Supervised and Unsupervised Methods for Transcriptional Sequencing Data

by

Z. L. Maas

B.A., University of Colorado Boulder, 2019

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

2024

Committee Members:

Robin Dowell, Chair

Mary Ann Allen

Daniel Larremore

Ryan Layer

John Rinn

Maas, Z. L. (Ph.D., Computer Science)

Supervised and Unsupervised Methods for Transcriptional Sequencing Data

Thesis directed by Prof. Robin Dowell

In human biology, understanding how the information encoded in an individual's genome regulates development and responds to environment and disease remains a key question. Characterizing the differential usage of functional elements of the genome remains a significant challenge, but has been advanced substantially through developments in sequencing protocols. Through a variety of sequencing protocols, we now have the ability to assay not only DNA sequence but also a wide variety of functional events and states, such as protein binding, chromatin accessibility, and RNA levels. Of particular interest are protocols measuring nascent transcription, which provide information on both the cell's actively used regions of DNA as well as a subset of transcription factor binding events. Chemical and biological limitations in the measurement of nascent transcription mean that it remains difficult to effectively dissect the heterogeneity present from biological variation in transcriptional sequencing data. In this thesis, I develop three distinct approaches using supervised and unsupervised machine learning to address this limitation. First, I propose a novel Bayesian re-framing of sample normalization that enhances normalization in samples with external controls and allows for approximation of normalization in samples at short time points without normalization controls. Next, I establish the feasibility of supervised deconvolution (estimating the mixing proportions of constituent cell types) in bulk nascent transcriptional sequencing data using established techniques, finding that non-coding regulatory regions enhance model accuracy but confound estimation when undifferentiated cell types are present in the mixture. Finally, I develop an unsupervised method for discovery of cell type specific regulatory motifs using approaches drawn from language modeling, deep learning, and mechanistic interpretability. Collectively, this work advances our ability to extract useful information from nascent transcriptional sequencing data and to better understand the heterogeneity within.

Dedication

Acknowledgements

I am deeply grateful to Dr. Robin Dowell for her exceptional guidance through the past several years, helping me grow from a chemist and mathematician into the competent computer scientist that I am now. I would like to thank Dr. Mary Allen for her constant enthusiasm and insatiable curiosity for biology, and for all of the expertise and brainstorming she provided throughout all of my projects. Without the Dowell and Allen (DnA) labs, I don't believe that I would have ever learned to love biology, the one topic that I absolutely could not stand when I was younger.

Thanks to all of the members of the DnA lab over the years for having the expertise that we all need to make an interdisciplinary lab work well. In particular, thank you to Dr. Jacob Stanley for the many hours debating the intricacies of Bayesian modeling and the state of science and to Dr. Rutendo Siguake for keeping things in perspective and maintaining all the lab plants throughout the years. I would also like to acknowledge Matthew Hynes Grace and the rest of the BIT team for their patience and expertise in maintaining our compute resources, without which none of this would have been possible.

To the IQ Biology program – thank you Kristin Powell and Stephanie Rauscher for entertaining every wild idea for making more community. To the external advisors who raised my expectations for what kind of science I could do – thank you to Dr. Larry Hunter, Dr. Boswell Wing, and Dr. Jeffrey Cameron.

Contents

Chapter

List of Tables

Table

List of Figures

Figure

Chapter 1

Introduction

1.1 Biological Context

A fundamental challenge in human biology is understanding how the information encoded in an organism's genome directs both development and responses to environmental stresses. Consider the human genome as an example. Each individual begins with a set of chromosomes containing all necessary genetic information for life, but this information must be selectively accessed and utilized at specific points throughout development and in response to environmental cues. During embryonic development, a single cell must proliferate into billions of cells, each acquiring a distinct cellular identity through unique patterns of differential gene expression. This selective activation of genomic regions enables cellular differentiation, resulting in specialized organs, each with unique cellular composition and gene expression patterns[1]. Differential gene expression mediates cellular responses across a spectrum of biological challenges, from acute stressors such as UV exposure and muscle injury to pathological conditions like oncogenic transformation[2]. Understanding these regulatory mechanisms is therefore crucial not only for developmental biology but also for the study of disease and therapeutic development. Effective mechanistic understanding of genetic regulation requires both structural characterization of the genome and functional analysis of how different genomic elements are used.

Since this is our problem, the next logical step then, would be to try to figure out what a genome sequence looks like, and then to see how different parts of that genome are used. The central dogma of molecular biology establishes that genetic information flows from DNA to RNA through transcription, and subsequently from RNA to protein through translation[3]. During transcription, DNA sequences

serve as templates for the synthesis of RNA molecules, which can be broadly categorized into coding and non-coding RNAs. Coding RNAs, specifically messenger RNAs (mRNAs), are then translated by cellular machinery into proteins, which serve as the primary functional molecules in cellular processes. This thesis focuses specifically on transcription, the process by which DNA-encoded information is converted into RNA molecules. While many transcribed RNAs ultimately serve as templates for protein synthesis through translation, a significant proportion of transcribed RNAs have regulatory functions. These regulatory RNAs form complex networks of transcriptional and post-transcriptional control, contributing to the sophisticated mechanisms of gene expression[4–6], but are often unstable and highly transient.

Transcription represents our most direct measurement of active genome utilization, making it valuable for two key areas of investigation. First, in developmental biology, transcriptional profiling quantifies both protein-coding gene expression and concurrent developmental regulation. The simultaneous capture of both coding and regulatory transcripts provides insight into the mechanisms controlling cell fate and differentiation. Second, studies of transcription provide substantial information on cellular responses to environmental stimuli. As with development, this approach captures both the immediate transcriptional response to environmental perturbations and accompanying regulatory changes associated with this response. The subsequent sections of this introduction detail both the biological and computational background of the study of transcription using sequencing technologies.

1.2 Data Generation and Protocols

Genomic sequencing protocols fundamentally involve the quantitative measurement of specific molecular populations within cellular systems. These measurements, when coupled with statistical modeling approaches, enable the inference of underlying biological phenomena. Understanding the historical development of sequencing technologies thus provides essential context for understanding current methodological approaches and their capabilities. To whit, we briefly review the history of these protocols. Initial applications of sequencing technologies primarily focused on genome characterization, aiming to simply determine the underlying sequence of an organism.

1.2.1 Technological Developments in Sequencing

The earliest efforts at understanding biology with sequencing data focused on understanding the genome, with the hope of letting us see the entirety of the sequence that defines an organism.

- **Shotgun and Sanger sequencing, and the human genome project:** In the heyday of the early 1990's, the human genome project set out on a mission to generate a full genome of a human. Initially, this project used Sanger sequencing which used either gels that were manually read[7] or, later, fluorescent labels and automated sequencers[8]. Initially the project focused on methodically mapping each read to a defined physical map, but as the project progressed, the application of Shotgun sequencing[9] significantly accelerated this work by foregoing the physical map in favor of leveraging computational tools for assembly. Despite the immense cost involved, this project succeeded in generating most of a human genome (called a complete genome), letting us see the entire sequence of a human for the first time[10, 11]. This advance was transformative — as prior to the genome studies typically focused on individual protein-coding genes whereas with a genome reference sequence, it became possible to conduct a variety of genome-wide studies. However, what we called 'a full human genome' at the time is not a full human genome, as it did not include a number of repetitive or difficult to sequence regions[11].

- **Short read sequencing, and decreasing costs:** Over the course of the 2000's, development of short read sequencing techniques resulted in a dramatic decrease in the cost of generating a genome. Short read sequencing techniques exacerbate the issues with tricky regions of the genome — they struggle with the same repetitive regions that the original human genome project did. However, the development of these techniques and their decreasing cost per base paved the way for functional genomics, with the ENCODE project[12] which had the goal of characterizing functional elements within the human genome.

- **Long read sequencing, looking forward to the future:** In the past decade, a new generation of techniques, dubbed long-read sequencing protocols, have emerged, which have helped to address the long-standing problems with difficult to sequence regions of the genome[13]. When we are

building a genome after sequencing, the key step that we care about is a process called assembly, which is where we take all our different reads that are gathered from across the entire genome and try to overlap them like puzzle pieces that fit together [14]. With longer read lengths, reads have less ambiguity (due to higher length and thus higher possibility of overlap) which allows for improved assembly quality of difficult regions[15]. These advances have and will continue to decrease the cost and improve the accuracy of our ability to build genomes for individuals, further expanding the kind and amount of data we have for understanding individuals and the population.

1.2.2 The limits of DNA-sequencing

With the successful sequence of a human genome generated, the subsequent ENCODE project set out to address a key unanswered question — now that we have a reference human genome, which parts of the genome play a functional role[12, 16]? To gain better understanding of the functional elements of the genome (the **why** of DNA use), we have developed a wide variety of high-throughput genomics techniques to measure as many things as possible that are associated with DNA (some examples below). The essential focus of most sequencing protocols (and of the ENCODE project) is this — we can use short or long read sequencing to measure DNA and/or RNA, and biochemistry has given us a huge toolkit to label and identify what things are attached to or near DNA and/or RNA, so by combining these two tools, we can measure almost anything in the central dogma prior to translation using sequencing. To understand what tools we've developed to try to gain a functional understanding of the genome, we review a variety of commonly used protocols — relevant to transcription regulation and this thesis — below.

1.2.3 ChIP-seq

What it does: sequences DNA that has some known protein attached to it.

Chromatin ImmunoPrecipitation Sequencing (ChIP-seq)[17], originally also referred to as location analysis, is a powerful tool to understand how and where proteins bind to DNA. The way that ChIP-seq

typically works is simple — we take a population of cells and cross-link them (typically with formaldehyde), which is to say, we run a chemical reaction that locks together things that are touching. Then, we can use an antibody that is specific to the protein of interest and extract that protein — which is now attached to whatever DNA it might have been touching (if any). We can then sequence that DNA and, by mapping it back to the reference genome, figure out where that protein was!

Since many proteins bind to DNA based on a known pattern of sequence (a motif)[18, 19], we can look at all the locations that the protein binds to and learn any sequence pattern that occurs more often than we would expect by random. This is a great, because it tells where certain regulatory proteins bind and what they bind to. For example, studies into chromatin architecture using ChIP-seq provided useful information on regulatory activity for both protein coding genes and their associated regulatory elements[20, 21]. Over time, a number of similar protocols (ChIP-nexus, ChIP-EXO, CUT&RUN) have been developed that improve on the sensitivity and/or resolution of protein-DNA interactions[22–24]. But, as with DNA sequencing, ChIP doesn't tell us **what** each regulatory protein is actually doing – changing the localized DNA environment (chromatin) or transcription. To answer that question, we need to move still closer to the actual processes driving how our cells use DNA.

1.2.4 RNA-seq

> **What it does:** sequences steady-state RNA that's floating around in the cell. Can either measure all messenger RNA, all RNA (infrequently used due to cost), or all RNA except for ribosomal RNA.

Because RNA is upstream of the production of proteins in the cell, measuring the messenger RNA (mRNA) content of cells is frequently used to examine a cell's response to a perturbation. The most widely used protocol today for measuring mRNAs is RNA-seq, which captures the totality of the steady-state messenger RNA present in a cell [25]. RNA-seq protocols enrich for mRNA typically by either positive selection using the polyadenylated tail of a fully processed messenger RNA or negative selection via depletion of ribosomal RNA (rRNA)[26]. Both of these steps serve to remove rRNA from the collected sample, which is typically the majority of RNA in a cell. As rRNA does not provide useful information

about the proteins that can be produced, these rRNAs are removed prior to sequencing. mRNAs are reverse transcribed into cDNA to enable sequencing. Since mRNAs have undergone splicing, reads in RNA-seq correspond to specific isoforms of a protein. Often, RNA-seq reads are sequenced as paired-end reads instead of single-end reads, such that both sides of the read are sequenced, providing additional information on isoform usage.

Steady-state RNA is, as a tool, much closer to getting us to that **what** question of functional genomics — if we observe a functional event, what does the RNA in a cell associated with that event look like? However, it is still an incomplete measure of transcription regulation. Because we are capturing all mRNA in the cell, what we are actually measuring is the steady state result of both transcription and the RNA maturation processes; which is to say, RNAs that have been produced by transcription and not yet destroyed by the process of degradation. Furthermore, when subjecting a cell to a perturbation (developmental change or environmental stimuli), these processes (transcription and transcript maturation) take time. Hence, with RNA-seq we are limited to only changes that are significant at this time point against the steady state background. Additionally, even when using a rRNA depletion protocol, the number of regulatory RNAs is significantly smaller than the number of messenger RNAs, so observing changes associated with those regulatory RNAs (and thus with transcription) remains challenging[**duttke_identification**].

1.2.5 ATAC-seq

What it does: measures DNA that is open and accessible.

Owing to its immense length relative to the size of the cell, DNA typically exists in a compact, highly wound state regulated by a set of molecules called chromatin. When some portion of the DNA is needed for transcription or protein binding, chromatin can be unwound, opening that region for easier access. This can be a good proxy for what portions of the genome are being used at any point in time — if DNA is accessible (e.g. free of chromatin), then it is often also used for transcription or protein binding. More succinctly, chromatin accessibility is a necessary but not sufficient condition for regulatory usage of the genome. We can measure chromatin accessibility using a protocol called ATAC-seq, which is elegant

and simple [27, 28]. Using an engineered protein from bacteria (a hyperactive TN5 transposase), we can take cells, extract the nuclei (where DNA lives), and then use our transposase to simultaneously fragment accessible DNA into chunks as well as labeling them with a tag for selection of only those open regions. Similar protocols, including MNase-seq and DNase-seq which were developed before ATAC-seq, also quantify nucleosome occupancy / positioning or open chromatin, respectively [**klen_genomic**]. Accessibility gets us close to the answer of what portions of the genome are being used at any moment in time, but fails to say what is actually functioning within each accessible region.

1.2.6 Nascent Run-On Protocols

> **What it does:** measure active transcription of RNA, with some extremely useful correlations with other biological processes.

At long last, we come to the set of protocols studied in the work done in this thesis. In contrast to the previously discussed protocols, nascent run-on sequencing protocols (which we will shorten to nascent protocols for much of this work) directly assay an active and ongoing biological process — transcription. While the previously discussed protocols focus on capturing some state of the cell as it exists at a moment in time (whether steady-state RNA, chromatin accessibility, or protein binding), nascent run-on protocols measure the active kinetic process of transcription using the incorporation of labeled nucleotides into RNA[29–31]. In so doing, we get a direct readout of exactly which portions of the genome are being actively transcribed, because our measurement is directly coupled to newly synthesized RNA. It turns out that when a transcription factor (a protein that binds DNA to regulate transcription) binds to DNA, we also simultaneously observe the transcription of short lived RNAs around its binding site (so-called enhancer associated RNAs or eRNAs) at a subset of these binding events. This means that by using nascent run-on sequencing protocols, we are able to simultaneously measure actively produced RNA and the regulatory activity associated with a subset of binding events[32]. This pair of readouts allows us to, from a single experiment, finally get to the **why** question of not just which portions of the genome are being actively used, but of the surrounding regulatory state driving that active process of transcription. For further clarity, we review below the two most common run-on nascent sequencing protocols at this point in time,

before taking a shift to focusing on the computational nature of this work.

1.2.6.1 Global Run-On Sequencing

What it does: measure active transcription of RNA using a BrU marked nucleotide.

GRO-seq (Global Run-On Sequencing) is the first high-throughput genome wide nascent sequencing protocol, and was developed contemporaneously with the first RNA-seq protocols (circa 2008). The GRO-seq protocol[29] leverages the incorporation of BrU tagged UTP (Br-UTP, a uridine analog), which can be selected for using immunoprecipitation with a Br-UTP specific antibody. The use of immunoprecipitation (similar to ChIP) allows for the repeated purification process that is needed to enrich the small amount of nascent RNA present in the cell at any given time. This incorporation of Br-UTP does not abort transcription, so fragmentation is also required generate fragments short enough for sequencing.

The development of the GRO-seq protocol revealed novel patterns of transcription activity in mammalian cells. At protein coding genes, nascent sequencing protocols show distinct peaks at the 5′ and to a lesser extent the 3′ ends of the gene. This 5′ peak (50+ bp) is associated with promoter proximal pausing while the 3′ peak is associated with polymerase dissociation (the last step of termination). Additionally, the development of GRO-seq also identified widespread divergent initiation (also called bidirectional transcription) across the genome. Bidirectional transcription occurs when RNA polymerase II (PolII) non-selectively loads onto either strand at binding sites in the genome. A subset of these loaded polymerases transition to productive elongation, producing pre-mRNA, most notably at annotated genes. Once this has occurred, PolII elongates through the body of the gene before aborting transcription. GRO-seq precision is on the order of 10s of base pairs (bp), which is useful in many cases but not for the study of transcriptional behavior on a smaller scale, when accuracy on the order of single nucleotide is desired. Unfortunately, in the mid-2010's GRO-seq started to suffer reproducibility issues after the quality of the Br-UTP antibody needed for purification steps in the protocol degraded[**rubin_DiscussionGROseqBrUTP_2018**].

1.2.6.2 Precision Run-On Sequencing

> **What it does:** measure active transcription of RNA using a marker that immediately stops transcription when incorporated, thus providing improved precision relative to GRO-seq.

Published in 2016, the PRO-seq protocol[33] improved on GRO-seq in a few key ways. Crucially, PRO-seq leverages the robust, specific, and sensitive binding of biotin to tag newly produced RNA RNA instead of using a BrUTP tag. Biotin is recognized by streptavidin in an exquisitely specific interaction that enriches for newly synthesized RNA. One consequence of this is that transcription is aborted when a biotinylated nucleotide is incorporated, giving single nucleotide read out of where exactly PolII was at that point in time. Originally, PRO-seq was run with all 4 NTPs (at a higher time and monetary cost) labeled with biotin, providing true single base pair resolution on the location of PolII[31]. To save money, most cases of PRO-seq are now run with only a single nucleotide marked with biotin. PRO-seq can also be run in a variant called PRO-cap which sequences from the $5'$ end instead of the $3'$ end of the RNA fragment, which can allow for more specific TSS information focused on the process of initiation.

1.2.7 Single Cell Protocols (RNA/ATAC/GRO)

> **What they do:** the same things as their parent protocols, but only using material from one cell rather than a bunch, letting us see cells as individuals rather than a population average.

We take one more diversion before discussing data analysis to focus on a set of increasingly common protocols that present significant potential for integration with the protocols discussed above. This has not yet been made explicit, but when working with the above protocols, they are performed in bulk, which is to say that they use an aggregate population of cells of the same type that are sequenced together and assumed to represent some average cellular state for that population[34, 35]. In most cases, this is necessary because the number of biological molecules to sequence is small and thus achieving an effective signal to noise ratio is challenging without aggregating a large number of cells. That has changed in recent years owing to advances in so-called single cell sequencing technologies, which do exactly what they claim. A single cell protocol uses some mechanism to physically isolate individual cells from a sample and assign a separate, uniquely defining barcode to them alongside whatever protocol is performed (typically

RNA-seq[36] or ATAC-seq[37]). With this barcode, we then have not just the sequences associated with all the cells we sequenced, but also a marker that allows for us to identify which cell those sequences came from. This allows for a much more effective view of heterogeneity in a cellular population — we can now aggregate (in silico) and analyze subpopulations within a single cell sequencing sample and compare it to other subpopulations.

While this is a powerful approach, it has a key limitation — by looking at only a single cell, we are usually only sequencing the most common RNAs in a sample, which in the case of RNA-seq is only the most highly expressed genes. Similarly, doing this with a nascent run-on sequencing approach, while possible, does not capture the same variety of useful intergenic regulatory information that bulk nascent sequencing protocols do. A key exception here, to the sampling biases introduced by single cell protocols, is single cell ATAC-seq. Since ATAC-seq looks at accessibility of chromatin, it can only sequence as many copies of a region as there are copies of that region on the chromosomes in that cell (typically 2 in human cells). This means that there is no equivalent of highly-expressed genes to dominate the single cell sequencing, and single cell ATAC-seq is thus more likely to give a representative view of cell state.

1.3 Analyzing Sequencing Data

Regardless of which protocol is used, there are some common analysis steps that are taken for all sequencing experiments, and these guide our approaches to understanding the data that is generated. Using sequence data to model and characterize functional characteristics (the focus of this thesis) fundamentally requires fully understanding the shape of the data, so here we provide a brief overview of the process of short read sequencing and typical initial analysis steps.

1.3.0.1 Sequencing a Library

To begin, the sequencing protocol of interest is performed and DNA is generated to be sequenced.

For the purposes of illustration, we briefly review the process of sequencing using an Illumina-style sequencer, which leverages fluorescence on a flow cell[38]. Because the Illumina sequencing platforms frequently used for our protocols of interest can only process relatively short reads, long molecules are

fragmented into shorter molecules (typically less than 500bp) prior to the rest of the protocol. After this, short adapters (known sequences separate from sequences encoding experimental data) are added to the fragmented sequences so that they can be attached to the sequencer's flow cell and actually sequenced. Once on the flow cell, the process of sequencing involves the repeated process of washing the cell with fluorescent nucleotides that mark when a certain base is incorporated as the complement of a base on the cell — this provides us with a readout of what sequences are in our sample. This fluorescence process has a degree of uncertainty, so some nucleotides are read with higher fidelity than others, and only high-quality, low-ambiguity reads are typically kept and used for downstream analysis. Long read protocols vary in protocol (typically using either fluorescence or voltage variation from a nanopore), with trade-offs in accuracy and read length depending on methodology[39] When the sequencing process is complete, a file containing all sequences and quality scores from the experiment is output.

1.3.0.2 Mapping Reads to a Reference Genome

Once the process of sequencing a sample is complete, we are left with a pile of reads that describe the samples used in our experiment. Before proceeding further, the next steps usually involves quality assessment, trimming and mapping, then further processing to convert the data into standard formats used by downstream processing tools. Quality control is done throughout this analysis process and involves steps ranging from checking the confidence in individual reads[40], to verifying the complexity of the sequenced library[41], to verifying the distribution of reads across the genome[42]. The initial reads from the experiment are often filtered by various quality control metrics and the adapter sequences from the sequencing process are removed (trimmed), then the reads are mapped to the reference genome of the organism of interest. Mapping is an extensive field in itself, but for the work done here we will not proceed deeply into discussing it[43]. Once reads are mapped they can be converted to a variety of formats useful for downstream tools, and then used for subsequent analysis. The most common step after mapping is to count all of the reads that map to specific regions of interest (ROIs).

1.3.0.3 Counts Tables as Matrices in $\mathbb{N}$

Of particular note in this context is the process of generating so-called counts tables from mapped reads. A key limitation in counting reads is that it requires that you **a priori** know the regions of interest (ROIs) that you want to count over. To quantify some characteristic of a functional element, you have to first know where that functional element is so that you can count it. To do this, either an existing reference can be used[44, 45], or a computational method can be used to identify these regions **de-novo**[46, 47]. Typically protein-coding genes are well annotated but regulatory RNAs are less well defined, particularly enhancer associated RNAs which are largely absent from public annotation resources.

Since many analyses (including those discussed here) are interested in the aggregate profile of reads of specific ROIs, reads are typically reduced from an alignment of mapped reads into a simple matrix consisting of the sum of all read alignments over (or reads mapping to) a region. Regions of interest can either be defined exogenously using standardized reference annotations, or endogenously using data-driven approaches to find features of interest[46–48]. To illustrate, consider counting the total number of reads in RNA-seq over a protein-coding gene. Reads that map to the gene's exons (the ROI) would be summed to obtain the total number of reads, a proxy measure for expression levels of the gene. In the case that there are reads that only partially overlap the ROI, a heuristic is typically chosen to determine when a read is included. The simplest of these is to say that a read is considered to be within a ROI when $> 50\%$ of the read lies within the ROI. By doing this in aggregate over all ROIs, data is reduced to a simple matrix format of size $m \times p$ where m is the number of ROIs and p is the number of samples that counting was performed over. The typical format of for these counts tables also includes information about region location and length so that normalization can be done if needed[49, 50].

With this set of counts, we can finally do the downstream analysis we want in a quantitative way! Dependent on the protocol, after all of this work, what we finally have is a set of reads mapped to a genome and a histogram of reads over the whole genome, which we usually subset in order to only look at a specific subset of regions of interest.

1.4 Using Data Effectively

Reductions in cost have resulted in unprecedented growth in data generation capabilities. This exponential increase in sequencing data, while offering impressive potential for advancing biological research, has also presented significant challenges in data analysis and interpretation. On a broader level, the work conducted here has focused on two distinct problems:

(1) How can we utilize the data we generate more effectively?

(2) How can we extract more information from existing datasets?

These questions arise from the paradoxical nature of modern sequencing experiments — while the cost of sequencing itself has decreased dramatically (and continues to drop), the expenses associated with experiment design, sample preparation, and researcher time remain substantial. Consequently, effectively using experimental data and developing appropriate tools for this task remains essential.

The analysis of sequencing data stands at the intersection of biology, chemistry, and computer science. This interdisciplinary nature necessitates an approach to algorithm development that considers not only the biological questions at hand, but also the chemical and physical constraints of sequencing protocols and how they might drive our interpretation of results. The rapid adoption of single-cell sequencing technologies in recent years has emphasized the importance of understanding cellular heterogeneity. However, technical limitations have hindered the application of single-cell approaches when studying transcription, as previously discussed. There are two distinct approaches that can address these problems while technologies improve to facilitate effective single cell transcriptional sequencing:

- **Single-Cell Backed Multi-Omic Approaches:** Combining bulk and single-cell techniques to complement the disadvantages of different protocols.
- **Enhanced Bulk Analysis:** Developing more advanced tools to extract fine-grained information from bulk data.

A substantial portion of the work discussed here was originally designed to focus on the first approach, with the idea of using single cell ATAC-seq to complement bulk nascent sequencing and to

develop new methodologies for approximating single cell nascent sequencing in an multi-omic way. However, as often happens when working with biology, this work was not yet feasible by virtue of the significant cost that single cell ATAC-seq still represents compared to bulk sequencing. Because of this data generation challenge, this work has instead focused on a computationally driven approach instead, using high quality published bulk nascent sequencing data as the basis of study and foundation for the algorithmic advancements presented here. To that end, this thesis focuses on three critical and connected aspects of transcriptional genomic data analysis:

(1) **Normalization:** Developing robust methods for data normalization for experiments with low-quality or absent external controls. Existing methods for differential expression analysis[51] often rely on assumptions that may not hold in complex experimental designs.

(2) **Supervised Deconvolution:** Dissecting the inherent heterogeneity present in bulk sequencing data in a supervised learning context. Current approaches do not fully capture the diversity of cell types and states in bulk transcription samples.

(3) **Unsupervised Information Extraction:** Uncovering latent information encoded across multiple experiments through unsupervised learning approaches. There is a need for more sophisticated methods to uncover regulatory patterns in transcriptional datasets.

Each of these areas represents a significant challenge in the field and offers opportunities for substantial methodological advancements.

1.5 Thesis Outline

This thesis presents novel methodologies and algorithms to address critical challenges in genomic data analysis, making significant contributions in three interconnected areas:

1.5.1 The Virtual Spike-In (VSI) Method for Normalization

Effective normalization is crucial to analysis of sequencing data, yet in many published nascent transcription sequencing experiments, normalization controls are of low quality or absent. The work

presented here represents an advance in our capability both to normalize and to think about normalization in situations with and without an external normalization factor.

- VSI treats spike-ins as random variables estimated using Markov Chain Monte Carlo (MCMC) methods, rather than as fixed constants.
- The VSI can effectively utilize the elongation of unperturbed RNA polymerases to determine normalization factors in the absence of external spike-ins.
- The method offers a flexible normalization strategy applicable across a variety of normalization conditions, improving these estimates independent of the presence of an external normalization control.
- VSI provides normalization bounds that improve on those estimated using standard tools like DESeq2, reducing the false positive rate in differential expression analysis of nascent sequencing data.

This approach represents a significant advancement in the field, offering more accurate normalization in scenarios where typical methods do not perform accuracy. By accounting for the inherent variability in spike-ins and leveraging underlying biological processes for internal normalization, the VSI approach enhances the analysis of transcription sequencing data across diverse experimental conditions at short time points.

1.5.2 Supervised Deconvolution of Bulk Transcription Data

Bulk sequencing samples are an ensemble representation of a population of cells that are assumed to be representative of the average cell in that population. True biological samples are typically heterogenous in some ways, and we are often interested in understanding these heterogenous system. For example, it can be more biologically relevant to study the population of mixed cell types that occur in a tissue or organ, rather than just looking at a single cell type isolated from that broader system. When looking at bulk sequencing, then, an important question is of how we can account for and understand the

heterogeneity within these mixed cell-type samples. While this work has been well established in other sequencing protocols, such as RNA-seq[52–59], it had not been studied prior in nascent transcription data. Here, this deconvolution approach is applied to nascent sequencing data for the first time, with adjustments for the unique regulatory characteristics provided by this data.

- A survey of well established supervised deconvolution strategies applied to nascent sequencing data shows that counterintuitively, simpler models outperform more complex state of the art models in this contest.
- This approach capitalizes on prior knowledge of cell type-specific transcriptional profiles to inform the deconvolution process.
- The addition of undifferentiated or partially differentiated cells to these mixtures confounds the supervised deconvolution process using all of the tested supervised deconvolution methods.

This work represents a crucial step in using cell type-specific information from bulk nascent transcription data. It demonstrates the feasibility of supervised deconvolution in this sequencing context, showing the feasibility of analyzing tissue composition directly from transcription. With this feasibility established, a wide variety of other projects are technically possible, which is discussed further in the conclusion of this work.

1.5.3 Unsupervised Learning of Sequencing Features

We present novel approaches for extracting and interpreting information from sequencing data across multiple experiments:

- Transformer based autoencoders can learn the transcriptional language of multiple cell types
- Applying modern interpretability techniques provides interpretable features that drive specific aspects of transcription.

- Using established tools alongside our interpretable features shows a diverse sequence language driving cell type, including known and novel sequence patterns, as well as general sequence regulatory elements associated with enhancers.

This work advances our ability to extract meaningful biological information from nascent sequencing datasets. By leveraging unsupervised learning techniques, we are able to learn a variety of known and novel sequence characteristics that drive a cell's transcriptional regime using a minimal amount of input data.

1.6 Thesis Structure

The remainder of this thesis is organized as follows:

- **Chapter 2:** Detailed description of the VSI method and its applications in transcription studies.
- **Chapter 3:** Deconvolution techniques for heterogeneous sample analysis.
- **Chapter 4:** Unsupervised learning approaches for extracting features from sequencing data.
- **Chapter 5:** Conclusions and future directions for research.

Through these chapters, this thesis presents a comprehensive body of work that advances the field of transcriptional data analysis, providing new tools and methodologies for researchers to extract deeper insights from nascent sequencing experiments.

Chapter 2

The Virtual Spike In

All of this chapter has been previously published as Maas, Z. L. & Dowell, R. D. Internal and External Normalization of Nascent RNA Sequencing Run-on Experiments. **BMC Bioinformatics 25,** 19. ISSN: 1471-2105. (2024) (Jan. 2024). Zachary Maas was responsible for the design of experiments, implementation of all project code, figure design, drafting and revision of the manuscript, and comments during peer review. Rutendo Sigauke provided preliminary data analysis of sequencing samples and curation of samples, associated with a separate manuscript[61]. Robin Dowell provided intellectual consultation, assistance on experimental design, and assistance during manuscript revision and during peer review. The supplemental information for this chapter is provided in Appendix A.

Abstract In experiments with significant perturbations to transcription, nascent RNA sequencing protocols are dependent on external spike-ins for reliable normalization. Unlike in RNA-seq, these spike-ins are not standardized and, in many cases, depend on a run-on reaction that is assumed to have constant efficiency across samples. To assess the validity of this assumption, we analyze a large number of published nascent RNA spike-ins to quantify their variability across existing normalization methods. Furthermore, we develop a new biologically-informed Bayesian model to estimate the error in spike-in based normalization estimates, which we term Virtual Spike-In (VSI). We apply this method both to published external spike-ins as well as using reads at the 3′ end of long genes, building on prior work from Mahat[62] and Vihervaara[63]. We find that spike-ins in existing nascent RNA experiments are typically under sequenced, with high variability between samples. Furthermore, we show that these high variability estimates can have significant downstream effects on analysis, complicating biological interpretations of results.

