8-2016

The design and statistical analysis of single-cell RNA-sequencing experiments

Faye H. Zheng
Purdue University

Graduate School Form
30 Updated 12/26/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Faye H. Zheng

Entitled
The Design and Statistical Analysis of Single-Cell RNA-Sequencing Experiments

For the degree of Ph.D.

Is approved by the final examining committee:

Rebecca W. Doerge

Hyonho Chun

Bruce A. Craig

Gayla R. Olbricht

Timothy L. Ratliff

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Rebecca W. Doerge

Approved by: Hao Zhang 7/7/2016

Head of the Departmental Graduate Program Date

THE DESIGN AND STATISTICAL ANALYSIS OF SINGLE-CELL RNA-SEQUENCING EXPERIMENTS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Faye H. Zheng

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2016

Purdue University

West Lafayette, Indiana

To my parents. The starting point of all my achievements lie with you.

ACKNOWLEDGMENTS

On the academic front, I can only begin by thanking my advisor, Rebecca W. Doerge, for being a guiding presence right from the beginning, when I entered as a doe-eyed first-year with everything to learn. Throughout the graduate school process, Rebecca has been an essential source of advice, trust, and friendship. I am also grateful to my committee members - Bruce Craig, Hyonho Chunh, Gayla Olbricht, and Tim Ratliff - for their helpful suggestions which provided direction for this work. The data which Dr. Ratliff graciously allowed me to use played an integral role in my research. I would also like to recognize professor James Eberwine and the dozen students I met at the 2014 Single Cell Analysis course at Cold Spring Harbor Laboratory. These were a formative few weeks in my understanding and engagement with the field; I could hardly believe how fun it was to wield a pipette and talk endlessly about biology.

I have benefited greatly from interactions among the RWD research group. Those who graduated before me - Jeremiah Rounds, Sanvesh Srivastava, Chee Chen, Doug Baumann, Tilman Achberger - set high standards in how to ask good questions, give spotless talks, and engage in academic citizenship. Present members - Patrick Medina, Emery Goossens, Ji Hwan Oh, Nadia Atallah, Yumin Zhang - always bring something interesting to the discussion table, and must be commended for sitting through endless practice talks. The graduate school experience would be nothing without good company. April, Kelly-Ann, Xiaosu, Kylie - my friends and my cohort, with whom I shared the long road of classes to quals to research. Nithin, Ana, Jeff, and Shaili - my fellow running, climbing, and 14er enthusiasts. Ruth and Brittany - fellow Women in Science Program members who met my worries with empathy and shared my successes with glee.

To my parents, who have always given me full independence to occupy my life with my own choices, even the offbeat ones all taken in stride. I have been greatly affected by your boundless confidence, no matter where I go or what I do. To my brother Shawn, whom I've had the distinct pleasure of watching develop into an exceptionally mature and hardworking young man. That we share the same genes I consider a wonder and something of a bragging right. To my chosen sisters - Robin, Anqi, and Rochelle - for our longstanding friendships that keep getting richer despite great distances. And finally to Benjamin, for bringing to my life such humor, steadiness, and perspective - I appreciate you every day.

Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABBREVIATIONS

BFGS	Broyden-Fletcher-Goldfarb-Shanno (algorithm)
BH	Benjamini-Hochberg
cDNA	complementary DNA
CML	conditional maximum likelihood
DE	differential expression, differentially expressed
DNA	deoxyribonucleic acid
FACS	fluorescence-activated cell sorting
FDR	false discovery rate
FPR	false positive rate
GLM	generalized linear model
mESC	mouse embryonic stem cell
mRNA	messenger RNA
NB	negative binomial
NGS	next-generation sequencing
PCA	principal component analysis
PrE	primitive endoderm (cell)
RNA	ribonucleic acid
RNA-seq	RNA-sequencing
ROC	receiver operating characteristic (curve)
SBS	sequencing by synthesis
scRNA-seq	single-cell RNA-seq
SVA	surrogate variable analysis
UMI	unique molecular identifiers
TPR	true positive rate
ZINB	zero-inflated negative binomial

ABSTRACT

Zheng, Faye H. Ph.D., Purdue University, August 2016. The Design and Statistical Analysis of Single-Cell RNA-Sequencing Experiments. Major Professor: R.W. Doerge.

Next-generation DNA- and RNA-sequencing (RNA-seq) technologies have expanded rapidly in both throughput and accuracy within the last decade. The momentum continues as emerging techniques become increasingly capable of profiling molecular content at the level of individual cells. One goal of this research is to put forward best practices in the design of single-cell RNA-sequencing (scRNA-seq) experiments, specifically as it relates to choices regarding the trade-off between sequencing depth and sample size. In addition to general guidelines, an interactive tool is presented to aid researchers in making experiment-specific decisions that are informed by real data and practical constraints. Further, a new approach to the modeling and testing of differential gene expression in scRNA-seq data is proposed, which notably incorporates salient features (*e.g.* highly zero-inflated expression values) of single-cell transcription that are otherwise obscured at the tissue level. As single-cell technologies offer an unprecedented window into cell-to-cell heterogeneity and its biological consequences, it is essential that suitable approaches are adopted for both the design and analysis of these experiments.

1. INTRODUCTION

Next-generation DNA- and RNA-sequencing technologies have expanded rapidly in both throughput and accuracy within the last decade. The momentum continues as emerging techniques become increasingly capable of profiling molecular content at the level of individual cells. Cell-to-cell heterogeneity and its biological consequences are now the focus of many unprecedented studies capable of illuminating the dynamic nature of single cells. Recent investigations have pushed the boundaries of understanding structural changes in cancer genomes, varying paths of cell differentiation, and finer mechanisms of cell regulation. Like many emerging technologies, the statistical analysis of single-cell data currently remains in the exploratory stage, but is poised to shift towards informative tests of specific hypotheses. Moving forward, thoughtful decisions regarding experimental design are essential if these experiments are to be maximally efficient, reproducible, and informative. One of the overarching goals of this research is to put forward best practices in the design of single-cell RNA-sequencing (scRNA-seq) experiments, specifically as it relates to choices regarding the trade-off between sequencing depth and sample size.

Aside from experimental design, the statistical analysis of scRNA-seq data itself invites a critical revisitation of standard RNA-seq methods. In particular, the modeling and testing of differential gene expression is currently addressed by implementing a variety of standard and available methods which incorporate salient features of tissue-level RNA-seq data. Because these current methods do not adequately extend to RNA-seq data from single cells, another goal of this work is the development of a novel approach for the detection of differential gene expression signatures between subpopulations of single cells; this is essential, given the great interest in understanding cell-to-cell heterogeneity.

1.1 History of Sequencing Technologies

The aim of DNA sequencing technologies is to decipher the order of nucleotides (*i.e.*, the adenine, thymine, cytosine, and guanine units, collectively called bases) in a DNA molecule, which constitutes the genetic code of an organism. Sanger sequencing marked the inception of these technologies, and culminated in the completion of the landmark Human Genome Project in 2001 [Lander et al., 2001]; this feat ushered in the age of genomics. The second wave of sequencing methods, beginning in 2004 and widely used today, brought with it substantial increases in speed and throughput. The parallel, automated nature of the process, commonly dubbed "next-generation sequencing" (NGS), produces millions of sequences concurrently, increasing throughput by many orders of magnitude [Metzker, 2010]. In addition, these high-throughput sequencing technologies have significantly decreased the cost of sequencing, which is now less than ten cents per megabase [National Human Genome Research Institute, 2015].

1.2 Next-Generation Sequencing: From Tissues to Cells

NGS procedures have become more affordable, ubiquitous, even routine, and yet the ceiling of optimization is being pushed still further. Capitalizing on well-established NGS platforms, recent technological advances have enabled a dramatic scaling down in the amount of genomic starting material required to produce sequence information. Indeed, it is now possible to sequence at the level of individual cells. In the past, genomic data generated by NGS procedures typically came from aggregating the entire population of thousands to millions of cells within a tissue (Figure 1.1), even though it is increasingly understood that genetic heterogeneity is the norm rather than the exception [Eberwine et al., 2014]. The bulk pooling of cell populations averages out differences between the behaviors of individual cells, blends together the patchwork composition of cells within certain tissues, and obscures the dynamic nature of cellular function. Sequencing at the single cell level allows for

the dissection of genetic heterogeneity with the intent of obtaining a much higher resolution of information.

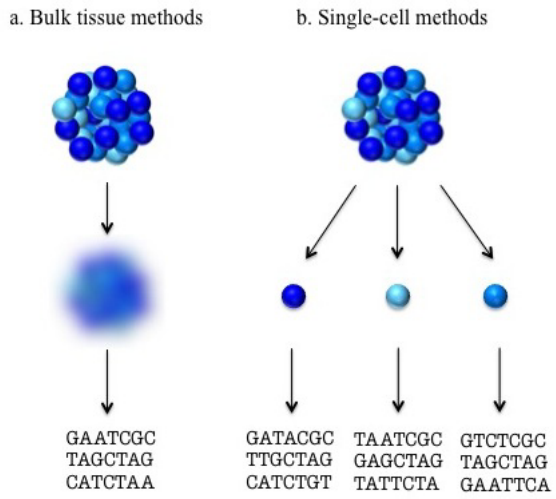

Figure 1.1. Bulk tissue sample (left) is used to obtain sequence information on an aggregate of the entire population of thousands to millions of cells within a tissue. A single-sampled cell (right) allows for genetic heterogeneity to be dissected by obtaining sequence information on each cell within the population of cells.

The ability to ask questions of individual cells has motivated a flood of research in pursuit of insights into both new and longstanding questions that previously could not be answered from bulk tissue analysis [Shapiro et al., 2013]. Living tissues are often comprised of a multitude of cell types with different lineages, stages of development, and function within the tissue. Cell lineage is particularly important in the study of intratumor heterogeneity; several single-cell sequencing studies have shown that tumor development occurs through a series of somatic mutations that drive groups of cells into distinct clonal subpopulations, each with its own mutational signatures and even drug response [Navin et al., 2011, Alexandrov and Stratton, 2014, Yates and Campbell, 2012]. Single-cell technologies have also made it possible to detect the presence of cancer by way of rare circulating tumor cells in blood specimens [Ramsköld et al., 2012, Cann et al., 2012]. Aside from cancer applications, the sensitivity of single-cell

sequencing allows for the isolation and characterization of complex microbes in the environment, offering a way to detect low-abundance and sometimes unculturable species [Yilmaz and Singh, 2012, Blainey, 2013]. The prevalence of somatic mosaic mutations in individual neurons of the human brain has recently been highlighted [McConnell et al., 2013], setting the stage for studying the roles of this mosaicism for neurodevelopmental diseases [Poduri et al., 2013]. Applications of NGS have even reached the realm of reproductive health, where single-cell sequencing has demonstrated its utility in diagnosing potential problems with *in-vitro* fertilized embryos prior to implantation, and in offering a viable non-invasive alternative for prenatal testing [Yan et al., 2013, Chandrasekharan et al., 2014]. Promising ventures have also been made into single-cell epigenomics [Lorthongpanich et al., 2013], proteomics [Willison and Klug, 2013], and metabolomics [Rubakhin et al., 2013], thus rounding out the astounding range of possibilities for single-cell NGS technologies.

1.3 RNA-Sequencing

1.3.1 Basics of RNA

RNA-sequencing (or RNA-seq) is one application of NGS high-throughput technologies, and is the primary focus of this work. RNA-seq is used to measure gene expression by sequencing and quantifying a sample's mRNA content. To fully understand the context of RNA-seq and what its measurements represent, it is instructive to review how genetic information flows from DNA to biological function, as explained by the classic Central Dogma of Biology (Figure 1.2) [Crick et al., 1970]. DNA, located in the nucleus of every cell, consists of a sequence of nucleotides that comprise the organism's genetic code. Genes are specific sections of DNA that encode for a particular protein or function. Through the process of transcription, the genetic information in DNA becomes copied into complementary strands of messenger RNA (mRNA). These aptly named mRNA deliver the copied genetic information to the

outside of the nucleus, where they are translated into proteins that ultimately carry out biological function.

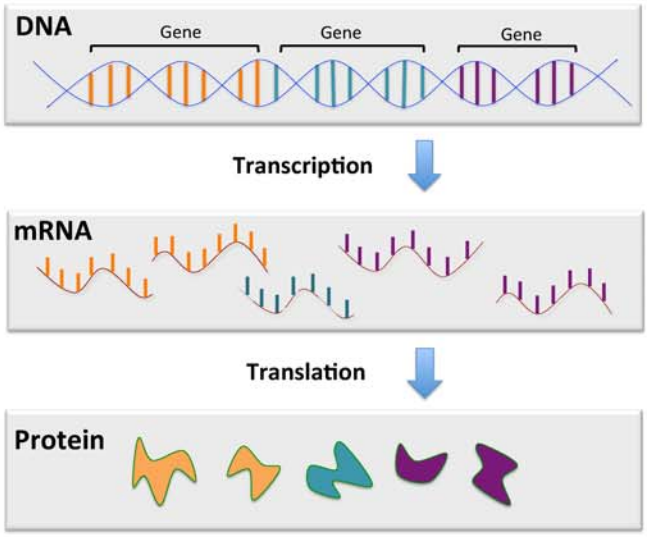

Figure 1.2. The Central Dogma of Biology [Crick et al., 1970] depicts the flow of genetic information. DNA is the genetic code; RNA (specifically, mRNA) are transcribed copies of genes located on the DNA; these mRNA are translated into proteins that carry out function.

The premise of using RNA-seq to measure gene expression is based on the important fact that transcription does not occur uniformly for all genes. In fact, regulatory proteins that exist in each cell constantly work to modulate transcription, so that some genes have many more mRNA copies made, and others fewer or none. Hence, obtaining the expression level of a gene boils down to measuring the quantity of its mRNA copies, that is, how many times the gene was transcribed. Thus, quantifying mRNA abundances across all genes, as RNA-seq does, paints an overall picture of the transcriptional machinery within the sample, whether that be a bulk tissue sample or a single cell.

1.3.2 The Process of RNA-Seq

The RNA-seq process begins with the extraction of mRNA from the given biological starting material. In original applications involving the analysis of bulk tissue, whole tissues are simply obtained by sampling, dissecting, or biopsying the organism of interest. For single-cell investigations, this primary tissue is first disassociated into its constituent cells, which must be isolated intact; cells that pass screening procedures for viability are finally submitted for further processing.

Single cell isolation does not yet have a single standard procedure and remains an active area of development and refinement [Saliba et al., 2014]. At the most rudimentary level, cells may be isolated by micromanipulation under a microscope using a patch pipette or nanotube. Despite the obvious limitations of low throughput, high risk of disruption, and the potential for experimental bias towards certain morphologies, manual handling is still employed for targeted applications, such as for rare cells [Shapiro et al., 2013].

The single cell gene expression applications considered here rely on the far more prevalent automated methods for isolating cells at high volume. For example, the technique known as fluorescence-activated cell sorting (FACS), which involves flow-sorting cells that are labeled with fluorescent antibodies [Shapiro, 2005], has achieved popularity due to its wide availability on commercial platforms [Saliba et al., 2014]. Another rapidly expanding and highly efficient approach is the use of automated microfluidic devices that compartmentalize cells into low-volume chambers and simultaneously screen them for viability. Current iterations of this technology offer standard plates with 96-well capacity for the parallel isolation of cells. However, 800-well plates are on the near horizon [Fluidigm, Inc., 2015], signaling the imminent need for statistical and computational tools that can accommodate this rapidly expanding scale of data.

Once the initial biological material has been obtained, whether from a tissue or individual cell, the mRNA that is extracted must be reverse-transcribed into comple-

mentary strands of cDNA. This step is required due to the fact that DNA molecules are much more biologically stable and resistant to degradation; in fact, for this reason all NGS technologies are solely designed for sequencing DNA, rather than RNA directly. This cDNA acts as input to the remaining RNA-seq protocol.

While there are several NGS sequencing platforms available that are capable of performing RNA-seq (*e.g.*, SOLiD, Roche 454, Pacific Biosciences, Ion Torrent), by far the most successful and widely adopted is the Illumina platform, whose "sequencing by synthesis" (SBS) chemistry has produced approximately 90% of global sequencing data, by the company's own accounts [Illumina, Inc., 2015a]. The general workflow, specific to the Illumina platform, can be broken into four basic steps: library preparation, cluster amplification, sequencing, and read alignment (Figure 1.3) [Illumina, Inc., 2015b].

Following reverse-transcription of the extracted mRNA, the resulting cDNA undergoes preparation for sequencing (Figure 1.3A). Specifically, the cDNA is randomly fragmented into millions of pieces, and specialized adapters are ligated to the ends of each piece. The resulting collection of fragments comprise the units which will get sequenced, and hence is termed the 'sequencing library'. The adapters on each fragment help attach them onto the surface of the flow cell within the sequencing machine (Figure 1.3B). Each fragment is copied thousands of times through many cycles of 'bridge amplification', creating distinct clusters containing identical copies of the same fragment. Sequencing by synthesis, specific to the Illumina platform, proceeds in the following manner (Figure 1.3C): fluorescently labeled nucleotides are washed onto the surface of the flow cell; as each nucleotide binds to a complementary base on a fragment cluster, its fluorescent signal is emitted and read as a digital image to identify the base; this wash-and-scan cycle is repeated one-by-one for each consecutive base, for all fragment clusters in parallel, to generate milions of sequenced reads of about 125 to 300 bases each in length. From this point in the workflow, computational tools are used to map each read to its appropriate location on a reference

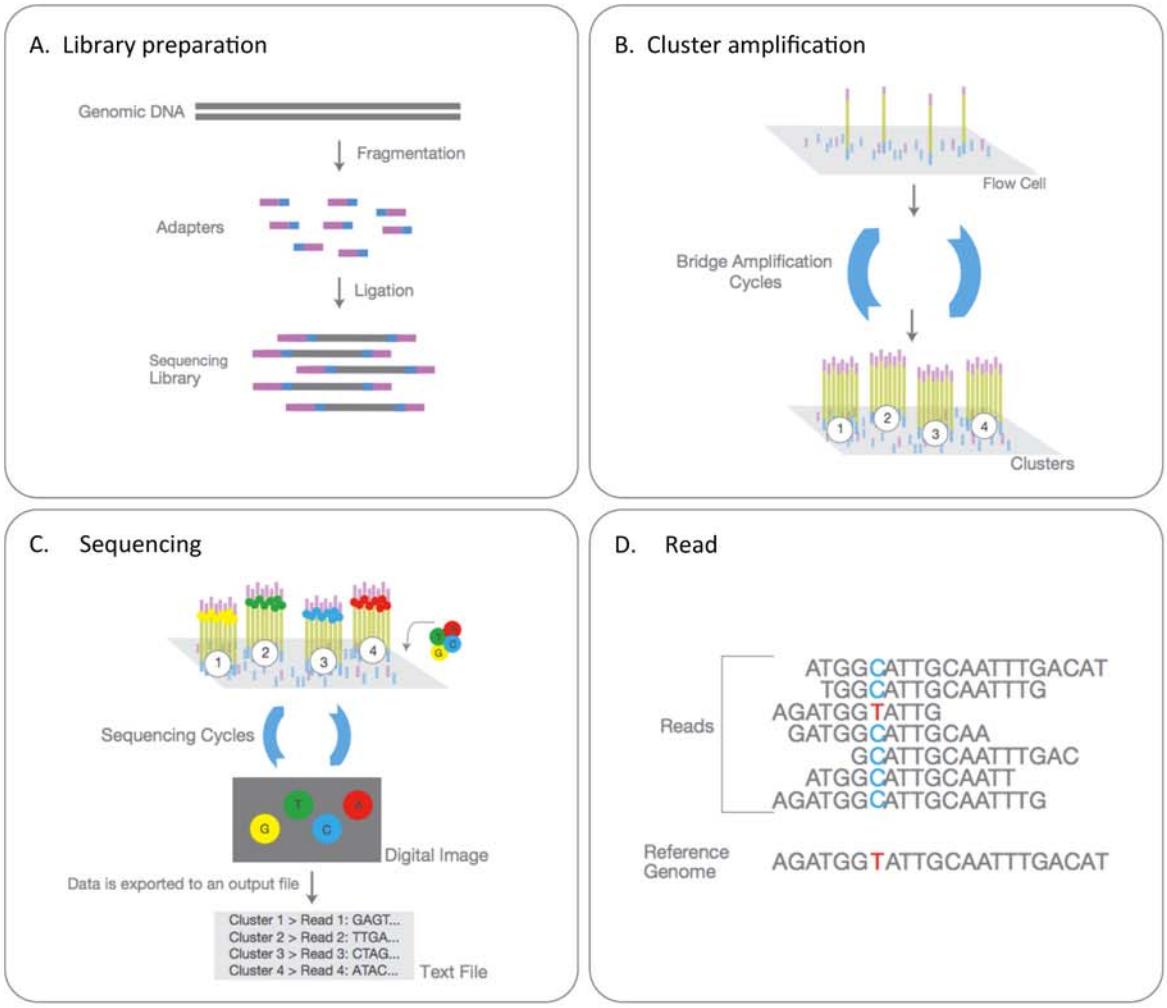

Figure 1.3. NGS workflow featuring the Illumina sequencing by synthesis (SBS) technology [Illumina, Inc., 2015b]. **A.** cDNA is randomly fragmented into millions of pieces and adaptors are ligated to the ends. **B.** The fragments are attached to the surface of the sequencing flow cell, then copied thousands of times, creating distinct clusters containing identical copies of the same fragment. **C.** SBS proceeds by washing fluorescently labeled nucleotides onto the flow cell, and using digital imaging to identify the bases as they are incorporated; this cycle is repeated one-by-one for each consecutive base until the desired lenth of sequenced reads is achieved. **D.** Reads are aligned to a reference genome using computational tools.

genome, a process called sequence alignment. If no reference genome is available, the reads can also be assembled *de novo* (Figure 1.3D).

Gene expression is finally quantified by counting the number of reads that map to the genomic feature of interest, *e.g.*, genes [Wang et al., 2009, Oshlack et al., 2010, Mortazavi et al., 2008]. The data are typically represented as a matrix, in which genes constitute row labels, samples constitute column labels, and values within the matrix are read counts representing the expression of a particular gene in a particular sample (Table 1.1). Samples can represent either bulk tissue samples or single-cell samples; in either case, the matrix representation is the same.

