# COMPUTATIONAL METHODS FOR INFERRING TRANSCRIPTOME DYNAMICS

Joshua D. Welch

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computer Science.

Chapel Hill
2017

Approved by:

Jan F. Prins

Alexander J. Hartemink

Corbin D. Jones

William F. Marzluff

Leonard McMillan

Jeremy Purvis

## ABSTRACT

Joshua D. Welch: Computational Methods for Inferring Transcriptome Dynamics
(Under the direction of Jan F. Prins)

The sequencing of the human genome paved the way for a new type of medicine, in which a molecular-level, cell-by-cell understanding of the genomic control system informs diagnosis and treatment. A key experimental approach for achieving such understanding is measuring gene expression dynamics across a range of cell types and biological conditions. The raw outputs of these experiments are millions of short DNA sequences, and computational methods are required to draw scientific conclusions from such experimental data.

In this dissertation, I present computational methods to address some of the challenges involved in inferring dynamic transcriptome changes. My work focuses two types of challenges: (1) discovering important biological variation within a population of single cells and (2) robustly extracting information from sequencing reads.

Three of the methods are designed to identify biologically relevant differences among a heterogenous mixture of cells. SingleSplice uses a statistical model to detect true biological variation in alternative splicing within a population of single cells. SLICER elucidates transcriptome changes during a sequential biological process by positing the process as a nonlinear manifold embedded in high-dimensional gene expression space. MATCHER uses manifold alignment to infer what multiple types of single cell measurements obtained from different individual cells would look like if they were performed simultaneously on the same cell. These methods gave insight into several important biological systems, including embryonic stem cells and cardiac fibroblasts undergoing reprogramming.

To enable study of the pseudogene ceRNA effect, I developed a computational method for robustly computing pseudogene expression levels in the presence of high sequence similarity that confounds sequencing read alignment. AppEnD, an algorithm for detecting untemplated additions, allowed the study of transcript modifications during RNA degradation.

To Jeanna Welch, my beautiful bride.

## ACKNOWLEDGEMENTS

I am thankful for the many people in my life–family, friends, and colleagues–who helped and supported me in many ways during my doctoral studies. I would like to acknowledge:

My lovely bride Jeanna, whose love and support changed my life–and graduate school experience–dramatically for the better. Before I married Jeanna, I had no first author publications from almost three years of graduate school. In the 2 years after we got married, I published 6 papers. Coincidence? I think not!

My parents Lonnie and Mary Welch. My mother educated me at home until I graduated from high school. I believe that Mom prepared me remarkably well for doctoral studies, which require the same sort of independence, intrinsic motivation, and perseverance that I learned at home. My father passed on his love for computer science and fascination with DNA. Dad also advised me in numerous ways as I navigated the process of applying for graduate school, getting a Ph.D., and finding an academic job. Both of my parents taught me that I could do anything as long as I was willing to work hard enough, and demonstrated by their example that this principle is true.

My advisor, Jan Prins, who gave me freedom to pursue my own ideas and provided invaluable support and encouragement. Jan always helped me to slow down, think things through step-by-step, and flesh out my ideas.

Bill Marzluff, who taught me most of what I know about histones, RNA biology, and biochemistry. Bill really took me under his wing and was like a second advisor to me, inviting me to his lab meetings, teaching me how to mentor undergraduate researchers, and helping me during the job search process.

Alex Hartemink, who taught a very inspiring and insightful class on computational genomics at Duke, gave very helpful suggestions during the development of SLICER and MATCHER, and was a great resource during the job search process.

Corbin Jones connected me with the biomedical side of campus as soon as I arrived at UNC, helped me to navigate an unexpectedly long and painful review process with the microrafts paper, and served as the co-mentor on my NIH F31 fellowship.

Praveen Sethupathy and Chuck Perou graciously helped me to complete my first paper in graduate school, providing invaluable insights into the biological interpretation of my computational results.

Jen Jen Yeh, Lindsay Williams, Matt DiSalvo, Nancy Allbritton, and Chris Sims introduced me to single cell genomics through our collaboration on the microrafts project.

Jeremy Purvis provided invaluable insights into the biology of stem cells and differentiation during the SLICER project and also helped coach me during the job search process.

My roommates: Jeremy Erickson, Alan Tubbs, Andy Garrison, Josh Fuchs, and Alex Main. The level of nerdiness and geekiness in our house made The Big Bang Theory TV show look normal, but I can't imagine a better set of roommates. You helped me keep perspective, maintain sanity, and take myself less seriously–especially during some stressful times early in graduate school.

My friends from Christ Community Church, especially my unofficial adoptive parents Brent and Dana Senior, John and Staci Reynolds, and Rick and Sara Hawkes.

My friends from the Graduate Intervarsity Fellowship at UNC: Hank Tarlton, Chris Turlington, Tyler Farnsworth, Rachael Fuchs, Megan Schertzer, Mimi Huang, Haley Vaseghi, Daniel Luckett, Andrew Chi, and many others.

My heavenly Father, who gave me life, breath, and everything else and marked out the times and places of my life so that I would seek and find Him.

## TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| ATAC-seq | Assay for Transposase-Accessible Chromatin with high throughput sequencing |
| DNA | Deoxyribonucleic acid |
| ChIP-seq | Chromatin immunoprecipitation sequencing |
| GPLVM | Gaussian Process Latent Variable Model |
| LLE | Locally Linear Embedding |
| PCA | Principal Component Analysis |

# CHAPTER 1
# Introduction

DNA is the blueprint for life. For over a half century, scientists have understood that an organism's DNA sequence functions as a control system that directs the intricate unfolding of its life processes through space and time (Hershey and Chase, 1952). But how does this happen? The secret of life lies not just in the sum total of the DNA sequence itself, but which portions of the genome are used in a particular cell. Although each cell in an organism has nearly identical DNA sequence, a given cell uses or "expresses" only a subset of its full gene complement, and it is the set of genes expressed within a particular cell that determine its properties. This is a fundamental concept in molecular biology, but the specific genes that underlie many important cellular properties are unknown.

The sequencing of the human genome (Lander et al., 2001) paved the way for a new type of medicine, in which diagnosis and treatment are informed by a molecular-level, cell-by-cell understanding of the genomic control system, how it functions under normal circumstances, how it goes awry in various diseases, and how genetic differences make each patient unique. Much work remains before genomic medicine reaches its full potential (Green and Guyer, 2011). A crucial part of realizing this potential is understanding the roles of genes in the diverse properties of human cells. But it is not enough to know which genes play roles in specifying cellular properties–we must also understand their regulation if we are to fix the genomic control system when it breaks and intervene in other ways.

Cells within the human body exhibit incredibly diverse forms and functions, varying based on location in the body, over the course of development, in response to stimuli, and between healthy and disease states. A whole host of important biomedical goals–understanding how the brain works, curing cancer, regenerating damaged tissue, learning how the human body develops from a single cell–all require understanding changes in gene expression. Examples of the types of questions related to gene expression that biomedical scientists ask include: What gene expression changes cause a specific disease? How do cells in a human embryo turn just the right genes on and off at just the right times to create a certain human organ? What genes do we need to modify to enable regeneration within a tissue that does not normally regenerate? Which genes enable a cell

to perform a specific molecular function such as secretion? How do genes specify the distinctive shapes, sizes, and morphologies of certain cell types? What genes are involved in how a given cell type responds to various stimuli? How do genes specify the precise 3D positions of individual cells within a spatially organized tissue?

One way that scientists explore these questions is to compare gene expression differences across different cell types, tissues, spatial locations, dynamic cell states, genetic perburbations, stages of development, and diseases. Characterizing the dynamics of gene expression across these comparisons allows researchers to begin to tease apart the roles of various genes in the genomic control system. The chemical properties of nucleic acids enable the measurement of gene expression by high-throughput sequencing using the RNA content, or transcriptome, of a cell as a proxy for gene expression. The raw outputs of these experiments measuring gene expression are millions of short DNA sequences, and computational methods are required to process, explore, interpret, and draw scientific conclusions from such experimental data.

Developing computational approaches for drawing conclusions from gene expression data is thus a crucial step in scientific progress toward genomic medicine. However, the task of developing such computational approaches is quite difficult. The data are noisy, subject to a number of biases and confounding factors, and high-dimensional. Algorithms are needed to align sequences to the genome and compute information about transcript abundance and other properties. The process of discovering interesting and significant changes across biological conditions requires sophisticated statistical and machine learning techniques. Often, the person performing the gene expression experiment seeks to generate hypotheses from the data in addition to confirming prior hypotheses; exploratory and unsupervised methods are important for such investigations.

The most significant biological insights derived from the field of computational biology increasingly occur at the intersection of three areas of expertise. The first concerns the rapidly changing technologies, experimental techniques, and sequencing protocols for measuring the transcriptome. The second type of expertise required is a deep understanding of what is known about important biological systems and what are the salient questions that researchers seek to answer about these systems. And of course, a computational biologist needs to possess breadth and depth in computational methods, including classical algorithmic techniques, statistical modeling, and machine learning.

My work occurs at this intersection of emerging experimental techniques, important biological questions, and computer science. In this dissertation, I present computational methods to address some of the challenges involved in inferring dynamic transcriptome changes from RNA sequencing data. The work presented here touches on a number of different types of tasks involved in turning transcriptome measurements into scientific

discoveries, including alignment and processing of sequencing data, careful treatment of experimental noise, discovery of latent structure underlying the variation among cells, and applications to specific biological systems.

## 1.1 Transcriptional and Post-Transcriptional Regulation of Gene Expression

Every cell within a multicellular organism accomplishes its specialized function through carefully coordinated spatiotemporal gene expression changes (Cooper, 2000). Gene expression is the process of making proteins specified by the genetic code contained in a particular gene. At a high level, the gene expression process involves the creation of an RNA copy of the gene to be expressed through a process called transcription, then the translation of the RNA copy into a protein. The proteins produced in this way carry out molecular functions within the cell. This simple abstraction is called the central dogma of molecular biology. But, as is often the case in molecular biology, reality is much more complicated than this simple model.

Gene expression is a complex, multi-step process (Cooper, 2000). Each step in the process can be regulated, allowing many different outcomes under different conditions. Understanding both the overall outcome of the gene expression process and the regulatory decisions made at each stage in the process is essential to deciphering the relationship between genes and cellular properties.

The process of transcription itself is regulated primarily by proteins called transcription factors, which bind to the genome near genes and either repress or activate transcription (Cooper, 2000). Different combinations of transcription factors bind in different cellular contexts, creating a complex transcriptional regulatory code that helps to specify which genes are transcribed in which cells and how many RNA transcripts are made.

After an RNA molecule is transcribed from a gene, a number of post-transcriptional regulatory mechanisms control the final outcome of the gene expression process. Transcript pieces called introns are removed and the remaining pieces, called exons, are spliced together (Nilsen and Graveley, 2010). The process of splicing can produce multiple distinct combinations of exons, each potentially encoding a different final protein product (Nilsen and Graveley, 2010). A tail consisting of adenine (A) nucleotides, called a poly(A) tail, must be added to the end of the transcript (Tian and Manley, 2016). Both the positions and lengths of poly(A) tails are regulated (Tian and Manley, 2016). At some point after a transcript has been fully processed, once it has served its cellular purpose, the trancript must be removed. Removing transcripts from

the cell is called degradation, and this process is yet another form of post-transcriptional regulation. The addition of uracil (U) nucleotides (Slevin et al., 2014) and the binding of small molecules called microRNAs, which play a role in determining RNA degradation (Filipowicz et al., 2008), are additional examples of post-transcriptional regulation.

Two additional complexities of the gene regulation process are worth mentioning here. First, RNA transcripts can function not only as protein-coding intermediaries, but also as non-coding molecular effectors (Lee, 2012). For example, some RNA transcripts do not encode proteins, but function instead to recruit protein factors to specific genomic locations. Some transcripts function in both protein-coding and non-coding roles, such as binding and sequestering regulatory factors that would otherwise bind elsewhere. Epigenetic regulation represents another layer of complexity in the gene expression process. The chemical and physical properties of the DNA chromosomes play a role in determining which genes are expressed in a given cell. For example, the addition of methyl groups to DNA (Jaenisch and Bird, 2003), chemical modifications of the histone proteins around which the DNA is wrapped (Lawrence et al., 2016), the tightness of DNA packing (Jaenisch and Bird, 2003), and the three dimensional arrangement of chromosomes in the nucleus are all known to contribute to regulation of gene expression (Dekker et al., 2013).

Most experimental measurements of gene expression focus on counting the number of RNA transcripts from each gene in a particular cellular context. Because RNA serves as the intermediate between DNA and proteins in the process of gene expression, these measurements of RNA abundance are often used as proxies for gene expression levels. The fact that RNA nucleotides participate in defined base pairing interactions (i.e., A pairs with U and C pairs with G) makes it possible to identify individual transcripts and determine which genes produced them. This property of sequence complementarity lies at the heart of high-throughput sequencing approaches (see next section). In contrast, identifying and quantifying abundance of proteins, the final products of the gene expression process, is much harder. Generally speaking, experimental measurements of protein abundance use antibodies that target known proteins. High-throughput protein surveys are possible using mass spectrometry, but such experiments still utilize defined protein databases, and it is difficult to get deep coverage. The ability of RNA sequencing experiments to perform unbiased surveys of the identity and quantities of expressed genes, without requiring pre-specified lists of genes, represents a distinct advantage over existing approaches for surveying proteins.

Because of post-transcriptional regulation, RNA abundances do not necessarily reflect protein abundances, and thus knowing both RNA and protein abundances is necessary for a complete picture of the gene expression

process. Nevertheless, knowing the identity and abundance of molecules in the transcriptome is very valuable. As we will see in the next section, transcriptome measurements based on high-throughput sequencing give insight into not only gene expression levels, but also the steps of post-transcriptional regulation.

## 1.2 Measuring the Transcriptome Using High-Throughput Sequencing

The ability to rapidly, cheaply, and accurately determine the nucleotide sequences of many DNA molecules is one of the foundational developments in modern molecular biology. The Human Genome Project (Lander et al., 2001) paved the way for the development of high-throughput DNA sequencing. Since then, high-throughput sequencing has become a ubiquitous "experimental subroutine" that enables numerous types of high-throughput molecular measurements. In particular, converting RNA molecules to DNA molecules and then performing high-throughput sequencing gives a quantitative, genomewide survey of the transcriptome in a certain cellular context (Wang et al., 2009).

The predominant technology for high-throughput sequencing is Illumina's sequencing by synthesis approach. As mentioned in the previous section, the complementary base pairing interactions of nucleic acids are at the heart of most sequencing approaches, including sequencing by synthesis. Sequencing by synthesis works as follows. DNA molecules are extracted by lysing a sample of cells, then broken into small fragments. A set of pre-specified sequences called sequencing adapters are then chemically joined to each DNA fragment. Many copies of the fragments are then made using polymerase chain reaction (PCR). The resulting set of DNA fragments with sequencing adapters attached is called a DNA library. The DNA library is loaded onto a microfabricated structure called a flow cell, which contains millions of copies of the adapter sequences fused to a glass plate. The adapter sequences attached to the flow cell are complementary to the adapter sequences on individual DNA fragments in the DNA library, anchoring each fragment to a specific location on the flow cell. The fragments are then copied locally to create spatial clusters, each consisting of copies from a single fragment in the DNA library. Next, the double-stranded molecules in each cluster are pulled apart (denatured) to give two strands; the second strand from each molecule is washed away, leaving only one strand on the flow cell. Finally, fluorescently labeled nucleotides–labeled with different colors for A, C, G, and T–are added to the flow cell and allowed to bind to each DNA strand. The crucial property that allows determination of the DNA sequence is that, due to the base pairing propensities of DNA, only the nucleotide complementary to the next nucleotide in the fragment under consideration will be incorporated.

The sequence of each fragment is determined based on the fluorescent color emitted from each cluster when a new nucleotide is synthesized. Because the flow cell contains millions of clusters, each from a different DNA fragment, this sequencing technique allows the rapid determination of millions of DNA sequences simultaneously.

A key limitation of sequencing by synthesis is that the fragments to be sequenced must be short–no more than about 600 nucleotides long. One reason for this limitation is that the sequencing by synthesis process relies on the proper incorporation of the correct nucleotide at the correct position during each synthesis cycle. However, an incorrect nucleotide or no nucleotide will be inserted with some probability at each cycle in a subset of the fragments within a cluster. Thus, as the number of cycles increases, the fragments within a cluster get increasingly "out of sync" with each other, making it more and more difficult to determine the fluorescent color of the correct nucleotide at each synthesis cycle. This synchronization issue also makes it nearly impossible to sequence through so-called homopolymer repeats consisting of the same nucleotide many times in a row, because the presence of the same nucleotide makes it much more likely that the correct nucleotide will be inserted at an incorrect position.

Although sequencing by synthesis is the most widely used sequencing method, other approaches have recently been developed in an attempt to address some of the shortcomings of sequencing by synthesis. In particular, so-called single-molecule sequencing approaches such as those developed by Pacific Biosciences and Oxford Nanopore can sequence very long DNA molecules without the need for fragmentation. However, single-molecule sequencing cannot yet match the accuracy, cost-effectiveness, and throughput of sequencing by synthesis. Such approaches hold promise for the future, but currently single-molecule sequencing is used only in applications that absolutely require knowing the entire sequences of long DNA molecules.

Once DNA sequencing became cheap, fast, and accurate, biologists rapidly developed numerous ingenious ways of preparing DNA libraries that, when sequenced, measure important cellular quanties. In particular, converting RNA transcripts into complementary DNA molecules allows biologists to perform genomewide surveys of the quantity and identity of RNA molecules in the cellular transcriptome (Wang et al., 2009). This procedure is called RNA sequencing (RNA-seq).

A typical RNA-seq experiment starts by extracting RNA from the cell. A large fraction of cellular RNA consists of many copies of a few transcripts called ribosomal RNA, which dominate the output of RNA-seq unless steps are taken to enrich RNA populations of interest. Therefore, the presence of a poly(A) tail, which ribosomal RNA molecules lack, is often used to select RNA before sequencing. Alternatively, steps

may be taken to selectively remove ribosomal RNA from the sample. In either case, the remaining RNA is then converted to complementary DNA (cDNA) using an enzyme called reverse transcriptase. After cDNA creation, library preparation and sequencing then proceed essentially as described above. The identity and number of the cDNA sequences provide quantitative information about the content of the transcriptome in a particular cellular context.

This standard RNA-seq protocol can be modified in numerous ways to highlight different post-transcriptional regulatory mechanisms. For example, instead of purifying transcripts with poly(A) tails, it is possible to purify microRNAs, which are short non-coding RNAs that regulate degradation of other transcripts. Sequencing cDNAs prepared in this way allows a survey of the microRNAs present in a biological sample. Another way to modify the standard RNA-seq protocol is to create cDNA molecules that start from the precise end of the transcript. This allows determination of the position and length of poly(A) (Chang et al., 2014) or U tails (Slevin et al., 2014; Welch et al., 2015; Lackey et al., 2016), as well as the dynamics of the degradation process in which transcripts are chewed up starting from the end (Slevin et al., 2014; Welch et al., 2015; Lackey et al., 2016).

## 1.3 Typical Steps in Computational Analysis of RNA-Seq Data

The volume and complexity of RNA-seq data require fast, robust computational methods to enable scientific discovery. For example, according to the current manufacturer specifications, the Illumina HiSeq 2500 produces between 300 million and 4 billion sequencing reads per run. Thus, computational methods represent an indispensable step in the process of determining transcriptome dynamics using RNA-seq data. This section summarizes the typical analysis steps involved in taking raw sequencing reads from an RNA-seq experiment and converting them into interpretable information (summarized in Fig. 1.1).

The first step in RNA-seq analysis is to determine the identities of the genes that produced the sequencing reads found in a biological sample. This can be accomplished in one of two ways: alignment or *de novo* assembly. In organisms for which a reliable reference genome is available, such as human or mouse, the most common strategy for mapping reads to genes is to align reads to the reference genome. Alignment of RNA-seq reads requires determining, for each read, the portion of the reference genome that best matches the sequence of the read. Several properties of sequencing reads and genomic DNA sequences make RNA-seq alignment an especially challenging problem. First, many alignments span non-contiguous portions of the

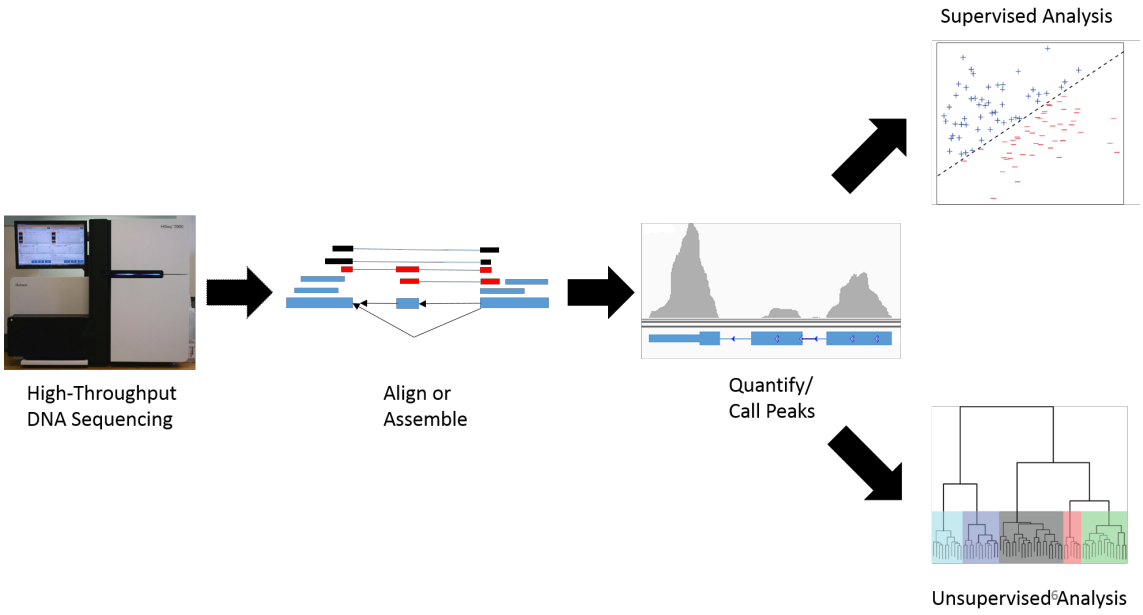



Figure 1.1: Typical steps in computational analysis of RNA-seq data. Sequencing reads must be aligned to a reference genome or assembled. Next, expression levels of genes and transcripts are estimated. Finally, gene and transcript expression levels are used for supervised or unsupervised analyses.

genome due to the splicing process that transcripts undergo. Second, the sequencing process can introduce errors into sequencing reads. Additionally, the cells from which RNA was extracted will have a genome that deviates from the reference genome, and thus reads may contain insertions, deletions, and substitutions relative to the reference genome sequence. Another complicating factor in the alignment process is that many genomes, including the human and mouse genomes, contain large numbers of repetitive sequences, such as transposable elements, viral insertions, pseudogenes, and duplicated genes. Most of these repetitive elements are longer than sequencing reads, making it difficult or impossible in some cases to unambiguously determine the genomic position from which a particular read originated.

Genomes are billions of nucleotides in length and RNA-seq datasets usually contain hundreds of millions of reads, so RNA-seq read alignment algorithms must be computationally efficient. The most common strategy to speed up sequence alignment is to use a pre-built genome index data structure that allows rapid querying of short sequences that match the reference genome almost exactly. These short sequence anchors or "seed" alignments are then refined using more computationally expensive strategies. TopHat (Trapnell et al., 2009; Kim et al., 2013) and MapSplice (Wang et al., 2010) use data structures based on the Burrows-Wheeler transform, and STAR uses a suffix array (Dobin et al., 2013).

When a reference genome is not available, RNA-seq reads can be assembled rather than aligned. Assembly-based approaches also provide a useful complement to alignment-based approaches even in cases when a reference genome is available. Individual genomic variation can create reference bias, which occurs when alignment fails to discover certain transcriptome features due to their deviation from the reference genome. Assembly-based approaches do not suffer from reference bias, and thus can sometimes discover features that alignment-based approaches miss.

Assembly works by grouping together similar reads based on the overlap between their sequences. The most common data structure underlying assembly-based approaches is the de Bruijn graph, in which vertices represent fixed-length substrings (k-mers) of reads and edges represent overlaps between k-mers. The topology of the de Bruijn graph is then used to reconstruct the full-length transcript sequences that could have produced the observed sequencing reads. The most commonly used transcriptome assemblers include Trinity (Grabherr et al., 2011) and Trans-ABySS (Robertson et al., 2010).

After assigning reads to genes using either alignment or assembly, these reads can be used to quantify gene and transcript expression levels. Cufflinks (Trapnell et al., 2010) and RSEM (Li and Dewey, 2011), two of the most widely used gene expression quantification approaches, use latent variable models to estimate expression levels. A key challenge in expression quantification using RNA-seq data is that, in general, transcripts are longer than the read length. Thus, in many cases it is impossible to tell which transcript a particular read came from. Worse still, complex alternative splicing patterns can result in isoforms for which no uniquely identifying class of reads exists, preventing determination of the relative expression levels of the indistinguishable transcripts (Steijger et al., 2013). Another challenging aspect of the quantification problem is that the number of sequencing reads produced from a particular transcript is influenced by both known and unknown sources of bias (Roberts et al., 2011). One known source of bias is GC content: because G-C base pairs are more energetically stable than A-T base pairs, the number of Gs and Cs in a stretch of sequence influences the number of reads that will be produced from that stretch of sequence. Additionally, the number of reads from a transcript depends on both the length of the transcript and the overall efficiency of the sequencing reaction in the sample as a whole. Cufflinks, RSEM, and related approaches therefore report normalized expression levels in units of transcripts per million (TPM) or fragments per kilobase per million mapped reads (FPKM).

Normalized gene and transcript expression levels can be used for both supervised and unsupervised downstream analyses. Supervised analyses use known categorical or numerical quantities associated with

biological samples to discover genes expression differences linked to these quantities. The most common supervised analyses are differential gene expression analysis and differential splicing analysis, which look for differences in overall gene expression or isoform usage linked to a variable such as disease status, treatment, or time point. Other common supervised analyses include building classification models and sparse regression models to predict known sample labels from gene expression levels.

There are two competing strategies for performing differential gene expression analysis. One approach is to use a gene quantification approach like Cufflinks or RSEM to estimate expression of a gene as the sum of the expression levels of its individual transcripts, then use a nonparametric test to assess the significance of the association between the gene expression level and the sample labels. This is the strategy that Cuffdiff uses (Trapnell et al., 2010). Another very popular approach is to simply count the number of reads that map to any transcript of a gene, then use a parametric statistical model for count data to assess the statistical significance of gene expression differences. DESeq is the most popular method in this category, using a negative binomial model to assess the significance of differential expression (Anders and Huber, 2010). Count-based approaches systematically underestimate the expression of genes with variable length transcripts and are confounded by alternative splicing differences between samples (Trapnell et al., 2010). Nevertheless, such approaches still enjoy wide usage because of their speed and simplicity.

There are two main strategies for identifying differential splicing analysis. The first strategy is to rely on expression level estimates for full-length transcripts, then use an approach like Cuffdiff as described in the previous paragraph, except using transcript expression levels instead of gene expression levels (Trapnell et al., 2010). Although this approach works well for gene-level differential expression analysis, transcript-level analyses are much less reliable because, as noted above, it is impossible to estimate the relative expression of transcripts for which no uniquely distinguishing short reads exist. A more sound strategy is to focus on individual alternative splicing events–the portions of transcripts which are guaranteed to be uniquely identifiable from short sequencing reads–rather than trying to estimate abundance of full-length transcripts. Miso (Katz et al., 2010), FDM (Singh et al., 2011), DiffSplice (Hu et al., 2013), and rMATS (Shen et al., 2014) are among the popular methods that use this strategy. Methods that assess differential splicing at the level of individual splicing events bring the additional advantage that their predictions are more interpretable and easier to validate, since they concern short-range variations in the coordinates of exon pairs, which can be validated using standard PCR experiments.

Unsupervised analyses are also frequently used in RNA-seq studies. Such analyses seek to characterize the structure of the biological variation present in a set of samples, rather than using known properties of the samples to guide the analysis. Clustering to discover intrinsic sub-groups and dimensionality reduction to visualize the dominant sources of variation are among the most frequently performed unsupervised analyses.

## 1.4 Single Cell Transcriptomics

The basic unit of the human body is the cell. Therefore, efforts to determine which genes underlie cellular properties would ideally use measurements from individual cells. However, measuring gene expression within single cells presents a significant experimental challenge, so traditional RNA-seq experiments measure the aggregate gene expression profile of a biological sample containing millions of cells. Such bulk measurements do not give any information about differences among cells within a sample. Nevertheless, bulk experiments are very useful, allowing measurement of gene expression differences between tissues, such as heart and brain tissue or cancerous and healthy tissue.

More recently, innovative experimental methods and new technologies have enabled biologists to measure gene expression levels in hundreds to thousands of individual cells. Single cell RNA-seq experiments allow investigation of gene expression differences within heterogeneous cell populations, opening a number of new biological questions for study (Shapiro et al., 2013). For example, single cell RNA-seq can be used to identify the cell types that make up complex tissues, such as those found in the brain or heart. Similarly, single cell resolution is useful for identifying differences among individual tumor cells. The ability to observe gene expression levels within individual cells is also very useful for studying dynamic changes in cell state that occur asynchronously within a cell population. For instance, within a population of dividing cells, cells will be at multiple stages of the cell division cycle at any given moment. A bulk measurement of such a cell population would reflect a weighted average of the gene expression levels present at different stages of the cell division process. Another benefit of single cell resolution is the ability to more precisely track the relationships among transcriptional and post-transcriptional gene regulatory steps. Because these gene regulatory mechanisms operate within individual cells, bulk measurements will tend to obscure relationships between various regulatory steps, such as the link between alternative splicing and polyadenylation.

Attempts to measure gene expression in single cells must overcome two key challenges: (1) the difficulty of manipulating and isolating many individual cells and (2) the tiny amount of RNA present in a single

cell. Currently, the most widely used cell isolation method is microfluidic sorting, in which cells are pushed through tiny pipes in a microfabricated chip, eventually landing in individual reaction chambers (Wu et al., 2014). The most popular microfluidic system is the Fluidigm C1, which accommodates up to 96 single cells per chip. An advantage of the Fluidigm C1 is that the chemical reagents for the RNA extraction and cDNA creation are concentrated in a very small volume, which increases the efficiency of the reaction (Wu et al., 2014). A key shortcoming is throughput–many studies require the profiling of more than 96 cells. Droplet-based methods, such as Drop-Seq (Macosko et al., 2015), inDrop (Klein et al., 2015), and the 10X Genomics Chromium system, encapsulate cells in individual oil droplets, each containing a bead with unique sequence tags. These isolation methods are extremely high-throughput, allowing the rapid and inexpensive capture of tens of thousands of cells at once. However, a key shortcoming of the droplet isolation methods is that they can profile only transcript ends, and thus do not give a full picture of the transcriptome. Microrafts (Gach et al., 2011; Welch et al., 2016b) (sold by Cell Microsystems) and microwells (sold by Benton Dickinson) allow cells to settle onto individual rafts or wells on a chip, then manipulate the rafts or wells individually. These methods are not yet as widely used, but have the advantage that specific cells of interest can be selected based on microscope imaging.

The bulk RNA-seq library preparation must be modified to generate a detectable DNA signal from the tiny amount of RNA in a single cell. First, rather than using separate steps to create cDNA and attach adapters, single cell RNA-seq protocols perform both steps simultaneously. This maximizes the number of transcripts that are successfully converted to cDNA molecules with attached adapters. Another crucial step in single cell RNA-seq library preparation is amplification–making many copies of the cDNA molecules. Performing many cycles of PCR amplification produces enough cDNA molecules for detection by high-throughput sequencing.

The tiny amount of RNA in a single cell also poses a challenge for the analysis of single cell RNA-seq data. The majority of transcripts in any given cell do not make it through the sequencing process; the capture efficiency of single cell RNA-seq has been estimated to be 10-40% (Wu et al., 2014). In addition, the many cycles of PCR amplification introduce significant technical variation, because some cDNA molecules are more highly amplified than others, a phenomenon known as amplification bias (Grün et al., 2014). The stochastic effects of low capture efficiency and amplification bias make it difficult to distinguish true biological variation in gene expression from variation caused only by technical variation (Brennecke et al., 2013). Another property of cell RNA-seq data that poses a challenge during data analysis is that the number of transcripts present in the cell varies widely. This variability in amount of starting RNA requires careful normalization to

ensure that the number of cDNA molecules reported by the sequencer is directly comparable among different cells (Buettner et al., 2015).

## 1.5 Contributions

In this dissertation, I present novel computational methods for using RNA sequencing data to infer dynamic transcriptome changes. My work occurs at this intersection of emerging experimental techniques, important biological questions, and computer science. To develop the techniques described here, I had to carefully study the properties of emerging experimental approaches for performing RNA-seq, including single cell measurements and 3' end measurements. Each of the methods I developed arose from a collaboration with one or more biomedical scientists seeking to answer specific biological questions. The computational techniques that I utilized include statistical modeling, manifold learning, and sequence alignment algorithms. I developed the methods to provide general tools that are useful in a variety of biological settings, but also applied them to discover novel insights about specific biological systems. Thus, my contributions include both the methods themselves and the discoveries that they enabled.

1. **SingleSplice, a computational method for detecting alternative splicing in single cells**
   SingleSplice uses a statistical model to detect biological variation in alternative splicing within a population of single cells. By learning the behavior of exogenous control transcripts, SingleSplice is able to distinguish between true biological variation and technical variation caused by stochasticity in the experimental process. We used SingleSplice to identify alternative splicing differences among mouse cells at different stages of the cell cycle and brain cells from the human neural cortex.

2. **SLICER, an algorithm for studying sequential gene expression changes using single cell data**
   SLICER reveals key properties of a sequential biological process by positing the process as a nonlinear manifold embedded in high-dimensional gene expression space. The approach constructs a manifold from single cell RNA-seq data in an unsupervised manner and uses the geometry of the manifold to order cells according to position in the process and to discover branches and loops in the process. SLICER is the first algorithm that can automatically determine the location and number of branches in a biological process, including multiple levels of branching. I used SLICER to study several important

biological processes, most notably reprogramming of cardiac scar tissue cells into heart muscle cells and aberrant tissue differentiation in cancer.

3. **MATCHER, a method for inferring single cell multi-omic profiles**

   MATCHER infers what single cell transcriptomic and epigenetic measurements obtained from different individual cells of the same type would look like if they were performed simultaneously on an individual cell. The method learns manifold structures underlying each type of data, then aligns these representations to allow direct comparison between different types of data. MATCHER is the first method designed specifically for performing this task, and the inferred multi-omic profiles that it generates allowed the first investigation of the connections among transcriptome changes and changes in chromatin accessibility and histone modifications at the single cell level. I applied MATCHER to mouse embryonic stem cells and induced pluripotent stem cells to gain new insights into how transcriptome and epigenome changes work together.

4. **A method for robust determination of pseudogene expression levels**

   I developed a method for reliably calculating pseudogene expression levels despite the difficulties involved in assigning short sequencing reads to pseudogenes or their parent genes. The method first identifies all unique sequences of a given length in the genome and annotated transcriptome, then uses these unique sequences to filter the reads before expression level calculation. I applied the method to RNA-seq data from breast cancer samples and identified novel pseudogenes that are differentially transcribed among breast cancer subtypes, as well as pseudogenes that are likely to function as competing endogenous RNAs.

5. **AppEnD, an algorithm for identifying untemplated nucleotide additions and precise transcript end positions**

   I developed AppEnD to analyze data from EnD-seq, a novel sequencing protocol for locating untemplated nucleotide additions during RNA degradation or biosynthesis. AppEnD uses dynamic programming alignment to find nucleotide additions as short as a single nucleotide, as well as the precise terminus of transcription. Using AppEnD, we found that histone transcripts undergo multiple uridylation events during degradation and discovered that these additions also have a novel function as a transcript end repair mechanism before degradation. Although I developed AppEnD specifically for

EnD-seq data, the approach is also useful for analyzing a variety of other types of sequencing data, and we applied it to PAS-seq, CLIP-seq, and short capped RNA-seq data.

## 1.6 Dissertation Roadmap

The dissertation is organized into 7 chapters. Chapters 2, 3, and 4 present SingleSplice, SLICER, and MATCHER and describe new biological insights generated by applying these methods to single cell data. Chapter 5 describes a method for robustly measuring pseudogene expression from RNA sequencing data, as well as biological results concerning the ceRNA effect in breast cancer. In Chapter 6, I present AppEnD, an algorithm for identifying untemplated additions and transcription termini, and describe how it gives insight into dynamics of RNA degradation. Chapter 7 discusses future applications of and extensions to the methods described here.

# CHAPTER 2

# Robust Detection of Alternative Splicing in a Population of Single Cells

## 2.1 Background and Related Work

Many eukaryotic genes exhibit alternative splicing, producing multiple types of transcripts with distinct exon combinations, which often result in distinct proteins with different functions (Nilsen and Graveley, 2010). Bulk RNA-seq experiments performed on populations of cells are commonly used to obtain an aggregate picture of the splicing changes between biological conditions (Wang et al., 2009). However, the recent development of single cell RNA-seq protocols enabled genomewide investigation of gene expression differences at the level of individual cells, opening many new biological questions for study (Sandberg, 2013; Shapiro et al., 2013). However, due to the technical limitations of nascent methods for single cell RNA-seq analysis, most single-cell studies have investigated cellular expression differences at the level of genes but not isoforms (Saliba et al., 2014; Stegle et al., 2015).

Single cell RNA-seq experiments possess several unique properties (summarized in Table 2.1), including high technical variation (Brennecke et al., 2013) and low coverage (Streets and Huang, 2014), requiring the use of methods different from bulk RNA-seq experiments (Stegle et al., 2015). A single cell possesses only a very small amount of RNA and the sequencing reaction is limited by the amount of starting material; consequently, variability in "cell size" (amount of biological RNA present) affects the sequencing results and must be taken into account during data analysis (Brennecke et al., 2013; Buettner et al., 2015). Note that technical variables such as global capture efficiency (Grün et al., 2014) can also cause differences in "cell size". The tiny amount of RNA in a single cell also means that much amplification is required, which introduces a high level of technical noise (Brennecke et al., 2013; Grün et al., 2014; Kharchenko et al., 2014). The single molecule capture efficiency is also low (Wu et al., 2014), making single cell experiments much less sensitive than bulk RNA-seq experiments; transcripts expressed at low levels may not be detected (Saliba et al., 2014).

Table 2.1: Differences Between Bulk and Single Cell RNA-seq

| **Bulk RNA-seq** | **Single Cell RNA-seq** |
|---|---|
| Number of reads is limiting | Tiny amount of starting RNA is limiting |
| Some coverage bias due to random hexamer priming | Strong 3' bias due to poly(A) priming |
| Not as much amplification needed | Large amount of amplification introduces noise |
| Generally small number of replicates | Usually want at least 80-90 cells (low coverage) |
| N-group design (tumor vs. normal, time course, etc.) | Often single group design |
| Highly sensitive; can detect transcripts present at very low concentrations | Low capture efficiency means that rare transcripts are often missed |

Single cell RNA extraction protocols prime reverse transcription using the poly(A) tail. During this process, the reverse transcriptase enzyme sometimes produces short cDNAs by falling off before reaching the 5' end of the transcript (Saliba et al., 2014). The probability of RT falloff increases with distance from the 3' end, resulting in read coverage biased toward the 3' end. In addition, most single cells are sequenced at low coverage to maximize the number of cells surveyed (Streets and Huang, 2014); as many as 96 cells are usually sequenced in a single HiSeq run (Treutlein et al., 2014), and emerging technologies are able to sequence thousands of cells at very low coverage (Macosko et al., 2015; Fan et al., 2015). Because RNA-seq produces reads that are much shorter than transcripts, inferring abundance estimates for full-length transcripts is not always possible even with bulk RNA-seq. The technical challenges of single cell RNA-seq data make abundance estimates for full-length transcripts highly unreliable (Stegle et al., 2015).

Another key difference is the experimental design; most bulk RNA-seq experiments use an $n$-class design, in which two or more biological groups are compared. The problem of identifying genes and isoforms that are differentially expressed is well studied for $n$-class designs. However, many single cell RNA-seq experiments use a single group design (Brennecke et al., 2013). A common problem is to identify genes that vary within a supposedly homogeneous population of cells. Because variation in the expression level of a gene can come from either technical noise or biological variation, a single group design requires modeling the technical noise of single cell sequencing protocol to determine genes whose variation exceeds that expected from noise (Brennecke et al., 2013; Grün et al., 2014; Kharchenko et al., 2014).

Recent papers have introduced models that describe the technical variation in expression levels of genes measured with single cell RNA-seq (Brennecke et al., 2013; Grün et al., 2014; Kharchenko et al., 2014).

These noise models are trained using spike-in transcripts added at known, constant amounts across a set of cells and can be used to identify genes with significant biological variation in excess of technical variation across populations of single cells. However, existing noise models are unable to detect isoform changes for two reasons: (i) an isoform switching event is a change in ratio, not necessarily absolute expression level and (ii) single cell RNA-seq data do not generally contain sufficient information to measure expression levels of full-length transcripts.

To understand the distinction between a ratio change and a change in absolute expression, consider a gene G that is transcribed into two different isoforms, A and B. If 30 transcripts of G are present in condition 1 and 60 in condition 2, G shows differential gene expression. But if the 30 copies of G in condition 1 consist of 10 A transcripts and 20 B transcripts, and the 60 copies of G in condition 2 consist of 20 A transcripts and 40 B transcripts, G does not undergo a change in isoform usage. In both conditions, isoform A makes up one-third of the transcripts from G and isoform B makes up two-thirds. To identify differences in isoform usage, we must look for a change in the proportions of the transcripts of G that come from A and B, independent of the overall gene expression level. Note that the situation may be more complicated if G has more than two isoforms; in this case, changes in isoform usage may change the contributions of multiple isoforms to the overall expression of G. However, any isoform usage change must result in a different ratio for at least one pair of isoforms. To detect differences in isoform usage, a distribution comparison metric like Jensen-Shannon Divergence can be used (Hu et al., 2013). Alternatively, the relative proportions of each pair of isoforms can be examined.

To overcome these difficulties, we developed a computational method, SingleSplice, which uses a statistical model to detect genes whose isoform usage varies more than expected from the effects of technical noise alone. Importantly, SingleSplice detects such isoform usage differences without attempting to infer expression levels for full-length transcripts. To the best of our knowledge, SingleSplice is the first method that can detect genes whose isoform usage shows significant variation across a set of single cells.

SingleSplice models the effects of technical noise on isoform ratios, allowing investigators to detect biological variation in isoform usage across a population of single cells. We discovered a set of 797 genes that show significant isoform usage differences in mouse embryonic stem cells. One can also use SingleSplice to identify alternative splicing between pre-specified groups of single cells, as we did with cells separated by experimentally determined cell cycle stage. Alternatively, the output of SingleSplice can be used to cluster cells by their isoform ratios to discover intrinsic cell types based on their isoform ratios.

With the development of SingleSplice, a number of interesting biological questions can be investigated using single cell RNA-seq data. For example, it is not known whether every cell within a tissue generally expresses all of the isoforms that are detected in a bulk RNA-seq sample. Preliminary studies suggest that populations of cells may display different "modes" of isoform usage that are blended together in bulk RNA-seq data (Shalek et al., 2013). Single cell studies can provide insight into the isoform usage differences that occur during dynamic biological processes, such as differentiation (Trapnell et al., 2014), immune cell activation (Shalek et al., 2014) or tumorigenesis (Patel et al., 2014). SingleSplice can also be used to investigate heterogeneity within healthy or diseased tissues, with the goal of characterizing previously unknown intrinsic subpopulations of cells defined by splicing differences. Ultimately, integrating other types of functional genomic assays such as single cell DNA sequencing (Dey et al., 2015), single cell Hi-C (Nagano et al., 2013), single cell ATAC-seq (Buenrostro et al., 2015) or single cell ChIP-seq (Rotem et al., 2015) with single cell RNA-seq will give insights into the connections between alternative splicing and other biological processes. Our analysis here indicates that deep coverage and use of spike-in transcripts are important prerequisites for careful and detailed future studies of alternative splicing at the single cell level. Combined with the robust detection method of SingleSplice, single cell RNA-seq studies promise to generate many new insights into basic RNA biology and the ways in which cells work together to enable complex multicellular life.

## 2.2 SingleSplice

The SingleSplice method consists of three main phases. In the first phase, we compute expression levels for the longest pieces of transcripts that can be unambiguously identified using short reads (Figure 2.1A). We accomplish this using the DiffSplice method (Hu et al., 2013). Briefly, we construct a directed, acyclic splice graph directly from read alignments so that possible transcripts correspond to paths through the graph. Using this splice graph, we identify single-entry, single-exit modules in the graph (Figure 2.1A). These single-entry, single-exit portions of the graph are called alternative splicing modules (ASMs), and each path through an ASM corresponds to a piece of one or more transcripts spanning two or more exons; there may be one or more ASMs per gene. ASMs possess the important property that any alternative splicing a gene undergoes will cause a change in the ratio of at least one pair of ASM paths.

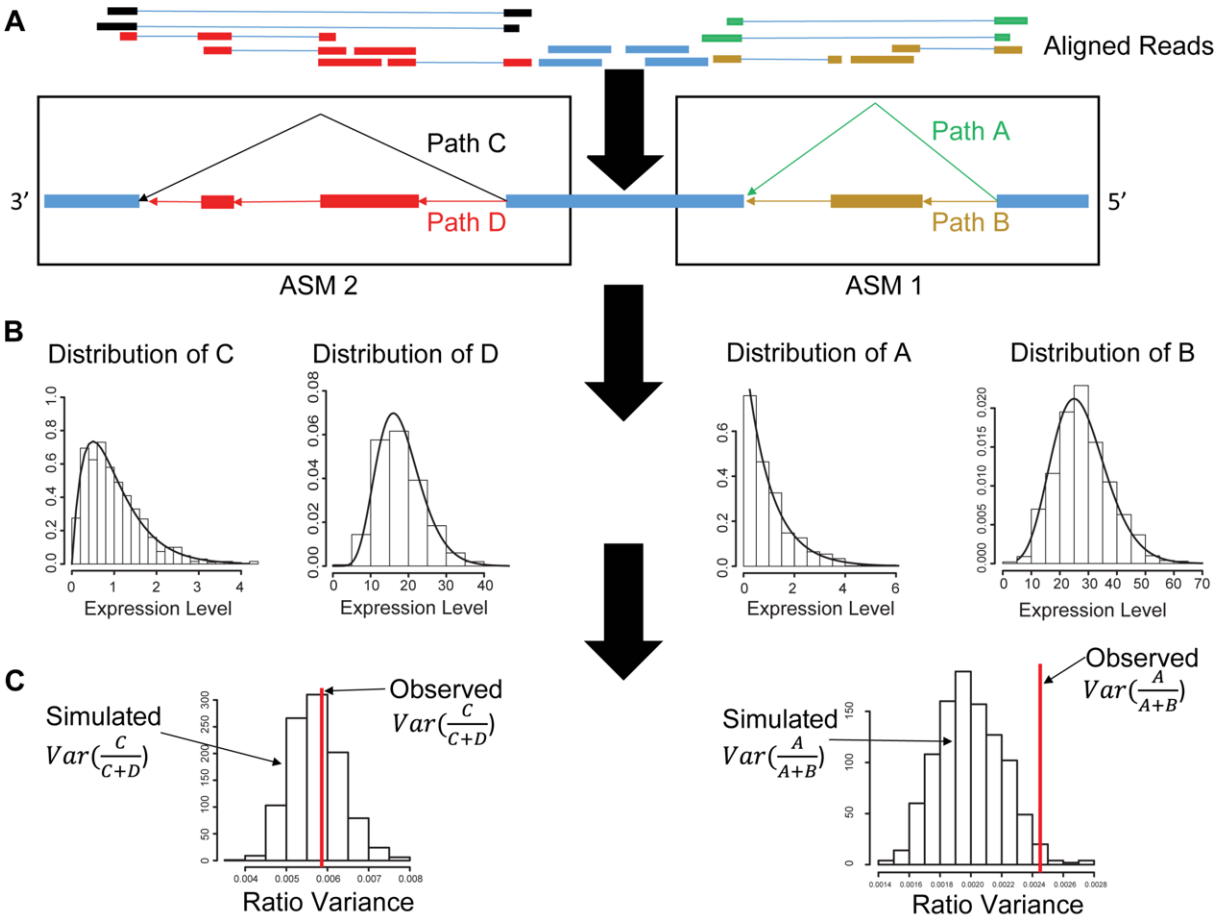



Figure 2.1: Overview of SingleSplice. (A) SingleSplice constructs an expression-weighted splice graph directly from aligned reads (top), then identifies alternative splicing modules (ASMs) and calculates the coverage on each ASM path (indicated in black, red, yellow and green). (B) For each ASM path, a distribution is fit to capture the expected variation in coverage due to technical noise. (C) SingleSplice computes the expected variation in isoform usage by sampling repeatedly from the fitted noise distributions. The resulting sampled values are used to compute an empirical P-value for the null hypothesis that the observed variation in isoform usage results from technical noise alone.

The second phase of SingleSplice fits distributions describing the expected expression variation of each ASM path due to technical noise (Figure 2.1B). In the third phase, to determine whether a gene shows significant splicing changes across a set of cells, we sample values from the fitted noise model of each ASM path to predict the variance of isoform ratios due to technical noise alone, then use these predicted values to assess the significance of the observed variation in isoform ratio (Figure 2.1C). Intuitively, performing this sampling procedure (a statistical technique known as parametric bootstrapping) is like sequencing the same set of cells repeatedly to see how the isoform usage changes from technical variation alone.

To identify genes that exhibit alternative splicing, we construct an expression-weighted splice graph (ESG) directly from the genomic read alignments. An ESG is a directed, acyclic graph in which vertices are genomic coordinates, edges represent splices or contiguous transcription, and the weight on each edge corresponds to its coverage (Hu et al., 2013; Singh et al., 2011; Pertea et al., 2015). Each gene has its own graph, and transcripts are represented as paths through the graph from a start site to an end site. A graph algorithm is subsequently used to identify ASMs (Hu et al., 2013). An ASM is defined as a subgraph of an ESG such that there is only one path into and out of the subgraph, and there is more than one path through the subgraph (Hu et al., 2013).

Intuitively, an ASM represents the longest portion of two or more distinct isoforms that can be each identified by at least one unique set of reads; an ASM path corresponds to the portion of the isoform that differs from other isoforms. To avoid isoforms expressed at very low levels, we used only splice junctions with 10 or more reads in at least 20 samples when identifying ASM structures. A probabilistic model is then fit using expectation maximization to estimate the coverage of each ASM path using the numbers of reads on both the exons and junctions of the paths (Hu et al., 2013). The strategy of identifying ASMs directly from the data as opposed to a simpler strategy such as that used by MISO (Katz et al., 2010) provides two important benefits: (i) discovery of isoforms incorporating unannotated splicing events and (ii) abundance estimation of the longest uniquely identifiable portions of transcripts rather than just the exons immediately adjacent to an alternative splicing event.

In order to predict the variation of isoform ratios caused by technical noise, we first needed a model for technical variation in measured expression level. The basic idea of our approach is to learn a mean-variance relationship from a set of spike-in transcripts, as has been shown to be effective in previous studies (Brennecke et al., 2013; Grün et al., 2014; Kharchenko et al., 2014). Once this mean-variance model is trained, the expected technical variation of any transcript (spike-in or endogenous) can be calculated from the mean of its measured expression levels.

Previous papers (Grün et al., 2014; Kharchenko et al., 2014) have used negative binomial models to predict the expression-dependent variation in read counts on genes. Note that a fundamental assumption of such approaches is that the level of technical noise depends on expression level, or more precisely the number of molecules present at the beginning of the sequencing process. To accurately reflect this assumption, we developed a model for the variation in coverage, not raw read counts, because we are comparing ASM paths that may be of different lengths, so we need to normalize read counts by length. The need to normalize by

length follows directly from the fact that read count is proportional to the number and length of transcripts sequenced. For a given isoform (or ASM path) $t$,

$$reads(t) \propto molecules(t) \times length(t) \tag{2.1}$$

$$coverage(t) = \frac{reads(t)}{length(t)} \tag{2.2}$$

Therefore, coverage (reads per base) is proportional to the number of transcripts present, and we model expression-dependent noise variation using coverage.

Since coverage is continuous rather than count data, we used a gamma distribution the continuous analog of the negative binomial distribution. When we attempted to fit gamma distributions to the spike-in data, we found that the gamma model worked well for highly expressed transcripts, but did not accurately predict the behavior of transcripts at low abundance. Testing the gamma fits using the KolmogorovSmirnov test showed that the fits were accepted for all highly expressed spike-ins but rejected for nearly all spike-ins expressed below 100 RPKMs. While looking at these low expression transcripts, we noticed frequent expression levels of 0 (a dropout event) (Grün et al., 2014), which has an undefined probability under the gamma distribution. Dropout events can occur because of the low capture efficiency of single cell RNA-seq protocols; transcripts expressed at low levels often fail to be captured and amplified (Grün et al., 2014). We thus chose to model technical variation using the following mixture distribution (where $I_{x=0}$ is 1 if $x = 0$ and 0 otherwise):

$$f_X(x) = pI_{x=0} + (1-p)\Gamma(k,\theta)I_{x>0} \tag{2.3}$$

The problem of fitting a noise model then reduces to finding values for $p$, $k$ and $\theta$. We accomplished this by using linear regression to predict the dropout probability $p$ and variance $\sigma^2$ from the mean expression level $\mu$. The variance is predicted using a generalized linear model of the gamma family (Figure 2.2A) and the dropout probability is predicted using logistic regression (Figure 2.2B). Once $p$, $\mu$, and $\sigma^2$ are known, $k$ and $\theta$ can be directly computed using the following equations (which can be easily derived from the expressions for the variance of a gamma distribution). Note that for $p = 0$ (i.e. in the absence of dropouts), these expressions reduce to the equations for gamma mean and variance in terms of $k$ and $\theta$.

$$k = \frac{\mu^2}{\sigma^2(1-p) - p\mu^2} \tag{2.4}$$

$$\theta = \frac{\sigma^2(1-p) - p\mu^2}{\mu(1-p)} \tag{2.5}$$

We performed the gamma regression using the glmgam.fit function from the statmod R package. Only spike-in transcripts with expression levels above a 10 RPKM certain threshold were used to fit the gamma model. Logistic regression was performed using the glm function in R.

Unlike bulk RNA-seq experiments, cellular variation in the amount of starting RNA (cell size) is significant in single cell RNA-seq experiments. Cellular differences like cell cycle stage can affect cell size (Figure 3A). Failure to account for this variation can result in artifacts such as the one shown in Figure 3C where two spike-in transcripts whose expression levels should vary randomly are instead correlated with cell size and with each other. Since spike-ins are added at known, constant amounts, we can use the ratio of biological reads to spike-in reads as a proxy for cell size. The total number of aligned reads per cell also varies independently of cell size variations due to differences in total sequencing depth, read quality, amount of non-polyadenylated RNA that was sequenced, etc. To account for these effects, we normalize coverage both by number of aligned reads and by cell size. To normalize by cell size, we compute a scale factor $s_i$ for each cell $i$ so that the expression levels of each cell are scaled to the median cell size:

$$s_i = \frac{median_j\{\text{aligned biological reads in sample } j/\text{total aligned reads in sample } j\}}{\text{aligned biological reads in sample } i/\text{total aligned reads in sample } i} \tag{2.6}$$

We normalize coverage by the total number of aligned reads, yielding a quantity similar to reads per kilobase length per million reads (RPKM), then multiply by the cell size scale factor:

$$c_{ij} = \frac{\text{coverage of ASM path}j \text{ in sample } i}{\text{total aligned reads in sample } i} \times s_i \tag{2.7}$$

The normalized coverage no longer shows the effects of cell size (compare Figure 2.3B and D).

We use a parametric bootstrapping approach to identify genes whose isoform usage varies more than expected based on technical variation. In the following discussion, we will refer to transcript abundance for convenience, but the values we work with are derived from ASM paths. After determining the parameters of a gamma distribution that predict technical variation in expression level of a pair of transcripts (as described above), we sample repeatedly from these distributions and calculate the proportions of each ASM path in the resulting samples. More formally, for transcript A expressed at an average level of $\mu_1$ and transcript B

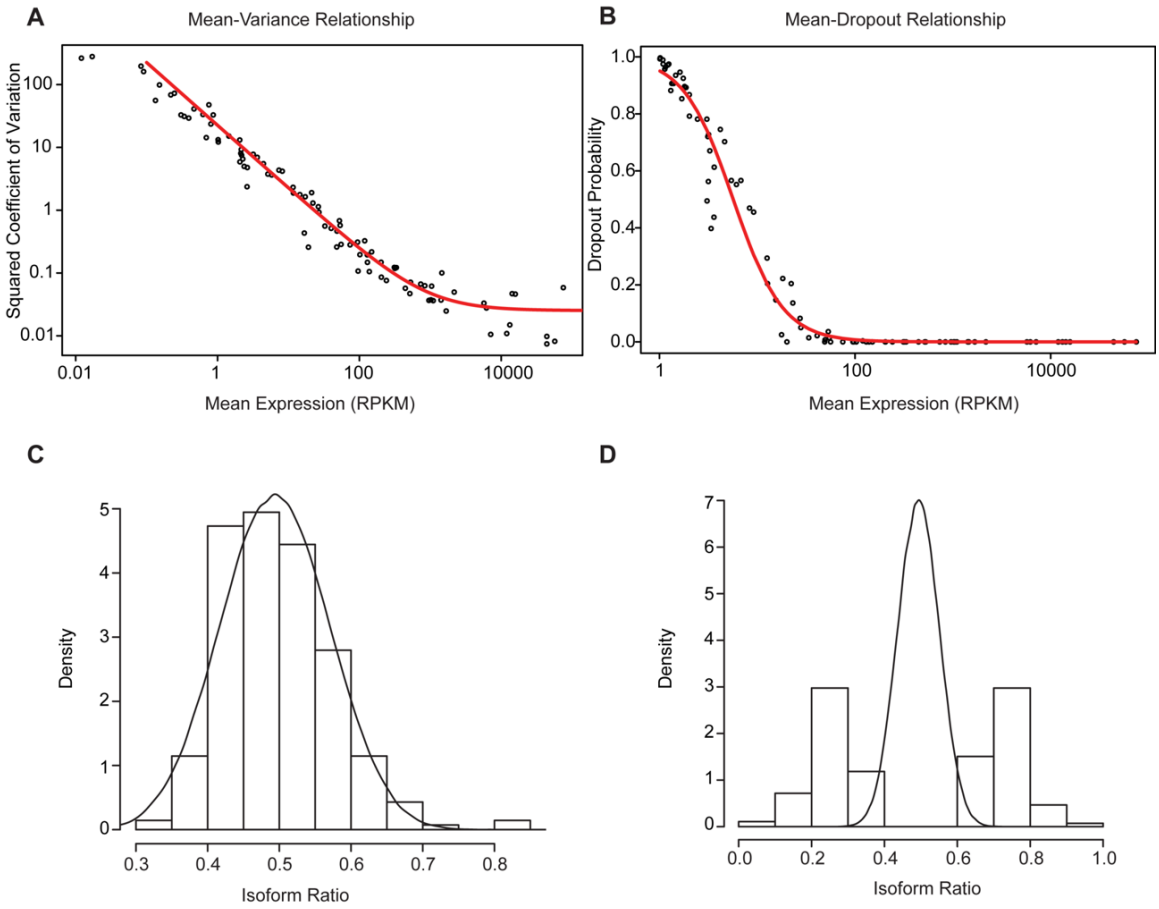



Figure 2.2: Fitting a technical noise model using spike-in transcripts. (A) Gamma regression model to predict variance in coverage as a function of mean expression level. The observed data are shown as black points and the gamma fit is drawn in red. (B) Logistic regression model predicting dropout rate as a function of mean expression level. The observed data are shown as black points, and the regression line is shown in red. (C) Expected (line) and observed (histogram) ratio distributions for a pair of spike-in transcripts showing no ratio change. Note that expectation and observation match very well in this case, indicating that the model effectively predicts the effects of technical noise. (D) Expected (line) and observed (histogram) ratio distributions for a pair of spike-in transcripts showing simulated isoform switching. Note that the observed ratio values differ significantly from what is expected based on technical noise alone.

