize the research that has gone into these topics.

1.1 Prevalence and Health Consequences of Down Syndrome

Down syndrome is the most common aneuploidy observed in the human population, affecting up to 1 in 700 births in the United States [7]. Down syndrome is caused by a total or partial triplication of chromosome 21, leading to increased dosage of 233 protein-coding gens and 423 non-protien coding genes[7]. Individuals with Down syndrome are afflicted with characteristic physical and mental disabilities, such as congenital heart defects, increased risk for Alzheimer’s disease, intellectual disability, and immune dysregulation. A recent review by Antonarakis et al summarizes the extent of these disorders well[7]. In the wake of the COVID-19 pandemic, researchers noted that individuals with Down syndrome who contracted COVID-19 had worse health outcomes than the population[32]. Perhaps paradoxically, these worse health outcomes happen in spite of the upregulation of several antiviral genes located on chromosome 21[29].

Because of the complex phenotypes associated with aneuploidy, it is difficult to attribute disease phenotypes to any single gene. As an example, late-stage COVID-19 infection induces Acute Respiratory Distress syndrome (ARDS) significantly more often in individuals with Down syndrome; however, sleep apnea and obesity (two disorders commonly associated with Down syndrome) are also risk factors for ARDS. Furthermore, as previously mentioned, many genes involved in immune system regulation are triplicated in Down syndrome, which may also contribute to ARDS risk. Separating the interactions of each relevant chromosome 21 gene from their downstream effectors and behavioral traits of the individual, is a daunting task. Recent efforts have thus focused on the primary interferon-specific responses within the innate immune system.

1.1.1 Early Responses in Interferon Response Pathways

In 1958, Jerome Lejeune first described the link between trisomy 21 and Down syndrome. After his discovery, early research quickly linked Down syndrome and immune dysregulation. For example, as early as 1969, researchers had established that Down syndrome was associated with autoimmune inflammation of the thyroid gland[1]. Recent research has further underlined this link: the advent of next generation sequencing and genome annotation has revealed that 4 of the 6 interferon receptors reside on chromosome 21[118, 57, 73], implicating these regions in Down syndrome immune dysregulation. As such, Down syndrome has been described as an interferonopathy. The molecular basis of Down syndrome-induced interferonopathy and its clinical implications have recently been reviewed by Malle et al[72].

The interferon (IFN) receptors convey cell-to-cell signals via the JAK/STAT pathway. In short, Type-I IFN binds to the IFNAR1 and IFNAR2 receptors, causing the receptors to dimerize and activate the associated kinases JAK1 and TYK2. The transcription factors STAT1 and STAT2 can summarily be phosphorylated. The phosphorylated forms interact with IRF9, to form the ISGF3 complex (see Fig. 1.1), which binds to DNA to activate the transcription of interferon-stimulated genes (ISGs)[88]. Different subtypes of IFN can result in the activation of other transcriptional pathways as well, such as the IFN-gamma activated pro-inflammatory gene set[118, 20]. In general, these pathways work in concert to shift the cell to an antiviral state, regulating cell cycle and apoptotic pathways, and stimulating the production of other antiviral and pro-inflammatory cytokines.

The antiviral gene products work to detect and silence viral agents within the cell. For example the Mx1 and Mx2 ISGs restrict viral genome transcription and replication, preventing the formation of viral RNA in the cell[123]. The OAS proteins activate RNase L to degrade all RNA in the cell upon detection of viral genetic material. Recently, a slew of miRNAs have also been identified as important factors within the immune system, which have both cellular and viral targets.

In particular, miR-155 and miR-125b, two such miRNAs associated with B-cell development, have previously been shown to be dysregulated in individuals with Down syndrome, leading to reduced B-cell subpopulations[33]. Altogether, these previous findings establish the possibility that Down syndrome has unique RNA dynamics throughout the interferon response pathway due to changes in DNA dosage.

1.1.2 Dosage Dependency in Disease

Gene dosage differences serve as the primary hypothesis to explain the disease phenotypes of Down syndrome. This hypothesis has been supported through the use of mouse models with partially triplicated chromosomes, which seek to isolate gene dosage effects to individual or subsets of chromosome 21 genes. Indeed, these dosage effects have been demonstrated for many genes of interest, including the interferon receptors[127, 18, 98]. As an example, Waugh et al. utilized the Dp16 mouse model, which has a segmental duplication of 120 genes including the interferon receptors. Dp16 mice exhibit many of the phenotypes of humans with Down syndrome, including mild interferonopathy and a more severe response to viral infection and viral genome mimetics. These phenotypes were rescued when one copy of each of the triplicated interferon receptors were knocked out, indicating the dosage dependency of these disorders[127]. Furthermore, this knockout improved mouse cognition and lessened the prevalence of congenital heart defects, a testament to the interactive nature of these genes beyond the interferon response.

However, owing to the complex nature of gene-gene interactions, gene dosage often cannot alone explain the observed disorder. Research from Malle et al highlights one such example within the interferon response pathway: while individuals with Down syndrome are more likely to have severe health consequences from viral infection, they are less likely to contract many viral infections in the first place[73]. Indeed, initial upregulation of the interferon genes leads to additional transcription of the negative regulator genes *USP18*, and *SOCS1* (located on chromosome 8 and 16, respectively). While, upregulation of interferon genes via increased receptor dosage logically renders the cell in a mildly antiviral state, Malle et al. contend that the downstream upregulation of the negative regulators renders the cell refractory to subsequent activation, resulting in unique vulnerability to viral propagation if the cell does get infected[73].

Furthermore, posttranscriptional regulation of RNA adds yet another wrinkle to the gene dosage hypothesis. While RNA may be produced at 1.5 times disomic levels (e.g. at DNA dosage levels), RNA degradation may also be shifted based on downstream effectors or cellular signals. Similar to shifts in transcription, shifts in degradation have rippling effects. Indeed, regulating RNA degradation is a pivotal step in properly timing cellular responses, including within the interferon response pathway[93].

1.2 RNA Degradation Regulation

RNA degradation is a multifaceted process which can be regulated through a myriad of cellular mechanisms. Broadly, degradation proteins can be classified into three categories: $5'$ decay factors, $3'$ decay factors, or endoribonucleases. Along with the combined activity of these three classes, RNA decay rates are further modulated by RNA modifications, chaperone proteins, protein phosphorylation, and sequence motifs. Each of these elements work in tandem to modulate the RNA's half-life[47, 36]. In this section, I will briefly review each of these classes, and some important factors which modulate their rates in the context of Down syndrome and the interferon response.

$5'$ Degradation: Enzymes which degrade starting at the $5'$ end are known as $5'$ exonucleases. These enzymes must contend with modifications at the $5'$ end. Specifically, mRNAs are protected by a m^7G cap, which shields the molecule from degradation from this class of nucleases[36].

The $5'$ exonuclease Xrn1 works in conjunction with decapping enzymes (Dcp1/Dcp2) and their associated activators to modulate the rate of degradation. When the mRNA is undergoing translation, the ribosome and its associated initiation factors block these enzymes, reducing the degradation rate of the RNA. Conversely, the miRNA-associated Argonaute proteins (Ago1-4) are known to recruit these decapping factors to stimulate degradation[83]. Interestingly, the *Xrn1* gene is upregulated during the antiviral response of the cell, and, perhaps paradoxically, has been shown to target some antiviral RNAs in the cell for degradation[67, 60]. It is unknown whether this trend

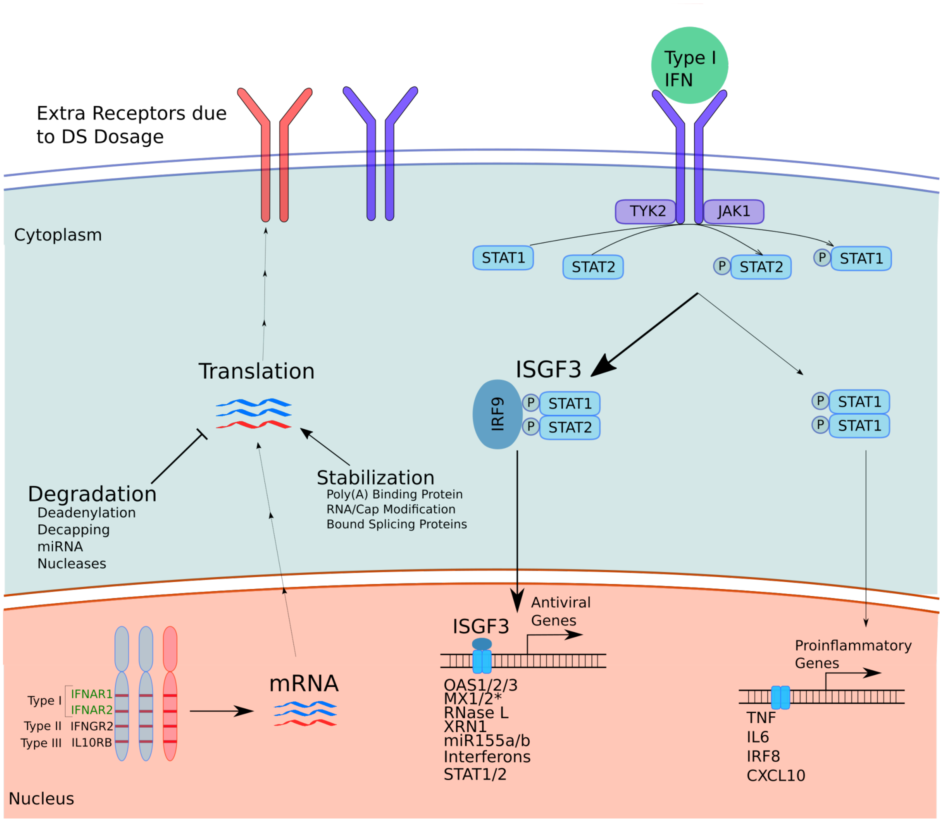


Figure 1.1: **Cartoon depicting interferon dysregulation in Down syndrome and RNA degradation modeling.** An extra copy of chromosome 21 results in the upregulation of several interferon genes, most critically the four interferon receptors. The additional mRNA results in additional receptors on the cell surface. The IFN response is mediated by the JAK/STAT pathway; specifically, type-I IFN binding results in the dimerization of the IFNAR1 and IFNAR2 receptor, which activates associated JAK1 and TYK2 kinases. These kinases phosphorylate STAT1 and STAT2. The STAT proteins then interface with each other as a transcription factor, governing the regulation of many downstream genes. In general, the type I interferon response is associated with ISGF3 activation and the antiviral response. Alongside the upregulated IFN receptors, chromosome 21 contains several other IFN-response genes (indicated with an asterisk). It is unknown whether this dysregulation is also detectable in RNA degradation pathways. Figure modified from Kim et al[55].

persists in Down syndrome as well.

3′ Degradation: The other classification of exonuclease instead targets the 3′ end for degradation. Mature mRNAs are protected by a ∼250 bp poly(A) tail, which is bound by stabilizing poly(A)-binding proteins (PABPs). These PABPs prevent degradation while they are bound; thus, the poly(A) tail must be removed via deadenylation for degradation to occur. The associated deadenylation rate is closely linked to the RNA's half-life[85]. As with decapping enzymes, miRNA-associated proteins recruit deadenylation enzymes (CNOT1-11, poly(A)-specific ribonucleases, and CAF1) to accelerate this process.

Importantly, outside of miRNA binding motifs, there are canonical motifs within the 3′ end that further dictate degradation rates. AU-rich elements are a known pattern which increases the degradation rate of the RNA by acting as a landing pad for destabilizing proteins, most importantly the exosome complex [119, 114]. The exosome complex is a large protein complex made up of several RNases and RNA-binding proteins (Rrp4, Rrp6, Rrp40-46, Csl4, and Mtr3) which target a wide selection of mRNAs for degradation. This action is blocked by PABP occupancy. Alternative splicing can result in more stable mRNA transcripts due to changes in the sequence composition of the 3′ UTR.

Outside of these constitutive decay factors, there are several 3′ exonucleases which are upregulated upon interferon treatment. For example, the ISG20 exonuclease acts primarily in an antiviral capacity, targeting viral RNA for degradation while sparing cellular RNA. Conversely, the interferon-stimulated gene product ZAP acts as an exonuclease for mainly viral transcripts, but also targets the anti-apoptotic cellular mRNA TRAILR4[120]. In both cases, these exonucleases are upregulated in Down syndrome, although the consequences of this upregulation on RNA half-life have not been investigated.

Endonuclease Cleavage: This last classification of degradation enzymes catalyze the hydrolysis of RNA from within the middle of the molecule. As the middle of the RNA is not protected by a cap or by a poly(A) tail, the activity of these molecules is mediated by protein binding, ribosome occupancy [85], and by activation of the endonuclease through other cellular signals. Within the interferon response, for example, the endonuclease RNase L is responsible for the global degradation of both host and viral RNA. However, it is translated as an inactive monomer if no virus is present within the cell. The activity of the OAS proteins (OAS1/2/3) catalyzes the dimerization of RNase L when double-stranded RNA is detected.

As with other interferon-stimulated genes, RNase L is upregulated in Down syndrome, both basally and in the interferon response. However, RNase L is inactive until viral dsRNA is encountered within the cell[66]. It is currently unknown whether this higher basal level and increased activation has consequences within the early interferon response to foreign invaders.

Each of these forms of degradation regulation can work alone or in concert with each other to precisely modulate the half-life of RNA within the cell[85]. While attributing changes to RNA half-life to any one of these components is difficult, measuring overall changes in half-life to each RNA species in the cell has been made possible through the advent of next-generation sequence technology. In the next section, I will address the canonical models and methods used to estimate RNA half-life.

1.2.1 Transcription Inhibition, Metabolic Labeling, and Tracking RNA

Early methods for measuring RNA degradation involved separating newly produced RNA from the initial population, and then measuring the decrease in the initial population over time. In early sequencing experiments, this was accomplished using a transcriptional inhibitor (for example, triptolide, actinomycin D, or flavopiridol)[44]. Once transcription is no longer occurring, the remaining population of RNAs decay over time; as such, RNA-seq libraries generated at different time points can be used to observe the decaying population across time. The RNA-seq signal over high-turnover RNA species will deplete quicker than the signal over more stable genes, giving a readout of relative degradation rates. Naturally, however, the use of a transcriptional inhibitor had its own background effects, including inducing the stress response and up-regulating some specific RNAs[14]. Additionally, this method decouples transcription and degradation, a situation which is not representative of the normal function of the cell.

Newer methods involve a metabolic label (e.g., a biotinylated nucleotide), introduced to the growth medium as a pulse before the experiment, which readily incorporates into the RNA as it is transcribed. The amount of labeled RNA can then be compared across time to estimate RNA half-lives, or time points when degradation rates unexpectedly change[102]. Unlike transcriptional inhibition, this method is less intrusive to the normal function of the cell, although there are some reports of cytotoxicity from different metabolic labels[35]. Notably, however, this method is sensitive to both the labeling time and concentration of the normal label. For lowly transcribed RNA, this poses a detection issue, as there may not be enough label to accurately track these transcripts across time. This is especially prevalent with noncoding RNAs, which tend to be both less abundant and less stable than their mRNA counterparts.

Lastly, the final set of methods do not directly measure RNA decay at all, but instead infer these rates by utilizing other types of high-throughput sequencing data (namely, steady state RNA-seq and nascent transcription assays). In the next sections, I will briefly describe key characteristics of these data types. Afterwards, I will describe how they can be used in conjunction with each other to infer RNA decay rates.

1.2.2 Run-on Sequencing and Steady-State Sequencing

RNA-seq library generation requires sampling a fraction of the total, steady state RNA within a population of cells. As such, the signal retrieved from this process does not capture every RNA species at once. Low abundance or transient RNAs are especially difficult to detect in these libraries. Thus, researchers have recently created new protocols to specifically sample from these more specific RNA populations of interest within the total pool of RNA. Of particular note here is the development of run-on sequencing protocols, which sample only the RNA which is actively being built by a polymerase engaged with the gene template[24, 71].

These run-on sequencing (RO-seq) protocols utilize a nucleotide analog which readily incorporates into the RNA being transcribed. This RNA makes up only a minuscule fraction of the total population (estimated at less than 1%[24]). The population of RNA bearing the labeled nucleotide can be enriched during subsequent library preparation steps. While this process is much more labor intensive than RNA-seq, RO-seq libraries are able to pick up a much more specific set of RNAs within the cell. As such, the signal in these libraries is dominated by the nascent, unspliced RNA fraction.

Comparing RNA-seq and RO-seq libraries shows stark contrasts in the read distribution throughout the genome. For one, RNA-seq signal is primarily from spliced, genic RNA, such that most reads pile up over exons. In RO-seq libraries, however, unspliced RNA dominates the signal, such that reads are dispersed throughout the gene body. Additionally, many more intergenic transcription sites are detected in RO-seq libraries. These two library types can also be used in conjunction with each other to model RNA turnover within the cell. In the next section, I will explore one such model.

1.2.3 The Bathtub Model

While RNA transcription and degradation are multifaceted processes with many different components, the entire scope of RNA turnover for a given gene can be conceptualized with the relatively simple “bathtub” model. In this model, the “water level” represents the total amount of RNA in the cell. This “water level” is modulated by two different sources: the rate of RNA production (the “faucet”) and the decay rate of the RNA (the “drain”). This simple representation summarizes the relationship between RNA levels, RNA production rates, and RNA degradation, and allows for the inference of any one of these values when the other two quantities are known.

One useful implementation of this model involves another assumption: if no net change in the total level of RNA is occurring, then the rate of RNA decay and RNA production must be equal. In mathematical terms, for a given gene i, the relationship between transcription and degradation can be described using the equation $M_i\alpha_i = B_i$, where the gene’s transcription rate is represented as B_i, its RNA steady state levels as M_i, and its per-molecule degradation rate as a_i. The gene’s total degradation rate is thus a function of total RNA levels ($M_i\alpha_i$). Critically, by rearranging the equation, the steady-state assumption allows for the calculation of α_i by finding the ratio of B_i and

M_i (Fig. 1.2).

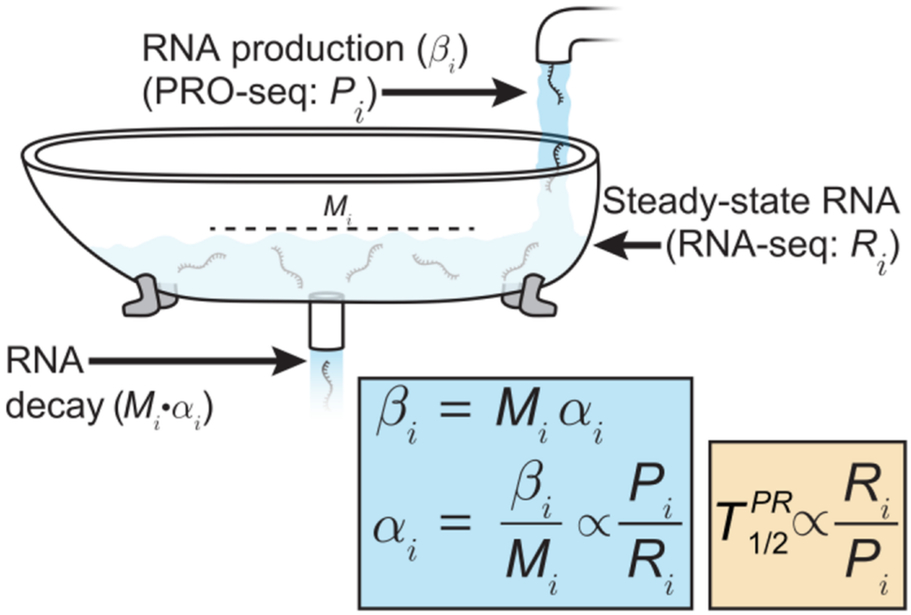

RNA production (β_i)
(PRO-seq: P_i)
M_i
Steady-state RNA
(RNA-seq: R_i)
RNA
decay ($M_i \cdot \alpha_i$)
$\beta_i = M_i \alpha_i$
$\alpha_i = \frac{\beta_i}{M_i} \propto \frac{P_i}{R_i}$
$T^{PR}_{1/2} \propto \frac{R_i}{P_i}$

Figure 1.2: **Cartoon depicting the Bathtub model of RNA turnover.** Summary diagram describing the "bathtub" model of RNA degradation. The turnover of RNA from any gene i can be broken down into three major steps: RNA production, RNA steady-state, and RNA degradation. Degradation can further be broken down into a per-molecule degradation rate (α_i) multiplied by the steady-state level of RNA (M_i). If no net change in RNA steady-state levels is occurring, RNA production (B_i) is equal to RNA degradation. As such, the per molecule degradation rate can be solved for. RNA-seq (R_i) and PRO-seq (P_i) levels can be used as stand-in values for M_i and B_i, respectively, allowing for a proportional estimate of the RNA's half-life $T_{(1/2)}$ by finding the ratio of R_i and P_i. Figure adapted from Blumberg et al[15].

Of course, this model theorizes an exact knowledge of transcription rates and steady-state levels. In reality, substitute values which approximate these quantities must be used. Previous studies have utilized a variety of stand-in approximates to accomplish this task. For example, Gaidatzis et al utilized intronic reads as a proxy for transcription rates, thereby enabling transcription and degradation estimates from total RNA-seq datasets, albeit with considerable bias towards genes with changing transcription rates[37]. Later, these estimates were improved by first determining the bias associated with the transcription rate of the gene[3]. Eventually, Blumberg et al circumvented this issue entirely by instead using PRO-seq signal as a proxy[15]. In each case, however, it is important to note that all of the quantities for the bathtub model are proxies built from sequencing data; as such, the per-molecule degradation values for each gene are proportional to their true rates, and can only be interpreted relative to each other.

On its face, the method proposed by Blumberg et al seems a good fit for investigating degradation throughout the interferon response in Down syndrome. However, this method faces several limitations when investigating a time series in these conditions. In the next section, I will cover these challenges in greater detail, and my strategies to overcome them.

1.3 Challenges in Studying RNA Turnover in Down syndrome

1.3.1 Protocol Variations in High-Throughput Sequencing Data

As more and more protocols enter the field of transcriptomics, integrative approaches run the risk of false conclusions due to an increase in technical noise from protocol variation. As an example, RNA-seq protocols come in two major flavors: poly(A)-tail selection, and ribosomal RNA (rRNA) removal[100, 23]. Because rRNA makes up the overwhelming majority of the RNA population in the cell, both protocols seek to enrich above this background rRNA signal. Poly(A) selection methods enrich specifically for coding mRNAs which have been polyadenylated, while rRNA removal methods seek to be less biased against other noncoding RNAs by instead directly degrading rRNA before library preparation. As such, comparisons of poly(A)-selected datasets to rRNA-depleted RNA-seq data will show a relative depletion of long noncoding RNAs, complicating comparative analyses which use both data types. Such differences have been extensively catalogued for RNA-seq protocols; however, no such comparison existed for nascent transcription protocols. Thus, before integrating nascent and steady-state assays into one comparison, I sought to first establish which technical considerations may affect the signal in nascent transcription protocols[49] (described in detail in Chapter 2).

1.3.2 Differential Analysis in Aneuploidy

Differential analysis pipelines are built with the explicit assumption that the majority of genes are not changing[69]. Logically, this assumption holds true when comparing trisomic samples to disomic samples; however, the additional information afforded by knowing the ploidy of the samples can be utilized to strengthen hypothesis testing. For example, the popular differential analysis program DESeq2 primarily weighs log-fold-change estimations of all genes towards 0 as part of the assumption that most genes do not change – implicitly assuming all genes exist at equal copy number (a disomy assumption). Without leveraging ploidy information, this assumption leads to inaccurate fold change estimations in trisomic backgrounds. These estimates have consequences when interpreting the results within their biological context; critically, a reduction in fold change estimations suggests that a subset of trisomic genes are dosage compensated back towards disomic expression levels.[56]. As such, before degradation analysis, I sought to adapt differential analysis pipelines to properly account for trisomy, and to question whether dosage compensation is truly occurring in Down syndrome or if previous reports were the result of a computational artifact (described in detail in Chapter 3).

1.3.3 Developing a Linear Model for Degradation Analysis Across Time

The steady-state approximation of RNA degradation is built off of the "bathtub" model, as previously described. Utilizing this model, degradation and transcription can be assumed to be equal to each other, such that the per-molecule degradation rate can be calculated with RNA-seq and PRO-seq alone, per the method pioneered by Blumberg et al[15]. However, this assumption is limited; if either transcription or degradation rates change across time, the steady-state assumption is violated. It is unknown how this change might affect the accuracy of degradation rate estimation at each time point.

One alternative is to instead allow for degradation and transcription to both change across time within the model. With this caveat, degradation is estimated from at least two time points using changes in both RNA-seq and PRO-seq. The principles remain the same: RNA-seq acts as a stand-in for total RNA levels and PRO-seq acts as a stand-in for the transcription rate. Degradation then can be approximated by estimating the change in RNA-seq across time, and subtracting the change associated with transcription (PRO-seq).

Using the Blumberg formulation as a point of comparison, I will show my progress in developing a linear approximation model of RNA degradation inference. Furthermore, I will demonstrate some key findings in determining changes to RNA degradation throughout the interferon response, and how RNA degradation rates differ between trisomic and disomic cells (for more details on the decay model and its conclusions, see Chapter 4). All of my findings will be summarized in the Conclusion, where I will also outline the future of these projects.

Chapter 2

Protocol Variations in Run-On Transcription Dataset Preparation Produce Detectable Signatures in Sequencing Libraries

The following chapter has been adapted from its publication in BMC Genomics, and was a collaborative effort between Rutendo Sigauke, Mary Allen, Robin Dowell, Jacob Stanley, and myself. Rutendo Sigauke, with assistance from Jacob Stanley, utilized a discrete wavelet transform to infer protocol information directly from the data. Mary Allen helped with the initial comparisons of the experiments and in the generation of several data sets. Robin Dowell assisted in coordinating the study, structuring the paper and proofreading. I performed all other analyses, generated datasets, and was the primary author for the manuscript. For more information, access the published article at: `https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-022-08352-8`

Abstract

Background: A variety of protocols exist for producing whole genome run-on transcription datasets. However, little is known about how differences between these protocols affect the signal within the resulting libraries.

Results: Using run-on transcription datasets generated from the same biological system, we show that a variety of GRO- and PRO-seq preparation methods leave identifiable signatures within each library. Specifically we show that the library preparation method results in differences in quality control metrics, as well as differences in the signal distribution at the $5'$ end of transcribed regions. These shifts lead to disparities in eRNA identification, but do not impact analyses aimed

at inferring the key regulators involved in changes to transcription.

Conclusions: Run-on sequencing protocol variations result in technical signatures that can be used to identify both the enrichment and library preparation method of a particular data set. These technical signatures are batch effects that limit detailed comparisons of pausing ratios and eRNAs identified across protocols. However, these batch effects have only limited impact on our ability to infer which regulators underlie the observed transcriptional changes.

2.1 Background

The transcriptome dictates much of a cell's identity and behavior. As such, tracking how transcription patterns change in response to a biological perturbation is a popular approach to understanding molecular regulatory mechanisms. In particular, newly transcribed RNAs provide a readout on the activity and regulation of cellular polymerases. Capturing and mapping these "nascent" transcripts provides a single base-pair resolution readout of the positions of all cellular RNA polymerases throughout the genome[24, 59, 71]. Changes in RNA polymerase behavior are associated with transcription factor activity[4, 11, 97], with a large portion of the changes occurring within enhancer regions. These enhancer RNAs (eRNAs) are unstable and thus not generally recovered by steady-state assays such as RNA-seq, which sample predominantly from the pool of stable transcripts such as mRNAs[96].

To capture all RNAs arising from cellular RNA polymerases, several run-on transcription capture protocols, such as global run-on sequencing (GRO-seq) and precision run-on sequencing (PRO-seq), have been developed[24, 124, 71, 53, 13]. These protocols, collectively known as RO-seq, follow roughly a two step process: first, the run-on RNA signal must be enriched above the background total RNA; second, the captured RNA is then converted into a sequencing-ready cDNA library[24]. For the first step, run-on protocols share the same basic strategy, namely use an enrichable nucleotide as a handle for distinguishing nascent RNA from previously produced RNA (Fig. 2.1A). Subsequently, sequencing adapters are added and the sample is reverse transcribed and amplified in preparation for sequencing. As these steps are somewhat modular, the process of enrichment is often interleaved with the various steps necessary for library preparation (Fig. 2.1B).

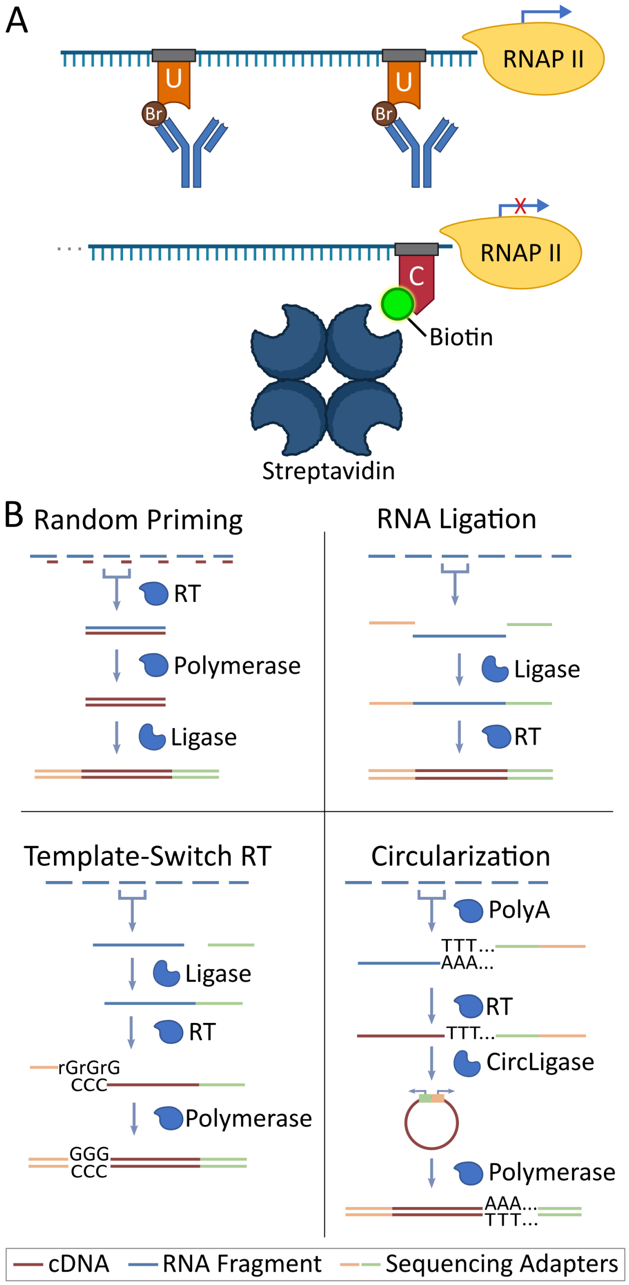


Figure 2.1: **Summary of Run-On Sequencing (RO-seq) data sets.** (A) Summary diagram indicating enrichment steps for Global Run-On (GRO-seq, top) and Precision Run-On (PRO-seq, bottom) reactions. (B) Summary diagram for library preparation reactions. Blue bars: RNA; brown bars: cDNA; yellow/green bars: sequencing adapters. Library preparation enzymes are labeled and represented by blue shapes at each step.

Similar to distinct RNA-seq library preparation methods, processing RNA through different

RO-seq protocols is thought to leave technical artifacts within the library[106, 100, 126]; however, the extent to which these artifacts influence the resulting analysis has not been thoroughly explored. In this study, we sought to identify specific signatures and biases inherent to the protocol (enrichment strategy) and library preparation methods typically employed in RO-seq methods. For this comparison, we generated data from HCT116 cells treated for 1 hour with the p53 activator Nutlin-3a or a DMSO control, a well studied perturbation[4, 6]. Using these matched datasets, we find specific and reproducible biases in each respective dataset that influence both the quality metrics and $5'$ distribution of reads. However, we find that these protocol and library specific effects do not strongly impact the inference of which transcription factor is driving the observed perturbation induced changes in transcription. These protocol-specific signals could enable an agnostic detection program to identify the protocols used; such programs could then be utilized to increase the validity of online sequence databases.

2.2 Results

2.2.1 Quality metrics are influenced by RO-seq transcription capture protocols

The ultimate goal of run-on protocols is to produce a dataset that accurately reflects the distribution of actively transcribing RNA polymerase [24, 2] genome wide. However, success in this endeavor depends greatly on the sequencing depth, library complexity, quality of enrichment, and transcription strength of the cell line [94]. To control for cell line differences, we generated run-on libraries from HCT116 cells using a previously employed perturbation strategy[4, 6]. Namely, we used global run-on (GRO) sequencing[24] with a Br-tagged UTP, and precision run-on (PRO) sequencing[59] with a Biotin to mark CTP [71] (Fig. 2.1A) as enrichment protocols. We then combined these enrichment protocols with one of four library preparation techniques: RNA adapter ligation (LIG)[24], Circularization (CIRC)[124], Random Priming (RPR)[84], or Template-Switching Reverse Transcription (TSRT)[53] (Fig. 2.1B) after either 1 hr DMSO control or 1 hr treatment with Nutlin-3a. Nutlin-3a is a molecule which interrupts p53 inhibition and leads to rapid transcription of downstream p53 targets (see Materials and Methods). Samples were subsequently sequenced on an Illumina NextSeq 500 platform (RTA version: 2.4.11, Instrument ID: NB501447) using a single end strategy (37, 50 or 75 bp lengths) to variable depths (summarized in Supplemental Table 1, see Materials and Methods).

The first noticeable differences between any two datasets (even with the same protocol/library preparation) are depth of sequencing and complexity of the library. The depth of our samples range from 20 million to 170 million reads. We correct for the disparity in sequencing depth by combining the technical replicates of low-depth samples, and by subsampling deeply sequenced samples. As such, all subsequent comparisons were performed at equivalent depth (with a minimum of 75 million reads).

In contrast, library complexity reflects data quality and cannot be corrected for computationally and ideally would be similar between library preparations before comparison. We use two metrics to assess complexity, the number of unique reads relative to the depth of the sample and the number of unique bases covered within the genome (Supplemental Table 1). While most of our libraries were comparably complex, we found that our libraries generated with a random-priming library kit were generally of lower complexity. The random-priming strategy is rarely used and thus, it is unclear whether the tendency of reduced complexity is a consequence of the library preparation method or a fault of our handling. However, public random primed datasets exhibited similar $5'$ read distributions to our datasets in spite of the differences in library complexity (Supplemental Figs. 1,2,3, Supplemental Table 1); therefore, we chose to include these libraries in our initial analyses to showcase possible technical signatures and potential biases, but refrained from using GRO-RPR libraries in further comparative analyses.

Notably, some library preparations result in clearly distinguishable sequence signatures within the acquired reads. In circularization (CIRC) libraries, regardless of the enrichment protocol, RNA is polyadenylated before reverse transcription, and the resulting cDNA is subsequently circularized via the enzyme circLigase[124]. As such, it is common to see many reads with long poly(A) tails before trimming (Fig. 2.2A). Additionally, the TSRT library preparation adds several C nucleotides to the end of each read[53]. Upon sequencing and adapter trimming, many read inserts showed an increased incidence of C nucleotides near the end of the read (Fig. 2.2A). In our samples, these sequence signals can effectively distinguish CIRC and TSRT libraries from the other library preparation methods. In contrast, LIG and RPR libraries show similar nucleotide composition across the reads. Likewise, GRO and PRO datasets constructed with matched library preparation methods are not distinguishable from sequence content signatures alone.

However, principal component analysis (PCA) of the read counts over all genes tightly clusters based on library preparation and enrichment protocol, suggesting there are additional protocol-distinguishing features not evident in the average nucleotide composition of the dataset (Fig. 2.2B). Therefore, we next sought to identify whether enrichment quality metrics could be used to distinguish between the protocols. Quality control pipelines offer a way of quantifying steady-state RNA contamination by calculating the ratio of reads over exons and introns for each gene. While the specific value expected for this ratio depends on how reads are counted, a comparatively lower exon-intron ratio is indicative of less mRNA contamination[109]. But is this exon-intron ratio influenced by the choice of protocol? To answer this, we calculated log-normalized exon-intron ratios for every gene in each HCT116 control (DMSO) library. On average, PRO libraries showed a slightly lower amount of mRNA contamination across all genes relative to GRO libraries, consistent with the relative strength of the two enrichment strategies (Fig. 2.2C). Additionally, both CIRC and LIG libraries showed lower mRNA contamination relative to RPR libraries (Fig. 2.2D).

Sequence composition (Fig. 2.2A) can be utilized to identify CIRC and TSRT library preparation protocols with high confidence, while LIG and RPR libraries were more similar in sequence composition, albeit with some differences in complexity and quality metrics (Fig. 2.2D, Supplemental Table 1). However, the differences between the enrichment protocols (GRO vs PRO) is less readily apparent from sequence composition or quality metrics alone (Fig. 2.2A,C, Supplemental Table 1). Yet, we wondered whether systematic signals exist within the data that could distinguish between the protocols. To this end, we applied a discrete wavelet transform (DWT) approach to the normalized coverage of each library (Fig. 2.2E). The DWT decomposes the signal in a region into low

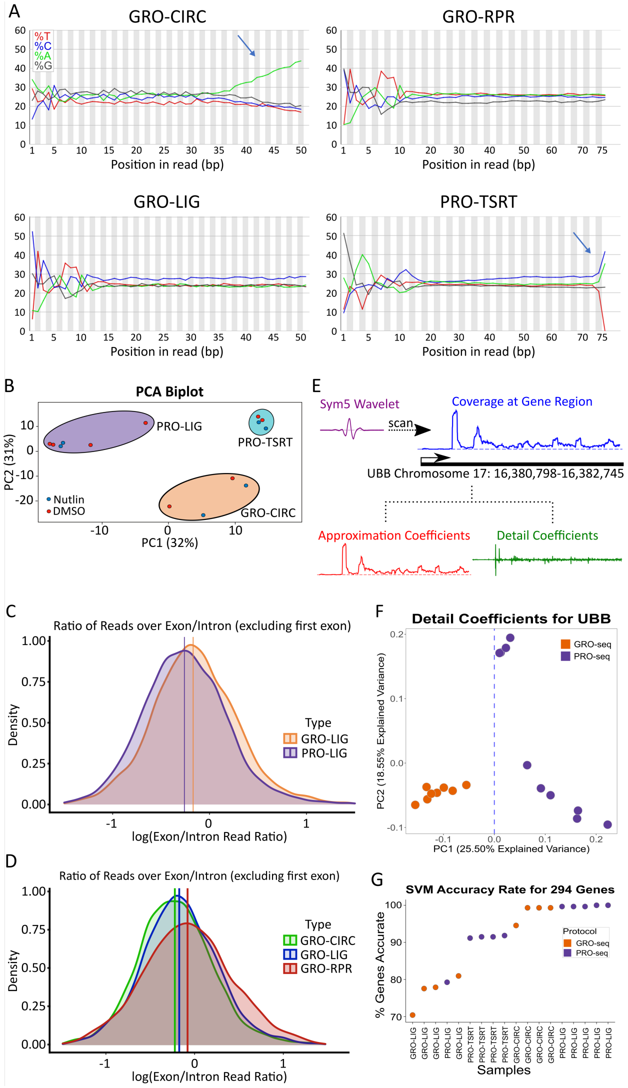


Figure 2.2: **Quality Control metrics for varying library preparation and enrichment techniques.** (A) Nucleotide distribution of DMSO samples are plotted indicating the percent nucleotide representation (y-axis) versus the position within each read (x-axis). Library specific signatures are identifiable in CIRC and TSRT libraries (blue arrows). (B) Principal-Component Analysis of assorted library preparation and enrichment methods. Each library was prepped using HCT116 cells treated with either DMSO or Nutlin-3a for 1 hour. Log-normalized density plots of exon/intron ratios for each gene for each (C) enrichment method and (D) library preparation method (GRO-seq samples shown), (GRO-LIG vs PRO-LIG: $p < .001$; GRO-CIRC vs GRO-LIG: $p < .05$; GRO-CIRC vs GRO-RPR: $p < .001$; GRO-RPR vs GRO-LIG : $p < .001$, K-S Test, n=1795). Mean indicated by vertical line for each respective distribution. (E) Schematic showing the wavelet transformation approach at the UBB locus. (F) Detail coefficients at UBB locus separates PRO and GRO libraries on PC1 (Low-biotin PRO-seq samples omitted, see Supplemental Table 1). (G) SVM classifier results for each tested library.

frequency signals (approximation coefficients) that capture consistent RNA polymerase signatures and high frequency signals (detail coefficients) that contain noise. The noise component captures both random noise and systematic noise. Because protocol specific signatures are a systematic source of noise, we reasoned that the high frequency signals may be able to distinguish between the protocols.

To test this hypothesis, we sought to evaluate the DWT on a set of genes where RNA polymerase signatures are the least influenced by library depth or complexity. Thus we identified a set of 294 highly transcribed genes that also had a low coefficient of variation across our datasets. Using the PyWavelets package in python, a symlet wavelet was scanned over the normalized coverage of each gene, effectively decomposing the signal into the two components (see Materials and Methods) (Fig. 2.2E)[27, 62]. Subsequently, we used principal component analysis (PCA) to cluster the detail coefficients. Overall, 117 genes (39.8%) separated the protocols (GRO vs PRO) directly on the first principle component whereas an additional 162 (55.1%) genes separated the protocols on a different plane within the PC1 and PC2 space (Fig. 2.2F, Supplemental Fig. 4, 5). These results suggested that the data sets contain a readily identifiable protocol signature. To confirm, we built a simple support vector machine classifier to determine whether the principle components of the wavelet analysis could be used to identify the protocol directly from the data (see Materials and Methods) (Supplemental Fig. 6). Using leave-one-out cross validation at the individual gene level, the classifier correctly identified the protocol >70% of the time (Fig. 2.2G, Supplemental Fig. 7). Furthermore, applying a simple majority rules voting scheme to the classifier results identified the protocol every time (100%), further confirming that each data set contains identifiable protocol specific signatures.

2.2.2 Enrichment and Library Preparation Methods Significantly Shift 5′ Distribution

To better understand the protocol specific signatures within the data sets, we next examined annotated, protein-coding genes for systematic differences in their read distributions. At protein-coding genes, the behavior of RNA polymerase II is well characterized[52] which leads to repeatable patterns of read distribution throughout the gene (Fig. 2.3A). Therefore, we sought to determine whether the protocol (GRO vs PRO) led to systematic differences in the detected 5′ initiation region or the elongation region. Counts across gene body regions suggested that elongation regions correlated well between protocol and library preparation differences (Supplemental Fig. 8, see also Materials and Methods); therefore, we subsequently focused our attention on the 5′ regions of genes.

