Approaches for Identifying Biases in Single-Cell RNA-Sequencing Data

by

Julie M. Deeke

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2019

Doctoral Committee:

Assistant Professor Johann A. Gagnon-Bartsch, Chair
Assistant Professor Yang Chen
Professor Jun Li
Professor Kerby Shedden

Julie M. Deeke
jghekas@umich.edu
ORCID iD: 0000-0002-3046-6181

for my family and friends, especially Alex and Sophie,
with many thanks for their unwavering love, support, and
patience
✝JMJ ✝

ACKNOWLEDGMENTS

First, I would like to thank Johann, without whom this work would not have been possible. I have learned so much from him over the past four years and appreciate all of his guidance. He served as a co-author on this work.

Second, I thank the members of my committee: Kerby, Yang, and Jun. Their comments and suggestions have substantially improved this dissertation, and I am grateful for their time and input.

Third, thanks go to the entire Michigan Center for Single-Cell Genomic Data Analytics. Their insights and comments have contributed to this dissertation, and I have enjoyed learning alongside them. Particular thanks to Anna Gilbert and Jun Li, and to Justin Colacino, Tasha Thong, Alex Vargo, Umang Varma, Yutong Wang, and Xiang Zhou.

Thanks also to Gregory Hunt for his comments, suggestions, and debugging help.

Table of Contents

List of Figures

Figure

List of Tables

Table

Abstract

Single-cell RNA-sequencing (scRNA-seq) involves the measurement of gene expression from isolated single cells, with the potential to illuminate cellular heterogeneity within complex tissue samples. However, scRNA-seq data are subject to a large number of technical effects. In this work, we present two approaches that can be applied for identifying technical effects in scRNA-seq data in pursuit of the ultimate aim of removing these effects in downstream analyses.

The first project introduces different concepts of stably expressed genes with respect to true biological expression. Different classes of stably expressed genes may capture different technical effects, assisting in the removal of these technical effects and increasing the biological interpretability of later results. We define different notions of stability explicitly and organize gene sets by structure to identify gene sets that are systematically enriched with different classes of stably expressed genes. We find that genes associated with the cytosolic ribosomal structure are enriched with genes that are stably expressed in proportion to the total RNA content of a cell. The cytosolic ribosomal genes can serve as a foundation for a set of negative control genes incorporated into normalization procedures to remove some technical effects associated with cell size.

The second project describes a procedure to estimate the efficiency with which genes are captured in scRNA-seq experiments. Unique molecular identifiers (UMIs) tag molecules before amplification and allow for later deduplication of reads originating from the same original molecule. Ordinarily, we are interested in the number of UMIs present to enumerate the gene expression of a cell. We consider the information typically discarded during deduplication by measuring the number of reads captured for each unique molecular identifier (rUMI). Features of reliable detection and genes with those features can be identified from the rUMI distributions and applied to downstream analyses, correcting for additional technical effects.

Specifically, we estimate the unobserved proportion of molecules that are not captured (0 rUMI) from the observed distribution of rUMI by extrapolation. We propose a framework for three possible zero-truncated integer-based distributions to model the rUMI distribution. Based on the models, we identify five estimators of the true proportion of reads that are lost during the experiment and result in 0 rUMIs. We derive and simulate asymptotic features of these estimators under different conditions and perturbations motivated by scRNA-seq data. We apply our estimators to a

scRNA-seq dataset, noting molecule characteristics like the proportion of GC bases in a transcript, that suggest reliable detection.

Together, these projects provide two approaches to identifying technical biases in scRNA-seq data and can be applied to later normalization procedures.

Chapter 1

Introduction

Single-cell RNA-sequencing (scRNA-seq) was first published in 2009, with many experimental and analytical advances developed since the first experiment. Scientists can capture thousands of cells at a low cost with droplet-based cell isolation protocols or can select individual cells for placement in well-based protocols. From each of these single cells, mRNA transcripts are captured, amplified, sequenced, and aligned, providing measures of the gene expression for a given cell. Errors can be introduced during any of these steps, and technical variation is fundamentally a problem in scRNA-seq data.

Various advances, both biological and analytical, have been introduced to correct and adjust for technical effects in scRNA-seq data. The introduction of unique molecular identifiers (UMIs) is one biological advancement, where the UMI serves as a tag used to deduplicate reads originating from amplification. External spike-ins provide another experimental means of measuring some of the technical effects in scRNA-seq experiments. Analytically, various procedures and normalizations have been developed and proposed to remove effects associated with biological and technological characteristics, including cell cycle and batch effects.

In the biologically motivated work presented here, we identify a set of genes and characteristics of genes that help capture some of the technical biases in scRNA-seq experiments and could help to normalize scRNA-seq data for technical effects. Specifically, we examine stably expressed genes and reliably detectable genes. The stably expressed genes could serve as negative controls in normalization procedures. Additionally, we can adjust a gene's observed expression measurement from its estimated proportion of molecules lost to technical effects. A low proportion of molecules estimated to be lost during a scRNA-seq experiment due to technical effects could be used to define reliably detectable genes.

In Chapter 2, we explore the concept of stable expression of genes. First, we define stably expressed genes; we argue that the concept of stable expression at the single cell level is ambiguous and provide different notions of stability. We propose measures with which to assess the stable expression of a gene. Further, we propose a model for measured expression at the single cell level and apply that model in the interpretation of the results from our stability measures. We explore

gene sets determined by the cellular structures in which a gene's final product is associated, identifying the cytosolic ribosomal genes as being enriched in proportionally stably expressed genes in scRNA-seq experiments performed by multiple labs using both Fluidigm C1 and 10x scRNA-seq platforms. Additionally, we find similar results from an analysis of bulk tissues gathered as part of the GTEx project, supporting that the cytosolic ribosomal genes are in fact enriched with proportionally stably expressed genes. We also generate a database of gene information summarizing various analyses for use by researchers in further refining sets of genes for later normalization. Supplementary information and analyses are found in Chapter 6.

In Chapter 3, we define rUMI as a measure of reliable detection for genes in scRNA-seq. The number of reads associated with a UMI (rUMI), which typically is removed during the deduplication of scRNA-seq results, can also provide a measure of the technical biases present in a scRNA-seq experiment. We propose an estimator for the parameter from a zero-truncated Poisson distribution. We apply this estimator to rUMIs obtained from a scRNA-seq experiment, exploring the relationship between different variables and rUMI. We find that an imbalance of GC bases compared to AT bases in both the full length of the transcript and the first half of a transcript results in less reliable detection of a gene. Chapter 7 provides supplementary information, including a generalized linear model to estimate rUMI.

Chapter 4 further develops estimators based on the zero-truncated distribution introduced in Chapter 3 and two additional zero-truncated distributions described in Chapter 4. We motivate the three special cases that describe potential rUMI distributions. We then propose five estimators for the unobserved probability that an rUMI is equal to zero, i.e. the proportion of molecules that attach to a UMI but fail to be captured as a read; these molecules are lost due to technical errors. We describe characteristics of these estimators analytically when possible and by simulation. We apply our estimators to subsets of the data with various transcript attributes to observe the relationship between transcript attributes and the unobserved probability of losing a transcript. We find that the robust estimator based on a zero-truncated Poisson distribution introduced in Chapter 3 performs well under most conditions but is rarely optimal. We also explore the relationship between the estimated expression level of a gene and the technical loss of that gene, finding that lowly expressed genes are estimated to experience marginally more technical loss. Chapter 8 provides supplementary theoretical results derived from estimator properties and additional evaluations of estimator performance from simulations.

Finally, Chapter 5 summarizes the findings of Chapters 2, 3, and 4 and suggests future avenues for continued exploration in this area.

Chapter 2

Stably Expressed Genes in Single-Cell RNA-Sequencing

2.1 Abstract

Motivation: In single-cell RNA-sequencing (scRNA-seq) experiments, RNA transcripts are extracted and measured from isolated cells to understand gene expression at the cellular level. Measurements from this technology are affected by many technical artifacts, including batch effects. In analogous bulk gene expression experiments, external references, e.g., synthetic gene spike-ins often from the External RNA Controls Consortium (ERCC), may be incorporated to the experimental protocol for use in adjusting measurements for technical artifacts. In scRNA-seq experiments, the use of external spike-ins is controversial due to dissimilarities with endogenous genes and uncertainty about sufficient precision of their introduction. Instead, endogenous genes with highly stable expression could be used as references within scRNA-seq to help normalize the data. First, however, a specific notion of stable expression at the single cell level needs to be formulated; genes could be stable in absolute expression, in proportion to cell volume, or in proportion to total gene expression. Different types of stable genes will be useful for different normalizations and will need different methods for discovery.

Results: We compile gene sets whose products are associated with cellular structures and record these gene sets for future reuse and analysis. We find that genes whose final product are associated with the cytosolic ribosome have expressions that are highly stable with respect to the total RNA content. Notably, these genes appear to be stable in bulk measurements as well.

Availability: The gene set database is available online through github (github.com/johanngb/sc-stable).

2.2 Introduction

Single-cell RNA-sequencing (scRNA-seq) experiments measure gene expression at the cellular level, capturing details at a resolution previously not possible. However, challenges arise due to

unwanted variation that scRNA-seq experiences. Some sources of unwanted variation include read depth, capture efficiency, amplification biases, batch effects, and cell cycle [Hicks *et al.*, 2018; Phipson *et al.*, 2017; Lun and Marioni, 2017; Dabney and Meyer, 2012; Kolodziejczyk *et al.*, 2015]. Methods have been developed to remove some sources of unwanted variation, often using certain sets of reference genes to aid in removing either specific or general effects. For example, Chen and Zhou [2017] use Bayesian methods to identify control genes that are unassociated with a factor of interest and adjust the target genes based on the control genes. Buettner *et al.* [2015] use genes that have been annotated as associated with the cell cycle to remove cell cycle effects from the data. Both Brennecke *et al.* [2013] and Grün *et al.* [2014] propose using external spike-in references to remove some of the technical noise present in the data. Finally, Lin *et al.* [2019b] propose using stably expressed genes as a form of negative controls to remove unwanted variation using a procedure called scMerge.

Commercially generated, synthetic, external spike-in references (often External RNA Controls Consortium (ERCC) spike-ins) can be used in bulk gene expression experiments [Baker *et al.*, 2005; Jiang *et al.*, 2011; Risso *et al.*, 2014; Pine *et al.*, 2016]. The ability to incorporate external spike-in references in scRNA-seq varies based on the cell isolation protocol. Spike-ins are not typically included in droplet-based isolation protocols but can be in plate- and well-based isolation methods including the Fluidigm C1 system [Macosko *et al.*, 2015; Bacher and Kendziorski, 2016; Lun *et al.*, 2016; Tung *et al.*, 2017].

Limitations to the use of spike-ins are not limited to scRNA-seq contexts. A critique in both single cell and bulk experiments is that the spike-ins possess qualities that are dissimilar to endogenous genes [Grün and vanOudenaarden, 2015; Tung *et al.*, 2017]. Spike-ins are designed to exhibit artificially wide ranging characteristics, like length and proportion of guanine and cytosine bases in the nucleic acid sequence, in order to understand how these characteristics might affect downstream results. Specific to scRNA-seq, the quantity of solution added for each cell is much smaller, so minor pipetting errors (with the Fluidigm C1 system) or other volume errors affect results much more than with a larger quantity of solution [Tung *et al.*, 2017]. The technical challenge of accurately introducing and measuring a smaller volume of spike-ins reduces their effectiveness as negative controls in scRNA-seq [Robinson and Oshlack, 2010].

Endogenous genes that are reasonably stably expressed have been proposed for use in normalization of microarray data [Eisenberg and Levanon, 2003, 2013; Gagnon-Bartsch and Speed, 2012]. However, it is unclear that the single cell expression of these same genes exhibit the same stability as in bulk experiments. For example, a gene may be expressed with a bursting mechanism, increasing its variability [Jiang *et al.*, 2017; Suter *et al.*, 2011; Fukaya *et al.*, 2016]. Bulk expression data might identify the gene as stably expressed, but that same classification at the single cell level would be inappropriate.

There is a need to discover single cell-specific stably expressed genes. Lin *et al.* [2019a] propose a method of creating an index at the single cell level for generating a set of stably expressed genes across all cell conditions. Desired characteristics of stably expressed genes from Lin *et al.* [2019a] include a distribution with a small proportion of measurements with low values and a small variance among the measurements with high values as estimated from parameters of a Gamma-Gaussian model.