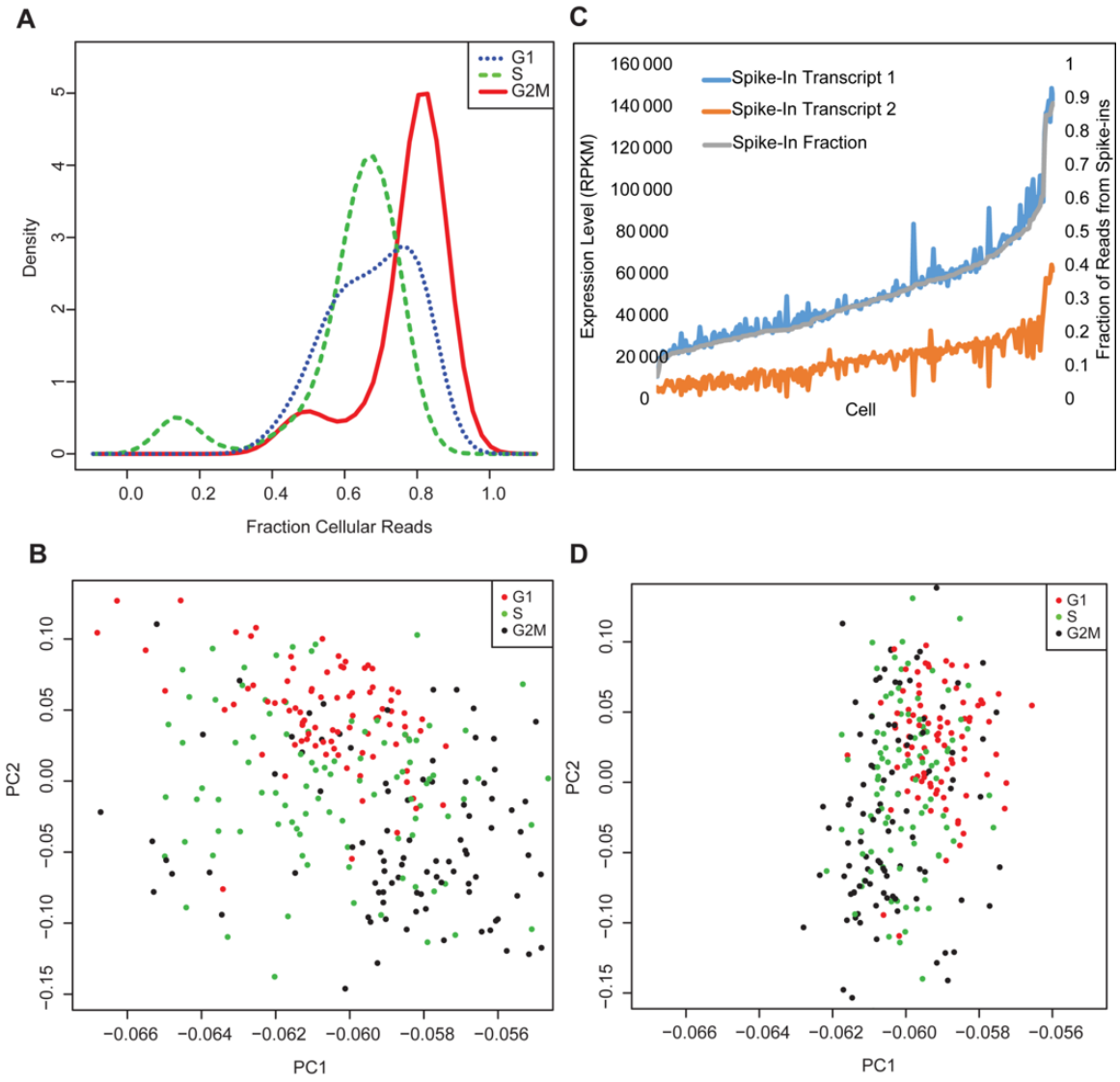



Figure 2.3: Accounting for effects of cell size. (A) Variation in the relative proportions of reads mapping to spike-in transcripts and cellular transcripts indicates that the amount of cellular RNA varies reproducibly during the cell cycle. (B) Since spike-in transcripts are added at constant amounts, their measured expression levels should vary randomly across the set of cells. Instead, PCA using only reads per kilobase length per million reads (RPKMs) from spike-in transcripts before cell size normalization predicts cell cycle stage. (C) Spike-in expression levels should fluctuate randomly due to technical noise, but instead spike-in expression levels before normalization are strongly correlated with each other and with cell size. Note how closely the blue, orange and grey lines trend together. (D) Normalizing for cell size using the fraction of reads that come from spike-in versus cellular RNA removes this effect.

expressed at an average level of $\mu_2$ in a set of $n$ cells, we sample n expression levels for each transcript and repeat this process 1000 times:

$$\mathbf{a} \sim p_1 I_{x=0} + (1 - p_1)\Gamma(k_1, \theta_1) I_{x>0} \tag{2.8}$$

$$\mathbf{b} \sim p_2 I_{x=0} + (1 - p_2)\Gamma(k_2, \theta_2) I_{x>0} \tag{2.9}$$

Then, for each of the 1000 sets of n values, we compute the sample variance of the isoform proportions:

$$s^2 = \sum_{i=1}^{n-1} (r_i - \bar{r})^2, \tag{2.10}$$

where $r_i = a_i/(a_i + b_i)$.

This gives the expected variation in isoform proportions due to technical noise. Intuitively, our parametric bootstrap samples simulate sequencing the same set of cells 1000 times to see how the results change due to technical noise alone. Using the set of $s^2$ values computed in this way, we determine an empirical P-value–for the null hypothesis that technical noise alone accounts for the observed changes in isoform proportions–by simply counting the number of times that variation at least as great as the experimental variation is present in our simulated $s^2$ values. Note that our parametric bootstrapping approach also gives an empirical distribution for isoform proportion $r$; these values can be compared to the distribution observed in the population of sequenced cells (as shown in Figures 2.2CD and 2.4A) using, for example, the KolmogorovSmirnov test. We therefore re-ran our true positive and true negative examples (Figure 2.5) using the KS test and found that the performance was very similar whether an empirical $P$-value for ratio variance or the KS test was used, although the KS test performed slightly worse (data not shown). Note that our method is designed to predict ratio variance for a pair of ASM paths. For an ASM with more than two paths, we compare all pairs of paths; in the case of an ASM with prohibitively many paths, we look only at the $k$ most highly expressed paths, where $k$ is a user-specified constant.

## 2.3 Validation Using Spike-In Transcripts

We used spike-in transcripts (Jiang et al., 2011) added at known, constant concentrations across a set of cells in a previously published data set (Buettner et al., 2015) to calibrate our model and test the sensitivity

and specificity of SingleSplice. Because we are comparing ASM paths that may be of different lengths, and the number of reads obtained from a particular ASM path depends on both initial number of molecules and length, we developed a model for the variation in coverage, not raw read counts (see previous section for a detailed discussion of this point). We used the gamma distribution–the continuous analog of the negative binomial distribution–to model coverage, since coverage is continuous rather than count data.

When we attempted to fit gamma distributions to the spike-in data, we found that the model did not accurately predict the behavior of transcripts at low abundance. These low expression transcripts frequently show expression levels of 0 (a dropout event) (Kharchenko et al., 2014), which has an undefined probability under the gamma distribution. We thus chose to model technical variation using a mixture of gamma and Bernoulli distributions (see Materials and Methods section for details). The problem of fitting a noise model then reduces to finding the parameters of this mixture distribution. We accomplished this by using logistic regression to predict dropout probability and gamma regression to predict variance from mean expression level (Figure 2.2 AB). Parametric bootstrapping using this noise model allows computation of the expected variation in ratio due to technical noise (Figure 2.2CD).

In addition, we found that it was necessary to normalize expression levels by cell size, the total amount of mRNA present in each cell. Since spike-in transcripts are added at known, constant amounts, the ratio of biological to spike-in reads can be used as a proxy for cell size. In the Buettner data set that we analyzed (Buettner et al., 2015), cells at different stages of the cell cycle show consistent differences in cell size (Figure 2.3A). As a result, PCA using only spike-in expression levels (which should show only stochastic variation across the set of cells) separates cells by cell cycle stage (Figure 2.3B), and the expression levels of pairs of spike-ins are strongly correlated with each other and with cell size (Figure 2.3C), even when total sequencing depth is taken into account. Normalizing expression levels by the proportion of reads that came from the cell rather than from spike-in transcripts removes this effect (Figure 2.3D).

To evaluate the performance of SingleSplice, we used two different kinds of tests constructed by pairing spike-in transcripts within each cell so that each spike-in represents an isoform of an alternatively spliced gene (Figure 2.4). We constructed true negative tests by simply pairing the measured expression levels of spike-in transcripts (Figure 2.4A). Because each spike-in transcript is added at a constant amount in every cell, the ratio between a given pair of spike-ins is also constant, technical noise being the only source of variation. The set of spike-ins consists of 96 separate transcripts, which gave 4186 pairs of spike-ins, each pair corresponding to an alternatively spliced gene, after omitting self-pairings and transcripts whose measured

expression was identically zero. SingleSplice correctly identified the majority (85% specificity) of these true negative spike-in pairs as showing no significant isoform ratio change at $p = 0.05$. Figure 2.4B shows the results of this test as a scatter plot, where the x-axis represents the ratio variance predicted by SingleSplice and the y-axis is the observed ratio variance of the spike-in pair. Each rectangle corresponds to a single pair of spike-ins, true negatives are colored green, false positives are colored black and the expression level is indicated by the size of the rectangle. Note that the SingleSplice model predicts the behavior of the isoform ratios quite well, as indicated by how the points generally lie along the dotted line.

We next devised a set of true positive tests in which we swapped half of the measured expression levels for pairs of spike-in transcripts (Figure 2.4C), mimicking isoform switching across a set of cells. In these examples, variation in the ratio of pairs of spike-in transcripts comes from technical noise and simulated isoform switching. As in the true negative case, we constructed 4186 pairs of spike-in transcripts. We found that SingleSplice again performed very well (86% sensitivity). Note that, unlike the true negative test cases, the observed ratio variance generally exceeds the variance expected from technical noise alone (indicated by the dotted line). This shows that SingleSplice accurately detects biological variation in excess of technical variation. Many of the false negatives come from pairs where the spike-ins were expressed at very low levels, as shown by the small boxes in Figure 2.4D that are also black. This effect may be due to a detection threshold below which isoform switching is simply undetectable due to the high level of technical noise (see also the discussion of Figure 2.5 below).

In addition, we note that the External RNA Controls Consortium (ERCC) spike-ins span a very wide range of concentrations, which for some spike-in pairs results in large abundance changes when we swap expression levels to simulate isoform switching. This wide range of spike-in concentrations allows us to assess the performance of SingleSplice across the full spectrum of ratio changes. However, by looking at subsets of the spike-ins we also confirmed that SingleSplice sensitivity shows graceful degradation as the expression levels of the swapped spike-ins approach each other. For spike-ins whose mean expression levels differ by at most a factor of five (mean $>$ 10 RPKMs), sensitivity is 85%. Similarly, for spike-in pairs with fold changes of at most four, three and two, the sensitivity values are 83%, 79% and 69%, respectively. Note also that these sensitivity values vary based on the actual expression level of each spike-in; i.e. isoform switching is much easier to detect between spike-ins with mean expression of 1000 and 2000 than mean expression of 10 and 20.

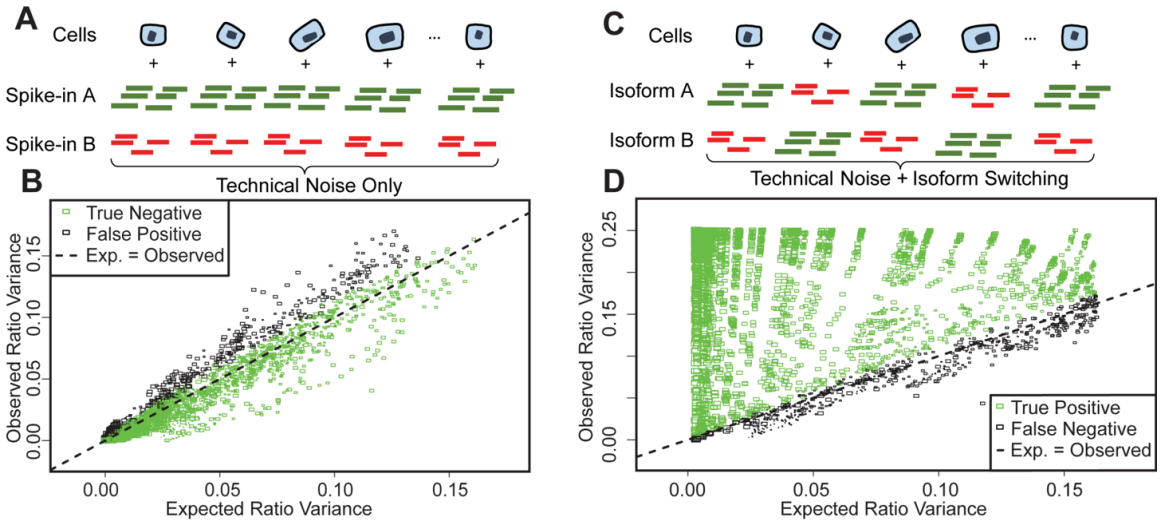



Figure 2.4: Testing the sensitivity and specificity of SingleSplice using spike-in transcripts. (A) True negative examples are created by pairing spike-in transcripts. Any variation in the ratio of these transcripts is due to technical noise. (B) Scatter plot showing expected (SingleSplice prediction) ratio variance versus observed ratio variance for true negative test cases. Each box represents a single pair of spike-ins, and area of the box is proportional to the mean expression level. Test cases where SingleSplice correctly identified the pair of spike-ins as showing no isoform variation are colored green. (C) True positive examples are created by swapping half of the measured expression levels of a pair of spike-in transcripts. Ratio variation in these examples comes from technical noise and simulated isoform switching. (D) Scatter plot showing expected versus observed ratio variance for true positive test cases. Test cases where SingleSplice correctly identified the pair of spike-ins as showing significant isoform variation are colored green.

To demonstrate the importance of the modeling strategies SingleSplice uses to capture expression-dependent noise behavior, we also compared the performance of SingleSplice to a baseline method. A reasonable first approach to identifying alternatively spliced genes would be to choose a threshold value $c$. This baseline method would then classify any genes with ratio variance greater than $c$ as showing significant alternative splicing, and all other genes as showing no significant change. For an appropriately chosen threshold value, the baseline method is fairly effective, achieving 92% sensitivity and 81% specificity across the full set of spike-in pairs described above for $c = 0.05$ (Figure 2.5A). The surprising effectiveness of this strategy is due to the separation between ratio variance for the true positive and true negative spike-in pairs (Figure 2.5B). However, a key shortcoming of the baseline method is its inability to account for differences in expected ratio variance due to expression level. Based on the mean-variance relationship that describes the behavior of technical noise (see Figure 2.2A), we expect that pairs of transcripts expressed at low levels will show much more ratio variance than highly expressed transcript pairs. Inspecting pairs of spike-ins where both transcripts are expressed at a low level (mean $<$ 10 RPKMs) compared to highly expressed spike-ins (mean ¿ 1000 RPKMs) shows that the ratio variance is strongly related to expression level (Figure 2.5B and 2.5C). This fact will systematically bias the baseline method toward calling low expression genes as alternatively spliced and identifying high expression genes as not alternatively spliced, the exact opposite of what is desirable when analyzing noisy, low coverage single cell data. For example, using the cutoff $c = 0.05$ on pairs of spike-ins where both transcripts have mean expression below 10 RPKMs gives a specificity of just 25%. In contrast, SingleSplice correctly identifies 86% of these low expression true negative pairs. Conversely, the cutoff $c = 0.05$ gives 71% sensitivity on highly expressed spike-in pairs compared to SingleSplice's sensitivity of 94% on the same pairs. By modeling the expected ratio variance as a function of expression level, we are able to remove the bias toward calling low expression genes as alternatively spliced. Instead, we determine the significance of splicing variation by the amount of variation expected based on the expression levels of the transcripts involved.

We also devised a set of tests to demonstrate SingleSplice's ability to detect alternative splicing in ASMs with more than two paths. To do this, we sampled random triples of spike-ins, then swapped half of the measured expression levels between two of the transcripts in the triple to mimic isoform switching. True negative examples were created as in the pairwise case by simply using the measured expression levels of the three chosen transcripts. Because there are more than 125000 possible spike-in triples, we randomly sampled 10000 rather than looking at all possible combinations as we did for the pairwise case. We then tested all

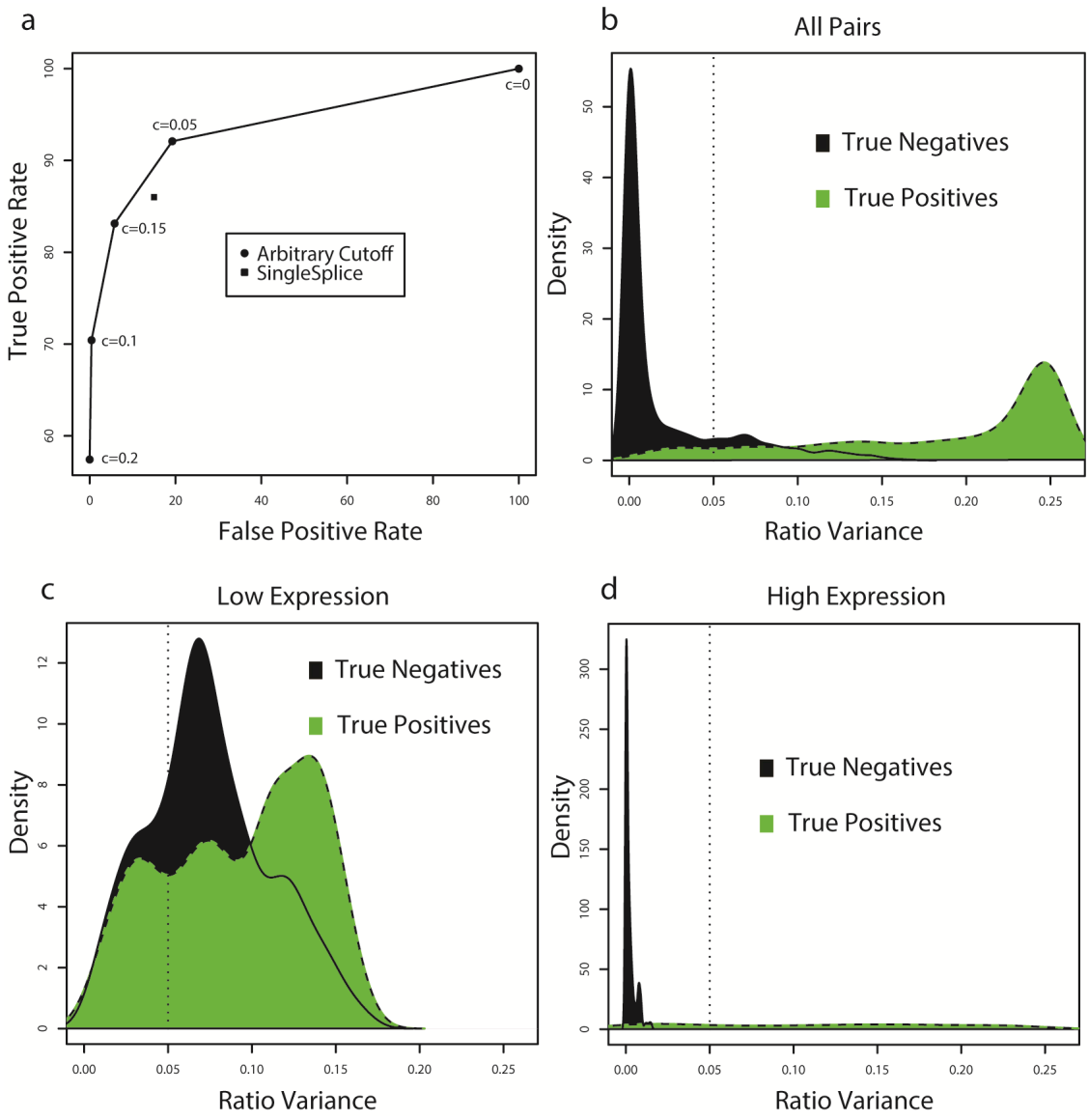



Figure 2.5: Comparison between SingleSplice and a baseline method. (a) Receiver operating characteristic (ROC) curve for the baseline method (choosing an arbitrary cutoff value to separate significant ratio change from no change). Each point on the curve indicates the true positive and false positive rates for a particular choice of the cutoff value. The performance of SingleSplice is indicated as a single point rather than a curve because there are no tunable parameters. (b) Plot showing the distributions of ratio variance for true negative (black) and true positive (green) test cases. The dotted line indicates the best cutoff ($c = 0.05$) according to the ROC curve in the previous panel. (c) Ratio variance for test cases in which both spike-in transcripts have mean expression less than or equal to 10 RPKMs. Note that in this range of expression levels, the fixed cutoff derived from the full set of spike-ins will show poor specificity, biasing the baseline method toward calling low expression pairs as alternatively spliced. (d) Ratio variance for test cases in which both spike-in transcripts have mean expression no smaller than 1000 RPKMs. Note that in this range of expression levels, the fixed cutoff derived from the full set of spike-ins will show poor sensitivity, biasing the baseline method away from calling high expression pairs as alternatively spliced.

$\binom{3}{2} = 3$ pairs of spike-ins for each triple and called the triple alternatively spliced if the P-value for any pair was significant. SingleSplice showed 87% sensitivity and 67% specificity on these tests. The reduction in specificity and the slight increase in sensitivity compared to the pairwise tests is likely due to the fact that a gene is called alternatively spliced if any pair shows a significant change. One strategy to mitigate the drop in specificity is to perform majority voting and call the gene as alternatively spliced only if a majority of the pairwise comparisons are significant. Using this voting strategy on the set of 10000 spike-in triples gives 91% specificity and 85% sensitivity. Our analysis of real data showed that most ASMs do not have more than two highly expressed paths, and SingleSplice allows the user to restrict analysis to the $k$ most highly expressed paths. In addition, SingleSplice outputs the result of the statistical test for each pair of ASM paths, allowing the user to choose whether to use majority voting when assessing if a gene truly shows alternative splicing.

## 2.4 Applications of SingleSplice

### 2.4.1 Alternative Splicing Changes During the Cell Cycle in Mouse Embryonic Stem Cells

Having verified the performance of SingleSplice using spike-in transcripts, we looked for genes with significant isoform usage variation across a set of mouse embryonic stem cells whose cell cycle stage had been determined experimentally before sequencing (Buettner et al., 2015). In the Buettner data set, SingleSplice identified 797 genes that showed significant biological variation in isoform usage (Figure 2.6A). Because the cells in this data set are all from the same cell line, this biological variation is most likely due to changes in the dynamic state of the cells rather than genetic differences. Thus, we would expect isoform usage variation to come from primarily (i) stochastic changes in transcription among cells or (ii) cell cycle differences.

To further investigate the source of the observed variation, we looked for genes whose isoform usage changes are linked to cell cycle phase. To do this, we compared the isoform proportions calculated by SingleSplice across cells in the G1, S and G2/M cell cycle phases. Using a KruskalWallis test and false discovery rate (FDR) correction, we identified 124 genes that show significant isoform usage differences among cell cycle stages, including three particularly interesting examples: *Hnrnpc*, *Snhg3* and *Rbm25* (Figure 2.6BD). *Hnrnpc* encodes an RNA binding protein that plays a role in mRNA splicing (König et al., 2010), nuclear export (Yang et al., 2013) and translational regulation (Kim et al., 2003). In addition, in human cells, the protein product is known to play a crucial role in cell cycle regulation through interaction with the long noncoding RNA *Malat1* (Yang et al., 2013); is differentially phosphorylated during the cell cycle

(Piñol-Roma and Dreyfuss, 1993); and modulates translation of the c-myc protein in a cell cycle dependent manner (Kim et al., 2003). Our SingleSplice analysis revealed that *Hnrnpc* uses an alternative 5' splice site that results in either a long or a short upstream exon, and the short upstream exon is used primarily in S-phase (Figure 2.6B, transcript structure above graph). *Snhg3* is a long non-coding RNA that is conserved between mice and humans but has not been extensively studied, and little is known about its function. *Snhg3* shows a cell-cycle-dependent alternative splicing change in which two short exons are replaced with a longer exon (Figure 2.6C). The relative abundance of the splice form containing two short exons (upper transcript structure in Figure 2.6C) steadily increases through G1 and S phase, peaking in G2/M phase. *Rbm25* is a spliceosome-associated RNA binding protein that has been shown to regulate apoptosis by modulating alternative splicing of the *BCL2L1* gene (Zhou et al., 2008). Our analysis showed that exon skipping in *Rbm25* produces two distinct splice variants (Figure 2.6D) with an expression pattern that differs strikingly between G2/M phase and G1 and S phase. Intriguingly, the distribution of these two splice variants across the set of single cells is bimodal, with modes at 0 and 1, indicating that most cells almost exclusively express either one form or the other (Figure 2.6D). The ASM path with two internal exons (lower transcript structure in Figure 2.6D) appears to be used with much greater frequency among cells that are in G2/M phase compared to the other cell cycle phases.

Principal component analysis (PCA) using only isoform proportions from these 124 genes separates cells by cell cycle stage, underscoring the strong relationship between cell cycle stage and isoform usage (Figure 2.6E). We also looked for gene ontology terms enriched in this set of genes to verify that the genes are involved in the cell cycle. A number of GO terms related to the cell cycle process, including regulation of DNA replication, nuclear division and maintenance of chromosome number, are enriched, lending further credence to the hypothesis that the mRNA splicing changes we observed are likely to play a role in the cell cycle. Interestingly, the set of 124 genes is also enriched for genes involved in RNA splicing and RNA processing, suggesting that global splicing regulation may change during the cell cycle.

Although we also investigated a different data set (Treutlein et al., 2014), we found fewer genes with multiple isoforms detected at appreciable levels, possibly due to lower sequencing depth. In contrast, the Buettner data set was sequenced to greater depth and showed many more splice variants. We found a roughly linear relationship between the read depth per cell and the number of ASM paths detected above 10 RPKMs (Figure 2.7). The majority of ASM paths that we detected occur in only a few cells, which suggests many alternative splicing events are relatively rare due to a combination of biological and technical variation. For

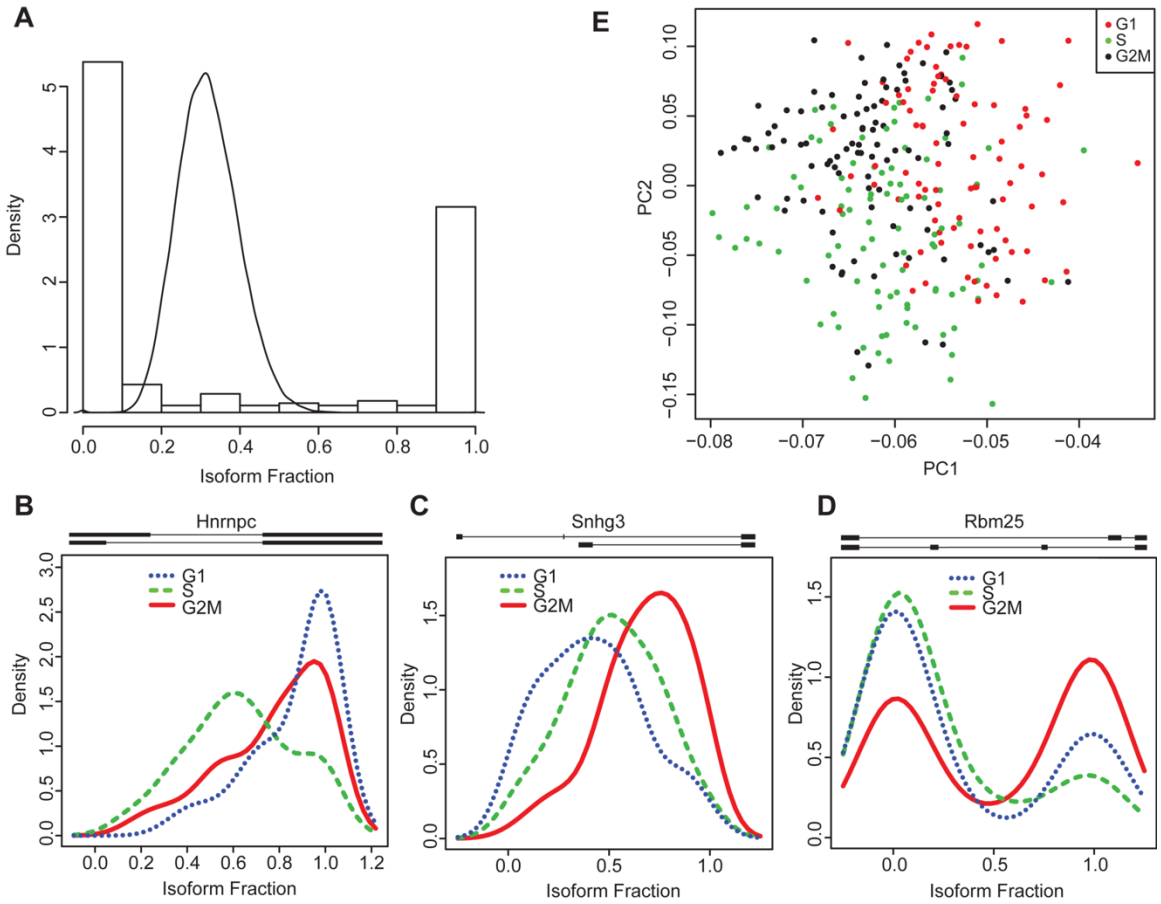



Figure 2.6: Discovery of splicing changes during the cell cycle. (A) Expected (line) and observed (histogram) ratio distributions for the Rbm25 gene. Note that the isoform usage differs significantly from what is expected based on technical noise alone. (BD) The Hnrnpc, Snhg3 and Rbm25 genes show isoform usage changes during the cell cycle. The exon-intron structure (5 to 3 in direction of transcription) of each pair of ASM paths is shown above the corresponding plot. The ratios shown in these panels are computed with respect to the top ASM path i.e. a ratio of 0 corresponds to only the bottom ASM path, and a ratio of 1 indicates only the top ASM path. (E) PCA using isoform ratios alone separates cells according to cell cycle stage.

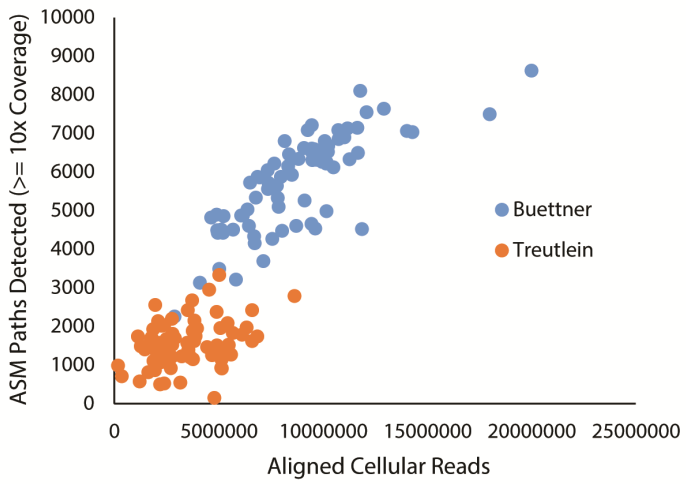



Figure 2.7: Evaluation of the influence of read depth on alternative splicing detection. This plot shows the number of ASM paths detected in each cell as a function of the number of reads in that cell. Note that there is an approximately linear relationship between read depth and number of ASMs detected. The number of reads in the Treutlein experiment is typical for single cell RNA-seq experiments, while the Buettner dataset has unusually deep coverage. The Buettner experiment also sequenced a larger number of cells, so we selected a random subset of cells the same size as the Treutlein dataset to make the two sets of cells as comparable as possible.

this reason, the number of cells sequenced will likely also influence the detection rate of ASM paths. In addition, sequencing more cells increases the statistical power for detecting alternative splicing across the set of single cells by giving more chances to observe a given splicing event. Furthermore, the number of ASM paths detected in each cell at low coverage is smaller than the number of genes detected in a typical single cell RNA-seq experiment, suggesting that many of the genes are not sampled deeply enough to reveal multiple isoforms. Thus, it appears that the low coverage typically used in single cell RNA-seq studies does not completely sample the complexity of the transcriptome, and experiments investigating alternative splicing may need to use increased sequencing depth.

### 2.4.2 Alternative Splicing Differences Among Cells from the Human Neural Cortex

We also applied SingleSplice to a collection of 466 single cells from adult and fetal human neural cortex (Darmanis et al., 2015). The initial analysis of this dataset identified multiple cell populations in the adult tissue, including neurons, astrocytes, oligodendrocytes, endothelial cells, oligodendrocyte precursor cells, and microglia (Darmanis et al., 2015). In addition, Darmanis et al. found that neurons in the fetal samples segregated into two clusters: fetal quiescent and fetal replicating cells. Figure 2.8a shows a 2D projection by

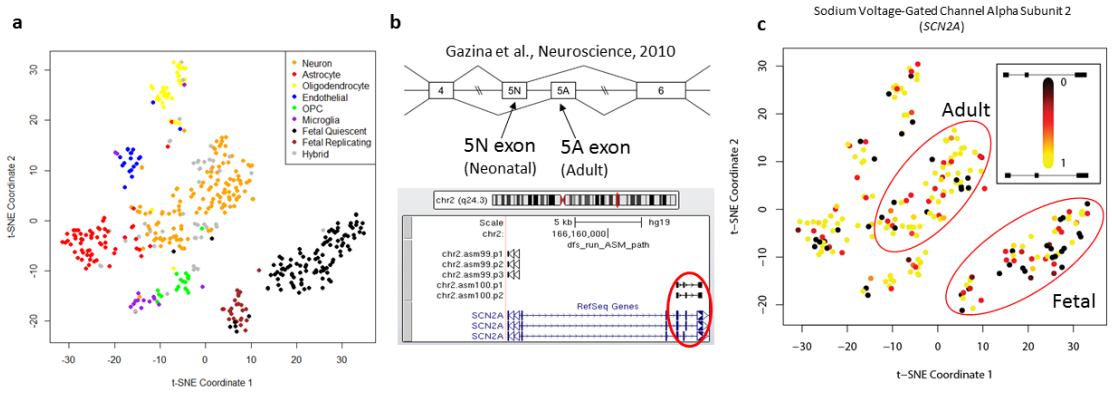



Figure 2.8: SingleSplice Results for the *SCN2A* Gene. (a) 2D projection (by t-SNE) of the gene expression profiles of the 466 cells, colored by the cell type assignments from Darmanis et al. (b) ASM identified by SingleSplice within the SCN2A gene. The splicing event involves two mutually exclusive exons. The 5N exon is used primarily in fetal neurons, and the 5A exon is used primarily in adult neurons. (c) Cell coordinates from panel a colored by ASM path ratio. Black indicates exclusive usage of the 5N exon, while yellow indicates exclusive usage of the 5A exon. Note that cells not expressing either splice form were omitted from the plot.

t-SNE (Van Der Maaten and Hinton, 2008) of the gene expression profiles of the 466 cells, colored by the cell type assignments from Darmanis et al.

SingleSplice identified a set of 985 genes that show significant biological variation (in excess of technical noise) across the population of cells. As in the previous section, we then used the Kruskal-Wallis test to identify a subset of genes that show variation among adult neurons, fetal quiescent neurons, and fetal replicating neurons. Two interesting examples are worth mentioning here.

At the top of the list of significant splicing differences among adult, fetal quiescent, and fetal replicating neurons, we found an alternative exon inclusion event in the *SCN2A* gene, which encodes a protein that performs a crucial function in neuronal action potentials. The ASM that SingleSplice detected involves two mutually exclusive exons (Fig. 2.8b). This exact splicing event was previously identified (Gazina et al., 2010) as a developmentally regulated change, in which the dominant splice form changes from the upstream exon (5N in Fig. 2.8b) to the downstream exon (5A in Fig. 2.8b). As Fig.2.8c shows, the 5N exon is much more prevalent in the fetal neurons. The change in exon usage results in an important change in the SCN2A protein, replacing an uncharged residue close to the voltage sensing domain with a negatively charged residue (Gazina et al., 2010). This change is thought to modulate neuronal excitability (Gazina et al., 2010).

Another interesting gene at the top of the list significant splicing differences among adult, fetal quiescent, and fetal replicating neurons is *NPM1*. The ASM identified by SingleSplice is an exon skipping event (Fig.

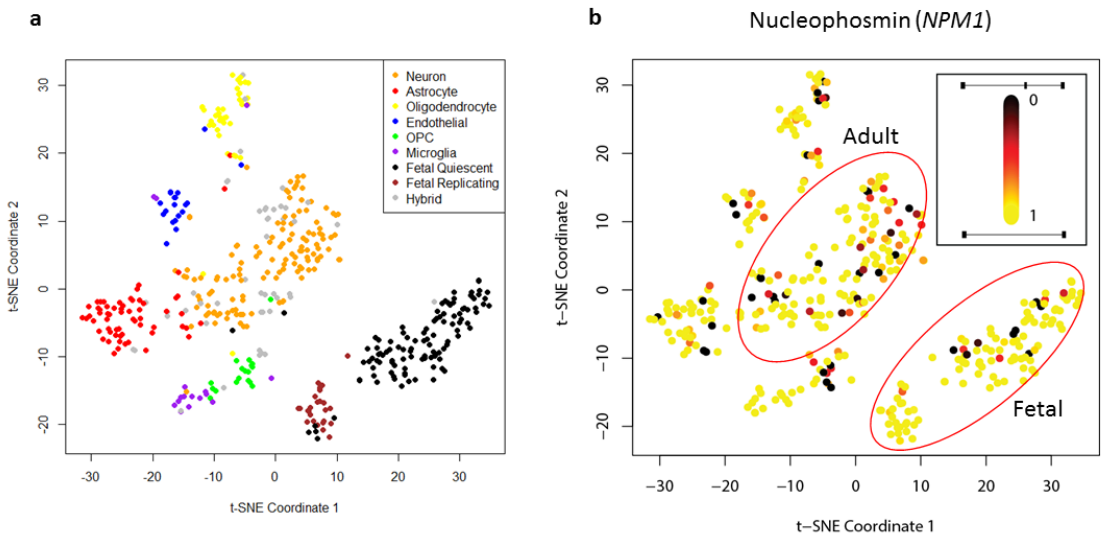



Figure 2.9: SingleSplice Results for the *NPM1* Gene. (a) 2D projection (by t-SNE) of the gene expression profiles of the 466 cells, colored by the cell type assignments from Darmanis et al. (b) Cell coordinates from panel a colored by ASM path ratio. The splicing event involves exon skipping. Black indicates exclusive usage of the exon inclusion splice form, while yellow indicates exclusive usage of exon skipping form. Fetal replicating neurons express the exon skipping splice variant almost exclusively, while the fetal quiescent and adult neurons express both splice variants. Note that cells not expressing either splice form were omitted from the plot.

2.9b). Both the fetal quiescent and adult neurons express a mixture of the exon inclusion and exon skipping isoforms, but the fetal replicating neurons express the exon skipping version almost exclusively (Fig. 2.9b). It appears that far less is known about the regulation of this splicing event during neural differentiation than in the previous example. However, the *NPM1* gene is known to play a role in regulating proliferation during neural stem cell differentiation (Qing et al., 2008), which lends credence to the differences between replicating and quiescent fetal neurons that we observed.

# CHAPTER 3
# Inferring Sequential Gene Expression Changes Using Single Cell Data

## 3.1 Background and Related Work

Understanding the dynamic regulation of gene expression in cells requires the study of important temporal processes, such as cell differentiation, the cell division cycle, or tumorigenesis. However, in such cases, the precise sequence of changes is generally not known, few if any marker genes are known, and individual cells may proceed through the process at different rates. These factors make it very difficult to externally judge where a cell is in the process. Additionally, bulk RNA-seq data may blur aspects of the process because cells sampled at a given point in time may be at different points in the process.

The advent of single cell RNA-seq enables the study of sequential gene expression changes by providing a set of time slices or "snapshots" from individual cells sampling different moments in the process (Trapnell et al., 2014; Bendall et al., 2014; Moignard et al., 2015). To combine these snapshots into a coherent picture, we need an "internal clock" that tells, for each cell, where it is in the process. Because one of the motivations for performing a single cell RNA-seq experiment is to conduct an unbiased, genome-wide study, we would like an unsupervised approach for inferring this internal clock, rather than relying on known marker genes or experiments starting from synchronized cells. Given these motivations, the internal state of a cell is the only reliable way to judge where it is in the process.

One way to approach this problem is to infer a low-dimensional manifold embedded in a high-dimensional space that captures the observed geometric relationships among the cells (Trapnell et al., 2014; Bendall et al., 2014). The modeling assumption behind this approach is that the main difference among cells is where they lie in the process, so that the sequence of gene expression changes traverses a "trajectory" through the sampled cells in high-dimensional space.

Several techniques to identify cellular trajectories have recently been developed. The Monocle tool (Trapnell et al., 2014) uses independent component analysis (ICA) to find a low-dimensional linear projection of the data and then constructs a minimum spanning tree in the resulting low-dimensional space to order

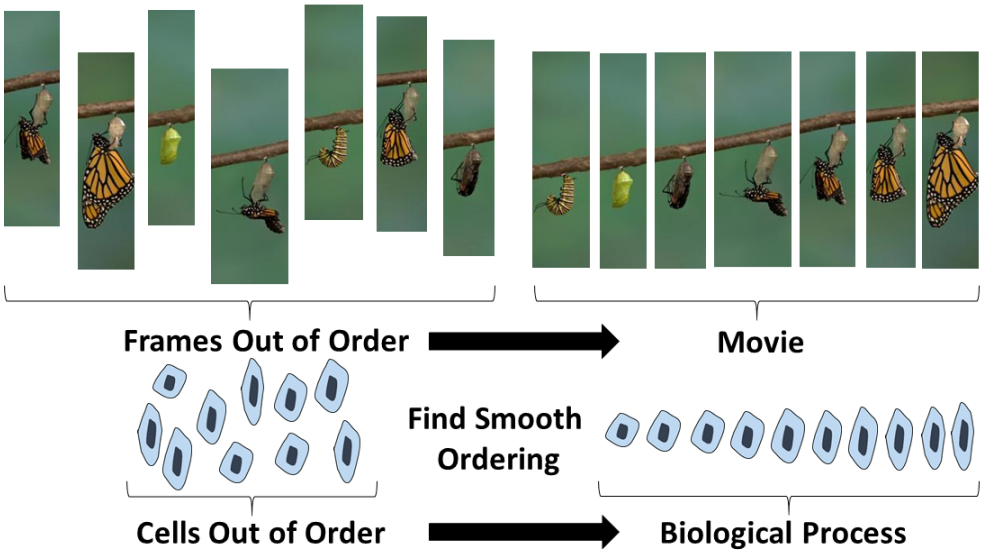



Figure 3.1: Inferring Sequential Gene Expression Changes from Single Cell Measurements

cells progressing through development. Another tool, Wanderlust, constructs an ensemble of $k$-nearest neighbor graphs directly in high-dimensional space without performing dimensionality reduction, then finds the shortest paths through the ensemble of graphs (Bendall et al., 2014). An advantage of Wanderlust is its ability to capture nonlinear behavior.

Monocle and Wanderlust have both been successfully applied to reveal biological insights about cells moving through a biological process (Trapnell et al., 2014; Bendall et al., 2014; Llorens-Bobadilla et al., 2015; Hanchate et al., 2015). However, a number of aspects of the trajectory construction problem remain unexplored. For example, both Monocle and Wanderlust assume that the set of expression values they receive as input have been curated in some way using biological prior knowledge. Wanderlust was designed to work on data from protein marker expression, a situation in which the number of markers is relatively small (dozens, not hundreds of markers) and the markers are hand-picked based on prior knowledge of their involvement in the process. In the initial application of Monocle, genes were selected based on differential expression analysis of bulk RNA-seq data collected at initial and final time points (Trapnell et al., 2014). In addition, Monocle uses ICA, which assumes that the trajectory lies along a linear projection of the data. In biological settings, this assumption may not hold. In contrast, Wanderlust can capture nonlinear trajectories, but it works in the original high-dimensional space, which may make it more susceptible to noise, particularly when given

thousands of genes, many of which are unrelated to the process being studied. Another challenging aspect of trajectory construction is the detection of branches. For example, a developmental process may give rise to multiple cell fates, leading to a bifurcation in the manifold describing the process. Wanderlust assumes that the process is non-branching when constructing a trajectory. Monocle provides the capability of dividing a trajectory into branches, but it requires the user to specify the number of branches.

In this paper, we present SLICER (Selective Locally Linear Inference of Cellular Expression Relationships), a new approach that uses locally linear embedding (LLE) to reconstruct cellular trajectories. SLICER provides four significant advantages over existing methods for inferring cellular trajectories: (1) the ability to automatically select genes to use in building a cellular trajectory with no need for biological prior knowledge; (2) the use of locally linear embedding, a nonlinear dimensionality reduction algorithm, for capturing highly nonlinear relationships between gene expression levels and progression through a process; (3) automatic detection of the number and location of branches in a cellular trajectory using a novel metric called geodesic entropy; and (4) the capability to detect types of features in a trajectory such as "bubbles" that no existing method can detect.

## 3.2 SLICER

Figure 3.2 summarizes the process by which SLICER infers cellular trajectories. SLICER takes as input a matrix of unfiltered gene expression levels. By computing a quantity we term "neighborhood variance," we choose a set of genes to use in building the trajectory (Fig. 3.2a). Intuitively, this method removes genes that show random fluctuation across the set of cells and selects only genes that vary incrementally from cell to cell in a systematic manner. Note that this gene selection method does not require either prior knowledge of genes involved in the process or differential expression analysis of cells from multiple time points. Next, the number of nearest neighbors $k$ to use in constructing a low-dimensional embedding is chosen so as to yield the shape that most resembles a trajectory, as measured by the alpha convex hull ($\alpha$-convex hull) of the embedding (Fig. 3.2a and Fig. 3.5). Alternatively, the user can specify $k$ to manually tune the trajectory. SLICER then uses a nonlinear dimensionality reduction algorithm, locally linear embedding (LLE), to project the set of cells into a lower dimensional space (Fig. 3.2b). The low-dimensional embedding is used to build another neighbor graph, and cells are ordered based on their shortest path distances from a user-specified starting cell. SLICER then computes a metric called geodesic entropy based on the collection of shortest

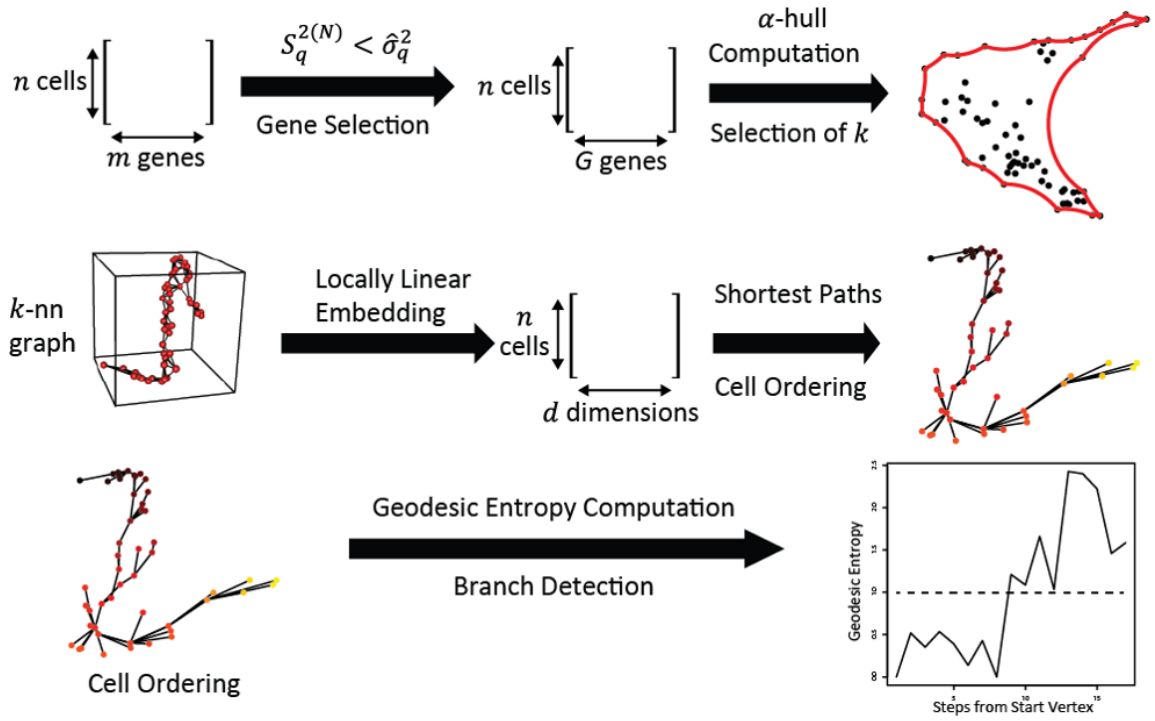



Figure 3.2: Overview of SLICER method. (a) Genes to use in building a trajectory are selected by comparing sample variance and neighborhood variance. Note that this gene selection method does not require either prior knowledge of genes involved in the process or differential expression analysis of cells from multiple time points. Next, the number of nearest neighbors k to use in constructing a low-dimensional embedding is chosen so as to yield the shape that most resembles a trajectory, as measured by the a-convex hull of the cells. (b) SLICER builds a k-nearest neighbor graph in high-dimensional space and then performs LLE to give a nonlinear low-dimensional embedding of the cells. The low-dimensional embedding is then used to build another neighbor graph, and cells are ordered based on their shortest path distances from a user-specified starting cell. (c) SLICER computes geodesic entropy based on the collection of shortest paths from the starting cell and uses the geodesic entropy values to detect branches in the cellular trajectory.

paths from the starting cell and uses the geodesic entropy values to detect the presence, number, and location of branches in the cellular trajectory (Fig. 3.2c, Fig. 3.6). The branch detection approach is based on the insight that the shortest paths along a non-branching trajectory will be highly degenerate, passing through only a small set of cells, in contrast with a branching trajectory which will use one or more distinct sets of cells (see Methods for details).

### 3.2.1 Trajectory Reconstruction

We use locally linear embedding (LLE) (Roweis and Saul, 2000) to reconstruct cellular trajectories. LLE belongs to the class of nonlinear dimensionality reduction techniques, which includes a number of methods, such as Isomap (Tenenbaum et al., 2000), Hessian LLE (Donoho and Grimes, 2003), Laplacian eigenmaps

(Belkin and Niyogi, 2003), and diffusion maps (Coifman and Lafon, 2006). Nonlinear dimensionality reduction techniques have been widely used on high-dimensional data to perform denoising and feature extraction for subsequent classification or regression. For example, such techniques were used to estimate head pose angle and age from images of human faces (Balasubramanian et al., 2007; Fu et al., 2007). We initially experimented with Isomap, Hessian LLE, Laplacian eigenmaps, and diffusion maps and found that LLE seemed to give the best results.

To infer a trajectory using LLE, we take as input a matrix of expression levels with n samples and m genes $\mathbf{E}_{n\times m} = (e_{ij})$, where $e_{ij}$ is the expression of gene $j$ in sample $i$. Then we perform LLE on $\mathbf{E}_{n\times m}$ to give a low-dimensional embedding $\mathbf{L}_{n\times d}$. Our analysis of synthetic and real datasets indicates that $d = 2$ is a reasonable choice (see the following paragraphs for a more detailed discussion of this point). LLE performs dimensionality reduction in two steps. First, a set of reconstruction weights $\mathbf{W}_{n\times d}$ is learned so that each point in high-dimensional space is represented as a linear combination of its $k$-nearest neighbors, where $k$ is a chosen constant:

$$\mathbf{W} = \arg\min_{\mathbf{W}} \sum_{i=1}^{n} |\mathbf{E_i} - \sum_{j=1}^{k} w_{ij}\mathbf{E_j}|_2^2 \tag{3.1}$$

The row sums of $W$ are constrained to 1 to ensure translational invariance (Roweis and Saul, 2000). Then, the weights are used to solve for the coordinates of each point in $d$-dimensional space:

$$\mathbf{L} = \arg\min_{\mathbf{L}} \sum_{i=1}^{n} |\mathbf{L_i} - \sum_{j=1}^{k} w_{ij}\mathbf{L_j}|_2^2 \tag{3.2}$$

The sum-to-one constraint on the reconstruction weights and the form of the weight equations ensure that the low-dimensional reconstruction preserves the high-dimensional geometry of the points (Roweis and Saul, 2000).

After embedding the data using LLE, we build a $k$-nearest neighbor graph in the low-dimensional space. Then we use Dijkstra's algorithm (Dijkstra, 1959) to find the single source shortest paths from a user-specified start point. These shortest paths can be thought of as geodesics that characterize the shape of the cell trajectory manifold, and the length of the shortest path to a particular point represents its geodesic distance from the source point. These geodesic distances can then be used to order the points according to their progress through a process.

The question of the best choice for dimensionality ($d$) is difficult to answer for the trajectory construction problem, because the ground truth cell ordering is unknown for the biological data, and the synthetic data are generated to yield a specific intrinsic dimensionality. While developing SLICER, we explored using intrinsic dimensionality estimators such as packing numbers (Kégl, 2002) and nearest neighbor estimation (Costa et al., 2005) to determine $d$, but tests on our synthetic data showed these methods to be unreliable and highly sensitive to noise. In addition, most of these methods require setting a scale parameter, which simply moves the problem of choosing the dimensionality parameter back one level. Most cell trajectory studies to date have used $d = 2$, and this seems to yield biologically meaningful results. To our knowledge, few studies have used $d > 2$ (Macaulay et al., 2016). For the datasets that we used here (see below for details), $d = 1$ will hide any branches in the trajectory (see Fig. 3.22), and $d = 3$ produces an embedding that is not qualitatively different than $d = 2$ (Fig. 3.3). We note that SLICER allows the user to specify the number of dimensions, and it works for $d \geq 2$.

### 3.2.2 Gene Selection

Selecting the genes to use when constructing a trajectory is a key step in the process. Both Monocle and Wanderlust require the pre-selection of genes based on some sort of prior knowledge. The Monocle paper selected genes that exhibited differential expression in bulk RNA-seq samples taken from the initial and final time points. However, in some cases, cells are collected at only a single time point, and furthermore it would be ideal to have a method for selecting genes without the need for prior knowledge provided by additional experiments.

We developed an approach for selecting genes based on a simple intuition: If a gene is involved in progression along a cellular trajectory, we expect to see gradual changes in the expression of the gene along the trajectory. Conversely, if a gene is not involved in the sequential progression, the gene should fluctuate in a manner independent of the trajectory. Because gene selection must be performed before trajectory construction, selecting genes directly based on whether they are related to the trajectory is not possible. Instead, we note that points close together in Euclidean space are likely to lie close together on the manifold, and we can thus use the similarity of genes in neighboring points to approximate the change in a gene moving along the trajectory. Specifically, for a gene $g$, we calculate the sample variance $\hat{\sigma}^2$ of $g$ across all samples.

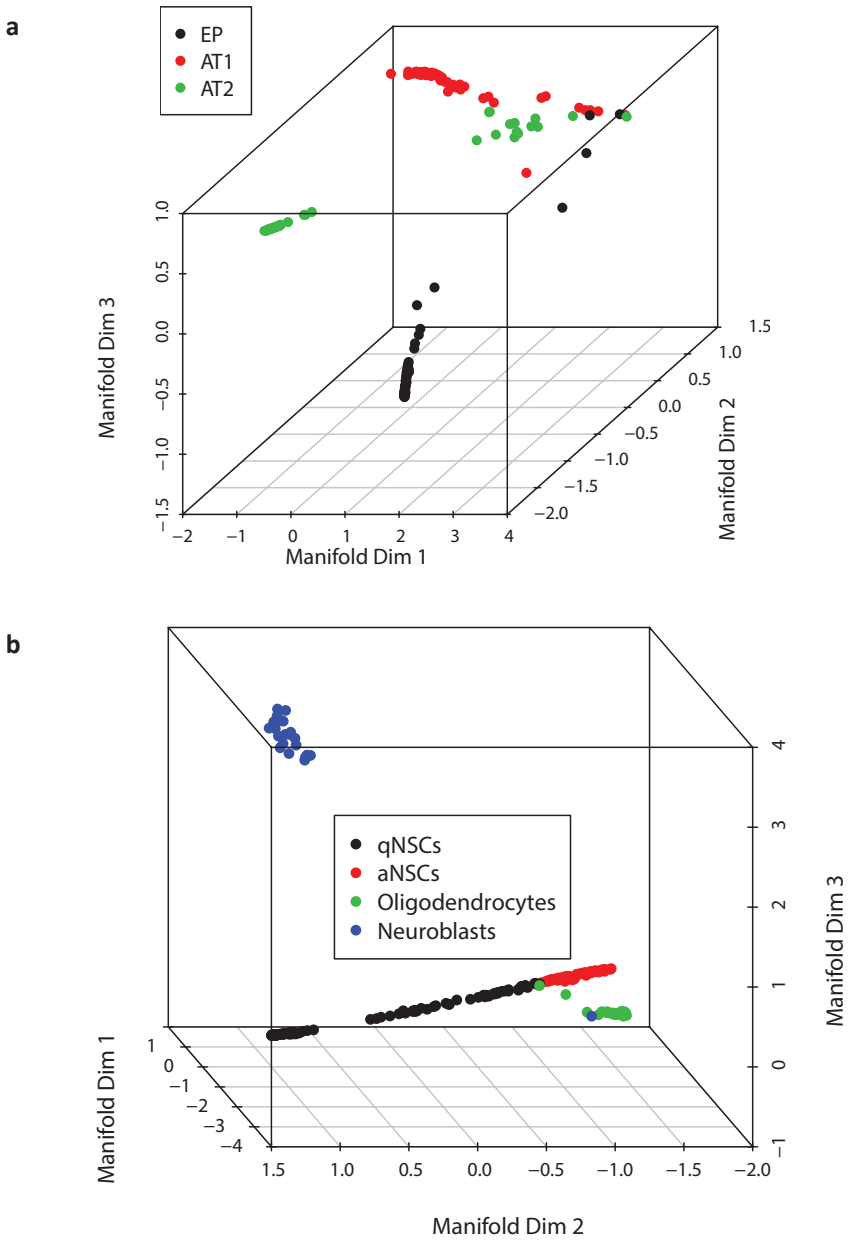



Figure 3.3: 3D embeddings of (a) mouse Lung and (b) neural stem cell datasets.

