Abstract

A Comparative Approach to Studying RNA Biochemistry using TimeLapse Chemistry and Labeling in Cell Culture (TILAC)

Meaghan Sullivan

2021

RNA sequencing is a sensitive, transcriptome-wide analysis of RNA levels, valuable for assessing changes in gene expression, or measuring RNA enrichment after immunoprecipitation or cellular fractionation. However, accurate quantification of reads in RNA sequencing datasets is challenging due to variance from variable handling during RNA purification, and biases introduced in downstream processing, like shearing and amplification. To address these challenges, I developed TimeLapse Chemistry Labeling in Cell Culture (TILAC), an internally controlled and normalized approach to compare RNA content between samples using RNA metabolic labeling that is analogous to the SILAC method from protein biochemistry.

TILAC utilizes the metabolic labels 4-thiouridine (s^4U) and 6-thioguanisine (s^6G) to differentially label RNA populations, allowing the samples to be pooled at the beginning of the experiment, prior to any of the handling steps that can introduce noise. TimeLapse chemistry recodes s^4U into a C analogue, and s^6G into an A analogue, thereby inducing mutations in sequencing reads. The ratios of RNAs from the two samples can be determined using the mutational content of the sequencing library.

I have used TILAC to study perturbations to both transcription and translation of RNA. First, I show that TILAC can capture the global downregulation that occurs when

cells are treated with the RNA polymerase II inhibitor, flavopiridol. Then, in the context of the *Drosophila* heat shock system, I show that TILAC captures both the high upregulation of heat shock responsive genes, as well as the more subtle downregulation across much of the transcriptome. I then turn my attention to translation, which can be studied by isolating polysomes by velocity sedimentation over a sucrose gradient. This method fractionates total cell lysate, and so polysome fractions contain contamination from non-ribosomal RNPs with similar coefficients of sedimentation. I use TILAC to characterize the background in sucrose sedimentation under normal conditions, and conditions of sodium arsenite stress. Armed with this knowledge, I used TILAC to study changes in translation, as well as in the total RNA pool in response to stress. Surprisingly, I found that there were few changes in the total RNA pool, but there was increased translation of a set of RNA helicases that are also found in stress granules. This puts forth an exciting hypothesis that the cell may need to upregulate these proteins to mitigate the potential damage caused by aggregating RNA in the cytoplasm. Through this set of experiments, I explore the advantages of using TILAC in RNA sequencing experiments and demonstrate how it can be used to advance biochemical studies, particularly in the context of global changes in RNA levels and in fractionations.

A Comparative Approach to Studying RNA Biochemistry using TimeLapse Chemistry and Labeling in Cell Culture (TILAC)

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Meaghan Sullivan

Dissertation Director: Matthew Simon
June 2021

Table of Contents

Index of Figures

Index of Tables

Acknowledgements

First and foremost, I need thank Dr. Matthew Simon for the great privilege of doing my PhD in his lab. Matt hired me as a technician when I left Caltech, and I continued on in his lab to work among some of the brightest and most inspiring scientists at Yale. I learned so much about chemical biology by training with Matt.

I want to thank my entire lab, past and present members, for the scientific conversations, and help troubleshooting – and also the extra watermelon slices at the West Campus BBQ's, the game nights, and the overwhelming love and support you've shown me during my PhD. I want to give a special thank you to Erin Duffy. Your mentorship and influence early on in graduate school helped set me on the course to success.

The Yale Molecular Biophysics and Biochemistry department has been very influential in my development. I have especially been fond of the Hall Seminars at the Medical School as a center for scientific and social engagement. I've gotten to watch so many graduate students develop their projects over the years, sitting at their talks every Friday morning. And importantly, people saw my projects develop and gave me a lot valuable insight.

In particular I want to thank my past and present thesis committee members – Drs. Yong Xiong, Wendy Gilbert, and Christian Schlieker for your constant scientific engagement and encouragement. As I wrote the outline for my thesis, I went back through my old committee meetings and realized that I've come quite a long way as a scientist. Thank you for your patience and commitment.

I would also like to thank my collaborator Rachel Niederer. Our collaboration was instrumental to me completing the parts of my project that were most exciting to me. I loved getting to dive into translation and stress. It was a pleasure to learn more about translation from you, and to get to know you better.

My parents have been a consistent source of support throughout my life. I am so thankful for their love, understanding, and for always being there for me. My brothers Alex and Matt have always been my best friends. Thank you for always being available to talk and make me laugh during my PhD.

My fiancé Edward, science is just more fun when you're around. I can't imagine getting through graduate school without you.

I am deeply grateful for the financial support of the MBB Department and the Cellular and Molecular Biology Training Grant.

A brief note on history

The world changed extremely rapidly over the last year of my PhD, during which most of this work was completed. In the early months of 2020, a novel coronavirus was emerging in Wuhan, China and President Donald Trump was being impeached. By March of 2020, the COVID19 pandemic had hit New York City. Businesses closed. Travel across state borders was restricted. Yale shut down all research operations. Life seemed to enter a suspended animation until May.

On May 25th, the murder of George Floyd in Minneapolis, MN by Derek Chauvin was filmed and put on the internet. It was another murder in a long history of a Black deaths at the hands of police. White people like me, sitting at home in relative safely during a pandemic, could no longer look away from the racism that destroys the lives of people living just a few blocks away. Protests continued for weeks. During the same week I returned to lab from the months-long shutdown, I also marched through New Haven with 5,000 other Connecticut citizens to protest police violence. The Yale MBB department participated in the Shutdown STEM Day, led by Wendy Gilbert.

The Democratic primary ended in the nomination of Joe Biden, and the historic choice of Kamala Harris as the first Black and India woman as a vice presidential nominee. A bitter battle over the US Post Office, mail in ballots, and election security ensued.

There was record turn-out for the 2020 Presidential election, including record use of mail-in ballots. I worked the polls as a tabulator, during which I stood by the door and made sure people properly submitted their ballots. It took until Saturday, Nov 7th, to call the election for Joe Biden and Kamala Harris.

On December 11th, the Pfizer COVID19 vaccine was approved, less than a year after development started. The Moderna vaccine was approved on December 18th. These are the first two mRNA vaccines to be approved by the FDA, and the culmination of decades of research by thousands of graduate students, post-docs, and faculty into the basic biochemistry of RNA.

President Donald Trump continued to claim election fraud. On January 6th, Trump supporters stormed the Capitol in an attempt to prevent the certification of the vote for President-elect Joe Biden and Vice President-elect Kamala Harris.

I submitted my last libraries to the YCGA on January 8th. These were 28 libraries to study how translation changes during and after stress.

Joe Biden and Kamala Harris were sworn into office on January 21st.

Today, Febrary 12th, as I finish my thesis, Donald Trump is in his second impeachment trial, related to the January 6th attack on the Capitol.

Over 475,000 people in the US have died from COVID19.

Chapter 1. Introduction

1.1. Central challenge in RNA sequencing

This work explores the challenges of quantifying reads in RNA sequencing. I approach this problem with an eye to the sources of noise that are inherent in every step, and with optimism that new technologies can greatly improve the accuracy with which all biologists and biochemists, not just RNA specialists, can describe their systems. Thus, I will introduce current biochemical and computational techniques in RNA sequencing, and compare them to what is available in protein biochemistry. Then I will discuss how nucleotide recoding technology developed in the Simon lab enabled me to develop an internally normalized method for performing RNA sequencing. The background given below justifies the need for a robust, comparative, internal normalization method for analyzing sequencing data, and presents several biological systems in which I applied the method. Subsequent chapters will present technical validation of the method, as well as its use in discovering a new set of transcripts that are preferentially translated during stress.

1.2. RNA Sequencing Technology

RNA sequencing has transformed RNA biology, and fields beyond, as a method to quickly profile both gene expression and differences in expression between conditions. It is a vital part of research programs from those studying the mechanisms of splicing, to the physician-researcher doing occasional sequencing experiments to understand a specific disease. With improving library preparation protocols and rapidly decreasing costs, it is a ubiquitous tool[1]. In RNA sequencing, RNA is isolated from cells and

converted into cDNA. The base sequence of the cDNA is read on a high-throughput sequencer. To date, sequencing has most often been done on an Illumina sequencer using short-read technologies, although long-read technologies like Pacific Biosciences sequencing and Oxford Nanopore sequencing are becoming increasingly common[2]. The data is then analyzed by aligning reads to a genome, quantifying transcripts covering genes, normalizing those reads, and calculating differential expression.

RNA-sequencing is often combined with other upstream biochemical steps in order to measure specific populations of RNA. Methods like formaldehyde RNP immunoprecipitation (fRIP) [3] or UV crosslinking and immunoprecipitation (CLIP) [4] are used to study the populations of RNAs that interact with a specific protein or protein complex. Fractionations that separate chromatin, nucleoplasm, and cytoplasm are used to ask about RNA processing status and localization over time[5,6]. Purification of polysomes, complexes of many ribosomes translating the same RNA, is used to study what is being actively translated[7,8]. These techniques are essential for RNA biochemistry and have provided an enormous wealth of invaluable biochemical knowledge. It is worth continuing the work to improve these methods to increase our capacity to study essential biochemical processes sensitively and accurately.

Technical Challenges in Performing RNA Sequencing

To perform short-read (Illumina) sequencing, the RNA population of interest must be enriched, then converted into a DNA library spanning about 150-500bp in size. Adapters are added that can preserve strandedness, serve as handles for sequencing platforms, and uniquely label individual libraries for pooling. Finally, libraries are

amplified to increase the amount of cDNA available for sequencing. I will go through each step to understand how it affects variance in RNA expression estimates.

The first source of noise is the selection of mRNAs to sequence. Many RNA sequencing experiments are designed to investigate how protein coding transcripts change. Since these RNA's make up only about 2-5% of the transcriptome[9], these transcripts have to be enriched to avoid wasting reads on ribosomal RNA. One solution to the problem is polyA+ selection. Since protein-coding genes are polyadenylated, they can be enriched from heterogenous RNA preparations using oligo-dT probes attached to beads that can be pulled down, either through magnetic separation or centrifugation[10]. As with any pulldown, the key to this step is to have the same efficiency in pulldown between comparable libraries. Another method is oligo-dT primed reverse transcription. By initiating reverse transcription with an oligo-dT primer, poly-A+ mRNA is selectively converted into DNA. However, this enriches for 3' fragments over the rest of the transcript length. These primers can also prime at A-rich motifs present in some RNAs, causing artifactual amplification of some transcripts over others [9,11,12].

By enriching for polyA+ RNA, other non-coding RNAs that might be of interest, for example snRNAs or lncRNAs, are also depleted. Therefore, ribosomal depletion may be a preferred method to enrich for interesting RNAs. Many of these strategies are best done using kits, and most of these methods are variations on the basic concept that probes complementary to rRNA are added and allowed to hybridize. Sometimes these probes are biotinylated, and the rRNA can be removed by pull-down. Other strategies use probe-directed degradation, in which enzymes such as RNaseH degrade the nucleic acid hybrid[10,13].

For Illumina sequencing, RNA needs to be fragmented into short pieces less than 500nt in length. This helps decrease 5' bias[14]. Placing RNA into alkaline solution with divalent cations like Mg^{2+} or Zn^{2+} will encourage internal transesterification.[15] This process is structure dependent, and is often done at high temperatures to decrease secondary structure[14]. Degradation using RNase III will also have some structure bias, since it preferentially cleaves dsRNA[9]. The extent of shearing has additive effects downstream. The size and sequence context of these short pieces affects ligation and amplification of cDNA.

Right after shearing, RNA needs to be reverse transcribed into cDNA, often using random hexamer priming[14,16]. Adaptors are added either during the reverse transcription step, using template-switching oligos, or through direct ligation in a separate step[17]. The specific challenge in adding adaptors is keeping track of the specific strandedness of the cDNA fragment being sequenced. In the most simple solution, different adaptors are added onto the 5' or 3' ends using RNA ligase I or II[18]. However, the sequence context at the 5' and 3' ends can have large effects on the efficiency of the ligation[9].

Properly sized cDNA fragments are enriched through bead purification and PCR amplified to generate enough material for sequencing. Overamplification creates PCR duplicates in the biochemical sample, decreasing library complexity. It also selectively amplifies smaller over larger library fragments[19].

Each of the above steps can introduce both bias and noise into sequencing libraries. This is problematic when it comes time to make comparisons between two different libraries in order to assess differential gene expression. If RNAs are fragmented to different sizes in the first step, different portions of the RNA pool will be size selected

or selectively amplified by PCR. RNAs fragmented to different sizes may end up with enrichment of different 5' and 3' ends, have different adaptor ligation efficiencies, and end up with slightly different library composition. In additional, smaller-sized RNA or cDNA pools will have less effective ribosomal depletion. These all affect the composition of the library that is ultimately put on the sequencer.

Technical challenges in performing upstream RNA biochemistry experiments

In addition to the technical biases inherent in library preparation, biochemical steps upstream of sequencing carry their own unique challenges. RNA immunoprecipitation is used to study how a specific protein binds to RNA. Proteins are pulled down with either an antibody, or a protein tag. These pulldowns, which take place within a crowded cell lysate, suffer from the need to both wash away background interactions and stabilize true interactions, even while true interactions can be of varied strength[20-22]. Crosslinkers are a common and often necessary way to stabilize interactions, but also keep nonspecific background associated with the complex[3,4,23-25]. Higher stringency washes can wash away background, but could also eliminate true low-affinity interactions. Fractionations deal with similar background issues. Incomplete lysis of the plasma membrane, or premature lysis of the nuclear membrane, can contribute to contamination of fractions, and chromatin fractionations can suffer from polyA+ contamination, especially without extra depletion steps [26,27]. When samples are processed in parallel for comparison after sequencing, each sample goes through its own handling steps, and accumulates different types of the bias described above.

1.3. Common normalization strategies in RNA sequencing

Most often RNA sequencing experiments are asking about the difference in gene expression between two conditions, e.g. healthy and diseased. The actionable information a researcher wants is a list of genes, ordered by expression change, and assessed by multiple-test adjustment as to whether the difference between conditions is significantly different than the null-hypothesis that there is no change. For this analysis, expression levels need to be accurately inferred. Technical variation could be dealt with by increasing the number of replicates in order to amplify real biological signal over disperse noise, but this is rarely done in RNA sequencing due to the expense associated with library preparation and sequencing, and conclusions are routinely drawn from as few as 2-3 replicates. Additionally, RNA sequencing produces uneven read coverage, and genes with low expression often have highly variable read counts across samples, something which can be summarized as the variance being inversely dependent on the mean read count [28].

A variety of normalization approaches have been developed to control for both intra-sample bias and inter-sample bias. Intra-sample normalization methods control for gene-specific effects, such as gene length and GC content. Inter-sample normalization methods have to control for differences in sample preparation, as discussed above, including differences in sequencing depth. The simplest methods of normalization adjust for sequencing depth by dividing read counts over any one gene by the library size[14]. This can skew results if one sample's total read count is very sensitive to a few high-expressed genes. An extra layer of normalization can be added, by dividing the read count also by the gene length. However, due to non-uniform coverage over a gene, this metric is often

insufficient[29,30]. More sophisticated between-sample normalization methods are upper-quartile and full-quantile, but these methods cannot account for overdispersion, an issue discussed in detail below[29].

Statistical normalization of RNA sequencing reads

With low numbers of sequencing replicates, uneven read coverage, and a dependence of the variance on the mean read count, RNA sequencing data can be challenging to quantify, and error modeling improves inferences about gene expression by modeling both biological and technical variations[31]. Accounting for technical variability allows researchers to use a statistical test to ask if there are true differences between the populations[32]. We can assume that sequencing reads are selected randomly out of the pool of possible cDNAs in the library, so ideally the abundance follows a Poisson distribution. In reality, at very low read counts (i.e. low gene expression), the variance is very high since the addition of a couple extra reads could double the amount of observed expression, even though that difference is likely due to random sampling effects of RNA sequencing. The Poisson distribution has only 1 parameter that describes both the mean and variance, and it is not able to capture the inverse correlation between mean read count and variance. The problem of having the variance of the count of reads be large at a small read count is called over dispersion, in that the variance is beyond what can be captured by a Poisson distribution[28,32].

Computational methods to model over-dispersion use different statistical models that rely on the negative binomial distribution, which has parameters for both the mean count and dispersion, and so can more accurately estimate expression and expression

changes[31]. Incremental improvements to this method involve developing algorithms that more accurately estimate parameters of the negative binomial distribution. Importantly, when these methods normalize for library size, they assume that most transcripts do not change expression, and the library sizes should be the same between samples[28,32]. It therefore forces the two libraries to have the same number of reads, while allowing only a subset of genes to experience large changes in expression. This assumption is not always true, and some transcription factors increase RNA production overall[33,34], while other perturbations decrease overall transcription[35]. These types of global changes are difficult to capture using these sorts of statistical programs.

Spike-in normalization of RNA sequencing reads

An alternative biochemical approach is to use RNA spike-ins. Spike-in RNA is exogenous RNA, either from a different species or of a synthetic nature, that is added in the same amount to all samples. They go through all the same handling and library prep as extracted RNA from the biological samples. Sequencing results are normalized based on knowing that there should be the same spike-in each sample.

RNA from a different species is often used as a spike-in because it is easy get from another model organism in the researcher's own lab or a lab on the same floor. A different species' genome will have different gene-length and sequence composition, and therefore this RNA can still experience biases different than the experimental RNA. Some proposed solutions have included using chimp RNA, since it is most similar to human RNA[36], or to use RNAs synthetically generated to have similar GC content as humans but to not align to the genome[37]. A set of standards has been proposed and

endorsed by the National Institute of Standards and Technology. This set consists of 92 polyA+ RNAs designed to span the full range of length and GC content present in genomes[38,39]. While useful under some conditions, if used carefully, spike-in RNAs do still experience a variety of technical variation that can make them unreliable. Reads counts of spike-ins can be variable compared to their initial concentration in the spike-in pools and their presence or absence in the final sequencing library can be highly depended on steps like polyA+ isolation efficiency[40]. The use of spike-ins for normalization is a complementary method to statistical methods like DESeq2, but there is a need for a method that is robust under a broader variety of experimental situations.

Use of SILAC in advanced protein biochemistry

SILAC[41,42] is a popular approach to quantitative biochemistry that uses internal normalization. It uses incorporation of either light or heavy isotope amino acids into distinct cellular populations. The cells are combined and go through mass spec as one mixed sample. This decreases any variability that may result from slight variations in handling that occur when processing samples separately. By decreasing technical noise, it becomes easier to identify important biological effects.

To compare two different cellular states through internal normalization, it helps to have as much of each population labeled as possible, and to have even amounts of label between samples. For SILAC, cells are grown for at least 5 doubling times in SILAC media[42]. This is enough time for the proteome to turn over, meaning there will be about 97% incorporation of the heavy isotope amino acid, of which the mostly commonly used are lysine and arginine. Because peptides are digested into smaller peptides by trypsin

cleaving at lysine and arginine residues, using these labels ensure that almost every peptide fragment will contain a label[43]. Data is analyzed with a program called MaxQuant, that identifies pairs of peptides that differ in mass by integers of the potential labels, and quantifies the ratio between them[44,45].

In its simplest form, SILAC makes quantitative comparisons of protein levels in two different cells types. More often, it is combined with many other biochemical experiments. For example, it can be used to identify post-translational modifications[46], in protein immunoprecipitations to distinguish real from background interactions[47], or used with additional labels to make measurements at multiple timepoints[48]. SILAC has even been extend to challenging samples, like primary neurons, where complete labeling isn't possible. In these cases, SILAC can be performed but requires two separate and distinguishable metabolic labels[49].

1.4. Nucleotide recoding technologies

Metabolic labeling

Metabolic labeling by incorporating non-canonical RNA bases into newly transcribed RNA enables separation of new and old RNAs from the total pool of cellular RNAs. The metabolic label can serve as a handle to immunoprecipitate new RNA away from preexisting RNA{Duffy:2018gm, Schofield:2018ff, Schwalb:2016bc}[50]. These pulldown methods suffer from high background and challenging normalization, and often low recovery of labeled RNA due to limited biotinylation efficiency[51]. Recently developed nucleotide recoding technologies allow newly made RNA to be distinguished

from old RNA without the use of biochemical enrichment, by making the label visible to the sequencer.

TimeLapse-seq

TimeLapse-seq is a nucleotide recoding method developed in the Simon Lab, with the intention of gaining kinetic information about RNA synthesis and decay[52,53]. It uses oxidative-nucleophilic-aromatic substitution to convert the metabolic label 4-thiouridine (s^4U) into a cytidine analogue and 6-thioguanine (s^6G) into a guanine analogue. This induces T-to-C or G-to-A mutations, respectively, in the resulting sequencing reads. It does this by replacing the thiol group with an amine group, thus changing the hydrogen bonding pattern of the nucleobase (Figure 1: TimeLapse Chemistry Recoding of s4U and s6G). The presence or absence of mutations gives an indication of whether that read was made prior to, or after, the metabolic feed.

A key advantage to TimeLapse-seq, compared to pull-down techniques discussed above, is that it is "enrichment-free". RNAs are sequenced within their natural context of cellular RNAs. Kinetic information is taken from the same pool of RNAs as whole RNA sequencing, and is therefore internally normalized. There is no need to attempt to normalize enrichment to input to estimate concentration or contamination of labeled RNAs.

The ability to use two orthogonal labels that are converted in the same reaction conditions is a powerful tool to apply to many different types of experiments. Two metabolic labels enables complex kinetic measurements[53]. Most importantly, it provides

two independent and distinguishable metabolic labels to uniquely label two different RNA populations for internal normalization methods.

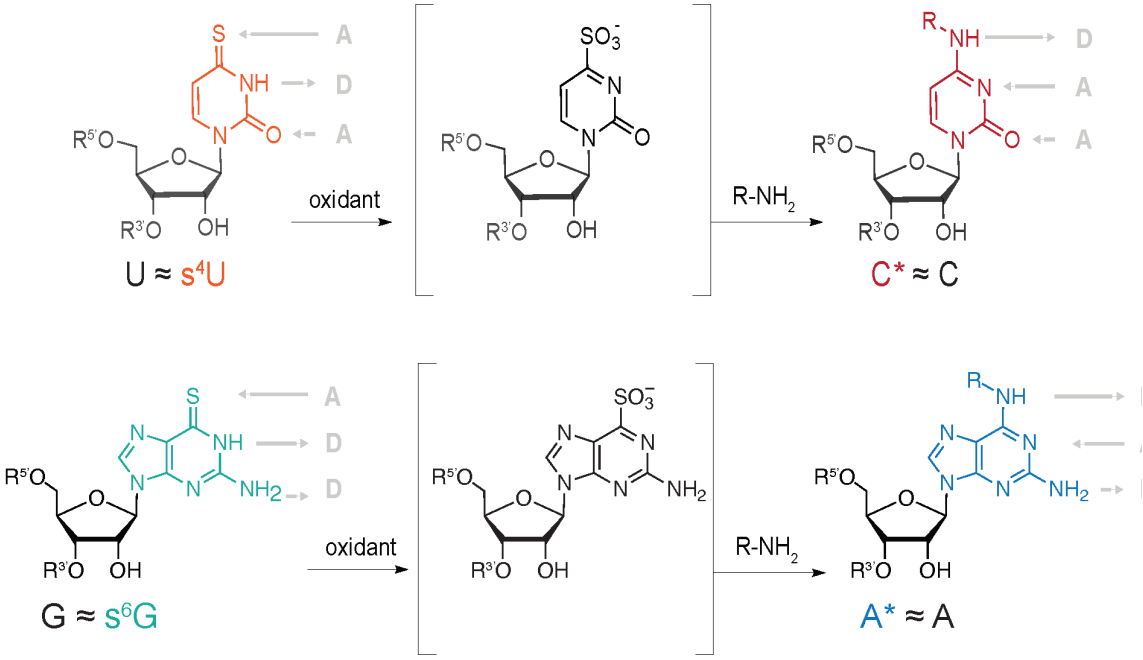

Figure 1: TimeLapse Chemistry Recoding of s4U and s6G

Top: TimeLapse chemistry recodes s^4U to a C analogue. *Bottom:* TimeLapse chemistry recodes s^6G into an A analogue.

Alternative Methods of Nucleotide Recoding

SLAM-seq is another chemistry that can induce T-to-C mutations that uses iodoacetamide to covalently attach a carboxyamindomethyl group to the thio group by nucleophilic substitution. This reaction does not recode the hydrogen pattern, but can still induce T-to-C mutations. However, the method has lower mutation rates than TimeLapse chemistry, and has not been applied to s^6G[54].

1.5. Biological systems studied using TILAC

To demonstrate the potential applications of TILAC, I applied it to several well-characterized systems. First, I studied transcriptional regulation using flavopiridol inhibition and heat shock treatment. Flavopiridol, I knew, would have strong effects on transcription that should be easily detected. The heat shock response would allow me to test TILAC on a more nuanced system, which needed to capture not only large changes in expression, but small downregulations that are challenging to detect. Studying translation is an important extension of TILAC, since it requires an additional fractionation step. Sucrose sedimentation separates polysomes (actively translating ribosomes and their mRNAs) from bulk lysate. Like any fractionation, this experiment is susceptible to variance due to its numerous handling steps. Properly identifying and controlling for this background allowed me to uncover a set of transcripts that I think could be translationally upregulated in response to stress.