Table 1.1
RNA-seq data are typically represented as a matrix of the following form. The values y_{ig} represent the expression of gene g in sample i. The library sizes, $L_i = \sum_{g=1}^{G} y_{ig}$, are the total number of reads aligned to sample i across all genes.

	Sample 1	Sample 2	...	Sample N
Gene 1	y_{11}	y_{21}	...	y_{N1}
Gene 2	y_{12}	y_{22}	...	y_{N2}
...	...	...	...	...
Gene G	y_{1G}	y_{2G}	...	y_{NG}
	L_1	L_2	...	L_N

1.3.3 Bulk Tissue vs. Single Cell Protocols

The workflow described in Figure 1.3 is shared between the RNA-sequencing of bulk tissues and that of individual cells. However, the single-cell procedure requires one important extra step: additional amplification of the cDNA during library prepa-

ration. Specifically, this amplification is applied to the fragments of genomic cDNA prior to adapter ligation (Figure 1.3A), and is done repeatedly until the DNA concentration matches the requirements of the sequencing technology. The amount of amplification required can often be around one million-fold, substantially more magnitudes beyond what is necessary for bulk tissue sequencing. This is a direct consequence of the scant amount of biological material that single cells provide.

Amplification comes with the unfortunate cost of biases that compromise quantitative accuracy, most often in the form of nonlinear distortions of transcript abundance and preferential amplification of certain sequence patterns. Amplification biases have previously been noted with traditional bulk RNA-seq, and a number of methods exist to correct the resulting data. The additional cDNA amplification of single-cell quantities exacerbates this already-existing problem and has required this issue to be revisted. Recently, Islam et al. [2014] developed an inventive technique in which unique labels are attached to each single-cell cDNA molecule prior to amplification. These labels, called unique molecular identifiers (UMIs), mark as distinct each molecule originally present in the sample. Following amplification, one can quantify gene expression by counting only the number of distinct UMIs aligned to each genomic feature, rather than counting all the amplified reads that are aligned. Since this method effectively counts only the original, unamplified molecules, amplification noise may be avoided altogether.

Amplification biases, combined with the delicate process of isolating single cells and the technical difficulty of sequencing a miniscule pool of transcripts, contribute substantially to the high levels of technical noise seen in scRNA-seq data. While not addressed directly in this work, these challenges that set single-cell sequencing apart are important considerations in other aspects of the design and statistical analysis of single-cell experimental data.

2. EXPERIMENTAL DESIGN OF SCRNA-SEQ EXPERIMENTS

Two leading questions are central to the design of a scRNA-seq experiment: the depth at which to sequence each cell, and how many cells to sequence. These decisions are affected by the biological question being considered, and by the tradeoffs imposed by practical financial constraints.

2.1 Sequencing Depth and Replication

The currently accepted definition of sequencing coverage originated from Lander and Waterman [1988]. This work first defined theoretical coverage as LN/G, where L is the length of each sequencing read, N is the number of high-quality reads aligned to the genome, and G is the total number of bases in the genome. In other words, this is the expected number of times that a given base is covered by a read. It is often reported as a technical specification of a sequencing experiment (*e.g.*, samples were sequenced at $1\times$ or $30\times$ coverage). The terms coverage, depth, and depth of coverage, all referring to this definition, are used interchangeably in the literature. In practice, particularly in RNA-seq, is often thought of as simply the total number of reads that are mapped to the genome and then counted as gene expression measurements (Figure 2.1). This will be the usage of the term subsequently adopted here.

The higher the sequencing depth, the more accurate the quantification of gene expression. This stems from the imperfect nature of the sequencing technology, in which reads are short and contain errors. At higher sequencing depths, alignment tools are better able to distinguish a base that is sequenced in error to a base that is a true variant from the reference genome. For example, a base that is covered by twenty reads, of which the base call consistently varies from the reference genome in a majority of those reads, is much more likely to be a true genetic variant than

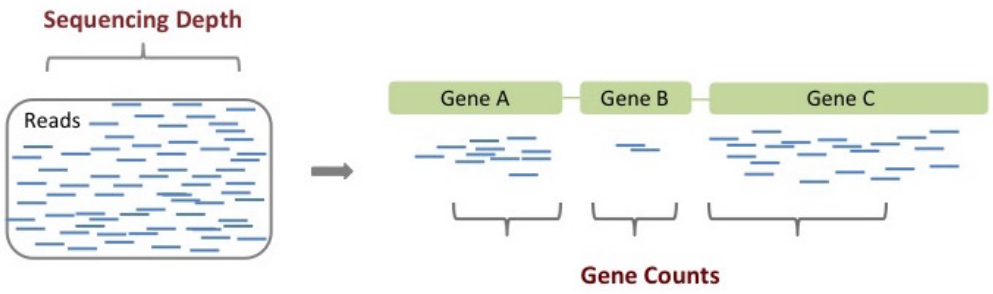

Figure 2.1. In the context of RNA-seq, sequencing depth most commonly refers to the total number of reads that are mapped to the genome and subsequently quantified as gene expression measurements.

a sequencing error. At lower depths, this distinction is harder to make [Sims et al., 2014]. In addition, there exist genes with low expression levels that are hence represented by fewer mRNA transcripts in the biological pool. Higher sequencing depths increase the likelihood that even these rare transcripts are sequenced. To illustrate, in Figure 2.1, a reduction to the pool of reads could lead to gene B being missed completely, whereas the remaining sequencing real estate becomes concentrated in the more highly expressed genes A and C.

Despite the clear benefits of sequencing at sufficiently high depths, researchers would be remiss to simply sequence as much as possible. Higher sequencing depths are accompanied by higher costs, as sequencing machines can accommodate only a limited number of reads per expensive run. Moreover, it has been shown that there exists a point of diminishing returns at which continuing to increase the sequencing depth fails to yield substantially more genomic information. This is demonstrated in so-called 'saturation curves', which plot the number of genes detected in a given sample against an increasing number of reads. The saturation curves in Figure 2.2, generated from the `R` package `NOISeq` [Tarazona et al., 2011], demonstrate this property using randomly chosen samples from a real scRNA-seq data set on human prostate cell lines (described in Section 2.2).

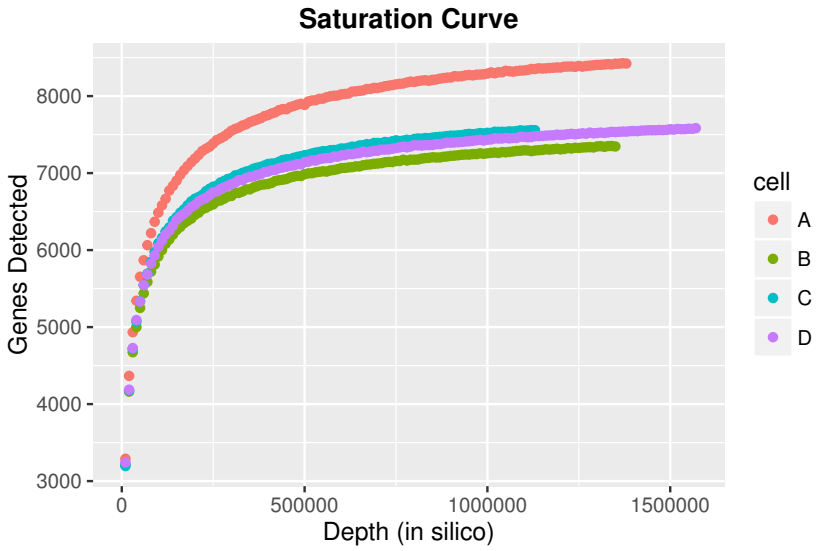

Figure 2.2. Saturation curves of randomly chosen samples from a real scRNA-seq data set on human prostate cancer cell lines. Each curve plots the number of detected genes, defined as genes with counts greater than 3, against against sequencing depth for a cell sample. For a full description of how this plot was generated, see Tarazona et al. [2011]. As the number of reads increase, the number of genes detected also increases, but begins to taper off. This pattern is typical of both bulk and single-cell RNA-seq data.

Sequencing depth is but one piece of the puzzle when designing a scRNA-seq experiment. A second consideration of great practical interest to a researcher is the optimal number of cells to sequence. The pricing structure of the sequencing technology links these two choices of depth and replicates. As mentioned previously, the sequencing reaction occurs on the surface of a flow cell within the machine. In practice, these flow cells are composed of multiple independent lanes, with a limit to the number of reads that can be sequenced per lane. For example, the various Illumina systems can accommodate between 80 to 200 million reads per lane, depending on the choice of read length. Importantly, the largest cost of a sequencing experiment is in the price per lane. Therefore, given the financial constraints of an experiment, there exists a tradeoff between sequencing fewer cells with more reads each or more

cells each at lower depths. The optimal balance point is the question of interest, specifically in the context of detecting differential expression.

The question of optimizing the trade-off between sequencing depth and biological replicates has been asked previously of bulk RNA-seq tissues [Liu et al., 2014]. However, here the focus has changed, in keeping with the shift in context going from designing experiments for bulk samples as opposed to single cells. While bulk RNA-seq studies often limit themselves to around a dozen samples (sometimes more but often less) over two or more treatments, single-cell studies are seen to involve hundreds to occasionally thousands of cells that are considered biological replicates. This is partially due to the enormous reduction of labor and cost involved in isolating single cells as opposed to collecting tissue samples from whole organisms. It is also partially a necessity; a large number of cells is needed to characterize cell populations and to counter the variability of each individual cell. Hence, while in bulk RNA-seq the popular recommendation is to always obtain as many biological replicates as possible, for single-cell applications, the question remains whether there may be a saturation point beyond which more replicates is not necessary. With respect to sequencing depth, the standard is to use around 30 million reads per sample for bulk RNA-seq differential expression studies. By contrast, the number of reads per single cell, though markedly less than what is used for bulk samples, still varies substantially between studies [Stegle et al., 2015]. For example, Jaitin et al. [2014] used around $20,000$ reads per cell for over $1,500$ cells, while Mahata et al. [2014] sequenced an average of 16 million reads per cell for each of around 90 cells.

There is currently no accepted rule of thumb or guide that either empirically or theoretically instructs researchers about the optimal choice of sequencing depth and/or replicate number. Certainly, understanding this relationship has great value when designing a scRNA-seq experiment. Here, a simulation study attempts to provide guidance by investigating the effect of different combinations of sequencing depths and numbers of replicates on the detection of differential gene expression.

2.2 Procedure for Simulating scRNA-seq Data

The starting point of any simulation study is the choice of how to generate simulated data in a way that adequately mimics real data. Throughout this work, count data are all simulated from a zero-inflated negative binomial (ZINB) distribution, with gene-wise parameters extracted from a real scRNA-seq dataset. Specifically, for each gene g, take $\bar{y}_g$ to be the mean of the non-zero counts; $\lambda_g = \bar{y}_g / \sum_g \bar{y}_g$ is then the proportion of all read counts originating from gene g, and can be considered the baseline rate of expression from gene g. The NB dispersion is calculated as $\phi_g = (s_g^2 - \bar{y}_g)/\bar{y}_g^2$, where s_g^2 is the variance of the non-zero counts. Also recorded are the gene-wise proportions of zero counts, p_g. Parameters are sampled in gene-wise triplets of $\{\lambda_g, \phi_g, p_g\}$ to be used to generate simulated gene counts.

To introduce differential expression on a select proportion of genes, coefficients β_g reflecting the group effect are drawn as $\beta_g \sim \text{logNormal}(2, 1)$ for differentially expressed genes, while coefficients for non-DE genes are set to 0. The log-linear model for the expected count μ_{gi} of gene g in sample i is

$$\log(\mu_{gi}) = \lambda_g + x_i \beta_g + \log(m_i) \tag{2.1}$$

where x_i indicates the group membership of the i^{th} sample and m_i is the sample i library size included as an offset. Finally, counts are simulated as $y_{gi} \sim \text{NB}(\mu_{gi}, \phi_g)$, with a proportion of counts for each gene set to zero with probability p_g to mimic the zero-inflation prevalent in real scRNA-seq data. Figure 2.3 visually demonstrates that the mean-variance plot of the simulated data suitably mimics that of the original data from which the simulation parameters were sampled.

Real scRNA-Seq Prostate Dataset

All simulated datasets in this work are based on a real scRNA-seq dataset from an experiment involving human prostate cancer cell lines, henceforth referred to simply as the "prostate" dataset. The dataset is comprised of a treatment group containing

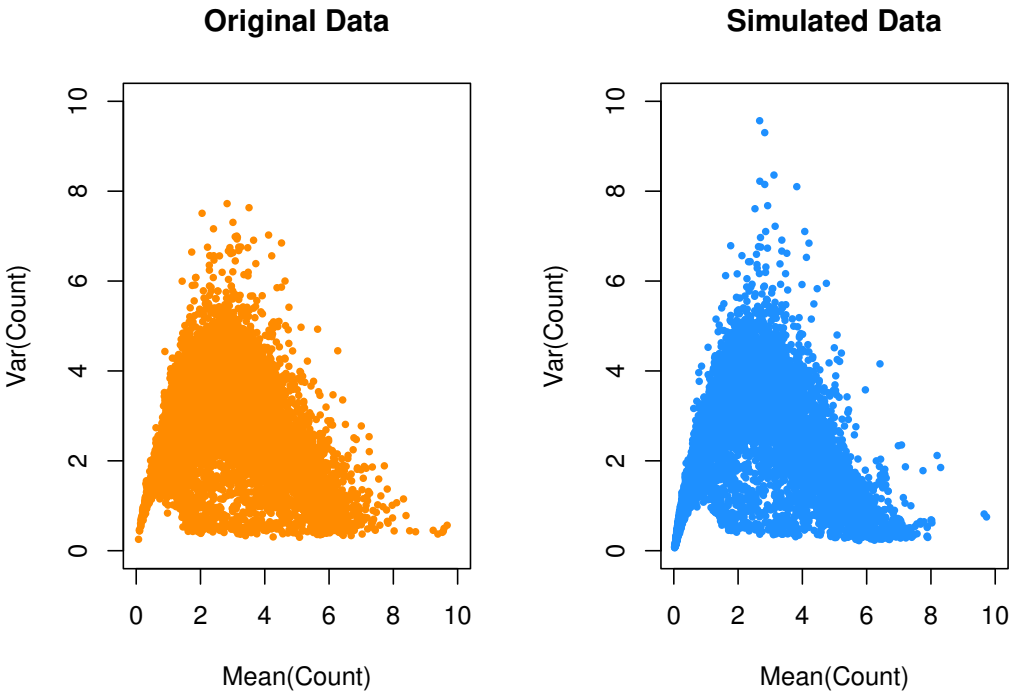

Figure 2.3. Count data from a real scRNA-seq experiment (left) provide the parameters that are used to generate simulated gene counts (right). Mean-variance plots of the simulated gene expression data suitably mimic that of the original data from which the simulation parameters were sampled.

65 cells in which a gene implicated in prostate cancer was knocked out, and the negative control group consisting of 76 cells to which no treatment was applied. Each cell underwent the standard process of cell capture, viability screening, and reverse-transcription. Paired-end libraries were prepared and sequenced on an Illumina HiSeq 2500 machine at an average depth of 1 million reads per cell. Following quality control, alignment, and expression quantification, the resulting count data exhibited a middle-50% of library sizes ranging from 0.85 to 1.3 million read counts. The original data comprised 36,135 sequenced genes, many of which exhibit very low expression levels.

Adopting a standard practice in the literature, only the genes that have average counts of at least 5 across all cells are considered, resulting in 10,854 remaining genes.

Sequencing Depth Resampling

In simulations that follow, sequencing depth is treated as an experimental feature whose effect on the outcome of interest is to be studied. To this end, it is necessary to vary this parameter between otherwise comparable datasets. This process is referred to as "resampling" in general; specifically, "downsampling" or "subsampling" describes the process of generating datasets to lower depths.

Sequencing depth generally refers to the number of reads that are sequenced in an experiment. Recall that the library size is the total number of sequencing reads that are successfully mapped to a sample. The observed difference between raw sequencing depth and the final library size is due to a number of factors that cause a proportion of reads to be discarded. These factors include quality-control filtering, removal of non-mRNA reads (*e.g.*, ribosomal RNA or other artifacts), and reads that fail to map unambiguously to the reference genome. Library sizes are therefore not equivalent to sequencing depth; however, they can reasonably act as a proportionate proxy. It has been argued by Robinson and Storey [2014] that in simulation applications that require subsampling reads in order to perform identical analyses on each subsample, it is functionally identical and substantially more computationally efficient to directly subsample the read count matrix (Table 1.1) as opposed to the raw unaligned reads. Hence, in subsequent simulations, the term 'sequencing depth' is used to refer to the 'library size' as opposed to the 'number of raw sequencing reads'.

Mulitnomial sampling in the following manner is used to obtain samples of desired depths D based on a set of original samples. Let $\boldsymbol{y_i} = \{y_{ig}\}_{g=1}^{G}$ be counts for the G genes from sample i of the original dataset, with corresponding library size $L_i = \sum_g y_{ig}$. Let $\boldsymbol{x_i} = \{x_{ig}\}_{g=1}^{G}$ be the sample generated from $\boldsymbol{y_i}$ with desired depth D_i. The gene counts for $\boldsymbol{x_i}$ are drawn from a multinomial distribution with size D_i

and probabilities $\{y_{ig}/L_i\}_{g=1}^{G}$. This results in simulated samples $\boldsymbol{x_i}$ with the same probability distribution of gene counts as the originating $\boldsymbol{y_i}$, but with the new depth of D_i.

2.3 Simulations

In order to study the relationship between sequencing depth and replicate number and their combined effect on detecting differential expression, datasets of varying depths and replication levels were generated from the original prostate dataset. The depths considered for the simulation are $D = \{0.1, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8\}$ million reads. The number of replicates per treatment group considered are $N = \{10, 20, 30, 40, 50, 60\}$; that is, N replicates were randomly chosen from each of the treatment and control groups. 50 datasets were generated for each combination of $D \times N$ (Table 2.1). The `R` package `edgeR` was applied to each simulated dataset to test for differentially expressed genes. The genes considered 'truly' DE are those testing significant in the most 'robust' simulation scenario, *i.e.*, the dataset with the highest sequencing depth ($D = 1.8$ million reads) and replicates per group ($N = 60$), at a false discovery rate (FDR) cutoff of 0.001. Using these 'true' DE genes as the gold standard for comparison, the statistical power of each experimental design may be calculated as the number of true DE genes that are also detected as DE (true positives), divided by the total number of true DE genes (positives).

Figure 2.4 (top) shows the statistical power to detect DE genes as a function of depth and replicates. Increasing the number of replicates substantially and consistently increases statistical power. By contrast, increasing the sequencing depth has a much smaller effect on power, and plateaus off after a point. Figure 2.4 (bottom) depicts the number of differentially expressed (DE) genes detected for the various combinations of sequencing depths and replicates per group. Consistently more DE genes are called as the number of replicates increases, particularly at higher sequencing depths. Increasing the sequencing depth has little to no effect on calling DE genes

Table 2.1
Simulated datasets are generated from the real prostate dataset, by randomly selecting N replicates per experimental group in the real data, and resampling counts to the desired depth D. 50 datasets are simulated for each combination of $D \times N$, and `edgeR` is applied to obtain lists of differentially expressed genes detected in each setting.

Reps per Group (N) \ Depths (D)	0.1M	0.2M	...	1.8M
10				
20		×50		
...				
60				

at lower replication levels, but has increasing effects at higher replication levels. The ROC curves in Figure 2.5 depict the effect of sequencing depth on the accuracy of DE testing for each level of replication. At lower replication levels, increasing the number of reads has some effect on accuracy, most notably moving away from the very lowest depths. However, as replication levels increase, more reads hardly contributes at all to increasing the accuracy of the test. In general, the area under the ROC curve improves as more replicates are included.

The results depicted in Figures 2.4 and 2.5 are limited in the maximum number of replicates per group that they are able to show, as they are based on direct subsampling of a real prostate dataset consisting of only 64 replicates for its smaller experimental group. The effects of greater sample sizes may be observed through generating synthetic data containing higher replicate numbers. This was accomplished by simulating datasets based on parameters extracted from the same human prostate scRNA-seq data (as described in Section 2.2). The intended effect was to mimic the real data in distributional parameters, but with group replicate numbers ranging from

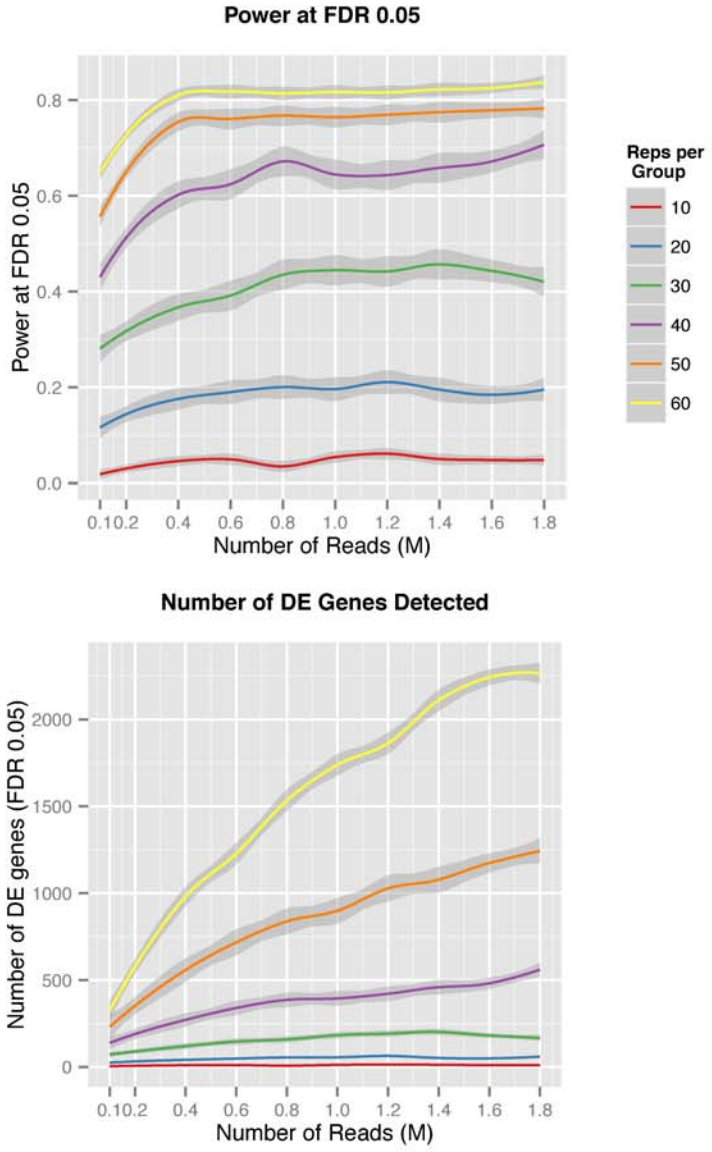

Figure 2.4. Results from datasets subsampled from the real prostate data, to varying sequencing depths and replicates per group. Plots depicted show the statistical power to detect DE genes (top) and the number of differentially expressed (DE) genes detected (bottom). Lines represent replication levels and x-axis depicts sequencing depths. The width of the gray line corresponds to the 95% confidence interval of the mean over multiple simulations.