Then, we compute the "neighborhood variance":

$$S_g^{2(N)} = \frac{1}{nk_c - 1} \sum_{i=1}^{n} \sum_{j=1}^{k_c} (e_{ig} - e_{N(i,j)g})^2 \tag{3.3}$$

where $e_i j$ is the expression level of the $j$th gene in the ith sample, $N(i, j)$ is the $j$th nearest neighbor of sample $i$, and $k_c$ is the minimum number of neighbors needed to yield a connected graph. Intuitively, the quantity $S_g^{2(N)}$ is like a sample variance computed with respect to neighboring points rather than the mean, and it measures how much $g$ varies across neighboring samples. To select the genes that are most likely to be involved in the trajectory, we pick $g$ such that $\hat{\sigma}^2 > S_g^{2(N)}$. These are genes that show more gradual variation across neighboring points than at global scale. In biological datasets, genes often cluster into co-expressed modules, so an important question is how our gene selection method handles co-expressed genes. Because the variance and neighborhood variance are computed for each gene separately, genes related to the trajectory will be selected whether or not they are co-expressed. Conversely, genes that are unrelated to the trajectory will not be selected even they are co-expressed. Examining the correlation matrix of selected genes from the two biological datasets shows that there is a high degree of co-expression, with genes clustering into co-expressed modules (Fig. 3.4). We also note that our simulations include genes that show strong co-expression because they are generated from a handful of functions simulating shared gene regulatory mechanisms. Our simulation results indicate that the gene selection approach works well for these co-expressed genes (Fig. 3.7, Fig. 3.8 and Fig. 3.10).

### 3.2.3 Choosing the Number of Neighbors

Previous approaches for selecting the number of neighbors for LLE have relied upon similarity metrics comparing the relative distances of points in the full space and the embedded space (Olga Kouropteva, Oleg Okun, 2002; Bayro-Corrochano and Eklundh, 2009). We initially tried such approaches and found that they work fairly well on the simulated data but tend to recommend improbably large values for $k$ when run on real data. Consequently, we developed an alternate method that is tailored to the particular manifold shape that we expect to see in this problem. In particular, we expect a trajectory to resemble a long, narrow shape rather than an amorphous point cloud.

To formalize this intuition, we use the notion of alpha convex hull (Edelsbrunner et al., 1983). The $\alpha$-hull of a set of points is the intersection of all closed discs with radius a that contain all of the points. For a given

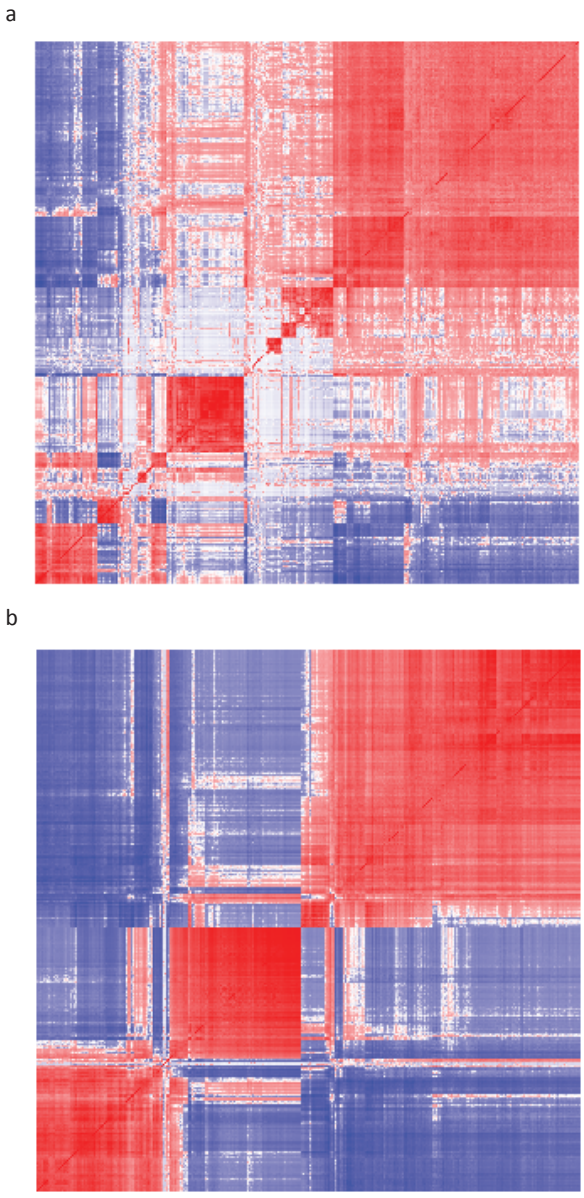



Figure 3.4: Correlation matrices for genes selected by SLICER

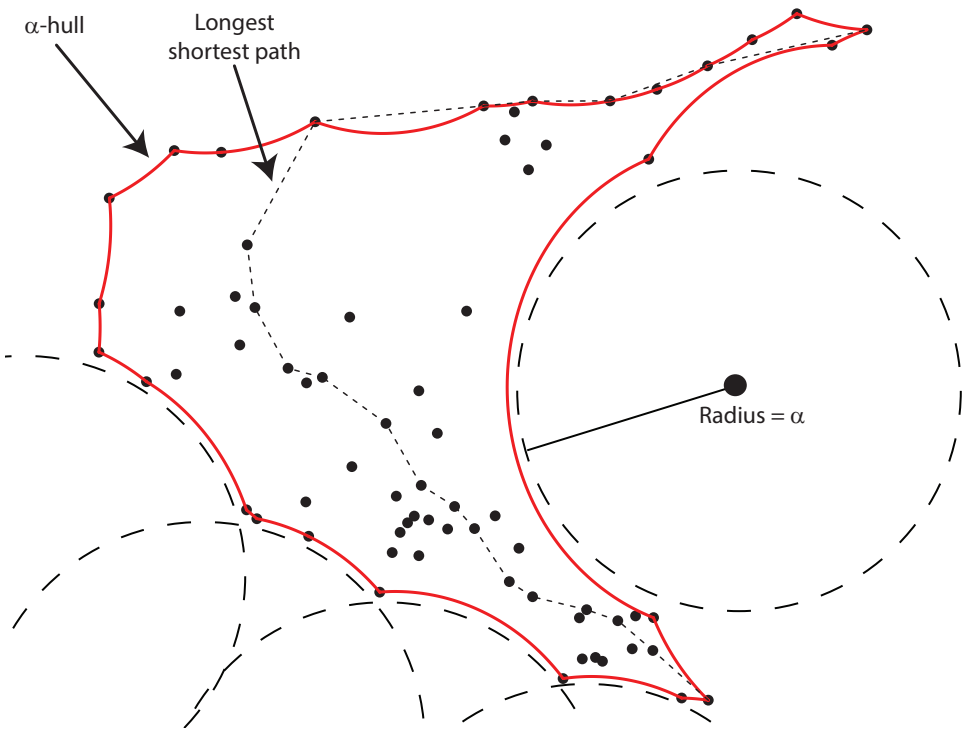



Figure 3.5: Example Illustrating Alpha-Hull Calculation

$k$, we perform LLE and compute the length $l$ of the longest shortest path (see Trajectory reconstruction). We then find the area a of the $\alpha$-hull with $\alpha = l/10$. This choice of a corresponds to the fraction of the length that contains roughly 10% of the data points. Using the area of the a-hull allows us to compute the width of the embedding: $w_k = a/l$. The quantity $w_k$ quantifies how much the embedding resembles a trajectory, and we choose $k = argmin_k w_k$. Figure 3.5 shows an example of the longest shortest path and $\alpha$-hull for a two-dimensional LLE embedding.

### 3.2.4 Detecting Branches

In some cases, the manifold describing a cellular trajectory possesses important properties such as branches. For example, the Monocle paper found a branch in the trajectory corresponding to a split in development resulting in two different cell fates (Trapnell et al., 2014). We developed a novel approach for characterizing the branching structure of a manifold. Our approach can detect the location and number of branches. In addition, we can readily distinguish branches from convergences and bubbles. To do this, we take as input the set of shortest paths used to characterize the trajectory (see Trajectory reconstruction) and

use them to compute a metric that we term geodesic entropy. Intuitively, our approach lines up the shortest paths from the start point to all other points and asks whether the paths use similar vertices.

Let $t_i = s = v_1, ..., v_k, ..., v_l = i$ be the shortest path along the manifold from the starting point $s$ to point $i$ that passes through the $l$ points $v_1, ..., v_k, ..., v_l$. Denote the $k$th vertex on the shortest path from $s$ to $i$ by $t_i(k)$. Consider the set $S$ of shortest paths to each point on the manifold. Then:

$$f_{jk} = \sum_{i}^{n} I[t_i(k) = j] \tag{3.4}$$

is the number of these paths that pass through point $j$ at distance $k$, where $I[]$ is an indicator function. The fraction of all paths in $S$ that pass through vertex $j$ at distance $k$ is:

$$p_{jk} = \frac{f_{jk}}{\sum_{i=1}^{n} f_{ik}} \tag{3.5}$$

Finally we define $H_k$ as the Shannon entropy of $p_k$:

$$H_k = -\sum_{i=1}^{n} p_{ik} \log_2 p_{ik} \tag{3.6}$$

We refer to the quantity $H_k$ as *geodesic entropy* because it describes the vertex composition degeneracy of the shortest paths along the manifold (geodesics). If most of the paths are similar in the first $k$ vertices, then the geodesic entropy $H_k$ will be low (approximately zero), indicating that the manifold does not branch. High geodesic entropy, on the other hand, indicates that multiple distinct vertices are being used along the shortest paths. In fact, following the information theoretic interpretation of entropy as the number of bits needed to transmit a message across a channel, a geodesic entropy of $H_k$ means that there are approximately $2^{H_k}$ distinct paths $k$ vertices from the start point.

Figure 3.6 shows an example of a branching trajectory and illustrates how the geodesic entropy at $k = 10$ steps from the starting cell is computed for this example. To compute $f_{ik}$ and $f_{jk}$ , count the number of shortest paths that contain $i$ and $j$ at position $k$; these numbers are 8 and 9 respectively. This means that the probability of seeing vertex $i$ at position $k$ is $8/17$, and the probability of seeing vertex j at position $k$ is $9/17$. If we treat $(p_{ik}, p_{jk})$ as a probability distribution, we can calculate the geodesic entropy to obtain $H_k \approx 1$.

We use geodesic entropy to detect the location and number of branches and to assign points to branches as follows. Choose $d$ as the smallest value of $k$ such that $H_k \geq 1$. This represents the number of steps from

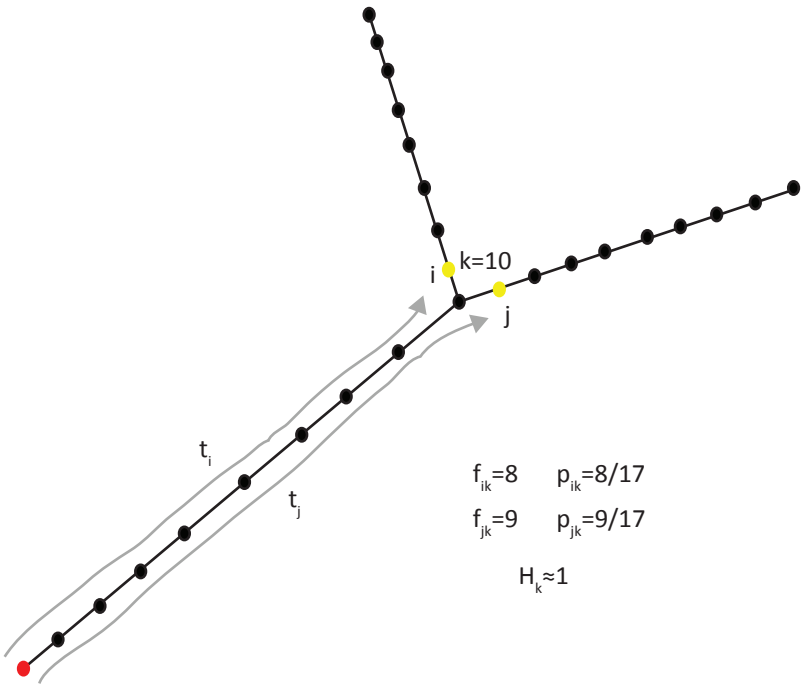



Figure 3.6: Example Illustrating Geodesic Entropy Calculation

the start point along the manifold geodesics at which at least two branches are first detected. The approximate number of branches at $d$ is given by $2^{H_d}$. Now decrement $d$ until you reach a value $c$ such that only one value of $p_{ic}$ is positive (or greater than some $\epsilon$; we used $\epsilon = 0.05$). This represents a vertex at which there is still only one path but beyond which the branch occurs. Now take $b = c + 1$ as the location of the branch and pick the $2^{H_d}$ "distinguishing points" with the highest $p_{ib}$ values. A point $i$ can then be assigned to a branch based on the value of $t_{ib}$ , that is, which of the "distinguishing points" is used at position $b$ in the shortest path to $i$. Points with shortest paths containing fewer than $b$ vertices fall before the branch. As a practical detail, the geodesic entropy will sometimes be high if very few cells are under consideration. For example, at the end of a trajectory, if the shortest path from the start passes through a single cell $c$ and ends at each of the $k$ neighbors of $c$, the geodesic entropy will be $\log_2 k$ even though there is not really a branch at $c$. This problem can easily be addressed by ignoring any branches with less than some number $t$ of cells (SLICER uses $t = 10$ by default).

In addition to detecting branches, geodesic entropy can be used to infer other interesting geometries, such as "bubbles" (see Fig. 3.9). A bubble is a branch that subsequently converges to a single path and can be detected as a spike in $H_k$ such that points on the distinct branches after the spike are connected downstream of

the branch. Complex structures with multiple branches can be unraveled by recursively computing geodesic entropy using the subgraph corresponding to each branch.

We can detect a bubble as follows. We first detect a branch as described above. If the branches identified in this way are connected through the $k$-nearest neighbor graph downstream of the branch point, then this indicates that the branches converge to form a bubble. However, the branches may not be of exactly equal lengths; if they are not, then the shortest paths from the start point will continue past the end of the shorter branch and wrap around the bubble. In such a case, there will be another branch at the end of the bubble, where one set of shortest paths continues around the bubble and the other set exits the bubble (see Fig. 3.9 for an example of such a case). We can detect this second branch by recursively computing geodesic entropy on the shorter of the two initial branches. The location of the second branch then indicates the end of the bubble. In the case of initial branches that are exactly the same length, the point at which they connect after the initial branch point indicates the end of the bubble.

## 3.3 Validation Using Synthetic and Real Data

We constructed a set of simulated trajectories to assess the performance of SLICER on inputs with known solutions. To do this, we generated simulated expression levels for genes in such a way that the expression levels are a function of a "process time" parameter $t$. We simulated five different "pathways" using distinct families of functions; the genes generated by a single family of functions are analogous to co-regulated genes in a biological pathway that all change in response to a common regulatory mechanism. For the simulations shown in Fig. 3.7, we used the following five functions:

$$f_1(t) = 5c_1 cos(t/5) + 8 + \epsilon_1 \tag{3.7}$$

$$f_2(t) = 5c2sin(t/5) + 8 + \epsilon_2 \tag{3.8}$$

$$f_3(t) = c_3\sqrt{t} + \epsilon_3 \tag{3.9}$$

$$f_4(t) = 12c_4(t/20)^2 + \epsilon_4 \tag{3.10}$$

$$f_5(t) = 14c_5(16 - (t/20)^2) + \epsilon_5 \tag{3.11}$$

where $c_i \sim N(1, 0.01)$ and $\epsilon_i \sim N(0, \sigma^2)$. For the simulations in Figure 3.8b-d, we used $f_6(t) = c_1(t/5) + 8$, $f_7(t) = 5\log(t+1) + 8$, and $f_3$, $f_4$, $f_5$ as defined below. The genes used in the simulation are generated by multiplying the value of the corresponding function $f(t)$ by a normally distributed random variable $c_i$. For the actual values of t, we used the sequence of 801 values $0, 0.1, 0.2, ..., 79.9, 80$. Because each simulated gene depends on $t$, points simulated in this way lie along an essentially one-dimensional manifold (a trajectory) in high-dimensional space. Because in the real data setting we do not know in advance which genes are involved in a trajectory, we also devised a means to simulate genes that are unrelated to the process. To do this, we randomly permute the simulated values of some genes, thus removing their relationship with $t$. The number of such randomly reshuffled genes is controlled by a parameter $p$. As genes are simulated, we pick a set of five genes (one from each pathway) to reshuffle. A group of 5 is reshuffled in this way with probability $p$. Randomly permuting the genes (rather than simply sampling from a Gaussian, for instance) ensures that the values lie in the exact same range as the related genes, yet have no relationship with $t$.

To measure the performance of a trajectory reconstruction algorithm, we use the algorithm to produce an ordering of the points, then compare the ordering to the true value of $t$ used to generate it. We measure the "percent sortedness" of a list by computing the following quantity:

$$1 - s/\binom{n}{2} \times 100\% \tag{3.12}$$

where $s$ is the number of pairs of items in the list that are out of order. We chose to use percent sortedness rather than a metric related to distance along the trajectory because dimensionality reduction re-scales the data, which makes it difficult to compare methods that perform dimensionality reduction with those that do not. We used the percentage of points assigned to the correct branch as a metric for evaluating SLICERs branch detection algorithm.

We constructed a set of simulated trajectories to assess the performance of SLICER on inputs with known solutions. To do this, we generated simulated expression levels for genes in such a way that the expression levels are a function of a "process time" parameter $t$. We simulated five different "pathways" using distinct families of functions; the genes generated by a single family of functions are analogous to co-regulated

genes in a biological pathway that all change in response to a common regulatory mechanism. Because each simulated gene depends on $t$, points simulated in this way lie along an essentially one-dimensional manifold (a trajectory) in high-dimensional space. Since, in the real data setting, we do not know in advance which genes are involved in a trajectory, we also devised a means to simulate genes that are unrelated to the process. To do this, we randomly permute the simulated values of some genes, thus removing their relationship with $t$. The number of such randomly reshuffled genes is controlled by a parameter $p$.

To measure the performance of a trajectory reconstruction algorithm, we use the algorithm to produce an ordering of the points, then compare it to the true ordering specified by parameter $t$. We used "percent sortedness", the percentage of pairs of items out of order in a list, as a metric for assessing trajectory reconstruction.

Using the synthetic data generated in this way, we compared SLICER to Wanderlust, a previously published method that can reconstruct nonlinear trajectories. Wanderlust requires the user to specify a value for $k$, the number of nearest neighbors; to ensure a fair comparison, we ran Wanderlust for all values of $k$ in $[5, 10, ..., 45, 50]$ and chose the $k$ that gave the best value. We evaluated SLICER in the same way (testing a sequence of $k$ values) and compared the best $k$ to the $k$ that SLICER automatically selected using our $\alpha$-convex hull approach. To test the importance of using a nonlinear method, we also used ICA, a method that finds a linear projection, to perform dimensionality reduction, then performed the same shortest path algorithm that SLICER uses to order the points in the resulting low-dimensional space. For a baseline method, we randomly permuted the elements in the trajectory and measured the sortedness of the result.

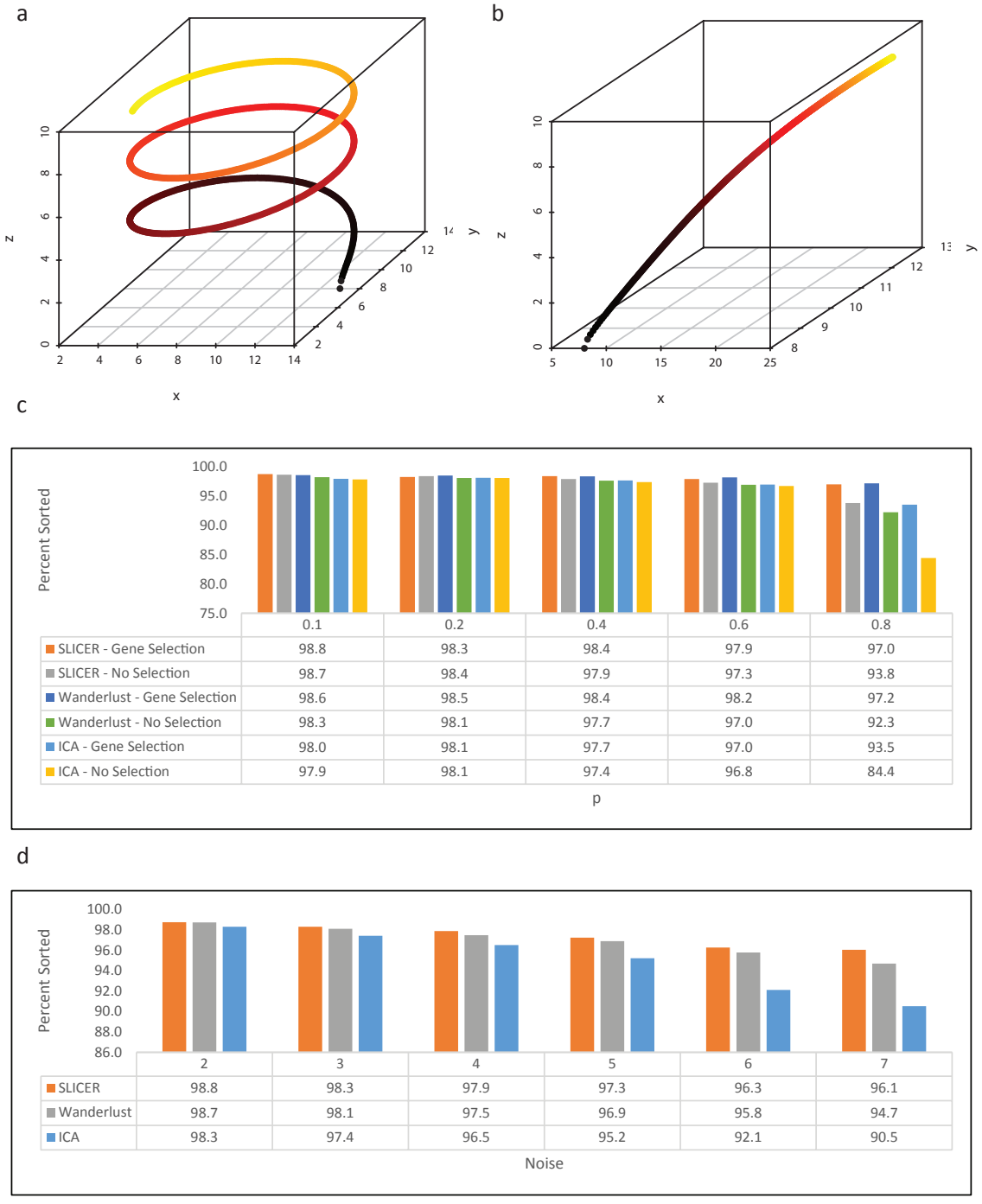
Figure 3.7 shows the results of this comparison. Several things are important to note about these results. First, Wanderlust performs well when the majority of genes are related to the trajectory, but the performance begins to degrade as more unrelated genes are added. This performance degradation may stem from the fact that Wanderlust operates in the original, high-dimensional space, so a large number of irrelevant features begin to compromise the result. In contrast, both SLICER and ICA are fairly stable in the presence of irrelevant genes. However, the ability to capture nonlinear behavior appears to be important, as the performance of ICA is far worse than that of SLICER and Wanderlust (though still better than a random strategy). Finally, the $\alpha$-hull approach for automatic selection of $k$ appears to work well.

The large performance gap between SLICER and the other methods in Fig. 3.7 is due in part to the highly curved shape of the trajectory and the use of gene selection. ICA performs poorly on this example because of the large departure from linearity, and both Wanderlust and ICA suffer from the noise added by


Figure 3.7: Evaluation of SLICER on synthetic data. (a) Comparison of performance of SLICER, Wanderlust, ICA, and random shuffling. The synthetic datasets were generated as described in the text using 500 genes, $\sigma = 2$ ($\sigma$ is the noise level), and increasing values of p. A higher p corresponds to an increased probability that a gene will be randomly reshuffled, removing its relationship with the simulated trajectory. To assess the effectiveness of automatic determination of k, SLICER was run both with and without automatic selection of k. Performance was evaluated by counting the number of inversions in the resulting sorted list of cells. (b) Histogram of percent sortedness values from 1000 random permutations of the simulated trajectory used in panel a. Note that the distribution of values is sharply peaked around 50% sortedness

irrelevant genes. We note, however, that the abilities to automatically select relevant genes and reconstruct highly nonlinear trajectories are key benefits of SLICER compared to existing methods. When we simulated a less highly curved trajectory and fed the genes selected by SLICER to the other methods (Figure 3.8), the gap between methods was much smaller. SLICER with gene selection and Wanderlust with SLICERs selected genes were very similar as the proportion of irrelevant genes increased, although Wanderlust performed slightly better in some cases (Fig. 3.8c). Both methods generally performed better than ICA, with the gap widening as the proportion of irrelevant genes increased (Figure 3.8c). SLICER with no gene selection consistently outperformed the other approaches without gene selection (Figure 3.8c), highlighting the robustness that LLE provides. We also compared SLICER with the other methods for increasing levels of noise with $p = 0$, that is, no irrelevant genes (Fig. 3.8d). This comparison showed that the performance of SLICER degrades slightly less rapidly than the other methods in the presence of increasing noise (Fig. 3.8d), once again indicating the robustness of LLE.

To demonstrate SLICER's ability to detect branches and "bubbles," we simulated a trajectory in which a single initial path splits into two branches that subsequently converge to a single path (Fig. 3.9a). We also created a simulated trajectory with a single branch (Fig. 3.10). We created three families of genes in a manner similar to what is described in the Methods section and used 300 genes, noise level $\sigma = 0.5$, and $p = 0$. Note




| | 0.1 | 0.2 | 0.4 | 0.6 | 0.8 |
|---|---|---|---|---|---|
| SLICER - Gene Selection | 98.8 | 98.3 | 98.4 | 97.9 | 97.0 |
| SLICER - No Selection | 98.7 | 98.4 | 97.9 | 97.3 | 93.8 |
| Wanderlust - Gene Selection | 98.6 | 98.5 | 98.4 | 98.2 | 97.2 |
| Wanderlust - No Selection | 98.3 | 98.1 | 97.7 | 97.0 | 92.3 |
| ICA - Gene Selection | 98.0 | 98.1 | 97.7 | 97.0 | 93.5 |
| ICA - No Selection | 97.9 | 98.1 | 97.4 | 96.8 | 84.4 |

| | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| SLICER | 98.8 | 98.3 | 97.9 | 97.3 | 96.3 | 96.1 |
| Wanderlust | 98.7 | 98.1 | 97.5 | 96.9 | 95.8 | 94.7 |
| ICA | 98.3 | 97.4 | 96.5 | 95.2 | 92.1 | 90.5 |

Figure 3.8: Synthetic data example with less curved trajectory.

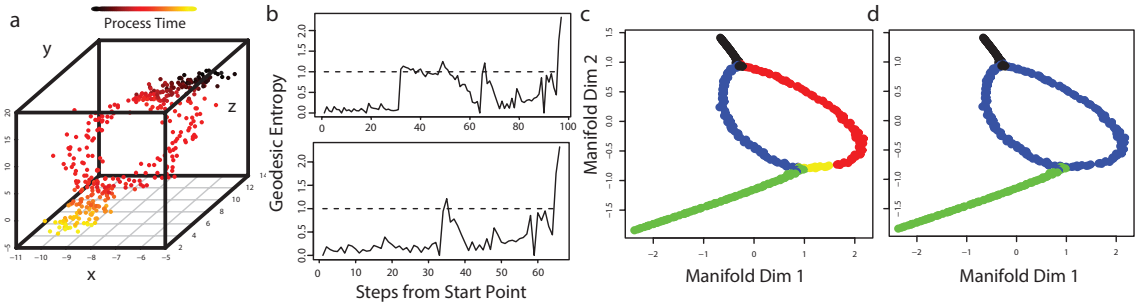



Figure 3.9: Synthetic data example showing that SLICER can detect branches and bubbles. (a) Three simulated genes showing the bubble structure. (b) Geodesic entropy computed for the trajectory (top) and recursively for the longest branch (bottom). The dotted line in each plot represents an entropy of 1, which indicates the beginning of a branch. (c) LLE embedding with branches colored. Black is the initial path that splits into two branches (red and blue). The shorter arm of the initial branch then branches again (yellow and green) at the end of the bubble. (d) Plot showing the boundaries of the bubble (blue) as detected by SLICER.

that the effect of the noise level depends on the relative magnitudes of the genes and the mean of the normal distribution used to add noise. The functions used to simulate the bubble example (Fig. 3.9) have a much smaller range than the genes used in Fig. 3.7, and thus a noise level of 0.5 represents a significant challenge (note the level of noise present in Fig. 3.9a). Figure 3.9a contains an example of the three different gene "shapes" used in the simulated dataset.

We also tested the robustness of SLICERs branch detection in the presence of increasing noise and proportion of irrelevant genes (Fig. 3.10). We used the percentage of cells assigned to the correct branch as a metric for the accuracy of branch detection. This analysis showed that SLICER is able to identify the correct branch assignment for cells even in the presence of irrelevant genes (Fig. 3.10b) and noise (Figure 3.10c), although it appears that noise affects the branch assignment more than irrelevant genes.

## 3.4 Pitfalls in Cell Trajectory Inference

SLICER relies on a key assumption about the data points that it takes as input: that the main source of variation among the data points is position in a one-dimensional, sequential process. In this section, we present simulations to illustrate several scenarios in which this assumption is violated and there is no underlying trajectory, but the output of SLICER looks deceptively similar to a trajectory. These simulations highlight pitfalls that could fool the unwary. The overall message of this section is that it is important to think carefully about the data that one feeds into SLICER, rather than blindly applying the method. We created the

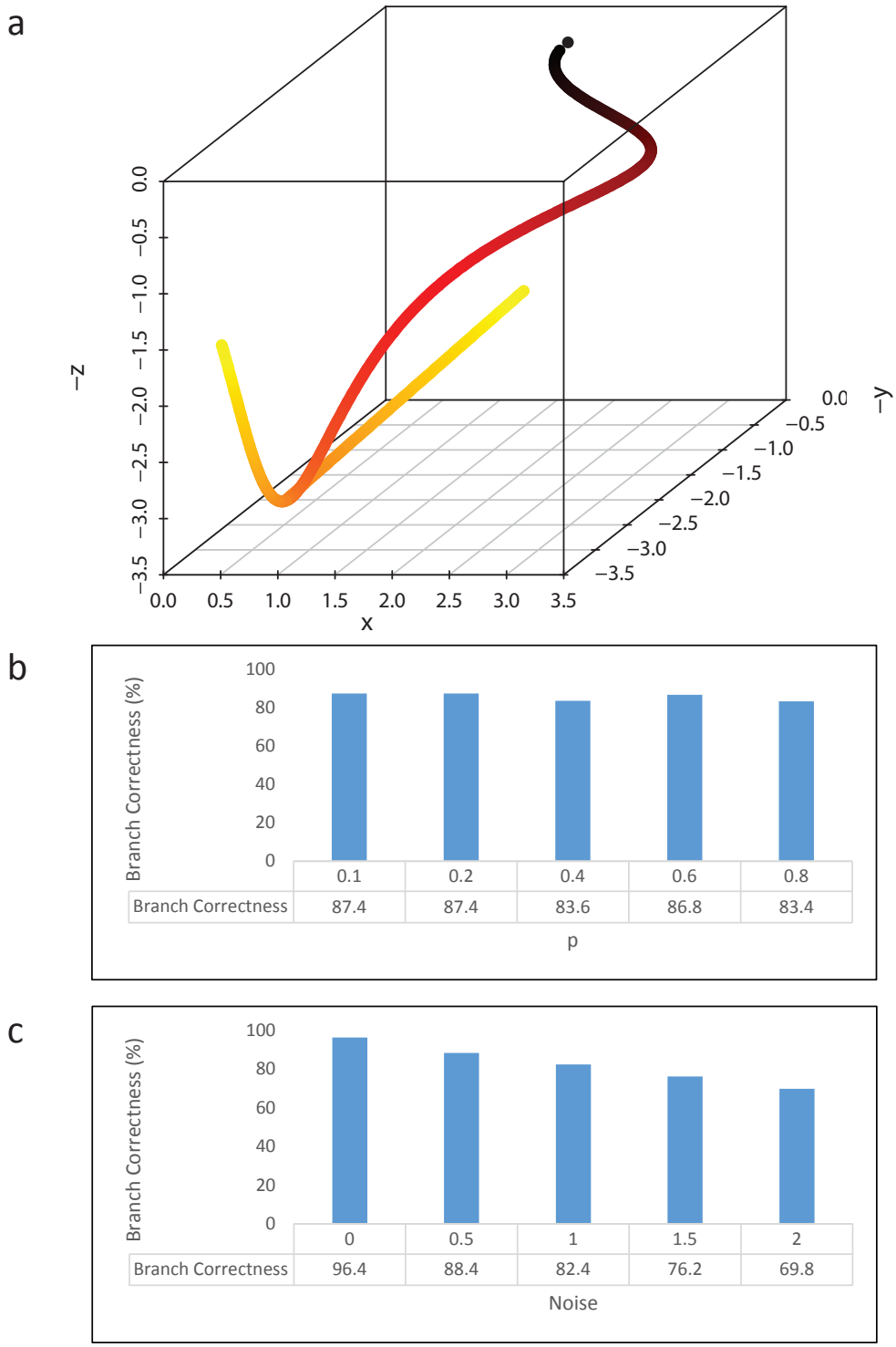



Figure 3.10: Robustness of branch detection in the presence of increasing noise and increasing proportion of irrelevant genes.

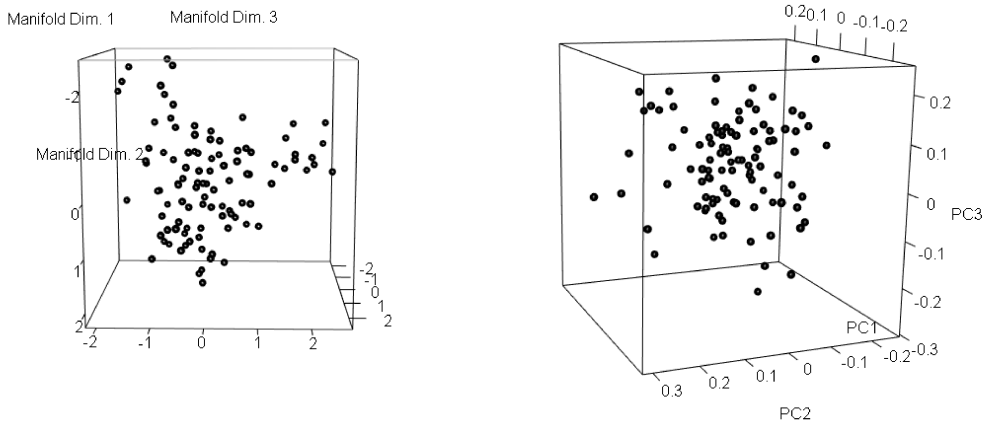



Figure 3.11: Simulation 1: LLE and PCA of Samples from 500-dimensional spherical Gaussian

simulated data sets described below by sampling from high-dimensional spherical Gaussian distributions, resulting in point clouds that lack the sort of low-dimensional structure that SLICER assumes. It is worth noting here that we also used SLICER's gene selection approach on the simulated datasets described below, but in all cases, the method selected 0 genes. This behavior is encouraging, because the simulated datsets lack any sort of trajectory structure, and we would hope that our gene selection approach would not be fooled by these datasets. Additionally, the LLE results shown below are robust to the choice of number of nearest neighbors $k$, and we tried to show the worst case behavior by picking adversarial values of $k$.

We first sample values from a 500-dimensional spherical Gaussian ($\mathcal{N}(\mathbf{0}, \mathbf{I})$). Performing dimensionality reduction on the simulated dataset using PCA or LLE gives an amorphous point cloud (Fig. 3.11). This is a good sanity check, because in this example, the data do not satisfy the assumptions of SLICER, and the amorphous shape of the low-dimensional embedding clearly reflects this fact.

If we now change the simulation slightly and sample from 3 different spherical Gaussians, we get a misleading result. Dimensionality reduction using LLE or PCA produces a plot that appears to contain three separate trajectories (Fig. 3.12). It would be tempting to use this embedding to infer three sequential orderings, but such orderings would be purely artifactual. In this case, the dominant source of variation in the data set is the spherical Gaussian from which a given point is sampled. The more prudent course of action would be to construct separate embeddings for each of the three apparent trajectories, at which point it would

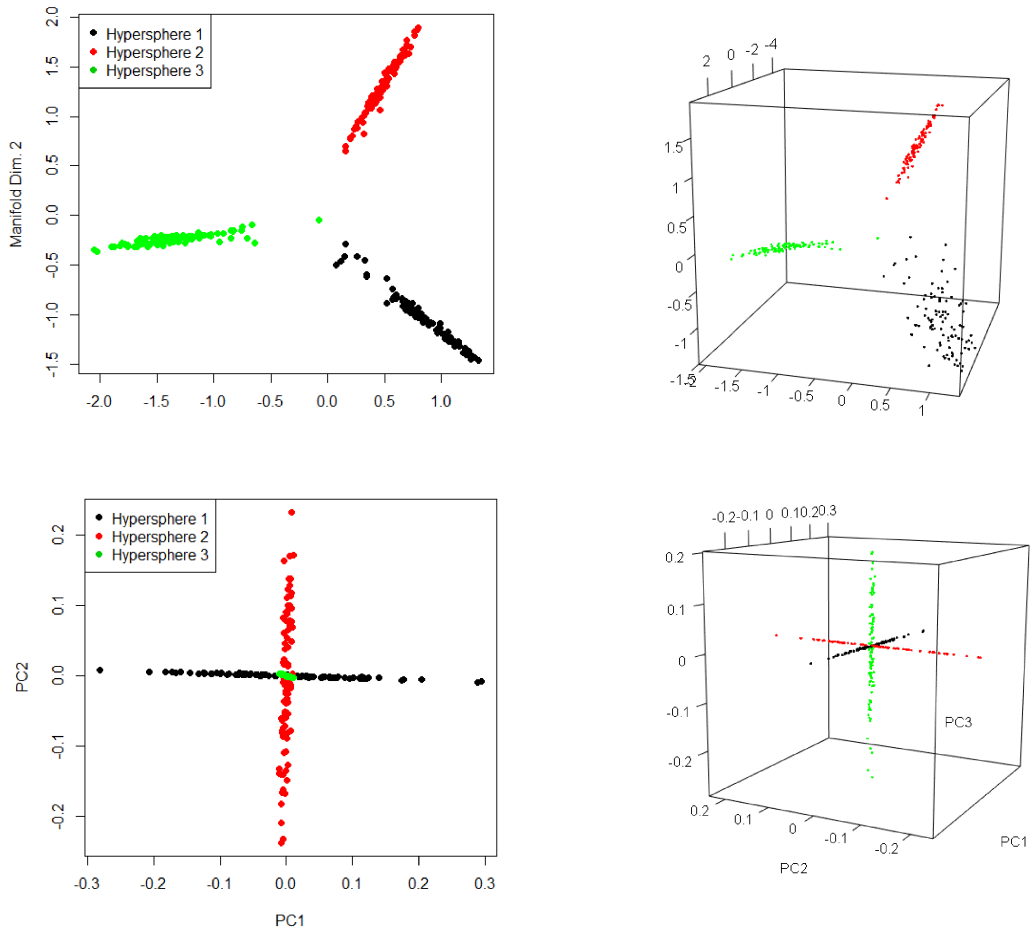



Figure 3.12: Simulation 2: LLE and PCA of Samples from 3 spherical Gaussians

become obvious, as in the previous simulation, that there is no sequential process underlying the differences among data points.

If we add a fourth spherical Gaussian that encompasses the first three, we get another interesting artifact (Fig. 3.13). To a researcher eagerly searching for a cell trajectory, the LLE embedding might appear to be a single starting population that branches in three separate directions. Once again, however, analyzing each of these sets of points separately would reveal that the apparent ordering of the points is illusory. Interestingly, in this case, the first 2 or 3 principal components do not show the same type of structure as the LLE projection.

We can see another artifact if we change the mean of the 4th spherical Gaussian (Fig. 3.14). If an unwary researcher looked only at the 2D LLE embedding shown in Fig. 3.14, he might conclude that the data points fall along a trajectory with a single branch. The 3D LLE projection shows that this apparent branch is an

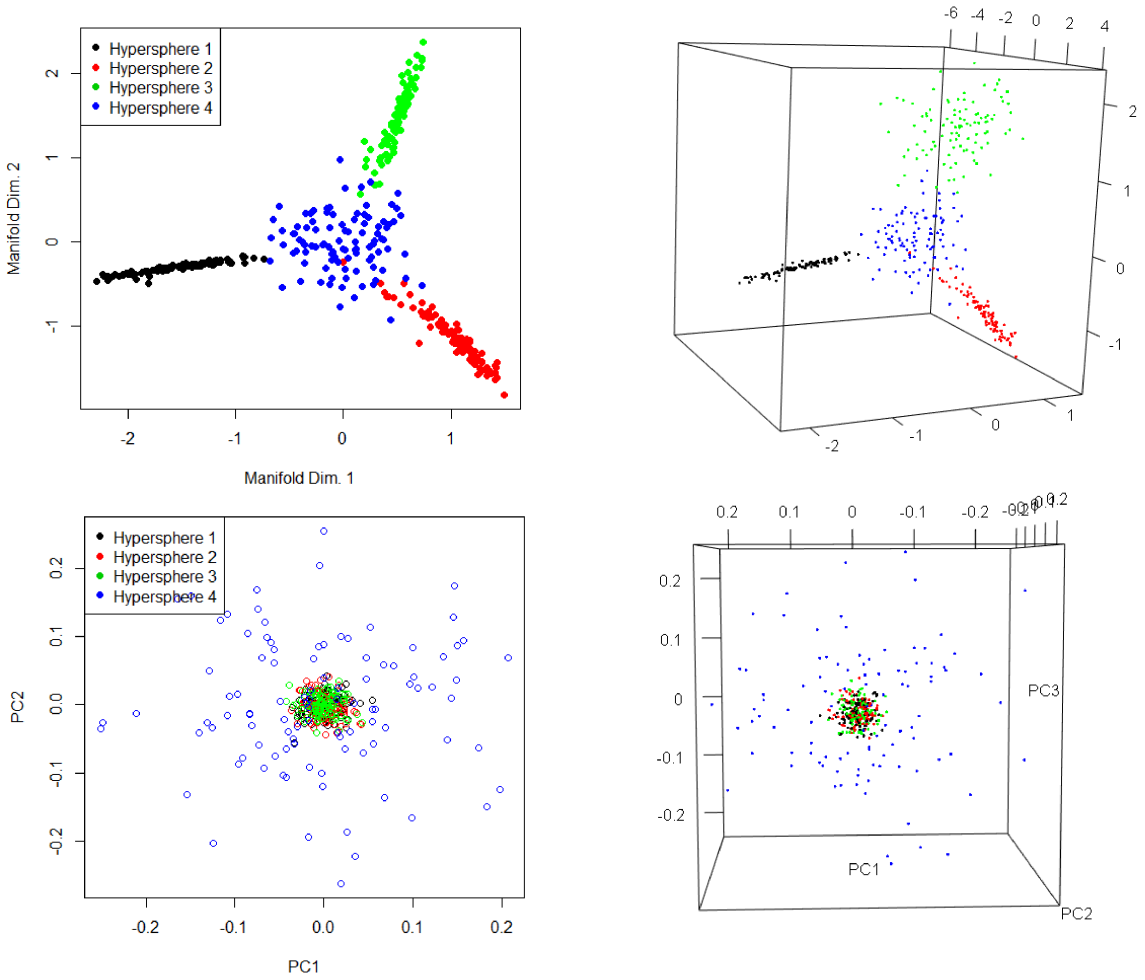



Figure 3.13: Simulation 3: LLE and PCA of Samples from 4 spherical Gaussians

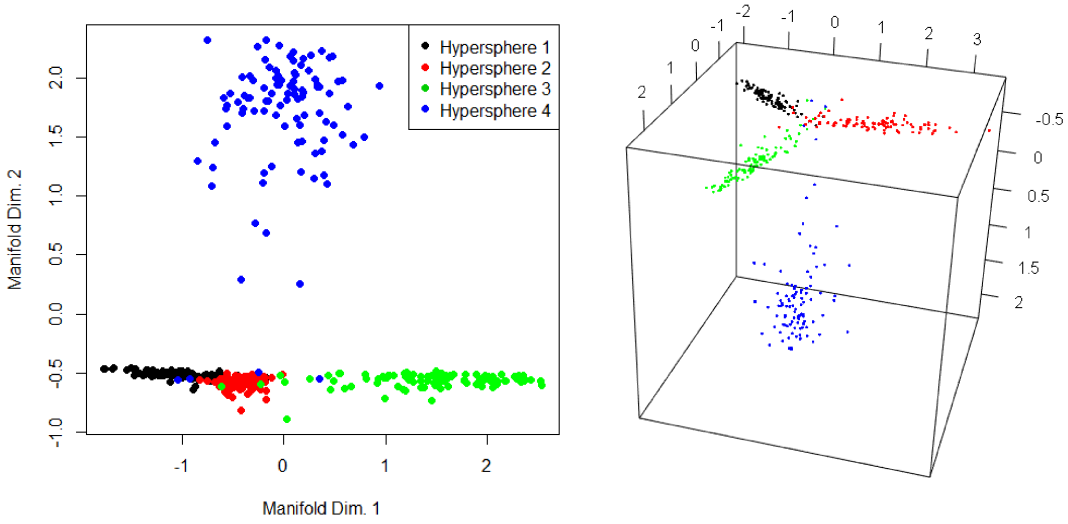



Figure 3.14: Simulation 4: LLE and PCA of Samples from 4 spherical Gaussians (one with different mean)

artifact of squashing the data set into 2D. Even in 3D, however, the same problem noted in the previous two simulations appears–it looks like there are multiple separate trajectories present.

Another way in which cell trajectory inference can produce misleading results is through a mixture of discrete states. As a biological example, bulk RNA-seq samples might be extracted from tissues with varying proportions of two distinct cell types. The main source of variation among the samples would then be the relative proportions of each cell type, but it might appear as though the samples spanned a continuum of states. To illustrate this point, we simulated data points by taking samples from two distinct spherical Gaussians, then drawing a uniformly distributed number $p$ and taking a weighted sum of the two samples: $p\mathcal{N}(\mathbf{0}, \mathbf{I}) + (1-p)\mathcal{N}(\mathbf{0}, \mathbf{I})$. As Fig. 3.15 shows, the LLE embedding of this data set resembles a perfect trajectory. However, the order of the points is determined solely by the value of $p$, as indicated by the color gradient.

## 3.5 Applications of SLICER

### 3.5.1 Developing Mouse Lung

We next ran SLICER on previously published data from developing mouse lung cells (Treutlein et al., 2014). The data were generated as follows: cells from the developing bronchio-alveolar epithelium were

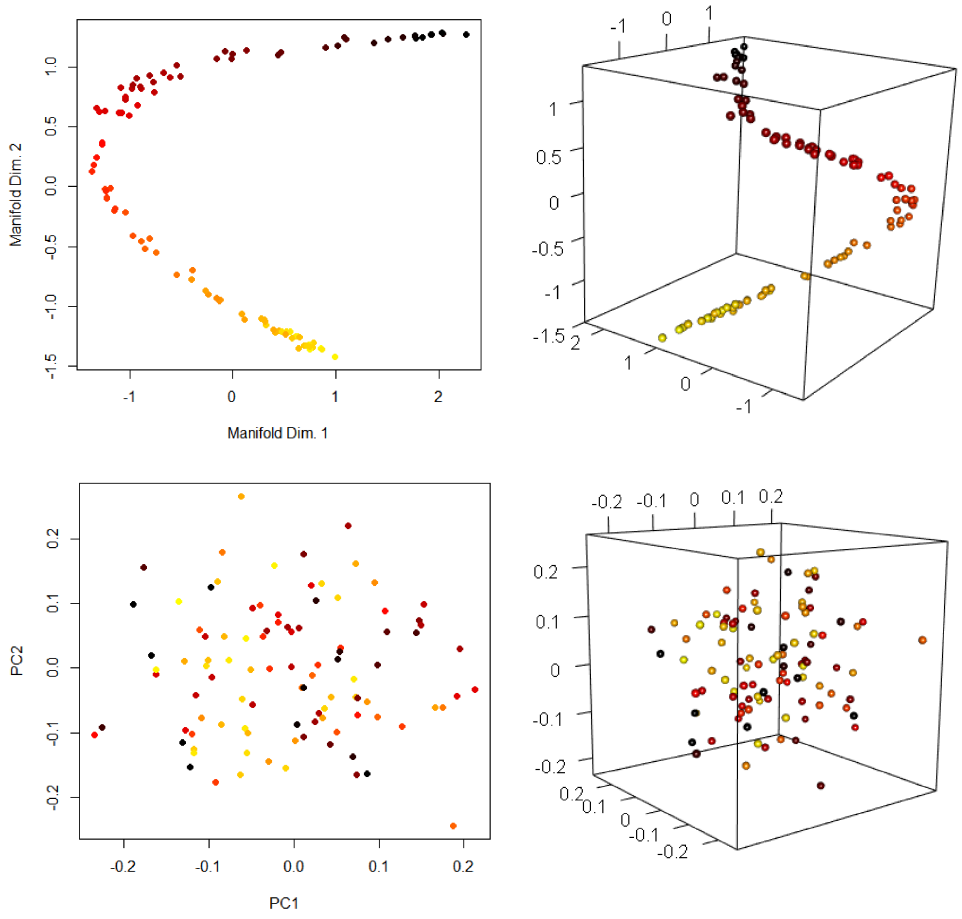



Figure 3.15: Simulation 5: LLE and PCA of Samples from a Mixture of Two Distinct Hyperspheres

extracted from embryonic mice on days E14.5, E16.5, and E18.5. The developing lung epithelium during this stage of development contains progenitor cells, intermediates, and cells committed to one of two specialized cell fates (alveolar type 1, AT1 and alveolar type 2, AT2) (Desai et al., 2014). AT2 cells from adult mice (postnatal day 107) were also extracted and sequenced for comparison. We computed gene expression levels using RSEM v. 1.2.8 and the UCSC mm10 gene annotations. Cells with less than 1000 genes detected at or above 1 FPKM were omitted from further analysis, leaving 183 out of 198 cells. We then log-transformed the expression levels but did not filter the genes in any way.

Each cell in this dataset represents a "snapshot" observation of the sequential process of gene expression changes required for differentiation. Our goal is to investigate the precise sequence of changes, which are not completely understood, although some marker genes for the AT1 and AT2 cell types are known. Differentiation may proceed at different rates across the set of cells, necessitating the use of an internal clock for monitoring differentiation progress, rather than relying strictly on the time point. We therefore would like to construct a trajectory that captures the sequential relationships among the cells undergoing the differentiation process. In addition, this dataset represents an excellent test for the branch detection capabilities of SLICER, because the cells are differentiating toward one of two cell fates, each with a handful of known marker genes.

To determine a set of genes to use in building the trajectory, we selected genes whose expression level variance exceeded their neighborhood variance. This method produced a list of 660 genes. We next computed a two-dimensional embedding of the data using LLE (Fig. 3.16a). We then picked a starting cell, constructed a nearest neighbor graph in the low-dimensional space, and found single-source shortest paths from the starting cell using Dijkstra's algorithm.

As Fig. 3.16a shows, the trajectory reconstructed by SLICER places cells in an order that is clearly related to the day of development. Based on the labels indicating the days on which the cells were extracted, starting at the bottom of the figure and moving to the top and then left or right seems to correspond to progress through development. In this ordering, the cells separate well by day of development. However, there are some exceptions: cells from days E14.5 and E16.5 overlap significantly, indicating that few changes occur during that two-day period. In contrast, there is a wide separation between day E18.5 and the fully differentiated AT2 cells from post-natal day 107. Another salient feature of the SLICER trajectory is that there appears to be a branch, with some cells positioned to the left of the early progenitors approaching the AT2 cells and some to the right of the early progenitors.

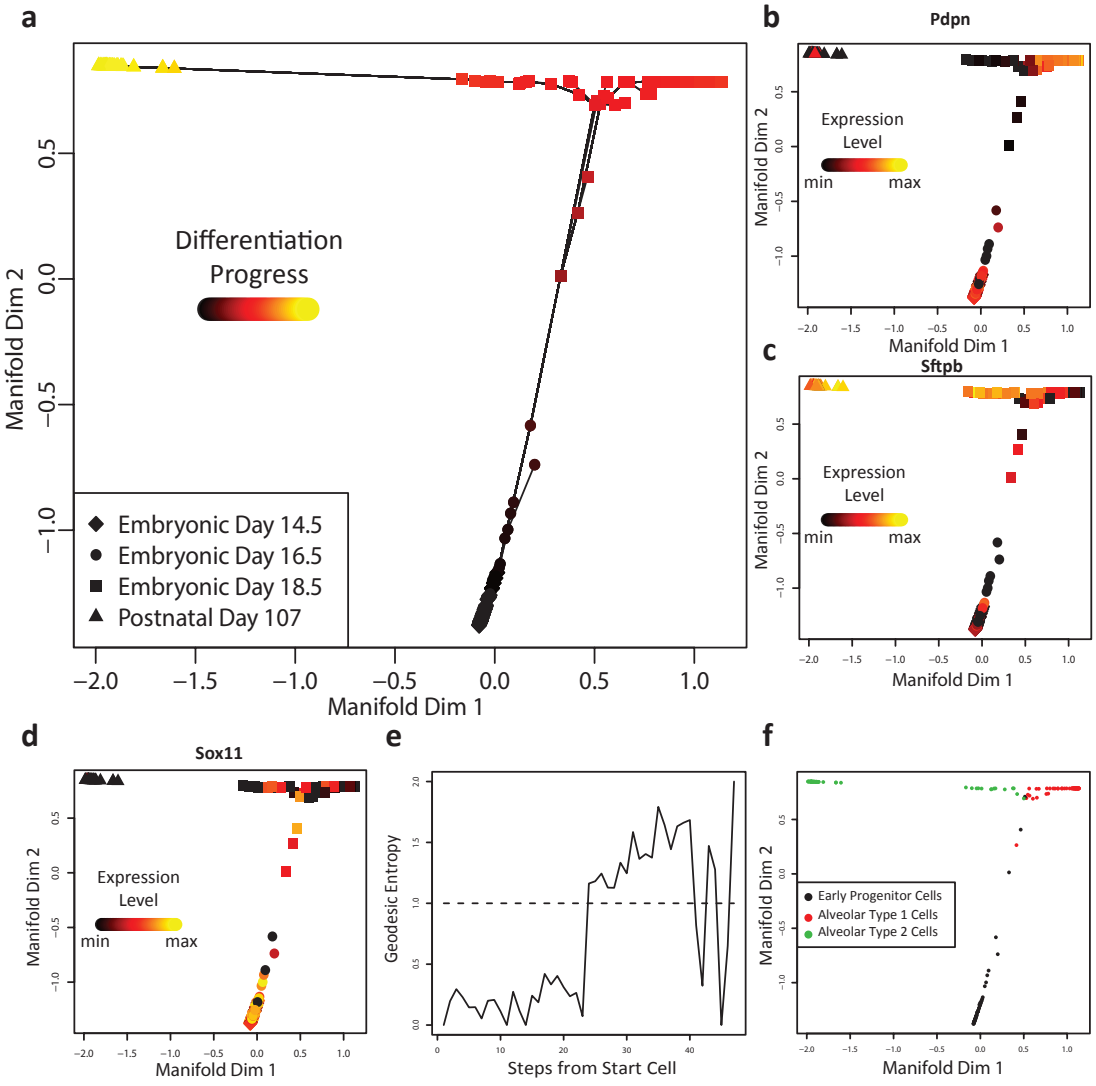



Figure 3.16: SLICER applied to cells from the developing mouse lung. (a) Cellular trajectory inferred by SLICER. The shape of each point indicates the time point (note that this information is used only after the fact for assessing whether the trajectory makes sense, not for constructing it). Color corresponds to inferred geodesic distance from the start cell (differentiation progress). The lines indicate edges used in the shortest paths to each point. Panels (b) through (d) show the expression levels of marker genes in each cell, with the cells ordered by developmental time. Panel b shows a marker for alveolar type 1 cells, c is an alveolar type 2 marker, and d is a marker for early progenitor cells. e Geodesic entropy plot for the trajectory shown in panel a. The dotted line represents an entropy value of 1, the threshold for branch detection. (f) Cells colored according to the branches that SLICER assigned using geodesic entropy. Note that no annotations were used in assigning cells to branches; instead, the interpretations indicated in the legend (AT1, AT2, or EP) were deduced based on marker genes such as those shown in panels b-d after branch assignment

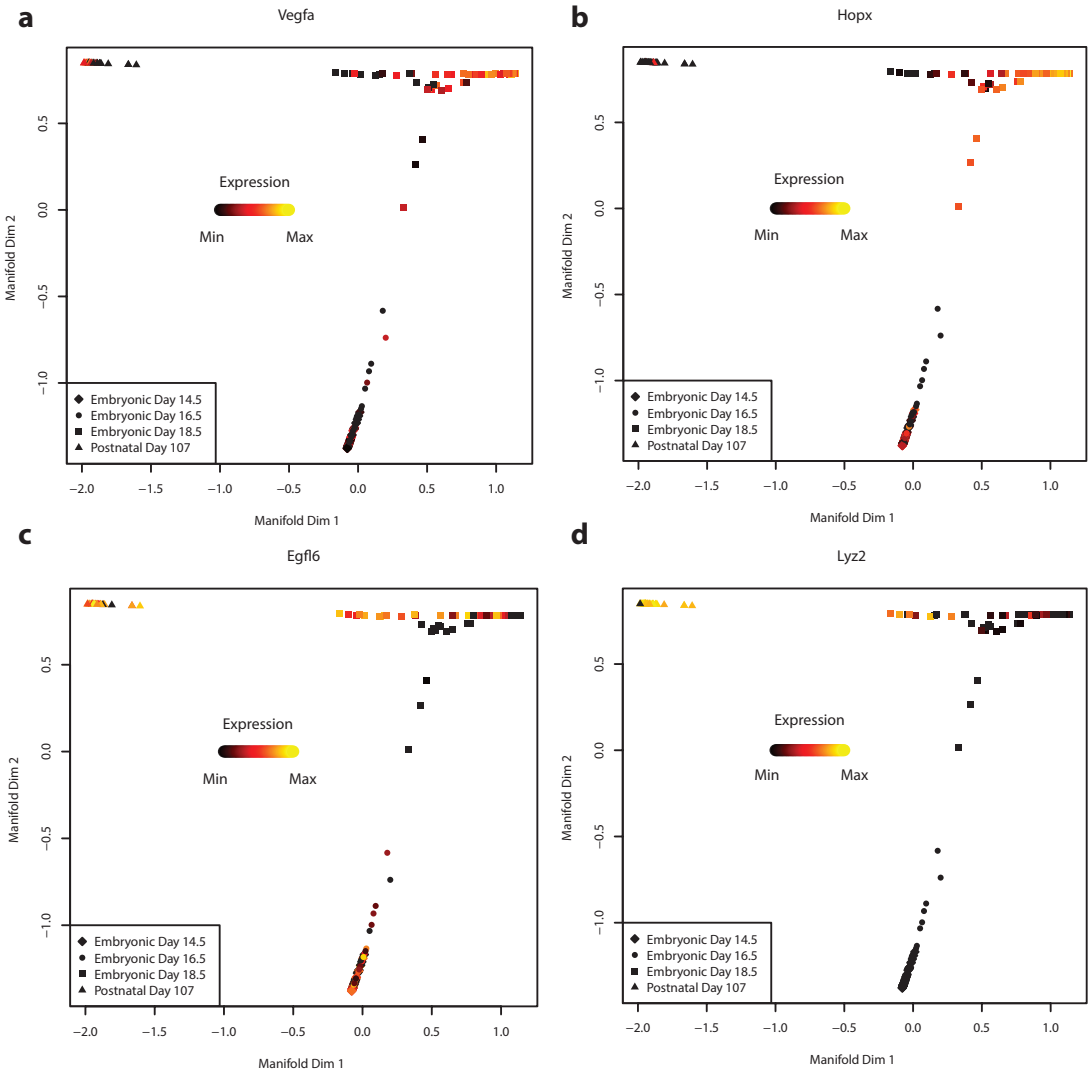



Figure 3.17: Additional marker genes for mouse lung data.