To assess the differences in the 5′ distribution across protocols, we examined the read distribution of GRO and PRO libraries prepped from DMSO-treated HCT116 cells, with an otherwise similar library preparation protocol (LIG). Metagenes revealed a shift in coverage near many transcription start sites (TSS) in PRO libraries that is not present in GRO libraries (Fig. 2.3B, Supplemental Fig. 9). GRO and PRO libraries differ in the nucleotide analog used to enrich for nascent RNA. In GRO-seq, bromouridine-triphosphate is used to mark newly transcribing RNAs which can then be detected by anti-BrdU antibodies. In contrast, PRO-seq uses Biotin-NTPs which also terminate transcription upon their incorporation into the nascent RNA. Streptavidin then efficiently isolates newly transcribed RNAs. The original PRO-seq strategy marked all four nucleotides to maximize precision[24], but for cost efficiency, subsequent efforts only marked a single nucleotide[71]. Notably, both the efficiency of pull down and the termination of transcription results in PRO-seq giving a more precise readout on the position of RNA polymerases relative to GRO-seq[59]. However, at the 5′ end this precision also results in short unmappable reads, leading to gaps in coverage near the TSSs[71]. In an attempt to mitigate these 5′ read coverage gaps, subsequent variations in the PRO-seq protocol include a ratio of Biotin-NTP/NTP to the run-on mixture [71].

We theorized that the shift in the 5′ region observed in our PRO libraries arose from early incorporation of Biotin-NTP near the TSS which leads to short, truncated reads that are not well mapped. As such, we reasoned that generating new libraries with a different ratio of Biotin-NTP/NTP in the initial run-on mixture would result in more reads captured around the 5′ end (Supplemental Fig. 10). Metagenes indeed show a smaller shift with lowered Biotin-NTP concentration, although GRO-LIG libraries continued to show more signal in these regions than any

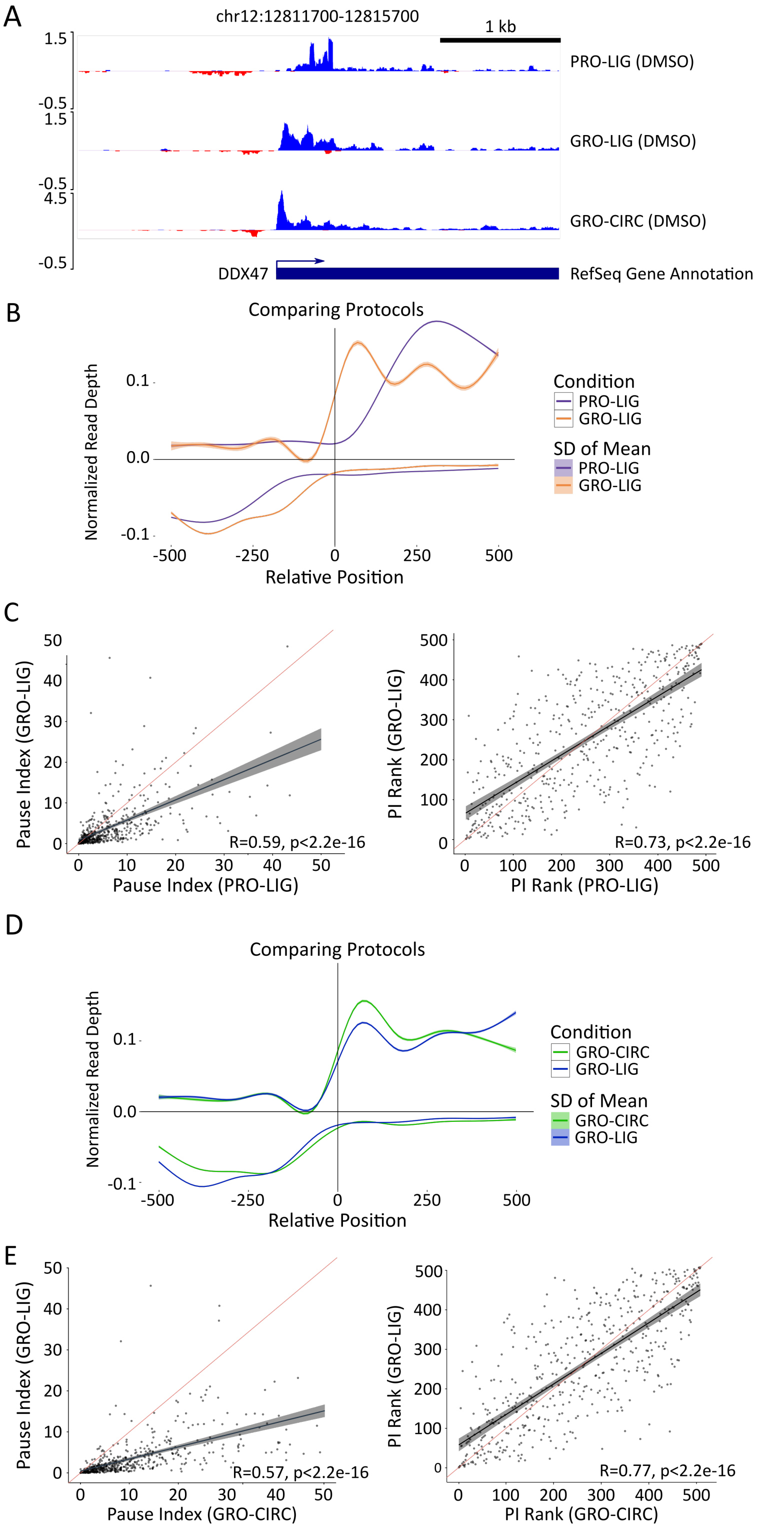


Figure 2.3: **Analysis of gene transcription start sites among different protocols and library preparations.** (A) 5′ end distribution among various library preparation and enrichment methods. Negative read depth (red) represents reads found on the minus strand. **continued ....**

Figure 2.3: **continued**(B) Metagenes constructed from GRO-seq (orange) and PRO-seq (blue) libraries (ligation library preparation, HCT116, DMSO 1hr). Genes shorter than 2000 bp, with significant signal 2 kb upstream (>1% of upstream bases covered), and with low coverage (TPM < .01) were removed (n=2527). Vertical line (0) is annotated TSS in RefSeq database, distance (x-axis) in bp, read depth normalized by counts-per-million (CPM). (C) Pausing index calculations for top 500 transcribed genes in GRO-seq and PRO-seq libraries, presented with Pearson (left) and Spearman (right) correlations (red line: y=x, black line: best fit). Pausing region defined as -50 bp to 250 bp from annotated TSS (See Materials and Methods). (D) Metagenes constructed from GRO-seq ligation (blue), and circularization-based (green) libraries (HCT116, DMSO 1 hr). Genes filtered and graphed as in (B). (E) Pausing index calculations for circularization and ligation based libraries (GRO-seq, HCT116, DMSO 1 hr), graphed as in (C).

PRO library.

To ensure that our findings generalize to other data sets, we next examined publicly available datasets. While these data sets likely have larger batch effects arising from their preparation in distinct laboratories and cell types, we reasoned that the overall trend in $5'$ end patterns should still be noticeable, albeit subject to more variance. GRO and PRO libraries obtained from other labs showed that the peak of PRO-seq libraries was noticeably further downstream than their GRO-seq counterparts; however, this comparison (using a consistent mapping and analysis strategy, see Materials and Methods) uncovered a broad range of peak positions (from +40 bps to +250 bps) with seemingly no linear relationship between the Biotin-NTP/NTP ratio and peak position (Fig. 2.3B Supplemental Fig. 11, 12, 10).

Therefore, we reasoned that there must be further underlying protocol influences on the $5'$ read distribution. Differences in size selection, read fragmentation, and gene filtering criteria were all hypothesized to influence the distribution. To evaluate these criteria, we took an *in silico* approach and simulated reads arising near a TSS from each protocol configuration (see Materials and Methods). Briefly, positions of potential polymerase occupancy were sampled from a simulated gene, including both initiation and elongation regions. For each polymerase position, we extended the hypothetical RNA based on the gene template downstream of the polymerase position, with the designated probability of incorporating a Biotin-NTP and halting extension. The subsequent read was then filtered by size selection and plotted to generate simulated metagene traces (Supplemental Fig. 13). Using these simulations, we found that the 5′ peak position was influenced by both the Biotin-NTP run-on ratio and the size selection criteria.

To validate our *in silico* findings, we returned to the data and examined the distribution of short reads (less than 30 bps) relative to transcription start sites. We reasoned that short fragments would consist of a combination of TSS associated fragments truncated by Biotin-NTP incorporation and small fragments arising from sample handling, which should be randomly distributed throughout the genome. Hence the ratio of short reads near TSS relative to all short reads should be indicative of the ratio of labeled and unlabeled NTPs used in the run-on reaction. Indeed, the short read ratio does shift along the Biotin-NTP ratio, but not as a monotonically increasing function (Supplemental Fig. 14). Consistent with our simulations, intermediate Biotin-NTP/NTP ratios returned the highest fraction of mappable TSS associated short reads. Our results indicate that several library preparation elements, such as size selection, Biotin-NTP run-on ratios, and mappability strongly influence the 5′ distribution. Importantly, this work also suggests that the ideal run-on scenario is a balance between producing reads that are long enough to escape size selection and map effectively; yet remain short enough to accurately report on the position of RNA polymerase.

We next reasoned that the observed differences in the detected 5′ read distribution at genes would commensurately affect the pausing index (PI), measured as the ratio of reads in the initiation region relative to the gene body[28]. We defined the initiation region as 50 bp upstream from the annotated TSS to 250 bp downstream of the TSS; gene body regions were defined as 251 bp downstream of the TSS to the annotated cleavage site. Using these regions, we calculated the PI for the longest isoform of each gene in both libraries. Consistent with our findings above, PI for individual genes were reasonably consistent across replicates (Supplemental Fig. 15) but showed significant disparities between GRO and PRO libraries (Fig. 2.3C, $R = 0.59$, $p < 2.2e-16$). Spearman rank correlations for PI in both libraries were marginally higher ($R = 0.73$, $p < 2.2e-16$). These overall trends were also observed within PI distributions when we extended this analysis to publicly available data (Supplemental Fig. 12). While the PI is known to depend on the method used to define the paused region[2], we found that the trends across protocols remained consistent even with different pause windows and read counting software (Supplemental Fig. 16).

Next, we evaluated the effects of library preparation on the 5′ end. To accomplish this, we constructed metagene summaries of our GRO-CIRC, GRO-LIG, and GRO-RPR libraries (Fig. 2.3D). While CIRC and LIG libraries showed a similar distribution near the 5′ end, GRO-RPR libraries show a shift in coverage that leaves a significant gap near the annotated start site (Supplemental Fig. 2). While it is unknown what leads to this shift, we theorize that random priming has a length bias that is a contributing factor (i.e. the longer a RNA is the more likely a primer is to anneal to it).

Additionally, we found that the pause ratio is sensitive to which method is used to prepare the RNA. We compared pause index calculations for GRO-CIRC and GRO-LIG libraries. We found that, for each gene, pause indices tended to be larger for GRO-CIRC libraries compared to GRO-LIG libraries (Fig. 2.3E, $R = 0.57$, $p < 2.2e\text{-}16$). To assay whether this shift was systematic, we also computed the Spearman rank-correlation for these indices. Rank correlation between GRO-LIG and GRO-CIRC libraries was stronger than Pearson correlation; however, there were still many genes that showed disparate rankings across our datasets (Fig. 2.3E, $R = 0.77$, $p < 2.2e\text{-}16$).

2.2.3 Changing library enrichment methods shifts intergenic read distributions and active enhancer detection

The bidirectional transcription typical of RNA polymerase initiation regions at the 5′ end of genes is also present at enhancers[54], albeit typically at much lower transcription levels. Therefore, we asked whether the patterns of enhancer transcription varied across protocols or library preparations. As a first pass inquiry that avoids reliance on enhancer annotations, we first compare the fraction of reads recovered from RefSeq annotated gene regions to reads recovered in intergenic regions for each data set. To ensure more statistical rigor, we included several publicly available datasets of different cell lines, along with six libraries we previously generated from MCF10A cells prepped with PRO-TSRT (See Supplemental Table 1). When comparing GRO and PRO libraries (irrespective of cell type or library preparation method), we found that GRO libraries showed significantly more reads over gene regions compared to PRO libraries (Fig. 2.4A, $p < .01$, See Materials and Methods). Conversely, we found no significant differences when comparing library preparation methods (Fig. 2.4B).

The disparity in the gene-to-intergenic reads ratio in GRO and PRO libraries suggest their respective enrichment strategies may capture signal in unannotated regions at different rates. In particular, we were curious whether the capture of eRNAs would be affected by the choice of protocol. To investigate this possibility, we first examined annotated enhancers in the HCT116 cell line acquired from the FANTOM database (converted to hg38 coordinates using the online UCSC tool liftOver)[38]. The level of transcription between these enhancers was largely consistent between our datasets (Supplemental Fig. 17). However, FANTOM annotated enhancers represent the comparatively stable enhancer transcripts arising from Cap Analysis Gene Expression (CAGE) data[105].

Therefore, we next sought to identify enhancers directly from the data using their characteristic bidirectional transcription signal[19]. Two algorithms have been developed to identify transcribed regulatory elements based on their bidirectional signal, dREG[26] and Tfit[12]. We employed both methods to annotate sites of bidirectional transcription in our GRO-CIRC, GRO-LIG, and PRO-LIG libraries. Strikingly, the identified regions varied substantially across protocol and library preparation for both algorithms (Supplemental Fig. 18). We hypothesized that these differences may be exaggerated by the sequencing depth, as eRNAs are lowly transcribed and therefore these regions are only consistently detectable at high sequencing depth. To this end, we combined replicates for PRO-LIG libraries to an effective depth of approximately 200 million reads, and replicates of GRO-CIRC libraries to an effective depth of approximately 300 million reads. Transcribed regions identified in these combined libraries remained inconsistent; while many strong enhancers were called in both of these two deep data sets, other regions were exclusively found in only one (Fig. 2.4C, Supplemental Fig. 19).

This suggested the existence of transcribed regions whose signal is strongly dependent on the underlying experimental protocol. To confirm this possibility, we next sought to identify the set of transcribed regions with apparent differential transcription across protocols or library preparations. To compare enrichment protocols, we combined Tfit regions from PRO-LIG and GRO-LIG libraries (Fig. 2.4D, Supplemental Fig. 17, see Materials and Methods), while library preparation methods were compared by combining Tfit regions from GRO-LIG and GRO-CIRC libraries (Fig. 2.4E). In every case, regions were combined using **muMerge**[97] and differential read signal was assessed with DESeq1 analysis (Fig. 2.4F). We then constructed metagenes from set of regions with differential signal (Fig. 2.4G, H, Supplemental Fig. 20) and observed strong bidirectional signal in only one of the two datasets, while the other dataset showed signal only slightly above background. Manual inspection confirmed that these transcribed regions were only effectively captured by one library, even at high depths (Fig. 2.4C).

2.2.4 Biological response to p53 activation is preserved across run-on transcription capture protocols

The protocol-specific nature of both pausing ratios and eRNA recovery led to concerns about whether the choice of experimental preparation influences commonly conducted downstream analyses, such as identifying which genes respond to a perturbation[4] and which transcription factors drive those changes[42, 43, 11, 97]. As such, we used the competitive MDM2 inhibitor Nutlin-3a, which has a known, specific, robust transcription response in human cells induced by the subsequent activation of the transcription factor p53[103, 4, 6].

First, we sought to determine the reproducibility of detecting differential gene transcription within our libraries. The precise identity of which genes respond to 1 hour of p53 activation is expected to vary across protocols and library preparations – as similar batch effects have been observed for RNA-seq libraries[116]. Thus, we focused specifically on whether the core p53 response program, i.e. the known targets of p53, was captured efficiently in each dataset. To this end we utilize the Gene Set Enrichment Analysis (GSEA) - Preranked[78, 117] tool on ranked, signed p-values obtained from DESeq2[69] (See Materials and Methods). Additionally, we expected that a

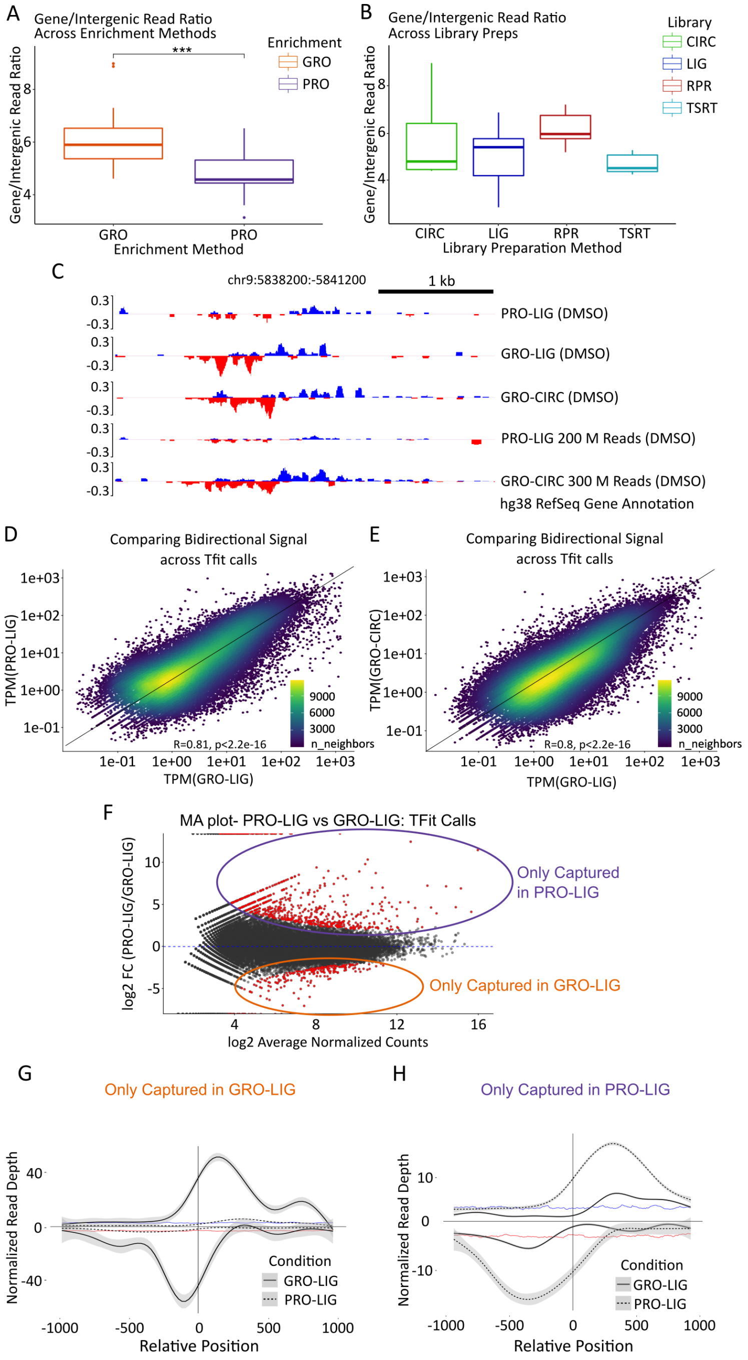


Figure 2.4: **Analysis of enhancer elements in multiple datasets.** (A,B) Number of reads counted over RefSeq annotated gene regions divided by the number of reads counted over intergenic (unannotated) regions, for each dataset analyzed. The datasets represented here are all those listed in Supplemental Table 1, including public datasets. **continued ...**

Figure 2.4: **continued** Datasets were first analyzed by enrichment method (GRO-seq (n=23) vs. PRO-seq (n=21), $p < .01$), then by library preparation method (LIG (n=17) vs CIRC (n=10) vs TSRT (n=10) vs RPR (n=7), $p > .05$). We note that the RPR boxplot includes 3 of our lower quality datasets; however, we chose to include them here owing to the scarcity of RPR datasets in the RO-seq database. These are otherwise excluded from further analysis. (C) Example section representing disparate representation of reads from our in-house datasets over an enhancer, even at high depths. (D, E) Scatterplots representing reads over Tfit (enhancer) calls (calls combined by MuMerge, counts normalized by TPM). (F) MA plot of calls found in (D). Red dots are significant ($p < .05$). (G, H) Metagenes of significant hits found in (F). Vertical line indicates the approximated center of the bidirectional transcripts as determined by Tfit. Distance from the center of the bidirectional is in bp, read depth was normalized by counts-per-million (CPM). (G): Calls that were differentially captured in GRO-LIG (n=1350). Background signal on the plus strand is indicated by the blue trendline, while background signal on the minus strand is indicated by the red trendline. (H): Calls that were differentially captured in PRO-LIG (n=3050), with the background signal depicted as in panel G.

substantial amount of variation between two libraries generated from different protocols would arise from the gene initiation region (Fig. 2.3). To confirm this, we subsequently examined two distinct methods of calculating differential gene transcription: the commonly used elongation-region-only approach and the full annotated gene region (Fig. 2.5A). Across all libraries and counting methods, the p53 pathway was the top hit in the GSEA-Preranked module (FDR q-val < 0.001, Fig. 2.5B, Supplemental Fig. 21), suggesting that each protocol, library preparation and counting method was capable of detecting the underlying biological perturbation in spite of technical signals introduced by protocol differences.

Next, we compared the correlation of the ranks of the genes in the Hallmark p53 pathway used by GSEA. We found that the majority of enriched genes were common between each of the libraries (58.3% in GRO-LIG vs GRO-CIRC, 57.1% in GRO-LIG vs PRO-LIG) (Fig. 2.5B,C, Supplemental Fig. 22). However, there remained several genes that were only enriched in one of libraries. When only the elongation region was considered, the overlap improved (68.3% in GRO-LIG vs GRO-CIRC, 58.9% in GRO-LIG vs PRO-LIG), consistent with the $5'$ initiation regions being the most variable portion of the gene between protocols. These results add further support to the most common method of assessing differential transcription from run-on sequencing protocols, namely excluding

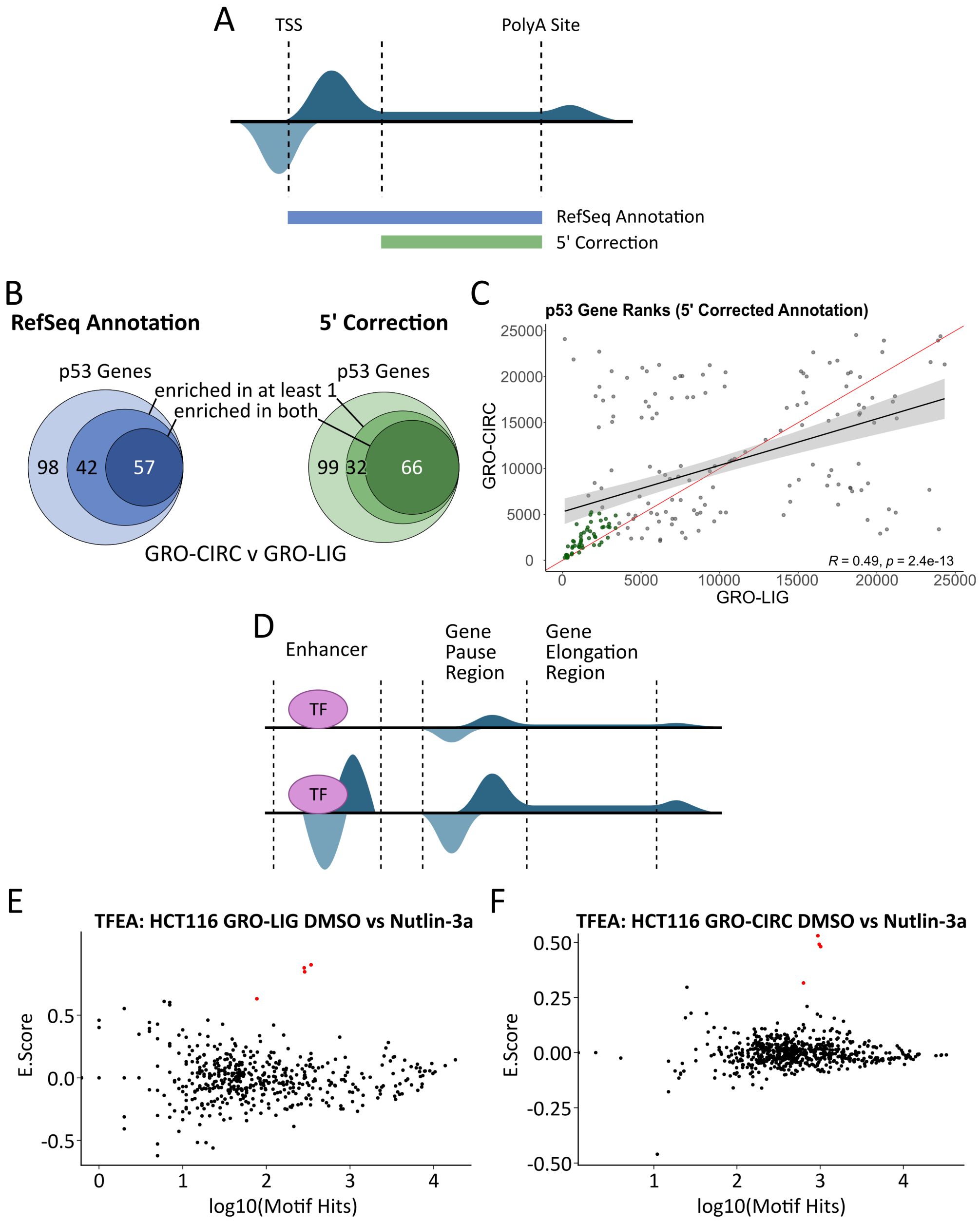


Figure 2.5: **TFEA and DESeq2 analyses of library preparation methods.** (A) Cartoon schematic demonstrating uncorrected (RefSeq Annotation) and 5′ corrected counting methods. (B) GSEA gene rank comparison of HALLMARK_P53 Gene set. Overlap is shown as genes that enrich in both datasets, genes that enrich in only one dataset, and genes that do not enrich in either dataset (Left: Uncorrected annotation, hypergeometric test p-value=4.32e-15; Right: Corrected annotation, hypergeometric test p-value=9.03e-22). (C) Scatterplot of comparative gene ranks for all p53 genes. Points in green indicate significant enrichment, as in (B). (Red line: y=x trendline, black line:line of best fit). (D) Representation of nascent transcription data set. Bidirectional transcripts occur at active enhancer sites and gene start sites. Enhancer transcription co-occurs with upregulated gene transcription, indicating transcription factor activation. (E) TFEA results for GRO-LIG (Left) and GRO-CIRC (Right). p53 family (p53, p63, p73) highlighted by red dots.

the $5'$ initiation regions[77, 70, 30, 17].

The second typical use of run-on sequencing data is to infer which regulators are driving observed patterns of differential transcription[54, 26, 11]. Alterations in transcription factor activity can be detected by changes in the locations and levels of sites of bidirectional transcription[11, 97], the majority of which reside at enhancers[19]. Therefore we next sought to determine whether the alterations observed in eRNA detection (Fig. 2.4) impacted TF activity inference[97].

To this end, we used the Transcription Factor Enrichment Analysis (TFEA) tool to evaluate which transcription factor motifs are enriched at transcription initiation sites with altered transcription levels in response to Nutlin-3a[97]. In all cases, TFEA correctly identifies the p53 family (TP53, TP63, and TP73) as significantly upregulated, independent of the protocol and library prep used to generate the dataset (Fig. 2.5E and F, Supplemental Fig. 23). Upon closer inspection, 94.59% of p53-responsive enhancers responded similarly across protocols, but 5.41% of p53-responsive enhancers were unique to a particular protocol (Supplemental Fig. 24, 25).

2.3 Discussion

We used multiple protocols and library preparations on HCT116 cells exposed to Nutlin-3a and determined that these experimental choices influence the signal of run-on sequencing libraries in systematic and often predictable ways. The shape of the characteristic gene initiation peak is strongly influenced by the underlying protocol, while the signal at gene elongation regions remain largely consistent across protocols. Likewise, the recovery of many intergenic regions was protocol specific, even when at high sequencing depths. Despite these differences, the ability to detect p53 activation was unaffected by the choice of enrichment or library preparation protocol.

Promoter proximal pausing is a pervasive feature of RNA polymerase II activity[2]. Pausing is often quantified through calculations of the pausing index, the ratio of reads within the initiation region relative to the elongation region. While PI values are known to depend on the choices of windows used to define these regions[2], our work demonstrates that they also depend on the underlying protocol even when the details of the PI index calculation are held constant. Furthermore, genes sometimes appear to have an additional pause site downstream of the annotated TSS (Fig. 2.3E)[9]. However, we have found that these second pause sites are protocol dependent; as changes in the library preparation method shift or ablate the signal of this second peak. While more work is necessary to fully characterize how protocol choices influence the precise location of the $5'$ peak, it is clear that care must be taken when comparing $5'$ distributions across experiments, as batch effects strongly influence this region.

Given the uniform activity of RNA polymerase II[5], the $5'$ end protocol specific patterns we observed at genes should also impact enhancer associated transcripts. The most highly transcribed eRNAs (e.g. those annotated by FANTOM) are detected equally well by each protocol, but many eRNAs are lowly transcribed. Indeed, we observe that some enhancers with relatively high read coverage in one library are not detectable using a different protocol. We were surprised that increased depth did not resolve many of these protocol specific eRNAs. The variability in eRNA detection has likely hampered efforts to answer an outstanding question in the field; namely, how many eRNAs exist throughout the genome? Combining results from many different protocols and cell types may help alleviate this issue.

This disparity in eRNA signal raises an intriguing question: which aspects of the protocols and resulting libraries contribute to the difference in eRNA capture rates? The slightly higher exon to intron ratio (Fig. 2.2D) of GRO-seq suggests this protocol contains a higher level of contaminating mRNA[125], consistent with Br-UTP antibody enrichment being a less efficient pull down method than Biotin-streptavidin enrichment. This bias also explains why GRO-seq has a higher gene to intergenic ratio compared to PRO-seq (Fig. 2.4A). These features may lead to some lowly transcribed eRNAs being more readily detectable with PRO-seq. In contrast, the use of Biotin halts polymerase elongation in PRO-seq, giving it a higher precision on RNA polymerase position[59]. However, this also results in short, unmappable fragments near the $5'$ end of transcripts, which may limit the ability of PRO-seq to capture some shorter eRNAs. This phenomenon would explain why certain eRNAs are only captured in GRO-seq. Likewise, other factors probably contribute to the recovery of eRNAs[130], including sequence composition and biological variability.

Despite the observed protocol specific differences, our downstream analysis was consistent in detecting the underlying p53 perturbation. At genes, it is customary to exclude the initiation peak from differential gene transcription analysis[77, 70, 30, 17], and our work indicates this is a wise choice, as counting reads only over elongation regions gave more consistent results across the protocols. Yet even when using only elongation regions, protocol specific batch effects determine which exact genes appear to respond, a problem also seen with RNA-seq[75, 100]. Likewise, detection of enhancer associated RNAs showed similar protocol specific batch effects. Importantly, despite the specifics of individual genes (and eRNAs) being not fully consistent, the large scale conclusion (p53 is activated by Nutlin-3a) remained consistent. Thus nascent transcription remains a powerful approach for understanding the immediate responses to perturbations including compounds and drug activity[30, 11, 112, 97].

2.4 Conclusion

Protocol and platform differences have long been recognized as batch effect variables that introduce non-trivial experiment specific signals within high throughput sequencing data[64, 39]. Numerous efforts have focused on correcting batch effects, but it is always difficult to do so without some loss of biological signal[110, 132]. On the other hand, the distinct signals we detect raise an intriguing possibility that protocol and library preparation information can be inferred directly from the data itself. The noise component of the data can reliably differentiate between GRO- and PRO-seq datasets with remarkable accuracy, while sequence and quality signatures can often identify the library preparation methods used to prepare the dataset. Thus an automatic detection approach could be built to confirm or correct experimental information within the short read archive, at least for run-on assays[99]. Regardless, knowing the experimental details and managing associated batch effects is necessary when comparing in house data to previously published data sets.

2.5 Materials and Methods

2.5.1 Cell Culture Conditions

HCT116 and MCF10A[65] cells were cultured in DMEM media supplemented with 10% FBS, 100 units/mL penicillin and 100 μg/mL streptomycin, at 37°C with 5% CO2. Cells were grown to a confluency of 60-70% in 15 cm culture dishes before passaging. Cells were passaged twice before harvesting, using PBS to wash and 0.05% w/v trypsin to detach the cells from the plate. Cells were aspirated and treated with media containing 10 μM Nutlin-3a (or DMSO) for 1 hour before harvest.

2.5.2 Nuclei Isolation

Post-treatment, cells were placed on ice and washed three times with ice-cold PBS. Cells were incubated on ice in 10 mL ice-cold Lysis Buffer (10 mM Tris-HCl pH 7.5, 2 mM $MgCl_2$, 3 mM $CaCl_2$, 0.5% IGEPAL, 10% Glycerol, 2 U/mL SUPERase-IN, brought to volume with 0.1% DEPC DI-water, filtered before use) for 10 minutes. Cells were scraped and collected into 50 mL Falcon tubes, and centrifuged with a fixed-angle rotor at 1000 x g for 10 minutes at 4°C. Cells were resuspended with Lysis buffer with a wide-opening P1000 tip, and washed twice with 10 mL Lysis buffer (centrifuged at 1000 x g for 5 minutes at 4°C). After the second Lysis buffer wash, the samples were resuspended with 1 mL Freezing Buffer (50 mM Tris-HCl pH 8.3, 5 mM $MgCl_2$, 40% Glycerol, 0.1 mM EDTA pH 8.0, brought to volume with 0.1% DEPC DI-water, filtered before use). Nuclei were centrifuged at 1000 x g for 5 minutes at 4°C, and resuspended with 500 μL Freezing Buffer. Nuclei were then centrifuged for 2 minutes at 2000 x g, 4°C, and resuspended in 110 μL Freezing Buffer. 10 μL was retained for counting nuclei, while the remaining sample was snap-frozen in liquid nitrogen and stored at -80°C until use.

2.5.3 GRO-seq and Library Preparation Methods

2.5.3.1 Ligation (LIG)

Run-on reactions were performed as in [24]. In brief, ice-cold isolated nuclei (100 μL) were added to 37°C 100 μL reaction buffer (Final Concentration: 5 mM Tris-Cl pH 8.0, 2.5 mM $MgCl_2$, 0.5 mM DTT, 150 mM KCl, 10 units of SUPERase In, 0.5% sarkosyl, 500 μM rATP, rGTP, and Br-UTP, 2 μM rCTP). The reaction was allowed to proceed for 5 min at 37°C, followed by the addition of 23 μL of 10X DNAseI buffer, and 10 μL RNase free DNase I (Promega). RNA was extracted twice with Trizol, washed once with chloroform, and precipitated with 3 volumes of ice-cold ethanol and 1-2 μL GlycoBlue. The pellet was washed in 75% ethanol before resuspending in 20 μL of DEPC-treated water. Libraries were prepared as in [24]. In brief, nascent RNA was extracted and fragmented by base hydrolysis in 0.2 N NaOH on ice for 10–12 min, and neutralized by adding a 1× volume of 1 M Tris-HCl pH 6.8. Fragmented nascent RNA was purified using Anti-BrdU beads and ligated with reverse 3′ RNA adaptor (/5Phos/GAUCGUCGGACUGUAGAACUCUGAAC/3InvdT/), and BrdU-labeled products were enriched by a second round of Anti-BrdU bead binding and extraction. For 5′ end repair, the RNA products were treated with tobacco acid pyrophosphatase (Epicenter) and T4 polynucleotide kinase (NEB). 5′ repaired RNA was ligated to reverse 5′ RNA adaptor (5′ UGGAAUUCUCGGGUGCCAAGG) before being purified by a final round of Anti-BrdU bead binding and extraction. RNA was reverse transcribed using 25 pmol of RP1 primer (5′AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA). The product was amplified 15 ± 3 cycles and products >150 bp (insert > 70 bp) were size selected with 1X AMPure XP beads (Beckman) before being sequenced.

2.5.3.2 Random Priming (RPR)

Run-on reactions were performed as in [24]. In brief, ice-cold isolated nuclei (100 μL) were added to 37°C 100 μL reaction buffer (10mM Tris-Cl pH 8.0, 5 mM $MgCl_2$, 1 mM DTT, 300 mM KCl, 20 units of SUPERase In, 1% sarkosyl, 500 μM ATP, GTP, and Br-UTP, 2 μM CTP).

The reaction was allowed to proceed for 5 min at 30°C, followed by the addition of 23 μL of 10X DNAseI buffer, and 10 μL RNase free DNase I (Promega). RNA was extracted twice with Trizol, washed once with chloroform, and precipitated with 3 volumes of ice-cold ethanol and 1-2 μL GlycoBlue. The pellet was washed in 75% ethanol before resuspending in 20 μL of DEPC-treated water. Libraries were prepared based on the NEBNext Ultra II Directional Library Preparation Kit. In brief, nascent RNA was extracted and fragmented by base hydrolysis in 0.2 N NaOH on ice for 10–12 min, and neutralized by adding a 1× volume of 1 M Tris-HCl pH 6.8. Fragmented nascent RNA was purified using Anti-BrdU beads (Santa Cruz Biotech, Santa Cruz, CA) 3 times. Samples were reverse-transcribed using random hexamers, and sequencing adapters added by PCR. The product was amplified 15 ± 3 cycles and products >150 bp (insert > 70 bp) were size selected with 1X AMPure XP beads (Beckman) before being sequenced.

2.5.4 PRO-seq and Library Preparation Methods

2.5.4.1 Ligation (LIG)

Run-on reactions were adapted from [71]. In brief, ice-cold isolated nuclei (100 μL) were added to 37°C 100 μL reaction buffer (Final Concentration: 5 mM Tris-Cl pH 8.0, 2.5 mM $MgCl_2$, 0.5 mM DTT, 150 mM KCl, 10 units of SUPERase In, 0.5% sarkosyl, 125 μM rATP, 125 μM rGTP, 125 μM rUTP, 25 μM biotin-11-CTP (additionally, two libraries generated with 25 μM biotin-11-CTP, 250 μM rCTP, see Supplemental Table 1). The reaction was allowed to proceed for 5 min at 37°C. RNA was extracted twice with Trizol, washed once with chloroform, and precipitated with 3 volumes of ice-cold ethanol and 1-2 μL GlycoBlue. The pellet was washed in 75% ethanol before resuspending in 20 μL of DEPC-treated water. Nascent RNA was extracted and fragmented by base hydrolysis in 0.2 N NaOH on ice for 10–12 min, and neutralized by adding a 1× volume of 1 M Tris-HCl pH 6.8. Fragmented nascent RNA was purified using streptavidin beads and ligated with reverse 3′ RNA adaptor (/5Phos/GAUCGUCGGACUGUAGAACUCUGAAC/3InvdT/), and biotin-labeled products were enriched by a second round of streptavidin bead binding and extraction. For 5′ end repair, the RNA products were treated with tobacco acid pyrophosphatase (Epicenter) and T4 polynucleotide kinase (NEB). 5′ repaired RNA was ligated to reverse 5′ RNA adaptor (5′ UGGAAUUCUCGGGUGCCAAGG) before being purified by a final round of streptavidin bead binding and extraction. RNA was reverse transcribed using 25 pmol of RP1 primer (5′AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA). The product was amplified 15 ± 3 cycles and products >150 bp (insert > 70 bp) were size selected with 1X AMPure XP beads (Beckman) before being sequenced.

2.5.4.2 Template-Switch Reverse Transcription (TSRT)

Template-Switch Reverse Transcription protocol (also known as uPRO), was adapted from [53]. Nuclei were incubated in the nuclear run-on reaction condition (5 mM Tris-HCl pH 8.0, 2.5 mM $MgCl_2$, 0.5 mM DTT, 150 mM KCl, 0.5% Sarkosyl, 0.4 units / l of SUPERase-In) along with biotin-NTPs and rNTPs (125 μM rATP, 125 μM rGTP, 125 μM rUTP, and 25 μM biotin-11-CTP) for 5 min at 37°C. Run-On RNA was extracted using TRIzol, and fragmented with 0.2 N NaOH for 10-12 min on ice. Fragmented RNA was neutralized with 1 M Tris-HCl pH 6.8, and buffer exchanged by passing through P-30 columns (Biorad). 3′ RNA adaptor (/5Phos/GAUCGUCGGACUGUAGAACUCUGAAC/3InvdT/) is ligated at 5 μM concentration for 1 hour at room temperature using T4 RNA ligase (NEB), and nascent RNA was enriched twice with streptavidin beads. Extracted RNA was converted to cDNA using template switch reverse transcription with 1 μM RP1-short RT primer (5′ GTTCAGAGTTCTACAGTCCGA), 3.75 M RTP-Template Switch Oligo (5′ GCCTTGGCACCCGAGAATTCCArGrGrG), 1x Template Switch Enzyme and Buffer (NEB) at 42°C for 30 min. Resulting product was size selected with AMPure XP beads, and the cDNA was PCR amplified using primers compatible with Illumina Small RNA sequencing (TruSeq Small RNA primers RP1 and RPIn).

2.5.5 Trimming, Mapping, Visualization, Quality Control

Resulting FASTQ files were trimmed and mapped to the GRCh38/hg38 reference genome and prepared for analysis and visualization through our in-house pipeline. In short, resulting FASTQ read files were first trimmed using bbduk (v38.05) to remove adapter sequences, as well as short or low quality reads. Reads were mapped with HISAT2 (v2.1.0), and resulting SAM files converted to BAM files using Samtools (v1.8). Reads with a mapping quality less than 5 were removed, which consequently also removed multi-mapping reads. BedGraph files were generated using Bedtools (v2.25.0), and converted to TDF files for visualization using IGVtools (v2.3.75). Quality metrics were generated with FastQC (v0.11.8), Preseq (v2.0.3), RSeQC (v3.0.0), with figures generated through MultiQC (v1.6). For further version information and specific input information, see NextFlow pipeline found at https://github.com/Dowell-Lab/Nascent-Flow.git.

2.5.6 Exon/Intron Ratio

RefSeq annotations were used to define exonic and intronic boundaries for each gene. The first exon of each gene was excluded (to avoid the initiation peak signal) in each calculation. Reads were counted using featureCounts from the R-Subread package (v1.6.0). Exonic and intronic reads were summed and normalized by RPKM, and a ratio for each gene is calculated. These ratios were log-normalized and the median ratio calculated for each set of libraries analyzed.