I present background to these systems here, and will discuss the results in the next chapter.

Transcription Inhibition with Flavopiridol

RNA transcription is carried out by one of three RNA polymerases, numbered I, II, and III. RNA polymerase I transcribes ribosomal RNA, which makes up the majority of transcription in the cell[55]. RNA polymerase III makes 5s rRNA as well as a number of other small, noncoding RNAs like tRNA, 7SK, 7SL1, some miRNAs and snoRNAs[56]. The majority of the transcriptome is transcribed by RNA polymerase II (RNAPII) [57]. As the polymerase responsible for making protein coding genes, it has evolved to be highly

regulated, through both post-translational modifications and association with accessory proteins. The regulation is mediated in part through the RNAPII C-terminal domain (CTD). This domain is unique to RNAPII compared to I and II. It consists of a long, flexible tail made up of repeats of the consensus sequence YSPTSPS. The serine residues are differentially phosphorylated at initiation, elongation, and pausing throughout the gene body[22,23].

After initiation and before elongation, RNAPII goes through a step called promoter proximal pausing. It stops about 20-60bp downstaream of the transcription start site. This poised state is characterized by interaction with DRB-sensitivity-inducing factor (DSIF) and negative elongation factor (NELF). Release from pausing requires the kinase Positive transcription elongation factor b (P-TEFB), which phosphorylates members of the paused PolII complex[58]. Flavopiridol is a very potent inhibitor (K_i = 3nM) of P-TEFb, preventing any polymerases from entering elongation and shutting down PolII transcription[19,20]. PolI and PolIII are not inhibited, since they do not need P-TEFb for transcription elongation[59].

Heat Shock

In response to experiencing elevated temperatures (37°C for *Drosophila* cells) the transcriptional activator heat shock factor (HSF) trimerizes and binds to highly conserved DNA sequences called heat shock elements[60,61]. This transcriptionally upregulates a set of heat shock response genes, which code for protein chaperones necessary to deal with misfolded proteins and the resulting aggregation[62,61,63]. This rapidly increases the number of chaperone proteins available to mitigate the effects of protein misfolding and

aggregation. In *Drosophila*, this activation occurs within 30-120 seconds following heat exposure for genes such as *hsp70, hsp22, hsp26,* and *hsp27*[64]. Many more transcripts are downregulated compared to upregulated, and the extent of this downregulation is minor compared to the upregulation of stress-response transcripts[65]. Transcription of these repressed genes shuts off with the same speed as those upregulated[64,66]. In our lab, we have seen the bulk transcription can take up to 10-15 minutes to totally shut down (Zimmer, unpublished).

There isn't good agreement about the extent to which heat shock genes are upregulated[67]. Methods such as PRO-seq[68] and GRO-seq[69] get immediate snap-shots of RNA Polymerase II position, and therefore what is being actively transcribed. This gives a more faithful view of active transcription than steady-state RNA sequencing. However, it can only provide a momentary measurement of transcription during a time when the cell is not at steady state, and levels of transcription are changing on the time scale of minutes to hours. Under these conditions, a better measure for cellular response could be the accumulation of mature mRNAs. However, measurement of total cellular mRNA can make it hard to see small changes in transcription.

Several investigations of genome-wide changes to transcription in *Drosophila* heat shock system have employed DNA microarrays, RNA Polymerase and HSF ChIP, and PRO-seq. PRO-seq has identified active transcription happening or repressed across broad swathes of genome[67]. PolII ChIP-seq studies in S2 cells have also seen a decrease in genome-wide PolII occupancy during heat shock, often associated with changes in nucleosome turnover and chromatin accessibility[70].

Sucrose sedimentation background and puromycin treatment

Ribosomes and polysomes can be isolated over a sucrose gradient by sedimentation velocity ultracentrifugation. The movement of particles through the gradient towards higher densities is interpreted using hydrodynamic theory, which relates its size, shape, and interactions to its sedimentation coefficient. The spinning rotor creates a centrifugal force outward, which is resisted by the viscous resistance of the media the particle is moving through. For a particle with a given size and density, its sedimentation coefficient is proportional to its buoyant molar mass. Sucrose gradients are generally run in buffer conditions that preserve the native state of the molecules. Ribosomes can still dissociate to some extent during centrifugation[71], and so cycloheximide is used to trap ribosomes on their associated mRNAs[7].

Ribosome are isolated by sedimentation through a 10-50% sucrose gradient, which separates them into monosomes (RNAs with 1 ribosome) and subsequent polysome fractions (mRNAs with 2, 3, 4… ribosomes on them). Since they are part of a complex mixture, other RNPs with a similar buoyant mass will co-sediment. It's not possible to distinguish this contamination from what is truly being translated on ribosomes. To control for this background, a separate sample needs to be prepared that lacks translating ribosomes. This was done using puromycin treatment when studying the translation regulation that contributes to the *Drosophila* oocyte-to-embryo transition. This study estimated that in oocytes, 25% of the mRNAs were likely contamination, and 9% in activated eggs[72].

Ribosomes are dissociated by treating lysate with puromycin[72,73]. Puromycin is a potent translation inhibitor made by *Streptomyces alboniger* through the enzymatic

conversion of ATP[74]. It consists of a modified adenosine base, linked to a modified tyrosine amino acid by a peptide bond [75]. This is structurally very similar to a charged tyrosyl, with the most consequential difference being the peptide linkage between the amino acid and base, as opposed to a more labile, ester bond in a canonical tRNA. The consequence of this is that puromycin can enter the ribosomal A site and be covalently attached to the 3' end of the elongating peptide chain. However, an incoming tRNA cannot hydrolyze the peptide bond to continue the chain. This triggers translation termination and dissolution of the ribosome, mRNA, peptide complex[75].

While puromycin can be fed directly to live cells, translation termination *in vivo* has consequences for the living cell. As translation stops, accumulation of free mRNAs induces the formation of cellular compartments called stress granules and activates cellular stress responses[76]. I wanted a way to assess background in sucrose sedimentation gradients, without activating the cellular stress response, so I followed protocols for *in vitro* ribosome dissociation. To dissociate ribosomes in vitro, cell lysate is incubated with puromycin at 0°C. This causes release of the nascent peptide chain, but not of the tRNA or associated mRNA. Upon reheating to 37°C, the ribosomes completely dissociate [77].

Cellular Stress induced by Sodium Arsenite

Cells respond to stress by activating the integrated stress response using one of the four eIF2α kinases (HRI, PKR, PERK, and GCN2) to phosphorylate eIF2α, preventing the formation of additional ternary complex and quickly halting translation[78]. As the currently translating ribosomes terminate, unbound RNA accumulates in the cytoplasm, promoting the formation of stress granules[78,79]. Stress-responsive transcripts

such as that transcription factor ATF4 escape translational downregulation in many stresses, including ER stress, leading to accumulation of this transcription factor[97].

Studies of cellular stress commonly focus on stress granules and their accumulation and dissolution during and after stress. Since these compartments contain ribosome-free mRNA and translation initiation factors and form upon phosphorylation of eIF2α, the field speculates that they could sequester untranslated RNA, and that interactions between stress granules and the often associated P-bodies could promote sorting and degradation of some RNAs[78,80]. One hypothesis is that stress granules store RNA for translation after stress, and protect it from degradation that could be happening during stress, both in P-bodies and the cytoplasm. However, like translation, deadenylation and RNA turnover appear to be stalled during stress[81,82]. In addition, single-molecule microscopy has indicated that RNA turnover for some ribosomal transcripts does not resume for 2 hours after stress. To address the idea that only RNAs store in stress granules are being translated, the same single-molecule study assessed how many transcripts are being translated during recovery from stress. They saw a higher quantity of transcripts being translated than just what is sequestered in stress granules, indicating that storage for translation is not the main role of these bodies[84,85]. Independent studies of translation are necessary.

Knowledge of composition of stress granules could still be useful in understand the cell biological response to stress, and help interpret the results of translation studies. Proteomics data has revealed that, in addition to translation initiation factors, stress granules consist of many DEAD-box helicases, heat shock proteins, and the MCM DNA/RNA helicase complex[86]. This and other evidence showing the role of the eIF4A

helicase in stressed cells indicates that RNA helicases may play a key role in controlling RNAs that are no longer part of their canonical RNP complexes[87].

This investigation has begun with several translational studies. One used genome-wide, ribosome profiling of sodium arsenite stress in 293T cells, but focused on the potential role of 5'UTR's, without a thorough description of translational changes over time[88]. Subsequent studies have used single-molecule microscopy and focused on individual transcripts, chosen to be representative of different classes of RNA, and shown different behaviors for different classes[84,85]. These studies need to be extended to a whole-transcriptome understanding of translation at several timepoints during the course of stress and return to homeostasis.

1.6. Summary

RNA sequencing is a powerful tool for understanding gene expression and regulation, as well as translation and RNA association with RNPs. Conducting well-controlled sequencing experiments is an ongoing challenge, and new methods are always being developed to try to improve the accuracy with which RNA levels can be measured. I contribute to this effort by developing TimeLapse Chemistry Labeling in Cell Culture (TILAC), an internally controlled and normalized approach to compare RNA content between samples. Below, I describe it applications to several transcriptional and translational systems, and discuss the details of its implementation.

Chapter 2. TILAC

2.1. Author Contributions

I performed all experiments, with assistance from Rachel Niederer, who performed the sucrose sedimentation. RNA sequencing data was processed using the TimeLapse pipeline written by Matthew Simon, Martin Machyna, and Josh Zimmer. I performed the bioinformatic analysis with assistance from Isaac Vok. Simulations code was written by Isaac Vok, and I adapted it to fit my statistical model.

2.2. Summary

I describe a method (TimeLapse Labeling in Cell Culture) to internally normalize and control RNA-sequencing for comparative studies. Internal normalization requires a way to uniquely label two different populations, and to achieve this I chose to use the metabolic labels 4-thiouridine (s^4U) and 6-thioguanine (s^6G) in combination with TimeLapse chemistry. TimeLapse chemistry is a nucleotide recoding technology that can be used with high-throughput sequencing to generate a sample of sequencing reads containing induced mutations. The mutational content of the sample can be used to infer what proportion of the sample came from either of the two mixed conditions.

I apply this method to study transcription and translation in cells. I first validate my method against several well-known systems: (1) Transcription inhibition using the Cdk9 inhibitor flavopiridol (2) The heat shock response massively upregulates heat-shock response transcripts, while inhibiting transcription across much of the rest of the transcriptome and (3) Sucrose sedimentation is used to enrich polysomes and study active translation. Using the sucrose sedimentation experiment I establish in (3), I study how

translation changes during and after cellular stress with sodium arsenite. I provide a global characterization of translational dynamics and identify a set of RNA helicases as translationally upregulated during stress.

2.3. Experimental set-up

To perform TILAC, cell populations need to be independently labeled, then mixed. Mixed samples experience all of the same biochemical manipulations, TimeLapse chemistry, and library preparation (Figure 2a). The mutational content of the resulting sequencing reads allows one to infer what portion of the signal is from the experimental or control samples (Figure 2b). Once these fractions are known, the experimental signal can be divided by the control signal, to get a ratio of expression levels, or TILAC ratio. Taking the $\log_2$ of the TILAC ratio results in an internally normalized measurement of differential gene expression.

To control for the effects of the metabolic labels, two types of TILAC samples are collected. Experiments where the experimental plate of cells is fed s^4U and the control plate is fed s^6G are called "forward" samples. When the opposite combination is made, and the experimental plate of cells is fed with s^6G and the control plate is fed with s^4U, it is called a "reverse" sample (Figure 2c). This terminology is taken from SILAC[89].

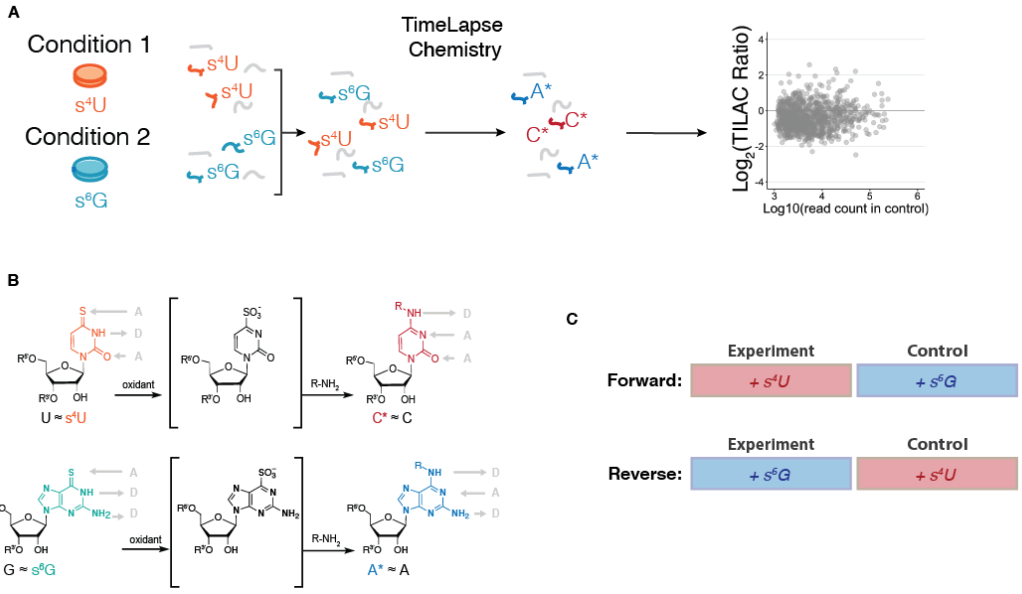

Figure 2: TILAC experimental design.

a, internal normalization is achieved by independently labeling two different cellular populations, then mixing them for downstream processing and sequencing. **b,** TimeLapse chemistry recodes the hydrogen bonding pattern of thiolated nucleobases, inducing mutations in sequencing reads. **c,** TILAC is performed in both "forward" and "reverse" combinations to control for any potential effects of the labels.

This chemistry is thoroughly validated for metabolic labeling experiments that are completed over the course of minutes to hours[52,53]. For these lengths of time, at moderate concentrations of nucleotide analogue (100uM to 1mM), there are minimal toxicity side-effects from the nucleotides, and labeling does not significantly affect transcriptome-wide gene expression. This allows one to get kinetic information on the time-scale of the average half-life of an RNA. In contrast, methods like SILAC or TILAC often want to answer a question about the steady-state levels of protein or RNA. In these cases, it is beneficial to have the entire proteome or transcriptome labeled. To perform a canonical SILAC experiment, cells are fed for at least 5 days to ensure that all peptides are labeled. Feeds of this length are impossible with s^4U and s^6G, since cells do start experiencing

toxicity effects after 24 hours of a feed with s^6G. Therefore, any TILAC experiment will need to deal with the technical challenge of inferring levels of steady state RNAs with incomplete labeling of the transcriptome.

2.4. Statistical Method

Bayesian method of analyzing data

In order to analyze TILAC data, we need to determine which reads came from which population, either the s^4U fed or s^6G fed. The simplest strategy is, for each sample, to separate reads into those with T-to-C mutations or with G-to-A mutations. There are two technical challenges that make this approach insufficient. First, due to short labeling times and incomplete labeling of the transcriptome, there are many reads with no mutations, and we cannot uniquely assign them to either condition. Secondly, mutations occasionally arise naturally during sequencing. Just because a transcript has a T-to-C mutation, does not necessarily mean it is from the s^4U fed sample. There is a small chance that the read is from the s^6G fed sample and was unlabeled due to the short feed time, but a T-to-C mutation arose as an artefact of Illumina sequencing. Additionally, reads may have both T-to-C and G-to-A mutations, where one might be induced by a metabolic label, and the other is sequencing error. I cannot assign a read like this confidently to either population. For these reasons, I needed to develop a more sophisticated statistical method to analyze TILAC data.

To develop this statistical method, I built on the existing TimeLapse chemistry analysis strategy that is used for singly-labeled samples[52]. I will describe that model first, and then extend its logical foundations to modeling TILAC data. The goal of the

TimeLapse model is to determine the fraction of sequencing reads that are labeled, based on the observed number of mutations. It models reads as coming from two populations, RNAs that existed before the labeling feed, with a number of mutations arising based on the background mutation rate, and RNAs made after the labeling feed started, that have a number of mutations arising from the mutation rate induced by labeling and TimeLapse chemistry (labeled mutation rate). That description includes several parameters that need to be estimated: (1) the fraction of labeled reads, (2) the expected number of mutations in a read from the unlabeled population, and the (3) expected number of mutations in a read from the labeled population. The number of mutations arising in a read can be modeled by a Poisson distribution, with an expected rate of mutations defined as λ. We write this Bayesian statistical model in the probabilistic programming language Stan[90], which infers maximum likelihood estimates of the parameters using a Markov chain Monte Carlo algorithm called the No-U-Turn sampler. We use the maximum likelihood estimate of the posterior as the value of the fraction labeled.

In a TILAC experiment, the same mutation kinetics and modeling strategy hold true, and I now just need to keep track of T-to-C and G-to-A mutations, and how they influence the posterior distribution of the fractions labeled in experimental or control samples, based on whether the data is from a forward or reverse TILAC experiment. This can be accomplished with careful indexing and a slight extension of the current model.

For each forward and reverse TILAC sample in an experiment, reads are aggregated into groups which align to the same gene, are of the same mutation type, and have the same number of mutations per read. Unlabeled controls are used to determine the background expected rate of mutations from sequencing error. Induced rates of

mutations for each label are modeled as a mixture of two Poisson distributions, one describing TimeLapse-induced mutations, and the other describing mutations arising from sequencing noise. These distributions are parameterized on a log scale. Fraction of reads from each sample (experimental or control) is inferred by indexing data as to whether it has come from the forward or reverse experiment and using the appropriate mutational content and mixed Poisson distributions to update the log likelihood. As an example – the number of reads containing T-to-C mutations from the forward experiment will influence the fraction of reads from the experimental sample, as well the number of reads containing G-to-A mutations from the reverse experiment.

The probability mass function describing this is:

$$f(y_m|\,\lambda_{u,m},\lambda_{l,m}) = \theta_c PoissonLog\left(y;\,\lambda_{l,m}\right) + (1-\theta_c)PoissonLog(y\,|\lambda_{u,m})$$

Where $\lambda_{u,m}$ is the rate of mutations in unlabeled reads, $\lambda_{l,m}$ is the rate of mutations in the labeled reads, for m = mutation type (T-to-C or G-to-A). y_m is the number of mutations per read of the given mutation type (m). θ_c is the fraction of labeled transcripts for the condition, either experimental or control. The condition is determined by considering the label combination (forward or reverse) and the mutation type (T-to-C or G-to-A). As briefly discussed above, T-to-C mutations from the forward replicates and G-to-A mutations from the reverse replicates are used to update the likelihood of the fraction of reads from the experimental condition. G-to-A mutations from the forward replicates and

T-to-C mutations from the reverse replicates are used to update the likelihood of the fraction of reads from the control condition.

To estimate the above global parameters, as the proportion of the dataset from experimental or controls samples per gene, I wrote the model in the Bayesian modeling software RStan, which uses No-U-turn Markov Chain Monte Carol (MCMC) sampling[90]. For this model, we chose global parameters with weak priors for expected mutation rates and fraction labeled.

Global parameters:

$$log(\lambda_{u,m}) \sim Normal(-2, 2)$$

$$\log(\lambda_{l,m}) = \log(\lambda_{u,m} + T_m)$$

$$T_m \sim \exp(0.5)$$

$$I_s = \begin{cases} 0, s \in controls \\ 1\ otherwise \end{cases}$$

$$g \in \{1, 2, \ldots, n_{\text{genes}}\}$$

Gene-specific priors:

$$logit(\theta_{exp,g}) \sim Normal(0, 1.5)$$

$$logit(\theta_{cntl,g}) \sim Normal(0, 1.5)$$

For read $i \in \{1, 2, \ldots, n_g\}$

$$f_{gm}(y_{g,m} \mid \theta_c, \lambda_{u,m}, \lambda_{l,m}) = \prod_{i=1}^{n_g} (I_s \theta_{c,g}\ Pois(y_{i,m}, \lambda_{l,m}) + (1 - I_s \theta_{c,g}) Poisson(y_{i,m} \mid \lambda_{u,m})$$

Validation of method using simulations

Before applying the model to complex sequencing data, I first wanted to understand how it works on a smaller dataset. In R, I simulated a TILAC experiment with 75 genes, each gene with its own unique expression pattern across the experimental and control samples. The experiment I simulated contained 3 samples, 1 forward TILAC mix, 1 reverse TILAC mix, and 1 control, unfed sample.

To perform the simulation, I first need to choose the expression pattern of each of those 75 genes, which is the fraction of reads in those samples unambiguously attributable to the experimental condition (F_e) or control condition (F_c), as well as the fraction that cannot be uniquely assigned to either condition. First, I chose the fraction of the sample attributable to the control condition (F_c) randomly from a uniform distribution between 0 and 0.8. From this, I calculated the corresponding fraction attributable to the experimental condition (F_e) by subtracting $1 - F_c$, and multiplying by a random downscaling factor that I drew, again, from a uniform distribution between 0 and 0.8. I multiplied by this random downscaling factor to make sure the fraction experimental and control did not add to 1. The remaining reads belong to the group that cannot be unambiguously assigned. I then calculated the fold change by taking the log-2 of the experimental divided by control fractions. This is the "ground truth" against which I will measure how my statistical model performs.

To simulate data for these 75 genes with their various fractions, we need to know the T-to-C and G-to-A mutation rates, both in reads that are labeled or that unlabeled. From our sequencing data, we observe a 5% induced mutation rate and 0.1% background mutation rate for T-to-C mutations. For G-to-A mutations, there is a smaller induced

mutation rate (2%) and a slightly higher background mutation rate (0.4%). Rather than using the same set mutation rates for each simulated gene, we let the mutation rate vary slightly between them. This should capture the fact that biologically genes might have different mutation rates due to noise. For each gene, a unique set of mutation rates are chosen from a set of normal distributions, each with a mean equal to the percentages described above, and a standard deviation of 0.2.

For each gene, we simulated 5000 reads, each 200nt long. They were split into 3 bins, labeled with s^4U, labeled with s^6G, and unlabeled. The number of U's or G's per read is drawn from a binomial distribution with a chance of being a U or G each set to 0.25 (there are 4 possible bases). Based on the bin, the number of TC or GA mutations per read was determined using a binomial distribution for the numbers of U or G in a read and the corresponding mutation rates.

The resulting data included information about the sample, gene, mutation type, number of mutations of that type per read, and the number of reads observed with those qualities. This data was analyzed using the Poisson model described above. As can be seen in Figure 3, there is very good agreement between these values, indicating that the Bayesian model is accurately inferring the gene expression differences.

Figure 3 TILAC Bayesian Model Validation.

Fold changes in gene expression between two conditions (experimental and control) are inferred using a Bayesian Poisson model. The known fold change in expression was set, and given a set of experimental parameters, a small sequencing dataset was simulated. This was then analyzed using the Bayesian Poisson model. The known and inferred fold changes in expression are plotted against each other.

Additionally, there are some model validation parameters that need to be examined at the end of each model run to determine if it converged. The first is the Gelman-Rubin convergence diagnostic, R, which should be less than 1.01 if the model has converged. Another metric is the ratio of effective number of samples to actual sampling steps (abbreviated here as n_{eff}). This is essentially a metric for how randomly the model iterated while it ran, and anything at or above 1 means the model was sufficiently random. Metrics for this simulation are shown in Figure 4.

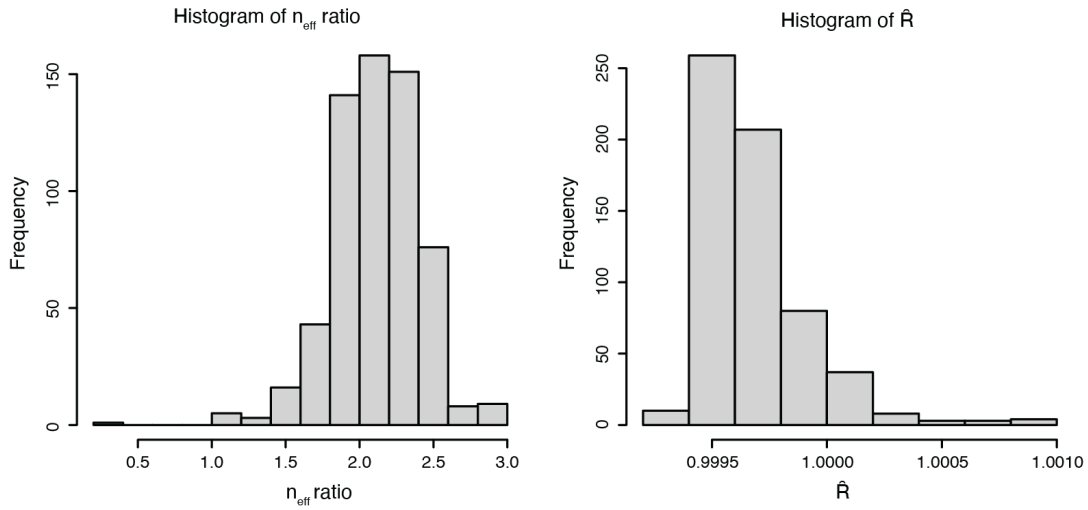

Figure 4: Rhat and n_eff metrics for simulated data

Two additional metrics evaluate individual model parameters. The pairs plot shows the correlation between inferred values. Parameters such as the expected rate of mutation during labeling (λ_l) for s^4U, s^6G, and the background rate of mutations in unlabeled reads (λ_u) rates should be uncorrelated. Since their posterior distributions are normal with respect to each other, we can see that they are indeed independent and well-converged. Lastly, we can look at the trace plots. The trace plots show how the model explores the parameter space at each iteration. A trace plot that oscillates rapidly up and down indicates that the model is rapidly exploring the whole parameters space, as indicated in Figure 5: Pairs plots and trace plots of simulated data.