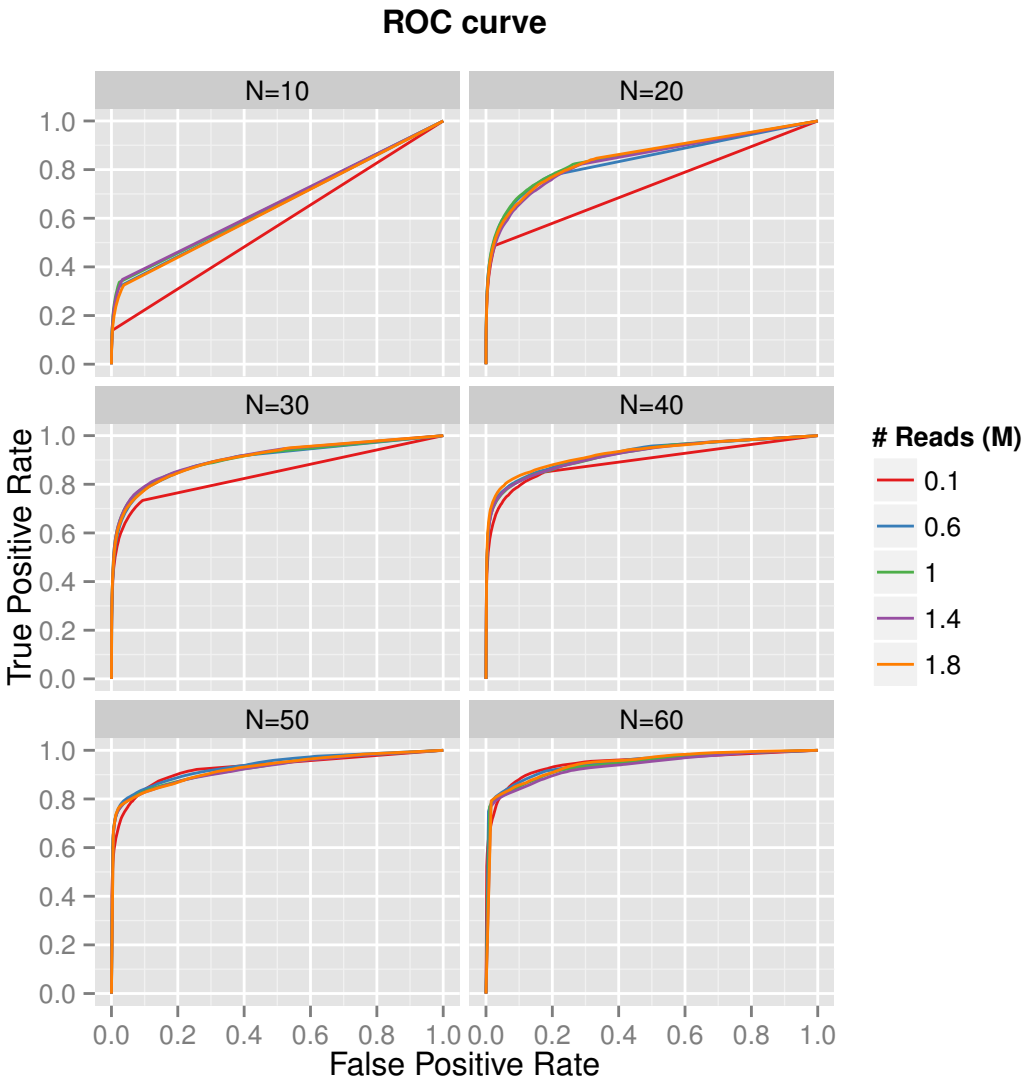

Figure 2.5. Results from datasets subsampled from the real prostate data, to varying sequencing depths and replicates per group. A separate ROC curve is depicted for each of the considered replication levels per group, and individual lines represent sequencing depths.

20 to 200. As expected, as the number of replicates per group continue to increase, diminishing returns are observed in both the number of DE genes detected as well as the statistical power to call true DE genes (Figure 2.6), particularly as the number of replicates per group reaches into the hundreds. ROC curves for the synthetic data

(Figure 2.7) demonstrate similar patterns to those in Figure 2.5, in that higher replicates consistently lead to higher areas under the curve, and increasing sequencing depth has little positive effect beyond the lowest depths.

Observations made here are consistent with what has been suggested previously with Liu et al. [2014] in bulk RNA-seq studies; that is, the number of biological replicates has a markedly more positive effect than sequencing more deeply. This said, for both variables, more is always better, but only up to a point.

2.4 Guiding the Choice of Optimal Experimental Design

Recommendations as to the choice of sequencing depth and replicate number may offer general guidelines in the way of experimental design. However, the real practical interest for researchers is in the ability to make experiment-specific decisions that are informed by the real or expected variability in their data, as well as constraints such as desired statistical power and budgetary limits. As part of this investigation, an interactive tool was implemented in a Shiny web application called `scDesignApp`, which may be accessed at `https://fayezor.shinyapps.io/scDesignApp/`. It is accompanied by an associated `R` package called `scDesign`, which may be installed from GitHub at `https://github.com/fayezor/scDesign`. Given pilot data, typically based on real data, this tool calculates statistical power and estimates costs for each of a user-specified range of experimental designs.

The general workflow for the implementation of the `scDesign` tool is as follows. First, the user provides a pilot dataset from which parameters will be estimated in subsequent calculations. These pilot data may be either a small-scale portion of the planned experiment or related prototype data from similar previous experiments. Recommendations for pilot data best practices are proposed in Section 2.4.3. To compare a variety of hypothetical experimental designs, users must specify a range of sequencing depths and a range of replication levels to be considered. Each combination of depth and replication level constitutes an experimental design. Other

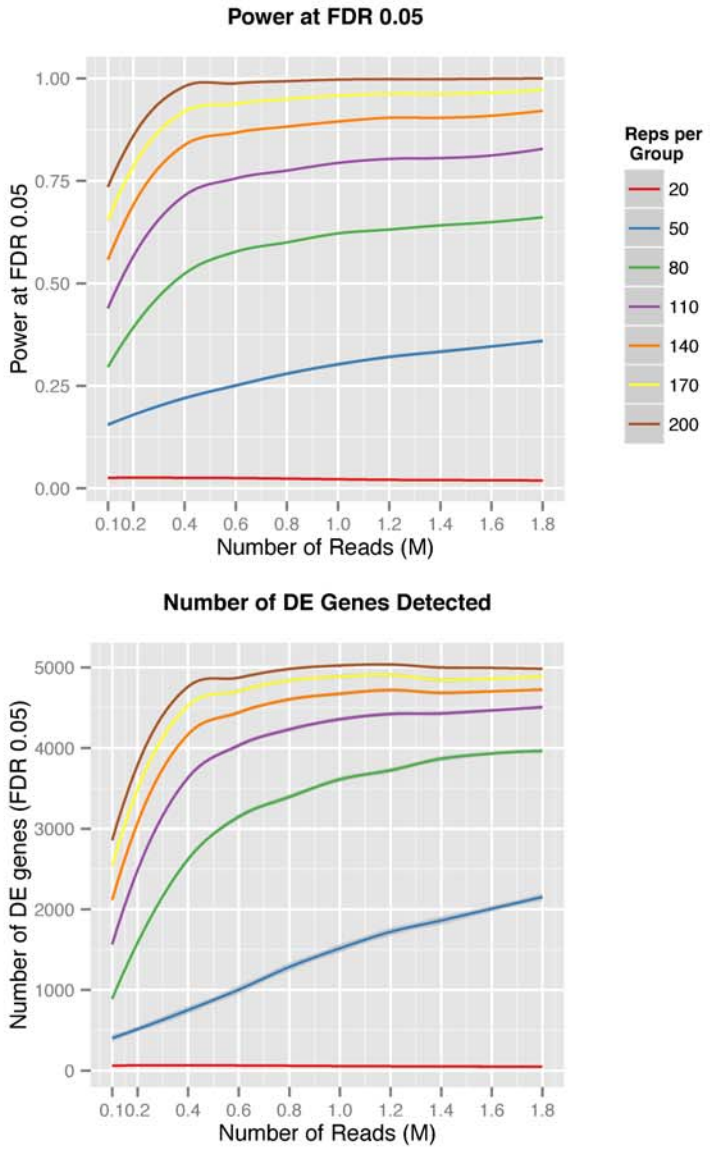

Figure 2.6. Statistical power to detect DE genes (top) and number of differentially expressed (DE) genes detected (bottom), for varying combinations of sequencing depths (in millions of reads) and replicates per group. The width of the gray line corresponds to the 95% confidence interval of the mean number of DE genes over simulation replicates at each setting. Plots were generated from synthetic data simulated using distributional parameters extracted from the human prostate scRNA-seq dataset.

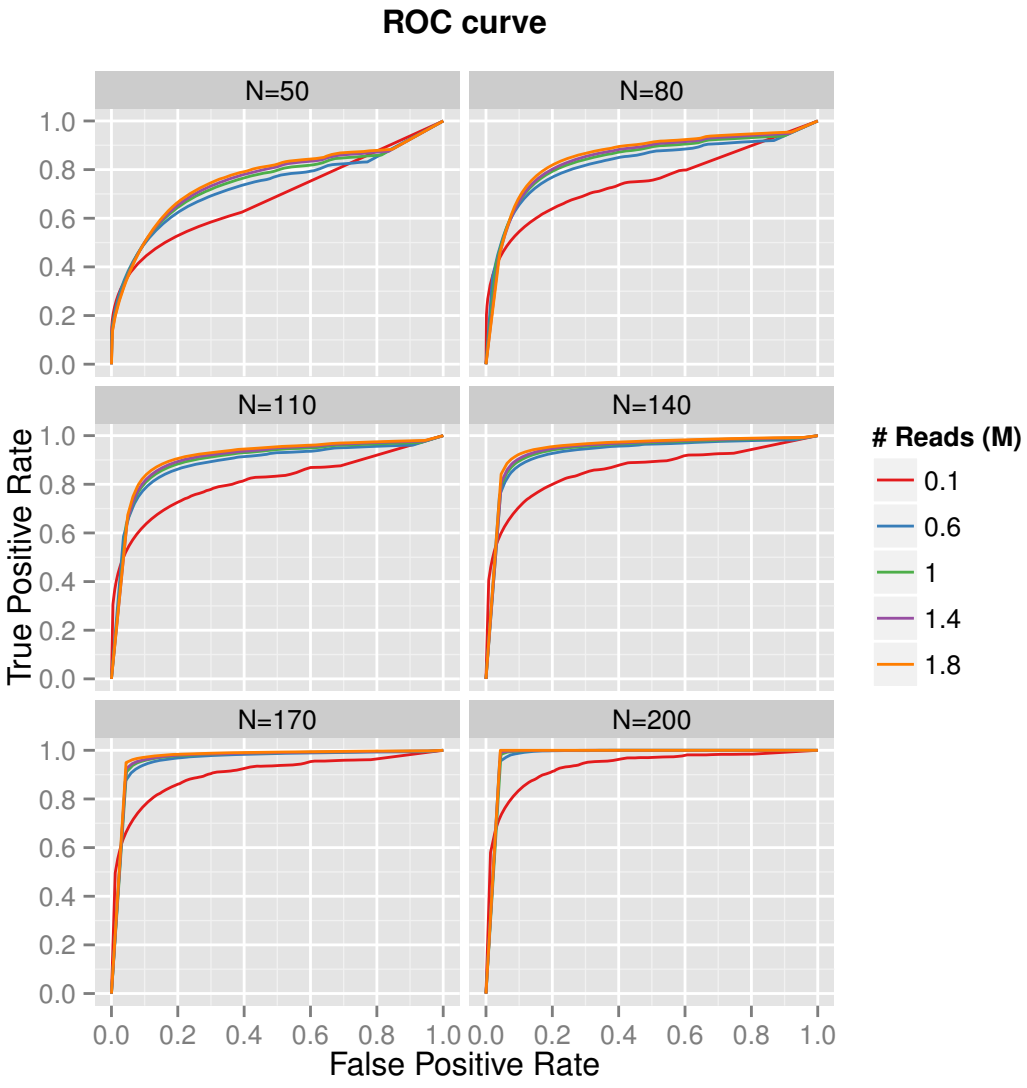

Figure 2.7. ROC curves for each of the considered replicate numbers per group. Lines in each curve represent the sequencing depth. Plots were generated from simulated data generated using distributional parameters extracted from the human prostate scRNA-seq dataset.

inputs from the user may include the desired statistical power, budget constraint, false-discovery rate (FDR) to be controlled, and anticipated cost parameters.

Gene-specific parameters are estimated from the user-provided pilot data, and statistical power is calculated for each experimental design. Statistical power is ad-

dressed in two ways. First, an empirical power calculation is performed by simulating datasets for each design and recording the observed levels of statistical power obtained. Second, a theoretical method implements the power-estimation procedure of Bi and Liu [2016]. Both statistical power calculation methods are described in Section 2.4.1 in greater detail. Finally, the cost of each experimental design is also projected, based on the formula and default cost assumptions provided in Section 2.4.2.

2.4.1 Statistical Power Calculation

The utility of the interactive tool comes from employing a dataset that is either a subset of or representative of a full experiment, and obtaining estimates of what the statistical power might be to detect differential expression if the researcher were to carry out a full experiment of a specified size. It is therefore important to choose with care the method of experiment-wide statistical power estimation.

Several methods exist for RNA-seq statistical power calculations that are performed on a gene-by-gene basis, with varying assumptions about the distribution of the true underlying expression values. For example, Fang and Cui [2011] propose a formula based on the Wald test for single-gene differential expression analysis, while treating the data as Poisson. Hart et al. [2013] treat the data as negative binomial and derive a formula based on a score test, highlighting the relationship between technical and biological variability, and using empirical justifications for how to choose certain parameters of the formula. Busby et al. [2013] uses a non-central t-distribution to approximate the statistical power of an experiment, arguing that a normal approximation is reasonable for RNA-seq data; however, such justifications are generally not accepted in the general literature, as RNA-seq count data are often known to have distributions too skewed to be modeled as normal.

What is of most interest, however, is not merely the statistical power of a single gene, but of experiment-wide power over the tens of thousands of genes measured in an RNA-seq experiment. Single gene statistical power calculations are often accompanied

by suggestions for how to pool per-gene powers over an experiment; typically this involves taking the average, with or without allowing parameters to vary between genes. However, in situations involving many simultaneous tests as with RNA-seq, it is necessary to account for this multiple testing using error criterion such as the false discovery rate (FDR) [Benjamini and Hochberg, 1995].

One procedure for calculating experiment-wide statistical power while controlling for FDR is proposed in Li et al. [2013a], consisting of a single gene formula for computing statistical power based on several test statistics, and an extension of that formula to incorporate FDR control. However, the procedure is based on a Poisson distribution, which is inappropriate for the overdispersion present in RNA-seq experiments involving many biological replicates. The authors try to address this in Li et al. [2013b] by assuming a negative binomial distribution for the expression counts, and using a statistical power calculation based on the exact test as used in `edgeR` for testing differential expression between two groups. However, several features of the procedure render it extremely conservative; for example, statistical power is computed by setting the fold change parameter to be the minimum fold change observed across all genes deemed differentially expressed, and similarly setting the dispersion to the maximum observed. This likely limits the practicality of the procedure.

It is evident that experiment-wide statistical power considerations for RNA-seq data while controlling FDR is underdeveloped. Reflecting on the microarray literature, Liu and Hwang [2007] calculate statistical power at a specified FDR level by finding the rejection region for the test procedure. The authors use t-tests to model microarray data. The recent Bi and Liu [2016] takes the machinery of Liu and Hwang [2007] and makes it applicable to RNA-seq data by applying the `voom` method of the `limma` package to first transform the count data into normalized log-counts. This circumvents the direct use of the negative binomial distribution, for which there exist no analytical relationships between statistical power and sample size, as there are no closed-form solutions for the maximum likelihood estimate of the NB dispersion. Due to the applicability of Bi and Liu [2016] to RNA-seq data, its control of FDR to

account for multiple testing, and its avoidance of computationally-heavy simulations, this procedure is implemented for the theoretical statistical power calculation in our experimental design tool. The details of the procedure are as follows, as originally described in Liu and Hwang [2007] and Bi and Liu [2016].

Theoretical Calculation of Statistical Power

Let H_0 and H_1 be indicators that the null or the alternative hypothesis is true, respectively; let Γ be the rejection region of a given test statistic T; and let π_0 be the assumed proportion of true nulls. Table 2.2, originally shown in Benjamini and Hochberg [1995], is popularly used to categorize the different outcomes of testing G hypotheses.

Table 2.2
Table of outcomes when testing G simultaneous hypotheses, π_0 of which are true nulls.

	Declared non-significant	Declared significant	Total
H_0 is true	U	V	$\pi_0 \cdot G$
H_1 is true	T	S	$(1-\pi_0) \cdot G$
Total	$G-R$	R	G

The false discovery rate (FDR) is defined as the expected proportion of false positives among the rejected hypotheses Benjamini and Hochberg [1995]. That is,

$$FDR = E\left(\frac{V}{R}\middle| R>0\right) P(R>0). \tag{2.2}$$

Storey [2003] offered a slight modification of this to the *positive* false discovery rate, defined as

$$pFDR = E\left(\frac{V}{R}\middle| R>0\right). \tag{2.3}$$

Since it may safely be assumed in genomic studies that there will be at least one rejection, *i.e.* that $R > 0$, pFDR and FDR will be used interchangeably here. By Bayes rule, (2.3) can be written[1]as

$$P(H_0|T \in \Gamma) = \frac{P(T \in \Gamma|H_0) \cdot \pi_0}{P(T \in \Gamma|H_0) \cdot \pi_0 + P(T \in \Gamma|H_1) \cdot (1 - \pi_0)} \tag{2.4}$$

In order to control FDR at a given level α, setting equation (2.4) to be less than or equal to α yields the following relationship with some simple algebra.

$$\frac{\alpha}{1-\alpha}\frac{1-\pi_0}{\pi_0} \geq \frac{P(T \in \Gamma|H_0)}{P(T \in \Gamma|H_1)} \tag{2.5}$$

On the right-hand side, Type I error is in the numerator and statistical power is in the denominator. The task is to find the rejection region Γ so that equation (2.5) is satisfied, hence controlling FDR at level α; statistical power may be computed once the rejection region is known.

The original application of Liu and Hwang [2007] was intended for microarrays, in which the data were appropriately assumed to be normal and t-tests could be applied for two-sample comparisons. However, the method is not directly applicable to commonly applied tests for RNA-seq data involving a negative binomial distribution, as there are no closed-form solutions for calculating $P(T \in \Gamma|H_0)$ and $P(T \in \Gamma|H_1)$. As mentioned earlier, Bi and Liu [2016] extended the method to be used for RNA-seq data by first transforming the data to a normalized log-counts per million (log-cpm) value, as part of the method called `voom` implemented in the `R` package `limma` (Linear

[1]The Bayesian interpretation of the pFDR (2.3) is detailed and proven in Theorem 1 of Storey [2003]. Briefly, given G identical tests of the null hypothesis H_0 with accompanying test statistics $T_1, ...T_G$ and a given rejection region Γ, $pFDR$ may be rewritten as

$$pFDR = E\left(\frac{V(\Gamma)}{R(\Gamma)}\middle| R(\Gamma)\right),$$

where $V(\Gamma) = \#\{\text{null } T_i|T_i \in \Gamma\}$ and $R(\Gamma) = \#\{T_i|T_i \in \Gamma\}$. $P(H_0|T \in \Gamma)$ represents the probability of a false positive, given a significant test statistic. For the case when $G = 1$, $V(\Gamma)/R(\Gamma)$ must be either 0 or 1, so it easily follows that $pFDR = P(H_0|T \in \Gamma)$. Storey [2003] show, with proof, that this result is the same for when $G > 1$.

Models for Microarray Data). This approach allows for the derivation of a t-test based test statistic formula that can subsequently be used in the application of the original method as before.