To further investigate the trajectory inferred by SLICER, we examined the expression levels of several genes that were previously validated (Treutlein et al., 2014) as markers of mouse lung development (Fig. 3.16b-d and Fig. 3.17). The AT1 marker gene *Pdpn* should show moderate expression in early progenitor cells, high expression in AT1 cells, and low expression in AT2 cells (Treutlein et al., 2014). As Fig. 3.16b shows, *Pdpn* expression gradually increases along the continuum from early progenitor cells to AT1 cells, matching the expected pattern. Similarly, the AT2 marker *Sftpb* shows increasing expression moving along the trajectory from early progenitors to adult AT2 cells but not AT1 cells (Fig. 3.16c). Additionally, the transcription factor *Sox11*, which plays a role in tissue remodeling during early lung development (Treutlein et al., 2014; Sock et al., 2004), shows decreasing expression levels with increasing distance from the start of the trajectory (Fig. 3.16d). Collectively, the expression patterns of *Pdpn*, *Sftpb*, and *Sox11* confirm that the SLICER trajectory represents a continuum of cells ordered by differentiation progress from early progenitor cells to either AT1 or AT2 cells.

We also used the branch detection capability of SLICER to infer the presence and location of a branch in the differentiation process. Approximately 25 steps from the starting cell, the geodesic entropy of the trajectory exceeds 1, indicating the beginning of a branch (Fig. 3.16e). Based on the above investigation of known marker genes, this location appears to represent a decision point for a differentiating cell, after which a cell proceeds toward either the AT1 or AT2 cell fate. After detecting the existence and location of a branch in the trajectory, we used SLICER to assign each cell to a branch (Fig. 3.16f).

### 3.5.2 Mouse Neural Stem Cells

We also ran SLICER on previously published data from mouse adult neural stem cells (Llorens-Bobadilla et al., 2015). In this study, cells were harvested from the subventricular zones of adult mouse brains with the goal of determining how gene expression changes during neural stem cell activation after a brain injury (Llorens-Bobadilla et al., 2015). Only one cell fell below the cutoff of 1000 genes detected, leaving 271 out of 272 cells.

We again selected genes by comparing sample variance and neighborhood variance. This yielded a list of 661 genes. Figure 3.18a shows the resulting trajectory. The embedding has a clear trajectory-like shape, with most of the cells lying along a horizontal path. There are also two clusters of cells, one close to the main group of cells along the horizontal axis, and one in the upper right corner of the plot.

SLICER has the ability to detect such clusters directly from the low-dimensional $k$-nearest neighbor graph, allowing the user to include or omit certain cell types from trajectory construction (Fig. 3.18b). For example, in the initial analysis of this dataset, the authors discovered the presence of oligodendrocytes, mature neural cells that were extracted at low levels due to overlap with the markers used to isolate neural stem cells. Based on our analysis of marker genes distinguishing oligodendrocytes and neural stem cells (see below), the green cell type in Fig. 3.18b corresponds to oligodendrocytes. SLICER thus gives the ability to easily exclude oligodendrocytes from further analysis, although we chose to retain them because they provide a good example of a trajectory with multiple branches (see below).

To investigate whether the trajectory produced by SLICER is related to the activation of neural stem cells, we examined the expression of known marker genes (Fig. 3.18c-g and Fig. 3.19). The Mki67 gene was previously shown to be a marker for active neural stem cells (aNSCs), and the transcription factor Sox9 is associated with quiescent neural stem cells (qNSCs) (Llorens-Bobadilla et al., 2015). When we colored the trajectory with the expression levels of these marker genes, we found that cells along the x-axis in Fig. 3.18a show gradual variation, with high qNSC marker expression on the right and high aNSC marker expression on the left (Fig. 3.18c-d). This suggests that these cells represent a continuum of states from quiescent to active neural stem cells. The expression of Dcx, a neuroblast marker that is also responsible for the proper migration of differentiating neurons (Llorens-Bobadilla et al., 2015; Ocbina et al., 2006), is expressed at high levels in the cluster of cells near the horizontal axis, indicating that this cluster of cells corresponds to neuroblasts (Fig. 3.18e). The cluster of cells that is far removed from the others shows high expression of the oligodendrocyte transcription factor Sox10 (Llorens-Bobadilla et al., 2015), indicating that these cells are oligodendrocytes (Fig. 3.18f). The Dlx1 gene encodes a neuroblast-associated transcription factor, and it was observed in (Llorens-Bobadilla et al., 2015) that some of the aNSCs also expressed this marker, indicating the initiation of a differentiation program in the aNSCs. Our analysis confirms this result (Fig. 3.18g).

One of the key advantages of SLICER is the ability to identify multiple levels of branches automatically using geodesic entropy, as the synthetic data example in Fig. 3.9 showed. The neural stem cell dataset provides an excellent opportunity to demonstrate this capability on real data because of the presence of three distinct cell types. The geodesic entropy profile of the trajectory indicates a branch about 50 steps from the starting cell. This branching event corresponds to the distinction between aNSCs and neuroblasts (Fig. 3.18h-i). We next computed geodesic entropy recursively on each of the top-level branches identified

(red and green cells shown in Fig. 3.18i). SLICER identified a second branch separating neuroblasts from oligodendrocytes but did not detect a branch in the aNSCs (Fig. 3.20).

Because the cost of single cell RNA-seq depends strongly on the number of cells to be sequenced, the number of cells required to construct a trajectory is an important question. However, the number of cells needed depends strongly on the biological process under consideration. Factors such as the number of branches, relative size of each branch, and extent of the changes across the sampled set of cells all can affect this number. With these caveats in mind, we have addressed this question by investigating, for both of our biological datasets, how much the trajectory changes when SLICER is given a random subset of the cells rather than the full dataset (Fig. 3.21). The results indicate that, for both datasets, the ordering of the cells is relatively stable even with as few as 20% of the cells. The assignment of cells to branches is stable down to 20% of the cells for the distal lung epithelium dataset, but the assignment accuracy steadily declines for the neural stem cell dataset. The reason for this difference is most likely that there are more branches in the neural stem cell dataset, and a smaller proportion of cells occur after the branch points. In contrast, the single branching event in the lung dataset is roughly an even split and occurs midway through the trajectory. Thus, in this case the separation between the cell fates is maintained even when only a few cells are used to build the trajectory.

Comparison with other methods In order to assess the performance of SLICER in relation to other approaches, we ran ICA and Wanderlust on the lung and neural stem cell data and compared the results from all three approaches. We used the set of genes selected by SLICER to ensure that the results from all three approaches were directly comparable. We also set the number of nearest neighbors for Wanderlust to the same values used by SLICER.

The ICA embedding of the mouse lung data in Fig. 3.22a resembles the trajectory inferred by SLICER, detecting a single main path with a prominent branch. However, the arrangement of the points in the embedding is noticeably more diffuse and less "trajectory-like" than the SLICER result shown in Fig. 3.16. In addition, the geometric relationship between early progenitor cells and AT2 cells is somewhat different than that inferred by SLICER (compare Fig. 3.16a and Fig. 3.22a). It appears that tracing a shortest path from early progenitor cells to AT1 cells in Fig. 3.22a would pass through AT2 cells, while the SLICER branching analysis and marker gene expression suggest that these cells should fall on different branches. The ICA embedding of the neural stem cells shows a similar overall shape to the SLICER trajectory (compare Fig. 3.18a and Fig. 3.22b). Once again, however, the overall shape of the ICA embedding is much more

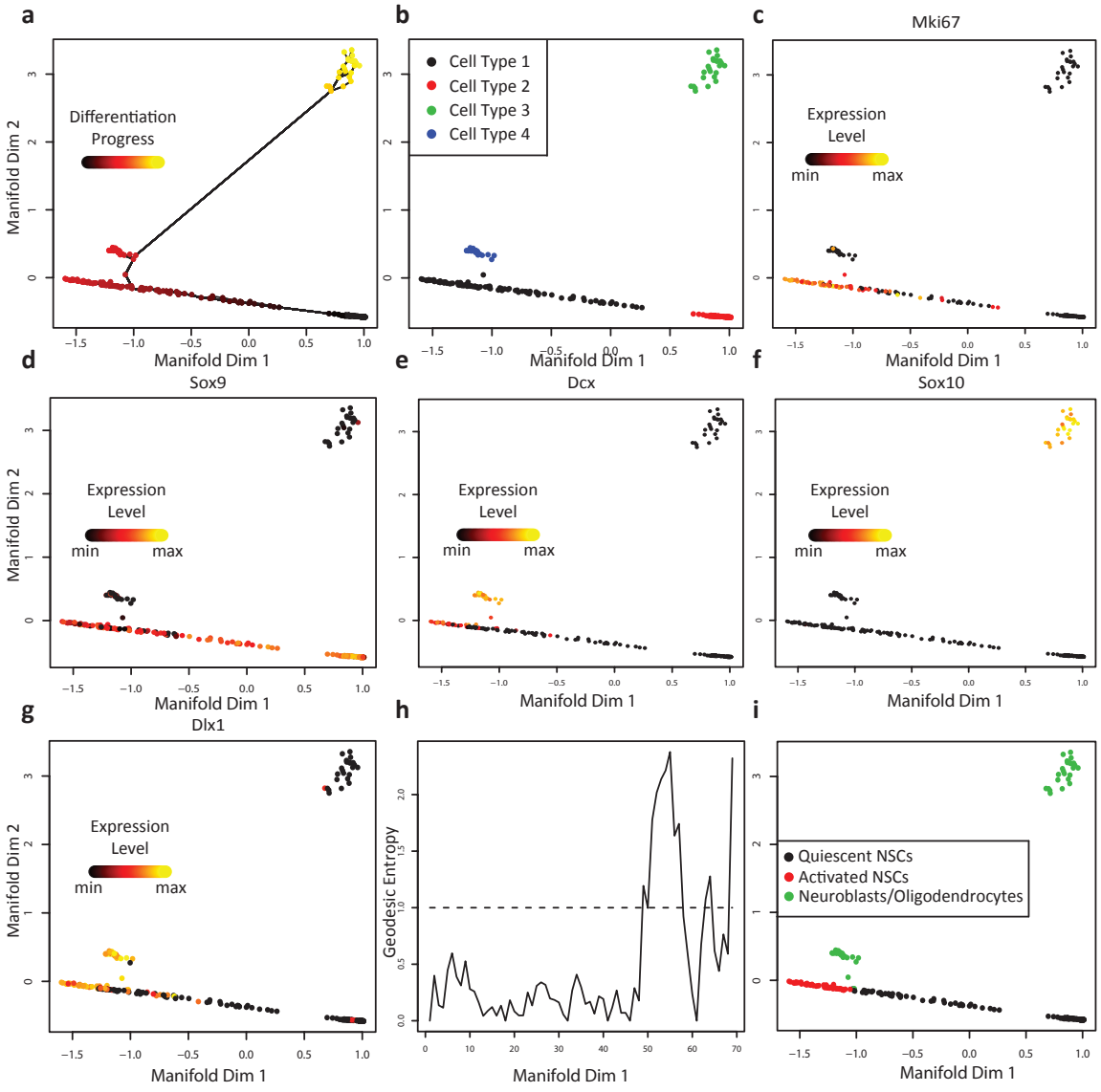



Figure 3.18: SLICER applied to mouse neural stem cells. (a) Cellular trajectory inferred by SLICER. Color corresponds to inferred geodesic distance from the start cell (differentiation progress). The lines indicate edges used in the shortest paths to each point. (b) Clustering using the connected components in the low-dimensional $k$-nearest neighbor graph before trajectory construction identifies four cell types. SLICER provides the option to select which cell types to include when building a trajectory. Panels (c) through (g) show the expression levels of marker genes for different cell types: (c) active neural stem cells, (d) quiescent neural stem cells, (e) neuroblasts, (f) oligodendrocytes, and (g) neuroblasts. (h) Geodesic entropy plot for the trajectory shown in panel (a). The dotted line represents an entropy value of 1, the threshold for branch detection. (i) Cells colored according to the branches that SLICER assigned using geodesic entropy. The interpretations indicated in the legend were deduced based on marker genes such as those shown in panels (c)-(g) after branch assignment

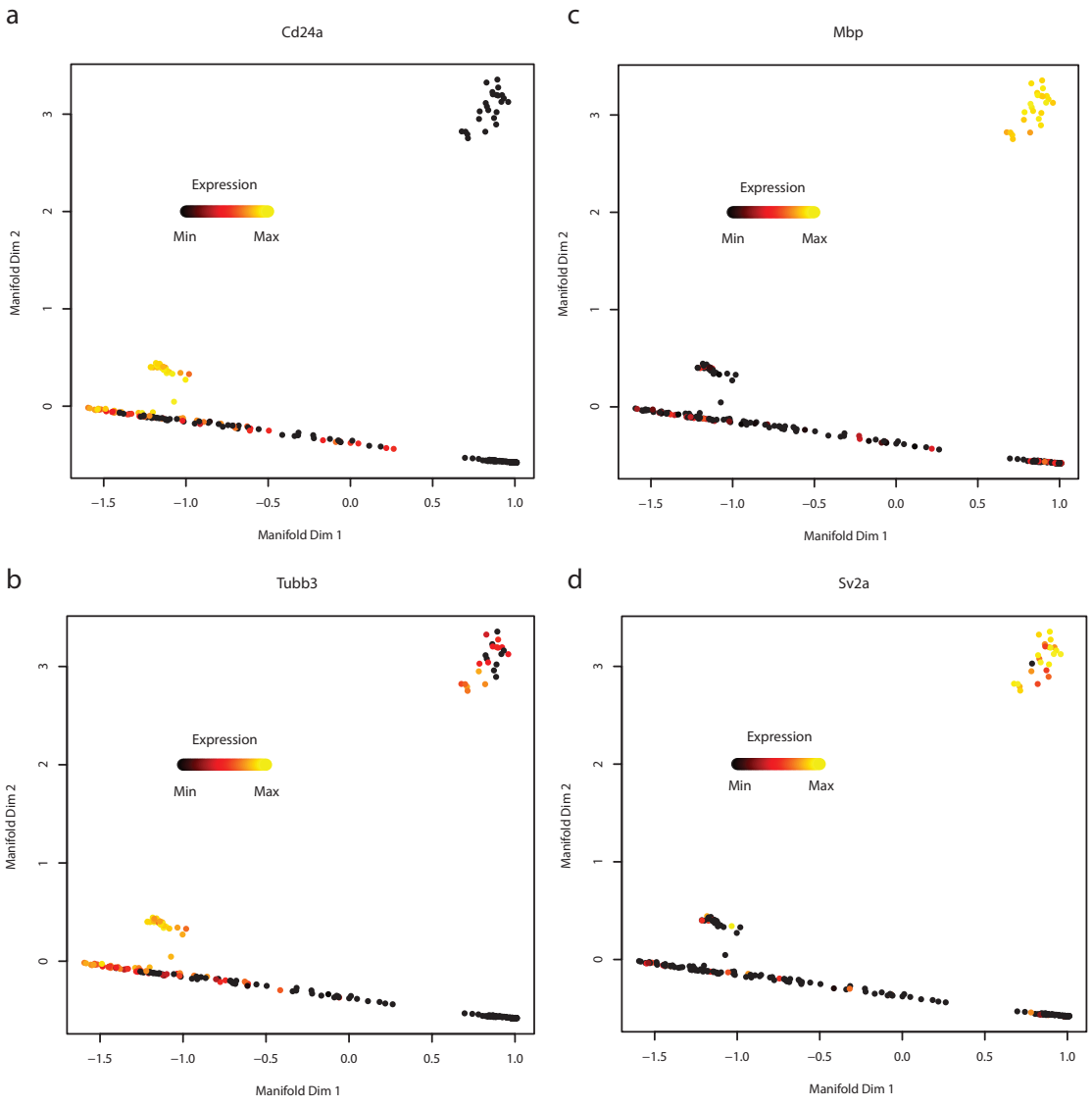



Figure 3.19: Additional marker genes for neural stem cell activation

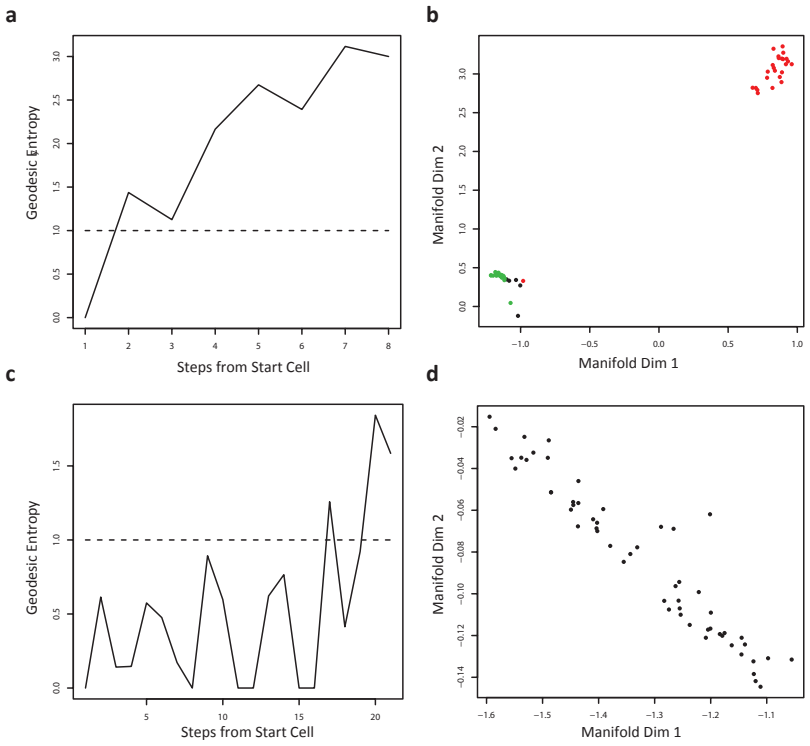
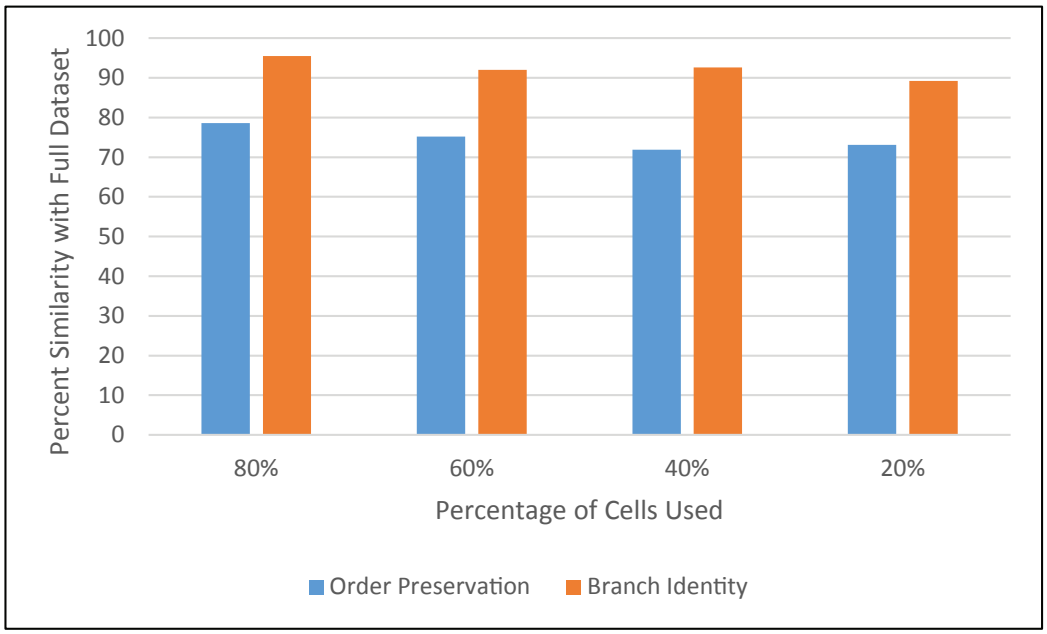



Figure 3.20: Nested branch detection in neural stem cell data.

amorphous, and an ordering of the cells from quiescent to active is much less apparent than in the SLICER trajectory shown in Fig. 3.18a.

Because Wanderlust produces only a one-dimensional ordering of cells rather than a two-dimensional embedding, we plotted the Wanderlust ordering of cells against the SLICER geodesic distance (Fig. 3.22c-d). The two tools agree on the relative ordering of mouse lung cell types, with early progenitor cells preceding AT1 cells and most AT2 cells (Fig. 3.22c). However, it is important to note that because Wanderlust assumes that a trajectory does not branch, the Wanderlust ordering suggests that the lung differentiation process moves from early progenitor cells to AT1 cells, then AT2 cells. In addition to obscuring the true sequence of events in the differentiation process, the existence of multiple cell fates is lost in this approach, underscoring the importance of detecting branches in a trajectory. The Wanderlust ordering of neural stem cells agrees with SLICER on the relative ordering of qNSCs, aNSCs, and neuroblasts (Fig. 3.22d). One exception to note, however, is that Wanderlust places the oligodendrocytes in the middle of the ordering, interleaving them with aNSCs.

a




b

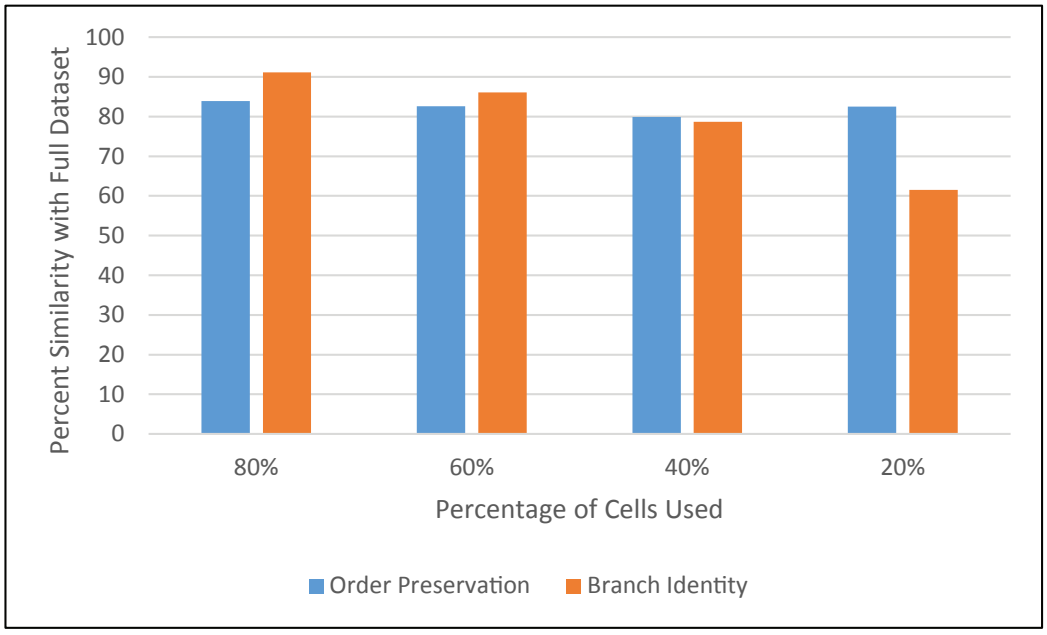



Figure 3.21: Subsampling experiments to estimate the number of cells required for trajectory inference and branch detection.

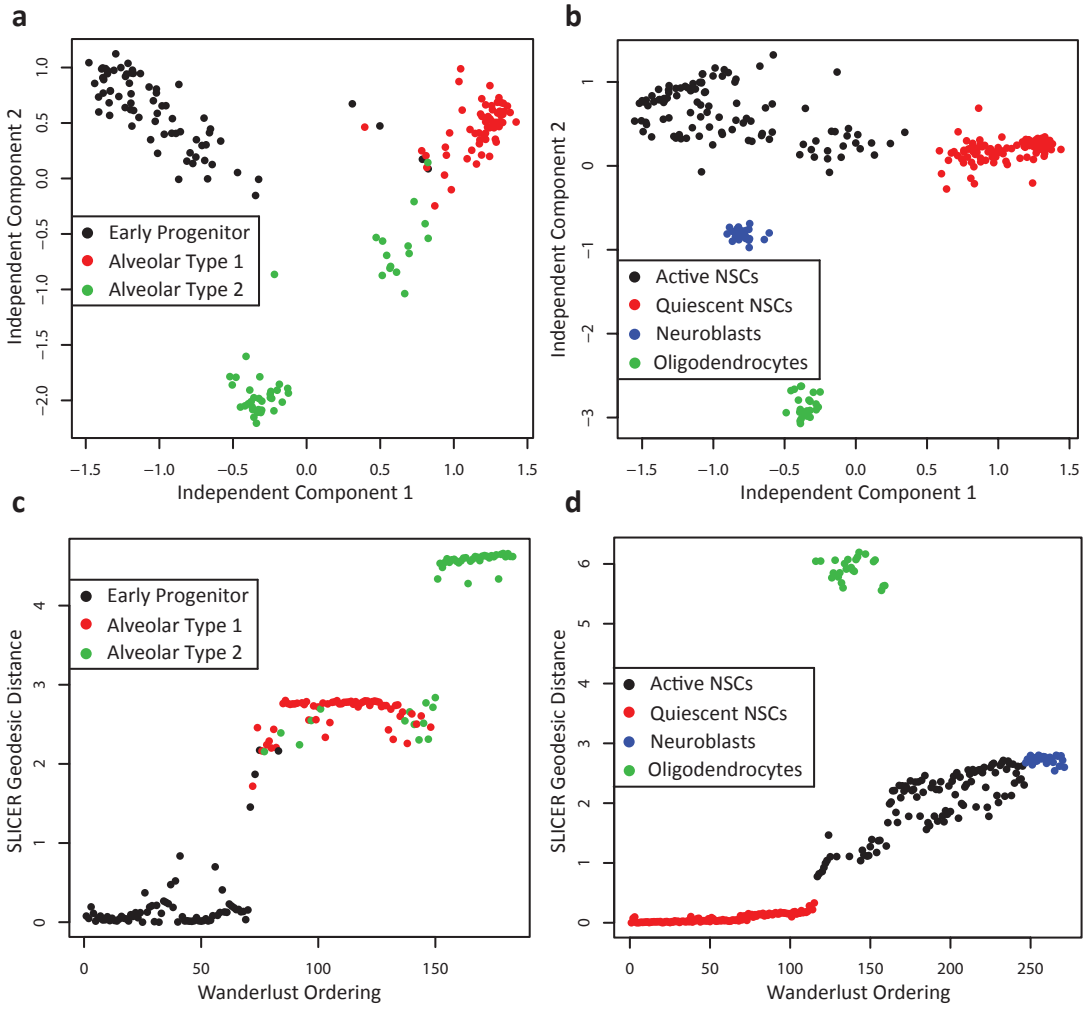



Figure 3.22: ICA and Wanderlust results from mouse lung and neural cells. Note that the genes selected by SLICER were used as input to both ICA and Wanderlust to ensure an accurate side-by-side comparison. (a) ICA embedding of mouse lung cells. The colors correspond to the branch assignments from SLICER. (b) ICA embedding of mouse lung cells. Colors correspond to the SLICER cell type assignments from Fig. 3.18b. (c) Comparison of one-dimensional Wanderlust ordering (x-axis) and SLICER geodesic distance (y-axis) for mouse lung cells. (d) Comparison of one-dimensional Wanderlust ordering (x-axis) and SLICER geodesic distance (y-axis) for mouse neural cells

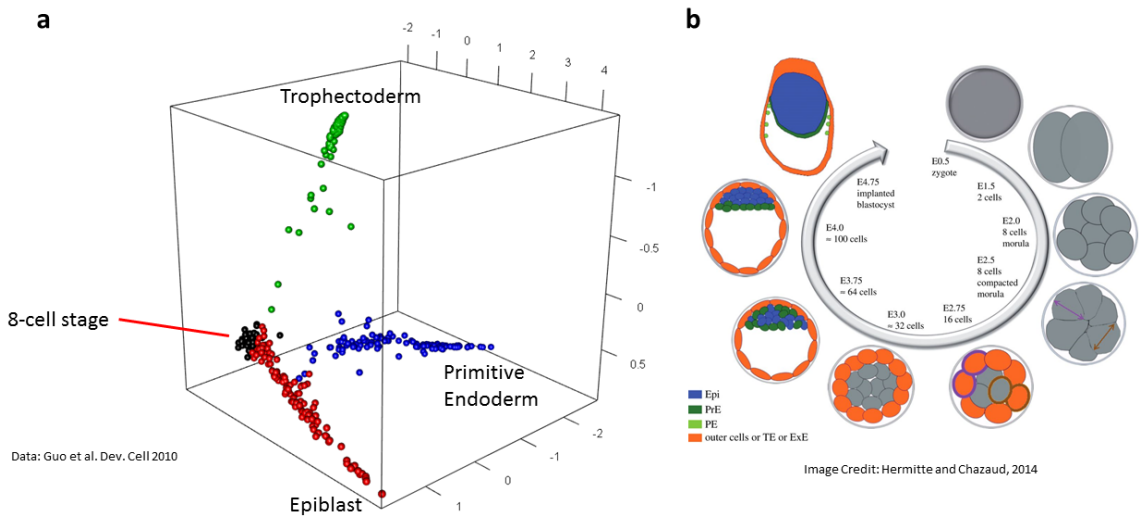



Figure 3.23: SLICER Trajectory and Branch Detection on Differentiating Cells from the Mouse Blastocyst

### 3.5.3 Differentiating Cells from the Early Mouse Embryo

As an additional example of SLICER's branch detection capabilities, we analyzed single cell PCR data from the early mouse embryo (Guo et al., 2010). The dataset contains measurements of 48 genes across 442 single cells from the 8-, 16-, 32-, and 64-cell stages of embryonic development. At the 8-cell stage, cells are completely undifferentiated and have not begun to commit to any particular cell fate (Hermitte and Chazaud, 2014). During the subsequent round of cell division, two distinct types of cells begin to form: the inner cell mass, which will become the embryonic tissue, and the trophectoderm, which will develop into extra-embryonic tissue, including the placenta. By the 64-cell stage, the inner cell mass has specialized into two additional types of cells–epiblast and primitive endoderm (Hermitte and Chazaud, 2014).

SLICER identifies a trajectory that clearly reflects this series of progressive specialization events (Fig. 3.23). Starting from the 8-cell-stage cells, SLICER identifies a branching event that separates trophectoderm cells (green branch) from the inner cell mass (red branch). SLICER also detects a second level of branching, in which the inner cell mass subsequently branches again into epiblast (red) and primitive endoderm (blue) cell types.

### 3.5.4 Direct Cardiac Reprogramming

Heart disease is a leading cause of death nationwide. In North Carolina, heart disease caused 18,474 deaths in 2015, or 21% of all deaths in the state (American Heart Association, 2015). The economic burden of treating heart disease is staggering: total hospital charges for heart disease in North Carolina exceed $4.1 billion annually and have increased, even after inflation adjustment, by nearly 80% since 1995 (Tchwenko, 2012). A key reason why heart disease is so deadly is that heart muscle tissue cannot regenerate, preventing the repair of tissue damage such as that caused by a heart attack. In addition, as heart disease leads to the death of heart muscle cells, scar tissue forms around large patches of dead cells. This scar tissue is less flexible than muscle tissue and cannot beat along with the surrounding muscle tissue, placing additional strain on the heart and often causing dangerous arrhythmias.

One approach to address heart tissue damage is direct cardiac reprogramming (Ieda et al., 2010; Qian et al., 2012), in which the introduction of certain factors induces heart scar tissue cells to turn directly into heart muscle cells. Because it provides a source of new heart muscle cells, direct cardiac reprogramming offers a promising solution to the muscle cell loss caused by heart disease. An additional benefit is that reprogramming of scar tissue cells into heart muscle cells also removes problematic patches of scar tissue. A current limitation of the direct cardiac reprogramming approach is that only a small percentage of cells exposed to the reprogramming factors actually become heart muscle, while the majority remain scar tissue. The relative inefficiency of the reprogramming process restricts the regenerative potential of this therapy. In turn, improving reprogramming efficiency requires precise mechanistic understanding of the changes required to convert a connective tissue cell into a heart muscle cell. We know that, in general, scar tissue genes must be gradually turned off and heart muscle genes must be gradually turned on during reprogramming. But it is not clear which genes must be turned on and off or in what sequence, nor is it understood how the reprogramming factors elicit this sequence of changes.

After developing SLICER, we began a collaboration with Prof. Li Qian from the UNC Pathology Department investigating the gene expression changes involved in the process of direct cardiac reprogramming. The Qian lab performed single cell RNA-seq on mouse cardiac fibroblast cells before and three days after treatment with the reprogramming factors. After filtering, we retained a total of 454 single cells for further analysis.

Using SLICER, I constructed a trajectory that summarizes the changes cells undergo during reprogramming. A manuscript describing our analysis is still under review, and we cannot disclose the details of these results without jeopardizing this publication.

### 3.5.5 Tumor Subtypes Related to Stages of Normal Differentiation

Cancer has long been understood to be a broad category of heterogeneous maladies, but recent large-scale tumor sequencing efforts have identified intrinsic cancer subtypes based on molecular properties. For example, the Cancer Genome Atlas project (TCGA) has identified intrinsic cancer subtypes in bladder, breast, head/neck, lung cancer, and many others based on tumor gene expression, DNA mutations, miRNA expression, and DNA methylation. Now that these tumor subtypes have been identified, an important question is how they arise. One hypothesis is that at least some of these subtypes reflect distinct stages in normal cellular differentiation processes. For example, the classic French-American-British (FAB) leukemia classification system is based on the differentiation stage of the malignant blood cells, as determined by visible cell morphology (Fig. 3.24b). Each of the leukemia subtypes arises from a mutational event that blocks blood cell differentiation at a specific point in normal development (Somasundaram et al., 2015).

There is also evidence that the breast cancer subtypes correspond to stages in normal mammary development (Prat et al., 2009). Figure 3.24a summarizes this model. Microarray analysis of bulk tissue samples purified using flow cytometry showed that the basal subtype closely resembles a mammary progenitor cell type (basal cell) that gives rise to the epithelial cell type that lines the lumen of the milk duct (luminal cell). Similarly, the luminal A and luminal B subtypes also resemble the more differentiated luminal cells. No cell population directly corresponding to the Her2 subtype was observed in the microarray analysis, but Prat and Perou showed that its expression profile is consistent with a developmental intermediate between basal and luminal cells. The rare claudin-low subtype appears to be most similar to a mammary stem cell population (Prat et al., 2010, 2013).

Many of the most common types of cancer occur in epithelial tissues, including lung, skin, bladder, breast, colorectal, kidney, head/neck, ovarian, cervical, and endometrial cancers. Many of these tissues have a similar physical structure (stratified epithelium), with some sort of basal layer and luminal layer. In addition, recent pan-cancer analyses have indicated that bladder, breast, head/neck, and lung cancers have similar subtypes and look alike at the molecular level, to the extent that some subtypes from one tissue may more closely resemble subtypes from another tissue than subtypes from the same tissue (Hoadley et al.,

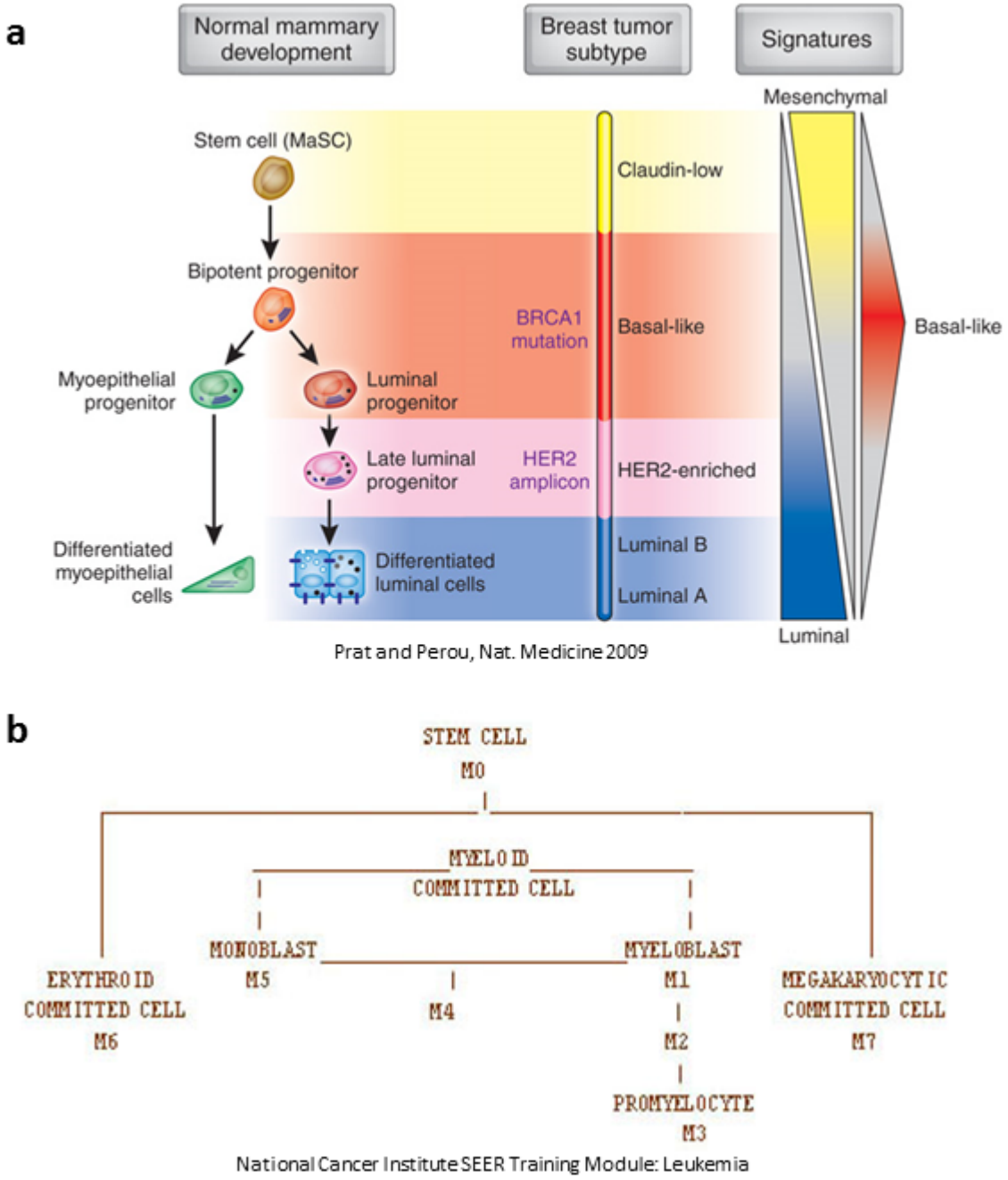



Figure 3.24: SLICER Results from TCGA Breast Cancer and Leukemia Bulk RNA-seq Data

2014). These results suggest the intriguing possibility that a similar developmental cascade may underlie the subtypes in each of these diseases. If this hypothesis turns out to be true, it would provide a unifying explanation for cancer development across multiple tissue types. In addition, this hypothesis predicts that common mutational events linked to stages of differentiation could occur across tissue types. In what may involve a bit of wishful thinking, another possibility is that common treatment options could prove successful per differentiation stage, largely independent of tissue.

We explored this hypothesis by using SLICER to construct trajectories from bulk RNA-seq data generated by TCGA. Note that a bulk RNA-seq sample contains an aggregate signal from millions of cells, and thus is not the sort of data SLICER was originally designed to analyze. However, the preceding discussion suggests that it may be reasonable to look for a trajectory in bulk tumor data. The assumption behind such an analysis is that the dominant source of variation in a collection of tumor samples from individuals with cancer will be the differentiation stages of the initial cell populations that gave rise to the tumors.

We used SLICER to construct a trajectory from over 800 TCGA breast cancer samples (Fig. 3.25a). Consistent with the model proposed by Prat and Perou, the trajectory begins with basal tumors, passes through Her2 tumors, and ends with luminal tumors. There are no claudin-low tumors in this dataset, but we re-analyzed an older microarray dataset and confirmed that SLICER predicts a very similar trajectory, with claudin-low samples preceding basal samples (data not shown). Interestingly, SLICER predicts a branch in the trajectory, with the branch somewhat separating luminal A and luminal B subtypes. The biological interpretation of this branch is not completely clear, although a related analysis suggests that the branch may be related to proliferative capability (Prat et al., 2013). Note that the branch in the SLICER trajectory does not correspond to the myoepithelial/luminal distinction in Fig. 3.24a; for incompletely understood reasons, it is very rare for tumors to develop from mammary myoepithelial cells, and there are no such tumors in the TCGA dataset that we analyzed. As an interesting aside, SLICER selected many more genes in this case than in the preceding single cell analyses. We suspect that this may be a combination of the much higher sensitivity of bulk RNA-seq and the aggregate nature of the data (the differentiation signal is likely not as pure as in the single cell case).

We also analyzed 200 samples from patients with acute myeloid leukemia (Fig. 3.25b). The points are colored according to the FAB classification assigned by the pathologist who examined each patient sample. There is a clear relationship between the SLICER trajectory and the tumor labels. The M0 samples (black) are in the middle of the trajectory, with the M1/M2 samples to the left and M3/M4 samples to the right. This

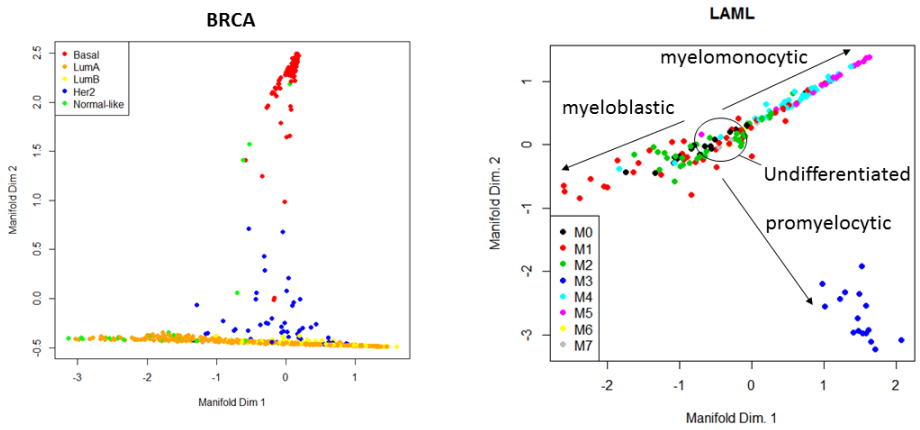



Figure 3.25: SLICER Results from TCGA Breast Cancer and Leukemia Bulk RNA-seq Data

pattern closely matches the established myeloid differentiation hierarchy shown in Fig. 3.24b: M1 and M2 designations are assigned to malignancies with cells differentiating into myeloblasts, while the M3 and M4 designations recognize the presence of cells differentiating into monoblasts. Note that very few M6 and M7 samples are present in the dataset (6 total). These would likely show up as additional branches diverging from the M0 cells, because M6 and M7 designations recognize the presence of erythrocytes and megakaryocytes, respectively, which are distinct types of blood cells. A surprising result is that the promyelocytic (M3) leukemias do not show up at the end of the continuum beyond the M2 samples. Instead, they form a distinct cluster of their own. The reason for this pattern is unclear. To rule out the possibility that the trajectory-like shape of the data is an artifact caused by distinct clusters (as shown by simulation in the Pitfalls section above), we repeated the trajectory analysis without the M3 samples and obtained nearly identical results. Thus, our results suggest that the developmental history of promyelocytic leukemia is somehow distinct from the M0-M2 and M4-M5 cancers. Further work is needed to ascertain the nature of this difference.

We also performed similar analysis on TCGA bladder, cervical, endometrial, head/neck, lung, and ovarian cancers. We hoped to find evidence of developmental processes underlying tumor subtypes in these types of cancers, and ultimately to explore the possibility of common epithelial differentiation patterns across cancer types. Our results (data not shown) suggest that differentiation cascades may help explain

the properties of tumor subtypes in these cancers, as well, and even hint at cross-cancer gene signatures of epithelial differentiation. However, the lack of a single-cell-resolution understanding of the gene expression profiles of the corresponding normal differentiation processes for each of these tissue types prevents us from drawing any solid conclusions. For example, the gene expression profiles of human lung cell types and how they develop during normal differentiation have not been systematically measured. This makes it very difficult to determine what cell type or stage of differentiation could give rise to a particular lung cancer subtype. Additionally, our lack of knowledge about which normal cell types look most similar to a particular subtype of lung cancer make it very difficult to identify the most similar cell type and corresponding tumor subtype in mammary tissue. In short, this direction of research seems promising, but cannot produce definitive results until our understanding of normal tissue composition and differentiation advances.

# CHAPTER 4
# Enabling Single Cell Multi-Omics Using Manifold Alignment

## 4.1 Background and Related Work

Understanding the mechanisms that regulate gene expression across space and time is a fundamental challenge in biology. Epigenetic modifications such as DNA methylation, histone marks, and chromatin accessibility are known to regulate gene expression, but the precise details of this regulation are not well understood. Single cell genomic technologies reveal heterogeneity within populations of cells, including complex tissues, tumors, and cells undergoing temporal changes (Sandberg, 2013; Shapiro et al., 2013). Furthermore, because bulk data consist of measurements averaged across a population of cells, single cell genomic data enable, in principle, much more precise study of how epigenetic changes and gene expression vary together.

Single cell RNA-seq has recently been applied with great success to the study of sequential cellular processes such as differentiation and reprogramming (Trapnell et al., 2014; Llorens-Bobadilla et al., 2015; Macaulay et al., 2016; Hanchate et al., 2015; Treutlein et al., 2014). In such experiments, each sequenced cell is assumed to be at one point in the process, and sequencing enough cells can reveal the progression of gene expression changes that occur during the process (Kolodziejczyk et al., 2015a; Welch et al., 2016a). More recently, several experimental techniques for performing single cell epigenetic have been developed (Nagano et al., 2013; Smallwood et al., 2014; Rotem et al., 2015; Buenrostro et al., 2015; Angermueller et al., 2016; Jin et al., 2015; Zhu et al., 2017; Mooijman et al., 2016), and several studies have demonstrated that single cell epigenetic data can be also used to elucidate the series of changes in a sequential process (Zhu et al., 2017; Farlik et al., 2015; Corces et al., 2016).

Identifying correlations among epigenome and transcriptome dynamics would allow more complete understanding of the sequential changes that cells undergo during biological processes. Measuring multiple genomic quantities from a single cell, or multi-omic profiling [20,21], would be the best way to identify such correlations. Unfortunately, performing single cell multi-omic profiling is very difficult experimentally,

because an assay on chromatin or RNA destroys the respective molecules and only tiny amounts of DNA and RNA are present in a single cell. In certain cases, it is possible to assay RNA and DNA (Angermueller et al., 2016; Dey et al., 2015; Macaulay et al., 2015; Hou et al., 2016) or RNA and proteins (Darmanis et al., 2016; Genshaft et al., 2016) from the same single cell, but experimentally performing multiple assays on either chromatin or RNA from the same cell is extremely challenging.

Our knowledge of epigenetic regulation suggests that any large changes in gene expression, such as those that occur during differentiation, are accompanied by epigenetic changes. Therefore, it should be possible, in principle, to infer sequential changes in cellular epigenetic state during a process. Furthermore, if cells undergoing a common process are sequenced using multiple genomic techniques, examining any of the genomic quantities should reveal the same underlying biological process. For example, the main difference among cells undergoing differentiation will be the extent of their differentiation progress, whether you look at the gene expression profiles or the chromatin accessibility profiles of the cells.

We reasoned that this property of single cell data could be used to infer correspondence between different types of genomic data. To infer single cell correspondences, we use a technique called manifold alignment (Ham et al., 2005; Chang Wang and Sridhar Mahadevan, 2009). Intuitively, manifold alignment constructs a low-dimensional representation (manifold) for each of the observed data types, then projects these representations into a common space (alignment) in which measurements of different types are directly comparable. To the best of our knowledge, manifold alignment has never been used in genomics. However, other application areas recognize the technique as a powerful tool for multimodal data fusion, such as retrieving images based on a text description, and multilingual search without direct translation (Chang Wang and Sridhar Mahadevan, 2009). We refer to our method as MATCHER (Manifold Alignment to CHaracterize Experimental Relationships). Using MATCHER, we identified correlations between transcriptomic and epigenetic changes in single mouse embryonic stem cells and single human induced pluripotent stem cells.

## 4.2 Overview of MATCHER

Manifold alignment is an approach for integrating multiple types of data that describe different aspects of a common phenomenon. For example, a video of a person speaking, an audio recording of the speech, and a written transcript of the words uttered all describe a common set of events from different perspectives. The key idea of manifold alignment, as initially proposed by Ham et al. (Ham et al., 2003), is to integrate

multiple data types by discovering the common manifold structure that underlies them. In many real-world settings, the assumption of a common underlying manifold generating multiple data types is a reasonable one. There are two main types of manifold alignment, distinguished by whether they require examples of precisely corresponding measurements to align manifolds (manifold alignment with correspondence) (Ham et al., 2003) or simply use geometric information (manifold alignment without correspondence) (Wang et al., 2009). Gaussian process latent variable models have also been used to perform manifold alignment by learning completely (Ek, 2009; Eleftheriadis et al., 2015) or partially (Damianou et al., 2012) shared latent representations of high-dimensional, multimodal data. Given a set of images and corresponding text descriptions, manifold alignment can be used to identify a low-dimensional representation that allows the prediction of a caption for a new image. This somewhat analogous to the problem of retrieving a corresponding epigenetic measurement for a given single cell transcriptome. However, in the context of single cell genomic data, correspondence information is not generally available to train a model, because it is impossible in most cases to measure more than one quantity on a single cell. Therefore, we developed a novel approach for manifold alignment without correspondence that leverages the unique aspects of this problem. We assume that:

1. Single cell genomic data from cells proceeding through a biological process lie along a one-dimensional manifold. Another way of saying this is that the variation among cells can be explained mainly by a single latent variable ("pseudotime") corresponding to position within the process.

2. Each of the genomic quantities under consideration changes in response to the same underlying process.

3. The biological process is monotonic, meaning that progress occurs only in one direction. Processes that alternate between forward and backward progress or repeat cyclically would violate this assumption.

4. The cells in each experiment are sampled uniformly at random from the same population, process, and cell type.

Given these assumptions, there are only three possible types of differences among the one-dimensional manifold representations of each data type: orientation, scale, and "time warping" (Fig. 4.1a). We can perform manifold alignment without correspondence information by accounting for these three types of differences. Differences in orientation can occur if the biological process corresponds to increasing manifold coordinates for one type of genomic data but decreasing coordinates for another data type. We can reconcile

different orientations by simply reversing the order of one set of manifold coordinates. It is not possible to infer the correct orientation from data, so we use biological prior knowledge to choose the correct orientation for the manifold inferred from each type of data. To address scale differences, we can normalize the manifold coordinates to lie between 0 and 1. Time warping effects can occur if different genomic quantities change at different rates. For example, gene expression changes may occur slowly at the beginning of a process and gradually speed up, while changes in chromatin accessibility may show exactly the opposite trend during the process (Fig. 4.1a). We account for time warping effects by learning a monotonic warping function for each type of data (see below for details).

We use a Gaussian process latent variable model (GPLVM) to infer pseudotime values separately for each type of data. A GPLVM is a nonlinear, probabilistic, generative dimensionality reduction technique that models high-dimensional observations as a function of one or more latent variables (Lawrence, 2004). The key property of a GPLVM is that the generating function is a Gaussian process, which allows Bayesian inference of latent variables nonlinearly related to the high-dimensional observations (Titsias and Lawrence, 2010; Damianou et al., 2016). The nonlinear nature of this model makes it more flexible and robust to noise than a linear model such as principal component analysis (PCA). In fact, PCA can be derived as a special case of a GPLVM in which the Gaussian process generating function uses a linear kernel (Lawrence, 2004). Importantly, GPLVMs are also generative models, meaning that they can answer the counterfactual question of what an unobserved high-dimensional datapoint at a certain location on a manifold would look like. The generative nature of GPLVMs is particularly important to our approach: We use this property to infer correspondence among single cell genomic quantities measured in different ways. We note that GPLVMs have previously been used to infer latent variables underlying differences among single cell gene expression profiles (Buettner et al., 2015; Reid and Wernisch, 2015; Campbell and Yau, 2016); our approach differs from these previous approaches in that we use GPLVMs to perform manifold alignment and generate measurements from unobserved cells to integrate multiple types of single cell measurements.

After inferring pseudotime separately for each type of data, we learn a monotonic warping function (Fig. 4.1b-c) that maps pseudotime values to "master time" values, which are uniformly distributed between 0 and 1 (Fig. 4.1d). This is equivalent to aligning the quantiles of the pseudotime distribution to match the quantiles of a uniform random variable. Master time values inferred from different data types are then directly comparable, corresponding to the same points in the underlying biological process.

The model (Fig. 4.1e) that we use to infer master time values allows us to generate corresponding cell measurements even from datasets where the measurements were performed on different single cells. The different types of measurements may produce datasets with cells from different positions in the biological process, and even different numbers of cells (Fig. 4.1e). To generate a corresponding measurement for a cell, we take the master time value inferred for a given cell, such as one measured with RNA-seq. Then we map this master time value through the warping function to a pseudotime value for a different type of data, such as ATAC-seq. Using the GPLVM trained on ATAC-seq data, we can output a corresponding cell based on this pseudotime value. As Fig. 4.1f shows, the generative nature of the model allows MATCHER to infer what unobserved cells measured with one experimental technique would look like if they corresponded exactly to the cells measured using a different technique. These corresponding cell measurements can then be used in a variety of ways, such as computing correlation between gene expression and chromatin accessibility.

Although it is very difficult in general to measure multiple genomic quantities on the same single cell, one particular protocol (scM&T-seq) has been developed for measuring DNA methylation and gene expression in the same single cell (Angermueller et al., 2016). It is possible that future protocols will enable other joint measurements. In such cases, MATCHER can perform manifold alignment with correspondence using a shared GPLVM (Ek, 2009) to infer a shared pseudotime latent variable for both data types.

## 4.3 MATCHER Method Details

### 4.3.1 Inferring Pseudotime

We infer pseudotime using a Gaussian process latent variable model (GPLVM) with a single latent variable $\mathbf{t}$. For a more thorough introduction to Gaussian processes and GPLVMs, see Rasmussen (Rasmussen et al., 2006) or Damianou (Damianou et al., 2016). Under our model, the observed high-dimensional data (RNA-seq, ATAC-seq, ChIP-seq, DNA methylation, etc.) are generated from $\mathbf{t}$ by a function $f$ with the addition of Gaussian noise:

$$\mathbf{Y} = f(\mathbf{t}) + \epsilon \tag{4.1}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. The key property of a GPLVM is that the prior distribution of $f$ is a Gaussian process:

$$f(\mathbf{t}) \sim GP(\mathbf{0}, k(\mathbf{t}, \mathbf{t}')) \tag{4.2}$$

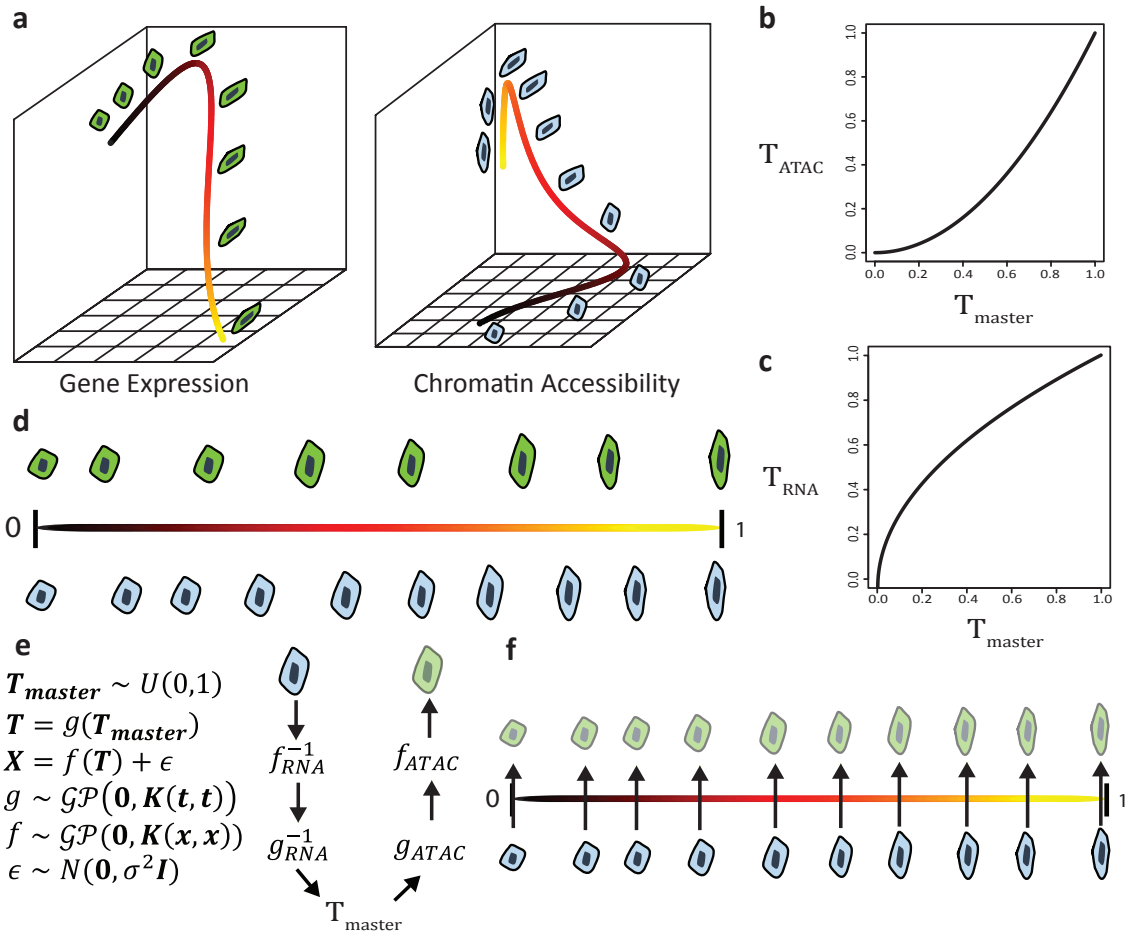



Figure 4.1: MATCHER Method Overview (a) We infer manifold representations of each dataset using a Gaussian process latent variable model (GPLVM). However, the resulting "pseudotime" values from different genomic data types are not directly comparable due to differences in orientation, scale, and "time warping". Both the color of the curve (black to yellow) and cell morphology (blob to oblong) indicate position within this hypothetical process. (b)-(c) To account for these effects, pseudotime for each kind of data is modeled as a nonlinear function (warping function) of master time using a Gaussian process. (d) MATCHER infers "master time" in which the steps of a biological process correspond to values uniformly distributed between 0 and 1 and are comparable among different data types. However, different datasets are measured from different physical cells, and thus may sample different points in the biological process and even different numbers of cells. (e) Diagram showing how MATCHERs generative model can infer corresponding cell measurements. The generated cell is drawn with transparency to indicate that this is an inferred rather than observed quantity. (f) Applying MATCHER to multiple types of data provides exactly corresponding measurements from observed cells and unobserved cells (indicated with transparency) generated by MATCHER.