2.5.7 Discrete Wavelet Transform

Samples with high coverage were used for this analysis. This included samples from the GRO-LIG, PRO-LIG, GRO-CIRC and PRO-TSRT libraries. The coverage over a gene transcript was normalized to 0-1 scale as show below:

$$c_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

Where $x = (x_i, ..., x_n)$ represents read counts over a genomic location n, and c_i is the normalized coverage per genomic location. As we sought to identify protocol influences independent of biological gene variability, we limited our analysis to ubiquitously transcribed genes with low coefficient of variation (CV) across all samples. Thus, a total of 294 genes with a CV less than 0.55 and average transcripts per million (TPM) greater than 150 were selected. Using the PyWavelet (version 1.0.3) API in python (version 3.6.3), the symlet 5 mother wavelet was scanned across the 294 genes, returning wavelet coefficients (approximation coefficient and detail coefficients) (Fig. 2.2E) [27, 62, 122]. After the first pass of wavelet transform, the detail coefficients were used as input for principal component analysis (PCA) using scikit-learn (version 0.20.2) [86]. So, for each gene and each sample, PC1 and PC2 values were returned. Genes were split into categories based on whether the protocols could be split on PC1 and PC2 or whether the gene could not separate the protocols in PC space. The above process was then repeated for a larger set of 669 genes (CV less than 0.85 and average TPM greater than 100). Plots were generated with matplotlib (version 3.3.4), ggplot2 (version 3.3.3) and cowplot (version 1.1.1) [48, 128, 129]. Code for the DWT analysis can be found on github (https://github.com/Dowell-Lab/Protocol-Comparisons).

2.5.8 Support Vector Machine

Principal component analysis values (from PC1 and PC2) derived from the wavelet transform analysis pipeline were used as input to a support vector machine (SVM). In order to verify the performance of the classification, the leave-one-out cross validation (LOOCV) criteria was used (Supplemental Fig. 6). A linear kernel was chosen for the SVM using the e1071 (version 1.7-4) package in R (R version 3.6.0) [76, 92]. The folds for the LOOCV were created with the caret package (version 6.0-86) in R (version 3.6.0) and accuracy for each fold and gene was calculated [58]. A total of 18 folds were created, where each of the 18 samples was held out one at a time as the test sample in the SVM, while the remaining samples were used as a training set. This was done for all the genes analysed and the evaluation determined the number of genes accurately predicting the protocol for each of the 18 samples. Plots were made using ggplot2 (version 3.3.3) and cowplot (version 1.1.1) [128, 129]. The jupyter notebook for the SVM LOOCV analysis can be found on github (https://github.com/Dowell-Lab/Protocol-Comparisons).

2.5.9 Pause Index Calculations

Refseq annotations were used as the basis for pause index calculations. Counts were generated either from bedtools multicov (v2.28.0). The paused region was defined as -50 bp to 250 bp from the annotated TSS [28], and the elongation region was defined as 251 bp from the TSS to the annotated PolyA site. Reads from the same strand as the annotated gene were counted for the paused and elongation region, and calculated the index as follows:

$$\text{pausing index}(pi) = \frac{ReadCount(Pausing\ Region)/L1}{ReadCount\,(Gene\ Body)\,/L2}$$

Where L1 is the length of the pausing region (300 bp) and L2 is the length of the elongation region, measured from 251 bp past the TSS to the annotated cleavage site found in RefSeq. Only pause index values from a gene's longest isoform were considered. Genes shorter than 2000 bp were removed.

The above analysis was repeated using featureCounts (v1.6.2) in the R-Subread package (v1.6.0), where the paused region was defined as -20 to +80 from the annotated TSS, and the elongation region as +81 from the TSS to -1000 from the annotated PolyA site. Genes shorter than 2000 bp were filtered out. These results are available in Supplemental Fig. 16.

2.5.10 Simulation of reads near transcription start sites

We generated 2000 base gene template with equal proportions of A, C, G, and T. Using these templates, we then simulated RNA polymerase activity similar to a previously established mathematical framework[12]. Briefly, the model assumes a position for reads to start (the transcription start site) and a polymerase distribution around the TSS determined by a normal distribution. We sampled 10,000 initiation polymerases and 5,000 elongating polymerases randomly. Each polymerase was then allowed to run-on with a random change to terminate transcription based on the sequence identity and biotin-NTP/NTP ratio specified. Transcript lengths, e.g. reads, were then determined using the difference between the TSS and the termninated location of the polymerase. To mimic Ampure bead size selection, reads were then subjected to a size selection cutoff determined by an exponential distribution proportional to their length, resulting in an average cutoff of approximately 25 bases. The resulting read pool was subsequently used to generate metaplots of our synthetic template (Python v. 3.6.3, Numpy v.1.15.4, Pandas v. 0.23.4. Jupyter Notebook available at https://github.com/Dowell-Lab/Protocol-Comparisons).

2.5.11 Short Read Ratio Comparison

All reads greater than 30bp were filtered out of PRO-seq libraries to analyze the location of short reads within the genome. Each library was first assigned an Unlabeled/Labeled NTP ratio based on the run-on reaction concentrations of biotin-NTP relative to unlabeled NTPs reported by the authors for each dataset. GRO samples SRR14355674, SRR14355673, SRR14355662, SRR14355655 were included as a reference point. All PRO-seq libraries indicated in Supplemental Table 1 were considered for this analysis. Public samples SRR8033049, SRR8033050, SRR8033051, SRR8033052, SRR8033053, SRR8033054, SRR8033055, SRR8033056, SRR8033057, SRR8033058, SRR6205688, SRR6205689, SRR4041365, SRR4041366, SRR4041367, SRR4041368, SRR4041369, SRR4041370, SRR4041371, SRR4041372, SRR4041373, SRR5364303, and SRR5364304 were also included in this analysis, but were excluded from Supplemental Table 1 as they were not part of other analyses within this study.

Reads within 20 bp of the RefSeq TSS were considered to be near the TSS; we then calculated the ratio of these reads relative to all small reads found throughout the genome. The resulting ratio was plotted relative to the run-on reaction NTP ratio using R (version 3.6.3). Plots were made using ggplot2 (version 3.3.3) and cowplot (version 1.1.1) [128, 129].

2.5.12 Gene/Intergenic Reads Ratio Calculation

Genic and intergenic regions were determined by RefSeq (hg38, release number 109, downloaded August 14, 2019 from UCSC genome browser) annotation. Genic and intergenic read proportions were calculated by RSeQC (v3.0.0) read_distribution.py. Genic regions were defined as those overlapping a RefSeq annotation, including introns and untranslated regions. Intergenic regions were calculated as the remainder of reads not mapping to a gene region. The reads ratio of genic and intergenic regions can be found for each sample in Supplemental Table 1.

2.5.13 Tfit

Tfit was used to identify regions of bidirectional transcription in each of our run-on sequencing libraries. Resultant BedGraph files from our samples were used as the input for the –bedgraph flag of the Tfit prelim module. The resultant preliminary region file was used as the –segment flag input for the Tfit model module, resulting in the final bidirectional calls used for analysis (see also https://github.com/Dowell-Lab/Tfit.git). Calls between replicates and treatments were combined using **muMerge**, generating a set of combined calls for each set of conditions (GRO-LIG, PRO-LIG, and GRO-CIRC). To compare library preparation methods, the above GRO-CIRC and GRO-LIG sets were combined together through bedtools merge (v2.28.0). Likewise, to compare enrichment methods, PRO-LIG and GRO-LIG sets were combined via bedtools merge (v2.28.0).

2.5.14 dREG

We used dREG to identify regions of bidirectional transcription in each of our run-on sequencing libraries. Resultant BAM files from our samples were first converted to BigWig files compatible with dREG (see https://github.com/Danko-Lab/RunOnBamToBigWig.git). Using the online dREG portal, these files were used to generate dREG calls for bidirectional regions (https://django.dreg.scigap.org). Calls between replicates and treatments were combined using **muMerge**, generating a set of combined calls for each set of conditions (GRO-LIG, PRO-LIG, and GRO-CIRC). For comparative analyses between any of these sets, each set combined by **muMerge** was concatenated and used as the input for bedtools merge (v2.28.0), generating a consensus set of regions for those two sets.

2.5.15 Differential Transcription Analysis

Differential transcription was performed using the DESeq2 (v1.26.0) R package (R version 3.6.3). DESeq2 no longer allows differential calls without replicates; thus, when comparing libraries where treatments and replicates were combined, the DESeq (v. 1.38.0) R package was used instead. Gene counts were generated using featureCounts (v1.6.2) from the R Subread package (v1.6.0), counting over the entire gene body from RefSeq Annotations (release number 109, downloaded August 14, 2019 from UCSC genome browser). For featureCounts, BED6 region files were converted to SAF format with the following command: awk -F "\t" -v OFS="\t" 'print{$4, $1, $2, $3, $6}' region.bed > region.saf. Only the highest transcribed isoform of each gene was considered. Counts over Tfit, dREG, or FANTOM calls were generated with featureCounts.

2.5.16 GSEA

DESeq2 gene results were ranked based on -log(P-value)/sign(Fold-Change). These ranked lists were used as the input for GSEA-preranked module (v4.1.0). The Hallmark v7.4 gene sets were used as the input database. Results were generated using 1000 permutations. Gene symbols were not collapsed.

2.5.17 TFEA

Resulting Tfit bidirectional calls were used as the input for TFEA for each experiment (summarized in Supplemental Table 1). Calls were combined using **muMerge**. Transcription factor motifs were identified using FIMO (MEME Suite v5.1.1), using full human HOCOMOCO (version 11) motifs.

2.6 Abbreviations

RO-seq: Run-On sequencing. PRO-seq: Precision Run-On sequencing. GRO-seq: Global Run-On sequencing. CIRC: Circularization based library preparation. LIG: Ligation based library preparation. RPR: Random Priming based library preparation. TSRT: Template Switching Reverse

Transcriptase based library preparation. DWT: Discrete Wavelet Transform. PCA: Principal Component Analysis. SVM: Support Vector Machine. LOOCV: Leave-One-Out Cross Validation. TSS: Transcription Start Site. eRNA: Enhancer RNA. GSEA: Gene Set Enrichment Analysis. TFEA: Transcription Factor Enrichment Analysis.

2.7 Declarations

2.7.1 Competing interests

Dr. Dowell is founder of Arpeggio Biosciences, the other authors declare that they have no competing interests.

2.7.2 Acknowledgements

We thank artist David Deen for figure composition and refinement assistance. We thank Chi Zhang and Nuria Morral for their contributions to PRO-LIG library generation. We also thank the BioFrontiers Institute Next-Gen Sequencing Core and the Biochemistry Shared Cell Culture Facility for their invaluable contributions to this study.

2.7.3 Author's contributions

This study was conceived by RDD, MAA and SH. Discrete wavelet transform analyses was conducted by RFS with guidance from JTS. GRO-seq libraries were generated by MAA. PRO-seq libraries were generated by SH and MAA. The scripts for *in silico* read generation and metaplot formation were written by MAA. All other analyses and initial manuscript was written by SH. All authors reviewed and revised the manuscript.

2.7.4 Funding

This work was funded by a National Science Foundation (NSF) ABI grant number 1759949. We acknowledge the BioFrontiers Computing Core at the University of Colorado Boulder for providing high-performance computing resources (NIH 1S10OD012300) supported by BioFrontiers' IT staff.

2.7.5 Availability of data and materials

The datasets used in this study are summarized in Supplemental Table 1. Datesets generated for this study are available through the Sequence Read Archive, under the accession PRJNA722106.

Chapter 3

Normalization of Aneuploid Sequencing Data for Differential Analysis

3.1 Author's contributions

This chapter is adapted from a submitted manuscript. Dr. Mary Allen was responsible for experimental data generation, initial concept and analysis, and normalization techniques. I was responsible for data simulations, final data analyses, figure generation, SNP analysis, and drafting the manuscript. Robin Dowell was responsible for manuscript proofreading and support on data analysis.

3.2 Introduction

Trisomy 21 (T21, also known as Down syndrome) is the most prevalent aneuploidy in the human population[46]. Individuals with Down syndrome have a number of common physical features as well as some degree of intellectual disability[7]. In addition, they have an increased risk for certain health problems (such as congenital heart defects) and a decreased risk of others (such as solid tumor formation). This altered risk profile arises primarily from the effect of higher levels of transcription of chromosome 21 genes[72, 90, 50]. The increase in DNA copy number and transcription has led to the DNA dosage hypothesis: gene transcription levels are proportional to DNA dosage [107, 121].

Dosage compensation, on the other hand, is any mechanism that exists to modulate gene expression in order to compensate for increased DNA dosage. The most famous and well studied dosage compensation mechanism involves X inactivation as a means of balancing sex chromosome levels. In contrast, no dosage compensation mechanism is known to exist for an entire mammalian autosomal chromosome. Given that in Down syndrome an autosomal chromosome is triplicated, there has been tremendous interest in whether dosage compensation exists for any genes on chromosome 21.

Numerous studies of gene expression in Down syndrome cells have reported gene expression levels that do not strictly follow DNA dosage [74, 87, 91, 111, 8, 56]. Others have contradicted these finding, arguing instead that most genes follow the expected 1.5 fold increase, with only a few genes showing lower than expected expression levels[72, 50]. In part, the differences in opinion arise from how each study defines dosage compensation in practice. For example, a permissive definition regards every gene below the DNA dosage expected 1.5 median fold change as dosage compensated[56], but this effectively ignores statistical variation inherent to these kinds of measurements. A more statistically principled approach looks for deviations from the 1.5 median fold change that exceed statistical expectation[50]. Regardless of the methodology used for identifying dosage compensation, all studies identify at least a small number of genes that are expressed below expectation, suggesting they are dosage compensated.

We sought to identify the sources of apparent dosage compensation, both technical and molecular. To this end, we focus on a family of individuals where one child has Down syndrome. We reasoned that if dosage compensation of a single transcript did occur, it would either occur via inhibition of transcription, or via increase in the RNA degradation of that transcript. Thus we examined both steady state RNA levels via RNA-seq and nascent transcription with global run-on sequencing (GRO-seq). As a first step, we preformed traditional differential expression analysis on these data sets.

All studies of differential expression use analysis tools, such as DESeq2 to determine the list of genes that are differentially expressed. These tools seek to quantify changes between conditions that exceed within-condition variability using principled models of read count data[21]. Overall, these pipelines provide a systematic, reproducible approach which seek to maximize the number of accurate differential gene calls while minimizing false positives by accounting for the underlying noise inherent to these datasets. With the advent of next-generation sequencing technologies, these differential analysis tools have become ubiquitous.

However, when we performed our analysis we quickly realized that traditional analysis assumes that differential expression tools work equally well on trisomy data. To date no study has examined how the presence of an extra copy of chromosome 21 impacts the typical differential expression analysis pipeline. Therefore, we dissect the the typical analysis pipeline to identify issues that could lead to erroneous identification of dosage compensation. To this end, we created simulated transcription data sets for both a disomic and trisomic individual where no dosage compensation is present by design. The simulated data allowed us to create a trisomy-aware differential expression analysis pipeline that when applied to simulated trisomy 21 samples accurately reflects the truth of the underlying data. When we applied our trisomy-aware analysis pipeline to the GRO-seq and RNA-seq data, many fewer chr21 genes are expressed lower than the 1.5x expectation.

For those few genes, we leverage the family structure and DNA sequencing to examine patterns of transcription associated with individual alleles to determine whether the set of genes with lower than expected transcription levels can be explained by allelic variation. Research suggests inter-individual variation maybe a stronger contributor to differential expression in T21 studies than sex or aneuploidy status[50, 72]. Consistent with this, tremendous variability in expression levels exist within the population of typical humans[61]. For example, large scale studies of gene expression among typical humans find that 83% of genes are differentially expressed between subsets of individuals[115]. In fact, expression quantitative trait loci (eQTL) studies seek to identify loci genome-wide that contribute to variations in gene expression levels[80, 40]. Therefore, we hypothesize that apparent dosage compensated genes could arise from genetic variations that lead to lower expression levels (eQTLs). In other words, some genes assumed to be dosage compensated may merely reflect the allele identity within the individuals.

Overall, our findings show that there is no dosage compensation at either transcription or steady state RNA levels. Furthermore, our work shows that natural genetic variation that contributes to expression levels can explain genes that appear to be dosage compensated. Finally, based on our simulated data sets, we create guidelines for accurate differential expression analysis in trisomy cells, which we call trisomy-aware transcription analysis.

3.3 Results

3.3.1 A naive analysis raises red flags suggesting technical issues

As an initial baseline, we first examined the typical differential analysis pipeline that leverages DESeq2[69]. In this naive analysis, we make no adjustments to the defaults inherent to programs within the pipeline. In RNA-seq, the median fold change of all genes on chromosome 21 is 1.41, with 57.6% of individual genes having a median fold change below the expected 1.5 fold change. The overall trends in GRO-seq are similar with an overall chromosome 21 median fold change of 1.38 with 48.8% below the expected 1.5 fold change.

When considering a cutoff for dosage compensation, we first examined the range of all possible fold change cutoffs. We generated a cumulative distribution function of fold change for all chromosome 21 genes in RNA-seq and GRO-seq in order to identify whether clear cutoff values could be identified (Supplemental Fig. 28). We identified a prevalence of genes expressed near 1.5 times disomic levels; however, no other cutoff was immediately clear. As such, these results contend that identifying dosage compensated genes using a cutoff is an arbitrary classification.

Alternatively, other researchers have contended that the variance of the distribution should be used to inform this cutoff. For example, Hwang et al. contend that genes which fall below an FDR q-value of 0.01 should be considered dosage compensated. As such, their research found that very few genes met this criteria, providing evidence against compensatory mechanisms for regulating gene dosage. While this approach is perhaps more justified by the distribution of the data, one could argue that this approach is too conservative, as only statistical outliers will survive the filter.

In any case, when we compare these two methods with real data, we find discrepancies between results found in GRO-seq and RNA-seq. The result is paradoxical as nascent transcription has a lower median fold change across the chromosome suggesting a larger magnitude of apparent dosage compensation in nascent transcription than in steady state RNA. Yet the central dogma requires a gene to be transcribed before being processed into mature RNA, leading to a contradiction.

One additional observation from the naive analysis is worth mentioning. The presence of a trisomic sample can in and of itself affect parameter estimation within the differential analysis pipeline, even when the trisomic sample isn't part of the final comparison. For example, comparing the D21 son to his father results in two different sets of significant gene calls, depending on whether the trisomic sample was included in the upstream processes (Supplemental Fig. 26).

Overall, the naive analysis leads to contradictory and incorrect conclusions. The discrepancy between GRO-seq and RNA-seq is likely technical. A much larger fraction of the genome is transcribed than is stable, which leads to GRO-seq having a lower coverage per position than an equivalently sequenced RNA-seq data set.

Finally, the fact that differential expression results vary depending on whether the T21 individual is included in the analysis or not suggests that the trisomy data specifically is undermining the analysis pipeline at some step. Thus we turn our attention to dissecting the analysis pipeline, using DESeq2 as a representative technique, to identify how T21 influences the algorithm's results.

3.3.2 Simulations reveal technical basis of reduced fold change calculations in trisomic datasets

To examine carefully the impact of trisomy data on the standard differential expression analysis pipeline, we needed to know *a priori* the correct answer. To achieve this, we simulated T21 and D21 datasets, using the D21 datasets as a reference. Briefly, we used the D21 child and the modeling and parameters (such as variance) utilized by DESeq2 to create artificial gene count tables (Fig. 3.1B, see Materials and Methods). The simulated T21 individual was generated in the same way, but now all genes on chromosome 21 are at a 1.5x increase from the simulated D21 individual. We then run our analysis pipeline (Fig. 3.2A) on the simulated data. Briefly, we subjected these samples to the DESeq2 pipeline to calculate significance and fold changes for all genes, and then calculated distributions and median fold change values for each chromosome. Moreover, we ran the simulation several times modifying the read depth, replicate number, and variance of the counts to test the effect of these variables on both differential expression and fold change estimation.

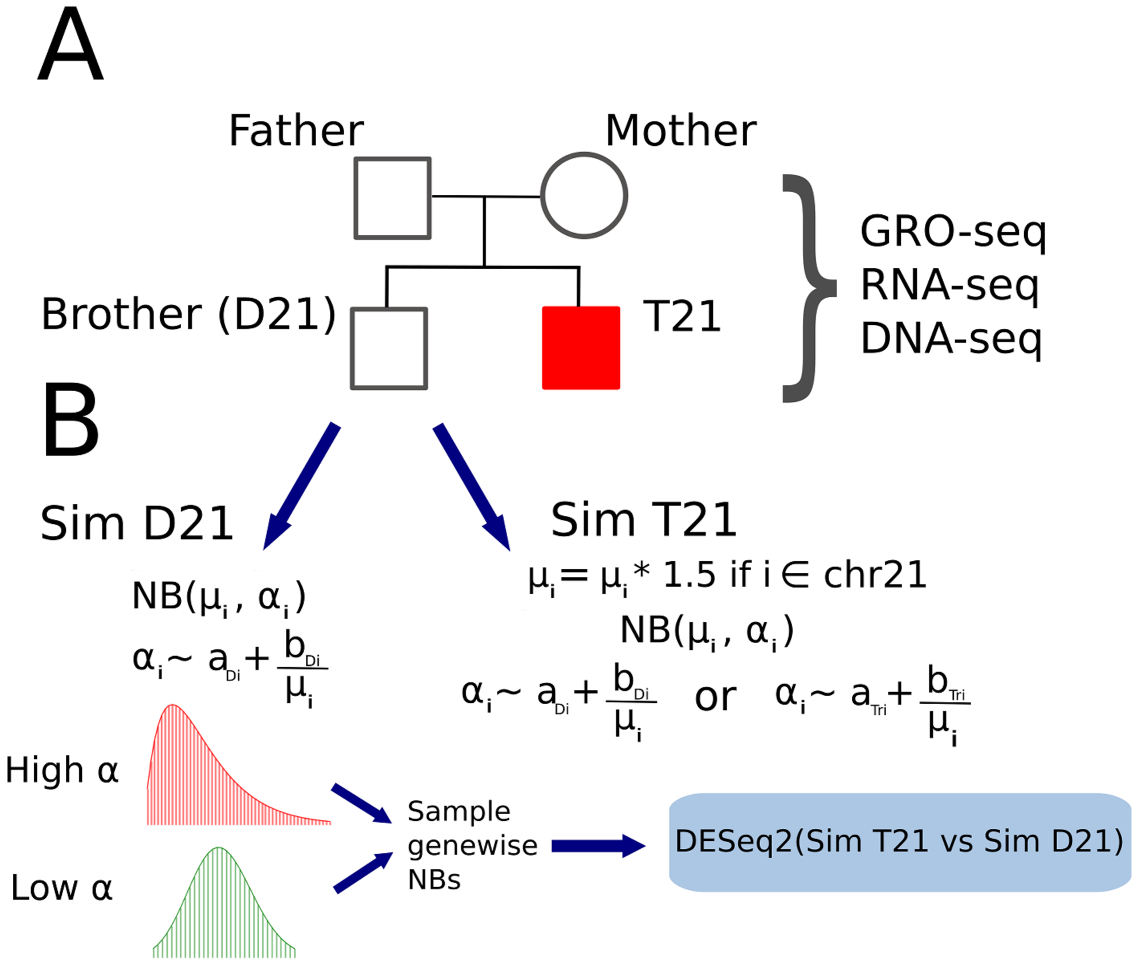

A
Father
Mother
GRO-seq
RNA-seq
DNA-seq
Brother (D21)
T21
B
Sim D21
Sim T21
$\mu_i = \mu_i * 1.5$ if $i \in$ chr21
$NB(\mu_i, \alpha_i)$
$\alpha_i \sim a_{Di} + \frac{b_{Di}}{\mu_i}$
$NB(\mu_i, \alpha_i)$
$\alpha_i \sim a_{Di} + \frac{b_{Di}}{\mu_i}$ or $\alpha_i \sim a_{Tri} + \frac{b_{Tri}}{\mu_i}$
High α
Low α
Sample genewise NBs
DESeq2(Sim T21 vs Sim D21)

Figure 3.1: **Summary of cell line generation and data simulation** (A) Pedigree depicting the relationship of our samples. Lymphoblastoid cell lines (LCLs) were derived from each of the individuals. Libraries for GRO-seq, RNA-seq, and DNA-seq were generated from these cell lines for downstream analysis (B) Simulations generated from the D21 child. The RNA-seq datasets from this individual were averaged together to inform the mean counts (μ) for each gene i. The hyperparameters a (termed asymptotic dispersion) and b (termed extra-Poisson noise) are used to inform the genewise dispersion of each negative binomial (NB) distribution. New read datasets for each gene were then generated by random variate sampling from these distributions. For trisomic genes, μ is first multiplied by 1.5, ensuring that calculated fold change estimates between trisomic and disomic genes should yield an expected distribution around 1.5, modulated by dispersion. Varying hyperparameters were used to generate multiple simulated datasets.

3.3.2.1 Sequencing depth and read counting methodologies

We first noticed that the most dramatic instances of apparent dosage compensation in our naive analysis disappear in the simulations. Specifically, highly expressed chromosome 21 genes with a large number of genomic repeats, such as ribosomal genes, often appeared to be at typical expression levels in our real datasets (Supplemental Fig. 35), but not in our simulated data. This suggests that repeat regions shared between chromosome 21 and other chromosomes are sensitive to the mapping strategy. These reads can be sponged away from chromosome 21 genes, resulting in a lower fold change estimation. This is, of course, dependent on the employed mapping strategy and how multiple mapped reads are handled. We found that a combination of a minimum read cutoff and masking repeat regions or removing multi-mapping reads effectively removed many of these genes as false positives. We suggest that users filter these genes from their list prior to differential expression analysis, either by masking repeat regions before counting reads or by manually removing genes with a high number of genomic repeats before subsequent analysis.

As mentioned previously, nascent transcription has typically lower overall counts per gene than RNA-seq when the two protocols are sequenced to roughly equivalent depths. This suggests that overall read depth may be an important factor. Consistent with this, we noticed that the lower fold change estimates correlated with low expression levels (Supplemental Fig. 36). Therefore we simulated datasets with varying depth, ranging from 0.1 times to 3 times the depth of the original datasets. We found that decreasing the depth of these simulated datasets resulted in a decreased fold change estimation for many genes, resulting in reduced MFC estimates (Fig. 3.2C). While increasing sequencing depth can help alleviate this effect, additional sequencing is not an option when reanalyzing public data sets. Furthermore, increased sequencing depth can be cost prohibitive and is not expected to fully alleviate the issue, as with increased depth unexpressed genes are more likely to have reads assigned to them due to noise. Thus we suggest users employ an additional minimum coverage filter to remove these genes as potential false positives.

3.3.2.2 Size factor calculation for sample normalization

The first step after counting reads is to normalize the data between libraries. Normalization accounts for differences in sequencing depth between samples and is crucial to proper differential analysis. DESeq2 utilizes a median-of-ratios method to find a normalizing "size factor" for each library[69]. In short, DESeq2 utilizes the assumption that the majority of genes are similarly expressed from sample to sample. By calculating the ratio between the count of a gene in one sample versus the mean count in all samples, and by finding the median of these ratios, samples can be effectively normalized to those genes which are most likely to remain unchanged in all samples. We thus wondered if this normalization method was being influenced by the trisomic condition of one of our samples, which would subsequently affect fold change calculations.

To investigate the impact of trisomy on size factor estimation, we removed chromosome 21 from both the real and simulated data sets. In both cases, chromosome 21 genes had only a minimal effect on size factor calculation (Supplemental Fig. 31), consistent with the relatively small size of chromosome 21 (1%) compared to the rest of the genome. Importantly, this result was robust to overall sequencing depth, which we showed by modulating the sequencing depth of our simulated datasets. So while the empirical result suggests including chromosome 21 genes in size factor estimation has little effect on the results, we nevertheless recommend removing chromosome 21 genes for the size factor calculation, as this is more consistent with the theory behind the median-of-ratios method.

3.3.2.3 Dispersion estimation and sample replication

We next investigated the effects of the model fitting process of DESeq2 on fold change estimation in T21 cells. Gene expression is estimated by fitting a negative binomial distribution with two parameters: the mean, and the dispersion (both of which inform the variance of the distribution). Both of these values can be inferred directly by first calculating the mean expression level for each group of replicates, and determining the value of dispersion which provides the best negative binomial fit for the data (a process known as maximum likelihood estimation, or MLE). However, this method is susceptible to error at low expression genes or with a low number of replicates, as systemic noise begins to dominate. More optimal methods (such as DESeq2) employ a Bayesian process which allows for information sharing across multiple genes and replicates. In particular, DESeq2 uses the assumption that genes with similar expression levels will exhibit similar dispersions. Information can thus be shared across genes and samples to more accurately estimate gene-wise dispersion, which increases the fidelity of fold change estimation and dispersion calls even within noisier datasets. However, if this assumption fails (i.e., if a cluster of similarly expressed genes has higher dispersion than expected), the resulting fits will not accurately reflect the underlying biology at these sites. We thus endeavored to answer this question: does the presence of a T21 sample affect the model estimation steps of differential analysis?

The default method used in DESeq2 for calculating gene-wise dispersion involves starting with the maximum likelihood estimate of dispersion, plotting these values against expression levels, and then fitting an asymptotic curve in the form of $y = a + b/x$ [69]. Here, a and b are the fitting parameters, y is the dispersion estimate, and x is the gene expression level. The parameters a and b control the two ends of the curve; for low-expression genes, the parameter b will be more meaningful, leading to higher dispersion estimates for the fitted curve at these genes (a phenomenon referred to in this study as extra-Poisson noise). For high expression genes, the b/x term trends toward 0, and thus the fitted curve will asymptotically approach the value of parameter a (a trend referred to in this paper as asymptotic dispersion). The resulting fitted curve is then used to inform one more round of dispersion estimation, effectively shrinking gene-wise dispersions towards the fitted curve. An increase in either parameter increases the value of the initial dispersion estimate, but the effects are felt asymmetrically depending on the expression level and the amount of information available to each gene (i.e., replicates and sequencing depth).

To determine the effects of T21 on dispersion estimation, we extracted the fitting parameters used in dispersion estimation from our datasets, with and without chromosome 21 genes, and when an equivalent number of random genes were removed from the other chromosomes (see Materials and Methods). We observed an increase in both fitting parameters relating to gene-wise dispersion when trisomic genes were included (Supplemental Fig. 34). But how does this increase in the initial dispersion fit affect differential calls and fold change estimates?

We sought to quantify and demonstrate how this increase in dispersion would affect differential analysis for genes on chromosome 21; as such, we generated datasets sampled from negative binomial distributions with the same means, but varying dispersion values for chromosome 21 genes in the T21 simulations. Our simulated datasets showed a MFC near 1.5, though no chromosome 21 genes were considered significant (Fig. 3.2C, padj $< .01$). We note that even when using a simulated dataset expressly designed to produce a MFC of 1.5, 54.1% of genes were found below expected levels (142/262 chromosome 21 genes below 1.5 FC), a percentage comparable to that observed in our real dataset comparisons. Furthermore, when we compared across a wide range of parameter values to model changes in dispersion estimates, we noted higher dispersion parameters resulted in decreased fold change estimations for chromosome 21 genes (Fig. 3.2C), consistent with the results observed in our real data (see Fig. 3.2B, Materials and Methods). In particular, genes with low expression levels are greatly affected; fold changes were consistently distributed around 1.0, both in simulated and real datasets for these genes (Supplemental Fig. 36).

When higher values for asymptotic dispersion and extra-Poisson noise were employed, the distribution of fold-changes drastically shifted towards 1.0, albeit with a wider spread (MFC $=$ 1.28) (Supplemental Fig. 38). This too resulted in no significant calls for chromosome 21 genes. Both fold change estimation and significance calls could be partially compensated for in simulated data when more replicates and depth were added to the simulation (Supplemental Fig. 38, Fig. 3.2C). Even with higher dispersion values, an increase in replicates and read depth resulted in a MFC of 1.43, and 126/262 differentially expressed chromosome 21 genes in RNA-seq simulations. In general, more replicates and read depth will allow users to overcome the effects of systemic noise. We also note that this serves as a potential explanation for the disparate results reported in previous studies; recent research which reported no dosage compensation integrated several available datasets, resulting in more confident fold change estimations[50].

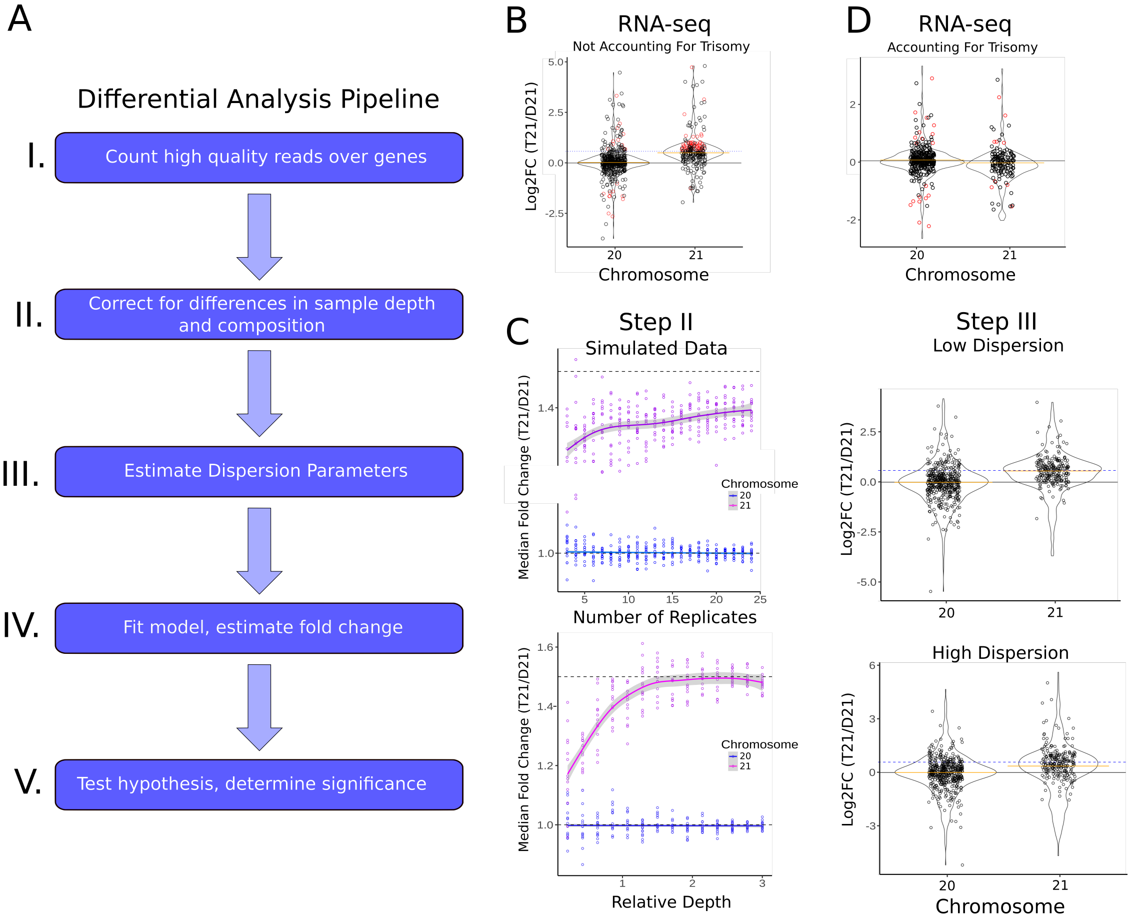


Figure 3.2: **Fold Change distributions of RNA-seq and GRO-seq datasets**(A) Pipeline of differential analysis. Variations at any step has the potential to increase or decrease fold change calculations for chromosome 21 genes (See also Supplemental Fig. 30, 29,34) (B) Naive differential analysis of Real T21 and D21 datasets. For chromosome 21 genes, the median fold change is 1.41, slightly below the expected level of 1.5 (C) Effects of shifting parameters of simulated datasets. Top-Left: simulated datasets with varying numbers of replicates (asymptotic dispersion=.01, extra-Poisson noise=1). Bottom-Left: simulated datasets with varying levels of depth (asymptotic dispersion=.01, extra-Poisson noise=1). Top-Right: Violin plots showing fold-changes of simulated datasets when dispersion parameters are low (asymptotic dispersion=.01, extra-Poisson noise=1). Bottom-Right: Violin plots showing fold-changes of simulated datasets when dispersion parameters are high (asymptotic dispersion=.05, extra-Poisson noise=30). (D) Real data violin plots showing fold-changes after applying adjustments for each step in the pipeline. Results are consistent with no dosage compensation in T21 datasets

3.3.2.4 Fold change shrinkage and hypothesis testing

The final steps of differential expression analysis is fold change estimation and hypothesis testing. Here, DESeq2 provides the option to once again utilize a Bayesian method for fold change estimation, using a prior distribution centered around a fold change of 1. In effect, this method shrinks fold change estimates around 1. As with dispersion estimation, the resulting shrinkage effect is stronger for low expression genes. These estimates (known as maximum a posteriori, or MAP estimates) are generally considered more reliable than MLE calculations of fold change for low expression genes[69]. However, this assumption can fail if the prior distribution does not represent the underlying biology, as is the case for trisomic genes. Researchers who utilize MAP estimates will thus note more genes which appear dosage compensated, although this apparent "compensation" is mainly due to the effects of the prior distribution. In general, we recommend users exercise caution when interpreting MAP estimates as evidence for dosage compensation; genes which experience strong fold-change shrinkage should be filtered out from analysis, or MLE calculations should be used instead.

Hypothesis testing in differential analysis is performed using the default null hypothesis that each gene's expression levels are equal in both test groups. Significance calls in our original analysis were thus of limited use for identifying dosage compensated genes, as the significance test only identified genes which were unlikely under the original null hypothesis. However, for chromosome 21 genes, we would expect that the fold-change should be centered around 1.5, such that significant genes represent those which are unlikely to occur within the trisomic background[7]. Identifying dosage compensated genes thus requires adjusting either the null hypothesis or the input data. We present both of these alternatives as potential avenues for future researchers.

The first method is to change the alternative hypothesis of the fold change for chromosome 21 genes to 1.5, such that significant genes are those which deviate from this expectation. Under these tests, significant gene calls which fall below a fold change of 1.5 are potential candidates for dosage compensation. In both our real and simulated datasets, we find that the majority of chromosome

21 genes are not considered significant when using this method (Supplemental Fig. 29). In other words, even for genes which have a fold change below 1.5, only significant calls can be interpreted as potentially dosage compensated. We note that this method cannot reliably utilize the MAP estimates of fold change however, as the prior distribution does not reflect the new alternative hypothesis.

The second method is to perform another normalization step prior to differential analysis. In DESeq2, the ploidy number of each gene in each sample can be loaded into a normalizing matrix. The resulting read counts are normalized both by the size factor of the library and the ploidy number of the gene. Subsequent fold change shrinkage and hypothesis testing can thus utilize the default parameters, as even trisomic genes are expected to exhibit a fold change of 1.0 under these conditions. Significant genes with a fold change less than 1.0 are then potentially dosage compensated. Chromosome 21 genes exhibited a MFC of 0.96 under these new conditions in RNA-seq (Fig. 3.2D). Furthermore, 1009 genes were considered differentially expressed, of which only 5 were on chromosome 21 (out of 143 total chromosome 21 genes, after read count filtering). In GRO-seq, chromosome 21 genes had a MFC of 0.97, with 3820 genes differentially expressed, of which 20 genes fell on chromosome 21 (out of 144 total chromosome 21 genes, after filtering) (Supplemental Fig. 32, 27). We note that these proportions are similar even when we compare two disomic individuals; the distribution of fold-changes and the number of differential genes are consistent so long as the reads are normalized by the ploidy number (Supplemental Fig. 26). In our simulated data, no such genes were detected, even in datasets with high systemic noise (Supplemental Fig. 37). Normalizing by ploidy is additionally advantageous, as MAP estimates of fold change can be utilized for visualization and downstream analysis.

3.3.3 Reduced fold change on chromosome 21 are consistent with identified eQTLs

After these analyses, we concluded that both transcription levels (GRO-seq) and expression levels (RNA-seq) of nearly all trisomic genes were proportional to DNA dosage. These results further underline recent research asserting that dosage compensation is not prevalent in mammalian trisomy events. However, we noted that there remained a small number of high expression chromosome 21 genes that were present at typical levels in RNA-seq and GRO-seq despite each of our corrections (Fig. 3.2D, Supplemental Fig. 37, 27). Due to the small number of dosage compensation candidates which survived our technical corrections and the lack of a known molecular mechanism for dosage compensation, we reasoned that there may be a genetic basis for the reduced expression of these leftover genes. We hypothesized that a reduction of expression due to the presence of an eQTL may serve as one possible explanation for these genes. To explore this possibility, we performed DNA-seq and SNP identification for all our family members, and compared these results with expression data catalogued in dbSNP[104].

We reasoned that eQTLs which reduce chromosome 21 gene expression could be present in the T21 sample. In this scenario, we expect the variant to be identifiable in one of the parent samples (Fig. 3.3A). In RNA-seq, we found that the reduced expression of three of these genes (*CLIC6*, *ITSN1*, *C2CD2)* could potentially be explained by these polymorphisms (Fig. 3.3B). These SNPs are also consistent with the transcription levels identified in GRO-seq (Fig. 3.3B).

While the remaining genes did not have explanatory SNPs identified in dbSNP, it is possible that this is simply due to a lack of data; additional polymorphisms may later be identified as eQTLs as new data is generated. We explored this possibility by comparing the transcription and expression levels of the parental samples. We reasoned that differences in these two samples were potentially indicative of allelic variation in transcription levels which were subsequently inherited by one of the two children. In GRO-seq We found that 9 of the remaining 20 genes were also differentially transcribed between the parental samples, suggesting that the T21 transcription levels may additionally be modulated by an inherited parental haplotype (Supplemental Fig. 32).

Altogether, we found reasonable explanations for nearly every gene which fell below expected levels in T21 (Fig. 3.3C). These results were consistent in both RNA- and GRO-seq (Supplemental Fig. 32), and contend that dosage compensation is not a primary mechanism of regulation for T21 genes.