2.1 Introduction

Effective normalization is essential for rigorous analysis of high throughput sequencing data. In sequencing data, normalization identifies a set of features that are expected to be invariant between two data sets and leverages these to counteract the effects of systematic experimental bias and technical

variation. Broadly, there are only two possibilities for the source of these invariant features: external spike-in controls or an internal invariant set[51, 64]. Whenever possible, external spike-in controls are preferred[65], as they control for more sources of variation by adding a presumably invariant set of data across samples. However, not all data sets contain external spike-ins and they cannot be added **post-facto**. Consequently, a variety of internal normalization methods have been developed [65, 66] which assume some internal feature of the data – typically a set of genes – is invariant between data sets. While most of these techniques were developed for microarrays or RNA-seq, they have been broadly applied to a variety of sequencing assays.

One set of protocols in particular — nascent RNA sequencing methods — are prone to large amounts of technical variation [67]. Nascent RNA sequencing protocols, such as global run-on sequencing (GRO-seq)[29], precision run-on sequencing (PRO-seq)[33] and their variations [31, 68], isolate small quantities of recently produced RNAs from actively engaged RNA polymerases[32]. Nascent RNA sequencing samples have a distinct profile relative to RNA-seq (Figure 2.1A), resulting from the different phases of the RNA life cycle that they capture. RNA-seq samples from the pool of stable, messenger RNAs (mRNAs) which are predominantly spliced and polyadenylated. These RNAs originate from a relatively small fraction of the genome (exons and UTRs). In contrast, nascent RNA sequencing protocols capture RNA that is still actively engaged with RNA polymerases, meaning the RNAs are pre-splicing and need not be stable. As much of the genome is actively transcribed, nascent transcription protocols recover reads from much larger proportion of the genome (not only exons and introns, but also numerous intergenic regions). Consequently, if both assays are sequenced to the same depth, the equivalent nascent transcription data would have a lower per position depth.

External spike-ins in nascent RNA sequencing are also inherently different than in RNA-seq, leading to more uncertainty in the normalization process (Figure 2.1A). The gold standard for spike-ins in RNA-seq is an External RNA Controls Consortium (ERCC) library, which uses a fixed amount of known RNAs which are added to the sample to quantify the variation introduced during sample handling, library preparation and sequencing. Crucially, this RNA spike-in library is introduced in known quantities prior to the experiment. Run-on centric nascent RNA protocols seek to identify the locations of actively engaged

RNA polymerases by using marked nucleotides and a run-on reaction. Hence the ERCC spike-ins, by virtue of being mature RNAs, are incompatible with the run-on reaction. Instead, fixed amounts of nuclei from an external organism are typically added to the sample nuclei and then the run-on reaction is employed on the combination of cell types. Thus, the quantity of RNA from the external spike-in is determined by not only the efficiency of the protocol and sequencing, but also the efficiency of the run-on reaction. A necessary but potentially flawed assumption, then, is that all of the run-on reactions have the same efficiency, allowing the reads mapping to the spiked-in nuclei to be treated with the same reliability as an ERCC spike-ins. If an external spike-in is not used, many off-the-shelf RNA-seq tools are used directly for internal normalization[51, 69–71].

Critical to the effectiveness of internal or external normalization are the assumptions about what remains invariant. Notably, when run-on reactions are performed in the presence of a perturbation, nascent RNA sequencing contains a unique internal set of invariant data. RNA Polymerase II loads at the 5′ end of a gene and then proceeds through the gene with a relatively consistent processivity[32]. Thus, as first described by Mahat[62], at short time points after a perturbation, changes in transcription are not expected to have reached the 3′ end of long genes. Prior work on 3′ end normalization applied linear regression to the set of 3′ invariant ends and showed this approach was similar to other, presumably invariant, internal gene sets[62, 63]. However, they did not directly compare the approach to external spike-in controls or establish uncertainty bounds on their estimates.

In this work, we set out to compare run-on based 3′ normalization to external spike-ins. To this end, we developed a method for quantifying error in the estimation of spike-in normalization. Using this method we compare external spike-ins to internal invariant sets, focusing on the 3′ subset. We uncovered that most external spike-ins in nascent RNA assays are under-sequenced and potentially unreliable. Additionally, we find that when external spike-ins are of adequate depth and the assumptions of the 3′ normalization approach are met, the two methods show high correspondence.

2.2 Results

2.2.1 An algorithm to quantify error in spike-in normalization estimates

When normalizing between samples, there are different approaches to computing normalization factors from the invariant set, whether that set is an external spike-ins or internal[65] (Figure 2.1B). The most naive of these is to take the simple ratio of reads mapping to the invariant set between two samples and use that as a normalization factor (Figure 2.1B, left). However, this reduces the information contained within the set to a single summary value. The alternative approach is linear regression, where estimates of counts per invariant entity, typically genes, are used as data points for the fitting algorithm and the resulting slope is used as an the normalization factor[65] (Figure 2.1B, middle). In this way, transcription levels across different orders of magnitude can be leveraged to give a more accurate normalization factor. Thus, prior work in nascent transcription has often used naive linear regression to estimate normalization factors instead of a simple point estimate[62, 63]. However, to use linear regression, a sample's spike-ins must be of sufficient depth that a linear relationship exists in the count data. Additionally, naive linear regression does not provide error bounds.

To quantify the error inherent in estimating a normalization factor from data, we developed a hierarchical Bayesian version of the linear regression framework (Figure 2.1B, right). Typically linear regression is formulated as:

$$y \sim mx + b$$

which describes the relationship between counts in two samples x and y in terms of two variables (m and b) the slope and intercept, respectively. In this framework, the slope (m) is interpreted as the best normalization factor between the two samples. In the naive context of normalizing to a spike-in (without considering the error of the estimate), this typically works well, as counts span multiple orders of magnitude and typically form a linear relationship between samples[65]. However, in standard linear regression only a single point estimate for the parameters is obtained.

To quantify the error in the estimated normalization factor, we extend the naive linear model above

to incorporate an estimation of the error in log-space, backed by biologically informed count distributions. In the simplest terms, we generate a linear model whose mean is a normal distribution defined by the log-transformed ratio of our read counts[72, 73], plus an intercept term. By using the log-transformed ratio of read counts, we can assume the slope is normally distributed:

$$\mu_{\text{slope}} \sim \text{Normal}\left(\text{mean} = \log_2 \frac{Y}{X}, \sigma_{\text{mean}}\right)$$

Where μ_{slope} is the desired mean (normalization factor) and the resulting variance estimate σ_{mean} is then used as an estimate on the error of that normalization factor.

Our model is shown formally as a plate diagram in Figure 2.1C. To fully specify the model, we assume the intercept follows a Normal distribution, Intercept ~ Normal. The input data for this model is formally a counts matrix, M where $M_{i,j}$ represents the number of reads in sample i in region j. For all samples M_i, we select a single reference sample M_r to normalize against. We first model the count data over regions of interest as a Negative Binomial Distribution, as we expect the count distribution to be over-dispersed. This yields two variables – X and Y which describe to the count distribution of each sample input to the model. Priors for σ variables are selected to be uninformative using the conjugate InvGamma(1, 1) [74], while priors for X and Y are defined as $\mu_X = \text{mean}(M_i + 1)$ and $\mu_Y = \text{mean}(M_r + 1)$ to reflect the log-transformed ratio of Laplace smoothed count data.

We call our new method Virtual Spike-In (VSI) and leverage Markov Chain Monte Carlo (MCMC) methods to fit the underlying distributions. The input to the model is a set of data points between two samples, thus this model can also be applied to both external spike-ins and internal invariant sets of regions, such as the unperturbed 3′ end of long genes, or to any other set of invariant regions shared between two samples that behave as count data. A technical discussion of implementation details for this model is available in the Methods section of this paper.

2.2.2 Confidence in normalization factor estimates depends on adequate spike-in depth

To assess the correctness of our VSI implementation and approach, we first compare the method to the standard linear regression approach. To this end, we processed samples of human cells with

Drosophila spike-ins from a number of previously published studies employing nascent RNA sequencing data[75–88]. After filtering for samples with replicates and a nonzero number of reads mapping to the dm6 Drosophila genome, we were left with $n = 180$ samples (Table S1, see Methods for complete details on data processing).

When running the VSI model on external spike-ins from published data[75–88], we find that it reliably recapitulates the results of naive linear regression (Figure 2.2A), but now provides error bars on these estimates. In the regime of small normalization factors (values near zero), both linear regression and the VSI model perform essentially identically. Importantly, when the absolute value of linear regression estimates are large, the VSI approach tends to recover a comparatively lower normalization factor, likely a consequence of the model being more resistant to noise and extreme values than linear regression alone. However, large normalization factors suggest extreme differences in sample efficiencies which should call into question whether the data and spike-in are of sufficient depth and quality to be trusted. A detailed examination of the posterior distribution variance shows higher variability at low spike-in sequencing depth (Figure 2.2B). The posterior variance (the variance of the estimated normalization factor after fitting the model) generally improves at depths greater than 10X the dm6 reference transcriptome, using a Drosophila transcriptome length of 30Mb[89]. Unfortunately, the majority of published samples are below this spike-in depth (Figure S1). This suggests that most published nascent RNA sequencing experiments using external spike-ins are under-sequenced, which may be a consequence of either an ineffective run-on reaction or a choice to prioritize sample read depth over spike-in read depth.

2.2.3 Evaluation of error in external and internal normalization

Normalization across invariant regions need not be limited to a spike-in, although an external spike-in is typically preferred. In theory, any set of invariant regions in a sequencing data set that follow a count distribution can be used to estimate a normalization factor between samples. This makes the Virtual Spike-In a versatile and widely useful model for quantifying normalization error across invariant regions.

As an example, our model can leverage reads at the $3'$ end of long annotated genes, building on prior

work[62, 63] (Figure 2.3A). Nascent RNA assays survey engaged RNA polymerases genome-wide, which for any singular time point can be anywhere along the gene. However, in the presence of a perturbation, changes in transcription levels must originate at the 5′ end of genes, either by altering RNA polymerase II's loading and/or release from pausing. Once released, RNA polymerase II then proceeds through the gene at a relatively consistent rate[32, 63]. For example, in human cells RNA polymerase II has an elongation rate of roughly $2-3\frac{\text{kb}}{\text{min}}$[90–94], although this rate can be highly variable. Therefore, at short time points, there is insufficient time to alter RNA polymerase II profiles at the 3′ ends of a long gene (see Figure 2.3A).

Under this model, we note that RNA polymerase II profiles at genes past **Length Threshold** = **Elongation Rate**·**Time Point** should retain a consistent level of baseline transcription unperturbed by the experiment. Using this assumption, the invariant 3′ gene regions can be used for normalization between samples. Previous work[62, 63, 95], used a simple linear regression model to determine a normalization factor, defined by the slope of the best fit line, between the two samples using 3′ regions. However, these models did not establish uncertainty bounds on the accuracy of their normalization factors and did not compare their methodology to external biological spike-ins to quantify its effectiveness.

We leveraged the VSI approach to compare the 3′ normalization to external spike-in controls (Figure 2.3B). For consistency of comparison between different experiments, and considering the typical timelines used, we selected a 180kb ($60\text{min}\cdot 3\frac{\text{kb}}{\text{min}}$) threshold for all samples when looking at the 3′ invariant region. We also exclude the last 500bps of the annotated gene from our normalization to reduce variance from the characteristic 3′ bump associated with termination in nascent RNA sequencing experiments. This results in 1198 3′ invariant regions used for normalization by the VSI model (roughly 10% of annotated RefSeq genes). Using this set, we found that the correspondence between the 3′ normalization approach and external spike-ins (Figure 2.3B) showed extensive variation. In fact, the internal and external normalization factors were only rarely the same (diagonal line). Thus, we next sought to determine which factors influence the 3′ normalization method's fidelity.

We first consider time points below the 60 minute threshold utilized. As the posterior estimate of the normalization factor varies dramatically below 10X spike-in coverage (Figure 2.2B), we first consider only samples with stable estimates (spike-in coverage > 10X). For these samples, there is generally good

concordance — small differences as most points are near the origin — between the 3′ normalization and external spike-in approach (Figure 2.3C). Notably, two data sets show strikingly lower concordance between the two methods. These two data sets were samples where NELF (negative elongation factor) was depleted and the cells were subjected to heat shock[75]. The lack of concordance between the methods suggests that the depletion of NELF may have had genome-wide effects on RNA polymerase, a condition that calls into question the invariant nature of any internal set.

At low external spike-in depth, inadequate spike-in data may exist for confidence in linear regression. Consistent with this notion, low depth spike-in samples have higher posterior estimate variance (Figure 2.2B). However, despite this increased uncertainty, we found good concordance between the spike-in and the 3′ normalization estimates (Figure 2.3D).

Importantly, the 3′ normalization approach inherently assumes that portions of genes are unreachable at the specified time point of the experiment. By using a uniform 60 minute assumption, we could determine whether the concordance between the 3′ approach and external spike-ins breaks down at longer time points, when the assumed invariant regions can no longer be assured to be unchanged. As expected, when the internal set contains regions that could be varying between the samples (e.g. the time point is longer than the 60 minute assumption), there was increasing discordance between the two normalization methods (Figure 2.3D,E), particularly when long time points co-occurred with low coverage (Figure 2.3F). Intriguingly, even in the data that fail to meet our assumptions (low depth + long time, Figure 2.3F) we observe a small cluster of samples close to the origin of the plot. In these scenarios, we achieve concordance between internal and external spike-ins even when all assumptions are violated, as in these cases the perturbation happens to not strongly impact the long gene set used by the VSI normalization.

Collectively, these results suggest that the 3′ internal normalization approach gives results similar to the linear approximation of external spike-ins when the assumptions of the model are met. This is particularly true when the normalization factors are small (e.g. near the origin in Figure 2.3B-F). When the assumptions of the VSI model are violated, either with long time points or disruptions that alter RNA polymerase itself, the two models strongly disagree.

To further characterize this pattern, we next turned our attention to the examination of a single

high quality data set that contains multiple time points and roughly average spike-in sequencing depth (GSE96869)[79]. In this study, Dukler et al. treated K562 cells with the natural drug Celastrol, which activates mammalian heat shock response[79]. Cells were then assayed at several time points including 10 min, 20 min, 40 min, 60 min and 160 min. This PRO-seq data set has spike-in sample depth ranging from 0.7 to 1.1X Drosophila transcriptome coverage. Importantly, the cells undergo replicative arrest around the 40 minute time point. As before, we employ a 180kb ($60\text{min} \cdot 3\frac{\text{kb}}{\text{min}}$) threshold for all samples when looking at the $3'$ invariant region. For each sample, we compared normalization results using the $3'$ internal normalization to external spike-ins, using both linear regression (VSI) and the ratio based point estimate.

We observe that the VSI model shows good concordance between internal ($3'$) and external spike-in estimates of the normalization factor, particularly at early time points (Figure 2.4). After the onset of replicative arrest (t=40 min), the internal and external normalization factors begin to diverge, though only modestly in both the 40 minute and one of the 60 minute time point replicates. As expected, the largest deviations between the $3'$ and external spike-in are observed at 160 minutes, when the time point is well beyond the 60 minutes assumed by the internal normalization. At all time points, the single point estimate of the external spike-in deviates substantially from both the linear model estimate of external spike-in and the $3'$ approach, consistent with prior work on normalization approaches[65].

2.2.4 Downstream effects of normalization

Normalization factors are crucially important in downstream analyses of high throughput sequencing data. To that end, we next compared the results of differential expression analysis on the Dukler data set[79]. For differential expression analysis, we used DESeq2[51], which uses an internal normalization approach. Specifically, DESeq2 calculates a size factor as the median ratio of counts over every gene in the sample divided by the geometric mean of counts at that gene over all samples. The result is an effective method for normalization that implicitly assumes that most genes are unchanged across the comparison.

We sought to compare the default DESeq2 size factor approach to the $3'$ internal normalization method. For this comparison, we performed differential expression analysis between the 0 minute and 60

minute time points (Figure 2.5A, 40 minute comparison shown in Figure S3).We observed that the posterior point estimate for the normalization factor recapitulate a strict subset of genes called as differentially expressed by the automatically estimated size factors (Figure 2.5). In simpler terms, it appears that 3′ internal estimated normalization factors are more conservative, effectively decreasing the set of genes called as significant. Arguably the VSI set is both more conservative and based on a biologically principled invariant set of data compared to the DESeq2 method.

In both cases, a single normalization term is calculated and presumed to be correct. Our earlier comparison to external spike-ins (Figure 2.3) suggests two estimators may reach similar but not quite the same normalization factor. Therefore, we next sought to ascertain the extent to which minor, plausible fluctuations in the calculated normalization factor might influence differential expression analysis. To this end, we use a sampling approach. We ran 1000 simulations sampling normalization factors from the posterior distribution estimated by VSI for each of the 4 samples (10 min, 60 min; 2 replicates at each time point). We then ask how often a particular gene is called as significant across the samples. We observe that many of the genes called by DESeq2 as differentially expressed (red dots in Figure 2.5A) have relatively low reproduciblity across the range of plausible normalization factors (Figure 2.5B) and are therefore potential false positives. Notably, the genes with the highest reproducibility are those found by the VSI 3′ point estimate (purple dots in Figure 2.5A correspond to red dots in Figure 2.5B).

2.3 Discussion

We present Virtual Spike-In, a novel approach that uses a hierarchical Bayesian regression model to calculate normalization factors and quantify their uncertainty for nascent transcription datasets. We use this method to compare 3′ end normalization in run-on based nascent RNA sequencing experiments to external spike-in controls. We find that while the internal and external normalization rarely perfectly agree, the 3′ end normalization shows high concordance to external spike-in controls when assumptions of the method are met. Additionally, normalization is known to have strong effects on analysis results[65], and our work further supports this conclusion.

While external spike-ins are typically assumed to be the gold standard for normalization of se-

quencing samples, we find that external spike-ins in published run-on based nascent RNA sequencing experiments are typically under sequenced. Importantly, external spike-ins in nascent RNA sequencing are not the same as those in RNA-seq. This makes the entire normalization process significantly more challenging. Using spike-in nuclei inherently assumes that for every sample, the efficiency of the run-on in the spike-in nuclei closely matches the efficiency of the run-on in the experimental nuclei. There is no reliable mechanism to determine if this assumption is correct. This problem is exacerbated by the relatively low read depth of most external spike-ins in nascent RNA assays. It is critically important that any normalization technique be based on adequate data, as even the best normalization model is limited by the available data.

The alternative to external spike-ins is to use an internal invariant set. Run-on based nascent transcription coupled to a perturbation has a unique invariant set in the 3′ ends. While 3′ end normalization is powerful, it has a number of important limitations compared to an external spike-in. First, the elongation rate of RNA polymerase II in the organism must also be known. At any given elongation rate and time point, a reasonable proportion of genes in the genome must be sufficiently long that invariant regions exist at the time point of interest. While this works well in the human genome, it is likely not the case for organisms with smaller genes and genomes. Even in the human genome, when the normalization factor is estimated on later time points, it is based on increasingly smaller quantities of data, leading to less certainty. With that said, the use of a Bayesian model in this context does make the model robust to a small number of genes to be normalized against. Finally, the 3′ end approach cannot be used in the absence of a perturbation or if the perturbation could alter previously loaded RNA polymerase.

In addition to the assumptions made about the model, it is also important to consider the assumptions made about the selected 3′ regions if performing normalization internally. First and foremost, low expression is a persistent concern across all experiments and must be considered here. Undersequencing is, in general, a problem for normalization (of external spike-ins or of 3′ invariant regions) and downstream analysis. Consequently, if a sample is of low sequencing depth, either generally or particularly at the 3′ end of genes, we recommend it be excluded from further analysis for quality concerns. Likewise, our 3′ assessment depends on the accuracy of gene annotations and the presumption that long genes are not in

some way atypical. Finally, the presence of intronic bidirectional signals (e.g. same strand overlapping transcription) could be problematic if the bidirectionals both reside within the invariant 3′ region and are themselves differentially transcribed. Despite these caveats, one benefit of 3′ end normalization is that it can be applied to many previously published run-on based nascent RNA sequencing data sets where an external spike-in is not present.

There are a number of nascent transcription assays that do not use a run-on step, and normalization for these assays present distinct challenges. Metabolic labeling approaches expose live cells to marked nucleotides over some time frame before the experiment[68, 96]. As such, both the profile and signal to noise characteristics of the data are influenced by the time and efficiency of the labeling process. In contrast, mammalian native elongating transcript sequencing (mNET-seq)[97] uses an antibody to pull down a component of the RNA polymerase II complex. As such, normalization of mNET-seq data is conceptually similar to ChIP-seq and should account for antibody efficiency. Further work is needed to characterize both internal and external normalization strategies for metabolic labeling and antibody oriented nascent transcription assays.

The Virtual Spike-In model is versatile. As the input to normalization is counts over a collection of regions, the VSI method can be applied to both internal invariant sets, such as the 3′ end normalization used here, and to external spike-in controls. Another notable advantage to the VSI technique is that it establishes error bounds on the calculated normalization factors, an important but often overlooked aspect of the data analysis. Effectively quantifying error in the point estimations of normalization factors is an important addition over the naive linear model. Quantification of error is essential to analyzing nascent RNA sequencing data rigorously. Ultimately, nascent RNA sequencing experiments appear to need a more reliable mechanism for external normalization, which is challenging given the limitations of the underlying protocols.

2.4 Methods

Our model is implemented in the Python programming language using the pymc3 MCMC library[98]. Inference is performed using an adaptive sampler, combining the No-U-Turn Sampler[99]

(NUTS) for continuous variables with a Metropolis-Hastings Sampler[100, 101] for discrete variables, using 25,000 iterations after a burn-in period of 2,500 samples. The number of iterations can be increased if a greater assurance of convergence is desired. A larger number of iterations are required for convergence of the discrete distribution due to the use of a Metropolis sampler instead of NUTS (Figure S5). Source code is available at `https://github.com/Dowell-Lab/virtual_spike_in`.

For both the human cell lines and **Drosophila** spike-in, reads were mapped to the hg38 and dm6 reference genomes respectively using the Nascent-Flow pipeline[102]. Counts were determined for all genes using featureCounts[103], considering only the maximally expressed isoform and counting reads per gene including exons and introns.

Figures

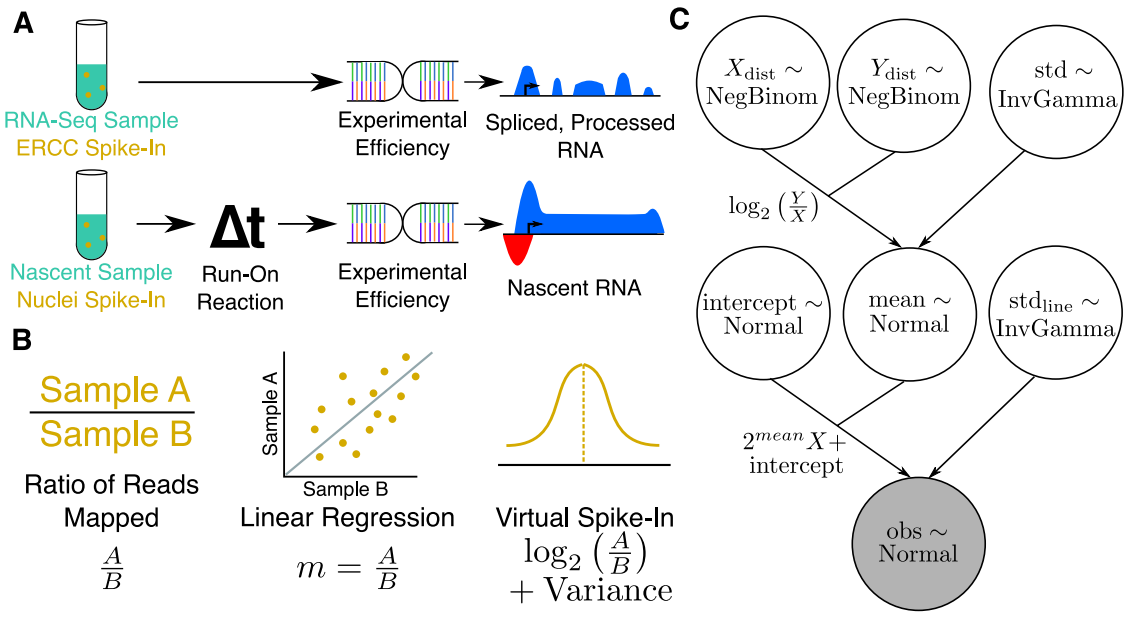


Figure 2.1: **A Bayesian model describing normalization data for nascent RNA sequencing data. A:** Schematic showing typical external control, handling, and resulting data profile differences between RNA-seq (top) and run-on nascent RNA sequencing assays (bottom). Note that run-on efficiency is assumed to be equivalent between spike-in nuclei and experimental nuclei. **B:** Quantifying a normalization factor is accomplished either by a naive ratio of total reads approach (left), linear regression (middle), or by the Bayesian model proposed here (right). Linear regression (middle) is more resistant to noise and outliers, but does not provide a reliable way to measure the variance of the normalization estimate. The Bayesian model (right) converts the slope $m = \frac{A}{B}$ to log space, converting the multiplicative nature of the normalization factor to a linear one, for which normalization factors can be readily inferred as a normal distribution with variance. **C:** A plate diagram showing the VSI model as implemented in pymc3. Briefly, we estimate our count distributions X and Y (top row) with a negative binomial. The ratio of two negative binomial distributions is approximately log-normal, so we derive a normal distribution called **mean** (middle) as the log of the ratio of Y and X with some variance (top right), estimated as an inverse gamma distributed random variable. With the estimation of the mean established, we then add additional parameters to describe the intercept, and variance of the actual line of best fit. This is done so that the parameter **mean** is estimating an error in log-transformed space, as discussed in Panel **B**.

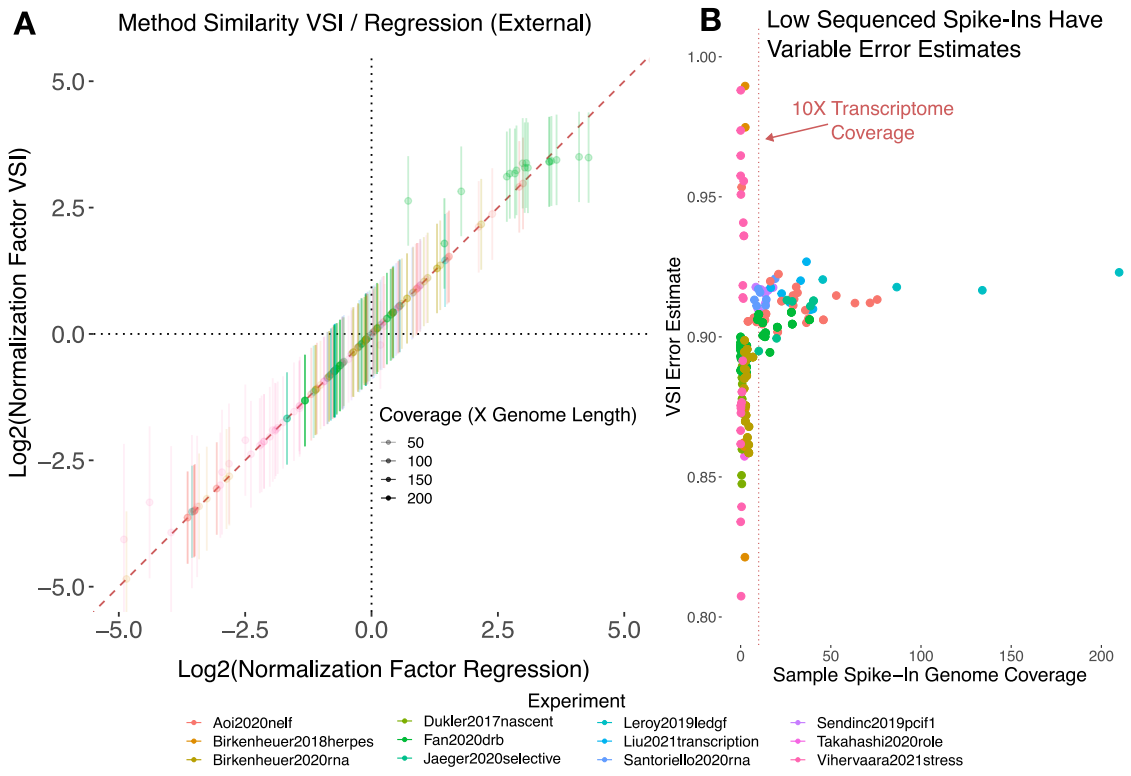


Figure 2.2: **Spike-ins have unusual behavior at the extremes.** To assess where our model diverges in behavior from linear regression, we ran the VSI model on data from a number of published experiments[75–88]. Within each experiment, samples were grouped by condition and analyzed within those groups. All samples had Drosophila spike-ins, so annotated Drosophila genes were selected as the invariant set to count over. **A:** Comparison of regression factors inferred by linear regression (x-axis) to those inferred by the Bayesian VSI model (y-axis). Estimates are shown along with an error bound of $\pm\sigma$. Notably, the regression estimate (x-axis) and VSI estimate (y-axis) deviate most dramatically when the absolute value of the normalization factor is large. **B:** When we plot the depth of coverage of the spike-in (x-axis) against the VSI error estimate (y-axis) shows samples with less than 10× spike-in transcriptome coverage are less consistent than those above this threshold (dotted red line). Of note, error estimates range between 0.8 and 1.0, but when applied to the data they must be converted out of log2 space and multiplied by the normalization factor. Hence the impact of the error will scale with the normalization factor size. In a biological context, this is good — samples with large normalization factors have less confidence indicating poorer experimental efficiency and reproducibility.

