New Tools for Measuring and Analyzing Bacterial Gene-Expression Dynamics

by

Mirae Leigh Parker

B.S. Physics
Stanford University 2018

Submitted to the Program of Computational and Systems Biology in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computational and Systems Biology

at the

Massachusetts Institute of Technology

February 2024

Signature of Author……
Computational and Systems Biology Graduate Program
11.30.2023

Certified by……
Gene-Wei Li
Associate Professor of Biology

Accepted by……...
Christopher Burge
Professor of Biology
Co-Director, Computational and Systems Biology Graduate Program

New Tools for Measuring and Analyzing Bacterial Gene-Expression Dynamics

by

Mirae Leigh Parker

Submitted to the Program of Computational and Systems Biology on 11.30.2023 in
Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy in Computational and Systems Biology

Messenger RNAs (mRNAs) are essential targets of gene regulation. The cell adapts and grows by changing its gene-expression profile, which it can achieve by manipulating the rates of mRNA initiation and decay and thus changing the relative abundances of transcripts. To understand the biological significance of these transcriptomic changes it is useful to observe how these changes correlate with emergent downstream behaviors and phenotypes. To manipulate and predict transcriptomic changes, it is also helpful to identify the sites of RNA regulation (transcription initiation, termination, and decay). By observing these sites of regulation, and how they change across different environmental and genetic contexts we can learn to recognize the sequence determinants of these processes and anticipate under which circumstances they will modulate gene expression changes. In order to record transcriptomic histories and facilitate the correlation of these archives with emergent cellular behaviors I have developed a molecular time capsule (MTC). An MTC is a self assembling protein capsule which captures highly reproducible snapshots of the full cellular transcriptome for delayed retrieval and analysis. These encapsulated records remain stable, even while the host transcriptome undergoes major remodeling. These records are also cleanly separable from non-encapsulated RNAs originating either from the host cell itself, or from other cells in a heterogenous population. To facilitate the identification of transcript ends across multiple conditions, I have also developed analysis tools for end-enriched RNA sequencing data (Rend-seq). Rend-seq is a variation of RNA-sequencing that aids in the interference of *in vivo* 5′ and 3′ ends in addition to determining their abundances. The tools I have developed for processing and identifying transcript ends from these data have been bundled into an open source Python package: rendseq. This package also powers an interactive website, rendseq.org, which increases the accessibility of published Rend-seq datasets.

Thesis Supervisor: Gene-Wei Li

Title: Associate Professor of Biology

Acknowledgements

I would like to start by thanking my advisor, Gene-Wei Li, for guiding me through my PhD. Gene, you are a brilliant scientist and an excellent writer. I was often impressed by the speed with which you ingest information and generate meaningful questions, as well as your creativity in coming up with new ideas. I would also like to thank my committee, Joseph (Joey) Davis and Paul Blainey. I am especially grateful for all the practical advice and suggestions I received during our committee meetings, not just on the scientific problems I was grappling with, but also on how to navigate the world of scientific publishing. Thank you also to Michael Baym who generously agreed to serve on my thesis defense committee - I am looking forward to sharing my science with all of you!

Next, I would like to thank all the members of the Li lab who I shared my PhD with. The original Li lab members who were present when I joined had created, along with Gene, a really exciting and welcoming scientific community. I am particularly thankful for the culture of curiosity they established (which persists to this day) of asking lots of questions during presentations, and remaining engaged in each other's work despite the fact that as a lab many of us work on disparate topics. To the newer members of the Li lab who joined with me and after me: thank you to you as well! Overall I feel extremely fortunate that I got to work with such thoughtful, supportive and interesting people.

I would like to thank both Grace and Dylan who oversaw my rotation and my first months in the lab, and helped me get established. Thanks also to Cindy and Ariel, who together have kept our lab running smoothly throughout my PhD, and have always been quick to help me personally if I need it. Thank you also to Jean for sharing with me some of his principles of effective quantitative biology. The ideas we discussed still shape how I think about my work today. I would also like to extend special thanks to Lydia. You were generous with your time and insights, and you taught me most of what I needed to know to be a wet lab biologist. Your friendship greatly improved my experience of the Covid-restrictions. Thank you also to Matthew, who was my original bay-mate. You taught me a lot about how to navigate the lab, introduced me to some great books and kept it real. It was a great pleasure towards the end of my time here to begin mentoring Jack. It has been so wonderful to collaborate with someone so enthusiastic and motivated! Thank you for your work on the MTC-project and your great questions and suggestions. And lastly, special thanks to Scott. I admire your enthusiasm for science, your quirkiness and your kindness. Your friendship made some of the difficult parts of my PhD much better and I will always be grateful for your encouragement to pursue my interests!

I am also very grateful to the sktime community for hosting me for Google Summer of Code. The experience of working on open source software as part of the sktime team was so exciting and inspiring, and I am really thankful to have been connected with such an engaging community. In particular I would like to thank my 3 mentors: Franz Kiraly, Lukasz Mentel and Guzal Bulatova who gave me many hours of their time and in turn saved me many hours of frustration with their helpful suggestions, feedback and ideas!

To the friends I have made during the PhD and to the ones I have from before: I am so glad our paths intersected - and I appreciate your support and companionship over the past few years. Sameed: how fantastic that you were one of the first people I met here and as fellow CSB students and Edgerton residents we were able to spend so much time together! I especially appreciate all the conversations we have had over the years. Molly: stat mech was hard but I am glad it brought us together! Thank you for pushing me during our stairs workouts, introducing me to some of my favorite books and doing so many fun things around Boston together. I was also grateful for all the memories we have made together with Sangeon and Kevin as well. Yukio: thank you also for all of the fun times together. I am proud of all your hard work and what you were able to accomplish at MIT! Yunsie: despite some of the challenges you faced, you are always such a wonderful and cheerful person to spend time with. I respect what a hard-working and accomplished scientist you are! To my other friends during my time here (Alex Konradi, Samira, Sachit, Alex Wu, Ifrah, Noor) - thanks for making life more interesting and meaningful!

Mom, Dad and Ema, I feel so lucky I ended up with you! Mom and Dad, you were the first scientists I ever knew - and I feel so lucky that you raised me in an environment where we spent so much time outdoors, or reading and talking about ideas. Dad: thank you for the time you spent working on projects with me and discussing fanciful ideas, and for what you taught me about being an open and accepting person. Mom: thank you for giving me such a fun and creative childhood. I admire your endless wonder at the natural world and your sense of beauty. I am also very grateful to my father's wife Kiera. You are such a kind, intelligent and funny person. Thank you for all the joy you have brought my Dad (and my cat). Lastly - thank you Ema, my wonderful little sister. I am so grateful to have had you as my companion since childhood and I am incredibly proud of you, both of how hard you work and of the human being you have become. Best of luck on your own PhD journey.

Lastly, to Kevin, I am so grateful we share a life together. Your hard work, honesty, curiosity and kindness dazzle me. Our partnership and friendship is one of the greatest joys of my life. Thank you for everything.

Table of Contents:

List of Figures

List of Tables

Chapter I

Introduction

Global Overview

Life as we know it perpetuates by converting genetic code into proteins. This process, known as the central dogma of molecular biology, has been the subject of intense study. It is not a simple one-step operation: segments of DNA must first be transcribed into messenger RNAs (mRNAs) which are then translated into proteins. The cell exerts regulatory control over this process by setting the rates of each of these steps, and by tuning the abundances of each component in this pipeline. In this work I focus exclusively on transcriptional regulation, highlighting some areas where our knowledge of transcriptional control is incomplete and sharing new tools for precise quantitative measurements to help fill these gaps. This introduction provides the background which contextualizes these methods.

In steady state the relative abundance of mRNAs correlates strongly with the relative abundance of their gene products (Balakrishnan et al., 2022). RNA production can be tuned over several orders of magnitude by adjusting the promoter activity. The rate of degradation is also transcript specific and can depend on factors as diverse as the ribosome occupancy, its interaction with small RNAs or proteins, or the presence or absence of various endonucleolytic processing sites (Mohanty & Kushner, 2016). Control of these rates, which can be highly dynamic and context dependent, determines the final abundance of a transcript. By manipulating the rates of RNA initiation and decay the cell can respond to changes in the environment, commit to a chosen developmental program, or initiate complex behaviors such as cell division. However, which gene-expression changes precipitate a cell's commitment to a particular developmental trajectory is not always easy to ascertain.

There exist varied methods for quantifying gene-expression changes over time; from microscopy, to full genome RNA-sequencing, to inference methods, and more. Each of these

methods has its own use cases where it excels. We will review these methods, as well as consider the pros and cons of each in uncovering the transcriptomic origins of developmental program selection.

Equally relevant are tools to help uncover the mechanism and identity of the biological processes which set transcript abundance, i.e., transcription initiation, termination, RNA processing, and decay. A better understanding of the controls of these processes will yield new design tools to predict and control gene-expression levels, and thus also help predict and control cell behavior. In this introduction we will explore what is known about these processes in bacteria, as well as some of the methods and techniques used to study their regulation.

Introduction for the Molecular Time Capsules (MTCs)

MTC Overview: Why measure the changing transcriptome?

Cellular phenotypes are dynamic and can change in response to new environments, stresses, and opportunities. Phenotypes are influenced by concurrent and proceeding changes in the proteome and the transcriptome. These gene-expression profiles are also dynamic and vacillate in response to internal and external stimuli and signals. It is through these changes that cells can adapt and thrive. To better understand and manipulate cell behavior, we can start by better understanding and manipulating the relationship it has with gene-expression.

There are several motivating examples where the connection between gene-expression profiles and phenotypic display is poorly understood. Antibiotic persistence, the

non-heritable transient reduced susceptibility to antibiotics (Huemer et al., 2020), is one such example. A better understanding of the expression to phenotype relationship is required and could inform better drug treatment plans or drug design. Immune response is another example. It is known that immune cell heterogeneity has an enormous impact on clinically relevant behavior, but the suspected transcriptomic origins of this heterogeneity can be challenging to identify (Hu et al., 2023; Liu & Cao, 2015). Both of these research problems, as well as many others, are united in the need for better tools to measure transcriptomic dynamics over time, and to probe how these dynamics relate to phenotypes of interest.

To understand, predict, control, and design cellular behaviors, it is necessary to gather the right data about gene-expression profiles and the phenotypes they correlate with. In particular, we need to observe what gene-expression changes precipitate interesting developmental trajectories. In certain cases, we may also need to expand our observations of gene-expression profiles outside of controlled laboratory conditions or else risk missing complex and unexpected behaviors that occur in the wild.

Making these sorts of measurements can be technically challenging. Cells may temporarily be inaccessible, prohibiting harvesting and measuring of their gene expression profiles. For example: bacteria during their transit of the human digestive tract. These profiles can also be challenging to capture if by the time one can distinguish which subpopulation of cells should be measured the transient transcriptomic signatures of interest are already gone.

In this section of the introduction, I will provide an overview of the following:

- Techniques for measuring transcriptomic profiles.
- The particular challenges associated with measuring transcriptomic dynamics

- Solutions to these challenges - with a summary of their relative merits and shortcomings.

This information will contextualize and motivate the Molecular Time Capsule as a worthy addition to the molecular biological toolkit, and hopefully provide a clearer sense of what problems it can be used to investigate.

Measuring Gene-Expression Profiles

To quantify expression, one must first be able to identify RNA. This is achieved by determining the sequence of nucleic acids in an RNA molecule. Once a process is developed to identify the sequence of an RNA molecule, it can be used to quantify the abundance of all such molecules in a sample. Where possible, this technique can even be applied to all RNAs in a sample - yielding to a full-transcriptome measurement.

In the early days, the method for identifying the sequences of RNA relied on running that RNA on a gel. Subsequent hybridization of this gel to a sequence specific labeled probe, and analysis via scintillation counter (Eisenstein, 2005) would allow for RNA identification and quantification. This method was ultimately improved upon in 1977 by the development of Northern Blotting (Noyes & Stark, 1975), a precise and highly sensitive method for quantifying RNAs in a sample that remains the gold standard for RNA quantification. The design and sensitivity of this methodology means that individual components of an RNA molecule (for example the 3′ end, the 5′ end or a splicing juncture) can be individually and quantitatively probed. The drawbacks in this approach lie in the fact that it is inherently low-throughput. Each gel can only support so many probes simultaneously without leading to

channel bleedthrough between different probes, practically limiting the ability to scale up this technique to generate quantitative full-transcriptome sequencing.

In 1985 a new technique for RNA quantification was unveiled: qPCR (Gödecke, 2018; Saiki et al., 1985), which allows for the quantification of RNA by converting it first to cDNA and then amplifying that cDNA with target specific primers using PCR. qPCR then measures the abundance of these amplified targets across amplification cycles to facilitate inference about the relative abundance of the original RNA molecules. While initially quantified by Southern Blot, there are now dedicated machines which can process and quantitatively assess the concentration of dozens to hundreds of qPCR reactions simultaneously (Taylor et al., 2019). However, qPCR remains highly sensitive to errors, and requires informed choices about what genes to use for informative normalization. It is easier to increase the throughput of qPCR measurements than it is for northern blots, however each gene-target still necessitates the design and testing of the amplification efficiency and specificity of primers. Each sample must also be mixed with its intended qPCR primers separately, making it challenging to generate multiplexed measurements. While qPCR is a useful tool for assessing the changes in a specific gene-target across conditions, it is not ideally suited to delivering full-transcriptome gene-expression profiles.

The limitation of these low-throughput approaches for assessing multiple sequences at once inspired the development of highly parallelizable assays for gene-expression quantification in the 90s (Niemitz, 2007). These efforts culminated in the establishment of the microarray as a highly parallelizable, multiplexed approach for assessing the abundance of a multitude of transcripts simultaneously (Pease et al., 1994; Southern et al., 1992). The number of sequences which could be simultaneously assayed was limited only by the diffraction limit of

light. Though many individualized and specific hybridization probes needed to be assembled on the glass slide.

In modern times RNA sequencing (RNA-seq) has become the successor to microarray technologies, becoming the preferred method for generating full transcriptome gene-expression profiles since its development in the early 2000s (Emrich et al., 2007; Lister et al., 2008). Technological advancements in flow cell technology for DNA sequencing (Holt & Jones, 2008) allowed for simultaneous sequencing of millions of transcripts, while circumventing some of the biases of microarray technologies (Zhao et al., 2014). Continued advances in detector technology and flow-cell fabrication have increased the throughput of RNA sequencing to the point where in 2023 billions of reads can be quantified in a single several hour run (Pervez et al., 2022).

As sequencing technology, library preparation, and analysis pipelines are all improved - the challenge of collecting the *right* RNA sample to answer a specific research question remains. This challenge has inspired an ongoing renaissance in molecular biological innovation. Cells can be sorted into relevant populations before sequencing to help establish sequence-function relationships (Peterman & Levine, 2016). Physical separation of single cells powers single-cell sequencing (Hwang et al., 2018; Shalek et al., 2014) and the recapitulation of spatial transcriptomic maps (Crosetto et al., 2015; Marx, 2021). We are constantly acquiring more tools for controlling and ascertaining the precise origins and identities of the RNAs we sequence.

There exist other modern methods for assessing RNA concentration in cells - for example microscopy/fluorescence-based approaches which were not discussed here. We will

briefly cover these techniques in an upcoming section entitled “Measuring Transcriptomic Dynamics with Microscopy”.

Challenges in Mapping Gene-Expression Dynamics

One source of particularly motivating and challenging questions about gene-expression profiles is how they change in time. How do changes in the transcriptome impact cellular behavior, and which changes are the most influential? What did the *in-situ* transcriptome look like when these cells were inaccessible, i.e., impossible to harvest and observe? These questions and others like them drive us towards finding a solution to the problem: how can we measure gene-expression dynamics over time?

Capturing the changes in transcriptomes is challenging as most methods for full-transcriptome quantification are destructive end-point measurements. This presents a problem as the emergence of a novel phenotype of interest may lag significantly behind the fluctuating signal of the changing transcriptome. By the time we are able to observe which cells display phenotype, its proceeding transcriptomic signature may no longer be observable. Conversely, once we destroy a cell in order to measure it, it becomes impossible to know what behaviors that cell would have gone on to express next.

There are also times where one may not be able to harvest cells when needed. The most obvious example of this is for bacteria during their transit of mammalian digestive tracts. We might be very interested in their behavior in these contexts, but without the ability to study them *in situ* we remain blind to their gene-expression dynamics. One solution to this research conundrum could be a technique for recording the transcriptome during the time of

inaccessibility, so that later when the cells are once again retrievable it is possible to observe their transcriptomic histories.

Both of these sets of research constraints require tools to help visualize transcriptomic histories and/or allow for longitudinal monitoring of the transcriptome. By gaining access to the historical transcriptomic record, it becomes possible to observe what transcriptome patterns precede the emergence of a phenotype of interest. It also facilitates the interrogation of transcriptomes *in situ* in contexts that would otherwise be difficult or impossible to access.

Mapping Phenotype to Transcriptomic Relationships Using Interventions

One solution to the challenge of mapping gene-expression to phenotype relationships is to perform interventions on the transcriptome and record the effects (Steinmetz et al., 2002; Winzeler et al., 1999). Genetic screens, which can under-express, over-express or mutate a particular target, have been instrumental in identifying the causal structure of gene-expression networks (Bakal et al., 2007; Echeverri & Perrimon, 2006).

This is the general strategy that has motivated recent developments of CRISPR-based screening methods (Bock et al., 2022) such as Perturb-seq (Dixit et al., 2016) and CROP-seq (Datlinger et al., 2017), which use CRISPR Cas9 to interfere with induction of a gene, also known as CRISPR interference or CRISPRi (Qi et al., 2013). These approaches make use of CRISPR tools to edit or repress large numbers of targets within a library of cells. This library can then be screened according to some phenotype of interest (Doench, 2018). Fluorescence-Activated Cell Sorting (FACS) can be used to sort single cells based on characteristics such as protein and RNA content, size, chromatin organization, methylation state

and accessibility (X. Chen et al., 2018) and even spatial origin (Satija et al., 2015) of single cells. These cells can even be sorted using multiple such filtering criteria, allowing for the separation of very specific phenotypes of interest. After sorting, each cell has a barcoded readout which allows for the identification of which guide RNA, and thus which target is repressed, in each cell.

This approach has obvious appeal for, in contrast to purely observational methods, it can yield conclusions about causal connections and structure in gene-expression networks. It has also been shown to be capable of uncovering surprising complexity in gene-expression networks. For example, it was deployed *in vivo* to successfully identify the point of action of Autism Spectrum Disorder, and Developmental Delay risk genes (Jin et al., 2020), suggesting new functions for these genes. It was also deployed to successfully study both single and combinatorial perturbations that influence the Unfolded Protein Response in Mammalian cells (Adamson et al., 2016), a study which revealed a branching complex network coupled with this cellular behavior, and highlight how Perturb-seq, and related methods, can effectively be used to study complex networks.

One uniting factor in both of these examples, as well as in most successful applications of CRISPR-mediated screens, is knowledge of a relevant set of target genes in advance. Due to the large number of genes in the cells, in order to gain sufficient statistical power for the relationship between all potential genes and the phenotype of interest, a large number of cells needs to be screened. Often, one might need to study multi-gene/combinatorial effects, in which case the number of conditions to test becomes astronomical and practically impossible. These limitations mean that CRISPR screens can be an impractical solution, both in terms of cost and time for questions which lack a predetermined set of target molecules to investigate. Efforts have been made to improve the cost efficiency of Perturb-seq like methods, for example Targeted

Perturb-seq, or TAP-seq, though ultimately these optimizations still necessitate prior knowledge of which targets of interest are worth observing (Schraivogel et al., 2020). The choice to only look at a small set of genes makes it challenging to uncover new biology, potentially glossing over alternative factors or pathways which can also contribute to a phenotype of interest, or perhaps obscuring aspects of the regulatory network that remain unobserved, or which rely on unobserved partners.

There is also variation in temporal dimension to consider. For example - is there a time window between the inhibition of a specific promoter by CRISPR is most relevant? One could imagine that within a single screen the time between inhibition, and sorting may vary. It is possible that steady state inhibition of a specific target gives rise to different phenotypes (for example a stress phenotype) than transient, biological repression would give rise to.

Practical constraints demand prior knowledge of what transcripts and targets should be interrogated. In practice this means this approach needs to scale up immensely in order to combat biases which keep unexpected factors hidden and obscure higher order multi-factor effects. Furthermore, it may demand knowledge of biologically relevant concentrations of certain transcripts - for example to maintain correct stoichiometric ratios of gene-products (Lalanne et al., 2018; G.-W. Li et al., 2014).

While intervention and observation of effect remain the gold-standard for establishing causal relationships, these techniques suffer when applied to large spaces of hypothesis. This reality can lead to conclusions which are potentially underpowered, or else which run the risk of unintentionally missing key elements. What these techniques would benefit from is some orthogonal method which can serve as a hypothesis generator for which gene targets, and what expression levels are worth testing for their relationship to phenotypes of interest.

Measuring Transcriptomic Dynamics with Microscopy

Fluorescence microscopy (Zou & Bai, 2019) is a well-established approach for quantifying transcript abundances over time and correlating those measurements with a variety of visually-readable phenotypes. Quantitative Fluorescence Time-Lapse Microscopy (QFTM) tracks fluorescent microscopy images over time - facilitating longitudinal tracking of fluorescently labeled targets in single cells (Muzzey & van Oudenaarden, 2009; Young et al., 2011). This approach is appealing because it allows one to directly quantify which transcripts are present in a cell almost continuously. By genetically modifying organisms to fluorescently tag mRNAs of interest, their expression levels can be quantified real time in living cells. Here we will first overview methods to assess live-cell dynamics of transcription using fluorescence microscopy, then we will discuss their relative merits with respect to the problem of assessing genotype-to-phenotype relationship mapping.

One approach for quantifying mRNAs in living cells is by attaching a label to the transcript via the affinity of a fluorescently tagged protein to a genetically encoded structure. This approach was first demonstrated in 1998, using the bacteriophage derived MS2-tag system (Bertrand et al., 1998; Ferguson & Larson, 2013; Peña et al., 2015), but has also been demonstrated with similar systems such as the λN_{22} derived system (Schönberger et al., 2012), and the BglG system (J. Chen et al., 2009). Each of these approaches place a stem-loop structure into the target RNA and expresses a partner stem loop binding protein. This stem loop binding protein is modified to be fluorescent. When target RNA and fluorescent protein are expressed simultaneously, the protein binds to the RNA. The fraction of bound protein can then be assessed by observing the diffusion patterns of the fluorescent molecule. From this analysis a

concentration of the target molecule of interest can be ascertained, and can be correlated with the phenotype of the cells.

Another class of fluorescently labeled proteins which can bind RNAs are those which are CRISPR/Cas derived. Both fluorescently labeled Cas 9 (Nelles et al., 2016) and Cas13a (Yang et al., 2019) have demonstrated capacities for fluorescently labeling target transcripts *in vivo* in a programmable fashion. This means that by changing the target sequence of the CAS molecule one can selectively label certain RNAs of interest and track their abundance in the cell.

In 2011 RNA aptamers were designed which bind small molecules and give rise to fluorescence (Paige et al., 2011; Strack et al., 2013). This discovery opened the path for genetically encoding fluorescence into RNAs in much the same way that proteins can be genetically tagged for fluorescence. This is achieved by adding the aptamer tag to the end of the transcript, and supplying the small molecule to the media which will aid fluorescence when bound to the aptamer tag. These classes of tags have been improved on (X. Li et al., 2020; Sunbul et al., 2021), using intensity or lifetime of the fluorophore to track multiple RNA species at once (Sarfraz et al., 2023).

Lastly another common class of approaches for fluorescently labeling RNA relies on the delivery of fluorescently labeled materials to the cell. Micro-injection of fluorescent full length transcripts (Glotzer et al., 1997), or injection of fluorescently labeled dUTPs which are then synthesized into fluorescent transcripts (Sinha et al., 2008) are two such examples of this. While these two approaches are relevant to the study of transcription dynamics within living cells, they are not useful for questions regarding gene-expression of specific targets in the cell, and thus we don’t consider them further here.

What all of these approaches for labeling transcripts have in common is that they are limited to labeling only a few genes at once. This means that even with the most up-to-date protocols for multi-channel fluorescent microscopy, researchers often need to have at least some prior knowledge of which genes or proteins to target for the most informative observations. Without prior knowledge, it can be difficult to deploy these techniques to identify new or unexpected transcriptome dynamics. Once targets of interest are known, any combination of these RNA labeling strategies would be an effective method for orthogonally verifying that specific transcript abundance changes proceed or seed developmental trajectories. Much like the intervention-based approaches discussed earlier however, these sets of techniques still rely on hypothesis generation in advance in order to limit the set of observed targets to a tractable size.

Inferring Transcriptomic Dynamics by Rates of Change:

Recently there has been a surge of research into methods for leveraging RNA-seq data to infer transcriptomic dynamics over longer time scales. This can be achieved by modifications to the analysis or preparation of a single time point sample that allow one to estimate the rate of change in abundance of a transcript in addition to its current quantity. This rate of change, or velocity, refers to whether the abundance of a particular gene is changing in time, either increasing or decreasing, as well as how fast that change is occurring.

Once the current levels of gene-expression and gene-expression change are established - the goal becomes to create a map in the high-dimensional landscape of gene-expression, and to predict flows within it. By creating such a map, or vector field, one can begin to perform inference about the complex temporal dynamics of gene-expression and to predict future expression levels or cellular phenotypes given the current gene-expression levels.

There are currently two popular methods for establishing this rate of change. The first of these methods infers the rate of transcriptomic change from a ratiometric analysis of unspliced and spliced mRNAs (Gray et al., 2014; La Manno et al., 2018; Zeisel et al., 2011). Unspliced mRNA, or pre-mRNA, levels can be predictive of the future levels of spliced, or mature, mRNAs. The levels of each population of transcript can be assessed by sequencing across the boundary of introns and exons to establish the percentage of transcripts which are spliced.