One prominent feature of scRNA-seq data is the presence of a large number of zero measured expressions for genes, sometimes called dropouts. The prevalence of zero counts can reach 90% in some datasets. Ideally, any stably expressed genes used for normalization would also be reliably detected and have a relatively low proportion of zero counts.

The goals of this Chapter are: (1) to clarify the notion of "stable expression" at the single cell level, and in particular to define multiple such notions, (2) to propose a method in which to identify a set of genes that exhibit stable expression, (3) to organize sets of genes based on the cellular component with which the final gene product is associated, and (4) to suggest the set of cytosolic ribosomal genes as stably expressed with respect to total RNA content.

2.3 Approach

2.3.1 Notions of Stability

We first consider explicitly what it means for a gene to be stable. The idea of stable expression has previously been addressed either implicitly or without much elaboration. However, the notion of stability at the single cell level is inherently ambiguous and requires a precise definition. In the following paragraphs we consider multiple notions of stability. Importantly, these notions refer to the true gene expressions within a cell, not to the measured expressions (data), a point we return to below.

One notion of a stably expressed gene would be that the gene is expressed at a constant absolute level. In other words, the number of RNA molecules present within each cell should be approximately constant, e.g., each cell has about 10 RNA molecules of that gene. We refer to this notion of stability as "absolute stability." Genes that are absolutely stable could replace the external spike-ins, as they are expected to be present at a fixed absolute amount in each cell. Like spike-ins, these genes could be used to pick up certain technical effects, such as reaction efficiency.

A second notion of stable expression would be that genes are expressed at a constant proportion with respect to cell volume; that is, they are stable in terms of concentration. We refer to this as "stable in concentration." In addition to picking up technical effects, genes that are stable in concentration could also pick up and adjust for effects that are associated with cell size.

Yet another notion of a stably expressed gene would be that the gene is expressed at a constant proportion with respect to the total RNA content of the cell from all genes. We refer to this as "stable in proportion to total RNA content," or, when it is clear from context, simply as "proportionally stable." In practice, we expect that sets of genes that are stable in concentration will be similar to sets of genes that are proportionally stable, provided that total RNA scales with cell size; in that case, both are likely to pick up cell size effects.

The notions of stability described above are biological in nature; they make no reference to *measurements* of gene expression, such as those provided by scRNA-seq data. In a hypothetical, extremely high quality dataset, these biological notions of stability would map clearly to features of the data. An absolutely stable gene would have a small coefficient of variation in terms of raw counts; a proportionally stable gene would have a small coefficient of variation after dividing the raw counts by total cell count.

Real data, however, is subject to technical factors that strongly affect the observed counts. In particular, some factors, like reaction efficiency, have a strong global effect on all counts for a given cell, effectively introducing a random scaling factor for each cell. Thus in real data, genes that are absolutely stable will not necessarily appear particularly stable in terms of their raw counts.

To further clarify these ideas, consider the following model:

$$y_{ij} \sim \text{Pois}(e_{ij} t_i^{\mathbb{I}(j \in \text{endogenous genes})} u_i^{\mathbb{I}(j \in \text{spike-ins})} v_j w_{ij}). \tag{2.1}$$

Here y_{ij} represents the measured expression (read count) for gene j in cell i and e_{ij} the true expression (number of transcripts) of gene j in cell i. Technical factors that affect endogenous genes and spike-ins are included in the terms t_i and u_i, respectively. A gene-specific scaling factor v_j could capture factors including amplification biases, and w_{ij} is an error term. Vallejos *et al.* [2017] provide a model for the expected read count as a function of the expression level and a cell-specific scaling factor that incorporates features such as capture efficiency and amplification efficiency (Box 1). Our model here is similar, with the expected read count corresponding to our Poisson parameter above, except that we have explicitly included one cell scaling factor for endogenous genes (t_i) and another for spike-ins (u_i). Some cell-specific factors will only affect endogenous genes (e.g., lysis efficiency), while others will affect only spike-ins (e.g., minor variations in the quantity spiked-in). Other factors (e.g., amplification, sequencing) are experienced by both endogenous genes and spike-ins and will therefore appear in both t_i and u_i. Note, importantly, that for the spike-ins to be useful for normalization purposes, the t and u should be reasonably correlated.

We can represent cell totals in terms of Model 2.1. Specifically, let the endogenous total of a cell be $E_{i,\text{endo}} = \sum_{k \in \text{endogenous}} e_{ik}$. Likewise, let $Y_{i,\text{endo}} = \sum_{k \in \text{endogenous}} y_{ik}$. Similarly, we denote the ERCC total expression as $E_{i,\text{ERCC}} = \sum_{k \in \text{ERCC}} e_{ik}$ and the measured total expression as $Y_{i,\text{ERCC}} = \sum_{k \in \text{ERCC}} y_{ik}$.

The notions of stability defined above map onto terms in Model 2.1. A gene j that is absolutely stable would appear with a small coefficient of variation in e_j, the true expressions of that gene. However, y_j may be quite variable due to the variation in t, obscuring the absolutely stable expression from e_j.

Additionally, we can illustrate proportional stability from Model 2.1. We denote the proportional expression of a gene as $p_{ij} = e_{ij}/E_{i,\text{endo}}$, or the expression of gene j in cell i divided by the total expression of cell i. Genes that are proportionally stable would appear with a small coefficient of variation in the vector p_j. The cell scaling factor t_i in Model 2.1 could plausibly be roughly proportional to the cell volume or $E_{i,\text{endo}}$. Thus, genes with small coefficient of variations in $y_{ij}/Y_{i,\text{endo}}$ could be approximately proportionally stably expressed or stable in concentration.

Normalizing the raw counts by the total cell count, $Y_{i,\text{endo}}$ can help adjust for the scaling factor t in Model 2.1, and such normalizations are common (e.g., RPKM is a variant of this[1]). After normalizing by total cell count, the notion of stability that is most relevant is proportional stability. That is, genes that appear stable in the normalized data would be those genes that are proportionally stable, and genes that are in truth absolutely stable would not necessarily appear stable in the data. Indeed, for this reason – that normalization by total cell count is effectively necessary to adjust for cell-specific scaling factors, but that normalization by cell total also obscures absolutely stable expression – absolutely stable genes are especially difficult to identify. Note also that similar comments apply to efforts to discover stably expressed genes in bulk tissue.

Importantly, note that a stably expressed gene can be viewed as the opposite of a differentially expressed gene or a highly variable gene. That is, the notion of stability, whether implicit or explicit, also implies the notion of instability or variability. In practical terms, the normalization that is applied to the data may not simply "clean" the data, but also alter the biological interpretation of the data, and determine which biological questions can (and cannot) be answered by the data. For example, normalizing the data against a set of absolutely stable genes would allow one to identify "absolutely differentially expressed genes," while normalizing the data against a set of proportionally stable genes would allow one to identify "proportionally differentially expressed genes." These two sets of differentially expressed genes could be quite different. Thus, finding sets of genes that exhibit different notions of stability would allow for different types of normalization, which provide different biological insights.

[1]RPKM is defined as [(gene count) / (total cell count)] * [1,000,000 / (length of gene in KB)]. The [1,000,000 / (length of gene in KB)] term is a rescaling to aid interpretability and adjust for gene length. For a given gene this rescaling is identical across cells, and thus a gene that appears stable in terms of [(gene count) / (total cell count)] will also appear stable in terms of RPKM.

2.3.2 Localization of Stable Genes

The final product of a gene (protein, ribosomal RNA, etc.) is often localized to specific structure(s) within the cell, e.g., nucleus, cell membrane, etc. We may therefore associate a gene with the location(s) where that gene's final product(s) are active.

We hypothesize that certain structures may be enriched with genes that are absolutely stable, while other structures may be enriched with genes that are stable in concentration. For example, because each cell has one nucleus, there may be a set of nuclear genes that are constant in absolute expression. In contrast, there may be a set of genes enriched in the cytosol that are reasonably constant in concentration.

The hypothesis that structures may be important for identifying stably expressed genes motivates our analysis. We create gene sets for each cell structure and assess the stability of each gene set as a whole.

2.4 Methods

2.4.1 Mapping Genes to Cell Structures

The Gene Ontology Consortium maintains a database that specifies the cellular component(s) with which each gene's final product, often a protein, is associated [Xin *et al.*, 2016; Wu *et al.*, 2013]. Many of the annotations provided by the Gene Ontology Consortium are highly detailed (e.g., "mitochondrial respiratory chain complex I"); we coarsen the 1,629 unique cellular components into ten categories corresponding to major cellular structures; see Section 6.2.1 for details. The ten categories are: nucleus, endoplasmic reticulum, Golgi body, cytoplasm, membrane, ribosome, mitochondria, mitochondrial ribosome, ribonucleoprotein complex, and cytosolic ribosome. The Gene Ontology Consortium returns zero to 41 cellular components for each gene we consider. Therefore, we allow a single gene to be associated with more than one cellular structure, and some genes are not associated with any structures. We will refer to the sets of genes as "nuclear genes," "cytoplasm genes," etc. Thus, nuclear genes are those genes whose final products are associated with the nucleus or some subpart of it, etc.

Table 2.1 shows the distribution of the cell structures amongst different gene sets within our scRNA-seq data. Note that genes can be associated with more than one structure; while the cytosolic ribosomal genes are associated with the cytosolic ribosome, they are not exclusively associated with that structure. We also show the distribution of genes of Eisenberg and Levanon [2003] and of Lin *et al.* [2019a].

Table 2.1: Distribution of Cell Structures Amongst Different Sets of Genes

Structure	All Genes	E & L	Lin et al.	Cyto. Ribo.
Number of Genes	8,605	470	967	103
Nucleus (4,280)	50%	57%	70%	68%
Endoplasmic Reticulum (934)	11%	14%	10%	16%
Golgi Bodies (850)	10%	11%	9%	2%
Cytoplasm (2,600)	30%	36%	32%	32%
Ribosome (177)	2%	8%	8%	100%
Mitochondria (1,187)	14%	22%	14%	13%
Mitochondrial Ribosome (75)	1%	1%	1%	4%
Ribonucleoprotein (185)	2%	5%	6%	17%
Membrane (3,160)	37%	52%	36%	64%
Cytosolic Ribosome (103)	1%	7%	4%	100%

Table 2.2: C1 Data Information

GEO sample	Author	Tissue type	Number cells	Number genes
77288	Tung *et al.* [2017]	Stem cells	2,592	20,425
84686	Das *et al.* [2017]	T cells	96	25,462
79130	Arguel *et al.* [2017]	Kidney	47	18,730
89235	Arguel *et al.* [2017]	Kidney	76	17,594
89236	Arguel *et al.* [2017]	Nasal epithelium	96	16,160
89237	Arguel *et al.* [2017]	Nasal epithelium	96	15,670

2.4.2 Expression Data

2.4.2.1 Fluidigm C1 Data

We downloaded six datasets containing human cells processed by three different lab groups and from four different tissue types from Gene Expression Omnibus [Edgar *et al.*, 2002; Arguel *et al.*, 2017; Das *et al.*, 2017; Tung *et al.*, 2017]. We selected these datasets because they were conducted with the Fluidigm C1 platform with human cells and included ERCC spike-ins. Note that four of these datasets were generated by one lab while performing an experiment [Arguel *et al.*, 2017]. Note also that all but one of these datasets make use of unique molecular identifiers (UMIs) [Kivioja *et al.*, 2012]. We filter to the genes that are expressed with at least one UMI in all datasets to ensure that genes are expressed in a wide variety of tissue types and to compare the datasets more directly. Table 2.2 provides general information on each of these datasets.

One cell was removed from each of GSE89236 and GSE89237. These cells had two and zero

Table 2.3: 10x Data Information

Sample	Tissue type	Number cells	Number genes
Fresh 68k PBMCs	Blood	68,579	20,387
AML027 Pre-transplant BMMCs	Bone marrow	3,933	15,705
Jurkat Cells	T cells	3,258	17,753
293T Cells	Kidney (embryonic)	2,885	18,760
293T Cells from mixture	Kidney (embryonic)	482	16,715
3T3 Cells from mixture	Fibroblast	538	13,573

UMIs measured, respectively.

GSE77288 reported genes by Ensembl identifiers, which we transformed to gene names with getGenes from `mygene` to match the other five datasets [Wu *et al.*, 2013; Xin *et al.*, 2016]. We removed genes that could not be identified with a gene name, and summed Ensembl identifiers associated with the same gene name.