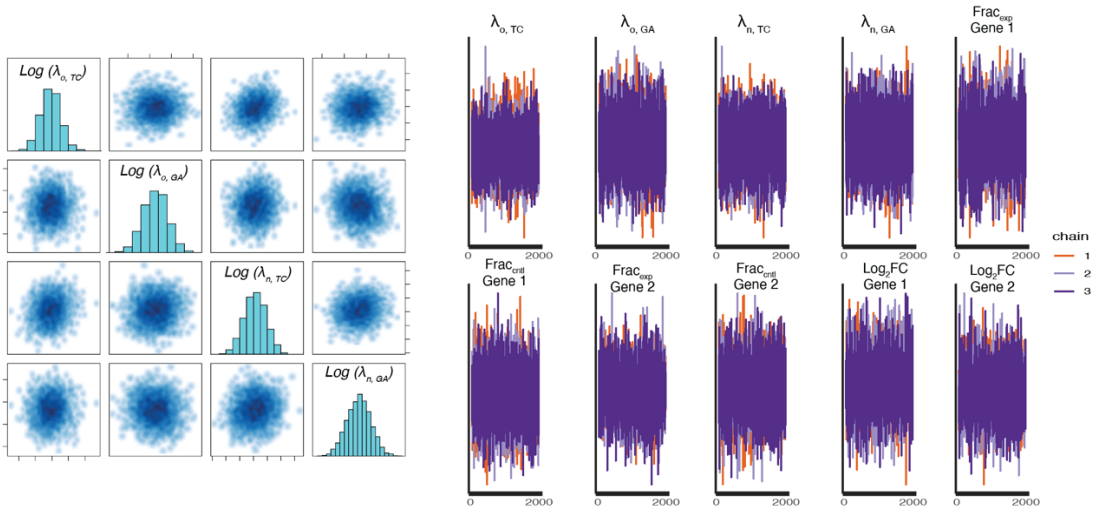

Figure 5: Pairs plots and trace plots of simulated data

Normally distributed pairs plots indicate independence of estimates. Chains oscillate rapidly up and down, which is an indication that the model is rapidly exploring the potential parameter space.

2.5. Application to transcriptional inhibition by flavopiridol

Flavopiridol inhibits cyclin-dependent kinase positive transcription elongation factor b (P-TEFb) [35]. This shuts down transcription across most of the genome. These huge changes in transcription should be easily detected by TILAC.

To perform this experiment, I fed the cells (100uM nucleotide) at the same time as we treated cells with flavopiridol (500nM). After two hours, I harvested cells, mixed, and performed TILAC. In addition to the forward and reverse TILAC experiments, I collected unmixed samples, both fed and unfed, to analyze with conventional RNA sequencing analysis tools. After harvesting and extracting RNA, I performed qPCR to look for transcriptional shutdown. I saw that 500nM flavopiridol had robust downregulation of CDKN1B pre-mRNA compared to mature mRNA, indicating that transcription was shutdown. In addition, I saw that slowly made and slowly degraded

RNAs like ACTB had less robust downregulation compared to MIR17HG, which is quickly synthesized and degraded, and therefore is highly downregulated.

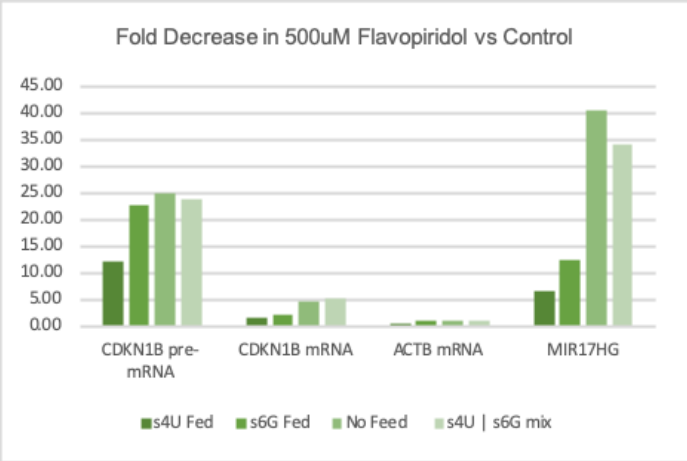

Figure 6: qPCR validation of flavopiridol treatment

Decrease in transcription after 2 hours of flavopiridol treatment. Fast-turnover RNAs such as CDKN1B pre-mRNA and MIR17HG show much higher fold decreases in transcription compared to slow-turnover RNAs like CDKN1B mRNA or ACTB. Only one replicate was performed. We did not try to draw any conclusions beyond that the flavopiridol treatment was working.

Before applying our statistical model, we analyzed the raw data to get a sense of whether we were capturing changes in gene expression. We considered the forward and reverse experiments separately, calculated the T-to-C and the G-to-A mutation rate, and used that as a proxy for expression levels in either condition. We plotted the mutation rates against each other for each label combination (Figure 7). In samples where flavopiridol-treated cells were fed with s^4U, there is a depletion of reads with T-to-C mutations, where as there is a depletion of G-to-A mutations in samples where the flavopiridol-treated samples were fed with s^6G. The depletion of s^4U reads in the forward experiment, where the s4U fed cells are treated with flavopiridol, (Figure 7, left)

is much less obvious. This is due to the low s^6G incorporation rate compared to the high s4U incorporation rate.

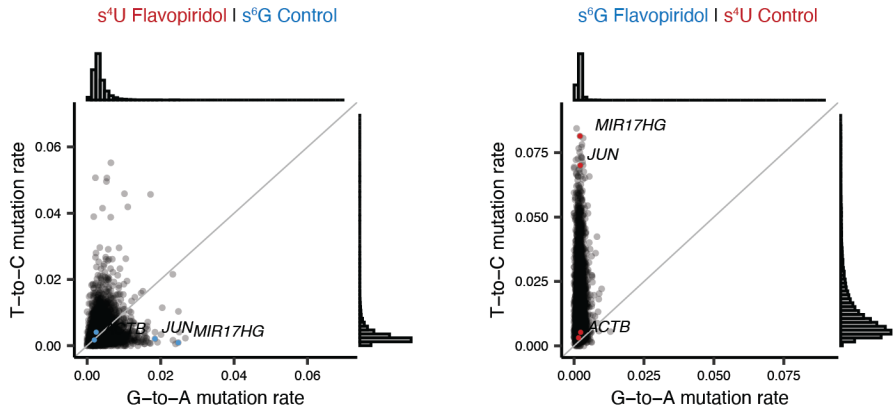

Figure 7: Comparisons of mutation rates in transcription inhibition experiments

Left: Comparison of TC and GA mutation rates when s4U is fed to flavopiridol-treated samples. s^6G has a lower mutation rate, and there is incomplete labeling. The histograms accentuate the spread of the data, and show a broader distribution than for TC mutations. *Right:* Comparison of TC and GA mutation rates in the reverse experiment. Here, efficient s4U incorporation captures all active transcription.

Despite this technical challenge, s^6G is still catching important experimental trends. I evaluated how well both labels were capturing experimental data by comparing the mutational content of TILAC samples to that of their singly-labeled controls. In Figure 8, I look at how the s^4U content of the forward and reverse TILAC experiments correlates with a singly-labeled s^4U fed and flavopiridol treated sample and with a singly-labeled s^4U fed control sample. There are correlations among similar treatments. This same analysis for s^6G in Figure 9 showed similar trends. If the s^6G label were not capturing any experimental signal, I would expect no correlation between any of the samples. Therefore, while s^6G has significantly lower mutation rates I can safely conclude that it is functioning well as a metabolic label in TILAC experiments. We have

observed that s^{6}G works well with both high and low mutation rates, but the next steps in this project will include a full investigation of the sensitivity and specificity of TILAC at different mutation rates and sequencing depths.

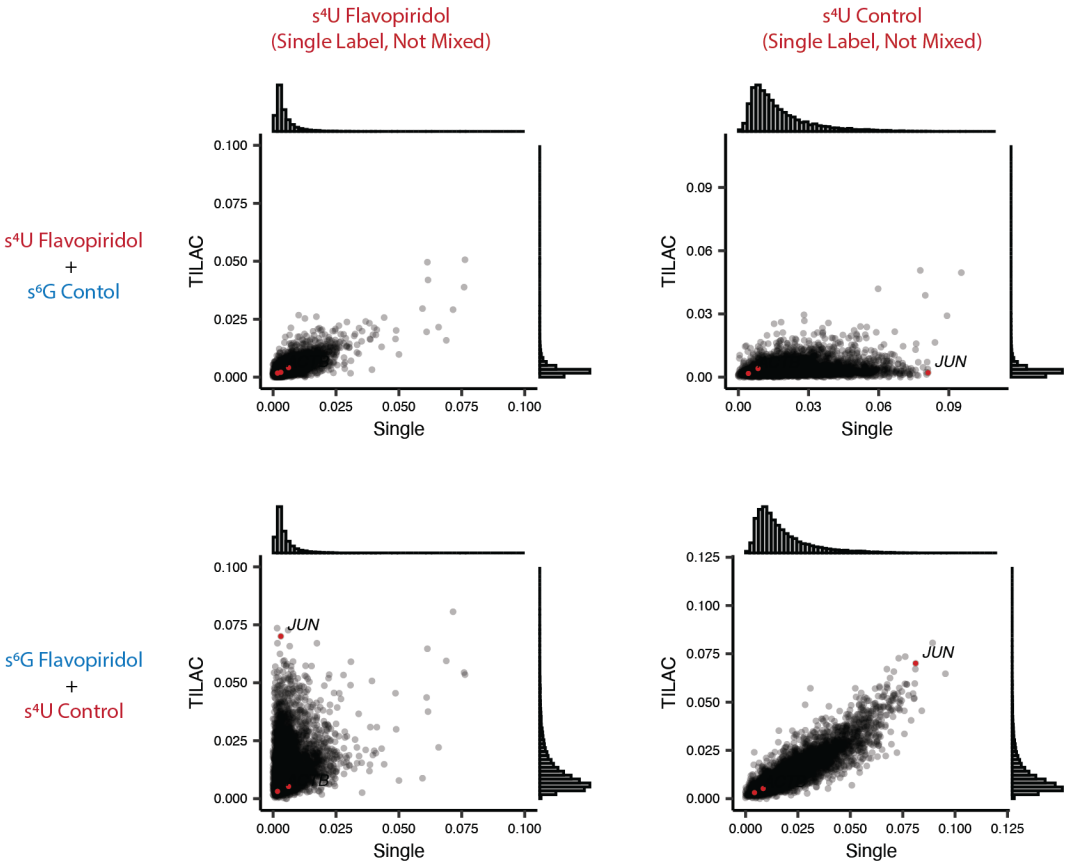

Figure 8: Comparing TILAC samples to singly-labeled controls

The T-to-C mutational content of the TILAC experiments is compared to the T-to-C mutational content of samples unmixed, and fed only with s^{4}U.

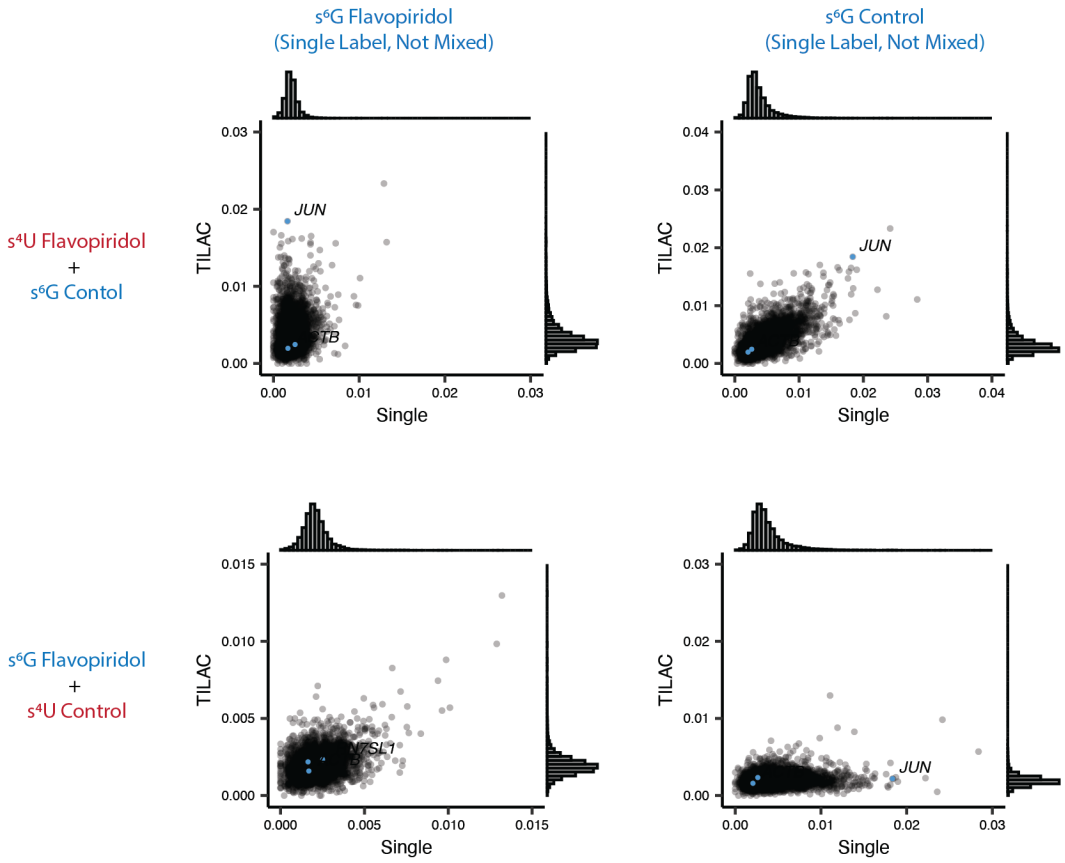

Figure 9: Comparing TILAC samples to singly-labeled controls

The G-to-A mutational content of the TILAC experiments is compared to the G-to-A mutational content of samples unmixed, and fed only with s^6G.

Another way to evaluate how well the labeling is working is by looking at sequencing tracks. The Simon Lab creates sequencing tracks with reads colored by the number of mutations in the read. This type of analysis allows me to understand how the labeling is affecting individual genes, and is an important complement to the abstraction of examining global mutation rates. Both s^4U and s^6G are efficiently incorporated into MYC transcripts. MYC is a high-turnover gene, meaning it is synthesized and degraded quickly. Over the course of a 2-hour experiment, many of its old transcripts will be degraded and replaced with newly synthesized, and hopefully labeled, mRNA. MYC has

an enrichment of reads with G-to-A mutations in the forward experiment (s^4U flavopiridol and s^6G control) and an enrichment of reads with T-to-C mutations in the reverse experiment (s^6G flavopiridol and s^4U control). In yellow is highlighted background mutations, either T-to-C in the forward experiment, or G-to-A in the reverse experiment (Figure 10). Incorporation is easily distinguishable, and there is low background from other mutations.

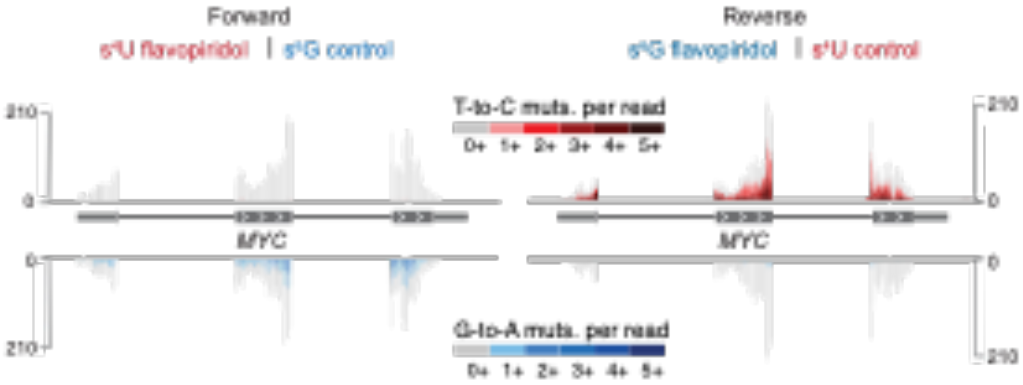

Figure 10: Label is efficiently incorporated into Jun transcripts

Top: In the forward experiment, s^4U fed cells are treated with flavopiridol, so most reads are expected to come from the s^6G fed control sample. *Bottom:* The opposite is true, and the s^6G fed cells are treated with flavopiridol. Reads are expected to predominantly be form the s4U fed sample. Reads of the opposing type are highlighted in yellow, and they constitute a minimal proportion of the reads covering the gene.

I applied my Bayesian statistical method to estimate fold-changes in expression between flavopiridol treated and untreated cells. Below, I show model validation for this full dataset, as first outlined in the TILAC simulation (pg 27). I show this data to demonstrate that the model works with the same precision on a real, complex dataset as it does on the simulated data. I will not show model validation for subsequent experiments,

as it becomes rather repetitive. In Figure 11, I show that the *R* metric converges to values around 1, and distribution of n_{eff} ratio values are at 1 or larger.

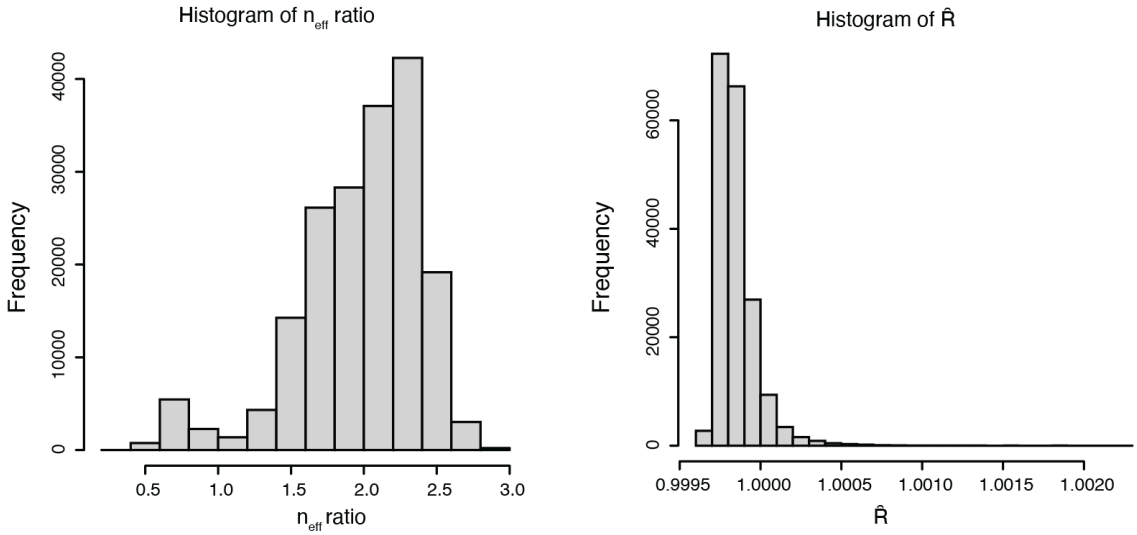

Figure 11: Rhats and N effective for flavopiridol transcription inhibition

R values converge to values less than 1.1 and the n_{eff} values are predominantly above 1, indicating the model sampled efficiently.

In Figure 12a, the posterior distributions that describe the are largely normally distributed, and in Figure 13b, 4 different chains explore the likelihood in efficiently and in a random manner.

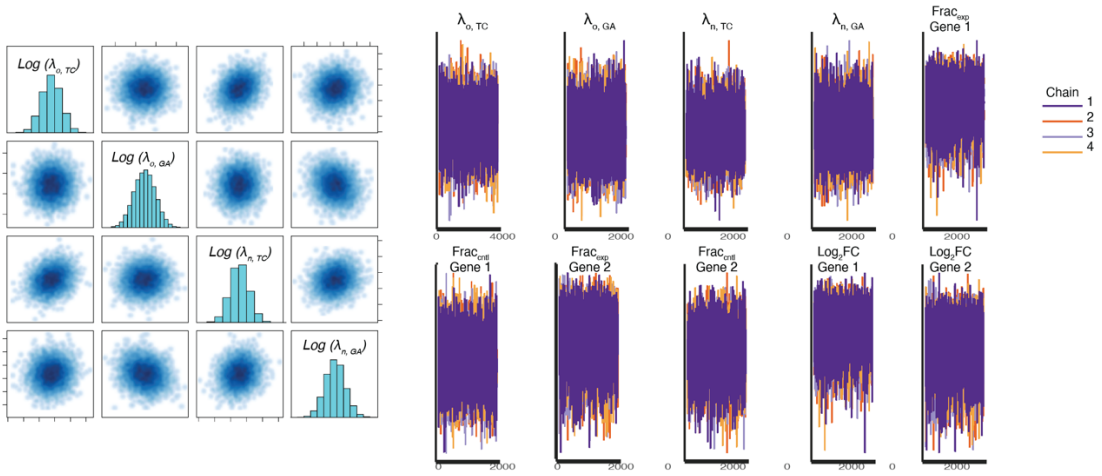

Figure 12: Pairs plots and trace plots of transcription inhibition data

Normally distributed pairs plots indicate independence of estimates. Chains oscillate rapidly up and down, which is an indication that the model is rapidly exploring the potential parameter space.

The TILAC analysis pipeline does capture downregulation of a significant portion of the transcriptome. As it preserves the ratio of RNA between the two samples, it achieves this without the use of spike-ins or additional statistical assumptions. Out of 23,090 measured genes, 9364 are downregulated. The model is not able to tell if 13,704 genes significantly changed in either direction. Transcripts made by Pol I or Pol III, for example 7SL1, we believe do not change expression between samples. Other are made by PolII, but are known to be resistant to flavopiridol treatment, such as IKBKI, PLCB4, BRWD3, and PHLPP[148]. Many of the other nonsignificant genes have quite low read counts (< 100 reads/gene) (Figure 13). Since the variance of count data is inversely correlated with the mean, we expect high variability in read count data over lowly expressed transcripts, and would hope that a good statistical method will not assign value to these genes, when we should be rightly skeptical of them. Visual examination of these

genes confirms that many of these have sparse or variable coverage. Therefore, we conclude that TILAC is adept at catching wide-spread changes in gene expression.

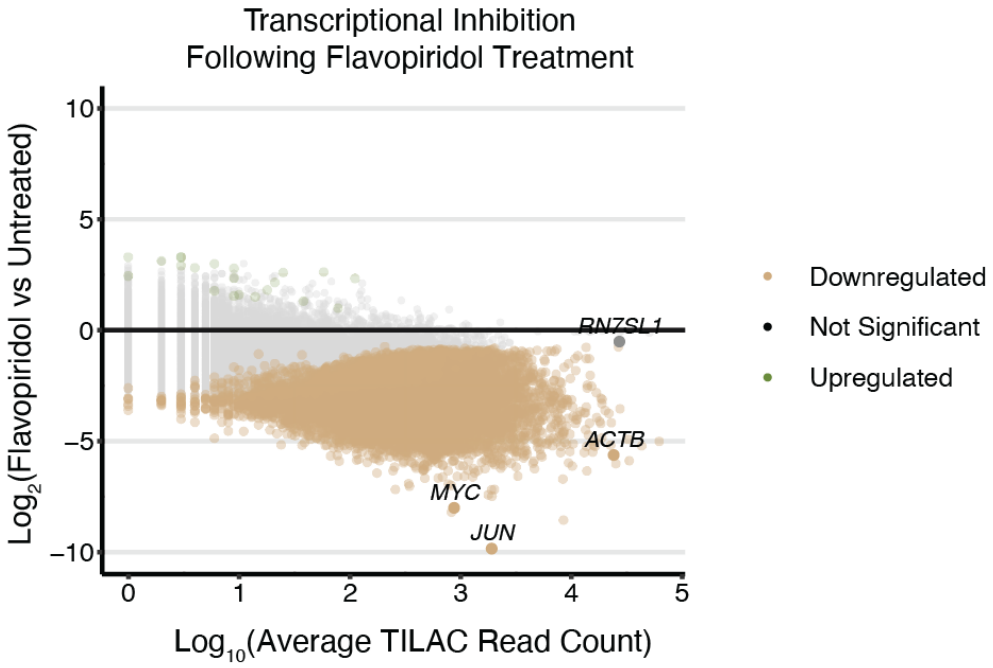

Figure 13: TILAC captures transcriptional inhibition by flavopiridol

Cells were treated with 500nM flavopiridol, and combined with untreated cells for TILAC analysis. This plot shows large downregulation of the transcriptome.