It is also possible to use metabolic pulse labeling to distinguish between newly synthesized mRNAs and older mRNAs (Herzog et al., 2017; Q. Qiu et al., 2020; Rabani et al., 2011), as has been demonstrated by methods such as SLAM-seq, TimeLapse-seq, TUC-seq. This approach allows for the interrogation of prokaryotic mRNA rates of change, which don't typically undergo splicing. To label transcripts using this method a nucleotide analog, most commonly 4-thiouridine (4sU), is added to the media of cells before harvesting. As this metabolic label is taken up by the cells, it is incorporated into their newly synthesized RNAs. Upon harvesting, the extent of this modification to a given transcript can then be assessed, either by inducing mutation at the site of incorporation (for example a conversion from 4sU to C), or by biotinylated recovery of the metabolically labeled RNAs and subsequent sequencing.

While both of these approaches can be applied to bulk/population level measurements, unless there is some sort of synchronicity in the population this technique is most informative when applied to single cells (La Manno et al., 2018; Q. Qiu et al., 2020). Each individual cell can be thought of as a single point in the "phase diagram" of gene expression levels and expression change velocities. After obtaining enough estimates of these points the analysis problem becomes the inference of paths within the phase-diagram (Bergen et al., 2020; X. Wang & Zheng, 2021; Weng et al., 2021). These analysis packages might take in splicing data,

metabolically labeled data or a combination of the two and use them to recreate vector maps. They seek to explain how a cell at one confluence of expression and expression change will flow towards another point of expression and expression change (X. Qiu et al., 2022).

This reliance on inference to interpret transcriptomic dynamics does have its limitations however. For example, when looking at RNA velocity inferred from pre mRNA and mature mRNA ratios, only a small number of transcripts can be described with the simple kinetics assumed in descriptions of RNA velocity (Bergen et al., 2021). In fact the phase diagrams created using these assumptions are typically inferred from only a handful of genes (Bergen et al., 2020).

These limitations mean that more complex models must be employed in order to extract dynamical information for the vast majority of genes in the genome, which themselves might come with many different parameters (Gorin et al., 2022). The tuning and selection of these different parameters is itself often a challenging and underpowered task, and should reduce the confidence in any gene-specific inferences. Thus while RNA velocity may be a useful tool for describing the relationships between different cells, it does not yet have the inference power to be used as a reliable tool for pinpointing which transcripts precipitate developmental trajectories of interest in all cases. Methods for longitudinal tracking of variation may in certain cases still be preferable or complimentary to this approach.

Continuous Harvesting of RNA for Monitoring:

Longitudinal tracking of cells can be achieved by simply taking multiple measurements of RNA over time. It has been shown that multiple samples of RNA can be extracted from a

living cell without causing death using several recently characterized techniques. These techniques allow for the removal of a small fraction of the cellular cytoplasm, which can then be used for RNA sequencing and analysis. Collecting multiple such cytoplasmic samples over time yields a measurement of gene-expression changes over time.

One recent system, COURIER (Horns et al., 2023) (Controlled output and update of RNA for interrogation, expression and regulation), allows for the controllable active export of RNA molecules from the cell, allowing both for communication between cells and also for recovery and analysis of excreted RNA payloads. Though cells naturally shed RNA over the course of their lives, this system increases the amount of RNA that is excreted from any cell - boosting both the sensitivity and information content of signals.

These packages of RNA are collected from the supernatant of the media and are filtered and clarified before the RNA is separated and sequenced. One issue that arises with using this method however is that it is challenging to identify which purified RNA package originated from which cell. Without physical separation of sub-populations of cells, as a means to help with the identification of which COURIER packet originated from which cell, this type of measurement becomes akin to a bulk time series measurement. Unlike bulk series measurements the cells being monitored are not destroyed in the process of harvesting the sample. Without further modifications to the protocol it seems that the inference that can be performed on these samples is identical to the forms of inference that could be performed on bulk time series measurements alone.

As such, used in its current form COURIER does not provide any inference advantages over simply taking multiple harvested samples from the population over time. This point does not discount its potential usefulness as an observational tool in situations where any cellular

destruction is undesirable, or as a method to facilitate inter-cell communication (as the authors discuss and demonstrate in their paper). This lack of identifiability prohibits the ability to correlate excreted RNA from one time point with a downstream behavior or phenotype that might emerge in only a subpopulation of cells meaning it is ill-suited to study the subtle transcriptomic origins of developmental trajectory commitment.

Another technique for continuous, non-destructive monitoring of transcriptomics is the direct removal and labeling of cytoplasm from living cells. The removed sample can then be processed and analyzed in a similar manner to single cell RNA sequencing samples. This process was recently demonstrated in a technique called Live-Seq (W. Chen et al., 2022) where researchers were able to recapitulate scRNA-seq clusters based on cell type and state. This approach had clear advantages over vanilla scRNA-seq because it allows for longitudinal tracking of cells, and the correlation of downstream phenotypes with upstream expression profiles.

The main challenge preventing Live-seq from being adopted more widely for the study of developmental trajectories is a scaling and accessibility issue. Cells must be operated on by a specialized instrument in order to recover their RNA, a process which has very low throughput. In addition, there is not wide-spread accessibility/knowledge of the mechanical mechanism in order for it to be put to wide use. Using Live-seq to study transcriptomic dynamics in hard to access locations is also clearly not tractable as any environment where it is impossible to recover RNA for regular sequencing would also prohibit any use of a microscope/cytoplasmic removal.

Genomic Recording Techniques:

By recording a snapshot of the transcriptome in the genome of a living cell, one can correlate that captured transcriptomic history with phenotypes expressed after its recording. These forms of transcriptomic records remain with the cell and its lineage, meaning that by separating cells based on emergent phenotypes one can successfully recover only those transcriptomic records which preceded the sorting phenotype. There has been a recent growth in research towards techniques which facilitate genomic recording of gene-expression information in living cells. To record transcriptomic histories into DNA, these methods must utilize a technique to reverse the flow of genetic information and use RNA abundance information to change the genome (Ishiguro et al., 2019; Sheth & Wang, 2018). Due to the increased stability of the genome, these records can then be maintained in the cell and may be recovered later by targeted sequencing of the site of recording. This record is also heritable, making it possible to trace transcriptomic histories along cellular lineages. There exist several different methods which have demonstrated the use RNA abundances to trigger changes in the DNA

One such method uses tracrRNAs, trans-activating CRISPR RNAs, to target specific transcripts. When combined with gene-specific guide RNAs these designed tracrRNAs trigger mutations in an accompanying and complementary recording construct. The resultant accumulation of mutations over time serves as a readout of the concentration of the various targeted transcripts over the recording window. This method has been demonstrated via a technique known as TIGER (Jiao et al., 2023) (transcribed RNAs inferred by genetically encoded records).

While this approach represents an exciting method for converting transcript abundances into genomic records, it does come with several limitations. For instance, it can only be used to interrogate a select number of transcripts within a single cell. In addition, as the

authors demonstrated, any target will require extensive engineering of the tracrRNA/gRNA combos, which will then have to be tested before deployment in order to allow for informative and reliable normalization of the readout. Lastly the reliance on mutation as a readout of abundance suggests the existence of a saturating abundance of the target. Beyond such a saturating abundance it would be challenging to infer RNA levels from the mutated record as mutations may start to either overwrite one another, or the efficacy of mutation in the region may decrease. These limitations make TIGER unsuitable for studying the origins of developmental trajectories. However it is worth noting that this technique is still useful for applications which require integrated measurement of a known set of target genes in difficult to reach contexts. For example, applications which demand sentinel record-keeping cells to track evidence of specific gene-expression programs.

Another demonstrated method achieves genomic recording by tagging specific transcripts of interest with retron-inspired RNA barcodes. In combination with a properly calibrated CRISPR-Cas system these RNA barcodes can be captured and recorded in a CRISPR array when their corresponding RNAs are induced. Because CRISPR arrays add new records at the end of the array it is even possible to infer the order of expression of the various targets by observing the order of barcoded spacers in the array. This technique has been demonstrated via the Retron-Cascorder (Bhattarai-Kline et al., 2022) and a related method called TRACE (Sheth et al., 2017). These methods have similar pros/cons as TIGER, meaning that they are limited with their scope of recording, making them also unsuitable for studying unknown origins of phenotypic heterogeneity.

Lastly, a compelling method for recording RNA in the genome comes from systems which can record arbitrary transcripts, as a function of their abundance in the cell, into a CRISPR

spacer. This was first demonstrated in 2016 when it was shown that a naturally occuring Reverse-Transcriptase-Cas1 fusion could facilitate CRISPR spacer acquisitions derived from RNA (Silas et al., 2016, 2017). The RT part of the hybrid converts random transcripts into DNA, and the CRISPR-Cas1 inserts them into the CRISPR array as a spacer. However, it was discovered that this system was unable to maintain its functionality in *E. coli*, leading to the testing of a different system of RT-Cas1 and Cas2 from *F. saccharivorans*. This system proved to be heterologously maintained in E. coli. Record-seq is a genome recording method which makes use of this novel system to direct the acquisition of *E. coli* RNA derived spacers and recover these recorded transcriptional histories by targeted deep RNA sequencing (Schmidt et al., 2018). In addition to the initial proof-of-concept experiments, Record-seq has demonstrated success recording the transcriptomes of *E. coli* as they pass through the digestion systems of mice (Schmidt et al., 2022; Tanna et al., 2020).

This method changes the scope of questions that can be answered using genome-recording techniques. As it can capture a record of full transcriptome, rather than being limited to a select set of targets, it can be used for novel hypothesis generation of transcripts of interest. While the initial results and applications with this technique demonstrate its potential as a tool for studying genotype to phenotype relationships, it currently has several areas which need to be improved upon before it can be used to study the subtle gene-expression changes thought to drive sub-population phenotypic heterogeneity.

To begin with each cell has a very small probability per unit time of acquiring an RNA-derived spacer. It is estimated that only $1.9\ 10^{4}$ E coli will acquire a spacer during a recording window of at least 12 hours. The low rate of spacer acquisition may require that this method be used over very long recording times, reducing the temporal resolution of any

transcriptional history that is recorded. In addition it implies that a large number of E. coli cells may need to be sequenced in order to yield sufficient information about the recorded gene-expression profile. Lastly, and likely due in part to the long recording times used by this technique, Record-seq is not highly reproducible across different biological replicates, with the highest reported Pearson coefficient between two samples being 0.786 (Schmidt et al., 2018). With this low of a Pearson coefficient it will be challenging to ascertain whether specific transcripts which seem to be “upregulated” in one Record-seq sample versus another are truly due to underlying biological differences between the two samples or are due to noise in the measurement.

MTC Background Summary

How transcriptomes change over time, and the relationship between transcriptomic profile dynamics and phenotypic expression, are motivating topics of study. By better understanding genotype-to-phenotype relationships we will be more prepared to predict, control, and design bacterial behaviors. However, the transcriptomic origins of many phenotypes remain challenging to uncover. Many techniques for detecting or intervening on the abundances of transcripts over time, such as fluorescent microscopy, CRISPR guided screens, or targeted recording of specific transcripts can not address this problem on their own. As these approaches are limited in the number of targets they can simultaneously record or effect, they require tools for hypothesis generation. Full transcriptome recording techniques do offer a promising avenue for increased coverage, but they may lack the reproducibility, the temporal resolution and the throughput to be appropriate for applications requiring subtle gene-expression differences.

What is needed is a method to generate a list of candidate targets which are known to correlate with phenotypes of interest. We can then interrogate this list using directed quantitative approaches, such as Perturb-seq, in order to establish causal links between gene-expression levels and phenotypes of interest. Such a method would need to be highly reproducible - any genes which appear to be different between recording conditions ought to reflect differences in the host transcriptome during recording rather than reflecting noise in the recording apparatus. These records ought to be cleanly separable from the current host transcriptome - such that signals from later time-points don't contaminate the historical records of the transcriptome. Lastly it is essential that the transcriptomic records are static over time - allowing for a variable time from snapshot collection and harvesting. A system with these combined properties could serve as a powerful, low-cost hypothesis generator for gene-expression and phenotype networks.

Rendseq introduction

rendseq Overview: What Are Transcripts:

The prevailing direction of information flow in living organisms is from DNA to RNA to protein. In this simplified view, DNA serves as the heritable genetic code, RNA as a message passing intermediate, and protein as the substance which forms the physical forms of cells and performs most of the critical processes of life. mRNA transcripts, the focus of this work, are the central players in this pipeline. mRNAs are essential targets for gene-regulation. Their levels can be tuned over many orders of magnitude, and in steady state mRNA levels serve as strong predictors of protein products (Balakrishnan et al., 2022).

mRNA transcripts are often merely thought of as substrates for the protein-making machinery of the cell. This apparent simplicity obscures the potential for complex regulation of the transcript before the translation process is ever initiated, for example via regulation of the steady state levels of each transcript. The rate of transcription initiation from any transcription start site (TSS) can be highly variable and subject to feedback loops and other forms of regulation by transcription factors (G. Wang et al., 2015) or alternative sigma factors (Kazmierczak et al., 2005). These transcripts are then actively destroyed by the cell - a process which can be influenced by the 3′ structure of transcripts, relative abundance of nuclease motifs (Agarwal & Kelley, 2022), and the rate of translation (Wu et al., 2019). Together these tunable rates of production and decay set the steady state levels of a transcript across many orders of magnitude.

Transcripts that share the same TSS may end at different sites, depending on the strength and mechanism of downstream terminators. Similarly, any given terminator can terminate transcripts originating from any number of TSSs. These overlapping transcripts with disparate starts and ends are known as transcription isoforms. In prokaryotes, the identities and relative abundances of different transcript isoforms is a critical mechanism for controlling the relative abundances of genes in a polycistronic mRNA (Cao et al., 2015). Eukaryotic RNAs encoding for the same gene can also exist as transcription isoforms, and can be recombined in combinatorially complex patterns. This process, known as RNA splicing, (Berget et al., 1977; Wilkinson et al., 2019) is essential for understanding gene expression control in Eukaryotes.

All of these factors - where a transcript starts, where it ends, how it is rearranged, and the speed and mechanism of RNA decay all contribute to what we mean when we talk about a transcript's identity, and the influence it has over the production of protein products. In the

following sections I will provide a condensed review of what is known about each of these processes. I will provide a very high level view of each process, as well as include references to various review articles which provide more detail than I intend to go into here. Then I will introduce a technique specifically designed for the purpose of identifying RNA ends *in vivo* (Rend-seq) as well as current analysis challenges faced by this technique.

What Defines Transcript Boundaries?

A full length transcript starts at a Transcription Start Site (TSS) and ends at the point of termination. These original boundaries can be changed as the RNA is processed or decayed. We will now go through a quick review of the processes of transcription initiation, transcription termination and factor mediated RNA decay in prokaryotes. We will apply special focus to the issues of where these processes happen (as well as suspected regulatory triggers of these processes) and what sort of signature they leave on the distribution of transcript ends in a living cell.

Transcription Initiation

Transcription initiation is the first step in the process of converting DNA to RNA. It is a highly regulated process that can significantly impact the final abundance of transcripts (Häkkinen & Ribeiro, 2016). It describes the process through which RNA polymerase (RNAP) is loaded onto a specific sequence location and transcription begins. Here I will provide a quick overview of some of the major players in the process of transcription initiation and what factors are thought to shape their actions. The players and interactions which facilitate this process are

illustrated in Figure 1.1 which can serve as your guide as you read the following descriptions. I will provide a quick overview of RNAP, promoters, transcription factors, and sigma factors, as well as the role each places in the process of transcription initiation in prokaryotes.

RNA Polymerase

The RNA Polymerase (RNAP) is the machinery which performs the work of transcription. It is a multisubunit enzyme. The core enzyme is composed of the α (2 copies), □, □′ and ω subunits (Sutherland & Murakami, 2018). These subunits are joined by the modular σ subunit (discussed in more detail below) to form the version of RNAP fully prepared to initiate transcription: the holoenzyme.

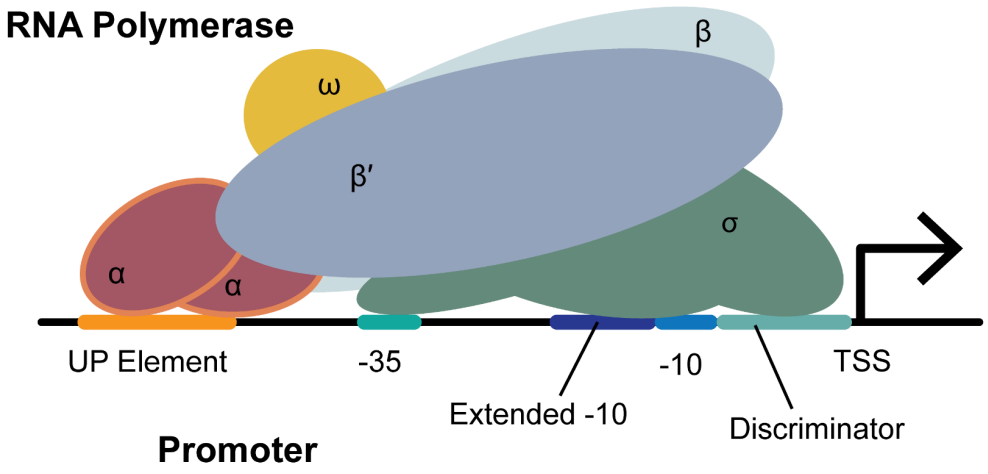

Fig 1.1: A cartoon illustration of key players in the transcription initiation process. The RNAP holoenzyme, composed of the α, □, □′ and ω subunits is bound to a sigma (σ) factor. This complex demonstrates a potential binding conformation between the RNAP a promoter,

with the α subunits interacting with the UP element, and the sigma factor interacting with the -10, the -35, the extended -10 and the discriminator sequence (Mejía-Almonte et al., 2020). Not all interactions will happen for all promoters.

Promoters

The promoter refers to the stretch of DNA where RNAP binds and initiates transcription. The holoenzyme bound to the promoter is a structure known as the "closed complex", while once the DNA double strand is unwound it transitions to being an "open complex" (Browning & Busby, 2004). The promoter is situated just upstream of the transcription start site (TSS). The particular sequence and structural features of the promoter determine its affinity to Sigma factors or transcription factors, and thus it plays a pivotal role in determining the regulation of when a gene is turned on, and the loading rate of RNAP across different conditions.

From analyzing the transcription initiation sites of many bacterial promoters, a "consensus" structure containing a -10 and -35 region was established to describe the structure of bacterial promoters (Hawley & McClure, 1983). This simple model has since been expanded on - for example via the discovery of an additional element called the UP or upstream element that is believed to influence transcription initiation by interactions with the α subunit of RNAP (Ross et al., 1993; Yan & Fong, 2017). There is also an additional extended -10 element, a 3-4 base pair motif located upstream of the -10 unit which can facilitate additional recognition and interactions with the σ subunit of the holoenzyme (Browning & Busby, 2004). Some promoters also contain a GC-rich sequence known as the discriminator (Josaitis et al., 1995; Travers et al., 1986) which

facilitates additional interactions with certain sigma factors and is believed to have special regulatory significance for amino-acid starvation responses.

Sigma factors

As mentioned before, sigma factors are a necessary component of the holoenzyme - the form of RNAP that performs transcription initiation. Sigma factors depart as the enzyme transitions into elongation mode, leaving behind the core enzyme. Sigma factors are unable to bind DNA while in their free state. When bound to RNAP however, the sigma factor undergoes a conformational change, exposing DNA-binding domains (Feklístov et al., 2014). In this way they serve as exchangeable modular subunits of RNAP which determine its sequence specificity.

Sigma factors act as a type of transcription factor in bacteria, one that interacts directly with RNAP. When RNAP binds to specific sigma factors, it elicits an allosteric shift in the sigma factor to expose DNA-binding domains. These different DNA binding domains have different sequence or structure specificity which defines the set of targets each sigma factor will recognize. These binding domains serve to guide RNAP to different locations in the genome. In addition to guiding RNAP to the correct location in the genome for transcription initiation, sigma factors can also perform strand separation of the dsDNA once they arrive. Because of their modular nature they can easily be swapped out for one another prompting a change in the behavior and specificity of RNAP. Sigma factors orchestrate sweeping changes in the transcriptome when their levels change in the cell. Each sigma factor often has specificity to many different locations in the genome, meaning that altering the expression level of a given sigma factor will often lead to a cascade of other genes being up or down regulated.

There are several different families of sigma factors, classified by their conserved homology to one another (Feklístov et al., 2014). A well known group of sigma factors are

known as the σ^{70} family of factors (Paget & Helmann, 2003) - named for their similarity to the housekeeping sigma factor which was originally discovered. The most abundant type of sigma factor are actually the recently discovered ECF sigma factors (Helmann, 2002). The diversity of sigma factors hints at the diversity of possible recognition motifs that accompany them. Rapid and robust identification of true transcript ends in different conditions is one fruitful avenue toward discovering more of the functions and specificity of these factors.

Transcription Factors

In addition to sigma factors, bacteria have transcription factors which activate or repress transcription of specific genes. These factors can have either a narrow or broad effect on the transcriptome, and are categorized as local or global based on the extent of their effect (Seshasayee et al., 2011). They will bind to the area of DNA near the TSS of their target genes. Depending on the proximity and orientation of this binding site to the TSS, this binding event can either repress or promote the rate of transcription initiation (Seshasayee et al., 2011).

LacI is the most famous “local” Transcription Factor. Even though LacI is such a well characterized promoter, yet even this well known transcription factor is still the subject of active research into its sequence to expression landscape (Swerdlow & Schaaper, 2014). A more streamlined discovery process of potential promoters powered by rapid identification of TSSs in a variety of different conditions provides new sequence contexts for understanding how promoter sequence and structure exert control over gene expression. Identification of the potential emergence of alternative promoters (which can emerge within the body of another transcript and can be challenging to identify with traditional techniques) or the change in position or RNAP loading rate of a particular promoter across different conditions can facilitate the identification of

where new transcription factors initiate, or how changes in the genetic or environmental context may change the behavior of individual promoters.

Transcription Termination

Transcription termination is another essential process for tuning transcript abundances. The position and method of termination allows for the timely recycling of RNAP (Kang et al., 2020) and can inform the stability of transcripts and their susceptibility to exo-nucleolytic decay (Richards et al., 2008) as well as informing the relative abundances of gene products in the same operon (Lalanne et al., 2018). In prokaryotes there are 2 known forms of transcription termination which give rise to programmed and repeatable termination: intrinsic termination and factor-mediated termination. These two modes of termination result in disparate ends of transcripts and each generates its own unique signature which can be identified by looking at the distribution of transcription ends.

Intrinsic Termination

Intrinsic termination is characterized by a unique sequence/structure signature in the 3′ end of RNA transcripts. This structure canonically consists of a GC-rich hairpin structure followed by a downstream Uracil-rich region (Gusarov & Nudler, 1999). Oftentimes the region of sequence upstream of the hairpin is also enriched for Adenines and it is thought that the downstream U-tract may also pair, either before or after transcription termination, with this upstream A-tract. The hairpin structure and downstream U-tract are the most conserved features of intrinsic terminators across species.

Intrinsic Termination is thought to occur when the RNAP is positioned over the U-tract - an area with lower than average strength of the RNA/DNA hybrid. It is thought that

while being positioned at this spot the upstream hairpin undergoes a nucleation event, which ultimately causes disassociation of the elongation complex and freeing of the RNAP (Ray-Soni et al., 2016). There is some evidence that external factors - NusA and NusG also contribute to pausing at the point of termination (Guo et al., 2018; Mandell et al., 2021, 2022) which is thought to increase the likelihood of successful termination. In certain cases, such as for certain B. subtilis intrinsic terminators, there is evidence that even Rho plays a role in stimulating transcription termination (Mandell et al., 2022), evidence which blurs the distinction between intrinsic and "Rho-mediated" transcription and demonstrating that there is more to be discovered about the regulation and performance of intrinsic termination. Successful intrinsic termination of a transcript results in a 3′ end without a free phosphate (Vasilyev et al., 2019). Intrinsic termination repeatedly happens at the same location over and over again.

Intrinsic termination will often occur over a very localized distribution of sequence positions, either at a single position, or over several adjacent positions. This means it gives rise to a very localized distribution of RNA ends, and that many transcripts which use an intrinsic terminator for termination will share a common 3 end.

Factor Mediated Termination

The other main form of termination in bacteria is Rho-mediated termination. Though there may be other factors which are as yet poorly characterized (especially across diverse bacteria), the canonical example of a factor which leads to termination is Rho. Rho is known to function in bacteria as diverse as *B. subtilis* and *E. coli*, though it is not essential in all organisms (including *B. subtilis* and *Staphylococcus aureus*) and is missing from other organisms altogether (such as *Cyanobacteria*). Rho can also be conditionally essential for some organisms,

for example it becomes essential for *Caulobacter crescentus* under oxidative stress (Mitra et al., 2017).

Rho-mediated termination plays several biologically important roles. Some bacteria, such as *E. coli*, experience transcription and translation coupling (Johnson et al., 2020). In these bacteria Rho is thought to terminate transcripts which are not being actively translated, as there is no trailing ribosome to protect RNAP from termination. Even in bacteria which experience runaway transcription, Rho still seems to play an important role in nonsense mediated decay (Bidnenko et al., 2017).

Though the exact mechanism of Rho termination is as yet unknown, as evidenced by current conflicting models of Rho termination (Hao et al., 2021; Molodtsov et al., 2023; Rashid & Berger, 2023; Said et al., 2021), there are established signatures of Rho termination. For example - Rho termination is thought to occur at/near Rho utilization site (Rut sites).

RNA Decay

Another mechanism for the regulation of mRNAs is via their rates of degradation. The rapid degradation of mRNA in most bacterial species determines their short half lives (Richards et al., 2008), which in turn facilitates the rapidity with which mRNAs levels are able to reach steady state after changes to the rate of production. This in turn means that bacteria can rapidly adapt their gene-expression levels in response to changes in the environment, in addition to getting rid of mRNAs whose protein products are no longer needed, and returning the ribonucleotides of a given RNA to the cellular pool for reuse.