A linear kernel yields a model equivalent to probabilistic PCA, but if we choose the kernel function $k$ to be nonlinear, the GPLVM can infer nonlinear relationships between $\mathbf{t}$ and $\mathbf{Y}$. We use the popular radial basis function (RBF) kernel, also called the squared exponential kernel.

$$k(t_i, t_j) = \sigma_{rbf}^2 \exp\left(-\frac{1}{2l^2}(t_i - t_j)^2\right) \tag{4.3}$$

Because a Gaussian process is a collection of random variables for which the covariance of any finite set is a multivariate Gaussian, we have:

$$P(Y|t, \sigma^2, {}^2_{rbf}, l) = \mathcal{N}(\mathbf{Y}|\mathbf{0}, K_{ff} + \sigma^2 I) \tag{4.4}$$

where $K_{ff}$ is the covariance matrix defined by the kernel function $k$. A simple approach to inferring the latent variable $\mathbf{t}$ would be to find the values that maximize the posterior distribution:

$$\mathbf{t_{MAP}} = \arg\max_t P(\mathbf{Y}|\mathbf{t})P(\mathbf{t}) \tag{4.5}$$

Instead of MAP estimation, we use the method of Damianou (Damianou et al., 2016), which estimates the posterior using a variational approximation. A key advantage of this approach is that it provides a distributional estimate of the latent variables rather than just a point estimate. The approximation relies on the introduction of auxiliary variables called inducing inputs to derive an analytical lower bound on the marginal likelihood. Inference is then performed by maximizing the lower bound with respect to the inducing inputs and the hyperparameters $\sigma^2$, $\sigma_{rbf}^2$, and $l$. We used 10 inducing inputs for all of our analyses, although we confirmed that the results are robust to the number of inducing inputs used. We used the Bayesian GPLVM model implemented in the GPy package, with the default initialization setting, which uses PCA to determine the initial values for the latent space before optimization. To infer shared master time from simultaneous measurements (such as scM&T-seq or sc-GEM), we first use a shared GPLVM (Ek, 2009) to infer pseudotime, then proceed to infer a warping function in the same way as for pseudotime values inferred from a regular GPLVM (see next section for details). The shared GPLVM model extends the regular GPLVM by assuming that multiple types of high-dimensional data (such as gene expression and DNA methylation measurements)

$\mathbf{Y^{(1)}}$,$\mathbf{Y^{(2)}}$ are generated from a shared latent space through different mapping functions:

$$\mathbf{Y^{(1)}} = f_1(\mathbf{t}) + \epsilon_\mathbf{1} \tag{4.6}$$

$$\mathbf{Y^{(2)}} = f_2(\mathbf{t}) + \epsilon_\mathbf{2} \tag{4.7}$$

As with the regular GPLVM, we used an RBF kernel k to calculate covariance among points in the latent space; however, for the shared GPLVM, each data type has a separate set of hyperparameters $\sigma^2$, $\sigma^2_{rbf}$, and $l$. The shared GPLVM model is a special case of a more general technique called manifold relevance determination, in which latent dimensions can be weighted differently in the covariance function for each data type (Damianou et al., 2012). The manifold relevance determination model uses an automatic relevance determination (ARD) kernel with a separate weight for each latent dimension. For example, for the RBF automatic relevance determination kernel is:

$$k(\mathbf{t_i}, \mathbf{t_j}) = \sigma^2_{rbf} \exp\left\{-\frac{1}{2}\sum_k w_k(t_{ik} - t^2_{jk})\right\} \tag{4.8}$$

Using a separate set of weights $\mathbf{w^{(1)}}$ and $\mathbf{w^{(2)}}$ for each data type allows the model to assign the latent dimensions weights that differ between data types. We use the manifold relevance determination model implemented in GPy but constrain the model to use an ordinary RBF kernel rather than an ARD kernel. This model is thus equivalent to a shared GPLVM. The GPy implementation of manifold relevance determination uses a variational approximation to estimate the posterior, optimizing the evidence lower bound with respect to separate hyperparameters $\sigma^2$, $\sigma^2_{rbf}$, and $l$ for each data type. We use the default initialization provided in GPy, which initializes the value of the latent space by performing PCA on the concatenated datasets.

### 4.3.2 Learning Warping Functions

To learn warping functions from pseudotime to master time, we compute the sample quantiles of pseudotime for a specified number of quantiles, then align these sample quantiles with the theoretical quantiles of a uniform (0,1) random variable. More precisely, we treat the sample quantiles of pseudotime as the independent values of an unobserved function and the theoretical quantiles of a uniform (0,1) random variable as the dependent values of the function. Then we use either Gaussian process regression or linear interpolation to approximate the warping function that maps a pseudotime value to a master time value. We

used 50 quantiles for all analyses in the manuscript, but found that the warping functions are robust to the number of quantiles used. Gaussian process regression is an attractive choice for learning a warping function due to the capability to capture nonlinear effects and uncertainty, but Gaussian processes are not theoretically guaranteed to be monotonic. In practice, we found that the mean of the Gaussian process fit is monotonic in most cases, because the training data are monotonically increasing quantiles. For cases when the mean of the Gaussian process is not monotonic (as is the case for the single cell ChIP-seq data), we use linear interpolation. The monotonicity of the quantiles guarantees that the linear interpolation will be monotonic.

## 4.4 Data Description and Processing

Several high-throughput single cell versions of epigenetic assays have been developed, including single cell bisulfite sequencing (DNA methylation) (Angermueller et al., 2016), ATAC-seq (chromatin accessibility) (Buenrostro et al., 2015), and ChIP-seq (histone modification) (Rotem et al., 2015). Each of the initial studies that pioneered these methods applied them to mouse embryonic stem cells (mESCs) grown in serum, a classic model system of stem cell biology. Cells in this system are heterogeneous, differing depending on where they are located along a spectrum ranging from a pluripotent ground state to a differentiation primed state (Kolodziejczyk et al., 2015b). Note that mESCs grown in serum have different properties than mESCs cultured in 2i medium, which are much more homogeneous and differ primarily in their cell cycle stage (Buettner et al., 2015; Kolodziejczyk et al., 2015b).

We also analyzed single cell gene expression and DNA methylation data generated by sc-GEM (Cheow et al., 2016), a protocol that measures DNA methylation and gene expression in the same cells, from human cells undergoing reprogramming to induced pluripotent stem cells (iPSCs). We collected the publicly available data from these papers. In total, we have four kinds of single cell data from a total of 5,151 cells: 250 cells with gene expression data only (Kolodziejczyk et al., 2015b), 238 with DNA methylation and gene expression (Angermueller et al., 2016; Cheow et al., 2016), 76 with chromatin accessibility (Buenrostro et al., 2015), and 4,587 with H3K4me2 ChIP (Rotem et al., 2015).

The processing of single cell epigenetic data is more difficult than RNA-seq, because the epigenetic data are nearly binary at each genomic position (apart from allele-specific effects and copy number variations) and extremely sparse, with only a few thousand reads per cell in many cases. This makes it very difficult to extract any meaningful information at base pair resolution from a single cell. Instead, we followed the data

processing steps laid out in each of the respective papers that developed these techniques and aggregated the reads across related genomic intervals. For example, we followed the authors' lead in summing the chromatin accessibility data values from ATAC-seq in a given cell across all of the binding sites for a given transcription factor. Doing this for each of 186 transcription factors results in a matrix of 186 chromatin accessibility signatures across the set of cells. The DNA methylation data and H3K4me2 ChIP-seq data were aggregated in a similar way. We obtained the processed DNA methylation and ChIP-seq data from the initial publications. The processed ATAC-seq data are not publicly available, so we processed the data by implementing ourselves the pipeline described in the paper. We found that the DNA methylation data was the least sparse of any of the single cell epigenetic data types; the ChIP-seq data was the sparsest. Consequently, it was sufficient to aggregate the DNA methylation data over relatively small genomic intervals such as individual promoters or CpG islands.

## 4.5 Single Cell Transcriptome and Epigenome Data Show Common Modes of Variation

It seems likely that gene expression, DNA methylation, chromatin accessibility, and histone modifications will all change during the transition from pluripotency to a differentiation primed state. However, to the best of our knowledge, no one has yet demonstrated that it is possible to build a cellular trajectory from single cell epigenetic data.

To test our hypothesis that each of these epigenetic data types are changing over the course of a common underlying process, we first attempted to construct a cell trajectory for each type of data. Using SLICER, a method we previously developed (Welch et al., 2016a), we visualized each type of data as a two-dimensional projection and inferred a one-dimensional ordering for the cells. The 2D projections show that each type of data resembles a one-dimensional trajectory rather than a 2D blob of points (Fig. 4.2a-d). Note that these 2D projections do not force the data into a one-dimensional shape; the plots could look like a diffuse point cloud, and the fact that they instead resemble trajectories shows that the differences among cells are predominantly one dimensional. Furthermore, the projections of each kind of data are strikingly similar visually (Fig. 4.2a-d).

We further investigated these trajectories to determine whether they correspond to the same underlying process. The trajectory built from RNA data shows decreasing expression of pluripotency genes such as *Sox2*, consistent with previously published analyses (Kolodziejczyk et al., 2015b) (Fig. 4.2e). DNA methylation

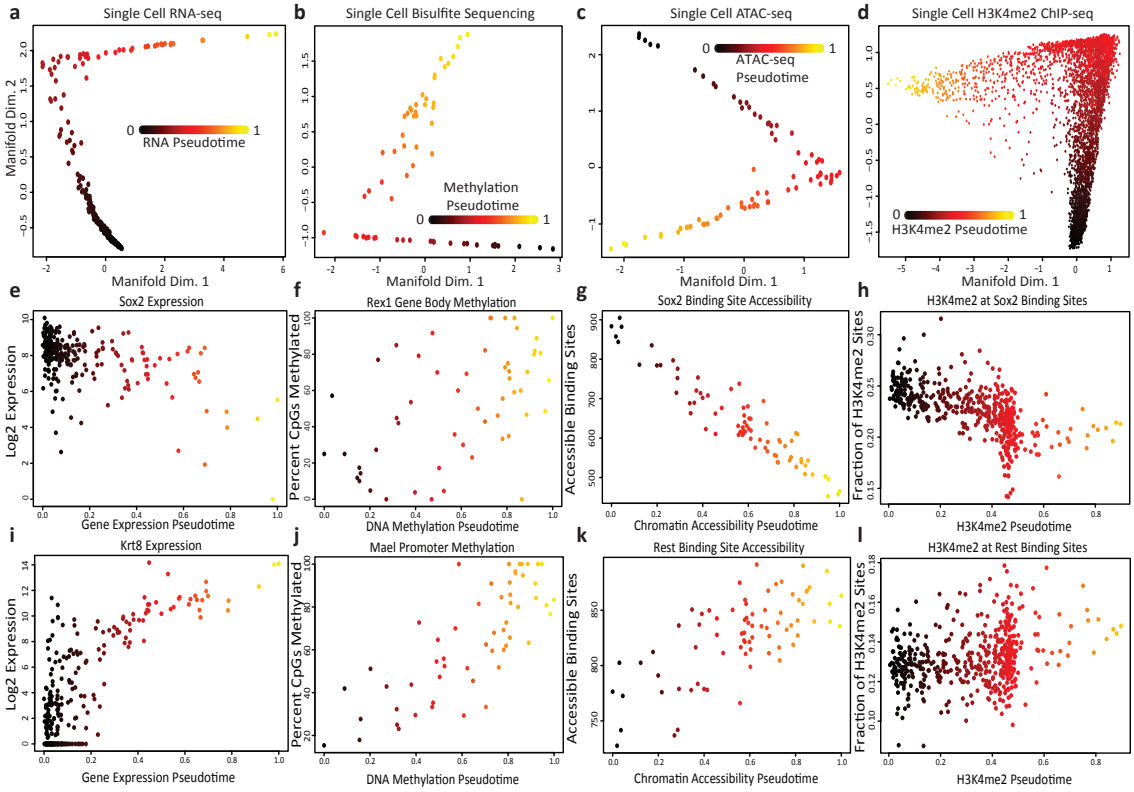



Figure 4.2: Single cell transcriptome and epigenome data show common modes of variation. (a)-(d): Single cell trajectories constructed by SLICER from RNA-seq, bisulfite sequencing, ATAC-seq, and H3K4me2 ChIP-seq of mouse embryonic stem cells grown in serum. (e)-(l) Levels of important gene expression, DNA methylation, chromatin accessibility, and H3K4me2 markers across the trajectories. We used SLICER for the analysis in this figure because it is a previously published method for constructing cell trajectories that allowed us to investigate the hypothesis that single cell transcriptome and epigenome measurements share common sources of variation.

of the gene body of *Rex1*, a gene that is shut off during the transition from pluripotency to differentiation priming(Singer et al., 2014), increases during the process (Fig. 4.2f). The single cell ATAC-seq data show that the chromatin accessibility of binding sites for the *Sox2* transcription factor decreases over pseudotime (Fig. 4.2g). Similarly, the levels of H3K4me2, a histone modification associated with active enhancers and promoters, decrease at *Sox2* binding sites (Fig. 4.2h). The RNA-seq data show increasing expression of previously identified differentiation markers(Kolodziejczyk et al., 2015b) such as *Krt8* (Fig. 4.2i). DNA methylation of the promoter for *Mael* increases, consistent with previous findings (Singer et al., 2014) (Fig. 4.2j). Both the chromatin accessibility (Fig. 4.2k) and H3K4me2 levels (Fig. 4.2l) at *Rest* binding sites increase, consistent with the known role of *Rest* in repressing key lineage-specifying genes (Jørgensen et al., 2009; Dietrich et al., 2012). In summary, our analysis indicates that each type of single cell data varies along a trajectory, establishing a continuum that ranges from pluripotency to a differentiation primed state.

We used SLICER to perform this initial exploratory analysis, but for the rest of this study, we use MATCHER, which is completely separate from SLICER and does not rely on the method in any way. We did confirm, however, that the master time values inferred by MATCHER are highly correlated with the pseudotime values inferred by SLICER (Fig. 4.3).

## 4.6 Validation Using Simulated and Real Data

To evaluate the accuracy of MATCHER, we generated synthetic data for which ground truth master time is known. We generated data by sampling 100 master time values uniformly at random from $[0, 1]$, then mapping these to pseudotime values through a warping function. Using the resulting pseudotime values, we generated 600 "genes" each following a slightly different "expression pattern" (function of pseudotime). Normally distributed noise was added to each gene expression value. We then used MATCHER to infer master time from these simulated gene expression values, and measured accuracy as the correlation between true and inferred master time values. Note that we use Pearson rather than Spearman correlation because we expect true and inferred master time to be linearly related (equal, in fact), and a nonlinear relationship would indicate that the inference process is inaccurate. The results of our simulations indicate that MATCHER accurately infers master time across a range of different warping functions and noise levels (Figs. 4.2-4.6). The method is very robust to noise in the simulated genes, yielding a correlation of 0.92 at a noise level of $\sigma = 9$, which is greater than $50\%$ of the range of the simulated features.

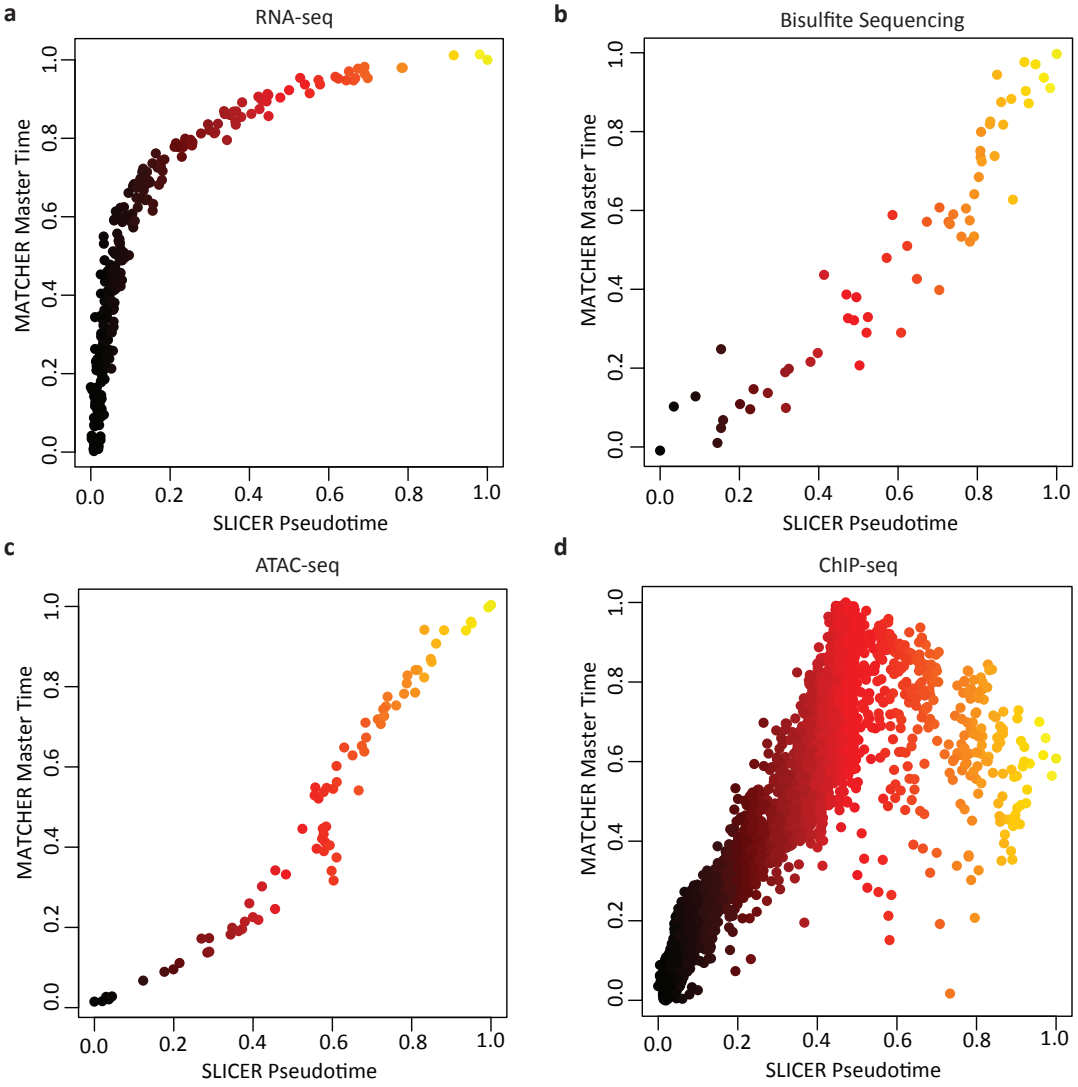



Figure 4.3: MATCHER master time is strongly correlated with SLICER pseudotime. Scatterplot of SLICER pseudotime versus MATCHER master time for (a) RNA-seq, (b) bisulfite sequencing, (c) ATAC-seq, and (d) H3K4me2 ChIP-seq. The points are colored by SLICER pseudotime.

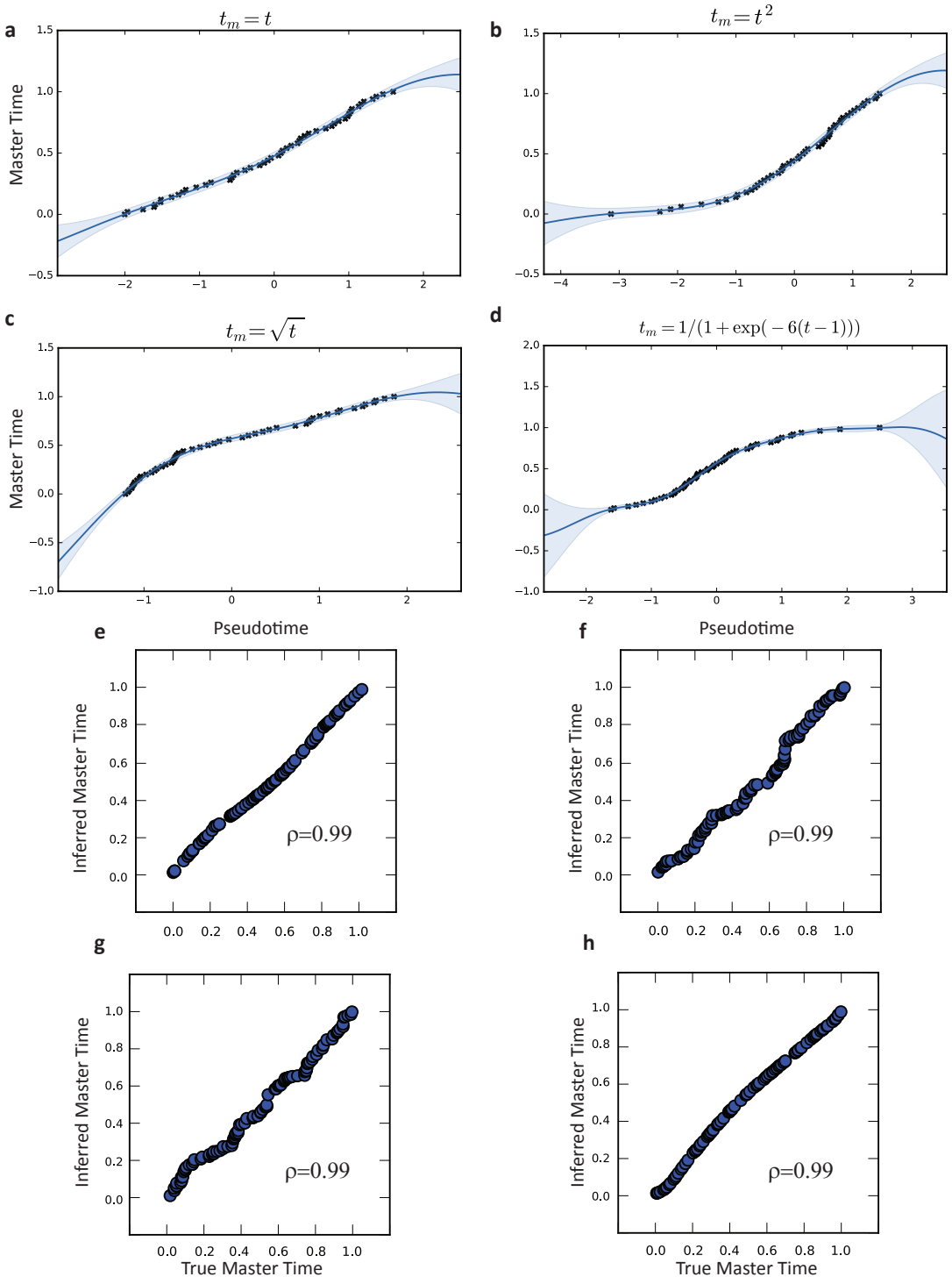



Figure 4.4: Results from synthetic data generated from different underlying warping functions. Inferred warping functions for (a) linear, (b) square root, (c) quadratic, and (d) logit true underlying warping functions. (e)-(h) Scatterplot of true vs. inferred master time for the corresponding warp functions of panels (a)-(d).

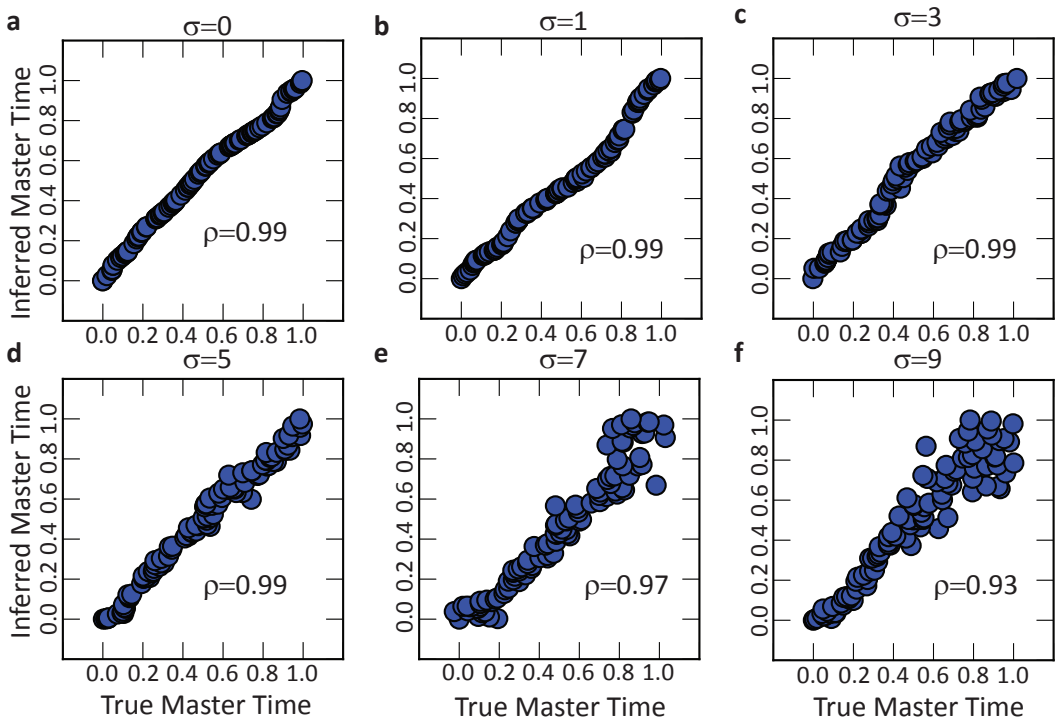



Figure 4.5: Synthetic Data Results for Increasing Noise Levels

We also tested MATCHER on real data. We used scM&T-seq data, in which DNA methylation and gene expression are measured in the same single cells (Angermueller et al., 2016), so that the true correspondence between single cell measurements is known. Note that we used the known cell correspondence information for validation only, not during the inference process. We first checked the relationship between master time inferred by MATCHER from RNA-seq and DNA methylation data by calculating the correlation between inferred master time values for corresponding DNA methylation and RNA-seq cells. This showed that the master time values, although not identical, are highly concordant (Pearson $\rho = 0.63$).

Predicting covariance of multiple genomic quantities across single cells is one of the key applications of MATCHER. Therefore, as an additional test, we investigated whether MATCHER can accurately infer correlations between DNA methylation events and gene expression. Here, we used Spearman correlation because we are interested in both linear and nonlinear relationships. We selected a set of genes and proximal methylated loci that showed statistically significant correlation in the original analysis of the scM&T data (Angermueller et al., 2016). Angermueller et al. grouped these pairs according to the type of region where the methylation site occurred. We selected the three types of regions with the largest number of significant

pairs (low methylation regions, H3K27me3 peaks, and P300 binding sites). Then, for each significant pair, we compared the true correlation (calculated using true cell correspondences) and correlation inferred by MATCHER (calculated using inferred cell correspondences). We also used MATCHER to compute correlations for the same gene-locus pairs using a single cell RNA-seq dataset published by a different lab (Kolodziejczyk et al., 2015b). In this dataset, the cells measured using RNA-seq are the same cell type, but not the same physical cells as those assayed for DNA methylation by Angermueller et al. In both cases, the inferred correlations closely match the true correlations (Fig. 4.6). The mean absolute deviation between true and observed correlations in the Angermueller dataset is 0.16. The correlations computed using the Kolodziejczyk data show slightly less concordance with the ground truth (mean absolute deviation = 0.27), likely due to the inevitable biological and technical variation that occur when different labs repeat an experiment. Even so, the vast majority of inferred correlations have the correct sign, and the relative magnitude of correlations tends to be preserved.

## 4.7 Correlations among single cell gene expression, chromatin accessibility, and histone modifications

We next used MATCHER to investigate the relationships among gene expression, chromatin accessibility, and histone modifications during the transition from pluripotency to a differentiation primed state. To our knowledge, this is the first time that investigation of the relationship among these three genomic quantities has been performed in single cells. We performed this analysis with two primary goals: (1) to confirm that the correlations among gene expression, chromatin accessibility, and H3K4me2 agree with what is known from bulk analysis (Fig. 4.7a, c); and (2) to demonstrate some of the unique insights that can be derived by correlating these quantities across individual cells (Fig. 4.7b, d-f). All of the correlation analyses described below are computed by taking the vector of values for a gene or set of genomic regions (such as binding sites for Sox2) across a set of single cells and correlating this vector with the values for another gene or set of genomic regions (such as binding sites for Oct4) across the set of single cells. Because the gene expression, chromatin accessibility, and H3K4me2 measurements that we are analyzing were performed on different single cells, this analysis is possible only because of MATCHERs ability to infer corresponding measurements. In summary, although some of the results that we describe recapitulate previous results from

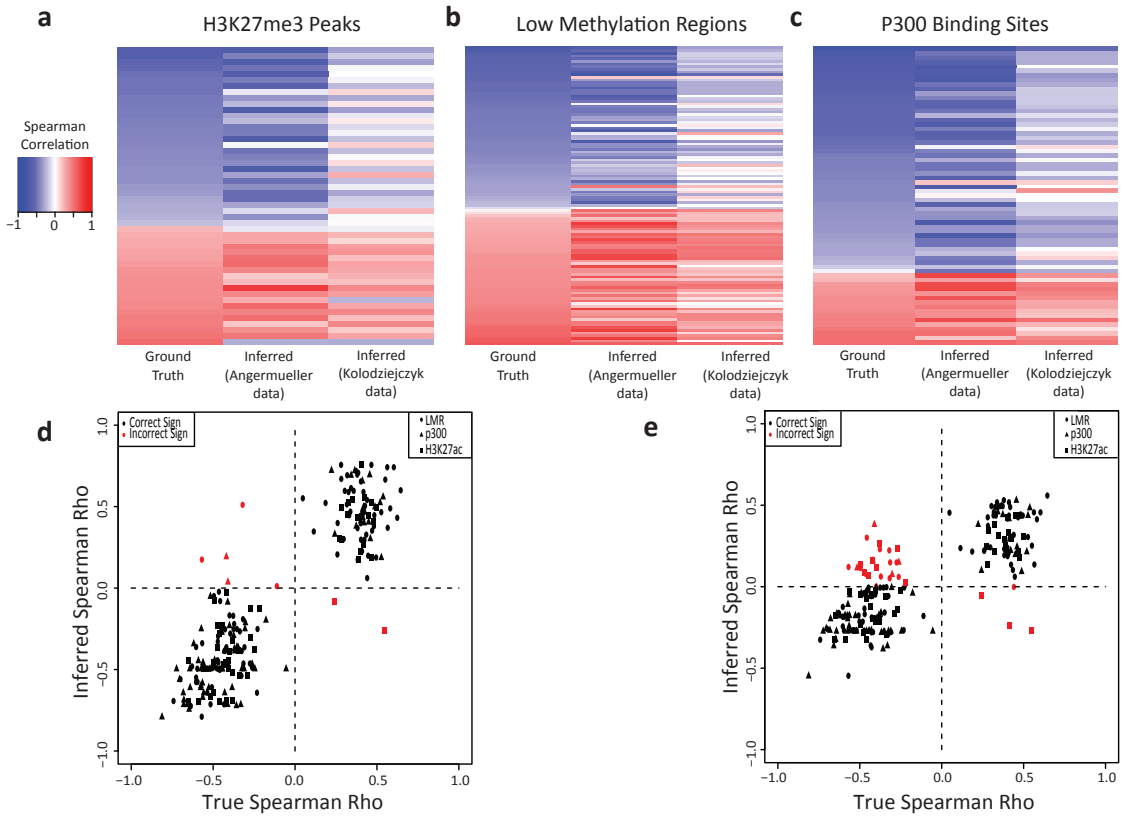



Figure 4.6: MATCHER accurately infers known correlations between DNA methylation and gene expression. (a)-(c) Heatmaps comparing true correlations between gene expression and DNA methylation of related regions (H3K27me3 peaks, LMRs, and P300 binding sites). The first column of each heatmap shows the true correlation based on known correspondence information, the second column shows the correlation inferred by MATCHER in the same dataset, and the third column is correlation inferred by MATCHER using a completely different single cell RNA-seq dataset from mESCs grown in serum. (d)-(e) Scatterplot representation of the results shown in (a)-(c). Panel (d) contains correlations computed using the Angermueller data; panel (e) is correlations computed from the Kolodziejczyk data. Each point represents the true and inferred correlation for a single gene-site pair; ideal results would lie along the y=x line. Note that the sign of the inferred correlation is correct for the vast majority of pairs.

analysis of bulk data, all of our analyses here are novel in that correlations are computed across individual cells within a heterogeneous population.

As an initial sanity check, we tested whether H3K4me2 and chromatin accessibility values within corresponding sets of genomic regions are positively correlated across the set of single cells (Fig. 4.7a). Because H3K4me2 is a histone modification associated with promoter and enhancer activation, we expect levels of the modification to correlate positively with chromatin accessibility. We confirmed this is, indeed, the case by inferring correlations between chromatin accessibility and H3K4me2 at the respective regions bound by 186 transcription factors and DNA binding proteins (Fig. 4.7a). For example, we correlated the chromatin accessibility at SOX2 binding sites across cells with the H3K4me2 levels at SOX2 binding sites across cells. The vast majority of these correlations are positive, consistent with previous findings from bulk data and with the role of H3K4me2 as an activating chromatin mark.

While investigating the correlation between H3K4me2 and chromatin accessibility, we found that the genomic binding regions clustered into two main groups: (1) pluripotency transcription factors and the NuRD complex and (2) chromatin remodeling factors that repress or activate lineage specific genes (Fig. 4.7b). Rotem et al. noted a similar relationship in the H3K4me2 data (Rotem et al., 2015). The accessibility of binding sites for *Oct4* (also known as *Pou5f1*), *Nanog*, and *Sox2*, well-established pluripotency transcription factors, is strongly anticorrelated with the accessibility of binding sites for *Ezh2*, *Ring1b*, and *Suz12*, which are Polycomb Group proteins (PcG) (Margueron and Reinberg, 2011). The targets of the transcription factor *Yy1*, which recruits PcG proteins (Basu et al., 2014), show a similar trend to the PcG proteins. Given that PcG proteins play a key role in repressing neuronal lineage genes in pluripotent cells (Surface et al., 2010), this anticorrelation suggests that chromatin is being remodeled to prime lineage-specific genes while shutting down regions associated with pluripotency. *Rest* and *CoRest* show a similar pattern to the PcG proteins; these proteins are known to co-associate with the polycomb repressive complex (PRC2) and also to repress key lineage specific genes in pluripotent cells (Jørgensen et al., 2009; Dietrich et al., 2012). Interestingly, the targets of *Usf1*, which is known to recruit Trithorax Group (TrxG) proteins (Deng et al., 2013), also show a pattern of increasing chromatin accessibility. The TrxG proteins are chromatin activators that regulate lineage differentiation genes (Surface et al., 2010; Deng et al., 2013; Bernstein et al., 2006), suggesting that the activation of certain differentiation genes is occurring while their repression by PRC2 is being lifted. Finally, targets of *Lsd1*, *Mi2*, *Hdac1*, and *Hdac2*, components of the NuRD complex, show positive correlation with targets of pluripotency factors. The NuRD complex contains chromatin remodeling proteins that remove

histone methylation and histone acetylation marks and function to "decommission" pluripotency enhancers during early differentiation (Whyte et al., 2012). In summary, our analysis of correlation between chromatin accessibility and H3K4me2 marks indicate that the overall trend in both types of data is toward chromatin changes that shut off pluripotency and begin to lift lineage repression in preparation for differentiation.

As an additional sanity check, we investigated whether chromatin accessibility and H3K4me2 are positively correlated with the expression of genes within the corresponding regions. For this analysis, we chose to focus specifically on the binding regions for EZH2, RING1B, TCF3, OCT4, SOX2, and NANOG. Because of the way we aggregated genomic regions when analyzing chromatin accessibility and ChIP-seq data, we needed a comparable way to aggregate the expression of genes within these regions. After locating genes whose promoters overlapped each of these binding regions, we filtered the sets of genes to remove genes that occurred in multiple binding regions. We then normalized the expression of each gene (zero mean, unit variance) and calculated the aggregate expression for each set of genes. These aggregate expression levels of genes whose promoters occur within the binding regions of each of the six proteins are then directly comparable with the chromatin accessibility and H3K4me2 from the same set of binding regions within each cell. Note again that we are correlating these quantities across single cellseach cell has six aggregate expression values and corresponding chromatin accessibility and H3K4me2 values. As expected, the aggregate expression of these sets of genes correlates well with the chromatin accessibility and H3K4me2 of the gene promoters (Fig. 4.7c-d), with the exception of Oct4. The expression of Oct4 targets are only weakly correlated with the aggregate chromatin accessibility and H3K4me2. Figures 4.8 and 4.9 show the corresponding values inferred by MATCHER for gene expression, chromatin accessibility, and H3K4me2 values in the same single cells.

To demonstrate that MATCHER can reveal unique insights not possible with bulk data, we investigated how the gene expression levels of key pluripotency factors and chromatin remodeling proteins correlate with the chromatin accessibility of their binding sites during the transition from nave to primed pluripotency (Fig. 4.7e-g). In this analysis, we made use of the fact that MATCHER tells us both (1) the relationship between chromatin accessibility and gene expression in individual cells and (2) the trends of both of these quantities over master time. This allowed us to begin to tease apart how different regulatory mechanismsboth chromatin and expressionoperate during a sequential biological process. Using the same transcription factors and DNA binding proteins as in Fig. 4.7a, we calculated the correlation between the expression level of each gene and the overall chromatin accessibility of the sites where its protein product binds to the genome. For

example, we correlated the vector of expression levels for Sox2 across the set of single cells with the vector of chromatin accessibility for the targets of SOX2 across the set of cells. Note that we are looking at the accessibility of the targets of these DNA binding proteins, not the promoters of the genes that encode these factors (although, in some cases, a protein may target the promoter of the gene that encodes it).

The pluripotency transcription factors *Esrrb*, *Nanog*, *Pou5f1*, and *Sox2* each show positive correlation between expression and chromatin accessibility, with both expression and chromatin accessibility showing an overall decreasing trend over master time (Fig. 4.7). This indicates that the expression of these genes is being shut off at the RNA level at the same time as the binding of the factors is shut off at the chromatin level. Interestingly, *Tcf7l2* expression shows strong negative correlation with the chromatin accessibility of its targets. We speculate that this negative correlation may be due to the fact that *Tcf7l2* functions primarily as a transcriptional repressor(Sokol, 2011), and thus increased expression will lead to more repression of its targets.

In contrast to the pluripotency factors, the expression of genes involved in chromatin remodeling show weak negative correlation with the accessibility of their binding sites (Fig. 4.7e). The chromatin accessibility of these factors targets shows an increasing trend over master time, but the expression of the chromatin remodeling factors does not vary significantly over master time. The inferred corresponding values for Yy1 are shown as an example in Fig. 4.7. Thus, changes in the chromatin accessibility of the targets of these chromatin remodeling complexes occurs without accompanying changes in the gene expression levels of the remodelers, indicating that regulation is occurring primarily at the chromatin level in this case. The one exception is the Rest gene, whose expression decreases over master time and shows strong negative correlation with the accessibility of its binding sites. The fact that these correlations are nearly zero indicates that changes in the chromatin accessibility of the targets of these chromatin remodeling complexes occurs primarily without accompanying changes in the gene expression levels of the remodelers. The one exception is *Rest*, whose expression shows strong negative correlation with the accessibility of its binding sites.

To understand the advantages of using MATCHER in this way to analyze a combination of omics data from single cells, it is instructive to imagine a comparable bulk experiment and what insights it might yield. One could perform bulk RNA-seq, ATAC-seq, and ChIP-seq on separate populations of embryonic stem cells. However, the cellular differences that we have observed here occur among stem cells grown in a common culture environment. A comparable bulk analysis would require some sort of purification (FACS, MACS, etc.) to isolate populations of nave and primed cells grown in serum. Even if such populations were purified,

they would likely still contain a mixture of cells at various points on the spectrum from ground state to primed pluripotency. Furthermore, such an experiment would allow only early and late comparisons, rather than examination of the continuous trends that MATCHER provides. Consequently, one could identify genes with higher population expression in ground state vs. primed cells and regions of chromatin that are generally more accessible in ground state vs. primed cells, but not any of the intermediate changes in expression or chromatin that occur during the transition from ground state to primed pluripotency. The point of this discussion is not to disparage bulk sequencing experiments, which are extremely useful, but rather to argue that there is also a place for the sort of integrative single cell multi-omic analysis that we performed here. We believe that, just as trajectory analysis of single cell RNA-seq data has proven useful for studying many important biological processes, MATCHER will reveal novel biology when applied to future single cell transcriptomic and epigenomic data.

## 4.8 Relationship between DNA methylation and gene expression during transition from ground state to primed pluripotency

We next used MATCHER to investigate the interplay between gene expression and DNA methylation in mouse ES cells. We first examined the relationship between master time inferred from gene expression and master time inferred for the same cells using DNA methylation (Fig. 4.10a). (Note that here we are using the known correspondences available from scM&T-seq to compare master time values inferred in two different ways for identical cells.) This analysis showed an intriguing relationship: DNA methylation and gene expression master time track together quite well until a specific point in RNA master time, around master time = 0.3. After that point, the degree of coupling suddenly decreases. This result is consistent with the results of the initial analysis of the scM&T-seq data, which found variability in the strength of coupling between gene expression and DNA methylation across the set of cells (Angermueller et al., 2016).

To assess the significance of the apparent partial decoupling between DNA methylation and gene expression, we computed separate Pearson correlation values for cells with gene expression master time less than 0.3 and greater than 0.3. Then we performed Fishers $r$-to-$z$ transformation on the correlations and computed a $p$-value for the null hypothesis that the correlation before master time = 0.3 is less than or equal to the correlation after master time = 0.3 (one-tailed test). The p-value was 0.037, indicating a significant difference at p=0.05. We also performed a permutation test, in which we sampled (without repetition) a

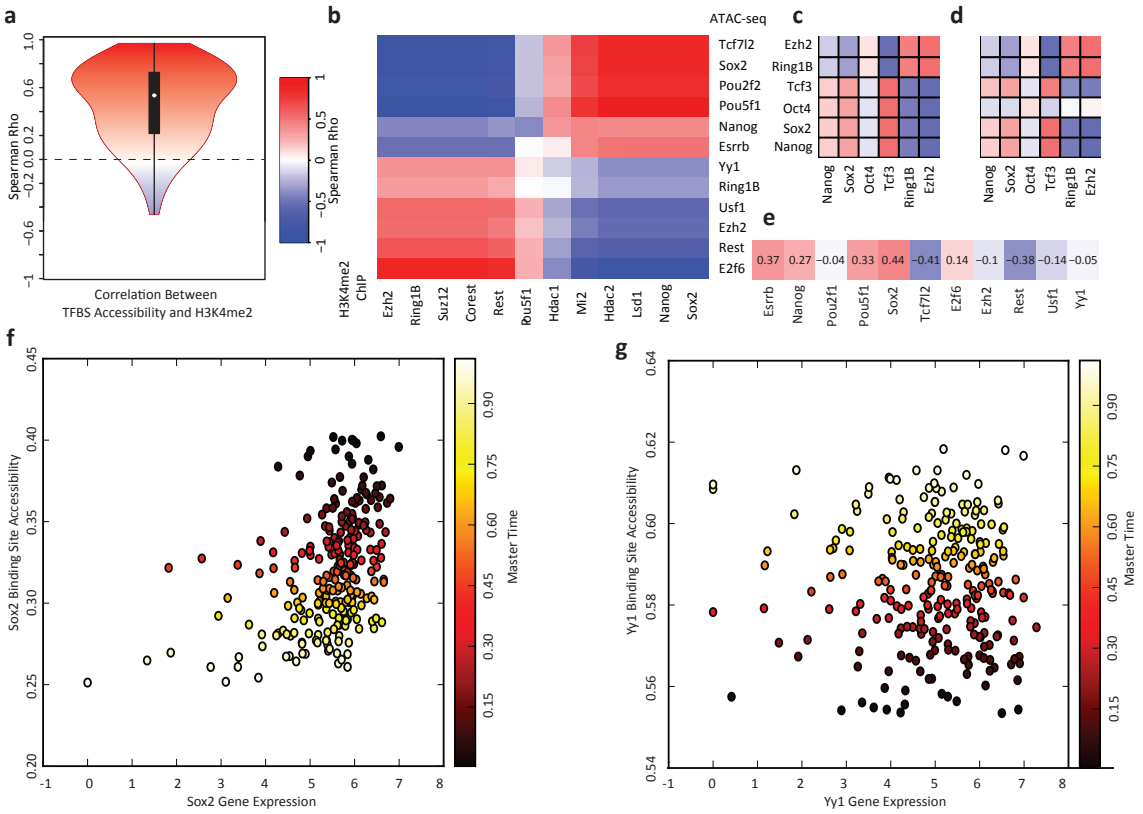



Figure 4.7: Correlations among single cell gene expression, chromatin accessibility, and histone modifications. (a) Violin plot of correlations among chromatin accessibility and H3K4me2 of transcription factor binding sites for 186 transcription factors. Note that most correlations are strongly positive. (b) Correlation between chromatin accessibility and H3K4me2 data reveals that targets of pluripotency factors/NuRD complex and targets of Polycomb Group/Trithorax Group proteins are anticorrelated in single cells. (c) Correlation between gene expression signatures and chromatin accessibility signatures. (d) Correlation between gene expression signatures and H3K4me2 signatures. (e) Correlation between gene expression of DNA binding proteins and chromatin accessibility of their targets. (f) Inferred corresponding values of Sox2 gene expression and chromatin accessibility of SOX2 binding sites. Each point represents inferred correspondence from a single cell. The x-axis shows the value of the gene expression signature in that cell, and the y-axis shows the value of the chromatin accessibility signature. The points are colored by inferred master time. (g) Inferred corresponding values of Yy1 gene expression and chromatin accessibility of YY1 binding sites.

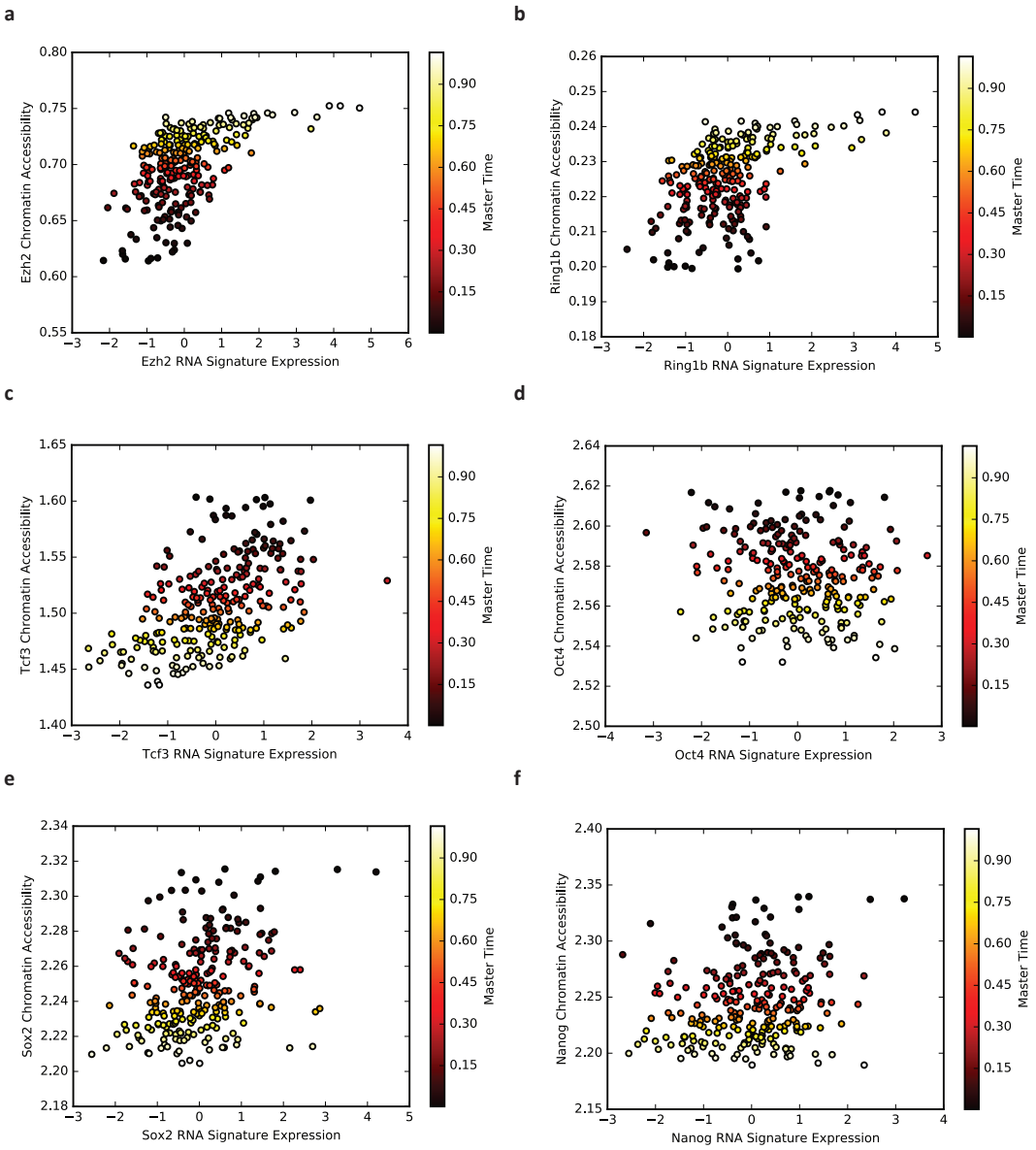



Figure 4.8: Corresponding values inferred by MATCHER for gene expression and chromatin accessibility signatures. Each point represents inferred correspondence from a single cell. The x-axis shows the value of the gene expression signature in that cell, and the y-axis shows the value of the chromatin accessibility signature. The points are colored by inferred master time. Note that these are the data used to generate the values on the diagonal of the heatmap in Fig. 4.7c.

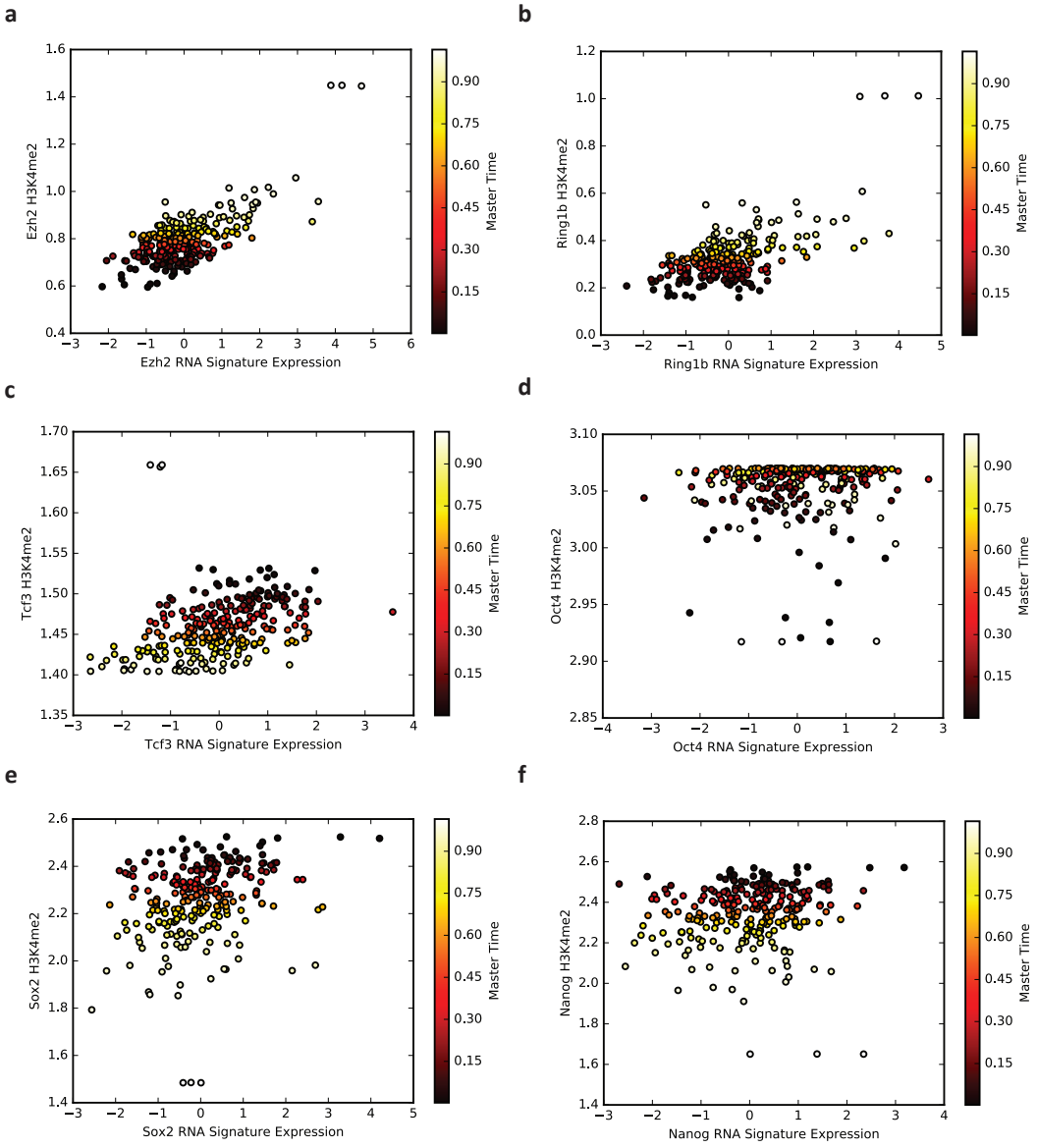



Figure 4.9: Corresponding values inferred by MATCHER for gene expression and H3K4me2 signatures. Each point represents inferred correspondence from a single cell. The x-axis shows the value of the gene expression signature in that cell, and the y-axis shows the value of the H3K4me2 signature. The points are colored by inferred master time. Note that these are the data used to generate the values on the diagonal of the heatmap in Fig. 4.7d.

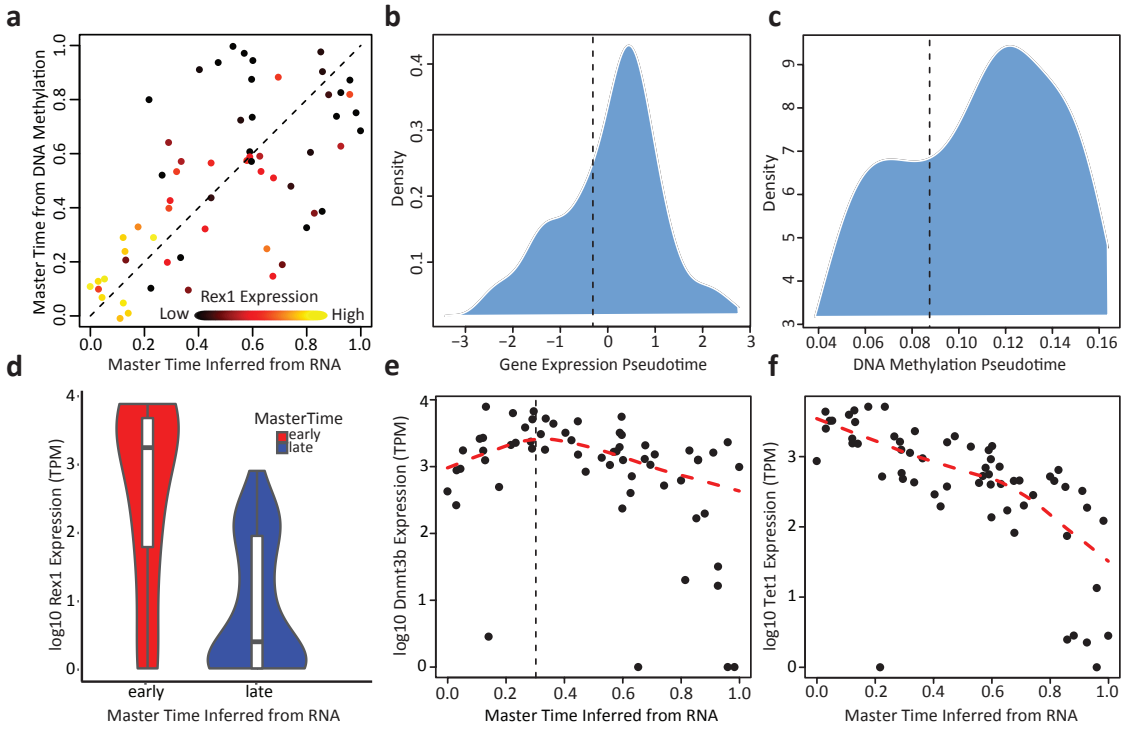



Figure 4.10: Relationship between DNA methylation and gene expression during transition from ground state to primed pluripotency. (a) Scatter plot showing the relationship between master time inferred from gene expression and master time inferred from DNA methylation. Points are colored by the log10 expression of Rex1. The dotted line is the y=x line. Note that the gene expression and DNA methylation master time values are more correlated before master time = 0.3 than after. (b)-(c) Density plots showing the distribution of pseudotime inferred from (b) gene expression and (c) DNA methylation. The vertical dotted line indicates the 30th percentile of pseudotime (master time = 0.3). (d) Violin plot showing the distribution of Rex1 expression in cells before master time = 0.3 ("early") and after master time = 0.3 ("late"). (e) Expression of Dnmt3b as a function of gene expression master time. The red line is a loess smoothing function indicating the overall expression trend. The black vertical line indicates master time = 0.3. (f) Expression of Tet1 as a function of gene expression master time. The red line is a loess smoothing function indicating the overall expression trend.

random division of the cells into two groups consisting of approximately 30% and 70% of cells, calculated the Spearman correlation between the gene expression and DNA methylation master time values in the two groups separately, and subtracted the two correlation values. Repeating this sampling procedure 100,000 times gave an empirical $p$-value of 0.0025 for the null hypothesis that the correlation before master time = 0.3 is less than or equal to the correlation after master time = 0.3. We also confirmed that both analyses are robust to the choice of division point in master time: the difference in correlations is also significant ($p < 0.05$) if master time = 0.5 is used as the dividing line.