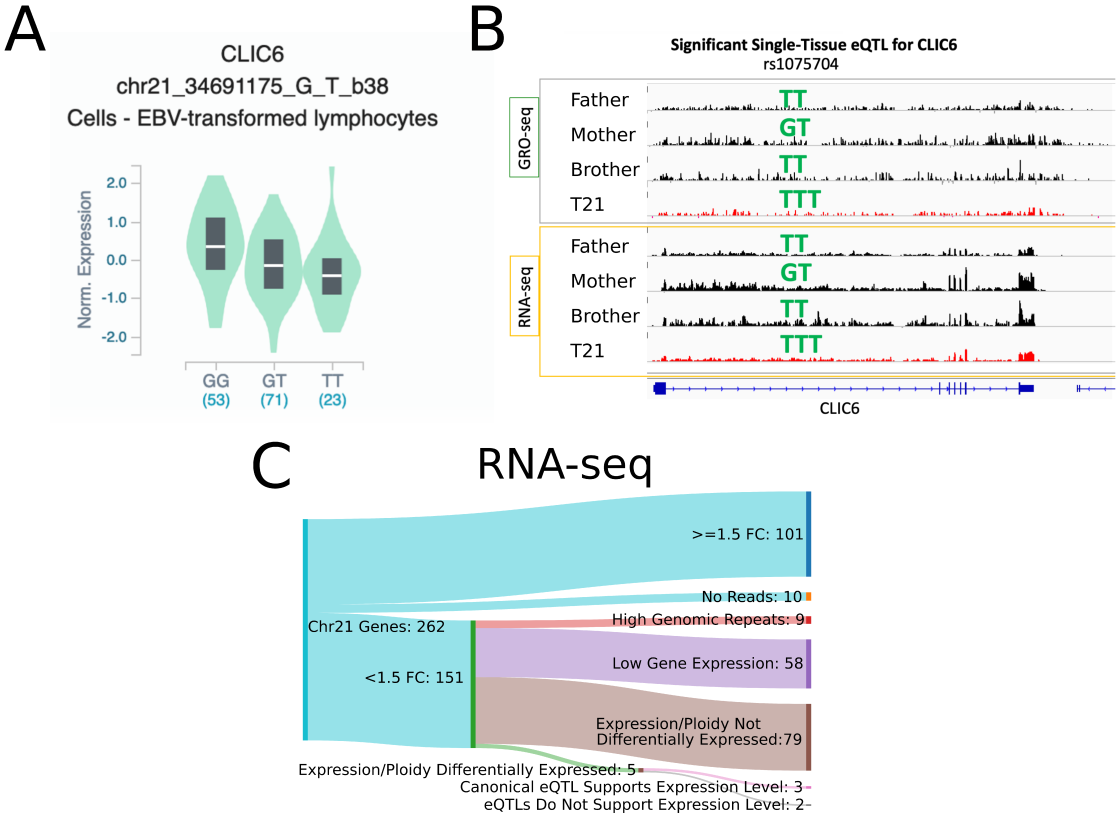


Figure 3.3: **Alternative explanations to disparate fold change estimates** (A) Example boxplot indicating relative expression of the gene *CLIC6* with one eQTL. (B) Genome viewer tracks for the gene *CLIC6* for all four family members, in GRO-seq (top) and RNA-seq (bottom). The T21 track is indicated in red. The allelic makeup of the eQTL in (A) is indicated by the green text above each track. (C) Sankey diagram depicted the filtering process of our RNA-seq analysis. The initial 151/262 genes identified as potentially dosage compensated can alternatively be explained by genomic repeats, high variance from low expression genes, or technical artifacts related to failing to normalize the data to the ploidy number. Remaining genes can be explained by the presence of eQTLs (See also Supplemental Fig. 32).

3.4 Discussion

Our results uniformly suggest that dosage compensation in T21 is rare, if not completely absent, in transcriptional and post-transcriptional regulation. Simulated read counts generated from real data showed that low sequencing depth and replication, along with increased dispersion estimates of trisomic count data, can skew fold change estimates and lead to inaccurate significance calls. These two effects are especially prevalent in low expression genes. When these effects are accounted for, the remaining candidate genes for dosage compensation are more readily explained by allelic variation which gives rise to reduced transcription and expression levels. Our research thus agrees with recent studies which suggest that there is no reduction in RNA expression levels in T21, and further extends this conclusion to RNA transcription rates as well.

Our analysis relied on adjusting the standard differential expression analysis to accommodate for the presence of trisomic samples. Many analysis pipelines (including the popular DESeq2 algorithm) use a null hypothesis that the fold change between two samples is 1, and are thus ill-equipped for fold change estimation and differential analysis in trisomic backgrounds. These native settings can be easily adjusted, leading to a more reliable analysis. In summary:

(1) Apply a minimum read coverage filter, depending on read depth. (a read depth filter of 30 was used in this study)

(2) Mask repeat regions or remove multi-mapping reads for read counting

(3) Remove chromosome 21 genes for size factor calculation

(4) For noisy samples, increase sequencing depth or replication

(5) Adjust null hypothesis, or normalize read count by ploidy number

While T21-driven dosage compensation may be rare, comparisons at the population level may erroneously suggest otherwise. Our research here was limited to only one family of related individuals; as such, sequence variability between these individuals is more limited. In this microcosm, we observed only a minority of genes whose reduced expression could be explained by parental SNPs. However, previous research found that population estimates between disomic and trisomic individuals indeed indicate that many genes are expressed below a 1.5 fold change expectation[89]. These genes are possibly subject to more selection pressure in trisomic backgrounds, as overexpression can lead to embryonic lethality. We anticipate that future studies can expand the analysis performed here to identify gene expression levels and SNPs which are more prevalent in individuals with Down syndrome. As the available data increases, we hope that future research employs the recommendations here to avoid inaccurate conclusions about fold change estimates.

3.5 Materials and Methods

3.5.1 GRO-seq and Library Preparation Methods

Run-on reactions were performed as in [24]. In brief, ice-cold isolated nuclei (100 μL) were added to 37°C 100 μL reaction buffer (Final Concentration: 5 mM Tris-Cl pH 8.0, 2.5 mM $MgCl_2$, 0.5 mM DTT, 150 mM KCl, 10 units of SUPERase In, 0.5% sarkosyl, 500 μM rATP, rGTP, and Br-UTP, 2 μM rCTP). The reaction was allowed to proceed for 5 min at 37°C, followed by the addition of 23 μL of 10X DNAseI buffer, and 10 μL RNase free DNase I (Promega). RNA was extracted twice with Trizol, washed once with chloroform, and precipitated with 3 volumes of ice-cold ethanol and 1-2 μL GlycoBlue. The pellet was washed in 75% ethanol before resuspending in 18 μL of DEPC-treated water. Libraries were prepared similar to [49]. In brief, RNA was treated with 2ul NEB Fragmentation Buffer at 94 degrees for 5 min. The RNA was then buffer exchanged via BioRad P-30 (or a G-25) column per manufacturer's protocol. Next, 2 ul DNaseI and 5 ul of 10X RQ1 DNase buffer and water was added to create a 1X final concentration of the buffer. After , incubatetion at 37ºC for 10min, 5ul DNAse stop solution was added and the reaction was place at 65ºC for 5 min. Beads were prepared by washing in pre-wash buffer. Nascent RNA was extracted and fragmented by base hydrolysis in 0.2 N NaOH on ice for 10–12 min, and neutralized by adding a 1× volume of 1 M Tris-HCl pH 6.8. Fragmented nascent RNA was purified using Anti-BrdU beads and ligated with reverse 3′ RNA adaptor (/5Phos/GAUCGUCGGACUGUAGAACUCUGAAC/3InvdT/), and BrdU-labeled products were enriched by a second round of Anti-BrdU bead binding and extraction. For 5′ end repair, the RNA products were treated with tobacco acid pyrophosphatase (Epicenter) and T4 polynucleotide kinase (NEB). 5′ repaired RNA was ligated to reverse 5′ RNA adaptor (5′ UGGAAUUCUCGGGUGCCAAGG) before being purified by a final round of Anti-BrdU bead binding and extraction. RNA was reverse transcribed using 25 pmol of RP1 primer (5′AATGATACGGCGACCACCGAGATCTA CACGTTCAGAGTTCTACAGTCCGA). The product was amplified 15 ± 3 cycles and products >150 bp (insert > 70 bp) were size selected with 1X AMPure XP beads (Beckman) before being sequenced.

3.5.2 Cell Culture of LCLs

The lymphoblastoid cell lines (LCLs) of 4 individuals were obtained from Translational Nexus Biobank (COMIRB 08-1276), University of Colorado School of Medicine, JFK Partners. Each member of a family was given a pseudoname. The blood from this family was not available, so we used lymphoblastiod lines derived from blood were seeded in upright T-25 suspension flasks with 10 ml RPMI (10% FBS, 1X L-Glutamine, 1X Penicillin/Streptamycin). These were passaged approximately every 2 to 3 days, by pelleting the cells via centrifugation (300 x g, 5 min) and resuspension. Cells were grown to an approximate density of 1 million cells per ml, before being harvested for analysis.

3.5.3 RNA-seq

RNA was isolated from the cells via Trizol extraction. NEBNext rRNA Depletion kit was used to remove rRNA. NEBNext Ultra II RNA was used to transform the RNA into an RNA-seq library. The product was amplified 15 ± three cycles and products >150 bp (insert > 70 bp) were size selected with 1X AMPure XP beads (Beckman) before being sequenced.

3.5.4 Whole genome sequencing

DNA was isolated from LCLs, fragmented, size selected, and prepared for sequencing via adapter ligation. Subsequent libraries were sequenced on a Highseq 2000 to an approximate depth of 40x. For specific sample preparation information, see Supplemental Table 2 at https://github.com/Dowell-Lab/DS_Normalization. Conversion to fastq: Some samples were sequenced at Illumina were received as one bam file per person, which were split by Read Group (RG) into individual lane bam files using samtools (version 0.1.19) view. Those files were then converted to fastq using bedtools (version 2.16.2) bamtofastq. All sample files were received as fastq files.

Reads were mapped to hg38 using bowtie(2.0.2) with the setting –very-sensitive and using SM, PU, RGID, PL. SAM files were converted to sorted BAM files using samtools(1.2) view and sort. All files for one individual were merged using Picard tools(1.72) MergeSamFiles. Bam files were sorted and dupiciates were marked using Picard tools(1.72) SortSam and MarkDuplicates.

3.5.5 Whole genome variant calling

GATK Version 3.3-0 was used for variant calling. BAM files were realigned using IndelRealigner with optional flags –known Mills_and_1000G_gold_standard.indels.hg38.vcf-known 1000G_phase1.indels.hg38.vcf [Note: the realignment table was created using all the merged files with RealignerTargetCreator optional flags -known Mills_and_1000G_gold_standard.indels.hg38.vcf -known 1000G_phase1.indels.hg38.vcf]. Then, BaseRecalibrator and PrintReads was used to recalibrate the bases (optional arguments –knownSites Mills_and_1000G_gold_standard.indels.hg38.vcf -knownSites 1000G_phase1.indels.hg38.vcf -knownSites dbsnp_138.hg38.vcf). HaplotypeCaller was used to call haplotypes two times; once with the optional flags -nct 4 –emitRefConfidence GVCF –dbsnp dbsnp_138.hg38.vcf –variant_index_type LINEAR –variant_index_parameter 128000 and another time with those flags and the flag -ploidy 3. Then vcf files were created for each trio (mother, father, child) using GenotypeGVCFs and the diploid gvcf files. There was not a straight forward way to created vcf files that had chr21 as triploid in the children. Therefore we first created vcf files for each trio (mother, father, child) using GenotypeGVCFs and the diploid gvcf parent files and the triploid gvcf child files, but this makes all chromosomes triploid. So, to create a VCF that contained chr21 as triploid in the children, we created our own code in python to combine the two types of family vcf files. This program combined the family vcf files by taking any lines that started with “chr21_” or “chr21” from the vcf files with the triploid child variants, and all lines that did not start with “chr21_” or “chr21” from the family diploid vcf files.

3.5.5.1 Discovering de novo mutations

VariantRecalibrator was used to create tables to recalibrate the SNPS using the optional flags -resource:hapmap, known=false, training=true,truth=true, prior=15.0 hapmap_3.3.hg38.vcf -resource:omni, known=false, training=true, truth=true, prior=12.0 1000G_omni2.5.hg38.vcf -resource:1000G, known=false, training=true, truth=false, prior=10.0 1000G_phase1.snps.high_confidence.hg38.vcf -resource:dbsnp, known=true,training=false,truth=false,prior=2.0 dbsnp_138.hg38.vcf -an QD -an MQ -an MQRankSum -an ReadPosRankSum -an FS -an SOR -an DP -mode SNP -tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0. Then ApplyRecalibration was used at each of the ts_filter_level. Then those vcfs were used to apply VariantRecalibrator to the indels using the optional flags -nt 4 –maxGaussians 4 -resource:mills,known=false,training=true, truth=true,prior=12.0 Mills_and_1000G_gold_standard.indels.hg38.vcf -resource:dbsnp, known=true, training=false, truth=false, prior=2.0 dbsnp_138.hg38.vcf -an QD -an DP -an FS -an SOR -an ReadPosRankSum -an MQRankSum -mode INDEL -tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0. Then ApplyRecalibration was used at each of trance levels for INDELs. CalculateGenotypePosteriors was used with the flag –skipPopulationPriors to label the de novo variants. (Without popultation was used becuase population could not be used with ploidy of 3 in GATK.) VariantFiltration was then used to remove variants with a genotype quality of less than 30 (–genotypeFilterExpression "GQ < 20") VariantAnnotator was used to annotate potiential de novo variants and SelectVariants was used to create vcf files of potential de novo locations.

3.5.5.2 Fold change and allele percentage graphs

To determine if the children had an extra copy of chromosome 21, the number of reads on chromosome 21 and 22 were calculated using pysam's get_index_statistics. We then did fold change of reads for each family's child/mother. To determine the reads per 1000 kb window we counted reads using bedtools coverage. We then calcuated the FPKM for each window and did fold change graphs over the WT mother.

To determine the allele percentages on chromosome 21, 22 and M, scikit-allele was used to extract the all variants as seen by VariantRecalibrator (unfiltered variants). Tranch 90 was used for Chromosome 21 and 22. Tranch 100 was used for Chromsome M because using any other tranch resulted in all Chromosome M variants filtered due to high depth. The heat map was plotted in python3 with plotly.

To determine the if the Maternity and Paternity of the samples was correct the SNPs found on M or Y were used. Again tranch "90.0" was used for chrY, but we had to use tranch "100.0" for chrM as all other tranches removed all chrM positions. First, we created the "world" of positions for each percentage by determining the all variants from the chromosome that had a defined genotype for both the parent and the child (regardless of what that genotype was). For the denominator we calculated all non-reference variants that were in the world. For the numerator we determined the number of non-reference variants that exist in the child that were also non-reference in the parent. For chrY paternity graphs we went the extra step of removing any variants from the calculation that were found in any of the 6 mothers. This is because many SNPs are artificially called in women on ChrY most likely due to mapping errors.

3.5.6 Simulation of trisomy and disomy datasets

Reads were simulated using the negative binomial model as part of the scipy (v1.8.0) python package (v3.6.3). The negative binomial means for each gene were estimated by averaging the counts for the D21 child samples in RNA-seq. For the disomic simulations, these means were used to directly inform the negative binomial parameters. For the trisomic simulations, the means for chromosome 21 genes were first multiplied by 1.5. The negative binomial instance is parameterized as follows:

$$
\begin{aligned}
NB(n,p) : p &= \mu/\sigma^2 \\
n &= \mu^2/\sigma^2 - \mu \\
\sigma^2 &= \mu + \alpha\mu^2 \\
\alpha &\sim a + b/\mu
\end{aligned}
$$

Where a and b are controllable hyperparameters. For the disomic genes, we used values of .01 and 1 for a and b, respectively. For trisomic genes, we varied these values from .001 to 1.2 (for a) and 1 to 100 (for b). The negative binomials were also scaled based on the depth of the original biological samples, from 0.1 times the depth to 3 times the depth. Each simulated counts file was generated with a minimum of three replicates for the disomy 21 sample and a minimum of three replicates for the trisomy 21 sample. For details, see the git repository for these scripts at https://github.com/Dowell-Lab/DS_Normalization.

3.5.7 Mapping and visualization

FASTQ files were trimmed and mapped to the GRCh38/hg38 reference genome and prepared for analysis and visualization through our in-house pipelines. In short, resulting FASTQ read files were first trimmed using bbduk (v38.05) to remove adapter sequences, as well as short or low quality reads. Reads were mapped with HISAT2 (v2.1.0), and resulting SAM files converted to BAM files with Samtools (v1.8). Multimapped reads were filtered from these files. BedGraph files were generated using Bedtools (v2.25.0), and converted to TDF files for visualization in IGV using IGVtools (v2.3.75). Quality metrics were generated with FastQC (v0.11.8), Preseq (v2.0.3), RSeQC (v3.0.0). Figures were generated through MultiQC (v1.6).

3.5.8 Differential expression analysis

Differential transcription was performed using the DESeq2 (v1.26.0) R package (R version 3.6.3). Gene counts were generated using featureCounts (v1.6.2) from the R Subread package (v1.6.0), counting over the gene body region (+150 from transcription start site to annotated transcription end site) to avoid the $5'$ peak. In both RNA-seq and GRO-seq analyses, reads were counted over the whole gene, such that results between these two experiments should be comparable. Annotations were downloaded from RefSeq (release number 109, downloaded August 14, 2019 from UCSC genome browser). Only annotations with both RNA-seq and GRO-seq signal were considered, again to keep both analyses comparable. For featureCounts, BED6 region files were converted to SAF format with the following command: awk -F "\t" -v OFS="\t" 'print{$4, $1, $2, $3, $6}' region.bed > region.saf. Only the highest transcribed isoform of each gene was considered.

For our corrective analysis, we made use of DESeq2's normMatrix parameter. The normalization matrix was generated by assigning each gene in the analysis its ploidy number divided by 2. For all genes not on chromosome 21, this number was thus 1. For genes on chromosome 21 samples in trisomy, this number was 1.5. We also removed reads within regions of genomic repeats, set the betaPrior parameter to False, and set an expression level cutoff at the second quintile of the baseMean counts for all genes.

The script and full table of results are available at https://github.com/Dowell-Lab/DS_Normalization

3.5.9 GTEx datasets/SNP analysis

We downloaded the GTEx (v8) database of eQTLs and their associated genes for all available tissues[68]. We merged these databases and filtered out all SNPs identified in our DNA-seq experiments that were not present in the merged GTEx database. We compared the expected effects of each SNP from GTEx with the allelic ratios of each of our datasets. We then asked whether there was at least one SNP which could explain the expression level we observed in the real data when each parent was compared to the child with T21. The script and full table of results are available at https://github.com/Dowell-Lab/DS_Normalization

Chapter 4

Inferring RNA Turnover Rates in Trisomy 21 Interferon Response Pathways

The following chapter is a collaborative effort between Jacob Stanley and myself. Jacob Stanley originally came up with the linear approximation model and wrote the grant that funded this work (NIH R03HD103995), whereas I carried out the dataset generation and analyses. This chapter is adapted from a work-in-progress manuscript to be submitted 2023.

4.1 Introduction

Cells exist in an ever changing environment. This environment can rapidly shift from hospitable to harmful under viral infection. An organism's health and survival is contingent upon each cell's ability to rapidly and accurately respond to these infections through the production and detection of interferon. Interferon regulation is an extremely delicate and complex process. Signals propagated from external interferon must be interpreted and conveyed to the response machinery within the cell's interior. This response machinery includes a slew of transcription factors, which instigate a core interferon transcription program made up of antiviral and pro-inflammatory genes[88, 118]. Improper activation of this program is hypothesized to result in autoimmune disorders, such as systemic lupus erythematosus and psoriasis[10]. Conversely, if the cell is unable to activate these pathways, the organism is rendered uniquely susceptible to viral insults[7, 25].

The RNA levels from these interferon-stimulated genes (ISGs) are delicately balanced to ensure optimal cellular health. The cell has two general strategies for modulating its levels of

RNA: production and degradation (Fig. 4.1A). RNA production from these genes is primarily regulated by the JAK/STAT signaling pathway, which is activated when the interferon receptor is bound with interferon [88]. Conversely, RNA degradation is a multifaceted mechanism of gene regulation; transcriptional, post-transcriptional, and translational forces can all serve to modulate the degradation rate of some or all RNAs in a cell [47, 79, 51].

Recently, RNA production and degradation have become of interest when studying Trisomy 21 in humans. The triplication of chromosome 21 leads to a commensurate increase in RNA levels from the affected genes[50]. In particular, the triplication of four interferon receptor genes encoded on chromosome 21 has been attributed to the mild interferonopathy phenotypes observed in many individuals with Down syndrome[118, 73, 57]. The response pathways to interferon activation have been well characterized in Down syndrome; however, the dynamics of RNA production and degradation throughout the interferon response have not yet been analyzed.

Determining primary and secondary effects on RNA production and degradation throughout the interferon response necessitates a time-series analysis. However, while total RNA levels and RNA production rates can be easily measured across time, there is considerable need for improvement in methods to assess degradation across time. The current “gold standard” methods (e.g., SLAM-seq or TimeLapse-seq) require a lengthy metabolic labeling period utilizing a nucleotide analog[35]. The assay’s sensitivity for detecting low abundance or high turnover RNA species is contingent upon the concentration of the metabolic label and the length of the labeling period[102, 16].

Rather than measuring degradation directly by labeling and tracking RNA, degradation rates can be modeled and estimated mathematically. Previously, the commonly used paradigm is the “bathtub” model, in which RNA degradation is equal to total RNA levels (the “water level”) multiplied by a per-molecule degradation rate (the “drain”). Additionally, if transcription rates are known (the “faucet”), the full scope of RNA turnover can be approximated[37, 3, 15]. Knowing any two of these values across time, or when the system is at equilibrium, allows for the calculation of the third value. Naturally, however, the true values of RNA levels and transcription rates in a cell are unknowable; as such, proxy values from sequencing read count data are used instead.

These proxy values yield a proportional estimate of the true rate; as such, these estimations are interpretable relative to each other only within the same experiment.

Previous implementations of the “bathtub” model utilized read counts over exonic regions as a proxy value for total RNA levels, and read counts over intronic regions as a proxy value for transcription rates[37, 3]. While this method allowed for estimation of degradation rates from a single RNA-seq dataset, it comes with significant caveats: first, estimates are considerably biased when genes are differentially transcribed across time; second, it requires deeply sequenced RNA-seq datasets to pick up enough intronic reads for accurate rate estimation; third, this method is unsuitable for any RNAs which lack intronic reads (such as many lncRNAs). Given these considerations, this method often disagrees with estimates determined by direct metabolic labeling experiments, limiting its usefulness[15].

Recently, Blumberg et al proposed another utilization of the “bathtub” model to improve on these issues. By using RNA-seq as a stand-in for total RNA levels, and PRO-seq as a stand-in for transcription rates, Blumberg et al demonstrate that a more consistent degradation rate estimation can be achieved at coding and noncoding RNAs alike[15]. However, this method assumes no net change in total RNA, meaning its degradation calculation is amenable to steady-state conditions only. As such, the transitional dynamics of RNA degradation are not accounted for within this model. Thus, this method must be adapted in order to assess whether there are changes in RNA dynamics in response to changes in the environment.

Here, I explore an expanded model which allows for estimating RNA dynamics over time. As in the previously established steady-state model (SSM) formulated by Blumberg et al, I utilize PRO-seq as a proxy for transcription rates, and RNA-seq as a proxy for total RNA levels. By assessing the changes in these two values across time, I show that RNA decay rates can be consistently estimated throughout a time course (Fig. 4.1B). I will refer to this method as the “linear approximation” model, or LA.

Using these two models, I sought to answer two questions: first, are there significant differences in degradation estimation between the LA and SSM methods? Second, are there differences in degradation between T21 and D21 throughout the interferon response? Using interferon beta (IFN-β) in Trisomy 21 (T21) and Disomy 21 (D21) lymphoblastoids sourced from two brothers, I showcase some potential differences in degradation between the two cell lines. Furthermore, I demonstrate that the SSM and LA methods may have some disagreements at key time points.

4.2 Results

4.2.1 Identification of genes undergoing changing degradation rates

I generated data from six time points of IFN-β treatment in T21 and D21 lymphoblastoid cell lines (Fig. 4.1C). In order to study RNA dynamics throughout the primary and secondary interferon response, I focused on the early time points (<= 2 hours). I reasoned that these time points were sufficient to capture primary effects and the beginning of the secondary response pathways, such that different RNA decay pathways should manifest in the PRO and RNA read signal.

I first set out to identify candidate genes of interest which may exhibit changing degradation rates across time. I reasoned that degradation rates for any given gene would shift substantially under two main circumstances. First, when RNA-seq signal trends upwards across time while PRO-seq signal trends downward, it suggests a decrease in degradation (e.g. an increase in RNA stability). Second, if RNA-seq signal for a gene trends downward while PRO-seq trends upward, then this suggests an increase in degradation. In both cases, the SSM and LA methods should yield differing results across time, as the underlying assumption of the SSM is violated in both of these circumstances[15].

To identify candidate genes which fit one of the two criteria above, I subjected the samples to differential analysis using DESeq2[69]. I first compared all time points to their respective baseline sample (T21 or D21 0m). Genes were then broadly classified as Early-Upregulated (LFC > 0) or Early-Downregulated (LFC < 0) based on their PRO-seq signal at 30 minutes (Fig. 4.2A). Using the 120 minute PRO-seq datasets, genes were then classified as either Late-Upregulated or Late-Downregulated. Because RNA-seq signal is only substantially shifted after approximately 45

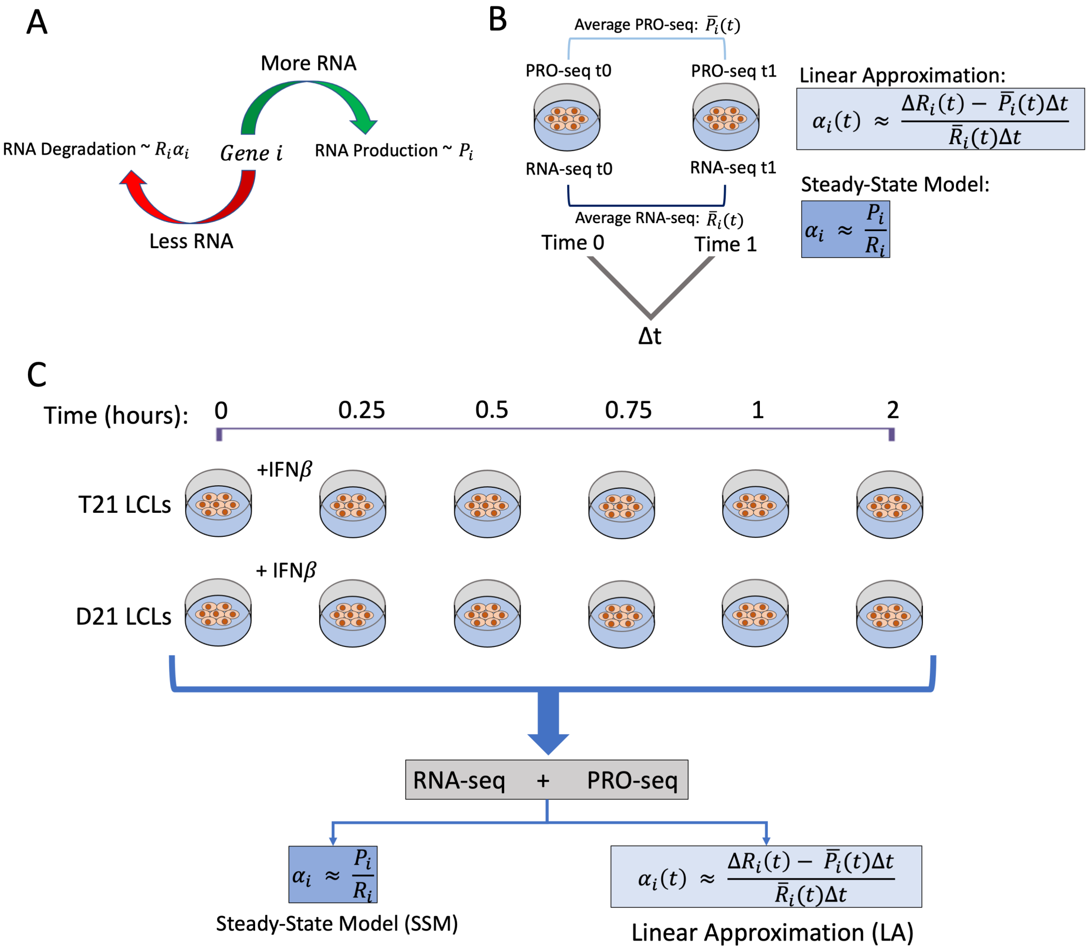

A
More RNA
RNA Degradation ~ $R_i\alpha_i$
$Gene\ i$
RNA Production ~ P_i
Less RNA
B
Average PRO-seq: $\bar{P}_i(t)$
PRO-seq t0
PRO-seq t1
RNA-seq t0
RNA-seq t1
Average RNA-seq: $\bar{R}_i(t)$
Time 0
Time 1
Δt
Linear Approximation:
$\alpha_i(t) \approx \frac{\Delta R_i(t) - \bar{P}_i(t)\Delta t}{\bar{R}_i(t)\Delta t}$
Steady-State Model:
$\alpha_i \approx \frac{P_i}{R_i}$
C
Time (hours):
0
0.25
0.5
0.75
1
2
T21 LCLs
+IFNβ
D21 LCLs
+ IFNβ
RNA-seq + PRO-seq
$\alpha_i \approx \frac{P_i}{R_i}$
Steady-State Model (SSM)
$\alpha_i(t) \approx \frac{\Delta R_i(t) - \bar{P}_i(t)\Delta t}{\bar{R}_i(t)\Delta t}$
Linear Approximation (LA)

Figure 4.1: **Summary of dataset generation and degradation rate calculations.** (A) RNA abundance can be regulated via production or degradation. Production can be approximated using PRO-seq (P), and total degradation can be approximated using a per-molecule degradation rate (α) and RNA-seq as a stand-in for total RNA levels (R). (B) Summary comparison of the two models for per-molecule degradation rate calculation. The linear approximation (LA) method utilizes the differences in RNA levels from two time points (ΔR), minus the amount of RNA produced during that time ($\bar{P}\Delta$t), normalized by average RNA levels across time ($\bar{R}\Delta$t). The steady-state model (SSM) assumes no net change in RNA across time, such that degradation can be represented as a ratio of PRO-seq and RNA-seq signal. (C) Summary diagram of experiment. Lymphoblastoid cell lines (LCLs) derived from two brothers (T21 or D21) were treated with IFN-β in a time series. RNA-seq and PRO-seq datasets were generated from each batch of cells, and degradation rates were estimated and compared using both models (LA vs SSM).

minutes, the results from the 60 and 120 minute RNA-seq datasets were joined together to classify genes as either RNA-Upregulated or RNA-Downregulated by the same criteria (see Materials and Methods).

Once these labels were assigned to each gene, I generated an UpSet plot to determine which genes were likely undergoing secondary degradation effects (Fig. 4.2B). Of the total set of genes, I identified 328 potential genes which likely had increased degradation rates, and 92 genes which likely had decreased degradation rates.

4.2.2 Are there significant differences in decay rate estimates between the LA and SSM methods?

I next subjected each of our candidate genes identified above to degradation analysis. Using the normalized count data output by DESeq2, I utilized both the SSM and LA approaches to estimate degradation for each time point (Fig. 4.1B, 4.2C). The LA model requires two time points to estimate degradation rates whereas the SSM can estimate the rate from a single time point. Consequently, to plot rates at equivalent time points, I averaged steady-state estimates for each neighboring time point. To compare the two models, I first chose to compare only the D21 degradation rates between the LA and SSM methods, using only the time intervals utilized in differential analysis (0 to 30, 30 to 60, and 60 to 120).

The example gene *Tbx21* was estimated by the pre-filtering step as undergoing a decrease in degradation rates by minute 120 (Fig. 4.2B, C). Indeed, the LA model matched this expectation well, while the SSM estimated no net decrease in degradation across time. Interestingly, both models predicted an increase in degradation near minute 45 (Fig. 4.2D).

4.2.2.1 Adjusting time points for linear approximations

When using only three intervals, differences between the SSM and LA approaches were apparent at late time points, but the models were consistent at the early intervals. However, it is possible that further profile characteristics exist when more time intervals are utilized. Thus, I

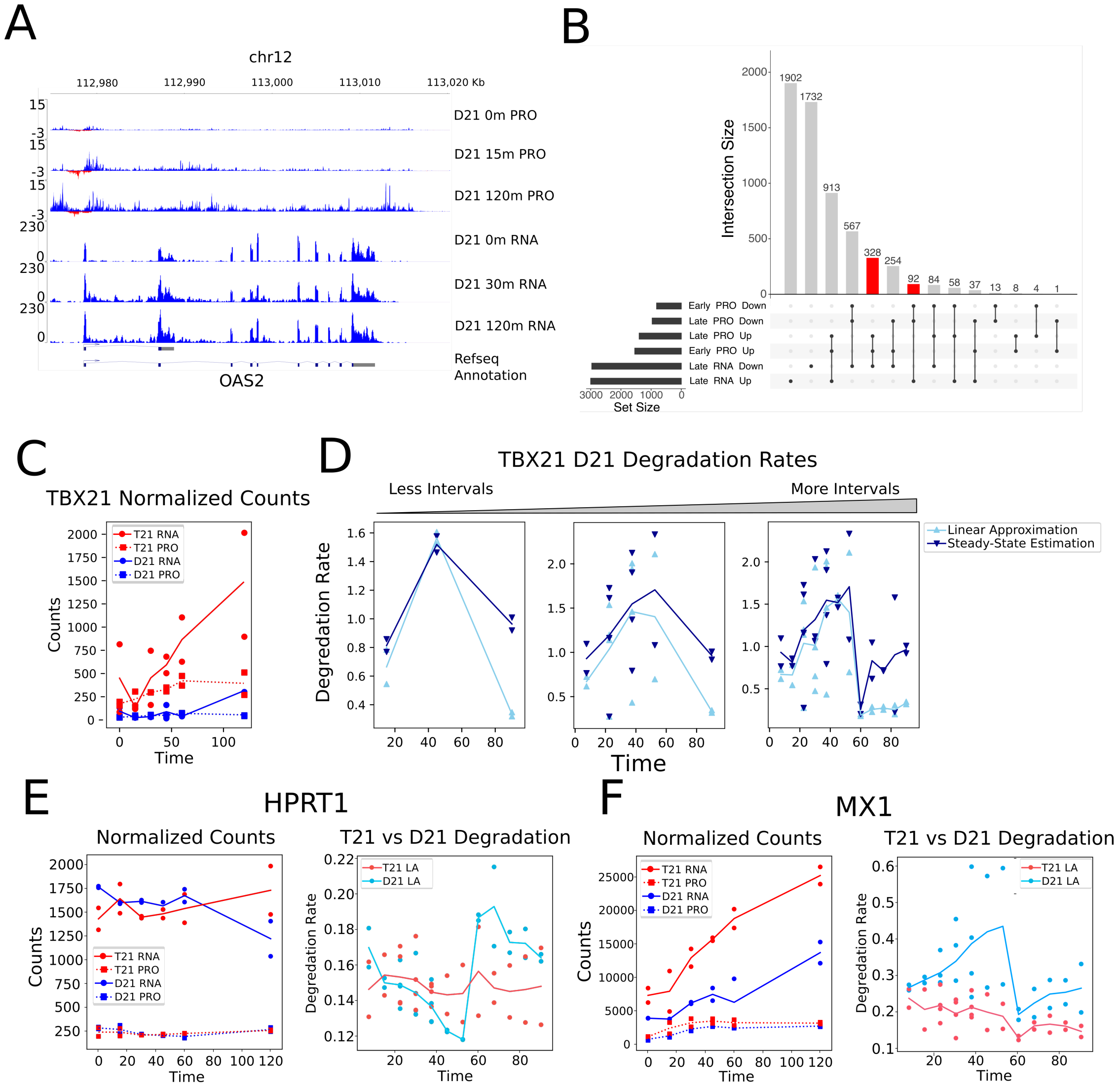


Figure 4.2: **Degradation of T21 and D21 LCLs using two competing models.** (A) Screenshot of IFN response gene *OAS2* across time. PRO-signal shifts as early as 15 minutes, whereas RNA-seq signal is more consistent across exonic reads at 30 minutes. (B) UpSet plot indicating candidate genes for degradation analysis. Gene which are likely undergoing secondary degradation dynamics will have differences between early (<45 minutes) and late (120 minutes) time points in RNA-seq. Furthermore, if the trend between PRO-seq and RNA-seq is reversed, degradation is more likely as an explanatory factor. Candidate gene sets are highlighted in red. **continued ...**

Figure 4.2: **continued** (C) Example gene counts for T21 (red) and D21 (blue) RNA-seq (solid line) and PRO-seq (dotted line) samples, plotted with replicates at each timepoint. Rising RNA-seq levels at minute 120 are not explained by changes in PRO-seq signal, suggesting differences in degradation.(D) Degradation rate calculations for *Tbx21* in D21 cells. LA (light blue) and SSM (dark blue) estimates show different estimates at later time points, but not earlier time points. This trend is more drastic with longer intervals used, but still present even with more intervals. (E) Two example genes (*HPRT1*, housekeeping gene, and *Mx1*, interferon-response gene on chromosome 21) and their respective normalized counts and degradation profiles (LA). T21 and D21 cells show potential differences in degradation at both genes. Further statistical rigor is needed to determine whether these differences are significant or a consequence of data quality.

increased the number of intervals considered, utilizing smaller time steps and more of the time series data (e.g., 0 to 15, 0 to 30, 0 to 45, and so on, see Materials and Methods)

Differences between the two methods persisted at later time points (>60 minutes), but estimates remained remarkably consistent for all earlier time points. Regardless of which intervals were used, several features of the profile of degradation rates remained consistent. Rate estimates peak near 45-60 minutes regardless of which model is used, while later estimates are noticeably lower in the LA model only. As the SSM assumes no net change in RNA across time, this result provides evidence that the different calculations are due to a secondary change in degradation.

4.2.3 Are there differences in degradation between T21 and D21 throughout the interferon response?

I next turned my attention to degradation estimates between T21 and D21 cells using all time intervals in the LA method (Fig. 4.2E). Here, I picked two candidate genes (*HPRT1* and *Mx1*) for analysis, both of which showed differences in counts between T21 and D21. The *Mx1* gene in particular is a chromosome 21 gene which is upregulated in interferon treatment, whereas *HPRT1* is a “housekeeping” gene which should remain relatively stable across time.

Interestingly, *HPRT1* showed an increase in degradation at minute 60 in D21 cells, while T21 estimates were consistent across time. Similarly, *Mx1* showed a drop in degradation at minute 45 in D21 cells alone. It is possible that differences in T21 and D21 calculations are due to differences in quality of the data, however. A more formalized statistical framework is required to determine whether there are significant differences between these or any other candidate genes.

4.2.4 Confirming consistency of both models at non-candidate genes

Lastly, to confirm that both models predicted the same values for genes which undergo no changes in degradation rates, I investigated decay rate estimates for genes which trended in the same direction across time in both RNA-seq and PRO-seq. In theory, degradation should remain relatively consistent for these genes across time. Indeed, when I investigate selected genes from this pool, I found that SSM and LA estimations were remarkably close, if not equal, for all timepoints. For example, the gene *PGK1* remains consistently highly expressed throughout the time-course. Indeed, decay rates are extremely consistent across time in both models, ranging from 0.02 to 0.07. Furthermore, the models deviate from each other by no more than 0.005 at any given time point (Supplemental Fig. 39). This result underlines the consistency of these estimates when no cellular changes to the RNA's degradation rate has occurred.

4.3 Future Directions

Thus far, I have identified a number of genes which are equivalently estimated between the steady-state and linear models, some of which deviate at later time points. In the future, I will expand this analysis to profile and cluster all candidate genes based on their degradation rates across time. Gene clusters may have unifying characteristics, allowing for the potential to infer which biological factors are driving changes in degradation (e.g., miRNA binding sites or $3'$ UTR secondary structures).

Interestingly, the differing results between the SSM and LA models suggest that the two models could be used in conjunction with each other in a statistical analysis. As the SSM assumes no net change, it can serve as a null hypothesis for degradation rates across time. The LA can then be used as the alternative hypothesis, allowing for more formalized conclusions to be drawn about degradation rates for a given gene.

All degradation rates have been inferred computationally thus far. As I expand this study, these rates will be validated using a "gold-standard" method, such as TimeLapse-seq[16, 102]. However, even after validating the LA method, there is no current statistical method for comparing degradation rates between two groups of samples. Degradation rates do not follow the same distribution as count data, given that their values are continuous (non-integer) and positively bounded. Evaluating differences between T21 and D21 degradation rates will require the construction of a statistical pipeline which accurately models these degradation rates. As part of this study, I plan to construct this pipeline, styled after other successful models such as DESeq2[69].

In all, our findings provide justification for the use of the LA model, and for the possibility of differential degradation programs between D21 and T21 cells. Such findings could further our understanding of gene dysregulation in Down Syndrome, along with identifying potential genes which further explain Down Syndrome phenotypes in the interferon response pathway.

4.4 Materials and Methods

4.4.1 Cell culture conditions

Lymphoblastoid cell lines for both T21 and d21 cells were cultured in 24 ml RPMI media supplemented with 15% FBS, 100 units/ml penicillin and 100 μg/ml streptomycin, at 37°C with 5% CO2. Cells were grown to a concentration of approximately 700,000 cells/ml in T-75 flasks before passaging. Lymphoblastoid cells grow in suspended clumps; as such, cells were dissociated by swirling the flask before counting or treatment. Cells were passaged at least twice before harvesting, by spinning and washing with PBS (300 x g, RT, 5 minutes). Cells were treated with 1 ml media containing IFN-β to a final concentration of 100 ng/ml in the flask. For the 0 minute samples, 1 ml of untreated media was added and the cells were dissociated before harvesting.

4.4.2 Nuclei isolation

Post-treatment, cells were placed on ice and washed three times with ice-cold PBS. Cells were incubated on ice in 10 ml ice-cold Lysis Buffer (10 mM Tris-HCl pH 7.5, 2 mM $MgCl_2$, 3 mM $CaCl_2$, 0.5% IGEPAL, 10% Glycerol, 2 U/ml SUPERase-IN, brought to volume with 0.1% DEPC DI-water, filtered before use) for 10 minutes. Cells were scraped and collected into 50 ml Falcon tubes, and centrifuged with a fixed-angle rotor at 1000 x g for 10 minutes at 4°C. Cells were resuspended with Lysis buffer with a wide-opening P1000 tip, and washed twice with 10 ml Lysis buffer (centrifuged at 1000 x g for 5 minutes at 4°C). After the second Lysis buffer wash, the samples were resuspended with 1 ml Freezing Buffer (50 mM Tris-HCl pH 8.3, 5 mM $MgCl_2$, 40% Glycerol, 0.1 mM EDTA pH 8.0, brought to volume with 0.1% DEPC DI-water, filtered before use). Nuclei were centrifuged at 1000 x g for 5 minutes at 4°C, and resuspended with 500 μl Freezing Buffer. Nuclei were then centrifuged for 2 minutes at 2000 x g, 4°C, and resuspended in 110 μl Freezing Buffer. 10 μl was retained for counting nuclei, while the remaining sample was snap-frozen in liquid nitrogen and stored at -80°C until use.