There are multiple mechanisms which facilitate degradation in bacteria. Here we will disregard pathways associated with nonsense mediated decay and focus purely on factor mediated decay of mature transcripts and their decay products. We will also focus purely on prokaryotic mRNA degradation.

Ribonucleases perform the work of RNA decay in the prokaryotic cell. These are composed of Endo and Exo (both 3 ′ -> 5′ and 5′ -> 3′) Nucleases. The role these nucleases play on turnover and regulation of mRNAs in the cell has been the focus of decades of study since the late 1970s (Apirion, 1973). Much of the pioneering work on nucleases took place in *E coli*, which initially limited the understanding of the potential diversity of prokaryotic nucleases (for example E. coli lacks any 5′ -> 3′ Exonucleases (Deutscher, 1985)). More modern work has revealed that while some of the nucleases first identified in *E. coli* seem to be conserved across many species (Huch et al., 2023), others are not, and there exist types of nucleases in other organisms which are not found in *E coli*. These discoveries underline the point that there still remains a diversity in mechanisms of RNA decay across bacterial phyla waiting to be fully characterized.

Identifying Transcript Ends:

Studying the processes of RNA modification and regulatory control necessitates first determining the precise identity of which transcripts are actually present in a given sample. Unfortunately it is much easier to identify the beginning and end of a protein from looking at its sequence than it is to find the start and end of a transcript. This is because proteins have a small, canonical set of start and stop codons, whereas the sequence features which determine where a

transcript starts and ends are not completely determined and vary more widely from species to species.

Oftentimes one might often wish to not only identify the starts and ends of transcripts - but to also make quantitative predictions about a relative transcripts abundance in different conditions. Without knowledge of what features define transcript boundaries, and presumably also set transcript abundance by informing the rates of transcription initiation and decay, it becomes challenging to make models capable of quantitative prediction.

Because these features can be difficult to predict *in-silico*, it is oftentimes preferable to measure the transcript boundaries from *in-vivo* samples. Regular RNA-sequencing is not capable of doing this with high precision however. The ends of transcripts are often represented as blurry slopes from which extracting a single nucleotide position of transcription start/end becomes challenging.

Instead, there is a sequencing technique, called Rend-seq, which overcomes this issue by enriching the ends of transcripts (Lalanne et al., 2018). It works by sparsely fragmenting RNA then applying a size selection. Fragments which end at one of the original transcript ends have a higher likelihood of ending up in the size-selected pool as they require only one fragmentation event at the correct location to achieve the correct size, whereas internal reads require two. See Figure 1.2 for an illustration of the protocol.

This technique has been shown to deliver single nucleotide resolution identification of transcript boundaries coupled with transcript abundance. This high resolution of transcript boundaries can then be used to see the boundaries and relative abundances of transcriptional isoforms. This insight into the beginning and end of transcription can uncover bioinformatic signatures hinting at the nature of transcriptional control across different

organisms. For example - the observation that termination of transcription happens very close to the site of translation termination served as the inspiration that led to the discovery of uncoupled transcription and translation in *B. subtilis*.

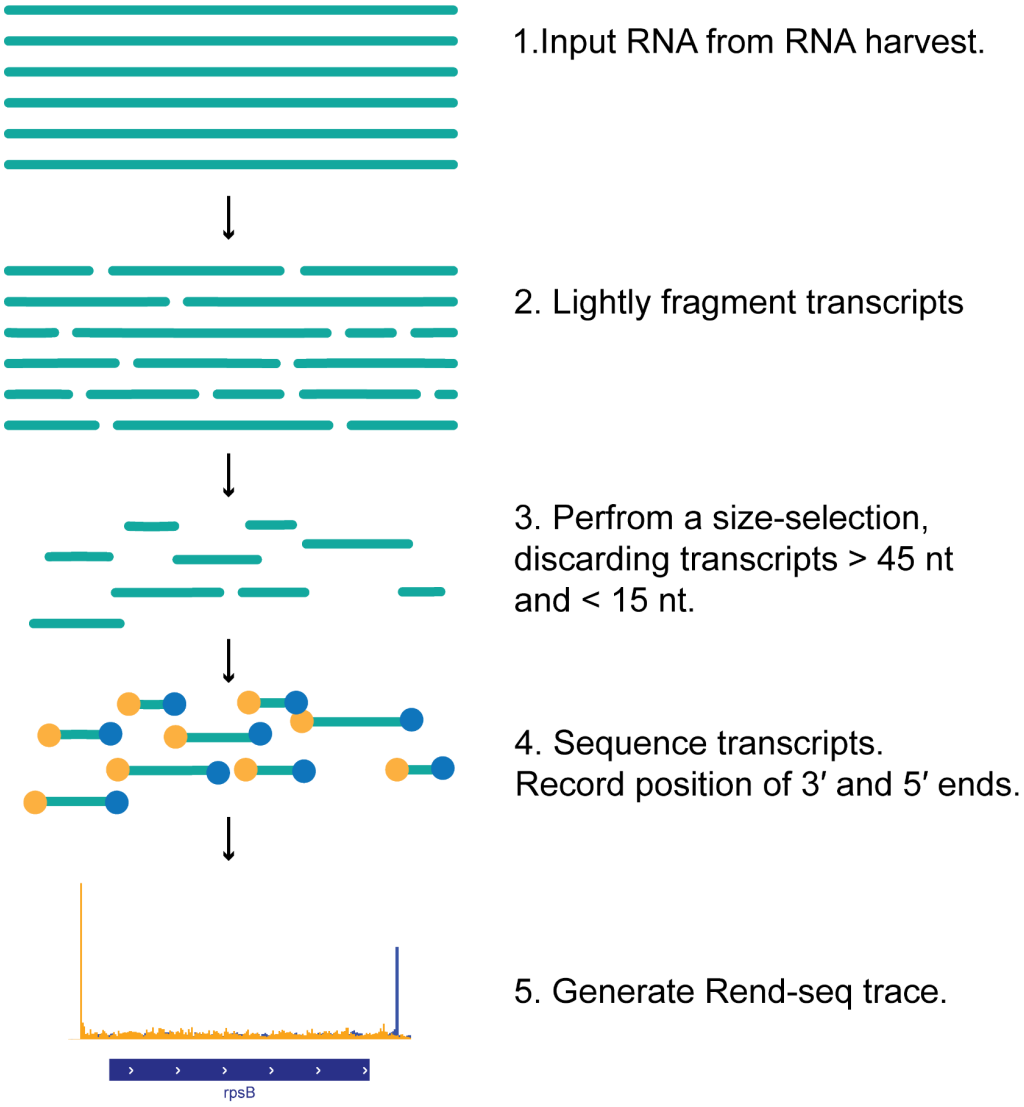

Figure 1.2: A conceptual illustration of the Rend-seq protocol. RNA harvested from a sample (1) is first subjected to a light fragmentation process (2). These fragmented transcripts are subjected to a size selection, after which only fragments longer than ~15 nt and shorter than ~45

nt will remain in a pool (3). This pool is sequenced and the 5′ and 3′ ends of each sequenced read are recorded (4). This process generates (5) a Rend-seq track. Section 5 demonstrates a real Rend-seq track for the *B. subtilis* gene rpsB.

Data Processing Techniques for Transcript End Finding:

As can be seen in Figure 1.2 Rend-seq data is characterized by long stretches of a fluctuating baseline level of read density that can shift in their mean by several orders of magnitude. These stretches of read density are periodically interrupted by sharp peaks in read density corresponding to a transcript boundary in the original sample. The main analysis goal of Rend-seq data is the rapid, automated and accurate identification of these peaks, but this task - easy to achieve by visual inspection, can be none-the-less challenging to achieve for all locations in the transcriptome.

The first challenging aspect with working with this sort of data is that it can vary in abundance over several orders of magnitude, meaning no simple global solution, such as a threshold, will suffice. To get around this limitation it is necessary to first pre-process the data to convert it to a form where high abundance regions are comparable to low abundance regions.

In the past practitioners of Rend-seq have performed a localized-z-score normalization of the data. This is done by taking the window of reads surrounding every sequenced position and calculating its mean and standard deviation. These two values are then

used to normalize the reads density at a given position to instead reflect a Z-score, with respect to the assumed approximately gaussian context which surrounds it.

The problem with this approach is that the symmetric approximately gaussian assumption does not hold for large stretches of Rend-seq data. Take for example the boundary of a transcript. At the boundaries the distribution undergoes a jump in read density such that the mean of the upstream and downstream distributions may differ by several orders of magnitude. Treating these two populations as the same will inevitably contribute to an over-inflated estimate of the standard deviation, potentially suppressing the z score of true peaks.

Another case where this assumption breaks down is with respect to transcription units that may be close together, as can be the case for example when there are multiple transcription start sites close together. In these cases the issue one runs into is that the peak corresponding to one end of a distribution can now show up in the “background distribution” of another peak , potentially suppressing the z scores of both true peaks due once again to an overestimation of the standard deviation.

Both of these edge cases underscore the need for a method for pre-processing/normalizing raw read counts that doesn’t miss true peaks by the breakdown of its assumption of normality. Furthermore, it is likely that even after this normalization of reads it could be worth exploring other methods that transform these normalized data into annotated peaks.

Rendseq Summary

The study of gene-expression regulation requires precise measurements of the sites of regulation. This necessitates the ability to accurately identify where transcripts begin and end as sequence encoded signals of transcriptional regulation in bacteria are often concurrent with the sites of transcript ends. An expanded set of transcript boundary locations across a variety of conditions will inform models which can predict and design gene expression levels from sequence alone. It will also aid in the understanding of how different sequence contexts inform the regulation of transcripts across a variety of conditions. A database of transcription start sites, RNA decay processing sites and transcription termination sites could also serve as modular building parts in a bioengineering tool box to construct transcripts with sequence to expression relationships matching the problem at hand.

Methods, such as Rend-seq, exist which can aid in the inference of 3′ and 5′ ends of transcripts *in vivo* across a variety of genetic and environmental conditions. While the current method for identifying transcript ends works well for most use cases, it struggles in cases where transcript ends are too close together, or when the step in density is too high. Furthermore, the information embedded in Rend-seq datasets is not easily accessible to the broader scientific community, limiting its ability to power predictive models for gene-expression or inform researchers about the precise start and end position of transcripts they are studying. Both of these current shortcomings of the Rend-seq analysis ecosystem suggest a need for a more thorough approach to end identification for all transcripts in Rend-seq data-set as well as a more streamlined and approachable method for accessing and viewing Rend-seq datasets.

Introduction References

Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., Pak, R. A., Gray, A. N., Gross, C. A., Dixit, A., Parnas, O., Regev, A., & Weissman, J. S. (2016). A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*, *167* (7), 1867–1882.e21.

Agarwal, V., & Kelley, D. R. (2022). The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biology*, *23* (1), 1–28.

Apirion, D. (1973). Degradation of RNA in Escherichia coli. *Molecular & General Genetics: MGG*, *122* (4), 313–322.

Bakal, C., Aach, J., Church, G., & Perrimon, N. (2007). Quantitative morphological signatures define local signaling networks regulating cell morphology. *Science*, *316* (5832), 1753–1756.

Balakrishnan, R., Mori, M., Segota, I., Zhang, Z., Aebersold, R., Ludwig, C., & Hwa, T. (2022). Principles of gene regulation quantitatively connect DNA to RNA and proteins in bacteria. *Science*, *378* (6624), eabk2066.

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., & Theis, F. J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, *38* (12), 1408–1414.

Bergen, V., Soldatov, R. A., Kharchenko, P. V., & Theis, F. J. (2021). RNA velocity-current challenges and future perspectives. *Molecular Systems Biology*, *17* (8), e10282.

Berget, S. M., Moore, C., & Sharp, P. A. (1977). Spliced segments at the 5′ terminus of adenovirus 2 late mRNA*. *Proceedings of the National Academy of Sciences*, *74* (8), 3171–3175.

Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S. M., Singer, R. H., & Long, R. M. (1998). Localization of ASH1 mRNA particles in living yeast. *Molecular Cell*, *2* (4), 437–445.

Bhattarai-Kline, S., Lear, S. K., Fishman, C. B., Lopez, S. C., Lockshin, E. R., Schubert, M. G., Nivala, J., Church, G. M., & Shipman, S. L. (2022). Recording gene expression order in DNA by CRISPR addition of retron barcodes. *Nature*, *608* (7921), 217–225.

Bidnenko, V., Nicolas, P., Grylak-Mielnicka, A., Delumeau, O., Auger, S., Aucouturier, A., Guerin, C., Repoila, F., Bardowski, J., Aymerich, S., & Bidnenko, E. (2017). Termination factor Rho: From the control of pervasive transcription to cell fate determination in Bacillus subtilis. *PLoS Genetics*, *13* (7), e1006909.

Bock, C., Datlinger, P., Chardon, F., Coelho, M. A., Dong, M. B., Lawson, K. A., Lu, T., Maroc, L., Norman, T. M., Song, B., Stanley, G., Chen, S., Garnett, M., Li, W., Moffat, J., Qi, L. S., Shapiro, R. S., Shendure, J., Weissman, J. S., & Zhuang, X. (2022). High-content CRISPR screening. *Nature Reviews Methods Primers*, *2* (1), 1–23.

Browning, D. F., & Busby, S. J. (2004). The regulation of bacterial transcription initiation. *Nature Reviews. Microbiology*, *2* (1), 57–65.

Cao, J., Arha, M., Sudrik, C., Mukherjee, A., Wu, X., & Kane, R. S. (2015). A universal strategy for regulating mRNA translation in prokaryotic and eukaryotic cells. *Nucleic Acids Research*, *43* (8), 4353–4362.

Chen, J., Nikolaitchik, O., Singh, J., Wright, A., Bencsics, C. E., Coffin, J. M., Ni, N., Lockett, S., Pathak, V. K., & Hu, W.-S. (2009). High efficiency of HIV-1 genomic RNA packaging and heterozygote formation revealed by single virion analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *106* (32), 13535–13540.

Chen, W., Guillaume-Gentil, O., Rainer, P. Y., Gäbelein, C. G., Saelens, W., Gardeux, V., Klaeger, A., Dainese, R., Zachara, M., Zambelli, T., Vorholt, J. A., & Deplancke, B. (2022). Live-seq enables temporal transcriptomic recording of single cells. *Nature*, *608* (7924), 733–740.

Chen, X., Teichmann, S. A., & Meyer, K. B. (2018). From Tissues to Cell Types and Back: Single-Cell

Gene Expression Analysis of Tissue Architecture. *Annual Review of Biomedical Data Science*, *1* (1), 29–51.

Crosetto, N., Bienko, M., & van Oudenaarden, A. (2015). Spatially resolved transcriptomics and beyond. *Nature Reviews. Genetics*, *16* (1), 57–66.

Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D., & Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, *14* (3), 297–301.

Deutscher, M. P. (1985). E. coli RNases: Making sense of alphabet soup. *Cell*, *40* (4), 731–732.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., & Regev, A. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, *167* (7), 1853–1866.e17.

Doench, J. G. (2018). Am I ready for CRISPR? A user's guide to genetic screens. *Nature Reviews. Genetics*, *19* (2), 67–80.

Echeverri, C. J., & Perrimon, N. (2006). High-throughput RNAi screening in cultured cells: a user's guide. *Nature Reviews. Genetics*, *7* (5), 373–384.

Eisenstein, M. (2005). Getting the message. *Nature Methods*, *2* (12), 996–996.

Emrich, S. J., Barbazuk, W. B., Li, L., & Schnable, P. S. (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, *17* (1), 69–73.

Feklístov, A., Sharon, B. D., Darst, S. A., & Gross, C. A. (2014). Bacterial sigma factors: a historical, structural, and genomic perspective. *Annual Review of Microbiology*, *68*, 357–376.

Ferguson, M. L., & Larson, D. R. (2013). Measuring transcription dynamics in living cells using fluctuation analysis. *Methods in Molecular Biology* , *1042*, 47–60.

Glotzer, J. B., Saffrich, R., Glotzer, M., & Ephrussi, A. (1997). Cytoplasmic flows localize injected oskar

RNA in Drosophila oocytes. *Current Biology: CB*, *7* (5), 326–337.

Gödecke, A. (2018). qPCR-25 years old but still a matter of debate. *Cardiovascular Research*, *114* (2), 201–202.

Gorin, G., Fang, M., Chari, T., & Pachter, L. (2022). RNA velocity unraveled. *PLoS Computational Biology*, *18* (9), e1010492.

Gray, J. M., Harmin, D. A., Boswell, S. A., Cloonan, N., Mullen, T. E., Ling, J. J., Miller, N., Kuersten, S., Ma, Y.-C., McCarroll, S. A., Grimmond, S. M., & Springer, M. (2014). SnapShot-Seq: a method for extracting genome-wide, in vivo mRNA dynamics from a single total RNA sample. *PloS One*, *9* (2), e89673.

Guo, X., Myasnikov, A. G., Chen, J., Crucifix, C., Papai, G., Takacs, M., Schultz, P., & Weixlbaumer, A. (2018). Structural Basis for NusA Stabilized Transcriptional Pausing. *Molecular Cell*, *69* (5), 816–827.e4.

Gusarov, I., & Nudler, E. (1999). The mechanism of intrinsic transcription termination. *Molecular Cell*, *3* (4), 495–504.

Häkkinen, A., & Ribeiro, A. S. (2016). Characterizing rate limiting steps in transcription from RNA production times in live cells. *Bioinformatics* , *32* (9), 1346–1352.

Hao, Z., Epshtein, V., Kim, K. H., Proshkin, S., Svetlov, V., Kamarthapu, V., Bharati, B., Mironov, A., Walz, T., & Nudler, E. (2021). Pre-termination Transcription Complex: Structure and Function. *Molecular Cell*, *81* (2), 281–292.e8.

Hawley, D. K., & McClure, W. R. (1983). Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucleic Acids Research*, *11* (8), 2237–2255.

Helmann, J. D. (2002). The extracytoplasmic function (ECF) sigma factors. In *Advances in Microbial Physiology* (Vol. 46, pp. 47–110). Academic Press.

Herzog, V. A., Reichholf, B., Neumann, T., Rescheneder, P., Bhat, P., Burkard, T. R., Wlotzka, W., von

Haeseler, A., Zuber, J., & Ameres, S. L. (2017). Thiol-linked alkylation of RNA to assess expression dynamics. *Nature Methods*, *14* (12), 1198–1204.

Holt, R. A., & Jones, S. J. M. (2008). The new paradigm of flow cell sequencing. *Genome Research*, *18* (6), 839–846.

Horns, F., Martinez, J. A., Fan, C., Haque, M., Linton, J. M., Tobin, V., Santat, L., Maggiolo, A. O., Bjorkman, P. J., Lois, C., & Elowitz, M. B. (2023). Engineering RNA export for measurement and manipulation of living cells. *Cell*, *186* (17), 3642–3658.e32.

Huch, S., Nersisyan, L., Ropat, M., Barrett, D., Wu, M., Wang, J., Valeriano, V. D., Vardazaryan, N., Huerta-Cepas, J., Wei, W., Du, J., Steinmetz, L. M., Engstrand, L., & Pelechano, V. (2023). Atlas of mRNA translation and decay for bacteria. *Nature Microbiology*, *8* (6), 1123–1136.

Huemer, M., Mairpady Shambat, S., Brugger, S. D., & Zinkernagel, A. S. (2020). Antibiotic resistance and persistence-Implications for human health and treatment perspectives. *EMBO Reports*, *21* (12), e51034.

Hu, Y., Liu, C., Han, W., & Wang, P. (2023). A theoretical framework of immune cell phenotypic classification and discovery. *Frontiers in Immunology*, *14*, 1128423.

Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, *50* (8), 1–14.

Ishiguro, S., Mori, H., & Yachie, N. (2019). DNA event recorders send past information of cells to the time of observation. *Current Opinion in Chemical Biology*, *52*, 54–62.

Jiao, C., Reckstadt, C., König, F., Homberger, C., Yu, J., Vogel, J., Westermann, A. J., Sharma, C. M., & Beisel, C. L. (2023). RNA recording in single bacterial cells using reprogrammed tracrRNAs. *Nature Biotechnology*, *41*, 1107–1116.

Jin, X., Simmons, S. K., Guo, A., Shetty, A. S., Ko, M., Nguyen, L., Jokhi, V., Robinson, E., Oyler, P., Curry, N., Deangeli, G., Lodato, S., Levin, J. Z., Regev, A., Zhang, F., & Arlotta, P. (2020). In vivo

Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science*, *370* (6520). https://doi.org/10.1126/science.aaz6063

Johnson, G. E., Lalanne, J.-B., Peters, M. L., & Li, G.-W. (2020). Functionally uncoupled transcription-translation in Bacillus subtilis. *Nature*, *585* (7823), 124–128.

Josaitis, C. A., Gaal, T., & Gourse, R. L. (1995). Stringent control and growth-rate-dependent control have nonidentical promoter sequence requirements. *Proceedings of the National Academy of Sciences of the United States of America*, *92* (4), 1117–1121.

Kang, W., Ha, K. S., Uhm, H., Park, K., Lee, J. Y., Hohng, S., & Kang, C. (2020). Transcription reinitiation by recycling RNA polymerase that diffuses on DNA after releasing terminated RNA. *Nature Communications*, *11* (1), 450.

Kazmierczak, M. J., Wiedmann, M., & Boor, K. J. (2005). Alternative sigma factors and their roles in bacterial virulence. *Microbiology and Molecular Biology Reviews: MMBR*, *69* (4), 527–543.

Lalanne, J.-B., Taggart, J. C., Guo, M. S., Herzel, L., Schieler, A., & Li, G.-W. (2018). Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry. *Cell*, *173* (3), 749–761.e38.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., … Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, *560* (7719), 494–498.

Li, G.-W., Burkhardt, D., Gross, C., & Weissman, J. S. (2014). Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, *157* (3), 624–635.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, *133* (3), 523–536.

Liu, J., & Cao, X. (2015). Regulatory dendritic cells in autoimmunity: A comprehensive review. *Journal*

of Autoimmunity, *63*, 1–12.

Li, X., Kim, H., Litke, J. L., Wu, J., & Jaffrey, S. R. (2020). Fluorophore-Promoted RNA Folding and Photostability Enables Imaging of Single Broccoli-Tagged mRNAs in Live Mammalian Cells. *Angewandte Chemie* , *59* (11), 4511–4518.

Mandell, Z. F., Oshiro, R. T., Yakhnin, A. V., Vishwakarma, R., Kashlev, M., Kearns, D. B., & Babitzke, P. (2021). NusG is an intrinsic transcription termination factor that stimulates motility and coordinates gene expression with NusA. *eLife*, *10*. https://doi.org/10.7554/eLife.61880

Mandell, Z. F., Zemba, D., & Babitzke, P. (2022). Factor-stimulated intrinsic termination: getting by with a little help from some friends. *Transcription*, *13* (4-5), 96–108.

Marx, V. (2021). Method of the Year: spatially resolved transcriptomics. *Nature Methods*, *18* (1), 9–14.

Mejía-Almonte, C., Busby, S. J. W., Wade, J. T., van Helden, J., Arkin, A. P., Stormo, G. D., Eilbeck, K., Palsson, B. O., Galagan, J. E., & Collado-Vides, J. (2020). Redefining fundamental concepts of transcription initiation in bacteria. *Nature Reviews. Genetics*, *21* (11), 699–714.

Mitra, P., Ghosh, G., Hafeezunnisa, M., & Sen, R. (2017). Rho Protein: Roles and Mechanisms. *Annual Review of Microbiology*, *71*, 687–709.

Mohanty, B. K., & Kushner, S. R. (2016). Regulation of mRNA Decay in Bacteria. *Annual Review of Microbiology*, *70*, 25–44.

Molodtsov, V., Wang, C., Firlar, E., Kaelber, J. T., & Ebright, R. H. (2023). Structural basis of Rho-dependent transcription termination. *Nature*, *614* (7947), 367–374.

Muzzey, D., & van Oudenaarden, A. (2009). Quantitative time-lapse fluorescence microscopy in single cells. *Annual Review of Cell and Developmental Biology*, *25*, 301–327.

Nelles, D. A., Fang, M. Y., O'Connell, M. R., Xu, J. L., Markmiller, S. J., Doudna, J. A., & Yeo, G. W. (2016). Programmable RNA Tracking in Live Cells with CRISPR/Cas9. *Cell*, *165* (2), 488–496.

Niemitz, E. (2007). The microarray revolution. *Nature Reviews. Genetics*, *8* (1), S15–S15.

Noyes, B. E., & Stark, G. R. (1975). Nucleic acid hybridization using DNA covalently coupled to cellulose. *Cell*, *5* (3), 301–310.

Paget, M. S. B., & Helmann, J. D. (2003). The sigma70 family of sigma factors. *Genome Biology*, *4* (1), 203.

Paige, J. S., Wu, K. Y., & Jaffrey, S. R. (2011). RNA mimics of green fluorescent protein. *Science*, *333* (6042), 642–646.

Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P., & Fodor, S. P. (1994). Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *91* (11), 5022–5026.

Peña, E., Heinlein, M., & Sambade, A. (2015). In Vivo RNA Labeling Using MS2. In M. Heinlein (Ed.), *Plasmodesmata: Methods and Protocols* (pp. 329–341). Springer New York.

Pervez, M. T., Hasnain, M. J. U., Abbas, S. H., Moustafa, M. F., Aslam, N., & Shah, S. S. M. (2022). A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. *BioMed Research International*, *2022*, 3457806.

Peterman, N., & Levine, E. (2016). Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics*, *17*, 206.

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., & Lim, W. A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, *152* (5), 1173–1183.

Qiu, Q., Hu, P., Qiu, X., Govek, K. W., Cámara, P. G., & Wu, H. (2020). Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. *Nature Methods*, *17* (10), 991–1001.

Qiu, X., Zhang, Y., Martin-Rufino, J. D., Weng, C., Hosseinzadeh, S., Yang, D., Pogson, A. N., Hein, M. Y., Hoi Joseph Min, K., Wang, L., Grody, E. I., Shurtleff, M. J., Yuan, R., Xu, S., Ma, Y., Replogle, J. M., Lander, E. S., Darmanis, S., Bahar, I., … Weissman, J. S. (2022). Mapping transcriptomic vector

fields of single cells. *Cell*, *185* (4), 690–711,e45.

Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., Amit, I., & Regev, A. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature Biotechnology*, *29* (5), 436–442.

Rashid, F., & Berger, J. (2023, January 25). *Protein structure terminates doubt about how transcription stops*. Nature Publishing Group UK. https://doi.org/10.1038/d41586-023-00121-1

Ray-Soni, A., Bellecourt, M. J., & Landick, R. (2016). Mechanisms of Bacterial Transcription Termination: All Good Things Must End. *Annual Review of Biochemistry*, *85*, 319–347.

Richards, J., Sundermeier, T., Svetlanov, A., & Karzai, A. W. (2008). Quality control of bacterial mRNA decoding and decay. *Biochimica et Biophysica Acta*, *1779* (9), 574–582.

Ross, W., Gosink, K. K., Salomon, J., Igarashi, K., Zou, C., Ishihama, A., Severinov, K., & Gourse, R. L. (1993). A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science*, *262* (5138), 1407–1413.

Said, N., Hilal, T., Sunday, N. D., Khatri, A., Bürger, J., Mielke, T., Belogurov, G. A., Loll, B., Sen, R., Artsimovitch, I., & Wahl, M. C. (2021). Steps toward translocation-independent RNA polymerase inactivation by terminator ATPase ρ. *Science*, *371* (6524). https://doi.org/10.1126/science.abd1673

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., & Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, *230* (4732), 1350–1354.

Sarfraz, N., Moscoso, E., Oertel, T., Lee, H. J., Ranjit, S., & Braselmann, E. (2023). Visualizing orthogonal RNAs simultaneously in live mammalian cells by fluorescence lifetime imaging microscopy (FLIM). *Nature Communications*, *14* (1), 867.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of

single-cell gene expression data. *Nature Biotechnology*, *33* (5), 495–502.

Schmidt, F., Cherepkova, M. Y., & Platt, R. J. (2018). Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature*, *562* (7727), 380–385.

Schmidt, F., Zimmermann, J., Tanna, T., Farouni, R., Conway, T., Macpherson, A. J., & Platt, R. J. (2022). Noninvasive assessment of gut function using transcriptional recording sentinel cells. *Science*, *376* (6594), eabm6038.

Schönberger, J., Hammes, U. Z., & Dresselhaus, T. (2012). In vivo visualization of RNA in plants cells using the λN_{22} system and a GATEWAY-compatible vector series for candidate RNAs. *The Plant Journal: For Cell and Molecular Biology*, *71* (1), 173–181.

Schraivogel, D., Gschwind, A. R., Milbank, J. H., Leonce, D. R., Jakob, P., Mathur, L., Korbel, J. O., Merten, C. A., Velten, L., & Steinmetz, L. M. (2020). Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nature Methods*, *17* (6), 629–635.

Seshasayee, A. S. N., Sivaraman, K., & Luscombe, N. M. (2011). An overview of prokaryotic transcription factors : a summary of function and occurrence in bacterial genomes. *Sub-Cellular Biochemistry*, *52*, 7–23.

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N., Schwartz, S., Fowler, B., Weaver, S., Wang, J., Wang, X., Ding, R., Raychowdhury, R., Friedman, N., Hacohen, N., … Regev, A. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, *510* (7505), 363–369.

Sheth, R. U., & Wang, H. H. (2018). DNA-based memory devices for recording cellular events. *Nature Reviews. Genetics*, *19* (11), 718–732.

Sheth, R. U., Yim, S. S., Wu, F. L., & Wang, H. H. (2017). Multiplex recording of cellular events over time on CRISPR biological tape. *Science*, *358* (6369), 1457–1461.

Silas, S., Makarova, K. S., Shmakov, S., Páez-Espino, D., Mohr, G., Liu, Y., Davison, M., Roux, S.,

Krishnamurthy, S. R., Fu, B. X. H., Hansen, L. L., Wang, D., Sullivan, M. B., Millard, A., Clokie, M. R., Bhaya, D., Lambowitz, A. M., Kyrpides, N. C., Koonin, E. V., & Fire, A. Z. (2017). On the Origin of Reverse Transcriptase-Using CRISPR-Cas Systems and Their Hyperdiverse, Enigmatic Spacer Repertoires. *mBio*, *8* (4), e00897–17.

Silas, S., Mohr, G., Sidote, D. J., Markham, L. M., Sanchez-Amat, A., Bhaya, D., Lambowitz, A. M., & Fire, A. Z. (2016). Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science*, *351* (6276), aad4234.

Sinha, D. K., Banerjee, B., Maharana, S., & Shivashankar, G. V. (2008). Probing the dynamic organization of transcription compartments and gene loci within the nucleus of living cells. *Biophysical Journal*, *95* (11), 5432–5438.

Southern, E. M., Maskos, U., & Elder, J. K. (1992). Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: Evaluation using experimental models. *Genomics*, *13* (4), 1008–1017.

Steinmetz, L. M., Scharfe, C., Deutschbauer, A. M., Mokranjac, D., Herman, Z. S., Jones, T., Chu, A. M., Giaever, G., Prokisch, H., Oefner, P. J., & Davis, R. W. (2002). Systematic screen for human disease genes in yeast. *Nature Genetics*, *31* (4), 400–404.

Strack, R. L., Disney, M. D., & Jaffrey, S. R. (2013). A superfolding Spinach2 reveals the dynamic nature of trinucleotide repeat-containing RNA. *Nature Methods*, *10* (12), 1219–1224.

Sunbul, M., Lackner, J., Martin, A., Englert, D., Hacene, B., Grün, F., Nienhaus, K., Nienhaus, G. U., & Jäschke, A. (2021). Super-resolution RNA imaging using a rhodamine-binding aptamer with fast exchange kinetics. *Nature Biotechnology*, *39* (6), 686–690.

Sutherland, C., & Murakami, K. S. (2018). An Introduction to the Structure and Function of the Catalytic Core Enzyme of Escherichia coli RNA Polymerase. *EcoSal Plus*, *8* (1). https://doi.org/10.1128/ecosalplus.ESP-0004-2018

Swerdlow, S. J., & Schaaper, R. M. (2014). Mutagenesis in the lacI gene target of E. coli: Improved analysis for lacId and lacO mutants. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *770*, 79–84.

Tanna, T., Schmidt, F., Cherepkova, M. Y., Okoniewski, M., & Platt, R. J. (2020). Recording transcriptional histories using Record-seq. *Nature Protocols*, *15* (2), 513–539.

Taylor, S. C., Nadeau, K., Abbasi, M., Lachance, C., Nguyen, M., & Fenrich, J. (2019). The Ultimate qPCR Experiment: Producing Publication Quality, Reproducible Data the First Time. *Trends in Biotechnology*, *37* (7), 761–774.

Travers, A. A., Lamond, A. I., & Weeks, J. R. (1986). Alteration of the growth-rate-dependent regulation of Escherichia coli tyrT expression by promoter mutations. *Journal of Molecular Biology*, *189* (1), 251–255.

Vasilyev, N., Gao, A., & Serganov, A. (2019). Noncanonical features and modifications on the 5'-end of bacterial sRNAs and mRNAs. *Wiley Interdisciplinary Reviews. RNA*, *10* (2), e1509.

Wang, G., Wang, F., Huang, Q., Li, Y., Liu, Y., & Wang, Y. (2015). Understanding Transcription Factor Regulation by Integrating Gene Expression and DNase I Hypersensitive Sites. *BioMed Research International*, *2015*, 757530.

Wang, X., & Zheng, J. (2021). Velo-Predictor: an ensemble learning pipeline for RNA velocity prediction. *BMC Bioinformatics*, *22* (Suppl 10), 419.

Weng, G., Kim, J., & Won, K. J. (2021). VeTra: a tool for trajectory inference based on RNA velocity. *Bioinformatics* , *37* (20), 3509–3513.

Wilkinson, M. E., Charenton, C., & Nagai, K. (2019). RNA Splicing by the Spliceosome. *Annual Review of Biochemistry*, *89*, 359–388.

Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., Chu, A. M., Connelly, C., Davis, K., Dietrich, F., Dow, S. W., El

Bakkoury, M., Foury, F., Friend, S. H., Gentalen, E., … Davis, R. W. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*, *285* (5429), 901–906.

Wu, Q., Medina, S. G., Kushawah, G., DeVore, M. L., Castellano, L. A., Hand, J. M., Wright, M., & Bazzini, A. A. (2019). Translation affects mRNA stability in a codon-dependent manner in human cells. *eLife*, *8*, e45396.

Yang, L.-Z., Wang, Y., Li, S.-Q., Yao, R.-W., Luan, P.-F., Wu, H., Carmichael, G. G., & Chen, L.-L. (2019). Dynamic Imaging of RNA in Living Cells by CRISPR-Cas13 Systems. *Molecular Cell*, *76* (6), 981–997.e7.

Yan, Q., & Fong, S. S. (2017). Study of in vitro transcriptional binding effects and noise using constitutive promoters combined with UP element sequences in Escherichia coli. *Journal of Biological Engineering*, *11*, 33.

Young, J. W., Locke, J. C. W., Altinok, A., Rosenfeld, N., Bacarian, T., Swain, P. S., Mjolsness, E., & Elowitz, M. B. (2011). Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. *Nature Protocols*, *7* (1), 80–88.

Zeisel, A., Köstler, W. J., Molotski, N., Tsai, J. M., Krauthgamer, R., Jacob-Hirsch, J., Rechavi, G., Soen, Y., Jung, S., Yarden, Y., & Domany, E. (2011). Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Molecular Systems Biology*, *7*, 529.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS One*, *9* (1), e78644.

Zou, F., & Bai, L. (2019). Using time-lapse fluorescence microscopy to study gene regulation. *Methods* , *159-160*, 138–145.

Chapter II

Molecular Time Capsules Enable Transcriptomic Recording in Living Cells.

- *Note the work in this chapter is also currently under submission for publication. I would like to acknowledge my co-authors in this work:*
- Jack Rubien (Department of Biology, Massachusetts Institute of Technology; Cambridge USA)
- Dylan McCormick (Department of Biology, Massachusetts Institute of Technology; Cambridge USA. and the Whitehead Institute for Biomedical Research; Cambridge, USA)
- Gene-Wei Li (Department of Biology, Massachusetts Institute of Technology; and the corresponding author for this work: gwli@mit.edu)

Abstract

Live-cell transcriptomic recording can help reveal hidden cellular states that precede phenotypic transformation. Here we demonstrate the use of protein-based encapsulation for preserving samples of cytoplasmic RNAs inside living cells. These molecular time capsules (MTCs) can be induced to create time-stamped transcriptome snapshots, preserve RNAs after cellular transitions, and enable retrospective investigations of gene expression programs that drive distinct developmental trajectories. MTCs also open the possibility to uncover transcriptomes in difficult-to-reach conditions.

Introduction

Gene expression is dynamic and shapes future cellular states, but such temporal trajectories remain challenging to map at scale. Many important biological processes, such as cellular differentiation and disparate survivability under stress, are seeded by subpopulations of cells with distinct transcriptomic signatures(Ackermann, 2015; Moris et al., 2016; Niepel et al., 2009; Norman et al., 2015). Mapping these time-dependent trajectories requires capturing the transcriptome specific to the transient and founding population before a distinguishable phenotype has emerged. However, most methods to probe global gene expression necessitate immediate destruction of the cell, preventing longitudinal tracking of subpopulations. Although single-cell transcriptomics provide an opportunity to dissect populational heterogeneity(Kolodziejczyk et al., 2015; Wagner et al., 2016), they are also end-point measurements and must rely on inference methods, such as RNA velocity(Bergen et al., 2020, 2021; La Manno et al., 2018; Qiu et al., 2022), to postulate temporal dynamics.

Meanwhile, live-cell methods for monitoring gene expression remain limited in coverage, throughput, or temporal-resolution. Even with advanced multicolor fluorescence microscopy, there are insufficient channels to simultaneously monitor the entire genome, therefore necessitating prior knowledge on marker gene selection to investigate expression-to-phenotype trajectories(Lin et al., 2023; Linghu et al., 2023; Young et al., 2011). Live-cell continual RNA extraction using force microscopy (e.g., Live-Seq(Chen et al., 2022)) can achieve full genome coverage and longitudinal tracking of single cells, though it currently has a limited throughput of 100s of cells and relies on specialized equipment.

Recent breakthroughs in gene-expression recording by DNA-editing(Bhattarai-Kline et al., 2022; Jiao et al., 2023; Schmidt et al., 2018; Shipman et al., 2016; Silas et al., 2016; Tang & Liu, 2018) provide another promising avenue for interrogating time-trajectories with a sequencing readout. These genome-based recording techniques currently require hours of active recording to accumulate dozens of events per lineage or are limited to recording a handful of pre-selected gene targets. We reasoned that a whole-transcriptome recording method that can be induced on-demand in a large population of cells could help elucidate the transitory expression changes that drive cellular differentiation. A tool with this capability could also enable the investigation of cell physiology in difficult-to-reach conditions, such as microbes in animal guts(Schmidt et al., 2022; Shalon et al., 2023).

In this work, we present an RNA-preservation method for transcriptome-wide recording inside living cells. Inspired by RNA viruses that use protein capsids to preserve their RNA genomes, we used engineered proteins that can self-assemble into nanoscale cages(Butterfield et al., 2017), encapsulating a fraction of cytoplasmic RNAs nonspecifically. We demonstrated that these “molecular time capsules” can stably maintain time-stamped transcriptomic records even

after changes in cellular states, and that the RNA contents can be extracted reproducibly and unambiguously. We anticipate that this concept will be broadly applicable to different types of scenarios, from studying the origin of heterogeneous responses to drug treatment for bacteria and cancer cells to elucidating the transcriptome trajectories that underlie cellular differentiation.

To establish such a method, we demonstrate that self-assembled protein capsules can be used to encapsulate and preserve a fraction of the cytoplasm in living cells (Fig. 2.1a). Capsule assembly can be controlled on-demand via an inducible promoter, creating a time-stamped record that is maintained in the cell. At a later time, such as when cells become phenotypically distinguishable or accessible, the capsules can be isolated via affinity purification and their content analyzed (Fig. 2.1b). We focus on their RNA content in this study, although these "molecular time capsules" (MTCs) could in principle be also used for proteomic or metabolomic studies.

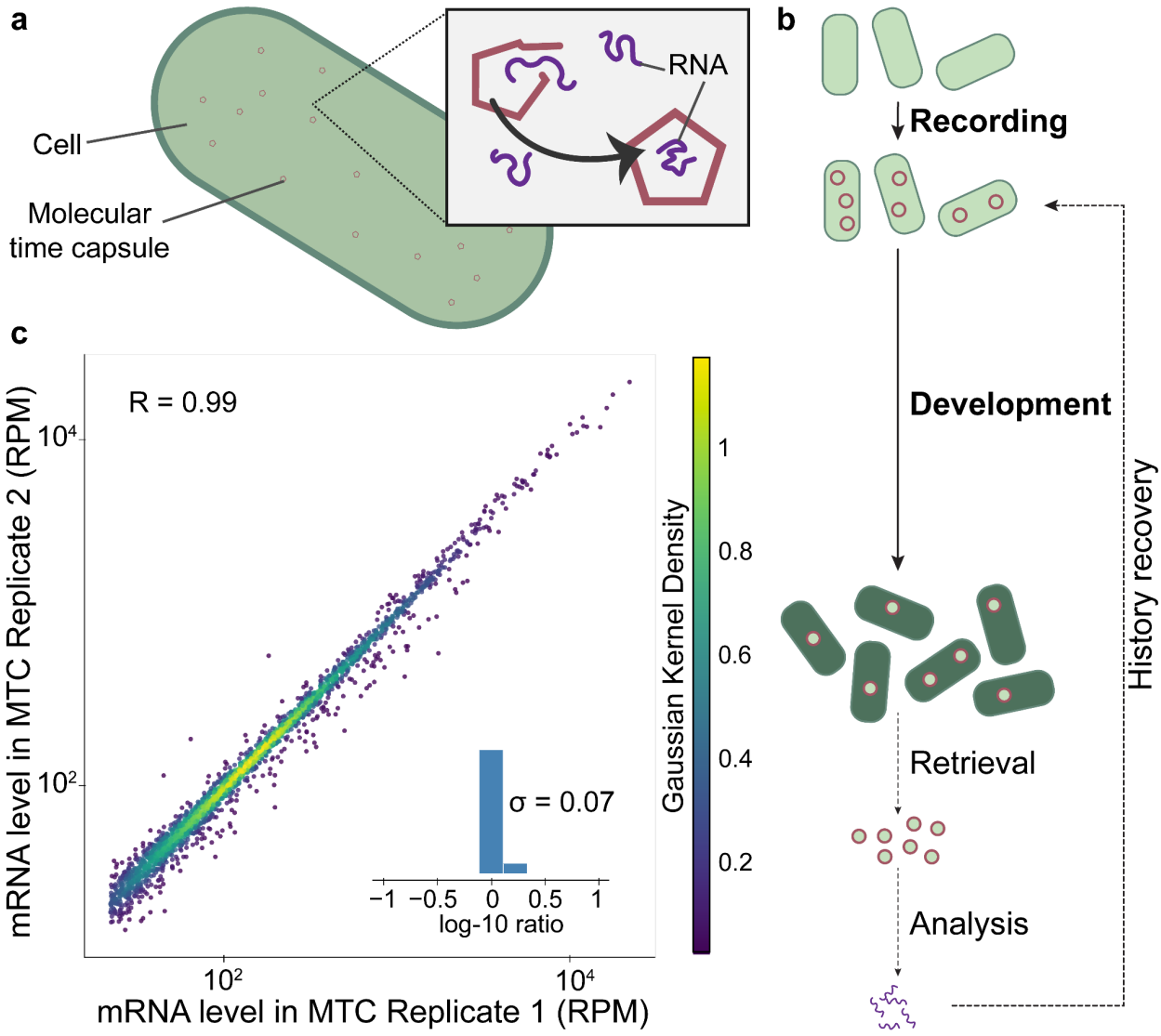

Fig. 2.1 Molecular time capsules provide reproducible transcriptome recording in living cells.

a, Molecular time capsules (MTCs) are genetically encodable protein capsules that can self-assemble within living cells. During assembly, MTCs can encapsulate RNAs (see inset). **b,** MTCs capture and protect snapshots of transcriptomes for delayed retrieval and analysis. After a "recording" period, MTC production is shut off, and the cells are allowed to continue along their developmental trajectory. After the phenotype of interest has emerged, MTCs are retrieved by affinity purification and their encapsulated RNA is extracted. Sequencing analysis is performed on the RNAs, facilitating historical recovery of the transcriptomic record that was captured during the

recording window. **c,** MTC-captured snapshots of the transcriptome across 2 biological replicates plotted in units of reads per million (RPM), where each dot corresponds to a gene with sufficient sequencing coverage. (Pearson correlation coefficient of log-transformed RPM values: R = 0.993, n = 2,302 genes, only genes with > 100 reads are considered). The inset plots the distribution of log-10 fold-change between the two replicates. The standard deviation, **σ**, of the fold-change distribution = 0.07, or 1.2-fold.

Several criteria are required for faithful transcriptome recording by MTCs. First, the captured RNA content must be reproducible. Second, the retrieved RNA content must have minimal contamination from non-encapsulated RNAs, whether from the host cell or from other cells in the sample. Third, the RNA record must be stably preserved over time despite changes in the host transcriptome. Here we demonstrate that these criteria are met using a *de novo* designed protein capsule(Butterfield et al., 2017) (13.5-nm radius) expressed with a poly-histidine tag in *Escherichia coli*, with the recording timing controlled via an Isopropyl ß-D-1-thiogalactopyranoside (IPTG)-inducible promoter.

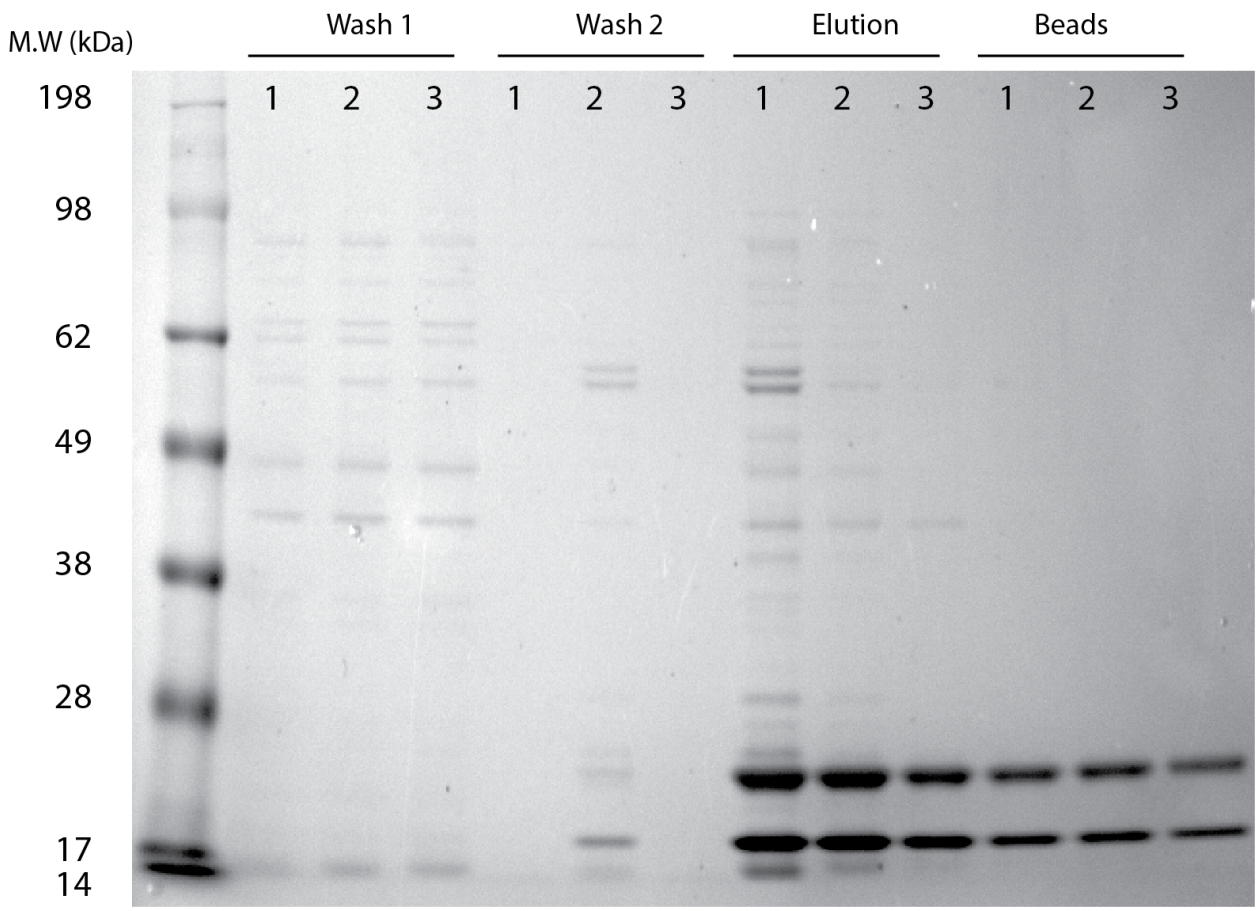

Fig. 2.2 Molecular time capsules can be cleanly purified by protein purification.

Elution of the MTC protein complex is very clean. This protein gel contains samples from various stages of the protein purification prep across 3 different testing conditions. Samples marked "3" represent the purification conditions used throughout this paper. 1 and 2 represent other conditions being tested. 1 = resuspension of cell pellets in 20mM Imidazole and a wash buffer concentration of 20mM Imidazole. There is still significant contamination of the eluate under these conditions. Condition 2 = resuspension in 20mM Imidazole and wash in 150 mM Imidazole. 3 = purification condition used for the work in this paper, resuspend in 150 mM Imidazole and wash in 150 mM Imidazole. All elutions were done at 500 mM Imidazole. Note

that there remains capsules stuck to the beads in our current purification protocol, indicating an avenue for improved yield from this protocol.

Molecular Time Capsule Characterization and Reproducibility

First, we assessed the reproducibility of MTC-based recording by comparing the encapsulated RNA content from three biological replicates. Poly-histidine-tagged MTCs were purified using a Ni-NTA column (Figure 2.2) from exponential-phase cells that have been expressing capsule proteins steadily (for >10 generations). Expression of MTCs only slightly increases the population doubling time (from 20 minutes to 28.5 minutes) (Figure 2.3). When expressed in this experiment, MTC transcripts represent 0.14% of all sequenced transcripts (transcripts per million (TPM) = 1350 +/- 230). Our purification and RNA extraction methods extract RNA specific to MTCs, as we are not able to recover a detectable amount of RNA from wild-type *E. coli* cells subjected to the same protocol.