I then compared our results to those obtained by DESeq2, the conventional software for calculating expression changes. Normalization methods such as DESeq2 often assume that a small subset of genes experience changes in expression levels, while most transcription remains unchanged. This assumption does not hold when cells are treated with the RNA polymerase II inhibitor flavopiridol. As expected from the assumptions discussed above (pg 7), DESeq2 shrinks fold change values towards zero, and identifies only 748 downregulated genes (Figure 14: Transcription inhibition measured by DESeq2). That is an order of magnitude fewer transcripts called

downregulated than TILAC. DESeq2 identifies 15,132 transcripts as not significantly changed, out of which 11,658 genes have read counts too low to make conclusions. This number is fairly similar number to what is found by TILAC.

Comparisons against DESeq2 are a useful foil to put TILAC results into context, and will be used again in the heat shock experiment. Importantly, TILAC and DESeq2 work under different experimental conditions and assumptions. It is not valid to draw more extensive comparisons. A major difference is the use of the metabolic label. Since TILAC labeling is done over the course of the 2 hour flavopiridol experiment, TILAC is largely evaluating the difference in expression over that two hour time period of transcriptional shutdown. DESeq2 is performing a true bulk analysis on total cellular RNA.

In conclusion, TILAC does capture global downregulation of transcription, and is capable of identifying when it can and cannot be confident in its ability to estimate differential expression. TILAC is useful in that it captures more transcriptional downregulation than DESeq2, and this is due both to difference in statistical assumptions, and also in experimental design. TILAC is useful for assess acute changes in transcription, especially when the feed time is concurrent with the drug treatment.

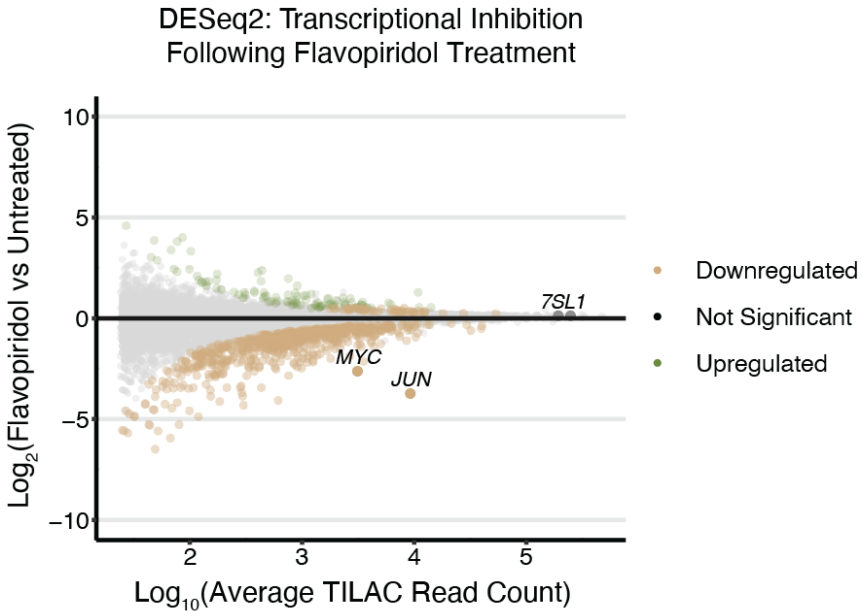

Figure 14: Transcription inhibition measured by DESeq2

Unmixed samples were collected at the same time as Figure 13 and analyzed using one common differential expression software.

2.6. Heat Shock

We next applied TILAC to the Drosophila heat shock response (pg 14), in which heat shock responsive genes are highly transcriptionally upregulated, while much of the genome is transcriptionally repressed[67,70,91]. While s^4U has been used in Drosophila experiments previously[52], no thiolated G analogue has been tried. We learned through work done by Jeremy Schofield while developing TimeLapse (data unpublished) that these nucleotide analogues are not well incorporated into HeLa cells, indicating that incorporation is highly dependent on cell type. Therefore, I tested the two G analogues previously used in TimeLapse, s^6G and 6-TG, to see if they are sufficiently incorporated into *Drosophila* RNA for a TILAC experiment. Both s^4U and s^6G are incorporated into

Drosophila RNA, although about 4-fold less than into 293t cells. 6-TG is not incorporated at all (Figure 15). Because there was still sufficient incorporation, we decided to proceed with the experiment.

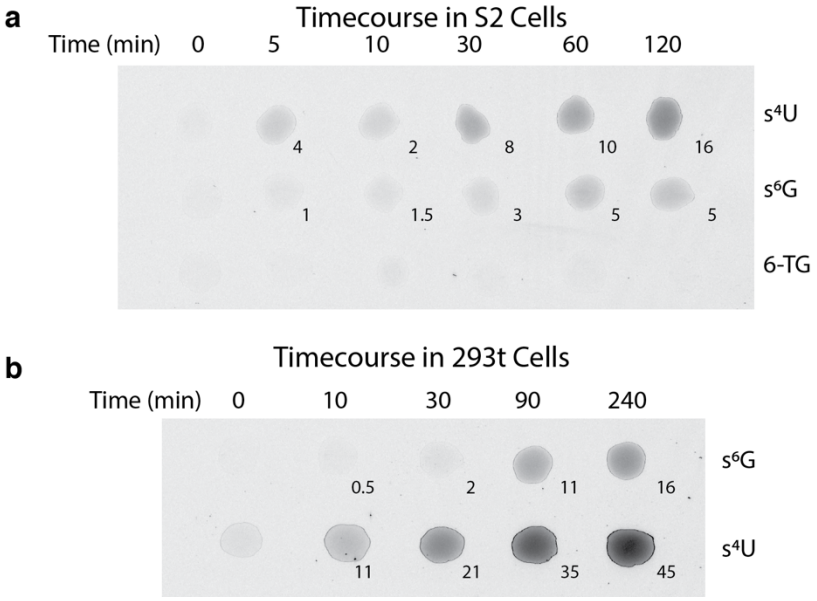

Figure 15: Dot blot to assess nucleotide incorporation in S2 cells

a, thiolated nucleotides were fed to *Drosophila* S2 cells for indicated times. This is comparing s^4U, which has been previously validated in S2 cells [52], to s^6G and 6-TG, since this is the first time the Simon Lab has fed these nucleotides to *Drosphila*. 6-TG is not incorporated. s^6G is incorporated, but less than s^4U. **b,** Data from Lea Kiefer, taken to compare incorporation between *Drosophila* and mammalian cells. There is less incorporation of nucleotides over time in *Drosophila* cells. However, trends agree. There are dose-dependent increases in nucleotide incorporation, and s^4U is more extensively incorporated than s^6G.

The experimental set-up was the same as for transcription inhibition with flavopiridol. I collected duplicates of the two TILAC label combinations. In the "forward" experiment, heat-shock samples were fed s^4U, while in the "reverse" experiment, s^6G was fed to the heat-shock samples. Heat shock conditions and protocol

were taken from lab member Martin Machyna[92]. We again examined mutation rates in both the forward and reverse experiments. In the forward direction, there is the expected enrichment of T-to-C mutations in heat shock transcripts. In the reverse experiment, G-to-A mutations are enriched in the heat shock transcripts (Figure 16). The transcriptional downregulation is more subtle in this experiment compared to flavopiridol treatment, and more challenging to see in this preliminary analysis, especially with relatively weak s^6G incorporation. This underscores the need for a robust statistical method for TILAC data analysis.

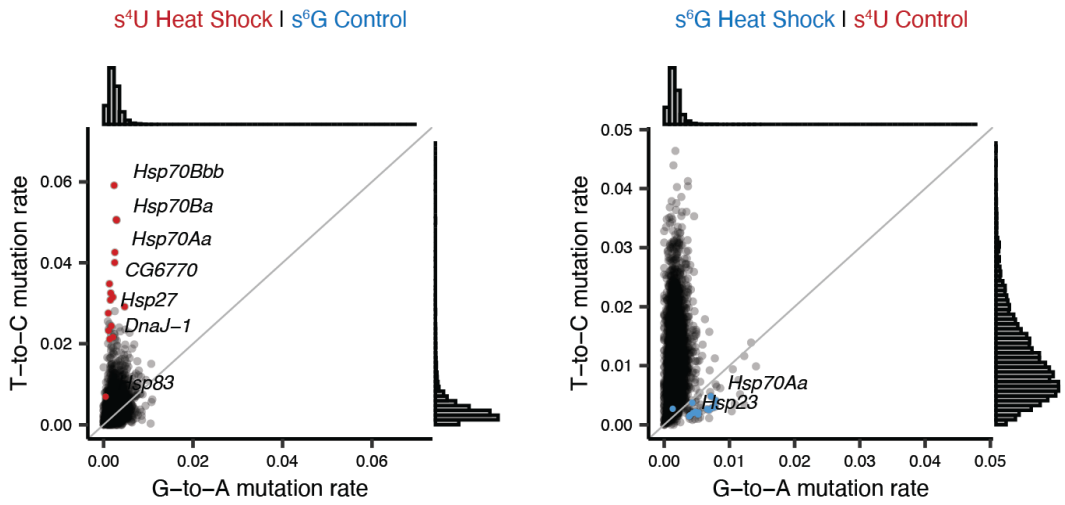

Figure 16: Comparisons of mutation rates in heat shock experiments

Left: Comparison of TC and GA mutation rates when s4U is fed to heat shocked samples. *Right:* Comparison of TC and GA mutation rates in the reverse experiment.

Again, we applied the TILAC Bayesian method and plotted the inferred fold changes as an MA plot. The most highly differentially regulated genes are also some of the most highly expressed, and are plotted in the top right-hand corner of the graph. These include the canonical heat shock proteins - Hsp26, Hsp23, Hsp68, Hsp70Ba, Hsp70Bc, Hsp27, Hsp70Ab, and Hsp83. It also includes known accessory proteins such

as DnaJ-1, Hsc70-3, Hsc70Cb, Hsc70-5, and Hsc70-4. In addition, 1109 transcripts are identified as being downregulated.

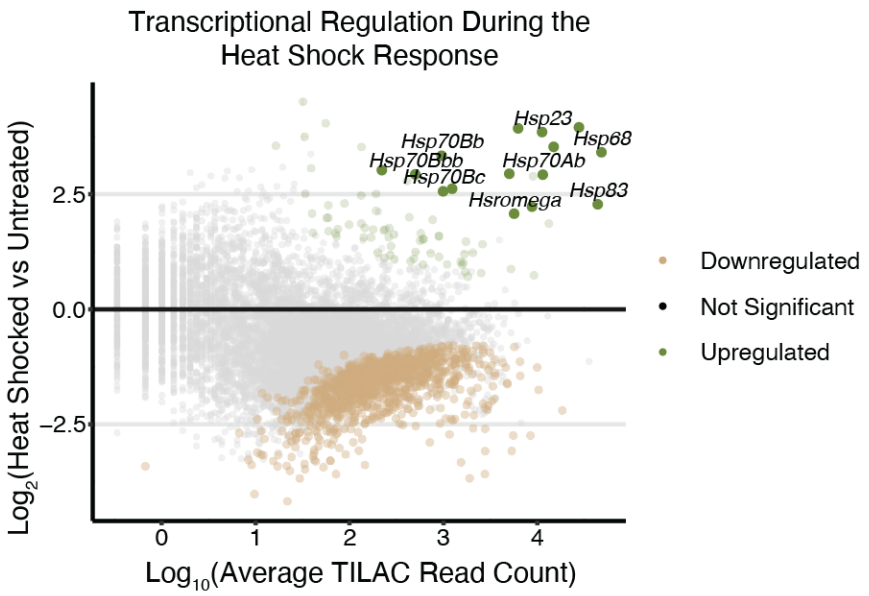

Figure 17: TILAC captures transcriptional regulation during 1 hour of heat shock

TILAC identifies the highly upregulated heat shock proteins, and also general transcriptional downregulation.

The TILAC result is generally in line with the literature. Early examination of transcription on polytene chromosomes and Pol II ChIP-seq have both observed genome-wide decreases in Pol II throughout gene bodies[70,91,93]. By DNA microarray, 508 transcripts are found to be downregulated[65]. In contrast, Duarte et al. uses PRO-seq to profile transcription at 20 minutes of heat shock and finds 2300 transcripts to be downregulated. Using DESeq2 to analyze her inputs, Duarte does not see the significant decrease in transcription. Neither do I in my study, in which DESeq2 reports 137 transcripts downregulated. Duarte also cites the challenges of normalization in the heat shock system as a reason she cannot identify changes in the transcriptome by bulk RNA

sequencing, and uses this to motivate her use of PRO-seq. My study and the Duarte et al. 2016 study cannot be directly compared because they were performed at different timepoints in heat shock, and because the methods measure different aspects of the response. This point about what the methods measure is an important way to consider how TILAC fits into the body of available techniques to study transcriptional changes. While DESeq2 is limited to assumptions on total RNA, PRO-seq can only provide a snapshot of what is being actively transcribed over a very short period of time. PRO-seq cannot measure how much of that transcription goes on to be a functional and stable mRNA in the cell. With metabolic labeling and internal normalization, TILAC can measure precisely what was transcribed over the feed time and identify broad transcriptional downregulation.

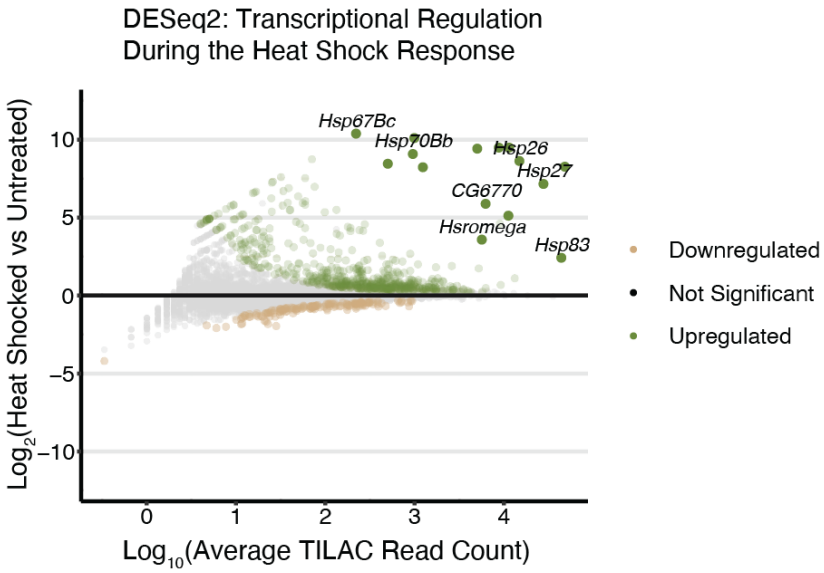

Figure 18: DESeq2 analysis of transcriptional changes during heat shock

DESeq2 captures transcriptional upregulation of heat shock proteins, but not the subtle downregulation that accompanies it.

2.7. Puromycin

Actively translating ribosomes and their associated RNAs have to be purified in order to learn what proteins are being synthesized in the cell. This is done by fractionating cell lysate over a sucrose sedimentation gradient. Since sucrose sedimentation separates cellular lysate by size and density and is not a clean immunoprecipitation, there could be other RNP background in the fractions containing ribosomes[72]. In some systems, up to 25% of the isolated transcripts are from contaminating RNPs[72], making this a significant challenge in accurately quantifying. The background was identified by treating cell lysate with puromycin, which dissociates ribosomes from their associated mRNAs, and comparing what was in the polysome fractions to that in normal, untreated lysate. This analysis requires a reliable spike-in control[94]. TILAC can improve this experiment, since it will control for the numerous handling steps at which biases could be introduced, and eliminates the need to devise clever spike-in methods. Therefore, I performed a TILAC experiment in which I mixed untreated cell lysate, or cell lysate treated with puromycin to dissociate polysomes[8,72,77]. Transcripts that showed decreased polysome enrichment upon puromycin dissociation are truly associated with ribosomes, while anything that becomes either enriched or is not significantly changed is considered potential background.

Cells were fed for 4 hours with 100uM s^4U or s^6G, harvested, and lysate was treated with puromycin (Figure 19a). After puromycin treatment, TILAC samples were mixed, and ribosomes were isolated by sucrose sedimentation. As controls, unmixed untreated and unmixed puromycin treated samples were also analyzed. Polysome

fractions collapse in the puromycin-treated samples, indicating that the treatment worked (Figure 19b,c).

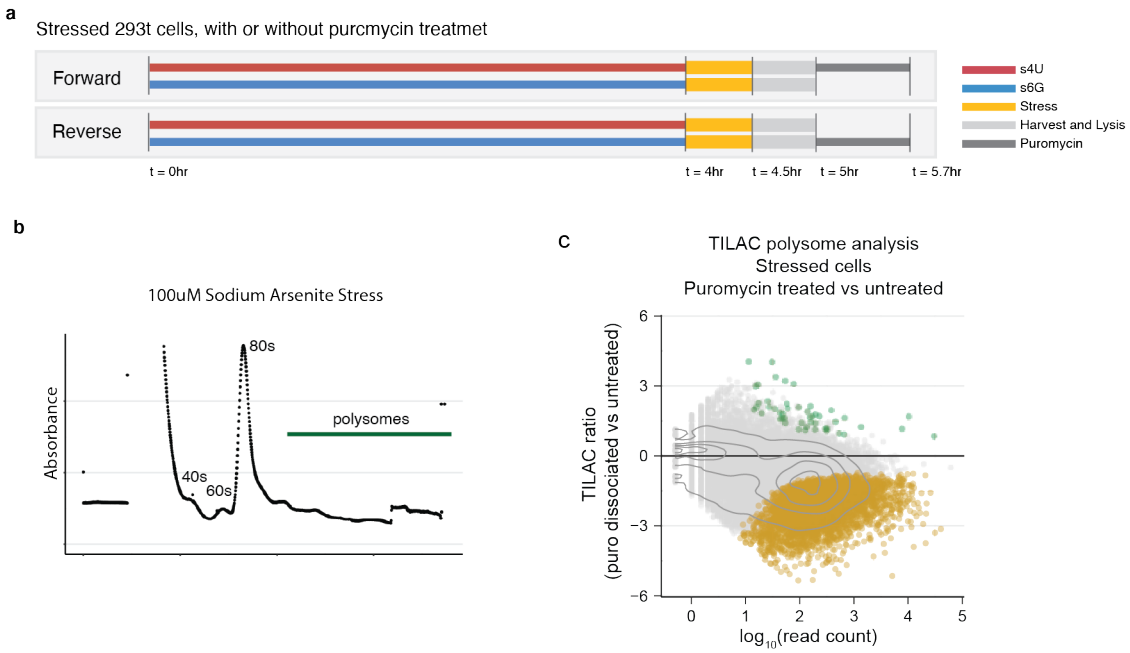

Figure 19 Sedimentation traces of lysate with or without puromycin treatment

a, Experimental set-up of puromycin treatment for forward and reverse TILAC samples. **b,** Untreated cell lysate shows canonical polysome oscillations. **c,** Puromycin treated samples lack polysome signal.

The resulting sequencing data was analyzed with the TILAC analysis pipeline. Of the 21,286 genes analyzed, 11,856 transcripts are confidently downregulated and 9430 transcripts would be considered background contamination. Since almost 45% of transcripts could be background, it is important to run this control in new systems before drawing conclusions about translational regulation.

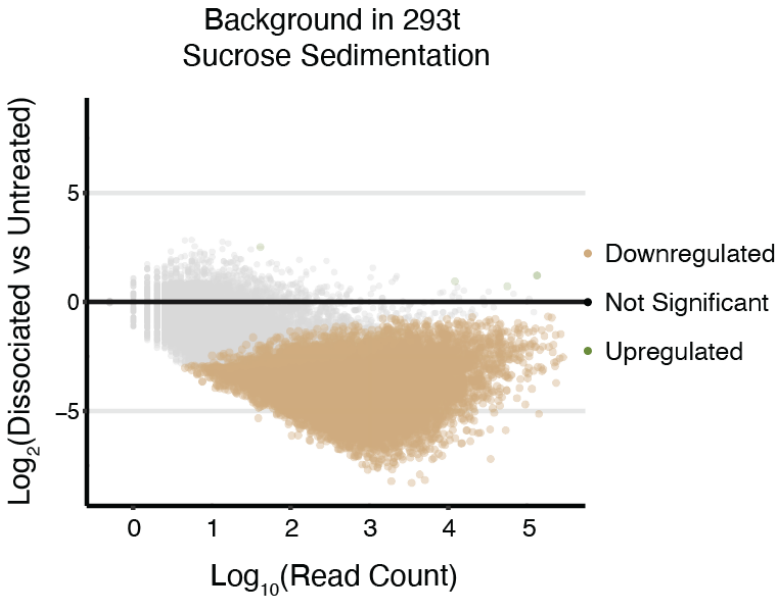

Figure 20: Puromycin treatment of 293t cells

Upon ribosome dissociation due to puromycin treatment, 11,856 transcripts become depleted from polysomes, indicating they are being actively translated. There are 9430 transcripts that are not depleted, and are considered contaminating background.

2.8. Stress response

Translational regulation is central to the integrated stress response cells use to respond to a variety of potentially toxic situations, such as ER stress and oxidative stress. One of the most commonly studied stress systems is sodium arsenite treatment. The cell biology of this system has been fairly well described, and is characterized by the formation of cytoplasmic stress granules. The biochemical details around how translation starts and stops has not been thoroughly described.

Armed with the knowledge of what is background in polysome fractions from the previous experiment treating cells with puromycin (pg 46), I was uniquely set up to rigorously describe translational changes. Therefore, I decided to stress 293t cells with

sodium arsenite and to profile translation using three comparisons: (1) TILAC comparisons of stressed and unstressed cells, (2) TILAC comparisons of stressed cells and cells 1 hour after recovery, and (3) TILAC comparisons of stressed cells and cells 6 hours after recovery. In order to do this experiment, I first validated the sodium arsenite treatment in 293t cell. Then I performed a puromycin experiment to characterize the background in sucrose sedimentation during stress conditions.

Optimizing sodium arsenite stress in 293t cells

The concentration of sodium arsenite and treatment time vary quite a bit from paper to paper. One study examined translational restart of a single reporter transcript in U2OS cells by treating with 100uM sodium arsenite for 30 minutes[84]. I verified these conditions were appropriate in 293T cells using a puromycin incorporation assay, which has been used previously to assess translation during sodium arsenite stress[80]. This assay does indicate that translation plateaus at a minimum around 30 minutes into treatment (Figure 21a). Translational recovery after stress was examined using the same assy. After 30 minutes of sodium arsenite stress, media was replaced and puromycin incorporation seemed to be increasing 1 hour after recovery from stress (Figure 21c).

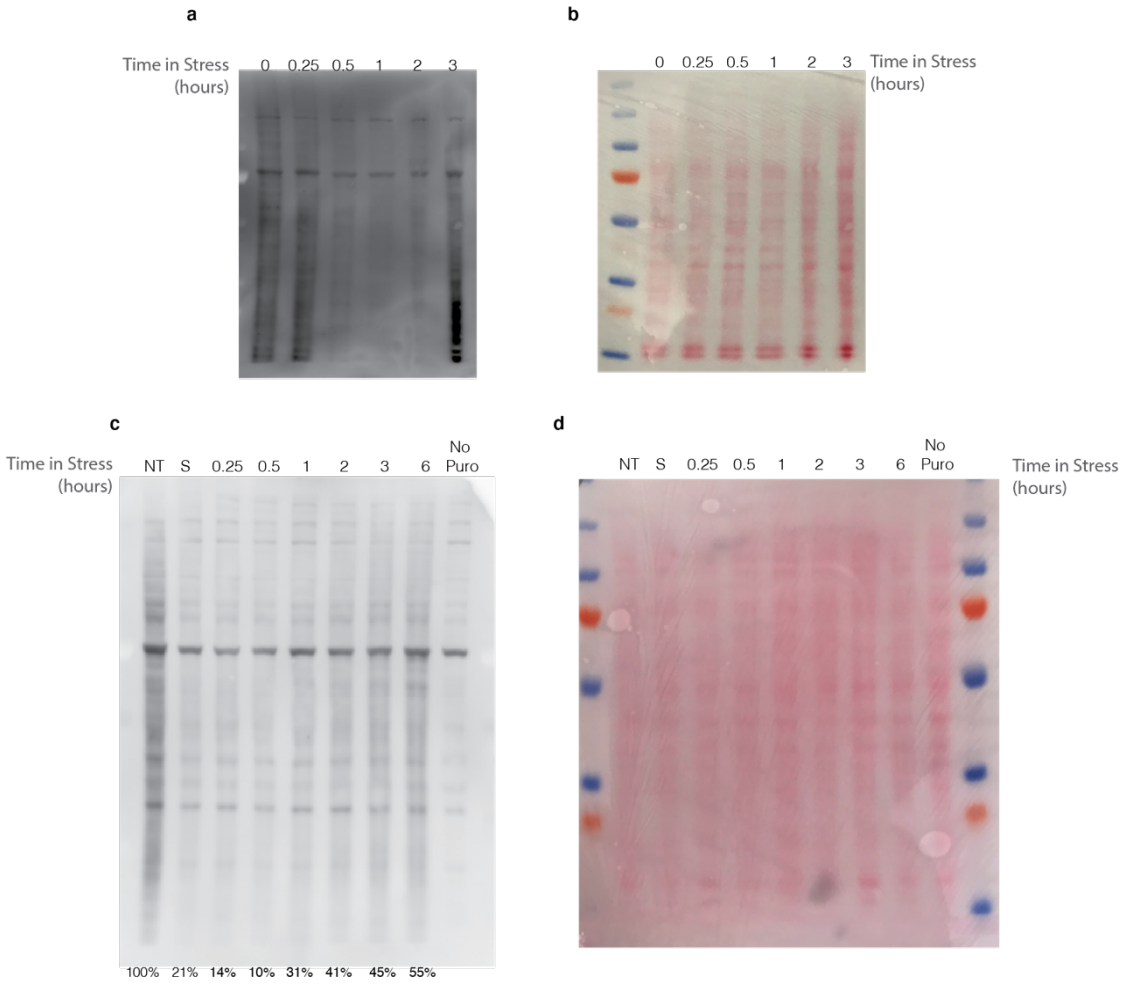

Figure 21 Puromycin incorporation assay to assess translational shutdown and restart

a, Cells were treated with sodium arsenite for the indicated amount of time. The last 15min included a puromycin feed to assess the extent of active translation. **b,** Ponceau staining of (a) to verify equal protein loading. **c,** The same puromycin incorporation experiment is used to assess when translation restarts. The red S indicates 30min of 100uM sodium arsenite stress. d**,** Ponceau staining of (c) to verify equal protein loading.