In a two-sample comparison, where the interest is to find differentially expressed genes between two experimental groups, the hypothesis to test for each gene g is

$$H_0^g : \mu_{g1} = \mu_{g2} \tag{2.6}$$

$$H_1^g : \mu_{g1} \neq \mu_{g2}, \tag{2.7}$$

where μ_{g1} and μ_{g2} are means of the normalized counts in each group. The t-test statistic for gene g is

$$T_g = \frac{\Delta_g}{s_g\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{2.8}$$

where Δ_g is the scaled effect size, defined as the weighted mean difference of log-cpm values between groups, and s_g is the pooled standard deviation. To accommodate the practical situation where genes may exhibit different parameters, assume that the effect size Δ_g for each gene follows a normal distribution

$$\Delta_g \sim N(\mu_\Delta, \sigma_\Delta^2), \text{ denoted by } \pi_1(\Delta_g), \tag{2.9}$$

and the variance of log-cpm values follows an inverse gamma distribution

$$\sigma_g^2 \sim InvGamma(a, b), \text{ denoted by } \pi_2(\sigma_g). \tag{2.10}$$

The average statistical power across all genes may be written as an integral over these distributions,

$$P(T \in \Gamma | H_1) = \iint P(T \in \Gamma | H_1, \Delta_g, \sigma_g)\pi_1(\Delta_g)\pi_2(\sigma_g)d\Delta_g d\sigma_g. \tag{2.11}$$

Putting (2.11) into equation (2.5), it follows that FDR is controlled at level α when

$$\begin{aligned}\frac{\alpha}{1-\alpha}\frac{1-\pi_0}{\pi_0} &\geq \frac{P(T\in\Gamma|H_0)}{P(T\in\Gamma|H_1)} \\ &= \frac{P(T\in\Gamma|H_0)}{\iint P(T\in\Gamma|H_1,\Delta_g,\sigma_g)\pi_1(\Delta_g)\pi_2(\sigma_g)d\Delta_g d\sigma_g} \\ &= \frac{P(|T_g|>c|H_0)}{\iint P(|T_g|>c|H_1,\Delta_g,\sigma_g)\pi_1(\Delta_g)\pi_2(\sigma_g)d\Delta_g d\sigma_g}\end{aligned} \tag{2.12}$$

Using the knowledge that T_g is distributed as a central t-distribution under the null and a non-central t-distribution under H_1, the denominator in (2.12) is

$$\begin{aligned}1 - &\iint T_{n_1+n_2-2}(c|\theta_g)\pi_1(\Delta_g)\pi_2(\sigma_g)d\Delta_g d\sigma_g \\ &+ \iint T_{n_1+n_2-2}(-c|\theta_g)\pi_1(\Delta_g)\pi_2(\sigma_g)d\Delta_g d\sigma_g,\end{aligned} \tag{2.13}$$

where θ_g is the non-centrality parameter defined as

$$\theta_g = \frac{\Delta_g}{\sigma_g\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \tag{2.14}$$

and the numerator in (2.12) equals

$$P(T\in\Gamma|H_0) = P(|T_g|>c|H_0) = 2\cdot T_{n_1+n_2-2}(-c). \tag{2.15}$$

Once the critical value c has been obtained which satisfies the relationship in (2.12) for a given level α and proportion of nulls π_0, statistical power may be calculated from equation (2.13) for a specific sample size.

The practical implementation of this method based on pilot data is achieved by first simulating a scRNA-seq count dataset as described in Section 2.2, with simulation parameters drawn empirically from the user-submitted pilot data. These counts are normalized to log-cpm by applying `voom`/`limma` as previously described, and are used to estimate the hyperparameters $\mu_\Delta, \sigma_\Delta, a$, and b which characterize $\pi_1(\Delta_g)$ and $\pi_2(\sigma_g)$. These characterized distributions can finally be used to solve the integrals necessary for computing statistical power.

Empirical Calculation of Statistical Power

In order to double-check how reasonable the theoretical statistical power calculation is, `scDesign` also implements an empirical power calculation based fully on simulations. This is done by first extrapolating the given pilot data to each desired experimental design setting. That is, parameters drawn from the pilot data are used to simulate a new dataset with the desired sequencing depth and number of replicates per group and containing a known set of differentially expressed genes. The `R` package `edgeR` is then applied to test for differential expression, and the resulting adjusted p-values are used to obtain the statistical power. Specifically, statistical power is calculated as the number of genes determined significant at a specified FDR that are truly DE (true positives), divided by the total number of truly DE genes in the simulated dataset (positives). This is repeated a number of times, and the average of statistical powers in each iteration is taken to be the empirical calculation of statistical power for the given experimental design. Figure 2.8 shows that for a given depth, the theoretical and empirical estimates of statistical power are similar, with empirical calculations being slightly more conservative. While a statistical power of 0.8 is a typical standard target for experiments, such levels of power are harder to achieve for scRNA-seq data, which are often zero-inflated with higher variability among replicates even of the same treatment group.

A few remarks bear noting. First, the empirical statistical power calculation inevitably reflects the power of the method chosen to test differential expression, in this case `edgeR`. Other methods, for example `DESeq2`, `SCDE`, or `limma`, among others, will likely yield different power estimates. `edgeR` was chosen for its useability on data simulated to mimic scRNA-seq data; other attempted methods either yielded substantially lower statistical power or were computationally intractable for large numbers of replicates. Second, the empirical power calculation is significantly more computationally intensive than the theoretical power calculation, as it involves applying `edgeR` to each experimental design a repeated number of times. Hence, while

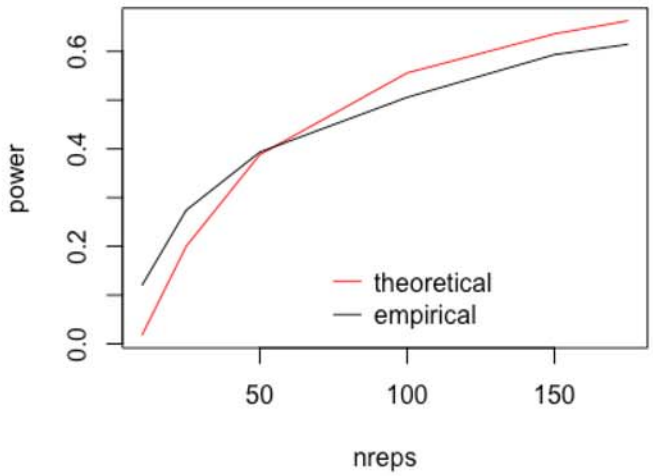

Figure 2.8. At a given depth of two million reads, the theoretical and empirical estimates of statistical power are similar, with empirical calculations being slightly more conservative.

it is given as an option in our experimental design `R` package, the Shiny application will only display results from the theoretical option.

2.4.2 Cost Function

Aside from statistical power, the projected cost of an experimental design is a valuable metric for judging what levels of replication and sequencing depth are appropriate, given the reality of budgetary constraints. Liu et al. [2014] propose a very simple function for the cost of an RNA-seq experiment, consisting of a per-sample cost c for each of N samples, and an overall fixed cost d.

$$\text{cost} = (c \times N) + d \tag{2.16}$$

Per-sample costs include materials for library preparation as well as labor costs. Overall fixed costs to the experiment depend on variables such as sequencing depth, mul-

tiplexing, and equipment. Attolini et al. [2015] incorporate read-specific costs in the following equation:

$$\text{cost} = (c_0 \times N) + (c_1 \times 2rDN), \tag{2.17}$$

where N is the number of samples, each associated with a fixed cost c_0. $2r$ denotes the read length for paired-end experiments, D the number of reads per sample, and c_1 the cost per read.

Building on (2.17) to incorporate costs specific to single-cell experimental designs, the following cost function is proposed. Let N, as before, denote the number of samples, in this case individual cells, with a per-cell cost of c_{cell}. These cells are captured onto plates which hold a default 96 cells at a time, at a per-plate cost of c_{plate}. This may be adjusted to a higher number of cells per plate, which may soon increase to as many as 800 cells per plate, as high-throughput cell capture protocols become more widely available. In addition to cell and plate costs, there are also per-lane costs c_{lane}, where each lane can accommodate a maximum number of reads, max. Finally, there may be other miscellaneous costs to be captured in c_{fixed}.

$$\text{cost} = (c_{cell} \times N) + (c_{plate} \times \lceil N/96 \rceil) + (c_{lane} \times \lceil ND/max \rceil) + c_{fixed} \tag{2.18}$$

A practical adjustment to the cost function is to account for inefficiencies in the capturing of cells and the sequencing of reads. That is, there may be some proportion of cells on the 96-well capture plates that are captured incorrectly or do not pass a viability screen. If $p_{capture}$ denotes the capture efficiency, an input of N cells will result in $N \times p_{capture}$ cells being used in the analysis. In addition, there is typically some proportion of sequenced reads that do not get aligned and are hence not quantified; reasons for this include the filtering of reads that do not pass quality control, ambiguously mapping reads, or reads from artifacts such as rRNA rather than genomic features of interest. Let p_{seq} denote the sequencing efficiency, so that $D \times p_{seq}$ is the

amount of sequenced reads that align successfully and thus quantified. The following cost function incorporates these adjustments.

$$\begin{aligned} \text{cost} = (c_{cell} \times N) + (c_{plate} \times \lceil N/96 \times p_{capture} \rceil) \\ + (c_{lane} \times \lceil ND/max \times p_{sample} \times p_{seq} \rceil + c_{fixed} \end{aligned} \tag{2.19}$$

The costs of cell, plate, and lane may be chosen with real experiments as a guide. Table 2.3 presents some typical costs associated with various stages of the scRNA-seq workflow, from cell capture to sequencing; these numbers are based on the actual costs of the prostate data described in Section 2.2. Per-cell costs may include kits for cDNA dilution and library prep; per-plate costs may cover plate reagents as well as the plate itself; per-lane costs comprise the costs of the sequencing itself, depending on read length and paired- or single-end sequencing; and fixed costs may include items such as assay tubes, viability kits, and labor. The `scDesign` tool allows for the specification of the following default parameters: costs per cell, plate, and lane are respectively \$1200, \$20, and \$2000; fixed costs per experiment are \$1200; and the maximum number of reads per lane is 96 million. Again, these values are based on the observed costs of the prostate dataset in particular, but may vary widely across different cell types, experimental platforms, and sample preparation protocols.

2.4.3 Pilot Data

Researchers often elect to first sequence a handful of replicates at a lower depth to get a sense of their data before committing to an expensive full experiment. The utility of our tool is that it requires only the pilot data to estimate what the statistical power would be in a full imagined experiment; it does so by taking parameters learned from the pilot dataset to extrapolate the data to "full" size, and using the full data as a basis for calculations. A natural question might be what is the effect of the size of the pilot data on the ability of the extrapolations to accurately recover properties of the full dataset.

Table 2.3
Some typical costs associated with various stages of the scRNA-seq workflow, from cell capture to sequencing; these numbers are based on the actual costs of the prostate data described in Section 2.2.

	Item	Cost
Cell Isolation	96-well plates	\$700/ea
	Instrument reagents	\$440/plate
	Viability kit	\$400
Library Prep	Library prep kit	\$12.50/cell
	Other (reagents, tubes)	\$305
Sequencing	HiSeq Rapid PE 100bp	\$1990/lane
	Multiplexing	\$300/two lanes

To investigate this, a full-size dataset was simulated with the approach described in Section 2.2, and relying on parameters taken from the prostate scRNA-seq data. The experimental design of these full data consist of two hundred replicates per group at depths of two million. There are ten thousand genes, one thousand of which exhibit true differential expression. `edgeR` was applied to detect DE genes, with results serving as a baseline for comparison in later analyses of pilot datasets; 647 genes were detected as DE, with a true positive rate (TPR) of 0.625 and false positive rate (FPR) of 0.002.

Pilot datasets were obtained from the full-sized dataset, by down-sampling to a range of smaller experimental designs in a similar fashion as Section 2.2. Specifically, the number of replicates per treatment group considered are $N = \{5, 10, 25, 50, 10, 150, 200\}$, and the depths considered are $D = \{0.1, 0.25, 0.5, 1, 1.5, 2\}$ million reads. 20 datasets were generated for each combination of $N \times D$. To extrapolate each pilot dataset back to full size while keeping any original differential expression patterns, the simulation procedure of Section 2.2 was adapted to allow the estimation of group-

specific means from the pilot data. Figure 2.9 demonstrates that extrapolations made from pilot data of various sizes exhibit variance-mean plots that look reasonably similar to the original full dataset of 200 replicates and depth of 2 million.

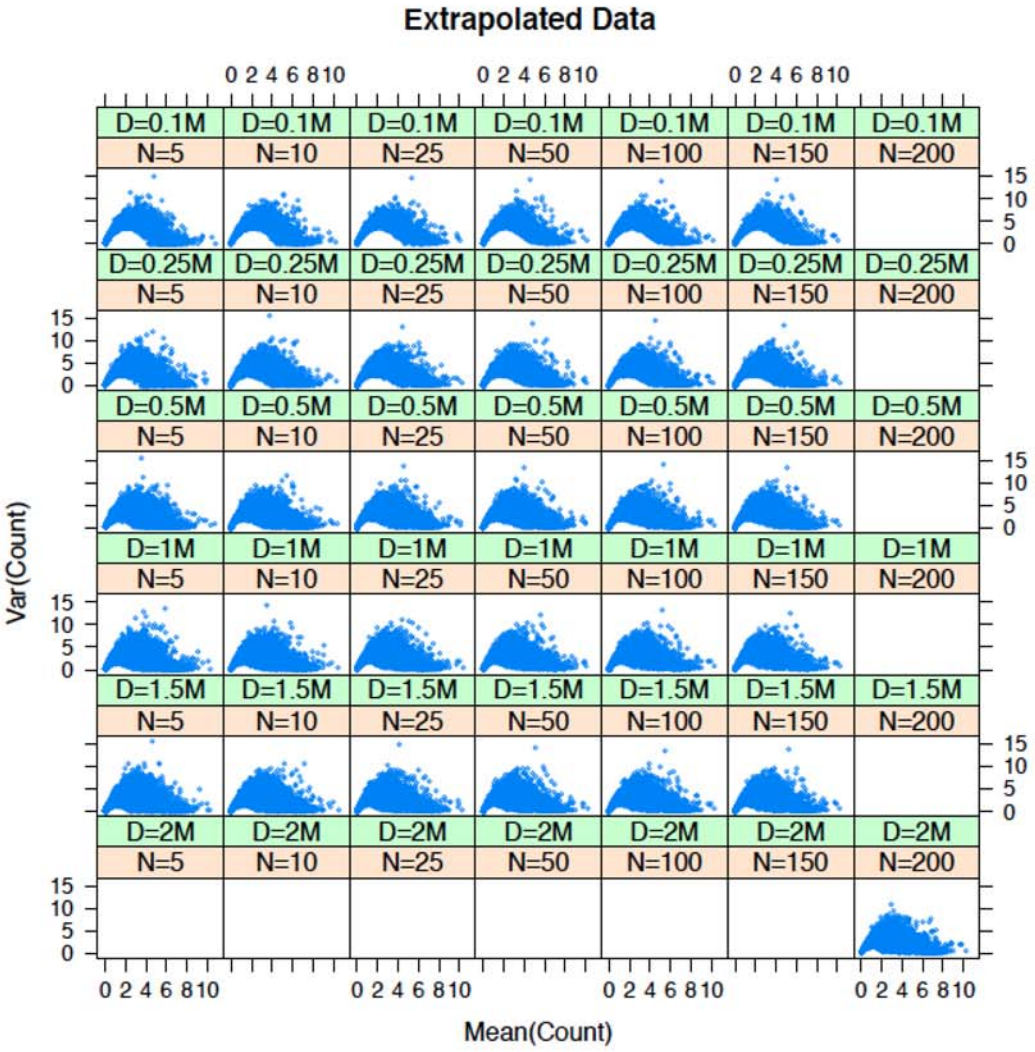

Figure 2.9. Extrapolations made from pilot prostate data of various sizes exhibit variance-mean plots that look reasonably similar to the original full dataset of 200 replicates and depth of 2 million (lower right).

Finally, `edgeR` was applied to each extrapolated dataset obtained from the pilots, and results were examined for their fidelity with results from the full data. The ROC plots of Figure 2.10, one for each replication level with lines representing sequencing

depth, show how accurately extrapolations from pilot datasets of each size recover the true DE genes simulated in the full dataset. As expected, extrapolations perform much better for pilot datasets with higher replication levels, without much difference between depths.

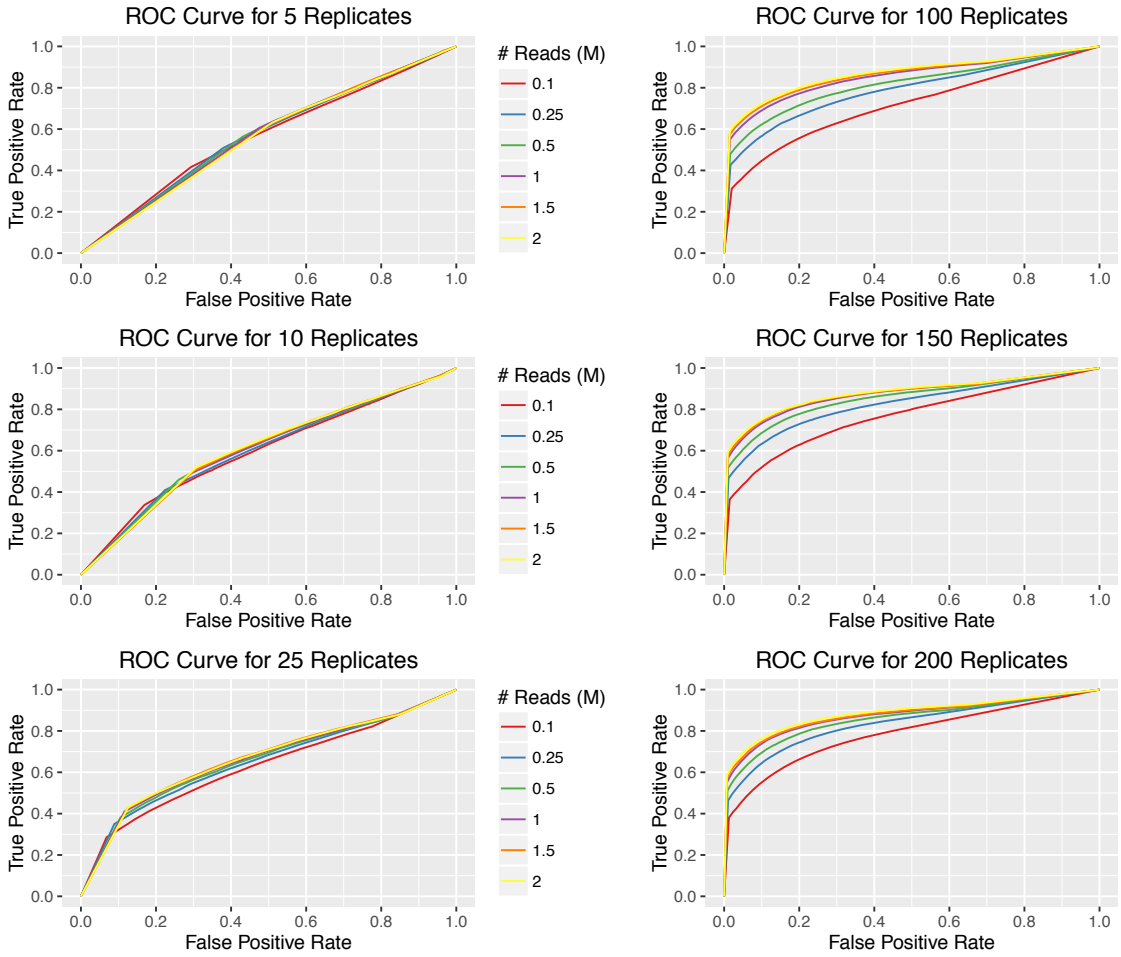

Figure 2.10. ROC plots, one for each replication level with lines representing sequencing depth, show how accurately extrapolations from pilot datasets of each size recover the true DE genes simulated in the full dataset.

Concordance plots shown in Figure 2.11 depict the fraction of matching genes in a list of top k genes, identified in the extrapolated datasets as compared to the full dataset, with k from 1 to 100. Each plot shows results for one depth setting, with

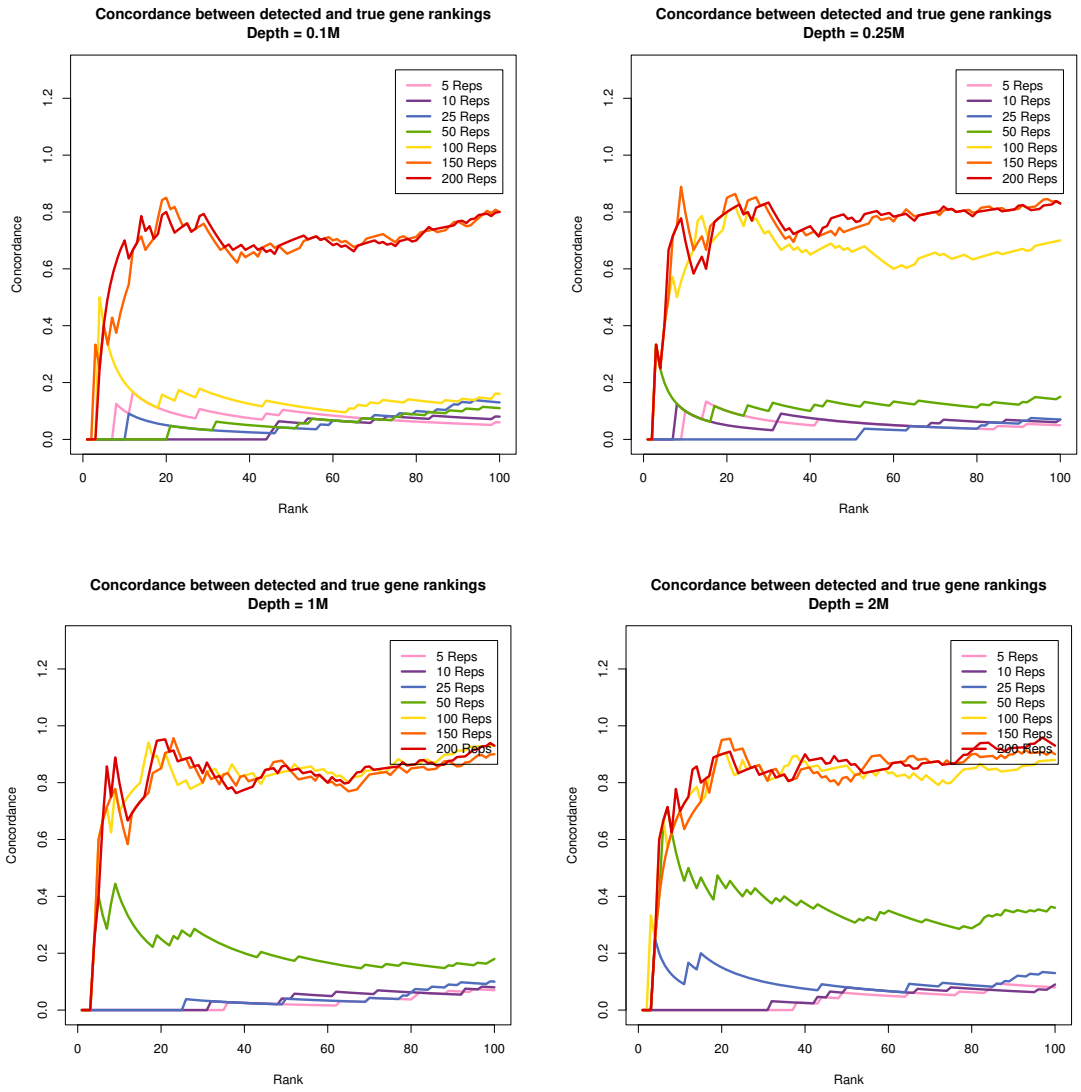

Figure 2.11. Concordance plots depict the fraction of matching genes in a list of top k ranking genes, identified in the extrapolated datasets as compared to the full dataset. Each plot shows results for one depth setting, with lines representing the numbers of replicates per group.

lines representing the numbers of replicates per group. The interpretation of this plot is that the higher the concordance, the better the extrapolated dataset was at reproducing the top gene rankings of the full dataset. Notice the clear separation between the abilities of replication levels above 100 as compared to lower levels,

with respect to recovering the "true" differential expression patterns of the full data; this effect is apparent in both the ROC and concordance plots, but is more striking in the latter. Sequencing depth plays a lesser role in this regard, although it is observed that higher sequencing depths compensate somewhat for lower replication levels. For example, the concordance plots show that datasets with 50 replicates per group may achieve greater concordance with true gene rankings when depths are above one million as compared to lower.