4.4.3 RNA-seq

RNA-seq samples were prepared using the KAPA RNA Hyperprep kit with Ribo-depletion. In short, RNA was first trizol/chloroform-extracted from pelleted cells using 500 μl trizol and 130 μl chloroform. Samples were washed once with 360 μl chloroform, and ethanol-precipitated using 700 μl 100% ethanol. Resulting samples were analyzed and quantified using a qubit and a Tapestation. 1 μg of input sample was used, along with ERCC spike-in. All samples had similar input and ERCC amounts. Samples were depleted of rRNA per the KAPA protocol, using complementary oligonucleotides and RNase H. Samples were then size selected with a 2.2x Ampure bead cleanup, and leftover hybridization DNA digested with a DNase cleanup. Samples were again cleaned with a 2.2x Ampure bead selection. RNA was then fragmented in KAPA Fragmentation buffer for 6 minutes at 85°C. First and second strands were then synthesized per the protocol, and adapters ligated with 7 μM adapter stock concentration. Samples were then cleaned twice with a 0.63x and 0.7x Ampure bead selection. Finally, samples were amplified for 6 cycles before a 1.0x Ampure bead cleanup and sequenced using an Illumina Novaseq 6000, with a 2x150 sequencing strategy.

4.4.4 PRO-seq

Run-on reactions are adapted from [71]. Ice-cold isolated nuclei (100 μl) were added to 37°C 100 μl reaction buffer (Final Concentration: 5 mM Tris-Cl pH 8.0, 2.5 mM $MgCl_2$, 0.5 mM DTT, 150 mM KCl, 10 units of SUPERase In, 0.5% sarkosyl, 125 μM rATP, 125 μM rGTP, 125 μM rUTP, 25 μM biotin-11-CTP (additionally, two libraries generated with 25 μM biotin-11-CTP, 250 μM rCTP). The reaction ran for 5 min at 37°C. RNA was extracted twice with Trizol, washed once with chloroform, and precipitated with 3 volumes of ice-cold ethanol and 1.5 μl GlycoBlue. The pellet was washed in 75% ethanol before resuspending in 20 μl of DEPC-treated water. Nascent RNA was extracted and fragmented by base hydrolysis in 0.2 N NaOH on ice for 10 minutes, and neutralized by adding a 1X volume of 1 M Tris-HCl pH 6.8. Fragmented nascent RNA was purified using streptavidin beads and ligated with reverse 3′ RNA adaptor (/5Phos/GAUCGUCGGACUGUAGAACUCUGAAC/3InvdT/), and biotin-labeled products were enriched using another streptavidin bead binding and extraction step as before. For 5′ end repair, the RNA products were treated with tobacco acid pyrophosphatase (Epicenter) and T4 polynucleotide kinase (NEB). 5′ repaired RNA was ligated to reverse 5′ RNA adaptor (5′ UGGAAUUCUCGGGUGCCAAGG) before being purified by a final round of streptavidin bead binding and extraction. RNA was reverse transcribed using 25 pmol of RP1 primer (5′AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA). The product was amplified 15 ± 3 cycles and products >150 bp (insert > 70 bp) were size selected with 1X AMPure XP beads (Beckman) before being sequenced, using an Illumina NextSeq 2000 with a 1x50 strategy.

4.4.5 Trimming, mapping, visualization, quality control

Resulting FASTQ files were trimmed and mapped to the GRCh38/hg38 reference genome and prepared for analysis and visualization through our in-house pipeline. In short, resulting FASTQ read files were first trimmed using bbduk (v38.05) to remove adapter sequences, as well as short or low quality reads. Reads were mapped with HISAT2 (v2.1.0), and resulting SAM files converted to BAM files using Samtools (v1.8). Reads with a mapping quality less than 5 were removed, which consequently also removed multi-mapping reads. BedGraph files were generated using Bedtools (v2.25.0), and converted to TDF files for visualization using IGVtools (v2.3.75). Quality metrics were generated with FastQC (v0.11.8), Preseq (v2.0.3), RSeQC (v3.0.0), with figures generated through MultiQC (v1.6). For further version information and specific input information, see NextFlow pipeline found at https://github.com/Dowell-Lab/Nascent-Flow.git.

4.4.6 Differential transcription analysis

Differential transcription was performed using the DESeq2 (v1.26.0) R package (R version 3.6.3). Gene counts were generated using featureCounts (v1.6.2) from the R Subread package (v1.6.0), counting over the entire gene body from RefSeq Annotations (release number 109, downloaded August 14, 2019 from UCSC genome browser), subtracting the promoter peak in PRO-seq samples (+250 bp from annotated start site). For featureCounts, BED6 region files were converted to SAF format with the following command: awk -F "\t" -v OFS="\t" 'print{$4, $1, $2, $3, $6}' region.bed > region.saf. Only the highest transcribed isoform of each gene was considered.

For batch correction, the program ComBat-Seq was used[132], utilizing replicate and ploidy information of the samples as covariates. DESeq2 was then run using batch-corrected samples with the following design formula: ploidy + time + ploidy:time. Results were fetched from the output file using the contrasts as specified in the text: for late time points, results were generated comparing the 120m and 60m samples to each other. For early time points (15,30 and 45 minute time points), all samples were compared to the 0 time point.

4.4.7 Degradation analysis

Time intervals were calculated using any pair of time points in our series. For example, the 0m, 15m, 30m samples can generate the intervals 0m-15m, 0m-30m, and 15m-30m. I utilized each interval as described in the text to calculate degradation.

For the SSM, I calculated degradation for each gene and each replicate by finding the ration of PRO-seq signal and RNA-seq signal. For a given time interval, I calculated the average degradation estimate using each end of the interval, such that the values were comparable to LA calculations.

The LA values were calculated using the following equation:

$$\alpha(t) = (\Delta R(t) - \bar{P}(t)\Delta t)/(\bar{R}(t)\Delta t)$$

Where α is the per-molecule degradation rate, Δt is the size of the time interval (in minutes), $\bar{P}$ is the average PRO-seq signal from each end of the interval, $\bar{R}$ is the average RNA-seq signal from each end of the interval, and ΔR(t) is the difference in RNA-seq signal between each end of the interval. The value for degradation was approximated as the rate at the average time of each end of the interval (e.g., the interval 0m-15m would yield the degradation rate at 7.5m), to ensure that these values would be comparable to SSM estimates.

For details of degradation calculations and plotting, see the accompanying jupyter notebook: https://github.com/Dowell-Lab/Degradation_Modeling

Chapter 5

Conclusion

Throughout the course of my studies, I explored both the technical considerations and the biological basis of RNA degradation analysis in Down syndrome. I first explored the origin of technical variance in degradation analysis: specifically, the technical noise inherent to nascent transcription protocols, and the prior considerations needed for differential analysis in aneuploidy. These two studies were used to inform the analysis pipeline for my RNA degradation analysis. The preliminary results of this decay analysis show the viability of using a linear approximation method for inferring degradation rates from RNA-seq and PRO-seq. Furthermore, there appeared to be a select number of genes which showed differences in RNA decay throughout the interferon treatment. In all, these results suggest potential differences in RNA dynamics in Down syndrome.

5.1 Identifying sources of technical error

5.1.1 Analysis of variations in run-on sequencing protocols

Throughout out the course of this study, I demonstrate that technical aspects of protocol and analysis can have rippling effects on biological conclusions drawn from the data. Importantly, this technical noise can generally be detected and accounted for, which can mitigate its effects on analysis[23]. In run-on sequencing data, conclusions regarding the $5'$ end of transcription are particularly susceptible to even slight variations in protocol, while elongation regions are more consistent.

Due to difficulties in modeling the $3'$ end, it is still unknown how the signal in this region changes due to technical aspects of the protocol. A new model — pioneered by Dr. Jacob Stanley — has only recently made it possible to accurately profile transcription across the entire gene, including both the $5'$ and $3'$ ends. As such, in the future, utilizing this model will enable a more rigorous protocol comparison specifically in the $3'$ end.

Interestingly, intergenic regions also showed a degree of dependency on the run-on sequencing protocol used. In this case, however, entire enhancer regions seemed to be differentially captured, especially when comparing GRO-seq and PRO-seq protocols. Thus far, it does not appear that this differential capture is dictated by sequence motifs, GC content, or the length of the enhancer, although this does not rule out that some combination of these and other factors may explain this result. Furthermore, while the identity of enhancers captured varies between protocols, when the cells were treated with a p53 activator, both protocols readily captured p53-activated enhancers. Importantly, these seemingly protocol-specific enhancers may instead be due to biological variance between my samples, or they may be representative of enhancers which have a low signal-to-noise ratio, and would thus be more subject to sampling issues. As more run-on datasets are generated, future studies will be better able to estimate the effects of biological variation, leaving the door open to more accurately detect which intergenic features are truly protocol-specific.

5.1.2 Accounting for trisomy in differential analysis

When studying Down syndrome, it is important to realize that many computational methods were developed implicitly with typical, disomy data in mind. Thus, the biological nature of the trisomy leads to unique data analysis challenges, as the assumptions of commonly used analysis techniques are not trisomy aware. For example, the data and hypothesis testing must be adjusted in order for proper conclusions to be made regarding the relationship of gene dosage and expression. Failing to do so can make biological and technical variance look like a reduction in gene expression levels for the triplicated chromosome. Indeed, it is possible that some reports of dosage compensation in trisomy 21 are sourced from this error, as our data only show a scant few genes which are significantly below expectation[50]. Furthermore, most of the genes with lower than expected expression arise from specific lowly expressed alleles that circulate in the population normally. It remains an unanswered question as to whether there is selection for the lowly expressed alleles in Down syndrome at the population level.

Differential expression analysis with aneuploid samples becomes significantly more complex when considering an integrative approach of multiple high throughput sequencing dataset types. In my study, I found that GRO-seq and RNA-seq signal was highly correlated at genes, but that fold change estimates at highly variable, low expression genes were dependent on the protocol type. Indeed, when comparing D21 and T21 samples, fold change estimates at these genes on chromosome 21 tended more towards 1.0x specifically in GRO-seq. Furthermore, I discovered that this bias is due to technical variance, which is exacerbated within run-on sequencing protocols specifically. If these two protocols are to be interpreted together, proper normalization and hypothesis testing will be required to account for the trisomic background.

5.2 Inferring RNA turnover from paired RNA-seq and GRO-seq datasets

5.2.1 Linear approximation of RNA decay rates

Utilizing PRO-seq and RNA-seq, and correcting for their respective technical signatures, I was able to identify changes in degradation throughout the course of the interferon response at a select number of genes. By using a linear approximation model, I sought to capture time points at which changes in degradation took effect, an occurrence which is unaccounted for under the steady-state assumptions of previous models[15]. Indeed, at several candidate genes, I observed disagreements between the two models, suggesting a shift in decay rates across time. Interestingly, this result raises the possibility of utilizing both models in a statistical pipeline. While the steady-state model did not capture differential changes over long time courses, it could be used to approximate the null hypothesis of a consistent, invariant decay rate. As this study is expanded, I plan to develop this test to more rigorously detect changes in degradation across time.

One factor remains pertinent for the development of a differential degradation pipeline:

differential analysis of count data deals with a degree of uncertainty in both the data and the null hypothesis[69, 45, 95]. Thus, as the linear approximation model deals with both PRO-seq and RNA-seq count data, as well as the possibility of sharing information from similar genes throughout the time course, frequentist statistics would be insufficient for differential degradation analysis. These factors suggest that a Bayesian approach is most appropriate. However, unlike count data, it is unknown which distribution would best model degradation rates. Thus far, I've found that simulated decay data appear to approximate an inverse-Gamma distribution; as more data are generated, this can be confirmed in real data as well.

Finally, both the steady-state and linear-approximation models have yet to be validated with an orthogonal method when a time series is involved. Although correlations have been established with untreated samples[15, 37], this fails to interrogate whether changes across time can be equivalently detected using these models. Crucially, many "gold-standard" degradation assays are based off of RNA-seq; as such, it is already known that certain unstable, lowly transcribed RNA species (e.g., many lncRNAs) are not captured by these techniques[37, 35, 102, 16]. It remains to be seen whether our inferential method will serve as an improvement in this regard.

5.2.2 Determining mechanism of degradation

While identifying degradation rates is possible using either metabolic labeling or these inferential methods, one missing component is assigning causality to an observed increase or decrease in degradation at a given gene. Degradation is mediated though a myriad of pathways and regulatory molecules[108]. Within the immune system, for example, several miRNA-mediated effects of interest have already been identified[33]. Viral infection itself is also known to trigger generalized RNase L mediated decay[66]. And quality control pathways such as nonsense-mediated decay are consistent factors in RNA turnover. Each of these act together along with many other components to form the final degradation rate of a given RNA species. As such, separating their individual effects requires identifying when and to what degree each separate pathway has been activated. Each of these pathways has an associated expected timing, component proteins, and — in some cases —

associated motifs. Thus, much like identifying primary and secondary transcriptional effects, gene decay responses can be separated by time. Based on this timing, potential decay pathways can be assigned for each gene and subsequently confirmed experimentally.

5.2.3 RNA degradation and Down syndrome

By comparing degradation between D21 and T21 cells, I identified several possible instances where the dysregulation of the T21 interferon response also resulted in changes to turnover rates. Naturally, as described above, these findings will require a statistical pipeline and a validating method to fully confirm. If confirmed, however, these findings would suggest that Down syndrome causes dysregulation not only at the transcriptional level, but also at the level of RNA degradation.

RNA degradation rates can be modulated at many different points within the regulatory cycle of the molecule[108]. Importantly, degradation can elicit a response extremely rapidly, especially in cells primed to trigger a global degradation response (as in the viral response pathway)[108, 47, 33]. The findings here could represent a crucial aspect of Down syndrome pathology- this dysregulation at critical time points would have rippling effects throughout the interferon response. Once I have identified which genes are undergoing differential degradation, a future study will be able to explore the biological consequences of this dysregulation on normal gene activity.

5.3 Conclusion

RNA degradation serves a pivotal role in the cell, acting as another layer of rapid fine-tuning for gene regulation[108]. However, analysing RNA decay has been fraught with difficulty in the past. Early high-throughput methods involving transcriptional inhibitors that trigger a stress response in the cell, while metabolic labeling approaches are subject to sensitivity and toxicity issues based on the concentration of the nucleotide analog. I provide here an alternative, which allows for the inference of degradation rates from a combination of PRO-seq and RNA-seq data. Naturally, the logistical benefit of obtaining decay rates, nascent transcription, and total RNA levels together cannot be overstated. These three intertwining signals allow for a detailed profile of transcriptional and post-transcriptional regulation across time. Alongside this, I provide a framework for anticipating sources of technical bias coming from each of these different protocols.

Due to the nature of time series data, I anticipate that future work may be able to utilize this data to infer causal networks for RNA regulation. For example, miRNA-mediated degradation has already been predicted by motif analysis in previous studies[113, 131]; combined with PRO-seq signal over miRNAs and calculated degradation rates, the predictive power of these networks grows. Regardless, these data already show potential differences in transcription and degradation in the interferon response pathway between D21 and T21. As the study expands, these genes hold potential for explaining discrepancies within the T21 interferon response.

Bibliography

[1] D. Aarskog. Autoimmune thyroid disease in children with mongolism. Archives of Disease in Childhood, 44(236):454–460, 1969.

[2] Karen Adelman and John T. Lis. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat Rev Genet, 13(10):720–731, 10 2012.

[3] Rached Alkallas, Lisa Fish, Hani Goodarzi, and Hamed S. Najafabadi. Inference of RNA decay rate from transcriptional profiling highlights the regulatory programs of alzheimer's disease. Nature Communications, 8(1):909, Oct 2017.

[4] Mary A. Allen, Hestia Mellert, Veronica Dengler, Zdenek Andryzik, Anna Guarnieri, Justin A. Freeman, Xin Luo, William L. Kraus, Robin D. Dowell, and Joaquín M. Espinosa. Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. eLife, 3:e02200, 2014. doi: 10.7554/eLife.02200.

[5] Robin Andersson, Peter Refsing Andersen, Eivind Valen, Leighton J. Core, Jette Bornholdt, Mette Boyd, Torben Heick Jensen, and Albin Sandelin. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. Nat Commun, 5, 11 2014.

[6] Zdenek Andrysik, Matthew D Galbraith, Anna L Guarnieri, Sara Zaccara, Kelly D Sullivan, Ahwan Pandey, Morgan MacBeth, Alberto Inga, and Joaquín M Espinosa. Identification of a core TP53 transcriptional program with highly distributed tumor suppressive activity. Genome research, 27(10):1645–1657, 2017.

[7] Stylianos E. Antonarakis, Brian G. Skotko, Michael S. Rafii, Andre Strydom, Sarah E. Pape, Diana W. Bianchi, Stephanie L. Sherman, and Roger H. Reeves. Down syndrome. Nature Reviews Disease Primers, 6(1):9, Feb 2020.

[8] Francesca Antonaros, Rossella Zenatelli, Giulia Guerri, Matteo Bertelli, Chiara Locatelli, Beatrice Vione, Francesca Catapano, Alice Gori, Lorenza Vitale, Maria Chiara Pelleri, Giuseppe Ramacieri, Guido Cocchi, Pierluigi Strippoli, Maria Caracausi, and Allison Piovesan. The transcriptome profile of human trisomy 21 blood cells. Human Genomics, 15(1):25, May 2021.

[9] Yuki Aoi, Edwin R. Smith, Avani P. Shah, Emily J. Rendleman, Stacy A. Marshall, Ashley R. Woodfin, Fei X. Chen, Ramin Shiekhattar, and Ali Shilatifard. Nelf regulates a promoter-proximal step distinct from RNA pol II pause-release. Molecular Cell, 78(2):261 – 274.e5, 2020.

[10] Robert C. Axtell and Chander Raman. Janus-like effects of type I interferon in autoimmune diseases. Immunological Reviews, 248(1):23–35, 2012.

[11] Joseph G Azofeifa, Mary A Allen, Josephina R Hendrix, Timothy Read, Jonathan D Rubin, and Robin D Dowell. Enhancer RNA profiling predicts transcription factor activity. Genome Research, Feb 2018.

[12] Joseph G Azofeifa and Robin D Dowell. A generative model for the behavior of RNA polymerase. Bioinformatics, 33(2):227–234, 09 2016.

[13] Elisa Barbieri, Connor Hill, Mathieu Quesnel-Vallieres, Yoseph Barash, and Alessandro Gardini. Rapid and scalable profiling of nascent RNA with fastGRO. bioRxiv, 2020.

[14] Olivier Bensaude. Inhibiting eukaryotic transcription. which compound to choose? how to evaluate its activity? Transcription, 2(3):103–108, 2011. PMID: 21922053.

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

[16] Etienne Boileau, Janine Altmüller, Isabel S Naarmann-de Vries, and Christoph Dieterich. A comparison of metabolic labeling and statistical methods to infer genome-wide dynamics of RNA turnover. Briefings in Bioinformatics, 22(6), 07 2021. bbab219.

[17] Gregory T. Booth, Pabitra K. Parua, Miriam Sansó, Robert P. Fisher, and John T. Lis. CDK9 regulates a promoter-proximal checkpoint to modulate RNA polymerase ii elongation rate in fission yeast. Nature Communications, 9(1):543, Feb 2018.

[18] Véronique Brault, Thu Lan Nguyen, Javier Flores-Gutiérrez, Giovanni Iacono, Marie-Christine Birling, Valérie Lalanne, Hamid Meziane, Antigoni Manousopoulou, Guillaume Pavlovic, Loïc Lindner, Mohammed Selloum, Tania Sorg, Eugene Yu, Spiros D. Garbis, and Yann Hérault. Dyrk1a gene dosage in glutamatergic neurons has key effects in cognitive deficits observed in mouse models of MRD7 and down syndrome. PLOS Genetics, 17(9):1–34, 09 2021.

[19] Joseph F Cardiello, Gilson J Sanchez, Mary A Allen, and Robin D Dowell. Lessons from eRNAs: understanding transcriptional regulation through the lens of nascent RNAs. Transcription, 11(1):3–18, Feb 2020.

[20] Kun Chen, Juan Liu, and Xuetao Cao. Regulation of type i interferon signaling in immunity and inflammation: A comprehensive review. Journal of Autoimmunity, 83:1–11, 2017. The Pathogenesis of Autoimmunity: Epigenomics and Beyond.

[21] Hussain Ahmed Chowdhury, Dhruba Kumar Bhattacharyya, and Jugal Kumar Kalita. Differential expression analysis of RNA-seq reads: Overview, taxonomy, and tools. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17(2):566–586, 2020.

[22] Annique Claringbould and Judith B. Zaugg. Enhancers in disease: molecular basis and emerging treatment strategies. Trends in Molecular Medicine, 27(11):1060–1073, Nov 2021.

[23] Luis A. Corchete, Elizabeta A. Rojas, Diego Alonso-López, Javier De Las Rivas, Norma C. Gutiérrez, and Francisco J. Burguillo. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. Scientific Reports, 10(1):19737, Nov 2020.

[24] Leighton Core and John Lis. Transcription regulation through promoter-proximal pausing of RNA polymerase II. Science, 319:1791, 2008.

[25] Yanick J. Crow and Daniel B. Stetson. The type I interferonopathies: 10 years on. Nature Reviews Immunology, 22(8):471–483, Aug 2022.

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

[27] Ingrid Daubechies. Ten lectures on wavelets. SIAM, 1992.

[28] Daniel S. Day, Bing Zhang, Sean M. Stevens, Francesco Ferrari, Erica N. Larschan, Peter J. Park, and William T. Pu. Comprehensive analysis of promoter-proximal RNA polymerase II pausing across mammalian cell types. Genome Biology, 17(1):120, Jun 2016.

[29] Ilario De Toma and Mara Dierssen. Network analysis of Down syndrome and SARS-CoV-2 identifies risk and protective factors for COVID-19. Scientific Reports, 11(1):1930, Jan 2021.

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

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

[32] Joaquin M. Espinosa. Down syndrome and COVID-19: A perfect storm? Cell Reports Medicine, 1(2), May 2020.

[33] Chiara Farroni, Emiliano Marasco, Valentina Marcellini, Ezio Giorda, Diletta Valentini, Stefania Petrini, Valentina D'Oria, Marco Pezzullo, Simona Cascioli, Marco Scarsella, Alberto G Ugazio, Giovanni C De Vincentiis, Ola Grimsholm, and Rita Carsetti. Dysregulated miR-155 and miR-125b are related to impaired B-cell responses in Down syndrome. Front. Immunol., 9:2683, November 2018.

[34] Alistair R. R. Forrest, Hideya Kawaji, Michael Rehli, J. Kenneth Baillie, Michiel J. L. de Hoon, Vanja Haberle, Timo Lassmann, Ivan V. Kulakovskiy, Marina Lizio, Masayoshi Itoh, Robin Andersson, Christopher J. Mungall, Terrence F. Meehan, Sebastian Schmeier, Nicolas Bertin, Mette Jørgensen, Emmanuel Dimont, Erik Arner, Christian Schmidl, Ulf Schaefer, Yulia A. Medvedeva, Charles Plessy, Morana Vitezic, Jessica Severin, Colin A. Semple, Yuri Ishizu, Robert S. Young, Margherita Francescatto, Intikhab Alam, Davide Albanese, Gabriel M. Altschuler, Takahiro Arakawa, John A. C. Archer, Peter Arner, Magda

Babina, Sarah Rennie, Piotr J. Balwierz, Anthony G. Beckhouse, Swati Pradhan-Bhatt, Judith A. Blake, Antje Blumenthal, Beatrice Bodega, Alessandro Bonetti, James Briggs, Frank Brombacher, A. Maxwell Burroughs, Andrea Califano, Carlo V. Cannistraci, Daniel Carbajo, Yun Chen, Marco Chierici, Yari Ciani, Hans C. Clevers, Emiliano Dalla, Carrie A. Davis, Michael Detmar, Alexander D. Diehl, Taeko Dohi, Finn Drabløs, Albert S. B. Edge, Matthias Edinger, Karl Ekwall, Mitsuhiro Endoh, Hideki Enomoto, Michela Fagiolini, Lynsey Fairbairn, Hai Fang, Mary C. Farach-Carson, Geoffrey J. Faulkner, Alexander V. Favorov, Malcolm E. Fisher, Martin C. Frith, Rie Fujita, Shiro Fukuda, Cesare Furlanello, Masaaki Furuno, Jun-ichi Furusawa, Teunis B. Geijtenbeek, Andrew P. Gibson, Thomas Gingeras, Daniel Goldowitz, Julian Gough, Sven Guhl, Reto Guler, Stefano Gustincich, Thomas J. Ha, Masahide Hamaguchi, Mitsuko Hara, Matthias Harbers, Jayson Harshbarger, Akira Hasegawa, Yuki Hasegawa, Takehiro Hashimoto, Meenhard Herlyn, Kelly J. Hitchens, Shannan J. Ho Sui, Oliver M. Hofmann, Ilka Hoof, Fumi Hori, Lukasz Huminiecki, Kei Iida, Tomokatsu Ikawa, Boris R. Jankovic, Hui Jia, Anagha Joshi, Giuseppe Jurman, Bogumil Kaczkowski, Chieko Kai, Kaoru Kaida, Ai Kaiho, Kazuhiro Kajiyama, Mutsumi Kanamori-Katayama, Artem S. Kasianov, Takeya Kasukawa, Shintaro Katayama, Sachi Kato, Shuji Kawaguchi, Hiroshi Kawamoto, Yuki I. Kawamura, Tsugumi Kawashima, Judith S. Kempfle, Tony J. Kenna, Juha Kere, Levon M. Khachigian, Toshio Kitamura, S. Peter Klinken, Alan J. Knox, Miki Kojima, Soichi Kojima, Naoto Kondo, Haruhiko Koseki, Shigeo Koyasu, Sarah Krampitz, Atsutaka Kubosaki, Andrew T. Kwon, Jeroen F. J. Laros, Weonju Lee, Andreas Lennartsson, Kang Li, Berit Lilje, Leonard Lipovich, Alan Mackay-sim, Ri-ichiroh Manabe, Jessica C. Mar, Benoit Marchand, Anthony Mathelier, Niklas Mejhert, Alison Meynert, Yosuke Mizuno, David A. de Lima Morais, Hiromasa Morikawa, Mitsuru Morimoto, Kazuyo Moro, Efthymios Motakis, Hozumi Motohashi, Christine L. Mummery, Mitsuyoshi Murata, Sayaka Nagao-Sato, Yutaka Nakachi, Fumio Nakahara, Toshiyuki Nakamura, Yukio Nakamura, Kenichi Nakazato, Erik van Nimwegen, Noriko Ninomiya, Hiromi Nishiyori, Shohei Noma, Tadasuke Nozaki, Soichi Ogishima, Naganari Ohkura, Hiroko Ohmiya, Hiroshi Ohno, Mitsuhiro Ohshima, Mariko Okada-Hatakeyama, Yasushi Okazaki, Valerio Orlando, Dmitry A. Ovchinnikov, Arnab Pain, Robert Passier, Margaret Patrikakis, Helena Persson, Silvano Piazza, James G. D. Prendergast, Owen J. L. Rackham, Jordan A. Ramilowski, Mamoon Rashid, Timothy Ravasi, Patrizia Rizzu, Marco Roncador, Sugata Roy, Morten B. Rye, Eri Saijyo, Antti Sajantila, Akiko Saka, Shimon Sakaguchi, Mizuho Sakai, Hiroki Sato, Hironori Satoh, Suzana Savvi, Alka Saxena, Claudio Schneider, Erik A. Schultes, Gundula G. Schulze-Tanzil, Anita Schwegmann, Thierry Sengstag, Guojun Sheng, Hisashi Shimoji, Yishai Shimoni, Jay W. Shin, Christophe Simon, Daisuke Sugiyama, Takaaki Sugiyama, Masanori Suzuki, Naoko Suzuki, Rolf K. Swoboda, Peter A. C. 't Hoen, Michihira Tagami, Naoko Takahashi, Jun Takai, Hiroshi Tanaka, Hideki Tatsukawa, Zuotian Tatum, Mark Thompson, Hiroo Toyoda, Tetsuro Toyoda, Eivind Valen, Marc van de Wetering, Linda M. van den Berg, Roberto Verardo, Dipti Vijayan, Ilya E. Vorontsov, Wyeth W. Wasserman, Shoko Watanabe, Christine A. Wells, Louise N. Winteringham, Ernst Wolvetang, Emily J. Wood, Yoko Yamaguchi, Masayuki Yamamoto, Misako Yoneda, Yohei Yonekura, Shigehiro Yoshida, Susan E. Zabierowski, Peter G. Zhang, Xiaobei Zhao, Silvia Zucchelli, Kim M. Summers, Harukazu Suzuki, Carsten O. Daub, Jun Kawai, Peter Heutink, Winston Hide, Tom C. Freeman, Boris Lenhard, Vladimir B. Bajic, Martin S. Taylor, Vsevolod J. Makeev, Albin Sandelin, David A. Hume, Piero Carninci, Yoshihide Hayashizaki, The FANTOM Consortium, the RIKEN PMI, and CLST (DGT). A promoter-level mammalian expression atlas. Nature, 507(7493):462–470, 2014.

[35] Caroline C. Friedel and Lars Dölken. Metabolic tagging and purification of nascent RNA: implications for transcriptomics. Mol. BioSyst., 5:1271–1278, 2009.

[36] Dominique Gagliardi and Andrzej Dziembowski. $5'$ and $3'$ modifications controlling RNA degradation: from safeguards to executioners. Philosophical Transactions of the Royal Society B: Biological Sciences, 373(1762):20180160, 2018.

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

[38] Tianshun Gao and Jiang Qian. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Research, 48(D1):D58–D64, 11 2019.

[39] Wilson Wen Bin Goh, Wei Wang, and Limsoon Wong. Why batch effects matter in omics data, and how to avoid them. Trends in Biotechnology, 35(6):498–507, 2021/05/04 2017.

[40] Consortium GTEx. The GTEx consortium atlas of genetic regulatory effects across human tissues. Science, 369(6509):1318–1330, Sep 2020.

[41] Zhenxing Guo, Ying Cui, Xiaowen Shi, James A. Birchler, Igor Albizua, Stephanie L. Sherman, Zhaohui S. Qin, and Tieming Ji. An empirical bayesian approach for testing gene expression fold change and its application in detecting global dosage effects. NAR genomics and bioinformatics, 2(3):lqaa072–lqaa072, Sep 2020. PMC7671412.

[42] Nasun Hah, Charles¬†G. Danko, Leighton Core, Joshua J. Waterfall, Adam Siepel, John T. Lis, and W.Lee Kraus. A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. Cell, 145(4):622 – 634, 2011.

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

[44] Shengli Hao and David Baltimore. The stability of mrna influences the temporal order of the induction of genes encoding inflammatory molecules. Nature Immunology, 10(3):281–288, Mar 2009.

[45] Thomas J. Hardcastle and Krystyna A. Kelly. bayseq: Empirical bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics, 11(1):422, Aug 2010.

[46] Terry Hassold and Patricia Hunt. To err (meiotically) is human: the genesis of human aneuploidy. Nature Reviews Genetics, 2(4):280–291, Apr 2001.

[47] Jonathan Houseley and David Tollervey. The many pathways of RNA degradation. Cell, 136(4):763–776, Feb 2009.

[48] J. D. Hunter. Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3):90–95, 2007.

[49] Samuel Hunter, Rutendo F. Sigauke, Jacob T. Stanley, Mary A. Allen, and Robin D. Dowell. Protocol variations in run-on transcription dataset preparation produce detectable signatures in sequencing libraries. BMC Genomics, 23(1):187, Mar 2022.

[50] Sunyoung Hwang, Paola Cavaliere, Rui Li, Lihua Julie Zhu, Noah Dephoure, and Eduardo M. Torres. Consequences of aneuploidy in human fibroblasts with trisomy 21. Proceedings of the National Academy of Sciences, 118(6):e2014723118, 2021.

[51] Stefanie Jonas and Elisa Izaurralde. Towards a molecular understanding of microrna-mediated gene silencing. Nature Reviews Genetics, 16(7):421–433, Jul 2015.

[52] Iris Jonkers, Hojoong Kwak, and John T Lis. Genome-wide dynamics of pol ii elongation and its interplay with promoter proximal pausing, chromatin, and exons. eLife, 3:e02407, apr 2014.

[53] Samantha Sae-Young Kim, Alexis Dziubek, Seungha Alisa Lee, and Hojoong Kwak. Nascent RNA sequencing of peripheral blood leukocytes reveal gene expression diversity. bioRxiv, 2019.

[54] Tae-kyung Kim, Martin Hemberg, Jesse M. Gray, Allen M. Costa, Daniel M. Bear, Jing Wu, David A. Harmin, Mike Laptewicz, Kellie Barbara-Haley, Scott Kuersten, Eirene Markenscoff-Papadimitriou, Dietmar Kuhl, Haruhiko Bito, Paul F. Worley, Gabriel Kreiman, and Michael E. Greenberg. Widespread transcription at neuronal activity-regulated enhancers. Nature, 465(7295):182–187, May 2010.

[55] You-Me Kim and Eui-Cheol Shin. Type I and III interferon responses in SARS-CoV-2 infection. Experimental & Molecular Medicine, 53(5):750–760, May 2021.

[56] Shihoko Kojima and Daniela Cimini. Aneuploidy and gene expression: is there dosage compensation? Epigenomics, 11(16):1827–1837, Dec 2019. PMC7132608.

[57] Xiao-Fei Kong, Lisa Worley, Darawan Rinchai, Vincent Bondet, Puthen Veettil Jithesh, Marie Goulet, Emilie Nonnotte, Anne Sophie Rebillat, Martine Conte, Clotilde Mircher, Nicolas Gürtler, Luyan Liu, Mélanie Migaud, Mohammed Elanbari, Tanwir Habib, Cindy S. Ma, Jacinta Bustamante, Laurent Abel, Aimé Ravel, Stanislas Lyonnet, Arnold Munnich, Darragh Duffy, Damien Chaussabel, Jean-Laurent Casanova, Stuart G. Tangye, Stéphanie Boisson-Dupuis, and Anne Puel. Three copies of four interferon receptor genes underlie a mild type i interferonopathy in down syndrome. Journal of Clinical Immunology, 40(6):807–819, Aug 2020.

[58] Max Kuhn. Building predictive models in R using the caret package. Journal of Statistical Software, Articles, 28(5):1–26, 2008.

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

[60] Hui-Chi Lai, Alexander James, John Luff, Paul De Souza, Hazel Quek, Uda Ho, Martin F. Lavin, and Tara L. Roberts. Regulation of RNA degradation pathways during the lipopolysaccharide response in macrophages. Journal of Leukocyte Biology, 109(3):593–603, 2021.

[61] Tuuli Lappalainen, Michael Sammeth, Marc R. Friedländer, Peter A. C. 't Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G. Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G. MacArthur, Monkol Lek, Esther Lizano, Henk P. J. Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B. Montgomery, Peter Donnelly, Mark I. McCarthy, Paul Flicek, Tim M. Strom, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Ángel Carracedo, Stylianos E. Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G. Gut, Xavier Estivill, Emmanouil T. Dermitzakis, and The Geuvadis Consortium. Transcriptome and genome sequencing uncovers functional variation in humans. Nature, 501(7468):506–511, 2013.

[62] Gregory R Lee, Ralf Gommers, Filip Waselewski, Kai Wohlfahrt, and Aaron O'Leary. Pywavelets: A python package for wavelet analysis. Journal of Open Source Software, 4(36):1237, 2019.

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

[64] Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics, 11(10):733–739, 2010.

[65] Cecilia B. Levandowski, Taylor Jones, Margaret Gruca, Sivapriya Ramamoorthy, Robin D. Dowell, and Dylan J. Taatjes. The Δ40p53 isoform inhibits p53-dependent eRNA transcription and enables regulation by signal-specific transcription factors during p53 activation. PLOS Biology, 19(8):1–33, 08 2021.

[66] Yize Li, Shuvojit Banerjee, Yuyan Wang, Stephen A. Goldstein, Beihua Dong, Christina Gaughan, Robert H. Silverman, and Susan R. Weiss. Activation of RNase L is dependent on OAS3 expression during infection with diverse human viruses. Proceedings of the National Academy of Sciences, 113(8):2241–2246, 2016.

[67] Yen-Chin Liu, Bobo Wing-Yee Mok, Pui Wang, Rei-Lin Kuo, Honglin Chen, and Shin-Ru Shih. Cellular $5'$-$3'$ mRNA exoribonuclease XRN1 inhibits interferon beta activation and facilitates influenza a virus replication. mBio, 12(4):e00945–21, 2021.

[68] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Harold Magazine, John Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, Mark Cosentino, Norma Diaz-Mayoral,

Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand, Molly Donovan, Yan Meng, George Grant, Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J. Cox, Dan L. Nicolae, Eric R. Gamazon, Hae Kyung Im, Anuar Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothèe Flutre, Xiaoquan Wen, Emmanouil T. Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalin, Mike Feolo, Nataliya Sharopova, Anne Sturcke, Justin Paschal, James M. Anderson, Elizabeth L. Wilder, Leslie K. Derr, Eric D. Green, Jeffery P. Struewing, Gary Temple, Simona Volpi, Joy T. Boyer, Elizabeth J. Thomson, Mark S. Guyer, Cathy Ng, Assya Abdallah, Deborah Colantuoni, Thomas R. Insel, Susan E. Koester, A. Roger Little, Patrick K. Bender, Thomas Lehner, Yin Yao, Carolyn C. Compton, Jimmie B. Vaught, Sherilyn Sawyer, Nicole C. Lockhart, Joanne Demchok, and Helen F. Moore. The genotype-tissue expression (GTEx) project. Nature Genetics, 45(6):580–585, Jun 2013.

[69] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology, 15(12):550, 2014.

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

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

[72] Louise Malle and Dusan Bogunovic. Down syndrome and type I interferon: not so simple. Current Opinion in Immunology, 72:196–205, 2021. Allergy and hypersensitivity * Host Pathogens.

[73] Louise Malle, Marta Martin-Fernandez, Sofija Buta, Ashley Richardson, Douglas Bush, and Dusan Bogunovic. Excessive negative regulation of type I interferon disrupts viral control in individuals with Down syndrome. Immunity, 55(11):2074–2084.e5, 2022.

[74] Rong Mao, Xiaowen Wang, Edward Spitznagel, Laurence Frelin, Jason Ting, Huashi Ding, Jung-whan Kim, Ingo Ruczinski, Thomas Downey, and Jonathan Pevsner. Primary and secondary transcriptional effects in the developing human Down syndrome brain and heart. Genome Biology, 6(13):R107, 2005.

[75] John Marioni, Christopher Mason, Shrikant Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research, 18:1509, xx 2008.

[76] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, 2021. R package version 1.7-6.

[77] Irene M. Min, Joshua J. Waterfall, Leighton J. Core, Robert J. Munroe, John Schimenti, and John T. Lis. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. Genes & Development, 25(7):742–754, 2011.

[78] Vamsi K. Mootha, Cecilia M. Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, Nicholas Houstis, Mark J. Daly, Nick Patterson, Jill P. Mesirov, Todd R. Golub, Pablo Tamayo, Bruce Spiegelman, Eric S. Lander, Joel N. Hirschhorn, David Altshuler, and Leif C. Groop. Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics, 34(3):267–273, Jul 2003.

[79] Jeffrey S. Mugridge, Jeff Coller, and John D. Gross. Structural and molecular mechanisms for the control of eukaryotic $5'$–$3'$ mRNA decay. Nature Structural & Molecular Biology, 25(12):1077–1085, Dec 2018.

[80] Alexandra C. Nica and Emmanouil T. Dermitzakis. Expression quantitative trait loci: present and future. Philosophical Transactions of the Royal Society B: Biological Sciences, 368(1620):20120362, 2013.

[81] Einari A. Niskanen, Marjo Malinen, Päivi Sutinen, Sari Toropainen, Ville Paakinaho, Anniina Vihervaara, Jenny Joutsen, Minna U. Kaikkonen, Lea Sistonen, and Jorma J. Palvimo. Global sumoylation on active chromatin is an acute heat stress response restricting transcription. Genome Biology, 16(1):153, Jul 2015.

[82] Gonçalo Nogueira, Rafael Fernandes, Juan F. García-Moreno, and Luísa Romão. Nonsense-mediated RNA decay and its bipolar function in cancer. Molecular Cancer, 20(1):72, Apr 2021.

[83] Jacob O'Brien, Heyam Hayder, Yara Zayed, and Chun Peng. Overview of microRNA biogenesis, mechanisms of actions, and circulation. Frontiers in Endocrinology, 9, 2018.

[84] Andrea Orioli, Viviane Praz, Philippe Lhôte, and Nouria Hernandez. Human MAF1 targets and represses active RNA polymerase III genes by preventing recruitment rather than inducing long-term transcriptional arrest. Genome Research, 26(5):624–635, March 2016.

[85] Roy Parker and Haiwei Song. The enzymes and control of eukaryotic mRNA turnover. Nature Structural & Molecular Biology, 11(2):121–127, Feb 2004.

[86] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[87] Maria Chiara Pelleri, Chiara Cattani, Lorenza Vitale, Francesca Antonaros, Pierluigi Strippoli, Chiara Locatelli, Guido Cocchi, Allison Piovesan, and Maria Caracausi. Integrated quantitative transcriptome maps of human trisomy 21 tissues and cells. Frontiers in Genetics, 9, 2018.

[88] Leonidas C. Platanias. Mechanisms of type-I- and type-II-interferon-mediated signalling. Nature Reviews Immunology, 5(5):375–386, May 2005.

[89] Konstantin Popadin, Stephan Peischl, Marco Garieri, M Reza Sailani, Audrey Letourneau, Federico Santoni, Samuel W Lukowski, Georgii A Bazykin, Sergey Nikolaev, Diogo Meyer, Laurent Excoffier, Alexandre Reymond, and Stylianos E Antonarakis. Slightly deleterious genomic variants and transcriptome perturbations in Down syndrome embryonic selection. Genome Res., 28(1):1–10, January 2018.

[90] Rani K. Powers, Rachel Culp-Hill, Michael P. Ludwig, Keith P. Smith, Katherine A. Waugh, Ross Minter, Kathryn D. Tuttle, Hannah C. Lewis, Angela L. Rachubinski, Ross E. Granrath, María Carmona-Iragui, Rebecca B. Wilkerson, Darcy E. Kahn, Molishree Joshi, Alberto Lleó, Rafael Blesa, Juan Fortea, Angelo D'Alessandro, James C. Costello, Kelly D. Sullivan, and Joaquin M. Espinosa. Trisomy 21 activates the kynurenine pathway via increased dosage of interferon receptors. Nature Communications, 10(1):4766, Oct 2019.