Identifying contamination background in sucrose sedimentation gradients

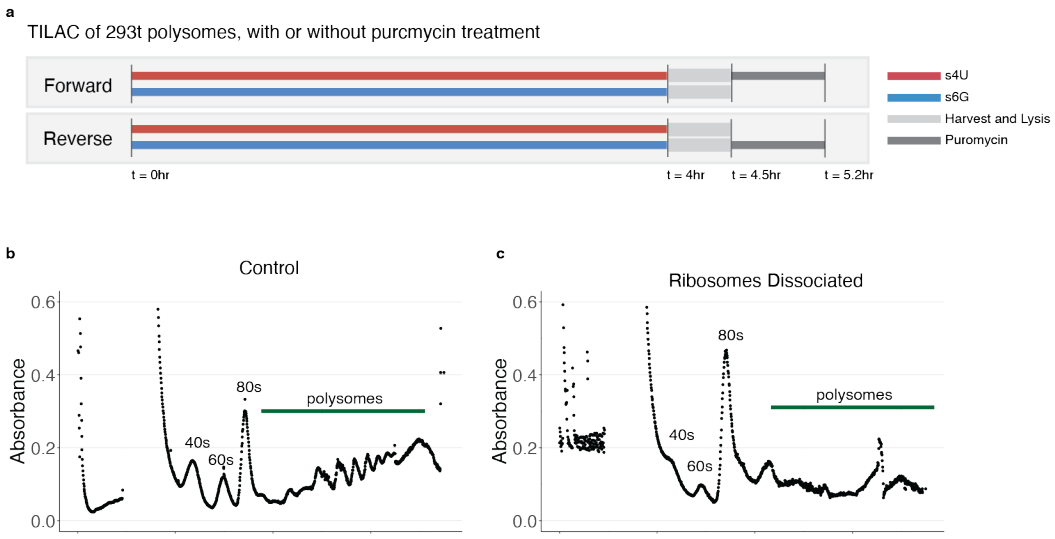

Figure 22: TILAC is used to look for background in both stressed and unstressed cells

a, Cells are fed for 4 hours, then treated with sodium arsenite for 30 minutes. Cells are harvested and lysate is kept on ice or treated with puromycin. **b,** Absorbance trace for TILAC sample combining stressed cell lysate with similar lysate that has been treated with puromycin. Compare polysome fractions to Figure 19.

To look for contamination in polysome fractions during stress, I performed the same experiment as described above (pg 46), but with the addition of a 30 minute treatment with sodium arsenite (Figure 22). After TILAC analysis, there are 48 transcripts upregulated, 3671 downregulated, and 18,830 are not able to be significantly

called either way. To use both puromycin experiments as controls, I made a combined list of all transcripts that were either more present in the puromycin sample, or had no statistically significant downregulation in either the stressed or unstressed comparisons. I eliminated these transcripts from analysis in any of my experimental samples. Fortunately, there were very few transcripts identified in these controls that had to be filtered out of the experimental results (Table 1).

	Status in Polysomes	Total	Present in Control	Confident
Stress vs No Stress	**Up**	64	11	53
	Down	1541	7	1534
1 hour Recovery vs Stress	**Up**	5	1	4
	Down	1	0	1
6 hour Recovery vs Stress	**Up**	1452	5	1447
	Down	33	6	27

Table 1: Translational regulation and contamination during stress

Experimental Design

I assayed translation at 4 different time points, (1) no treatment, (2) stressed, (3) 1 hour of recovery, and (4) 6 hours of recovery. I followed the same basic outline of what was performed for the two puromycin experiments. I fed cells for 4 hours to build up a pool of labeled RNAs, then treated and combined samples as outlined in Figure 23.

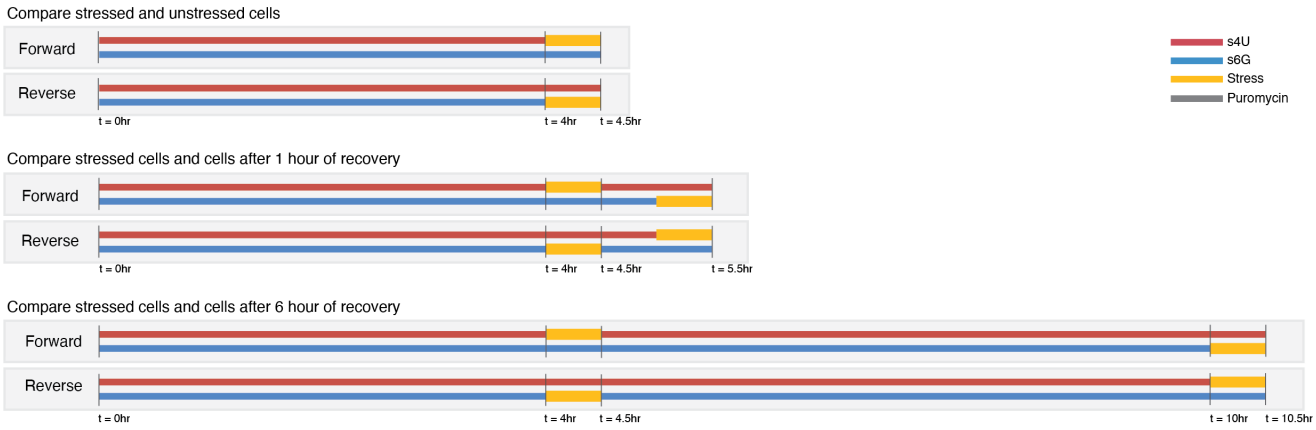

Figure 23: Experimental design to study translation during stress

Three different samples were collected to compare cells that are unstressed, 1 hour into recovery, and 6 hours into recovery to cells experiencing stress.

To analyze the resulting sequencing data, I used a false-discovery rate of 0.05, and compared results to the null hypothesis that the absolute value of the fold change is equal to or less than 0.5. Using these parameters to compare stressed and unstressed conditions and accounting for likely background, I identified 59 transcripts that are more translated, 1,541 that are downregulated, and 21,144 that cannot be significantly called up or down (Figure 24a).

GO analysis of the transcripts that have increased presence in polysomes indicate enrichment for proteins that could be related to the stress response. This includes the RNA helicases DHX30, DDX5, DDX24, UPF1, the U5 helicase SNRNP200, and also components of the DNA helicase MCM Complex. Many of these proteins are enriched in stress granules, indicating there may be a functional reason the cells needs to produce more of these proteins[86]. Recent work has suggested that RNA helicases may be important for managing stress by acting as ATP-dependent RNA chaperones[95]. Several heat shock proteins were identified (HSP90AB1, HSPA1A, HSPA8) and eEF2 which

plays a role in phosphorylating eIF2α during ER stress[96]. Dropping the fold change cutoff to 0.3 increases the number of helicase transcripts called translationally upregulated, and includes all components of the MCM complex.

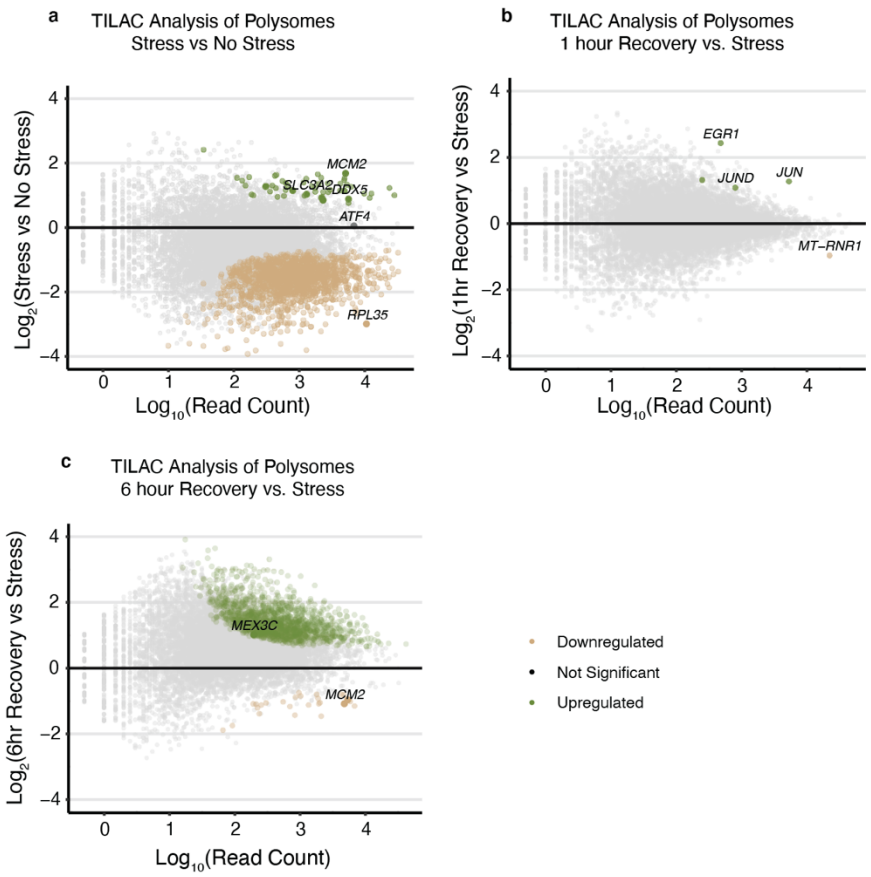

Figure 24: Translational upregulation during cellular stress

a, Enriched transcripts include RNA helicases. Depleted transcripts include RNA transcription factors and several eIF's. **b**, There are almost no changes in polysome association. **c**, About 1500 transcripts are more polysome associated. About half are transcripts that became depleted in (a). The other half are enriched for transcripts encoding ubiquitin/proteasome associated proteins, but most of these are transcriptionally up as well.

ATF4 is a ubiquitously expressed mRNA, but the protein is present at very low levels in cells at homeostasis. Hypoxia, which induces eIF2α phosphorylation, increases ATF4 levels without any transcriptional response. ATF4 transcripts shifts from mostly being partially in monosome fractions and largely in low polysome fractions to higher polysome fractions in mammalian cells under ER stress[97]. Since I am combining all polysomes fractions, I'm unlikely to see this translational shift from low to high polysome fractions. Another study performed ribosome profiling on 293t cells, and did not find significant upregulation in translation efficiency. They did see global downregulation of ribosome footprints across the transcriptome, showing that ATF4 ribosome-footprints and translation efficiency were significantly up relative to other similarly expressed transcripts. Our data agree in that there are many transcripts translationally down in polysomes compared to ATF4. I additionally identify translationally upregulated transcripts with functions linked to stress survival. I based my experiment on one that examined translational recovery after stress using a single-molecule reporter system and smFISH[84]. This paper analyzes the translation of a reporter construct containing the 5'UTR of RPL35. It sees translation shutdown after 30 minutes of stress, which we also see.

The transcripts that are lost from polysomes during stress are enriched for the eIF's (4H, 1, 1B, 5, 3G, 2B1 and several others) and RNA Pol II regulators and transcription factors. This fits with the general transcription and translation downregulation that occurs during stress.

Translated at 1 hour after stress

There are very significant changes in what is enriched in polysomes between stress and 1 hour of recovery. Five transcripts are more expressed – JUN, EGR1, JUND, and DUSP1. The only downregulated transcript encodes a mitochondrial protein, RNR1. This result generally agrees with results from my puromycin incorporation results and literature, which indicate there is little translation happening 1 hour after stress[80] (Figure 24b). This is in contradiction to the single molecule studies of the reporter construct with the 5'UTR of RPL35, which seems to have resumed translation 30 minutes after release from stress[84]. RPL35 is translated 6 hours after release from stress.

Translated at 6 hours after stress

At 6 hours after stress, 1452 transcripts are translationally upregulated. Of these, 50% (749 transcripts) were translationally downregulated during stress (Figure 24c). Of those that are newly upregulated, GO analysis reveals they are highly enriched for several categories of proteins involved in the ubiquitin-proteasome system (p=2.82E-4 to 1E-11)[98]. The majority of these are also transcriptionally upregulated.

How input RNA changes between stages of stress and recovery

There are relatively few changes in the population of total RNA between the unstressed and stressed conditions. Going into stress, 40 transcripts appear to be translated less, among them 15 mitochondrial transcripts. Five transcripts appear to be translated more. Two of these are JUN and DUSP1, which will become translationally

upregulated 1 hour after recovery from stress. ATF3 is a stress-responsive transcript (Figure 25c).

Between stress and 1 hour of recovery from stress, there are no significant changes in the transcriptome. The 4 upregulated transcripts are JUN, EGR1, DUSP1 and FOS (Figure 25b).

After 6 hours of recovery, 7338 transcripts appear to be significantly enriched in total RNA, compared to the stressed condition. Since we have already observed that there is no significant RNA turnover during stress or after 1 hour of recovery, I conclude that this increase in labeled RNA comes from the cell highly upregulating transcription between 1 and 6 hours after recovering from stress. From this experiment, it is impossible to distinguish the effects of upregulated transcription or upregulated degradation, but this is an interesting future direction (Figure 25c).

These results follow the same general trend as observed with an RNA-reporter construct using single molecule FISH. This reporter experienced a halt in transcription and degradation of mRNAs for about 2 hours after removal of stress[84]. I cannot say precisely when transcription and degradation restart based on my experimental results, but they do support the idea that there is a relatively long pause of RNA kinetics during and following sodium arsenite stress.

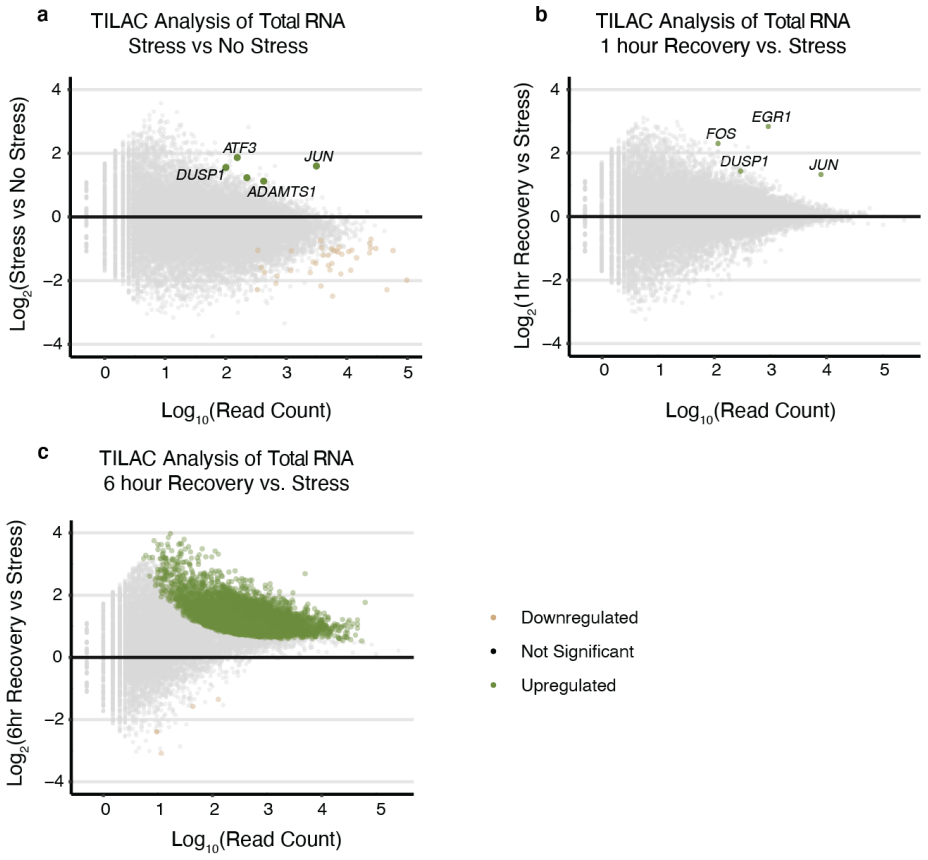

Figure 25: Transcriptional changes during sodium arsenite stress

a, There are few transcriptional changes during stress. **b**, There are few transcriptional changes between stress and 1 hour recovery. **c**, About 7338 transcripts are enriched in the 6 hour recovery samples.

2.9. Conclusions and Discussion

TILAC is a metabolic labeling and analysis method that can be combined with any type of biochemical experiment that would benefit from internal normalization. I have shown how it can be used to measure RNA levels in whole cell lysate or after fractionation. In addition to the experiments described here, TILAC would be beneficial when combined with a variety of other subcellular fractionations and formaldehyde RNA immunoprecipitaitons. A TILAC experiment can be performed with variable feed times, allowing customization of the method for each individual experiment. The experiments performed in this section used labeling times that spanned 45 minutes to 10.5 hours. Experiments cannot be conducted for timespans greater than 24 hours, due to s^6G toxicity effects. Analysis is done with a statistical model that can be set up to be relatively user friendly, with little to no customization needed to analyze a variety of experiments.

Chapter 3. Chromatin-Associated RNAs

3.1. Author Contributions

I performed all experiments. DNA origami was made by John Powell in the Lin Lab, and he also took some negative stain images. RNA sequencing data was processed using the TimeLapse pipeline written by Matthew Simon, Martin Machyna, and Josh Zimmer. Bioinformatics analyses are my own, with advice from Jeremy Schofield.

3.2. Introduction

Over the past decade, it has been revealed that mammalian genomes are broadly transcribed, including intergenic regions. Transcription over regions that do not code for proteins produces non-coding RNAs, many of which are function. Examples of small non-coding RNAs include snoRNAs and snRNAs involved in ribosome biogenesis and splicing, and also miRNAs involved in gene silencing. Long non-coding RNAs (lncRNAs), on the other hand, appear to be localized to and function predominantly on chromatin. Two such lncRNAs are involved in regulating gene expression during dosage compensation. roX2 is a *Drosophila* lncRNA that is part of the male sex-lethal complex. It localizes to the X-chromosome in males and upregulates transcription so that males have equal numbers of transcripts to females with two X-chromosomes[99]. Mammals have developed a similar system using the lncRNA Xist, which silences one of the X-chromosomes in females[100,101].

Xist is necessary but not sufficient for XCI, and acts by binding to the X-chromosome, spreading in cis, and triggering silencing of transcription[102]. While several models have been proposed to explain various aspects of this process [103-106], we still do not know the biochemical and structural details underlying the molecular mechanism of

XCI. Structural studies of Xist have been hampered by the challenges of working with a flexible 18 kb RNA. Work to identify and characterize proteins in the Xist ribonucleoprotein (RNP) has produced conflicting results [107-109].

I took inspiration from XCI in mammalian cells to pursue the first two projects of my thesis. In this section, I will tell you about my efforts to determine the structures of small, modular, structured regions of Xist. Then, I will tell you how I was inspired to investigate the Xist binding protein hnRNP-U, and its role in retaining RNAs on the chromatin.

3.3. A method to probe three-dimensional structures of long-noncoding RNAs

My original thesis project examined the mechanism of XCI by interrogating the structure of Xist. Current structural data came from chemical probing data from our lab and several others[110-113]. I used these secondary structure predictions as an important starting point for 3D structural studies. I integrated RNA chemical probing, DNA nanotechnology and cryo-electron microscopy in pursuit of a new method that would harnesses the modularity of lncRNAs to provide the first high-resolution structures of small, independently folding portions of Xist that are ~50-300 nt long.

Structures of small RNAs using cryoEM

The field is still working to describe the molecular events leading to the transcriptional silencing and heterochromatization of one X-chromosome during XCI. In

XCI, the X-chromosome that expresses Xist is transcriptionally silenced and heterochromatinized. Xist is necessary for XCI, and generally thought to act in three steps: binding to the chromatin, spreading in cis, and silencing transcription[105,106]. The extent to which Xist functions independently in XCI or acts as a scaffold for helper proteins is unknown. Efforts to characterize Xist's protein partners have produced conflicting results, identifying anywhere from 10 to greater than 200 interactors depending on the study[107-109].

I aimed to understand the role that Xist structure plays in XCI. In order to function, RNAs fold into specific structures. LncRNAs have been shown to be composed of smaller modules that fold and function independently[114,115]. Xist likely has similar modules (Figure 27), and understanding the structure and function of each of these would shed light on how the whole molecule coordinates two partially separable functions, 1) associating with the X-chromosome and spreading in cis, and 2) triggering silencing of transcription.

At roughly 18 kb, the size of Xist makes it hard to purify or *in vitro* transcribe, and thus the type of careful biochemical dissection necessary has been essentially impossible. Genetics experiments have indicated function for only several small regions of Xist, the A repeat region (repA) and the C repeat region (repC) (Figure 26) [106]. Both regions are predicted to contain modular, highly structured segments of Xist[110], indicating that these regions require further *in vitro* characterization. *In vivo* deletion analysis demonstrated that (repA) is important for silencing. This segment of Xist also associates with chromatin, and mutational analysis indicates that the structure of repA is responsible for this interaction[116]. If we could understand those mutations in the context of a tertiary

structure model of repA, we could identify specific nucleotides that may be involved in associating with the chromatin.

RepC was also identified by deletion analysis and is associated with spreading[106]. This was confirmed by studies using variations on standard oligonucleotides, such as locked nucleic acids and protein nucleic acids, that bind to specific regions of Xist and disrupt the function of that region. These studies found that targeting repC with LNAs disrupted Xist's interaction with the X-chromosome(Figure 26)[110,117,118]. One model to explain these finding is that the oligonucleotides disrupt a structure that allows repC to interact with the chromatin or with helper proteins. Tertiary structure determination would confirm that the structure of this region is important.

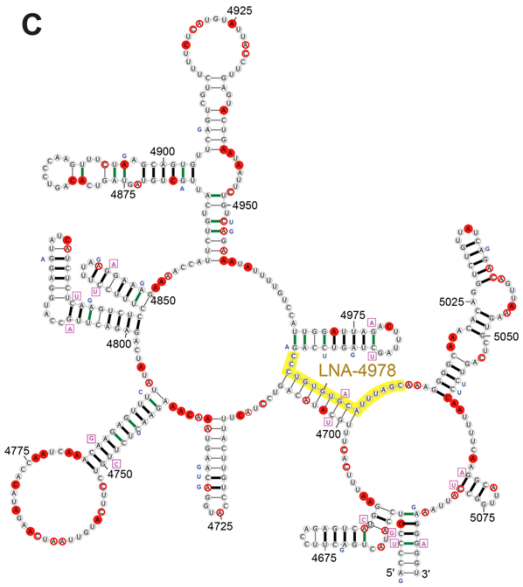

Figure 26: Xist repC region

The repC region of Xist is both functional and folded. When targeted by locked-nucleic acid probe at the region highlighted in yellow, Xist is knocked from the chromatin.

In addition to the two regions above, the Simon Lab has used computational folding algorithms to predict 26 modular regions of Xist ranging from 70 – 300nt in size (Figure 27) [110]. These regions are too large for NMR, too flexible for crystallography, and slightly too small for current methods of single particle cryoEM. I set out to develop a new single particle cryoEM method that is easily engineered and versatile, which would allow me to investigate a variety of RNAs of different sizes with moderate throughput. I proposed to use a DNA origami frame to anchor these RNA modules, restrict conformational heterogeneity, and assist in orientation and localization of smaller RNA molecules in single particle cryoEM structure determination.