2.4.2.2 10x Data

Additionally, we downloaded data from five experiments generated on the 10x platform from the 10x website [Zheng *et al.*, 2017]. We selected datasets that appeared to have a large number of cells and a fairly high sequencing depth, providing additional tissue types and heterogeneity. Note that all datasets consist of human cells except the 3T3 cells, which are mouse cells. Note also that Jurkat, 293T, and 3T3 cells are cell lines, and the PBMC and BMMC cells were obtained from live samples. We analyzed the Gene / cell matrix (filtered) using gene names for the specified datasets, including only those genes with at least one read. Table 2.3 summarizes characteristics of the 10x datasets.

For the Fresh 68k PBMCs dataset, we randomly sampled 20,000 cells to make the calculations more computationally tractable. Additionally, we subsetted the genes to those that were considered in our Fluidigm C1 analyses and expressed in each of the four 10x datasets, resulting in 8,172 genes.

2.4.3 Zero Counts

Sometimes called dropouts, zero counts occur when no copies of a given gene's transcripts are measured for a specific cell. Zero counts are a prominent feature of scRNA-seq data. For the six Fluidigm C1 datasets, the proportion of zeros ranges from 41% to 82% of the entries. Ideally, any stably expressed genes that we discover would have a low proportion of zero counts. Figure

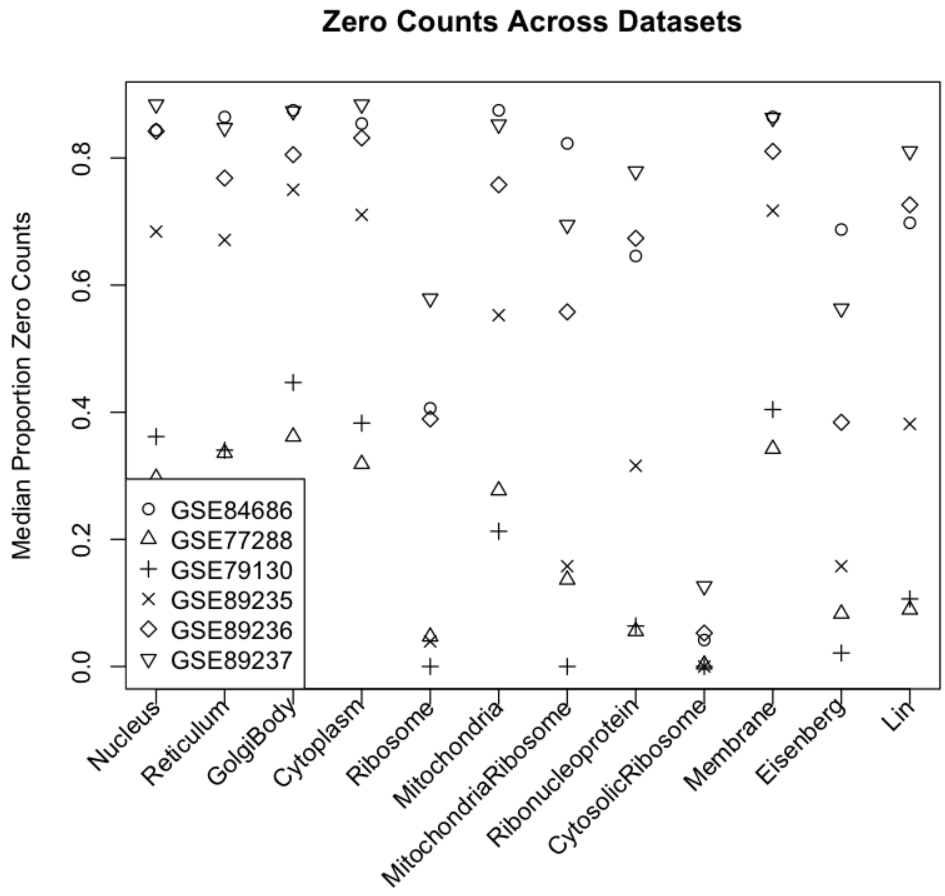

Figure 2.1: The median proportion of zero counts within each gene set, with symbols representing the different Fluidigm C1 datasets. Cytosolic ribosomal genes have an especially low proportion of zero counts across datasets.

2.1 displays the proportion of zero counts for each of the structural gene sets from each dataset. The proportion of zero counts is smaller for the ribosomal genes and dramatically smaller for the cytosolic ribosomal genes.

2.4.4 Absolute Stability

Genes that are absolutely stable would ideally appear stable in the data in terms of raw counts; however, as noted in Section 2.3, due to technical factors that result in strong variation in library size from well to well, simply looking for genes that are stable in terms of raw counts is not a feasible way to discover absolutely stable genes. Instead, we leverage the stable absolute "expres-

sion" of the ERCC spike-ins and attempt to find absolutely stable genes by looking for those genes that have a high correlation with the ERCCs. We assume that the ERCCs have been appropriately spiked in at a fixed quantity for each cell, so that the ERCC "expression" is absolutely stable across cells. While technical factors may impede this assumption, the ERCCs provide the best option for a known absolutely stable quantity.

More specifically, for each of the six Fluidigm C1 datasets we perform the following analysis. We begin with raw counts of UMIs. We transform the raw counts by log + 1. In addition, for each cell we also compute the sum of the raw UMIs of all ERCCs; we refer to this as the "ERCC total." Finally, for each gene, we compute Pearson's correlation (across all cells) between that gene's (log + 1) expression and the log of the ERCC total.

Ordinarily, one would expect the "ERCC total" $Y_{i,\text{ERCC}}$ to be approximately stable across all cells, because the ERCCs are spiked-in at a constant level (i.e., $E_{i,\text{ERCC}}$ is approximately constant). Thus, the variability among the ERCC measurements should ideally be small, rendering the gene-ERCC total correlations difficult to interpret. Indeed, this would be the case if we were using the true ERCC totals $E_{i,\text{ERCC}}$ rather than the measured ERCC totals $Y_{i,\text{ERCC}}$. However, since the ERCC measurements are subjected to the experiment and its corresponding technical biases (e.g., u in Model 2.1), there is additional variability introduced to the ERCC totals. In fact, it appears that the factors captured by u are quite large. The correlation between a gene and the ERCC total is approximately the correlation between $\log(u_i) + \log(E_{i,\text{ERCC}})$ and $\log(t_i) + \log(e_{ij})$, ignoring w_{ij} and the Poisson variability in Model 2.1 and assuming the +1 transformation is negligible. Therefore, high correlations with the ERCC totals indicate that a given gene appears to be absolutely stably expressed assuming t and u are correlated. In contrast, a low correlation with the ERCC totals indicate that a given gene is not absolutely stably expressed or t and u are uncorrelated.

We then summarize the correlations by finding, for each gene, the mean of the correlations across the six datasets. Thus, for each gene, we now have an average correlation of that gene's expression with the ERCC total, and we regard this as a measure of that gene's absolute stability. To see if any cellular structures are enriched for stably expressed genes, we plot histograms of the correlations subsetted by structure, and inspect these histograms to see if the gene sets from any structures have especially high correlations.

2.4.5 Proportional Stability

We also attempt to find genes that are proportionally stable. Our method is similar to the one for absolute stability, but we replace the ERCC total with a total of endogenous genes. For a given structure S, we sum a cell's overall raw measured expression after removing the set of all genes associated with structure S; we refer to this as an "adjusted cell total". For each gene, we compute

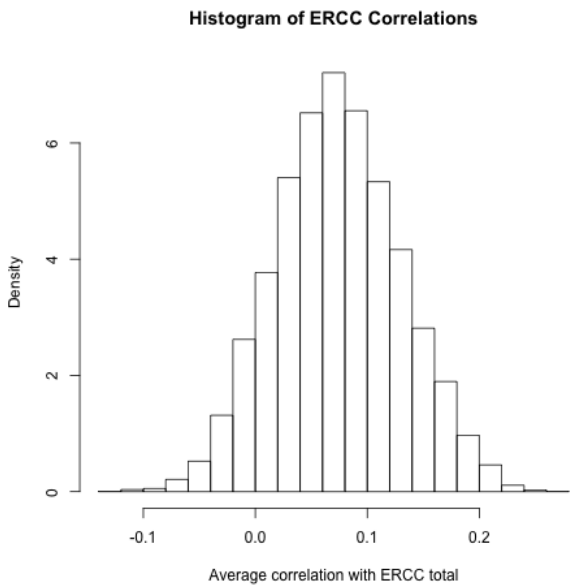

Figure 2.2: Histogram of the correlations of all genes with the ERCC totals. Figures 6.1 and 6.2 contain detailed histograms with separation based on the structure.

Pearson's correlation (across all cells) between that gene's (log + 1) expression and the log of the adjusted cell total.

Again, we summarize correlations by finding, for each gene, the mean of the correlations across the six datasets. For each structure, we plot a histogram of that structure's correlations with that structure's adjusted cell total.

The log-transformed cell total according to Model 2.1 is approximately $\log(t_i) + \log(E_{i,\text{endo}})$. Individual genes (with expected measured expression on the log scale of approximately $\log(t_i) + \log(e_{ij})$, as modeled by Model 2.1) with high correlations with the cell total indicate that the gene's measurements appear to be proportional to the measured cell total, ignoring Poisson variability. Conversely, genes with small gene-cell total correlations indicate poor proportional stability.

2.5 Results

2.5.1 Absolute Stability

Figure 2.2 shows a histogram of the correlations with the ERCC total; histograms of correlations separated by gene sets can be found in Figures 6.1 and 6.2. The most notable observation is that all correlations are smaller in absolute value than 0.3. One possible explanation for the weak correlations is that no gene is absolutely stable. However, an alternative explanation is that the ERCCs

and endogenous genes are affected by different technical artifacts, and we suspect in particular that the spike-ins are not exposed to some strong technical effect(s) that do affect the endogenous genes. In other words, it appears that the t and u from Model 2.1 are quite different, likely from the different capture and reverse transcription efficiency as described by Vallejos *et al.* [2017]. Figure 2.4 shows that, on the log scale, the endogenous cell totals and ERCC totals ($Y_{i,\text{endo}}$ and $Y_{i,\text{ERCC}}$, respectively) have a low correlation, further indicating that t and u are indeed different. As cells need to be lysed and RNA needs to be extracted from the cells, we believe that this may introduce technical factors which affect the measurements of endogenous genes while the spike-ins do not experience the same variability.

Supporting these claims, the spike-ins exhibit much lower variability than the endogenous genes within our datasets. We compare the ERCC cell total to the endogenous cell total on a log scale for each of the datasets (Figure 2.3). From Model 2.1, this approximately corresponds to comparing $\log(u_i) + \log(E_{i,\text{ERCC}})$ and $\log(t_i) + \log(E_{i,\text{endo}})$, respectively. If we suppose that the spike-ins are appropriately added to the experiment, then $E_{i,\text{ERCC}}$ should be stable. Similarly, since all of our cells are of the same type, it is reasonable to assume that $E_{i,\text{endo}}$ could be relatively constant. Following from these assumptions, Figure 2.3 allows us to compare the variability in u_i to t_i. We see that the variability for the ERCC measurements is considerably smaller than the variability for the endogenous gene measurements, sometimes by orders of magnitude.

Figure 2.4 displays the logged totals of the endogenous genes and the ERCC spike-ins. Following the modeling assumptions discussed above, the scatterplots allow us to estimate the size of technical effects unique to t_i and u_i relative to any shared technical effects, like amplification and sequencing biases. If the shared technical effects contained within t_i and u_i were much larger than those unique to the endogenous genes or spike-ins, then the scatterplots would have strong, linear relationships. However, the scatterplots show poor correlations, indicating that the technical factors affecting the endogenous genes and spike-ins separately dominate the cell scaling factors. The variability of the ERCC measurements appear to capture different technical effects than those that affect the endogenous genes. Figure 2.5, which examines the mean and standard deviation of each of the endogenous genes and ERCC spike-ins separately, demonstrates that, again, the ERCC spike-ins have much lower variability when compared to endogenous genes with similar overall expression.