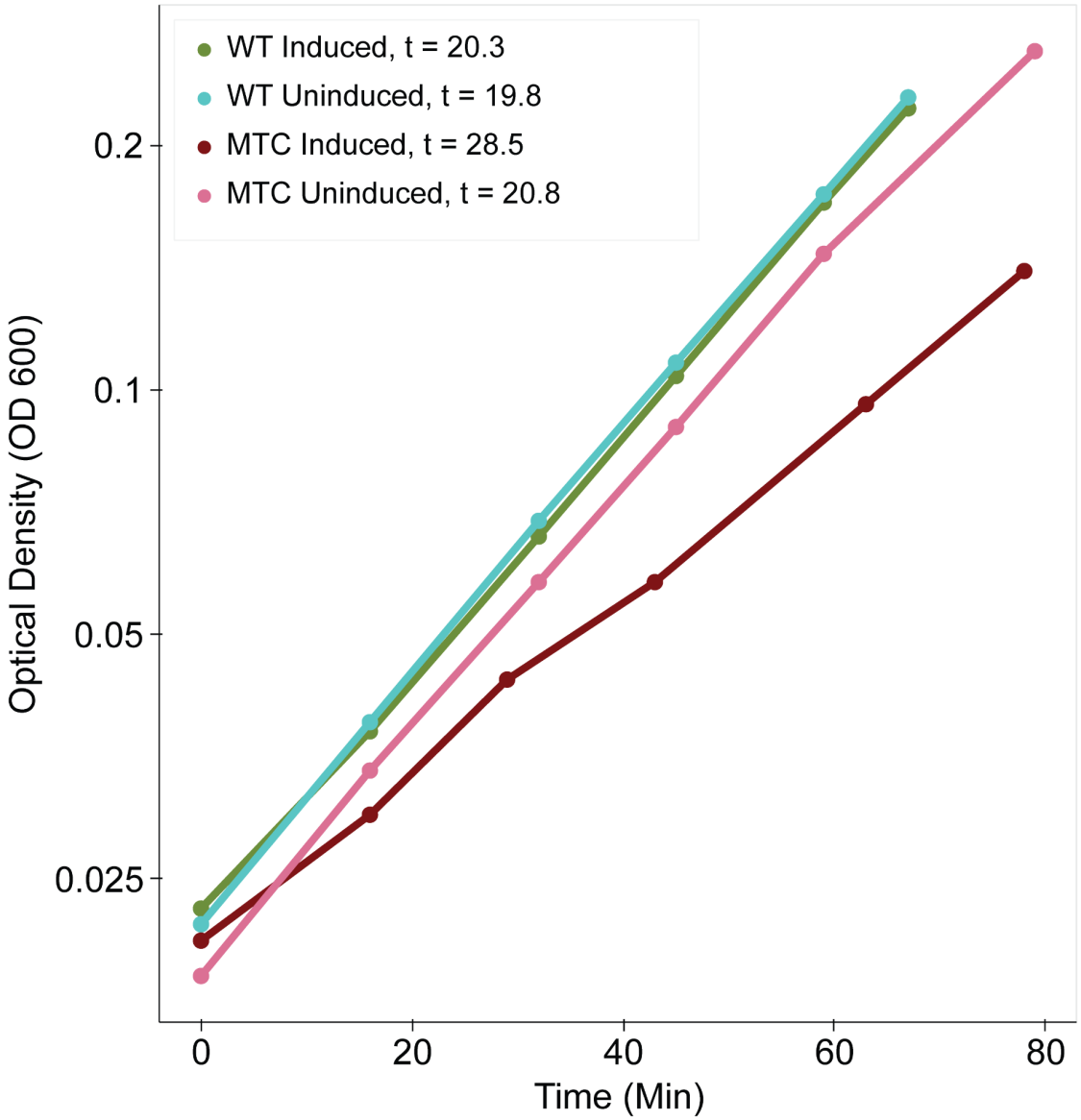

Fig. 2.3 Doubling Time Effects of Extended Induction of MTCs. Inducing the MTCs has a slight effect on growth rate. This is the semi-log plot of optical density (measured at 600nm) of cultures versus time (Minutes). WT corresponds to wild type MG1655 *E. coli*, while the MTC refers to the wild type MG1655 *E. coli* which has pMP026, the MTC plasmid. Induced samples had 1mM IPTG added at the moment of back-dilution from overnight culture (>10 doublings before a detectable OD 600 value). The "t" in the legend above corresponds to the

calculated doubling time for each sample. The induced MTC sample has a doubling time of 28.5, which is ~ 8.5 minutes slower than WT samples.

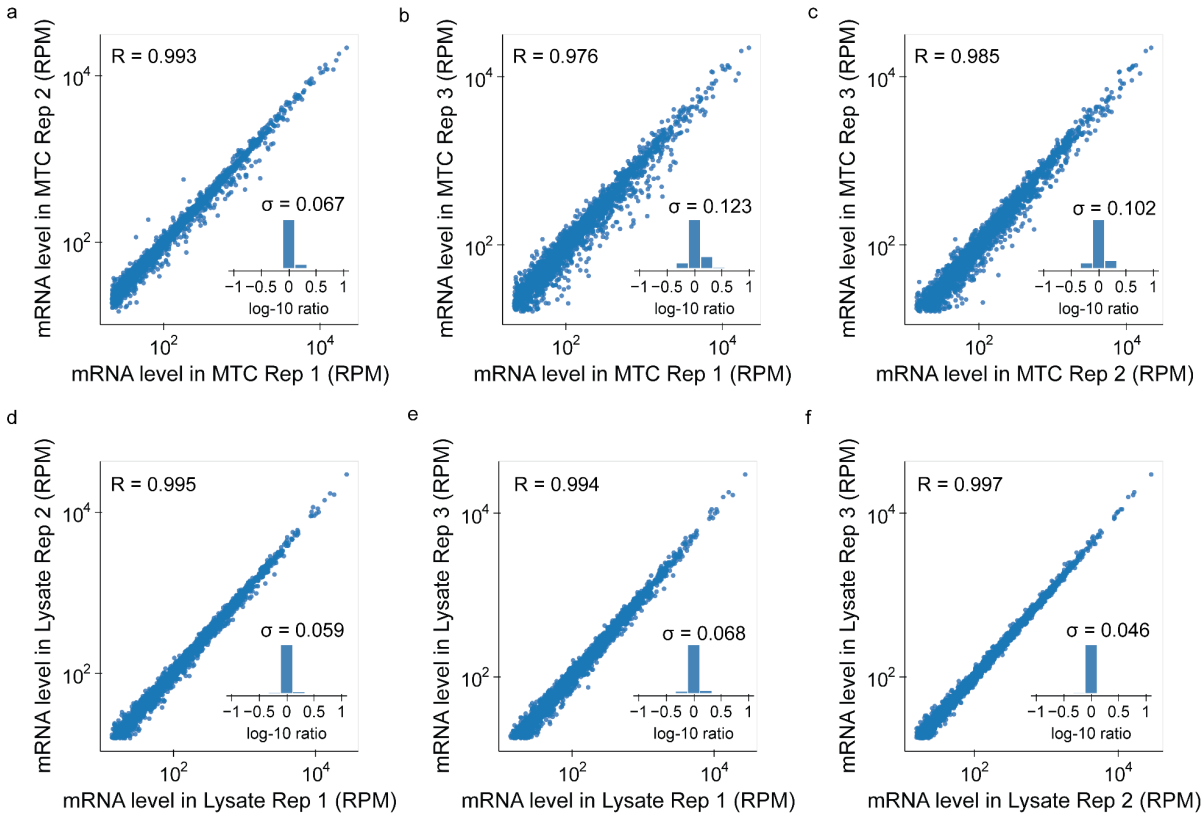

Fig. 2.4 MTC-encapsulated per gene RPM and general Lysate RNA per Gene RPM across 3 biological replicates. MTCs have highly reproducible capture. Subfigures a-c demonstrate the high reproducibility on a gene-by-gene basis of the read per million (RPM) values of MTC-encapsulated mRNAs across 3 biological replicates (between replicates 1 and 2 P = 0.993 (n = 2302); 1 and 3 P = 0.976 (n = 2277); 2 and 3 P = 0.985 (n=2395)). d-f demonstrates the corresponding gene-by-gene RPM values of mRNAs taken from the general lysate (between replicates 1 and 2 P = 0.995 (n = 2430); 1 and 3 P = 0.994 (n = 2374); 2 and 3 P = 0.997 (n=2364)). The high reproducibility across both sets of samples suggests first that the recorded

transcriptome was similar across replicates, and that the MTC-captured transcripts are highly reproducible across similar biological conditions. Note that while the samples are highly similar - there are slight differences between the MTC samples which might point to transient upregulation of certain regulons during the growth period. Previous transient transcriptomic variations would not necessarily be visible at the final time point measurement (i.e. the general lysate samples) so it is not surprising to see a higher degree of agreement between the Lysate samples than across MTC-encapsulated samples. All Pearson coefficients (R) are calculated on log-transformed values (as is plotted) for reads with > 100 sequencing reads. The "n" reported for each Pearson coefficient corresponds to the number of genes used in that calculation.

Overall, the levels of encapsulated mRNAs, as measured by RNA-seq, agree across biological replicates on a gene-by-gene basis. The median Pearson correlation coefficient of log-transformed mRNA levels between replicates is R = 0.985 (Fig. 2.1c and Figure 2.4). This reproducibility suggests that MTCs can facilitate differential expression analysis between cellular states. The mRNA levels in the total cell lysate also correlate with MTC samples, albeit less well compared to the reproducibility of MTC capture across biological replicates (median R = 0.784, Figures 2.4 and 2.5). Interestingly, the majority of RNA fragments recovered from MTCs are short (< 200 nt), suggesting that MTC-based capture leverages the abundant RNA decay intermediates recently reported to dominate the transcriptome(Herzel et al., 2022). (Figure 2.6).

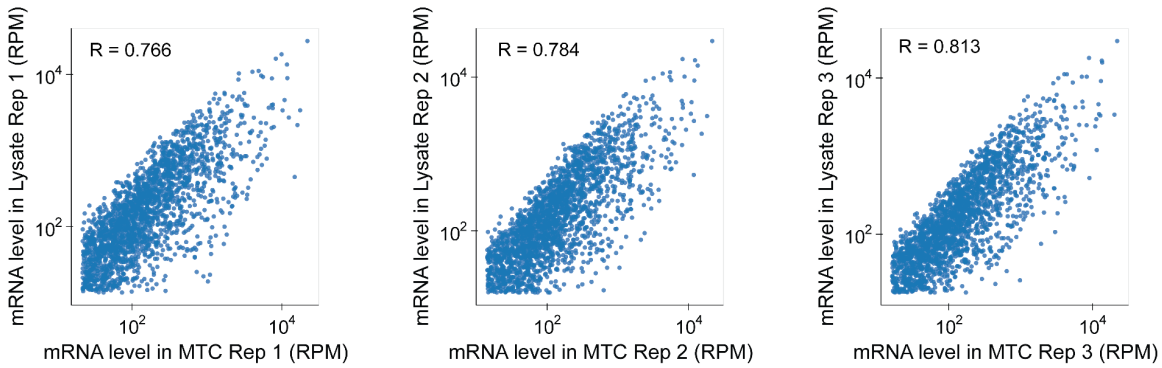

Fig. 2.5 MTC-encapsulated transcripts versus Lysate transcripts across 3 biological replicates. Three biological replicates of MTC-encapsulated mRNAs and lysate purified mRNAs plotted on a gene-by-gene basis with read per million (RPM) value for each sample (between replicates 1 and 2 P = 0.766 (n = 2147); 1 and 3 P = 0.783 (n = 2235); 2 and 3 P = 0.813 (n=2168)). While this represents a good agreement between MTC-encapsulated RNAs and lysate RNAs, it is important to note that the variance which contributes to the spread is highly reproducible - as is demonstrated in Extended Figure 3. This high reproducibility means that in using the MTCs it is best to compare MTCs between two conditions or populations of interest to detect potential up-regulated or downregulated genes rather than looking at the MTC versus lysate RPM values. All Pearson coefficients (R) are calculated on log-transformed values (as is plotted) for reads with > 100 sequencing reads. The "n" reported for each Pearson coefficient corresponds to the number of genes used in that calculation.

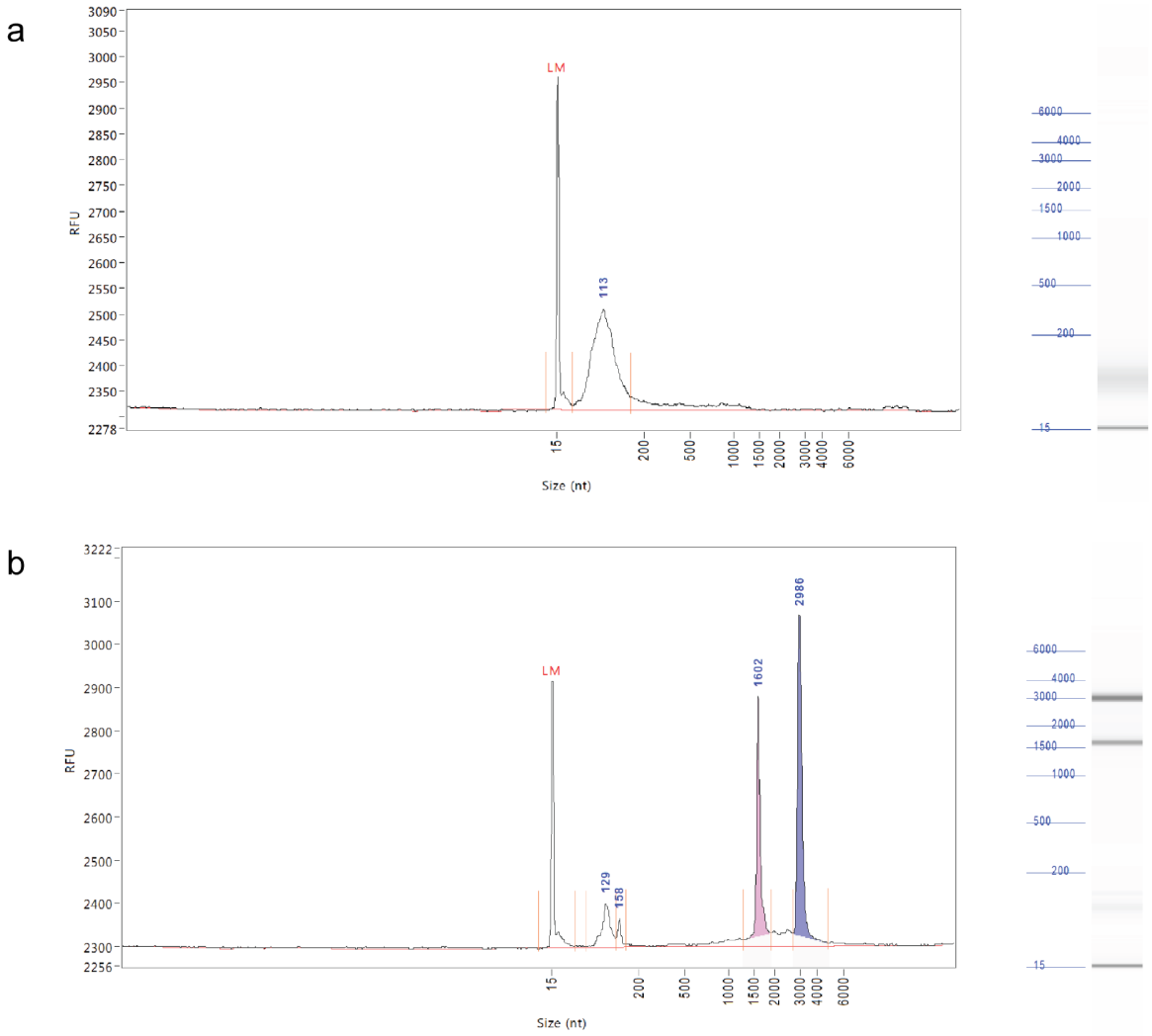

Fig. 2.6 Length Distribution of MTC Encapsulated RNAs in Comparison to General Lysate RNAs. MTCs are dominated by short transcripts. Shown here is the output of the length distribution measured via Fragment Analyzer. **a**, represents the read length distribution of MTC-protected transcripts taken before the typical 200nt size selection step in library prep. **b**, represents the length distribution of the corresponding total lysate sample, which is by contrast dominated by the expected rRNA bands at ~16k nt and ~29-30k nt.

MTC-encapsulated RNA recovery and stability

We next demonstrated that MTC-encapsulated transcripts can be cleanly recovered without substantial contamination from non-encapsulated RNAs. To do so, we purified MTCs from a heterogeneous cell culture in which MTCs are only expressed in one subpopulation (*E. coli,* 50% of cells) but not in the other (*Bacillus subtilis*, 50% of cells) (Fig. 2.7a). These bacterial species have distinct genomes, facilitating the identification of which species a particular RNA originates from RNA-seq analysis.

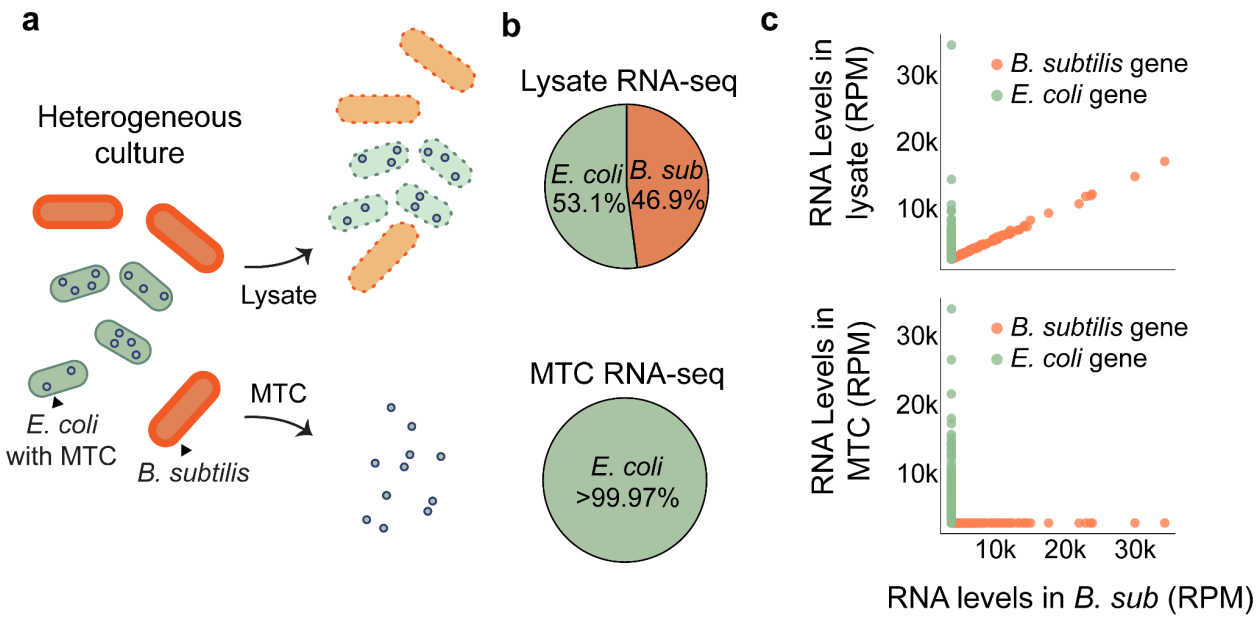

Fig. 2.7 MTC-encapsulated RNAs can be recovered without detectable contaminations.

a, Total lysate and MTC samples are removed from a mixed "barnyard" sample consisting of both *B. subtilis* (orange) and MTC-containing *E. coli* (green). The MTC sample refers to MTCs purified from the mixed population. **b,** The percentage of reads in the MTC sample mapping to the *E. coli* genome is >99.97% (out of n = 5.2 x 10^6

uniquely mapping reads), the mixed lysate sample has 53.1% (out of n = 3.2 x 10^6 uniquely mapping reads) of reads uniquely mapping to *E. coli* genome. **c,** The gene-by-gene scatter plot in units of reads per million (RPM) for samples purified from pure *B. subtilis* lysate (x axis) versus the mixed lysate sample (top y-axis) and the MTC sample (bottom y-axis). *E. coli* genes are colored in green, whereas *B. subtilis* genes are colored in orange.

Although 46.9% of uniquely mapping reads from the heterogeneous lysate mapped to the *B. subtilis* genome, MTCs extracted from the same lysate contain almost exclusively *E. coli* RNAs (99.97%) (Fig. 2.7b). Every *B. subtilis* gene is depleted in MTCs (Fig. 2.7c). The level of *B. subtilis* mapping RNAs (0.03%) in the MTC sample is close to the baseline level estimated using an *E. coli* RNA sample sequenced in the same lane (0.1%). This result indicates that the MTC-purification procedure specifically captures the encapsulated RNAs, and that the transcriptome of a targeted subpopulation can be successfully isolated using MTCs without contamination from non-encapsulated transcripts.

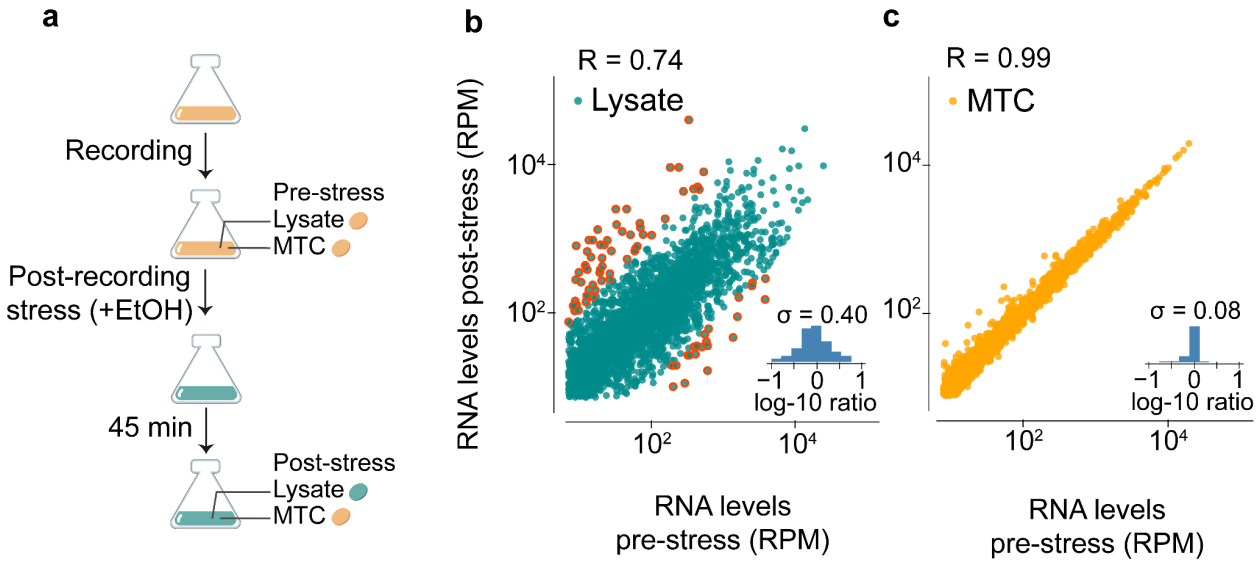

Fig. 2.8 MTCs preserve mRNA records after transcriptome changes.

a, MTC encapsulated RNA samples and lysate samples are collected from before and after a stressful treatment of 4% ethanol. **b and c,** Both plots contrast the reads per million (RPM) of all genes with sufficient sequencing coverage, with each point corresponding to a separate gene. The scatter plot places a dot for each gene corresponding to its pre- (x) and post- (y)stress time-points. Only genes with more than 100 reads in both samples are considered. Each plot has an accompanying inset showing the log-10 fold distribution between the pre- and post-stress samples. **b**, The lysate pre- and post samples (points in blue). For the lysate samples, the Pearson correlation coefficient of log-transformed RPM values is R = 0.744, and the standard deviation (**σ**) of the log-10 fold-change distribution = 0.40, which corresponds to 2.5-fold (n = 2477 genes). Genes with greater than or equal to an order of magnitude change in expression between the pre- and post-stress samples are circled in red. **c,** The pre- and post-stress MTC samples (points in orange). The Pearson correlation coefficient of the log-transformed RPM values is R = 0.991, and the **σ** of the fold-change distribution = 0.08, which corresponds to 1.2-fold, (n = 2543 genes)

Finally, we demonstrated that MTCs provide stable transcriptome storage inside host cells. Many factors could contribute to loss of MTC fidelity over time: the protein capsules could disassemble and reassemble, RNA within MTCs could decay, and expression of MTC proteins may be leaky after recording. To quantify the combined effect of all these potential sources, we examined the maintenance of MTC-encapsulated RNAs before and after shifting their host cells into a different environment that alters their transcriptomes. We first briefly induced MTC expression for one hour by adding IPTG (Isopropyl ß-D-1-thiogalactopyranoside) to Luria Broth. After washing off IPTG to shut off MTC expression, we introduced ethanol stress (4%) to the cells (Fig. 2.8a). The stress led to substantial remodeling of the host transcriptome, with 81 genes changed by >10-fold and an overall Pearson correlation coefficient of R=0.79 between mRNA levels pre- and post-stress (Fig. 2.8b). By contrast, the MTC samples collected before and after stress have almost identical RNA levels on a gene-by-gene basis, with an overall Pearson

correlation coefficient of R=0.99 (Fig. 2.8c). The most extremely shifted gene in the host cell, *tnaA*, changed by 124-fold, whereas this same gene only changed by 1.5-fold across the MTC samples. These results demonstrate that MTC contents remain static despite large contextual changes in the cell state.

Discussion

In summary, we established that MTCs can capture and preserve high-fidelity snapshots of the transcriptome in living cells. This approach opens several avenues of future applications. First, it will help elucidate the gene expression heterogeneities that precede distinct phenotypic outcomes during development, cellular differentiation, or stress survival. To do so, cells carrying pre-assembled MTCs can be sorted based on their final phenotypes, allowing their prior transcriptomes to be analyzed. This is facilitated by the fact that MTC-records are physically maintained inside the cell lineage (Horns et al., 2023). MTCs can therefore nominate candidate genes for targeted studies. Second, MTCs can help elucidate cellular states adopted in hard-to-access locations, where in situ sample collection is difficult. The comparatively short recording time will enable precise capture of transitory responses, such as bacterial cells at specific locations inside an animal (Schmidt et al., 2022). As MTC formation simply requires the self-assembly of two protein subunits, we anticipate that it will be generalizable to new systems, both eukaryotic and prokaryotic.

Methods:

Strains: *B. subtilis* strains were taken directly from *Bacillus subtilis* subsp. Subtilis str. 168. *E. coli* strains were generated from *Escherichia coli*, K-12 MG1655. This includes our wild-type control, which was taken directly from *Escherichia coli*, K-12 MG1655 and an MTC-containing strain which consists of *Escherichia coli*, K-12 MG1655 with the MTC-containing plasmid (described below). The plasmid was constructed using the method described below and transformed first into DH5α cells before being transformed into competent MG1655 cells using the method described below.