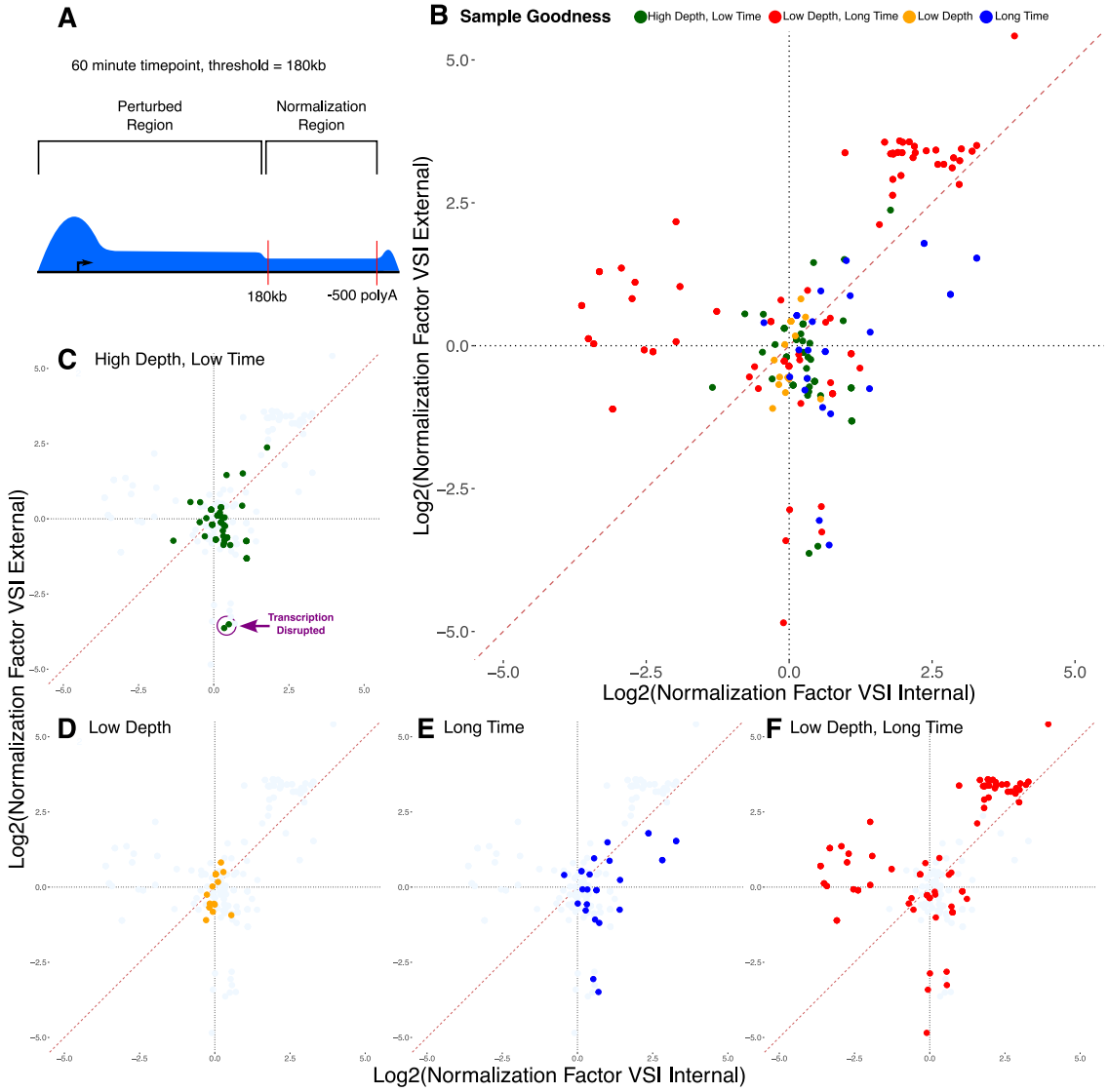


Figure 2.3: 3′ **Normalization estimates depend on assumed polymerase elongation behavior and sequencing depth A:** A cartoon showing the characteristic shape of nascent RNA sequencing samples after a perturbation. RNA polymerase II loads at the 5′ end of genes, thus after a perturbation alterations in transcription levels can only reach a distance that depends on the processivity of RNA polymerase II. In this work we assume 3kb/min and hence for a 60 minute experiment the perturbation influences the first 180kb ($60\text{min} \cdot 3\frac{\text{kb}}{\text{min}}$). **B:** We compared external spike-ins (y-axis) to 3′ internal normalization across a large collection of previously published data. Samples are colored by whether they **C:** meet both time point and depth assumptions (green), **D:** have low sequencing depth (< 10X spike-in transcriptome) (orange), **E:** have time points beyond the 3′ assumed 60 minutes (blue), or **F:** meet neither assumption, being of both low spike-in depth and long time point (red). Notably, two samples in (circled in **C**) meet the coverage and time constraints of the 3′ normalization approach but involve depletion of NELF under heat shock conditions, which likely alters RNA polymerase elongation.

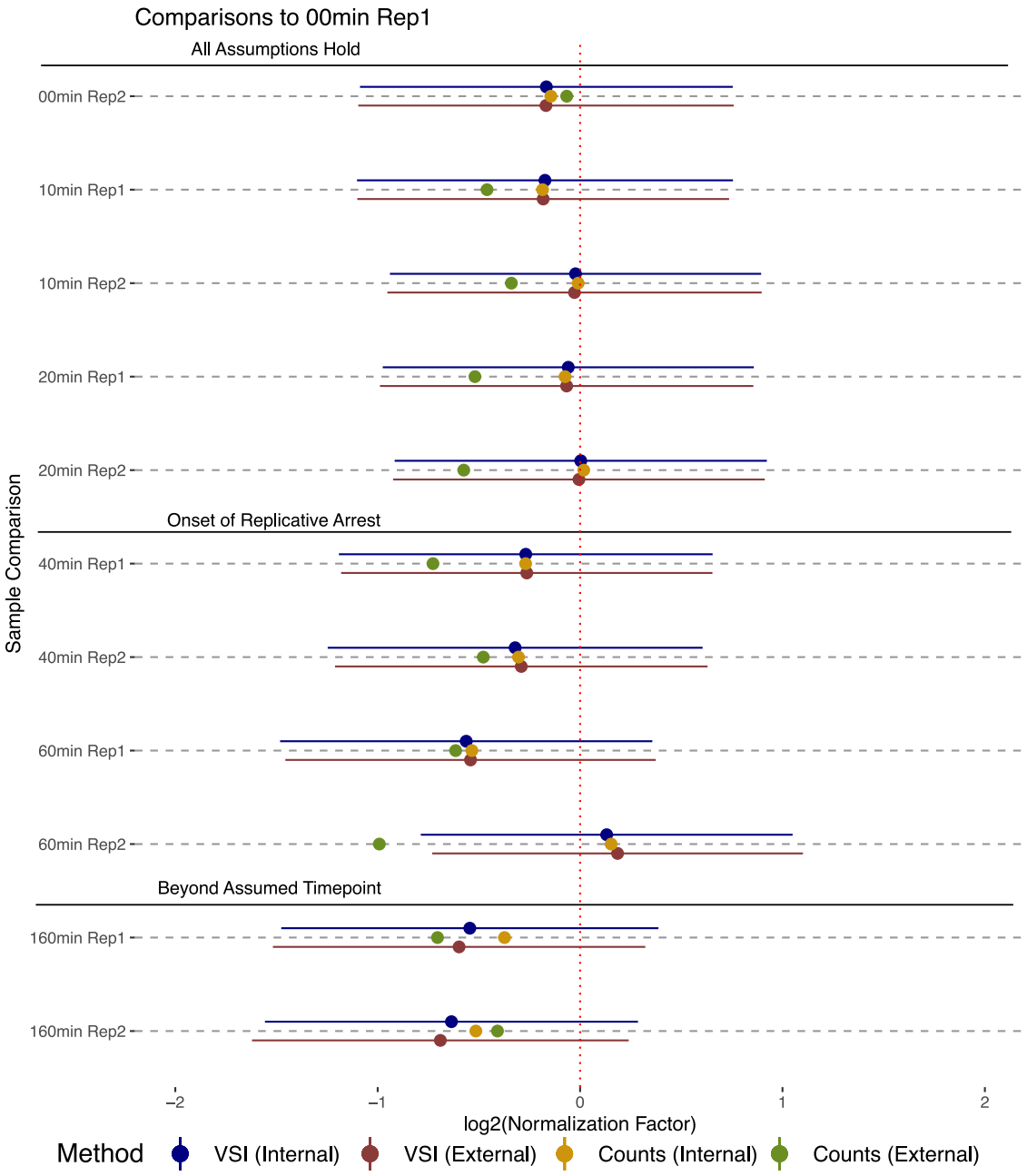


Figure 2.4: **Comparison of all normalization methods on good quality data.** We compare normalization factors on a high quality data set[79] (GSE96869) computed by four distinct methods: VSI applied to the internal 3′ invariant gene set (blue), VSI applied to an external Drosophila spike-in (red), the ratio approach applied to the 3′ invariant gene set (yellow), and the ratio approach applied to the Drosophilia spike-in (green). Error bars are shown for the VSI estimates. The 3′ invariant set uses a threshold of 180 kb (60 min), regardless of the data time point. For orientation, we note the normalization factor of zero (red dotted line), the onset of biological replication arrest and the assumed time point for the 3′ invariant gene set.

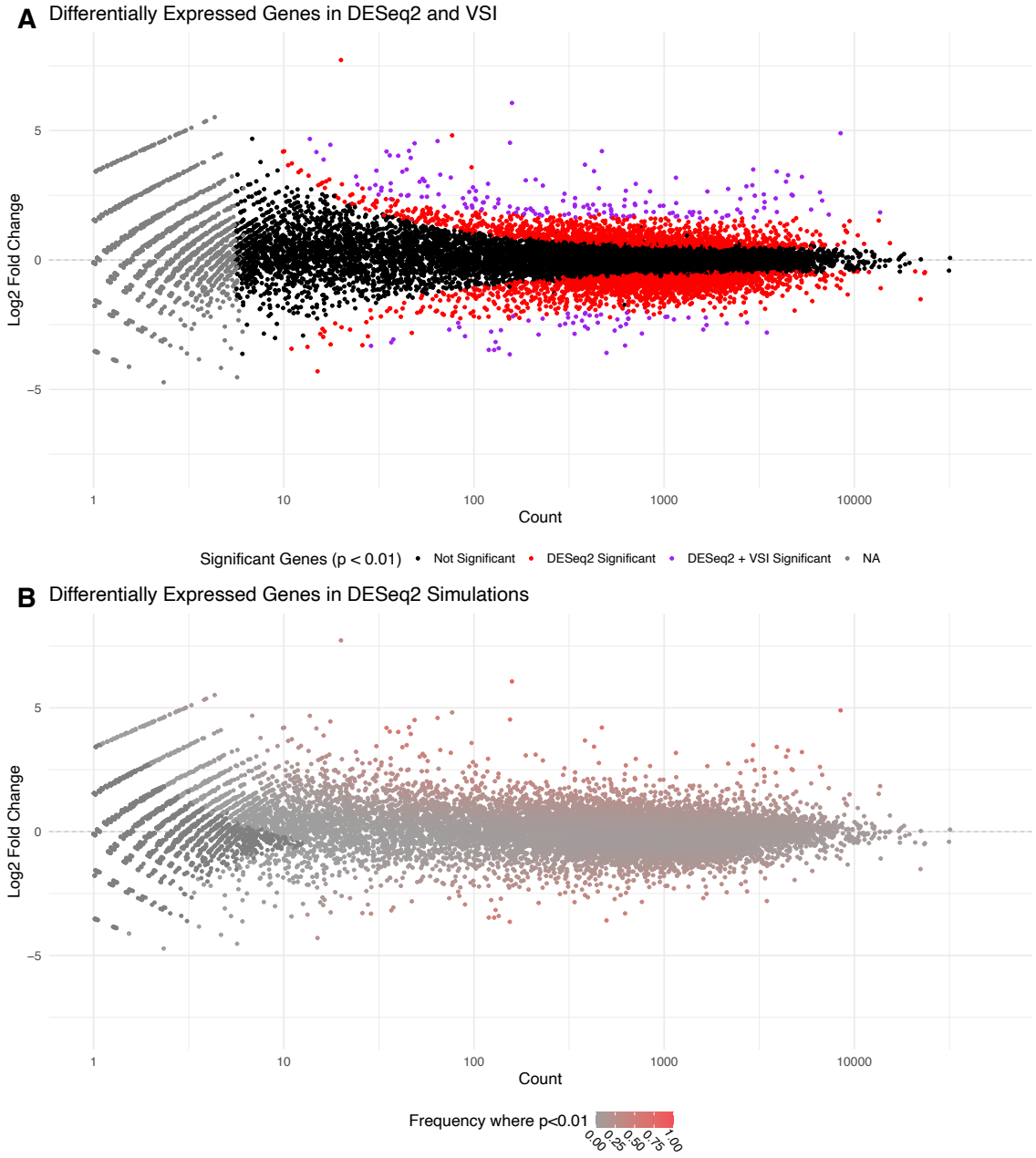


Figure 2.5: **Estimated Normalization Factors Provide Strict Cutoffs for DESeq2 A:** Differential expression analysis by DESeq2 (adj. p-val < 0.01) using size factors estimated from DESeq2 (red) and the VSI model (purple) on 3′ invariant regions. Note that DESeq2 calls normalization factors "size factors". The more conservative VSI identified set (purple) is a strict subset of the DESeq2 identified significant set. **B:** Consistency of differential expression calls across a broad range of plausible normalization factors. Genes are colored based on the reproducibility of statistically significant differential expression (DESeq2, adj. p-val < 0.01) across 1000 iterations where normalization factors were sampled from the posterior distribution estimated by VSI. Points that appear as significant most often are also those that are called as significant using both DESeq2 size factors and VSI 3′ normalization (Panel A, purple).

Chapter 3

Deconvolution

All of this chapter has been previously published as Maas, Z. **et al. Deconvolution of Nascent Sequencing Data Using Transcriptional Regulatory Elements** Preprint (Bioinformatics, Oct. 2023). (2023). Zachary Maas was responsible for the design of experiments, implementation of all project code, figure design, drafting and revision of the manuscript, and comments during peer review. Robin Dowell provided intellectual consultation, assistance on experimental design, and assistance during manuscript revision and during peer review.

Abstract The problem of microdissection of heterogeneous tissue samples is of great interest for both fundamental biology and biomedical research. Until now, microdissection in the form of supervised deconvolution of mixed sequencing samples has been limited to assays measuring gene expression (RNA-seq) or chromatin accessibility (ATAC-seq). We present here the first attempt at solving the supervised deconvolution problem for run-on nascent sequencing data (GRO-seq and PRO-seq), a readout of active transcription. Then, we develop a novel filtering method suited to the mixed set of promoter and enhancer regions provided by nascent sequencing, and apply best-practice standards from the RNA-seq literature, using **in-silico** mixtures of cells. Using these methods, we find that enhancer RNAs are highly informative features for supervised deconvolution. In most cases, simple deconvolution methods perform better than more complex ones for solving the nascent deconvolution problem. Furthermore, undifferentiated cell types confound deconvolution of nascent sequencing data, likely as a consequence of transcriptional activity over the highly open chromatin regions of undifferentiated cell types. Our results suggest that while the problem of nascent deconvolution is generally tractable, stronger approaches integrating other sequencing protocols may be required to solve mixtures containing undifferentiated celltypes.

3.1 Introduction

One key problem of interest when studying transcription is the ability to capture the heterogeneity that exists in true biological samples[56]. Bulk sequencing samples from cells are an aggregate across a

cellular population, and thus average out differences between individual cells to capture only an ensemble profile of a given sample. Notably in the case of samples taken from tissues composed of heterogeneous constituent cells, any celltype specific differences are not necessarily discernible in the heterogeneous mixture of expression data.

To some extent, this problem has been at least partially solved in the context of RNA-seq with the emergence of single cell RNA-seq protocols which allow for RNA content at the level of the individual cell to be measured[35]. However, the relatively high cost of sampling deeply limits the use of scRNA-seq in many contexts. Consequently, a great deal of work has been done to separate samples into constituent cell types **in silico**. This task is interchangeably referred to as deconvolution or microdissection. Deconvolution has been studied extensively in the context of both microarray data and in RNA-seq[53, 54, 56, 58, 59], but has seen only limited application to other high throughput genomic data.

Nascent transcription protocols[29, 33] are of particular interest for studies into transcriptional regulation[105, 106]. Nascent sequencing protocols profile active RNA Polymerase II activity, which captures enhancer associated RNAs (eRNAs), short unstable transcripts that are often associated with transcription factor binding sites[107]. These eRNA transcript have proven to be highly informative markers of transcription factor activity[105, 106, 108–112]. Unfortunately RNA-seq, whether bulk or single cell, does not capture enhancer associated transcripts due to the fact they are unstable and not polyadenylated[107]. For this reason, the theoretical possibility of single cell measures of nascent transcription has tremendous potential for understanding regulation and transcription factor activity in key biological processes including development and disease progression.

Today, nascent sequencing protocols still operate only on the bulk level, largely because nascent protocols are relatively onerous, taking up to a week to process a set of samples[29, 33, 67]. Because nascent protocols capture RNA production, many of the signals arise from lowly abundant, highly unstable RNAs [107]. Furthermore, with current biochemical efficiencies, a single cell nascent sequencing protocol is likely infeasible, and thus deconvolution is needed to dissect nascent transcription profiles within tissues.

Nascent transcription data has relatively unique properties compared to RNA-seq. First, RNA-seq

measures steady state mature, stable RNA levels which tend to be of relatively high abundance. In contrast, nascent sequencing protocols cover a much larger proportion of the genome (~ 40% as opposed to ~ 8%)[67]. The consequence is that the average sequencing depth per transcript is typically lower in nascent data, in spite of often sequencing samples to a higher depth. Second, many transcripts measured in nascent protocols are unannotated, lowly transcribed, unstable eRNAs (Figure 3.1A)[67, 107]. In development, enhancer activities are the first changes detectable when a cell undergoes state change, suggesting their associated eRNAs have high potential as cell type markers[113]. Furthermore, enhancer associated RNAs tend to be more cell type specific than protein coding genes[114]. However, their low transcription levels lead to issues of reliable detection[67]. Thus methods developed for RNA-seq must be appropriately adapted to use with nascent sequencing data.

Here, we use standardized methods for supervised deconvolution to nascent sequencing data, applying a newly developed filtering technique to solve problems presented by nascent data in the deconvolution context. We show that deconvolution of nascent sequencing data works reliably, albeit with different model performance than in RNA-seq. We find that eRNAs present an informative set of information for deconvolution that can be inferred without a reference annotation. Furthermore, we find that undifferentiated celltypes confound deconvolution of nascent sequencing data, likely because their transcriptional expression resembles that of an aggregate of different differentiated celltypes.

3.2 Results

The problem of supervised deconvolution with sequencing data is formulated as follows: **Given sequencing samples from homogenous cell types and a heterogenous sample made up of those cell types, can we estimate the mixing proportions of those constituent cell types?** The problem of supervised (or partial) deconvolution is typically formulated as a linear system (Equation 3.1) [52, 53].

$$X = AS \tag{3.1}$$

Here, X is a single-row matrix with one column per region of interest (ROI) ($1 \times g$), A is a single row matrix with one column per reference homogenous cell type ($1 \times s$), and S is a matrix with one row per sample

and one column per ROI ($s \times g$). In most contexts, regions of interest (ROIs) correspond to annotated genes.

This is a overdetermined linear system, since the number of ROIs far exceeds the number of constituent cell types. Additionally, because these are biological values sampled from a noisy process, the key challenge is minimizing errors when solving the system. Most work in the literature has sought to solve the issues of this system in the context of RNA-seq or microarray[52–59] data, with limited applications of this approach to other kinds of sequencing data.

For RNA-seq, a large variety of tools and approaches have been developed[52–59], which approach the problem using different models, constraints, and regularization approaches, as well as different ways to shrink the linear system. Many of these approaches claim to be the state-of-the-art, with most tools providing good performance. Consequently, we first examine the deconvolution problem on nascent sequencing using annotated genes and methods developed for RNA-seq.

3.2.1 Deconvolution on annotated genes

To evaluate existing deconvolution methods on nascent sequencing data, we first identified a number of high quality nascent sequencing data sets from a variety of cell types (see Table 3.2.1). Samples were processed using a standardized analysis pipeline[115] which includes quality control, mapping and bidirectional transcript identification. These bidirectional transcripts originate from both gene start sites and regulatory elements such as enhancers (Figure 3.1A). The non-gene associated bidirectionals are often referred to as enhancer associated RNAs, or eRNAs.

As a first test, we examined only annotated protein coding genes to mimic deconvolution analysis typically done in RNA-seq. Notably, nascent data differs from RNA-seq in that splicing information is not present in nascent sequencing experiments, as RNA is collected pre-splicing. Furthermore, consistent with standards in nascent transcription analysis[67], we exclude the +300 initiation region of each gene when using featureCounts[49] to count reads (see Figure 3.1A), as this avoids the 5′ bidirectional peak.

To simulate a mixed sample, we generated 128 randomly mixed samples by subsampling reads from each reference sample. Samples used for all **in-silico** experiments in this paper were mixed proportionally

Study	GEO Accession	SRR	Cell Type
Samples used in Figure 3.1–3.3			
Jiang 2018[117]	GSM3025555	SRR6789175	HCT116
Fei 2018[118]	GSM3100195	SRR7010982	HeLA
Andrysik 2017[119]	GSM2296635	SRR4090102	MCF7
Dukler 2017[79]	GSM2545324	SRR5364303	K562
Zhao 2016[120]	GSM2212033	SRR3713700	Kasumi-1
Danko 2018[121]	GSM3021718	SRR6780907	CD4+-T-cell
Chu 2018[122]	GSM3309955	SRR7616132	Jurkat-T-cell
Samples added for Figure 3.4			
Core 2014[123]	GSM1480326	SRR1552485	GM12878
Smith 2021[124]	GSM4214080	SRR10669536	ESC
Ikegami 2020[125]	GSM4207079	SRR10601203	BJ5ta

Table 3.2.1: Samples used in this study.

from raw reads using samtools[116], and are listed in Table 3.2.1. With these randomly mixed samples, we then performed supervised deconvolution using 4 different methods which are commonly discussed in the literature — Nonnegative-Least Squares Regression (NNLS), Ridge Regression, LASSO Regression, and ϵ-Support Vector Regression (SVR). For all methods tested, we apply a nonnegativity constraint (all mixing proportions must be at least zero) and a sum-to-one constraint (all mixing proportions must sum to one), as suggested in prior work[56]. These constraints serve to make results from various deconvolution procedures interpretable as mixing weights for the linear deconvolution system. Code and supplemental materials for this project are available at `https://github.com/Dowell-Lab/DeconvolutionNascent`. We find that these methods provide generally good accuracy on deconvolution on our 128 randomly generated mixtures, although it appears that regularized methods perform more poorly than naive NNLS (Figure 3.1B,C) in certain celltypes across these mixtures. In this context, it appears that regularization does not improve accuracy at the cost of significant computational slowdowns relative to NNLS. Given these promising initial results, we next sought to shift the focus away from annotated genes to the unannotated bidirectional transcripts present at both promoters and enhancers.

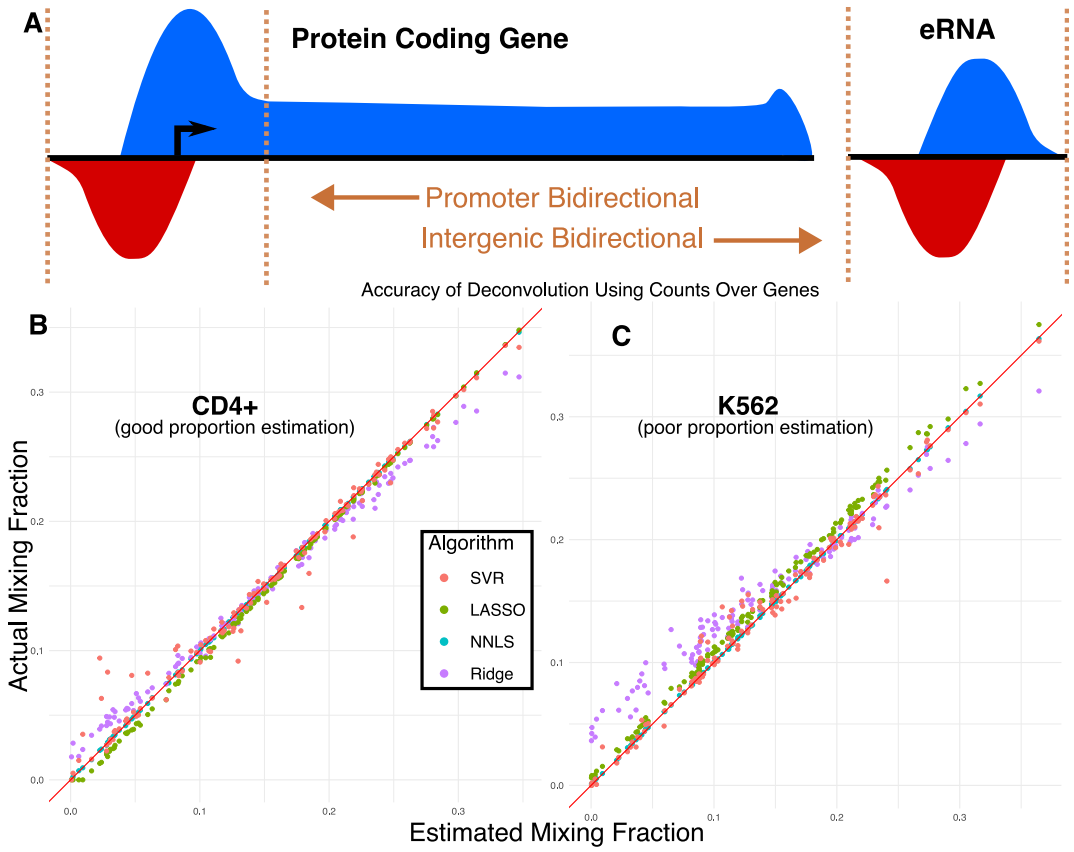


Figure 3.1: **A:** Nascent transcripts accumulate in a known bidirectional pattern around promoter sites as well as at enhancers[29, 47]. These bidirectional regions are counted by convention around ±300bp from the site of RNA Polymerase initiation (roughly the center of the bidirectional)[47, 105, 106]. For annotated genes, we exclude the initiation peak by counting +300 to the annotated transcription end site. **B:** Deconvolution was performed on random mixtures of cells from Table subsection 3.2.1. Some celltypes show highly accurate estimation of mixing proportion when doing deconvolution over all annotated genes, with most methods showing good linearity in their estimation. **C:** Other celltypes confound the regularized models used here, suggesting a systematic failure of regularization for proper estimation mixing proportion in this naive analysis. This failure appears to be more pronounced with L2 regularized methods and appears in all analyses conducted in this work, to some extent.

3.2.2 Identifying bidirectionals as regions of interest

In addition to transcription at annotated genes, nascent transcription data contains bidirectional transcription at both promoters and regulatory elements. While annotated genes are widely studied and the typical target for this class of deconvolution algorithms, the study of enhancer associated RNAs is important for understanding the regulatory landscape of the cell. Various methods exist to identify sites of bidirectional transcription [46–48, 126] and to combine them across different samples[106]. As such, bidirectionals are an additional region of interest that we now consider in our deconvolution framework.

To this end, we use a combined set of 485,688 bidirectionals, identified by Tfit and dReg within the Nascent-flow framework, capturing both enhancer RNAs and promoter regions, for all samples in Table 3.2.1 [47, 48]. Notably, this system is significantly larger than the set of protein coding genes (approximately 490,000 vs 20,000). In this work, we use the following terminology in reference to subsets of this system — Bidirectionals refers to any site of RNA polymerase II initiation and generally includes both promoters and enhancers; any bidirectionals whose $5'$ end (+/-300bp annotated TSS) overlaps an annotated $5'$ gene in the RefSeq hg38 annotation is called a promoter; all other bidirectionals are called enhancers. Given the large size of this system, we next turn our attention to filtering the set of bidirectionals, to shrink the size of the overdetermined system to make deconvolution more computationally feasible.

3.2.3 Filtering methods are useful for shrinking the system

In traditional deconvolution contexts like microarray and RNA-seq, patterns of differential expression are often leveraged to shrink the system. For example, CIBERSORT[127] uses an adaptive filtering method based on DESeq2 to find genes most indicative of specific celltypes. In the context of nascent sequencing data, however, tools like DEseq2 are problematic. The relatively low read coverage and cell type specificity of bidirectionals (e.g. inherent variability) leads DESeq2 to distrust these regions. To counter this, we developed a naive filtering scheme, selecting a fixed number of ROIs defined by the user for each homogenous reference sample where the reads for that sample were most different compared to all other samples. More formally, we define an algorithm for pruning the system of ROIs to a tractable

level:

- Filter all ROIs to restrict them to regions where all celltypes have counts lower than the 99th percentile of reads in the sample. We do this to remove outliers whose extreme values could break the assumptions of a linear system.

- Generate transformed ratio T such that for each ROI (row), for each celltype (column), that entry is the log2 ratio of the count at that ROI over the maximum count for that ROI not in that celltype. This step generates a log2 transformed list of the ROIs that are the most specific to a single celltype.

- Order this list by the largest log2 ratio in any celltype in any ROI. Then, walk down this list keeping ROIs such that the number of ROIs for each celltype is approximately equal, up to some limit of elements. This generates a subset of the full system with the most celltype specific elements for each cell. The number of ROIs is approximate because the number of celltype specific elements varies per-celltype and can be exhausted at larger system sizes.

3.2.4 Most linear methods perform with high accuracy on synthetic nascent data

Given that bidirectional regions have distinct transcription characteristics compared to more robustly transcribed annotated genes, we first sought to assess deconvolution methods on the filtered bidirectional set. Using this set, we find that deconvolution achieves a high degree of accuracy (Figure 3.2A). Unexpectedly, we observe that across all sizes of system tested (including systems far in excess of the total number of genes in the human genome), non-negative least squares (NNLS) regression performs with the highest degree of accuracy. LASSO (L1 regularized linear regression) has a close second in performance. This is likely because LASSO regularization will only drop out cell types that are unlikely to be present in the mixture. In contrast, Ridge Regression (L2 regularized linear regression) performs worse than all other tested methods for most system sizes. Similarly, ϵ-Support Vector Regression (ϵ-SVR) with L2 regularization also performs relatively poorly compared to NNLS, but relatively well compared to Ridge regression. Despite these differences in accuracy, all models perform reasonably well on our synthetic mixtures, achieving accuracy to within a few percent on randomized mixtures. This is notable because

these deconvolution methods perform well both on systems much smaller and much larger than those typically used for deconvolution of RNA-seq data.

Interestingly, we find our subsetting method consistently selects a mixture of enhancers and promoters that does not significantly differ from the distribution expected by random chance (Figure 3.2B). Consequently, this procedure captures mostly eRNAs and not promoters, since the number of eRNAs far outnumbers the number of promoters. This suggests that certain enhancer-driven regulatory elements are highly informative in identifying celltype.

We next sought to determine which ROIs were most informative to the deconvolution problem. To answer this question, we utilized NNLS, the best performing method in our prior tests. Using NNLS, we compared the performance on bidirectionals (as in Figure 3.2A), to annotated genes (as in Figure 3.1B,C) and a combination of these features – selected using our region filtering approach (Figure 3.3). We find that these methods achieve high accuracy for both genes and bidirectionals across a number of system sizes, with somewhat reduced accuracy when combining these two sets of ROIs. This reduction in accuracy could be a result of colinearity in the combined set of ROIs, as some bidirectionals may be intronic and thus they are not a strictly non-overlapping set relative to annotated genes.

For the data tested and the size of system used, we found that certain methods in the literature were prohibitively slow for the large linear systems we tested. For example, a ν support vector regression (ν-SVR) approach as suggested by CIBERSORT[127] was too computationally expensive to test or benchmark reliably, taking more than 24 hours to do deconvolution on a single mixture of cells at large system sizes (approximately 100k ROIs or more). Due to these poor scaling characteristics, we instead chose to use an optimized implementation of the primal version of ϵ-SVR. This was chosen instead of a dual formulation to maintain computational tractability for the large number of samples relative to the number of features. In the context of nascent sequencing data, NNLS is likely the best model to use based on our benchmarking.

3.2.5 Undifferentiated celltypes confound deconvolution of mixtures

In the course of testing our model, we observed that certain celltypes strongly confounded all deconvolution models tested when using bidirectionals. To understand this puzzling behavior, we

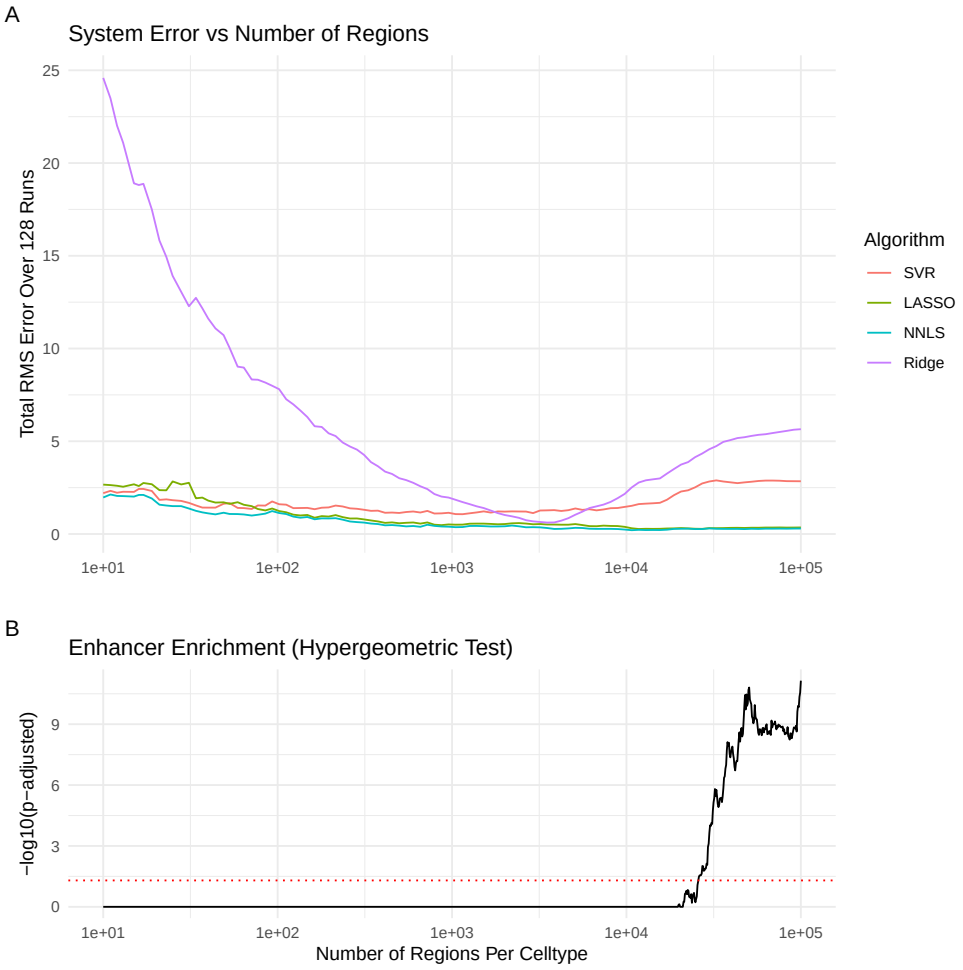


Figure 3.2: **A:** Models were tested using standard library implementations on a set of 128 randomly generated sets of mixing parameters. Each model was tested on 100 different subsets of ROIs selecting 10^n-many points for $n \in [1,5]$ using linear spacing between subsequent n. Most models perform well in the intermediate region of 10^3–10^4 points selected per-sample, but diverge outside of that regime. For each set of ROIs selected, the same 128 randomly generated sets of mixing parameters were used as in Figure 3.2. We observe that for essentially all points, NNLS outperforms more complex models. **B:** To understand the selection process of our subsetting algorithm, we tested whether enhancers were selected from the full ROI set at a greater rate than would be expected by random. To do so, we performed a hypergeometric test with Bonferroni correction over all trials of our ROI subsets. We observe that for smaller system sizes the enhancer/promoter sampling ratio does not differ dramatically from that expected by random sampling. When the system size increases, enhancers become preferentially selected over promoters ($p < 0.05$), but this increase in the rate of enhancer selection does not correlate with the accuracy of any model.