For each gene set, we plot the mean correlation with the total ERCC spike-ins from a dataset in Figure 2.6. One feature to note is that the mean correlation for GSE84686 is negative for each of the gene sets and appears substantially smaller than other five. This negative relationship is driven by an overall correlation of -0.28 between the log of the total UMIs and the log of the total ERCC spike-ins (Figure 2.4). We suspect that this relationship exists as a result of the ERCC measurements overpowering the transcripts for a given cell, especially with the high levels of ERCCs,

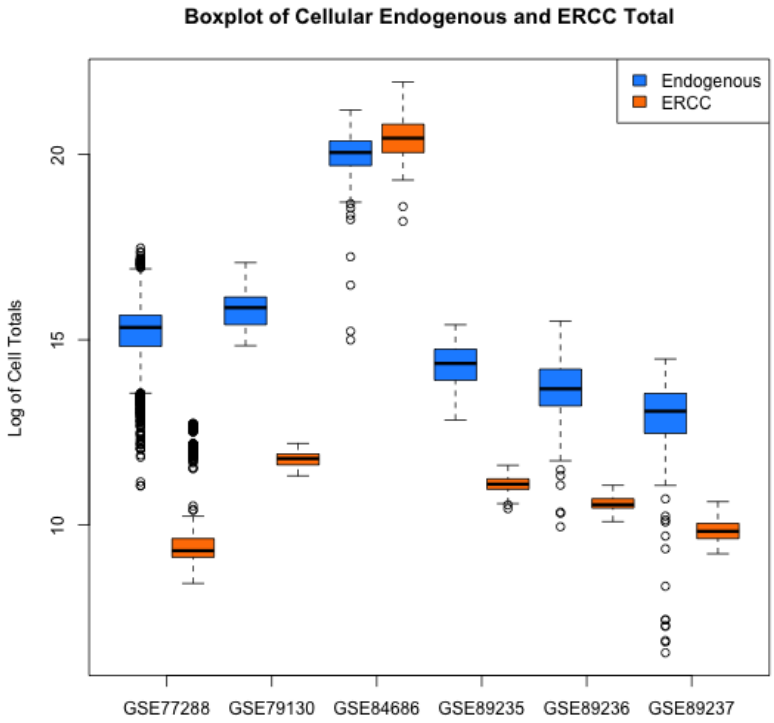

Figure 2.3: Boxplots of the ERCC and endogenous gene totals from each cell. The blue boxplots are of the log (base 2) transformed sum of the endogenous expression for each cell, while the orange boxplots are of the log-transformed sum of the ERCC measurements. Note that the range of endogenous totals is often several units on the log scale, indicating variability that spans orders of magnitude. Specifically, for a cell total at 2^{15}, the associated standard deviation from Poisson sampling would extend only 0.008 on the log-scale. The expected variation from the Poisson sampling is dwarfed by the overall variability of the cell totals. Note also that GSE84686 is an outlier in many respects.

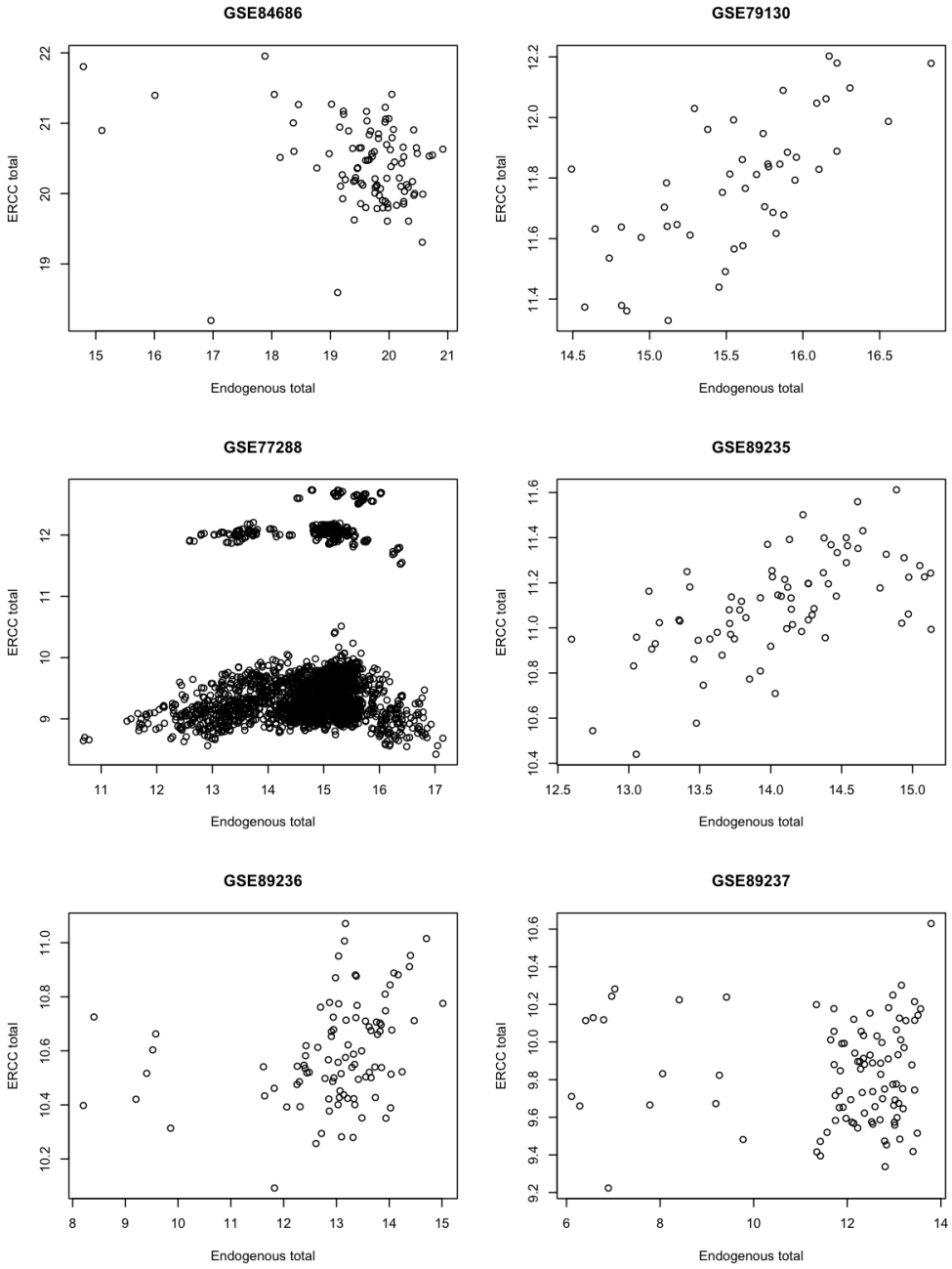

Figure 2.4: Scatterplots of the endogenous gene and ERCC totals, on the log (base 2) scale, from each cell. Note that GSE84686 has a negative correlation. Note also that the ERCC measurements from GSE77288 vary, consistent with the original publication that an experimental error occurred [Tung *et al.*, 2017].

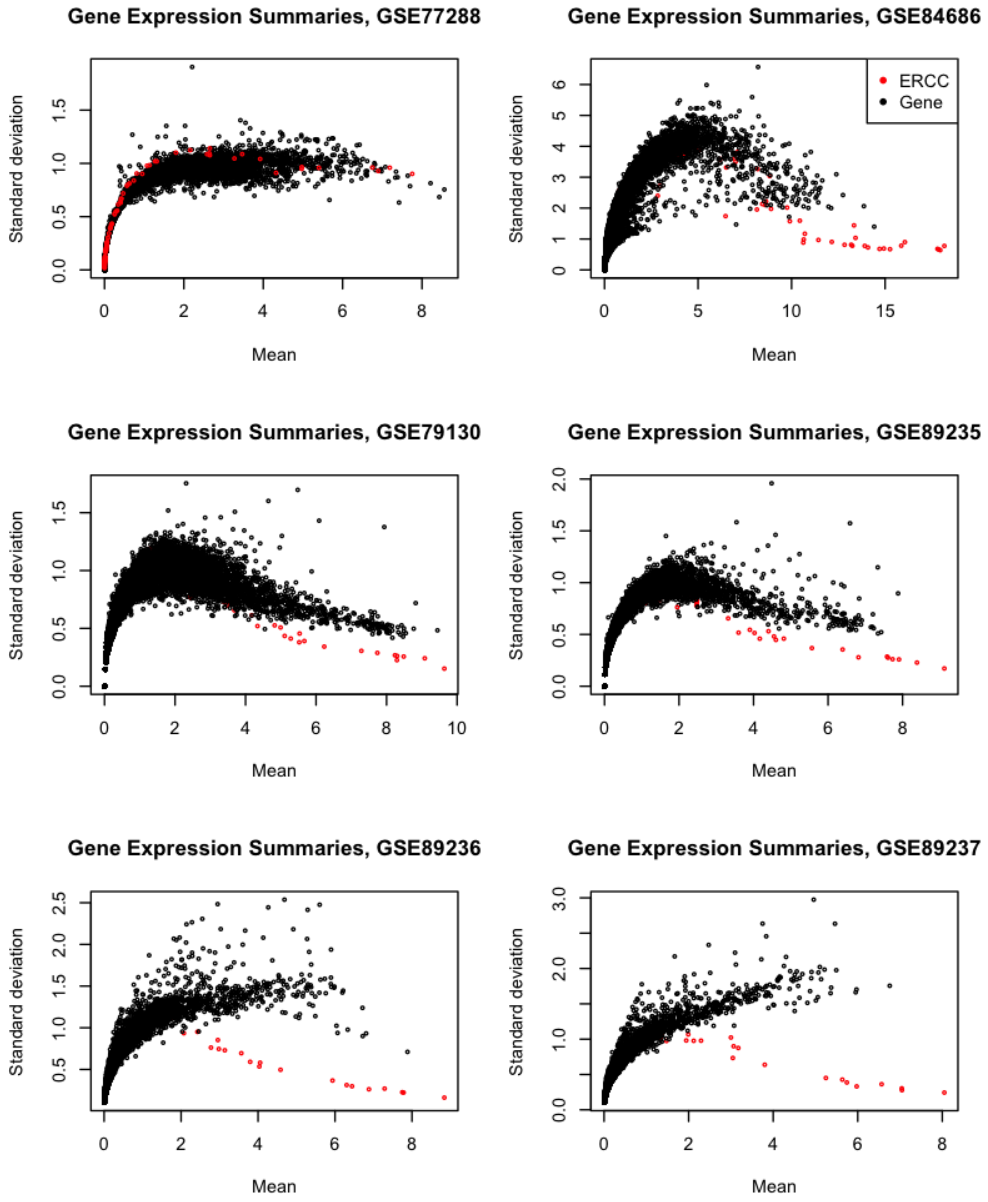

Figure 2.5: Scatterplots of the mean and standard deviation of ERCC spike-ins and the endogenous genes on the log scale for each of the six datasets. From these graphs, it appears that the variance exhibited by the spike-ins is smaller than those for endogenous genes, after accounting for the mean measured expression.

as seen in Figure 2.3. We see that highly expressed sets of genes, including the cytosolic ribosomal and ribosomal genes, have the lowest mean correlations in GSE84686, further supporting this hypothesis.

Overall, the spike-ins appear to have measurements that are far more stable than the endogenous genes. Since each dataset captures one type of cell, and because the range of endogenous cell totals is so large (orders of magnitude), we suspect that technical effects likely contribute to much if not most of the variation observed in the endogenous totals. Moreover, the technical factors affecting the endogenous cell totals appear stronger than technical factors affecting the spike-ins, suggesting in particular that the spike-ins may not be capturing the same technical effects. If so, using spike-ins to identify genes that are absolutely stably expressed is inappropriate.

Model 2.1 further supports the interpretation that the technical factors that affect the ERCC spike-ins are different than those that affect the endogenous genes. Specifically, there are some experimental processes that differ for spike-ins. Spike-ins are incorporated into the experiment in a different manner and sometimes at a different time point than endogenous genes, which introduces experimental inconsistencies in protocols, like being pipetted in at a specific volume and not being lysed from a cell. Additionally, the spike-ins have synthetic characteristics, including their manufactured nature and their varying properties designed to capture technical effects. Finally, the amounts with which the spike-ins are introduced into the experiment differ. The spike-ins are often incorporated at an amount inconsistent with the expression level of the endogenous genes. Any factors affected by overall expression level would disproportionately affect the spike-ins. The differences between the spike-ins and the endogenous genes are represented by t and u in Model 2.1 and appear to be different from the data. The endogenous technical effects may not be captured well by the spike-ins. Therefore, spike-ins are unlikely to be helpful for removing the technical effects from the endogenous measurements.

2.5.2 Proportional Stability

We examine the measures of proportional stability in Figures 2.7, 6.3, and 6.4. Unlike the correlations with the ERCC measurements, we see large correlations, with values ranging from -0.03 to 0.95. The cytosolic ribosomal genes appear especially enriched in proportionally stable genes.

We plot the mean correlation of each gene set with the unadjusted cell total by dataset (Figure 2.8). The spread and the ordering of the datasets are similar for each structure. Thus, it appears that mean correlations are an appropriate summary for a gene's correlation across the six datasets.