Based on these results, a recommendation can be made that, when submitting pilot data to serve as the basis for extrapolation-based statistical power calculations, researchers should strive to provide data that have a moderately high number of replicates, even if this requires sacrificing sequencing depth. The accuracy of statistical power calculations may be substantially improved by increasing the number of replicates in the pilot data, but only marginally affected by increasing sequencing depth.

3. MODELING DIFFERENTIAL GENE EXPRESSION FROM SCRNA-SEQ DATA

Statistical analyses of scRNA-seq data up to this point have been largely limited to exploratory data analysis tools such as principal component analysis (PCA) and hierarchical clustering. Preliminary investigations into the capabilities of scRNA-seq have tended to favor such methods in large part for offering interpretable visualizations of the patterns and underlying structures among collections of individual cells, without any *a priori* assumptions or expectations. This kind of bottom-up approach, which characterizes exploratory data analysis, is par for the course in the early stages of new technologies. For example, after microarrays were invented in the 1990s, enabling the first gene expression profiling experiments on a genomic scale [Schena et al., 1995, Brown and Botstein, 1999], clustering analyses dominated the results sections of early manuscripts [Quackenbush, 2001]. In similar fashion, PCA and clustering are now regularly applied to single-cell gene expression profiles to, for example, detect how a population of cells may be separated into subpopulations of distinct cell types [Shalek et al., 2013, Dalerba et al., 2011, Islam et al., 2011].

Certainly, exploratory data analysis has an important place in the scientific procedure [Tukey, 1977]. However, it is classically understood that to be truly comprehensive, statistical analyses of experiments must iterate between preliminary investigations involving exploratory tools and confirmatory tests of concrete hypotheses; it is from the latter that scientific conclusions and predictions can be made. For example, a major goal of statistical inference for bulk RNA-seq data over the past decade has been the identification of genes whose levels of expression differ between phenotypes or experimental conditions. The objective of these experiments is to test, for each given gene, whether an observed difference in read counts between groups is statisti-

cally significant in the presence of biological and experimental variation. There are many reasons to expect scRNA-seq studies to rapidly move forward into similarly targeted tests of concrete hypotheses, by way of controlling well-defined conditions for different groups of single cells. For example, drug testing applications are often interested in the response of different cell types to varying drug treatments; cancer researchers may like to investigate the effect of knocking out an oncogene on the rest of the transcriptome; others may want to detect the most important genes that drive the transcriptional changes between cells at separate stages of differentiation. Applications such as these abound in the scRNA-seq literature, but presently remain largely caught in the realm of clustering and hypothesis generation.

3.1 The Stochastic Nature of Gene Expression

Most of the existing scRNA-seq studies that attempt to formally test for differential expression between experimental conditions have simply adopted standard methods developed in the context of bulk cell RNA-seq. Unfortunately, there are caveats to be raised regarding the rote application of bulk tools to single-cell data in a one-size-fits-all fashion. Bulk RNA-seq aggregates gene expression across whole populations of cells (Figure 1.1), obscuring some unique features that only manifest at the cellular level. In particular, gene expression in individual cells is inherently a dynamic process with unknown rates of activity [Elowitz et al., 2002, Raj and van Oudenaarden, 2008, McAdams and Arkin, 1997]. This phenomenon, dubbed 'stochasticity' in scientific jargon, refers to how transcription occurs not uniformly, but often in bursts of individual genes or of coordinated gene networks [Kaufmann and van Oudenaarden, 2007, Marinov et al., 2014, Munsky et al., 2012, Sanchez and Golding, 2013, Wills et al., 2013]. Rather than exhibit constant gene expression, levels of mRNA molecules monitored in real time have been found to fluctuate as if the genes themselves were randomly and unpredictably switching back and forth between active and inactive states [Golding et al., 2005]. Single cells that are captured for sequenc-

ing are invariably a snapshot of these stochastic fluctuations. As a result, scRNA-seq data are frequently seen to exhibit genes that may have moderate to strong expression in some cells, but very little to no expression in other cells (Figure 3.1). This feature has been previously referred to as 'dropout events' [Kharchenko et al., 2014] or 'bimodal expression' [Shalek et al., 2013].

In addition to gene expression stochasticity driving observed bimodality in the data, there is also a technical component contributing to the effect. Single cells offer tens of thousands of times less input RNA than bulk tissue samples, often in the range of picograms rather than the usual nanograms or micrograms. These minimal amounts of starting RNA can lead to transcripts either being missed in the reverse-transcription stages of sample preparation, or not being present in sufficient quantity to be detected by the sequencing machine. In addition, extremely low levels of input material result in samples being much more likely to degrade or to be perturbed by any of the many stages of the experimental process, from sample preparation to sequencing. Altogether, the effect of both biological stochasticity and of technical challenges is that many genes, while potentially expressed in truth, do not become represented in the data in expected quantities.

A second important feature that becomes manifest in single cell data is the effect of the cell cycle, defined as a series of steps that define the life span of the cell. These steps are defined by the following phases: the first and longest growth phase (G1) when cells grow larger and increase their production of proteins and ribosomes in preparation for DNA synthesis; the synthesis phase (S) when cells replicate a complete copy of their DNA; the second growth phase (G2) when cells continue to prepare metabolically for mitosis; and finally, mitosis (M) during which active cell division occurs (Figure 3.2). Given the highly regulated and controlled nature of this process, the cell cycle stage is known to affect the transcriptional activity of cells in global, non-trivial ways. The impact this has on observed gene expression in captured cells may pose a substantial confounding effect when testing other factors of interest. There is evidence that the high variability found in mRNA levels of single

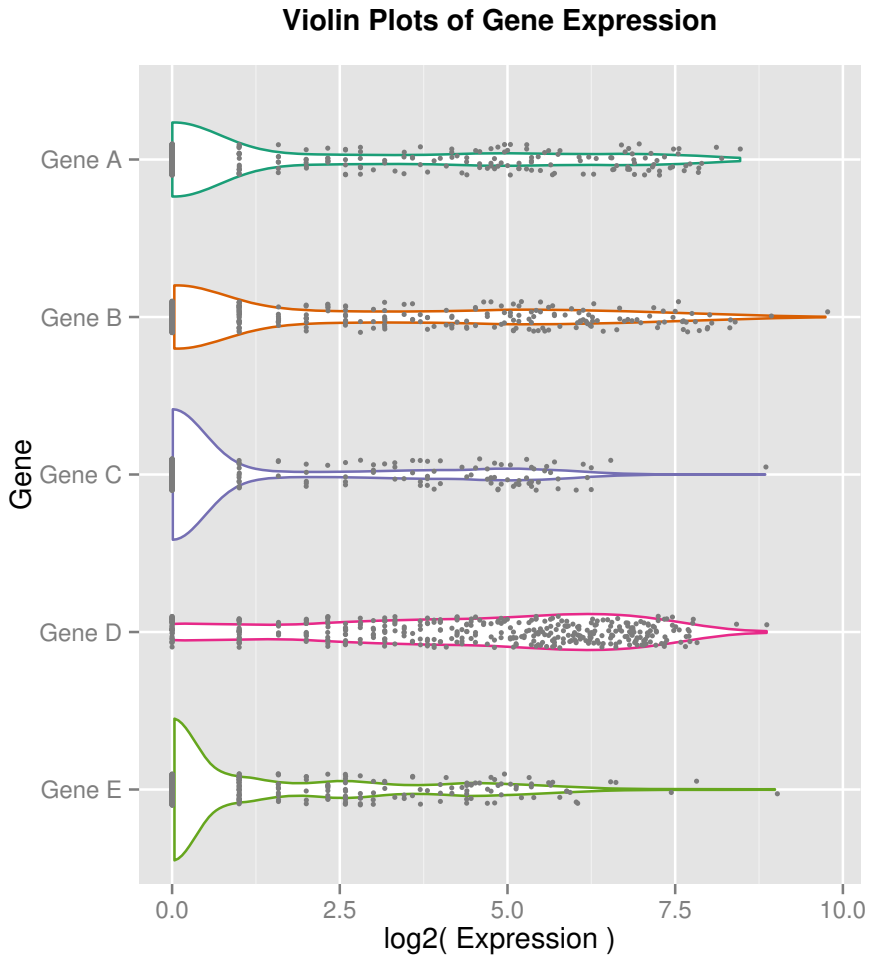

Figure 3.1. Violin plots of log counts for five randomly selected genes across 399 human prostate cells demonstrate the bimodality, or 'dropout events', commonly seen in scRNA-seq data. This bimodality may arise from both biological as well as technical sources. Each point represents a cell's expression value for a given gene, with a vertical jitter added for visual clarity. The lines display a smoothed kernel density for visualizing the overall distribution of expression values.

yeast cells is driven not only by stochastic bursts of gene expression, but in no small part also by transcriptional differences between phases [Zopf et al., 2013]. Deliberate perturbations of the cell cycle by inhibition of cell cycle regulator proteins have been observed to substantially affect phenotypes such as nuclear and cell morphology [Chen

et al., 2013]. Moreover, cell cycle has been linked to fundamental biological processes such as differentiation [Singh et al., 2013, Pauklin and Vallier, 2013] and oncogenesis [Bar-Joseph et al., 2008, Kastan and Bartek, 2004]. Altogether, cell cycle is a key driver of cell-to-cell heterogeneity.

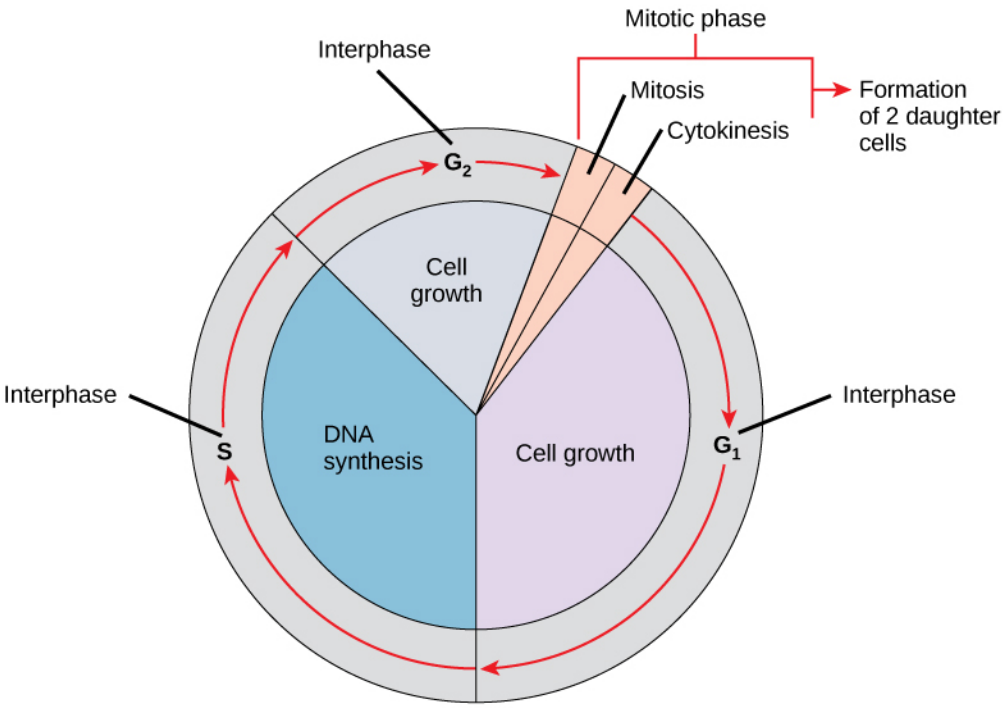

Figure 3.2. The cell cycle is a series of steps that define the life span of the cell, and are divided into the following phases: the first and longest growth phase (G1) when cells grow larger and increase their production of proteins and ribosomes in preparation for DNA synthesis; the synthesis phase (S) when cells replicate a complete copy of their DNA; the second growth phase (G2) when cells continue to prepare metabolically for mitosis; and finally, mitosis (M) during which active cell division occurs.

3.2 Existing Methods for Bulk RNA-Seq Data

The differential expression analysis of tissue-level RNA-seq data has proliferated in the scientific literature, and is now a routine procedure applied to a wide variety of

organisms for the purpose of testing a large range of biological and experimental effects. In large part, the ubiquity of these analyses is owed to the success and useability of several well-established statistical methods for testing differential gene expression; the most notable of these are implemented in the `R` packages `edgeR` [Robinson et al., 2010] and `DESeq` [Anders and Huber, 2010].

The statistical methods that comprise `edgeR` were originally presented in Robinson and Smyth [2008]. The authors model the read counts of each gene across multiple samples using a negative binomial distribution, a departure from previously used Poisson models which fail to account for the increased variance observed with RNA-seq data. The negative binomial dispersion parameter is estimated using a modification to the conditional maximum likelihood (CML) method, based on an adjustment of the data to being of equal library sizes. This adjustment allows the CML machinery to be used for dipersion estimation, and also affords a test for differential expression between experimental conditions using an exact test. Robinson and Smyth [2007] build on this by allowing for both a global estimation of a common dispersion across all genes, as well as a per-gene disperison estimation using conditional weighted likelihood to shrink each individual dispersion towards the global value. McCarthy et al. [2012] extends these negative binomial methods for differential expression to generalized linear models (GLMs), allowing for the testing of more complex experimental designs.

Anders and Huber [2010], the authors of `DESeq`, also model the counts of each gene via a negative binomial distribution; however, they take a different approach to estimating the dispersion parameter. Rather than assuming the typical variance-mean relationship of the negative binomial distribution, as $\sigma^2 = \mu + \phi\mu^2$ where ϕ is the dispersion parameter (as done in `edgeR`), the authors take a data-driven approach to link the variance and mean. Specifically, the variance and mean are instead linked by a smooth local regression. Differential expression testing proceeds via an exact test, similar to that of Robinson and Smyth [2008]. The methods are updated in Love

et al. [2014] to include shrinkage estimation using empirical Bayes priors as well as a GLM framework, and implemented in the `R` package `DESeq2`.

3.3 Accounting for Bimodality

Regardless of the source of bimodality in single-cell gene expression measurements, whether technical or biological, it is important to account for this structure of the data when testing for differential expression. Previously, Kharchenko et al. [2014] developed a three-component mixture model to describe the dropout prevalence of scRNA-seq data, and then employed a Bayesian approach to test for differential expression. However, this method is limited to testing between two groups, restricting its applicability to more complex designs or experiments with multiple conditions. McDavid et al. [2013] accounted for bimodality also by using a mixture model, this time in a GLM framework, but the model was developed for the continuous measurements generated from single-cell quantitative PCR expression measurements, and do not extend to the count data of scRNA-seq.

The approach proposed here is to employ a zero-inflated negative binomial (ZINB) model to account for the salient features of scRNA-seq data while testing for differential gene expression. In particular, the zero-inflated component attempts to capture the bimodality of scRNA-seq data which often manifests as a prevalence of excess zeros. The remaining observations are modeled as a separate component using the negative binomial distribution, appropriate for RNA-seq count data. Finally, the GLM testing framework allows for the testing of covariates and hence can accommodate complex experimental designs.

As depicted in Table 1.1, RNA-seq data are represented as a matrix of read counts, and are frequently modeled using a negative binomial (NB) distribution. The NB model appropriately accommodates the overdispersion typically seen in count data from biological applications, in which the observed variation is greater than would be

expected under a more restrictive Poisson model. One common parametrization of the NB probability mass function is

$$f(y;\mu,\phi) = P(Y=y) = \frac{\Gamma(y+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y+1)}\left(\frac{1}{1+\mu\phi}\right)^{\phi^{-1}}\left(\frac{\mu}{\phi^{-1}+\mu}\right)^{y}, \tag{3.1}$$

where ϕ is the dispersion parameter, $E[Y] = \mu$, and $\text{Var}[Y] = \mu + \phi\mu^2$. Notice that when $\phi = 0$, this reduces to the Poisson distribution. In the specific context of RNA-seq, this distribution is used to model the counts y_{gi} of a single gene g across n samples as

$$y_{gi} \sim NB(\mu_{gi} = m_i\lambda_{gi}, \phi_g),\ i = 1, ..., n. \tag{3.2}$$

The mean parameter μ_{gi} is a product of m_i, the total number of reads mapped to sample i (the library size), and λ_{gi}, the fraction of all reads from sample i that originate from gene g.

In experimental contexts with non-normally distributed response data, it is useful to use generalized linear models (GLMs) to model the dependence of the observed data on a vector of covariates [Nelder and Wedderburn, 1972, McCullagh and Nelder, 1989]. This dependence is described via the log-linear model

$$\log(\mu_{gi}) = \boldsymbol{x}_i^T\beta_g + \log(m_i), \tag{3.3}$$

where $\boldsymbol{x}_i$ is the covariate vector for the i^{th} sample, β_g is the vector of regression coefficients for gene g, and the library size m_i is used as an offset.

The previously described bimodality seen in scRNA-seq data manifests itself as an excess of zeros which cannot be fully explained by the NB distribution alone, even with the sophisticated dispersion estimation methods of `edgeR` and `DESeq`. The proposed alternative is to employ a zero-inflated count model [Lambert, 1992], which specifies two components which may give rise to zero observations: a point mass of zeros arising with probability π, and a count distribution $f_c(y)$ by which zeros may also be observed. $f_c(y)$ is NB for these purposes but in general may be any

count distribution such as Poisson or binomial. The zero-inflated count model with regressors may be written as a mixture model of the form

$$f(y; x, \beta) = \pi \cdot I_{\{0\}}(y) + (1 - \pi) \cdot f_c(y; x, \beta), \tag{3.4}$$

where $I_{\{0\}}(y)$ is an indicator of a zero observation. The corresponding mean of these observations is

$$\begin{aligned} E[f(y; x, \beta)] &= \pi \cdot 0 + (1 - \pi) \cdot E[f_c(y; x, \beta)] \\ &= (1 - \pi) \cdot \exp(x^T \beta), \end{aligned} \tag{3.5}$$

using the canonical log-link of the count component GLM. Model (3.4) may equivalently be written in the following form, which more clearly demonstrates the two possible sources of zeros:

$$f(y; x, \beta) = \begin{cases} \pi + (1 - \pi) \cdot f_c(0; x, \beta) & \text{if } y = 0 \\ (1 - \pi) \cdot f_c(y; x, \beta) & \text{if } y > 0. \end{cases}$$

The probability of the zero component π may be either set as a constant, or modeled with its own GLM $g(\pi) = w^T \gamma$, most often with a logit link. The covariates w for the zero component are not necessarily distinct from the covariates β of the count component, and in the simplest case consists of only an intercept.

Note that the zero-inflated count model is not the same as the hurdle model originally proposed by Mullahy [1986], though they may appear similar at first glance. The hurdle model also consists of two components, but in this case, the zero component is used to model all the zeros, leaving a truncated count component to describe the rest of the positive observations. Thus, the distinction lies in the interpretation of how zeros are generated. While the zero-inflated model allows for a both a 'structural' source as well as a "sampling" source of zeros, the hurdle model assumes all the zero observations to originate from a 'structural' source and that the remaining 'sampling' process is strictly positive. The nature of scRNA-seq data favors the interpretation of the zero-inflated model; hence, this is the method of choice.

To set up the log-likelihood function for gene g, first define z_{gi} as an indicator of a zero count for sample i. That is,

$$z_{gi} = \begin{cases} 1 & \text{if } y_{gi} = 0 \\ 0 & \text{if } y_{gi} > 0. \end{cases}$$

The likelihood of a single observation y_{gi} may be written

$$f(y_{gi}) = [\pi_g + (1 - \pi_g) f_c(0)]^{(1-z_{gi})} \times [(1 - \pi_g) f_c(y_{gi})]^{z_{gi}} . \tag{3.6}$$

Given regression parameters γ_g and β_g for the zero and count components, respectively, such that $\pi = \pi(\boldsymbol{x}_i, \boldsymbol{\gamma}_g)$ and $f_c(y_{gi}) = f_c(y_{gi}; \boldsymbol{x}_i, \boldsymbol{\beta}_g)$, the log-likelihood is

$$\begin{aligned} \mathcal{L}(\boldsymbol{\gamma}_g, \boldsymbol{\beta}_g) = &\sum_{i=1}^{n} (1 - z_{gi}) \cdot \log \left[\pi_g(\boldsymbol{x}_i, \boldsymbol{\gamma}_g) + (1 - \pi(\boldsymbol{x}_i, \boldsymbol{\gamma}_g)) \, f_2(0; \boldsymbol{x}_i, \boldsymbol{\beta}_g) \right] \\ &+ \sum_{i=1}^{n} z_{gi} \cdot \log \left[(1 - \pi(\boldsymbol{x}_i, \boldsymbol{\gamma}_g)) f_2(y_{gi}; \boldsymbol{x}_i, \boldsymbol{\beta}_g) \right] . \end{aligned} \tag{3.7}$$

For the purposes presented here of testing differential gene expression, the zero component $\pi(\boldsymbol{x}_i, \boldsymbol{\gamma}_g)$ is modeled as intercept-only, and the count component $f_c(y_{gi}; \boldsymbol{x}_i, \boldsymbol{\beta}_g)$ is taken to be a negative binomial regression model with single covariate β_g corresponding to the treatment group. For a particular gene g, the hypothesis of interest is

$$\begin{aligned} H_0 &: \beta_g = 0 \\ H_a &: \beta_g \neq 0. \end{aligned} \tag{3.8}$$

All parameters are estimated by maximum likelihood using numerical optimization methods such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [Broyden, 1970]. Significance tests for each parameter may be carried out using the Wald test, or for nested models using a likelihood ratio test [McCullagh and Nelder, 1989]. Given the tens of thousands of individual gene tests for a complete gene expression dataset, false discovery rate (FDR) is controlled using the Benjamini Hochberg (BH) procedure [Benjamini and Hochberg, 1995].