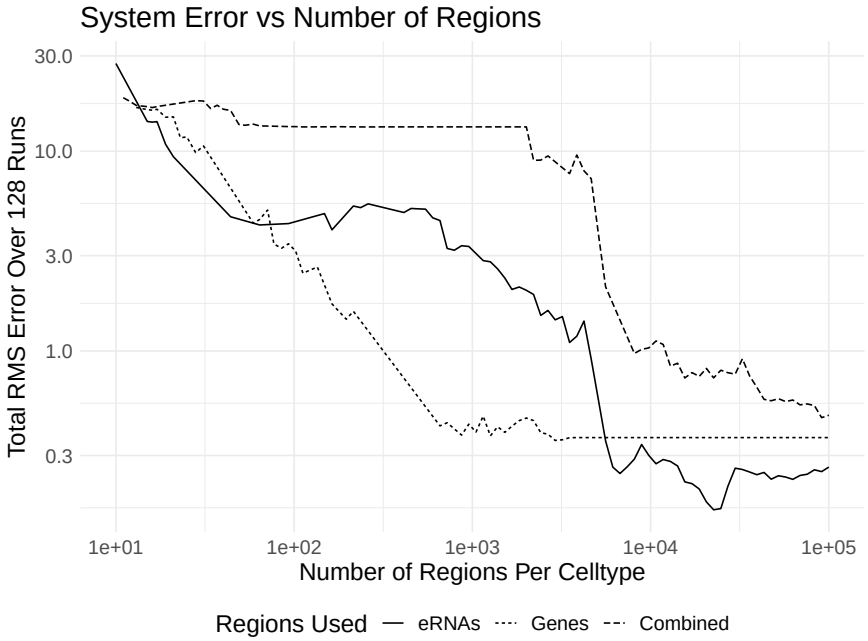


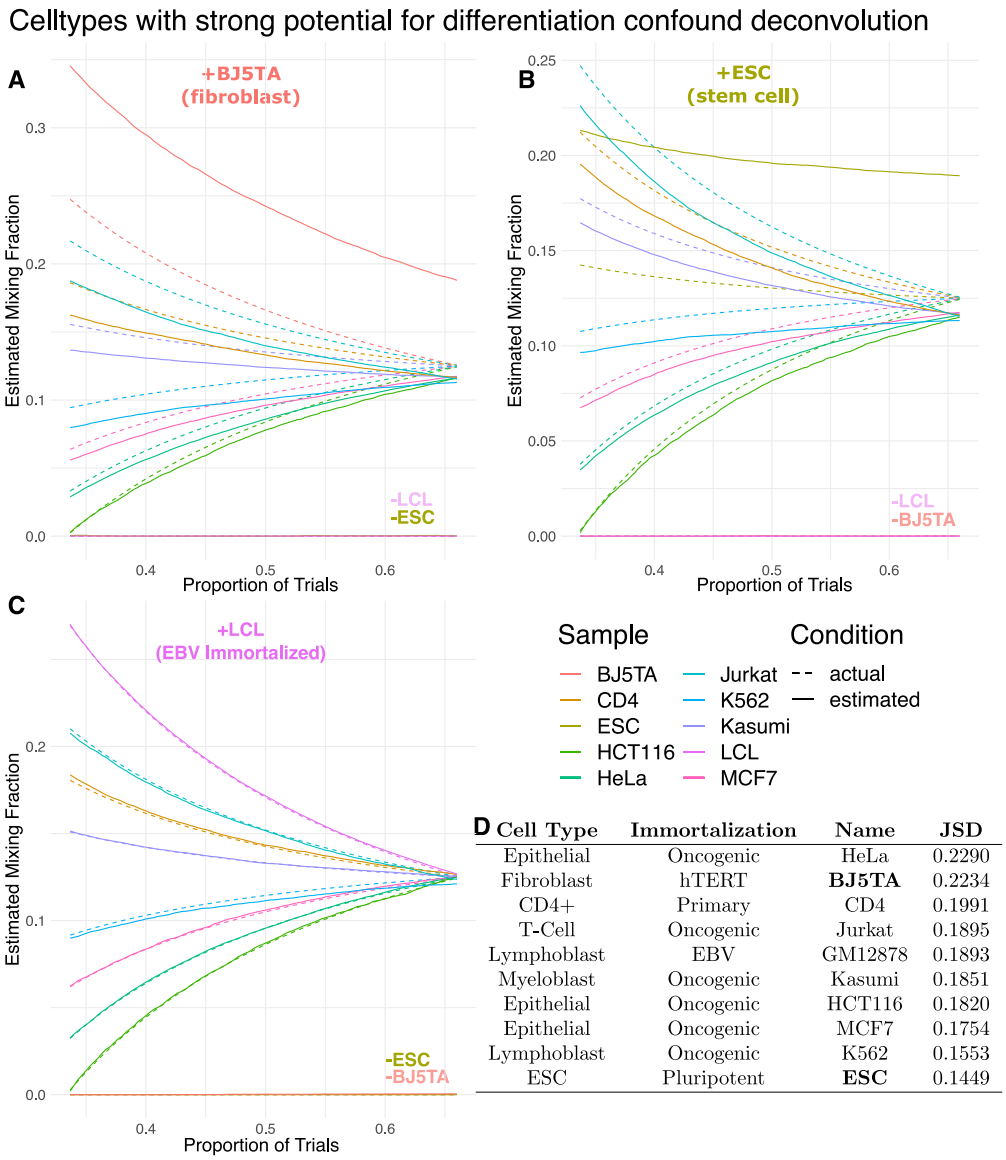
Figure 3.3: To compare the maximum theoretical accuracy of our system, we conducted the same analysis as in Figure 3.2 using either the region sets of bidirectionals, annotated genes, or a combination of the two, performing the same subsetting procedure as before. We observe that at smaller region sizes using genes alone provides a higher degree of accuracy than just bidirectionals, but that at larger sets of ROIs bidirectionals alone can achieve a higher absolute degree of accuracy. Somewhat unexpectedly, the combination of both sets of regions performs more poorly than each separate subset. Note that as system size increases, the accuracy of the set using annotated genes reaches a constant level purely because the total size of that system is exhausted by virtue of being an order of magnitude smaller than that of the bidirectionals or combined set.

examined deconvolution in the presence and absence of these cell types. To do deconvolution of this system, we generated a titration curve, mixing celltypes from distinct separate mixing proportions into equivalent proportions for all celltypes.

We observed that both ESC cell lines and BJ5TA cell lines caused deconvolution to fail (Figure 3.4A,B). Specifically, inclusion of either cell line results in an overestimation of the mixing proportion for those cell types. We carefully examined these two cell lines to identify distinguishing features relative to the other cell lines.

To determine whether the number of cell lines or cell line immortalization differences could be the source of the problem, we added lymphoblastoid cell lines immortalized (LCL) by EBV. Notably, LCLs do not confound the model and show excellent performance (Figure 3.4C). Both ESC (embryonic stem cells) and BJ5TA (fibroblast derived) are non-terminally differentiated and non-oncogenic (Figure 3.4D). Furthermore, we see that even without regularization, NNLS successfully removes non-present celltypes (Figure 3.4A-C), meaning that undifferentiated celltypes will not be inferred in the mixing proportion if they are not present at all in the mixture. Furthermore, regularization techniques are not required to accomplish this removal of celltypes that are absent.

One alternative hypothesis to the source of this problem is that heterogeneity in the population of undifferentiated celltypes is the source. However, this would suggest that more heterogeneous cell populations should perform worse in deconvolution, as should cells from similar tissue types. Yet based on this data, this seems unlikely, given that both CD4+ and Jurkat cells, both peripheral blood mononuclear cells (PBMC) derived, are present in the mixture and are successfully estimated by our models. Since the addition of a lymphoblast cell line immortalized using EBV (GM12878) does not result in system failure in the same way that is observed with the non-differentiated cell-lines, we suspect that differentiation is the key issue here as opposed to heterogeneity. Our work suggests that undifferentiated or partially differentiated cell types pose a key challenge to the deconvolution of nascent sequencing data when using enhancers because their regulatory profile, particularly that of their enhancer regions, resemble an ensemble profile of multiple differentiated celltypes. In support of this, the problem does not seem to occur when using genes alone, suggesting that undifferentiated cells may lack the same level of specificity



D

Cell Type	Immortalization	Name	JSD
Epithelial	Oncogenic	HeLa	0.2290
Fibroblast	hTERT	**BJ5TA**	0.2234
CD4+	Primary	CD4	0.1991
T-Cell	Oncogenic	Jurkat	0.1895
Lymphoblast	EBV	GM12878	0.1893
Myeloblast	Oncogenic	Kasumi	0.1851
Epithelial	Oncogenic	HCT116	0.1820
Epithelial	Oncogenic	MCF7	0.1754
Lymphoblast	Oncogenic	K562	0.1553
ESC	Pluripotent	**ESC**	0.1449

Figure 3.4: To interrogate the effect of undifferentiated and partially differentiated celltypes on the performance of deconvolution, we performed a titration experiment, estimating mixing parameters for 100 different mixtures of celltypes as mixing proportions were taken from maximally separated to equivalent. For each trial n, the mixing proportions are equally spaced points in $[n\frac{1}{n_{\text{tot}}}, 1]$ that are then rescaled to sum to one. Each subset (A,B,C) was generated by holding out one celltype from the full mixture and renormalizing the adjusted mixing proportions to sum to one. **A,B:** Adding either BJ5TA or ESC cells into the mixture causes a higher-than-true proportion of those cells to be estimated. Neither of these cell lines are terminally differentiated. **C:** Addition of EBV immortalized LCL cells into the mixture does not result in failure of the deconvolution model, suggesting that the observed failures are not a function of how cells were immortalized. **D:** To understand if this failure could be attributed to celltype specificity, we calculated the mean Jensen-Shannon Divergence for each sample compared to all others. The pluripotent ESC cells show the lowest celltype specificity while the partially differentiated BJ5TA cells show the highest celltype specificity, with the exception of HeLa cells.

at bidirectionals as terminally differentiated cell types.

Our results suggest either very low or very high celltype specificity when looking at these samples' bidirectional ROIs (Figure 3.4D). When looking at the mean Jensen Shannon Divergence for each sample compared to all others, we observe that our undifferentiated cell lines are either the least specific (ESC) or the most specific (BJ5TA). Although HeLa cells show the highest degree of celltype specificity by this measure, HeLa cells are not representative of human cells, exhibiting notably different expression patterns[128] which would lead to a high degree of cell type specificity. Past work has shown that ESC cell lines have genome-wide transcriptional hyperactivity[129] that narrows as differentiation progresses. Additionally, work in hematopoetic cells has suggested that these undifferentiated cell lines are characterized by a high degree of fluidity in chromatin modification[130]. More work is required to definitively establish that differentiation is the source of the breakdown of deconvolution in this system, and will likely require significant work outside the scope of this preliminary study.

3.3 Conclusion

This work is the first to examine supervised deconvolution of heterogenous mixtures of nascent sequencing data. Deconvolution is an essential tool for the study of heterogenous samples, whether cell lines or tissues. While most work on deconvolution of heterogenous samples has moved on to focusing on single cell protocols, a single cell nascent sequencing protocol currently seems infeasible. Thus, nascent sequencing is limited to bulk experiments, which appear to be reliably separable by supervised deconvolution. We present here the use of nascent sequencing data as a testbed for this supervised deconvolution problem. We integrate best practices from the literature and develop new techniques to handle characteristics in nascent sequencing data where assumptions from the RNA-seq deconvolution literature do not hold.

To benchmark various deconvolution algorithms, we first developed a new algorithm to filter ROIs to only use regions with the most celltype specific expression. We find that this selection process does not preferentially select enhancer or promoter ROIs. That said, the number of enhancer associated bidirectionals far exceeds annotated genes, providing ample features from which to select regions of

interest. Our proposed algorithm is simple, fast, and reliable, and establishes a strong first basis for the development of more specific ROI filtering tools for nascent deconvolution.

Using this algorithm, we compared standard methods used for solving the deconvolution problem. Specifically, we tested NNLS, Ridge, LASSO, as well as ϵ-SVR. We found that all methods reliably separate the nascent deconvolution system, with L2-regularized methods achieving comparatively poor performance to NNLS. Furthermore, we found that even a simple method like NNLS could reliably eliminate celltypes that were not present in the sample, suggesting regularization is not necessary for solving the deconvolution problem here. While we find that both annotated genes and bidirectionals can achieve high accuracy in supervised deconvolution (with bidirectionals having an edge in absolute accuracy), it is worth emphasizing that bidirectionals are distinctly advantageous in that they are annotation-independent and discovered **de-novo** for each sample.

We show that the addition of undifferentiated samples to a nascent deconvolution system results in highly skewed mixing estimates, with undifferentiated celltypes predicted as far more likely than their actual frequency in the mixture. One possible reason for this is that undifferentated celltypes tend to show regulatory patterns akin to a combination of the regulatory patterns of each constituent celltype. It appears to be a necessary condition for some amount of the undifferentiated celltype to be present in the mixture in order for the system to fail.

One key issue in this work is the lack of availability of diverse high quality nascent sequencing data to perform simulations against. Although a large amount of nascent sequencing data is available and published, the number of cell types available is somewhat limited. Protocols aimed at extending run-on sequencing to a broader base of samples, such as ChRO-seq[131] show promise in alleviating this bottleneck. Importantly, many of the earliest nascent data sets lacked replicates – which excluded their usage here. Data quality and availability is often a limiting factor in computational studies, and this work is not an exception to that rule.

In this work, and generally for the supervised deconvolution problem, we assume that all cells in a sample are taken from an approximately homogeneous population. This is sometimes a reasonable assumption but is often not. One future frontier that could be highly beneficial to this project is the

incorporation of single cell ATAC-seq (scATAC) as a secondary source of information to augment bulk nascent sequencing data. scATAC combines the chromatin accessibility readout provided by ATAC-seq (indicative of regions open to transcription) with the cell-specific information provided by modern single cell sequencing protocols. Tools are already well defined for clustering single cell sequencing data into constituent cell types, as individual cells can typically be separated using dimensionality reduction methods like PCA, tSNE, or UMAP[132, 133]. Because transcription occurs in regions of open chromatin, which is what ATAC-seq measures, mixing fractions and celltype specific transcripts could be estimated more reliably using combined data from both protocols. Future work combining pairing single cell ATAC-seq data and nascent sequencing data could leverage techniques used by existing tools[55] to do deconvolution on a more granular level for individual samples, providing a strong complementary tool to the bulk deconvolution discussed here. While single cell approaches remain comparatively expensive, this combination would be a powerful tool for looking at transcriptional regulatory networks at the level of sub populations of samples.

Nascent sequencing is a powerful tool for the assessment of transcriptional regulatory networks, and when paired with deconvolution tools will also facilitate deeper understanding of those regulatory networks in heterogeneous cell populations. Leveraging a transcription oriented sequencing approach instead of an expression oriented (e.g. steady state) one provides myriad benefits — more thorough coverage of the genome, understanding of regulatory elements, and a deep view of underlying transcriptional dynamics — all of which can be integrated with different sequencing protocols to great effect. Supervised deconvolution represents an important preliminary foothold into this space, and this work shows that nascent sequencing data is well suited for that class of problems.

Chapter 4

Encoder

All or portions of this chapter are currently in preparation for publication. Zachary Maas was responsible for the design of experiments, implementation of all project code, figure design, drafting and revision of the manuscript. Robin Dowell provided intellectual consultation, assistance on experimental design, and assistance during manuscript revision and during peer review.

Abstract Identifying cell-type specific transcription factors from sequencing data remains challenging, particularly when comparing multiple conditions. Existing deep learning tools for bulk sequencing can effectively find sequence characteristics that drive binding for a single transcription factor[134] or single cell type[135], but have not been extended to the case of comparing multiple conditions. To address this gap, we present here a novel unsupervised approach combining transformer-based autoencoders with sparse feature extraction to discover regulatory elements from nascent transcription data. Our method learns a representation of sequence and read data, then extracts interpretable features using a sparse autoencoder. Applied to nascent run-on sequencing from three cancer cell lines, this method can identify cell-type specific transcription factors with minimal input data, and finds known regulatory motifs inside of novel learned sequence context. This flexible framework can feasibly extend to various sequencing modalities and shows promise for multi-omic integration.

4.1 Introduction

The regulation of transcription plays a key role in the progression of both cellular identity through differentiation as well as in rapid response to environmental stressors and disease. A significant portion of transcriptional regulation occurs via the binding of transcription factors - proteins that bind to patterns (motifs) in the genomic sequence and modify the activity of RNA polymerase (up or downregulate transcription) at nearby sites. Active sites of transcription factor binding produce short unstable RNA transcripts referred to as enhancer RNA. By assaying the location of enhancer RNAs in run-on transcription

sequencing protocols, the collective patterns of eRNAs can be used to infer active transcription factor binding sites genome-wide[136]. Because enhancer RNA transcripts are produced at sites of transcription factor binding, these transcripts can subsequently be used to measure differential transcription factor activity for all identified transcription factors genome-wide in a single experiment[106].

While ChIP-seq has been instrumental in understanding regulatory processes, it is limited to measuring binding of a single protein at a time, requiring numerous experiments to comprehensively assess transcriptional regulation. Additionally, with a large enough corpus of high quality data, some of the regulatory elements that drive cell type can also be discovered[137]. In contrast, nascent run-on sequencing methods provide a more direct readout of transcriptional activity, capturing both genic and intergenic transcription simultaneously. This project aims to develop an exploratory tool using unsupervised learning approaches to learn likely markers of cell-type specific regulation that can be further investigated, with the ultimate goal of discovering potential targets by deciphering the genomic language of transcription. In the context of discovery of cell-type specific regulatory functional regulatory elements, the ideal data type is quantitative (meaning that differential signal is indicative of true differential effects in the underlying biology) and captures some sort of first-order regulatory effect (meaning that the used data type should capture an active biochemical process rather than a steady-state readout). Nascent run-on sequencing is well suited to this this case, providing a signal that is both “close" to signal (by virtue of assaying RNA directly engaged to RNA Polymerase II), and quantitative (meaning differential transcription likely reflects underlying transcriptional changes).

Recent work has tested the ability of neural network models to learn the sequence characteristics that drive the binding of various transcriptional regulatory elements. For example, BPNet[134] provides a model architecture that predicts sequencing reads from underlying sequence, using gradient based interpretability to find predictive sequence elements driving the binding of a single transcription factor. Curiously, this model approach is also capable of learning novel, previously unrecognized, flanking sequence characteristics at transcription factor binding site. Other recent work[135] has extended this approach to nascent transcription data using a single cell type (lymphoblastoids) sampled from a varied population of individuals, and in doing so was able to learn sequence characteristics and motifs that

underlie transcription initiation. These techniques are powerful and allow for supervised discovery of regulatory elements directly from sequencing data, providing a useful **de novo** method for discovery of both transcription factor binding sites (motif hits) as well as surrounding sequence context. Importantly, these methods learn important sequences without prior knowledge of recognition motifs – rather the known motifs are used to interpret the patterns learned. While BPNet derived approaches are flexible and powerful, they are typically only useful for the discovery of motifs for a single TF (in the case of BPNet[134] which leverages ChIP) or variation in transcription initiation associated with SNPs in a single cell type (in the case of [135]). Alone, this approach is useful but has a key limitation — it is not well suited to understanding the sequence elements that drive the differences between samples (like those in experiments with a perturbation or between distinct cell types).

Parallel to this work in discovery of sequence motifs [134, 138], a significant amount of progress has been made in improving interpretability of complex models driven by results from decoder-based large language models[139, 140]. It has been shown that sufficiently complex neural networks (in particular, transformer based language models) learn to represent features nearly-orthogonally as a high dimension randomized projection[141]. This phenomena — referred to as polysemanticity — refers to the ability of a single model neuron to encode multiple input features in superposition rather than just one. Recent work in interpretability has suggested that these polysemantic representations can be reduced to monosemantic (encoding a single feature) representations by training a sparse (L1 penalized) autoencoder on the latent representations learned by a model[139], showing promise in improving our ability to interpret large language models. Then, if we view sequencing data and the backing genomic sequence as encoding a genomic language, training a language-model style architecture and using interpretability tools may let us learn important biologically relevant features beyond those that drive the general biology (e.g. general transcription initiation) of whatever protocol is being studied. This approach, while theoretically distinct from gradient based approaches, is a powerful tool in learning interpretable features from complex data directly.

In this work, we attempt to bridge the gap between current models and differential data by proposing a flexible framework for interpretable discovery of the sequence elements driving differential transcrip-

tion. Using transcriptional run-on sequencing data, we develop a transformer based autoencoder model which learns the genomic language describing coupled transcription and sequence. Using this model, we are able to extract monosemantic features from the trained latents of the transformer using a sparse autoencoder and associate these features with both specific conditions as well as sets of differentially enriched sequence motifs within those conditions. Using the features learned by our sparse autoencoder, we apply per-neuron gradient attribution methods and uncover interesting learned patterns of sequence periodicity describing CpG islands. We find that our model reliably finds enriched transcription factors associated with cellular identity, using a comparatively minimal amount of data compared to other approaches[137].

4.2 Methods

We selected the use of nascent run-on sequencing data for our model after carefully evaluating other possible data types and considering the limitations of each (Appendix C), finding that run-on based nascent sequencing protocols served this purpose best. In addition to the selection of data type, we also discuss the variety of preliminary designs tested in this project prior to selection of the model discussed here.

Briefly, in order to learn sequence features that characterize cell type, we use a combination of a transformer-based autoencoder model to reconstruct input sequencing data (genomic sequence paired with the normalized histogram of reads over that region), and a sparse autoencoder (SAE) to extract interpretable features from the autoencoder model's latent representation of the data (Figure 4.1A-B). Using this SAE model's results, we then extract sequences that are highly activating for each neuron, search for transcription factor binding motifs that are enriched within that binding set (Figure 4.1C-D). Using this set of per-neuron enriched motifs, we find neurons that are overrepresented for regions from a specific cell type, and say that enriched motifs in over-represented neurons that do not appear enriched in other cell types are cell-type specific (Figure 4.2A-B). We selected this model architecture after extensive testing and attempts to extend both existing methodologies[135] and contrastive methods to our differential context, for which a more detailed description is available in the supplemental material (Appendix C).

More details on each step of the process are below.

4.2.1 Encoding Transcription

For this work, we used previously published data[61, 119] in 3 distinct cell types — HCT116, MCF7, SJSA — from distinct tissues of origin — colon cancer, breast cancer, and osteosarcoma (Appendix C). Using a consensus set of of transcribed regulatory elements (TRE) for all cell types used, we take a fixed window of ±128bp around the center of each region of interest (ROI). The approach used to generate these ROIs and identify their centers has been shown to be highly accurate[143] For each region (n=96,953 ROIs identifed in previous work[61]), this data is of dimensionality $[256x6]$, using one-hot encoded genomic sequence and the normalized histogram of sequencing depth for both strands of DNA.

Our model (Figure 4.1) processes input data represented as 6-dimensional vectors: 4 dimensions for one-hot encoded DNA sequence, and 2 dimensions for positive and negative strand read counts — reflecting bidirectional transcription (Figure 4.1A). Each input sequence has a length of 256 base pairs, chosen to capture local genomic context while maintaining computational efficiency.

Our model architecture starts with an input embedding layer that projects the 6-dimensional input into a higher-dimensional space, followed by sinusoidal positional encoding[144] to allow the subsequent transformer to understand the sequential nature of the data. The transformer encoder uses of 6 layers with 2 attention heads each, a configuration empirically determined to balance model capacity with computational constraints. This encoder learns to capture complex patterns and dependencies in the input data in a latent space.

For reconstruction, we employ a small two-layer multi-layer perceptron (MLP) decoder. This asymmetric architecture, with a deep encoder and shallow decoder, places the burden of learning on the encoder, ensuring that the latent space captures the most salient features of the data. The decoder's role is primarily to map this latent representation back to the original input space.

We intentionally use a small layer size and hidden dimension throughout the model. This design choice ensures that the latent representation is a compression of the input data, forcing the model to learn an efficient latent encoding. Through empirical testing, we found that this compact architecture

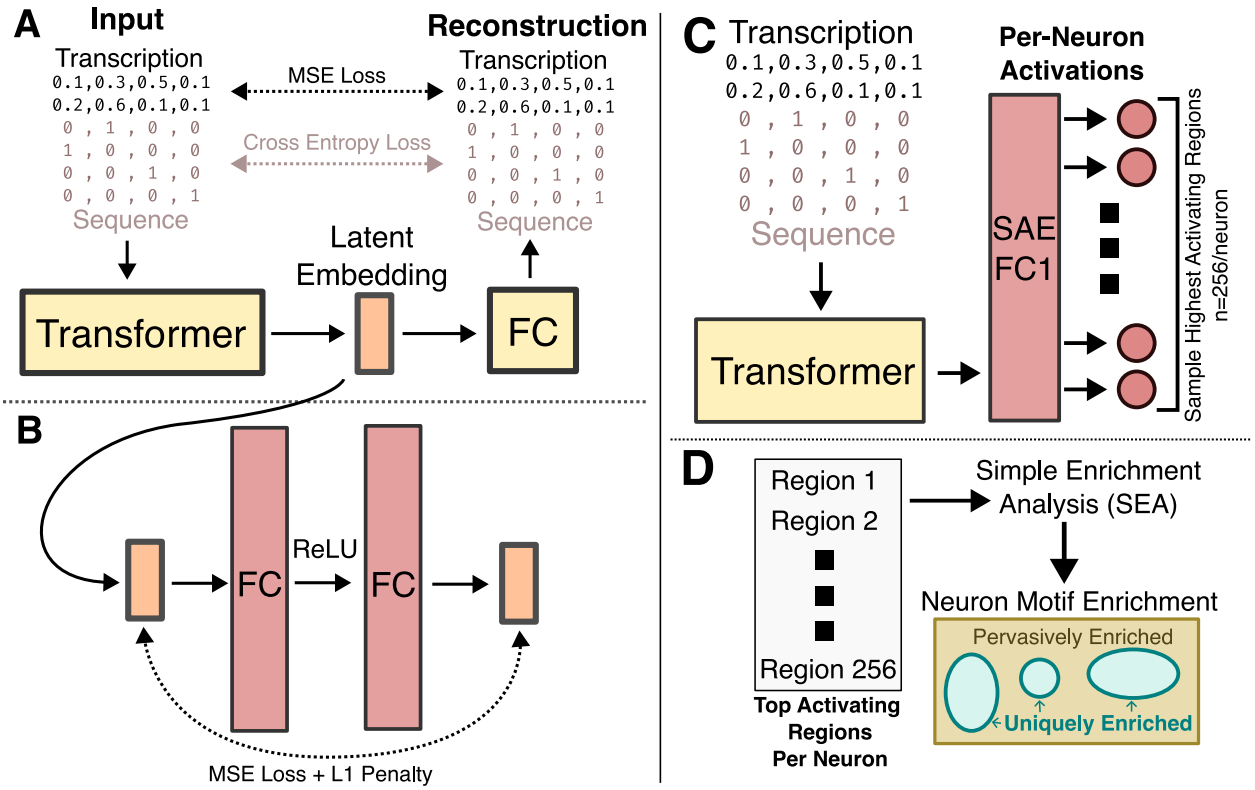


Figure 4.1: **A:** Using coupled transcription and read data over sets of bidirectional regions, we train a transformer based autoencoder on the task of reconstruction. In doing so, we learn a latent embedding that can be used as the input of a secondary model to attempt to extract monosemantic features from that embedding. **B:** To extract monosemantic features, we follow the framework established by[139], using a single layer fully connected autoencoder with ReLU activation after the encoder layer, with random reinitialization of dead neurons performed at each 10% step of training. **C:** Using the SAE model trained in **??**B, we extract the top activating regions for each neuron in both the positive and negative direction, which provides a set of sequences expected to be descriptive of a monosemantic feature. **D:** Separately for the top and bottom activating set of regions for each SAE neuron, we extract the underlying sequences and perform motif enrichment analysis per-neuron using MEME SEA[142]. Across all discovered neuron-specific motifs, we remove motifs that show global enrichment across all neurons, as short and GC rich TFs are enriched across most neurons since our input regions are known to be transcriptionally active.

achieved similar convergence to larger models tested models of this same architecture. We find that this architecture can robustly recreate transcriptional data across multiple cell types (n=3).

4.2.2 Automated Interpretability using SAEs

To extract features from our base model, we employed a L1 regularized Sparse Autoencoder (SAE) trained to reconstruct the latent representation learned by the encoder of our base model. Our SAE consists of a simple linear encoder/decoder architecture, with Rectified Linear Unit (ReLU) activation on the encoder layer[139] (Figure 4.1B).

The SAE is designed to learn a sparse representation of the base model's latent space. By imposing an L1 penalty during training, we encourage the SAE to activate only a small subset of its neurons for each input. This sparsity constraint forces the model to learn a more distributed and disentangled representation, where each neuron ideally corresponds to a single, interpretable feature of the data.

The principle behind this approach is that the sparsity constraint penalizes single neurons that encode multiple features simultaneously[141]. Instead, the model is incentivized to allocate distinct neurons to represent different aspects of the input data. This process of disentanglement produces in neurons that are more likely to be monosemantic — representing a single, coherent feature of the transcriptional landscape. By applying this two-step process of first compressing the input data with the base autoencoder's encoder, and then disentangling this representation with the SAE, we aim to extract interpretable, biologically relevant features from the data.

Throughout training, some neurons periodically 'die', failing to activate for nearly all training data and losing their representational capability[139]. To maintain effectiveness of our SAE's learning, we employ periodic reinitialization to prevent the 'dying neuron' problem. Every 10% of training steps, we evaluate the activation patterns of neurons across a random 25% subset of samples. Neurons that fail to activate for more than a minimal threshold (1e-5) of these samples are considered 'dead'. These dead neurons (Figure 4.2D) are then randomly re-initialized to normalized values derived from the input, scaled to match the average norm of currently active neurons. This reinitialization prevents the loss of representational capacity in the SAE by dead neurons and instead puts those neurons in a state where

they can potentially capture additional features in the data.