[91] Paola Prandini, Samuel Deutsch, Robert Lyle, Maryline Gagnebin, Celine Delucinge Vivier, Mauro Delorenzi, Corinne Gehrig, Patrick Descombes, Stephanie Sherman, Franca Dagna Bricarelli, Chiara Baldo, Antonio Novelli, Bruno Dallapiccola, and Stylianos E. Antonarakis. Natural gene-expression variation in Down syndrome modulates the outcome of gene-dosage imbalance. American journal of human genetics, 81(2):252–263, Aug 2007.

[92] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[93] Rachel E. Rigby and Jan Rehwinkel. RNA degradation in antiviral immunity and autoimmunity. Trends in Immunology, 36(3):179–188, Mar 2015.

[94] Thomas C Roberts, Jonathan R Hart, Minna U Kaikkonen, Marc S Weinberg, Peter K Vogt, and Kevin V Morris. Quantification of nascent transcription by bromouridine immunocapture nuclear run-on RT-qPCR. Nature protocols, 10(8):1198, 2015.

[95] Mark D. Robinson and Gordon K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics, 9(2):321–332, 08 2007.

[96] Gerson Rothschild and Uttiya Basu. Lingering questions about enhancer RNA and enhancer transcription-coupled genomic instability. Trends in Genetics, 33(2):143–154, Feb 2017.

[97] Jonathan D. Rubin, Jacob T. Stanley, Rutendo F. Sigauke, Cecilia B. Levandowski, Zachary L. Maas, Jessica Westfall, Dylan J. Taatjes, and Robin D. Dowell. Transcription factor enrichment analysis (TFEA): Quantifying the activity of hundreds of transcription factors from a single experiment. Nature Communications Biology, 2021.

[98] Ahmad Salehi, Jean-Dominique Delcroix, Pavel V. Belichenko, Ke Zhan, Chengbiao Wu, Janice S. Valletta, Ryoko Takimoto-Kimura, Alexander M. Kleschevnikov, Kumar Sambamurti, Peter P. Chung, Weiming Xia, Angela Villar, William A. Campbell, Laura Shapiro Kulnane, Ralph A. Nixon, Bruce T. Lamb, Charles J. Epstein, Gorazd B. Stokin, Lawrence S.B. Goldstein, and William C. Mobley. Increased <em>app</em> expression in a mouse model of Down's syndrome disrupts NGF transport and causes cholinergic neuron degeneration. Neuron, 51(1):29–42, Jul 2006.

[99] Matheus Sanitá Lima and David Roy Smith. Don't just dump your data and run. EMBO reports, 18(12):2087–2089, 2017.

[100] Dimitra Sarantopoulou, Soon Yew Tang, Emanuela Ricciotti, Nicholas F. Lahens, Damien Lekkas, Jonathan Schug, Xiaofeng S. Guo, Georgios K. Paschos, Garret A. FitzGerald, Allan I. Pack, and Gregory R. Grant. Comparative evaluation of RNA-seq library preparation methods for strand-specificity and low input. Scientific Reports, 9(1):13477, Sep 2019.

[101] Sarah K. Sasse, Margaret Gruca, Mary A. Allen, Vineela Kadiyala, Tengyao Song, Fabienne Gally, Arnav Gupta, Miles A. Pufall, Robin D. Dowell, and Anthony N. Gerber. Nascent transcript analysis of glucocorticoid crosstalk with TNF defines primary and cooperative inflammatory repression. Genome Research, 2019.

[102] Jeremy A. Schofield, Erin E. Duffy, Lea Kiefer, Meaghan C. Sullivan, and Matthew D. Simon. Timelapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. Nature Methods, 15(3):221–225, Mar 2018.

[103] Hong Shen and Carl G. Maki. Pharmacologic activation of p53 by small-molecule mdm2 antagonists. Current pharmaceutical design, 17(6):560–568, 2011.

[104] 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, 01 2001.

[105] Toshiyuki Shiraki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, Akira Watahiki, Mari Nakamura, Takahiro Arakawa, Shiro Fukuda, Daisuke Sasaki, Anna Podhajska, Matthias Harbers, Jun Kawai, Piero Carninci, and Yoshihide Hayashizaki. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proceedings of the National Academy of Sciences, 100(26):15776–15781, 2003.

[106] Haridha Shivram and Vishwanath R. Iyer. Identification and removal of sequencing artifacts produced by mispriming during reverse transcription in multiple RNA-seq technologies. RNA, 24(9):1266–1274, June 2018.

[107] Jake J. Siegel and Angelika Amon. New insights into the troubles of aneuploidy. Annual Review of Cell and Developmental Biology, 28(1):189–214, 2012.

[108] Boris Slobodin, Anat Bahat, Urmila Sehrawat, Shirly Becker-Herman, Binyamin Zuckerman, Amanda N. Weiss, Ruiqi Han, Ran Elkon, Reuven Agami, Igor Ulitsky, Idit Shachar, and Rivka Dikstein. Transcription dynamics regulate poly(a) tails and expression of the RNA degradation machinery to balance mRNA levels. Molecular Cell, 78(3):434–444.e5, 2020.

[109] Jason P. Smith, Arun B. Dutta, Kizhakke Mattada Sathyan, Michael J. Guertin, and Nathan C. Sheffield. Peppro: quality control and processing of nascent RNA profiling data. Genome Biology, 22(1):155, 2021.

[110] Judith Somekh, Shai S. Shen-Orr, and Isaac S. Kohane. Batch correction evaluation framework using a-priori gene-gene associations: applied to the GTEx dataset. BMC Bioinformatics, 20(1):268, May 2019.

[111] Georgios Stamoulis, Marco Garieri, Periklis Makrythanasis, Audrey Letourneau, Michel Guipponi, Nikolaos Panousis, Frédérique Sloan-Béna, Emilie Falconnet, Pascale Ribaux, Christelle Borel, Federico Santoni, and Stylianos E. Antonarakis. Single cell transcriptome

in aneuploidies reveals mechanisms of gene dosage imbalance. Nature Communications, 10(1):4495, Oct 2019.

[112] Iris Steinparzer, Vitaly Sedlyarov, Jonathan D. Rubin, Kevin Eislmayr, Matthew D. Galbraith, Cecilia B. Levandowski, Terezia Vcelkova, Lucy Sneezum, Florian Wascher, Fabian Amman, Renata Kleinova, Heather Bender, Zdenek Andrysik, Joaquin M. Espinosa, Giulio Superti-Furga, Robin D. Dowell, Dylan J. Taatjes, and Pavel Kovarik. Transcriptional responses to IFN-γ require mediator kinase-dependent pause release and mechanistically distinct CDK8 and CDK19 functions. Molecular Cell, 76(3):485–499.e8, 2019.

[113] Carsten Sticht, Carolina De La Torre, Alisha Parveen, and Norbert Gretz. miRWalk: An online resource for prediction of microRNA binding sites. PLOS ONE, 13(10):1–6, 10 2018.

[114] Georg Stoecklin, Min Lu, Bernd Rattenbacher, and Christoph Moroni. A constitutive decay element promotes tumor necrosis factor alpha mRNA degradation via an AU-rich element-independent pathway. Molecular and Cellular Biology, 23(10):3506–3515, 2003.

[115] John D. Storey, Jennifer Madeoy, Jeanna L. Strout, Mark Wurfel, James Ronald, and Joshua M. Akey. Gene-expression variation within and among human populations. The American Journal of Human Genetics, 80(3):502–509, 2015/04/23 2007.

[116] Zhenqiang Su, Paweł P. Łabaj, Sheng Li, Jean Thierry-Mieg, Danielle Thierry-Mieg, Wei Shi, Charles Wang, Gary P. Schroth, Robert A. Setterquist, John F. Thompson, Wendell D. Jones, Wenzhong Xiao, Weihong Xu, Roderick V. Jensen, Reagan Kelly, Joshua Xu, Ana Conesa, Cesare Furlanello, Hanlin Gao, Huixiao Hong, Nadereh Jafari, Stan Letovsky, Yang Liao, Fei Lu, Edward J. Oakeley, Zhiyu Peng, Craig A. Praul, Javier Santoyo-Lopez, Andreas Scherer, Tieliu Shi, Gordon K. Smyth, Frank Staedtler, Peter Sykacek, Xin-Xing Tan, E. Aubrey Thompson, Jo Vandesompele, May D. Wang, Jian Wang, Russell D. Wolfinger, Jiri Zavadil, Scott S. Auerbach, Wenjun Bao, Hans Binder, Thomas Blomquist, Murray H. Brilliant, Pierre R. Bushel, Weimin Cai, Jennifer G. Catalano, Ching-Wei Chang, Tao Chen, Geng Chen, Rong Chen, Marco Chierici, Tzu-Ming Chu, Djork-Arné Clevert, Youping Deng, Adnan Derti, Viswanath Devanarayan, Zirui Dong, Joaquin Dopazo, Tingting Du, Hong Fang, Yongxiang Fang, Mario Fasold, Anita Fernandez, Matthias Fischer, Pedro Furió-Tari, James C. Fuscoe, Florian Caimet, Stan Gaj, Jorge Gandara, Huan Gao, Weigong Ge, Yoichi Gondo, Binsheng Gong, Meihua Gong, Zhuolin Gong, Bridgett Green, Chao Guo, Lei Guo, Li-Wu Guo, James Hadfield, Jan Hellemans, Sepp Hochreiter, Meiwen Jia, Min Jian, Charles D. Johnson, Suzanne Kay, Jos Kleinjans, Samir Lababidi, Shawn Levy, Quan-Zhen Li, Li Li, Peng Li, Yan Li, Haiqing Li, Jianying Li, Shiyong Li, Simon M. Lin, Francisco J. López, Xin Lu, Heng Luo, Xiwen Ma, Joseph Meehan, Dalila B. Megherbi, Nan Mei, Bing Mu, Baitang Ning, Akhilesh Pandey, Javier Pérez-Florido, Roger G. Perkins, Ryan Peters, John H. Phan, Mehdi Pirooznia, Feng Qian, Tao Qing, Lucille Rainbow, Philippe Rocca-Serra, Laure Sambourg, Susanna-Assunta Sansone, Scott Schwartz, Ruchir Shah, Jie Shen, Todd M. Smith, Oliver Stegle, Nancy Stralis-Pavese, Elia Stupka, Yutaka Suzuki, Lee T. Szkotnicki, Matthew Tinning, Bimeng Tu, Joost van Delft, Alicia Vela-Boza, Elisa Venturini, Stephen J. Walker, Liqing Wan, Wei Wang, Jinhui Wang, Jun Wang, Eric D. Wieben, James C. Willey, Po-Yen Wu, Jiekun Xuan, Yong Yang, Zhan Ye, Ye Yin, Ying Yu, Yate-Ching Yuan, John Zhang, Ke K. Zhang, Wenqian Zhang, Wenwei Zhang, Yanyan Zhang, Chen Zhao, Yuanting Zheng, Yiming Zhou, Paul Zumbo, Weida Tong, David P. Kreil, Christopher E. Mason, Leming Shi, and S. E. Q. C. /. M. A. Q. C.-I. I. I. Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility

and information content by the sequencing quality control consortium. Nature Biotechnology, 32(9):903–914, Sep 2014.

[117] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43):15545–15550, October 2005.

[118] Kelly D Sullivan, Hannah C Lewis, Amanda A Hill, Ahwan Pandey, Leisa P Jackson, Joseph M Cabral, Keith P Smith, L Alexander Liggett, Eliana B Gomez, Matthew D Galbraith, James DeGregori, and Joaquín M Espinosa. Trisomy 21 consistently activates the interferon response. eLife, 5:e16220, jul 2016.

[119] Christopher Tiedje, Manuel D. Diaz-Muñoz, Philipp Trulley, Helena Ahlfors, Kathrin Laaß, Perry J. Blackshear, Martin Turner, and Matthias Gaestel. The RNA-binding protein TTP is a global post-transcriptional regulator of feedback control in inflammation. Nucleic Acids Research, 44(15):7418–7440, 05 2016.

[120] Tanya Todorova, Florian J. Bock, and Paul Chang. PARP13 regulates cellular mRNA post-transcriptionally and functions as a pro-apoptotic factor by destabilizing TRAILR4 transcript. Nature Communications, 5(1):5362, Nov 2014.

[121] Eduardo M Torres, Michael Springer, and Angelika Amon. No current evidence for widespread dosage compensation in *S. cerevisiae*. eLife, 5:e10996, mar 2016.

[122] Guido Van Rossum and Fred L. Drake. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA, 2009.

[123] Judith Verhelst, Eef Parthoens, Bert Schepens, Walter Fiers, and Xavier Saelens. Interferon-inducible protein Mx1 inhibits influenza virus by interfering with functional viral ribonucleoprotein complex assembly. Journal of Virology, 86(24):13445–13455, 2012.

[124] Dong Wang, Ivan Garcia-Bassets, Chris Benner, Wenbo Li, Xue Su, Yiming Zhou, Jinsong Qiu, Wen Liu, Minna U. Kaikkonen, Kenneth A. Ohgi, Christopher K. Glass, Michael G. Rosenfeld, and Xiang-Dong Fu. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature, 474(7351):390–394, Jun 2011.

[125] Jing Wang, Yue Zhao, Xiaofan Zhou, Scott W. Hiebert, Qi Liu, and Yu Shyr. Nascent RNA sequencing analysis provides insights into enhancer-mediated gene regulation. BMC Genomics, 19(1):633, Aug 2018.

[126] Lei Wang, Sara J. Felts, Virginia P. Van Keulen, Larry R. Pease, and Yuji Zhang. Exploring the effect of library preparation on RNA sequencing experiments. Genomics, 111(6):1752–1759, 2019.

[127] Katherine A. Waugh, Ross Minter, Jessica Baxter, Congwu Chi, Kathryn D. Tuttle, Neetha P. Eduthan, Matthew D. Galbraith, Kohl T. Kinning, Zdenek Andrysik, Paula Araya, Hannah Dougherty, Lauren N. Dunn, Michael Ludwig, Kyndal A. Schade, Dayna Tracy, Keith P. Smith, Ross E. Granrath, Nicolas Busquet, Santosh Khanal, Ryan D. Anderson, Liza L. Cox, Belinda Enriquez Estrada, Angela L. Rachubinski, Hannah R. Lyford, Eleanor C. Britton,

David J. Orlicky, Jennifer L. Matsuda, Kunhua Song, Timothy C. Cox, Kelly D. Sullivan, and Joaquin M. Espinosa. Interferon receptor gene dosage determines diverse hallmarks of Down syndrome. bioRxiv, 2022.

[128] Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer, 2016.

[129] Claus O. Wilke. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2', 2020. R package version 1.1.1.

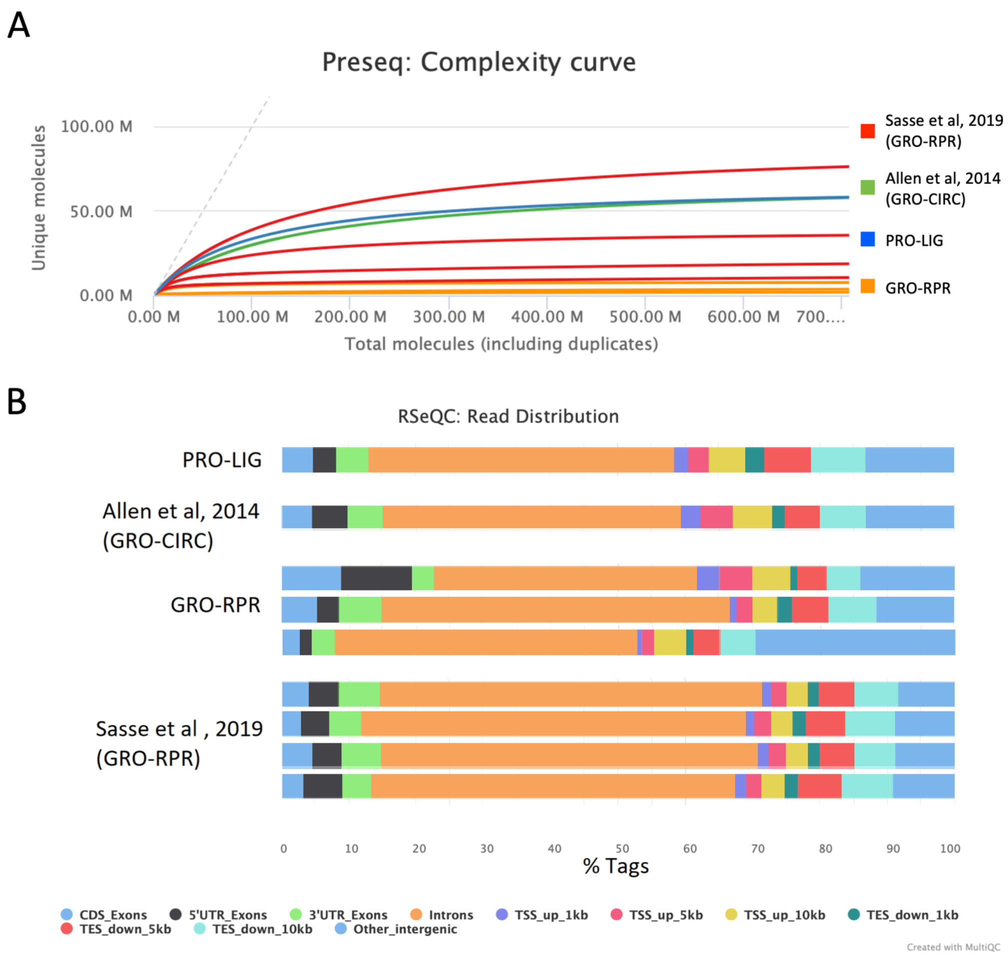
[130] Erin M. Wissink, Anniina Vihervaara, Nathaniel D. Tippens, and John T. Lis. Nascent RNA analyses: tracking transcription and its regulation. Nature Reviews Genetics, 20(12):705–723, Dec 2019.

[131] Nathan Wong and Xiaowei Wang. miRDB: an online resource for microRNA target prediction and functional annotations. Nucleic Acids Research, 43(D1):D146–D152, 11 2014.

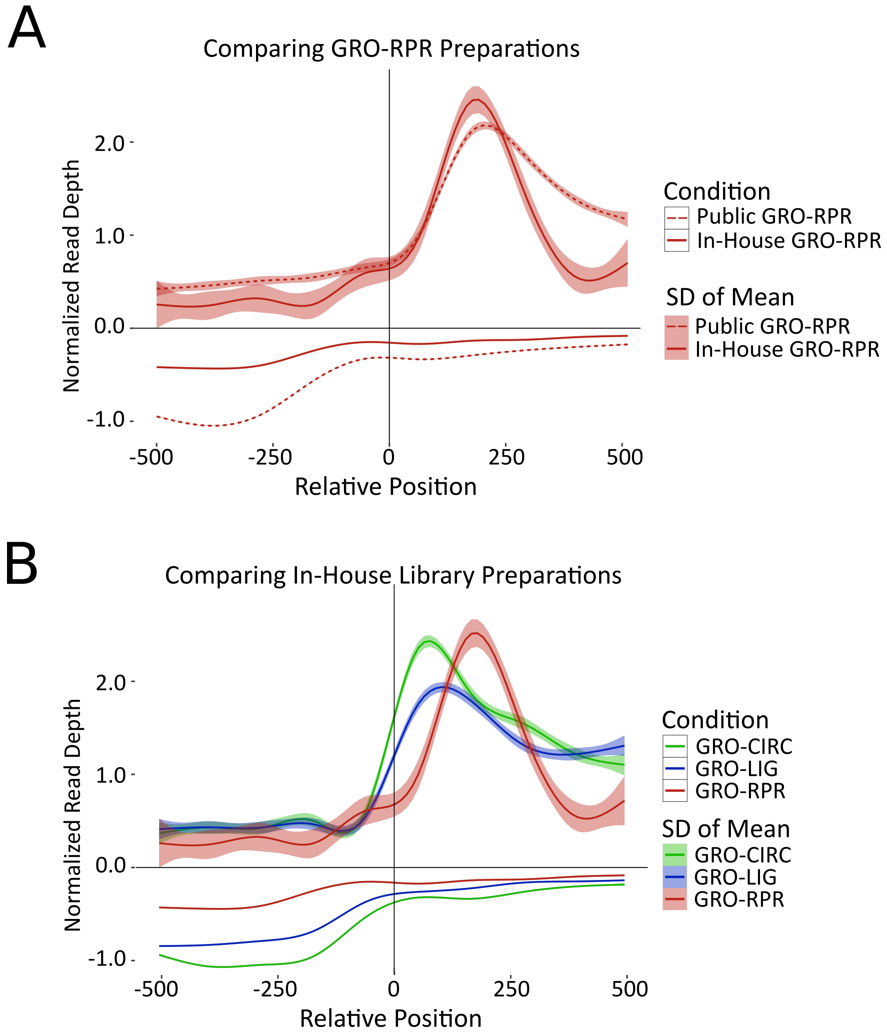
[132] Yuqing Zhang, Giovanni Parmigiani, and W Evan Johnson. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics and Bioinformatics, 2(3), 09 2020. lqaa078.

Appendix A: Chapter 2 Supplementary Material

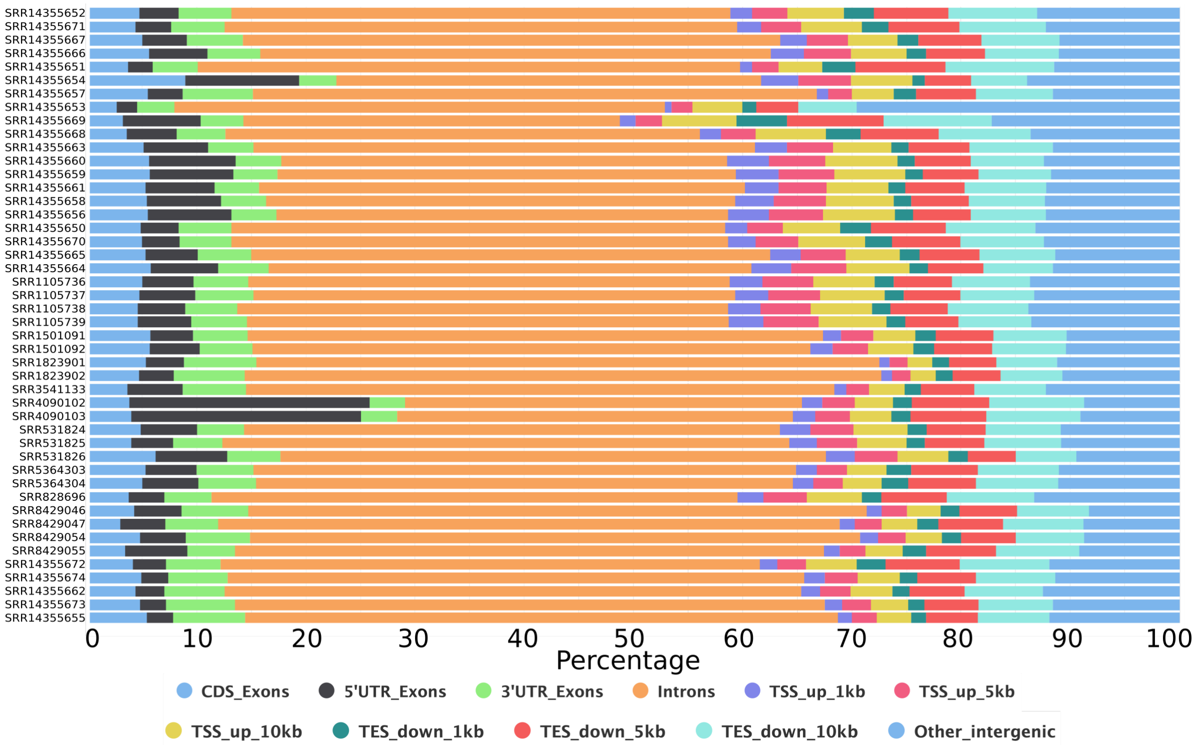
Supplemental Table 1 is available online at https://static-content.springer.com/esm/art\%3A10.1186\%2Fs12864-022-08352-8/MediaObjects/12864_2022_8352_MOESM1_ESM.xlsx



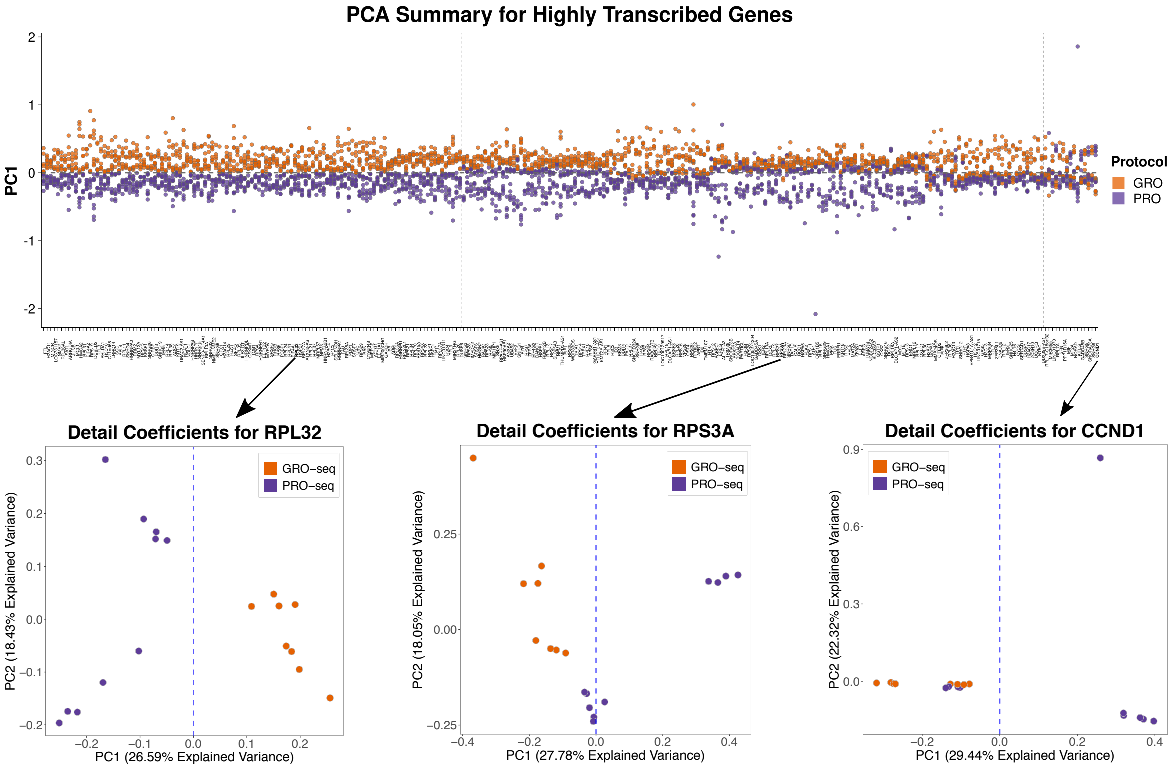
Supplementary Figure 1: **Preseq complexity curves and RSeQC Read Distribution Graphs of RPR Datasets** (A) Complexity curves of 4 publicly available GRO-RPR datasets ([101]: SRR8429046, SRR8429047, SRR8429054, SRR8429055), our in-house generated GRO-RPR datasets (see Supplemental Table 1, Materials and Methods. SRR14355654, SRR14355657, and SRR14355653), one PRO-LIG dataset (SRR14355672), and one publicly available GRO-CIRC dataset ([4]: SRR1105737). While the most complex library was from a GRO-RPR preparation, we found that the majority of these RPR datasets tended to be of lower complexity. Despite this trend, we contend that there is insufficient data to determine whether this is a fault of our handling or a feature of RPR library preparations with RO-seq datasets. (B) Read distribution plots of the datasets described in (A). While many regions were consistent regardless of protocol, was considerable variation in read distributions within the GRO-RPR datasets, especially comparing the proportion of reads found in 5′ UTR regions and intergenic regions. As such, we chose to summarize additional quality metrics and library characteristics for our GRO-RPR datasets (Fig. 2.2D,2.3D,2.4B, see also Supplementary Table 1), with the understanding that their poor quality influence these metrics. GRO-RPR datasets were otherwise not used for further comparative analysis. From top to bottom, the samples are as follows: SRR14355672 (PRO-LIG); SRR1105737 (GRO-CIRC); SRR14355654, SRR14355657, SRR14355653 (GRO-RPR); SRR8429046, SRR8429047, SRR8429054, SRR8429055[101].



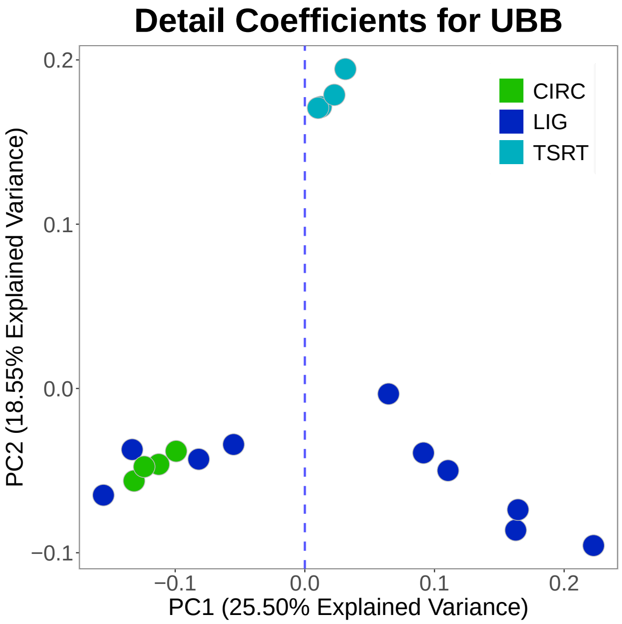
Supplementary Figure 2: **Metagenes of Public GRO-RPR and in-house libraries** (A) Metagenes of public GRO-RPR and in-house GRO-RPR libraries. All GRO-RPR datasets display a similar gap in coverage near the annotated TSS. Note that each public GRO-RPR library was subsampled to 20 million reads such that the comparison was performed at the same depth. (Public GRO-RPR data: SRR8429046, SRR8429047, SRR8429054, SRR8429055) (B) Metagenes of in-house libraries, including GRO-RPR libraries. Each library was subsampled to 20 million reads to match the lower depth of the GRO-RPR libraries. Additionally, we note that our GRO-RPR libraries are lower complexity. For both metaplots, genes shorter than 2000 bp, genes with significant signal 1 kb upstream (>1% of upstream bases covered), and genes with low coverage (TPM < .01) were removed. (n=1428) (GRO-CIRC: SRR1105736, SRR1105737. GRO-LIG: SRR14355673, SRR14355674. GRO-RPR: SRR14355653, SRR14355654, SRR14355657)



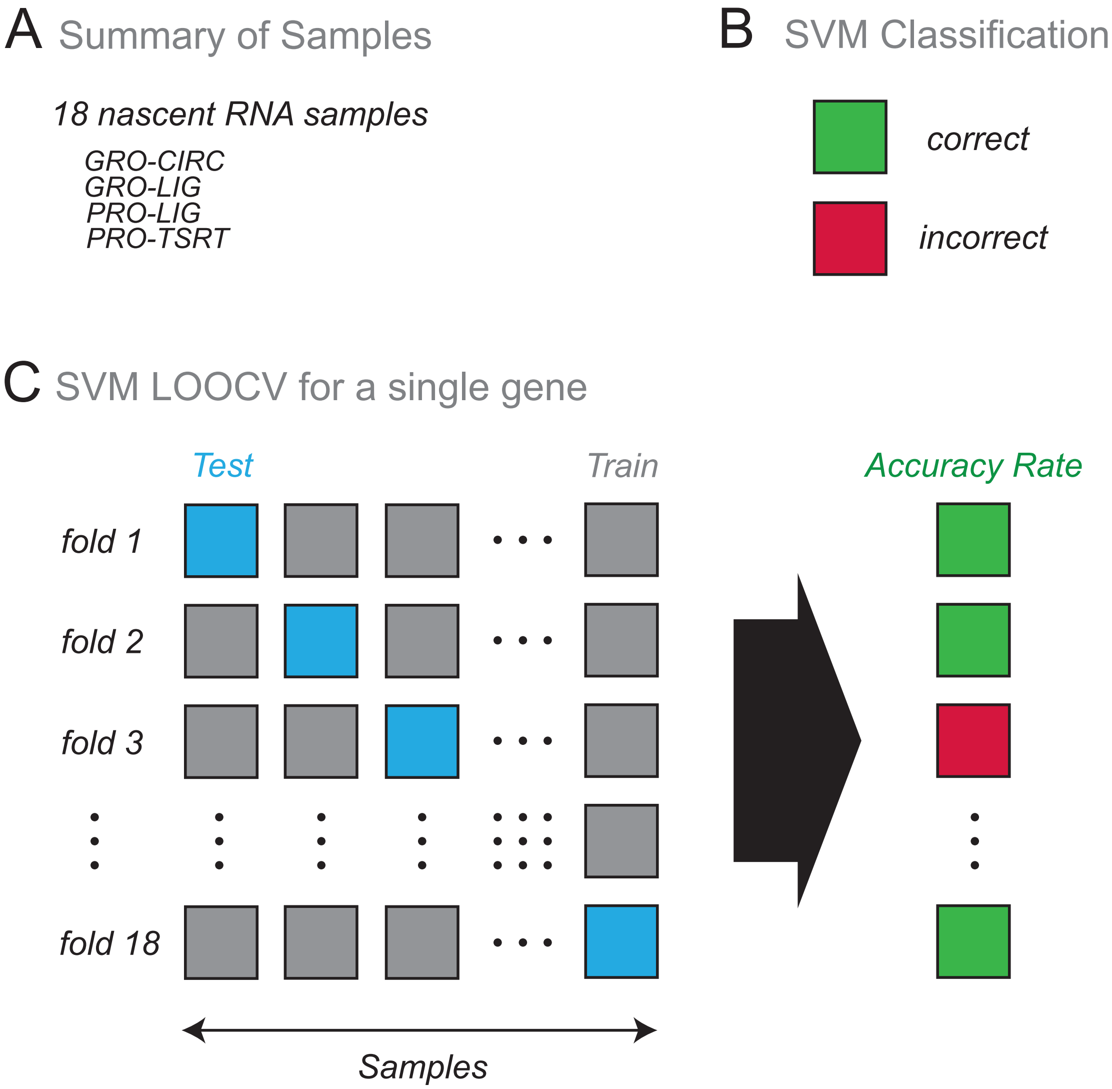
Supplementary Figure 3: **Read Distribution of all libraries in analysis** Read distributions were generated from RSeQC, see Materials and Methods, Supplemental Table 1.



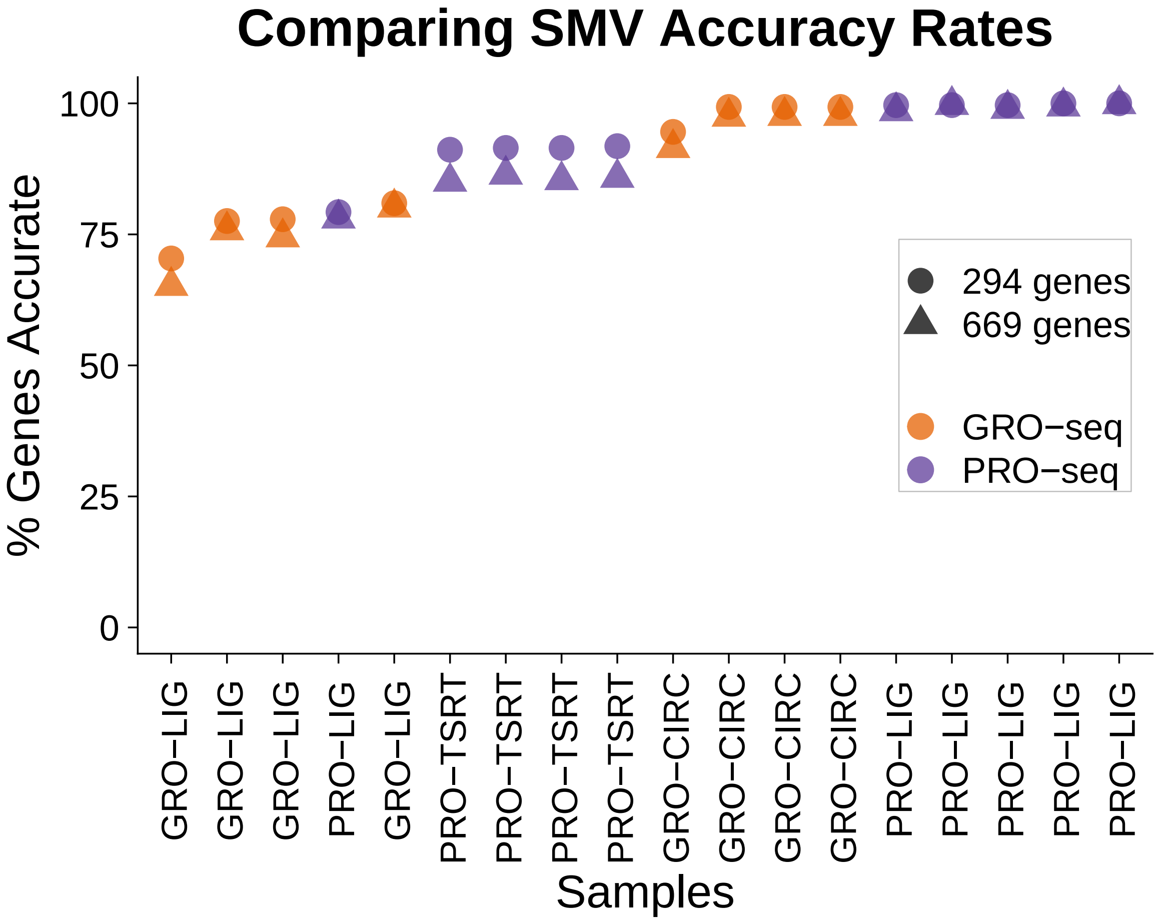
Supplementary Figure 4: **Discrete wavelet transform PCA results for 294 highly transcribed genes** (Top) PC1 effectively separates GRO and PRO libraries for 39.8% (117 genes) of the set of 294 highly transcribed genes while 55.1% (162 genes) of the genes separates the libraries on PC1 and PC2. (Bottom) PC1 and PC2 results for each library are shown for three example genes: RPL32 (separates on PC1), RPS3A (separates on a plane in the PC1/PC2 space), and CCND1 (not separable with these PC).



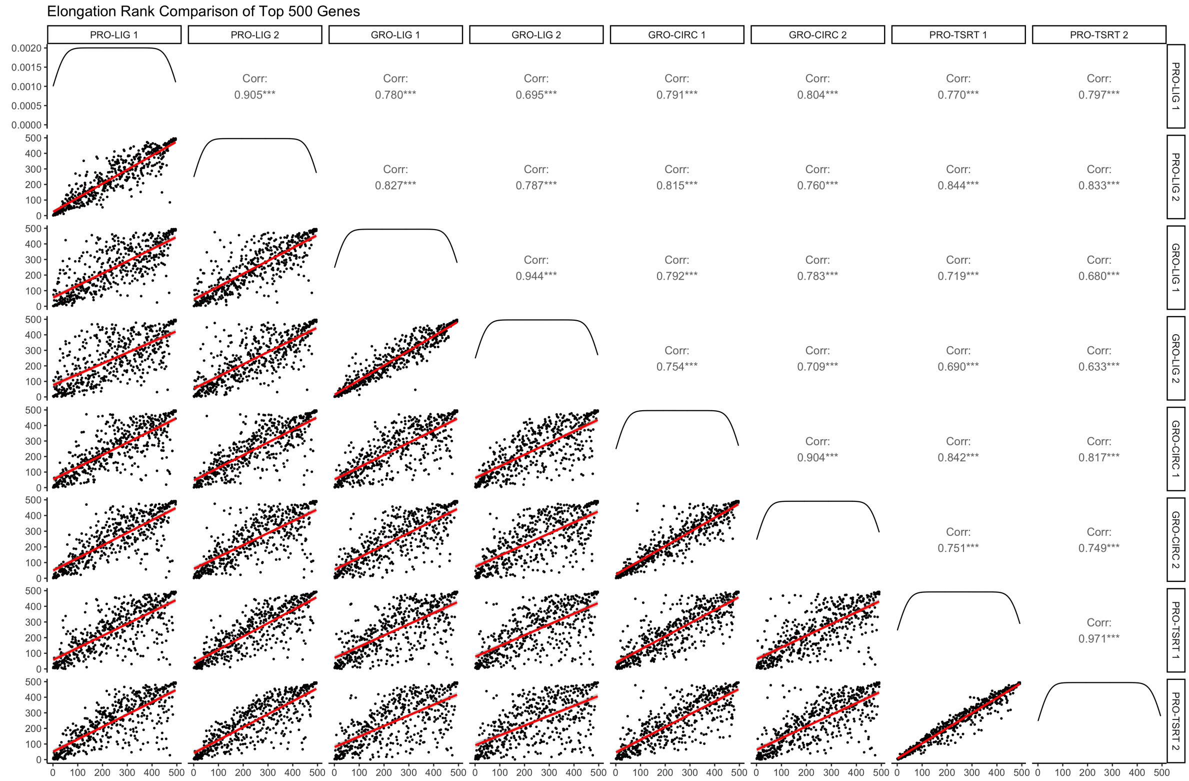
Supplementary Figure 5: **DWT PCA results of detail coefficients at UBB locus.** PCA results for UBB locus, as in Figure 2.2F. Results are colored by library preparation method. At this locus, the results cluster less distinctly by library preparation method, compared to the enrichment protocol.