3.4 Accounting for Unmeasured Cell Cycle Effects

One major challenge in characterizing the effect of cell cycle on observed gene expression is that of measuring the cell cycle stage itself. Experimental approaches to accomplish this are varied. One method is to induce cell-cycle arrest, either by depleting factors driving progression between stages, by chemical treatments, or by inhibiting key pathways [Meijer, 1996]. Other techniques include using centrifugation to stratify cells by size (and by proxy, their stage) [Ly et al., 2014], or using flow cytometry to measure DNA content based on retention of a dye [Nunez, 2001]. Major drawbacks to these approaches include the labor intensiveness of the procedures, as well as potential side effects that could disrupt the biological system in undesirable ways. For these reasons, most single-cell gene expression datasets that do not directly aim to study the cell cycle are generally not accompanied by cell cycle stage annotations; that is, it is still uncommon for experimenters to annotate cell cycle stages as a matter of course. Consequently, though it is generally recognized that cell cycle effects exist and may be substantial, the magnitude of cell-cycle distortions to gene expression has not been precisely characterized, nor is it well-understood which genes are affected. The drawbacks of experimental approaches to cell cycle characterization motivate the application of statistical tools that can achieve the same goals.

One apparent approach to inferring cell cycle information is by using unsupervised classification methods to explicitly predict the unobserved cell cycle stage on the basis of the cells transcriptome profile. To evaluate the effectiveness of a range of established methods toward this objective, Scialdone et al. [2015] applied each chosen method to the same scRNA-seq training dataset in which the cell cycle stages were known, and assessed their predictive performance on a variety of labeled test datasets. The methods evaluated include the following: a random forest classifier; logistic regression, both with and without a lasso penalty; support vector machines; PCA-based classification; and a custom algorithm based on the idea of selecting pairs of genes whose relative expression changes signs across cell-cycle stages. Rather than

building the algorithms using the full transcriptome consisting of all the genes, the authors construct a set of cell cycle marker genes whose variation in the training dataset exceed an established threshold of technical noise. The idea is that by using only the expression levels of the selected cell cycle marker genes as training data, the algorithms are built in a way that is informed primarily by the cell cycle rather than other artifacts. The results of Scialdone et al. [2015] could ostensibly be used to inform the choice of how best to assign cell cycle stages to each cell.

3.4.1 Methods Employing Control Genes

Often, however, the cell cycle itself is of no interest and is considered primarily a nuisance factor. In these cases, rather than explicitly assign cell cycle stages to cells, an alternative might be to estimate and then remove the effect of cell cycle variation from the data altogether. This approach was first proposed by Gagnon-Bartsch and Speed [2012] to correct for hidden factors present in microarray data, such as batch effects from sample processing or unwanted biological variation. The correction of these factors is presented as a means to more clearly identify differential expression signatures from a primary factor of interest, such as experimental conditions or biological groups. Crucially, the authors make use of negative control genes, which are defined as genes that are both uninfluenced by the primary factor of interest, while also being positively influenced by the unwanted factor(s). With this definition, it may be assumed that variation observed in the negative control genes is attributed to the unwanted factor, rather than the factor of interest.

Briefly, the method of Gagnon-Bartsch and Speed [2012] involves modeling the expression data Y as

$$Y = X\beta + Z\gamma + W\alpha + \epsilon, \tag{3.9}$$

where X, Z, and W are matrices whose columns represent the factors of interest, the (optional) observed covariates, and the unobserved covariates, respectively. When the estimation is restricted to the set of negative control genes, selected *a priori* on

the basis of being uninfluenced by the factors of interest X, the coefficients for β are by definition equal to zero and the corresponding term goes away. With or without Z, factor analysis is used to produce an estimate $\hat{W}$ for W, which may then be substituted back into the full model in order to estimate coefficients β for X.

The method of Gagnon-Bartsch and Speed [2012] is specifically designed for the purposes of differential expression testing in a microarray regression context. Along with that, the authors strongly recommend against using it naively towards a global adjustment of expression values. This latter goal has been addressed recently in Buettner et al. [2015], who draw on the idea in Gagnon-Bartsch and Speed [2012] of using negative control genes, but with the intent of explicitly recovering a corrected gene expression matrix free of unwanted variation. Most relevant for the purposes here, the method in Buettner et al. [2015] was developed to deal in particular with the presence of confounding cell cycle effects arising from single-cell gene expression data.

Specifically, in the first step of Buettner et al. [2015], the expression profiles of a set of annotated cell cycle genes are used to recover a covariance matrix Σ, which can be said to describe the cell-to-cell variation attributed to the cell cycle. In the second step, a linear mixed model (3.10) is fit to the expression values of each gene, breaking down sources of expression variance attributed to technical noise (δ_g^2), biological variability (v_g^2), and the unwanted factor(s) under consideration (σ_g^2; *i.e.* cell cycle).

$$y_g \sim N(\mu_g, \sigma_g^2 \Sigma_h + v_g^2 + \delta_g^2). \tag{3.10}$$

Under this variance component model, a residual expression dataset with the effect of the unwanted cell cycle factor removed may be obtained by employing the predictive distribution of the cell cycle component with mean $\hat{y_i}$; residual expression values are defined as $y_i^* = y_i - \hat{y_i}$. The suggested use of this corrected gene expression dataset is as an input to existing statistical methods for clustering, dimension reduction, and visualization.

A major caveat exists with the use of negative control genes in the previously described methods. That is, the validity of these strategies is entirely predicated on the negative control genes being both uninfluenced by the primary factor(s) of interest, a well as being indeed influenced by the unwanted factor. If the negative control genes are in fact influenced by the primary factor, their removal will result in an effect of "throwing the baby out with the bathwater"; that is, variation due to the primary factor would be removed along with the unwanted factor. Conversely, if the negative controls do not in fact exhibit the unwanted variation assumed of them, methods to detect that variation break down. In addition, even if the expression variance of a control gene is owed to the unwanted variation under consideration, there is no way to know if it is also influenced by other, non-cell-cycle effects. In practice, both of these conditions are difficult, if not impossible, to confidently verify, rendering these methods hazardous to use.

3.4.2 Surrogate Variable Analysis

Given the shortcomings of the previously discussed approaches and the need to account for cell cycle effects, we settle on a method called surrogate variable analysis (SVA), originally developed in Leek and Storey [2007] for microarray data and updated in Leek [2014] to accommodate RNA-seq count data. SVA identifies and estimates the unwanted effects of all unmeasured confounding factors directly from the data, and subsequently incorporates these "surrogate variables" into expression analyses. SVA differs from Gagnon-Bartsch and Speed [2012] and Buettner et al. [2015] in that it does not attempt to estimate effects from specific unwanted factors such as cell cycle, but from all unmeasured factors that exhibit substantial patterns of variation. Hence, control genes do not need to be specified and the associated challenges of their use are avoided as a result. In addition, the method is able to more flexibly pick up unwanted factors that haven't been considered. SVA is also unlike Buettner et al. [2015] in that it does not try to remove unwanted effects from

the expression data; rather, the estimated effects from unmeasured variables may be employed as covariates in gene-wise models. This allows the explicit quantification of the different effects that detected surrogate variables may have on different genes.

The SVA method may be broken down into three steps. First, the method begins by fitting a simple model containing only the measured variable of interest, given by

$$y_{ij} = \mu_i + f_i(x_j) + e_{ij}, \tag{3.11}$$

where μ_i is the baseline expression level of gene i, $f_i(x_j)$ is a function describing the relationship between the primary variable and the outcome, and e_{ij} is the random error term. In practice, $f_i(x_j)$ may often be taken as $\beta_i x_j$, where β_j is the linear regression parameter for the primary factor of interest x_j. The residual expression matrix R, with values $r_{ij} = y_{ij} - \hat{\mu}_i - \hat{f}_i(x_j)$, represents the variation that is left over after accounting for the primary variable. A singular value decomposition is applied to R to identify signatures of variation due to any unmodeled factors, in the form of singular vectors. By definition, these signatures are independent of the signal due to the primary variable, as they are derived from the residual matrix with the primary effect removed. A permutation test is used to determine which of these singular vectors exhibit significantly more variation than would be expected by chance. These are said to be significant signatures of residual unmodeled variation.

Next, for each significant signature, a list is obtained of genes that are each significantly associated with that signature. These subsets of genes are interpreted to be the drivers of the expression variability arising from that signature. Third, the original expression matrix is subset to the list of genes for each signature under consideration. This reduced expression matrix represents the expression of those genes estimated to contain the signature of expression heterogeneity. Another singular value decomposition on this reduced expression matrix returns an estimate of the surrogate variable for that signature.

Finally, all significant surrogate variables may be included as covariates into downstream regression models in the following manner:

$$y_{ij} = \mu_i + f_i(x_j) + \sum_{k=1}^{K} \lambda_{ki} \hat{h}_{kj} + e^*_{ij}, \tag{3.12}$$

where $\hat{\boldsymbol{h}}$ are the surrogate variables detected, with associated coefficients λ_{ki}.

3.5 Simulations

A simulation study was constructed to demonstrate how SVA may be applied in conjunction with ZINB to account for cell cycle effects and zero-inflation when testing differential gene expression in scRNA-seq data.

Data were generated to mimic scRNA-seq data using the method described in Section 2.2. As before, the distributions from which the count data were drawn are based on parameters sampled empirically from the real human prostate cancer cell dataset. The data were simulated to exhibit effects arising from a group factor of primary interest for testing differential expression, as well as a confounding factor, called the cell cycle factor for the purposes here. Genewise coefficients for both group and cell cycle effects were drawn as $\beta_g^G, \beta_g^C \sim \text{logNormal}(2, 1)$ for DE genes, while coefficients for non-DE genes were set to 0. The dataset of 10,000 genes consisted of 4000 (40%) genes with a non-zero group effect, and 4000 genes with a non-zero cell cycle effect. Each cell was randomly assigned to one of two levels of the group factor and one of three levels of the cell cycle factor. Groups and cell cycles were assigned such that these variables exhibited a specified amount of correlation, specifically $\rho = \{0, 0.25, 0.5, 0.8\}$. Datasets with $\rho = 0$, for example, exhibit independent assignment of group and cell cycle levels to samples, whereas a high correlation of $\rho = 0.8$ indicate greater challenges in differentiating between group and cell cycle effects in the testing stage. The number of replicates per treatment group considered are $N = \{5, 10, 25, 50, 100, 200\}$. Five datasets were generated for each combination of $N \times \rho$. From each dataset, genes whose average count is less than five were filtered out.

SVA was applied to each dataset to produce estimates for the cell cycle of each sample. Figure 3.3 depicts the SVA estimates of cell cycle across correlation settings, for selected replication levels of 25 and 200 replicates per group. When the cell cycle and group variables are uncorrelated, SVA estimates are able to clearly separate the true cell cycle levels. This becomes more difficult as the correlation between variables increase, although the estimates still remain fairly accurate. Even for relatively small sample sizes such as 25 replicates per group and for high correlations between variables, SVA performs relatively well.

To test for differential expression between levels of the primary group factor, `edgeR` and ZINB were applied to each dataset, both with and without SVA estimates of cell cycle incorporated into the design matrix. The methods of `DESeq2` and `SCDE` were also attempted, but these methods were found to be so computationally intensive as to be practically intractable for larger replicate sizes; hence, these comparisons were left out of the analysis.

3.5.1 Simulation Results

Figure 3.4 depicts ROC plots comparing the performance of `edgeR` and ZINB, with and without SVA adjustment. For higher levels of replication, 100 and above in the simulations, the methods are virtually indiscernible; however, for lower levels of replication, both variations of ZINB outperform both variations of `edgeR`. This suggests an advantage that ZINB has over `edgeR`, in that it explicitly accounts for the high prevalence of zeros that are characteristic of scRNA-seq data. As mentioned previously, the genes in these simulated data exhibit the same proportion of zeros as the real scRNA-seq dataset on which its distributional parameters were based. More replicates ostensibly compensate for zero-inflation, thus erasing the advantage from ZINB for larger sample sizes. With regards to the SVA adjustment, as correlations between the group and cell cycle variables increase, the SVA versions of both methods outperform their non-SVA counterparts. Higher correlations between group and cell

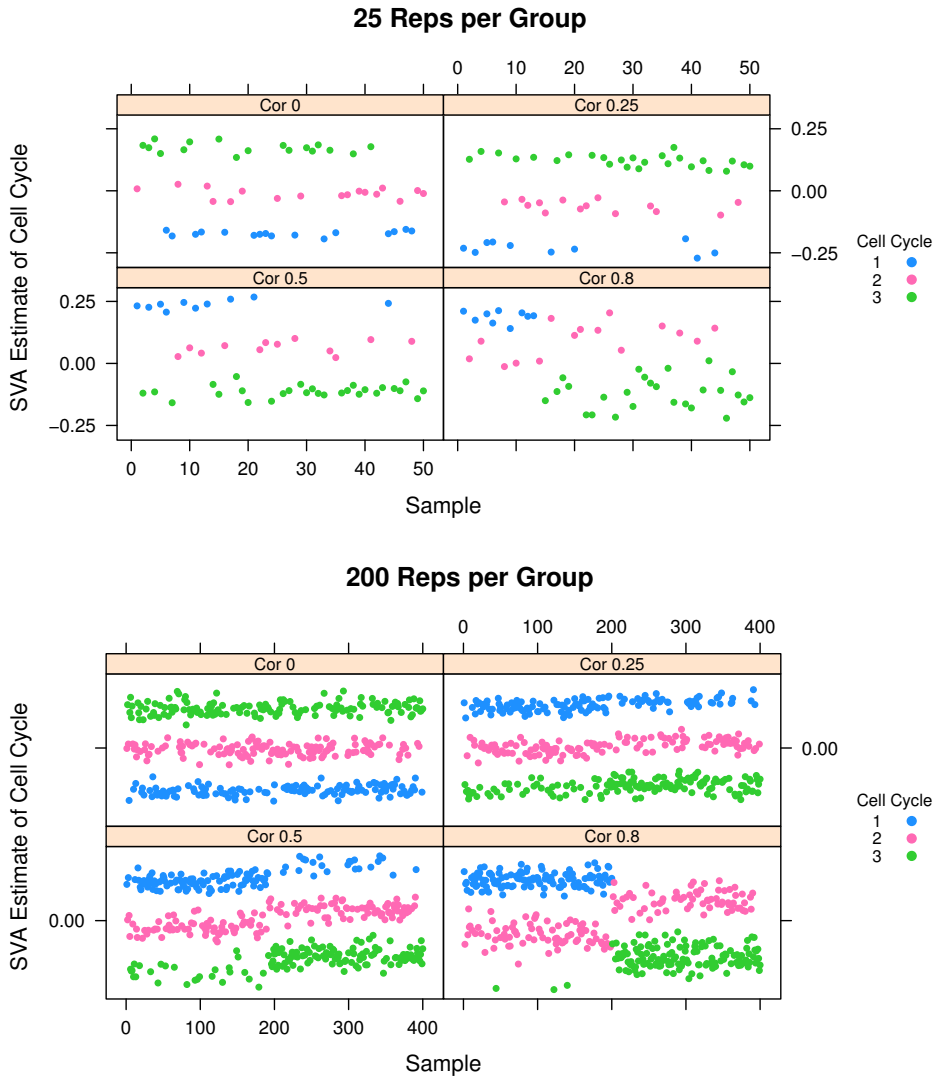

Figure 3.3. SVA estimates of cell cycle across correlation settings, for selected replication levels of 25 and 200 replicates per group. Each point represents the SVA estimate of a cell, and is colored by the true cell cycle.

cycles imply greater confounding between the effects of these variables. In turn, this leads to greater challenges in detecting differential expression with respect to the group factor of primary interest. Incorporating the SVA estimates into GLM-based methods such as `edgeR` and ZINB offers a way to adjust for these confounding effects in a way that improves the statistical power of the primary analysis. The concordance

plots of Figure 3.5 show results that are consistent with the ROC plots of Figure 3.4. That is, the advantages of ZINB are most clearly seen in lower replicate numbers, and the advantages of SVA are more apparent with high levels of correlation between group and cell cycle variables.

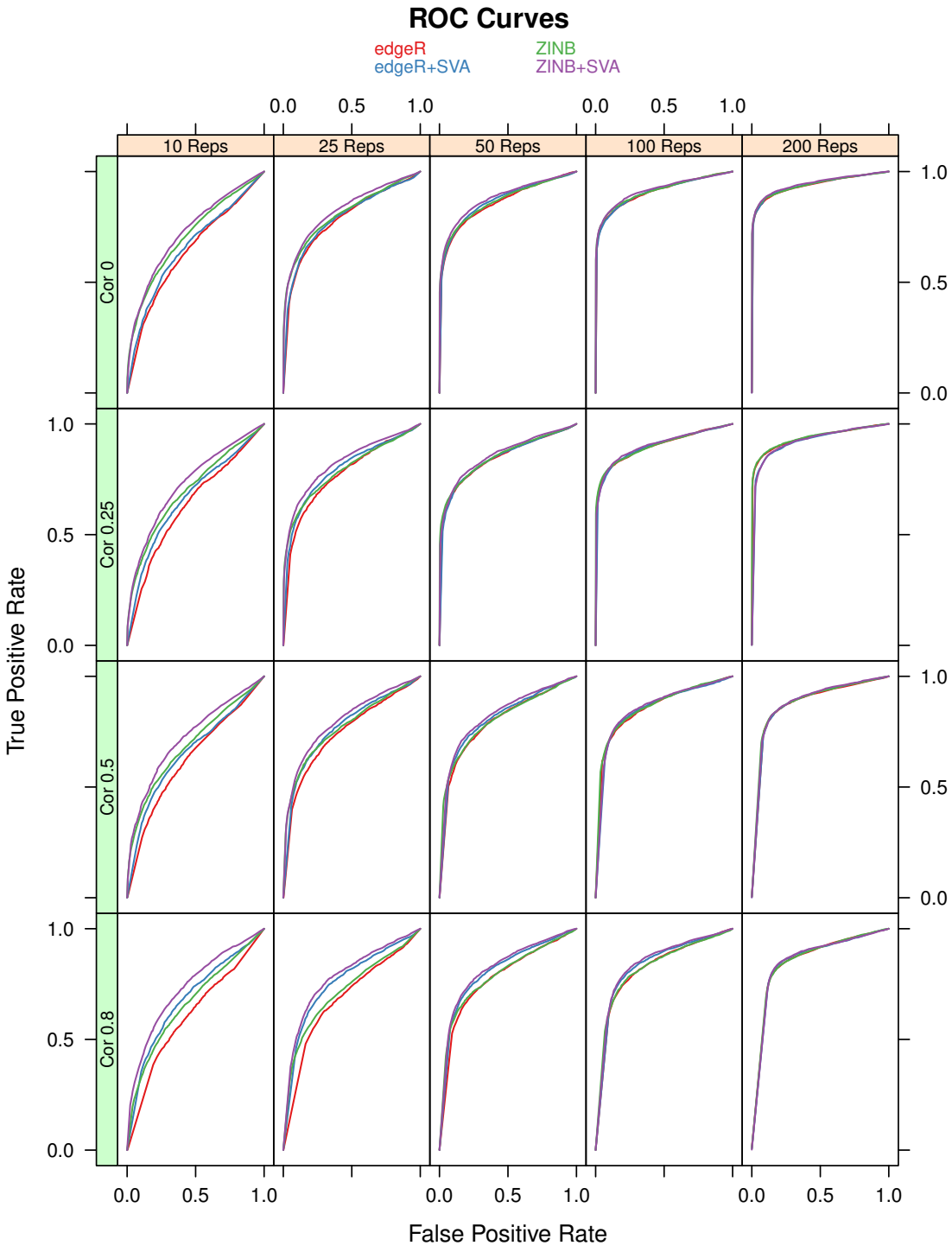

Figure 3.4. ROC curves that compare ZINB and `edgeR`, both with and without SVA adjustment, for each combination of replicates per group and level of correlation between the group and cell cycle variable.