Using the features learned by the SAE, we then extracted the top (n=256) most and least activating samples for each neuron (hereafter referred to as the activation sets), verifying that each neuron captured a distinct set of input samples (Figure 4.1C). Minimal overlap of top activating regions across SAE neurons was observed, with only 0.7% of input samples repeated across all top activating samples across every neuron. This activation set provides a sort of subset of a clustering of the original data, sampling only the samples that are most strongly associated with each feature learned by our SAE.

4.2.3 Motif Discovery from Sparse Neurons

We proceeded to use this collected set of samples to discover cell-type specific transcriptional regulatory elements on a per-cell-type basis. Using the underlying genomic sequence from each activation set, we performed motif enrichment analysis using SEA and motifs from HOCOMOCOv11. To determine cell type specificity of motifs, we first searched for activation sets that were enriched for samples from a single cell-type relative to random sampling (Figure 4.2B). Then, for all enriched motifs discovered across all enrichment sets, we filtered out those motifs that were enriched globally across every cluster (Figure 4.2C,E). This second filtering step is necessary to remove motifs that will show up broadly across any set of sampled transcriptionally active regions. Using this combination of a model to learn the interplay between sequence and transcription, a sparse interpretability model to reduce the complexity of the encoder in the first model, and classical sequence motif discovery techniques on the activation sets of our interpretability model, we are thus able to learn cell-type specific sequence regulatory elements in a **de-novo** unsupervised context.

4.2.4 Discovery of Potentially Novel Motif Syntax Using Interpretability

To further leverage the features discovered by our SAE, we next applied per-neuron GradientSHAP[145] to each neuron in the SAE encoder in order to extract positional attributions (Figure 4.3A). Curiously, the typical pattern of the mean absolute attributions for each cluster often shows a pattern of periodicity on the order of 2-4 base pairs. Individual sequences, however, show an expected pattern in attributions,

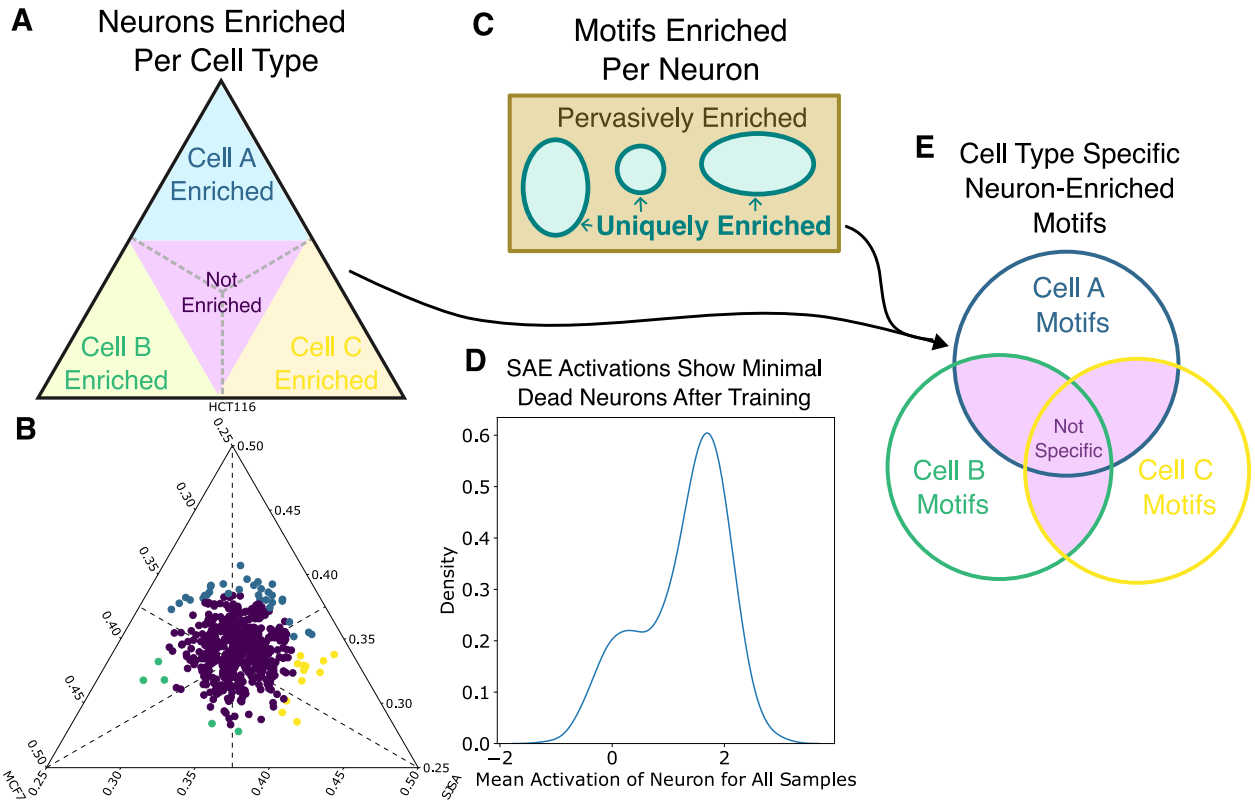


Figure 4.2: **A:** Using the set of positively activating regions for each neuron, we calculate whether each neuron is preferentially enriched for regions sampled from a specific cell type of interest. With these sets, we then extract discovered TFs from Figure 4.1D that are only found in neurons enriched for a specific cell type. This enrichment is performed using a χ-squared test to determine if top-activating regions for a given cell type are sampled within expectation given the number of input regions for that cell type — those that are not are assigned to a specific cell type. **B:** Performing this enrichment discovers a small number of cell-type enriched clusters, which we use to examine cell-type specific elements in downstream analysis. Note here that this plot is corrected after testing for the background frequencies of each cell type to center the visualization on the plot axis — without correction the plot skews towards cell types with higher number of input regions, but our testing approach compensates against this. **C:** We validate neuron-specific features using two distinct approaches. First, we use a traditional approach, taking the top-activating regions for each neuron and examining sequence enrichment using SEA to use information in the top-sequences only. Second, we use DeepSHAP and TF-MODISCO on a per-neuron basis to interrogate the defining per-neuron sequence characteristics predicted by our models, to use information provided by the model itself. **D:** After training with dead neuron re-initialization, our final models have a small cluster of 2-3 neurons after training that fail to maintain an "alive" state (when n=512). **E:** Using the enrichment results per-neuron and previous discovery of cell-type specific neurons, we look at the set of all discovered motifs for each cell type and those not associated with any specific cell type and generate a final list of high confidence cell-type regulators as the set of motifs associated with a cell type that do not occur in any other examined cell type. This is a strict method, as some cell type specific TFs may be shared between multiple TFs.

with a small number of base locations contributing to most of the attributions for the sequence. Using these positional activations, we applied TF-MODISCO[146] (tf-modisco-lite version 1.0.0) to each SAE neuron's attributions to determine the set of defining motifs for each feature. Briefly, TF-MODISCO uses a hierarchical clustering approach to identify recurring patterns in attributions of genomic sequence, extracting both general patterns and sub-clusters of variants of those patterns. Upon initial testing, we observed that TF-MODISCO was only able to identify short sequences consisting of high attribution single nucleotides centered around low attribution positions. To combat this, we applied a positionally aware smoothing transform to our attributions (see Methods) before input in order to reduce the impact of the most highly attributed input positions.

Across all SAE neurons, we observe consistent enrichment of SP/KLF as well as ZNF family transcription factors, which are GC rich motifs. We do not expect that these are real features, however, owing to the sequence bias in our input data. Because sites of transcription initiation are inherently GC rich, and because our model architecture does not explicitly account for the sequence bias of the input regions, we would expect most short, GC rich motifs to be found in most features regardless of actual biological relevance (Figure 4.3B). To combat this, we set the background sequence composition for TOMTOM in TF-MODISCO's downstream analysis to the average GC content of all input PSSMs.

To apply this attribution based approach to examine cell type specificity, we looked at our previously identified subset of cell type specific neurons (Figure 4.2). Simply, we take all TF-MODISCO identified PSSMs within a cell-type specific cluster and run TOMTOM to identify significant motifs in that cluster (discussed more extensively in Appendix C). We repeat this for the aggregate of all neurons outside of any cell-type specific cluster, hereafter referred to as the general transcription cluster. With these cell-type specific sub-sets identified, we further filter for uniqueness by removing TOMTOM matches from each cell type cluster that are identified in any other cell type cluster. Similarly, to determine generally transcription associated regulatory elements, we remove uniquely identified cell-type specific TOMTOM matches from the general transcription cluster. Compared to the approach discussed above looking directly at top-k sequences, this methodology discovers a wider set of potential cell-type associated motifs (Table C.2), and is founded in the actual learned characteristics of the model.

While this set almost certainly contains a wide set of false positives, a significant number of the identified cell-type specific targets have been previously validated in the literature. The full table of TFs identified for this analysis are presented in Table C.2 For HCT116 / colon cancer cells, MYC[147], NFKB1[148], ESR1[149], TFL7L2[150], GATA4[151], RUNX2[152], FOXC1[153], SOX17[154], and JUND[155] have literature confirmation. Similarly, MCF7 / breast cancer cells identify TFs for GCR[156, 157], HSF1[158, 159], SOX9[160], SMAD4[161–163], IRF7[164, 165], RELB[166], and HSF2[167]. Finally, SJSA shows some osteosarcoma associated TF motifs — BCL6[168] and TEAD1[169]. We also see a number of TFs that are likely false positives, likely arising from the limited number of cell types utilized here (3). For example, this approach identifies the general transcription factor TBP as specific to HCT116, which by virtue of being a general transcription factor is certainly not cell type specific. This could be a result either of the small number of cell types used or of the model capturing relative transcription levels of samples and attributing general transcription factor TFs to the highest depth sample. Improved filtering could improve the inference of cell-type specific features over those that learn sequence background.

In addition to identifying the sequence of identified motifs, our model's TF-MODISCO enriched sequence patterns show novel flanking patterns around many identified motifs, often as large as the recognized motif itself. We find useful and validate associations in SAE feature neurons, but the combination of TF-MODISCO and TOMTOM struggle with the repetitive patterns of attributions that we observe in our neurons (Figure 4.3C). Given the limitations discussed here, TF-MODISCO is not an obviously appropriate tool to apply to this model context, but is nonetheless a valuable first step towards achieving direct interpretability of our model.

4.2.5 Understanding SAE Neuron Characteristics

When compared to their application in language, interpretation of SAE features is less obvious. To interpret these features in the context of natural language, it is relatively easy to examine the most activating tokens within a corpus and use those to associate a SAE neuron with some feature of language (for example, words associated with mathematics or positivity). However, when applied to genetic sequence as we have here, this interpretation is less immediately obvious. Because it is no guarantee that our model

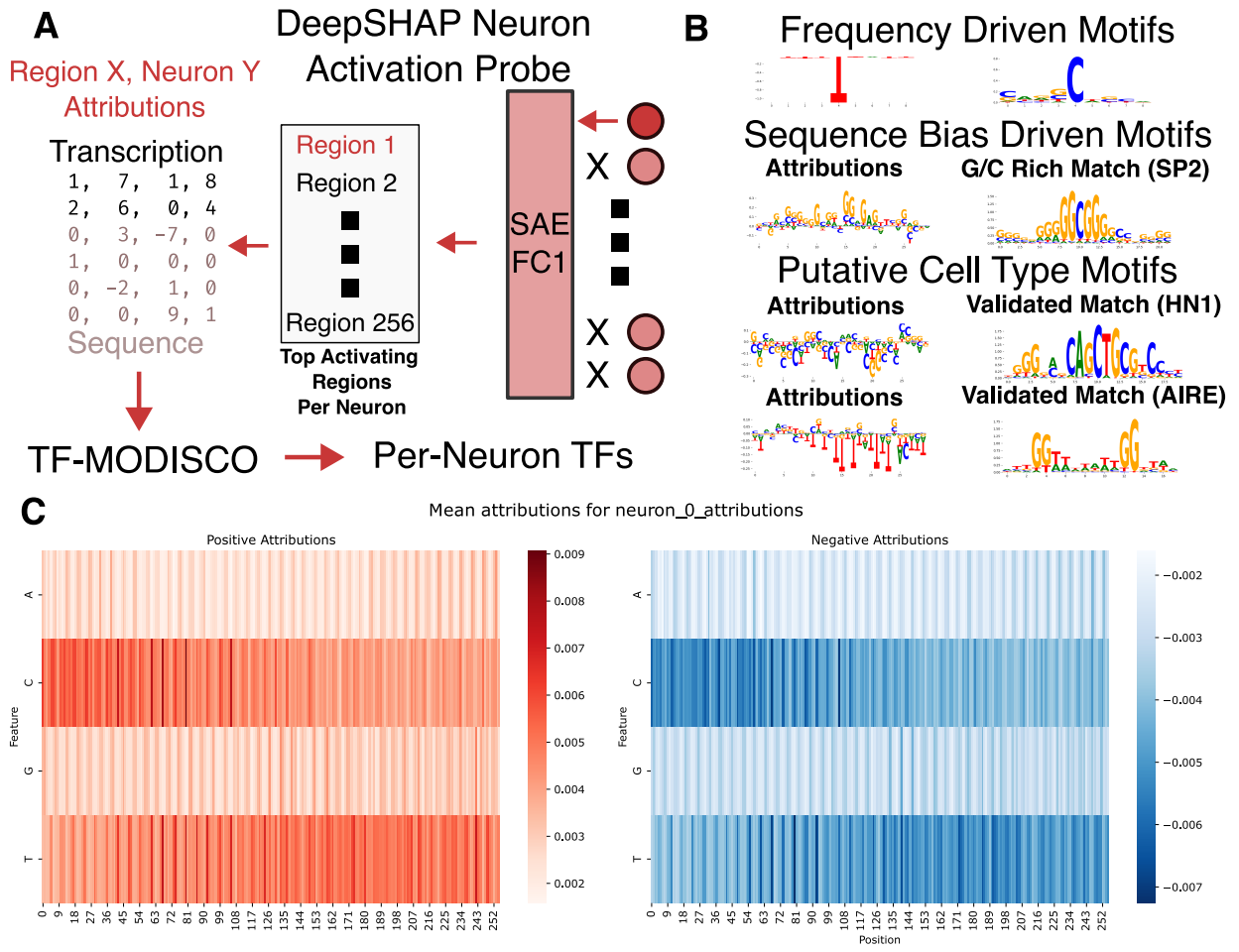


Figure 4.3: **A:** Using the trained SAE model, we extract per-neuron positional attributions for input data using neuron-specific DeepSHAP, which we then use as input for TF-MODSICO to perform discovery of **de-novo** motifs. Given the strong periodicity seen in our model (see panel C), we perform a local smoothing procedure before input into TF-MODISCO to nonlinearly reduce the outsized attribution contribution of this periodicity. **B:** This interpretability approach shows initial promise, but has limitations given the set of features learned by our model. Many TF-MODISCO discovered motifs embed the periodicity of our model (top), requiring correction. When using locally smoothed data, we observe two distinct categories of motifs identified by TF-MODISCO. First, nearly all neurons are ubiquitously enriched for SP/KLF class motifs, which resemble the sequence background of the transcriptionally active regions used as input for the model (middle). Outside of these, TF-MODISCO does identify a set of motifs in cell-type specific neurons that appear to match cell-type specific TFs with additional flanking sequence context. Compensating against the learned periodicity of small models such as the one used here is necessary for more robust TF-MODISCO analysis. **C:** Average DeepSHAP attributions split by greater than or less than zero for a sample SAE neuron. SAE neurons encode periodic patterns associated with specific base pairs, and typicallly learn a representation that is biased towards T and C base pairs. This particular neuron appears to encode for some sort of symmetric characteristic in activations around the center of the bidirectional peak (note the increased intensity on the left in T and right in C). Other neurons show various levels of symmetry or positional preference at certain base pairs.

learned TF binding sequence motifs alone, and given the periodicity seen in the average attributions for a neuron (Figure 4.3C), we further sought to investigate whether certain SAE neurons additionally encoded distinct sequence-specific characteristics beyond TF binding motifs. To investigate what other sequence-specific characteristics were learned, we sought to examine the frequency of trinucleotide and pentanucleotide repetition as well as CpG island frequency, using each SAE's aggregate attributions. In order to quantify this, we used the background sequence composition of each input to calculate the random expectation with which a given trinucleotide would be seen, testing whether any 3 or 5-mer occurred more frequently than this expectation within the neuron's attribution predictions. Here, we used the most likely base pair at each position based on GradientSHAP attributions. With this approach, all SAE neurons appear to have attributions whose repetitive pattern is enriched for CAG and CTG repeats, which are canonically associated with enhancer and promoter activity[170].

An alternative explanation for the repetitive sequence pattern, perhaps, is that these models may be learning some gapped k-mer representation in addition to the other learned sequence features. This learned periodicity is not surprising, as interpretability work in language modeling has shown that transformers frequently learn periodic features[171], which appears to extend to the context of transcription associated sequence features here.

4.2.6 Comparison to Existing Work

It bears drawing comparison of this work to similar recent work in inference of cell type associated transcription factor binding patterns[137]. Compared to previous work, we see disconcordant results in cell-type specific transcription factors identification, using largely different sets of data. Whereas in this project we suggest a model inferring cell type specificity from a minimal amount of data (2-3 replicates per cell type across 3 cell types), compared to 287 experiments across 20 different cell types [137]. Of the mismatches with previosuly published work, none were attributed to another to a cell type (colon, breast, or bone cancer). Instead, these differently attributed TFs were instead assigned[137] to developmentally distinct cell types – typically blood. This problem would likely disappear with use of a larger number of cell types.

4.3 Results and Discussion

We find that our modeling and interpretability process is able to extract clusters of features that are enriched for cell-type specific regulatory elements. In testing, the neurons in our sparse autoencoder showed minimal overlap on a per-cluster basis (0.7% of regions shared across any two neurons in Dataset A). This suggests that our SAE is performing effective feature extraction.

After controlling for transcription factors that are globally enriched across all clusters (ZNF, short motifs, GC rich motifs), we find a small subset of explainable features attributable to specific cell types using the most highly activating regions per-neuron as input for MEME (section 4.3). For example, we see HCT116 (colon cancer) cells are enriched for the TFs PDX1[172] and RELB[173], which are associated with coleorectal cancers, and negatively enriched for HBX4[174] which is associated with worsening of coleorectal cancers in the presence of specific environmental factors. Curiously, in the discovered consensus set, MCF7 (breast cancer) cells showed no uniquely enriched postive activating TFs, which is likely a result of the low number of neurons discovered for that cell type (Figure 4.2). The reason for this small number of discovered neurons is not obvious, but may be related to the relative depth of each sample, discussed more below. Finally, SJSA (bone cancer) cells are enriched for PBX1 (associated with bone development[175]), ISL1, MYB (a known prognostic marker for osteosarcoma[176], NKX32 (associated with bone development across organisms[177], and MBD2 (involved in response to load in osteocytes[178]. Compared to the other cell lines used in this study, the SJSA cell lines had lower sequencing depth, which may drive the difference in learned quantity of cell type specific features, if the model is primarily capturing relative differences in transcriptional level overall.

Together, these results indicate that our model is indeed learning cell-type specific regulatory factors characteristic of cell type, although these features appear to be more defining of the oncogenic state of the cell lines used. These results differ from the motifs discovered using TF-MODISCO, which is most likely attributable to TF-MODISCO not being well suited to the repetitive patterns learned by our models as discussed here and in the supplement of this work (Appendix C). Even with the local smoothing approach we have leveraged, the predictions of TF-MODISCO appear to be primarily driven by sequence

Cell Type	MEME Enriched Transcription Factors
HCT116	PDX1[172], RELB[173]
MCF7	N/A
SJSA	ZN329, PBX1[175], ZN490, ISL1, MYB[176], NKX32[177], MBD2[178]

Table 4.1: Table of differentially enriched transcription factors for Dataset A, using the stabilized set found using top-k sampling of enriched TFs. Factors for each cluster are associated with their cell type of origin and/or the oncogenic state of that cell line.

bias and repetitive patterns and disentangling underlying motifs from these is challenging.

To our knowledge, this work is the first attempt to use a deep learning approach to perform fully unsupervised discovery of cell type specific markers from transcription factor associated enhancer RNA transcription. While we have used this technique here to the end of unsupervised discovery of clusters of sequence regulatory elements that drive a given condition, the design of this method is such that it could also be flexibly applied to other protocols to discover similar clusters in other types of data. For example, using ATAC-seq or RNA-seq, this method could be extended for discovery of important characterizing elements in chromatin or steady-state RNA, respectively, in combination with transcriptional data.

We also apply well-established interpretablity techniques to this novel domain and model architecture, and uncover novel limitations of those models, as well as potential novel flanking sequence syntax associated with cell type specific transcription factors. While TF-MODISCO is a powerful technique, this approach suggests that it (or a similar approach) would benefit from the ability to normalize more robustly against sequence bias. In particular, it is unlikely that the observed ubiquitious SP/KLF/ZNF motifs discovered at transcribed regulatory elements in this work and others[135] are real, but rather a consequence of models learning that these regions are GC rich. The same is true in reverse, where some neurons show seemingly artificial enrichment of AT rich motifs in learned features that show AT bias. Notably, we also extended the full model interpretability technique of TF-MODISCO to a novel mechanistic interpretability, leveraging it to explain the characteristics of interpretable features learned by a sparse autoencoder. Adapting these learned feature neurons to work with TF-MODISCO required

the development of a local, context aware smoothing approach (described in Appendix C) to nonlinearly reduce the magnitude of large attributions relative to smaller attributions. In doing so, we extend the toolkit of interpretability for genetic deep learning by demonstrating that SAEs are effective feature learners that can leverage TF-MODISCO for explainability. Given the inherent periodicity observed in our learned SAE neurons, additional improvements on the TF-MODISCO technique are required to extend this methodology to function well in this context.

There are a number of unaddressed questions in generalization that should be explored by further work. First and foremost, our methodology has only been tested on a small number of conditions ($n = 2-3$), and while it performs well in this case, this approach may not generalize well to a significant number of conditions. For example, when presented with samples from a significant number of conditions or cell types, the small encoder in our base model may lack sufficient expressivity to fully encode the transcriptional language of such a variety of distinct regulatory regimes. Another point of concern for applying this method more broadly (at least in the context of nascent transcription) is that in Dataset A we appear to discover not only cell-type specific regulatory elements but also transcription factors that drive the oncogenic state of these immortalized cell lines. While this bodes positively for our method's ability to discover the elements driving differences between cells, it does complicate the ability to discover "true" cell type specific markers that are not downstream of cell type origin. With sufficient data from distinct primary cell lines, this could be probed further, but is currently challenging given data availability. In addition to the limitations discussed above, a key gap that future work could address is the integration of regulatory networks into the model's description or learning, using a graph neural network approach. While our method performs well at discovering transcription factors associated with specific cell types, it is unaware of the spatial relation between regions of interest in the input and thus is unaware of key interactions and correlations that could further inform the learned embedding.

The model architecture discussed here leverages the ability of attention-based models to simultaneously attend to all positions in the input data. Furthermore, it allows us to use the aforementioned SAE technique to extract interpretable features. However, the use of SAEs in this context for interpretability still requires an amount of caution and significant trust in this combined model architecture to extract

monosemantic features, particularly with the observed failures of explainability tools like TF-MODISCO to adapt to our model context. Nonetheless, because our methodology is supplemented by more traditional tools (STREME), even a failure of a given SAE neuron to represent is still likely to provide useful information on the regulatory identity of the transcriptional elements that it is associated with.

Scaling Issues

Further validation of the method here with this alternative dataset is an necessary next step, but has two principal risks that must be balanced against. The first challenge in using substantially more data is that a larger model is likely necessary both for the encoder model as well as the SAE. By significantly scaling the amount of input features to almost 10× the number of cell types with substantially more replication, a more expressive model will almost certainly be necessary to capture the transcriptional variation between cell types and not just general transcriptional patterns. The model described here (and other similar work) primarily learn sequence biased motifs – KLF/SP class motifs which are GC-rich, for example[135]. The interpretability approach pioneered here makes some progress on resolving that task by allowing for semantically distinct elements to be extracted from those elements that are ubiquitously identified, but requires further development to link individiual SAE neurons to function.

The second, and more fundamental challenge, is that significantly scaling up the number of cell-types will break assumptions required for the downstream analysis done here. While training a larger encoder model and SAE is technically straightforward (and has been verified in preliminary testing), extraction of cell type specificity will require a new approach. Our current method determines cell-type specific SAE neurons by fitting a normal distribution to SAE encoder activations for each neuron then drawing all regions that activate at above 3σ from the mean activation for that neuron. This set of top-activating regions is used for motif discovery in our activation probe, and we call a region cell-type specific if it is preferentially enriched for regions from one cell type as the null of a χ-squared test. Say, hypothetically, that we increase our number of classes from 3 to 30 but maintain our 3σ threshold for top-activating regions. For each neuron, we then have the same 800-1000 top-activating regions per neuron, but instead of being on the order of 300 regions per class (top-activating / number of cell types), we now have on the order of 30 regions per class for 29 degrees of freedom, which is infeasible for a

χ-squared test. The statistical problem that emerges is thus this — it is not clear how to balance specificity of motif discovery with specificity of cell type enrichment. We can increase our number of top-activating regions using a smaller threshold on the normal distribution (say, 2σ), giving us an order of magnitude more sampled regions. However, this reduces the specificity of the regions to that neuron substantially and risks learning general transcriptional characteristics rather than cell-type specific ones. If we instead keep the top-activating threshold constant, our χ-squared test instead becomes infeasible.

Alternative approaches are potentially feasible here, as a departure from this approach. For example, since the number of SAE neurons is relatively large, a "meta-neuron" clustering approach might be feasible. By grouping SAE neurons into aggregate "meta-neurons" on the basis of similarity in activation or attribution patterns, we could maintain the current statistical testing approach while using a smaller set of input regions. In a complementary manner, we could take a hierarchical clustering approach in neuron-to-cell assignment, first identifying neurons that share lineage-specific enrichment and then extracting constituent cell type information from there.

Nonetheless, scaling this approach is necessary and given the rich expressive capability of transformer based models, this technique will likely scale with appropriate modifications to the underlying approach.

4.4 Conclusion

We present here a new modeling approach aimed at the identification of cell-type specific regulatory elements from nascent transcriptional sequencing data. To learn cell-type specificity, instead of training a model to predict reads from underlying sequence (the predominant approach in the field currently[134, 135]), we train a small autoencoder to jointly predict sequence and reads from transcriptional data together. We present here a this novel unsupervised approach combining transformer-based autoencoders with sparse feature extraction to discover cell-type specific regulatory elements. Our method learns a representation of coupled sequence and sequencing data, then extracts interpretable features using a sparse autoencoder. Applied to nascent run-on sequencing from three cancer cell lines, we identify cell-type specific transcription factors with minimal input data, and finds known regulatory motifs inside

of novel learned sequence context. Using gradient based attribution, we perform motif discovery and observe that periodicity in learned model representations confounds the accuracy of TF-MODISCO, requiring local smoothing. Using attribution, in addition to cell-type associated motifs, we also observe that our model learns repetitive sequence context across our input regions associated with CpG islands. This flexible framework can feasibly extend to various sequencing modalities and shows promise for multi-omic integration, showing the potential of improved interpretability approaches for transcriptional sequencing data.

Chapter 5

Conclusion

The work discussed in this thesis has sought to advance our abilities around a single key question — how can we use nascent sequencing data to more effectively extract biologically relevant information from experiments. Even at the start of my graduate work, a key issue faced by our research community was the sheer scale of data being produced, and the question of whether or not we were using it effectively. Through this work, I have advanced the state of the art in methods that allow us to extract additional information from transcriptional sequencing data in both supervised and unsupervised manners.

5.1 Re-normalizing Data

The earliest work in this pursuit focused on the question of normalization chapter 2. In some sequencing protocols (like RNA-seq), we have the ability to easily add an external normalization control during an experiment. External controls (spike-ins) lets us control for the various experimental effects that may change a sample but that are not related to the underlying biology being studied. Naturally, this normalization is important. Without it, we cannot reliably make biological conclusions from our data. Unfortunately, in many experimental protocols, external normalization controls are less straightforward. For example, the nascent transcriptional protocols discussed here can use an external normalization, but these controls are typically miniature copies of the same protocol performed using cells from a different organism, rather than being a known fixed quantity to normalize to. While this is not ideal, it is the best method available now given limitations in the biology and the physics of the chemical reactions involved in these protocols. To complicate this situation further, many researchers do not add external

controls to transcriptional experiments, and when they do, they often under-sequence them, reducing their usefulness.

To combat these difficulties in normalization, we proposed a two-step solution to estimate normalization parameters both with and without external controls in transcriptional experiments[60]. Our first key insight was a re-framing of normalization parameters from a fixed value to a random variable, in a way that is applicable to exogenous spike-ins and the endogenous approach discussed here. Typically, normalization parameters are viewed as a fixed value per-sample, which accounts for relative efficiency of experiments compared to each other. For example, after selecting one experiment to be a reference sample, one sample may have 80% of the relative efficiency of the other, yielding a normalization factor of 0.8. Our approach instead views these normalization parameters as random variables, for which we can estimate both the value and the variability of that estimate. Using a hierarchical model, we are able to estimate normalization factors leveraging the expected underlying distributions of our sequencing data, and ultimately find a set of tighter normalization parameters that establishes more cautious bounds on differential analysis compared to previous methodologies. Using this approach, via Bayesian estimation, we are thus able to establish better bounds on the external controls used in these experiments, which are more variable than in other protocols.

Our second insight in this work was the extension of this method to work in situations when external normalization controls are not present. Because nascent transcriptional protocols measure transcription directly, and as a consequence of the structure of these protocols, at the start of an experiment, some number of RNA Polymerases are already bound to DNA and engaged in productive elongation of a transcript. Given that transcription proceeds at a known, finite rate, we then necessarily know that some portion of the data in our sample comes from transcription that has not been affected by any treatment given to the cells. Crucially, this approach is only feasible when the treatment applied to cells does not disrupt global transcription, and when the time point of an experiment is short enough that useful invariant regions exist. By extracting these presumptively invariant regions of nascent transcription from cells and applying the same normalization method we developed for external normalization controls, we are able to estimate normalization factors that are within error of those found using external controls. This

is remarkable — using previously unused data latent in our sequencing samples, we are able to determine effective normalization parameters even when an external control is not provided.

This work substantially advances the field of nascent transcription sequencing — it not only improves the quality of normalization, improving all downstream analysis, but also allows normalization to be performed on most previously published data where no such normalization was previously possible.

5.2 Separating Cells in Bulk Samples

My next project focused on establishing the feasibility of deconvolution analysis for nascent transcription data. One key limitation to the nascent transcription literature as it stands is the use of bulk sequencing assays — that is to say, some mix of cells from a population that are mixed together, sequenced, and treated as a single sample. While this is obviously not ideal, given the known heterogeneity of biological systems, it is still necessary, as efficient single cell nascent transcription protocols do not exist. At any given time, the amount of nascent transcription occurring in a single cell is tiny. It accounts for less than 1% of the total RNA in a cell, and after accounting for limitations in the efficiency of the reactions used for these protocols, there is very little useful nascent RNA left to extract information from. With this said, it is still of obvious biological interest to understand the transcriptional heterogeneity within a sequencing sample. To that end, my next work (chapter 3) examined the question of supervised deconvolution on transcriptional data. In sequencing experiments, deconvolution, also called microdissection, is the process of estimating the relative proportions of constituent cell types in a sample made up of a heterogeneous population of cells. Supervised deconvolution is a subset of this problem, and focuses on discovery of cell type mixing proportions when reference samples are available.

In this work, we established the feasibility of supervised deconvolution on transcriptional data, showing that published state-of-the-art methods do work reliably. Counterintuitively, we found that in the context of transcriptional data, simpler methods for deconvolution typically performed better than more complex ones, likely as a consequence of the additional regulatory information provided by transcriptional data. Remarkably, we also found that partially differentiated and undifferentiated

cells reliably confound this deconvolution process, suggesting that cells that are not fully differentiated resemble an ensemble of the transcriptional state of cells that they may eventually differentiate into. This work shows the feasibility of extending various deconvolution methods into transcriptional data where a wealth of untapped developmental and regulatory information is available in data.