Cytosolic ribosomal genes are typically highly expressed. We create a set of highly expressed single cell genes to assess how high expression affects stability. We find the top 100 highly expressed genes in each of the six Fluidigm C1 datasets. The union of these six lists produces a

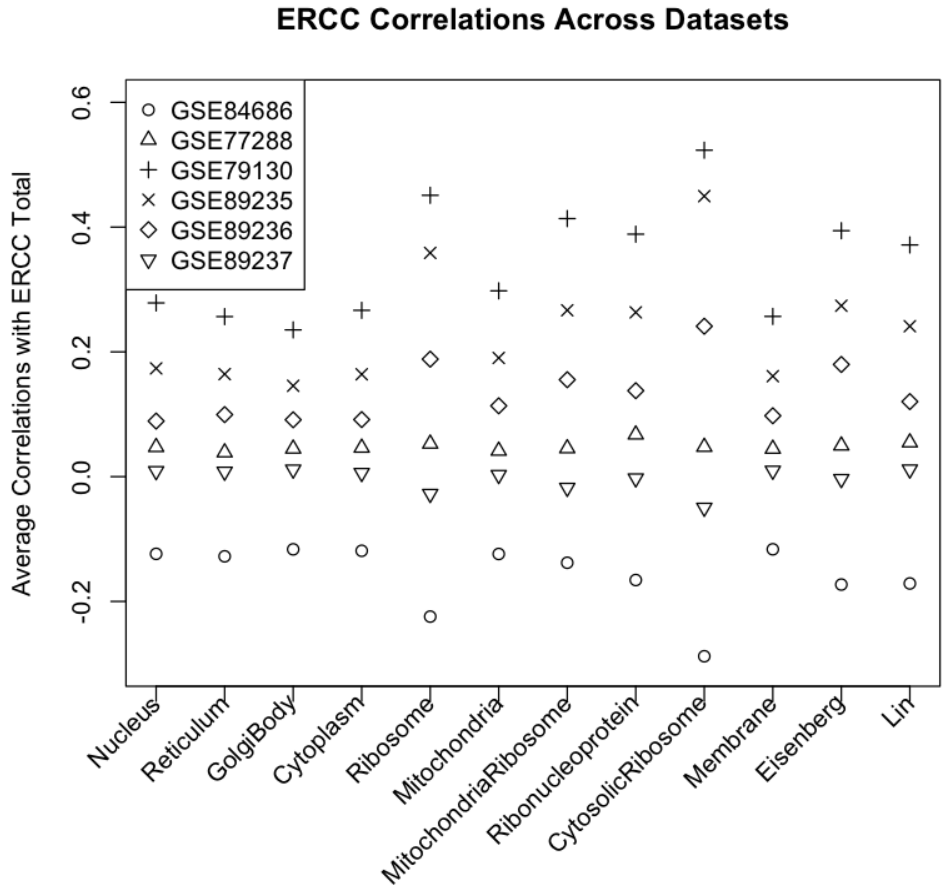

Figure 2.6: For each set of genes, we calculate the mean correlation with the ERCC total, on the log scale. The mean correlations are shown here for each of the six Fluidigm C1 datasets. These correlations exhibit a clear pattern based on dataset, indicating variation by dataset in the measures of absolute stability. The patterns of the distribution are similar across the structures, indicating that the average correlations is an appropriate summary of the six datasets.

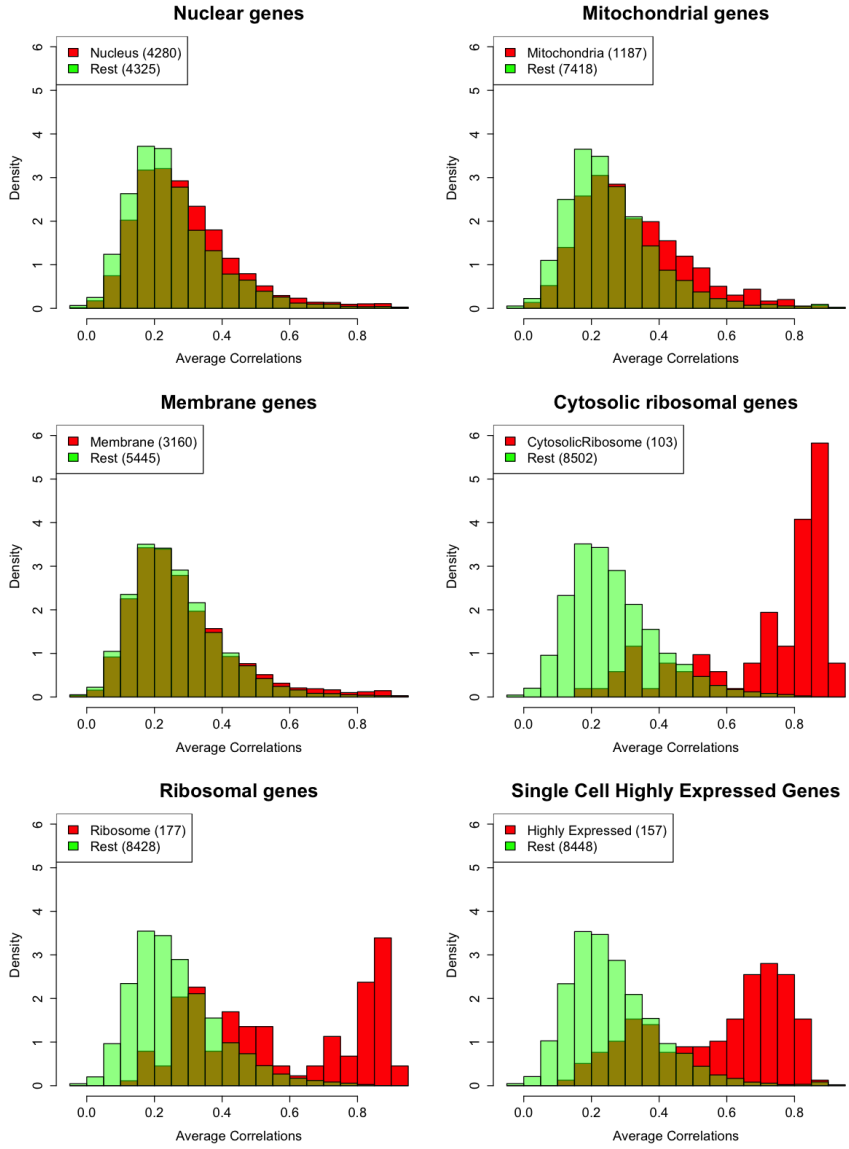

Figure 2.7: Histograms of the average correlations of each gene over the six Fluidigm C1 datasets considered, with comparisons between the set of genes of interest (red) and the remaining genes (green). Only some of the structures are shown here; the remaining structures can be found in Figures 6.3 and 6.4. The highly expressed genes (bottom right panel) consist of the union of the top 100 genes by average expression, excluding the cytosolic ribosomal genes. The correlations plotted here are averages across the six Fluidigm C1 datasets, but no single dataset seems to be contributing to the averages differently across the gene sets; see Figure 2.8.

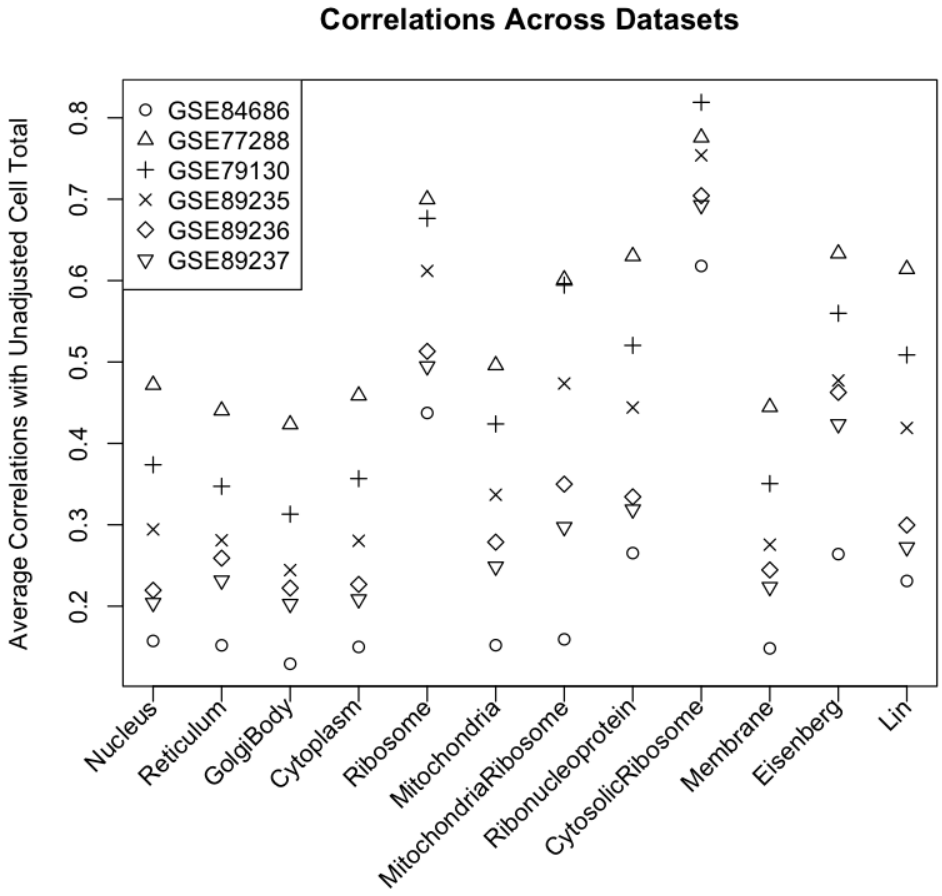

Figure 2.8: For each set of genes, we calculate the mean correlation with the unadjusted cell total. These mean correlations are shown here for each of the six Fluidigm C1 datasets. The patterns of the distribution are similar across the structures, indicating that the average correlations is an appropriate summary of the six datasets.

Table 2.4: Distribution of Cell Structures Amongst Highly Expressed Genes

Structure	All Genes	Highly Expressed
Number of Genes	8,605	157
Nucleus (4,280)	50%	66%
Endoplasmic Reticulum (934)	11%	11%
Golgi Bodies (850)	10%	5%
Cytoplasm (2,600)	30%	39%
Ribosome (177)	2%	0%
Mitochondria (1,187)	14%	27%
Mitochondrial Ribosome (75)	1%	0%
Ribonucleoprotein (185)	2%	9%
Membrane (3,160)	37%	43%
Cytosolic Ribosome (103)	1%	0%
Eisenberg & Levanon (470)	5%	34%
Lin et al. (967)	11%	20%

This table is essentially an extension of Table 2.1 to include the highly expressed single cell genes.

set of 231 genes. We remove the cytosolic ribosomal genes from this set, resulting in 157 highly expressed genes. The structure associations of the highly expressed single cell genes are shown in Table 2.4. We then perform the same correlation analysis as for the other gene sets, seen in the bottom-right panel of Figure 2.7. We can see that the cytosolic ribosomal genes are especially stable even relative to this set of other highly expressed genes.

Cytosolic ribosomal genes also have a low proportion of zero counts, but this does not seem to be driving the high correlations. We recomputed the correlation histograms after removing cells with zero counts for a given gene (Figures 2.9 and 6.5). Comparing these to Figures 6.3 and 6.4, we see a similar pattern in the correlations with the cytosolic ribosomal genes exhibiting the highest measures of proportional stability.

Ribosomal genes, from which the cytosolic ribosomal genes are a subset, have been previously identified as relatively stable. Thorrez *et al.* [2008] found that ribosomal genes were the most stable set of genes that they had encountered in microarray experiments, but Thorrez *et al.* [2008] also note that the ribosomal genes do exhibit tissue-dependent variation.

We compare the correlations of cytosolic ribosomal genes to those of GAPDH and Beta-actin, two genes that have previously been identified as stable genes. Both GAPDH and Beta-actin appear to have similar or smaller correlations as the cytosolic ribosomal genes (on average), indicating that the proportional stability is similar to two commonly used stable genes; the correlation for GAPDH is 0.82 and for Beta-actin is 0.75. For a comparison of correlations between these two genes and the cytosolic ribosomal genes, see Figure 2.10.