Plasmids: pGL002 is a Kanamycin (Kan) resistant plasmid that contains constitutively produced LacI and IPTG-inducible MTC. It has pMB1 origin of replication. The full plasmid map can be found on our GitHub. This plasmid was constructed using the method described below. It was transformed into the *E. coli* strain K-12 MG1655 using the method outlined in greater detail below. This pGL002-containing *E. coli* strain was the one used for all experiments in this work.

Plasmid Assembly: Plasmid components were either amplified using PCR from a pre-existing plasmid, followed by DpnI (NEB #R0176L) digestion, or were ordered directly as gene blocks from IDT. Plasmids were constructed by Gibson assembly, using the protocol and reagents associated with NEB #E2611 with a total volume of 4 µL. Plasmids were then transformed into Zymo Mix & Go DH5α Competent Cells (Zymo T3007) for verification and amplification before re-purification and transformation into MG1655 *E. coli*. All plasmid purification was done using overnight culture and the Zymo Zyppy Plasmid Miniprep Kit (Genesee 11-30).

Plasmids were verified using Sanger sequencing (performed by Quintara), or by whole plasmid sequencing (performed by Plasmidsaurus). The plasmid associated with this work: pGL002 (Figure 2.9), will be deposited to Addgene.

Fig 2.9 Illustration of the plasmid map used for this work. The I53-50-v4N genome represents the MTC transcripts, which comprises two parts: a pentameric and a trimeric subunit. The plasmid is resistant to Kanamycin. It has a pMB1 origin, and uses a copy of lac to repress production of the MTCs.

Media: All experiments were conducted using Luria Broth (LB).

Transformation: For transformation into Zymo Mix & Go DH5α competent cells (Zymo T3007), 100 µL of cells were thawed on ice per reaction. Then 1-4 µL plasmid DNA was added followed by gentle mixing. After 5-minute incubation on ice cells, 400 µL of prewarmed SOB medium was added, and the mixture was incubated in an Eppendorf tube for 1 hr at 37 °C with 300 rpm shaking. The mixture was then spread on pre-warmed agar plates with the appropriate antibiotic (50 µg/mL for MTC-plasmid-containing cells).

For transformation into non-competent wild-type MG1655 *E. coli*, the cells were made competent using the protocol described by Chung et al.(Chung et al., 1989). 3 mL of LB was inoculated with a colony from a fresh agar plate. Cells were incubated at 37 °C 230 rpm for 1.5 to 2 hrs. 200 µL of cells were added to 200 µL ice cold TSS buffer (LB broth containing 10% (wt/vol) polyethylene glycol, 5% (vol/vol) dimethyl sulfoxide, and 50 mM Mg2+ at pH 6.5) and 1 µL plasmid. Cells were vortexed and incubated on ice for 20 - 30 minutes. Then cells were incubated for 45 - 60 min at 37 °C on the thermomixer with 900 rpm shaking. Finally, cells were plated on the appropriate antibiotic (50 µg/mL for MTC-plasmid-containing cells).

Assessing MTC reproducibility. (Experiment shown in Figure 2.1c, Extended Figures 3 & 4) MG1655 containing pGL002 was streaked on a 50 µg/mL Kan marker plate and left to grow overnight at 37 °C. The next day, 3 single colonies were selected from the plates and incubated in separate test tubes with 5 mL LB and 50 µg/mL Kan for 2 hrs. These cultures were then back-diluted to allow for 12 doublings before reaching OD 0.3 in 500 mL of pre-warmed LB with 1 mM IPTG and 50 µg/mL Kan. Shortly before reaching OD 0.3, both the lysate and MTC samples were harvested. MTC samples were harvested by splitting the volume into four 50mL Falcon Tubes and spinning them down for 10 min at 4000 rpm at 4 °C in an Eppendorf 5810R

Centrifuge. The supernatant was discarded, and the cell pellets were frozen for future protein purification. The lysate sample was collected by placing 500 µL of culture directly into a prewarmed RNA extraction solution (see below) and proceeding with RNA extraction.

Assessing the Contamination from bulk lysate. (Experiment shown in Figure 2.7**)** - Two samples were harvested for this experiment - MG1655 containing pGL002 and Wild Type *B. subtilis* cells (BS168).

The MG1655 containing pGL002 cells were streaked from a glycerol stock onto an LB + 50 µg/mL Kan plate and grown overnight. A colony from this plate was selected the next day and added to 10mL of LB with 50 µg/mL Kan and grown overnight at 37 °C on a rotator drum spinning ~ 225 rpm. This overnight culture was then diluted into a volume of 500 mL to achieve 12 doublings before reaching OD 0.3 in pre-warmed LB with 50 µg/mL Kan. The culture was grown at 37 °C on a shaker plate - shaking at ~ 225 rpm. At OD 0.3 1mM IPTG was added. At OD 2 the culture was divided between 50mL Falcon tubes and spun for 10 minutes at 4000 rpm in a pre-chilled (4 °C) Eppendorf 5810R Centrifuge. The Supernatant was then discarded, and the cell pellets were flash frozen in liquid nitrogen and stored at -80 °C.

The *B. subtilis* cells (BS168) were inoculated directly from a glycerol stock into 5 mL LB, then grown for 2 hours at 37 °C on a rotator spinning ~ 225 rpm. This culture was then diluted to achieve 12 doublings before reaching OD 0.3 in pre-warmed LB. At OD 2 the culture was divided between 50 mL Falcon tubes and spun for 10 minutes at 4000 rpm in an Eppendorf 5810R Centrifuge. The Supernatant was then discarded, and the cell pellets were flash frozen in liquid nitrogen and stored at -80 °C.

Assessing the stability of encapsulated RNAs. (Experiment shown in Figure 2.8) - MG1655 containing pGL002 was streaked on a 50 μg/mL Kan marker plate and left to grow overnight at 37 °C. The next day single colonies were selected from the plates and incubated in a test tube with 5 mL LB and 50 μg/mL Kan for 2 hrs. This culture was then back diluted in a 500 mL volume with 50 μg/mL Kan to achieve 10 doublings before reaching OD 0.3. Once the cultures reached OD 0.3 MTC recording was initiated by the addition of 1 mM IPTG. After 1 hr, the capsule induction was shut off by filtering and washing the cells. The filter was washed with half the original cell volume of prewarmed LB. The cells were then resuspended in the original volume of LB without any antibiotics or IPTG. Pre-Stress Lysate and MTC samples were collected 15 minutes post-wash. The MTC sample was collected by filtering 2 volumes of 125 mL of culture. The cells were then resuspended in 2 50 mL Falcon tubes in pre-chilled (4 °C) LB. These tubes were then centrifuged for 10 min at 4000 rpm at 4 °C in an Eppendorf 5810R Centrifuge. The supernatant was discarded, and the cell pellets were frozen for future protein purification. The lysate sample was collected by placing 500 μL of culture directly into a prewarmed RNA extraction solution (see below) and proceeding with RNA extraction. To the remaining 250 mL of culture, we added a 4% ethanol stress, and allowed the cells to grow for 45 minutes before harvesting the post-stress MTC and lysate samples in a manner similar to the one described above.

Table 2.1 Summary and description of all samples collected and sequenced:

Sample Name	Sequencing Library ID	OD 600 of rehydrated cells before Protein Purification (*note - sample diluted before measurement)	RNA yeild (ng/uL) Measured by QuBit HS RNA kit	ng RNA per cell (Note - we assume 8*10^8 E coli cells per mL per OD of 1) We also use the fact that cells were resuspended to a total volume of 40mL before purification	Sample/Harvesting Notes - Note all cultures were grown in Luria Broth (LB). OD's reported ar OD 600 values (where cultures may have been appropriately diluted before measurment).
Lysate Rep 1	230724Li	-	96	-	Collected at OD 0.19
Lysate Rep 2	230724Li	-	128	-	Collected at OD 0.177
Lysate Rep 3	230724Li	-	100	-	Collected at OD 1.64
MTC Rep 1	230724Li	0.99	9.98	3.15E-10	Collected at OD 0.19, then RNA was extracted from purified MTCs
MTC Rep 2	230724Li	0.87	10.8	3.88E-10	Collected at OD 0.177, then RNA was extracted from purified MTCs
MTC Rep 3	230724Li	0.78	4	1.60E-10	Collected at OD 0.164, then RNA was extracted from purified MTCs
Early Cap	230712Li	6.88	6.16	2.80E-11	Induced at OD 0.264, 1 hr post induction cells were filtered and washed, then rehydrated. This sample was collected 15 minutes post resuspension in prewarmed LB.
Early Lysate	230712Li	-	60	-	Induced at OD 0.264, 1 hr post induction cells were filtered and washed, then rehydrated. This sample was collected 15 minutes post resuspension in prewarmed LB.
Late Cap	230712Li	9.84	6.18	1.96E-11	Induced at OD 0.264, 1 hr post induction cells were filtered and washed, then rehydrated. This sample was collected ~ 55 min post resuspension in prewarmed LB (40 minutes after the "Ealry or Pre stress sample)
Late Lysate	230712Li	-	54		Induced at OD 0.264, 1 hr post induction cells were filtered and washed, then rehydrated. This sample was collected ~ 55 min post resuspension in prewarmed LB (40 minutes after the "Ealry or Pre stress sample)

B. sub Lysate	230712Li	-	172		Collected at OD 2.0 by centrifugation, 4000 rpm for 10 min at 4C followed by flash freezing.
E. coli Lysate	230712Li	-	156		Induced at OD 0.3 with 1 mM IPTG, harvested at OD 2.0 by centrifugation, 4000 rpm for 10 min at 4C followed by flash freezing.
Mixture Lysate	230712Li	-	182		The "mixture sample" was made by mixing 9:1 B sub to E coli (based on their OD 600 values, which does not correspond to the ratio of cells due to difference in OD 600 to cell count conversion factors). OD600 of rehydrated E. coli was 6.77. The OD of the rehydrated B. subtilis was 5.56. The mixture lysate sample was taken as 250uL of this resuspended mixture, with 250uL buffer.
E. coli MTC	230712Li	6.77	79.8	3.68E-10	Induced at OD 0.3 with 1 mM IPTG, harvested at OD 2.0 by centrifugation, 4000 rpm for 10 min at 4C followed by flash freezing.
Mixture MTC	230712Li	0.68	4.2	1.93E-10	The "mixture sample" was made by mixing 9:1 B sub to E coli (based on their OD 600 values, which does not correspond to the ratio of cells due to difference in OD 600 to cell count conversion factors). OD600 of rehydrated *E. coli* was 6.77. The OD of the rehydrated *B. subtilis* was 5.56.

Protein Purification: Harvested Cell Pellets were rehydrated in 40 mL Lysis Buffer (150 mM Imidazole, 250 mM NaCl, 25 mM Tris-HCL, pH 8, with Protease Inhibitors) and the OD 600 values of each rehydrated solution were measured. Each sample was then sonicated using the 450W Ultrasonic Homogenizer (10 to 300 mL) from US Solid following a program of 2 seconds on at 25% power, followed by a 6 second off break for a total of 10 minutes. Lysate was then clarified by centrifugation at 4 °C using a Sorvall RC-5B Refrigerated Superspeed Centrifuge at 10,000 rpm for 45 min, after which the clarified supernatant was retained for loading on the protein column. To prepare the protein column we loaded 2 mL (1 mL column volume) of Nickel NTA resin onto the 5 mL Polypropylene columns from Qiagen and the resin buffer was allowed

to drain. The resin was then rinsed with 15 column volumes of ddH2O followed by 15 column volumes of wash buffer (150 mM Imidazole, 250 mM NaCl, 25 mM Tris-HCL, pH 8) to equilibrate the column. After this the clarified supernatant was loaded onto the column. The column was then washed with 15 column volumes of the wash buffer followed by an elution with the elution buffer (500 mM Imidazole, 250 mM NaCl, 25 mM Tris-HCL, pH 8). 3 column volumes of the elution buffer were used, but only the last 2 column volumes were kept. The purified proteins were then treated with RNase A (1 µL/mL 20 °C for 10 minutes) before proceeding with RNA extraction.

RNA Extraction: A prewarmed (65 °C) solution of 500 mL phenol acid chloroform with 29 µL 20% SDS was added to 500 mL of sample in a 1.5 mL Eppendorf (split between multiple tubes if needed). Samples were incubated at 65 °C for 5 minutes at 1,400 rpm, followed by a 5-minute incubation of the samples on ice. Samples were then spun at 20,000 g for 2 minutes after which the top aqueous layer was transferred to a new non-stick tube. 45 µL 3M NaAc (pH 5.4), 500 mL 100% Isopropanol and 1 µL GlycoBlue coprecipitant was added. The tubes were chilled at -80 °C for 30 min and then spun at 20,000 g for 60 min at 4 °C. The supernatant was discarded before 250 mL pre-chilled 80% Ethanol was added and the sample was spun again at 20,000 g for 5 min. The supernatant was discarded and the samples were resuspended in 100 µL 10 mM Tris 7. Samples were then cleaned using the modified version of the Zymo-5 RNA Clean and Concentrator columns to remove all RNAs < 200 nt (to remove tRNAs). Finally the sample was eluted in 85 µL DEPC-water and we proceeded immediately with DNase treatment.

DNase Treatment: We added 10 µL 10x Turbo DNase buffer and 5 µL Turbo DNase to 85 µL RNA (from previous step). The sample was incubated for 20-30 min at 37 °C in an Eppendorf thermomixer without shaking. We added 330 µL 100% Ethanol, 11 µL 3M NaAc (pH 5.4) and 1 µL GlycoBlue coprecipitant to the mixture. The tubes were then chilled at -80 °C for 30 min+ (possible stopping point) and then spun at 20,000 g for 60 min at 4 °C. The supernatant was discarded and then 250 mL pre-chilled 80% Ethanol was added and the mixture was spun at 20,000 g for 5 min. The supernatant was discarded and the pellet was then resuspended in 13.5 µL DEPC-water. The concentrations were then measured using the High Sensitivity RNA kit for Qubit, after which we proceeded with library preparation.

Library Preparation: Libraries in this paper were prepared using NEB's rRNA Depletion Kit (Bacteria) for rRNA removal with beads (NEB #E7860). Post rRNA removal samples were prepared for either Illumina or Singular Sequencing using NEBNext Ultra II RNA Library Prep Kit for Illumina with Beads (NEB#E7775). Singular specific primers for the libraries sequencing on Singular which had the S1/S2 handles instead of Illumina's p7/p5 handles.

Code Availability: Raw sequencing fastq files were trimmed using seqtk and cutadapt to remove bases of low quality and adapters. Reads were then aligned using bowtie (version 1)(Langmead et al., 2009), after which the density of the 5' ends was quantified using SAMtools(Li et al., 2009) and the CDS files for each genome were used to quantify how many transcripts were found within each gene. The complete scripts for raw analysis, as well as processed data files (in the format of counts per gene) can be found on our GitHub

(https://github.com/gwlilabmit/MTC_2023_Scripts). Jupyter notebooks also exist for each of the plotted subfigures and extended figures.

Data Availability: Raw fastq files associated with this work have been deposited to the NCBI's SRA database. These files are associated with Bioproject PRJNA1024409, and are publicly available at the link https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1024409, Processed data files and figure source data can be found on our GitHub (https://github.com/gwlilabmit/MTC_2023_Scripts), Non-sequencing data (i.e. the doubling rate data in Extended Figure 2) can also be found in the GitHub.

References

Ackermann, M. (2015). A functional perspective on phenotypic heterogeneity in microorganisms. *Nature Reviews. Microbiology*, *13*(8), 497–508.

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., & Theis, F. J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, *38*(12), 1408–1414.

Bergen, V., Soldatov, R. A., Kharchenko, P. V., & Theis, F. J. (2021). RNA velocity-current challenges and future perspectives. *Molecular Systems Biology*, *17*(8), e10282.

Bhattarai-Kline, S., Lear, S. K., Fishman, C. B., Lopez, S. C., Lockshin, E. R., Schubert, M. G., Nivala, J., Church, G. M., & Shipman, S. L. (2022). Recording gene expression order in DNA by CRISPR addition of retron barcodes. *Nature*, *608*(7921), 217–225.

Butterfield, G. L., Lajoie, M. J., Gustafson, H. H., Sellers, D. L., Nattermann, U., Ellis, D., Bale, J. B., Ke, S., Lenz, G. H., Yehdego, A., Ravichandran, R., Pun, S. H., King, N. P., & Baker, D. (2017). Evolution of a designed protein assembly encapsulating its own RNA genome. *Nature*, *552*(7685), 415–420.

Chen, W., Guillaume-Gentil, O., Rainer, P. Y., Gäbelein, C. G., Saelens, W., Gardeux, V., Klaeger, A., Dainese, R., Zachara, M., Zambelli, T., Vorholt, J. A., & Deplancke, B. (2022). Live-seq enables temporal transcriptomic recording of single cells. *Nature*, *608*(7924), 733–740.

Chung, C. T., Niemela, S. L., & Miller, R. H. (1989). One-step preparation of competent Escherichia coli: transformation and storage of bacterial cells in the same solution. *Proceedings of the National Academy of Sciences*, *86*(7), 2172–2175.

Herzel, L., Stanley, J. A., Yao, C.-C., & Li, G.-W. (2022). Ubiquitous mRNA decay fragments in E. coli redefine the functional transcriptome. *Nucleic Acids Research*, *50*(9), 5029–5046.

Horns, F., Martinez, J. A., Fan, C., Haque, M., Linton, J. M., Tobin, V., Santat, L., Maggiolo, A. O., Bjorkman, P. J., Lois, C., & Elowitz, M. B. (2023). Engineering RNA export for measurement and manipulation of living cells. *Cell*, *186*(17), 3642–3658.e32.

Jiao, C., Reckstadt, C., König, F., Homberger, C., Yu, J., Vogel, J., Westermann, A. J., Sharma, C. M., & Beisel, C. L. (2023). RNA recording in single bacterial cells using reprogrammed tracrRNAs. *Nature Biotechnology*, *41*, 1107–1116.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., & Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Molecular Cell*, *58*(4), 610–620.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., … Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, *560*(7719), 494–498.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* , *25*(16), 2078–2079.

Lin, D., Li, X., Moult, E., Park, P., Tang, B., Shen, H., Grimm, J. B., Falco, N., Jia, B. Z., Baker, D.,

Lavis, L. D., & Cohen, A. E. (2023). Time-tagged ticker tapes for intracellular recordings. *Nature Biotechnology*, *41*(5), 631–639.

Linghu, C., An, B., Shpokayte, M., Celiker, O. T., Shmoel, N., Zhang, R., Zhang, C., Park, D., Park, W. M., Ramirez, S., & Boyden, E. S. (2023). Recording of cellular physiological histories along optically readable self-assembling protein chains. *Nature Biotechnology*, *41*(5), 640–651.

Moris, N., Pina, C., & Arias, A. M. (2016). Transition states and cell fate decisions in epigenetic landscapes. *Nature Reviews. Genetics*, *17*(11), 693–703.

Niepel, M., Spencer, S. L., & Sorger, P. K. (2009). Non-genetic cell-to-cell variability and the consequences for pharmacology. *Current Opinion in Chemical Biology*, *13*(5-6), 556–561.

Norman, T. M., Lord, N. D., Paulsson, J., & Losick, R. (2015). Stochastic Switching of Cell Fate in Microbes. *Annual Review of Microbiology*, *69*, 381–403.

Qiu, X., Zhang, Y., Martin-Rufino, J. D., Weng, C., Hosseinzadeh, S., Yang, D., Pogson, A. N., Hein, M. Y., Hoi Joseph Min, K., Wang, L., Grody, E. I., Shurtleff, M. J., Yuan, R., Xu, S., Ma, Y., Replogle, J. M., Lander, E. S., Darmanis, S., Bahar, I., … Weissman, J. S. (2022). Mapping transcriptomic vector fields of single cells. *Cell*, *185*(4), 690–711,e45.

Schmidt, F., Cherepkova, M. Y., & Platt, R. J. (2018). Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature*, *562*(7727), 380–385.

Schmidt, F., Zimmermann, J., Tanna, T., Farouni, R., Conway, T., Macpherson, A. J., & Platt, R. J. (2022). Noninvasive assessment of gut function using transcriptional recording sentinel cells. *Science*, *376*(6594), eabm6038.

Shalon, D., Culver, R. N., Grembi, J. A., Folz, J., Treit, P. V., Shi, H., Rosenberger, F. A., Dethlefsen, L., Meng, X., Yaffe, E., Aranda-Díaz, A., Geyer, P. E., Mueller-Reif, J. B., Spencer, S., Patterson, A. D., Triadafilopoulos, G., Holmes, S. P., Mann, M., Fiehn, O., … Huang, K. C. (2023). Profiling the human intestinal environment under physiological conditions. *Nature*, *617*(7961), 581–591.

Shipman, S. L., Nivala, J., Macklis, J. D., & Church, G. M. (2016). Molecular recordings by directed CRISPR spacer acquisition. *Science*, *353*(6298), aaf1175.

Silas, S., Mohr, G., Sidote, D. J., Markham, L. M., Sanchez-Amat, A., Bhaya, D., Lambowitz, A. M., & Fire, A. Z. (2016). Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science*, *351*(6276), aad4234.

Tang, W., & Liu, D. R. (2018). Rewritable multi-event analog recording in bacterial and mammalian cells. *Science*, *360*(6385), eaap8992.

Wagner, A., Regev, A., & Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, *34*(11), 1145–1160.

Young, J. W., Locke, J. C. W., Altinok, A., Rosenfeld, N., Bacarian, T., Swain, P. S., Mjolsness, E., & Elowitz, M. B. (2011). Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. *Nature Protocols*, *7*(1), 80–88.

Acknowledgements and Funding for the MTCs

We would like to thank Lydia Herzel, Matthew Tien and Robert Battaglia for all their help with various protocols used in this work. We would also like to thank Katherine (Julia) Dierksheide, James (Scott) McCain, and James Xue for their helpful discussions and suggestions for framing of this work. Julian Stanley, Jenny Cascino, Manraj Gill, Andrew Savinov, Mandy Levine, Hannah LeBlanc, James Taggart, Grace Johnson and Cassandra Burgos-Schaening also provided useful feedback, suggestions, ideas and criticism over the course of many group meetings and presentations. Furthermore we would like to thank Joseph (Joey) Davis and Paul Blainey for their expert feedback and suggestions.

This work was made possible in collaboration with the MIT Biology Bio Micro Center, and in particular we are grateful to Stuart Levine, Christopher Hallee and Noelani Kamelamela for facilitating the sequencing and qPCR work reported in this paper.

This work was supported in part by the Koch Institute Support (core) Grant P30-CA14051 from the National Cancer Institute. Thanks also go to our funding sources: NIH grant R35GM124732, the NSF CAREER Award, the Smith Odyssey Award, and the Pew Biomedical Scholars Program.

Chapter III

rendseq and rendseq.org: an open source Python package and web platform for end-enriched RNA sequencing (Rend-seq) data.

- *Note the work in this chapter is also currently under submission for publication. I would like to acknowledge my co-authors in this work:*
- Julian Stanley (Department of Biology, Massachusetts Institute of Technology; Cambridge USA)
- Gene-Wei Li (Department of Biology, Massachusetts Institute of Technology; and the corresponding author for this work: gwli@mit.edu)

Abstract

Transcription initiation, termination, and RNA processing have critical regulatory roles for gene expression through their ability to tune mRNA abundances. The sequence features governing these processes can be studied by mapping the transcript ends that they define. We recently developed end-enriched RNA sequencing (Rend-seq) for identifying transcript ends in bacteria with single-nucleotide resolution. Here, we present both rendseq, an open-source Python package that automates end identification from Rend-seq datasets, and rendseq.org, an interactive web platform for exploration of published Rend-seq datasets. These tools facilitate the interrogation of *in vivo* RNA termini and can inform future studies of RNA regulatory control.

Introduction

Messenger RNAs (mRNAs) are targets of gene regulation. The sequence context of mRNA 5′ and 3′ ends contains much of the information needed to define the transcript's abundance by determining the rates of transcription initiation, termination, and mRNA processing. However, unlike proteins whose termini are straightforward to computationally determine from genomic sequences, it is challenging to predict the 5′ and 3′ ends of transcripts (Cassiano & Silva-Rocha, 2020).

In vivo measurement or inference of transcript ends can recover RNA 5′ and 3′ boundaries across multiple environmental, cell state, and genetic conditions. Building a comprehensive database of transcript ends could inform improved computational methods for predicting transcription initiation, termination, mRNA processing, and decay. Such a database

could also provide a set of regulatory building blocks, e.g., promoters, terminators, and RNA processing sites that could then be tested for their performance as modular components in informed transcript unit design.

End-enriched RNA sequencing (Rend-seq) (Lalanne et al., 2018) is a variation of RNA sequencing that can quantify the abundances of transcripts as well as identify their 5′ and 3′ ends with single-nucleotide precision. Rend-seq combines a sparse fragmentation of input RNA with a subsequent size selection to generate a pool of short transcripts for sequencing. Any fragment containing either the original 5′ or 3′ end of the pre-fragmentation transcript is enriched. This is because fragments containing one of the original ends only require one fragmentation event to end up in the final pool, whereas fragments with neither original end must undergo at least two events. Due to the size selection, both the 5′ and 3′ ends of each fragment can be recovered via sequencing. Rend-seq tracks are generated by recording these 5′-mapped and 3′-mapped reads, creating 4 tracks per experiment: a 5′-mapped track and a 3′-mapped track for both the forward and reverse strands. These are stored in a wiggle (.wig) format and can then be viewed in genome browsers (Fig 3.1a).

We present rendseq, an open-source Python package that automates the extraction of peaks (corresponding to transcript ends) from Rend-seq data. We have also made a web interface, rendseq.org, to make published Rend-seq datasets easily accessible and explorable. Users can explore a variety of Rend-seq datasets across different conditions to see how the end profile distribution changes between different genetic and environmental conditions. Lastly, we seek to use rendseq.org to ease the adoption of both the Rend-seq protocol and analysis platform by using it to host a growing set of learning resources.