This work establishes the feasibility of a new field of analysis of deconvolution of nascent transcriptional sequencing data, showing both the effectiveness of existing methods and new complicating factors that are not addressed by published literature.

5.3 Unsupervised Discovery of Cell Regulators

With my previous work establishing the feasibility of these deconvolution methods to nascent transcriptional data, I next sought to answer the question of developing methods for inferring cell-type specific regulators in an unsupervised way. The key idea to this work is this — transcriptional sequencing data, when paired with the underlying sequence information, contains important cell-type-specific regulatory information. However, this information is spread across a variety of regions across the genome, involves nonlinear responses to stimuli, and is not well annotated. In contrast to the previously studied problems, unsupervised learning appeared to be an effective approach to learning cell-type-specific regulatory information. Given the spatially encoded regulatory characteristics of transcription sequencing data, and recent advances in sequence modeling in deep learning, we chose to use a deep learning approach backed by modern mechanistic interpretability techniques. Using an autoencoder model, we learn a compact representation of transcription across a variety of cells, and then use interpretability methods to extract understandable features from this representation. In doing so, we learn a number of features generally associated with transcription and cell type specific markers of cell type, which are typically indicative of the oncogenic state of the cells used. Using interpretability techniques, we are able to extract interpretable features from model activations which are associated with cell types, providing a set of potential new flanking sequence regulatory sequences associated with known cell type regulators.

This research provides the first step in extending deep learning based sequence models into use in the study of cell-type specific nascent transcription and transcription regulation, providing a novel

and powerful tool for hypothesis generation.

5.4 Future Directions

While this work has resulted in a series of important and useful advances, the state of both machine learning and informatics mean that there is now a wide range of projects that were not feasible when this work started. The first area of research to extend on the work done here is the extension of these methods to a multi-omic context by the integration of single cell chromatin accessibility data (which continues to get less expensive). While not sufficient for transcription to occur, chromatin accessibility is thought to be a necessary condition for transcription to occur, simply because transcriptional proteins cannot bind to closed chromatin. Integrating chromatin accessibility and transcriptional data together is a deceptively simple idea. If you sequence transcription in a bulk population of cells, and then also find single cell chromatin accessibility on a sample taken from that same population of cells, you can use some clustering of the single cell accessibility data to estimate the most likely transcription patterns in those clusters. Immediately, this suggests two useful applications. First, if we are looking at a single population with single cell accessibility information, we can extend the deconvolution work done here to use the chromatin accessibility (plus some statistical estimation of counts distributions) as the reference samples for the deconvolution approach, letting us estimate the mixing proportion of your entire cellular population as well as the transcriptional patterns that underlie each subpopulation. Second, if we instead apply this deconvolution approach over a time course experiment (say, a differentiation experiment), we can use single cell tools to generate a trajectory of chromatin accessibility and generate overlapping clusters that cover the entire differentiation trajectory. By approximating transcription using these methods, we can then infer how a cell's transcriptional regulation changes over time. Using other tools developed in our lab, this method would allow for estimation of differential transcription factor activity over the entirety of a cell's differentiation trajectory (or within distinct cell type clusters in a sample), providing a rich view of the heterogeneity of transcription and how it differs both within cell populations and over time. This would go a long way towards helping us to better disentangle the confounding effect of heterogeneity on studying transcription.

In the vein of the VSI method developed here, an excellent future direction would be to extend the Bayesian modeling approach towards normalization into ATAC-seq data. Unlike protocols that extract RNA (where the amount of RNA produced in a cell is variable), ATAC-seq data has the nice property that each cell can only produce as many reads per-sample as there are copies of that chromosome. In a typical cell, this means that each cell can only produce 2 reads per location in the genome. Since the number of reads is discrete and bounded in a known quantity per-cell, the hybrid discrete/continuous model discussed in this work could be extended for accurate estimation of the number of input cells in an ATAC-seq sample and consequently of bounded normalization factors between samples. Ambitiously, this approach could also be extended towards the development of a more suitable differential expression (more accurately, differential accessibility) methodology for ATAC-seq data that more accurately leverages the underlying characteristics of the data. This kind of tool is desperately needed for current analysis and could also be extended to single cell data as well.

Additionally, the deep learning approaches proposed here have a huge number of possible applications, both in increasing the quality of the methods developed here as well as in furthering mechanistic interpretability of how these deep learning models are actually characterizing transcription and transcriptional changes. A substantial advance in the work done here is the use of a transformer based model over a deep convolutional neural network model. Notably, this transformer architecture can learn important sequence features using a much smaller number of parameters (on the order of 5000) compared to existing models (on the order of 50000 parameters). This efficiency is notable, but by scaling this model architecture up by one to two orders of magnitude, future researchers could instead train a model in the spirit of the one used here on the entirety of the samples in our database of nascent sequencing data. A larger model trained on a broad set of transcriptional data would instead learn the general characteristics of transcription across most human tissues, giving insight into the most defining sequence characteristics of transcription at both annotated and unannotated sites. Alternatively, a more effective approach, likely, is to view this problem as a translation problem instead of a reconstruction problem, which could allow this style of model to learn the characteristics driving differential behavior across treatments using a large corpus of data. Suppose that the process of transcription can be thought of as an expression of a

language that links sequence to cellular state (a metaphor that is common in literature studying this style of approach). Then, instead of focusing on learning reconstruction, as this work pioneered, we could use the same model architecture plus an added annotation of cell type or treatment, and train a model on the task of taking sequence + transcription from one cell type and predicting sequence + transcription for a second cell type on the exact same input region. With proper augmentation (e.g. training the task in both directions) such a model would learn an interpretable set of features where a fixed cell type could be provided and an arbitrary regulatory region in the genome, and the most influential sequence features that determine differences between cell type at a given locus would likely be learned. While this is speculation, such a model would most likely learn differential motif characteristics between cell types that reveal some sort of unique cell-type or treatment-response defining syntax.

Emerging work has also started to push towards the integration of transcriptional information with DNA sequencing data that captures genetic variation. Given that a substantial portion of SNPs in the genome are associated with noncoding regions including many that are accurately transcribed, further extending research into using these sorts of modeling approaches to describe disease will be important and valuable.

5.5 Closing Comments

More philosophically, starting research into machine learning for nascent transcription is now more feasible with far more possibilities than when this work started five years ago. The combination of the deep learning revolution, continuing decreases in sequencing costs, and the effective curation of databases of nascent transcriptional data[61], means that this research is more feasible than ever. These two subfields are a perfect complement, as interdisciplinary matches go. Deep learning remains deeply data starved, and sequencing data is often under-utilized and we produce more data than we are able to effectively use. It's a perfect fit — we can develop data hungry models using the surplus of data that we have, balancing out the weaknesses of both fields. Because biological research remains important, leveraging these tools and future advances to guide experimental design and hypothesis generation has the possibility not just to speed up future science, but also to decrease costs. We are just at the beginning

of our understanding of the total complexity of genetic regulation, and this work will remain important with substantial opportunities for novel and significant research.

Bibliography

1. Monk, M. Epigenetic Programming of Differential Gene Expression in Development and Evolution. **Developmental Genetics 17,** 188–197. ISSN: 1520-6408. (2024) (1995).
2. Liang, P. & Pardee, A. B. Analysing Differential Gene Expression in Cancer. **Nature Reviews Cancer 3,** 869–876. ISSN: 1474-1768. (2024) (Nov. 2003).
3. Crick, F. Central Dogma of Molecular Biology. **Nature 227,** 561–563. ISSN: 1476-4687. (2024) (Aug. 1970).
4. Lam, M. T. Y., Li, W., Rosenfeld, M. G. & Glass, C. K. Enhancer RNAs and Regulated Transcriptional Programs. **Trends in Biochemical Sciences 39,** 170–182. ISSN: 0968-0004. (2024) (Apr. 2014).
5. Sartorelli, V. & Lauberth, S. M. Enhancer RNAs Are an Important Regulatory Layer of the Epigenome. **Nature Structural & Molecular Biology 27,** 521–528. ISSN: 1545-9985. (2024) (June 2020).
6. Lewis, M. W., Li, S. & Franco, H. L. Transcriptional Control by Enhancers and Enhancer RNAs. **Transcription 10,** 171–186. ISSN: 2154-1264. (2024) (Oct. 2019).
7. Sanger, F. & Coulson, A. R. A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase. **Journal of Molecular Biology 94,** 441–448. ISSN: 0022-2836. (2024) (May 1975).
8. Smith, L. M. **et al.** Fluorescence Detection in Automated DNA Sequence Analysis. **Nature 321,** 674–679. ISSN: 0028-0836 (1986-06-12/0018).
9. Weber, J. L. & Myers, E. W. Human Whole-Genome Shotgun Sequencing. **Genome Research 7,** 401–409. ISSN: 1088-9051, 1549-5469. (2024) (May 1997).

10. Lander, E. S. **et al.** Initial Sequencing and Analysis of the Human Genome. **Nature 409,** 860–921. ISSN: 1476-4687. (2024) (Feb. 2001).

11. International Human Genome Sequencing Consortium. Finishing the Euchromatic Sequence of the Human Genome. **Nature 431,** 931–945. ISSN: 1476-4687. (2024) (Oct. 2004).

12. Elise Feingold. The ENCODE (ENCyclopedia Of DNA Elements) Project. **Science 306,** 636–640. (2024) (Oct. 2004).

13. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-Read Human Genome Sequencing and Its Applications. **Nature Reviews Genetics 21,** 597–614. ISSN: 1471-0064. (2024) (Oct. 2020).

14. Dobin, A. **et al.** STAR: Ultrafast Universal RNA-seq Aligner. **Bioinformatics 29,** 15–21. ISSN: 1367-4803. (2024) (Jan. 2013).

15. Jain, M. **et al.** Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads. **Nature Biotechnology 36,** 338–345. ISSN: 1546-1696. (2024) (Apr. 2018).

16. The ENCODE Project Consortium. Identification and Analysis of Functional Elements in 1% of the Human Genome by the ENCODE Pilot Project. **Nature 447,** 799–816. ISSN: 0028-0836, 1476-4687. (2024) (June 2007).

17. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. **Science 316,** 1497–1502. (2024) (June 2007).

18. Stormo, G. D. DNA Binding Sites: Representation and Discovery. **Bioinformatics 16,** 16–23. ISSN: 1367-4803. (2024) (Jan. 2000).

19. Castro-Mondragon, J. A. **et al.** JASPAR 2022: The 9th Release of the Open-Access Database of Transcription Factor Binding Profiles. **Nucleic Acids Research 50,** D165–D173. ISSN: 0305-1048. (2024) (Jan. 2022).

20. Heintzman, N. D. **et al.** Distinct and Predictive Chromatin Signatures of Transcriptional Promoters and Enhancers in the Human Genome. **Nature Genetics 39,** 311–318. ISSN: 1546-1718. (2024) (Mar. 2007).

21. Creyghton, M. P. **et al.** Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State. **Proceedings of the National Academy of Sciences 107,** 21931–21936. (2024) (Dec. 2010).

22. He, Q., Johnston, J. & Zeitlinger, J. ChIP-nexus Enables Improved Detection of in Vivo Transcription Factor Binding Footprints. **Nature Biotechnology 33,** 395–401. ISSN: 1546-1696. (2024) (Apr. 2015).

23. Rhee, H. S. & Pugh, B. F. ChIP-exo Method for Identifying Genomic Location of DNA-Binding Proteins with Near-Single-Nucleotide Accuracy. **Current Protocols in Molecular Biology 100,** 21.24.1–21.24.14. ISSN: 1934-3647. (2024) (2012).

24. Skene, P. J., Henikoff, J. G. & Henikoff, S. Targeted in Situ Genome-Wide Profiling with High Efficiency for Low Cell Numbers. **Nature Protocols 13,** 1006–1019. ISSN: 1750-2799. (2024) (May 2018).

25. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq. **Nature Methods 5,** 621–628. ISSN: 1548-7105 (July 2008).

26. Li, S. **et al.** Multi-Platform Assessment of Transcriptome Profiling Using RNA-seq in the ABRF next-Generation Sequencing Study. **Nature Biotechnology 32,** 915–925. ISSN: 1546-1696 (Sept. 2014).

27. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-binding Proteins and Nucleosome Position. **Nature Methods 10,** 1213–1218. ISSN: 1548-7105. (2022) (Dec. 2013).

28. Buenrostro, J., Wu, B., Chang, H. & Greenleaf, W. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. **Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.] 109,** 21.29.1–21.29.9. ISSN: 1934-3639. (2023) (Jan. 2015).

29. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. **Science (New York, N.Y.) 322,** 1845–1848. ISSN: 1095-9203 (Dec. 2008).

30. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. **Science (New York, N.Y.) 339,** 950–953. ISSN: 1095-9203 (Feb. 2013).

31. Wissink, E. M., Vihervaara, A., Tippens, N. D. & Lis, J. T. Nascent RNA Analyses: Tracking Transcription and Its Regulation. **Nature Reviews. Genetics 20,** 705–723. ISSN: 1471-0064 (Dec. 2019).

32. Cardiello, J. F., Sanchez, G. J., Allen, M. A. & Dowell, R. D. Lessons from eRNAs: Understanding Transcriptional Regulation through the Lens of Nascent RNAs. **Transcription 11,** 3–18. ISSN: 2154-1264 (Jan. 2020).

33. Mahat, D. B. **et al.** Base-Pair-Resolution Genome-Wide Mapping of Active RNA Polymerases Using Precision Nuclear Run-on (PRO-seq). **Nature Protocols 11,** 1455. ISSN: 1750-2799 (Aug. 2016).

34. Wang, Y. & Navin, N. E. Advances and Applications of Single-Cell Sequencing Technologies. **Molecular Cell 58,** 598–609. ISSN: 1097-2765. (2024) (May 2015).

35. Hwang, B., Lee, J. H. & Bang, D. Single-Cell RNA Sequencing Technologies and Bioinformatics Pipelines. **Experimental & Molecular Medicine 50,** 1–14. ISSN: 2092-6413. (2023) (Aug. 2018).

36. Macosko, E. Z. **et al.** Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. **Cell 161,** 1202–1214. ISSN: 0092-8674, 1097-4172. (2024) (May 2015).

37. Buenrostro, J. D. **et al.** Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation. **Nature 523,** 486–490. ISSN: 1476-4687. (2024) (July 2015).

38. Bentley, D. R. **et al.** Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry. **Nature 456,** 53–59. ISSN: 1476-4687. (2024) (Nov. 2008).

39. Wenger, A. M. **et al.** Accurate Circular Consensus Long-Read Sequencing Improves Variant Detection and Assembly of a Human Genome. **Nature Biotechnology 37,** 1155–1162. ISSN: 1546-1696. (2024) (Oct. 2019).

40. Andrews. **Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data** http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (2024).

41. Daley, T. & Smith, A. D. Predicting the Molecular Complexity of Sequencing Libraries. **Nature Methods 10,** 325–327. ISSN: 1548-7105. (2024) (Apr. 2013).

42. Bushnell, B. **BBMap: A Fast, Accurate, Splice-Aware Aligner** tech. rep. LBNL-7065E (Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), Mar. 2014). (2024).

43. Reinert, K., Langmead, B., Weese, D. & Evers, D. J. Alignment of Next-Generation Sequencing Reads. **Annual Review of Genomics and Human Genetics 16,** 133–151. ISSN: 1527-8204, 1545-293X. (2024) (Aug. 2015).

44. Harrow, J. **et al.** GENCODE: The Reference Human Genome Annotation for The ENCODE Project. **Genome Research 22,** 1760–1774. ISSN: 1088-9051, 1549-5469. (2024) (Sept. 2012).

45. Li, W. **et al.** RefSeq: Expanding the Prokaryotic Genome Annotation Pipeline Reach with Protein Family Model Curation. **Nucleic Acids Research 49,** D1020–D1028. ISSN: 0305-1048. (2024) (Jan. 2021).

46. Azofeifa, J., Allen, M. A., Lladser, M. & Dowell, R. **FStitch: A Fast and Simple Algorithm for Detecting Nascent RNA Transcripts** in (ACM, New York, NY, USA, Sept. 2014), 174–183. ISBN: 978-1-4503-2894-4. (2017).

47. Azofeifa, J. G. & Dowell, R. D. A Generative Model for the Behavior of RNA Polymerase. **Bioinformatics 33,** 227–234. ISSN: 1367-4803. (2017) (Jan. 2017).

48. Danko, C. G. **et al.** Identification of Active Transcriptional Regulatory Elements from GRO-seq Data. **Nature Methods 12,** 433–438. ISSN: 1548-7105. (2022) (May 2015).

49. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features. **Bioinformatics (Oxford, England) 30,** 923–930. ISSN: 1367-4811 (Apr. 2014).

50. Liao, Y., Smyth, G. K. & Shi, W. The R Package Rsubread Is Easier, Faster, Cheaper and Better for Alignment and Quantification of RNA Sequencing Reads. **Nucleic Acids Research 47,** e47. ISSN: 0305-1048. (2022) (May 2019).

51. Love, M. I., Huber, W. & Anders, S. Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2. **Genome Biology 15,** 550. ISSN: 1474-760X (Dec. 2014).

52. Gong, T. **et al.** Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples. **PLOS ONE 6,** e27156. ISSN: 1932-6203. (2021) (Nov. 2011).

53. Avila Cobos, F., Vandesompele, J., Mestdagh, P. & De Preter, K. Computational Deconvolution of Transcriptomics Data from Mixed Cell Populations. **Bioinformatics 34,** 1969–1979. ISSN: 1367-4803. (2021) (June 2018).

54. Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H. F. Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. **PLOS ONE 4,** e6098. ISSN: 1932-6203. (2021) (July 2009).

55. Pliner, H. A. **et al.** Cicero Predicts Cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. **Molecular Cell 71,** 858–871.e8. ISSN: 1097-2765. (2021) (Sept. 2018).

56. Mohammadi, S., Zuckerman, N., Goldsmith, A. & Grama, A. A Critical Survey of Deconvolution Methods for Separating Cell Types in Complex Tissues. **Proceedings of the IEEE 105,** 340–366. ISSN: 1558-2256 (Feb. 2017).

57. Erdmann-Pham, D. D., Fischer, J., Hong, J. & Song, Y. S. Likelihood-Based Deconvolution of Bulk Gene Expression Data Using Single-Cell References. **Genome Research 31,** 1794–1806. ISSN: 1088-9051, 1549-5469 (Oct. 2021).

58. Shen-Orr, S. S. **et al.** Cell Type–Specific Gene Expression Differences in Complex Tissues. **Nature Methods 7,** 287–289. ISSN: 1548-7105. (2022) (Apr. 2010).

59. Zhong, Y., Wan, Y.-W., Pang, K., Chow, L. M. & Liu, Z. Digital Sorting of Complex Tissues for Cell Type-Specific Gene Expression Profiles. **BMC Bioinformatics 14,** 89. ISSN: 1471-2105. (2021) (Mar. 2013).

60. Maas, Z. L. & Dowell, R. D. Internal and External Normalization of Nascent RNA Sequencing Run-on Experiments. **BMC Bioinformatics 25,** 19. ISSN: 1471-2105. (2024) (Jan. 2024).

61. Sigauke, R. F. **et al. Atlas of Nascent RNA Transcripts Reveals Enhancer to Gene Linkages** Dec. 2023. (2024).

62. Mahat, D. B., Salamanca, H. H., Duarte, F. M., Danko, C. G. & Lis, J. T. Mammalian Heat Shock Response and Mechanisms Underlying Its Genome-wide Transcriptional Regulation. **Molecular Cell 62,** 63–78. ISSN: 1097-2765 (Apr. 2016).

63. Vihervaara, A. **et al.** Transcriptional Response to Stress Is Pre-Wired by Promoter and Enhancer Architecture. **Nature Communications 8,** 255. ISSN: 2041-1723 (Aug. 2017).

64. Jiang, L. **et al.** Synthetic Spike-in Standards for RNA-seq Experiments. **Genome Research 21,** 1543–1551. ISSN: 1088-9051 (Sept. 2011).

65. Chen, K. **et al.** The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses. **Molecular and Cellular Biology 36,** 662–667. ISSN: 0270-7306 (Feb. 2016).

66. Evans, C., Hardin, J. & Stoebel, D. M. Selecting Between-Sample RNA-Seq Normalization Methods from the Perspective of Their Assumptions. **Briefings in Bioinformatics 19,** 776–792. ISSN: 1467-5463 (Feb. 2017).

67. Hunter, S., Sigauke, R. F., Stanley, J. T., Allen, M. A. & Dowell, R. D. Protocol Variations in Run-on Transcription Dataset Preparation Produce Detectable Signatures in Sequencing Libraries. **BMC Genomics 23,** 187. ISSN: 1471-2164 (Mar. 2022).

68. Schwalb, B. **et al.** TT-seq Maps the Human Transient Transcriptome. **Science (New York, N.Y.) 352,** 1225–1228. ISSN: 1095-9203 (June 2016).

69. Ritchie, M. E. **et al.** Limma Powers Differential Expression Analyses for RNA-sequencing and Microarray Studies. **Nucleic Acids Research 43,** e47. ISSN: 0305-1048 (Apr. 2015).

70. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The Sva Package for Removing Batch Effects and Other Unwanted Variation in High-Throughput Experiments. **Bioinformatics 28,** 882–883. ISSN: 1367-4803 (Mar. 2012).

71. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. **Biostatistics 8,** 118–127. ISSN: 1465-4644 (Jan. 2007).

72. Wu, H., Wang, C. & Wu, Z. A New Shrinkage Estimator for Dispersion Improves Differential Expression Detection in RNA-seq Data. **Biostatistics 14,** 232–243. ISSN: 1465-4644 (Apr. 2013).

73. Choi, Y., Coram, M., Peng, J. & Tang, H. A Poisson Log-Normal Model for Constructing Gene Covariation Network Using RNA-seq Data. **Journal of Computational Biology 24,** 721–731. ISSN: 1557-8666 (July 2017).

74. Gelman, A. **Bayesian Data Analysis** Third edition. ISBN: 978-1-4398-4095-5 (CRC Press, Boca Raton, 2014).

75. Aoi, Y. **et al.** NELF Regulates a Promoter-Proximal Step Distinct from RNA Pol II Pause-Release. **Molecular Cell 78,** 261–274.e5. ISSN: 1097-4164 (Apr. 2020).

76. Barbieri, E. **et al.** Rapid and Scalable Profiling of Nascent RNA with fastGRO. **Cell Reports 33,** 108373. ISSN: 2211-1247 (Nov. 2020).

77. Birkenheuer, C. H., Danko, C. G. & Baines, J. D. Herpes Simplex Virus 1 Dramatically Alters Loading and Positioning of RNA Polymerase II on Host Genes Early in Infection. **Journal of Virology 92,** e02184–17. ISSN: 1098-5514 (Apr. 2018).

78. Birkenheuer, C. H. & Baines, J. D. RNA Polymerase II Promoter-Proximal Pausing and Release to Elongation Are Key Steps Regulating Herpes Simplex Virus 1 Transcription. **Journal of Virology 94,** e02035–19. ISSN: 0022-538X (Feb. 2020).

79. Dukler, N. **et al.** Nascent RNA Sequencing Reveals a Dynamic Global Transcriptional Response at Genes and Enhancers to the Natural Medicinal Compound Celastrol. **Genome Research.** ISSN: 1088-9051, 1549-5469 (Oct. 2017).

80. Fan, Z. **et al.** CDK13 Cooperates with CDK12 to Control Global RNA Polymerase II Processivity. **Science Advances 6,** eaaz5041 (Apr. 2020).

81. Jaeger, M. G. **et al.** Selective Mediator Dependence of Cell-Type-Specifying Transcription. **Nature Genetics 52,** 719–727. ISSN: 1546-1718 (July 2020).

82. LeRoy, G. **et al.** LEDGF and HDGF2 Relieve the Nucleosome-Induced Barrier to Transcription in Differentiated Cells. **Science Advances 5,** eaay3068 (Oct. 2019).

83. Liu, N. **et al.** Author Correction: Transcription Factor Competition at the γ-Globin Promoters Controls Hemoglobin Switching. **Nature Genetics 53,** 586–586. ISSN: 1546-1718 (Apr. 2021).

84. Rao, S. S. **et al.** Cohesin Loss Eliminates All Loop Domains. **Cell 171,** 305–320.e24. ISSN: 00928674 (Oct. 2017).

85. Santoriello, C. **et al.** RNA Helicase DDX21 Mediates Nucleotide Stress Responses in Neural Crest and Melanoma Cells. **Nature Cell Biology 22,** 372–379. ISSN: 1476-4679 (Apr. 2020).

86. Sendinc, E. **et al.** PCIF1 Catalyzes m6Am mRNA Methylation to Regulate Gene Expression. **Molecular Cell 75,** 620–630.e9. ISSN: 1097-4164 (Aug. 2019).

87. Takahashi, H. **et al.** The Role of Mediator and Little Elongation Complex in Transcription Termination. **Nature Communications 11,** 1063. ISSN: 2041-1723 (Feb. 2020).

88. Vihervaara, A. **et al.** Stress-Induced Transcriptional Memory Accelerates Promoter-Proximal Pause Release and Decelerates Termination over Mitotic Divisions. **Molecular Cell 81,** 1715–1731.e6. ISSN: 1097-4164 (Apr. 2021).

89. Daines, B. **et al.** The Drosophila Melanogaster Transcriptome by Paired-End RNA Sequencing. **Genome Research 21,** 315–324. ISSN: 1088-9051 (Feb. 2011).

90. Jonkers, I., Kwak, H. & Lis, J. T. Genome-Wide Dynamics of Pol II Elongation and Its Interplay with Promoter Proximal Pausing, Chromatin, and Exons. **eLife 3.** ISSN: 2050-084X (Apr. 2014).

91. Mimoso, C. A. & Adelman, K. U1 snRNP Increases RNA Pol II Elongation Rate to Enable Synthesis of Long Genes. **Molecular Cell 83,** 1264–1279.e10. ISSN: 1097-2765. (2023) (Apr. 2023).

92. Noe Gonzalez, M., Blears, D. & Svejstrup, J. Q. Causes and Consequences of RNA Polymerase II Stalling during Transcript Elongation. **Nature Reviews Molecular Cell Biology 22,** 3–21. ISSN: 1471-0080. (2023) (Jan. 2021).

93. Fuchs, G. **et al.** 4sUDRB-seq: Measuring Genomewide Transcriptional Elongation Rates and Initiation Frequencies within Cells. **Genome Biology 15,** R69. ISSN: 1474-760X. (2023) (May 2014).

94. Muniz, L., Nicolas, E. & Trouche, D. RNA Polymerase II Speed: A Key Player in Controlling and Adapting Transcriptome Composition. **The EMBO Journal 40,** e105740. ISSN: 0261-4189. (2023) (Aug. 2021).

95. Fant, C. B. **et al.** TFIID Enables RNA Polymerase II Promoter-Proximal Pausing. **Molecular Cell 78,** 785–793.e8. ISSN: 1097-2765 (May 2020).

96. Herzog, V. A. **et al.** Thiol-linked alkylation of RNA to assess expression dynamics. **Nature Methods 14,** 1198–1204 (2017).

97. Nojima, T., Gomes, T., Carmo-Fonseca, M. & Proudfoot, N. J. Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. **Nature Protocols 11,** 413–428 (2016).

98. Salvatier, J., Wiecki, T. V. & Fonnesbeck, C. Probabilistic Programming in Python Using PyMC3. **PeerJ Computer Science 2,** e55. ISSN: 2376-5992 (Apr. 2016).

99. Hoffman, M. D. & Gelman, A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. **Journal of Machine Learning Research 15,** 1593–1623. ISSN: 1533-7928 (2014).

100. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. **The Journal of Chemical Physics 21,** 1087–1092. ISSN: 0021-9606 (June 1953).

101. Hastings, W. K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. **Biometrika 57,** 97–109. ISSN: 0006-3444 (1970).

102. Tripodi, I. J. & Gruca, M. A. Nascent-Flow (Dec. 2018).

103. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features. **Bioinformatics 30,** 923–930. ISSN: 1367-4803 (Apr. 2014).

104. Maas, Z., Sigauke, R. & Dowell, R. **Deconvolution of Nascent Sequencing Data Using Transcriptional Regulatory Elements** Preprint (Bioinformatics, Oct. 2023). (2023).

105. Azofeifa, J. G. **et al.** Enhancer RNA Profiling Predicts Transcription Factor Activity. **Genome Research 28,** 334–344. ISSN: 1088-9051, 1549-5469. (2019) (Mar. 2018).

106. Rubin, J. D. **et al.** Transcription Factor Enrichment Analysis (TFEA) Quantifies the Activity of Multiple Transcription Factors from a Single Experiment. **Communications Biology 4,** 1–15. ISSN: 2399-3642. (2021) (June 2021).

107. Cardiello, J. F., Sanchez, G. J., Allen, M. A. & Dowell, R. D. Lessons from eRNAs: Understanding Transcriptional Regulation through the Lens of Nascent RNAs. **Transcription 11,** 3–18. ISSN: 2154-1264. (2023) (Jan. 2020).

108. Kim, T.-K. **et al.** Widespread transcription at neuronal activity-regulated enhancers. **Nature 465,** 182–187 (2010).

109. Wang, Z., Chu, T., Choate, L. A. & Danko, C. G. Identification of Regulatory Elements from Nascent Transcription Using dREG. **Genome Research 29,** 293–303. ISSN: 1549-5469 (Feb. 2019).

110. Kaikkonen, M. U. **et al.** Remodeling of the Enhancer Landscape during Macrophage Activation Is Coupled to Enhancer Transcription. **Molecular Cell 51,** 310–325. ISSN: 1097-4164 (Aug. 2013).

111. Kristjánsdóttir, K., Dziubek, A., Kang, H. M. & Kwak, H. Population-Scale Study of eRNA Transcription Reveals Bipartite Functional Enhancer Architecture. **Nature Communications 11,** 5963. ISSN: 2041-1723. (2023) (Nov. 2020).

112. Bae, S. **et al.** RANKL-responsive epigenetic mechanism reprograms macrophages into bone-resorbing osteoclasts. **Cellular & Molecular Immunology 20,** 94–109 (2023).

113. Spitz, F. & Furlong, E. E. M. Transcription Factors: From Enhancer Binding to Developmental Control. **Nature Reviews Genetics 13,** 613–626. ISSN: 1471-0064. (2023) (Sept. 2012).

114. Lidschreiber, K. **et al.** Transcriptionally Active Enhancers in Human Cancer Cells. **Molecular Systems Biology 17,** e9873. ISSN: 1744-4292. (2023) (Jan. 2021).

115. Tripodi, I. J. & Gruca, M. A. **Nascent-Flow** Dec. 2018. (2023).

116. Li, H. **et al.** The Sequence Alignment/Map Format and SAMtools. **Bioinformatics 25,** 2078–2079. ISSN: 1367-4803. (2023) (Aug. 2009).

117. Jiang, W. **et al.** A Multi-Parameter Analysis of Cellular Coordination of Major Transcriptome Regulation Mechanisms. **Scientific Reports 8,** 5742. ISSN: 2045-2322. (2023) (Apr. 2018).

118. Fei, J. **et al.** NDF, a Nucleosome-Destabilizing Factor That Facilitates Transcription through Nucleosomes. **Genes & Development 32,** 682–694. ISSN: 0890-9369, 1549-5477. (2023) (May 2018).

119. Andrysik, Z. **et al.** Identification of a Core TP53 Transcriptional Program with Highly Distributed Tumor Suppressive Activity. **Genome Research 27,** 1645–1657. ISSN: 1088-9051, 1549-5469. (2023) (Oct. 2017).

120. Zhao, Y. **et al.** High-Resolution Mapping of RNA Polymerases Identifies Mechanisms of Sensitivity and Resistance to BET Inhibitors in t(8;21) AML. **Cell Reports 16,** 2003–2016. ISSN: 2211-1247 (Aug. 2016).

121. Danko, C. G. **et al.** Dynamic Evolution of Regulatory Element Ensembles in Primate CD4+ T Cells. **Nature Ecology & Evolution 2,** 537–548. ISSN: 2397-334X (2018).

122. Chu, T. **et al.** Chromatin Run-on and Sequencing Maps the Transcriptional Regulatory Landscape of Glioblastoma Multiforme. **Nature genetics 50,** 1553–1564 (2018).