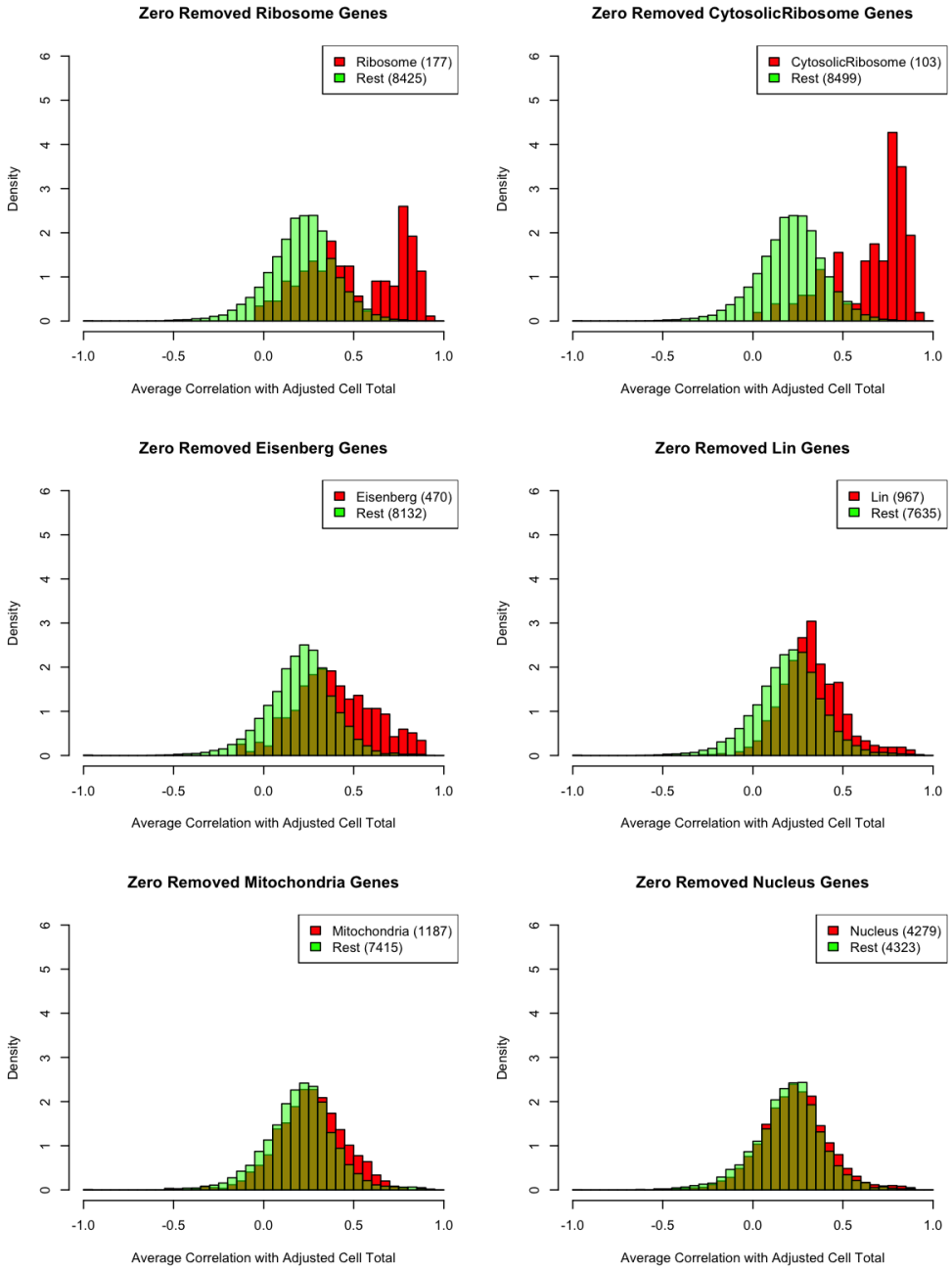

Figure 2.9: Histograms of the average correlations of each gene over the six Fluidigm C1 datasets considered, with comparisons between the set of genes of interest and the remaining genes. Cells with zero counts for individual genes were removed before calculating the correlations with the adjusted cell total. Continued in Figure 6.5.

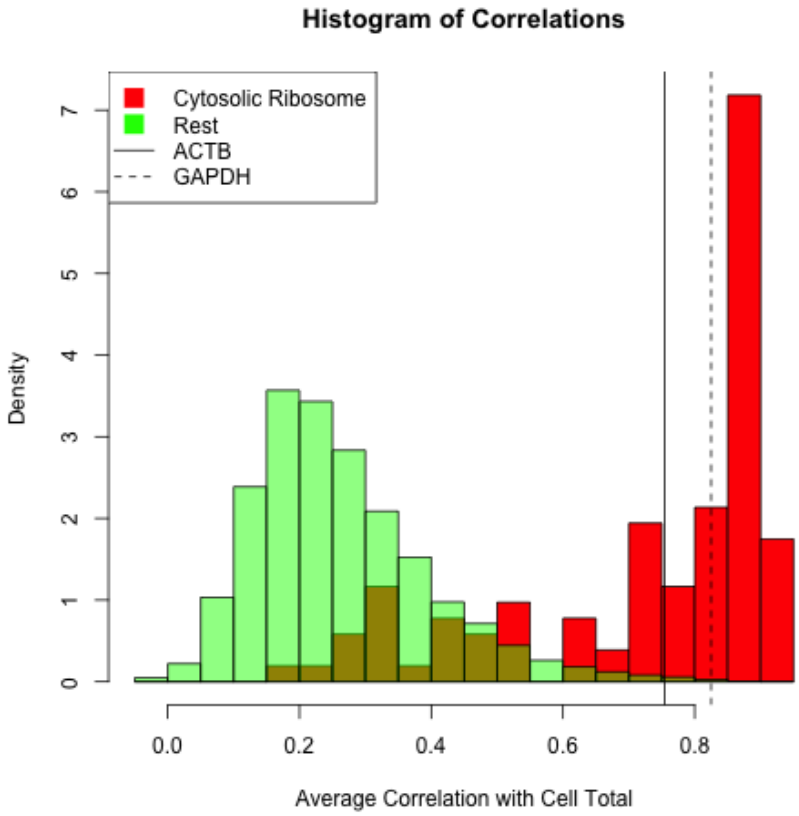

Figure 2.10: The average correlations of Beta-actin (ACTB) and GAPDH with the cell total over the six datasets. Note that, unlike Figure 2.7, the correlations here are with the unadjusted cell total rather than an adjusted cell total. This allows correlations between ACTB, GAPDH, and the cytosolic ribosomal genes to be comparable.

In addition to their high measures of proportional stability, we again note that the cytosolic ribosomal genes are highly expressed (Figure 2.11) and have a low proportion of zero counts (Figure 2.1), which further suggests they may be effective as reference genes.

2.5.3 Stability of Cytosolic Ribosomal Genes in the 10x Platform

2.5.3.1 Proportional Stability

We further evaluate the cytosolic ribosomal genes for proportional stability in datasets generated with the 10x platform [Zheng *et al.*, 2017], assessing the robustness of our results from the Fluidigm C1 platform. Using the same measures from the Fluidigm C1 data, the cytosolic ribosomal genes again appear to be enriched in proportionally stable genes, extending our findings of proportional stability to another platform (Figures 2.12 and 6.6).

Note that we mentioned that some characteristics of GSE84686 are unusual, including the measured ERCC spike-ins and the small correlations calculated with both the ERCC spike-ins and the biological cell totals. Both Jurkat cells and GSE84686 contain T cells. The size of the correlations obtained from the Jurkat cells are similar to those found in other datasets. In addition, measures of proportional stability for GSE84686 from the cytosolic ribosomal genes are higher than those measures for other gene sets. Together, these two facts suggest that the cytosolic ribosomal genes are proportionally stable, and that the size of the proportional stability measure for GSE84686 is likely influenced by other aspects of the data than by a lack of proportional stability of the cytosolic ribosomal genes.

2.5.3.2 Stability Across Species

The proportional stability of the cytosolic ribosomal genes seems to be robust based on species. We examine the additional 10x dataset that contains a mixture of human and mouse cells, 293T and 3T3 cells respectively. We did not create a full structural annotation dictionary for mouse genes as we did for human. Rather, we approximate the set of cytosolic ribosomal genes with those genes that begin with a prefix of either 'RPS' or 'RPL' in humans or with a prefix of 'Rps' or 'Rpl' in mice. We selected these genes based on the patterns observed in Section 6.2.2. Using this approximation, we capture 82 of 108 cytosolic ribosomal genes in the human 293T dataset and include an additional 15 genes that are not cytosolic ribosomal genes.

Using the approximated set of cytosolic ribosomal genes, we calculate the proportional stability measures with the correlation between genes and the adjusted cell total after removing the approximated set of cytosolic ribosomal genes. We repeat the same procedure for the 3T3 mice dataset. Figure 2.13 indicates that the proportional stability of cytosolic ribosomal genes, based on our approximated set of genes, appears similar in mice and humans.

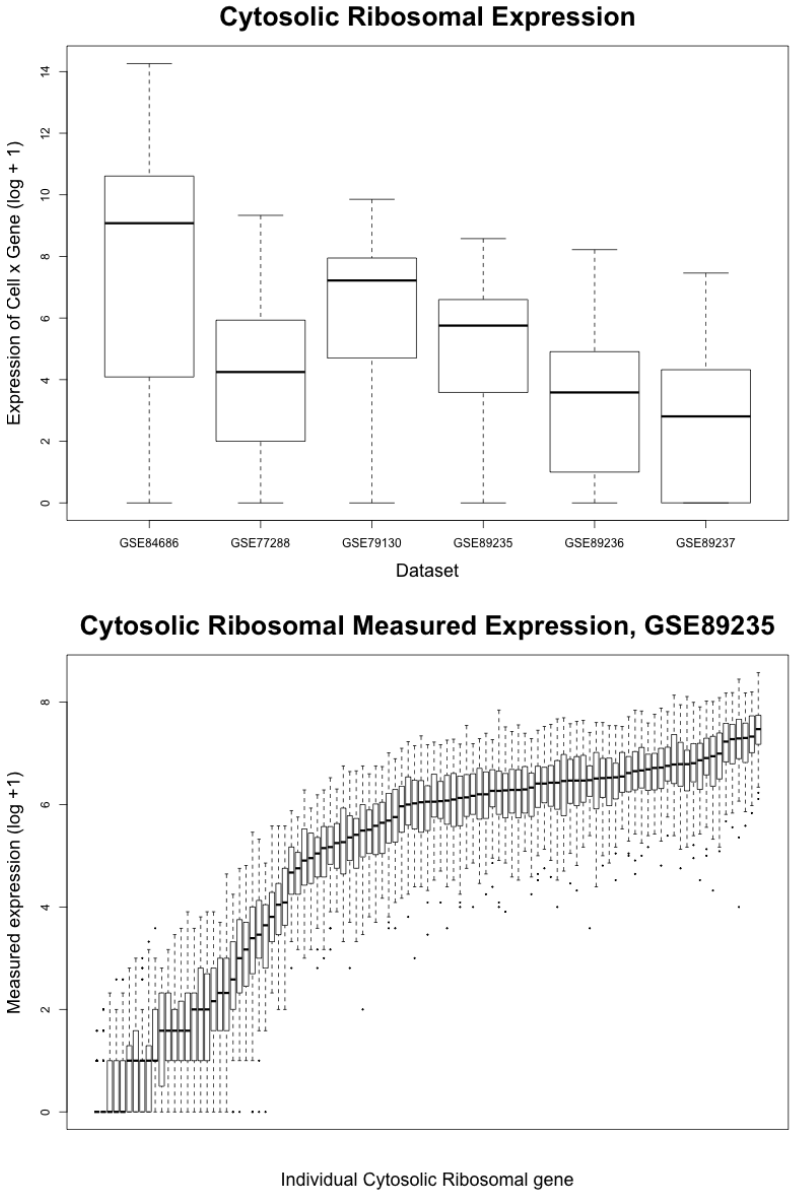

Figure 2.11: Summary characteristics of the expression of cytosolic ribosomal genes across cells and across datasets. The boxplot on the top shows all of the expression measurements by dataset for the cytosolic ribosomal genes, transformed with the log + 1. For comparison, the proportion of zeros in each of these datasets is 80%, 41%, 42%, 63%, 76%, and 82%, respectively and means (on the log + 1 scale) of 1.23, 1.26, 1.37, 0.71, 0.45, and 0.30, respectively. The boxplot on the bottom looks at the expression (log + 1) of each of the cytosolic ribosomal genes within GSE89235.