Implementation

We have developed rendseq, an open-source Python package available on both GitHub and PyPi, to facilitate the identification of transcript ends in Rend-seq data. The locations of these ends are identifiable by the peaks in Rend-seq tracks. Robustly labeling all the peaks in a sample can be a challenging data analysis task to automate. This is because the abundance of different transcripts can vary by several orders of magnitude. Together, this means that naive approaches, such as direct thresholding, cannot reliably identify transcript ends.

rendseq overcomes this challenge by pre-processing each track using a modified Z-score transformation. For each position, we compute the Z-score using data only from its immediate neighbors instead of all positions in the genome. Furthermore, the upstream and downstream read densities are treated as separate distributions to reflect the structure of Rend-seq data, i.e., peaks mark the transitions in RNA levels. Outliers are removed from these distributions to suppress the effect of other peaks (e.g., alternative transcription start sites) that might fall close to one another, thereby allowing us to identify closely spaced RNA ends. The modified score for each position is then set to be the smaller of the two Z-scores calculated with respect to the upstream and downstream distributions. These transformed data files can then be thresholded to find peaks. We provide a useful default for this threshold which performs well on Rend-seq datasets we have generated. For users who wish to change this default we also supply tools, such as plotting the distribution of transformed values, to help choose a new threshold value.

To make the Rend-seq datasets more accessible, we created an interactive database and web platform hosted at rendseq.org. The architecture of rendseq.org is shown in Fig 3.1b. Rend-seq wig tracks are uploaded internally to Google Cloud storage, triggering Google Cloud Functions to create a transformed Rend-seq file using the latest version of rendseq. Each

Rend-seq wig track is simultaneously compressed from wig to big-wig files to support lower latency viewing. Files are compressed by using the ENCODE project's wigToBigWig functionality (Hitz et al., 2023) run by a bash script on a Google Cloud Run triggered by EventArc. After processing is complete, the new Rend-seq dataset is added to the browsable list of public Rend-seq datasets. Both the raw Rend-seq file and the transformed Rend-seq file can then be referenced and displayed via the embedded genome viewer, powered by igv.js (Robinson et al., 2023). The transformed file appears directly below each raw Rend-seq track, and is zoomed in to display lines for all transformed values with a value of 10 or greater. Users can now browse our published Rend-seq datasets without the need for any data processing or software installation.

Features

Rendseq.org supports viewing of public Rend-seq datasets in the explore tab. Here, users can simultaneously view multiple stacked Rend-seq tracks from the same organism, allowing informative comparisons across growth and genetic conditions with an igv.js-powered (Robinson et al., 2023) genome browser. Tracks for individual genes can be accessed by using the search bar on the top of the plot. The browser supports viewing the datasets across multiple resolutions - from single nucleotide to full genome.

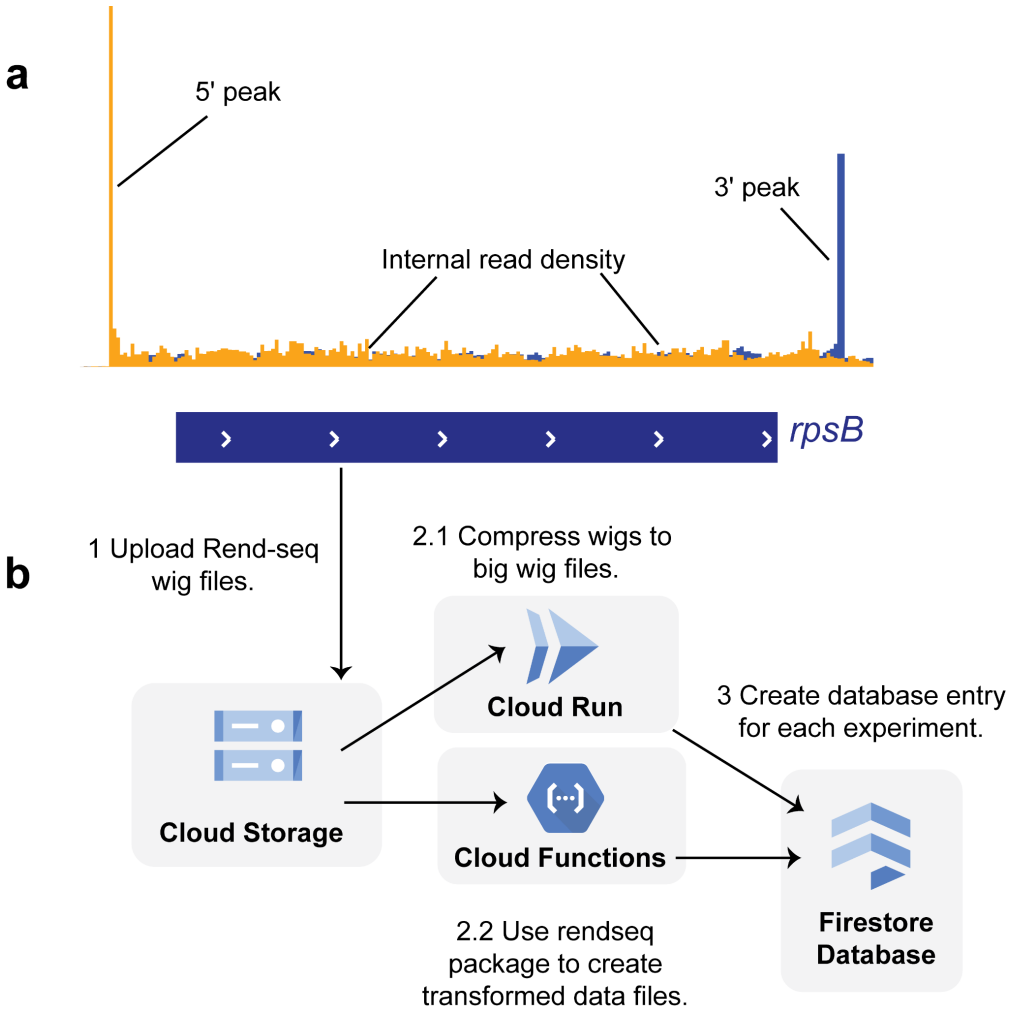

Fig 3.1 a) Example of Rend-seq data. Plotted are two overlaid wig tracks corresponding to the 3′-mapped (blue) and 5′-mapped (orange) reads of the forward-strand mapping reads from a Rend-seq experiment performed on *Escherichia coli*, zoomed in on the gene *rpsB*. b) Illustration of rendseq.org data-processing implementation. All images correspond to the image for each Google Cloud Platform™ service used. To start, (1) Rend-seq tracks are uploaded to Firebase Cloud Storage. This event triggers (2.1) a Cloud Run event which will compress each of the Rend-seq wig files into bigwig files using the wigToBigWig functionality. It also triggers (2.2) a Cloud Functions instance to begin analyzing each track using the rendseq package. Both the compressed data file location, and the location of the peaks file are (3) added to the Cloud

Firestore entry for this experiment, which can then be accessed by rendseq.org to display datasets in the viewer.

Rendseq.org also hosts learning materials for the Rend-seq protocol in the “Learn” tab. These materials are intended to help build intuition about how the Rend-seq protocol works, facilitate adoption of the Rend-seq protocol, motivate the assumptions behind our data analysis pipeline, and provide readers with use cases of Rend-seq. We welcome any new users of either the sequencing technique or the analysis package to notify us (via an email to rendseq ‘at’ mit.edu) of published results for inclusion in the rendseq.org database. In this way we also hope rendseq.org can grow to serve as the definitive database for all Rend-seq data sets and applications.

Future Work

Although the rendseq package can be used to identify peaks, not all transcripts end profiles carry this signature. For example, many genes in prokaryotes are known to undergo factor-mediated transcription termination, such as Rho-dependent termination (Banerjee et al., 2006). Unlike intrinsic termination, Rho-dependent termination does not always occur at the same genomic position, meaning that Rho terminated transcripts can end with a downward sloping ramp rather than a sharp peak (Johnson et al., 2020). Robustly identifying these features in Rend-seq data would facilitate the study of alternative modes of transcription termination, and is a functionality under development in rendseq.

Once ramps, which indicate that termination or degradation is taking place over a wide stretch of sequence context, can be robustly identified, it will then be possible to automatically

predict transcription isoform annotations from Rend-seq data. Transcription isoforms are especially important in prokaryotes which often encode related genes in polycistronic operons (Cao et al., 2015), whose relative levels are then tuned by changing the rates of transcription termination, initiation, and decay (Lalanne et al., 2018). We aim to use this planned functionality to add a new channel to the rendseq.org genome browser where users can visualize the annotated transcription isoforms alongside the Rend-seq datasets.

Observing the distribution of RNA ends *in vivo* can reveal new biology. For example, high-resolution mapping of transcript boundaries in exoribonuclease-deficient strains can facilitate identification of RNA processing sites (Taggart et al., 2023). By revealing that the precise 3′ ends of *Bacillus subtilis* transcripts are very close to translation termination sites, Rend-seq also provided the motivation that led to the discovery of runaway transcription in *B. subtilis (Johnson et al., 2020)*. By providing easy access to Rend-seq datasets and streamlining the data analysis process, we hope rendseq and rendseq.org will facilitate more discoveries into the biology of transcription and RNA regulation.

References

Banerjee, S., Chalissery, J., Bandey, I., & Sen, R. (2006). Rho-dependent transcription termination: more questions than answers. *Journal of Microbiology* , *44*(1), 11–22.

Cao, J., Arha, M., Sudrik, C., Mukherjee, A., Wu, X., & Kane, R. S. (2015). A universal strategy for regulating mRNA translation in prokaryotic and eukaryotic cells. *Nucleic Acids Research*, *43*(8), 4353–4362.

Cassiano, M. H. A., & Silva-Rocha, R. (2020). Benchmarking Bacterial Promoter Prediction Tools: Potentialities and Limitations. *mSystems*, *5*(4), e00439–20.

Hitz, B. C., Lee, J.-W., Jolanki, O., Kagda, M. S., Graham, K., Sud, P., Gabdank, I., Seth Strattan, J., Sloan, C. A., Dreszer, T., Rowe, L. D., Podduturi, N. R., Malladi, V. S., Chan, E. T., Davidson, J. M., Ho, M., Miyasato, S., Simison, M., Tanaka, F., … Cherry, J. M. (2023). The ENCODE Uniform Analysis Pipelines. *bioRxiv : The Preprint Server for Biology*. https://doi.org/10.1101/2023.04.04.535623

Johnson, G. E., Lalanne, J.-B., Peters, M. L., & Li, G.-W. (2020). Functionally uncoupled transcription-translation in Bacillus subtilis. *Nature*, *585*(7823), 124–128.

Lalanne, J.-B., Taggart, J. C., Guo, M. S., Herzel, L., Schieler, A., & Li, G.-W. (2018). Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry. *Cell*, *173*(3), 749–761.e38.

Robinson, J. T., Thorvaldsdottir, H., Turner, D., & Mesirov, J. P. (2023). igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics* , *39*(1), btac830.

Taggart, J. C., Lalanne, J.-B., Durand, S., Braun, F., Condon, C., & Li, G.-W. (2023). A high-resolution view of RNA endonuclease cleavage in Bacillus subtilis. In *bioRxiv* (p. 2023.03.12.532304). https://doi.org/10.1101/2023.03.12.532304

Acknowledgements and Funding for rendseq

We would like to acknowledge Jenny Cascino and James Xue for their use and feedback on this software package and website. We would also like to thank Connie New for her ideas and contributions to parts of this package still under development. We would also like to thank Jean-Benoit Lalanne, James Taggart and Lydia Herzel for their assistance with understanding the Rend-seq protocol, data analysis goals and for freely sharing data analysis pipelines and data with us. This work was supported by NIH grant R35GM124732, the NSF CAREER Award MCB-1844668, the Smith Odyssey Award, the MathWorks Science Fellowship Program, and the Pew Biomedical Scholars Program.

Chapter IV

Conclusion

Summary of Thesis Chapters

Molecular-biological tools empower our study of the microscopic dynamics which give rise to life. New tools can facilitate novel modes of observation as well as inspire original directions of inquiry and exploration. In this work I have presented some novel techniques to aid the study of the mechanisms and dynamics of transcriptomic regulation. Transcriptional regulation is a key process through which cells determine their developmental trajectories and respond to changes in the environment. By achieving better measurements of this process we can hopefully also achieve a greater understanding of the biology which governs it and better grasp the repercussions of certain transcriptomic changes. This expanded knowledge of transcriptional control and effect will complement efforts to predict, manipulate and design cellular behaviors.

Chapter I provides the summary and background necessary to contextualize the later chapters. It is divided into two broad sections: one focused on the study of transcriptomic dynamics in time and the second including a review of basic RNA regulatory processes and of Rend-seq, a sequencing method whose data can aid in the inference of *in vivo* transcript ends. I begin with an overview of why measuring transcriptomic variation in time and mapping it to cellular behavior or outcomes can be a challenging task. From interventional data collected on the single cell level, to fluorescence-microscopy tracking of RNA abundances over time to genomic recording of transcript abundances in the genome, we explored different methods for studying which gene-expression changes proceed and/or cause phenotypic behavior and development. What united many of these methods is the common need for effective hypothesis generation to reduce the space of tested genes and their expression levels. As this space is so large, consisting of the potentially combinatorial interactions of thousands of genes at analog

expression levels there are a huge number of conditions to test for their causal role in genotype to phenotype relationships.

In the next part of the introduction I went over a quick review of some of the processes which tune relative transcript abundance: transcription initiation, termination, and RNA decay. By reviewing these processes, as well as discussing how they can be governed by sequence context I aimed to provide motivation as to why identifying transcript ends with precision is a useful exercise for uncovering new modes of RNA-regulation. I also aimed to motivate how creating a corpus of sequence contexts of RNA ends across conditions can complement efforts to build better models of transcriptional control. After providing context for what is known about these biological processes I then explored a method for inferring from *in vivo* data where RNA transcripts end (Rend-seq) [1]. I concluded with a discussion of the existing analysis methods to extract information about transcript ends from Rend-seq data.

Chapter II introduces a new method for the elucidation of transcriptomic dynamics over time: the molecular time capsules (MTCs). These self-assembling protein capsules can be used to capture and protect a "snapshot" of the full-transcriptome, allowing it to be recovered later and analyzed. I demonstrated that the snapshot captured by the MTCs is highly reproducible. It can be cleanly isolated from non-encapsulated RNAs without contamination of the recovered transcripts. Furthermore, once captured, these snapshots are well maintained over time even as the host transcriptome undergoes significant remodeling. My co-authors and I foresee several applications where the MTCs can help elucidate fundamental biological questions: from understanding the transcriptomic heterogeneity of immune response [2–4] or bacterial antibiotic persistence [5,6], to uncovering the transcriptomic origins of cellular differentiation, we anticipate

that the MTCs will be able to serve as effective hypothesis generators and identify candidate genes for further studies into the link between transcriptomic variation and phenotypic effect.

Chapter III provides an overview of the analysis and data visualization platform I constructed for inferring transcript ends from end enriched sequencing (Rend-seq) data. This includes both an open source Python package, rendseq, which includes several methods for the identification of peaks in Rend-seq data. These peaks correspond to the inferred 5′ and 3′ ends of transcripts *in vivo*. This chapter also describes the web platform which hosts the interactive web-database of our published Rend-seq datasets and tools to learn more about the Rend-seq protocol: rendseq.org.

With new tools, both molecular biological tools, and analysis tools, it becomes easier to answer otherwise challenging biological questions. In this chapter: Chapter IV, I am providing a broad summary of the other chapters in this work, as well as highlighting what I believe to be the main takeaways, potential applications, and future avenues for improvement for these tools.

Molecular Time Capsules can serve as a Hypothesis Generation Tool for causal genotype-to-phenotype studies.

Organisms contain a large number of individual gene products with even minimal bacterial cells containing hundreds of seemingly essential genes [7]. Through tuning the expression levels of these genes, or more commonly by tuning the expression levels of several genes simultaneously, cells are able to adopt different phenotypes, respond to changes in the environment, or commit to different developmental pathways and behaviors. Establishing causality between gene-expression levels and cellular phenotypes can be cost and time prohibitive unless one has the means to limit the number of hypotheses to explore. An effective strategy for establishing a reduced set of candidate genes to test for their causal linkage to

phenotypes of interest is to first determine which gene expression changes correlate with the emergence of these phenotypes. Once a set of correlating genes has been established, one can perform experiments to ascertain which of these changes are both necessary and sufficient to explain the adopted cell behavior. Observation has an advantage over model-driven hypothesis generation in that it can capture unexpected correlative factors that a model with incorrect or insufficient assumptions may miss. This is because models often rely upon well-established biological principles and this bias can potentially obscure novel and interesting biology.

The characteristics of the MTCs mean they are well situated to serve as hypothesis generators. They can reproducibly capture snapshots of the transcriptome. This means one should be able to compare two MTC samples and infer that measured gene expression differences reflect real differences in the captured transcriptome between samples. These snapshots of MTC encapsulated RNAs can be cleanly separated from the host transcriptome. This lack of contamination during the MTC extraction process means one can be confident that the RNAs recovered from the MTC purification process are not merely remnants of the host-transcriptome or other forms of contamination which would reduce one's capacity to meaningfully interpret the MTC's contents. Lastly, once captured MTC snapshots remain stable over time. This observation should also increase confidence that MTC-encapsulated RNAs continue to serve as a representative snapshot of the recorded transcriptome up until they can be recovered from the cell.

Together these characteristics of reproducibility, clean recovery and stable storage suggest it should be possible to use MTCs to capture a transcriptomic snapshot of a population, then recover and analyze it later. One use case for this capability is to study transcriptome dynamics conditional on a phenotype which is only revealed after the transcriptomic history is

captured. This process would begin with a transient induction of MTC production, and thus transcriptomic history recording, for the entire population. After an appropriate delay, some fraction of that population may exhibit the phenotype of interest: for example resistance to stress, special reactions to the external environment or commitment to a specialized cell fate. By sorting the population based on this emergent phenotype and retrieving the MTC-encapsulated RNAs from the sorted subpopulation, one can selectively query the transcriptomic history of only the sub-population of interest. This technique can thus generate a profile of what the transcriptomic history looks like conditional on the downstream phenotypes.

Building on our belief that the MTCs are an effective tool for conditionally querying the transcriptome histories of interesting sub-populations, we are exploring ways to verify that it has this capacity, and seeking new systems to apply it to. To do this we are focused on using the MTCs in conjunction with Fluorescent Cell sorting (FACS). First we seek to demonstrate that the MTC is able to recapitulate known transcriptomic variations which correlate with interesting emergent phenotypes. It has been demonstrated that transient expression of genes involved in the acid stress response, such as gadX, leads to a non-genetic reduced susceptibility to ciprofloxacin [8]. To begin with we wish to characterize whether the MTC can indeed recover signatures of these known correlational relationships. We will do this by first capturing a transcriptomic snapshot of the cells using the MTCs, and then exposing the cells to ciprofloxacin stress. After treatment with the antibiotic we will recover those cells which survived antibiotic treatment using live/dead staining and FACS. We anticipate this experiment will demonstrate that the MTCs capture the anticipated evidence of gadX upregulation in the survivors.

Once we are able to demonstrate this proof of principle we plan to use the MTCs to probe the subtle transcriptomic origins of otherwise challenging to study biological processes - for

example other forms of bacterial stress resistance, commitment of bacteria to certain cell states or gene-programs and he origins of immune-cell heterogeneity. We believe that the MTCs should be able to be applied across many different organisms, both prokaryotic and eukaryotic. If the MTCs are able to generate a list of hypotheses we plan to establish causality between the set of observed correlators and the phenotypes of interest by orthogonal interventional techniques.

rendseq and rendseq.org can increase the accessibility and ease of use of Rend-seq data.

There exist motivating questions for which the knowledge about the precise starts and ends of transcripts leads to actionable information. Such information can help inform designs and directed mutations of transcript promoters and terminators, as well as help assess the steady state ratios of different transcription isoforms. Some transcripts are known to have cryptic internal promoters, which may or may not give rise to translatable transcripts. Identification of transcript ends can thus help design cleaner transcripts and assess the extent to which transcription of one product is influencing transcription of another. Many of these questions can be answered by making use of pre-existing Rend-seq datasets, such as those which have already been published by the Li Lab [1,9–11]. By making it easy to view Rend-seq data and identify potential transcript ends without the need to download or process any data, we believe that we have made our currently published Rend-seq datasets more available, and thus more useful. I am excited to share this resource with the broader scientific community and look forward to learning more from them about how to improve it further.

In addition to making currently public Rend-seq datasets more accessible I have also sought to make the Rend-seq protocol and analysis tools more widely available. To this end I have co-authored, with my labmate Jenny Cascino, an illustrated protocol hosted on rendseq.org

which provides step by step instructions on how to perform the Rend-seq protocol. I have also created a guide to help gain intuition about how the Rend-seq protocol is able to enrich the ends of transcripts. I have also created and maintained, along with the assistance of my co-authors, a Python package which handles all the stages of Rend-seq analysis after bowtie alignment. The backend of rendseq.org makes use of this package to facilitate automated analysis and feature extraction of new Rend-seq datasets when they are uploaded to the database. We are currently exploring the best way to facilitate new dataset uploads on the public portion of the website as well. By opening the site to user uploaded datasets, users of Rend-seq could take advantage of rendseq analysis without the need for downloading or running any analysis on their local machines. We believe this could accelerate both the adoption of Rend-seq as a technique and the analysis and sharing of Rend-seq datasets.

rendseq is a tool for uncovering transcriptomic ends from end-enriching sequencing data, but it captures some ends better than others. In particular the current version of rendseq assumes that all transcript ends will look like well-defined peaks. In reality we know many transcripts do not end in a well defined peak spread out over a handful of adjacent nucleotides. For example, in Rho-mediated transcription termination, termination actually takes place over an extended stretch of sequence leading to a gradual tapering off of the 3′ end density rather than a well defined peak. I, along with my collaborators, are exploring methods to robustly identify other forms of transcription termination using rendseq - with a focus on ramps. Once ramp-like end profiles can also be robustly identified it will be easier to automate construction of transcription isoform atlases (collections of all overlapping transcripts and their relative abundances). It will also be easier to identify which types of transcription termination prevail in certain organisms,

and how the balance of transcription termination mechanisms changes across genetic and environmental conditions.

Concluding Remarks

Transcriptional control is one of the main mechanisms through which cells determine what gene-products to produce and thus what processes they wish to execute. Precise control of transcription is essential, as protein production is often tightly coupled with RNA abundance, and the relative ratios of certain protein products must often be produced in ratiometric quantities in order to fulfill their biological roles [12,13]. When studying transcriptional control, as with studying most things, it is important to have precise and quantitative methods to facilitate measurements and analyses. The molecular time capsules are tools designed for the observation of transcriptomic dynamics over time in interesting sub-populations. I designed them with the hope they will improve our ability to probe the relationship transcriptional programs have with cellular phenotypes, and initial testing suggests they have the characteristics necessary to perform as required. The analysis tools I created for rendseq are intended to promote quantitative measurements of transcriptional control, and perhaps also aid in the identification of novel sites of regulatory control.

The usefulness of new tools should be measured by how much they assist the researchers who are willing to try them out. It is my sincere hope that these tools, the MTCs and rendseq and rendseq.org, help unveil new biological discoveries for those who choose to use them.

References

1. Lalanne, J.-B. *et al.* Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry. *Cell* **173**, 749–761.e38 (2018).
2. Satija, R. & Shalek, A. K. Heterogeneity in immune responses: from populations to single cells. *Trends Immunol.* **35**, 219–229 (2014).
3. Hu, Y., Liu, C., Han, W. & Wang, P. A theoretical framework of immune cell phenotypic classification and discovery. *Front. Immunol.* **14**, 1128423 (2023).
4. Liu, J. & Cao, X. Regulatory dendritic cells in autoimmunity: A comprehensive review. *J. Autoimmun.* **63**, 1–12 (2015).
5. Harms, A., Maisonneuve, E. & Gerdes, K. Mechanisms of bacterial persistence during stress and antibiotic exposure. *Science* **354**, (2016).
6. Huemer, M., Mairpady Shambat, S., Brugger, S. D. & Zinkernagel, A. S. Antibiotic resistance and persistence-Implications for human health and treatment perspectives. *EMBO Rep.* **21**, e51034 (2020).
7. Hutchison, C. A., 3rd *et al.* Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253 (2016).
8. Sampaio, N. M. V., Blassick, C. M., Andreani, V., Lugagne, J.-B. & Dunlop, M. J. Dynamic gene expression and growth underlie cell-to-cell heterogeneity in *Escherichia coli* stress response. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2115032119 (2022).
9. Taggart, J. C. *et al.* A high-resolution view of RNA endonuclease cleavage in Bacillus subtilis. *bioRxiv* 2023.03.12.532304 (2023) doi:10.1101/2023.03.12.532304.
10. Herzel, L., Stanley, J. A., Yao, C.-C. & Li, G.-W. Ubiquitous mRNA decay fragments in E. coli redefine the functional transcriptome. *Nucleic Acids Res.* **50**, 5029–5046 (2022).
11. Johnson, G. E., Lalanne, J.-B., Peters, M. L. & Li, G.-W. Functionally uncoupled

transcription-translation in Bacillus subtilis. *Nature* **585**, 124–128 (2020).

12. Taggart, J. C., Zauber, H., Selbach, M., Li, G.-W. & McShane, E. Keeping the Proportions of Protein Complex Components in Check. *Cell Syst* **10**, 125–132 (2020).
13. Taggart, J. C., Lalanne, J.-B. & Li, G.-W. Quantitative Control for Stoichiometric Protein Synthesis. *Annu. Rev. Microbiol.* **75**, 243–267 (2021).

Appendix A

Comparing other Recording Methods to the Molecular Time Capsules.

In recent years there has been a growth of techniques aimed at quantifying transcriptomic dynamics. In order to clarify the niche that the molecular time capsules (MTCs) hold in this ecosystem, we will now discuss some of the advantages of the MTC when compared to the most similar alternative method: genomic recording techniques.