123. Core, L. J. **et al.** Analysis of Nascent RNA Identifies a Unified Architecture of Initiation Regions at Mammalian Promoters and Enhancers. **Nature Genetics 46,** 1311–1320. ISSN: 1546-1718. (2023) (Dec. 2014).

124. Smith, J. P., Dutta, A. B., Sathyan, K. M., Guertin, M. J. & Sheffield, N. C. Quality Control and Processing of Nascent RNA Profiling Data. **bioRxiv 22,** 2020.02.27.956110. (2020) (Feb. 2020).

125. Ikegami, K., Secchia, S., Almakki, O., Lieb, J. D. & Moskowitz, I. P. Phosphorylated Lamin A/C in the Nuclear Interior Binds Active Enhancers Associated with Abnormal Transcription in Progeria. **Developmental Cell 52,** 699–713.e11. ISSN: 1534-5807. (2023) (Mar. 2020).

126. Zhao, Y. **et al.** Deconvolution of Expression for Nascent RNA-sequencing Data (DENR) Highlights Pre-RNA Isoform Diversity in Human Cells. **Bioinformatics 37.** ISSN: 1367-4803. (2021) (24 Aug. 2021).

127. Newman, A. M. **et al.** Robust Enumeration of Cell Subsets from Tissue Expression Profiles. **Nature Methods 12,** 453–457. ISSN: 1548-7105. (2022) (May 2015).

128. Landry, J. J. M. **et al.** The Genomic and Transcriptomic Landscape of a HeLa Cell Line. **G3: Genes|Genomes|Genetics 3,** 1213–1224. ISSN: 2160-1836. (2023) (Mar. 2013).

129. Efroni, S. **et al.** Global Transcription in Pluripotent Embryonic Stem Cells. **Cell stem cell 2,** 437–447. ISSN: 1934-5909. (2023) (May 2008).

130. Chung, Y. S. **et al.** Undifferentiated Hematopoietic Cells Are Characterized by a Genome-Wide Undermethylation Dip around the Transcription Start Site and a Hierarchical Epigenetic Plasticity. **Blood 114,** 4968–4978. ISSN: 0006-4971. (2023) (Dec. 2009).

131. Chu, T. **et al.** Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. **Nature Genetics 50,** 1553–1564 (2018).

132. van der Maaten, L. & Hinton, G. Visualizing High-Dimensional Data Using t-SNE. **Journal of Machine Learning Research 9,** 2579–2605. ISSN: 1532-4435 (2008).

133. McInnes, L., Healy, J. & Melville, J. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction** Sept. 2020. arXiv: 1802.03426 [cs, stat]. (2022).

134. Avsec, Ž. **et al.** Base-Resolution Models of Transcription-Factor Binding Reveal Soft Motif Syntax. **Nature Genetics 53,** 354–366. ISSN: 1546-1718. (2024) (Mar. 2021).

135. He, A. Y. & Danko, C. G. Dissection of Core Promoter Syntax through Single Nucleotide Resolution Modeling of Transcription Initiation. **bioRxiv,** 2024.03.13.583868. (2024) (Sept. 2024).

136. Azofeifa, J. G. **et al.** Enhancer RNA Profiling Predicts Transcription Factor Activity. **Genome Research 28,** 334–344. ISSN: 1088-9051. (2023) (Mar. 2018).

137. Jones, T. **et al. A Transcription Factor (TF) Inference Method That Broadly Measures TF Activity and Identifies Mechanistically Distinct TF Networks** Mar. 2024. (2024).

138. Shrikumar, A., Greenside, P. & Kundaje, A. **Learning Important Features through Propagating Activation Differences** in **Proceedings of the 34th International Conference on Machine Learning - Volume 70** (JMLR.org, Sydney, NSW, Australia, Aug. 2017), 3145–3153. (2024).

139. Bricken, T. **et al. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning** Oct. 2023. (2024).

140. Templeton, A. **et al. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet** May 2024.

141. Elhage, N. **et al. Toy Models of Superposition** Sept. 2022. arXiv: 2209.10652. (2024).

142. Bailey, T. L. & Elkan, C. Fitting a Mixture Model By Expectation Maximization To Discover Motifs In Biopolymer.

143. Yao, L. **et al.** A Comparison of Experimental Assays and Analytical Methods for Genome-Wide Identification of Active Enhancers. **Nature Biotechnology 40,** 1056–1065. ISSN: 1546-1696 (July 2022).

144. Vaswani, A. **et al. Attention Is All You Need** Aug. 2023. arXiv: 1706.03762. (2024).

145. Lundberg, S. M. & Lee, S.-I. **A Unified Approach to Interpreting Model Predictions** in **Advances in Neural Information Processing Systems 30** (Curran Associates, Inc., 2017). (2024).

146. Shrikumar, A. **et al. Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) Version 0.5.6.5** Apr. 2020. arXiv: 1811.00416. (2024).

147. Rennoll, S. & Yochum, G. Regulation of MYC Gene Expression by Aberrant Wnt/β-Catenin Signaling in Colorectal Cancer. **World Journal of Biological Chemistry 6,** 290–300. ISSN: 1949-8454. (2024) (Nov. 2015).

148. Lind, D. S. **et al.** Nuclear Factor-κB Is Upregulated in Colorectal Cancer. **Surgery 130,** 363–369. ISSN: 0039-6060. (2024) (Aug. 2001).

149. Tsiambas, E. **et al.** Significance of Estrogen Receptor 1 (ESR-1) Gene Imbalances in Colon and Hepatocellular Carcinomas Based on Tissue Microarrays Analysis. **Medical Oncology 28,** 934–940. ISSN: 1559-131X. (2024) (Dec. 2011).

150. Wenzel, J. **et al.** Loss of the Nuclear Wnt Pathway Effector TCF7L2 Promotes Migration and Invasion of Human Colorectal Cancer Cells. **Oncogene 39,** 3893–3909. ISSN: 1476-5594. (2024) (May 2020).

151. Hellebrekers, D. M. **et al.** GATA4 and GATA5 Are Potential Tumor Suppressors and Biomarkers in Colorectal Cancer. **Clinical Cancer Research 15,** 3990–3997. ISSN: 1078-0432. (2024) (June 2009).

152. Slattery, M. L. **et al.** Associations between Genetic Variation in RUNX1 , RUNX2 , RUNX3 , MAPK1 and eIF4E and Risk of Colon and Rectal Cancer: Additional Support for a TGF-β-signaling Pathway. **Carcinogenesis 32,** 318–326. ISSN: 0143-3334. (2024) (Mar. 2011).

153. Li, D., Li, Q., Zhuo, C., Xu, Y. & Cai, S. Contribution of FOXC1 to the Progression and Metastasis and Prognosis of Human Colon Cancer. **Journal of Clinical Oncology 33,** 636–636. ISSN: 0732-183X. (2024) (Jan. 2015).

154. Goto, N. **et al.** SOX17 Enables Immune Evasion of Early Colorectal Adenomas and Cancers. **Nature 627,** 636–645. ISSN: 1476-4687. (2024) (Mar. 2024).

155. Chang, Y. **et al.** USP7-mediated JUND Suppresses RCAN2 Transcription and Elevates NFATC1 to Enhance Stem Cell Property in Colorectal Cancer. **Cell Biology and Toxicology 39,** 3121–3140. ISSN: 1573-6822. (2024) (Dec. 2023).

156. Vilasco, M. **et al.** Glucocorticoid Receptor and Breast Cancer. **Breast Cancer Research and Treatment 130,** 1–10. ISSN: 1573-7217. (2024) (Nov. 2011).

157. Moutsatsou, P. & Papavassiliou, A. G. The Glucocorticoid Receptor Signalling in Breast Cancer. **Journal of Cellular and Molecular Medicine 12,** 145–163. ISSN: 1582-4934. (2024) (2008).

158. Jiang, S. **et al.** Multifaceted Roles of HSF1 in Cancer. **Tumor Biology 36,** 4923–4931. ISSN: 1423-0380. (2024) (July 2015).

159. Carpenter, R. L. **et al.** Combined Inhibition of AKT and HSF1 Suppresses Breast Cancer Stem Cells and Tumor Growth. **Oncotarget 8,** 73947–73963. ISSN: 1949-2553. (2024) (May 2017).

160. Jana, S. **et al.** SOX9: The Master Regulator of Cell Fate in Breast Cancer. **Biochemical Pharmacology 174,** 113789. ISSN: 0006-2952. (2024) (Apr. 2020).

161. Stuelten, C. H. **et al.** Smad4-Expression Is Decreased in Breast Cancer Tissues: A Retrospective Study. **BMC Cancer 6,** 25. ISSN: 1471-2407. (2024) (Jan. 2006).

162. Liu, N.-n. **et al.** SMAD4 Is a Potential Prognostic Marker in Human Breast Carcinomas. **Tumor Biology 35,** 641–650. ISSN: 1423-0380. (2024) (Jan. 2014).

163. Li, Q. **et al.** Smad4 Inhibits Tumor Growth by Inducing Apoptosis in Estrogen Receptor-α-positive Breast Cancer Cells *. **Journal of Biological Chemistry 280,** 27022–27028. ISSN: 0021-9258, 1083-351X. (2024) (July 2005).

164. Bidwell, B. N. **et al.** Silencing of Irf7 Pathways in Breast Cancer Cells Promotes Bone Metastasis through Immune Escape. **Nature Medicine 18,** 1224–1231. ISSN: 1546-170X. (2024) (Aug. 2012).

165. Lan, Q. **et al.** Type I Interferon/IRF7 Axis Instigates Chemotherapy-Induced Immunological Dormancy in Breast Cancer. **Oncogene 38,** 2814–2829. ISSN: 1476-5594. (2024) (Apr. 2019).

166. Costa, T. D. F. **et al.** PAK4 Suppresses RELB to Prevent Senescence-like Growth Arrest in Breast Cancer. **Nature Communications 10,** 3589. ISSN: 2041-1723. (2024) (Aug. 2019).

167. Pessa, J. C. **et al. Dynamic HSF2 Regulation Drives Breast Cancer Progression by Steering the Balance between Proliferation and Invasion** June 2024. (2024).

168. Yang, J. **et al.** Downregulation of miR-10b Promotes Osteoblast Differentiation through Targeting Bcl6. **International Journal of Molecular Medicine 39,** 1605–1612. ISSN: 1107-3756. (2024) (June 2017).

169. Chai, J., Xu, S. & Guo, F. TEAD1 Mediates the Oncogenic Activities of Hippo-YAP1 Signaling in Osteosarcoma. **Biochemical and Biophysical Research Communications 488,** 297–302. ISSN: 0006-291X. (2024) (June 2017).

170. Pachano, T. **et al.** Orphan CpG Islands Amplify Poised Enhancer Regulatory Activity and Determine Target Gene Responsiveness. **Nature Genetics 53,** 1036–1049. ISSN: 1546-1718. (2024) (July 2021).

171. Geva, M., Schuster, R., Berant, J. & Levy, O. **Transformer Feed-Forward Layers Are Key-Value Memories** Sept. 2021. arXiv: 2012.14913. (2024).

172. Ballian, N., Liu, S.-H. & Brunicardi, F. C. Transcription Factor PDX-1 in Human Colorectal Adenocarcinoma: A Potential Tumor Marker? **World Journal of Gastroenterology : WJG 14,** 5823–5826. ISSN: 1007-9327. (2024) (Oct. 2008).

173. Zhou, X. **et al.** RelB Plays an Oncogenic Role and Conveys Chemo-Resistance to DLD-1 Colon Cancer Cells. **Cancer Cell International 18,** 181. ISSN: 1475-2867. (2024) (Nov. 2018).

174. Wang, L. **et al.** HOXB4 Mis-Regulation Induced by Microcystin-LR and Correlated With Immune Infiltration Is Unfavorable to Colorectal Cancer Prognosis. **Frontiers in Oncology 12,** 803493. ISSN: 2234-943X. (2024) (Feb. 2022).

175. L, S. **et al.** Requirement for Pbx1 in Skeletal Patterning and Programming Chondrocyte Proliferation and Differentiation. **Development (Cambridge, England) 128.** ISSN: 0950-1991. (2024) (Sept. 2001).

176. Říhová, K. **et al.** Transcription Factor C-Myb: Novel Prognostic Factor in Osteosarcoma. **Clinical & Experimental Metastasis 39,** 375–390. ISSN: 1573-7276 (Apr. 2022).

177. LETTICE, ., HECKSHER-SØRENSEN, . & HILL, .. The Role of Bapx1 (Nkx3.2) in the Development and Evolution of the Axial Skeleton. **Journal of Anatomy 199,** 181–187. ISSN: 0021-8782. (2024) (2001).

178. Hum, J. M., Day, R. N., Bidwell, J. P., Wang, Y. & Pavalko, F. M. Mechanical Loading in Osteocytes Induces Formation of a Src/Pyk2/MBD2 Complex That Suppresses Anabolic Gene Expression. **PLOS ONE 9,** e97942. ISSN: 1932-6203. (2024) (May 2014).

179. Barron, D. N. The Analysis of Count Data: Overdispersion and Autocorrelation. **Sociological Methodology 22,** 179–220. ISSN: 0081-1750 (1992).

180. Esser, P., Rombach, R. & Ommer, B. **Taming Transformers for High-Resolution Image Synthesis** June 2021. arXiv: 2012.09841. (2024).

181. Loshchilov, I. & Hutter, F. **Decoupled Weight Decay Regularization** Jan. 2019. arXiv: 1711.05101. (2024).

182. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying Similarity between Motifs. **Genome Biology 8,** R24. ISSN: 1474-760X. (2024) (Feb. 2007).

183. Kulakovskiy, I. V. **et al.** HOCOMOCO: Towards a Complete Collection of Transcription Factor Binding Models for Human and Mouse via Large-Scale ChIP-Seq Analysis. **Nucleic Acids Research 46,** D252–D259. ISSN: 0305-1048. (2024) (Jan. 2018).

184. Wang, L., Zhang, M.-X., Zhang, M.-F. & Tu, Z.-W. ZBTB7A Functioned as an Oncogene in Colorectal Cancer. **BMC Gastroenterology 20,** 370. ISSN: 1471-230X. (2024) (Nov. 2020).

185. Rothzerg, E. **et al.** Upregulation of 15 Antisense Long Non-Coding RNAs in Osteosarcoma. **Genes 12,** 1132. ISSN: 2073-4425. (2024) (July 2021).

186. Somarelli, J. A. **et al.** Mesenchymal-Epithelial Transition in Sarcomas Is Controlled by the Combinatorial Expression of MicroRNA 200s and GRHL2. **Molecular and Cellular Biology 36,** 2503–2513. ISSN: 0270-7306. (2024) (Sept. 2016).

187. Liu, W., Hao, Y., Tian, X., Jiang, J. & Qiu, Q. The Role of NR4A1 in the Pathophysiology of Osteosarcoma: A Comprehensive Bioinformatics Analysis of the Single-Cell RNA Sequencing Dataset. **Frontiers in Oncology 12,** 879288. ISSN: 2234-943X. (2024) (July 2022).

188. Smeester, B. A. **et al.** Implication of ZNF217 in Accelerating Tumor Development and Therapeutically Targeting ZNF217-Induced PI3K-AKT Signaling for the Treatment of Metastatic Osteosarcoma. **Molecular Cancer Therapeutics 19,** 2528–2541. ISSN: 1538-8514 (Dec. 2020).

189. Saito, H. **et al.** TG-interacting Factor 1 (Tgif1)-Deficiency Attenuates Bone Remodeling and Blunts the Anabolic Response to Parathyroid Hormone. **Nature Communications 10,** 1354. ISSN: 2041-1723. (2024) (Mar. 2019).

190. Tan, J. **et al.** Identification and Analysis of Three Hub Prognostic Genes Related to Osteosarcoma Metastasis. **Journal of Oncology 2021,** 6646459. ISSN: 1687-8450. (2024) (Jan. 2021).

191. Lang, T. **et al.** NFATC2 Is a Novel Therapeutic Target for Colorectal Cancer Stem Cells. **OncoTargets and therapy 11,** 6911–6924. ISSN: 1178-6930. (2024) (Oct. 2018).

192. Zhang, Y. **et al.** ZNF8 Promotes Progression of Gastrointestinal Cancers via a P53-Dependent Mechanism. **Cellular Signalling 123,** 111354. ISSN: 1873-3913 (Nov. 2024).

193. Yu, J., Liu, M., Liu, H. & Zhou, L. GATA1 Promotes Colorectal Cancer Cell Proliferation, Migration and Invasion via Activating AKT Signaling Pathway. **Molecular and Cellular Biochemistry 457,** 191–199. ISSN: 1573-4919 (July 2019).

194. Xiao, Y., Liu, Y., Sun, Y., Huang, C. & Zhong, S. MEIS2 Suppresses Breast Cancer Development by Downregulating IL10. **Cancer Reports 7,** e2064. ISSN: 2573-8348. (2024) (May 2024).

195. El Dika, M. **et al.** Epigenetic-Mediated Regulation of Gene Expression for Biological Control and Cancer: Fidelity of Mechanisms Governing the Cell Cycle. **Results and problems in cell differentiation 70,** 375–396. ISSN: 0080-1844. (2024) (2022).

196. Li, Y.-J., Yang, Z., Wang, Y.-Y. & Wang, Y. Long Noncoding RNA ZNF667-AS1 Reduces Tumor Invasion and Metastasis in Cervical Cancer by Counteracting microRNA-93-3p-dependent PEG3 Downregulation. **Molecular Oncology 13,** 2375–2392. ISSN: 1574-7891. (2024) (Nov. 2019).

197. Yu, C. **et al.** The lncRNA ZNF667-AS1 Inhibits Propagation, Invasion, and Angiogenesis of Gastric Cancer by Silencing the Expression of N-Cadherin and VEGFA. **Journal of Oncology 2022,** 3579547. ISSN: 1687-8450. (2024) (July 2022).

198. Brechka, H., Bhanvadia, R. R., VanOpstall, C. & Vander Griend, D. J. HOXB13 Mutations and Binding Partners in Prostate Development and Cancer: Function, Clinical Significance, and Future Directions. **Genes & Diseases 4,** 75–87. ISSN: 2352-3042. (2024) (Feb. 2017).

199. Li, W.-f. **et al.** The Transcription Factor PBX3 Promotes Tumor Cell Growth through Transcriptional Suppression of the Tumor Suppressor P53. **Acta Pharmacologica Sinica 42,** 1888–1899. ISSN: 1745-7254. (2024) (Nov. 2021).

200. Gilmore, T. D. & Gerondakis, S. The C-Rel Transcription Factor in Development and Disease. **Genes & Cancer 2,** 695–711. ISSN: 1947-6019. (2024) (July 2011).

201. Huang, S., Hou, Y., Hu, M., Hu, J. & Liu, X. Clinical Significance and Oncogenic Function of NR1H4 in Clear Cell Renal Cell Carcinoma. **BMC Cancer 22,** 995. ISSN: 1471-2407. (2024) (Sept. 2022).

202. Demicco, E. G. **et al.** EXTENSIVE SURVEY OF STAT6 EXPRESSION IN A LARGE SERIES OF MESENCHYMAL TUMORS. **American journal of clinical pathology 143,** 672–682. ISSN: 0002-9173. (2024) (May 2015).

203. Zhang, Q. **et al.** INSM1 Expression in Mesenchymal Tumors and Its Clinicopathological Significance. **BioMed Research International 2022,** 1580410. ISSN: 2314-6133. (2024) (Dec. 2022).

204. Xuan, C. **et al.** miR-218 Suppresses the Proliferation of Osteosarcoma through Downregulation of E2F2. **Oncology Letters 17,** 571–577. ISSN: 1792-1074. (2024) (Jan. 2019).

205. Li, B., Huang, Q. & Wei, G.-H. The Role of HOX Transcription Factors in Cancer Predisposition and Progression. **Cancers 11,** 528. ISSN: 2072-6694. (2024) (Apr. 2019).

206. Guo, J., Zhang, T. & Dou, D. Knockdown of HOXB8 Inhibits Tumor Growth and Metastasis by the Inactivation of Wnt/β-Catenin Signaling Pathway in Osteosarcoma. **European Journal of Pharmacology 854,** 22–27. ISSN: 0014-2999. (2024) (July 2019).

207. Zhou, Z. **et al.** Heat Shock Transcription Factor 1 Promotes the Proliferation, Migration and Invasion of Osteosarcoma Cells. **Cell Proliferation 50,** e12346. ISSN: 0960-7722. (2024) (Apr. 2017).

208. Meng, X. **et al.** Silencing of the Long Non-Coding RNA TTN-AS1 Attenuates the Malignant Progression of Osteosarcoma Cells by Regulating the miR-16-1-3p/TFAP4 Axis. **Frontiers in Oncology 11,** 652835. ISSN: 2234-943X. (2024) (June 2021).

209. da Silva, R. A. **et al. HOXA** Cluster Gene Expression during Osteoblast Differentiation Involves Epigenetic Control. **Bone 125,** 74–86. ISSN: 8756-3282. (2024) (Aug. 2019).

210. Avnet, S. **et al.** Acid Microenvironment Promotes Cell Survival of Human Bone Sarcoma through the Activation of cIAP Proteins and NF-κB Pathway. **American Journal of Cancer Research 9,** 1127–1144. ISSN: 2156-6976. (2024) (June 2019).

211. Nasarre, P. **et al.** Overcoming PD-1 Inhibitor Resistance with a Monoclonal Antibody to Secreted Frizzled-Related Protein 2 in Metastatic Osteosarcoma. **Cancers 13,** 2696. ISSN: 2072-6694. (2024) (May 2021).

212. Gong, T., Su, X., Xia, Q., Wang, J. & Kan, S. Expression of NF-κB and PTEN in Osteosarcoma and Its Clinical Significance. **Oncology Letters 14,** 6744–6748. ISSN: 1792-1074. (2024) (Dec. 2017).

213. Mohan, S. & Kesavan, C. T-Cell Factor 7L2 Is a Novel Regulator of Osteoblast Functions That Acts in Part by Modulation of Hypoxia Signaling. **American Journal of Physiology - Endocrinology and Metabolism 322,** E528–E539. ISSN: 0193-1849. (2024) (June 2022).

214. Zhong, Q.-H., Zha, S.-W., Lau, A. T. Y. & Xu, Y.-M. Recent Knowledge of NFATc4 in Oncogenesis and Cancer Prognosis. **Cancer Cell International 22,** 212. ISSN: 1475-2867. (2024) (June 2022).

215. Zhang, L. **et al.** KLF8 Promotes Cancer Stem Cell-like Phenotypes in Osteosarcoma through miR-429-SOX2 Signaling. **Neoplasma 67,** 519–527. ISSN: 0028-2685 (May 2020).

Appendix A

Supplement: Virtual Spike-In

A.0.1 Analysis of Sequencing Data

All data for this paper was processed using the Nascent-Flow pipeline (, commit e20c72l), a standardized NextFlow pipeline for the analysis of nascent sequencing data. The general flow of this pipeline is as follows:

(1) Initial quality control is performed using fastQC

(2) Reads are trimmed using BBDuk

(3) Post-trimming quality control is done again with fastQC

(4) Reads are mapped to the reference genome using HISAT2.

(5) Mapped reads are converted into BAM/CRAM files using Samtools.

(6) Bedtools is used to generate bedGraph files from the mapped CRAM files.

For input to the VSI NextFlow pipeline developed for this work, we use the BAM and bedGraph files generated using the previous pipeline. Within the VSI pipeline, we take the following analysis steps:

(1) Genes within samples are filtered down to a single isoform based on the highest transcribed isoform after length normalization.

(2) A single reference isoform list is selected for the entire comparison.

(3) The isoform list / ROI list is filtered by length when requested for the analysis (example: the $3'$ methodology)

(4) featureCounts is used to count reads over the filtered and processed ROIs for all samples

(5) Read counts from each sample are merged into one unified count table

(6) The VSI algorithm is run on the merged count table

A.0.2 Characteristics of $3'$ regions

For our $n = 1198$ genes with suitable length for a 180kb threshold, we find that the median length is 288254bp, the mean length is 371005bp, while the minimum length region is 180201bp and the longest region is 2220164bp.

A.0.3 The variance distribution for the VSI is highly skewed

In our testing, we observe that the typical estimated variance for normalization factors is around 1, corresponding to a 2-fold change. However, this estimated variance is noisy, with the 94% Highest Density Interval (HDI) of the variance lying somewhere between 0.10 and 2 in $\log_2$ transformed space. This means the most likely variance on our normalization estimate is somewhere between 1.1-fold and 4-fold, which is a wide range. Because the estimated variance has a skewed distribution, our variance estimates are shifted to be relatively large because of the long tail in our data.

A.0.4 Analysis of Samples

For the sake of consistency of analysis, samples were only analyzed if they could be processed by our analysis pipeline without error or modification to the pipeline code. A full list of files that we used as input for our pipeline and if/why they were excluded can be found in Supplemental File 2. Samples were processed separately in groups of single-end and paired-end samples using the Nascent-Flow pipeline, then analyzed downstream. Once processed, samples were then grouped by experiment and analyzed using the VSI Nextflow pipeline. Two different analyses were performed — one on RefSeq annotated

genes in the dm6 genome and one on RefSeq annotated genes constrained to a 180kb / 60 minute $3'$ threshold. 60 minutes was selected among all experiments for the sake of consistency as well as to test the assumptions made in the choice of the $3'$ invariant region. A summary of the data sets used for this study can be found in Table A.1.

Experiment	Cell Type	SRP Project	GSE Accession Number
Aoi 2020[75]	DLD-1	SRP247346	GSE144786
Barbieri 2020[76]	THP-1	SRP242477	GSE143844
Barbieri 2020	HeLa	SRP242477	GSE143844
Barbieri 2020	iPSC	SRP242477	GSE143844
Birkenheuer 2018[77]	HEp2	SRP121447	GSE106126
Birkenheuer 2020[78]	HEp2	SRP193891	GSE130342
Dukler 2017[79]	K562	SRP102240	GSE96869
Fan 2020[80]	MV4-11	SRP234556	GSE141377
Jaeger 2020[81]	KBM7	SRP227189	GSE139468
Leroy 2019[82]	Myoblast	SRP153901	GSE117155
Liu 2021[83]	CD34+ Erythoblast	SRP261462	GSE150530
Rao 2017[84]	HCT116	SRP124968	GSE104334
Santoriello 2020[85]	A375	SRP188036	GSE128086
Sendinc 2019[86]	MEL624	SRP170033	GSE122803
Sendinc 2019	MEL624	SRP170034	GSE122803
Sendinc 2019	MEL624	SRP170035	GSE122803
Sendinc 2019	MEL624	SRP170036	GSE122803
Sendinc 2019	MEL624	SRP170037	GSE122803
Sendinc 2019	MEL624	SRP170038	GSE122803
Sendinc 2019	MEL624	SRP170039	GSE122803
Sendinc 2019	MEL624	SRP170040	GSE122803
Sendinc 2019	MEL624	SRP170041	GSE122803
Takahashi 2020[87]	HEK293T	SRP164752	GSE121024
Vihervaara 2021[88]	K562	SRP187541	GSE127844
Vihervaara 2021	K562	SRP187541	GSE154746

Table A.1: Accession Numbers for Analyzed Projects

A.0.5 MCMC convergence and autocorrelation

As discussed in the main text (See Figure A.5), the convergence of model parameters is dependent on the sampler used. The NUTS sampler used for the continuous variables is more efficient in exploring the sampling space than Metropolis-Hastings, which is only used to sample from the two discrete variables. Convergence of the discrete distributions is the limiting step in model convergence.

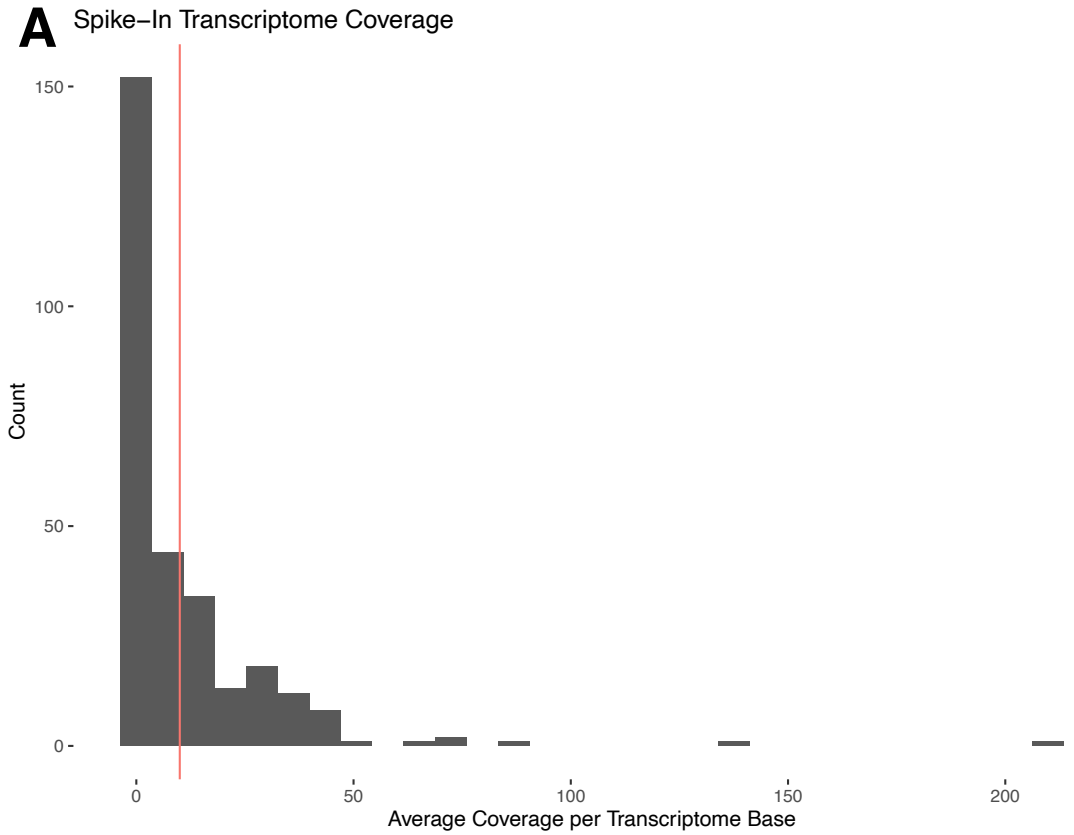


Figure A.1: **Depth of spike-in sequencing across data sets from the literature.** Red vertical line is at 10X transcriptome coverage of Drosophilia. Most data sets are under-sequenced relative to a minimum threshold of 10x coverage for use in normalization.