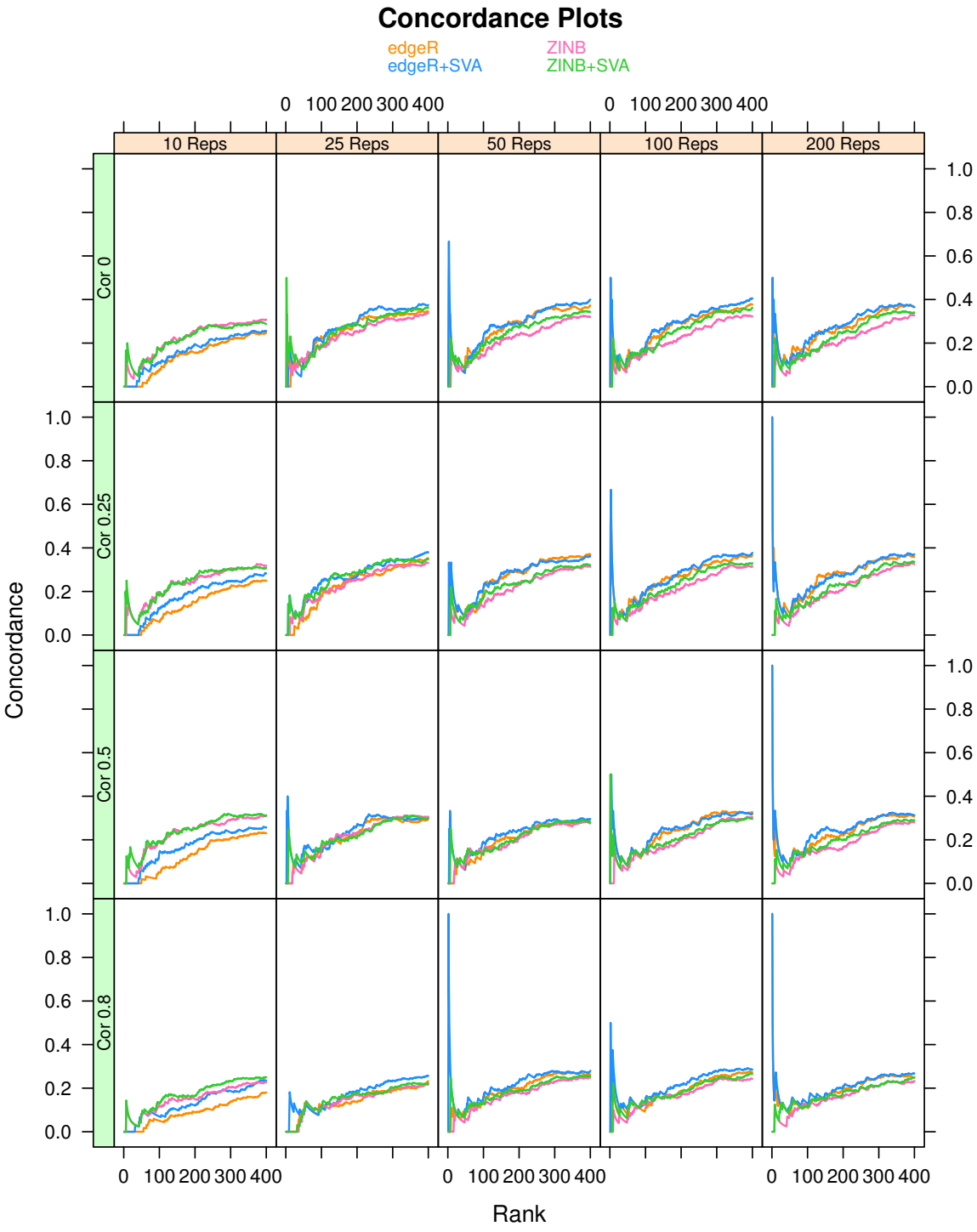

Figure 3.5. Concordance plots depicting the similarity in gene rankings between each method (ZINB and `edgeR`, both and without SVA adjustment), for each combination of replicates per group and level of correlation between the group and cell cycle variable.

3.6 Experimental Data

The performance of ZINB with a surrogate variable cell-cycle adjustment was tested on real scRNA-seq datasets. While ideally it would be preferable to apply the developed methods on datasets that have cell cycles annotated, in addition to another biological factor of interest to be tested for differential expression, such a dataset was not available at time of writing. This reflects the novelty in the scRNA-seq literature of formally testing differential expression between experimental groups, much less treating cell cycle as an explicit covariate when doing so. As an alternative, two datasets are employed, Sasagawa et al. [2013] and Buettner et al. [2015], both of which have cell cycles measured and annotated, and an artificial primary factor establishing differential expression between two groups *in silico* was simulated.

3.6.1 Dataset Descriptions

Sasagawa *et al.* 2013

The Sasagawa et al. [2013] dataset is comprised of 35 mouse embryonic stem cells (mESCs) whose cell cycles were sorted by DNA content using Hoechst staining, which enriches the cells in different stages of the cell cycle. This resulted in the identification of 20 cells in the G1 cycle, 7 in the S cycle, and 8 in the G2/M cycle. The original experiment also involved 12 primitive endoderm (PrE) cells that are directly differentiated from the embryonic stem cells. However, these belong exclusively to the G1 cycle, so were excluded in order to avoid unnecessary confounding with the mESC cell type. Each cell was prepared as a paired-end sequencing library, and all were sequenced in parallel using the Illumina HiSeq 2000 instrument.

Both the raw and processed data files are available on the public genomics data repository Gene Expression Omnibus (GEO), under accession code GSE42268. However, the only processed data that the authors provide had been normalized to FPKM (fragments per kilobase million) expression values, which are inappropriate for the

count data methods that would be applied. As it is not valid to simply convert FPKM values to count values, the raw sequencing files were obtained and the data processed into expression values. This entailed aligning raw sequencing reads to the *Mus musculus* reference genome using the command line tool `bowtie2`, sorting and indexing the aligned reads using the `samtools` utilities, and finally expression quantification using the `summarizeOverlaps` function in the `R` package `GenomicAlignments`. The resulting dataset consisted of count expression values of 22,421 genes in 35 cells. After filtering out genes with average counts of less than 5 across all cells, 13,095 genes remain for analysis.

Buettner *et al.* 2015

The Buettner et al. [2015] dataset is comprised of 288 mESCs, with 96 cells each from the G1, S, and G2/M cell cycle stages. Cell cycles were identified by Hoescht-staining and subsequently sorted using flow cytometry. Single-cell library preparation was performed using the Fluidigm C1 system, paired-end libraries generated using the Illumina Nextera XT kit, and the libraries were sequenced using the Illumina HiSeq 2000 sequencer. Mapping of raw reads was done using the `GSNAP/GMAP` program, to a custom mouse genome (mm10; Ensembl GRCm38.pl); mapped reads were quantified using the python package `HTSeq`. The resulting count data are publicly available at the ArrayExpress archive of functional genomics data, under accession code E-MTB-2805. The original count expression matrix consisted of 38,390 genes in 288 cells. Filtering out genes whose average counts across all cells were less than 5 resulted in 12,938 remaining for analysis.

3.6.2 Data Analysis

The experiments of both Sasagawa et al. [2013] and Buettner et al. [2015] are focused primarily on demonstrating experimental methods with respect to cell cycle specifically; Sasagawa et al. [2013] seeks to demonstrate the ability of their scRNA-

seq method Quartz-Seq to differentiate cell cycle phases, and Buettner et al. [2015] are concerned with removing effects of the cell cycle from the expression data to more robustly identify true cell subpopulations. Hence, the count data obtained from these experiments exhibit labels only for cell cycle, and not for any other experimental groups towards which tests for differential expression could be applied. To remedy this, artificial experimental groups are created by imposing differential expression *in silico*. Specifically, for each dataset, the cells are randomly divided into two experimental groups, and for 20% of the genes, a fold-change value drawn from a $N(\mu = 2, \sigma = 1)$ distribution is generated. This lends the specification of "true" differentially expressed genes, while keeping all other properties of the original scRNA-seq dataset intact, notably any potential cell cycle effects. SVA was applied to each dataset to produce estimates for the cell cycle of each cell. To test for differential expression between the artificial experimental groups, `edgeR` and ZINB are applied, both with and without SVA estimates of cell cycle incorporated into the design matrix.

3.6.3 Sasagawa *et al.* (2013) Results

Figure 3.6 shows the SVA estimates for each cell in the Sasagawa data, colored by the true cell cycle stage. There is some separation seen between G2M and the other stages, but the G1 and S cycles are not cleanly differentiated. This could be partly due to the fact that the SVA method does not look for latent variables due to a specific, defined factor; that is, the estimates may not be describing cell cycle at all, but other artifacts in the data. Hence, a clear separation of SVA estimates on the basis of cell cycle is not to be immediately expected in the real data, and in truth, it is unknown whether cell cycle effects are present at all.

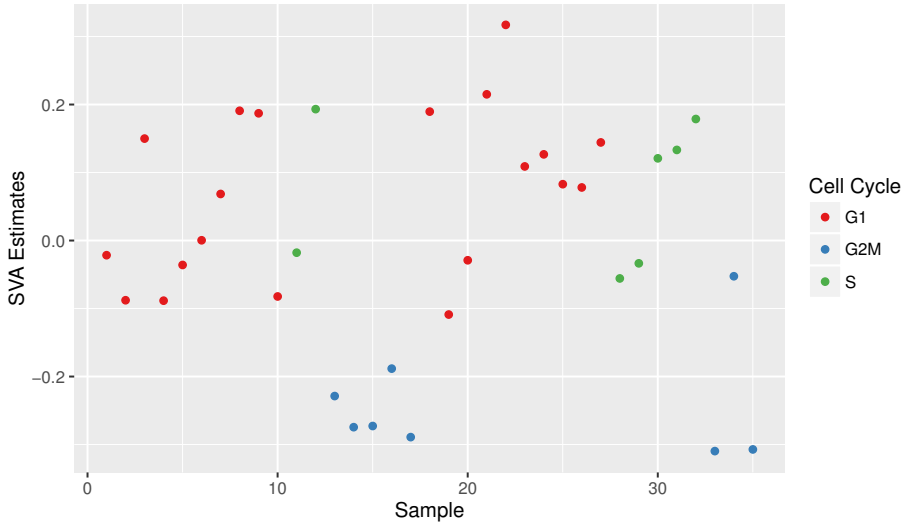

Figure 3.6. SVA estimates for each cell in the Sasagawa *et al.* (2013) data, colored by true cell cycle. See Figure 3.2 for cell cycle descriptions.

A ROC plot is shown in Figure 3.7, comparing the ability of `edgeR` and ZINB, with and without adjustment using the SVA variable, to detect the differential expression that were imposed onto the data. Both versions of ZINB perform better than both versions of `edgeR`, a result that is consistent with the simulations of Section 3.5 which suggest that ZINB exhibits more statistical power in datasets of smaller replicates per group when there are confounding variables present. SVA improves the performance of both `edgeR` and ZINB, though the advantage is slight for the latter. The concordance plot of Figure 3.8 shows similar conclusions; that is, the ranking of genes found by ZINB variations have higher fidelity with the true gene rankings than those found by variations on `edgeR`.

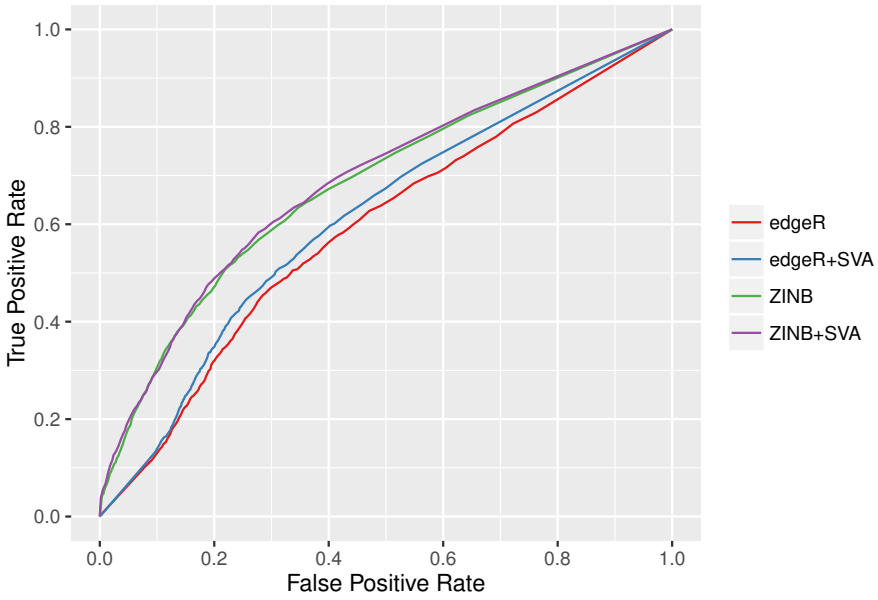

Figure 3.7. Sasagawa *et al.*(2013) results. ROC plot for detecting differential expression between experimental conditions.

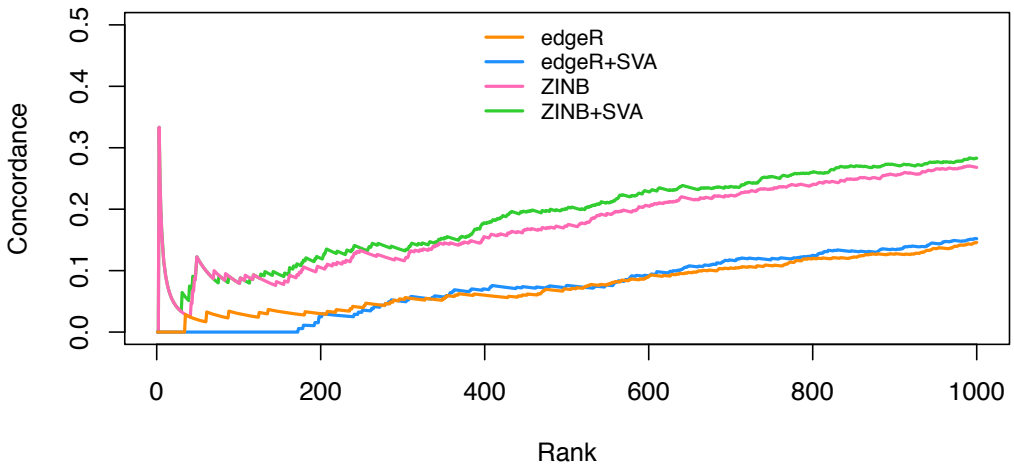

Figure 3.8. Sasagawa *et al.* (2013) results. Concordance plot depicting the similarity in gene rankings between each method and the true gene ranks.

3.6.4 Buetter *et al.* (2015) Results

Figure 3.9 shows the SVA estimates for each cell in the Buettner data, colored by the true cell cycle stage. For these data, the SVA estimates did not seem to detect the cell cycle at all. As with the Sasagawa *et al.*(2013) data, the most apparent explanation is that the cell cycle specifically is not the most influential driver of unwanted variation in the data. In fact, given the small values of the SVA estimates (mostly ranging from -0.1 to 0.1), it may well be that there do not exist systematic confounding variables with large enough effects to be picked up by SVA at all. This interpretation is consistent with the ROC curves (Figure 3.10) and concordance plot (Figure 3.11) of the Buettner *et al.* (2015) data, which do not display added benefits of adding the SVA adjustment to either method. Once again, however, both plots depict better performance of ZINB variations than edgeR variations, though the ZINB advantage is smaller here than in the Sasagawa data. This is likely due to the larger replicate sizes, an effect similarly observed in the simulations of Section 3.5.

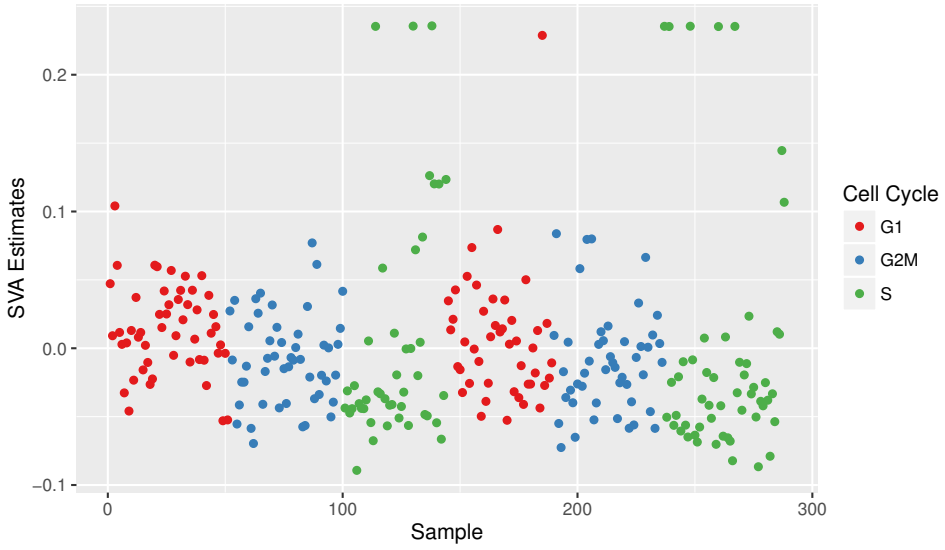

Figure 3.9. SVA estimates for each cell in the Buettner *et al.* (2015) data, colored by true cell cycle.

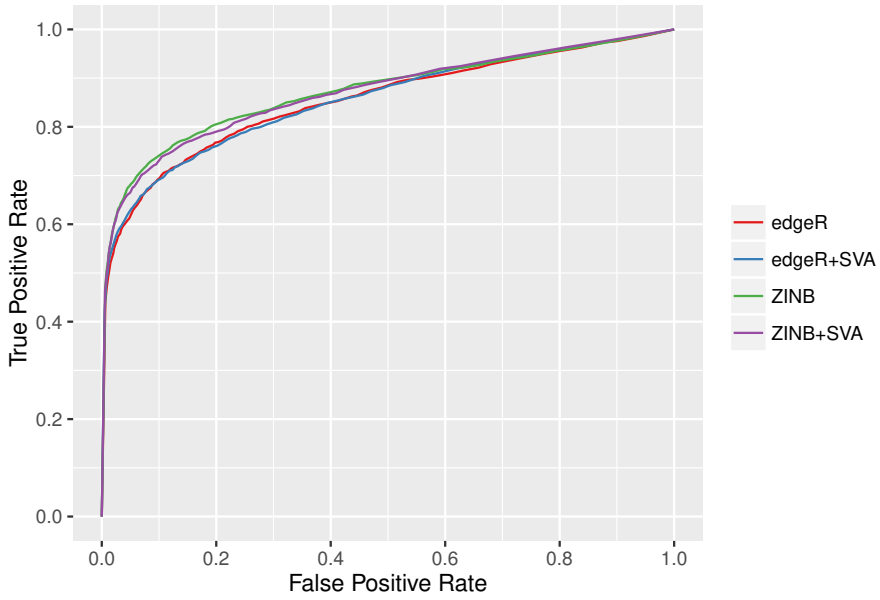

Figure 3.10. Buettner *et al.* (2015) results. ROC plot for detecting differential expression between experimental conditions.

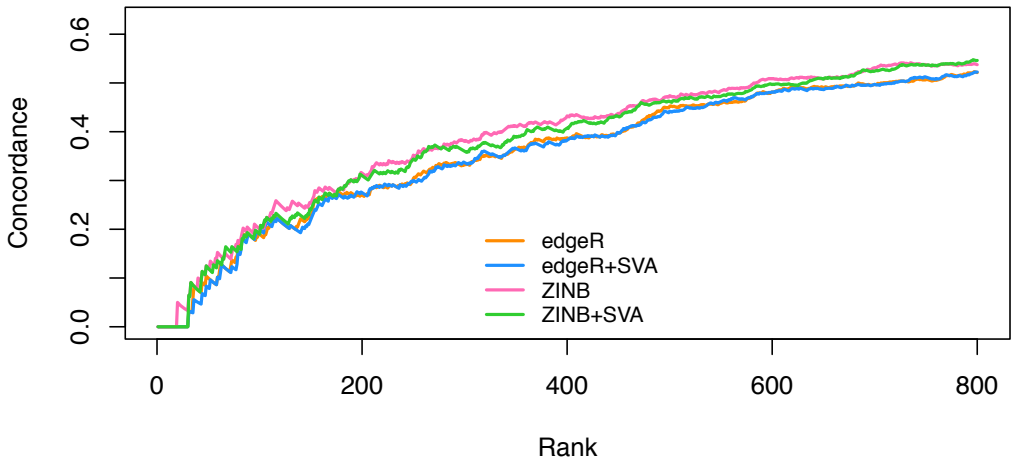

Figure 3.11. Buettner *et al.* (2015) results. Concordance plot depicting the similarity in gene rankings between each method and the true gene ranks.

3.7 Discussion

Single-cell RNA-seq data has been shown to be fundamentally distinct from tissue-level RNA-seq data in two important respects: highly zero-inflated expression values, and the confounding presence of cell cycle stage. Statistical methods for the detection of differential gene expression in scRNA-seq data is underdeveloped. Existing methods originally developed for tissue-level RNA-seq data fail to account for these anomalies, and their rote application to single-cell data consequently fall short. The two strategies posed here work in tandem to address these special features. Surrogate variables are estimated directly from the data, and serve as covariates that account for the unwanted variation from latent factors (*i.e.* cell cycle). Subsequent inclusion of these surrogate variables into a zero-inflated negative binomial model is used to estimate and test for differential gene expression, in light of both the excess of zero counts as well as the unwanted effects from factors such as cell cycle.

Through simulations, ZINB has been shown to outperform the standard bulk tissue method of `edgeR` when replicate sizes are lower, presumably as higher replication levels are able to compensate somewhat for zero-inflation. Furthermore, the incorporation of SVA into ZINB and `edgeR` models improves detection capabilities of both methods. This is particularly true when there is a high correlation between cell cycle stage and the primary group factor, a situation which otherwise poses distinct challenges to testing differential gene expression by confounding the primary signal of interest. It is to be noted that in these simulations, `edgeR` was chosen as the only competitor to ZINB, due to significant computational difficulties of existing methods that attempt to achieve the same goal, namely `DESeq` and `SCDE`, when replication levels are even moderately high. Hence, ZINB with the addition of SVA may be seen as a suitable alternative to existing bulk methods when analyzing scRNA-seq data.

4. SUMMARY

4.1 Summary of Work

Single-cell RNA-sequencing is a revolutionary new frontier, both for biologists as well as statisticians. For well over a decade, the measurement of genome-wide transcription information (RNA-seq) of populations of cells has driven an important part of genomic research. Today, RNA-sequencing of bulk tissues is increasingly affordable and ubiquitous, even routine, and enjoys a level of standardization that continues to push the rate and reproducibility of these experiments. The emerging ability to ask questions of individual cells has already shown tremendous promise in extracting new biological information that eluded scientists even just a few years ago. However, despite the recent surge in research investigations aiming to profile the molecular content of single cells, the field has yet to mature in many important aspects. The purpose of this dissertation is to shed light on critical issues in the design and statistical analysis of single-cell RNA-sequencing experiments, and to offer guidelines and strategies on how to proceed on both fronts.

4.1.1 Design of scRNA-Seq Experiments

Currently, no clear guidelines exist for the design of scRNA-seq experiments, specifically as it pertains to the choice of sequencing depth and replication levels. In Chapter 2, it is explained how the depth at which a sample is sequenced impacts the robustness of its gene expression quantification. Higher sequencing depths compensate for inadequacies of the technology and ensure that an adequate number of molecules are represented in the sequencing library. Despite these benefits, however, researchers would be remiss to simply sequence as much as possible. There exists a

point at which the gains from sequencing more deeply begin to taper off, as more reads fail to yield substantially more genomic information. In addition, there is a practical tradeoff between the number of biological replicates to include in an experiment and how deeply to sequence those replicates; that is, a choice must be made as to whether to sacrifice sequencing depth in favor of including more replicates, or vice versa. Simulations were carried out to shed light on this question, by investigating the effect of different combinations of sequencing depths and replication levels on the detection of differential gene expression. The simulations were carried out in two ways: down-sampling from a real scRNA-seq dataset consisting of two experimental groups, and generating fully simulated data to investigate the effects of larger sample sizes than were available from the real data. In both cases, it was found that increasing the number of replicates substantially and consistently increases statistical power; by contrast, increasing the sequencing depth has only a marginally positive effect beyond the lowest depths.