We hypothesized that the observed relationship may occur because specific de novo DNA methylation changes are required to trigger a key step in the process of gene expression changes during the transition from ground state pluripotency to a primed state, but after this point in the process, the sequential gene expression changes proceed somewhat independently from the DNA methylation changes. A previous single cell study of mouse embryonic stem cells grown in serum showed the existence of two metastable expression states, corresponding to ground state and primed pluripotency (Singer et al., 2014). The Rex1 gene was previously shown to be a marker for these metastable expression states, with high Rex1 expression in the ground state and low Rex1 expression in the primed state (Singer et al., 2014). Singer et al. also found that the transition between these two states is dependent on the activity of DNA methyltransferase (DNMT) enzymes, and knocking out DNMT activity greatly increases the proportion of cells in the Rex1-high state (Singer et al., 2014).

In support of this hypothesis, the cells in which DNA methylation and gene expression correlate strongly show high levels of Rex1 expression, while the remaining cells show much lower expression (Fig. 4.10a and d). We also found that the distributions of pseudotime values for both gene expression and DNA methylation are highly non-uniform and roughly bimodal (Fig. 4.10b-c). This pattern is consistent with the existence of two metastable states, suggesting that cells tend to accumulate toward the beginning and end of pseudotime and transition fairly rapidly in between. In further support of this model, the two modes of the distribution account for approximately 30% and 70% of cells, respectively (Fig. 4.10b-c); these proportions correspond to the divergence point (master time = 0.3) noted in Fig. 4.10a.

To further investigate the potential role of de novo methylation in the transition from the ground state to the primed state, we examined the expression trends of Dnmt3b, a gene encoding a DNA methyltransferase, and Tet1, a gene implicated in demethylation (Fig. 4.10e-f). Singer et al. previously found the expression of these two genes to be strongly negatively and positively correlated with Rex1 expression, respectively (Singer

et al., 2014). Intriguingly, we find that Dnmt3b shows a transient pulse of expression, with initially increasing expression that peaks, then steadily decreases (Fig. 4.10e). The peak of Dnmt3b expression occurs precisely at master time = 0.3, which fits well with the data in Fig. 4.10a-d and is also consistent with a model in which de novo methylation activity increases to help cells escape the Rex1-high state. Tet1 expression is highest at the beginning of master time and steadily decreases (Fig. 4.10f). These two observations together suggest that Tet1 actively maintains low methylation levels in the Rex1-high state but is gradually downregulated while a pulse of Dnmt3b expression occurs, leading to the accumulation of methylation and transition to the Rex1-low state. These results also suggest that de novo methylation is required primarily to transition away from the Rex1-high state, and both de novo methylation activity and demethylation gradually subside after this transition, stabilizing the DNA methylation profiles of the cells.

It is worth noting that the partial decoupling we have just described is not the same as complete decoupling. The master time values that MATCHER inferred separately from DNA methylation and gene expression are highly correlated ($\rho = 0.63$), and our results shows that the method accurately predicts the ground truth correlations between DNA methylation and gene expression in single cells (mean absolute deviation of 0.16). We have chosen to use the term partial decoupling to indicate that DNA methylation and gene expression are somewhat, but not completely, predictive of each other. MATCHER does not require that the measurements be completely coupled, and our analysis here shows that the method still performs well even in the presence of partial decoupling. It is perhaps not surprising that DNA methylation and gene expression do not perfectly predict each other, because gene expression is regulated by many factors in addition to DNA methylation. Our discovery of this partial decoupling does highlight the fact that simultaneous experimental measurements, such as scM&T-seq, provide additional information that MATCHER cannot infer. Nevertheless, MATCHER provides a useful tool for analyzing single cell transcriptomic and epigenomic data, whether or not experimentally determined cell correspondences are available.

## 4.9 Analysis of gene expression and DNA methylation changes during human iPS cell reprogramming

We used MATCHER to analyze data from sc-GEM, a protocol (distinct from scM&T-seq) that allows simultaneous measurement of pre-selected DNA methylation and gene expression markers in single cells using PCR (Cheow et al., 2016). Cheow et al. performed sc-GEM on human fibroblasts undergoing iPS cell

reprogramming. Unlike the mouse ES cell data that we analyzed above, the Cheow dataset contains multiple time points, from 0 to 24 days after the start of the reprogramming process. We downloaded the processed, normalized PCR data from the Cheow paper and did not perform additional processing.

When we used MATCHER to analyze the Cheow data, we found that, as with the mouse ES cells, the distribution of pseudotime inferred from both DNA methylation and gene expression was bimodal rather than uniform (Fig. 4.11a-b). This pattern suggests that only unprogrammed fibroblast cells and successfully reprogrammed iPS cells are stable; cells transitioning between states are relatively unstable, and thus transition relatively rapidly. Unlike in the case of ES cells, DNA methylation and gene expression master time values appear to be strongly correlated throughout the entire iPS reprogramming process (Fig. 4.11c).

Because sc-GEM provides measurements where the true correspondence between cells and correlation between DNA methylation and gene expression are known, this dataset provides an additional opportunity to assess the accuracy of MATCHER. To do this, we computed the true Spearman correlation between all pairs of genes and promoters assayed in the sc-GEM experiment. Then, we compared these true values to the values inferred by MATCHER. As with the scM&T-seq dataset described above, MATCHERs inferred correlations closely matched the true values (Fig. 4.11d-e), with a mean absolute deviation of 0.17.

The experimental design of the Cheow dataset, which contains multiple time points, allows us to utilize both temporal and pseudotemporal information. We therefore investigated whether we could use the time point information to learn anything about the relative ordering of DNA methylation and gene expression changes. Our analysis suggests that DNA methylation changes lag behind gene expression changes. As Fig. 4.11f shows, the day 0 (BJ) fibroblasts and day 8 fibroblasts span nearly identical portions of master time inferred from DNA methylation, and signs of reprogramming are apparent only at day 16 or beyond. In contrast, gene expression master time shows a continual, steady progression, with only a handful of cells overlapping the master time range of the previous time point (Fig. 4.11g). Thus, enough gene expression changes occur within 8 days of the reprogramming process to distinguish untreated cells and day 8 cells, but it takes more than 8 days for distinguishing DNA methylation changes to occur. In other words, the gene expression changes occur temporally prior to the DNA methylation changes. Consistent with this result, the relative height of the iPS cell mode in Fig. 4.11b is less than the relative height of the iPS cell mode in Fig. 4.11a, indicating that fewer cells have moved beyond the DNA methylation profile of the starting fibroblast state than have moved beyond the starting gene expression state. We note that sc-GEM experiment measured only a pre-selected subset of genes and promoters, so we cannot rule out the possibility

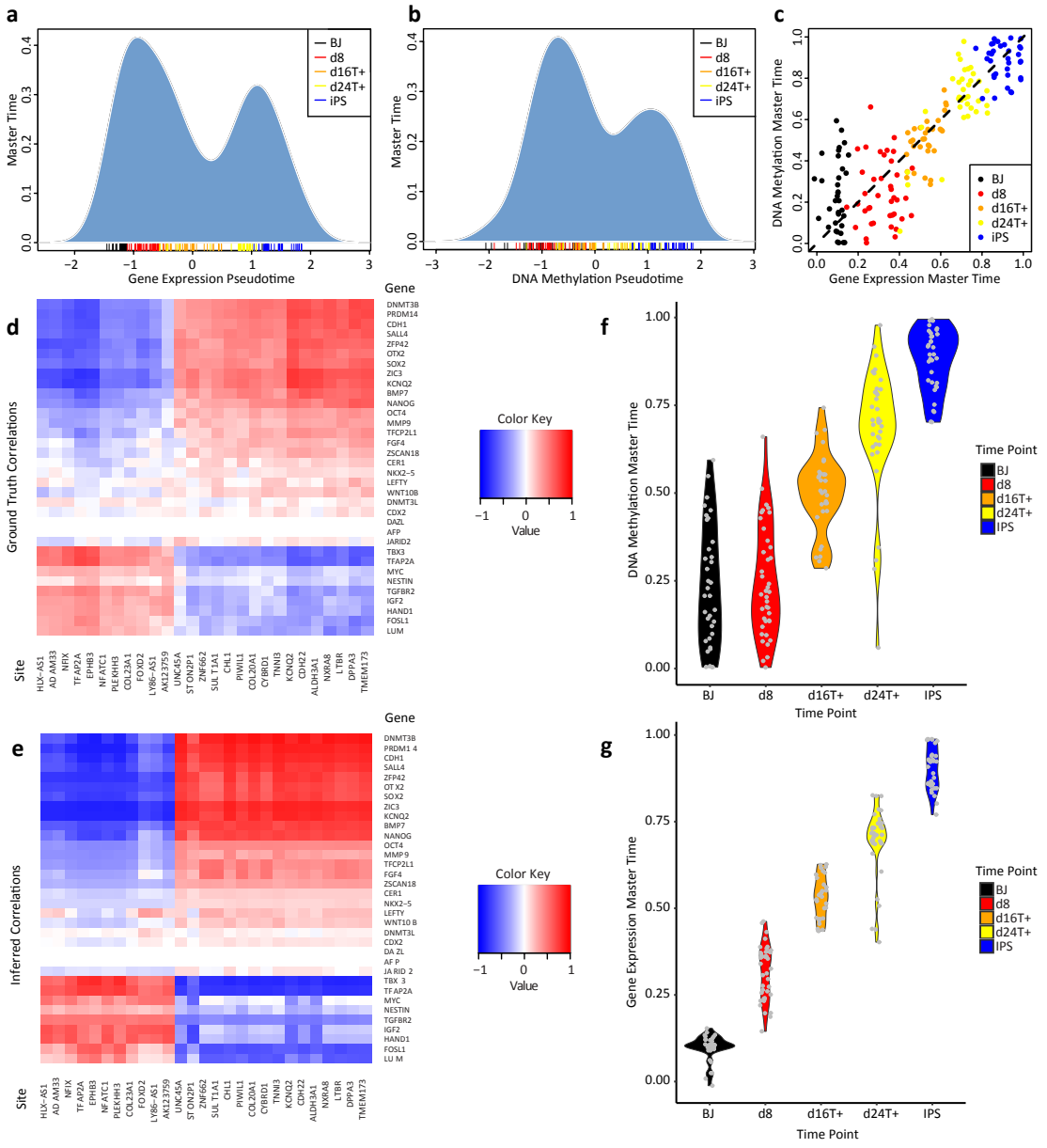



Figure 4.11: Analysis of gene expression and DNA methylation in human fibroblast cells undergoing reprogramming. (a)-(b) Density plots showing distribution of pseudotime inferred from (a) gene expression and (b) DNA methylation. The pseudotime values for individual cells are shown as a rug plot below the density plot; color indicates the time point. (c) Relationship between master time inferred from gene expression and master time inferred from DNA methylation. (d) Heatmap of ground truth correlation between expression of all genes measured in the sc-GEM experiment and DNA methylation level of all promoters measured. (e) Heatmap of correlation inferred by MATCHER from sc-GEM data. Note that MATCHER inferred these correlations without using the known correspondence among cells in any way. (f) Violin plot of the DNA methylation master time values for cells at each time point. Note that the distributions for untreated fibroblasts (BJ) and fibroblasts 8 days after treatment (d8) are virtually identical. (g) Violin plot of the gene expression master time values for cells at each time point.

that the DNA methylation status of other genomic loci could distinguish the untreated and day 8 fibroblasts. Nevertheless, our findings are consistent with a previous report that the vast majority of the DNA methylation changes in iPS reprogramming occur after day 9 (Polo et al., 2012).

One of the motivations for developing MATCHER was to enable integration of single cell datasets in which cells do not exactly correspond. Therefore, we performed additional analysis to demonstrate that, even though sc-GEM provides measurements from exactly corresponding cells, MATCHER does not require this information. To simulate datasets in which DNA methylation and gene expression were measured separately on distinct cells of the same type, we repeatedly sampled a random 75% or 50% of sc-GEM gene expression profiles and a random 75% or 50% of sc-GEM DNA methylation profiles. This analysis showed that we could reproduce the results in Fig. 4.11 using a dataset without exactly corresponding cells (4.12).

Finally, we note that the lagging behavior observed here does not violate the assumptions of MATCHER; in fact, this is an example of just the sort of time warping behavior that is shown in the imaginary example of Fig. 4.1a. Comparing the master time ranges for corresponding time points in Fig. 4.11f-g shows that the warping functions inferred by MATCHER are largely able to correct for this effect. For example, days 0-8 span the same master time range for both DNA methylation and gene expression. If MATCHER did not correct for time warping, day 8 DNA methylation measurements would be matched only with day 0 gene expression cells; day 16 DNA methylation cells would be matched only with day 8 gene expression measurements; and so on.

## 4.10 Incorporating known cell correspondence information to infer shared master time

So far, we have used MATCHER to infer separate master time values for each type of transcriptomic or epigenomic measurement. Our results demonstrate that such an approach can reveal important insights, whether the true cell correspondences are known or unknown. However, in cases where multiple measurements are performed simultaneously on the same cells, as with scM&T-seq and sc-GEM, it could also be informative to infer a shared cell ordering that indicates each cells overall progress in terms of both transcriptomic and epigenomic changes. We now demonstrate how to infer "shared master time" using MATCHER and give an example of how such analysis can be useful.

To infer shared master time, MATCHER uses a shared GPLVM (Ek, 2009) to infer pseudotime in place of a separate GPLVM for each data type. The shared GPLVM assumes that each type of measurement is

a

| | | | | | |
|---|---|---|---|---|---|
| mESCs - scM&T | 0.16 | | | | |
| mESCs - scRNA | 0.27 | | | | |
| iPSCs - Full | 0.17 | | | | |
| iPSCs - 75% | 0.18 | 0.18 | 0.18 | 0.18 | 0.16 |
| iPSCs - 50% | 0.19 | 0.19 | 0.17 | 0.19 | 0.18 |

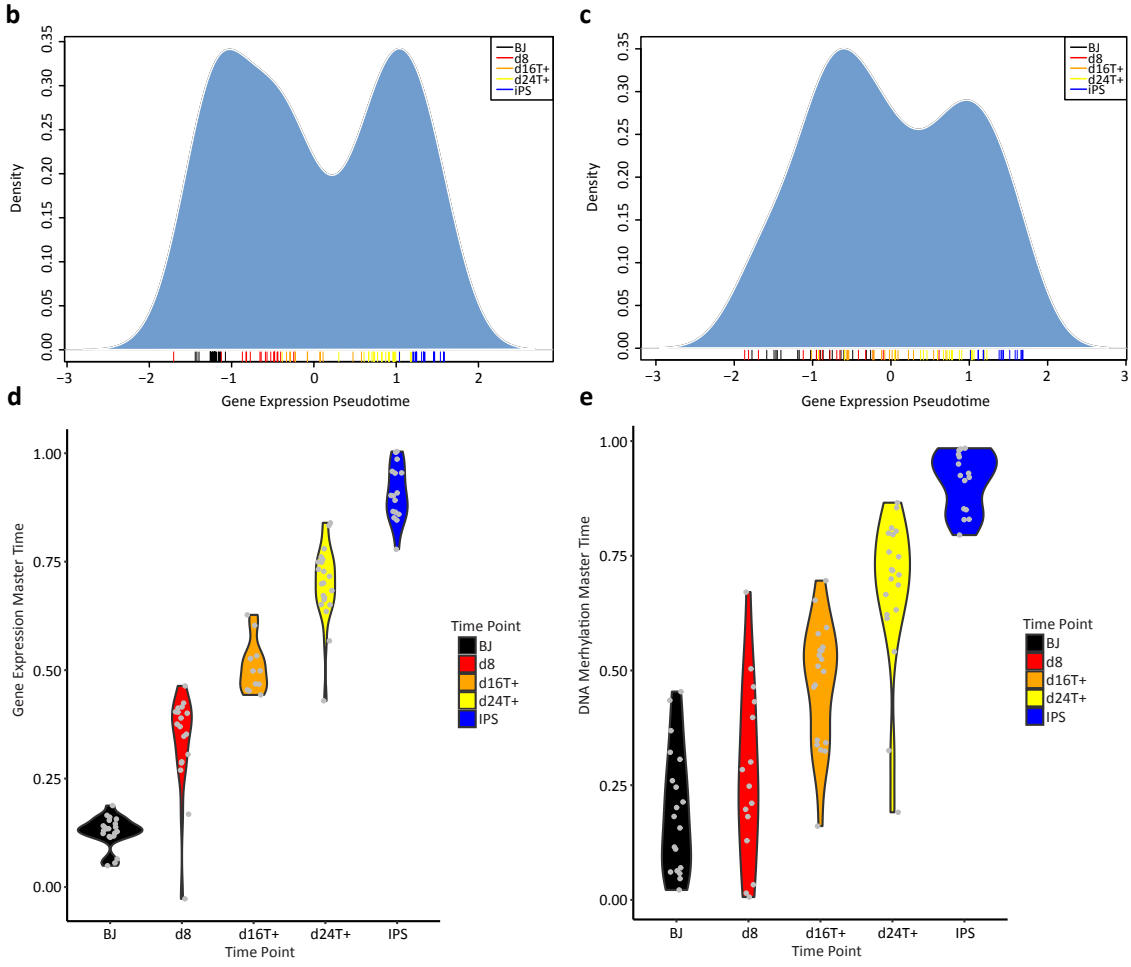



Figure 4.12: Subsampling analysis of sc-GEM data showing that MATCHER does not require corresponding cell measurements (a) Table of mean absolute deviation between ground truth and inferred correlations for scM&T-seq dataset (top row); scM&T-seq methylation data and Kolodziejczyk gene expression data (second row); the full sc-GEM dataset from Cheow; 5 random subsamples of 75% of cells from Cheow; and 5 random subsamples of 50% of cells from Cheow. (b)-(c) Density plots showing distribution of pseudotime inferred from (b) gene expression and (c) DNA methylation. The pseudotime values for individual cells are shown as a rug plot below the density plot; color indicates the time point. Compare panels (b)-(c) to Fig. 6 (a)-(b). (d) Violin plot of the DNA methylation master time values for cells at each time point. (e) Violin plot of the gene expression master time values for cells at each time point. Compare panels (d)-(e) to Fig. 6 (f)-(g).

generated, through different mappings, from a common (shared) latent space (Ek, 2009). After inferring pseudotime using a shared GPLVM, MATCHER uses Gaussian process regression to learn a warping function and infer master time values that are uniformly distributed between 0 and 1, in the same way as when pseudotime values are inferred separately for each data type.

We first used MATCHER to infer a shared master time value using both DNA methylation and gene expression data for each cell assayed with scM&T-seq (Fig. 4.13a-b). The resulting shared master time values reconcile the sequence of changes occurring in both genomic quantities. The Pearson correlation between DNA methylation master time and RNA master time is 0.63. In contrast, the correlation between DNA methylation master time and shared master time is 0.93 (Fig. 4.13a), and the correlation between RNA master time and shared master time is 0.84 (Fig. 4.13b).

We also inferred shared master time for cells assayed with sc-GEM (Fig. 4.13c-e). As an example of how this shared master time can be used, we identified "lagging cells" whose shared master time values overlap with the shared master time values of cells from an earlier time point (Fig. 4.13c). These cells lag behind other cells from the same time point in terms of both their gene expression and DNA methylation reprogramming progress. Using either gene expression (Fig. 4.13d) or DNA methylation (Fig. 4.13e) alone to identify lagging cells gives conflicting sets of cells; some cells whose gene expression lags show timely methylation changes and vice versa. Thus, it is not clear which of these cells should be considered lagging in the overall process of reprogramming both DNA methylation and gene expression. Shared master time provides a principled way to reconcile the two perspectives obtained from gene expression and DNA methylation measurements and determine the overall reprogramming progress of each cell.

## 4.11 Discussion

We used MATCHER to characterize the corresponding transcriptomic and epigenetic changes in embryonic stem cells undergoing the transition from pluripotency to a differentiation primed state. Interesting future directions of research include extending the model to align manifolds with dimensionality higher than one, as well as adapting the method for cell populations whose cells fall into clusters rather than along one continuous spectrum. In addition, our model does not explicitly account for branching trajectories, which can arise in biological processes with multiple outcomes (Trapnell et al., 2014; Welch et al., 2016a). A simple

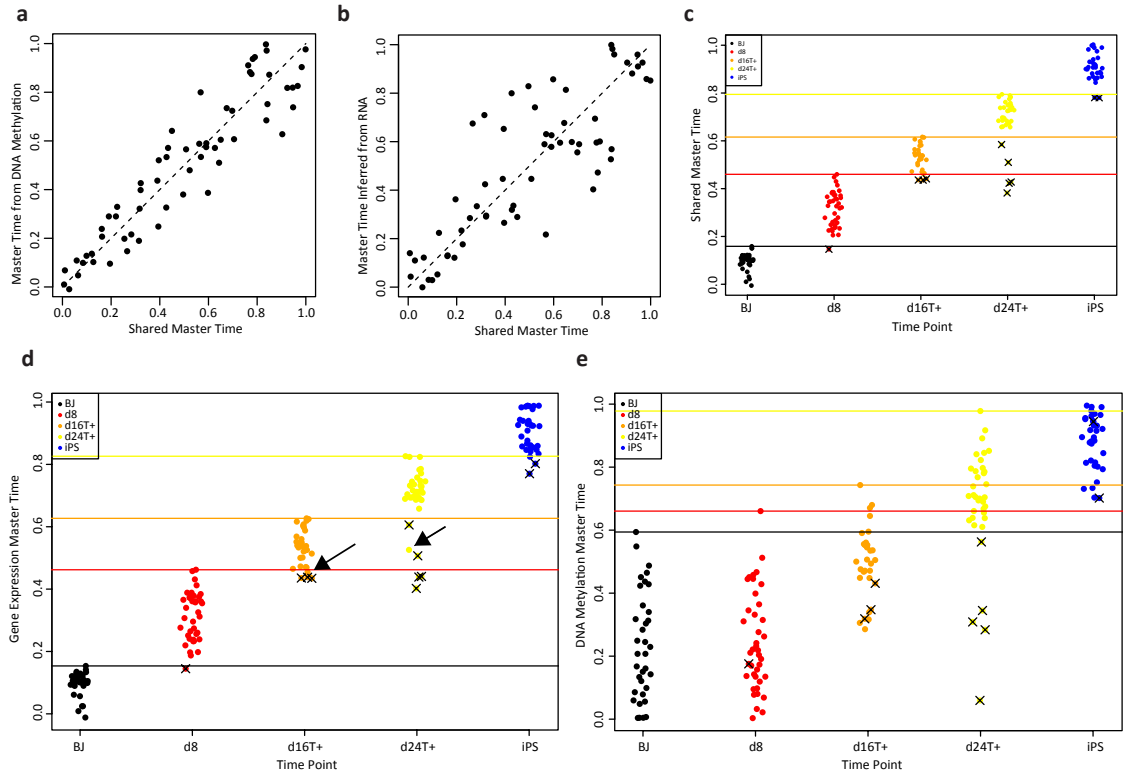



Figure 4.13: Incorporating known cell correspondence information to compute shared master time. (a) Scatterplot of shared master time inferred from both gene expression and DNA methylation (x-axis) and master time inferred using DNA methylation only (y-axis). (b) Scatterplot of shared master time inferred from both gene expression and DNA methylation (x-axis) and master time inferred using gene expression only (y-axis). (c) Plot showing "lagging cells" whose shared master time values overlap with the master time values of a previous time point. The x-values are jittered to mitigate overplotting. Colored horizontal lines indicate the maximum master time value for the corresponding time point. Lagging cells are indicated by "x" symbols. (d) Plot showing differences between lagging cells identified from shared master time and lagging cells identified from gene expression master time alone. The "x" symbols indicate lagging cells identified using shared master time. Arrows indicate two cells that are lagging based on gene expression master time along but not shared master time. (e) Plot showing differences between lagging cells identified from shared master time and lagging cells identified from gene expression master time alone. The "x" symbols indicate lagging cells identified using shared master time.

way to handle such situations would be to assign cells to branches before running MATCHER, and then perform manifold alignment on each branch separately.

Although the Hi-C protocol for measuring chromatin conformation has been adapted to single cells (Nagano et al., 2013), we did not include single cell Hi-C data in this study for two reasons. First, to the best of our knowledge, there are no published single cell Hi-C datasets from mouse embryonic stem cells. In addition, Hi-C data are a set of pairwise interactions (a matrix for each cell, rather than a vector), and it is not clear how to construct a trajectory from this type of data. Further work is necessary to investigate whether chromatin conformation shows sequential changes during biological processes, as well as the best ways infer such sequential changes and integrate them with other types of data.

One promising application of the method is aggregating single cell measurements into biologically meaningful groups. Cells can be grouped by their inferred master time values, and measurements within these groups can be aggregated. In experiments with thousands of cells, this will likely enable correlation between individual loci and related genes, which is currently impossible because of the extreme sparsity of the epigenetic data. Computational aggregation of measurements from many similar single cells may be the most immediate way to address the sparsity of single cell epigenetic measurements, although experimental protocols will likely improve over the long term.

MATCHER gives insight into the sequential changes of genomic information, allowing the use of both single cell gene expression and epigenetic data in the construction of cell trajectories. In addition, it reveals the connections among these changes, enabling investigation of gene regulatory mechanisms at single cell resolution. MATCHER promises to be useful for studying a variety of biological processes, such as differentiation, reprogramming, immune cell activation, and tumorigenesis.

# CHAPTER 5
# Quantifying Pseudogene Expression to Study the ceRNA Effect

## 5.1 Background

Recent studies have shown that some pseudogenes are transcribed and contribute to cancer when dysregulated. In particular, pseudogene transcripts can function as competing endogenous RNAs (ceRNAs). The high similarity of gene and pseudogene nucleotide sequence has hindered experimental investigation of these mechanisms using RNA-seq. Furthermore, previous studies of pseudogenes in breast cancer have not integrated miRNA expression data in order to perform large-scale analysis of ceRNA potential. Thus, knowledge of both pseudogene ceRNA function and the role of pseudogene expression in cancer are restricted to isolated examples.

To investigate whether transcribed pseudogenes play a pervasive regulatory role in cancer, we developed a novel bioinformatic method for measuring pseudogene transcription from RNA-seq data. We applied this method to 819 breast cancer samples from The Cancer Genome Atlas (TCGA) project. We then clustered the samples using pseudogene expression levels and integrated sample-paired pseudogene, gene and miRNA expression data with miRNA target prediction to determine whether more pseudogenes have ceRNA potential than expected by chance.

Our analysis identifies with high confidence a set of 440 pseudogenes that are transcribed in breast cancer tissue. Of this set, 309 pseudogenes exhibit significant differential expression among breast cancer subtypes. Hierarchical clustering using only pseudogene expression levels accurately separates tumor samples from normal samples and discriminates the Basal subtype from the Luminal and Her2 subtypes. Correlation analysis shows more positively correlated pseudogene-parent gene pairs and negatively correlated pseudogene-miRNA pairs than expected by chance. Furthermore, 177 transcribed pseudogenes possess binding sites for co-expressed miRNAs that are also predicted to target their parent genes. Taken together, these results increase the catalog of putative pseudogene ceRNAs and suggest that pseudogene transcription in breast cancer may play a larger role than previously appreciated.

Pseudogenes are genomic sequences sharing considerable sequence identity with protein-coding genes yet possessing features such as premature stop codons, deletions/insertions, or frameshift mutations that prevent them from producing functional proteins. There are three classes of pseudogenes: processed, duplicated, and unitary. A processed pseudogene lacks introns, resembling a spliced transcript that was inserted into the genome. A duplicated pseudogene is essentially a partial or complete copy of a protein-coding gene, including introns and sometimes even upstream regulatory elements. Thus, for any processed or duplicated pseudogene, there is an associated protein-coding gene called its parent gene that is highly similar in sequence. The third type of pseudogene is the unitary pseudogene, which arises when a protein-coding gene loses its coding potential through the accumulation of mutations. Unitary pseudogenes therefore do not have parent genes.

According to the GENCODE pseudogene annotations (v.17), there are nearly 15,000 human pseudogenes. Since their discovery in 1977, pseudogenes have generally been considered "biologically inconsequential" and non-functional (Jacq et al., 1977). Therefore, the discovery that a number of pseudogenes, such as PTENP1 (Fujii et al., 1999), are transcribed was somewhat surprising. The ENCODE project recently performed a survey of publicly available expression data to identify transcribed pseudogenes, and found over 800 pseudogenes with strong evidence of transcription (Pei et al., 2012). These transcribed pseudogenes showed both tissue-specific and constitutive expression profiles. In addition, many of the pseudogenes not found to be transcribed by ENCODE possessed properties indicative of transcription potential, including open chromatin, histone modifications that indicate transcriptional activity, transcription factor binding, and RNA polymerase II occupancy. Another recent study found evidence for over 2000 expressed pseudogenes in 13 different cancer and normal tissue types (Kalyana-Sundaram et al., 2012).

Although some pseudogenes are transcribed, this fact does not necessarily imply that pseudogene transcripts perform biologically important functions. However, recent research has revealed several mechanisms by which pseudogenes regulate gene expression. For example, in snail neurons, translation of the neural nitric oxide synthase mRNA is blocked by an antisense pseudogene transcript that binds to the mRNA (Korneev et al., 1999). Pseudogenes in mouse can form double-stranded RNA by base-pairing with their corresponding protein-coding genes and generate siRNAs to silence the expression of these genes (Tam et al., 2008). Pseudogenes may also compete with mRNAs for transcript stability factors, as in the case of the human HMGA1-p pseudogene (Chiefari et al., 2010).

The most recent function identified for pseudogenes is post-transcriptional regulation of mRNA levels by competing for miRNAs. This mechanism was first discovered in animals when it was shown that two human pseudogenes, PTENP1 and KRASP1, are transcribed and harbor miRNA response elements (MREs) for some of the same miRNAs that target their corresponding protein-coding genes, PTEN and KRAS, respectively (Poliseno et al., 2010). By binding and sequestering miRNAs that would otherwise bind and regulate PTEN or KRAS, the corresponding pseudogenes free the protein-coding genes from miRNA target repression. Thus, if the pseudogene is transcribed at a low level, more miRNAs will be able to target the parent gene transcripts, whereas an increase in pseudogene transcription will cause fewer miRNAs to target the parent gene. In this way, pseudogene RNA can compete with the parent gene RNA for miRNAs and thereby influence gene expression. This mechanism of regulation was first characterized in plants, where it was termed "target mimicry" (Franco-Zorrilla et al., 2007). Competition for miRNAs had also been used to create exogenous "miRNA sponges" containing specific MREs designed to soak up micro-ribonucleoprotein complexes and de-repress natural miRNA targets (Ebert et al., 2007). Salmena et al. coined the term competing endogenous RNA (ceRNA) to describe the function of PTENP1 and KRASP1 (Salmena et al., 2011). In theory, any type of RNA molecule, including mRNA, transcribed pseudogenes, and long non-coding RNA (lncRNA), can function as a ceRNA, provided the molecule shares at least one MRE with another RNA (Ebert and Sharp, 2010). A number of ceRNAs have been identified since the initial discovery of PTENP1 and KRASP1, including mRNAs (Tay et al., 2011; Sumazin et al., 2011; Karreth et al., 2011), and lncRNAs (Cesana et al., 2011). Non-coding transcripts may serve as more effective ceRNAs than mRNAs, since they are substrates for miRNA binding but are not translated. The absence of bound ribosomes on a non-coding transcript allows miRNAs to bind freely along the entire transcript rather than primarily in the regions that are outside the ribosome footprint as on mRNAs (Gu et al., 2009). Transcribed pseudogenes are especially strong ceRNA candidates because pseudogenes are identified by alignment with protein-coding genes, so by definition, they possess strong sequence similarity with their corresponding parent genes. This suggests that pseudogenes are likely to share MREs with their parent protein-coding genes. In fact, the sequence similarity between the PTEN coding gene and the PTENP1 pseudogene was one of the initial observations that led to the discovery of the ceRNA function of the PTENP1 pseudogene (Poliseno et al., 2010).

Interestingly, several transcribed pseudogenes play a key role in the development of cancer. PTENP1, KRASP1, and OCT4-pg4 are known to promote tumor progression through their roles as ceRNAs (Poliseno et al., 2010; Hayashi et al., 2015). The pseudogenes SUMO1P3 (Mei et al., 2013), ATP8A2-Ψ (Kalyana-

Sundaram et al., 2012), and Nanog-p8 (Uchino et al., 2012) have each been shown to enable cancer progression, but the mechanisms by which they do this are unknown. Ψ-PPM1K was shown to suppress oncogenic cell growth in hepatocellular carcinoma by generating endogenous siRNAs (Chan et al., 2013). ATP8A2-Ψ is an especially interesting case, because it is the first published example of a pseudogene that is differentially expressed among cancer subtypes, showing high expression in breast cancer samples with luminal histology but very little expression in basal samples (Kalyana-Sundaram et al., 2012). Also, ATP8A2-Ψ was shown to induce tumor progression when overexpressed in breast cancer cell lines (Kalyana-Sundaram et al., 2012).

Recently, a survey of RNA-seq data from The Cancer Genome Atlas project spanning seven cancer types showed that pseudogenes can be used to classify cancer samples into clinically relevant subtypes (Han et al., 2014). In particular, this study found that pseudogene expression alone separates endometrial cancer samples into groups corresponding to the major histological subtypes. Another interesting result from this study is that pseudogene-defined subtypes in kidney cancer show different patient survival rates. In addition, 547 pseudogenes with subtype-specific expression in breast cancer were identified. Finally, using miRNA expression data in conjunction with gene and pseudogene expression levels, they identified 38 pseudogenes with potential to function as ceRNAs in kidney cancer.

The pseudogenes that have been shown to participate in ceRNA interactions or play a role in cancer certainly represent provocative examples. However, the difficulty of reliably quantifying pseudogene expression and the lack of suitable datasets have hindered attempts to study these phenomena on a large scale. Therefore, it is not known whether pseudogenes like PTENP1 and ATP8A2-Ψ represent a few anomalous cases or point to a pervasive regulatory mechanism.

To begin to address this open and important question, we performed an investigation of the expression, subtype specificity, and ceRNA potential of transcribed pseudogenes in breast cancer using data from The Cancer Genome Atlas project (TCGA). The data include RNA-seq results for a total of 819 tumor and adjacent normal samples, along with sample-paired small RNA-seq. The dataset contains a representative sampling of breast cancer subtype, including 123 samples from the basal subtype, 60 her2 samples, 371 luminal A samples, and 170 luminal B samples. To the best of our knowledge, this study is the first to make use of sample-paired pseudogene and miRNA expression data to investigate the ceRNA mechanism in breast cancer.

## 5.2 Results

### 5.2.1 Reliable Quantification of Pseudogene Expression

Reliable quantification of pseudogene expression remains a challenging problem for a number of reasons. First, since parent genes and pseudogenes are highly similar in nucleotide sequence, short RNA-seq reads derived from one may align equally well to the other one. Such reads are fundamentally ambiguous in terms of their origin. Second, some reads may have nearly identical alignment to locations in the gene and pseudogene, and their mapping is often determined by the location with the least error in alignment. However, this strategy is unreliable in the presence of subject-specific variation with respect to the reference genome, or in the event of base call errors during sequencing, since these can result in an incorrect assignment of the read. Third, some aligners may follow a parsimony strategy in which a simple alignment is preferred to a complex (e.g. spliced) alignment. In the case of a processed pseudogene that lacks splices, this approach may erroneously bias the alignments to the pseudogene rather than the parent gene. Finally, in some cases, aligners report only a subset of possible alignments as a result of the heuristics used. For all of these reasons, studies of gene and pseudogene expression using existing tools are likely inaccurate without additional considerations.

A first approach to reliably studying pseudogene expression is to consider only the reads that are assigned to a single location by an aligner. However, the above confounding factors can result in reads that are uniquely aligned to the wrong positions (Figure 5.1). Any conclusions drawn from such reads in downstream analyses will be unreliable. One approach to addressing this problem is to identify and discard from the analysis reads that map to regions in the genome that are especially sensitive to these confounding factors. We have adopted this approach using the concept of transcriptome mappability, which we describe below.

Our approach for computing transcriptome mappability builds on the notion of genomic mappability. Mappability is a measure of the inherent distinctiveness of a genomic region; the more frequently a genomic region occurs, the less mappable it is. Although mappability can be defined as a continuous quantity (the reciprocal of $k$-mer frequencies, for example, as in (Derrien et al., 2012)), it is generally not very useful to know the degree to which a region is unmappable. If a $k$-mer occurs more than once in the genome, a read aligned there will be ambiguous. For this reason, we compute mappability as a discrete quantity-that is, a region is either mappable (unambiguous) or not mappable (ambiguous). Our notion of mappability also includes a "safety margin", so that a mappable region guarantees not only a unique alignment for the reads matching the sequence, but also that no read with one or two base call errors or SNPs relative to the reference

genome could be uniquely mismapped to this region. Mappability is important even if an aligner does not use heuristics and exhaustively enumerates read alignments. As demonstrated by Figure 5.1A, highly similar regions can produce uniquely mismapped reads as a result of genome variation and read errors in a way that no aligner can recognize (see Methods section for details).

If we restrict our attention to alignments in mappable regions, we ensure that the downstream analysis results are robust, even if the reference genome does not match the subject genome or the reads contain sequencing errors. Mappability is thus inversely related with sensitivity to genome variation and read errors.

Since RNA-seq reads may span multiple exons, the transcriptome contains additional $k$-mers beyond those found in the genome. To compute transcriptome mappability, we can align $k$-mers to the genome sequences crossing splice junctions. This transcriptome mappability scheme allows the computation of pseudogene expression levels using only reads uniquely aligned to mappable regions. Using these reliable reads, we compute pseudogene expression levels in units of Reads per Kilobase of Uniquely mappable transcript per Million reads (RPKUM). See the Methods section for a detailed description of the transcriptome mappability and RPKUM calculations.

We tested our RPKUM metric by comparing expression levels for protein coding genes computed with both RPKUM and RSEM (Li and Dewey, 2011), a commonly used transcript quantification method. We computed the mean expression level across the TCGA dataset for each protein-coding gene using both methods, then calculated the correlation between the expression levels from the two methods. The result showed good agreement between RPKUM and RSEM values (Spearman correlation$> 0.85$), indicating that RPKUM values provide a reliable method for quantifying expression levels.

An important question is whether RPKUM values computed from few mappable bases are trustworthy. To investigate the robustness of the RPKUM metric, we simulated RPKUM values by randomly sampling positions of genes that are completely mappable and then using these sampled bases as the only mappable bases of a gene in an RPKUM calculation. Genes spanning a wide range of expression levels from 1 to 200 RPKMs were used in the simulation. We performed the simulations with 500, 100, and 50 mappable bases per gene. RPKUM values computed from genes with as few as 50 simulated mappable bases showed very strong agreement with the true RPKM expression levels across the range of expression levels ($\rho = 0.95$). In addition, increasing the number of mappable bases slightly increases the correlation between RPKUM and RPKM levels ($\rho = 0.97$ for 100 mappable bases and $\rho = 0.99$ for 500 mappable bases).

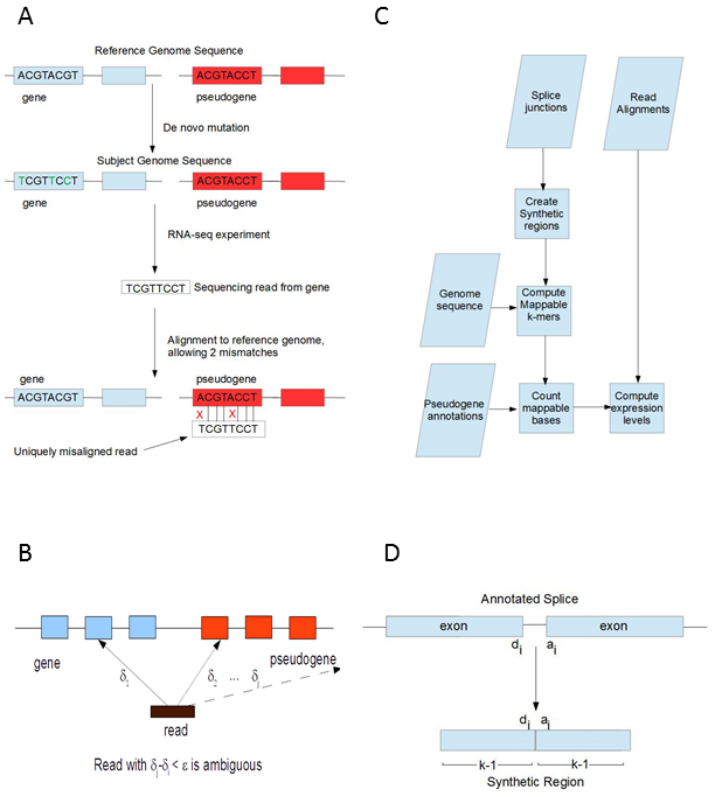



Figure 5.1: Reliable quantification of pseudogene expression. (A) Example showing that even an ideal aligner may produce uniquely misaligned reads in the presence of mutations and read errors if alignments to unmappable regions are considered trustworthy. The problem arises because the sequences of the gene and pseudogene are sufficiently similar that unique misalignment cannot be ruled out. (B) If a read has at least two alignments that are at distance $\delta_1$ and $\delta_2$ from the reference genome, respectively, then the true position of the read should be considered ambiguous unless $|\delta_1 - \delta_2| > \epsilon$ for some integer safety margin $\epsilon > 0$. (C) Pipeline for computing RPKUM expression levels for pseudogenes. (D) "Synthetic regions" around splice junctions are used to extend mappability to the transcriptome. A synthetic region is constructed by concatenating $k1$ nucleotides from the donor and acceptor exons on either side of a splice junction. Any $k$-mer that crosses the splice junction thus occurs in the synthetic region.

Figure 5.2A shows the distribution of transcriptome mappability for protein coding genes and GENCODE v. 17 pseudogenes. As expected, pseudogenes are much less mappable than protein-coding genes; the median protein-coding gene mappability value is nearly 100% of gene length, and the vast majority of genes are almost completely mappable. In contrast, the median pseudogene mappability value is around 80% of pseudogene length. The distribution of pseudogene mappability is approximately bimodal, with peaks near 10% and 90%. A sizable fraction of pseudogenes are completely unmappable (2169 out of 14942). Nonetheless, the majority of pseudogenes possess a significant fraction of mappable bases and are thus accurately detectable using RPKUM expression levels.

As expected, restricting the set of reads aligned to pseudogenes to only those in mappable regions leads to a dramatic reduction in the number of reads (Figure 5.2B). On average, each sample contains nearly 10 million reads mapped to pseudogenes, but our filtering process leaves a set of just over 360,000 pseudogene reads per sample. The surviving reads comprise a high-confidence set that can be used to assess pseudogene transcription.

### 5.2.2 High-confidence breast cancer pseudogene transcripts

Using the GENCODE v. 17 pseudogene annotations, we identified 2012 pseudogenes with evidence of transcription, defined as genes with at least 50 mappable bases, 50 reads, and 1 RPKUM in at least 1 sample. The majority of these pseudogenes occurred in only a small number of samples (Figure 5.3A). However, a subset of the pseudogene transcripts occurs in a large number of samples, including 94 pseudogenes that are transcribed in over 95% ($n = 780$) of the samples. To investigate the pseudogenes that are most likely to play a role in cancer biology, we chose to focus the remainder of our analysis on pseudogenes that exhibited evidence of transcription in at least 10% ($n = 80$) of the samples; this set consists of 440 pseudogenes.

The GENCODE pseudogene decoration resource (psiDR v. 0), assembled from a recent genome-wide survey of pseudogenes using ENCODE data (Pei et al., 2012), provides useful information for an initial assessment of the transcriptional potential of our pseudogene set. Out of the set of 440 transcribed pseudogenes we identified, 287 pseudogenes are annotated in psiDR for a number of attributes, including pseudogene type, parent gene, transcription evidence, open chromatin, histone modifications that indicate activity, transcription factor binding, RNA polymerase II occupancy, and evolutionary constraint (Pei et al., 2012). Although the functional genomics annotations come from the ENCODE cell lines, not from breast

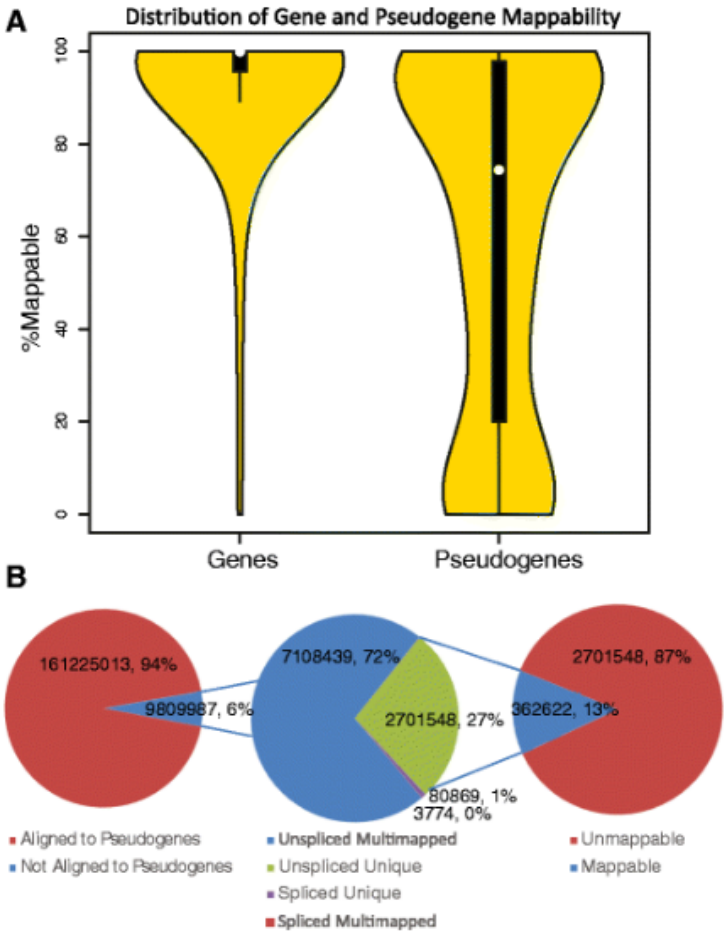



Figure 5.2: Pseudogene mappability and read alignments. (A) Violin plot showing the distribution of gene and pseudogene mappability as a percentage of gene length. The dot in the middle of each plot represents the median, and the black box is the interquartile range. (B) Pie charts showing how many reads are removed by mappability filtering. From left to right: Fraction of all aligned reads that map to pseudogenes; fraction of reads aligned to pseudogenes that are uniquely aligned; and fraction of reads uniquely aligned to pseudogenes that are also mappable.

cancer tissue, they nonetheless serve as a reasonable starting point for assessing the transcriptional activity of the pseudogenes we identified.

Examining the collection of psiDR annotations for these 287 transcribed pseudogenes shows that they possess a number of properties that indicate transcriptional activity (Figure 5.3B). Nearly half ($n = 125$) of the 287 pseudogenes were reported by psiDR to be transcribed. The remainder ($n = 162$) represent potentially novel pseudogene transcripts not annotated in psiDR. The pseudogenes producing these unannotated transcripts show strong evidence of transcriptional activity. Compared to the full set of more than 11,000 pseudogenes annotated by psiDR, the set of 287 is significantly enriched for active chromatin, Pol II occupancy, and transcription factor binding ($p < 0.002$,$\chi^2$ test). In addition, 20 of these pseudogenes display fewer substitutions compared to chimp and mouse orthologs than expected by chance. Interestingly, duplicated and unitary pseudogenes are also enriched within the set of 287. This may be due in part to the fact that duplicated pseudogenes are thought to be more likely to possess upstream regulatory elements similar to those of the parent genes. Also, unitary pseudogenes are likely to be more mappable, and thus are easier to detect from short-read RNA-seq data. In short, the diverse data types from the ENCODE project provide strong support for the transcriptional activity of the pseudogenes that we have detected in breast cancer tissue.

It is worth noting that PTENP1 and KRASP1, the two initial examples of pseudogene ceRNAs, are present (though at low levels) in the breast cancer samples we study here. Our method of computing RPKUM expression levels is thus capable of detecting these important pseudogenes, but their expression levels fall below the cutoff that we used to define our set of highly-expressed pseudogene transcripts, and therefore they were not considered for further analysis. The set of 748 breast-cancer pseudogene transcripts provided by Han et al. (Han et al., 2014) does not contain PTENP1 or KRASP1, confirming the low expression of these pseudogenes in breast cancer.

### 5.2.3 Hierarchical clustering shows association with known cancer subtypes

The four molecular subtypes of breast cancer possess a number of distinguishing characteristics, including estrogen/progesterone receptor status, response to chemotherapy drugs, and gene expression profile (Perou et al., 2000). A common method of studying the differences among these subtypes is to use unsupervised clustering techniques to group samples together based on their gene expression patterns. Unsupervised clustering using protein-coding genes results in four distinct clusters corresponding to the subtypes (Perou et al., 2000). To investigate the relationship between pseudogene transcription and breast cancer disease

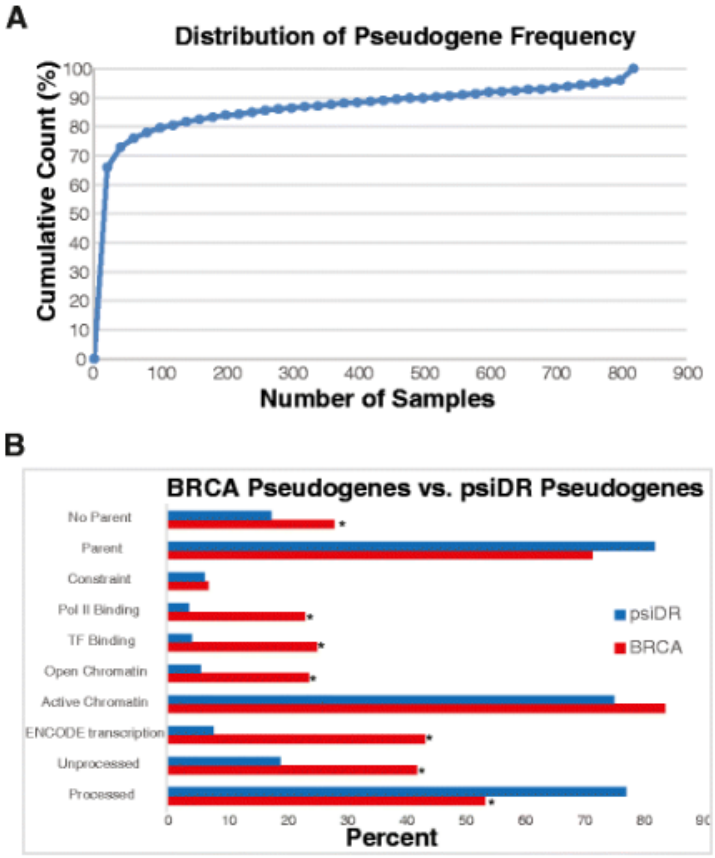



Figure 5.3: Pseudogene occurrence in the TCGA breast cancer samples and overlap with ENCODE functional genomics annotations. (A) Cumulative distribution function showing how many samples pseudogenes occur in. Approximately 65% of the 2,012 transcribed pseudogenes occur in fewer than 20 samples. Roughly 25% of the pseudogenes occur in at least 80 samples. (B) Bar chart comparing the set of 287 pseudogenes transcribed in breast cancer with the full psiDR v. 0 annotation set. The asterisks indicate categories that are significantly enriched in the set of 287 pseudogenes compared to the full set ($p < 0.002$, $\chi^2$ test).

state, we performed hierarchical clustering using the high-confidence set of 440 pseudogenes. Unsupervised clustering based solely on these pseudogene expression levels effectively separates tumor and normal samples (Figure 5.4A). However, since the normal samples are extracted from tumor adjacent breast tissue that contains a different cell type composition than the tumor itself, the ability to distinguish tumor from normal is likely due in large part to tissue specificity rather than tumor biology. Nonetheless, this result shows that pseudogene expression varies considerably between the cell types that make up the tumor and adjacent normal samples.

We also removed the adjacent normal samples and clustered solely on the tumor samples. As Figure 5.4B shows, the unsupervised clustering algorithm successfully separates the basal samples from the other subtypes. However, the pseudogene expression profiles for the luminal and Her2 subtypes are not sufficiently distinct to consistently separate samples from these subtypes. Basal tumors grow more rapidly and have significantly different histology than the other subtypes (Perou et al., 2000), and this may be why basal/luminal and basal/Her2 separation stands out more clearly than the luminal/Her2 separation. The fact that pseudogene expression alone can identify the basal subtype shows that pseudogene expression has a strong, non-random association with specific pathways and cellular environments. This suggests that previous findings, such as the pseudogene ATP8A2, which is more highly expressed in luminal compared to basal samples (Kalyana-Sundaram et al., 2012), are not isolated examples.

To identify the pseudogenes with the most strong subtype-specific expression profiles, we performed a multi-class differential expression analysis using the SAM tool (Tusher et al., 2001). This analysis yielded 309 pseudogenes with significant subtype-specific expression (FDR$< 1\%$). Several interesting pseudogenes are at the top of this list. For example, the second pseudogene on the list is ATP8A2-$\Psi$, a pseudogene that has been found to be upregulated in luminal subtypes and shown to induce tumor progression (Kalyana-Sundaram et al., 2012). The expression profile found here reflects this pattern, showing strong upregulation in luminal samples compared to basal.

Three other interesting examples are shown in Figure 5.5. A pseudogene of CASP4, a member of the caspase family known to initiate apoptosis under certain conditions (Hitomi et al., 2004), is expressed at higher levels in basal samples and downregulated in luminal A samples (Figure 5.5A). Interestingly, the expression of the CASP4 pseudogene is lower in tumor samples than normal, which is the expression profile expected for a ceRNA that promotes CASP4 expression. Additionally, the CASP4 pseudogene was found to be transcribed in the ENCODE analysis (Pei et al., 2012). Another interesting property of this unprocessed

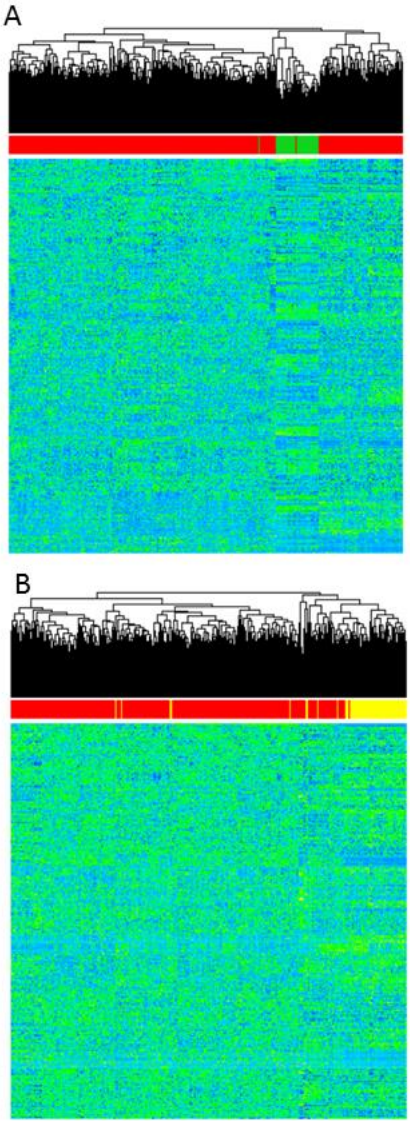



Figure 5.4: Hierarchical clustering based on pseudogene expression shows pseudogene association with breast cancer subtypes. (A) Heatmap showing pseudogene expression profiles in tumor and adjacent normal samples. High expression levels are shown in light green, and low expression levels are shown in light blue. Tumor samples are highlighted in red along the top of the plot; adjacent normal samples are highlighted in green. (B) Heatmap of pseudogene expression profiles in tumor samples. Samples belonging to the basal subtype are highlighted in yellow along the top of the plot.

pseudogene is that it shows alternative splicingthere appear to be multiple isoforms represented in the reads covering the pseudogene locus. Intriguingly, our analysis of potential ceRNA interactions also indicated that the CASP4 pseudogene is positively correlated ($\rho = 0.3$) with expression of its parent gene and shares a miRNA target site for hsa-mir-203 (see next section for detailed summary of ceRNA investigation).

The CYP2F1 pseudogene is expressed at quite high levels compared to most pseudogenes in the dataset, and the average expression level in the luminal B subtype is nearly five times the average expression in the basal subtype. The pseudogene is a unitary pseudogene, with no clear parent protein-coding gene. However, it possesses strong sequence similarity with the cytochrome P450 family of genes. It was previously demonstrated that CYP2F1 is expressed in colorectal cancer and that expression in primary tumors correlated with corresponding metastatic tumors in lymph nodes (Kumarakulasingham et al., 2005). Like the CASP4 pseudogene, the CYP2F1 pseudogene shows evidence for multiple isoforms.

A pseudogene of the MSL3 gene shows nearly twice the expression level in basal compared to luminal A (Figure 5.5C). The processed pseudogene was found to be transcribed in the ENCODE analysis. The MSL3 protein is thought to play a function in chromatin remodeling and transcriptional regulation, and it has been reported as part of a complex that is responsible for histone H4 lysine-16 acetylation (Smith et al., 2005). Furthermore, expression of this pseudogene is correlated with the expression of its parent gene ($\rho = 0.3$), and it is predicted to share target sites for six different miRNAs (see next section for detailed summary of ceRNA investigation).

### 5.2.4 Analysis incorporating miRNA and gene expression levels reveals pseudogenes with ceRNA potential

A common hypothesis about ceRNA interactions is that if transcript A sequesters miRNA C away from transcript B, the expression levels of A and B will be positively correlated, while both A and B will be negatively correlated with C. To assess the possibility that the transcribed pseudogenes identified may function as ceRNAs for their parent genes, we performed an analysis integrating miRNA target prediction with pseudogene, gene, and miRNA expression levels. The miRNA expression levels were computed from sample-paired TCGA small RNA-seq data using a previously described small RNA-seq analysis pipeline (Baran-Gale et al., 2013). We computed expression levels for the parent genes of the pseudogenes using the same RPKUM method as for the pseudogenes.