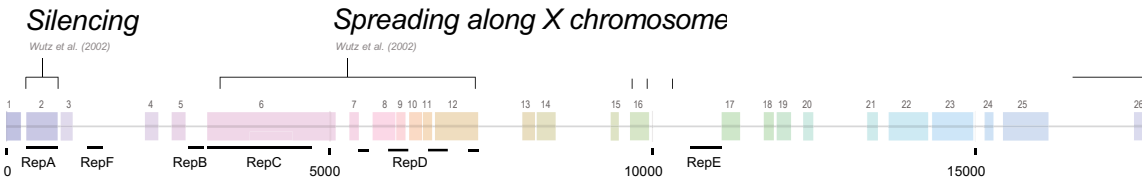

Figure 27: Structured Regions of Xist

Fang et al., 2015, used computational folding algorithms to on the primary sequence of Xist and identified 26 regions predicted to be module and structured.

Piloting DNA origami strategy using known RNAs

The modular, functional regions of interest in Xist, discussed above, are roughly 50 – 150 kDa in size. In 2016, these pieces of Xist were rather small for cryoEM structure determination. However, the cryoEM field was in a period of rapid technical advances in both hardware and software, and the first studies of small proteins were being published as I began working on this project[119-121]. To anchor small RNA domains, restrict conformational heterogeneity, and assist in orientation and localization of the molecules, I planned to put these pieces in the context of a large and easily oriented DNA origami frame. We designed this frame in collaboration with the Lin Lab at Yale, who are experts in DNA origami. The frame is made of interconnected DNA helices, with single-stranded DNA oligo handles that extend from the frame into the inner window that are complementary to single stranded handles on the RNA (Figure 28). This allows the RNA to be loaded into the frame by hybridization. While working on this project, the Scheres lab published a paper using a similar origami scaffold to determine the structure of a DNA binding protein to about 15 Å resolution[122].

Results of cryoEM structure determination

While my longer-term goal was to study Xist and other lncRNA fragments, my pilot work to develop this method focused on the c-di-GMP riboswitch. It is a good model for several reasons. First, at about 100nt, it is roughly the same size as some of the smaller regions of Xist. Second, preparing well-folded samples is aided by the fact that it has picomolar affinity for its ligand, ensuring a high percentage of RNA molecules will be properly folded. Finally, the Strobel Lab at Yale solved its structure by

crystallography[123]. I received technical support and mentoring in performing IVT, RNA folding, and radioactivity work from Caroline Reiss. Finally, I would be able to compare my structure to that determined by crystallography to ensure that my method is not introducing biases into the final structures.

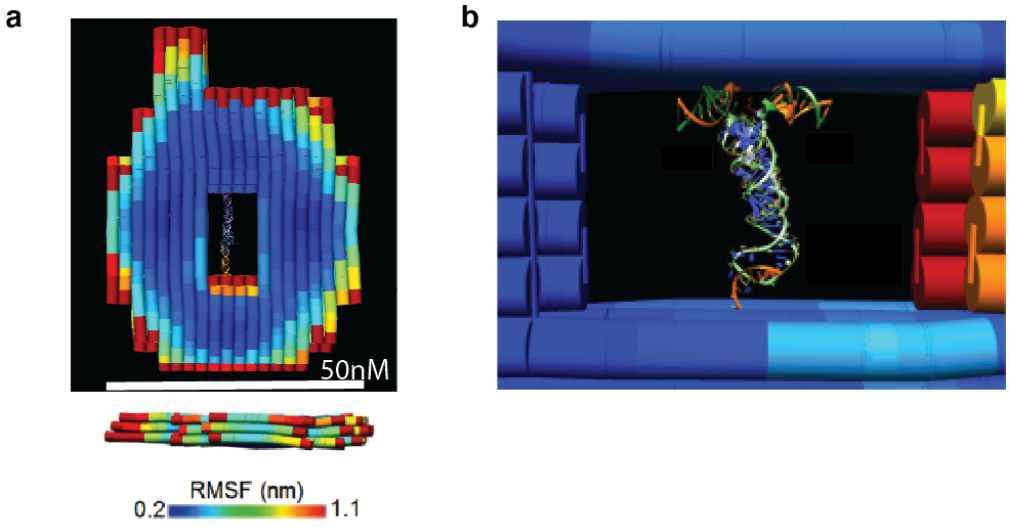

Figure 28: DNA origami frame used to scaffold small RNAs for cryoEM

DNA origami frame designed by John Powell in the Lin Lab. **a,** The DNA origami frame is roughly 50nm by 60nm, and 3 helices tall. It is shaped like a thumb, which gives it a unique view at every orientation and is appropriate for single particle reconstruction. Colors indicate the predicted flexibility, and overall we expected this frame to be fairly stable. Across the window of this frame is just a single-stranded piece of DNA, which is just to indicate an RNA could be loaded. **b,** Zoomed in view of the hole in the frame shown in a loaded with the c-di-GMP riboswitch. It is attached to the frame by 3 handles (orange), anchoring it in one plane relative to the rest of the frame.

I transcribed, folded, and purified two versions of the c-di-GMP riboswitch. One had 10nt long handles, and the other had 19nt long handles. My collaborator in the Lin Lab produced versions of the frame that accommodated either version. We reasoned that

a shorter linker (10nt) might keep the riboswitch more tightly oriented on a plane within the frame, but on the other hand we might need 19nt handles to get specific loading of the RNA into the frame. Both RNAs were able to be loaded into the frame (Figure 29).

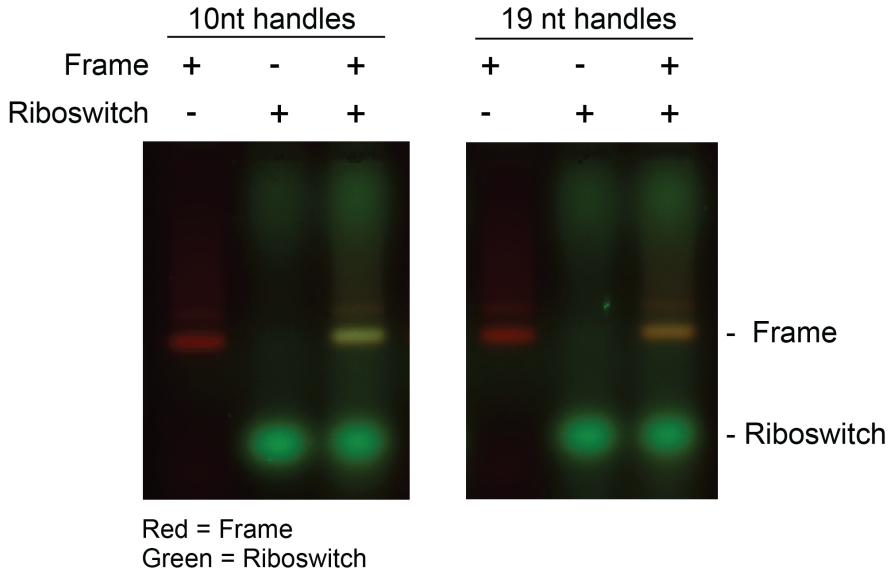

Figure 29: Loading the c-di-GMP riboswitch into the origami frame

Left: Loading a c-di-GMP riboswitch with 10nt long handles. *Right:* Loading a c-di-GMP riboswitch with 19nt long handles

Once the RNA was loaded, John collected and manually averaged some negative stain images, and we saw density in the frame that looked like it could correspond to a loaded RNA. I then screened cryo conditions on a FEI Technica T-12, and saw that I could get several orientations of the frame in ice (Figure 30).

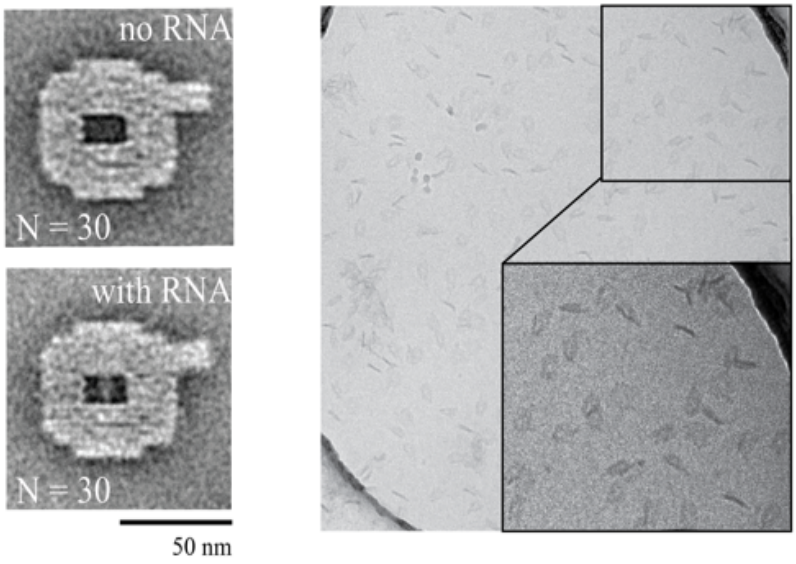

Figure 30: Electron microscopy of a riboswitch loaded into a DNA origami fram

Left: Negative stain TEM averages of the frame, with or without RNA loaded. *Right:* A cryoEM micrograph taken while screening freezing conditions on a T-12 with CCD camera. *Insert:* Image contrast enhanced in ImageJ.

Once I had dialed in on a range of freezing conditions that worked, I collected a preliminary dataset on a Krios, and performed class averages in RELION[124,125]. Unfortunately, those classes revealed several challenges. First, the frame was much more flexible than we had anticipated, and was liable to warping. It was unlikely to be able to be reconstructed to any resolution. Second, with my freezing conditions, I was not able to

get enough different orientations of the particle. Based on these results, it seemed that we would need to redesign the frame in order to move forward.

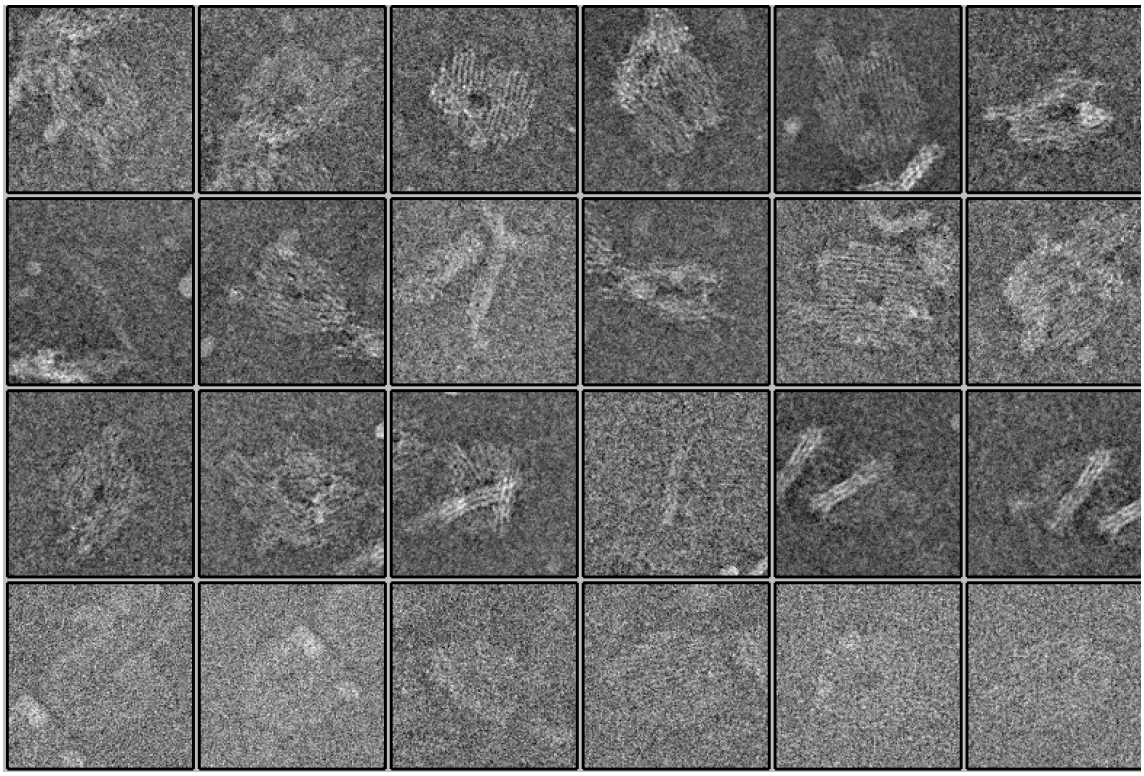

Figure 31: Class averages of cryoEM dataset of frame with loaded riboswitch.

An alternative to using the frame is to image the RNA directly. West Campus had just bought and installed a Krios with a K2 camera, which I used to collect the images in Figure 31. I chose to look at two RNAs by negative stain to assess their potential suitability for cryoEM. The first was the Xist structure region 8 (Figure 32a). However, based on the extended (squiggled) nature of the particles, it appears to be unstructured, in contrast to predictions. I also examined the c-di-GMP riboswitch, which does appear to be well folded by negative stain (Figure 32b). An occasional artefact with negative stain

of RNA that the particles becomes counterstained (is darker than the background) due to the high charge of the molecule. From this image, I concluded that the c-di-GMP riboswitch is well folded. I moved on to screening cryoEM grids on a FEI Talos L120C. The grids needed extensive freezing optimization, and generally showed poor consistency and reproducibility. However, in spots where I could find appropriate ice, there did appear to be some specks that could represent folded riboswitch (Figure 32c).

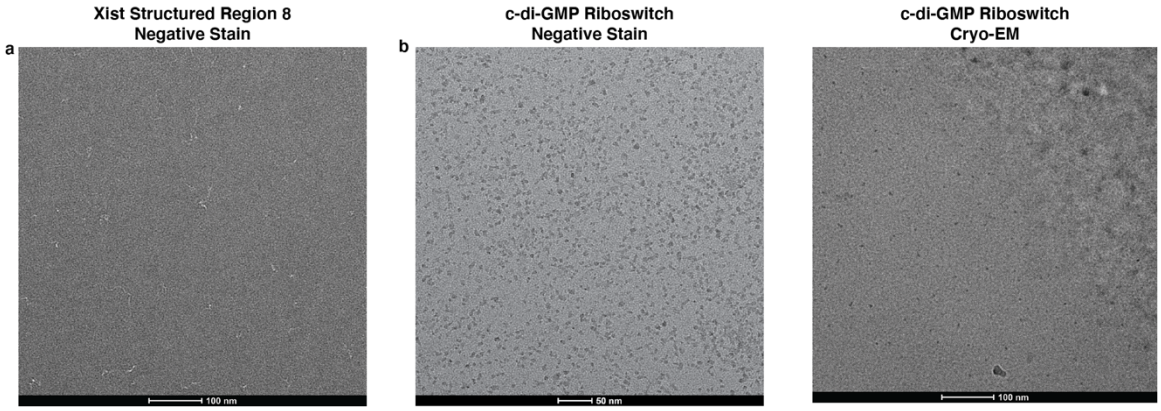

Figure 32: Electron microscopy of small RNAs

a, Negative stain images of Xist structured region 8. This RNA appears to be unstructured (light, extended features). **b,** Negative stain of c-di-GMP riboswitch, where the dark spots are thought to be the riboswitch. **c,** CryoEM image of the c-di-GMP riboswitch. Mottled area in the top right is where thick ice gives way to clean ice below and to the left. Dark spots could be folded c-di-GMP riboswitch.

Conclusions and Outlook

My progression through this project mirrors the progression of the cryoEM field. When I started, biochemical strategies to make particles more amenable for single-particle reconstruction seemed like the best strategy for working within the technology at the time. With the rapid advances in imaging and processing technology, the field was becoming increasingly ambitious with the size of reconstructed particles, in a race to atomic resolution. At the same time, many labs were discovering the challenges of

working with DNA origami. During my time at the Cold Sping Harbor Course on Electron Microscopy and the Three Dimensional Electron Microscopy Gordon Conference, I became more familiar with much of the unpublished research on DNA origami frames. Through my increasing knowledge of the cryoEM field and it's technological advance and my increasing familiarity with the DNA origami field, I realized reconstructions were going to be done on single RNA particles, without needing a frame.

Since I put this project down, several beautiful high-resolution structures of small RNAs have been published[126,127]. With better and better instrumentation, this should become a routine procedure in biochemistry[128,129].

3.4. hnRNP-U and chromatin-retained RNAs

hnRNP-U, also known as SAF-A, has a SAF DNA binding domain, an RGG RNA binding domain, and an AAA+ ATPase domain [130,131]. It has been implicated in a myriad of nuclear processes, including transcriptional and splicing regulation, RNA stability and chromatin structure [132-135], yet there is little consensus about the function or mechanism of hnRNP-U in these processes. The most evidence currently is for hnRNP-U acting as part of a matrix that anchors some RNAs to chromatin.

Early *in vitro* biochemical assays demonstrated that hnRNP-U could bind both DNA and RNA in the test tube [130,136,137]. In 2003, co-imaging of hnRNP-U and the X-chromosome, using DNA FISH, showed colocalization of hnRNP-U with heterochromatin X-chromosome territories. The association disappeared when hnRNP-U was mutated to lack its RGG domains. This experiment relied on overexpression of either

hnRNP-U or the mutant over endogenous hnRNP-U, and it was not possible to learn anything how this might affect Xist localization.

However, Xist localization to the X-chromosome is necessary for X-chromosome inactivation, and understanding the molecular interaction between RNA and chromatin is crucial for understand the epigenetic changes that take place during XCI. More detailed studies of XCI depleted hnRNP-U and saw that Xist is delocalized from the X-chromosome without hnRNP-U, and that this phenotype is dependent on both the SAF and RGG domains[104]. This finding is actively debated in the literature. Other labs have found that hnRNP-U is necessary for localization only in some cell types, rarely in primary cells, and find the RGG domain is not necessary for this function[138]. One confounding factor is that hnRNP-U has two poorly characterized homologs, hnRNP U-like1 and hnRNU U-like2 with different expression levels across cell types. These could be partially rescuing the hnRNP-U knock down. In addition, the RGG domain is not really a structure domain. It is a relatively disordered protein region enriched in argenine and glycine residues. Difference is the extent of RGG deletion could have variable effects on how hnRNP-U interacts with Xist[139]. In addition to Xist, hnRNP-U seem to be necessary to anchor CoT1 repetitive RNAs to the chromosome from which they are transcribed, similar to Xist[133].

One potential model for how hnRNP-U could affect chromatin structure integrates its SAF, RGG, and AAA+ ATPase domains. It proposes that hnRNP-U binds chromatin-associated RNAs, often nascent RNA, with its RGG domain. This activates an RNA-dependent AAA+ ATPase domain and induced oligomerization. Oligomers of hnRNP-U could regulate large-scale chromatin structures, potentially keeping areas of active

transcription euchromatic [131]. This model would need to be reconciled with the potentially conflicting ideas of how hnRNP-U functions in XCI.

These microscopy studies have been able to characterize hnRNP-U in relation to one RNA or genomic location at a time. I conducted genome-wide sequencing studies to ask what RNAs require hnRNP-U for their chromatin localization? What would be the characteristics of these RNAs? Are they long-lived, like Xist?

Results

I optimized an experiment to study RNA dynamics on chromatin. This experiment combined a chromatin fractionation technique[6] with TimeLapse chemistry developed in the Simon Lab (Schofield et al 2018). Briefly, cells were metabolically labeled with s^4U for 1 hour, after which I isolated chromatin-associated RNA. TimeLapse chemistry converted the s^4U to a C analogue, so the presence of s^4U in a read can be is evident in sequencing data as a U-to-C mutation. Mutational content was used to infer the fraction of new RNA for any gene.

To validate the quality of the chromatin fractionation, I performed a Western Blot of proteins characteristic of each fraction. In the cytoplasm I looked for actin, in the nucleus I looked for U1-70k, and in the chromatin I looked for histone H3. There is some U1-70k in the cytoplasmic fraction, indicating that there was likely a few nuclei that lysed during the first lysis step. However, the chromatin fraction is free of contamination. Additionally, these Western blots confirm that hnRNP-U is knocked-out in this cell line. Interestingly, the bulk of hnRNP-U is in the nucleoplasm, but there is a proportion that make up a small band in the chromatin fraction.

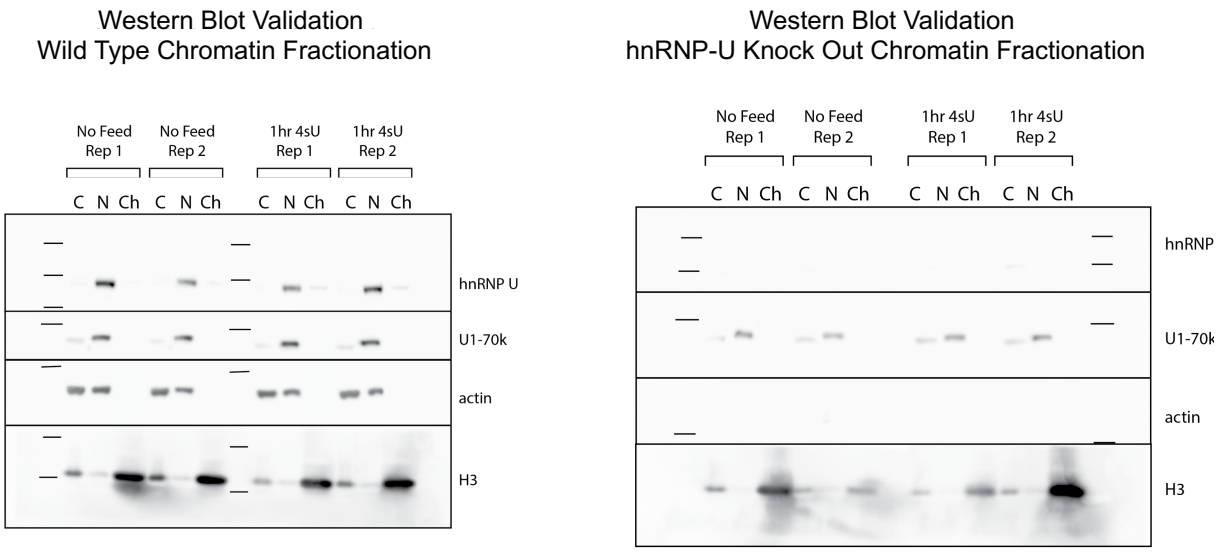

Figure 33: Chromatin fractionation validation by Western blot

Right: Western blot to confirm fractionation in wild type cells. *Left*: Western blot confirming fractionation in hnRNP-U knock out cells.

I also verified that the TimeLapse chemistry worked well by analyzing the presence of T-to-C mutations in the datasets. Figure 34 shows the proportion of reads that have mutations in red, while the proportion without mutations is in gray. Increasing shades of red make up increasing number of T-to-C mutations per read. The actin transcripts are slow turnover, meaning they are made slowly and degraded slowly. We expect that this population of RNAs will include relatively few reads that were labeled during a one-hour period. On the other hand, both Jun and Myc are high turnover, meaning they are synthesized and degraded quickly. They have much higher rates of s^4U incorporation, as indicated by the increase in proportion of the bar colored red. The chromatin fractions have high mutation content than the inputs, which makes sense since chromatin is where new RNAs containing s^4U are made.

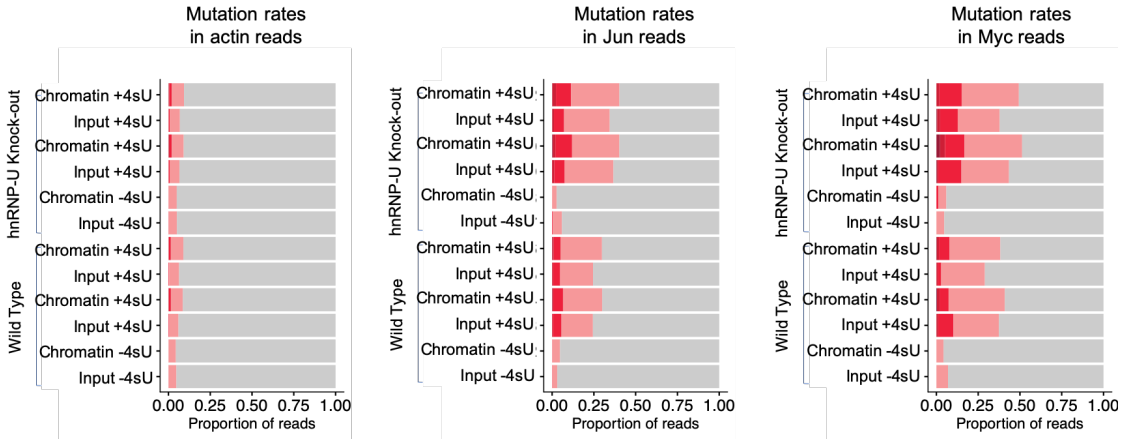

Figure 34: Analysis of T-to-C mutation content in fractionated samples

Extent of the red bars indicates the proportion of reads that have T-to-C mutations. Shown are input and chromatin fractions for all of the sequenced samples. The top six rows are the hnRNP-U knock out cells. The bottom six are wild type cells.