While these results offer general guidelines, also presented in Chapter 2 is an interactive tool, called `scDesign`, that makes experiment-specific recommendations that are informed by user-submitted pilot data. These pilot data may be either a small-scale portion of a planned experiment or related prototype data from similar existing experiments. In either case, the recommendation is for researchers to provide pilot data that contain a moderately high number of replicates to ensure that features of the data may be adequately captured, even if this requires sacrificing sequencing depth. For each of a range of experimental designs characterized by a sequencing depth and a replication level, `scDesign` estimates the experiment-wide statistical power in one of two ways: a theoretical procedure based on that of Bi and Liu [2016], and a simulation-based empirical calculation. In addition, the projected cost of each experimental design will be calculated, based on a cost function with parameters guided by real experiments. This tool is available both as the `R` package mentioned, and is also implemented for interactive use as a Shiny application, located at `https://github.com/fayezor/scDesignApp`.

4.1.2 Modeling Differential Gene Expression from scRNA-Seq Data

While the RNA-seq data of bulk tissue and single cells look structurally the same - that is, they both contain the expression measurements of tens of thousands of mRNA transcripts obtained across a number of biological replicates - the similarity proves superficial upon closer examination. In Chapter 3, several important ways in which single cell expression data differ from bulk expression measurements on populations of cells are detailed. First, single cells that are captured for sequencing are invariably snapshots of "stochastic" fluctuations in transcription. This phenomenon is masked in bulk data, but manifests itself as an abundance of zeros in single-cell data, where many genes exhibit moderate to strong expression in some cells, but drop out in other cells. Contributing to the observed zero-inflation is also a technical component: minimal amounts of starting mRNA in single cells can lead to transcripts being missed in sample preparation or undetected in the sequencing process. A second important feature that presents itself at the single-cell level is the effect of the cell cycle, which has been known to affect the transcriptional activity of cells in global, non-trivial ways.

The approach proposed in Chapter 3 is to employ a zero-inflated negative binomial distribution for the purpose of modeling differential gene expression in scRNA-seq data. The zero-inflated component would capture the prevalence of excess zeros, while the negative binomial component would model the remaining counts with a distribution appropriate for RNA-seq count data. To account for the unmeasured effects of cell cycle stage on gene expression, an application of surrogate variable analysis [Leek and Storey, 2007] was proposed. SVA is capable of estimating the effects from unwanted factors directly from the data, avoiding the risk inherent in specifying control genes, and allows the incorporation of estimated effects directly as covariates in subsequent models. A simulation study was performed to demonstrate how SVA may be applied in conjunction with ZINB to improve the detection of differential gene expression in data simulated to display the features in question. It was observed that

when the cell cycle is correlated with the group factor of primary interest, SVA estimates are able to clearly distinguish true cell cycles. Accordingly, the incorporation of SVA covariates into subsequent models improves differential expression detection capabilities over corresponding models without SVA adjustment. Both variations of ZINB, with and without SVA, outperform both variations of the accepted standard method of `edgeR`. However, this advantage disappears for higher levels of replication, presumably because larger samples are able to compensate for loss of statistical power due to zero-inflation.

4.2 Future Work

Experiments involving thousands, even tens of thousands, of cellular replicates will soon be the norm in the single-cell field, as technologies become ever more massively parallel and platforms find ways to exploit economies of scale. 96-well plates for the isolation and processing of cells is the current standard, but 800-well systems are already making their way into cutting edge facilities. Chapter 2 was focused on experimental design specifically as it pertains to the choice of sequencing depth and replicate number, parameters which impact the ability to extract genomic information from sequence data. These questions are necessary for establishing experimental standards for the budding technology, and to encourage thoughtful planning for researchers who face limits to their resources. However, as costs continue to plummet and scientists may be freed to think beyond the constraints of cellular replicates and sequencing depths, these considerations will give way in immediate importance to more foundational notions of experimental design. For example, the necessity of replication is driven by the presence of biological variability, which exists not just from cell to cell, but also from organism to organism and tissue to tissue. Currently, the unit of direct interest is limited to the individual cell, while other layers of biological variability, originating from the tissues and whole organisms from which those cells are selected, are routinely neglected. The most commonly reported experiments are

performed on samples of cellular replicates with no information on tissue or organism replicates; that is, the cells may as well originate from a single tissue in a single organism. As long as experimenters seek to characterize cells in reproducible ways, deliberate replication in these other layers is essential.

Throughout this work, the target task of statistical inference was the detection of genes that exhibit differential expression across experimental groups. The intended focus was on improving tests of specific hypotheses, contrasting these with the unsupervised exploratory data analysis procedures that are currently much more prevalent in the literature. However, the goals of identifying subpopulations of cells, currently accomplished through tools such as clustering and PCA, deserve to be revisited in their own right in order to accommodate them to the unique features of single cell data. Indeed, it should be recognized that there are pertinent biological questions that would treat the characterization of cell subpopulations as an end goal, rather than simply a middle step towards confirmatory statistical tests. Single cell data also provide unique opportunities to understand cell differentiation processes that could elucidate mechanisms behind cell renewal, disease development, and tissue generation. Current data collection methods can be thought of as providing snapshots of cells frozen in time, which is sufficient for singular characterizations but does not lend itself well to tracing differentiation behaviors over time. An open statistical problem is the question of how to make powerful inferences both across and within time points, and how to identify subsets of genes that follow similar differentiation patterns.

Aside from these opening questions, there will be continued clarification of biological goals as the field develops. Looking to the future, single-cell experimentation will inevitably continue to involve more cells from more tissues from more organisms, as well as more kinds of data that beg to be integrated. Needless to say, the computational weight will continue to rise. For cells easily numbering in the thousands, one must generate easily accessible raw data from the sequencing machines, process the raw data into sequence information using bioinformatics tools, and infer some knowledge about the biological property using appropriately developed statis-

tical methodology. Current data analysis pipelines meant for bulk tissues may not adequately account for the new errors, biases, and sources of variation that single-cell technologies will carry. Carefully characterizing the issues specific to single-cell experimentation is an important goal in future statistical analyses of these data.

LIST OF REFERENCES

LIST OF REFERENCES

Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

Michael L Metzker. Sequencing technologies: the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

National Human Genome Research Institute. DNA sequencing costs. `http://www.genome.gov/sequencingcosts/`, 2015. Accessed: 2015-11-13.

James Eberwine, Jai-Yoon Sul, Tamas Bartfai, and Junhyong Kim. The promise of single-cell sequencing. *Nature Methods*, 11(1):25–27, 2014.

Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14 (9):618–630, 2013.

Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, 2011.

Ludmil B Alexandrov and Michael R Stratton. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Current Opinion in Genetics & Development*, 24:52–60, 2014.

Lucy R Yates and Peter J Campbell. Evolution of the cancer genome. *Nature Reviews Genetics*, 13(11):795–806, 2012.

Daniel Ramsköld, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, 30(8):777–782, 2012.

Gordon M Cann, Zulfiqar G Gulzar, Samantha Cooper, Robin Li, Shujun Luo, Mai Tat, Sarah Stuart, Gary Schroth, Sandhya Srinivas, Mostafa Ronaghi, et al. mRNA-Seq of single prostate cancer circulating tumor cells reveals recapitulation of gene expression and pathways found in prostate cancer. *PloS One*, 7(11), 2012.

Suzan Yilmaz and Anup K Singh. Single cell genome sequencing. *Current Opinion in Biotechnology*, 23(3):437–443, 2012.

Paul C Blainey. The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiology Reviews*, 37(3):407–427, 2013.

Michael J McConnell, Michael R Lindberg, Kristen J Brennand, Julia C Piper, Thierry Voet, Chris Cowing-Zitron, Svetlana Shumilina, Roger S Lasken, Joris R Vermeesch, Ira M Hall, et al. Mosaic copy number variation in human neurons. *Science*, 342(6158):632–637, 2013.

Annapurna Poduri, Gilad D Evrony, Xuyu Cai, and Christopher A Walsh. Somatic mutation, genomic variation, and neurological disease. *Science*, 341(6141), 2013.

Liying Yan, Mingyu Yang, Hongshan Guo, Lu Yang, Jun Wu, Rong Li, Ping Liu, Ying Lian, Xiaoying Zheng, Jie Yan, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structural & Molecular Biology*, 20(9):1131–1139, 2013.

Subhashini Chandrasekharan, Mollie A Minear, Anthony Hung, and Megan A Allyse. Noninvasive prenatal testing goes global. *Science Translational Medicine*, 6 (231), 2014.

Chanchao Lorthongpanich, Lih Feng Cheow, Sathish Balu, Stephen R Quake, Barbara B Knowles, William F Burkholder, Davor Solter, and Daniel M Messerschmidt. Single-cell DNA-methylation analysis reveals epigenetic chimerism in preimplantation embryos. *Science*, 341(6150):1110–1112, 2013.

Keith R Willison and David R Klug. Quantitative single cell and single molecule proteomics for clinical studies. *Current Opinion in Biotechnology*, 24(4):745–751, 2013.

Stanislav S Rubakhin, Eric J Lanni, and Jonathan V Sweedler. Progress toward single cell metabolomics. *Current Opinion in Biotechnology*, 24(1):95–104, 2013.

Francis Crick et al. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Research*, 2014.

Howard M Shapiro. *Practical Flow Cytometry*. John Wiley & Sons, 2005.

Fluidigm, Inc. `https://www.fluidigm.com/press/high-throughput-single-cell-mrna-sequencing-preparation-comes-to-the-fludigm-c1-system`, 2015. Accessed: 2016-06-22.

Illumina, Inc. mRNA Sequencing. `http://www.illumina.com/techniques/sequencing/rna-sequencing/mrna-seq.html`, 2015a. Accessed: 2015-11-02.

Illumina, Inc. An introduction to next-generation sequencing technology. `http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf`, 2015b. Accessed: 2015-11-02.

Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

Alicia Oshlack, Mark D Robinson, and Matthew D Young. From RNA-seq reads to differential expression results. *Genome Biology*, 11(12):1, 2010.

Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, 2008.

Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166, 2014.

Eric S Lander and Michael S Waterman. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2(3):231–239, 1988.

David Sims, Ian Sudbery, Nicholas E Ilott, Andreas Heger, and Chris P Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–132, 2014.

Sonia Tarazona, Fernando García-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. Differential expression in RNA-seq: a matter of depth. *Genome Research*, 21(12):2213–2223, 2011.

Yuwen Liu, Jie Zhou, and Kevin P White. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, 30(3):301–304, 2014.

Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3): 133–145, 2015.

Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779, 2014.

Bidesh Mahata, Xiuwei Zhang, Aleksandra A Kolodziejczyk, Valentina Proserpio, Liora Haim-Vilmovsky, Angela E Taylor, Daniel Hebenstreit, Felix A Dingler, Victoria Moignard, Berthold Göttgens, et al. Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell Reports*, 7(4):1130–1142, 2014.

David G Robinson and John D Storey. subSeq: determining appropriate sequencing depth through efficient read subsampling. *Bioinformatics*, 30(23):3424–3426, 2014.

Ran Bi and Peng Liu. Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. *BMC Bioinformatics*, 17(1):1, 2016.

Zhide Fang and Xiangqin Cui. Design and validation issues in RNA-seq experiments. *Briefings in Bioinformatics*, 2011.

Steven N Hart, Terry M Therneau, Yuji Zhang, Gregory A Poland, and Jean-Pierre Kocher. Calculating sample size estimates for rna sequencing data. *Journal of Computational Biology*, 20(12):970–978, 2013.

Michele A Busby, Chip Stewart, Chase A Miller, Krzysztof R Grzeda, and Gabor T Marth. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics*, 29(5):656–657, 2013.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

Chung-I Li, Pei-Fang Su, Yan Guo, and Yu Shyr. Sample size calculation for differential expression analysis of RNA-seq data under Poisson distribution. *International Journal of Computational Biology and Drug Design*, 6(4):358–375, 2013a.

Chung-I Li, Pei-Fang Su, and Yu Shyr. Sample size calculation based on exact test for assessing differential expression analysis in RNA-seq data. *BMC Bioinformatics*, 14(1):357, 2013b.

Peng Liu and JT Gene Hwang. Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*, 23(6): 739–746, 2007.

John D Storey. The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics*, pages 2013–2035, 2003.

Camille Stephan-Otto Attolini, Victor Peña, and David Rossell. Designing alternative splicing RNA-seq studies. Beyond generic guidelines. *Bioinformatics*, 2015.

Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, 1995.

Patrick O Brown and David Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21:33–37, 1999.

John Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6):418–427, 2001.

Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–240, 2013.

Piero Dalerba, Tomer Kalisky, Debashis Sahoo, Pradeep S Rajendran, Michael E Rothenberg, Anne A Leyrat, Sopheak Sim, Jennifer Okamoto, Darius M Johnston, Dalong Qian, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnology*, 29(12):1120–1127, 2011.

Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, 21(7):1160–1167, 2011.

John W Tukey. Exploratory data analysis. 1977.

Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

Arjun Raj and Alexander van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2008.

Harley H McAdams and Adam Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 94(3):814–819, 1997.

Benjamin B Kaufmann and Alexander van Oudenaarden. Stochastic gene expression: from single molecules to the proteome. *Current Opinion in Genetics & Development*, 17(2):107–112, 2007.

Georgi K Marinov, Brian A Williams, Ken McCue, Gary P Schroth, Jason Gertz, Richard M Myers, and Barbara J Wold. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Research*, 24(3):496–510, 2014.

Brian Munsky, Gregor Neuert, and Alexander van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187, 2012.

Alvaro Sanchez and Ido Golding. Genetic determinants and cellular constraints in noisy gene expression. *Science*, 342(6163):1188–1193, 2013.

Quin F Wills, Kenneth J Livak, Alex J Tipping, Tariq Enver, Andrew J Goldson, Darren W Sexton, and Chris Holmes. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature Biotechnology*, 31 (8):748–752, 2013.

Ido Golding, Johan Paulsson, Scott M Zawilski, and Edward C Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, 2005.

Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, 2014.

CJ Zopf, Katie Quinn, Joshua Zeidman, and Narendra Maheshri. Cell-cycle dependence of transcription dominates noise in gene expression. *PLoS Computational Biology*, 9(7), 2013.

Wei-Chiang Chen, Pei-Hsun Wu, Jude M Phillip, Shyam B Khatau, Jae Min Choi, Matthew R Dallas, Konstantinos Konstantopoulos, Sean X Sun, Jerry SH Lee, Didier Hodzic, et al. Functional interplay between the cell cycle and cell phenotypes. *Integrative Biology*, 5(3):523–534, 2013.

Amar M Singh, James Chappell, Robert Trost, Li Lin, Tao Wang, Jie Tang, Hao Wu, Shaying Zhao, Peng Jin, and Stephen Dalton. Cell-cycle control of developmentally regulated transcription factors accounts for heterogeneity in human pluripotent cells. *Stem Cell Reports*, 1(6):532–544, 2013.

Siim Pauklin and Ludovic Vallier. The cell-cycle state of stem cells determines cell fate propensity. *Cell*, 155(1):135–147, 2013.

Ziv Bar-Joseph, Zahava Siegfried, Michael Brandeis, Benedikt Brors, Yong Lu, Roland Eils, Brian D Dynlacht, and Itamar Simon. Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. *Proceedings of the National Academy of Sciences*, 105(3):955–960, 2008.

Michael B Kastan and Jiri Bartek. Cell-cycle checkpoints and cancer. *Nature*, 432 (7015):316–323, 2004.

Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10), 2010.

Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332, 2008.

Mark D Robinson and Gordon K Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 2012.

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

Andrew McDavid, Greg Finak, Pratip K Chattopadyay, Maria Dominguez, Laurie Lamoreaux, Steven S Ma, Mario Roederer, and Raphael Gottardo. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics*, 29(4):461–467, 2013.

JA Nelder and RWM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, pages 370–384, 1972.

Peter McCullagh and John A Nelder. *Generalized Linear Models*, volume 37. CRC press, 1989.

Diane Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

John Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365, 1986.

Charles George Broyden. The convergence of a class of double-rank minimization algorithms. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

Laurent Meijer. Chemical inhibitors of cyclin-dependent kinases. *Trends in Cell Biology*, 6(10):393–397, 1996.

Tony Ly, Yasmeen Ahmad, Adam Shlien, Dominique Soroka, Allie Mills, Michael J Emanuele, Michael R Stratton, and Angus I Lamond. A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells. *eLife*, 3, 2014.

Rafael Nunez. DNA measurement and cell cycle analysis by flow cytometry. *Current Issues in Molecular Biology*, 3:67–70, 2001.

Antonio Scialdone, Kedar N Natarajan, Luis R Saraiva, Valentina Proserpio, Sarah A Teichmann, Oliver Stegle, John C Marioni, and Florian Buettner. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61, 2015.

Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.

Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–160, 2015.

Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9), 2007.

Jeffrey T Leek. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research*, 2014.

Yohei Sasagawa, Itoshi Nikaido, Tetsutaro Hayashi, Hiroki Danno, Kenichiro D Uno, Takeshi Imai, and Hiroki R Ueda. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biology*, 14(4), 2013.

José A Robles, Sumaira E Qureshi, Stuart J Stephen, Susan R Wilson, Conrad J Burden, and Jennifer M Taylor. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics*, 13(1):484, 2012.

Diana W Bianchi, R Lamar Parker, Jeffrey Wentworth, Rajeevi Madankumar, Craig Saffer, Anita F Das, Joseph A Craig, Darya I Chudova, Patricia L Devers, Keith W Jones, et al. DNA sequencing versus standard prenatal aneuploidy screening. *New England Journal of Medicine*, 370(9):799–808, 2014.

Andrew McDavid, Lucas Dennis, Patrick Danaher, Greg Finak, Michael Krouse, Alice Wang, Philippa Webster, Joseph Beechem, and Raphael Gottardo. Modeling bi-modality improves characterization of cell cycle on gene expression in single cells. *PLoS Computational Biology*, 10(7), 2014.

Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Julianna McElrath, Martin Prlic, Peter Linsley, and Raphael Gottardo. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA-seq data. *bioRxiv*, 2015. doi: 10.1101/020842.

Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–640, 2014.

Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1095, 2013.

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, 2010.

Ronald A Fisher. The Design of Experiments. *Oliver and Boyd, Edinburgh*, 1935.

Gary A Churchill. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics*, 32:490–495, 2002.

M Kathleen Kerr and GARY A CHURCHILL. Statistical design and the analysis of gene expression microarray data. *Genetical Research*, 77(02):123–128, 2001.

M Kathleen Kerr and Gary A Churchill. Experimental design for gene expression microarrays. *Biostatistics*, 2(2):183–201, 2001.

M Kathleen Kerr, Mitchell Martin, and Gary A Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7(6):819–837, 2000.

Mei-Ling Ting Lee, Frank C Kuo, GA Whitmore, and Jeffrey Sklar. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences*, 97(18):9834–9839, 2000.

Paul L Auer and RW Doerge. Statistical design and analysis of RNA sequencing data. *Genetics*, 185(2):405–416, 2010.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Springer, 2013.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *International Conference on Database Theory*, pages 217–235. Springer, 1999.

Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1):1, 2009.

Yosef Buganim, Dina A Faddah, Albert W Cheng, Elena Itskovich, Styliani Markoulaki, Kibibi Ganz, Sandy L Klemm, Alexander van Oudenaarden, and Rudolf Jaenisch. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell*, 150(6):1209–1222, 2012.

Sean C Bendall, Erin F Simonds, Peng Qiu, D Amir El-ad, Peter O Krutzik, Rachel Finck, Robert V Bruggner, Rachel Melamed, Angelica Trejo, Olga I Ornatsky, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687–696, 2011.

Guoji Guo, Mikael Huss, Guo Qing Tong, Chaoyang Wang, Li Li Sun, Neil D Clarke, and Paul Robson. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental Cell*, 18(4):675–685, 2010.

Wenjun Ju, Casey S Greene, Felix Eichinger, Viji Nair, Jeffrey B Hodgin, Markus Bitzer, Young-suk Lee, Qian Zhu, Masami Kehata, Min Li, et al. Defining cell-type specificity at the transcriptional level in human disease. *Genome Research*, 23(11): 1862–1873, 2013.

Fuchou Tang, Catalin Barbacioru, Siqin Bao, Caroline Lee, Ellen Nordman, Xiaohui Wang, Kaiqin Lao, and M Azim Surani. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell*, 6(5): 468–478, 2010.

Richard Bellman, Richard Ernest Bellman, Richard Ernest Bellman, and Richard Ernest Bellman. *Adaptive Control Processes: A Guided Tour*, volume 4. Princeton University Press Princeton, 1961.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer Series in Statistics Springer, Berlin, 2001.

Marco Gerlinger, Andrew J Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine*, 366(10):883–892, 2012.

Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. Comparison of next-generation sequencing systems. *BioMed Research International*, 2012, 2012.

National Health Genome Research Institute. `http://www.genome.gov/11006943/`, 2014a.

National Health Genome Research Institute. `http://www.genome.gov/sequencingcosts/`, 2014b.

Min Yu, David T Ting, Shannon L Stott, Ben S Wittner, Fatih Ozsolak, Suchismita Paul, Jordan C Ciciliano, Malgorzata E Smas, Daniel Winokur, Anna J Gilman, et al. RNA sequencing of pancreatic circulating tumour cells implicates WNT signalling in metastasis. *Nature*, 2012.

Rickard Sandberg. Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods*, 11(1):22–24, 2014.

Tomer Kalisky and Stephen R Quake. Single-cell genomics. *Nature Methods*, 8(4): 311–314, 2011.

VITA

VITA

Faye Zheng was born in the city of Yongzhou in Hunan Province, China on August 4, 1987, and immigrated with her parents to the United States in 1991. She received her Bachelor of Arts degree from Northwestern University in 2009, where she majored in Mathematics and Economics as part of the Mathematical Methods in the Social Sciences (MMSS) honors program. She began her graduate studies at Purdue Unversity in 2010, and received her Masters degree in Applied Statistics there in 2012.