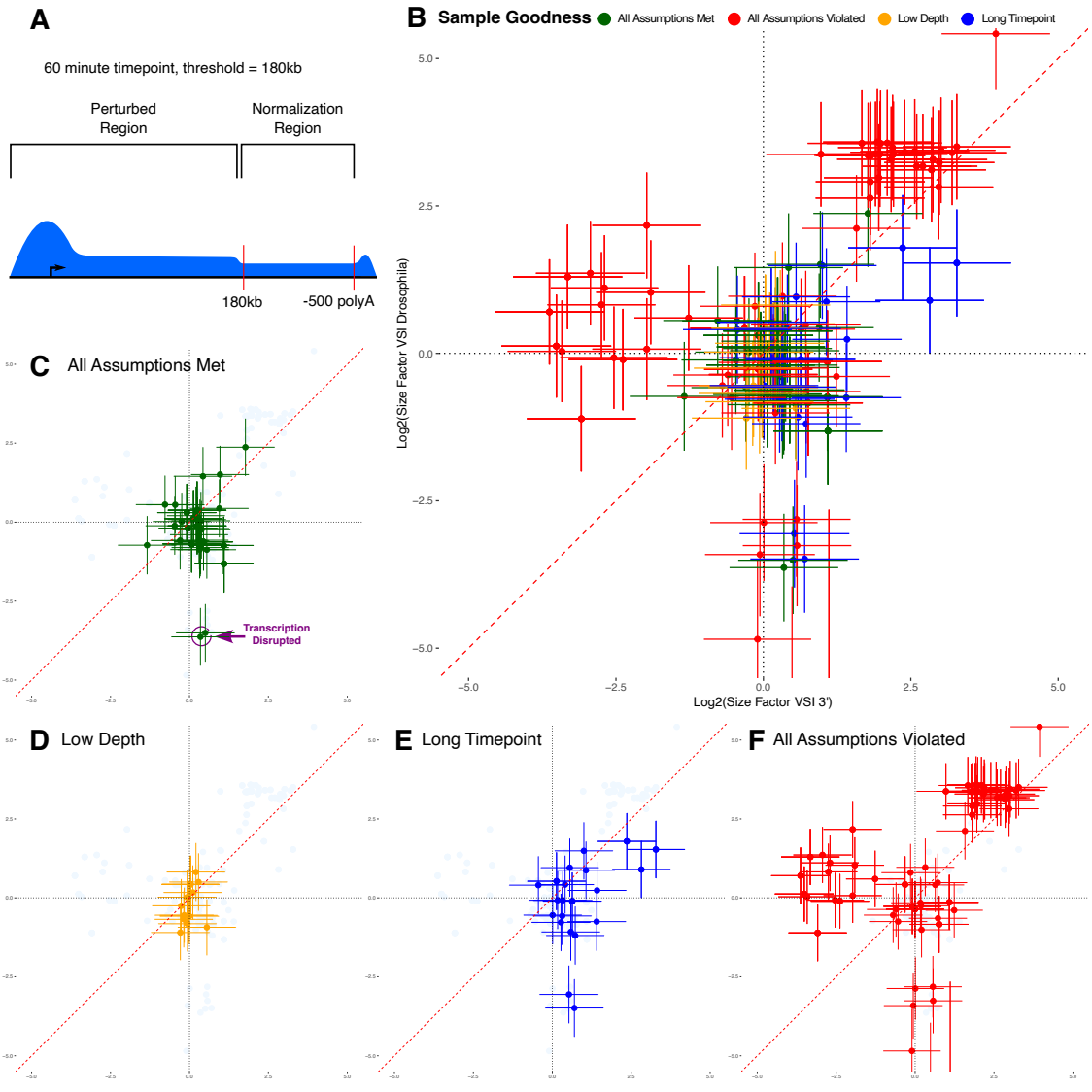


Figure A.2: **Comparison of internal 3′ normalization to exogenous spike-ins**. This figure is the same as Figure 3 but each panel B-F now with error bars.

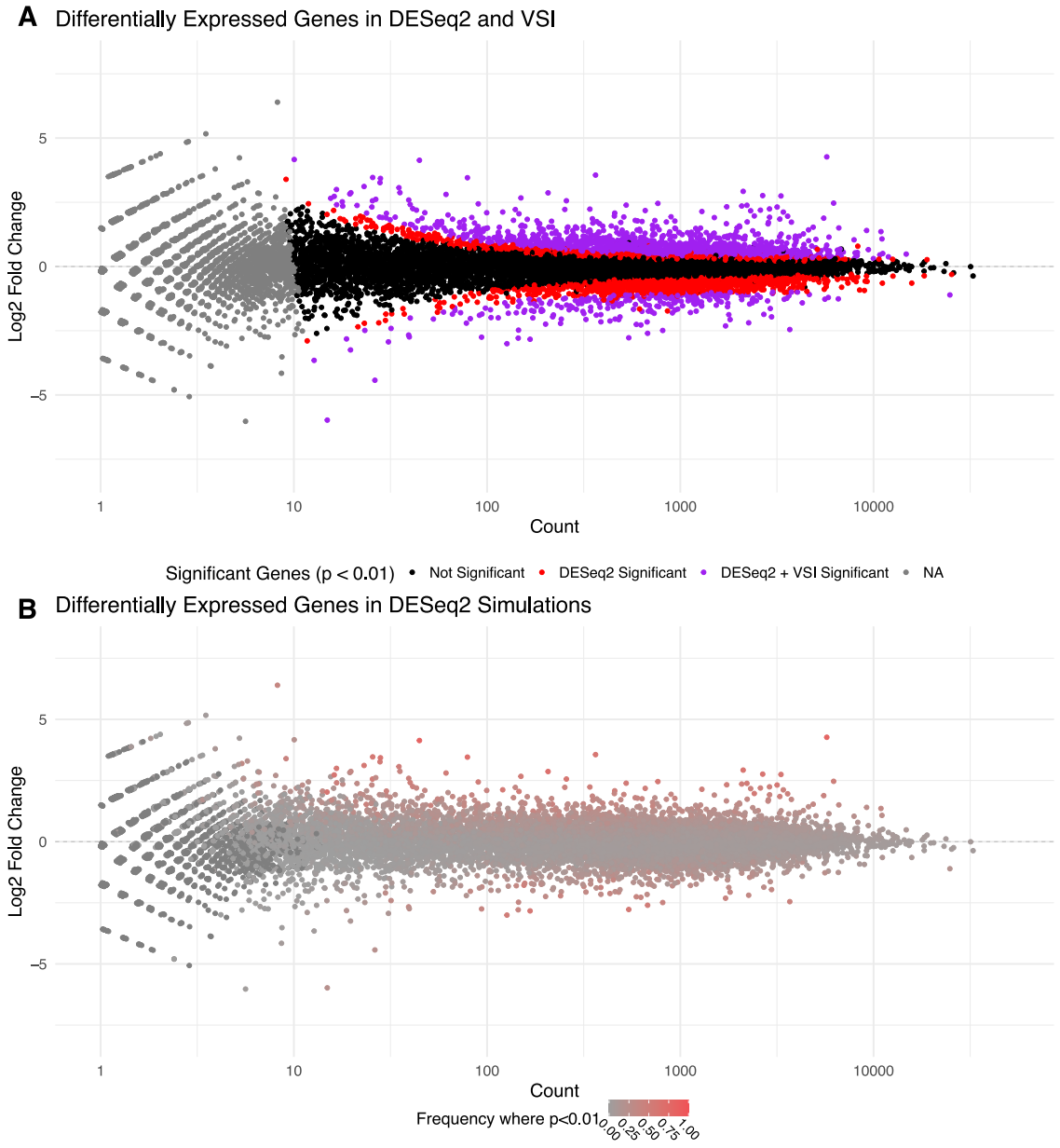


Figure A.3: The same as Figure 5, but using the 40 minute for comparison instead of the 60 minute timepoint.

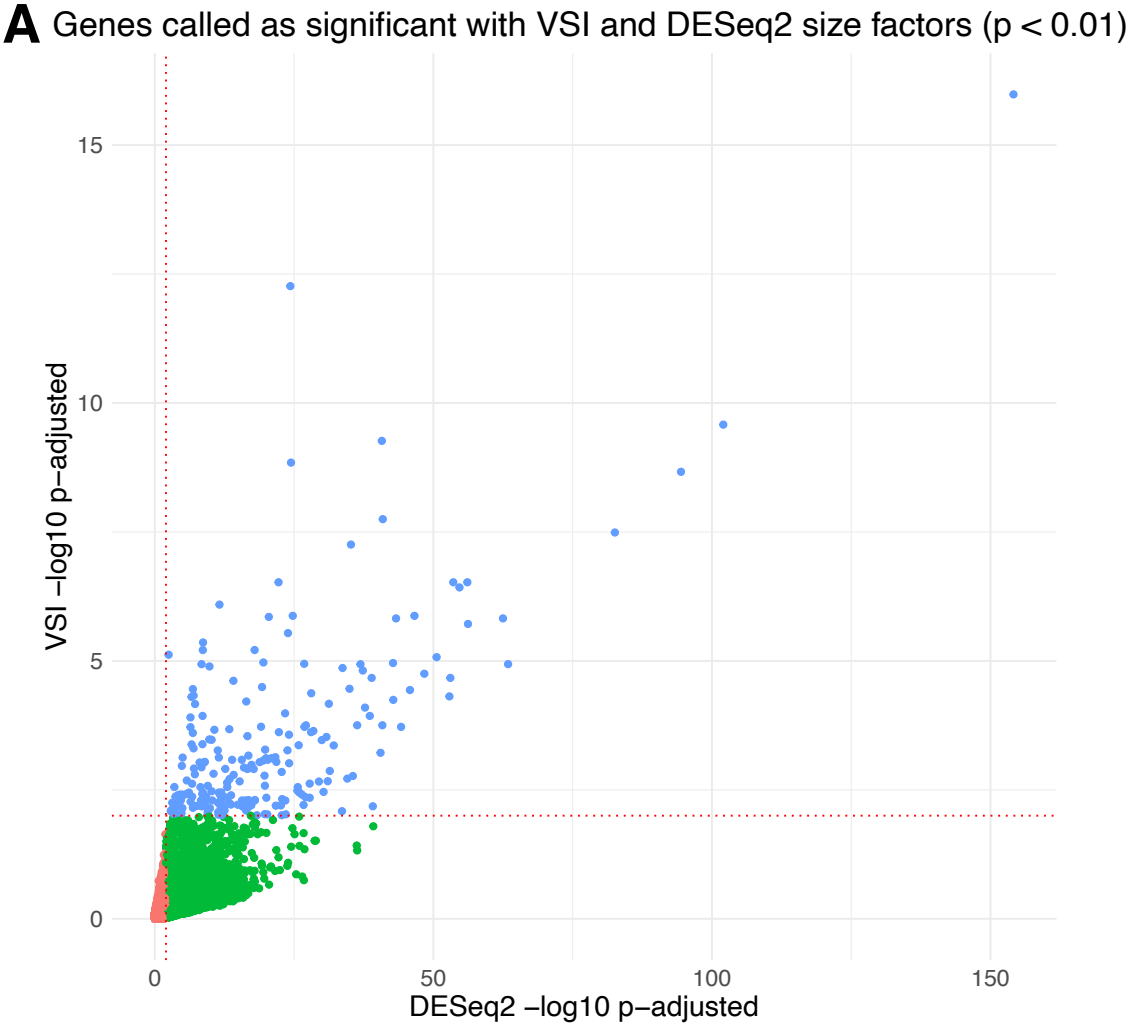


Figure A.4: **Comparison of DESeq2 to VSI normalization.** In the DESeq2 results shown in Figure 5A, using size factors estimated from the VSI 3′ internal approach instead of those naively estimated by DESeq2. Notably, for this data set, no stricter p-value cutoff on DEseq2 results (x-axis) would produce the same gene set as called by the VSI method (y-axis), yet VSI is a strict subset of the DESeq2 calls.

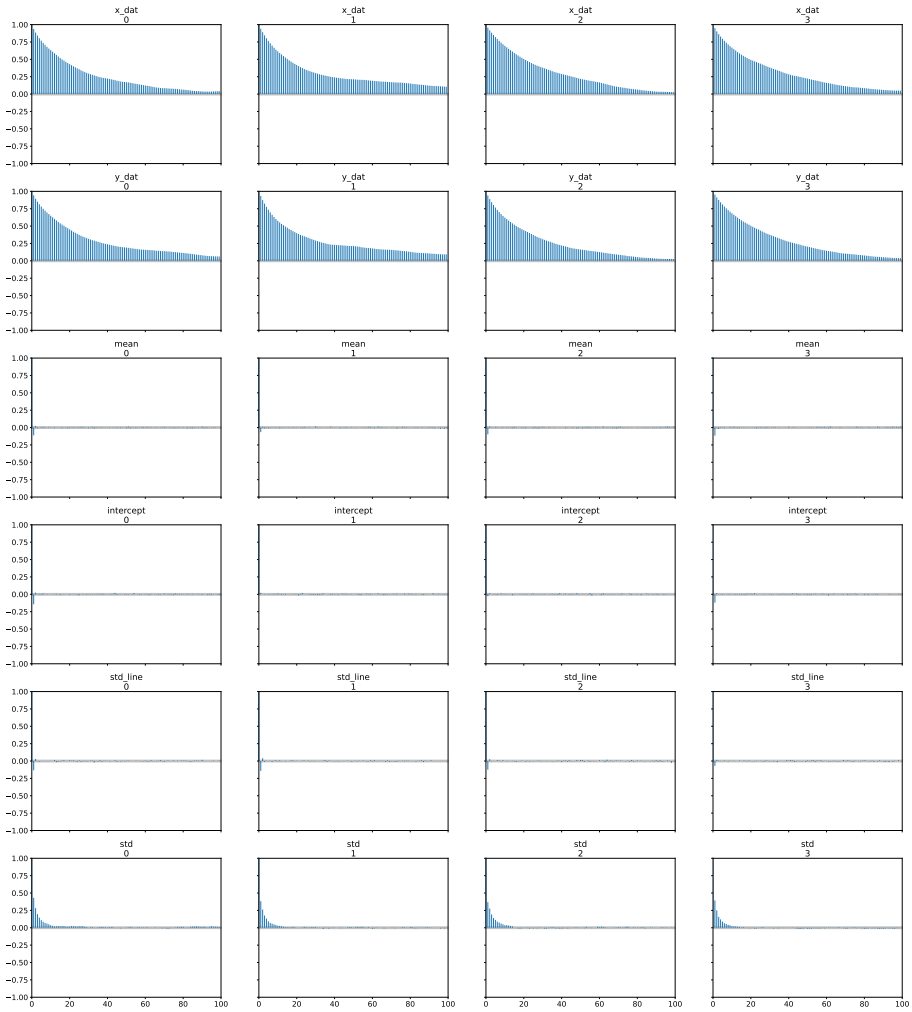


Figure A.5: Parameter autocorrelation across 4 runs (chains) (columns, left to right) for a single pairwise comparison (SRR5364303 vs SRR5364304) corresponding to the Dukler 2017[79] 0 minute replicates 1 and 2. Estimates of the negative binomial distributions (top half of plot) show autocorrelation for much longer than estimates of the mean and variance (bottom half of plot). Some degree of autocorrelation is expected when fitting count data under the assumption that the 3′ region is invariant, as the choice of an invariant region implies that values should be correlated across samples. Additionally, some degree of autocorrelation is likely unavoidable, because in count data both autocorrelation and overdispersion can have the same causes[179]

Appendix B

Supplement: Deconvolution

No supplemental material was generated for this work at publication, but code for the project is available at `https://github.com/Dowell-Lab/DeconvolutionNascent`.

Appendix C

Supplement: Encoder

C.0.1 Datasets Used

All data from this experiment was previously produced and published under GSE86222, using the following SRRs: SRR4090098, SRR4090099, SRR4090102, SRR4090103, SRR4090106, and SRR4090107. Bidirectional regions of interest were identified using previously published methodologies[61].

C.0.2 Statistical Methodology

Code for this project is available at . Training and evaluation for all models was performed using Pytorch 2.3 and Captum 0.7.

Our base autoencoder model uses a linear projection with sinusoidal positional encoding[144] followed by a transformer encoder with 2 heads, 6 layers, and hidden dimension of 4. The decoder model uses two layers, one that maintains the hidden dimension followed by one that projects the hidden dimension back to input dimension, with ReLU activation between these two layers. After the decoder, the reconstruction is normalized using Softmax per-base on the sequence prediction and then normalizing predicted reads to sum to one across each strand. Loss is calculated sum of the KL Divergence of the reconstructed sequence and the RMS Loss of the reconstructed reads. During training, the gradients for each loss are normalized against their magnitude to allow for balanced learning of each task without requiring manual tuning of a hyperparameter[180] Using our 256x6 input regions (one-hot encoded sequence and reads), training is performed using minibatches of 512 randomly shuffled inputs with the AdamW optimizer[181] for the deep model used. During training, 50% of inputs are randomly flipped

along the center of the region and reads from each strand are reversed, to account for bidirectional binding and initiation at enhancer loci[29]. The model is trained for 30 epochs using a one cycle cosine learning rate scheduler with a maximum learning rate of 0.001 to allow for effective warmup of the transformer layers.

The SAE model is trained to reconstruct the encoder layer outputs (latents) of the autoencoder model using a single layer architecture[139]. This architecture uses a single linear encoder layer of hidden dimension size ($n = 512$ here) followed by ReLU activation and a decoder layer of the same size to reconstruct the input latents with MSE loss. We train this model for 360 epochs using the same training data as the base autoencoder model, performing random re-initialization of dead neurons every epoch. We again use AdamW with a cosine learning rate scheduler and a maximal learning rate of 0.001.

For downstream analysis, the activation of each SAE neuron for all inputs in the dataset is calculated. Empriically, these activations form a normal distribution, so to select most-activating regions, we fit a normal distribution each neuron's activations and take all regions with activations more than three standard deviations above the mean. This yields in 800-1200 top-activating regions per SAE neuron. Using this set of top-activating regions for each neuron, we then perform analysis to extract cell-type specific and regulatory information.

First, we test if each SAE neuron's top activating regions are sampled preferentially from one cell type. To do so, we perform a Bonferroni corrected χ-squared test on each neuron's top activating regions, testing whether those regions were sampled within random expectation ($p < 0.05$). If they are not, we attribute that neuron to the most highly sampled cell-type.

Next, we use traditional motif identification approaches[142] to look for enrichment of motifs within the top-activating sequences for each neuron. Given the sequence bias in our input regions, we identify a set of ubiquitously present GC rich transcription factor motifs (SP/KLF) which are difficult to separate from the pattern in genetic background. To counteract this, for each set of cell-type attributed SAE neurons, we take the set of motifs identified within that neuron that do not occur in any neurons outside that cell's neurons. This provides a first list of potential cell-type specific motifs.

Finally, we use attribution based methods to understand the features driving the activation of each

SAE neuron. Using the Neuron GradientSHAP[145] implementation provided by the Captum library, we calculate per-feature attributions for each SAE neuron's top activating regions. Using these attributions, we then run TF-MODISCO[138] to identify motifs driving underlying model attributions. In initial testing, we observed that TF-MODISCO showed remarkably poor performance on our model's attributions, due to the nonlinear scale of the repetitive patterns observed in our attributions (Figure 4.3C). To compensate against this for TF-MODISCO analysis, we perform a position-aware transform to locally smooth these repetitive patterns. We implement this position-aware transform by dividing each feature of the attribution by the 95th percentile value of the surrounding 16 base pairs, maintaining local relative importance while smoothing global importance values at the same time. Despite this procedure, some neurons still show the same failure more of TF-MODISCO, suggesting that this model architecture is poorly suited to MODISCO's motif discovery approach given the model's learned periodicity. Using the motifs discovered by TF-MODISCO, we then use TOMTOM[182] to match these motifs to known regulatory motifs[183] With the TF-MODISCO identified motifs, we perform the same sub-setting procedure as with the previous analysis to identify potentially–attributable cell-type specific motifs. However, these identified motifs require caution given the observed difficulties with TF-MODISCO on this particular model architecture.

C.0.3 Filtering and Selection of Features

C.0.4 Alternative Approaches

Over the course of this work, we tried a lot of things that didn't end up working out as well as hoped.

- Models in the style of [134] or [135] were the first thing we tried but not well suited to this problem. While the BPNet architecture is effective for the problem of taking sequence and generating reads, it implicitly relies on a one-to-one relationship that is not present in our data. The same region in transcriptional data may be used by multiple distinct cell types for different transcriptional patterns, meaning that no model, no matter how rich, can generalize sequence to transcription across cell types unless that region is transcriptionally invariant across cell types.

- Contrastive learning between sequence and transcription in the style of CLIP sometimes worked

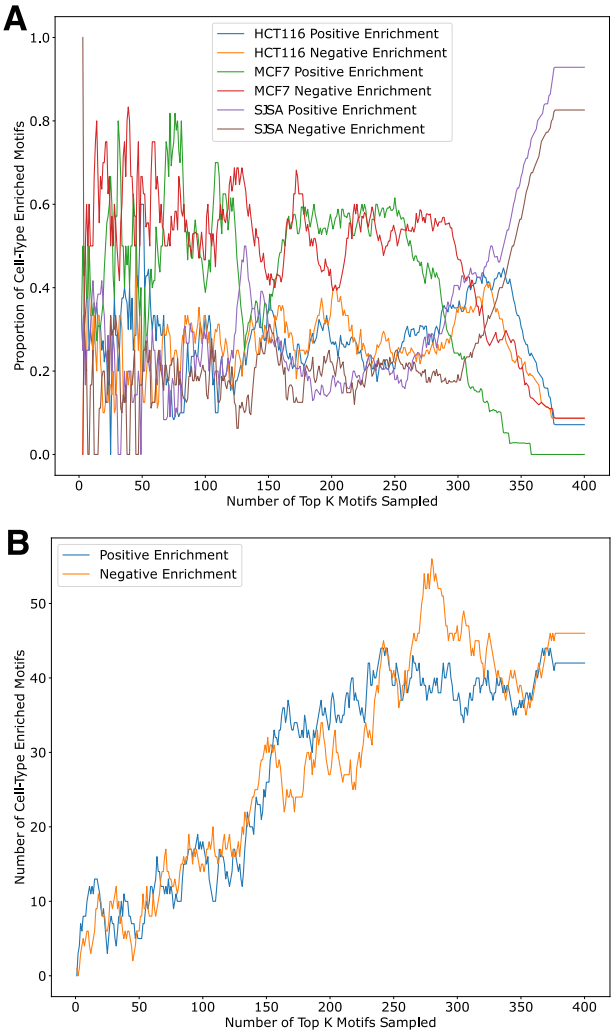


Figure C.1: **A:** Using top-k sampling for determination of features unique to cell-types provides the best overall results, but gives a non-obvious hyperparameter question of the ideal k required for effective sampling. For example, here $k = 65 - -75$ gives the set that by inspection appears to be the most "cell-type-appropriate" in terms of unique TFs, minimzing the effect of disproportionate pull from the SJSA cells showing a higher number of enriched neurons. **B:** Increased top-k sample number increases the number of discovered motifs, but as seen in **A**, this becomes imbalanced for large k values, for which the discovered TFs begin to have a bias for the neuron with the largest number of discovered neurons.

Cell Type	Enriched Transcription Factors
Positively Enriched	
HCT116	PDX1[172], RELB[173], ZBT7A[184]
MCF7	N/A
SJSA	AP2B, PBX3, REL, ZN329, RXRG, NKX61, ZN528[185], PBX1, ZN350, PBX2, TF65, MEIS1, GRHL2[186], GCR, HIC1, ZN490, HLF, OTX2, PRGR, NR4A1[187], ZBT14, ZN317[188], ZN708, HNF4G, ISL1, SRF, SOX9, HNF4A, MYB[176], ATOH1, ZNF8, ZN547, KAISO, NKX32, ZNF18, NDF1, MBD2, TGIF1[189], DBP[190]
Negatively Enriched	
HCT116	NFAC2[191], ZNF8[192], GATA1[193], HXB4[174]
MCF7	MEIS2[194], ZN329, HINFP[195]
SJSA	ZN667[196, 197], HXB13[198], PBX3[199], REL[200],NR1H4[201], STAT6[202], INSM1[203], NFIA, GRHL2[186], PDX1, E2F2[204], TGIF1[189], HXC9[205], ZN490, HXB8[206], ZN680, NDF2, HNF4G, ISL1, ZN382, HSF1[207], TFAP4[208], HXA1[209], HNF4A, RELB[210], PEBB, NFAC3[211], NFKB1[212], ZN274, TF7L2[213], ZN547, ZNF18, NDF1, NFAC4[214], MBD2, NKX25, KLF8[215], NOBOX

Table C.1: Table of unfiltered differentially enriched transcription factors for Dataset A, when considering all motifs for non-intersecting motifs between sets and not just those enriched by a χ-squared test. We observe that a substantially larger set of motifs is discovered, but that these motifs appear to have a looser relationship to the cell types of interest. Citations are provided that link motifs to either (the cell type of interest) or a different cell type or cancer (in the case of non-matching negative set motifs). This suggests that the process of filtering previously described is necessary to remove "noisy" motifs from the set of all motifs found by SEA. Our set of input regions is biased towards TF binding sites and transcriptional activity, meaning that short and GC-rich motifs will be found as enriched with high frequency.

to generate useful embeddings, but showed variable convergence characteristics and were highly sensitive to initialization. Additionally, the model size required and internal dimensionality made it challenging to work with the resultant embedding, while giving similar results to the work shown here.

- Different cell types appear to have different "strengths" in terms of how strong their cell type identity is relative to other cell types. For example, in our comparison with HCT116/MCF7/SJSA cells, SJSA cells are notably for TREs over other cell types despite using the same number of input samples and similar amounts of data.

- One low hanging idea that we spent a substantial amount of time on in the early phases of this project was the implementation of a classifier head on the latent output of the transformer encoder in the reconstruction model. Unfortunately, in our testing across a wide range of models (both autoencoder and contrastive), we observed that it was extremely difficult to train an added classifier head (or a classifier alone as the output of the model). This is a useful result – it indicates that, as expected, cellular identity on a global basis cannot be predicted from a single region or a small set of regions. This is the behavior we expect to see in the case of combinatorial control of transcriptional regulation.

C.0.5 Soft Binding Syntax?

By using a bottlenecked autoencoder in the base model, we learn a compressed representation of our input regions of interest. Curiously, it appears that across regions our model learns "soft" representations of certain positions more preferentially than others.

C.0.6 Other Model Designs

The approach in our model architecture described here appears to be relatively straightforward, but is the ultimate convergence of experimentation with a variety of model approaches and architectures that did not work well. Initially, we sought to solve this problem using a convolutional network in the style

of BPNet – that is to say, using sequence to predict reads. However, initial testing revealed poor model convergence even with a deeper convolutional architecture than used in BPNet, which is to be expected – our dataset contains replicated regions whose sequence is identical but with differences in transcription. Given this limitation, we subsequently tested a contrastive approach instead, training models to learn whether a sequence and a set of transcription signals were sourced from the same sample or not. This approach performed well in many training runs, giving similar results to the ones shown here, but showed poor consistency in reproducibility across runs. Given our desire to develop a model that can be applied flexibly and reliably to new input data, we did not select this approach due to its poor reproducibility. With the limitations of these two prior approaches considered, we instead chose to use an imbalanced autoencoder (deep encoder, shallow decoder) architecture to learn a latent representation of our input data from which we could extract interpretable features. In testing our autoencoder based models, our first attempts used deep convolutional networks, as implemented in prior work[134, 135]. However, we found that these models performed substantially worse at the reconstruction task than using a transformer for the encoder block. Consequently, and given recent research in interpretability based on transformer models, we opted for the use of a transformer model. Theoretically, we believe that transformer models can do a better job of simultaneously attending to all positions in the sequence/transcription space compared to a convolutional encoder network.

C.0.7 Choice of Data Type

As discussed in the main text, our preferred data type for exploring this modeling approach must be both quantitative and capture some sort of first-order regulatory effect. Without data for which differences in signal reflect some true differential biological behavior, it is unlikely that a differential approach will succeed. In addition to this, a protocol that provides a first-order measure of a biological phenomena is of interest, as these steps are more likely to involve immediate regulatory responses. Below, subsection C.0.7 provides a brief summary of which of these requirements each data type satisfies, with additional discussion of each data type afterwards.

Data Type	Quantitative	First-Order
ChIP-seq	No	Yes
ATAC-seq	No	Yes
RNA-seq	Yes	No
Immature-RNA Nascent	Yes	Yes (caveated)
Run-On Nascent	Yes	Yes

- ChIP-seq is ostensibly a good choice at first glance - it captures direct binding events. However, for our approach here, there are two issues. First, quantitation of ChIP is challenging and data is only semi-quantitative. For example, there are a number of blacklisted regions [cite/expand] for which sequence characteristics and/or technical artifacts render those regions falsely enriched across experiments. Additionally, differential binding of a single TF is a useful thing to learn, but limits experiments and modeling to a single TF at a time, rather than looking at global regulatory behavior.

- ATAC-seq is another potentially good data source for this modeling approach, but again, it is only a semi-quantitative protocol. In many ATAC-seq experiments, a substantial amount of sample variation is explained by sample handling alone [cite], complicating quantitation of the resultant data. Additionally, the question of whether differential peak depth between samples is a result of sequencing depth or from a true change in chromatin accessibility remains a difficult question to answer.

- RNA-seq data is, compared to the aforementioned data types, quantitative. However, in the case of the differential approach suggested here, RNA-seq has two substantial issues. First, it is not a first-order or direct measure of an active biological process. Poly-A enrichment RNA-seq (and to a lesser extend ribosomal depletion RNA-seq) captures the steady-state of mature RNA that exists in a cell. This makes associating RNA-seq peaks to an active regulatory process more challenging since differential behavior must be measured against the background of all steady-state RNA in the cell. Second, the choice of regions of interest for a differential method like this is non-

obvious. Regions from protein-coding genes are well characterized, and because only mature RNA is typically measured, the amount of informative intergenic / noncoding RNA available is minimal.

- Immature-RNA Nascent Sequencing protocols (meaning any nascent sequencing protocol that does not capture only RNA directly engaged with RNA Polymerase II) have the same advantages as our selected run-on sequencing protocols, but because of protocol specifics are less suited for the differential approach discussed here. Examples of this class of protocol include NET-seq and TT-seq. In particular, unlike with run-on protocols, the labeling strategies in these protocols are more likely to incorporate mature RNA and not capture the immediate first-order effects of transcription.

- Run-on Nascent RNA Sequencing protocols represent, in our assessment, the best data type for testing the differential approach used here. While other nascent protocols will likely work with the approach described here, run-on nascent sequencing protocols capture directly engaged RNA Polymerase II and many of these transcripts are associated with transcription factor binding sites[136]. This provides a clean first-order readout of transcription that is directly associated with important regulatory behavior, making this data well suited to the differential question explored here.

C.0.8 TF-MODISCO Results

Type	Unique TFs
HCT116	ERR3, SMCA1, ETV2, PO2F2, FOXC1, PIT1, NFKB1, GLI3, ZIC3, ZN260, MYC, SIX2, FOSL1, DLX3, GATA2, TFAP4, RUNX1, SOX10, NR1H3, SOX17, ESR1, PO3F1, ZN329, RORA, ZN274, USF1, ZN143, NR1H4, NFIC, NFYA, TFEB, HXB4, NOBOX, ZBT18, ETS2, NR2C2, HXA13, IKZF1, ZNF18, THA11, PEBB, RXRB, OSR2, RFX3, NR5A2, INSM1, SRF, ZNF76, PO2F1, FOSL2, ZN436, FEV, NR1I2, ZNF8, OZF, HXB13, STF1, TF7L2, NFE2, TYY1, SOX3, RARG, ETV4, GFI1, GATA4, SUH, TBP, TBX21, ZFP42, ETV1, COT2, RFX2, PDX1, ZN335, BACH1, JUND, MAX, RUNX2, ZN816, BACH2, CRX, ESR2, ARNT, MEIS1, NDF1, ELF2, HXB8
MCF7	HSF2, RELB, SMAD4, ZN317, IRF7, THA, GCR, HSF1, SOX9
SJSA	OTX2, BATF, FOXH1, EHF, MAFG, AP2A, GRHL2, PBX3, REST, CLOCK, FOXP2, TF7L1, TEAD1, ZN768, PKNX1, ELF3, BCL6, PPARA, CUX1, GFI1B, PO3F2
Background	KAISO, ZN490, E2F2, HNF4A, CEBPB, JUN, ELK1, STAT4, ATF4, NF2L1, TGIF1, P53, ZN549, ATF3, OVOL1, SNAI2, ELF1, ZN121, HNF4G, JUNB, ERR1, ZN502, ZN384, TF65, HXA1, ISL1, ZN257, BRAC, P63, IRF9, E2F5, ELK4, ZN582, MAF, NFIA, MTF1, HINFP, MEIS2, MYB, STAT3, ATF2, CEBPG, TFE3, MITF, ZN134, ZN667, RXRG, HIF1A, ZN382, DBP, STA5A, HLF, E2F3, STAT6, NFYB, MAFF, NR4A1, BMAL1, CREB1, FOXO3, NR4A2, ZN708, SMCA5, ZN140, FOXI1, FOSB, MAFB, ZN586, SMAD2, ZN528, ZN322, NFYC, NKX21, ATF1, AHR, ERR2, BHE40, TEAD4, COE1, NKX28, FOXM1, CREM, NF2L2, CEBPA, CEBPE, FOS, P73, ZKSC1, PAX6, CEBPD, EPAS1

Table C.2: Unique TFs for each cell type and background as identified using TF-MODISCO and smoothed DeepSHAP attributions, paired with TOMTOM for identification of motif matches.