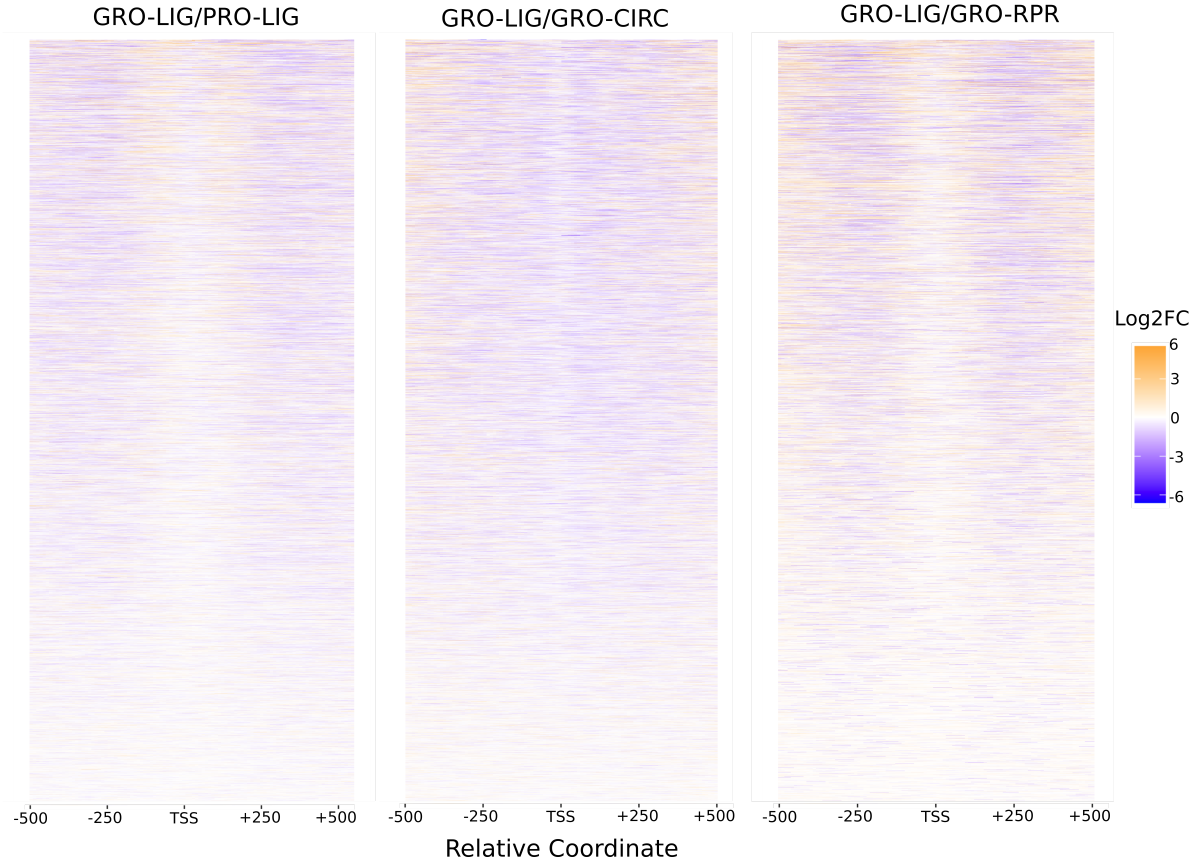
Supplementary Figure 6: **Schematic for the Support Vector Machine Leave one out cross validation analysis.** (A) Eighteen nascent RNA sequencing samples were used as input, from GRO-CIRC, GRO-LIG, PRO-LIG and PRO-TSRT libraries. (B) SVM classification was considered correct if the protocol was inferred from the data. (C) Given a gene, eighteen consecutive leave one out tests were performed. In each, one sample was selected as a test sample while the other samples were used as the training set. The SVM classification was subsequently evaluated for accuracy. Based on the SVM LOOCV method, a majority of the genes (>75%) accurately classified the protocol for the 18 samples.



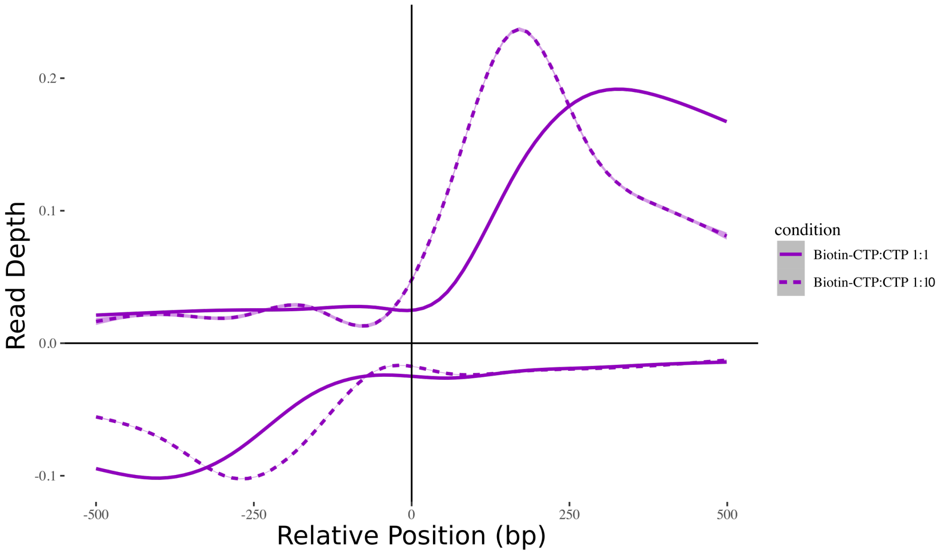
Supplementary Figure 7: **SVM results for highly transcribed genes.** The accuracy rate for the classifier remained mostly unchanged for both the top 294 and top 669 genes with high coefficient of variation (CV less than 0.85 and average TPM greater than 100).



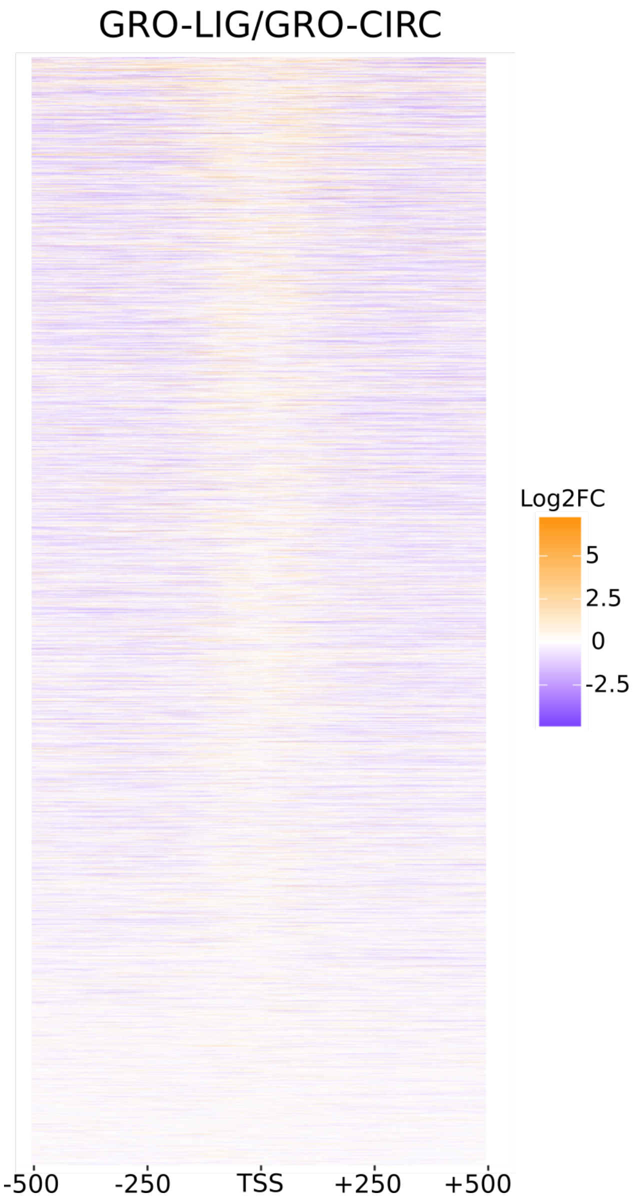
Supplementary Figure 8: **Scatterplot matrix of elongation regions.** Only the top 500 genes (by TPM) were considered. There is considerably more correlation in elongation regions versus pause regions at these genes, suggesting more variability occurs near the TSS across protocols. Each replicate dataset is a biological replicate (see Supplemental Table 1).



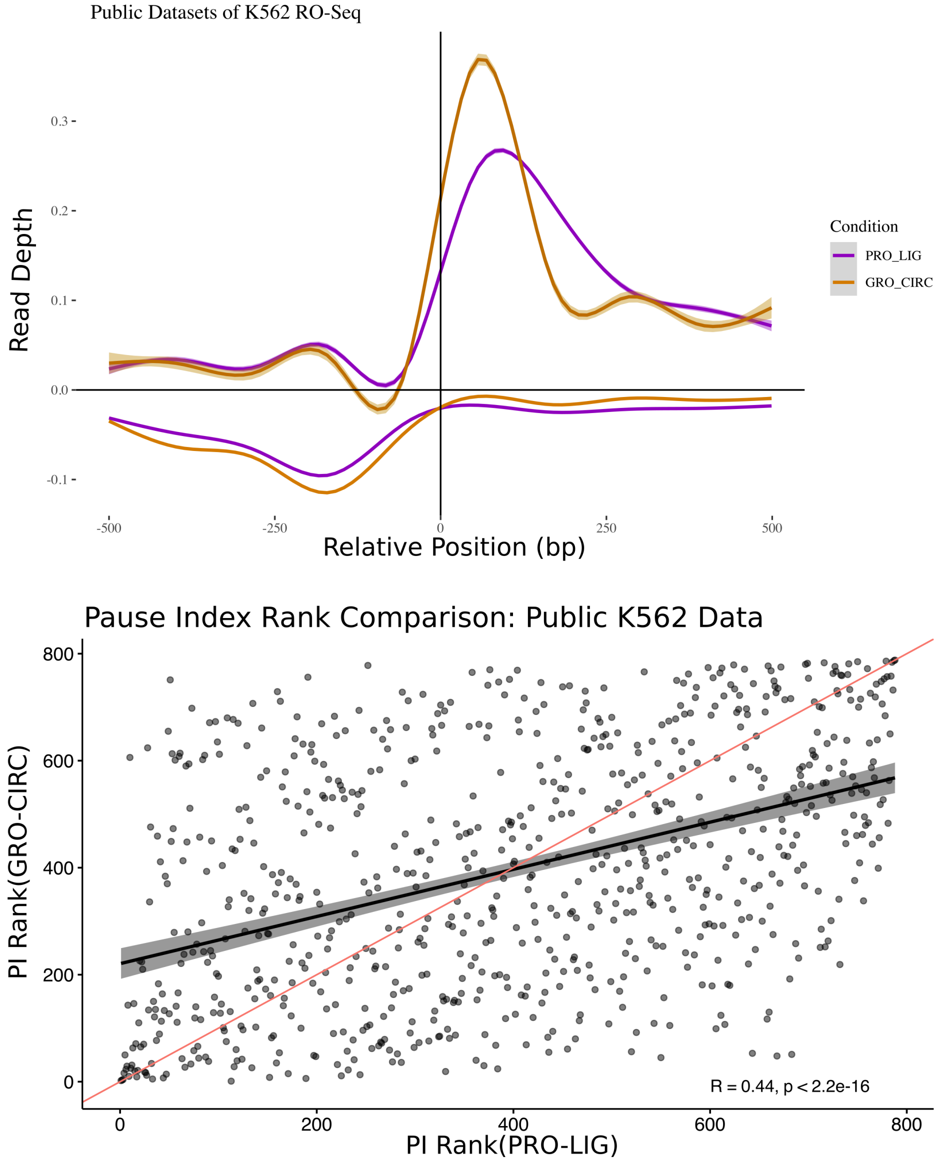
Supplementary Figure 9: **Heatmap of read ratios of pause regions in GRO-CIRC,GRO-LIG, GRO-RPR, and PRO-LIG libraries.** (TSS +/- 500 bp, 10 bp per window; RefSeq hg38 gene annotations were used.) Genes shorter than 2000 bp were not included. A pseudocount of 1 was added to all libraries to avoid undefined values. There is comparatively lower coverage near the TSS in many genes, representing the center of bidirectional transcription. This is especially prevalent in GRO-RPR and PRO-LIG libraries.



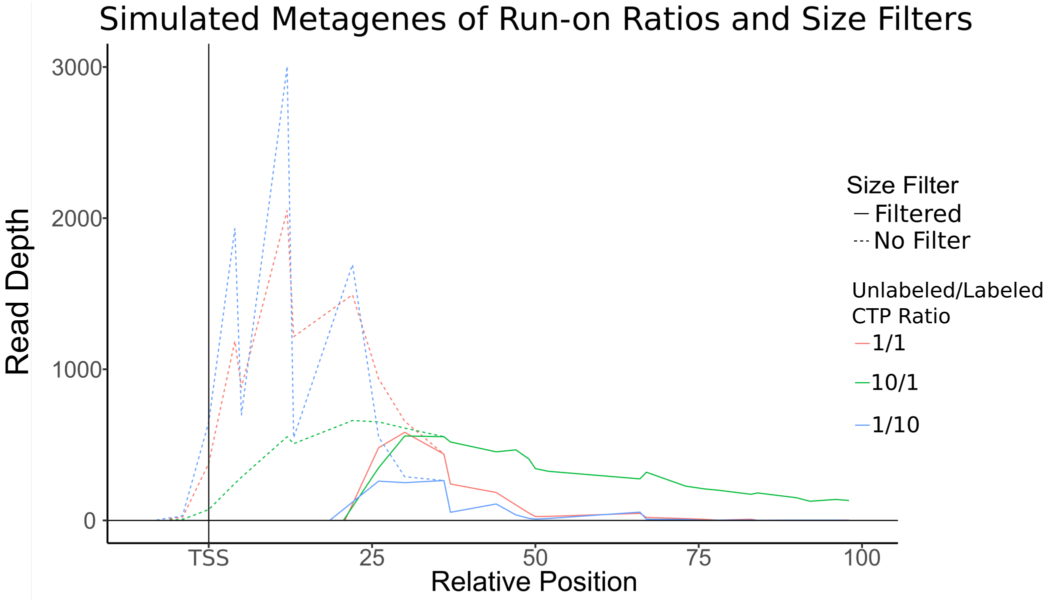
Supplementary Figure 10: **Metagenes of PRO-LIG libraries with varying Biotin ratios.** Libraries generated from HCT116 cell treated with DMSO, using the PRO-LIG protocol and library preparation strategies. Libraries differed only in the relative amounts of unlabeled CTP added (See Materials and Methods).



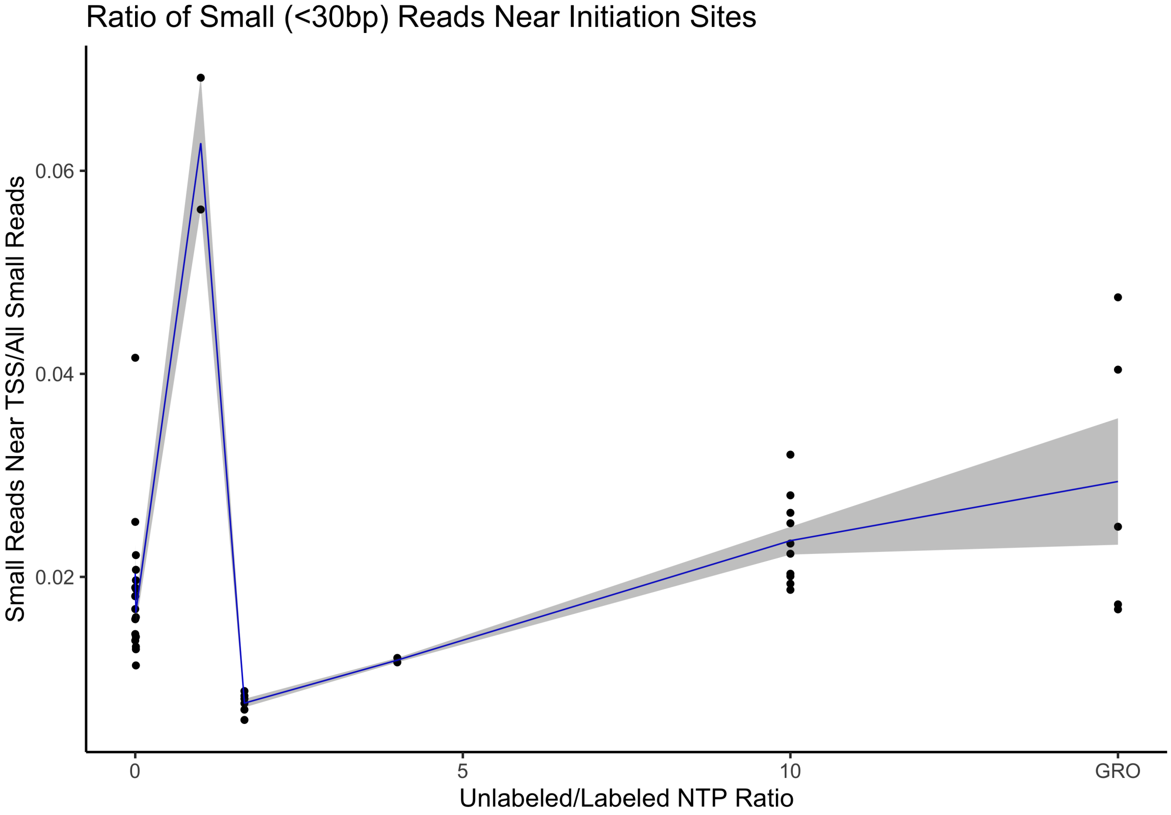
Supplementary Figure 11: **Ratio of reads near TSS in public datasets.** (TSS +/- 500 bp, 10 bp per window; RefSeq hg38 gene annotations were used.) Genes shorter than 2000 bp were not included. A pseudocount of 1 was added to all libraries to avoid undefined values. There is considerably more signal in the analyzed GRO-LIG library near the TSS, suggesting additional factors such as size selection contribute to disparities near these regions. (Public GRO-LIG: SRR1501091, SRR1501092; Public GRO-CIRC: SRR4090102, SRR4090103)



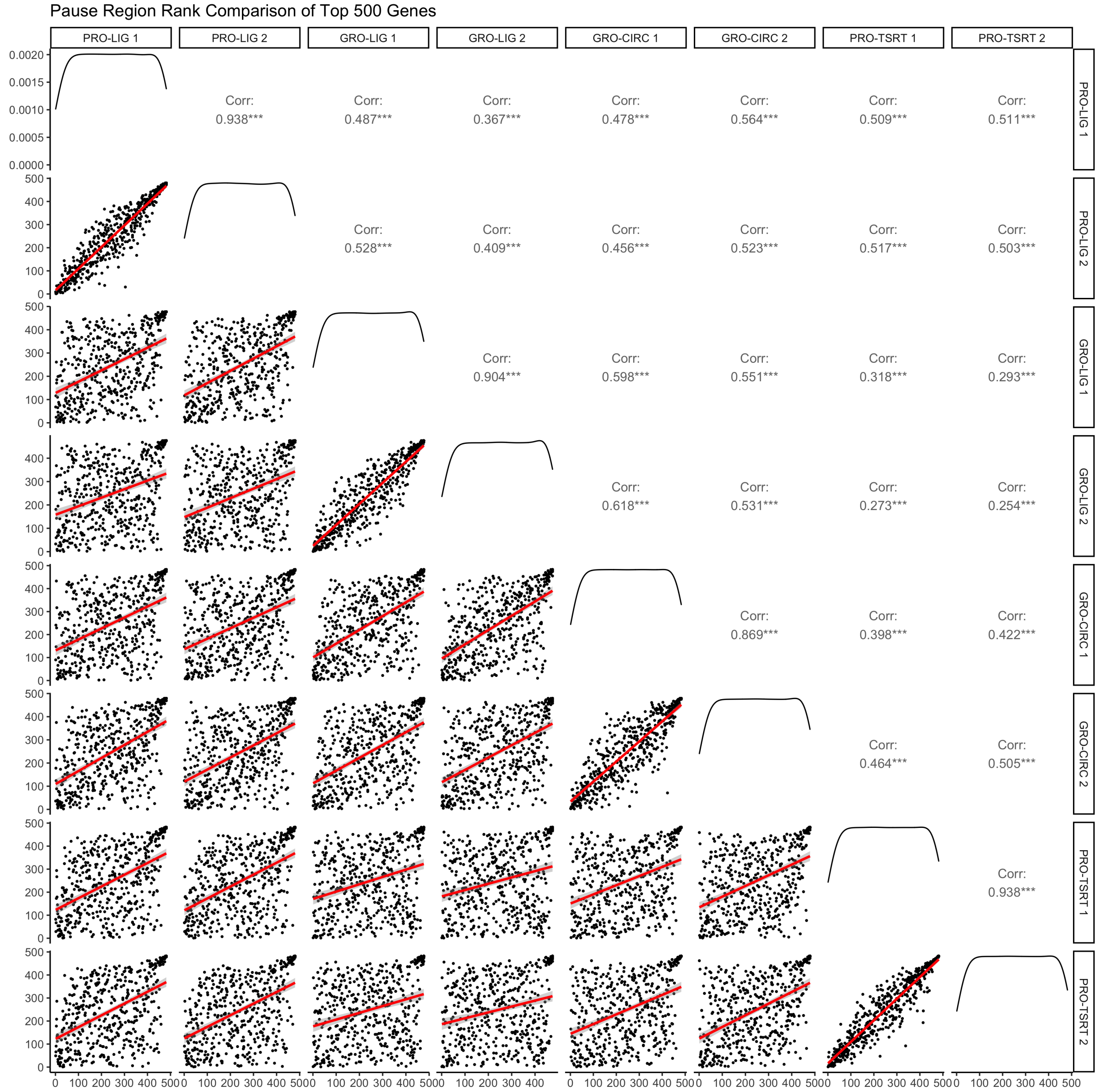
Supplementary Figure 12: **Metagene and Pause Index Comparison of Public K562 Data.** (Top) Metagenes of public datasets[81, 31]. Libraries were generated from K562 Cells treated with DMSO and prepped with either PRO-LIG or GRO-CIRC methods. PRO-LIG libraries were prepared with all 4 NTPs labeled with biotin during the run-on reaction. While the peak of these distributions occur at different relative locations than our datasets, we note that the PRO library still shows a peak that is further downstream than the comparative GRO library. (Bottom) Public data[81, 31] were subjected to analysis as in Fig. 2.3C, left (see Supplemental Table 1). PI regions were defined as in Fig. 2.3. Notably, the rank correlation remains low (R=0.44) consistent with PI differences being driven by protocol. Public GRO-CIRC: SRR1823901 and SRR1823902. Public PRO-LIG: SRR5364303 and SRR5364304, see Supplemental Table 1.



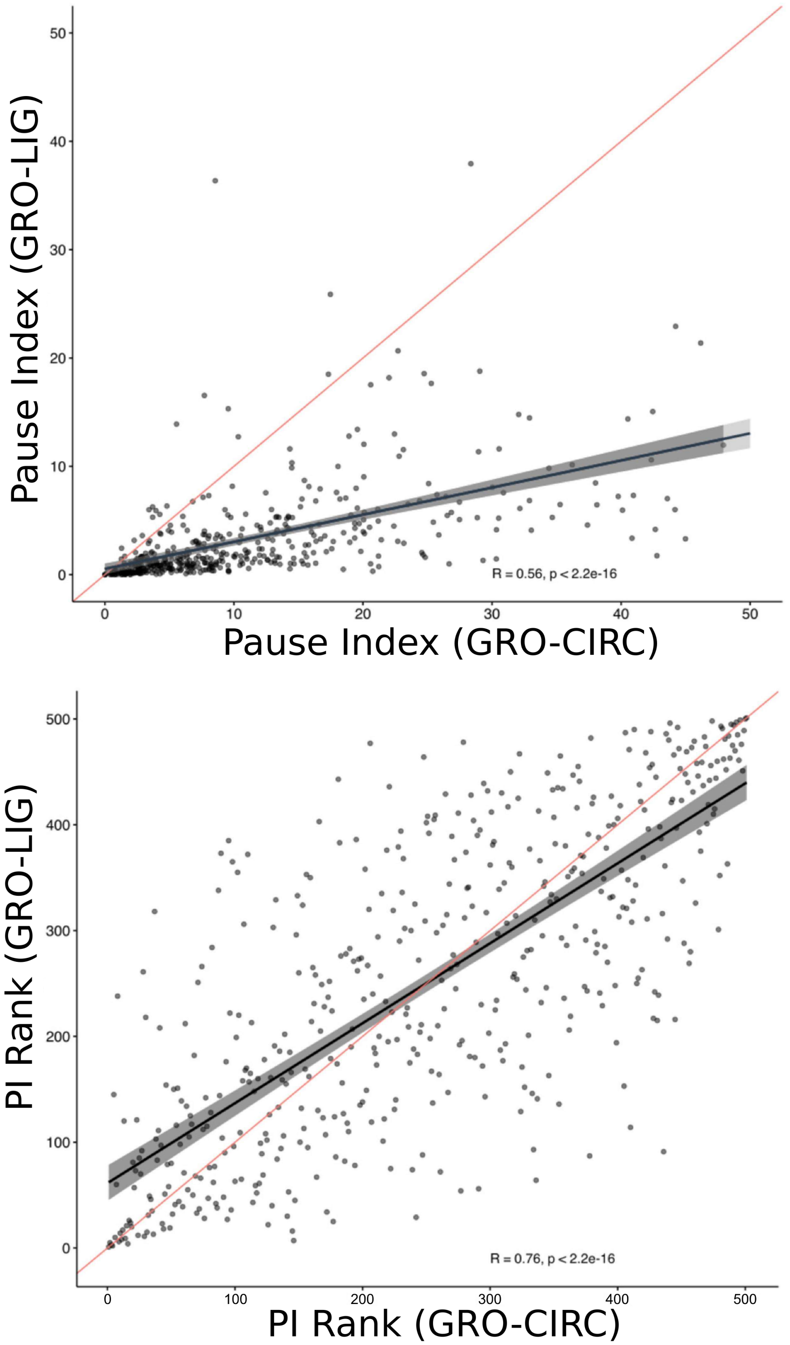
Supplementary Figure 13: **Simulated metagenes using different run-on ratios and size selection criteria.** Reads were generated *in silico* from a simulated gene template (see Materials and Methods), using run-on ratios to inform read positions and length. Small reads (approx. <25bp, see Materials and Methods) were either filtered out (solid lines), or kept in (dotted lines). As expected, with increasing NTP concentration the peak moves downstream (dashed lines). However, the size selection subsequently alters the location of the visible peak (solid lines) based on the proportion of the data that passes beyond the filter. In this way, the two protocol steps interact to influence the location observed for the 5′ peak. Here, for example, both the filtered 10/1 (green) and 1/1 (red) tracks report a 5′ peak near 28 bp, whereas the filtered 1/10 (blue) track reports a 5′ near 38 bp. Additionally, the read distribution in shifted towards the TSS in the filtered 1/1 track relative to the 10/1 track.



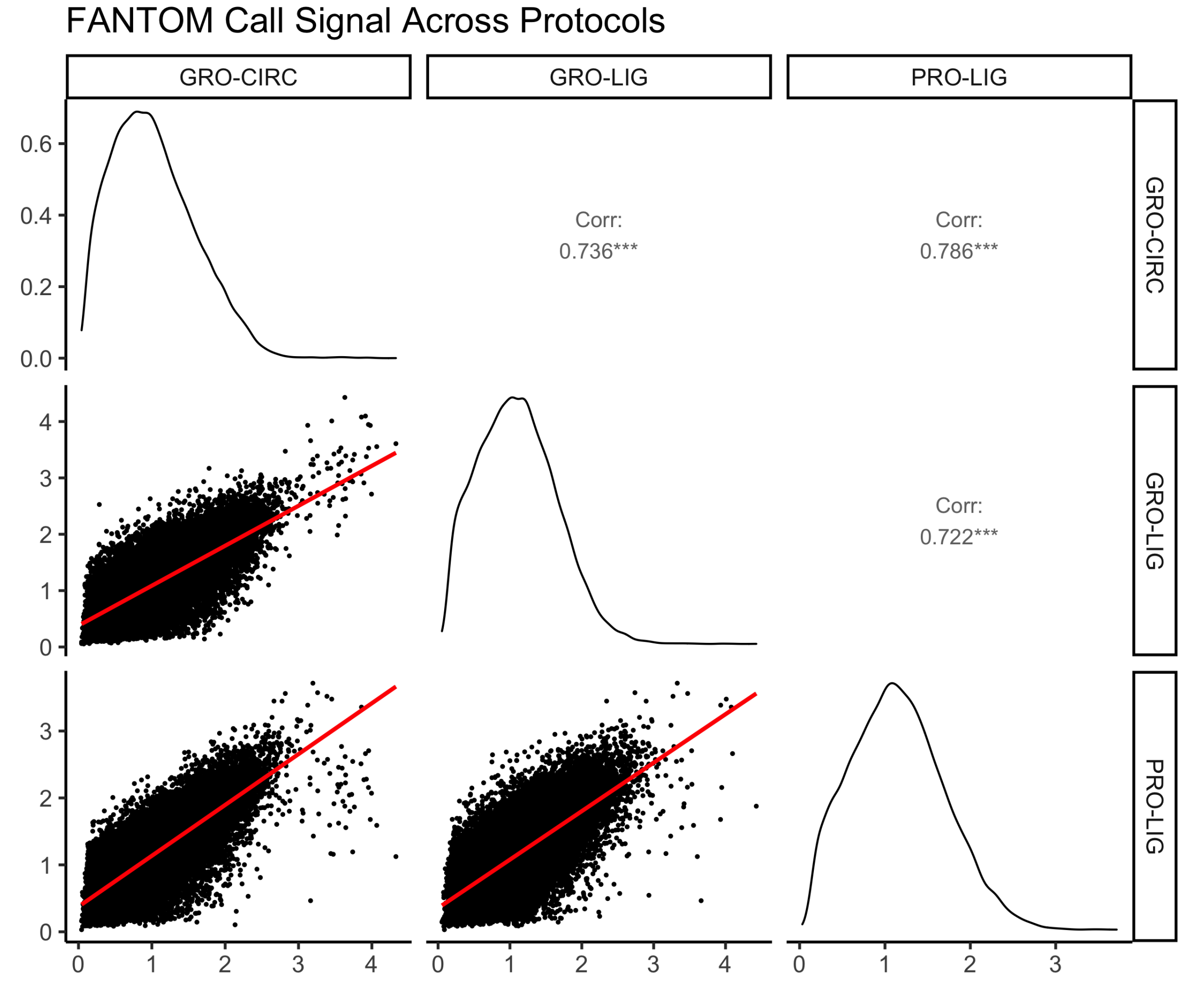
Supplementary Figure 14: **Ratio of small reads near TSS versus all small reads.** We reasoned that this ratio would be informative of the mixture of labeled and unlabeled NTPs in the run-on reaction. Based on publicly available data and our own in-house data (see Materials and Methods for full list of samples analyzed), there appears to be a trend in this ratio, although not a monotonically increasing function. The scarcity of different run-on ratios in public data do not warrant an estimate on an "ideal" ratio from these data; however, we note that these data are consistent with our in silico simulations.



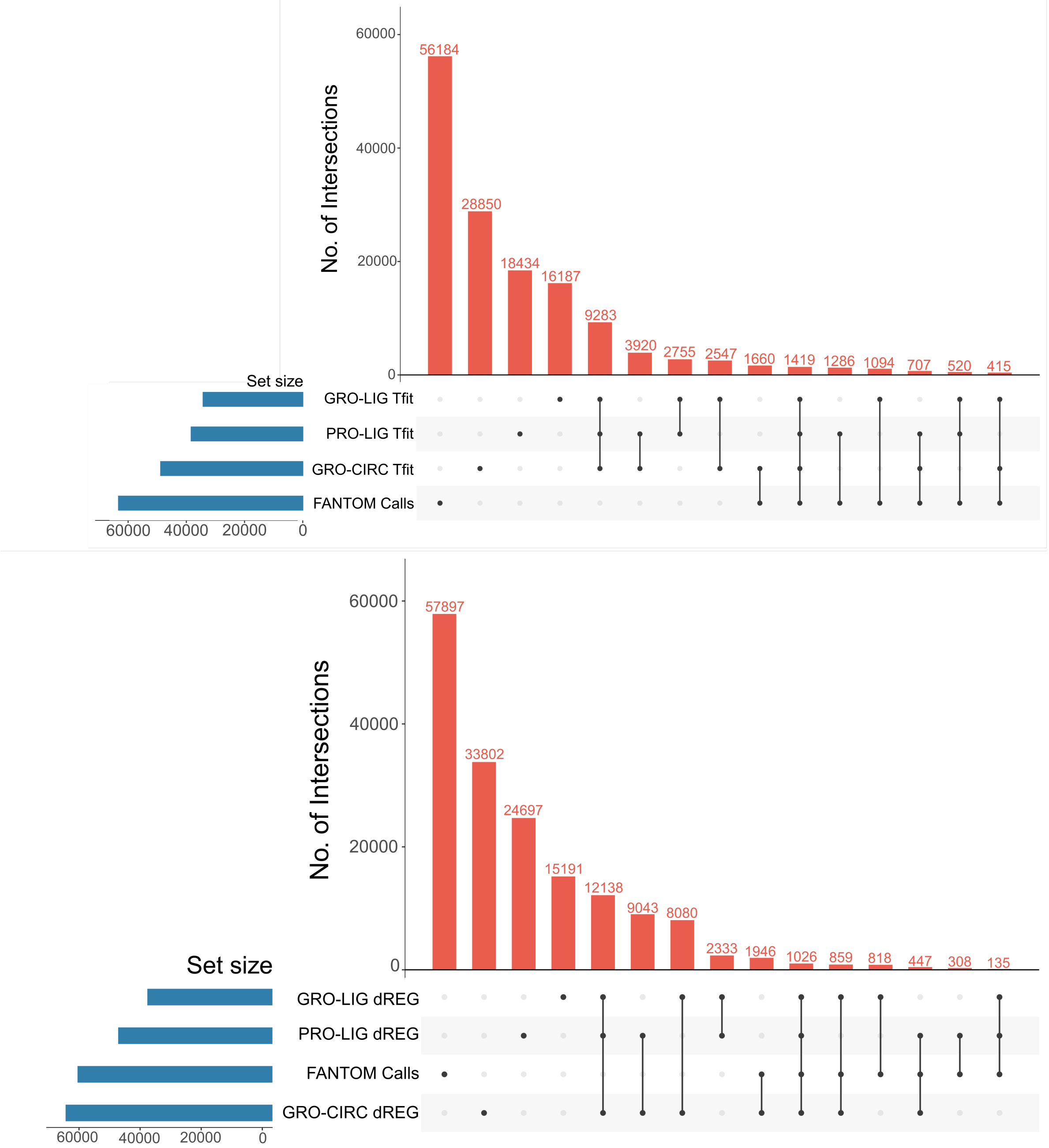
Supplementary Figure 15: **Scatterplot matrix of counts within the pause region of the top 500 genes.** (pause region:-50 to +250 from RefSeq hg38 TSS annotation). There is considerable variation between protocols at these regions. Replicates shown are biological replicates (see Supplemental Table 1).



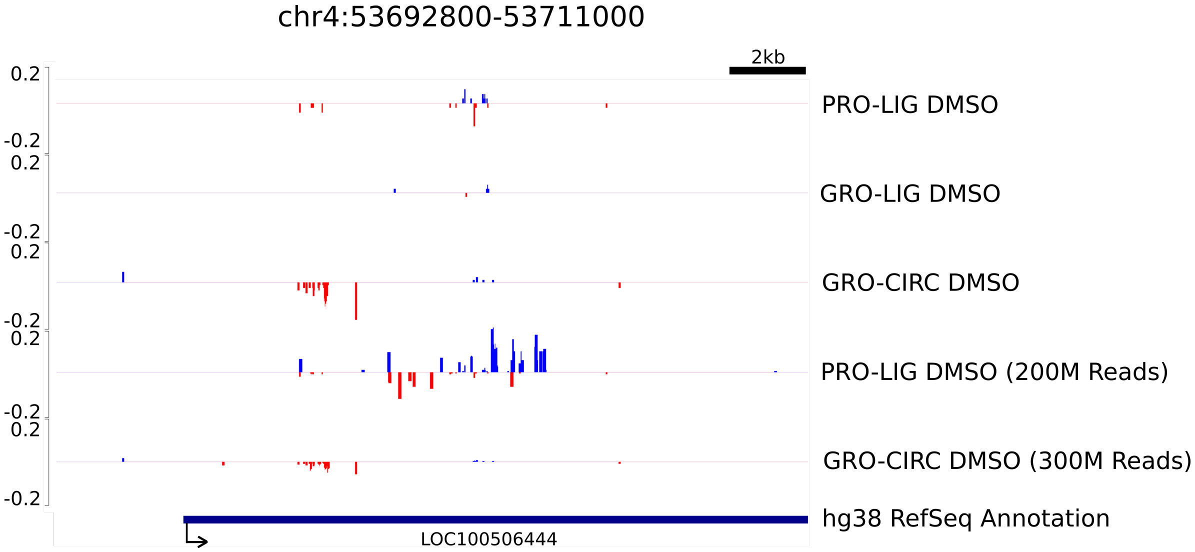
Supplementary Figure 16: **Pause index (PI) and rank correlation of PI generated from GRO-CIRC and GRO-LIG libraries.** Pause indices generated using a different pause region definition than Fig. 2.3E. Namely here the pause ratio is TSS to +80, elongation region +81:TES-1000 (genes shorter than 2000 bp were not included) and features were counted with featureCounts. In spite of using both a distinct interval and counting scheme, the pausing ratio remains poorly correlated (here Pearson R=0.56, Spearman R=0.76).



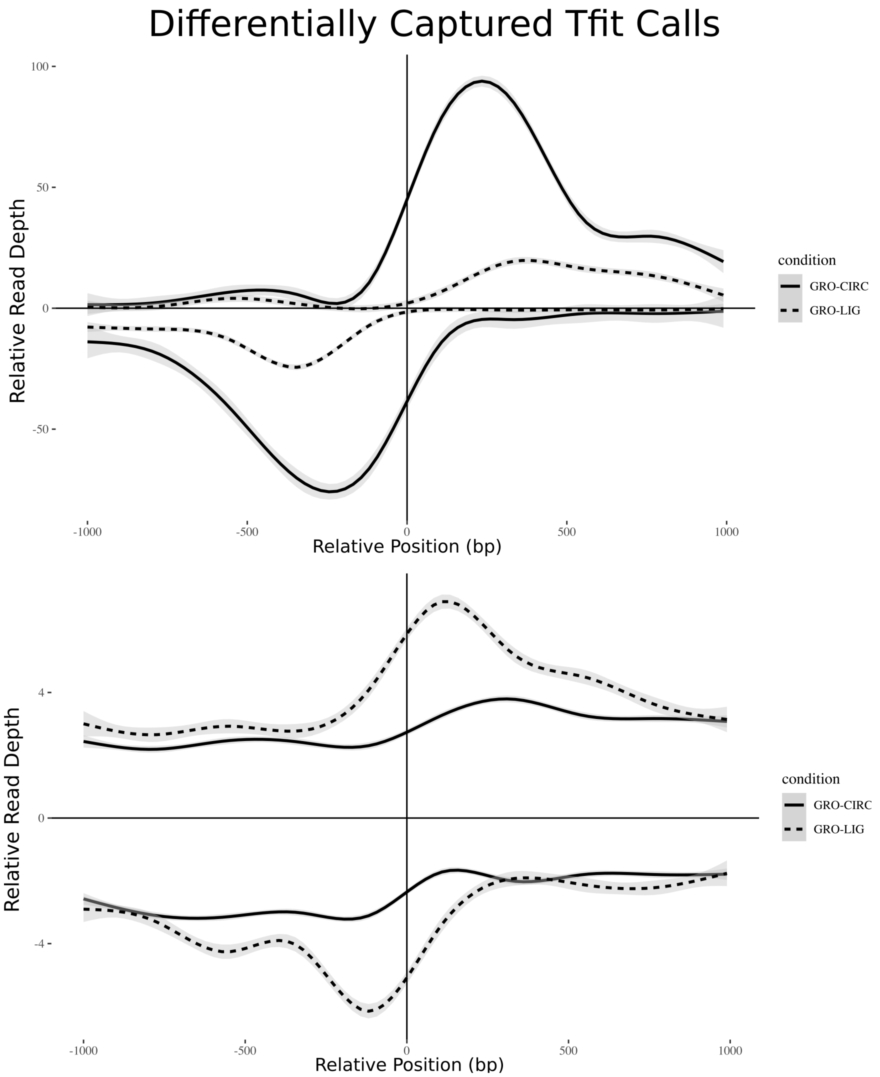
Supplementary Figure 17: **Scatterplot matrix of FANTOM regions.** FANTOM annotations [34] are generated from CAGE data, thus we reasoned that FANTOM annotated regions would be highly transcribed enhancers. Correlation levels are high between all protocols at these regions, albeit with considerable variation near select sites.



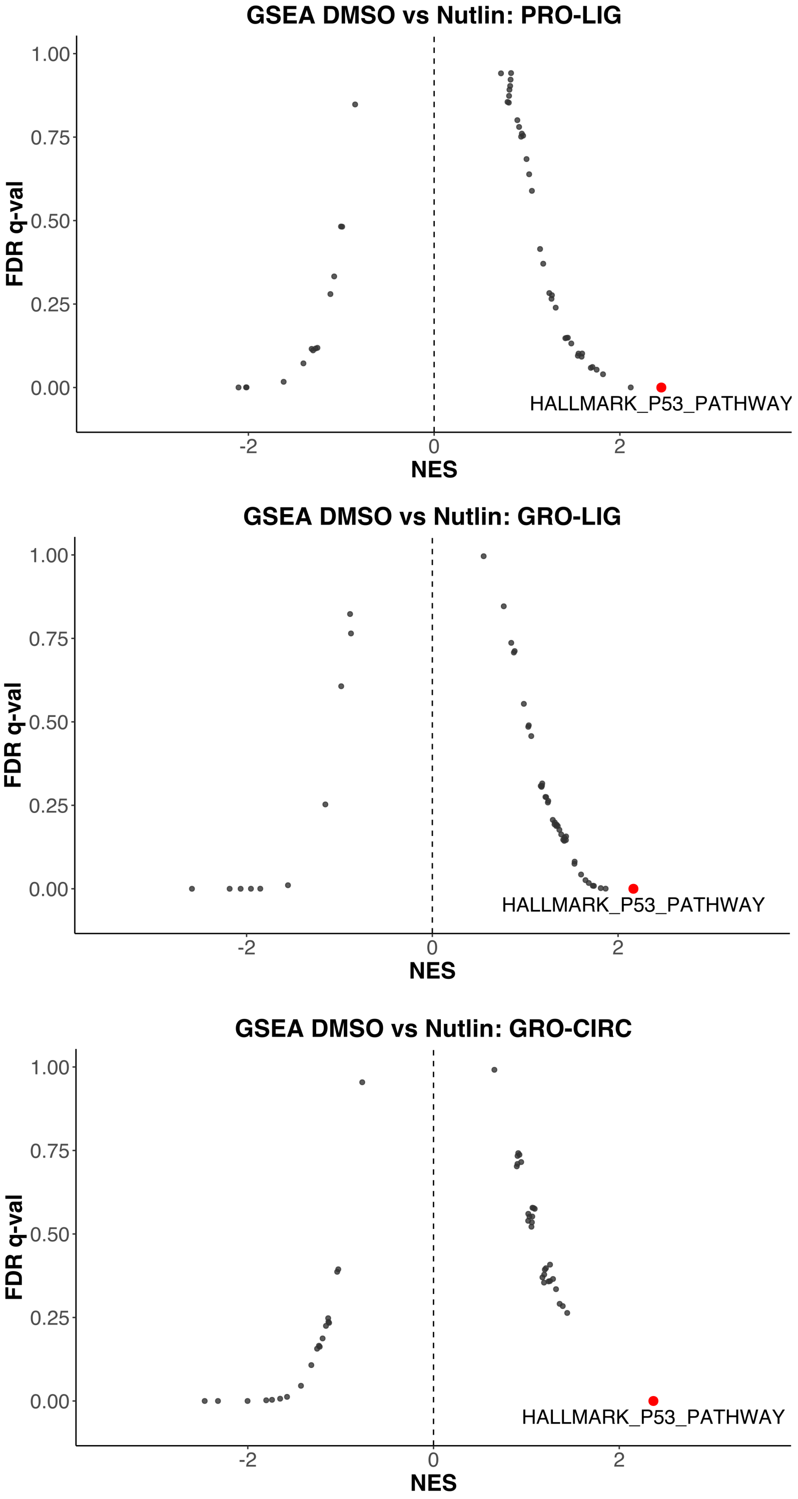
Supplementary Figure 18: **UpSet of Tfit/dREG calls among PRO-LIG, GRO-LIG, and GRO-CIRC libraries.** Bidirectional calls for equal numbers of DMSO-treated biological replicates were combined to form each set (PRO-LIG: n=2, combined depth 83.3 million reads (SRR14355652, SRR14355672); GRO-LIG: n=2, combined depth 108 million reads (SRR14355673, SRR14355674); GRO-CIRC: n=2, combined depth 212 million reads (SRR1105736, SRR1105737) (see Supplemental Table 1)). We observe frequent instances where each method does not call a region, despite the presence of bidirectional transcription, as shown in Fig. 2.4D,E. While this effect is depth dependent, there are notable regions where the strength of signal is strongly protocol dependent even after correcting for disparities in depth (Fig. 2.4G,H).