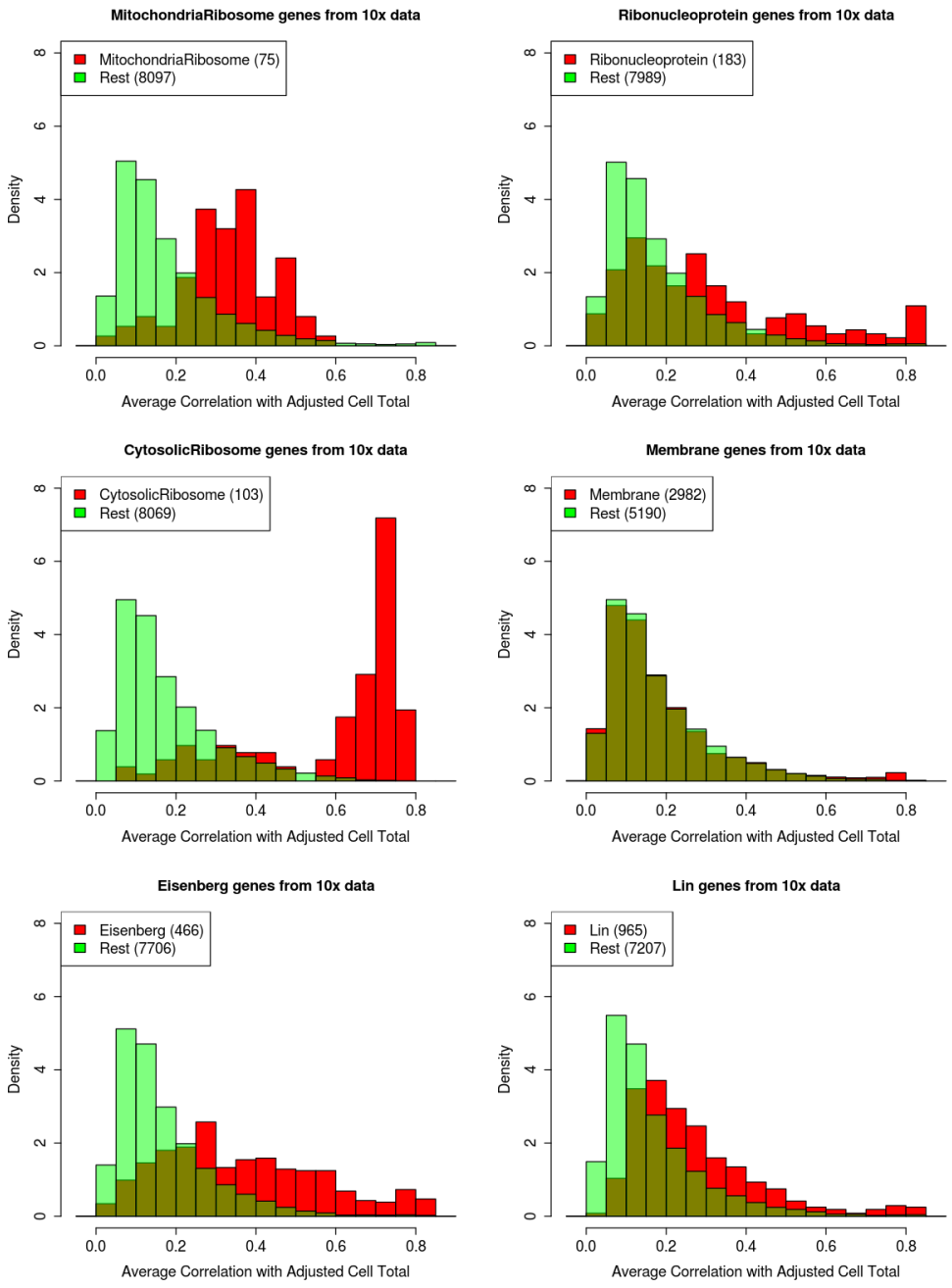

Figure 2.12: Histograms of the average correlations of each gene with the adjusted cell total over the four human 10x datasets, with comparisons between the gene sets of interest and the remaining genes. This figure is continued with Figure 6.6 with histograms for the other six gene sets considered.

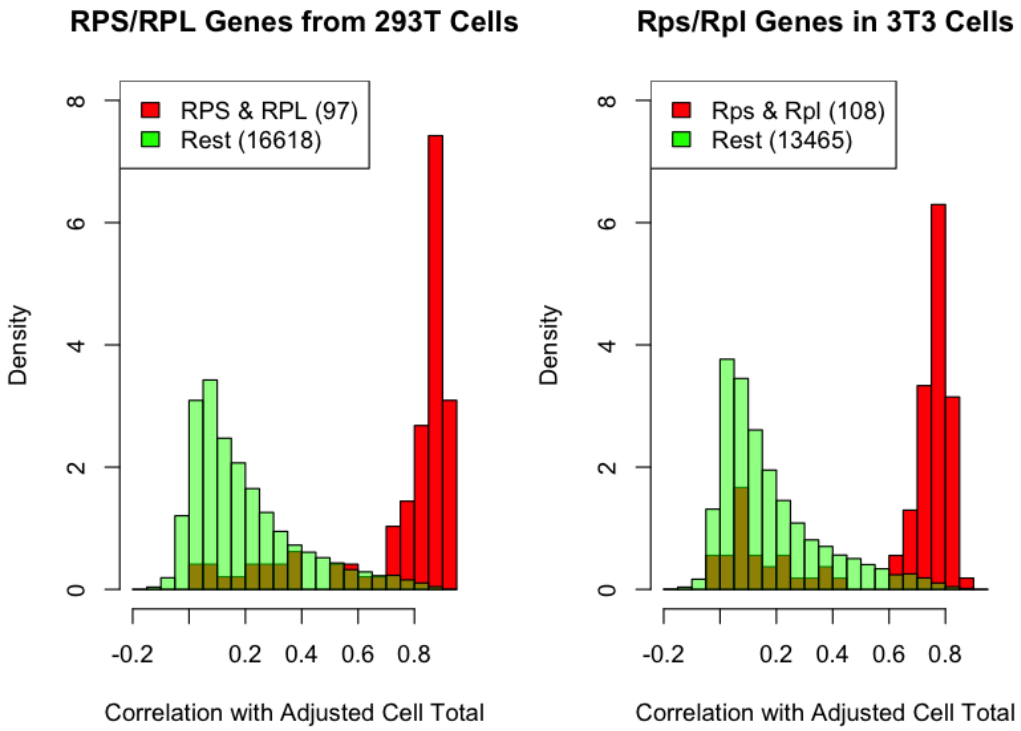

Figure 2.13: Comparison of correlation histograms of estimated cytosolic ribosomal genes in human and mouse. We approximated cytosolic ribosomal genes with those genes that contained a prefix of Rps or Rpl (for mouse cells) and RPS or RPL (for human cells). Based on the human gene information, these genes capture the set of cytosolic ribosomal genes fairly well.

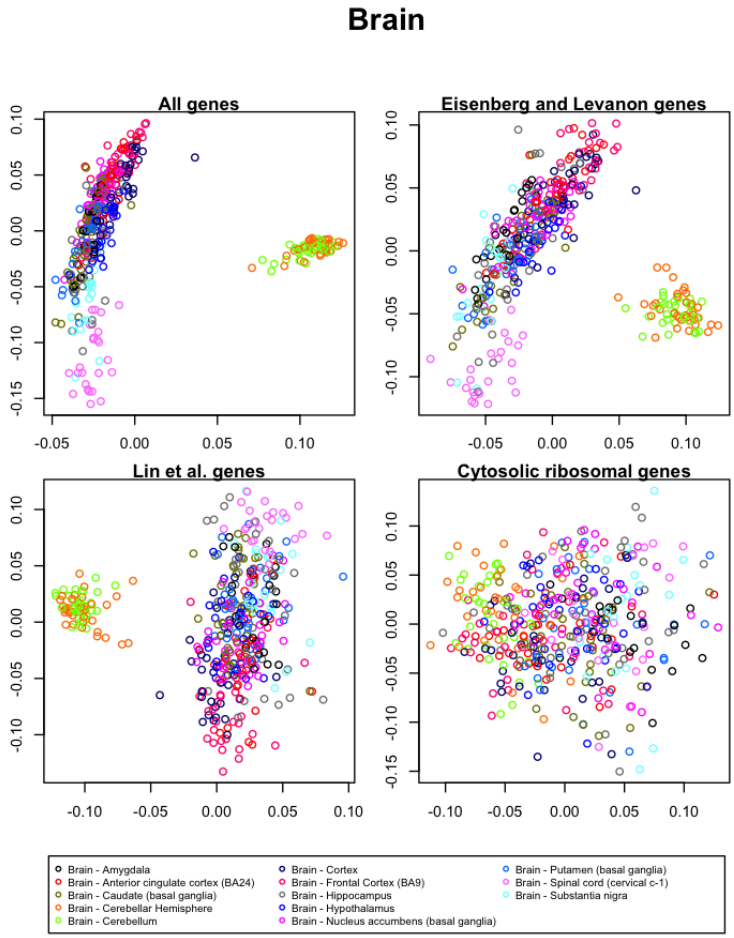

Figure 2.14: Singular Value Decomposition of GTEx data from the brain using different sets of genes. The overlap of the clusters denoted by the coloring indicate how stably expressed the genes are; the overlap is the strongest for the cytosolic ribosomal genes, indicating that they are the most stably expressed of the three options between the brain subtissue types. Note that the number of genes in each of the four SVD plots differ. More specifically, 89 cytosolic ribosomal genes are present in the available GTEx data, while more of the other three gene sets are present. To ensure that differences in the number of features are not accounting for differences in apparent differential expression, we repeat the analysis with random samples of 89 genes from each of the gene sets with similar results.

2.5.4 Stability of Cytosolic Ribosomal Genes in Bulk Tissues

To assess the stability of cytosolic ribosomal genes in bulk samples from many tissue types, we analyze data collected with bulk sequencing from the Genotype-Tissue Expression (GTEx) project [Carithers *et al.*, 2015]. The GTEx project systematically collects multiple tissue types from many individuals for genetic profiling and analysis. The tissue types are often subcategorized into addi-

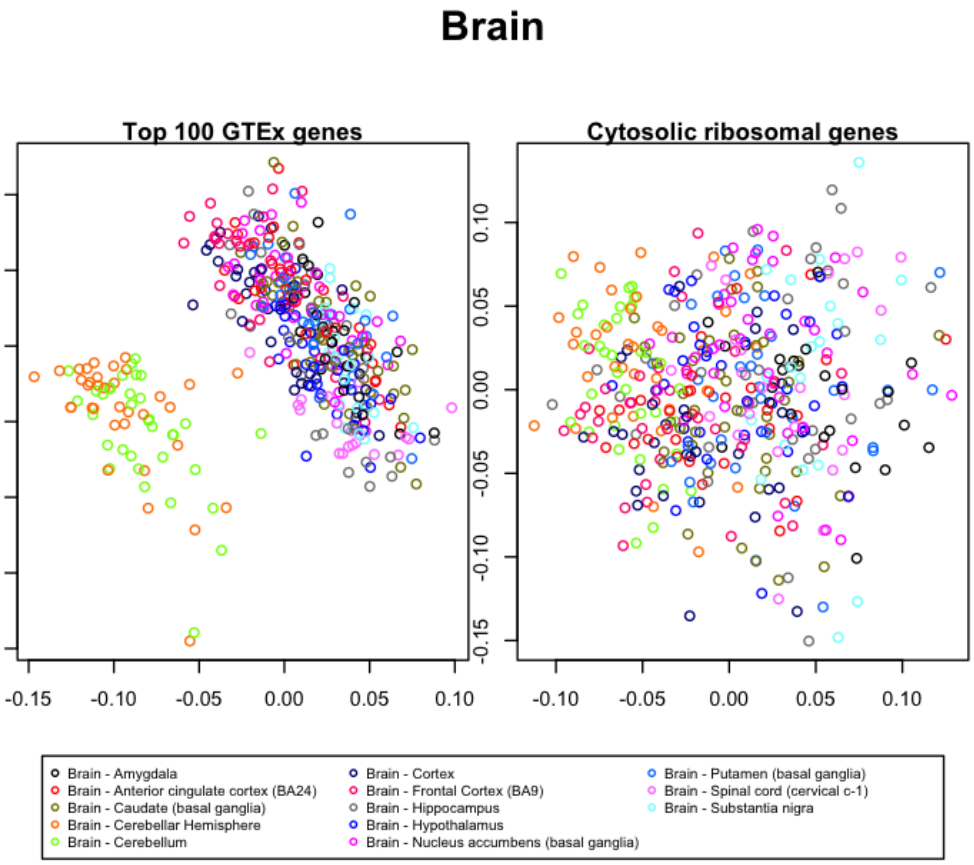

Figure 2.15: Singular Value Decomposition of GTEx data from the brain using the top 100 highly expressed GTEx genes and the cytosolic ribosomal genes. This figure extends the analysis in Figure 2.14 to incorporate the set of highly expressed genes.