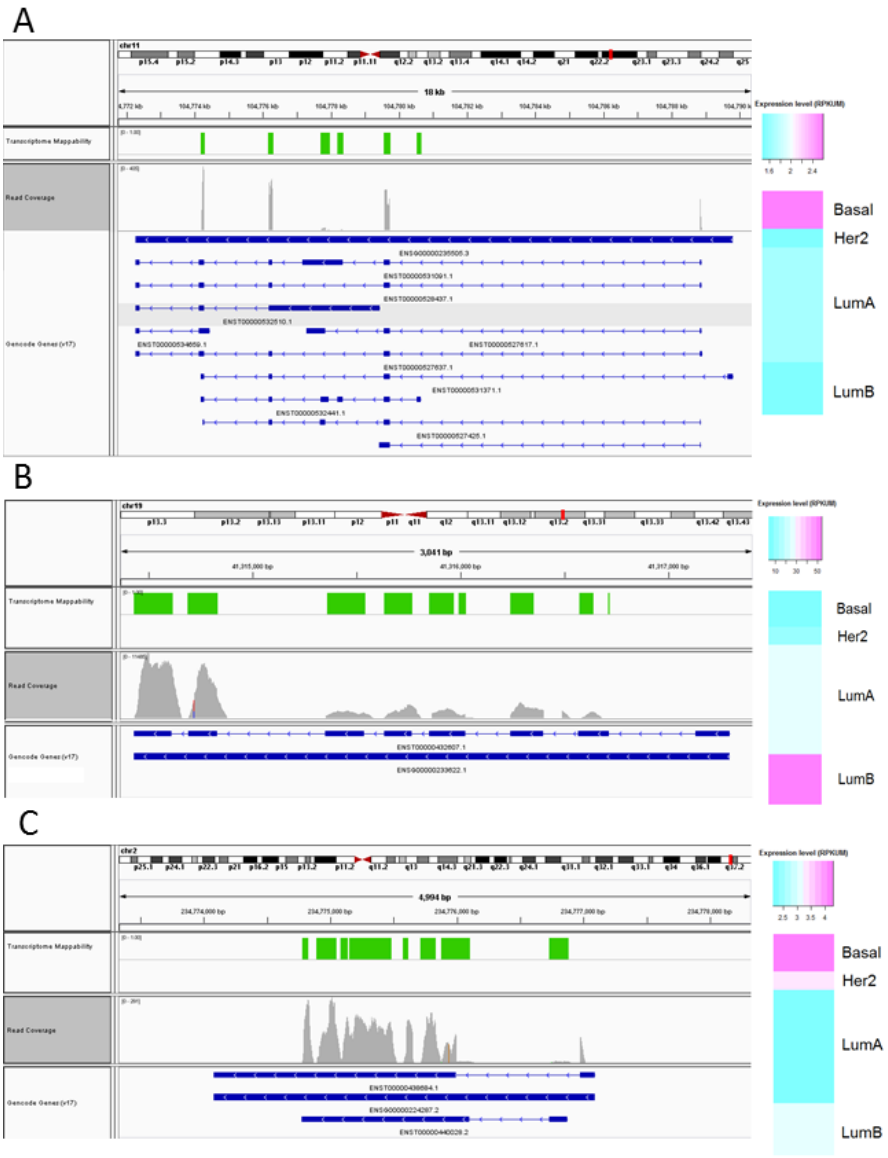



Figure 5.5: Read coverage, mappability, and tumor expression profile for (A) CASP4 pseudogene, (B) CYP2F1 pseudogene, and (C) MSL3 pseudogene.

Since pseudogenes are non-coding RNAs and are not densely bound by ribosomes, the vast majority of the transcribed region of a pseudogene is likely accessible for miRNA binding. However, if a pseudogene serves as a miRNA sponge for its parent gene, it is more likely that the shared miRNA binding site occurs in the 3' UTR of the parent gene than in the coding region. In addition, using a restricted region for prediction somewhat ameliorates the lack of specificity common to miRNA target prediction algorithms (Ritchie et al., 2009). We therefore chose to restrict our target prediction analysis to the portion of the pseudogene with sequence similarity to the 3' UTR of the parent genewhat might be termed the "pseudo-3' UTR". During the process of performing miRNA target prediction on pseudogenes, we noticed that the GENCODE pseudogene annotations often did not span the pseudo-3' UTR. Therefore, we used BLAST to identify the pseudo-3' UTRs of pseudogenes by aligning the GENCODE annotation and surrounding genomic context with the annotated 3' UTRs of the parent gene (see Methods section for details). TargetScan version 7 (Grimson et al., 2007) was used to predict target sites for only the top 100 miRNAs expressed in the TCGA breast cancer dataset. This analysis revealed 177 transcribed pseudogenes that are predicted to share at least one miRNA target site with their corresponding parent genes.

We computed Pearson correlation coefficients for each pseudogene-parent gene pair. As the plot in Figure 5.6 shows, the majority of pseudogene-parent gene pairs are uncorrelated. However, there is a positive skew to the distribution of correlations. To test whether the distribution of correlations differs significantly from expectation, we performed a permutation test. We constructed 5000 sets of gene-pseudogene pairs in which the genes and pseudogenes were randomly paired. The sets were of the same size as the set of pseudogene-parent gene pairs. For each random set, we computed the number of pairs with Pearson correlation above 0.3. In the 5000 random sets we generated, there were never more than 15 such pairs per set (Figure 5.6C). However, the set of correlations resulting from pairing pseudogenes and parent genes contains 55 pairs with correlation above 0.3. This indicates that the positive skew to the distribution of correlations shown in Figure 5.6A is very unlikely to be due to chance. We also tested an additional correlation threshold of 0.5 and observed a similar result, indicating that our findings are robust to the choice of correlation threshold.

We also computed the correlation between the expression level of each pseudogene and the miRNAs predicted to target it. The correlations observed for these pseudogene-miRNA pairs closely approximate a normal distribution, but show a slight negative trend (Figure 5.6B). A total of 180 pseudogene-miRNA pairs show strong negative correlation of less than 0.3. To test whether this number of pairs is significant, we

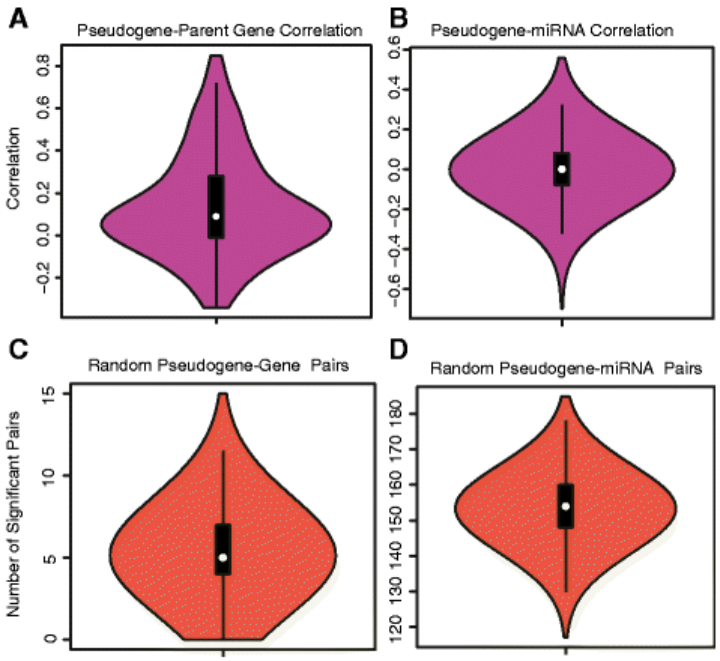



Figure 5.6: Violin plots summarizing pseudogene-parent gene and pseudogene-miRNA pairwise correlations. Correlations between (A) expressed pseudogenes and parent genes and (B) expressed pseudogenes and expressed miRNAs predicted to target them. Results of permutation analysis showing how many correlated pseudogene-parent gene pairs (C) and pseudogene-miRNA pairs (D) were found.

approximated a null distribution of pseudogene-miRNA correlations using the same permutation method we applied to the pseudogene-parent gene pairs. Randomly shuffling the pseudogene-miRNA pairs to create 5000 random sets (Figure 5.6D) showed only 5 permutations with at least as many strongly anti-correlated pairs as we observed in the data, which corresponds to an empirical p-value of 0.001. This supports the conclusion that the extent of negative correlations observed in the data cannot be attributed solely to chance, and is likely due to genuine miRNA target repression.

Next we sought to identify the pseudogene-parent gene-miRNA triples with the strongest ceRNA potential. To do this, we first identified expressed miRNAs predicted to target both a pseudogene and its parent gene. For each such triple, we computed the correlation between pseudogene and parent gene, pseudogene and miRNA, and parent gene and miRNA. We also computed p-values with Benjamini-Hochberg FDR correction for the miRNA correlations. In this way, we identified 17 pseudogene-gene pairs with strong ceRNA potential, which we defined as pseudogene-gene correlation greater than 0.3 and statistically significant miRNA anti-correlation.

Two of these pseudogenes stand out as especially interesting examples. A pseudogene of GBP1 and its parent gene show statistically significant anti-correlation with hsa-mir-199a, which has been shown to regulate autophagy in breast cancer cells (Yi et al., 2013). This pseudogene was also found to be transcribed in the ENCODE analysis (Pei et al., 2012). The parent gene GBP1 is known to be the mediator of the anti-proliferative effect of inflammatory cytokines in endothelial cells (Guenzi et al., 2001), and is implicated in several types of cancer according to GeneCards. In addition, the GBP1 pseudogene shows strong positive correlation with the expression of its parent gene across the TCGA dataset ($\rho = 0.82$). Another interesting pseudogene is SUZ12P1. This pseudogene and its parent gene both show strong anti-correlation to hsa-mir-28. SUZ12P1 also shows moderate positive correlation with its parent gene ($\rho = 0.41$). The parent gene, SUZ12, is a polycomb group protein and part of the PRC2/EED-EZH2 complex, an important epigenetic regulator that performs histone methylation (Cao and Zhang, 2004). This gene is also frequently translocated in endometrial stromal tumors, where it forms the JAZF1-SUZ12 oncogene (Amador-Ortiz et al., 2011).

An interesting question is whether the genes that have pseudogenes with ceRNA potential are functionally related. To investigate this question, we performed a Gene Ontology (GO) term enrichment analysis using three different sets of parent genes. The sets of genes used were parent genes strongly correlated with a pseudogene, parent genes whose pseudogenes was strongly anti-correlated with a shared miRNA, and parent genes participating in a putative gene-pseudogene-miRNA ceRNA interaction as defined above. For each of these sets of parent genes, we used the GOrilla tool with default settings to look for GO terms enriched in the set compared to the background list of all parent genes. No significantly enriched GO terms were found for any of the 3 sets of interest, indicating that there is no clear functional relationship among the parent genes in the sets that we have identified.

## 5.3 Discussion

The recent paper by Han et al. that investigated pseudogene expression in cancer (Han et al., 2014) identified 748 pseudogenes transcribed in breast cancer, 547 of which showed subtype-specific expression. Although the results of Han et al. partially overlap with our own, our study is distinct in two key ways: (1) we investigate the ceRNA potential of pseudogenes transcribed in breast cancer, but Han et al. do not and (2) we use a more detailed method for measuring pseudogene transcription, designed to maximize specificity. In an effort to avoid the artifacts that plague pseudogene transcription detection, we designed our analysis to be as

conservative as possible. Consequently, the set of pseudogenes detected by our method is somewhat smaller. However, our set of pseudogenes is not simply a subset of theirs. Out of the 440 pseudogenes we detect, only 174 were also found by Han et al. (Figure 5.7B). The remaining 266 represent novel pseudogene transcripts. In addition, 103 of the subtype-specific pseudogenes we identified overlap with the set of subtype-specific pseudogenes presented in Han et al. (Figure 5.7C).

To understand why our set of pseudogenes is substantively different from that of Han et al., we carefully analyzed how they computed pseudogene expression levels. They used 75-mers to compute mappability, and decided for each exon whether to include or exclude reads for the entire exon. One shortcoming of this approach is that it either includes or excludes reads for entire exons, rather than making decisions for individual reads. In our experience, small islands of similarity within an otherwise distinct exon are often enough to promote false positive read alignments. Conversely, small islands of distinct sequence within an exon can be used to detect the presence of pseudogene transcripts. As a result, our approach detected 266 pseudogenes with strong evidence of transcription that were overlooked in Han et al. (Han et al., 2014). Another limitation is that the analysis in (Han et al., 2014) did not account for the presence of splice junctions inserted into the genome. Processed pseudogenes containing concatenated exons are a major source of error in pseudogene RNA-seq alignments because RNA-seq aligners sometimes prefer unspliced alignments to spliced, particularly in the presence of SNPs. However, genomic mappability as used in (Han et al., 2014) cannot detect such artifacts.

A more serious problem is that although the RNA-seq reads from the TCGA BRCA data are 50 bases long, Han et al. use mappability based on 75-mers to decide which pseudogenes are mappable. Given that longer sequences are more likely to be distinct in the genome, this mismatch between read length and the k-mer size used to compute mappability means that an exon that appears completely mappable may nonetheless have many misaligned reads. Figure 7A shows the difference in mappability obtained from 75-mers without accounting for splice junctions inserted in the genome and 50-mers. In the first case, the median mappability as a percentage of gene length is 94%, but in the second case it is 74%. The use of 75-mers as in (Han et al., 2014) rather than 50-mers results in a loss of specificity. Thus, it is possible that some of the pseudogenes transcripts detected in this way are not actually transcribed, but are simply read alignment artifacts.

In summary, two major differences between the approach of Han et al. and our own method for computing pseudogene expression explain the differing lists of pseudogenes that were obtained. First, Han et al. either

kept or removed entire pseudogene exons, while we made the decision for each individual read; this explains why we detected some pseudogenes that they did not. Second, Han et al. used 75-mers to compute genome mappability, but we used 50-mers and accounted for processed pseudogenes containing splice junctions; consequently, our list of pseudogenes did not include some of theirs. We emphasized specificity in our algorithm in order to facilitate the identification of the highest confidence pseudogenes and candidate ceRNAs for further analysis. If the methods used to derive pseudogene expression levels do not properly account for misaligned reads, it is difficult to exclude the possibility that apparent pseudogene-based classification of subtypes are actually driven by improperly aligned reads from protein-coding genes with subtype-specific expression. Furthermore, such misaligned reads could bias toward stronger positive correlations between parent genes and pseudogenes.

In this paper, we undertook an initial investigation to address the important questions of how pervasive the pseudogene ceRNA mechanism is and how pseudogene transcription relates to breast cancer subtype. Careful scrutiny of RNA-seq evidence yielded a high-confidence set of pseudogene transcripts, a subset of which exhibit strong subtype-specific expression and are candidates for ceRNA function. Further experimental work is needed to examine these candidates; in particular, assays for miRNA binding and siRNA knockdown experiments can provide more conclusive evidence for ceRNA interactions in individual gene-pseudogene pairs. Follow-up studies are also needed to determine the nature of the relationship between pseudogene expression and subtype. Many of the subtype-specific pseudogene transcripts are likely passengers rather than drivers. However, some of these may play a role in the tumor progression of individual subtypes, as was demonstrated in the case of ATP8A2-$\Psi$.

The integration of pseudogene, gene, and miRNA expression data demonstrates that while not all pseudogenes may function as ceRNAs, the phenomenon is likely more pervasive than currently appreciated. One limitation of our approach is that ceRNA activity may not always be indicated by positive correlation between a pseudogene and its parent gene or negative correlation between a pseudogene and its targeting miRNA. For example, if the miRNA regulation of a pseudogene is very strong, leading to rapid and robust degradation of the pseudogene, this could produce a negative correlation between pseudogene and parent gene. Furthermore, it is well-known that regulatory network structures such as incoherent feed-forward loops can produce positive correlation between an mRNA and a targeting miRNA (Tsang et al., 2007). Even with this limitation, our results suggest that more pseudogenes than currently known likely function as ceRNAs, and more detailed experimental work is required to determine the physiological significance of this function.

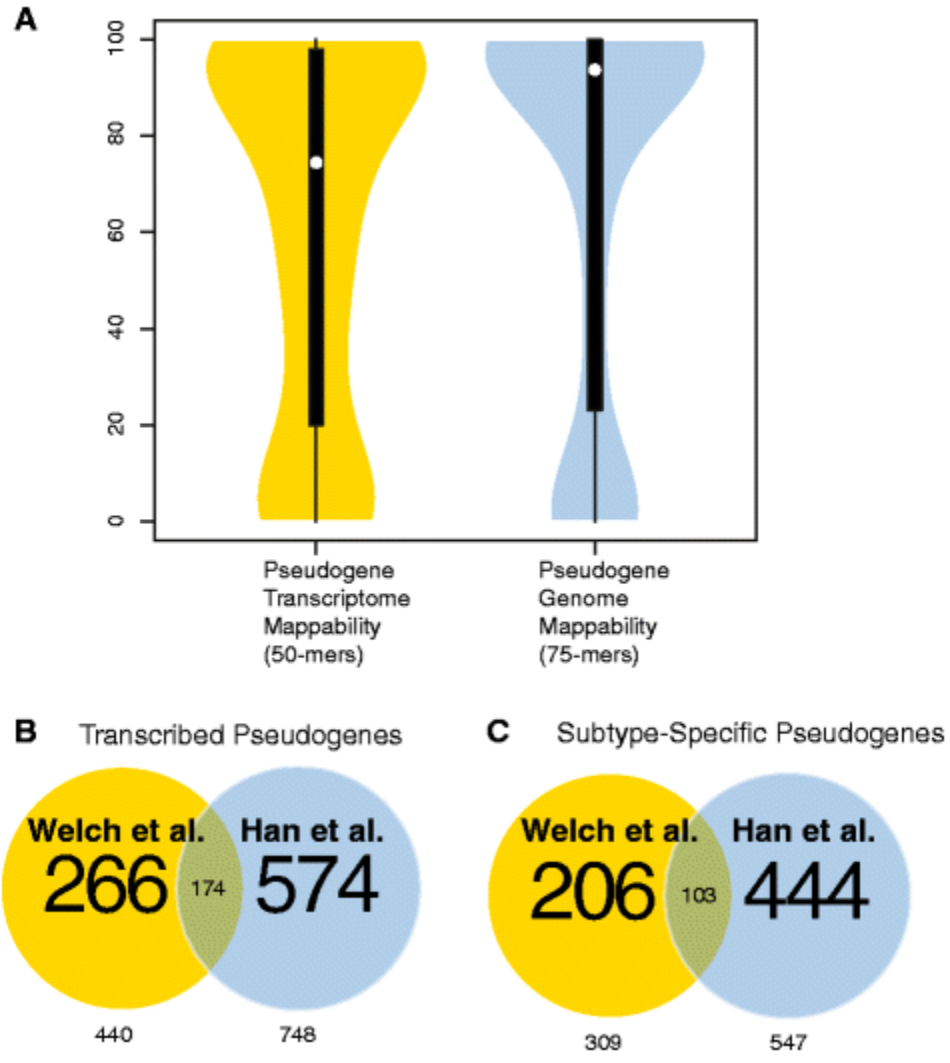



Figure 5.7: Comparison with the results of Han et al. (A) Violin plots showing the difference in pseudogene mappability when using 50-mers and accounting for splice junctions inserted in the genome (yellow) and 75-mers (blue). (B) Comparison with breast cancer pseudogene transcripts found by Han et al. (C) Comparison with breast cancer subtype-specific pseudogene transcripts found by Han et al.

## 5.4 Methods

### 5.4.1 Computing transcriptome mappability

A first approach to reliably studying pseudogene expression is to consider only reads that are assigned to a single location by an aligner. However, the confounding factors of SNPs, read errors and aligner heuristics can result in reads that are uniquely aligned to the wrong positions (Figure 5.1A). We refer to such reads as uniquely misaligned reads. Any conclusions drawn in the presence of uniquely misaligned reads in downstream analyses will be unreliable. In order to guard against this problem, we should distrust any reads for which there exist multiple possible alignments whose distance from the genome is less than some safety margin $\epsilon$ (Figure 5.1B). In such cases, there is sufficient ambiguity that we cannot rule out the possibility of unique misalignment.

To address the problem of read mismapping between genes and pseudogenes, we developed an approach based on the concept of mappability. Since RNA-seq reads may span multiple exons, the transcriptome contains additional $k$-mers beyond those found in the genome. In considering transcriptome $k$-mers, two cases arise that are particularly problematic for pseudogenes: processed pseudogenes with integrated splice junctions and duplicated pseudogenes that may have highly similar splice junctions to their parent genes. The former case is particularly problematic because RNA-seq aligners sometimes prefer direct alignments to spliced alignments, causing spuriously aligned reads to accumulate on processed pseudogenes. To compute transcriptome mappability, we consider $k$-mers from the genome and "synthetic regions" surrounding splice junctions (Figure 1D). The synthetic region around a splice junction is the concatenation of the immediately adjacent $k1$ bases from donor and acceptor exons. These regions thus contain any $k$-mers that span annotated splice junctions. For a given genome $G$, transcriptome $T$ (represented as $k$-mers from synthetic regions), position $i$, read length $k$ and error tolerance $\epsilon$, we define the mappability of position $i$ as a Boolean quantity:

$$M(G,T,i,k,\epsilon) = \begin{cases} 0 \text{ if } G_i...G_{i+k-1} \text{ is within Hamming distance } \epsilon \text{ of any other } k\text{-mer in } G \text{ or } T \\ 1 \text{ otherwise} \end{cases} \tag{5.1}$$

### 5.4.2 Finding transcribed pseudogenes

We filtered reads by requiring that either (1) the read has a unique, direct alignment to the genome starting at position $i$ and this position is mappable or (2) the read has a unique, spliced alignment and the spliced $k$-mer to which the read is aligned occurs exactly once in the genome and transcriptome. We refer to reads surviving this filtering as "mappable reads". Ensembl protein-coding gene annotations and GENCODE pseudogene v. 14 annotations were used to compute synthetic regions around splice junctions.

The number of mappable bases for each pseudogene was computed by constructing a "consensus pseudogene model" in which all annotated exons are merged into a nonredundant set of positions including all potentially transcribed regions from the gene model. We count a position within the resulting nonredundant set of transcript positions as mappable if either (1) the corresponding position in the genome is mappable or (2) a mappable spliced read occurs at that position.

Using the reliably mapped reads and mappable bases, compute pseudogene expression levels in units of Reads per Kilobase of Uniquely mappable transcript per Million reads (RPKUM):

$$\text{Expression level in RPKUM} = \frac{\text{Mappable reads from pseudogene} \times 10^9}{\text{Mappable bases in pseudogene} \times \text{total reads}} \tag{5.2}$$

The justification for computing expression levels in units of RPKUM instead of RPKM is that reads aligned to unmappable regions are not considered in the expression level calculation, so counting the total number of bases in the transcript would underestimate the expression level. One limitation of the RPKUM metric is when the regions used to determine pseudogene transcription are disjoint from a transcript isoform. In such a case the RPKUM expression measurement does not include the expression of the unmappable isoform. Out of 14,943 pseudogenes annotated by GENCODE v.17, only 89 pseudogenes have one or more unmappable transcript isoform (defined as¡50 mappable bases). Only 17 of these occur in the set of 440 that we analyze in the paper, and of this set of 17, only 5 have parent genes.

Figure 5.1C summarizes our pipeline for computing pseudogene expression levels. Our approach improves on the strategy used in (Pei et al., 2012) and (Kalyana-Sundaram et al., 2012). In (Tonner et al., 2012) a method was proposed that, as ours, tries to avoid uniquely misaligned reads and also included a measure of mappability. However, the method developed in (Tonner et al., 2012) applied only to processed

pseudogenes and could not be used for duplicated pseudogenes. Our method also accounts for the possibility of reads that cross splice alignments in defining mappability.

### 5.4.3 Hierarchical clustering and differential expression analysis

Tumor subtype classification was determined using the PAM50 score (Parker et al., 2009). Unsupervised hierarchical clustering was performed using the R function hclust. Expression levels were log transformed and normalized using the R scale function before clustering. We first performed clustering using both tumor and adjacent normal samples. Next, we omitted the adjacent normal samples and clustered only the tumor samples. To determine which pseudogenes showed significant subtype-specific expression, we used the Significance Analysis of Microarrays R package (samr) (Tusher et al., 2001). This approach uses a nonparametric test based on the Kruskal-Wallis statistic to assess the evidence for rejecting the null hypothesis that the expression levels do not differ among subtypes. The multiclass differential expression option of the samr package was used.

### 5.4.4 Prediction of miRNAs targeting pseudogenes and genes

Since pseudogenes are thought to be non-coding and thus not densely bound by ribosomes, the entire transcript can be targeted by miRNAs. Also, since pseudogenes are non-coding, 3' UTRs are not annotated for pseudogenes. However, if a miRNA targets both a pseudogene and its parent gene, the shared miRNA binding site is likely to be located in the 3' UTR of the parent gene and the corresponding "pseudo-3' UTR" of the pseudogene. In order to be more conservative and in an effort to reduce the number of false positives arising from the lack of specificity in miRNA target prediction algorithms, we chose to restrict our analysis to the pseudo-3' UTRs of pseudogenes; we therefore had to annotate these regions. Pseudo-3' UTRs were annotated by BLAST alignment to the 3' UTRs of the parent genes.

For each parent gene-pseudogene pair, we downloaded all annotated 3' UTRs for the parent gene. Next, we extracted the pseudogene locus according to GENCODE and 10 kb of genomic context on either side of the pseudogene. BLAST was then used to align the parent gene 3' UTRs against the pseudogene plus genomic context. The longest statistically significant alignment (based on the BLAST E-value) was taken to be the pseudo-3' UTR. Target prediction was performed on pseudo-3' UTRs and annotated gene 3' UTRs using TargetScan version 7 (Grimson et al., 2007). Only miRNA target seeds from the top 100 expressed miRNAs by average expression level across the samples were used in the target prediction. Isomirs (mature miRNAs

resulting from a shift in the annotated transcription start site of the same miRNA locus) were considered to be different miRNAs in this analysis. A miRNA was considered to be shared between a pseudogene and parent gene if TargetScan predicted that the miRNA could target both of them.

### 5.4.5 Correlation with protein-coding gene and miRNA expression levels

We computed Pearson correlation coefficients on log-transformed gene and pseudogene expression levels using the parent gene annotations from the ENCODE pseudogene decoration resource (psiDR v. 0). To avoid detecting spurious correlations due to predominantly low expression, we required at least 20 samples in which gene and pseudogene are present at 1 RPKUM or greater. Gene-pseudogene pairs with fewer than 20 such samples were omitted from the analysis. We used the miRNA targeting predictions from TargetScan (see "Prediction of miRNAs targeting pseudogenes and genes") to compute correlations between pseudogene and miRNA expression levels. Only the top 100 miRNAs by average expression level were used for this analysis. The pipeline described in Baran-Gale et al. (Baran-Gale et al., 2013) was used to compute miRNA expression levels from the TCGA small RNA-seq data. Correlations with miRNAs were assessed by computing p-values using a T-statistic for the null hypothesis that the correlation is no smaller than 0. False discovery rate correction using the method of Benjamini and Hochberg was performed with the R function p.adjust.

# CHAPTER 6
# Detecting RNA Degradation Intermediates and Untemplated Nucleotide Additions

## 6.1 Introduction

The synthesis, processing, and degradation of RNA are complex processes, with every stage of an RNA's lifetime, from transcription initiation to degradation, requiring careful control. Much attention has been focused on regulation of transcription and pre-mRNA processing, but the detailed pathways of mRNA degradation remain poorly understood. During exonucleolytic degradation of RNA some portions of the molecule are more difficult to degrade than others, resulting in accumulation of intermediates in regions that are degraded more slowly. Eukaryotic mRNAs can be degraded in either 5'-3' or 3'-5' directions, or in some cases in both directions (Mullen and Marzluff, 2008). Critical to understanding the pathway of degradation or modification of the mRNA is a method for determining the precise termini of RNA molecules. Here we describe a method to determine the 3' end of RNA molecules, which can be applied to mapping degradation intermediates generated during 3'-5' degradation. The presence of RNA binding proteins and secondary structure motifs may block the progress of 3'-5' degradation resulting in a spectrum of partly degraded transcripts that differ only at the 3' end (Fig. 6.1A). Additionally, the 3' ends of RNAs are often modified by the addition of short, nontemplated 3' tails, and we are just starting to appreciate the broad range of these modifications (Chang et al., 2014). For example, during degradation of mammalian histone mRNAs, there is oligouridylation of mature mRNA to initiate degradation (Mullen and Marzluff, 2008; Hoefig et al., 2012; Su et al., 2013) as well as uridylation of a large variety of degradation intermediates (Slevin et al., 2014).

Existing methods for studying RNA degradation intermediates or RNAs with nontemplated nucleotides are low-throughput and laborious, requiring cloning of individual degradation intermediates, limiting our ability to probe intermediates in mRNA degradation. Conventional RNA-seq techniques do not yield precise 3' ends of RNA molecules, since the sequences are generated using cDNA priming. As a result the first nucleotides identified are located internal to the 3' end of the molecule. A number of methods for locating alternative polyadenylation sites have been developed (Mayr and Bartel, 2009; Shepard et al., 2011; Lianoglou

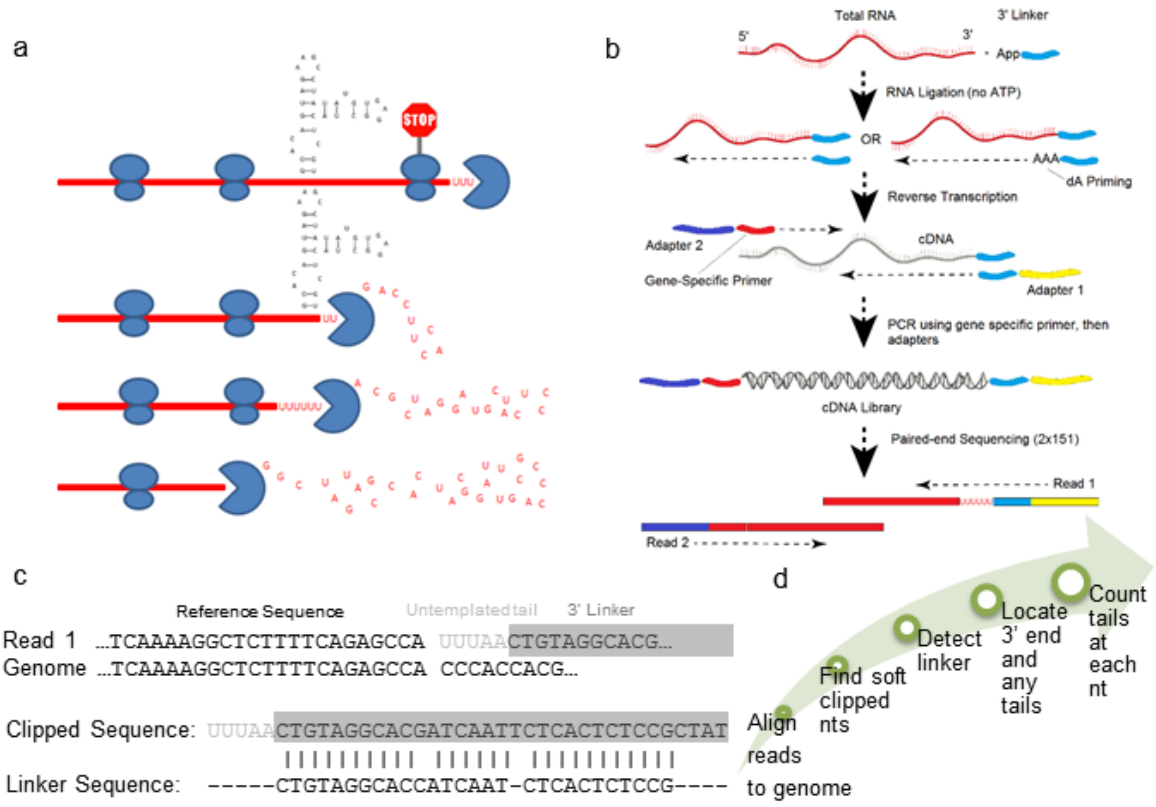



Figure 6.1: EnD-seq and AppEnD Strategy. (A) Schematic of the 3' end of a hypothetical RNA molecule, indicating potential intermediates in 3'-5' degradation resulting from bound proteins or RNA secondary structure that might slow 3'-5' exonuclease degradation. (B) EnD-seq sequencing strategy. (C) Examples of two sequences, one containing an untemplated tail and one containing a single U-tail. (D) Flow chart detailing how AppEnD works.

et al., 2013; Hoque et al., 2014; Masamha et al., 2014), some of which rely on sequencing the junction between the nontemplated poly(A) tail and the cleavage site to identify the precise nucleotide where poly(A) is added (Martin et al., 2012; Hoque et al., 2014; Yao and Shi, 2014).

A common approach to analyzing sequencing data containing nontemplated nucleotides is to strip homopolymers from raw reads before genomic read alignment (Henriques et al. 2013; Yao and Shi 2014). Such a prealignment read stripping approach is less than ideal, making restrictive assumptions about the length and nucleotide composition of the nontemplated additions.

The Marzluff lab developed EnD-Seq (Exonuclease Degradation sequencing; Slevin et al. 2014) and we developed AppEnD (Application for mapping EnD-Seq data), a customized high-throughput sequencing strategy and computational method for identifying 3' ends of RNA molecules, including any nontemplated additions, with no assumptions about sequence composition. Here we demonstrate the utilization of EnD-Seq

and AppEnD to identify nontemplated nucleotides as short as 1 nt, allowing us to define an unanticipated modification of the 3' end of histone mRNA after processing. We also use AppEnD to gain insight into 3' nontemplated additions from diverse types of sequencing data, including small capped RNA sequencing data and PAS-SEQ and A-SEQ polyadenylation data.

## 6.2 Results

The EnD-Seq protocol is designed to identify the 3' end of nonpolyadenylated RNA molecules, including degradation intermediates of polyadenylated mRNAs after deadenylation (Fig. 6.1A). The key to preserving 3' end information is the ligation of a 3' linker. (Note that conventional RNA-seq protocols use random priming, which is not guaranteed to cover the precise 3' end of the molecule.) The linker is then used to prime cDNA synthesis, generating cDNAs that contain the junction between the linker and the 3' end of the transcript. EnD-seq generates paired-end reads in which read 1 contains the linker and 3' end information, and read 2 is an upstream read used to aid in aligning read 1. The EnD-seq data that we analyzed was generated using PCR primers that specifically targeted the human histone mRNAs. Figure 6.1 summarizes EnD-seq and AppEnD.

To obtain information about the position of the transcript end and any nontemplated tails, we examined the first read which begins with the linker sequence, followed by a nontemplated tail or the 3' end with no tail. AppEnD aligns the paired-end reads to the genome using an RNA-seq aligner, e.g., bowtie2 (Langmead and Salzberg, 2012) for unspliced RNAs or MapSplice (Wang et al., 2010) for spliced RNAs. Since read 1 contains the linker sequence and any nontemplated 3 additions, the end of the read sequence that diverges from the genome is soft-clipped. We identify the linker sequence within the soft-clipped portion of the read using the NeedlemanWunsch algorithm (Needleman and Wunsch, 1970). Any nontemplated 3' additions are identified as nucleotides after the end of the linker in the soft-clipped portion of the read (Fig. 6.1C). After identifying the 3' ends at single nucleotide resolution, we plot the abundance of transcripts ending at each nucleotide. This gives the positional distribution of the last templated nucleotides and the pattern of nontemplated additions, if any are present.

### 6.2.1 Human histone mRNAs have modified 3' ends containing untemplated uridines

Histones are the proteins that make up nucleosomes, the "wire spools" around which DNA is wound. In addition to this essential role, histone genes are noteworthy in that their RNA transcripts are unspliced and

lack a poly(A) tail, ending instead in a stem-loop secondary structure (Marzluff et al., 2008). The Marzluff lab discovered several years ago that the histone transcripts have another unusual property: U tails get added to them during the process of degradation (Mullen and Marzluff, 2008).

Applying EnD-seq and AppEnD to the study of histone mRNAs in human cells revealed several interesting results about the positions, identities, and relative amounts of 3' ends and untemplated tails. Steady-state histone transcripts not actively undergoing degradation end mostly at the normal genomic coordinate, but surprisingly many of the transcripts end in 1 or 2 nucleotide untemplated tails (Fig. 6.2A). The really intriguing thing about these tails is that their pattern of addition is such that the transcripts with the additions still end at exactly the normal position (3 nts after the end of the stem loop; Fig. 6.2B-D). This suggests that the tails may serve as a "repair mechanism" to restore the histone transcripts to their normal length after they have been nibbled back. The tails are almost exclusively U nucleotides (Fig. 6.2E), and there are essentially no tails longer than 2 nucleotides before degradation begins.

After the cellular signal to degrade histone transcripts is given, longer tails (but still almost all Us) show up. The long tails accumulate at two main positions: 2-4 nts inside the stem loop and just 3' of the stop codon (Fig. 6.2F,H). Interestingly, these two locations correspond precisely to positions where proteins are known to be present. The histone stem loop is bound by a protein when not undergoing degradation, and the stop codon is often occupied by a ribosome during active translation. The position of the highest peak in the tail addition distribution occurs 15 nucleotides downstream of the stop codon, precisely the width of the ribosome. These results suggest that the addition of long U tails may be a way of "re-priming" the degradation machinery when it stalls out after hitting a protein bound to the transcript. In support of this hypothesis, the area in between the stem loop and the ribosome footprint is almost completely free of 3' ends and tails during degradation, suggesting that degradation occurs processively after the stem loop binding protein is removed and before the degradation machinery runs into the ribosome.

Because the long U tails represent a relatively small proportion of the observed 3' ends, the Marzluff lab devised a strategy to enrich for long U tails. Using a modified primer ending in 3 A nucleotides (the reverse complement of 3 Us) gives reads that are almost exclusively from U tails 3 nucleotides in length or longer. EnD-seq data generated in this fashion showed very similar patterns to the previous data but much deeper coverage (Fig. 6.1G-I). When analyzing the data generated from this modified protocol, one must be careful to distinguish between untemplated U tails and mispriming events caused by genomically

encoded U nucleotides (Fig. 6.2J-K). AppEnD automatically handles this issue, detecting mispriming events as untemplated tails 0-2 nucleotides in length, which are then rejected.

The 3A-primer also allowed us to analyze two histone mRNAs, HIST1H3H and HIST1H2AB, that were present at low abundance. The HIST1H3H gene had low coverage using the standard protocol due to its moderately low expression, but we were able to confirm that the pattern of coverage is the same between the two versions of the protocol (Fig. 6.3A-D). The HIST1H2AB gene is expressed at levels too low to allow the sensitive detection of tails using the standard EnD-seq protocol. The 3A-primed data allowed us to look at the pattern of tail addition on HIST1H2AB, confirming that the pattern is consistent with what we observed on the other histone genes that we investigated (Fig. 6.3E-F). We did obtain a small number of sequences that are clearly artifacts with this approach due to internal priming at U-rich sequences, including at UGU or UCU sequences (Fig. 6.3F), which results in apparent 2U tails (Fig. 6.3G). When we primed with the 3A-primer we used a cut-off of nontemplated tails 3 nt or greater in AppEnd. Thus, these artifactual "tails" were easily identified by their 1 or 2-nt length and removed by AppEnD. As expected, the abundant one or 2-nt tails at the end of the histone mRNA were not detected using the 3A primer.

### 6.2.2 Fly histone mRNAs also have untemplated additions

To investigate whether histone mRNA metabolism is conserved between human and fly, we performed EnD-Seq on histone mRNAs from Drosophila embryos and ovaries. The overall distribution of 3' end locations was similar to what we observed in human cells, indicating that the histone 3' end processing is conserved. Fig. 6.4A-B and Fig. 6.4D-E show the patterns for fly H2a and H3 mRNAs in the ovary and embryo, respectively. The main features of the human histone mRNAs can be observed here: most molecules are full length, with 1-2 nucleotide tails serving to "repair" transcripts that have been nibbled by a nucleotide or two. Also, the dominant degradation intermediates occur within the stem loop. Surprisingly, the untemplated nucleotides added in the ovary are almost exclusively As (Fig. 6.4C), whereas the tails in the embryo are almost all Us (Fig. 6.4F).

### 6.2.3 Mapping short capped RNAs using AppEnD

Although the EnD-seq experiments described here used gene-specific primers to specifically target histone genes, the method can readily be extended to allow a genome-wide analysis of RNAs with nontemplated nucleotides at the 3' end. In addition, AppEnD is also useful for detecting nontemplated tails in other types

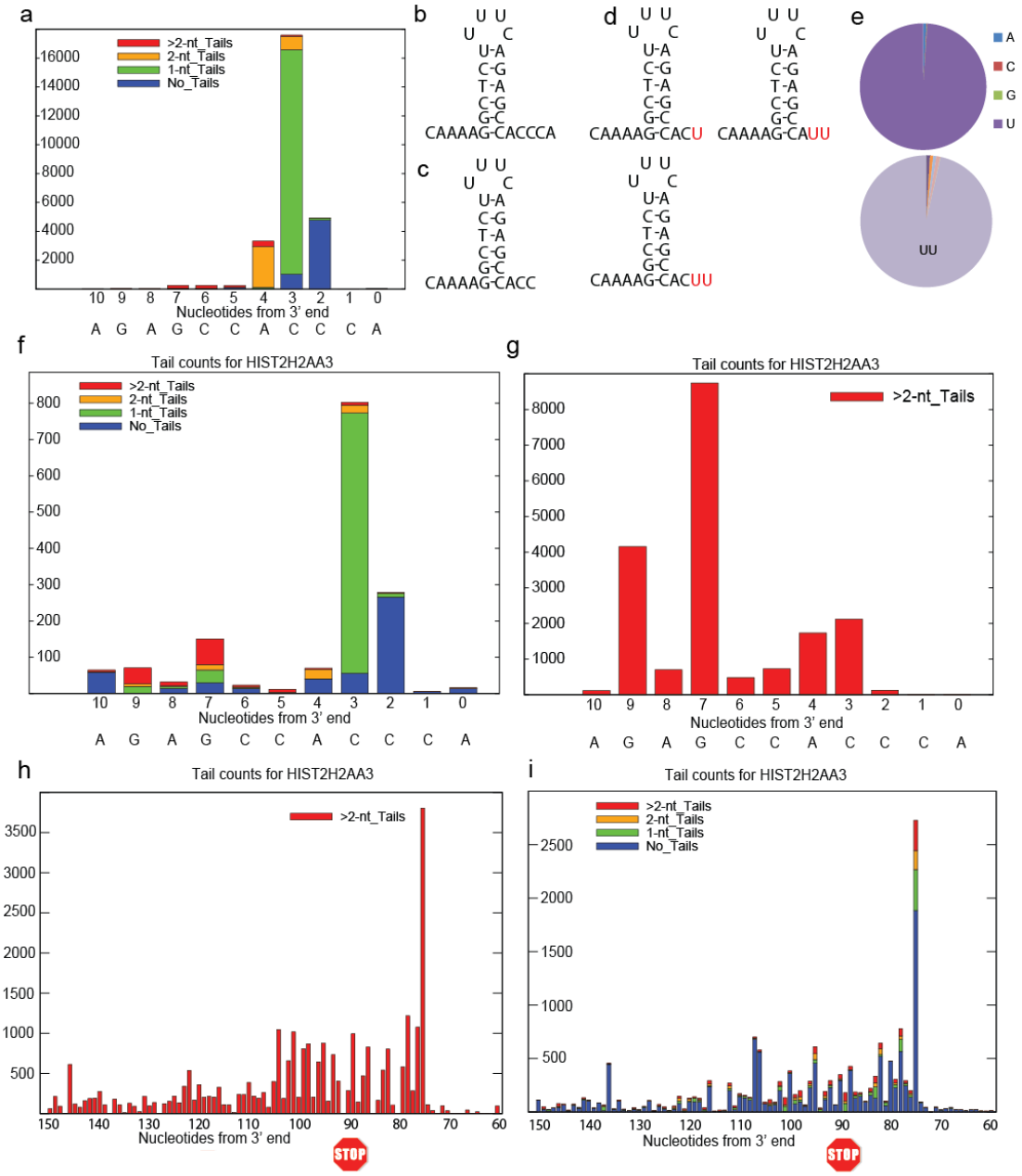



Figure 6.2: Using Standard or Oligo(dA) Priming to Detect Histone 3' Ends. (A) Graph of position and length of 3' untemplated additions observed on HIST2H2AA3 gene (blue indicates no tail). (B)-(D) Unprocessed, normal, and repaired histone 3' ends. (e) Pie charts showing the nucleotide compositions of one- and two-nucleotide tails. (F) Position and length of HIST2H2AA3 untemplated additions after degradation has begun. (G) Position and length of HIST2H2AA3 untemplated additions after degradation has begun, as determined by EnD-seq with a modified primer containing 3 As. (H) Position and length of HIST2H2AA3 untemplated additions after degradation has begun (internal portion of the gene). (I) Position and length of HIST2H2AA3 untemplated additions after degradation has begun (internal portion of the gene; modified primer containing 3 As).

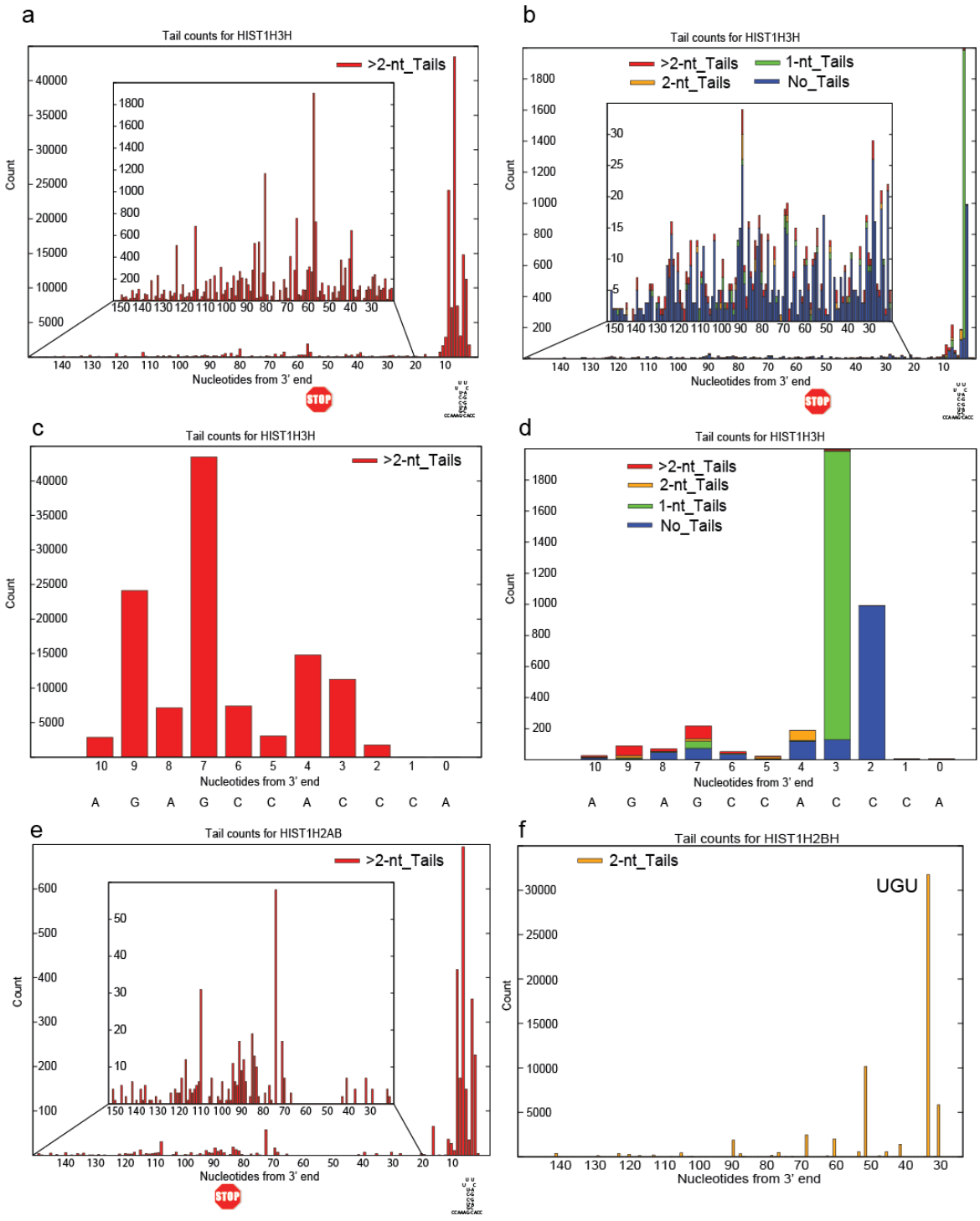



Figure 6.3: Priming with 3 As Enhances Detection of U Tails. (A) Position and length of HIST1H3H untemplated additions after degradation has begun (dA primed). (B) Position and length of HIST1H3H untemplated additions after degradation has begun (no dA priming). (C) Position and length of HIST1H3H untemplated additions after degradation has begun (dA primed; 3' end only). (D) Position and length of HIST1H3H untemplated additions after degradation has begun (3' end only). (E) HIST1H2AB untemplated additons (dA primed). (F) Mispriming event due to the presence of UGU in the genome sequence.

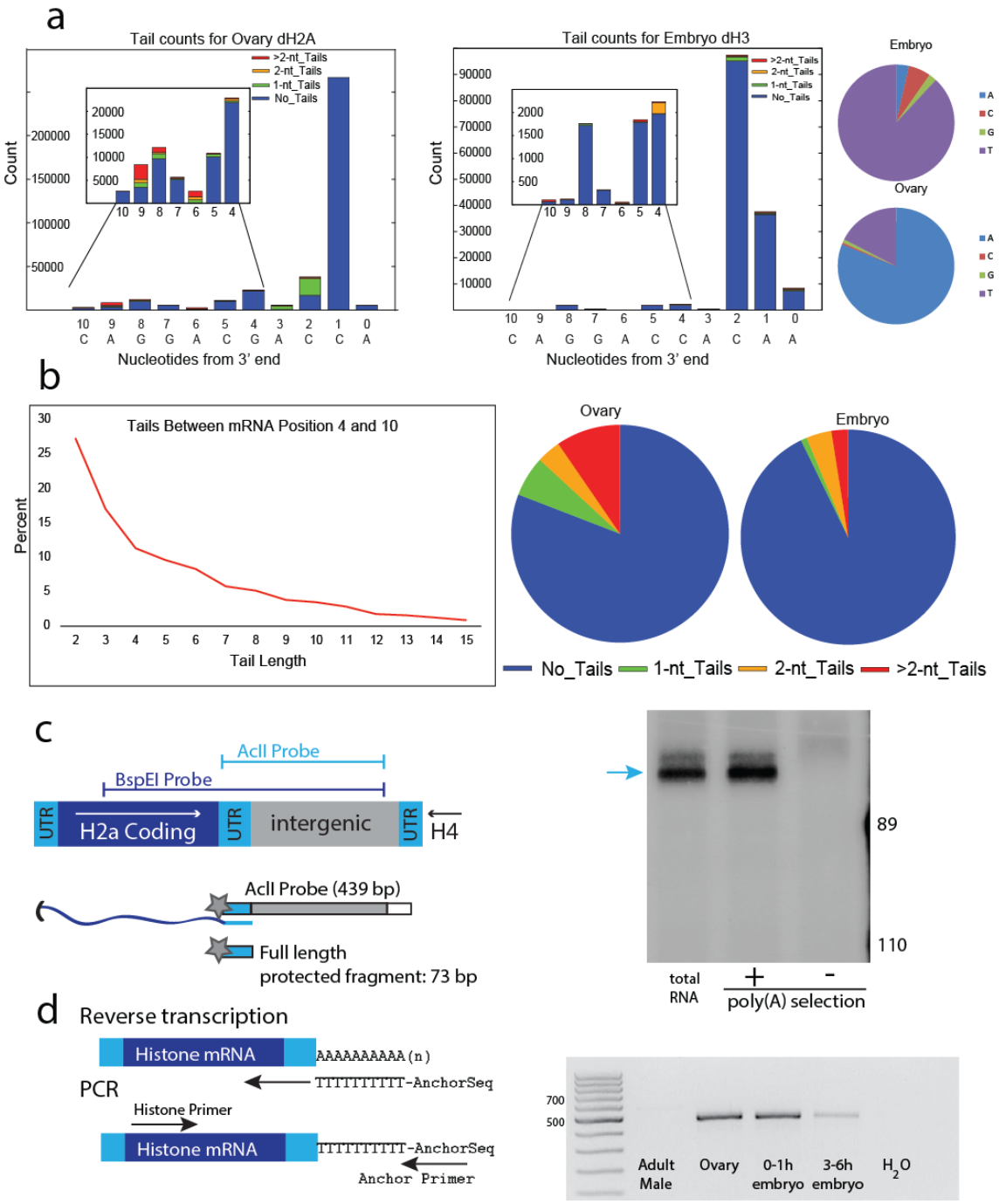



Figure 6.4: Analysis of Drosophila Histone mRNAs. (A) Untemplated tail counts for dH2a gene in fly ovary. (B) Untemplated tail counts for dH3 gene in fly ovary. (C) Untemplated tail compositions in fly embryo and ovary. (D) Untemplated tail counts for dH2a gene in fly embryo. (E) Untemplated tail counts for dH3 gene in fly embryo. (F) Tail length composition in fly ovary and embryo.

of sequencing data. A common approach to this problem is to strip homopolymers from raw reads before genomic read alignment (Henriques et al. 2013; Yao and Shi 2014). Such a prealignment read stripping approach has the shortcoming that it cannot distinguish between genomically encoded and nontemplated nucleotides. Such an approach also relies on knowing the sequence composition of the pattern to trim from the reads; consequently, the pattern must be sufficiently long that it is distinctive and must be known in advance.

In contrast to prealignment read stripping, AppEnD detects untemplated tails by direct comparison with the reference genome during the read alignment process. This strategy provides three advantages: (1) It allows the detection of nontemplated additions without any assumptions about tail composition; (2) it detects tails as short as one nucleotide, and (3) it more effectively distinguishes nontemplated homopolymers from repeated genomic bases by direct comparison with the genome.

To demonstrate the capability of AppEnD to map genomewide data, we mapped a data set of short capped RNAs from cultured Drosophila cells resulting from stalling of RNA polymerase II immediately after initiation produced by Adelman and coworkers (Henriques et al., 2013). These RNAs were sequenced from the 3' end. The initial study found that knocking down the RNA degradation machinery produced a large increase in the number of small RNAs with short untemplated tails. AppEnD successfully mapped these reads to the Drosophila genome and confirmed a sixfold to 10-fold increase in the number of oligo(A) tails after exosome knockdown (Fig. 6.5A). The Illumina primer was ligated directly onto the RNA, but the primer-transcript junction does not occur in the reads, descreasing the confidence with which we can detect untemplated additions. Thus, we could only unambiguously detect tails 3 nucleotides in length or longer (Fig. 6.5F). The fly histone genes are present as tandemly repeated units (Lifton et al., 1978), and thus reads from fly histone genes are commonly discarded due to multimapping issues. We created a custom Drosophila genome containing a single histone repeat (McKay et al., 2015), with which we were able to map the short capped RNAs expressed from the Drosophila histone genes (Fig. 6.5B). The tails in the samples with degradation machinery knocked down were almost exclusively oligo(A) tails with occasional substitutions of a U or C, ranging up to 15 nt in length (Fig. 6.5C). We show the results for the tails that mapped to two Drosophila histone genes, histone H2B and histone H4. Since there are 100 copies of each histone gene and large amounts of histone mRNA produced in a growing cell, we obtained a large number of short capped RNA reads. For all five histone genes, the major pause site was $\sim 40$ nt from the transcription start site (Fig. 6.5C, and data not shown). The paused tailed RNAs remaining after inhibition of transcription are generally

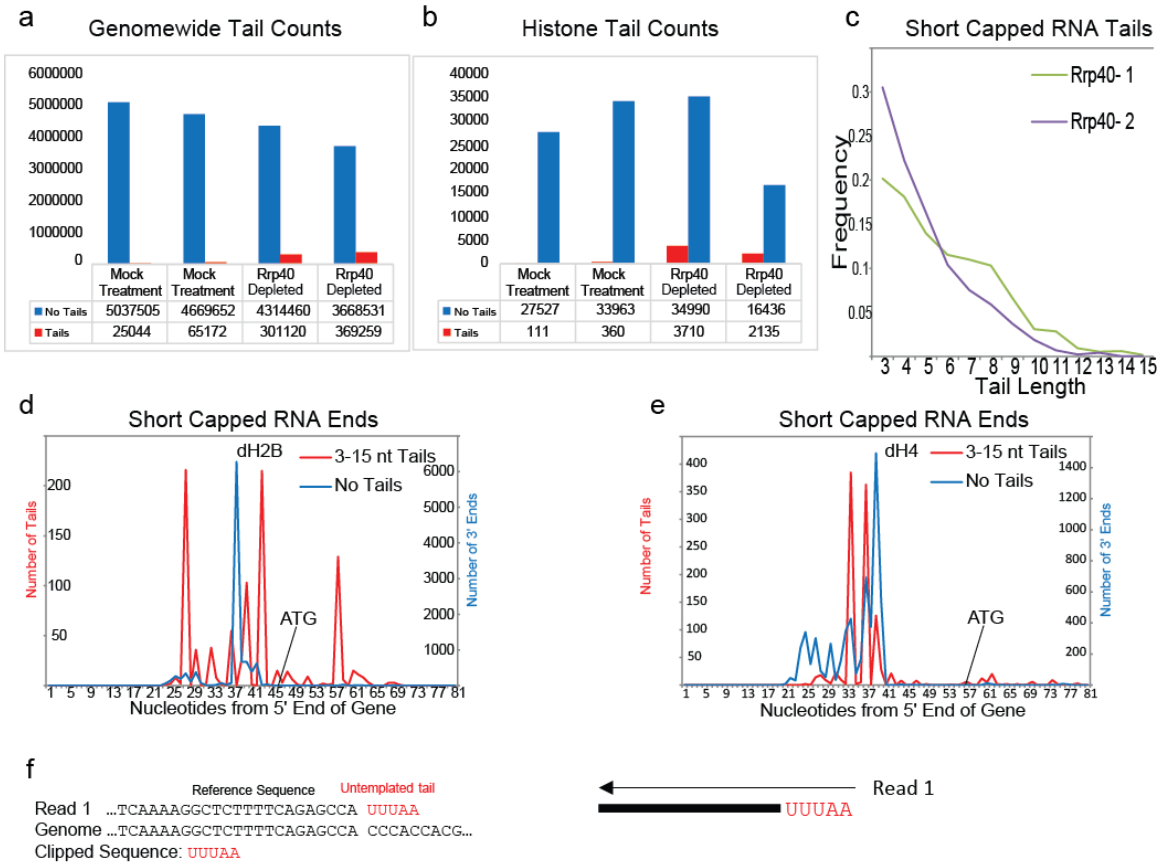



Figure 6.5: AppEnD Analysis of Nontemplated Tails on Short Capped Transcripts. (A) Number of untemplated tails detected in control and exosome knockdown samples. (B) Number of untemplated tails detected on histone genes in control and exosome knockdown samples. (C) Tail length distributions in knockdown samples. (D)-(G) Positional distributions of no tails and 3-15 nucleotide tails on 4 fly histone genes. (H)-(I) Sample reads showing how tails are detected from (H) short capped RNA-seq vs. (I) EnD-seq.

not at the major pause site (Fig. 6.5D-E). This suggests that when the polymerase falls off before or after reaching the major pause site, untemplated nucleotide addition may somehow specifically target these shorter or longer RNAs for degradation.