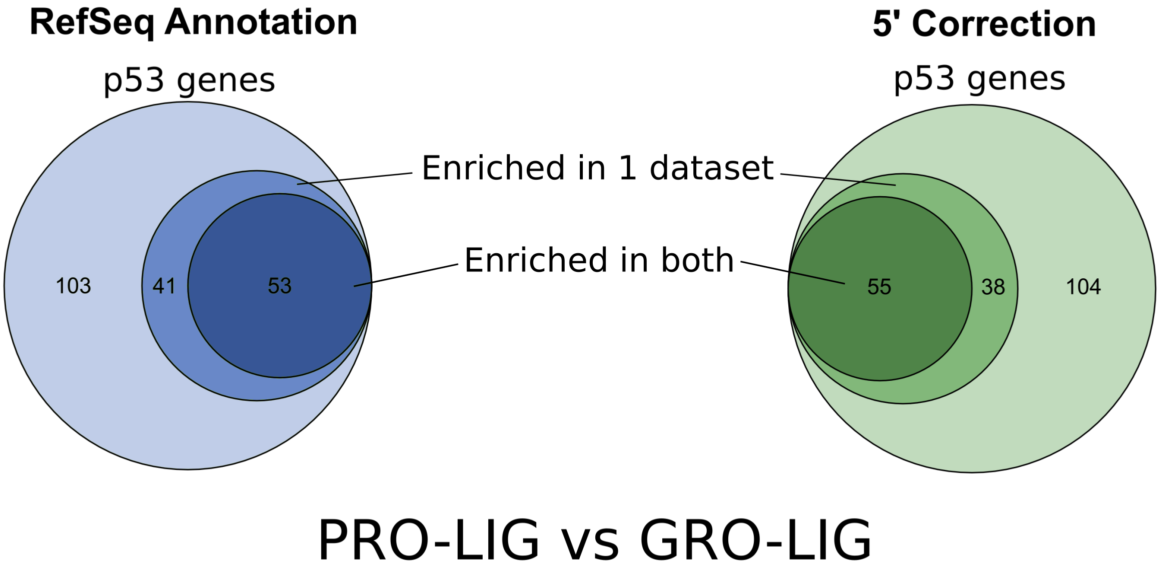
Supplementary Figure 19: **Example region indicating differences in enhancer transcription between protocols.** Read depths were normalized by CPM. Biological and technical replicates were combined to increase effective depth, as indicated in the bottom two read tracks (PRO-LIG: SRR14355650, SRR14355651 SRR14355652, SRR14355672; GRO-CIRC: SRR1105736, SRR1105737, SRR828696, see Supplemental Table 1).



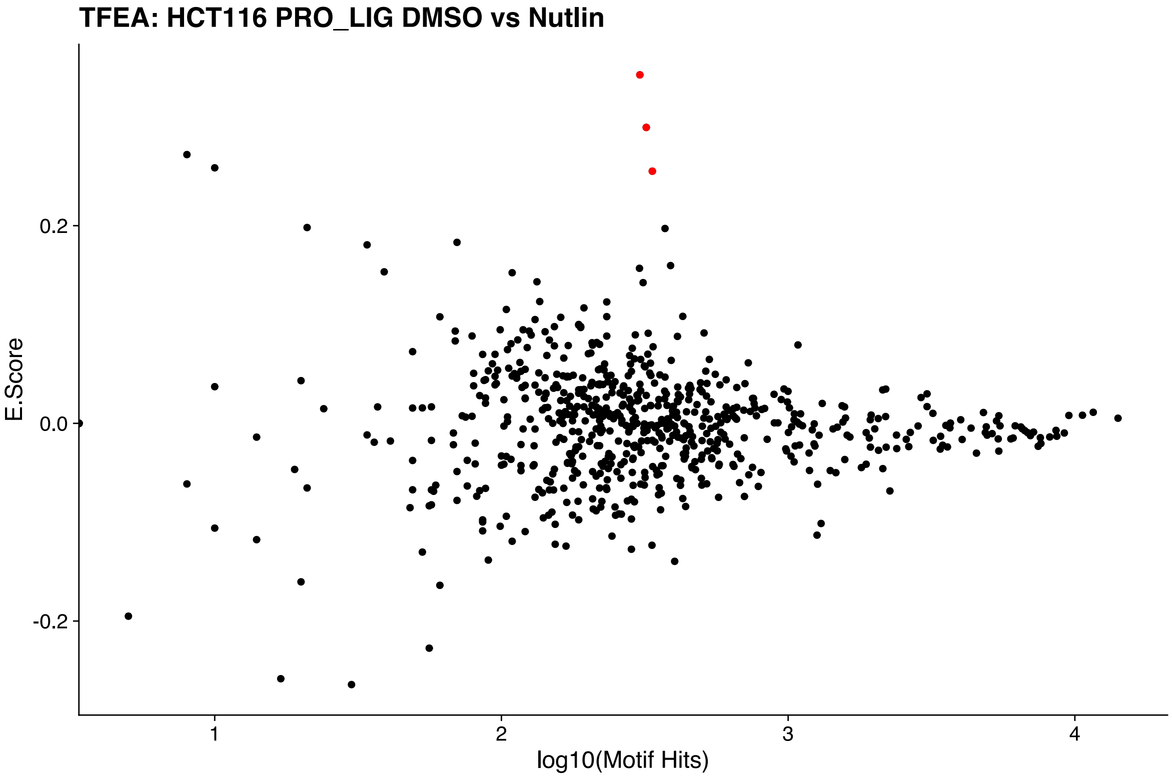
Supplementary Figure 20: **Metagene of enhancers differentially captured in either GRO-LIG or GRO-CIRC libraries.** Tfit calls across all replicates and treatments were combined together using **muMerge** for both GRO-LIG and GRO-CIRC libraries. Combined enhancers for GRO-LIG were then merged with combined enhancers for GRO-CIRC using bedtools merge (v2.28.0). Counts over these regions were used as input for DESeq1 (See also Materials and Methods). Differentially transcribed enhancers (Fig 2.4F, Materials and Methods) were used as inputs for metagene construction of GRO-CIRC (Top) and GRO-LIG (Bottom) preferentially obtained regions. Reads counts were normalized by CPM.



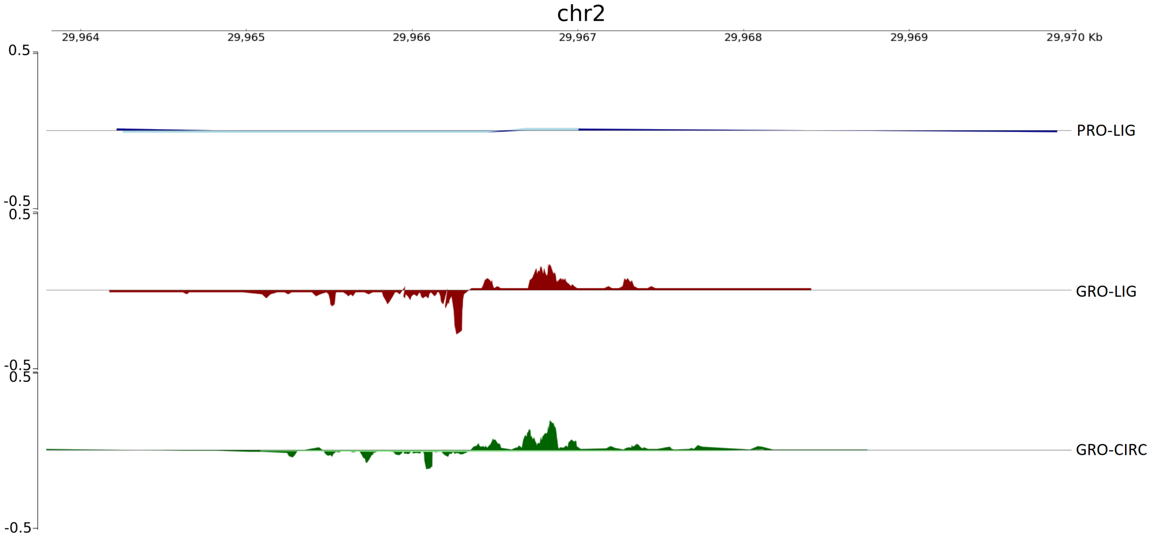
Supplementary Figure 21: **Enrichment plot of GSEA results for GRO-LIG, PRO-LIG, and GRO-CIRC libraries.** Gene region definitions were adjusted to exclude the 5′ pause peak, as per Fig 2.5A. In spite of library variations, the HALLMARK_P53_PATHWAY (red) is the strongest hit in all comparisons.



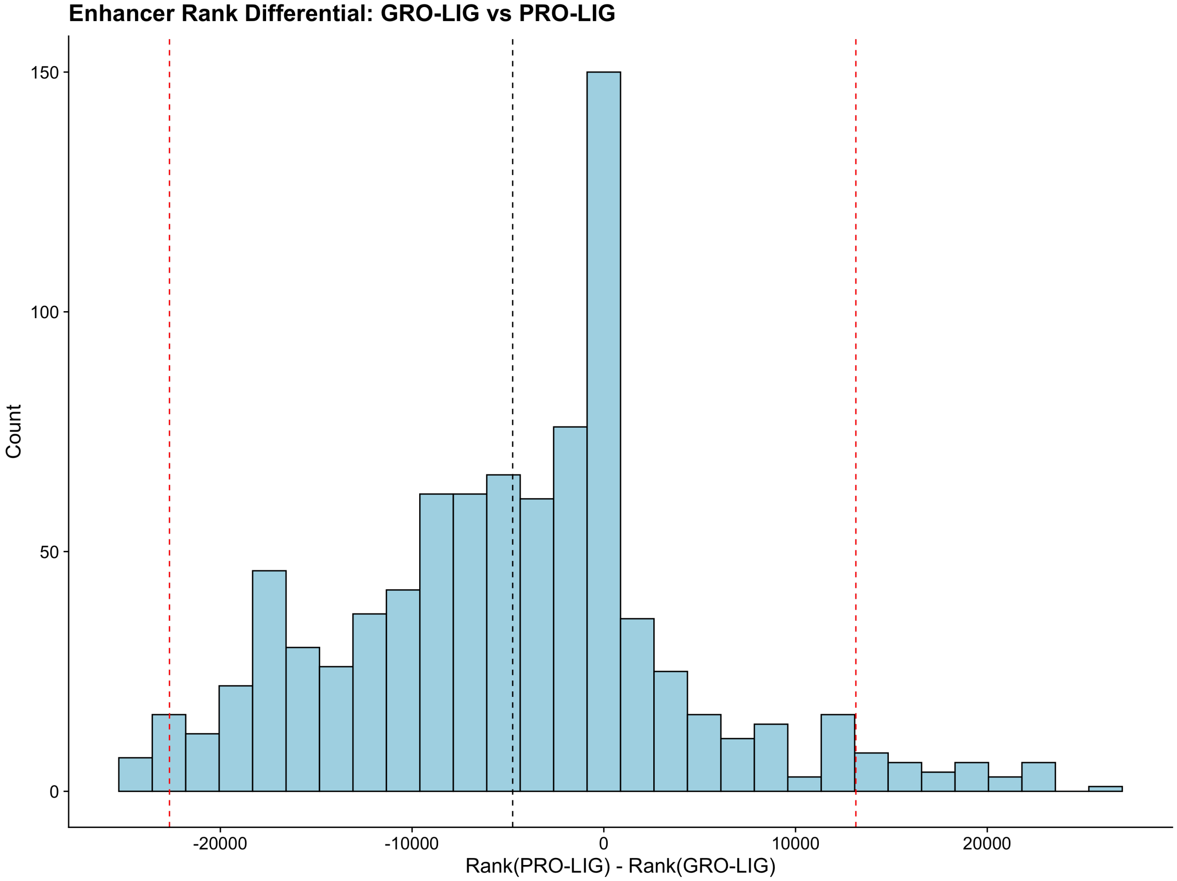
Supplementary Figure 22: **Overlap of GSEA p53 genes in GRO-LIG and PRO-LIG libraries.** Analysis was performed using counts over gene bodies (Left, hypergeometric test p-value=5.54e-15), and using a 5′ correction (Right, hypergeometric test p-value= 8.87e-17), as in Fig. 2.5A (see also Materials and Methods).



Supplementary Figure 23: **TFEA results for PRO-LIG libraries.** Regions were combined using **muMerge**, as in Fig. 2.5E,F. Red dots indicate transcription factors belonging to the p53 family (TP53, TP63, TP73).

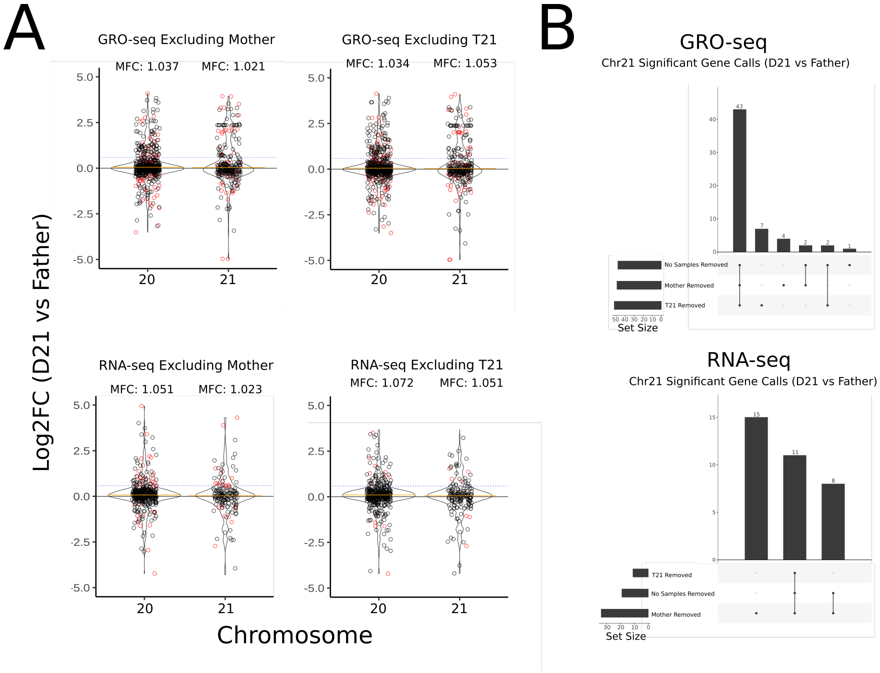


Supplementary Figure 24: **Example enhancer region where libraries disparately capture differential p53 enhancer activity.** Darker colors represent transcription level in Nutlin-3a treated libraries, while lighter colors represent levels found in DMSO-treated libraries. (Notably DMSO levels are nearly zero.) Read counts are normalized by CPM.

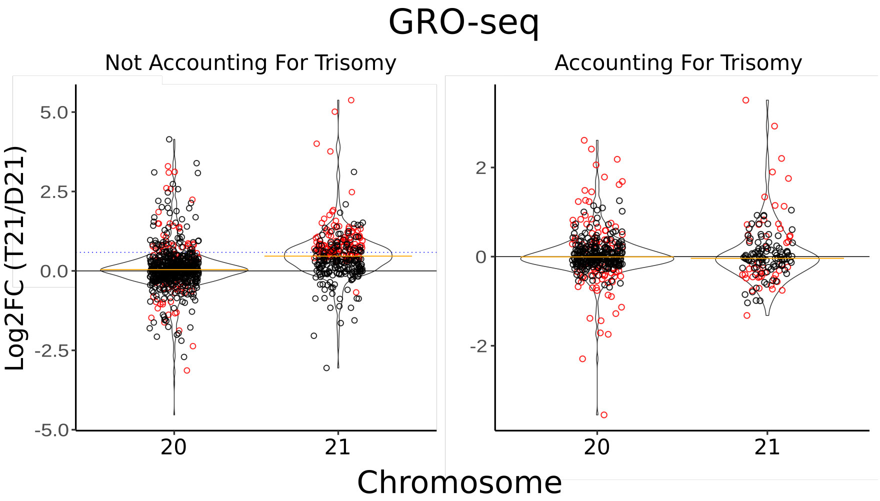


Supplementary Figure 25: **Rank differential of GRO-LIG and PRO-LIG enhancers.** Ranks were determined within TFEA through DESeq2. p53 enhancers which were more than 2 standard deviations (red dotted lines) from the mean (black dotted line) were considered to be differentially captured in GRO-seq or PRO-seq.

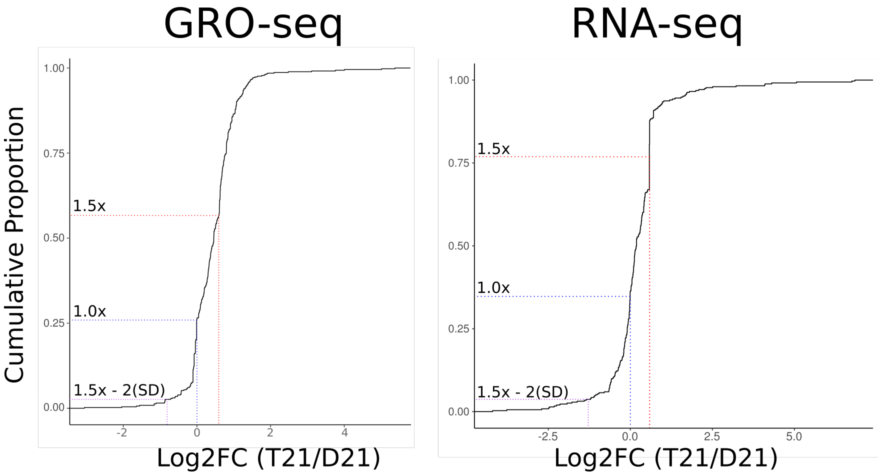
Appendix B: Chapter 3 Supplementary Material



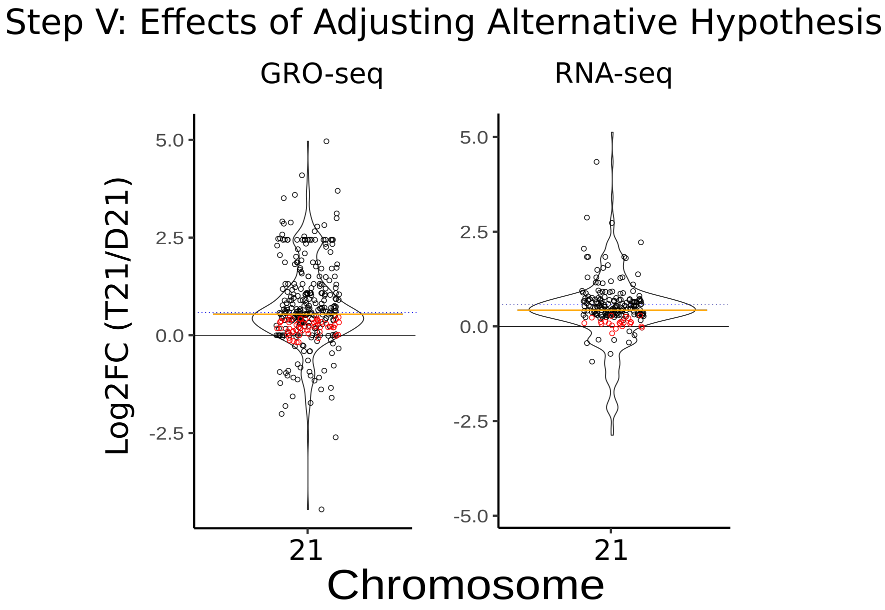
Supplementary Figure 26: **Comparisons of disomic individuals.** (A) Violin plots indicating fold change estimations between two D21 datasets, excluding either T21 or the mother samples from the pipeline. Excluding samples can cause differences in significance calls genome-wide, especially with T21 samples. Yellow line: Median Fold Change. Blue line: 1.5x Fold change. Red: Significant gene calls (padj<.01) (B) UpSet plots indicating overlap of significant gene calls on chromosome 21 (padj<.01).



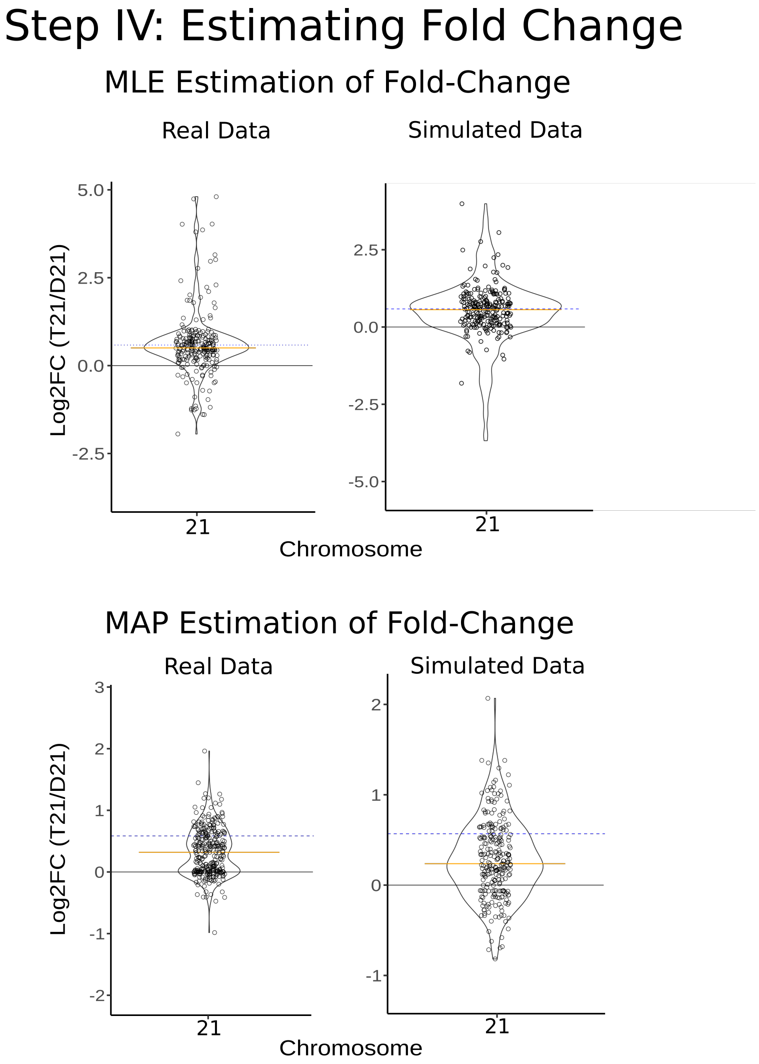
Supplementary Figure 27: **Violin plots depicting fold change estimates for chromosome 20 and 21 genes between T21 and D21 GRO-seq datasets.** (Left) Default analysis not accounting for the ploidy of the samples. (Right) Adjusted analysis correcting for ploidy differences between the samples.



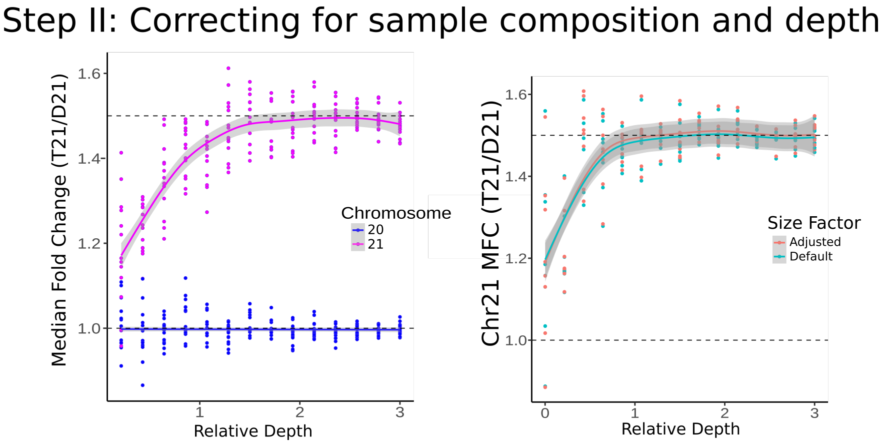
Supplementary Figure 28: **Cumulative distribution plot of fold changes of chromosome 21 genes.** Dosage compensated genes are often identified using a cutoff, indicated by each dotted line (1.5x fold-change, 1,0x fold-change, or 1.5x - 2 standard deviations); however, the reduced fold change estimations of these genes can also be explained by biological or technical variance. Left: GRO-seq. Right: RNA-seq



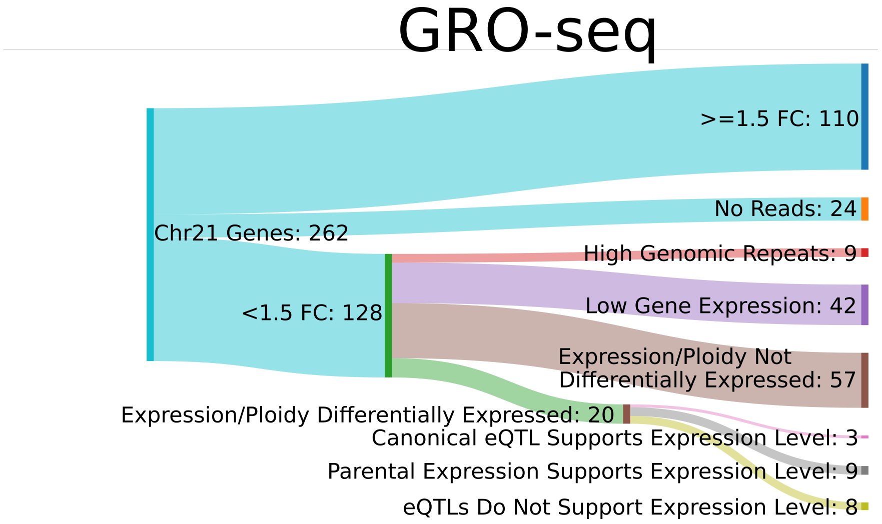
Supplementary Figure 29: **Differential Analysis with adjusted alternative hypothesis (|LFC| < log2(1.5)).** Significant genes are those which are significantly below the expected value of 1.5 (dotted line). Left: GRO-seq, 56 significant genes. Right: RNA-seq, 20 significant genes



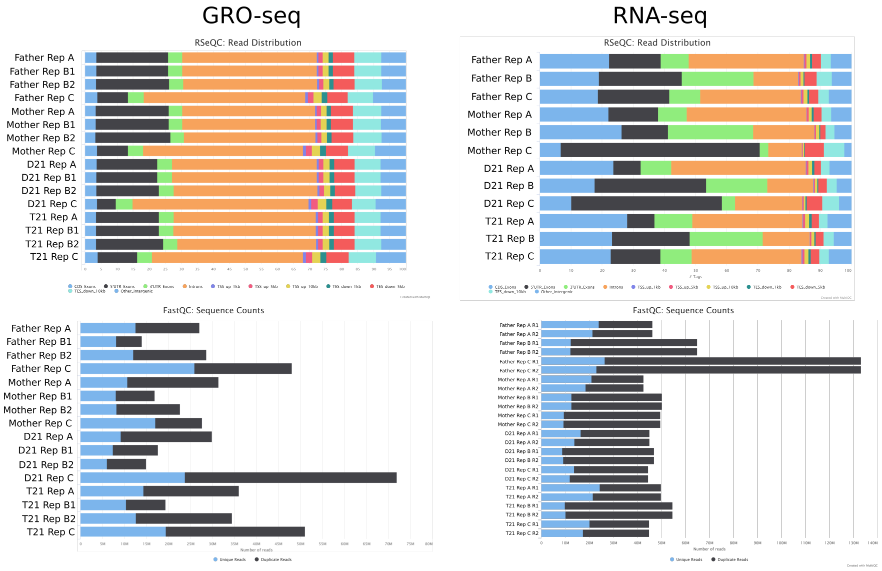
Supplementary Figure 30: **Violin plots of fold change of chromosome 21 genes in real RNA-seq and simulated datasets, using maximum a posteriori estimates of fold change.** Due to fold change shrinkage, median fold change estimates are further pushed towards 0 (RNA-seq MFC: 1.24, Simulateed Data MFC: 1.21)



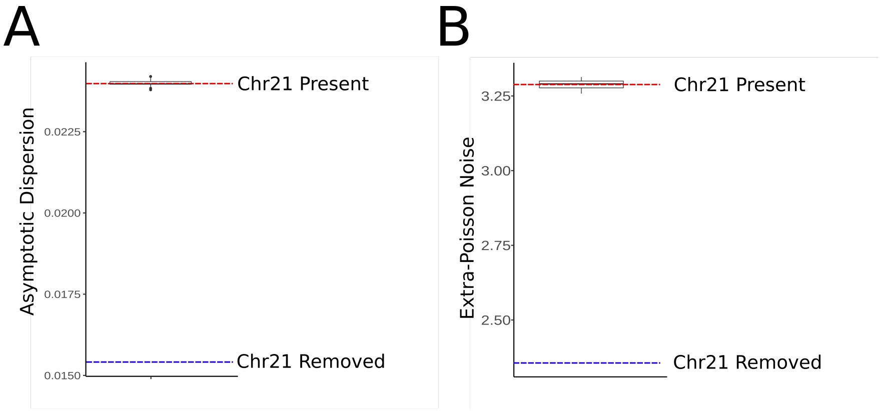
Supplementary Figure 31: **Fold change estimations in simulated datasets using two different size factor calculations.** Default: chromosome 21 genes are included in size factor calculation. Adjusted: chromosome 21 genes are excluded from size factor calculation. In general, the differences in fold change estimates between these two estimates is minimal.


GRO-seq
>=1.5 FC: 110
No Reads: 24
Chr21 Genes: 262
High Genomic Repeats: 9
<1.5 FC: 128
Low Gene Expression: 42
Expression/Ploidy Not
Differentially Expressed: 57
Expression/Ploidy Differentially Expressed: 20
Canonical eQTL Supports Expression Level: 3
Parental Expression Supports Expression Level: 9
eQTLs Do Not Support Expression Level: 8

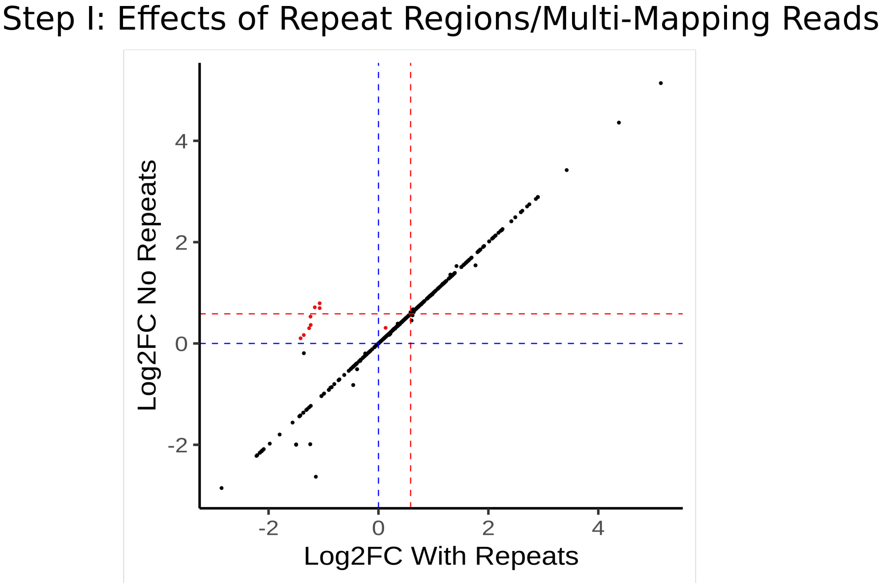
Supplementary Figure 32: **Sankey Diagram indicating explanations of fold changes for chromosome 21 genes in GRO-seq (See also Fig 3.3C).** Nearly all genes with expression lower than 1.5 can be explained using technical factors, leaving only 8 genes whose expression is below expectation.



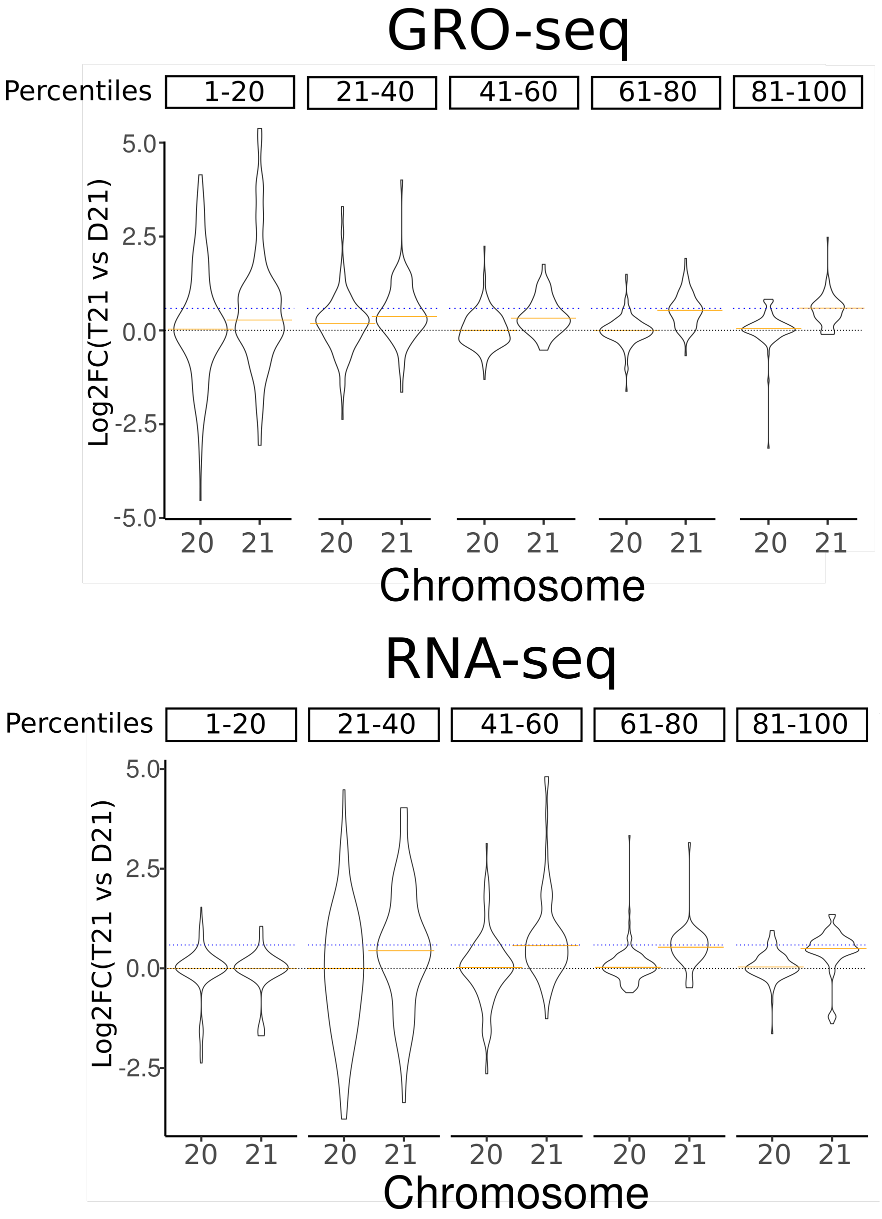
Supplementary Figure 33: **QC metrics of RNA-seq and GRO-seq datasets. (Top) RSeQC read distribution plots of all datasets, showing relative abundance of reads at each listed genomic feature.** (Bottom) FastQC plots showing number of total reads, and proportion of duplicate reads for each dataset



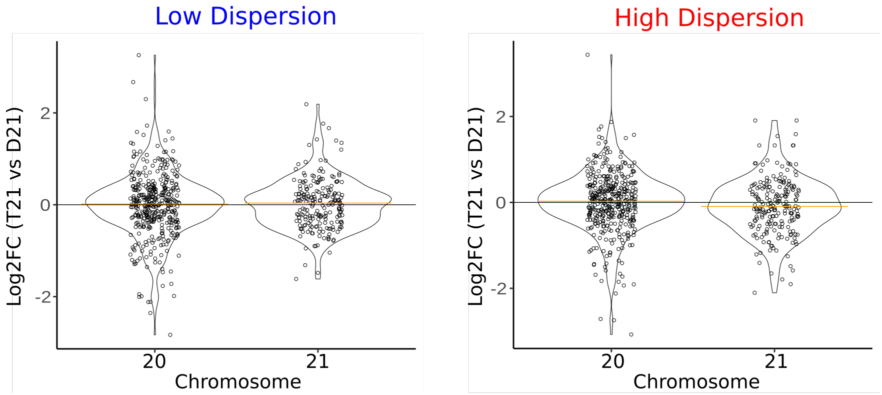
Supplementary Figure 34: **Dispersion fitting estimates calculated by DESeq2 including chromosome 21 genes (red) and excluding chromosome 21 genes (blue).** Additionally, an equivalent number of random genes were excluded as a control (boxplot). (Left) asymptotic dispersion estimates. (Right) extra-Poisson noise estimates. See Materials and Methods for information about how these values are estimated.



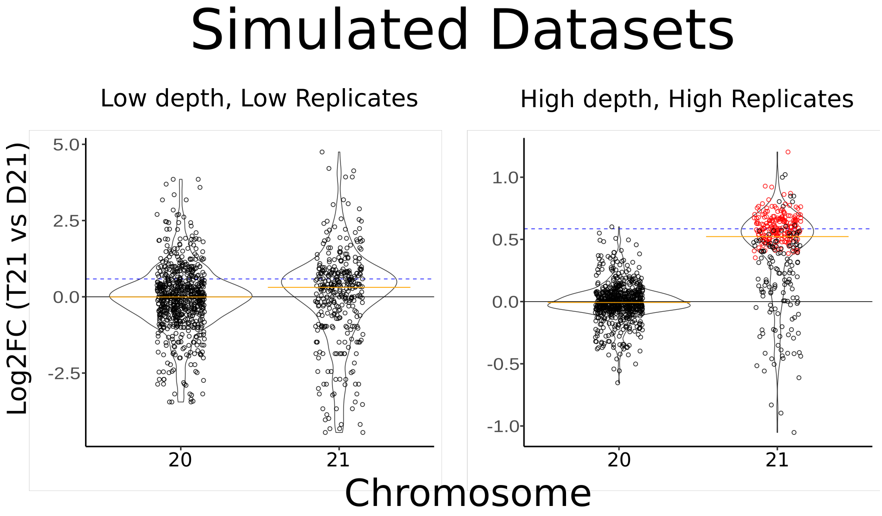
Supplementary Figure 35: **Effects of genomic repeats and multi-mapping reads on fold change estimates (T21 vs D21, RNA-seq).** Read counts were generated including genomic repeats and multi-mapping reads (x-axis), and excluding these reads (y-axis). Nine genes show a reduced fold change estimation when these reads are included (highlighted in red). Red dotted lines indicate 1.5x fold change. Blue dotted lines indicate 1.0 fold change



Supplementary Figure 36: **Fold change estimations (T21 vs D21) across expression levels in GRO-seq (top) and RNA-seq (bottom).** Genes are grouped by expression quantiles. Lower quantiles show lower fold change estimates for chromosome 21 genes.

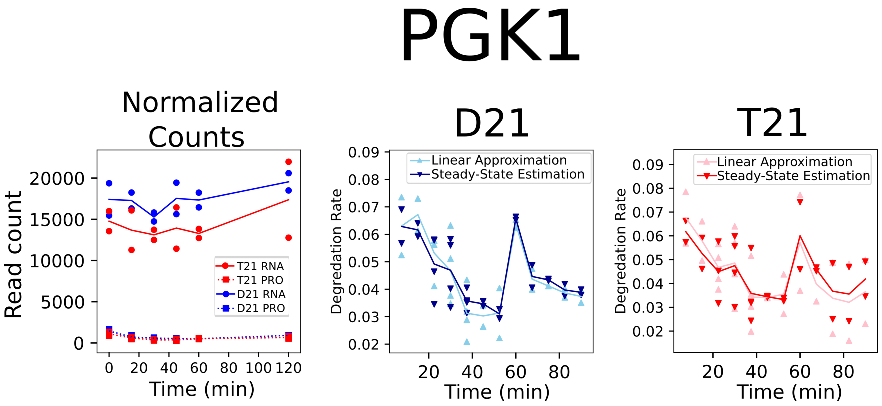


Supplementary Figure 37: **Fold-change estimates of simulated T21 and D21 datasets, using low and high dispersion estimates.** Fold-change estimates were adjusted to account for trisomy, as in Fig 3.2D. Low dispersion: a=.01, b=1. High dispersion: a=.05, b=30.



Supplementary Figure 38: **Fold-change distributions of simulated T21 and D21 datasets, high dispersion estimates (asymptotic dispersion=.08, extra-Poisson noise=8), and varying the depth and replication number of the samples.** Depth was changed relative to our D21 RNA-seq datasets (see also Supplemental Fig 33). Left: low depth (1x) and low replication (n=3). Right: high depth (3x) and high replication (n=12).

Appendix C: Chapter 4 Supplementary Material



Supplementary Figure 39: **Degradation calculations for PGK1, a gene which undergoes little change in RNA decay rates across time.** Both SSM and LA calculations are consistent for every time interval.