I performed several analyses on the data. First, I used the differential expression software DESeq2 to look for significant changes in gene expression between the total RNA in WT and knock cells. This analysis revealed thousands of genes affected by hnRNP-U knock out (Figure 35). Gene ontology analysis results indicated that these transcripts are predominantly involved in growth, intracellular signal transduction, and signaling pathways.

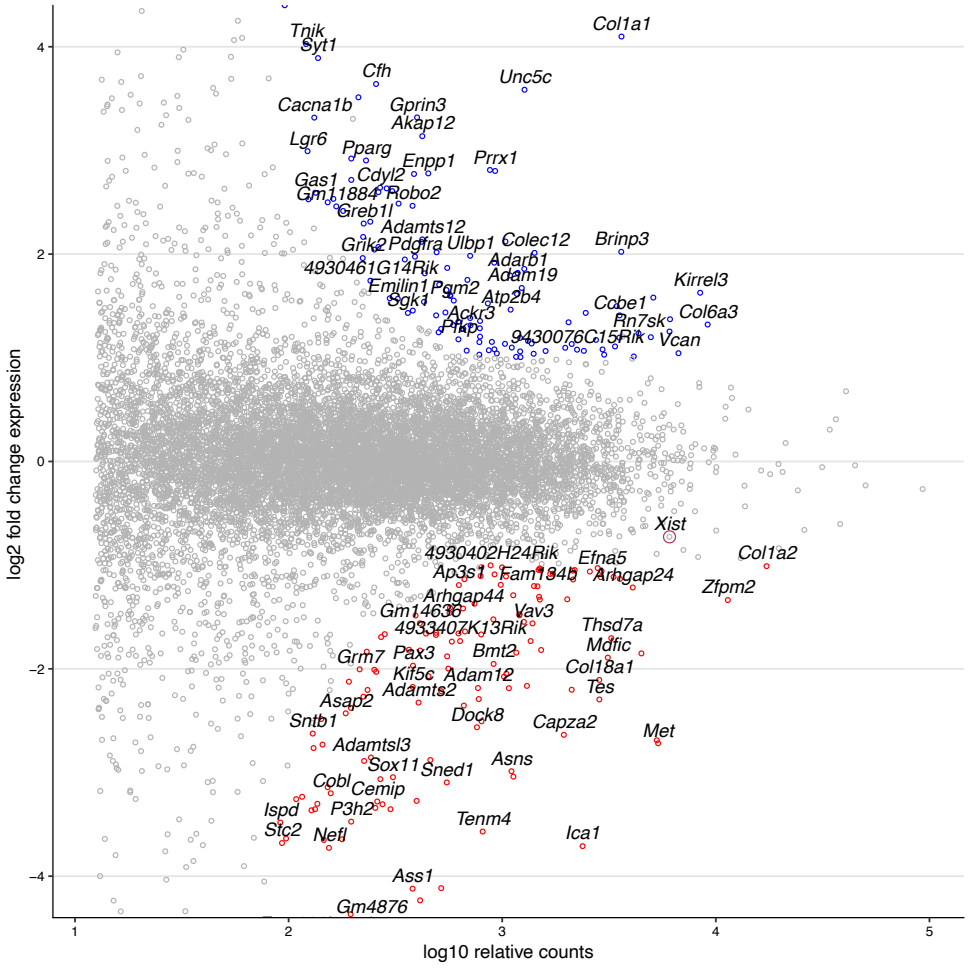

Figure 35: Thousands of transcripts are dysregulated upon hnRNP-U knock out

I then looked for what RNAs are chromatin enriched in WT cells. To distinguish between active transcription and retained, long-lived, RNAs enriched on chromatin, I used TimeLapse chemistry to identify transcripts with a longer half-life on chromatin. However, the list is not made up of lncRNAs, as I had hypothesized. I mostly find transcripts coding for proteins involved in the cytoskeleton, extracellular matrix, and development.

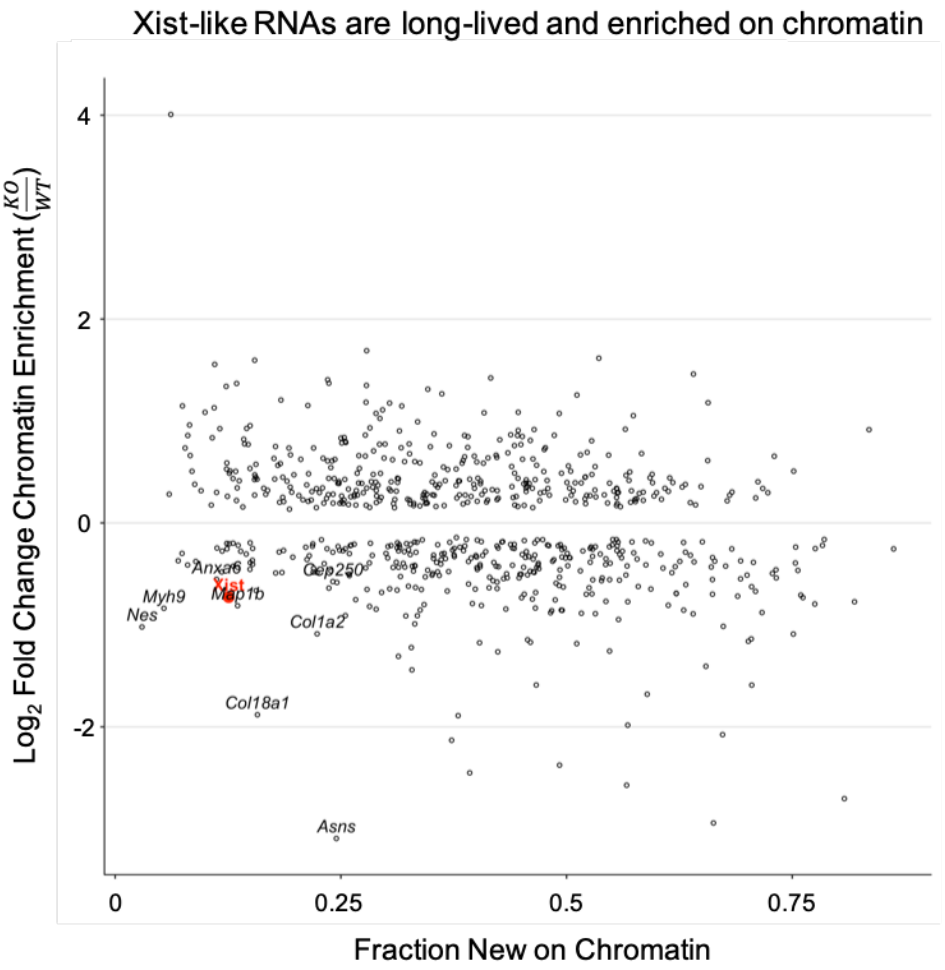

Figure 36: Chromatin retained, long-lived RNAs on chromatin are associated with cell strucure.

Finally, I examined which chromatin-enriched RNAs change between WT and KO cells. Changes in active transcription occur in transcription factors and RNA splicing regulators. Xist and other similar RNAs are more likely to be newly transcribed in hnRNP-U knock out cells. This indicates that they potentially are more highly transcribed, or more highly turned over. This result indicates that these transcripts, which are involved in mRNA splicing, transcription regulation, cellular differentiation, and the cell cycle, are more highly turned over in the hnRNP-U knock out cells.

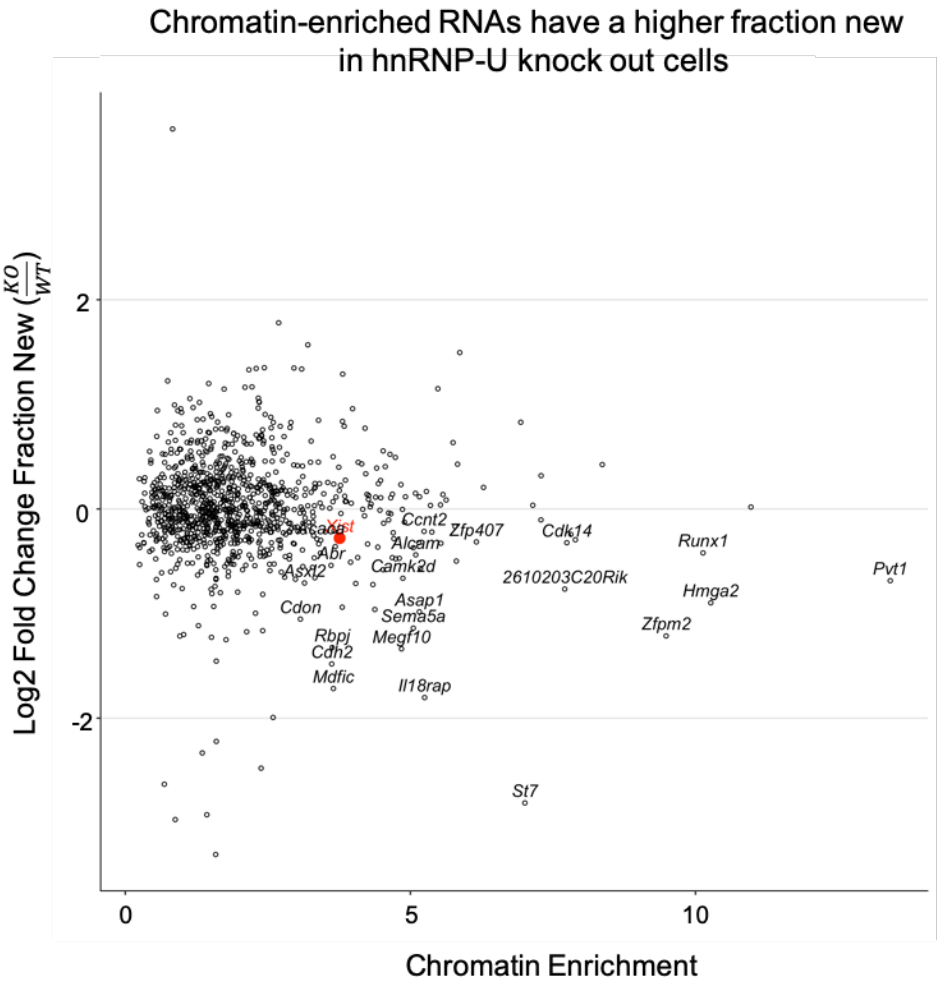

Figure 37: Chromatin-retained RNAs in hnRNP-U knock-out cells have shorter half-lives

Conclusion and outlook on chromatin-retained RNAs

I conclude that hnRNP-U does not seem to be involved in anchoring long non-coding RNAs to the chromatin, based on this experiment in C2C12 MEF cells. Both hnRNP U-like1 and hnRNP U-like2 are expressed in these cells, and these proteins could be compensating for the hnRNP U knock out. Future experiments should start by knocking down these homologs and repeating the experiment. Additionally, changes

should be included in the experimental design. The cytoskeletal transcripts that I'm finding could be contamination. The experiment should be repeated with polyA+ depletion or using TILAC to control for this. The one potentially interesting results is that RNAs on chromatin appear to be more likely to be newly made in hnRNP U knockout cells than in wild type cells. This could mean that hnRNP U is having an effect on RNA stability independent of any potential role in tethering RNA.

Chapter 4. Methods

4.1. TILAC Experimental Procedures

Cell culture

Drosophila S2 cells (gift from H. Lin, Yale) were cultured in Schneider medium (Lonza) supplemented with 10% heat-inactivated FBS (Invitrogen) and 1% penicillin–streptomycin (Millipore) and maintained at 27 °C. Cultures were split every 3 d to a concentration of 5×10^5 cells per ml.

MEF and HEK293t cells were grown at 37°C in DMEM – High Glucose media (Invitrogen) supplemented with 10% FBS (Invitrogen) and 1% penicillin–streptomycin (Millipore). Cells were split when they reached 70% confluence.

Regular mycoplasma tests (ATCC) were conducted to check quality for contamination.

Plating cells for TILAC experiments

To properly compare, the mixture needs to include the same number of cells from each labeled condition. Counting cells is extremely time consuming, especially when working with large numbers of samples (up to 18), and counting by eye has an average error of 20%[140]. To get as close to the same number of cells as possible, cells are split in a specific way. Cells for the experiment are only ever split 1:2 and plated 24 hours before it starts. This gives cells time to adhere, but is short enough to minimize any plate-to-plate variability in growth. To split, cells are harvested by trypsin, and all plates are mixed into a 50mL conical, spun down, and resuspended in media. Cells from this common stock are

distributed to the required number of experimental plates. The conical is capped and inverted every 3-4 plates to keep cells in a homogenous suspension.

Dot blots of nucleotide incorporation

RNA was labeled, isolated from TRIzol, and DNased according to protocol (pg 88). RNA (2-5ug) was reacted with 10% MTS-TAMRA (dissolved in dry DMF) by weight in 20m HEPES (pH7.4) and 1mM EDTA to a final volume of 50uL. The reaction was incubated for 1.5 hours at RT with rotation in the dark. The reaction was cleaned up by adding 50uL DEPC and using 100uL of 24:1 chloroform:isoamyl alcohol for one extraction. The aqueous phase was combined with 350uL RLT (from Quiagen RNease Mini Kit) and mixed thoroughly. Then 250uL 100% ethanol was added and mixed thoroughly. The solution was spun through an RNeasy column at 12,000xg for 30 seconds and flow-through. The column was washed with 500uL RPE (from RNeasy Mini Kit), then 500uL 80% ethanol, and finally dried for 2 min at max speed. RNA was eluted into 50uL DEPC and imaged by pipetting 5uL onto the glass of a Typhoon imager, using the TAMRA setting[51,53].

Heat shock

Drosophila S2 cells were grown in 6-well plates as described above (pg 80). To perform the heat shock experiment, cells were incubated at 37 °C for 1 hour. Cells were fed with 100uM nucleotide analogue for the last 45 minutes of heat shock. To harvest, cells were scraped into individual tubes if controls, or mixed if they were TILAC samples, and immediately placed on ice. Cells were spun down at 500x g at 4 °C, washed 1x with ice-cold PBS, and resuspended in 500uL of Trizol[92].

Puromycin-induced ribosome dissociation

HEK293t cells were grown to ~60% confluency in DMEM with 10% FBS and Penn/Strep in 15cm plates. Cells were fed with 100uM s^4U or s^6G for 4 hours. Plates were washed in ice-cold PBS and each plate was scraped into its own lo-bind tube. Control samples were lysed in cycloheximide lysis buffer (20mM Tris-HCl pH 7.5, 10mM $MgCl_2$, 200mM KCl, 1% Triton, 0.2mg/mL cycloheximide, 4mM EDTA) and passed 10x through a 26-guage needle to shear genomic DNA. Lysate was cleared at 20,000xg for 10 min at 4°C. Puromycin treated samples were resuspended in puromycin lysis buffer (20mM Tris-HCl pH 7.5, 5 mM $MgCl_2$, 200mM KCl, 1% Triton, 4mM EDTA), passed through a 26-guage needle, and cleared at 20,000xg for 10 min at 4°C. Puromycin was added to 2mM, and incubated on ice for 20 min, and then at 37°C for 20 min. $MgCl_2$ was added up to 10mM[8,72,77].

After puromycin treatment, samples were mixed equally to a final volume of 1mL. This was centrifuged through a 10-50% sucrose gradient[7]. Fractions were collected into phenol:chloroform:isoamyl alcohol. In total, two phenol extractions were performed, and then one additional chloroform extraction. RNA was ethanol precipitated, and DNA removed with TurboDNase. Finally, RNA was TimeLapse treated and libraries were prepared using the Clontech SMARTer Stranded Total RNA-seq v2 library prep kit.

Transcription inhibition with flavopiridol

Cells were grown in 6-wells plates to ~80% confluency. They were treated with 500nM flavopiridol and 100uM of either s^4U or s^6G for 2 hours. They were immediately washed in cold PBS and scraped into pre-chilled lo-bind Eppendorf tubes. TILAC samples were mixed during this step. Samples were pelleted at 1800xg at 4 °C and

resuspended in 500uL of Trizol. Protocol was developed in the Simon Lab by Joshua Zimmer (in submission, 2021).

Puromycin incorporation to examine active translation

Cells were split into 10cm plates so that they were 50-70% confluent at the start of the experiment. Cells were treated with 100uM sodium arsenite for the indicated time, or water control, by adding straight into the plate's media. For recovery experiments, media with sodium arsenite was replaced with fresh media at 37°C. During the last 15 minutes of treatment, cells were fed 7uM puromycin. Cells were washed with ice-cold PBS and harvested by scraping. Cells were pelleted at 750xg at 4°C for 5 minutes and washed once more with ice-cold PBS, and then flash frozen for storange at -20°C overnight.

Pellets were thawed and lysed in 100uL RIPA buffer, plus protease inhibitor (Roche Complete, EDTA-free), passing 10x through a 26 gauge needle. Protein concentration was determined using a BCA assay (ThermoFisher Pierce BCA Protein Assay Kit). Between 9 and 15 ug of protein was loaded onto a NuPAGE Novex 4-12% BisTris gel and run for 50 minutes at 200V in MOPS buffer, then transferred onto a PVDF membrane. Loading was assessed by Ponceau S staining, then destained with 1x TBST. Membrane was blocked in 5% milk/1x TBST for 1 hour at room temperature. Membrane was stained with primary antibody (Kerafast anti-puromycin, 3RH11, 1:1000) overnight at 4°C in 1% milk/1xTBST. The next day, the membrane was washed 3x 5min in 1xTBST. It was then stained in secondary antibody (goat anti-mouse peroxidase, Sigma-Aldrich A9917, 1:2000) at RT for 1 hour in 1% milk/1x TBST. It was washed 3x 5 min in 1x TBST and developed with SuperSignal West Femto Maximum Sensitivity

Substrate (ThermoFisher Scientific). It was visualized using chemiluminescent setting for between 5 and 10 minutes on an LAS 4000.

Studying translation during stress

293t cells were grown in 15 cm plates according to cell culture procedure (pg 80). Cells were fed with 100uM nucleotide from 4 hours before starting the sodium arsenite portion of the experiment. For cells undergoing stress treatment, they were treated with 100uM sodium arsenite for 30 minutes. For nucleotide chase experiments, the media was replaced with fresh media plus 100uM sodium arsenite[84], without nucleotide. For recovery experiments, the media was replaced with fresh media with the appropriate nucleotide feed. Puromycin treatments on stressed cells were performed as described above (pg 82).

Plates were washed in ice-cold PBS. Plates were scraped under 1mL of ice-cold PBS. Suspended cells were aliquoted into lo-bind tubes so that 500uL of each plate was mixed with it corresponding TILAC partner for ribosome isolation, 100uL was mixed with its corresponding TILAC partner for input sequencing analysis, and the final 400uL was saved separately in case it might be needed in the future. Samples for ribosome purification were lysed in cycloheximide lysis buffer (20mM Tris-HCl pH 7.5, 10mM $MgCl_2$, 200mM KCl, 1% Triton, 0.2mg/mL cycloheximide, 4mM EDTA) and passed 10x through a 26-guage needle to shear genomic DNA. Lysate was cleared at 20,000xg for 10 min at 4°C, and then flash frozen for transportation to collaborator for sucrose sedimentation. Samples for input sequencing were resuspended in TRIzol and downstream processing according to pg 88. The extra 400uL was also lysed in cycloheximide lysis buffer and clarified, then flash frozen.

Sucrose sedimentation

Sucrose sedimentation was performed by Rachel Neiderer in the Gilbert lab. Briefly, lysate was layered onto a 10%-50% (w/v) sucrose gradient (20mM HEPES pH 7.6, 100mM KCl, 5mM $MgCl_2$, 1mM DTT, and 100ug/mL cycloheximide). Sedimentaiton gradients were centrifuged for 2 hours at 36,000 RPM.

4.2. Cryo electron microscopy methods

***In vitro* RNA transcription**

To *in vitro* transcribe the c-di-GMP riboswitch, a 1mL reaction was set up with 75 mM Tris-HCl (pH7.5), 40mM $MgCl_2$, 2mM spermidine, 5mM DTT, 5mM each of ATP, UTP, CTP, GTP, 150pmol of DNA template, and 40 units of T7 RNA polymerase. In addition, 5uL of SuperasIN (RNase Inhibitor, Invitrogen). Reactions were incubated for 4-16 hours at 37°C. After that time, 20uL of Turbo DNase is added to a 1mL reaction, and it is incubated for another 1 hour at 37°C. Then, 20 uL of ProK is added and the reaction is incubated for another 1 hour at 37°C.

To isolate RNA, the reaction is run on a 8% denaturing polyacrylamide gel. The RNA band is visualized by UV shadowing and cut out, crushed, and covered with 300mM sodium acetate, andn incubated overnight at 4°C or at room temperature for 1 hour. The RNA is then ethanol precipitate and resuspended in DEPC-treated water.

Folding the c-di-GMP riboswitch

In a 50uL reaction, 2uM riboswitch is mixed with 5uM c-di-GMP, in a buffer containing 10mM $MgCl_2$, 10mM KCl, 10mM sodium cacodylate at pH 6.8. Heat to 70 deg C for 5 minutes and slow cool to fold.

Loading c-di-GMP riboswitch into the DNA origami frame

Combine 150nM folded c-di-GMP riboswitch, 10nM DNA origami frame, in 30mM HEPES (pH 7.5), 20mM MgCl2, and 100mM KCl. Incubate overnight at room temperature.

Grid preparation, imaging and analysis

DNA origami frame loaded with c-di-GMP riboswitch was frozen at a concentration of 10nM. C-Flat Holey Carbon Grids (CF-2/2-3C-Thick, 300 mesh, Copper) were coated with an amorphous carbon film. Before freezing, grids were glow-discharged for 30 seconds at 30mA. Freezing was performed on an FEI VitroBot with 3uL of 10nM DNA origami sample at 100% humidity for 5 seconds of blotting, with not blot offset. Negative stain images were taken on either a FEI Technica T-12 or on a FEI Talos L120C. Cryo images were take on the FEI TEchnica T-12, the FEI Talos L120C, and the Krios. Image processing and classification were performed using Relion[124].

4.1. hnRNP-U and chromatin associated RNAs

Metabolic labeling and cell culture

For metabolic labeling, 10cm plates of MEF cells were fed with s^4U for 1 hour. To evaluate the role of hnRNP-U knock out on chromatin-retained RNAs, WT MEFs were compared to CRISPR KO lines generated by Alec Sexton.

Chromatin fractionation protocol

To harvest, plates were washed 1x with ice cold PBS, then scraped under 1mL of ice cold PBS into Lo-bind microcentrifuge tubes. 10% of the suspension was taken as input. Cells were collected by centrifugation and washed with 1X PBS/1mM EDTA. Cells were resuspended in 200uL of ice-cold NP-40 lysis buffer (10mM Tris-HCl [pH7.5], 0.2% NP-40, 150mM NaCl, 1mM DTT), flicked to resuspend into a homogenous suspension, and left on ice for 5 minutes to lyse the plasma membranes. Lysate was aspirated with a P200 tip, but with the tip cut off to make a bigger opening, and avoid lysing the nuclei. It was layered on top of 500uL of sucrose cushion (24% sucrose in NP-40 lysis buffer) and centrifuged for 10 minutes at 4°C and 14,000xg. After centrifuging, there should be an opaque/white band of cytoplasmic content at the top of the supernatant. The supernatant was collected and saved as the cytoplasmic fraction.

Nuclei were washed with ice-cold 1xPBS/1mM EDTA, and resuspended in 100uL pre-chilled glycerol buffer (20 mM Trips-HCl [pH7.9], 75mM NaCl, 0.5mM EDTA, 1 mM DTT, 0.125mM PMSF, 50% glycerol) by flicking. An equal volume of ice-cold nuclei lysis buffer (10mM HEPES [pH 7.6], 1mM DTT, 7.5mM $MgCl_2$, 0.2mM EDTA, 0.3 M NaCl, 1M UREA, 1% NP-40) was added and tubes were vortexed for 2 x 2 sec, incubated for 2 min on ice, and then centrifuged for 2 min at 4°C and 14,000xg. The supernatant was collected as the nuclear fraction. Chromatin was gently rinsed with 1xPBS/1mM EDTA and resuspended into TRIzol, according the the TimeLapse RNA extraction protocol.

Western blot validation

To validate clean chromatin, each fraction was validated by Western Blot. Cytoplasm was marked with actin (sc-47778, 1:500), nucleoplasm by U1-70k (05-1588,

1:500) and chromatin by histone H3 (ab1791, 1:2000). HnRNP-U knock out was validated using the antibody sc-32315. Western blots run as described in puromycin incorporation protocol.

4.2. RNA Sequencing and Analysis

RNA sequencing

RNA extraction and TimeLapse chemistry were peformed as previously reported, with slight variations[52,53], described below.

RNA was isolated from TRIzol, and precipitated using isopropanol supplemented with 1mM DTT to prevent oxidation of the thiolated-bases. DNA was removed using TurboDNase, and RNA was purified using Agencour RNAClean XP beads. TimeLapse chemistry was performed by mixing RNA with TFEA (600 mM), EDTA (1 mM) and sodium acetate (pH 5.2, 100 mM) in water. A solution of $NaIO_4$ (10 mM) was then added dropwise, and the reaction mixture was incubated for 1.5 hours at 50°C. RNA was isolated using RNAClean beads. RNA then went through reducing treatment to remove any excess oxidant (100uM DTT, 100uM Tris pH 7.4, 10uM EDTA, 1M NaCl) and was cleaned up using RNAClean beads. The modifications to the protocol will be reported by Kiefer, Zimmer, and Schofield, manuscripts in preparation.