There are a variety of different genomic recording techniques. Some approaches use the presence or absence of an RNA to make records in the genome which can then be read and converted to abundances of specific transcripts by analysis/post processing. Because these techniques are targeted and do not cover the full transcriptome we limit our comparison to full-genome recording techniques. In particular, we will compare to the only class of full genome recording techniques demonstrated so far: those that reverse transcribe RNA into DNA and then use Cas systems to place this DNA into a CRISPR array. For clarity we will refer to solutions which make use of this approach as "full genome CRISPR recording systems" (or FGCRSs).

The FGCRS systems are currently rate limited by the probability per cell that an RNA is reverse transcribed and successfully inserted into a CRISPR array during the recording window. This low rate of spacer acquisition implies that many cells need to be used in order to capture sufficient information about the transcriptome for statistically significant analysis. Recent estimates of this number indicate that only 1 in ~1.9×10^4 *E. coli* will acquire a spacer during the recording window. By contrast, we estimate that we *recover* the equivalent of ~6.8 MTCs per cell. If we consider that each recovered MTC corresponds to 1 piece of information about the transcriptome (reasonable given that most MTCs contain nucleic acids) then this comparison suggests that the number of transcript spacers recovered per FGCRS would need to be improved by ~ 3.4×10^4 fold in order to match the recording capacity per cell currently achieved by the MTCs.

The second area of comparison is time required for recording. The MTCs have demonstrated a recording time as short as 1 hr. So far the minimal recording time demonstrated by FGCRS systems has been on the order of ~12 hours. Shorter recording times lead to higher temporal resolution of the transcriptomic record, which can in turn lead to better inference about the observed transcriptomic state. The transcriptomic recording times should be on the same scale as the events they wish to capture, which for bacteria whose RNAs have very short half-lives and which can double as frequently as every 20 minutes, may require recording times on the order of dozens of minutes.

The current long recording time of FGCRSs makes it difficult to assess their reproducibility. Because the transcriptome changes so much during the time of recording (~12 h during which the cell transitions from exponential to stationary) it is to be expected that the captured record does not look like the last time point of the transcriptome, and that even two biological replicates will may not look alike. This issue likely explains the lack of sample-to-sample reproducibility seen in FGCRS systems (The highest Pearson Coefficient seen between biological replicates using FGCRS recording is: 0.786, much lower than the lowest log-transformed Pearson Coefficient seen between biological replicates of MTC recording: 0.976 (n = 2276 genes)). With shorter recording times it would be easier to make direct and fair comparisons between the reproducibility of FGCRS systems and MTC based recording.

Appendix B

Rend-seq track pre-processing and peak identification.

Supplement: Data-Preprocessing and Peak Calling Algorithms in rendseq:

Preprocessing Rend-seq files

rendseq will assist in the normalization of files after they have been aligned to genomes. It takes an input either a sam file outputted from the bowtie assisted alignment of sequencing files to the genome of interest or a wig file of read density as a function of genome position. The first step in the rendseq analysis pipeline is to convert these aligned files into normalized Rend-seq tracks. There are four Rend-seq tracks for every experiment, one for each the 5′ and 3′ mapping reads on both the forward and the reverse strands. When converting files from the sam format rendseq will handle the occasional issue which arises when non-templated additions are made to the 5′ end of reads during the library preparation process, and will remove the nucleotides which do not match at the 5′ end before generating the 5′ end Rend-seq tracks. Each of these Rend-seq tracks is saved in a wiggle, or .wig format.

Each Rend-seq file is then preprocessed before peak calling can proceed. The step of pre-processing takes each Rend-seq track and converts it into a form where it is easier to make comparisons across disparate parts of the genome. This normalized Rend-seq track can also be saved in a wig format.

The first step of pre-processing is to pad the zero-values of the Rend-seq track. This padding is done using gaussian generated values, with a mean of 0 and a standard deviation of 1. It is important that they include a standard deviation, rather than just setting the values to 0. Setting all values to 0 will cause the local estimation of the standard deviation to be artificially low and could lead to a high false positive rate of peaks. Even though true reads densities do not exist as floats we found that by treating these areas of low read coverage as decimal values increased the performance of the peak-calling algorithm in areas with low read density.

After padding, normalization proceeds - replacing each genomic position in the data with its normalized value, referred to henceforth as a "score". This score is calculated via a function of the read density at that genomic position, and the surrounding read density. Only scores above a threshold of 5 are kept, as in practice scores of less than 5 do not correspond to true peaks. The score-generating function is a modified Z-score which considers the Z-score of the position with respect to both the (optionally normalized) upstream and downstream read-density distributions. It has several parameters: gap size (an integer), window size (an integer), percent trim (a float in the range [0, 1]), and winsorize (a boolean). I will explain each of these parameters in more depth and also explore the effects that each can have on score-generation.

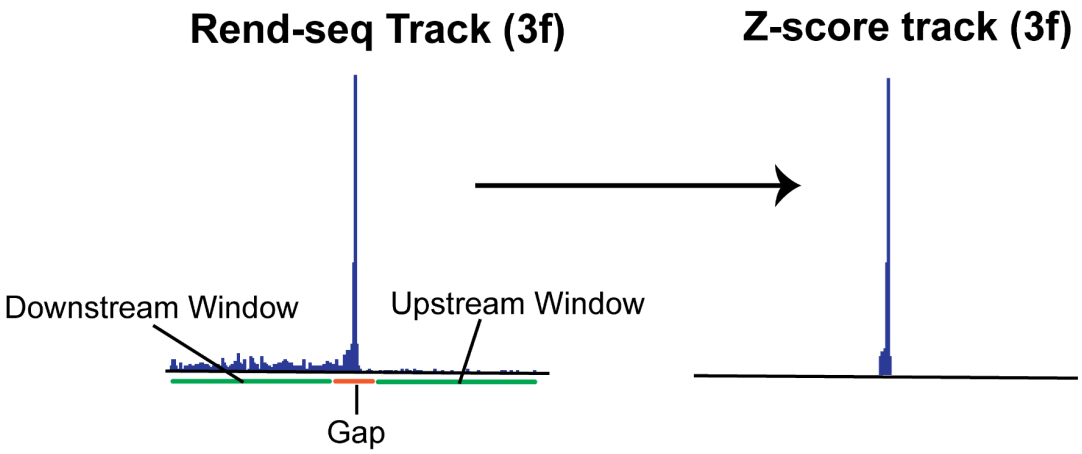

Figure A2.1 Illustration of a Rend-seq track, and its corresponding z-score (ie the pre-processed track). This example track is from the 3 ends of reads aligning to the forward strand. In the center of the track is a peak. The Upstream and Downstream windows around this peak have different read densities. The normalized score is calculated via a sliding window

approach that creates upstream and downstream distributions for reference as a function of the gap and window-size parameters.

The gap size is the distance upstream and downstream of the given genomic position to exclude from any distribution calculations. This exclusion is important as peaks can often be distributed across many read positions, meaning that a true peak's local density can artificially inflate the estimation of the local distribution's standard deviation, and may deflate the score we assign to true peaks. This deflation can then lead to a higher false negative rate in peak identification. The default value of the gap size is 5.

The window size refers to the total number of reads to include in the distribution. This number ought to be large enough to provide a useful estimate of the mean and standard deviation, but not so large as to cause it to include other genes or features in the distribution being used for the mean and standard deviation estimation. The default value for the window size is 50.

Percent trim aids in the trimming of the upstream and downstream distributions before the mean and standard deviation are called. It will remove the top x % of reads, based on their read density. The default is 0, though the user can supply a different percentage, which in certain cases may be useful in removing extremely high values, such as peaks. Setting a value greater than ~10% is not recommended however, as there are cases where a high percentage of trimming will allow other features, such as noise, or the shadows which can accompany peaks, to appear as peaks as well.

The boolean parameter winsorize serves a similar purpose as percent trim - although it facilitates the removal of reads based on standard deviation. If set to true then the program will normalize the upstream and downstream distributions before estimating their means and standard

deviations. For a given window - any read which exceeds a standard deviation of 1.5 will be dropped. We use a low standard deviation here, as it is important to remember extreme outliers can lead to a very large standard deviation. We have found that a threshold for winsorization of 1.5 strikes a good balance between aiding in the elimination of true outliers (other peaks) and leaving regions without extreme outliers undisturbed.

$$z\,score(x) = \frac{(x - \mu)}{\sigma}$$

Equation A2.1: The z-score of x with respect to a distribution with mean and standard deviation is the distance between the mean and the value normalized by the standard deviation.

These normalized distributions are created for the region both upstream and downstream of the peak location. Their mean and standard deviation are calculated, and these in turn are used to calculate the z-score of the original position. See Equation 1 for a description of how z-scores are calculated. The score which is assigned to that position is the minimum value of the upstream and downstream z-scores. The reason behind separating the two distributions is that peaks often accompany a step in read density, which can vary across several orders of magnitude. If one were to treat the left and right distributions as the same then this transition in the underlying generating function of the read density would lead to an artificially high standard deviation and may lead to the suppression of true peaks. Rendseq assigns each score to be the minimum of the z-scores calculated with respect to the right and left distributions. This choice was made because true peaks will obtain high scores relative to both their right and the left distributions, whereas other features, for example reads which are close to the step, would only have an extreme z-score with respect to either the upstream or the downstream distribution, not both.

Calling Peaks:

After normalized tracks have been generated they are used to call peaks. We currently support two methods for identifying peaks. The first is to simply choose a z-score threshold above which reads are considered peaks and below which reads are not considered peaks. The second method makes use of a statistical hamiltonian monte carlo method to assign each location in the genome a “peak” or “not peak” state.

The thresholding method is straightforward to apply. Once a threshold is selected, and transformed values above that threshold will be called peaks, and all those below will not. The trouble with this method lies in selecting the optimum threshold. Rendseq has a default threshold of 12. In practice this threshold performs well on our Rend-seq data sets. Across different organisms, or across different conditions it is possible that this threshold may not be appropriate. As such we support several methods which can support the automatic selection of a more appropriate threshold.

One method is to select a value for the threshold such that the number of values expected to have that value is much less than 1. This method however is not recommended as it will ultimately be a function of the length of the genome in addition to the underlying gaussian assumption (longer genomes will require a higher expected value cutoff, and shorter genomes a lower one). The total length of the genome should not affect whether a location is a peak or not. In practice however, because most bacterial genomes are of similar lengths, this approach may nonetheless be useful for setting a lower bound on the threshold.

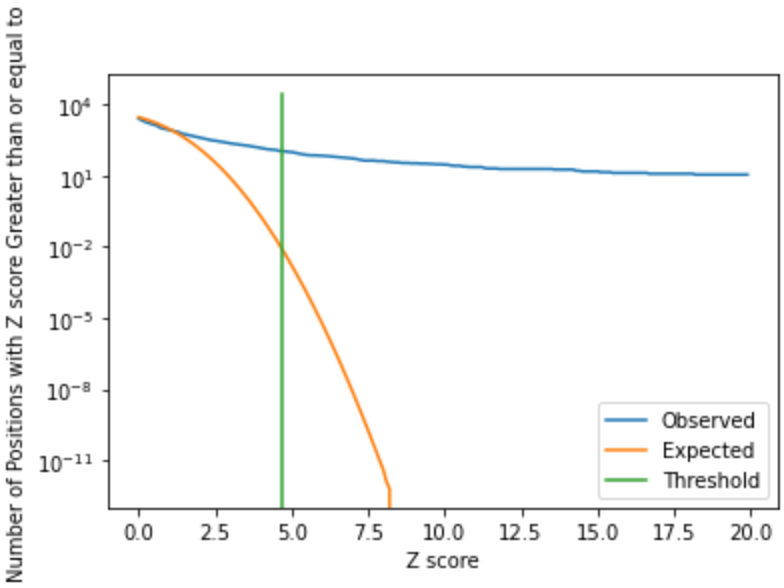

Figure A2.2 Expected vs observed scored Rend-seq values. An illustration of the divergence between the expected number of certain z-scores (which is a function of the length of the genome) and the observed number. Note that the observed number and the expected number have an extreme divergence - this is largely driven by the true peaks in the distribution which have extremely high scores. However, as the distribution is not perfectly gaussian the true curve of the "expected values" is in truth stretched to the right of what is represented here. None-the-less, one can still use this curve to either find a point of sufficient divergence between the expected and the observed values, or find the point such that the expected number of positions with a certain score falls below a certain threshold.

A more laborious, though effective method can be trying out a variety of different thresholds and manually reviewing which peaks are added/removed as the threshold changes in order to inform how confident one is in any given threshold.

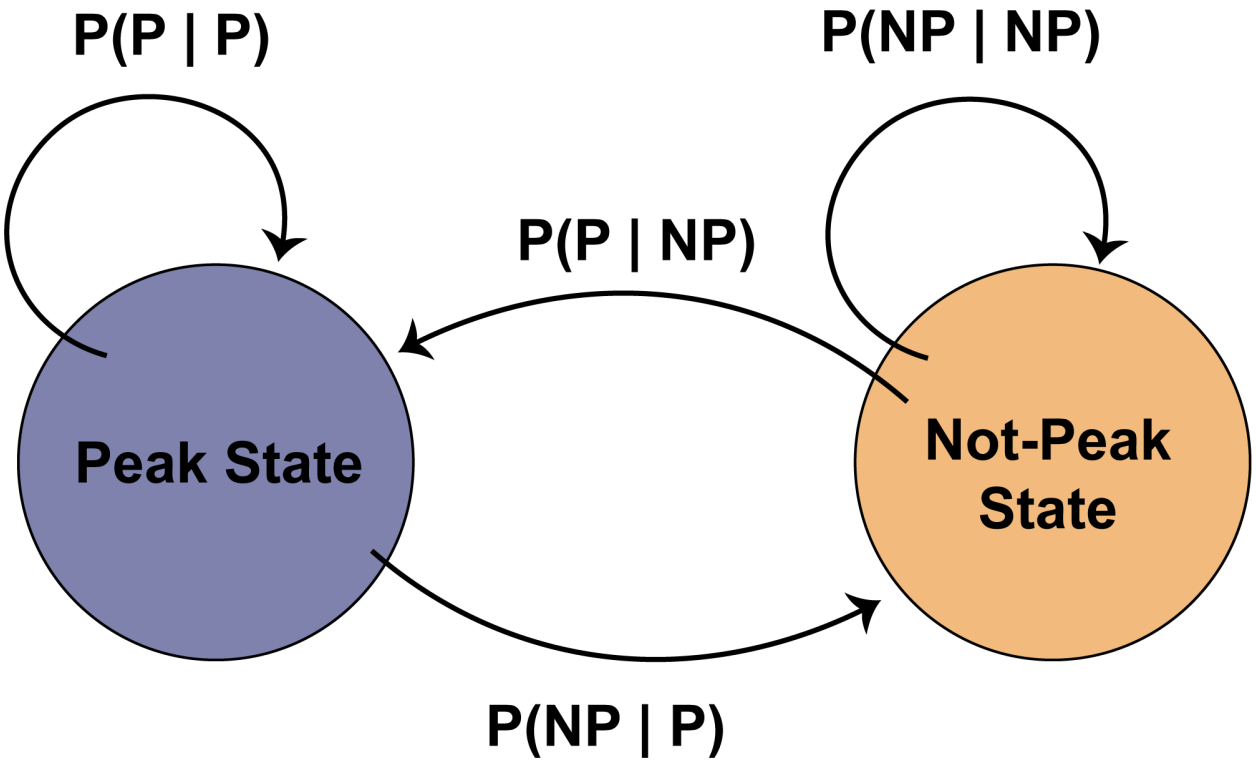

Figure A2.3 Illustration of HMM states and transition probabilities. There are two states in the model: Peak state and Non-Peak state. Each of these states has transition probabilities to each other as well as to themselves.

The other currently supported method for peak calling makes use of a hidden markov model (HMM) [1] to call peaks by creating a sequence of peak and non-peak which is then assigned to the sequence of read density. A diagram of the hidden markov model used in this work can be viewed in Figure A2.3. This method assumes there are two hidden states: "peak" and "not-peak". Each of these states has some probability of transitioning to the other and to themselves; probabilities collectively known as the transition matrix. Separately each of these states also has a probability distribution over the likely observed data-sets (in this case the scored Rend-seq data sets) known as the emission probabilities. Given the probabilities of transition and

the probabilities of emission the one can make use of the Viterbi algorithm to find and return the most likely sequence of hidden states for a particular Rend-seq dataset.

To test this method I employed a gaussian distribution to model both the emission probabilities of both the peak state (which was centered around higher z-score values), and for the non-peak state (which was centered around 0). By testing the mean and standard deviation of these parameters I was able to test how robustly the methods call the peaks.

The transition probabilities were very robust to changes, as is commonly seen with Hidden Markov Models. I set the value of the transition probability so that the mean length of continuously labeled peak states was 3, and the mean length of continuously labeled non-peak states matched a reasonable biological value for the average length of transcripts ~ 1000. The mean represents the mean of the geometric distribution generated from particular transition probabilities. As can be seen in figure A2.4, the number of peaks called as a function of what the peak to not-peak probability or the peak to peak probability is very flat across several real scored Rend-seq datasets. This suggests that the model is extremely robust to these parameters, and not much tuning is needed to these parameters in order to control the output of the model.

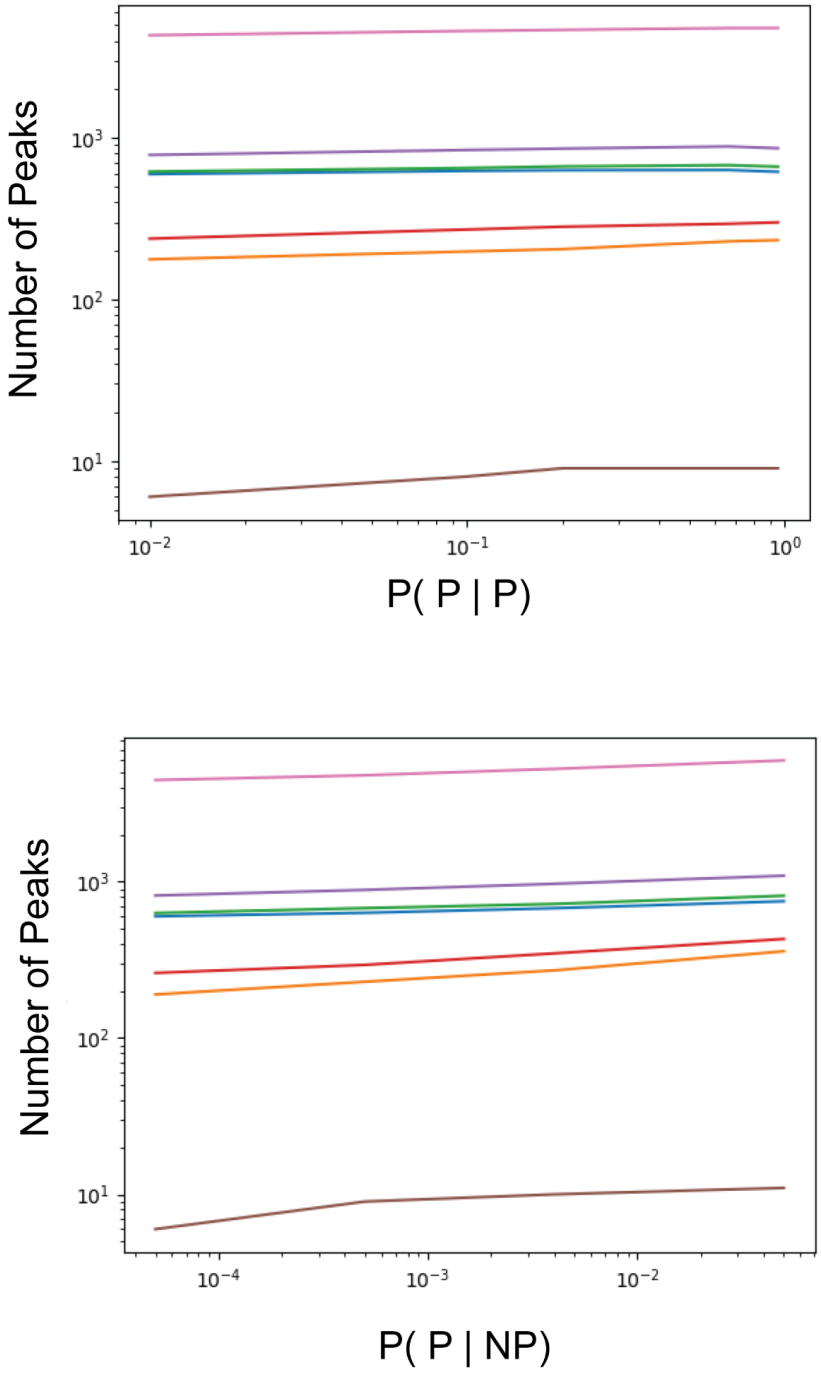

Figure A2.4 Robustness of the HMM to transition probabilities. Across several real data-sets (represented by the colored lines) the number of peaks called remains very flat across several

orders of magnitude values for both the Peak to Peak probability (P(P | P)) and the Not-Peak to Peak transition probability (P (P | NP)).

In contrast to the apparent robustness of the transition probabilities however, the two tunable parameters of the peak emission distribution, namely the mean and the standard deviation do not seem to be quite as robust. For the standard deviation, it seems that the number of peaks initially increases as the standard deviation increases before reaching a plateau. So long as the correct number of peaks identified can be found in the plateaued region the standard deviation will not need to be heavily tuned in order to fit the data.

The mean however seems to exhibit a linear decrease in the number of peaks as the value of the mean increases. This result is not surprising - as the mean increases some values which were initially classified as peaks are now better described as not-peaks. Overall the HMM seems to agree best with other published peak-calling techniques when the mean of its emission distribution is very high, and the standard deviation is also very high. There is still some degree of parameter tuning which may need to be undergone for each dataset, suggesting that the HMM-approach to peak calling may not have entirely solved the issues associated with selecting an appropriate threshold seen with the simple threshold calling method, though it may in some cases be less sensitive to parameter selection.

We plan to continue to add support and descriptions for both of these techniques, as well as to look into other methods to robustly identify transcript ends. All of the code which executes the methods described in this appendix can be found on this project's Github (**https://github.com/miraep8/rendseq**). The documentation for this section is hosted on read the docs! (https://rendseq.readthedocs.io/en/latest/).

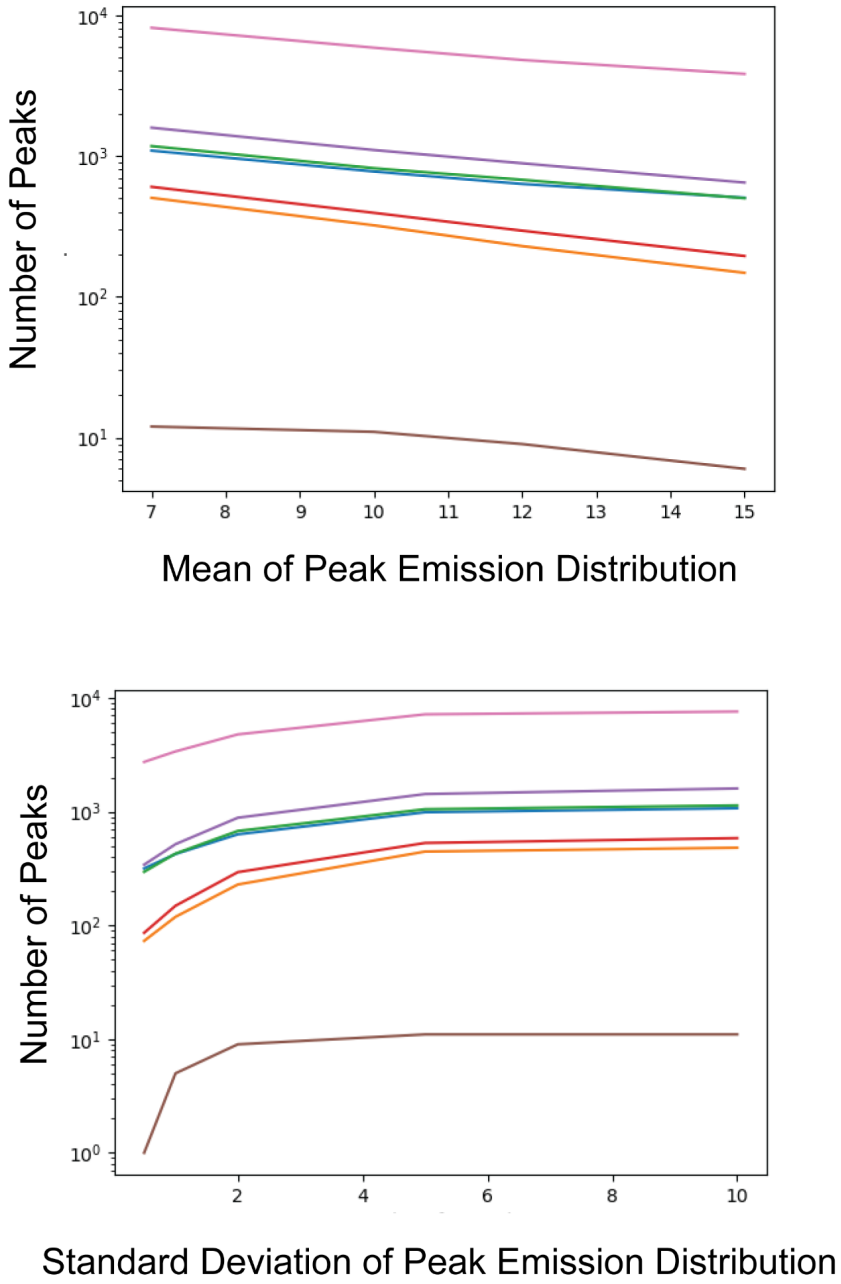

Figure A2.5 Robustness of the HMM peak emission distribution parameters. The peak emission probabilities are modeled as a simple gaussian distribution. As such it has two tunable parameters: the center or mean of the distribution and the standard deviation of the distribution. While the standard deviation seems to reach a plateau in terms of the number of peaks called,

suggestion that there is a threshold beyond which tuning the standard deviation does not influence the number of peaks called.

References

1. Models, M. An Introduction to Hidden Markov Models. http://ai.stanford.edu/~pabbeel/depth_qual/Rabiner_Juang_hmms.pdf.