### 6.2.4 Mapping alternative polyadenylation data using AppEnD

Similarly to detecting untemplated additions with EnD-seq, detecting alternative polyadenylation requires observing the junction between the templated transcript sequence and the untemplated poly(A) tail. Two polyadenylation sequencing methods, PAS-Seq (Shepard et al., 2011) and A-Seq (Yao and Shi, 2014), both prime cDNA synthesis with a oligo(dT) primer ending in random dinucleotides. Ideally, cDNA priming occurs at the junction between poly(A) tail and templated sequence. However, mispriming events can occur

within the poly(A) sequence or within the templated portion of the transcript. Thus, a challenge of the data analysis for these two methods is distinguishing mispriming at encoded (A) stretches from genuine poly(A) tails. AppEnD automatically detects internally primed A-tails obtained from the PAS-Seq or A-Seq methods, keeping only the true poly(A) tails.

We applied AppEnD to two PAS-Seq data sets generated by sequencing from the 3' UTR, through the oligo(A) tail from the primer, into the linker sequence (Fig. 6.6C). The results from two genes are shown, one, EBAG9, whose distribution of polyadenylation sites changed between the control and experimental conditions (Fig. 6.6A), and the other NET1, where the polyadenylation pattern remained constant (Fig. 6.6B). Figure 6.6C shows an example of a PAS-Seq read containing a poly(A) site. We also applied AppEnD to two A-Seq data sets from normal cells and cells with a polyadenylation factor knocked down, which were generated by sequencing from the 3' UTR into the anchor primer. Examples of genes showing a change (Fig. 6.6D) and no change (Fig. 6.6E) between A-Seq experimental conditions are shown in Figure 6.6. Figure 6.6F,G show examples of a mispriming event and a true poly(A) site found from our analysis of the A-Seq data. We found that only 16% of the A-Seq reads contained authentic polyadenylation sites, 77% were misprimed, and 6% of the reads were uninformative since they did not get to the poly(A) tail. These numbers underscore the importance of filtering mispriming events, which make AppEnD useful for analyzing these types of data.

The AppEnD method is applicable to any deep sequencing data set where the 3' ends are sequenced. This includes small RNA data sets, such as miRNAs and pre-miRNAs (Newman et al., 2011), or the capped paused transcripts made by Pol II (Henriques et al., 2013). The data can be mapped genome-wide and does not require knowledge of the nontemplated nucleotides on the pre-miRNAs or miRNAs. One constraint is that for accurate mapping of 1- or 2-nt nontemplated tails, the data have to be generated using an anchor primer on the 3' end to serve as the sequence that primes the cDNA, and the primer-RNA junction needs to be observed directly. If that is not the case, we found we could not reliably map nontemplated nucleotides of $< 3$ nt due to random heterogeneity at the end of many of the sequence reads. AppEnD is particularly applicable to the study of alternative polyadenylation, if the method used directly determines the sequence of the 3' end and the ligated primer. In such a case, AppEnD readily removes artifactual sequences resulting from internal priming at A-rich sequences. In conclusion, EnD-Seq provides a platform for determining the 3' end of RNA molecules together with any nontemplated nucleotides added to the transcript in a completely unbiased way, regardless of the length or composition of the nontemplated region. There are many potential applications

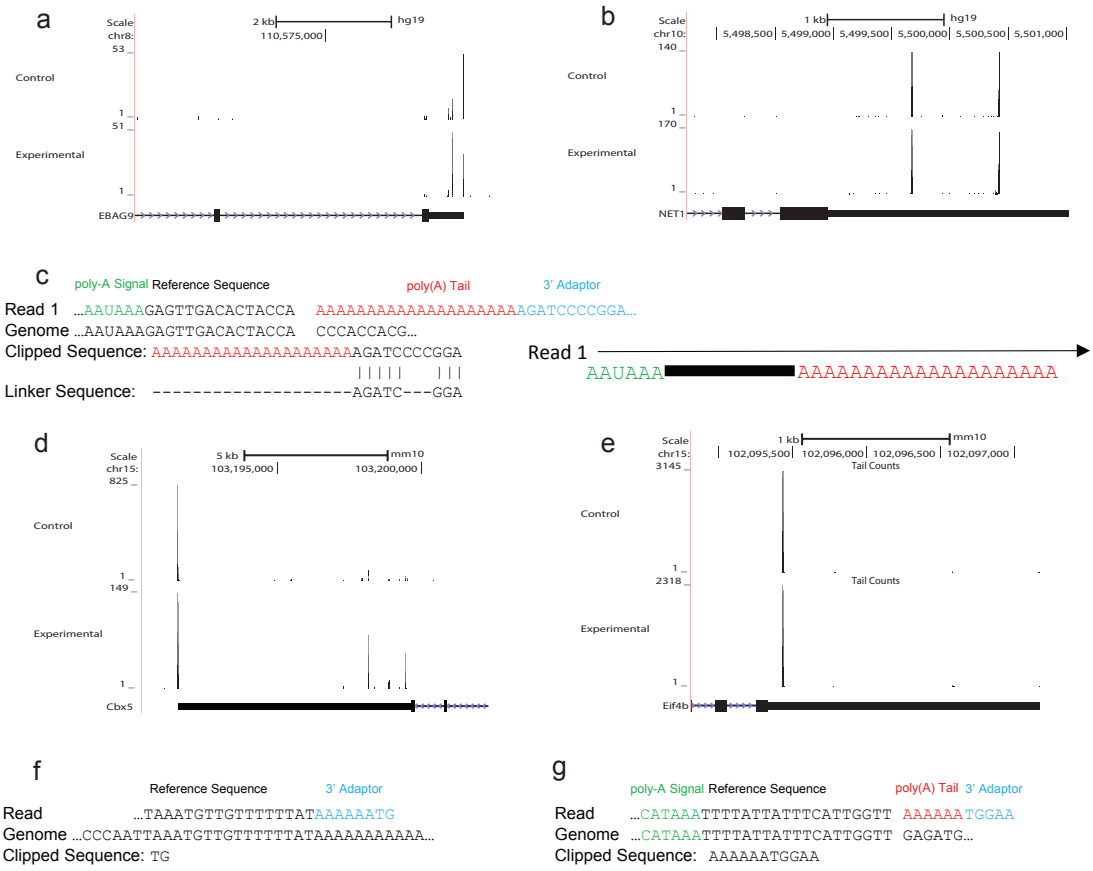



Figure 6.6: Mapping Alternative Polyadenylation Data with AppEnD. (A) Gene that shows alternative polyadenylation between control and experimental treatments, as detected by running AppEnD on PAS-Seq data. (B) Gene that does not show alternative polyadenylation between control and experimental treatments, as determined by running AppEnD on PAS-Seq data. (C) Example showing how AppEnD detects untemplated tail from PAS-seq data. (D) Gene that shows alternative polyadenylation between control and experimental treatments, as detected by running AppEnD on A-Seq data. (E) Gene that does not show alternative polyadenylation between control and experimental treatments, as determined by running AppEnD on A-Seq data. (F)-(G) Examples showing a mispriming event (F) and true poly(A) tail (G) as they appear in A-Seq data.

of this platform for identifying novel cleavages and modifications of the 3' ends of RNA molecules and for determining the details of RNA degradation proceeding in the 3'-5' direction.

## 6.3 Methods

### 6.3.1 Mapping EnD-seq data

We used bowtie2 in local alignment mode (Langmead and Salzberg, 2012) with default settings to map reads to either hg19 or dm3. A custom sequence including one copy of the histone repeat with the 5 histone genes (H1, H2A, H2B, H3, and H4) was added to the dm3 index, since the histone genes are not present in the dm3 assembly. Local alignment mode maximizes the alignment score of the whole read and will computationally remove ("soft clip") portions of the beginning or end of a read that does not match the genome. We use this feature to detect the portion of EnD-Seq reads containing the 3 ends of transcripts, including any nontemplated additions. Although spliced aligners are usually used for RNA-seq data, the histone genes are not generally spliced, so we chose to use bowtie2. A spliced aligner that performs soft clipping, such as Mapsplice (Wang et al., 2010) or Star (Dobin et al., 2013), could also be used.

Our EnD-seq sequencing strategy produces paired-end reads, although this is not essential, since sufficient information is present in the read that contains the 3' end. Read 1 contains the reverse complement of the ligated linker followed by the reverse complement of any nontemplated additions, then the genomic portion of the transcript. Read 2 provides additional genomic context to aid in aligning read 1 but does not generally contain 3' end information. We thus look for read 1 sequences whose alignments begin with a soft-clipped portion. To account for possible sequencing errors, we detect the linker within this soft-clipped portion by performing dynamic programming alignment to the known linker sequence using the NeedlemanWunsch algorithm. The remainder of the soft clipped portion of the read beyond the end of the linker as detected by this alignment represents a nontemplated addition. The end of the linker also indicates the precise position of the 3' end of the RNA molecule being sequenced and thus provides important information that aids in the computational identification of 3' nontemplated tails, and allows us to accurately determine nontemplated tails as short as 1 nt. Since the linker is at the beginning of read 1, this part of the read is generally of high quality. This represents a distinct advantage of EnD-seq over other sequencing strategies that either lack a 3' linker or sequence it at the end of the read.

### 6.3.2 Mapping short capped RNAs and polyadenylation sites

We used AppEnD to map short capped RNAs (Henriques et al., 2013), PAS-Seq, and A-Seq data. This demonstrates the usefulness of the method for mapping other types of data than just EnD-seq. The protocol used to sequence short capped RNAs in this case produced single-end reads starting with the 3' end of the RNA (Fig. 6.5H), since an Illumina linker was ligated onto the 3' end of the RNA. These data sets were a single direction read from the 3' end of the RNA. Unlike EnD-seq data, there is no linker present on the end of these short capped RNA reads, making it more difficult to distinguish short nontemplated additions from read errors. We therefore restricted analysis to nontemplated tails that were homopolymers at least 3 nt in length. We could not have reliably assigned reads that had shorter number of nontemplated bases or that had a mixed composition. This is in contrast to our ability to assign any nontemplated read regardless of length or composition with our EnD-Seq protocol.

The PAS-Seq protocol utilizes an anchored 20-nt dT primer ending in two random nucleotides to generate the cDNA, while the A-Seq strategy is similar but contains a 6-nt dT primer followed by a stemloop and an additional 14 dTs. Short cDNA fragments were sequenced from the 5' end of the mRNA producing a single end read with up to 20 nontemplated A's (PAS-Seq) followed by the complement of the anchor on the dT primer (Fig. 6.6C) or six nontemplated A's (A-Seq) followed by the sequencing adapter. In PAS-Seq, because the reads end with the sequencing adapter, the adapter sequence is generally of low quality, since it follows a long stretch of repeated A's. Nevertheless, the presence of the adapter in the reads provides useful information that indicate how many nucleotides of the poly(A) tail were nontemplated, which helps distinguish authentic poly(A) tails from mispriming events. One of the challenges in analyzing PAS-Seq or A-Seq data is detecting false-positive polyadenylation sites due to mispriming events that can occur when the PAS-Seq primer anneals to stretches of repeated genomic A's. We detected such false positives by requiring that the 5 nt immediately preceding the soft-clipped portion of the read were not all A's. This shows the clear advantage of our method compared with a commonly used strategy in which reads are stripped of repeated A's before alignment to the genome; to such a strategy, mispriming events appear the same as true positive poly(A) sites, and must be identified in a separate computational step. However, by locating the precise position at which a read stops matching the genome, we are able to effectively detect misprimed reads.

# CHAPTER 7
# Conclusion and Future Directions

The problem space that I have explored during this dissertation is incredibly rich and rapidly expanding. During the coming years, researchers will likely continue to make significant progress toward addressing a host of important biomedical problems. In my opinion, developments must continue in two key areas for this rapid advance to continue: (1) technologies and experimental protocols for high-throughput measurement of cellular properties and (2) computational approaches for analyzing, exploring, and modeling biomedical data. These two fronts will likely become increasingly intertwined, as computational approaches inspire new experiments and experimental developments drive fundamental computer science research. In this chapter, I describe directions for future work that arise naturally from the work described in this dissertation. I conclude with predictions about the future direction of the broader field that encompasses my work.

## 7.1 Extensions to SingleSplice

In its current implementation, SingleSplice depends critically on the presence of spike-in transcripts added in equal amounts to each cell in a dataset. The ability to use SingleSplice without spike-in transcripts would be very useful, because many experimental datasets do not contain spike-ins. One way of addressing this limitation would be to assume that most endogenous genes do not show biological variation across the set of single cells, and use the set of all genes to fit a technical noise model.

Another possible extension is to develop a way of detecting alternative 3' end usage, which is a second way (besides alternative splicing) that multiple transcripts are generated from a single gene. A number of single cell sequencing protocols prime reverse transcription using the poly(A) tail, then produce sequencing reads close to or precisely at the cleavage/polyadenylation site. Many of these protocols also incorporate unique molecular identifiers (UMIs), which allow the precise counting of the number of starting molecules that produced the observed sequencing reads (Kivioja et al., 2011; Islam et al., 2013). UMIs can be used to remove the effects of amplification bias and precisely determine the number of molecules per cell without

spike-ins, providing a different way of accounting for technical noise (Grün et al., 2014). However, sequencing data with UMIs does not allow detection of general alternative splicing changes because reads come only from the ends of the transcripts. Nevertheless, they may serve as a method to robustly detect alternative polyadenylation from single cell data, which would be very useful.

Existing approaches to modeling technical noise in single cell RNA-seq data focus mainly on modeling variance as a function of expression level. However, it is very likely that other sources of bias contribute to technical noise, such as the GC content of transcripts. Additionally, it is known that the ERCC transcripts most commonly used as spike-ins do not closely resemble human (or even eukaryotic) transcripts (Svensson et al., 2017). Therefore, an interesting future direction is to develop a model that takes into account all observed sources of bias, perhaps using a more realistic set of spike-ins such as the SIRV transcripts from Lexogen (Svensson et al., 2017).

Another promising direction of research is to increase sensitivity of single cell alternative splicing detection by pooling information from related cells. Because there is a large amount of stochasticity in which transcripts are missed in any given cell, a group of very similar cells should give a much more complete picture of the transcriptome than any individual cell. Such an approach would need to determine the best way to identify similar cells, how many cells should be pooled, and how to prevent any individual cell from exerting too much influence over the pooled result, among other things. Nevertheless, this seems like a promising direction to explore, and increased sensitivity would greatly improve the utility of SingleSplice.

## 7.2 Extensions to SLICER

As the number of single cell datasets from cells undergoing sequential processes of change increases, there will be a need for increasingly general cell trajectory models. For example, SLICER is not designed for a dataset in which there are multiple distinct trajectories. Multiple trajectories could arise from the presence of multiple cells types each undergoing the same process, multiple cell types undergoing different processes, or a single cell type undergoing multiple processes simultaneously. Each of these scenarios would require different solutions. Another interesting extension would involve relaxing the assumption that a given process proceeds through exactly one defined sequence of gene expression changes. It seems highly likely that there will rather be some stochasticity in the relative ordering of events during many biological processes, so that

gene A may sometimes be turned off before gene B and sometimes after gene B. Additional work is required to investigate such possibilities.

Although SLICER can detect branches and loops in biological processes, it relies on such structure being appropriately preserved during the dimensionality reduction process. It seems that a strategy that detects important topological features directly in the high-dimensional space may yield increased sensitivity compared to a general purpose dimensionality reduction approach like LLE. A related issue is how to determine the statistical significance of a branch or loop feature, particularly a small-scale feature that could arise from either noise or a rare cell population. More generally, the theoretical framework of computational topology may prove helpful for characterizing the salient features of the manifolds underlying biological processes. Persistent homology and the Morse theory of functions defined on manifolds seem especially relevant for the cell trajectory problem.

## 7.3 Extensions to MATCHER

MATCHER assumes that the dominant source of variation underlying multiple types of measurements is a continuous, one-dimensional, non-branching, non-cyclic sequential process. Additional work is required to relax these assumptions and model cyclic processes, branching processes, datasets containing discrete clusters, and higher-dimensional manifolds. Additionally, it will be interesting to extend the method to include other types of single cell measurements, such as protein expression from mass cytometry (Bendall et al., 2011) or single cell Hi-C. Single cell Hi-C data will be especially interesting to integrate with other single cell measurements, because Hi-C data measures pairwise information and is thus somewhat different from the 1D measurements that we analyzed. Finally, a key question that biologists want to answer concerns the relative ordering of transcriptional and epigenetic changes: Which happens first? Thus, determining the relative ordering of transcriptional and epigenetic events from single cell multi-omic data is an important future direction.

## 7.4 Next Steps for Pseudogenes, 3' Ends, and Post-Transcriptional Regulation in General

There is considerable utility in taking the computational methods we developed for pseudogene and 3' end studies and applying them to data generated from additional biological contexts. However, it seems that, at least for the moment, the primary limitation on this front is experimental in nature. New experimental

approaches for single molecule and single cell sequencing promise to give new insights into pseudogene transcription, untemplated nucleotide addition, and post-transcriptional regulation in general. As these experimental approaches become more widespread, there will be a need for new computational methods to analyze the new types of data.

Single molecule sequencers like the Pacific Biosciences Sequel instrument and the Oxford Nanopore MinIon generate long reads, allowing sequencing of entire transcripts. Currently transcripts are defined by the possible combinations of alternative splicing, alternative start sites and alternative polyadenylation. It is likely that many of these events are correlated (e.g. specific start sites may lead to specific splicing patterns and/or specific polyadenylation sites). Hints that this is the case are already in the literature. Currently there is no way to determine the structure of individual tranacripts, and long sequence reads (or single-molecule droplet sequencing) is a way that this information could be obtained. These long reads could also be used to more effectively disambiguate gene and pseudogene expression, due to the additional sequence context. If applied to the study of 3' end modifications, single molecule sequencing could link changes at the 3' end to other aspects of post-transcriptional regulation, such as splicing, capping, or 5'-3' degradation. The low throughput, high cost, and high error rate of single molecule sequencing are current barriers to such studies, but the situation will likely improve over the next few years.

Measuring pseudogene expression and 3' end modification at single cell resolution will also enable new insights into post-transcriptional regulation. As with single molecule measurements, there are significant experimental challenges involved in performing such measurements on single cells. But the ability to correlate, within individual cells, various post-transcriptional regulation mechanisms and other cellular quantities will greatly aid studies of gene regulation. For example, long noncoding RNAs often regulate epigenetic modification in *cis*. If we could correlate epigenetic marks and lncRNA expression within individual cells, we would undoubtedly shed new light on the roles of many lncRNAs.

## 7.5 The Future

One of the most exciting developments on the horizon is the human cell atlas project. As part of the ongoing efforts to understand the human genome for the purpose of enabling genomic medicine, researchers worldwide are now proposing an ambitious goal: to characterize every major cell type in the human body (Regev, 2016). Initial estimates indicate that 50-100 million cell measurements would be sufficient to

accomplish this goal (Regev, 2016). With the development of techniques such as Drop-Seq (Macosko et al., 2015) and inDrop (Klein et al., 2015) for simultaneously measuring gene expression in tens of thousands of single cells per experiment, such a project is now feasible. Indeed, hundreds of individual research groups worldwide have already begun performing high-throughput measurements of single cell gene expression using Drop-Seq and inDrop on their tissue of choice. For example, Macosko et al. used Drop-Seq to identify 39 cell populations in the retina (Macosko et al., 2015). Meanwhile, funding agencies, including NIH and the Chan Zuckerberg Foundation, are launching programs to fund comprehensive single cell characterization efforts. Thus, it seems that a comprehensive collection of single cell profiles from human tissues is not far away.

The existence of a human cell atlas would open many exciting new avenues of research. Such a resource would provide an unprecedented opportunity to learn the "gene expression code", linking genes and their expression patterns to specific cellular properties. A recent review paper referred to this idea as "revealing the vectors of cellular identity" (Wagner et al., 2016). One could even think of learning the gene expression manifold for the entire human body, characterizing the islands of cellular gene expression profiles in the vast, multidimensional sea of possible gene expression combinations. The computational difficulties involved in such a project will be enormous, and will require innovations at the frontier of computer science.

Knowing the gene expression profiles of healthy human cell types would also enable tremendous progress in understanding the genetic basis of disease. For example, single cell gene expression profiles can be used to identify the cell type–or the even dynamic cell state of a cell type–in which a deleterious mutation causes harm.

In addition to increasing number of single cells, I anticipate that future efforts, including the human cell atlas project, will focus on measuring more and more properties of single cells. There is currently great interest in retaining information about the spatial context of cells, and spatial information will likely be crucial to understanding cellular gene expression profiles. Another promising avenue of research measuring phenotypic and functional properties of individual cells, and pairing these measurements with other assays such as RNA-seq. Increasing integration of single cell measurements, by either experimentally measured or computationally inferred multi-omic profiles, is another direction in which I expect considerable progress in the near future. More information about cellular properties will increasingly allow computational biologists to harness the well-developed techniques of supervised machine learning.

Of course, a human cell atlas is not a panacea, and many questions will remain after successful completion of this undertaking. Even if we successfully measure every type of cell in the human body, we will still miss much of the variation in dynamic cellular states–during development, in response to stimuli, in disease conditions, and among individuals with different genotypes. An even bigger question is how the gene expression profiles of individual cells in complex tissues influence each other. Spatial contacts, cell-cell junctions, and intercellular signaling all play crucial roles in building tissues from cells. Finally, it is worth noting that measuring cellular quantities through sequencing is a very imperfect science even when performed in bulk on millions of cells, and the challenges (both experimental and computational) only get harder when one moves to the single cell level. Ingenious methods for measuring cellular properties in bulk and analyzing the resulting data are continuously being devised and improved.

A crucial goal beyond simply understanding the genomic control system is designing perturbations that move a cell from one point in gene expression space to another. As we start to understand the gene expression code, the focus will likely begin to shift from purely descriptive analysis to predictive modeling (Tanay and Regev, 2017). Researchers have already begun developing computational methods for predicting how to transdifferentiate any cell type into any other cell type (Rackham et al., 2016). Such predictive models promise to generate breakthroughs in regenerative medicine and the treatment of cancer and other genetic diseases.

Ultimately, we may never fully understand the genomic control system. But it's hard to imagine a more exciting task than studying the blueprint that directs the intricate unfolding of human life.

# BIBLIOGRAPHY


Amador-Ortiz, C., Roma, A. A., Huettner, P. C., Becker, N., and Pfeifer, J. D. (2011). JAZF1 and JJAZ1 gene fusion in primary extrauterine endometrial stromal sarcoma. *Human Pathology*, 42(7):939–946.

American Heart Association (2015). North Carolina Fact Sheet. Technical report.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S. A., Ponting, C. P., Voet, T., Kelsey, G., Stegle, O., and Reik, W. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, 13(3):229–232.

Balasubramanian, V. N., Ye, J., and Panchanathan, S. (2007). Biased Manifold Embedding: A Framework for Person-Independent Head Pose Estimation. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE.

Baran-Gale, J., Fannin, E. E., Kurtz, C. L., and Sethupathy, P. (2013). Beta Cell 5-Shifted isomiRs Are Candidate Regulatory Hubs in Type 2 Diabetes. *PLoS ONE*, 8(9):e73240.

Basu, A., Wilkinson, F. H., Colavita, K., Fennelly, C., and Atchison, M. L. (2014). YY1 DNA binding and interaction with YAF2 is essential for Polycomb recruitment. *Nucleic acids research*, 42(4):2208–23.

Bayro-Corrochano, E. and Eklundh, J.-O., editors (2009). *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volume 5856 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Belkin, M. and Niyogi, P. (2003). Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6):1373–1396.

Bendall, S. C., Davis, K. L., Amir, E.-A. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Shenfeld, D. K., Nolan, G. P., and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, 157(3):714–25.

Bendall, S. C., Simonds, E. F., Qiu, P., Amir, E.-a. D., Krutzik, P. O., Finck, R., Bruggner, R. V., Melamed, R., Trejo, A., Ornatsky, O. I., Balderas, R. S., Plevritis, S. K., Sachs, K., Pe'er, D., Tanner, S. D., and Nolan, G. P. (2011). Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science*, 332(6030).

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., and Lander, E. S. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, 125(2):315–326.

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., and Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature methods*, 10(11):1093–5.

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–160.

Campbell, K. R. and Yau, C. (2016). Order Under Uncertainty: Robust Differential Expression Analysis Using Probabilistic Models for Pseudotime Inference. *PLOS Computational Biology*, 12(11):e1005212.

Cao, R. and Zhang, Y. (2004). SUZ12 Is Required for Both the Histone Methyltransferase Activity and the Silencing Function of the EED-EZH2 Complex. *Molecular Cell*, 15(1):57–67.

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. (2011). A Long Noncoding RNA Controls Muscle Differentiation by Functioning as a Competing Endogenous RNA. *Cell*, 147(2):358–369.

Chan, W.-L., Yuo, C.-Y., Yang, W.-K., Hung, S.-Y., Chang, Y.-S., Chiu, C.-C., Yeh, K.-T., Huang, H.-D., and Chang, J.-G. (2013). Transcribed pseudogene PPM1K generates endogenous siRNA to suppress oncogenic cell growth in hepatocellular carcinoma. *Nucleic Acids Research*, 41(6):3734–3747.

Chang, H., Lim, J., Ha, M., Kim, V., Anders, S., Benes, V., Steinmetz, L., Brown, C., Bassell, G., Richter, J., and et Al. (2014). TAIL-seq: Genome-wide Determination of Poly(A) Tail Length and 3 End Modifications. *Molecular Cell*, 53(6):1044–1052.

Chang Wang and Sridhar Mahadevan (2009). A General Framework for Manifold Alignment. In *AAAI*.

Cheow, L. F., Courtois, E. T., Tan, Y., Viswanathan, R., Xing, Q., Tan, R. Z., Tan, D. S. W., Robson, P., Loh, Y.-H., Quake, S. R., and Burkholder, W. F. (2016). Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nature Methods*, 13(10):833–836.

Chiefari, E., Iiritano, S., Paonessa, F., Le Pera, I., Arcidiacono, B., Filocamo, M., Foti, D., Liebhaber, S. A., and Brunetti, A. (2010). Pseudogene-mediated posttranscriptional silencing of HMGA1 can result in insulin resistance and type 2 diabetes. *Nature Communications*, 1(4):1–7.

Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30.

Cooper, G. (2000). *The Cell: A Molecular Approach*. Sinauer Associates, Sunderland (MA), 2 edition.

Corces, M. R., Buenrostro, J. D., Wu, B., Greenside, P. G., Chan, S. M., Koenig, J. L., Snyder, M. P., Pritchard, J. K., Kundaje, A., Greenleaf, W. J., Majeti, R., and Chang, H. Y. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature genetics*, 48(10):1193–203.

Costa, J., Girotra, A., and Hero, A. (2005). Estimating Local Intrinsic Dimension with k-Nearest Neighbor Graphs. In *IEEE/SP 13th Workshop on Statistical Signal Processing, 2005*, pages 417–422. IEEE.

Damianou, A., Ek, C., Titsias, M., and Lawrence, N. (2012). Manifold Relevance Determination. *ArXiv*.

Damianou, A. C., Titsias, M. K., and Lawrence, N. D. (2016). Variational Inference for Latent Variables and Uncertain Inputs in Gaussian Processes. *Journal of Machine Learning Research*, 17:1–62.

Darmanis, S., Gallant, C. J., Marinescu, V. D., Niklasson, M., Segerman, A., Flamourakis, G., Fredriksson, S., Assarsson, E., Lundberg, M., Nelander, S., Westermark, B., and Landegren, U. (2016). Simultaneous Multiplexed Measurement of RNA and Proteins in Single Cells. *Cell Reports*, 14(2):380–389.

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., Hayden Gephart, M. G., Barres, B. A., and Quake, S. R. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences of the United States of America*, 112(23):7285–90.

Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, 14(6):390–403.

Deng, C., Li, Y., Liang, S., Cui, K., Salz, T., Yang, H., Tang, Z., Gallagher, P., Qiu, Y., Roeder, R., Zhao, K., Bungert, J., and Huang, S. (2013). USF1 and hSET1A Mediated Epigenetic Modifications Regulate Lineage Differentiation and HoxB4 Transcription. *PLoS Genetics*, 9(6):e1003524.

Derrien, T., Estellé, J., Marco Sola, S., Knowles, D. G., Raineri, E., Guigó, R., and Ribeca, P. (2012). Fast Computation and Applications of Genome Mappability. *PLoS ONE*, 7(1):e30377.

Desai, T. J., Brownfield, D. G., and Krasnow, M. A. (2014). Alveolar progenitor and stem cells in lung development, renewal and cancer. *Nature*, 507(7491):190–4.

Dey, S. S., Kester, L., Spanjaard, B., Bienko, M., and van Oudenaarden, A. (2015). Integrated genome and transcriptome sequencing of the same cell. *Nature Biotechnology*, 33(3):285–289.

Dietrich, N., Lerdrup, M., Landt, E., Agrawal-Singh, S., Bak, M., Tommerup, N., Rappsilber, J., Södersten, E., and Hansen, K. (2012). REST-mediated recruitment of polycomb repressor complexes in mammalian cells. *PLoS genetics*, 8(3):e1002494.

Dijkstra, E. W. (1959). A Note on Two Problems in Connexion with Graphs. *NUMERISCHE MATHEMATIK*, 1(1):269–271.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21.

Donoho, D. L. and Grimes, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596.

Ebert, M. S., Neilson, J. R., and Sharp, P. A. (2007). MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nature Methods*, 4(9):721–726.

Ebert, M. S. and Sharp, P. A. (2010). Emerging Roles for Natural MicroRNA Sponges. *Current Biology*, 20(19):R858–R861.

Edelsbrunner, H., Kirkpatrick, D., and Seidel, R. (1983). On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4):551–559.

Ek, C. H. (2009). *Shared Gaussian Process Latent Variables Models*. PhD thesis, Oxford Brookes University.

Eleftheriadis, S., Rudovic, O., and Pantic, M. (2015). Discriminative Shared Gaussian Processes for Multiview and View-Invariant Facial Expression Recognition. *IEEE Transactions on Image Processing*, 24(1):189–204.

Fan, H. C., Fu, G. K., and Fodor, S. P. A. (2015). Combinatorial labeling of single cells for gene expression cytometry. *Science*, 347(6222):1258367–1258367.

Farlik, M., Sheffield, N. C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., and Bock, C. (2015). Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell reports*, 10(8):1386–97.

Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature Reviews Genetics*, 2008(2):102–114.

Franco-Zorrilla, J. M., Valli, A., Todesco, M., Mateos, I., Puga, M. I., Rubio-Somoza, I., Leyva, A., Weigel, D., García, J. A., and Paz-Ares, J. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genetics*, 39(8):1033–1037.

Fu, Y., Xu, Y., and Huang, T. S. (2007). Estimating Human Age by Manifold Analysis of Face Pictures and Regression on Aging Features. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 1383–1386. IEEE.

Fujii, G. H., Morimoto, A. M., Berson, A. E., and Bolen, J. B. (1999). Transcriptional analysis of the PTEN/MMAC1 pseudogene, ΨPTEN. *Oncogene*, 18(9):1765–1769.

Gach, P. C., Wang, Y., Phillips, C., Sims, C. E., and Allbritton, N. L. (2011). Isolation and manipulation of living adherent cells by micromolded magnetic rafts. *Biomicrofluidics*, 5(3):32002–3200212.

Gazina, E., Richards, K., Mokhtar, M., Thomas, E., Reid, C., and Petrou, S. (2010). Differential expression of exon 5 splice variants of sodium channel $\alpha$ subunit mRNAs in the developing mouse brain. *Neuroscience*, 166(1):195–200.

Genshaft, A. S., Li, S., Gallant, C. J., Darmanis, S., Prakadan, S. M., Ziegler, C. G. K., Lundberg, M., Fredriksson, S., Hong, J., Regev, A., Livak, K. J., Landegren, U., and Shalek, A. K. (2016). Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. *Genome Biology*, 17(1):188.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29(7):644–52.

Green, E. D. and Guyer, M. S. (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature*, 470.

Grimson, A., Farh, K. K.-H., Johnston, W. K., Garrett-Engele, P., Lim, L. P., and Bartel, D. P. (2007). MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. *Molecular Cell*, 27(1):91–105.

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nature methods*, 11(6):637–40.

Gu, S., Jin, L., Zhang, F., Sarnow, P., and Kay, M. A. (2009). Biological basis for restriction of microRNA targets to the 3 untranslated region in mammalian mRNAs. *Nature Structural & Molecular Biology*, 16(2):144–150.

Guenzi, E., Töpolt, K., Cornali, E., Lubeseder-Martellato, C., Jörg, A., Matzen, K., Zietz, C., Kremmer, E., Nappi, F., Schwemmle, M., Hohenadl, C., Barillari, G., Tschachler, E., Monini, P., Ensoli, B., and Stürzl,

M. (2001). The helical domain of GBP-1 mediates the inhibition of endothelial cell proliferation by inflammatory cytokines. *The EMBO Journal*, 20(20):5568–5577.

Guo, G., Huss, M., Tong, G. Q., Wang, C., Li Sun, L., Clarke, N. D., and Robson, P. (2010). Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst. *Developmental Cell*, 18(4):675–685.

Ham, J., Lee, D., and Saul, L. (2003). Learning high dimensional correspondences from low dimensional manifolds. *Workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, (August 2003):34–41.

Ham, J., Lee, D. D., and Saul, L. K. (2005). Semisupervised alignment of manifolds. In *AISTATS*, pages 120–127.

Han, L., Yuan, Y., Zheng, S., Yang, Y., Li, J., Edgerton, M. E., Diao, L., Xu, Y., Verhaak, R. G. W., and Liang, H. (2014). The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nature Communications*, 5:R90.

Hanchate, N. K., Kondoh, K., Lu, Z., Kuang, D., Ye, X., Qiu, X., Pachter, L., Trapnell, C., and Buck, L. B. (2015). Single-cell transcriptomics reveals receptor transformations during olfactory neurogenesis. *Science*, 350(6265):1251–1255.

Hayashi, H., Arao, T., Togashi, Y., Kato, H., Fujita, Y., De Velasco, M. A., Kimura, H., Matsumoto, K., Tanaka, K., Okamoto, I., Ito, A., Yamada, Y., Nakagawa, K., and Nishio, K. (2015). The OCT4 pseudogene POU5F1B is amplified and promotes an aggressive phenotype in gastric cancer. *Oncogene*, 34(2):199–208.

Henriques, T., Gilchrist, D., Nechaev, S., Bern, M., Muse, G., Burkholder, A., Fargo, D., Adelman, K., Serrano, L., Meziane, O., and et Al. (2013). Stable Pausing by RNA Polymerase II Provides an Opportunity to Target and Integrate Regulatory Signals. *Molecular Cell*, 52(4):517–528.

Hermitte, S. and Chazaud, C. (2014). Primitive endoderm differentiation: from specification to epithelium formation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1657):20130537–20130537.

Hershey, A. D. and Chase, M. (1952). INDEPENDENT FUNCTIONS OF VIRAL PROTEIN AND NUCLEIC ACID IN GROWTH OF BACTERIOPHAGE. *The Journal of General Physiology*, 36(1).

Hitomi, J., Katayama, T., Eguchi, Y., Kudo, T., Taniguchi, M., Koyama, Y., Manabe, T., Yamagishi, S., Bando, Y., Imaizumi, K., Tsujimoto, Y., and Tohyama, M. (2004). Involvement of caspase-4 in endoplasmic reticulum stress-induced apoptosis and A$\beta$-induced cell death. *The Journal of Cell Biology*, 165(3):347–356.

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D., Niu, B., McLellan, M. D., Uzunangelov, V., Zhang, J., Kandoth, C., Akbani, R., Shen, H., Omberg, L., Chu, A., Margolin, A. A., van't Veer, L. J., Lopez-Bigas, N., Laird, P. W., Raphael, B. J., Ding, L., Robertson, A. G., Byers, L. A., Mills, G. B., Weinstein, J. N., Van Waes, C., Chen, Z., Collisson, E. A., Benz, C. C., Perou, C. M., Stuart, J. M., and Stuart, J. M. (2014). Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*, 158(4):929–944.

Hoefig, K. P., Rath, N., Heinz, G. A., Wolf, C., Dameris, J., Schepers, A., Kremmer, E., Ansel, K. M., and Heissmeyer, V. (2012). Eri1 degrades the stem-loop of oligouridylated histone mRNAs to induce replication-dependent decay. *Nature Structural & Molecular Biology*, 20(1):73–81.

Hoque, M., Li, W., and Tian, B. (2014). Accurate Mapping of Cleavage and Polyadenylation Sites by 3 Region Extraction and Deep Sequencing. In *Methods in molecular biology (Clifton, N.J.)*, volume 1125, pages 119–129.

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., Huang, Y., and Peng, J. (2016). Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Research*, 26(3):304–319.

Hu, Y., Huang, Y., Du, Y., Orellana, C. F., Singh, D., Johnson, A. R., Monroy, A., Kuan, P.-F., Hammond, S. M., Makowski, L., Randell, S. H., Chiang, D. Y., Hayes, D. N., Jones, C., Liu, Y., Prins, J. F., and Liu, J. (2013). DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic acids research*, 41(2):e39.

Ieda, M., Fu, J.-D., Delgado-Olguin, P., Vedantham, V., Hayashi, Y., Bruneau, B. G., and Srivastava, D. (2010). Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell*, 142(3):375–86.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2013). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166.

Jacq, C., Miller, J. R., and Brownlee, G. G. (1977). A Pseudogene Structure in 5S DNA of Xenopus laevis. *Cell*, 12(0):109–120.

Jaenisch, R. and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33(3s):245–254.

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome research*, 21(9):1543–51.

Jin, W., Tang, Q., Wan, M., Cui, K., Zhang, Y., Ren, G., Ni, B., Sklar, J., Przytycka, T. M., Childs, R., Levens, D., and Zhao, K. (2015). Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature*, 528(7580):142–6.

Jørgensen, H. F., Terry, A., Beretta, C., Pereira, C. F., Leleu, M., Chen, Z.-F., Kelly, C., Merkenschlager, M., and Fisher, A. G. (2009). REST selectively represses a subset of RE1-containing neuronal genes in mouse embryonic stem cells. *Development (Cambridge, England)*, 136(5):715–21.

Kalyana-Sundaram, S., Kumar-Sinha, C., Shankar, S., Robinson, D., Wu, Y.-M., Cao, X., Asangani, I., Kothari, V., Prensner, J., Lonigro, R., Iyer, M., Barrette, T., Shanmugam, A., Dhanasekaran, S., Palanisamy, N., and Chinnaiyan, A. (2012). Expressed Pseudogenes in the Transcriptional Landscape of Human Cancers. *Cell*, 149(7):1622–1634.

Karreth, F., Tay, Y., Perna, D., Ala, U., Tan, S., Rust, A., DeNicola, G., Webster, K., Weiss, D., Perez-Mancera, P., Krauthammer, M., Halaban, R., Provero, P., Adams, D., Tuveson, D., and Pandolfi, P. (2011). InVivo Identification of Tumor- Suppressive PTEN ceRNAs in an Oncogenic BRAF-Induced Mouse Model of Melanoma. *Cell*, 147(2):382–395.

Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods*, 7(12):1009–15.

Kégl, B. (2002). Intrinsic Dimension Estimation Using Packing Numbers. In *NIPS*.

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–2.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36.

Kim, J. H., Paek, K. Y., Choi, K., Kim, T.-D., Hahm, B., Kim, K.-T., and Jang, S. K. (2003). Heterogeneous nuclear ribonucleoprotein C modulates translation of c-myc mRNA in a cell cycle phase-dependent manner. *Molecular and cellular biology*, 23(2):708–720.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2011). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9(1):72–74.

Klein, A., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D., and Kirschner, M. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*, 161(5):1187–1201.

Kolodziejczyk, A., Kim, J. K., Svensson, V., Marioni, J., and Teichmann, S. (2015a). The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell*, 58(4):610–620.

Kolodziejczyk, A., Kim, J. K., Tsang, J., Ilicic, T., Henriksson, J., Natarajan, K., Tuck, A., Gao, X., Bühler, M., Liu, P., Marioni, J., and Teichmann, S. (2015b). Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell*, 17(4):471–485.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural {&} molecular biology*, 17(7):909–915.

Korneev, S. A., Park, J. H., and O'Shea, M. (1999). Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 19(18):7711–20.

Kumarakulasingham, M., Rooney, P. H., Dundas, S. R., Telfer, C., Melvin, W. T., Curran, S., and Murray, G. I. (2005). Cytochrome P450 Profile of Colorectal Cancer: Identification of Markers of Prognosis. *Clinical Cancer Research*, 11(10):3758–3765.

Lackey, P. E., Welch, J. D., and Marzluff, W. F. (2016). TUT7 catalyzes the uridylation of the 3' end for rapid degradation of histone mRNA. *RNA (New York, N.Y.)*.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson,

D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A., and Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature, Published online: 15 February 2001; — doi:10.1038/35057062*, 409(6822):860.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9.

Lawrence, M., Daujat, S., and Schneider, R. (2016). Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends in Genetics*, 32(1):42–56.

Lawrence, N. D. (2004). Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data. In *Advances in Neural Information Processing Systems 16*, pages 329–336.

Lee, J. T. (2012). Epigenetic Regulation by Long Noncoding RNAs.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12:323.

Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S., and Mayr, C. (2013). Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes & Development*, 27(21):2380–2396.

Lifton, R. P., Goldberg, M. L., Karp, R. W., and Hogness, D. S. (1978). The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications. *Cold Spring Harbor symposia on quantitative biology*, 42 Pt 2:1047–51.

Llorens-Bobadilla, E., Zhao, S., Baser, A., Saiz-Castro, G., Zwadlo, K., and Martin-Villalba, A. (2015). Single-Cell Transcriptomics Reveals a Population of Dormant Neural Stem Cells that Become Activated upon Brain Injury. *Cell Stem Cell*, 17(3):329–40.

Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M., Smith, M., Van der Aa, N., Banerjee, R., Ellis, P. D., Quail, M. A., Swerdlow, H. P., Zernicka-Goetz, M., Livesey, F. J., Ponting, C. P., and Voet, T. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, 12(6):519–522.

Macaulay, I. C., Svensson, V., Labalette, C., Ferreira, L., Hamey, F., Voet, T., Teichmann, S. A., and Cvejic, A. (2016). Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells. *Cell reports*, 14(4):966–977.

Macosko, E., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A., Kamitaki, N., Martersteck, E., Trombetta, J., Weitz, D., Sanes, J., Shalek, A., Regev, A., and McCarroll, S. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214.

Margueron, R. and Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. *Nature*, 469(7330):343–349.

Martin, G., Gruber, A., Keller, W., and Zavolan, M. (2012). Genome-wide Analysis of Pre-mRNA 3 End Processing Reveals a Decisive Role of Human Cleavage Factor I in the Regulation of 3 UTR Length. *Cell Reports*, 1(6):753–763.

Marzluff, W. F., Wagner, E. J., and Duronio, R. J. (2008). Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nature Reviews Genetics*, 9(11):843–854.

Masamha, C. P., Xia, Z., Yang, J., Albrecht, T. R., Li, M., Shyu, A.-B., Li, W., and Wagner, E. J. (2014). CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature*, 510(7505):412–6.

Mayr, C. and Bartel, D. P. (2009). Widespread Shortening of 3UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell*, 138(4):673–684.

McKay, D., Klusza, S., Penke, T., Meers, M., Curry, K., McDaniel, S., Malek, P., Cooper, S., Tatomer, D., Lieb, J., Strahl, B., Duronio, R., and Matera, A. (2015). Interrogating the Function of Metazoan Histones using Engineered Gene Clusters. *Developmental Cell*, 32(3):373–386.

Mei, D., Song, H., Wang, K., Lou, Y., Sun, W., Liu, Z., Ding, X., and Guo, J. (2013). Up-regulation of SUMO1 pseudogene 3 (SUMO1P3) in gastric cancer and its clinical association. *Medical Oncology*, 30(4):709.

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A. J., Tanaka, Y., Wilkinson, A. C., Buettner, F., Macaulay, I. C., Jawaid, W., Diamanti, E., Nishikawa, S.-I., Piterman, N., Kouskoff, V., Theis, F. J., Fisher, J., and Göttgens, B. (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature Biotechnology*, 33(3):269–276.

Mooijman, D., Dey, S. S., Boisset, J.-C., Crosetto, N., and van Oudenaarden, A. (2016). Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nature Biotechnology*, 34(8):852–856.

Mullen, T. E. and Marzluff, W. F. (2008). Degradation of histone mRNA requires oligouridylation followed by decapping and simultaneous degradation of the mRNA both 5' to 3' and 3' to 5'. *Genes & development*, 22(1):50–65.

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–53.

Newman, M. A., Mani, V., and Hammond, S. M. (2011). Deep sequencing of microRNA precursors reveals extensive 3' end modification. *RNA*, 17(10):1795–1803.

Nilsen, T. W. and Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–63.

Ocbina, P. J., Dizon, M. L. V., Shin, L., and Szele, F. G. (2006). Doublecortin is necessary for the migration of adult subventricular zone cells from neurospheres. *Molecular and cellular neurosciences*, 33(2):126–35.

Olga Kouropteva, Oleg Okun, M. P. (2002). Selection of the Optimal Parameter Value for the Locally Linear Embedding Algorithm. In *1st International Conference on Fuzzy Systems and*, pages 359—-363.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., and Bernard, P. S. (2009). Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167.

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., Louis, D. N., Rozenblatt-Rosen, O., Suvà, M. L., Regev, A., and Bernstein, B. E. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (New York, N.Y.)*, 344(6190):1396–401.

Pei, B., Sisu, C., Frankish, A., Howald, C., Habegger, L., Mu, X., Harte, R., Balasubramanian, S., Tanzer, A., Diekhans, M., Reymond, A., Hubbard, T. J., Harrow, J., and Gerstein, M. B. (2012). The GENCODE pseudogene resource. *Genome Biology*, 13(9):R51.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, Ø., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A.-L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3):290–295.

Piñol-Roma, S. and Dreyfuss, G. (1993). Cell cycle-regulated phosphorylation of the pre-mRNA-binding (heterogeneous nuclear ribonucleoprotein) C proteins. *Molecular and cellular biology*, 13(9):5762–5770.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W. J., and Pandolfi, P. P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, 465.

Polo, J. M., Anderssen, E., Walsh, R. M., Schwarz, B. A., Nefzger, C. M., Lim, S. M., Borkent, M., Apostolou, E., Alaei, S., Cloutier, J., Bar-Nur, O., Cheloufi, S., Stadtfeld, M., Figueroa, M. E., Robinton, D., Natesan, S., Melnick, A., Zhu, J., Ramaswamy, S., and Hochedlinger, K. (2012). A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell*, 151(7):1617–32.

Prat, A., Karginova, O., Parker, J. S., Fan, C., He, X., Bixby, L., Harrell, J. C., Roman, E., Adamo, B., Troester, M., and Perou, C. M. (2013). Characterization of cell lines derived from breast cancers and normal mammary tissues for the study of the intrinsic molecular subtypes. *Breast Cancer Research and Treatment*, 142(2):237–255.

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., He, X., and Perou, C. M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research*, 12(5):R68.

Prat, A., Perou, C. M., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., He, X., Perou, C. M., Abate-Shen, C., Shen, M., Brown, M., Viens, P., Xerri, L., Bertucci, F., Stassi, G., Dontu, G., Birnbaum, D., Wicha, M., McManus, R., Scherneck, S., Ponder, B., Ford, D., Peto, J., Stoppa-Lyonnet, D., Easton, D., Perou, C., and Mills, G. (2009). Mammary development meets cancer genomics. *Nature Medicine*, 15(8):842–844.

Qian, L., Huang, Y., Spencer, C. I., Foley, A., Vedantham, V., Liu, L., Conway, S. J., Fu, J.-d., and Srivastava, D. (2012). In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes. *Nature*, 485(7400):593–8.

Qing, Y., Yingmao, G., Lujun, B., and Shaoling, L. (2008). Role of Npm1 in proliferation, apoptosis and differentiation of neural stem cells. *Journal of the neurological sciences*, 266(1-2):131–7.

Rackham, O. J. L., Firas, J., Fang, H., Oates, M. E., Holmes, M. L., Knaupp, A. S., Suzuki, H., Nefzger, C. M., Daub, C. O., Shin, J. W., Petretto, E., Forrest, A. R. R., Hayashizaki, Y., Polo, J. M., and Gough, J. (2016). A predictive computational framework for direct reprogramming between human cell types. *Nature Genetics*, 48(3):331–335.

Rasmussen, C. E., Williams, C. K. I., Sutton, R. S., Barto, A. G., Spirtes, P., Glymour, C., Scheines, R., Schölkopf, B., and Smola, A. J. (2006). *Gaussian Processes for Machine Learning*. MIT Press MIT Press.

Regev, A. (2016). The Human Cell Atlas. Technical report.

Reid, J. E. and Wernisch, L. (2015). Pseudotime estimation: deconfounding single cell time series. Technical report.

Ritchie, W., Flamant, S., and Rasko, J. E. J. (2009). Predicting microRNA targets and functions: traps for the unwary. *Nature Methods*, 6(6):397–398.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., and Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome biology*, 12(3):R22.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K., She, R., Varhol, R., Kamoh, B., Prabhu, A.-L., Tam, A., Zhao, Y., Moore, R. A., Hirst, M., Marra, M. A., Jones, S. J. M., Hoodless, P. A., and Birol, I. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, 7(11).

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., and Bernstein, B. E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, 33(11):1165–1172.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science (New York, N.Y.)*, 290(5500):2323–6.

Saliba, A.-E., Westermann, A. J., Gorski, S. A., and Vogel, J. (2014). Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Research*, pages gku555–.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., Pandolfi, P. P., Qiao, Y., Chen, N., Sun, F., Fan, Q., Lapointe, E., et Al., Program, N. C. S., Center, B. C. o. M. H. G. S., Center, W. U. G. S., Institute, B., Institute, C. H. O. R., and et Al. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, 146(3):353–8.

Sandberg, R. (2013). Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods*, 11(1):22–24.

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., Trombetta, J. J., Gennert, D., Gnirke, A., Goren, A., Hacohen, N., Levin, J. Z., Park, H., and Regev, A. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–40.

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N., Schwartz, S., Fowler, B., Weaver, S., Wang, J., Wang, X., Ding, R., Raychowdhury, R., Friedman, N., Hacohen, N., Park, H., May, A. P., and Regev, A. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 509(7505):363–9.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature reviews. Genetics*, 14(9):618–30.

Shen, S., Park, J. W., Lu, Z.-x., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences of the United States of America*, 111(51):E5593–601.

Shepard, P. J., Choi, E.-A., Lu, J., Flanagan, L. A., Hertel, K. J., and Shi, Y. (2011). Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*, 17(4):761–772.

Singer, Z. S., Yong, J., Tischler, J., Hackett, J. A., Altinok, A., Surani, M. A., Cai, L., and Elowitz, M. B. (2014). Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Molecular cell*, 55(2):319–31.

Singh, D., Orellana, C. F., Hu, Y., Jones, C. D., Liu, Y., Chiang, D. Y., Liu, J., and Prins, J. F. (2011). FDM: a graph-based statistical method to detect differential transcription using RNA-seq data. *Bioinformatics (Oxford, England)*, 27(19):2633–40.

Slevin, M. K., Meaux, S., Welch, J. D., Bigler, R., Miliani de Marval, P. L., Su, W., Rhoads, R. E., Prins, J. F., and Marzluff, W. F. (2014). Deep sequencing shows multiple oligouridylations are required for 3' to 5' degradation of histone mRNAs on polyribosomes. *Molecular cell*, 53(6):1020–30.

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, 11(8):817–820.

Smith, E. R., Cayrou, C., Huang, R., Lane, W. S., Cote, J., and Lucchesi, J. C. (2005). A Human Protein Complex Homologous to the Drosophila MSL Complex Is Responsible for the Majority of Histone H4 Acetylation at Lysine 16. *Molecular and Cellular Biology*, 25(21):9175–9188.

Sock, E., Rettig, S. D., Enderich, J., Bösl, M. R., Tamm, E. R., and Wegner, M. (2004). Gene targeting reveals a widespread role for the high-mobility-group transcription factor Sox11 in tissue remodeling. *Molecular and cellular biology*, 24(15):6635–44.

Sokol, S. Y. (2011). Maintaining embryonic stem cell pluripotency with Wnt signaling. *Development (Cambridge, England)*, 138(20):4341–50.

Somasundaram, R., Prasad, M. A. J., Ungerbäck, J., and Sigvardsson, M. (2015). Transcription factor networks in B-cell differentiation link development to acute lymphoid leukemia. *Blood*, 126(2):144–52.

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–45.

Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Akerman, M., Alioto, T., Ambrosini, G., Antonarakis, S. E., Behr, J., and Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nature methods*, 10(12):1177–84.

Streets, A. M. and Huang, Y. (2014). How deep is enough in single-cell RNA-seq? *Nature biotechnology*, 32(10):1005–6.

Su, W., Slepenkov, S. V., Slevin, M. K., Lyons, S. M., Ziemniak, M., Kowalska, J., Darzynkiewicz, E., Jemielity, J., Marzluff, W. F., and Rhoads, R. E. (2013). mRNAs containing the histone 3' stem-loop are degraded primarily by decapping mediated by oligouridylation of the 3' end. *RNA (New York, N.Y.)*, 19(1):1–16.

Sumazin, P., Yang, X., Chiu, H.-S., Chung, W.-J., Iyer, A., Llobet-Navas, D., Rajbhandari, P., Bansal, M., Guarnieri, P., Silva, J., Califano, A., and et Al. (2011). An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell*, 147(2):370–81.

Surface, L. E., Thornton, S. R., and Boyer, L. A. (2010). Polycomb Group Proteins Set the Stage for Early Lineage Commitment. *Cell Stem Cell*, 7(3):288–298.

Svensson, V., Nath Natarajan, K., Ly, L.-H., Miragaia, R. J., Labalette, C., Macaulay, I. C., Cvejic, A., and Teichmann, S. A. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nature Publishing Group*, 14.

Tam, O. H., Aravin, A. A., Stein, P., Girard, A., Murchison, E. P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R. M., and Hannon, G. J. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, 453(7194):534–538.

Tanay, A. and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. *Nature*, 541(7637):331–338.

Tay, Y., Kats, L., Salmena, L., Weiss, D., Tan, S., Ala, U., Karreth, F., Poliseno, L., Provero, P., DiCunto, F., Lieberman, J., Rigoutsos, I., and Pandolfi, P. (2011). Coding-Independent Regulation of the Tumor Suppressor PTEN by Competing Endogenous mRNAs. *Cell*, 147(2):344–357.

Tchwenko, S. N. (2012). The Burden of Cardiovascular Disease in North Carolina. Technical report.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science (New York, N.Y.)*, 290(5500):2319–2323.

Tian, B. and Manley, J. L. (2016). Alternative polyadenylation of mRNA precursors. *Nature Publishing Group*, 18.

Titsias, M. K. and Lawrence, N. D. (2010). Bayesian Gaussian Process Latent Variable Model. In *AISTATS*, pages 844–851.

Tonner, P., Srinivasasainagendra, V., Zhang, S., and Zhi, D. (2012). Detecting transcription of ribosomal protein pseudogenes in diverse human tissues from RNA-seq data. *BMC Genomics*, 13(1):412.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–6.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9):1105–11.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–5.

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., Desai, T. J., Krasnow, M. A., and Quake, S. R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, 509(7500):371–5.

Tsang, J., Zhu, J., and van Oudenaarden, A. (2007). MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Molecular cell*, 26(5):753–67.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121.

Uchino, K., Hirano, G., Hirahashi, M., Isobe, T., Shirakawa, T., Kusaba, H., Baba, E., Tsuneyoshi, M., and Akashi, K. (2012). Human Nanog pseudogene8 promotes the proliferation of gastrointestinal cancer cells. *Experimental Cell Research*, 318(15):1799–1807.

Van Der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

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

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., MacLeod, J. N., Chiang, D. Y., Prins, J. F., and Liu, J. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic acids research*, 38(18):e178.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63.

Welch, J. D., Hartemink, A. J., and Prins, J. F. (2016a). SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biology*, 17(1):106.

Welch, J. D., Slevin, M. K., Tatomer, D. C., Duronio, R. J., Prins, J. F., and Marzluff, W. F. (2015). EnD-Seq and AppEnD: sequencing 3' ends to identify nontemplated tails and degradation intermediates. *RNA (New York, N.Y.)*, 21(7):1375–89.

Welch, J. D., Williams, L. A., DiSalvo, M., Brandt, A. T., Marayati, R., Sims, C. E., Allbritton, N. L., Prins, J. F., Yeh, J. J., and Jones, C. D. (2016b). Selective single cell isolation for genomics using microraft arrays. *Nucleic acids research*, 44(17):8292–301.

Whyte, W. A., Bilodeau, S., Orlando, D. A., Hoke, H. A., Frampton, G. M., Foster, C. T., Cowley, S. M., and Young, R. A. (2012). Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. *Nature*, 482(7384):221.

Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., Mburu, F. M., Mantalas, G. L., Sim, S., Clarke, M. F., and Quake, S. R. (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nature methods*, 11(1):41–6.

Yang, F., Yi, F., Han, X., Du, Q., and Liang, Z. (2013). MALAT-1 interacts with hnRNP C in cell cycle regulation. *FEBS letters*, 587(19):3175–3181.

Yao, C. and Shi, Y. (2014). Global and quantitative profiling of polyadenylated RNAs using PAS-seq. *Methods in molecular biology (Clifton, N.J.)*, 1125:179–85.

Yi, H., Liang, B., Jia, J., Liang, N., Xu, H., Ju, G., Ma, S., and Liu, X. (2013). Differential roles of miR-199a-5p in radiation-induced autophagy in breast cancer cells. *FEBS Letters*, 587(5):436–443.

Zhou, A., Ou, A. C., Cho, A., Benz, E. J., and Huang, S.-C. (2008). Novel splicing factor RBM25 modulates Bcl-x pre-mRNA 5' splice site selection. *Molecular and cellular biology*, 28(19):5924–5936.

Zhu, C., Gao, Y., Guo, H., Xia, B., Song, J., Wu, X., Zeng, H., Kee, K., Tang, F., Yi, C., and et Al. (2017). Single-Cell 5-Formylcytosine Landscapes of Mammalian Early Embryos and ESCs at Single-Base Resolution. *Cell stem cell*, 338(0):1622–1626.