Libraries were prepared using the Clontech SMARTer Stranded Total RNA-seq v2 library prep kit. Sequencing was performed on a NovaSeq using paired-end 100bp reads.

Sequencing Analysis:

Reads were filtered for unqiue reads using FastUniq[141], and adapators were removed using Cutadapt[142]. Sequencing samples were aligned to both the genome and transcriptome annotations using HISAT2[143] using default parameters and -mp4,2. Human

samples were aligned to GRCh38 genome, while Drosophila reads were aligned to the Dm6 genome. Reads were further processed with Picard tools (http://broadinstitute.github.io/picard/) including FixMateInformation, SortSam, and BuildBamIndex. Reads were filtered using samtools[144] to retain only those that mapped uniquely (flag: 83/163, 99/147), with MAPQ >= 2. Reads over genes were counted using HTSeq. The number of T's, G's, and associated mutations in each read were counted using Rsamtools (http://bioconductor.org/packages/release/bioc/html/Rsamtools.html) and a custom R script. Snp's were identified using bcftools[145] and samtools mpileup, and then filtered out of mutational analysis. Tracks were made using the STAR aligner[146] (inputAlignmentsFromBam mode, outWigType bedGraph). Tracks were converted to binary format (toTDF, IGVtools) and viewed in IGV[147].

Downstream data analysis was performed using custom R-scripts.

References:

1. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biology* 1–19 (2016). doi:10.1186/s13059-016-0881-8
2. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat Rev Genet* 1–26 (2019). doi:10.1038/s41576-019-0150-2
3. Hendrickson, D., Kelley, D. R., Tenen, D., Bernstein, B. & Rinn, J. L. Widespread RNA binding by chromatin- associated proteins. *Genome Biology* 1–18 (2016). doi:10.1186/s13059-016-0878-3
4. Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Meth* **13,** 508–514 (2016).
5. Baghirova, S., Hughes, B. G., Hendzel, M. J. & Schulz, R. Sequential fractionation and isolation of subcellular proteins from tissue or cultured cells. *MethodsX* **2,** 440–445 (2015).
6. Pandya-Jones, A. & Black, D. L. Co-transcriptional splicing of constitutive and alternative exons. *RNA* **15,** 1896–1908 (2009).
7. Floor, S. N. & Doudna, J. A. Tunable protein synthesis by transcript isoforms in human cells. *eLife* **5,** 1–25 (2016).
8. Clark, I. E., Wyckoff, D. & Gavis, E. R. Synthesis of the posterior determinant Nanos is spatially restricted by a novel cotranslational regulatory mechanism. *Current Biology* 1–4 (2000).
9. Wilhelm, B. T. & Landry, J.-R. RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48,** 249–257 (2009).
10. O'Neil, D., Glowatz, H. & Schlumpberger, M. Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity. *Current Protocols in Molecular Biology* **103,** 4.19.1–4.19.8 (2013).
11. Zhao, S., Zhang, Y., Gamini, R., Zhang, B. & Schack, von, D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Scientific Reports* **8,** 4781–12 (2018).
12. Zhao, W. *et al.* Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* **15,** 1–11 (2014).
13. Herbert, Z. T. *et al.* Cross-site comparison of ribosomal depletion kits for Illumina RNAseq library construction. *BMC Genomics* **19,** 199–10 (2018).
14. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Meth* **5,** 613–619 (2008).
15. Soukup, G. A. & Breaker, R. R. Relationship between internucleotide linkage geometry and the stability of RNA. *RNA* **5,** 1308–1325 (1999).
16. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research* **38,** e131–e131 (2010).

17. Li, X., Nair, A., Wang, S. & Wang, L. in *RNA Bioinformatics* **1269,** 137–146 (Humana Press, New York, NY, 2015).
18. Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Meth* **7,** 709–715 (2010).
19. Bansal, V. A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. *BMC Bioinformatics* **18,** 43–123 (2017).
20. Mili, S. & Steitz, J. A. Evidence for reassociation of RNA-binding proteins after cell lysis: Implications for the interpretation of immunoprecipitation analyses. *RNA* **10,** 1692–1694 (2004).
21. Riley, K. J., Yario, T. A. & Steitz, J. A. Association of Argonaute proteins and microRNAs can occur after cell lysis. *RNA* **18,** 1581–1585 (2012).
22. Wheeler, E. C., Van Nostrand, E. L. & Yeo, G. W. Advances and challenges in the detection of transcriptome-wide protein-RNA interactions. *WIREs RNA* **9,** e1436–11 (2017).
23. Nicholson, C. O., Friedersdorf, M. & Keene, J. D. Quantifying RNA binding sites transcriptome-wide using DO-RIP-seq. *RNA* **23,** 32–46 (2016).
24. Hafner, M. *et al.* Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell* **141,** 129–141 (2010).
25. Ule, J. *et al.* CLIP Identifies Nova-Regulated RNA Networks in the Brain. 1–5 (2003).
26. Holden, P. & Horton, W. A. Crude subcellular fractionation of cultured mammalian cell lines. *BMC Res Notes* **2,** 243–10 (2009).
27. Oesterreich, F. C., Preibisch, S. & Neugebauer, K. M. Global Analysis of Nascent RNA Reveals Transcriptional Pausing in Terminal Exons. *Molecular Cell* **40,** 571–581 (2010).
28. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 31–21 (2014).
29. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11,** 94–13 (2010).
30. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. in *Statistical Analysis of Next Generation Sequencing Data* **11,** 169–190 (Springer, Cham, 2014).
31. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2009).
32. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Nat Prec* 1–1 (2010). doi:10.1038/npre.2010.4282.1
33. Lin, C. Y. *et al.* Transcriptional Amplification in Tumor Cells with Elevated c-Myc. *Cell* **151,** 56–67 (2012).
34. Nie, Z. *et al.* c-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells. *Cell* **151,** 68–79 (2012).
35. Chao, S. H. & Price, D. H. Flavopiridol inactivates P-TEFb and blocks most RNA polymerase II transcription in vivo. *Journal of Biological Chemistry* **276,** 31793–31799 (2001).

36. Yu, H. *et al.* Normalization of human RNA-seq experiments using chimpanzee RNA as a spike-in standard. *Scientific Reports* 1–10 (2016). doi:10.1038/srep31923
37. Hardwick, S. A. *et al.* Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat Meth* **13,** 792–798 (2016).
38. The External RNA Controls Consortium. The External RNA Controls Consortium: a progress report. *Nat Meth* **2,** 731–734 (2005).
39. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Research* **21,** 1543–1551 (2011).
40. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology* **32,** 896–902 (2014).
41. Ong, S.-E., Mann, M., Blagoev, B. & Kratchmarova, I. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics*. 1–11 (2002).
42. Deng, J., Erdjument-Bromage, H. & Neubert, T. A. Quantitative Comparison of Proteomes Using SILAC. *Current Protocols in Protein Science* **95,** e74–14 (2018).
43. Ong, S.-E. & Mann, M. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat Protoc* **1,** 2650–2660 (2007).
44. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26,** 1367–1372 (2008).
45. Cox, J. *et al.* A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nat Protoc* **4,** 698–705 (2009).
46. Blagoev, B. *et al.* A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nature Biotechnology* **21,** 315–318 (2003).
47. Boldt, K., Gloeckner, C. J., Texier, Y., Zweydorf, von, F. & Ueffing, M. in *Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC)* **1188,** 177–190 (Humana Press, New York, NY, 2014).
48. Zhang, G. *et al.* In-Depth Quantitative Proteomic Analysis of de Novo Protein Synthesis Induced by Brain-Derived Neurotrophic Factor. *J. Proteome Res.* **13,** 5707–5714 (2014).
49. Zhang, G., Deinhardt, K., Chao, M. V. & Neubert, T. A. Study of Neurotrophin-3 Signaling in Primary Cultured Neurons using Multiplex Stable Isotope Labeling with Amino Acids in Cell Culture. *J. Proteome Res.* **10,** 2546–2554 (2011).
50. Rabani, M. *et al.* Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature Biotechnology* **29,** 436–442 (2011).
51. Duffy, E. E. *et al.* Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry. *Molecular Cell* **59,** 858–866 (2015).
52. Schofield, J. A., Duffy, E. E., Kiefer, L., Sullivan, M. C. & Simon, M. D. TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nature Publishing Group* **15,** 221–225 (2018).

53. Kiefer, L., Schofield, J. A. & Simon, M. D. Expanding the Nucleoside Recoding Toolkit: Revealing RNA Population Dynamics with 6-Thioguanosine. *J. Am. Chem. Soc.* **140,** 14567–14570 (2018).
54. Herzog, V. A. *et al.* Thiol-linked alkylation of RNA to assess expression dynamics. *Nat Meth* **14,** 1198–1204 (2017).
55. Moss, T., Langlois, F., Gagnon-Kugler, T. & Stefanovsky, V. A housekeeper with power of attorney: the rRNA genes in ribosome biogenesis. *Cell. Mol. Life Sci.* **64,** 29–49 (2007).
56. Dieci, G., Conti, A., Pagano, A. & Carnevali, D. Identification of RNA polymerase III-transcribed genes in eukaryotic genomes. *BBA - Gene Regulatory Mechanisms* **1829,** 296–305 (2013).
57. Vannini, A. & Cramer, P. Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Molecular Cell* **45,** 439–446 (2012).
58. Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* 1–12 (2012). doi:10.1038/nrg3293
59. Price, D. H. P-TEFb, a Cyclin-Dependent Kinase Controlling Elongation by RNA Polymerase II. *Molecular and Cellular Biology* **20,** 2629–2634 (2000).
60. Xiao, H. & Lis, J. T. Germline transformation used to define key features of heat-shock response elements. *Science* **239,** 1139–1142 (1988).
61. Guertin, M. J., Petesch, S. J., Zobeck, K. L., Min, I. M. & Lis, J. T. Drosophila Heat Shock System as a General Model to Investigate Transcriptional Regulation. *Cold Spring Harb Symp Quant Biol* **75,** 1–9 (2010).
62. DiDomenico, B., Bugaisky, G. & Lindquist, S. The Heat Shock Response Is Self-Regulated at Both the Transcriptional and Posttranscriptional Levels. 1–11 (2004).
63. Sorger, P. K. Heat Shock Factor and the Heat Shock Response. *Cell* **65,** 363–366 (1991).
64. O'Brien, T. & Lis, J. T. Changes in Drosophila Transcription after an Instantaneous Heat Shock. *Molecular and Cellular Biology* **13,** 3456–3463 (1993).
65. Sorensen, J., Nielsen, M. M., Kruhoffer, M., Justesen, J. & Loeschcke, V. Full genome gene expression analysis of the heat stress response in Drosophila melanogaster. **10,** 312–328 (2005).
66. Leemans, R. *et al.* Quantitative transcript imaging in normal and heat-shocked Drosophila embryos by using high-density oligonucleotide arrays. *Proc Natl Acad Sci USA* **97,** 12138–12143 (2000).
67. Duarte, F. M. *et al.* Transcription factors GAF and HSF act at distinct regulatory steps to modulate stress-induced gene activation. *Genes & Development* **30,** 1731–1746 (2016).
68. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science* **339,** 950–953 (2013).

69. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322,** 1845–1848 (2008).
70. Teves, S. S. & Henikoff, S. Heat shock reduces stalled RNA polymerase II and nucleosome turnover genome-wide. *Genes & Development* **25,** 2387–2397 (2011).
71. Spirin, A. S., Belitsina, N. V., letters, E. L. F.1972. On some artifacts of sucrose gradient sedimentation of ribosomes. *core.ac.uk*
72. Kronja, I. *et al.* Widespread Changes in the Posttranscriptional Landscape at the Drosophila Oocyte-to-Embryo Transition. 1–46 (2014). doi:10.1016/j.celrep.2014.05.002
73. Clarke, B. D., Roby, J. A., Slonchak, A. & Khromykh, A. A. Functional non-coding RNAs derived from the flavivirus 3′ untranslated region. *Virus Research* **206,** 53–61 (2015).
74. Tercero, J. A., Espinosa, J. C., Lacalle, R. A. & Jiménez, A. The Biosynthetic Pathway of the Aminonucleoside Antibiotic Puromycin, as Deduced from the Molecular Analysis of the pur Cluster of Streptomyces alboniger(âˆ—). *Journal of Biological Chemistry* **271,** 1579–1590 (1996).
75. Aviner, R. The science of puromycin: From studies of ribosome function to applications in biotechnology. *Computational and Structural Biotechnology* **18,** 1074–1083 (2020).
76. Kedersha, N. *et al.* Dynamic Shuttling of TIA-1 Accompanies the Recruitment of mRNA to Mammalian Stress Granules. 1–12 (2000).
77. Blobel, G. & Sabatini, D. Dissociation of Mammalian Polyribosomes into Subunits by Puromycin. *Proc Natl Acad Sci USA* **68,** 390–394 (1971).
78. Buchan, J. R. & Parker, R. Eukaryotic Stress Granules: The Ins and Outs of Translation. *Molecular Cell* **36,** 932–941 (2009).
79. Kedersha, N. L., Gupta, M., Li, W., Miller, I. & Anderson, P. RNA-binding Proteins TIA-1 and TIAR Link the Phosphylation of eIF-2a to the Assembly of Mammalian Stress Granules. 1–11 (1999).
80. Aulas, A. *et al.* Stress-specific differences in assembly and composition of stress granules and related foci. *Journal of Cell Science* **130,** 927–937 (2017).
81. Yoon, J.-H., Choi, E.-J. & Parker, R. Dcp2 phosphorylation by Ste20 modulates stress granule assembly and mRNA decay in Saccharomyces cerevisiae. *Journal of Cell Biology* **189,** 813–827 (2010).
82. Hilgers, V., Teixeira, D. & Parker, R. Translation-independent inhibition of mRNA deadenylation during stress in Saccharomyces cerevisiae. *RNA* **12,** 1835–1845 (2006).
83. Bley, N. *et al.* Stress granules are dispensable for mRNA stabilization during cellular stress. *Nucleic Acids Research* **43,** e26–e26 (2014).
84. Wilbertz, J. H. *et al.* Single-Molecule Imaging of mRNA Localization and Regulation during the Integrated Stress Response. *Molecular Cell* **73,** 946–958.e7 (2019).

85. Mateju, D., Eichenberger, B., Eglinger, J., Roth, G. & Chao, J. A. Single-molecule imaging reveals translation of mRNAs localized to stress granules. *bioRxiv* **6,** 43927–29 (2020).
86. Jain, S. *et al.* ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure. *Cell* **164,** 487–498 (2016).
87. Hilliker, A., Gao, Z., Jankowsky, E. & Parker, R. The DEAD-Box Protein Ded1 Modulates Translation by the Formation and Resolution of an eIF4F-mRNA Complex. *Molecular Cell* **43,** 962–972 (2011).
88. Andreev, D. E. *et al.* Translation of 5′ leaders is pervasive in genes resistant to eIF2 repression. *eLife* **4,** 1–21 (2015).
89. Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R. & Mann, M. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat Meth* **7,** 383–385 (2010).
90. Carpenter, B. *et al.* Stan: A Probabilistic Programming Language. *Grantee Submission* **76,** 1–32 (2017).
91. Jamrich, M., Greenleaf, A. L. & Bautz, E. K. Localization of RNA polymerase in polytene chromosomes of Drosophila melanogaster. *Proc Natl Acad Sci USA* **74,** 2079–2083 (1977).
92. Machyna, M., Kiefer, L. & Simon, M. D. Enhanced nucleotide chemistry and toehold nanotechnology reveals lncRNA spreading on chromatin. *Nature Structural & Molecular Biology 2010 17:7* 1–25 (2020). doi:10.1038/s41594-020-0390-z
93. Spradling, A., Penman, S. & Pardue, M. L. Analysis of drosophila mRNA by in situ hybridization: Sequences transcribed in normal and heat shocked cultured cells. *Cell* **4,** 395–404 (1975).
94. Wang, Y. J. *et al.* Lso2 is a conserved ribosome-bound protein required for translational recovery in yeast. *PLoS Biol* **16,** e2005903–39 (2018).
95. Tauber, D. *et al.* Modulation of RNA Condensation by the DEAD-Box Protein eIF4A. *Cell* 1–33 (2020). doi:10.1016/j.cell.2019.12.031
96. Pires Da Silva, J. *et al.* SIRT1 Protects the Heart from ER Stress-Induced Injury by Promoting eEF2K/eEF2-Dependent Autophagy. *Cells* **9,** 426 (2020).
97. Lu, P. D., Harding, H. P. & Ron, D. Translation reinitiation at alternative open reading frames regulates gene expression in an integrated stress response. *Journal of Cell Biology* **167,** 27–33 (2004).
98. Moon, S. L., Morisaki, T., Stasevich, T. J. & Parker, R. Coupling of translation quality control and mRNA targeting to stress granules. *Journal of Cell Biology* **219,** 803–23 (2020).
99. Alekseyenko, A. A. *et al. BioTAP-XL: Cross-linking/Tandem Affinity Purification to Study DNA Targets, RNA, and Protein Components of Chromatin-Associated Complexes*. 21.30.1–21.30.32 (John Wiley & Sons, Inc., 2001). doi:10.1002/0471142727.mb2130s109
100. Brockdorff, N. *et al.* Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. 1–3 (1991).
101. Heard, E. Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes & Development* **20,** 1848–1867 (2006).

102. Plath, K. *et al.* Developmentally regulated alterations in Polycomb repressive complex 1 proteins on the inactive X chromosome. *Journal of Cell Biology* **167,** 1025–1035 (2004).
103. Engreitz, J. M. *et al.* The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. *Science* **341,** 1237973–1237973 (2013).
104. Hasegawa, Y. *et al.* The Matrix Protein hnRNP U Is Required for Chromosomal Localization of Xist RNA. *DEVCEL* **19,** 469–476 (2010).
105. Simon, M. D. *et al.* High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature* **504,** 465–469 (2013).
106. Wutz, A., Rasmussen, T. P. & Jaenisch, R. Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat. Genet.* **30,** 167–174 (2002).
107. Chu, C. *et al.* Systematic Discovery of Xist RNA Binding Proteins. *Cell* **161,** 404–416 (2015).
108. McHugh, C. A. *et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521,** 232–236 (2015).
109. Minajigi, A. *et al.* A comprehensive Xist interactome reveals cohesin repulsion and an RNA- directed chromosome conformation. *Science* **349,** 1–14 (2015).
110. Fang, R., Moss, W. N., Rutenberg-Schoenberg, M. & Simon, M. D. Probing Xist RNA Structure in Cells Using Targeted Structure-Seq. *PLoS Genet* **11,** e1005668–29 (2015).
111. Liu, F., Somarowthu, S. & Pyle, A. M. Visualizing the secondary and tertiary architectural domains of lncRNA RepA. *Nat Chem Biol* **13,** 282–289 (2017).
112. Maenner, S. *et al.* 2-D Structure of the A Region of Xist RNA and Its Implication for PRC2 Association. *PLoS Biol* **8,** e1000276–16 (2010).
113. Smola, M. J., Rice, G. M., Busan, S., Siegfried, N. A. & Weeks, K. M. Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat Protoc* **10,** 1643–1669 (2015).
114. Novikova, I. V., Hennelly, S. P. & Sanbonmatsu, K. Y. Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic Acids Research* **40,** 5034–5051 (2012).
115. Somarowthu, S. *et al.* HOTAIR Forms an Intricate and Modular Secondary Structure. *Molecular Cell* **58,** 353–361 (2015).
116. Chigi, Y., Sasaki, H. & Sado, T. The 5' region of Xist RNA has the potential to associate with chromatin through the
A-repeat. 1–27 (2017).
117. Beletskii, A., Hong, Y.-K., Pehrson, J., Egholm, M. & Strauss, W. PNA interference mapping demonstrates functional domains in the noncoding RNA Xist. 1–6 (2001).
118. Sarma, K., Levasseur, P., Aristarkhov, A. & Lee, J. T. Locked nucleic acids (LNAs) reveal sequence requirements and kinetics of Xist RNA localization to the X chromosome. 1–8 (2010). doi:10.1073/pnas.1009785107/-/DCSupplemental/pnas.201009785SI.pdf

119. Kato, T., Goodman, R. P., Erben, C. M., Turberfield, A. J. & Namba, K. High-Resolution Structural Analysis of a DNA Nanostructure by cryoEM. *Nano Lett.* **9,** 2747–2750 (2009).
120. Khoshouei, M., Radjainia, M., Baumeister, W. & Danev, R. Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. *Nature Communications* **8,** 1–6 (2017).
121. Merk, A. *et al.* Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery. *Cell* 1–16 (2016). doi:10.1016/j.cell.2016.05.040
122. Martin, T. G. *et al.* Design of a molecular support for cryo-EM structure determination. *Proc Natl Acad Sci USA* **113,** E7456–E7463 (2016).
123. Smith, K. D. *et al.* Structural basis of ligand binding by a c-di-GMP riboswitch. *Nature Publishing Group* **16,** 1218–1223 (2009).
124. Scheres, S. H. W. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology* **180,** 519–530 (2012).
125. Fernandez-Leiro, R., Scheres, S. H. W.IUCr. A pipeline approach to single-particle processing in RELION. *Acta Crystallogr Sect D Struct Biol* **73,** 496–502 (2017).
126. Bonilla, S., Sherlock, M. E., MacFadden, A. & Kieft, J. S. A structured viral RNA uses molecular mimicry and conformational dynamics to coordinate multiple functions. *bioRxiv* **558,** 75–40 (2020).
127. Zhang, K. *et al.* Cryo-EM structure of a 40 kDa SAM-IV riboswitch RNA at 3.7Å resolution. *Nature Communications* 1–6 (2019). doi:10.1038/s41467-019-13494-7
128. Kappel, K. *et al.* Accelerated cryo-EM-guided determination of three-dimensional RNA-only structures. *Nat Meth* **17,** 699–707 (2020).
129. Wu, M. & Lander, G. C. How low can we go? Structure determination of small biological complexes using single-particle cryo-EM. *Current Opinion in Structural Biology* **64,** 9–16 (2020).
130. Fackelmayer, F., Dahm, K., Renz, A. & Richter, A. Nucleic-acid-binding properties of hnRNP-U/SAF-A, a nuclear-matrix protein which binds DNA and RNA in vivo and in vitro. *European Journal Biochemistry* 749–757 (1994).
131. Nozawa, R.-S. *et al.* SAF-A Regulates Interphase Chromosome Structure through Oligomerization with Chromatin- Associated RNAs. *Cell* **169,** 1214–1227.e18 (2017).
132. Xiao, R. *et al.* Nuclear Matrix Factor hnRNP U/SAF-A Exerts a Global Control of Alternative Splicing by Regulating U2 snRNP Maturation. *Molecular Cell* **45,** 656–668 (2012).
133. Hall, L. L. *et al.* Stable C0T-1 Repeat RNA Is Abundant and Is Associated with Euchromatic Interphase Chromosomes. *Cell* **156,** 907–919 (2014).
134. Yugami, M., Kabe, Y., Yamaguchi, Y., Wada, T. & Handa, H. hnRNP-U enhances the expression of specific genes by stabilizing mRNA. *FEBS Letters* **581,** 1–7 (2006).
135. Vizlin-Hodzic, D., Runnberg, R., Ryme, J., Simonsson, S. & Simonsson, T. SAF-A Forms a Complex with BRG1 and Both Components Are Required

for RNA Polymerase II Mediated Transcription. *PLoS ONE* **6,** e28049–9 (2011).
136. Kiledjian, M. & Dreyfuss, G. Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box. *The EMBO Journal* **11,** 2655–2664 (1992).
137. Fackelmayer, F. O. & Richter, A. Purification of Two Isoforms of hnRNP-U and Characterization of Their Nucleic Acid Binding Activity. 1–7 (1994).
138. Kolpa, H. J., Fackelmayer, F. O. & Lawrence, J. B. SAF-A Requirement in Anchoring XIST RNA to Chromatin Varies in Transformed and Primary Cells. *DEVCEL* **39,** 9–10 (2016).
139. Sakaguchi, T. *et al.* Control of Chromosomal Localization of Xist by hnRNP U Family Molecules. *DEVCEL* **39,** 11–12 (2016).
140. Ongena, K., Das, C., Smith, J. L., Gil, S. & Johnston, G. Determining Cell Number During Cell Culture using the Scepter Cell Counter. *JoVE* 1–5 (2010). doi:10.3791/2204
141. Xu, H. *et al.* FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *PLoS ONE* **7,** e52249–6 (2012).
142. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).
143. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Meth* **12,** 357–360 (2015).
144. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
145. Danecek, P. & McCarthy, S. A. BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* **33,** 2037–2039 (2017).
146. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
147. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14,** 178–192 (2013).
148. Keskin, H. *et al.* Complex effects of flavopiridol on the expression of primary response genes. *Cell Division* **7**, 11, 1-13 (2012).