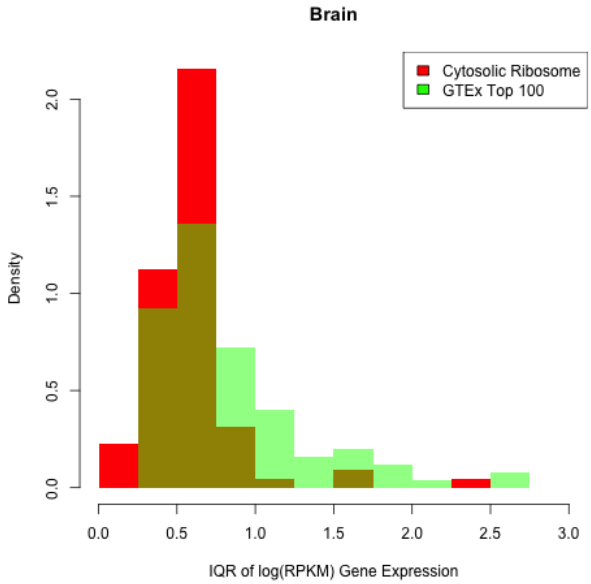

Figure 2.16: Histograms of the IQRs of cytosolic ribosomal genes and the top 100 highly expressed genes from the brain, with small IQRs indicating stable expression. Figures 6.11 to 6.17 show similar figures for additional tissue types and comparison gene sets.

tional subtissue types.

We examine the stability of different sets of genes by seeing how strongly their expressions vary across tissue type. More specifically, for a given set of genes, we compute the singular-value decomposition (SVD) and plot the first two singular vectors to see how strongly the samples cluster by tissue type. Prior to calculating the SVD, the GTEx samples are RPKM-normalized, transformed with log + 1, and centered by gene. In these plots, clustering by tissue type indicates that the genes examined have some differential expression between tissue types, whereas lack of clustering provides evidence of stable expression across tissue types. We examine the SVD plots to assess the stability of the cytosolic ribosomal genes compared to other sets of genes.

We perform the SVDs for both single tissues and combinations of two tissue types. We select the seven tissue types with the most samples (brain, skin, esophagus, blood vessel, adipose tissue, heart, and muscle); of these, six have subtissue types. We first plot the SVD for these six tissue types separately with coloring by subtissue type to examine how stably expressed genes are within a tissue type (Figure 2.14). We extend Figure 2.14 in Figure 2.15 by generating an SVD plot comparing the stability of the cytosolic ribosomal genes to highly expressed genes from the GTEx data. Our set of 100 highly expressed genes in the GTEx data are generated by calculating the top 151 genes according to their average expression of RPKM-normalized and log-transformed GTEx measurements and removing the 51 cytosolic genes within this set. The cytosolic ribosomal

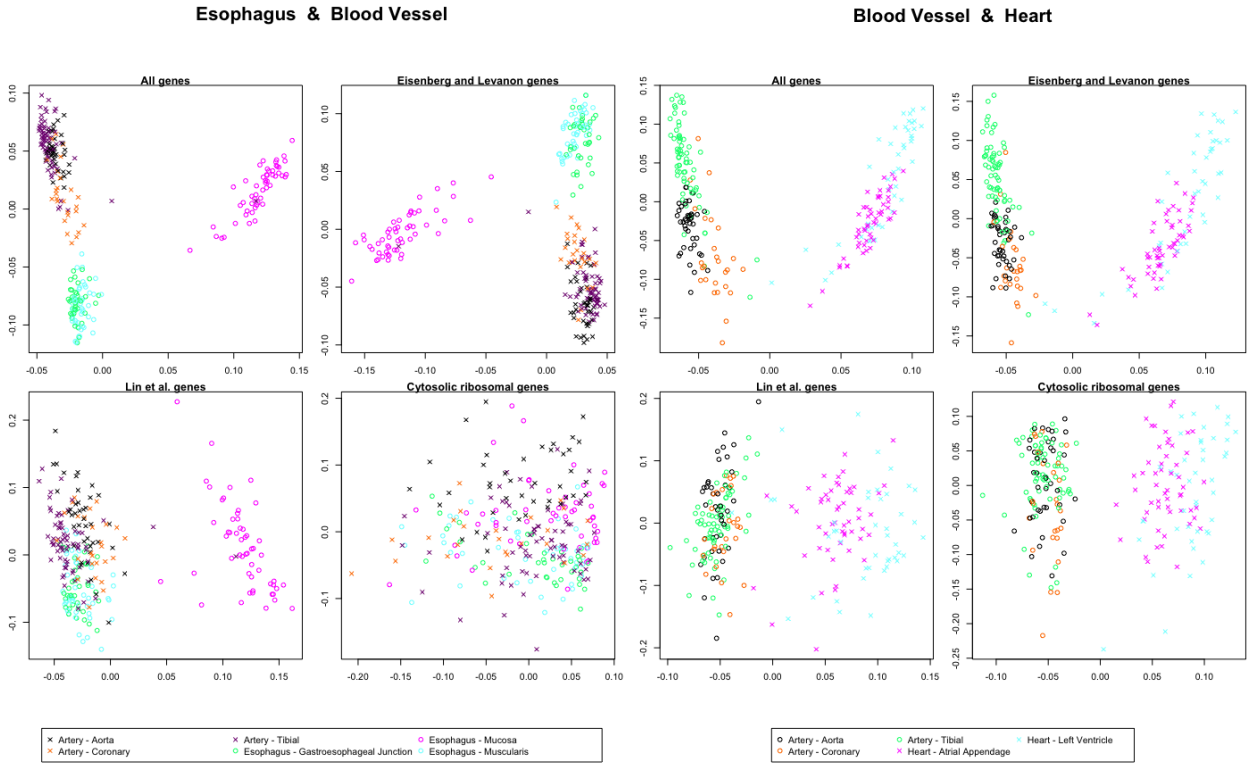

Figure 2.17: Singular Value Decompositions of GTEx data from the esophagus and blood vessel and blood vessel and heart, using different sets of genes as features. For each set of tissues, the first plot on the top left uses all genes as features; the plot on the top right uses genes proposed by Eisenberg and Levanon [2003] as features; the plot on the bottom left uses genes proposed by Lin *et al.* [2019a] as features; and the plot on the bottom right uses cytosolic ribosomal genes as features. The two tissue types are denoted by either the circle or cross symbol, and the tissue subtypes are denoted by the coloring. The esophagus and blood vessel overlap substantially for the cytosolic ribosomal genes, indicating that their expression is reasonably stable across these tissues and subtissues. Overlap is much smaller for the other three sets of genes, indicating less stable expression of these gene sets. The separation for the blood vessel and heart is clear, indicating that the four sets of genes are not stably expressed across the blood vessel and heart; this separation is one of the strongest that we observed. We performed SVDs for random samples of 89 genes from each gene set with similar results, indicating that the size of the gene set is not driving the results.

genes appear more stably expressed both within and across tissue types compared to the highly expressed GTEx genes. Thus, it appears that the high expression of cytosolic ribosomal genes is not driving their apparent stability, further supporting the results from the scRNA-seq data. Figure 2.16 assesses the stability of the cytosolic ribosomal genes, comparing their IQRs to those of the GTEx highly expressed genes.

We also plotted SVDs for each combination of two of the seven tissue types to examine how

stably expressed genes are across those two tissue types. Figure 2.17 shows two sets of two tissue types across which the cytosolic ribosomal genes are especially stable (left panel) and especially not stable (right panel). In general, the cytosolic ribosomal genes exhibit the most stable expression of the gene sets examined between subtissue types and across different tissue types. However, as also noted by Thorrez *et al.* [2008], the cytosolic ribosomal genes are not highly stable across all tissue types.

2.6 Table of Gene Information

On the whole, the cytosolic ribosomal genes appear to be relatively stable. However, they are not uniformly stable across all tissue types. Thus, although the cytosolic ribosomal genes may be a good starting point for a set of reference genes, in most cases a researcher may want to further refine this set. We have therefore generated a database that contains summary information about the genes in our Fluidigm C1 and GTEx datasets. This database contains our structural annotations; means, standard deviations, and correlations from the single cell datasets; and means, standard deviations, and F-statistics for the tissue types from the GTEx data. This information can be used to identify and exclude genes that are not especially stable or reliably detectable in a given tissue type, or to allow a researcher to impose more stringent stability requirements, or to discover other gene sets that may be suitable for a given application. Detailed information about the database can be found in Section 6.6, and the database can be accessed through GitHub (johanngb/sc-stable).

2.7 Discussion and Conclusions

We have found that cytosolic ribosomal genes, as a whole, appear to be relatively stably expressed in proportion to the total RNA content of a cell. Our findings are robust across many single-cell datasets, including those generated with both Fluidigm C1 and 10x platforms. Expression patterns of cytosolic ribosomal genes observed in bulk GTEx experiments further support the conclusion that cytosolic ribosomal genes are relatively stably expressed.

Our focus in this chapter has been limited to identifying sets of stable genes, and we have not yet discussed specifically how such genes might ultimately be used to perform a normalization. A full discussion is beyond the scope of this chapter, but we highlight here some key points. One of the primary challenges is that the cytosolic ribosomal genes, while on the whole relatively stable, are not perfectly so. This can be seen in Figure 2.7, where the histogram for the cytosolic ribosomal genes has a long left tail, suggesting that at least some of these genes are not particularly stable. It can also be seen in Figure 2.16, where we see that a small number of cytosolic ribosomal genes have large IQRs, and in Figure 2.17, where we see evidence of differential expression between

blood vessel and heart. Similar findings have been reported elsewhere in the literature [Islam *et al.*, 2011; Ilicic *et al.*, 2016; Thorrez *et al.*, 2008]. Indeed, our findings are similar to those of Thorrez *et al.* [2008], who reported that ribosomal genes were the most stable set of genes that they had encountered, but still do exhibit tissue-level dependence.

Nonetheless, we believe that even genes that are only approximately stable may still be potentially useful for normalization. For example, GAPDH and Beta-actin are routinely used for normalization in many contexts, despite the fact that these genes may also exhibit some degree of tissue-level dependence [Oyolu *et al.*, 2012; Barber *et al.*, 2005; De Jonge *et al.*, 2007; Islam *et al.*, 2011]. Indeed, we have found the stability of cytosolic ribosomal genes to be comparable to that of GAPDH and Beta-actin (Figure 2.10).

We see two main approaches to ensure a normalization based on cytosolic ribosomal genes is effective, given the imperfect stability of this set of genes. The first is methodological. Recent methods have been developed that exploit a set of reference genes to perform a normalization, but that are also reasonably robust to at least some instability within the reference set [Wang *et al.*, 2017; Gagnon-Bartsch *et al.*, 2013]. That is, these methods may still be effective even when the reference gene set is only approximately stable. In particular, the method of Wang *et al.* [2017] requires that only 50% of the designated reference genes actually be stable.

The second approach is to refine the set of reference genes for any given application, to ensure that the chosen set is as stable as possible. For this, we have created the table of gene information described in Section 2.6. In addition to the structural annotations, the table contains summary statistics from our analyses that allow one to quickly identify genes that are not stable in any of the single cell datasets or in a given tissue type (using GTEx data). For example, using this table, one can easily identify the genes in the long left tail of the histogram of correlations in Figure 2.7, or the genes in the long right tail of Figure 2.16, and remove them from consideration.

We have emphasized in this chapter that the notion of stability at the single cell level is ambiguous. We attempt to clarify specific notions of stability, and in particular consider absolute and proportional stability. We argue that the distinction is important, as the notion of stability implicitly defines the notion of variability (or differential expression, etc.), with implications for not just "data cleaning," but also biological interpretation.

While we were able to identify cytosolic ribosomal genes as proportionally stable, our attempts to identify absolutely stable genes were not successful. We attempted to leverage the (presumed) absolute stability of ERCC spike-ins to identify endogenous absolutely stable genes. However, we concluded this approach is infeasible because of strong technical factors that affect the endogenous genes, but not the spike-ins, rendering correlations between the two effectively meaningless. We proposed a model (Model 2.1) for the expression measurement that incorporates the technical effects that we suspect to help motivate future work. Nonetheless, our hope is that our work has

highlighted both the challenges and the importance of identifying such a set of genes and will encourage future work in that direction.

Chapter 3

Modeling Biases of Reads per UMI in Single-Cell RNA-Sequencing

3.1 Abstract

Single-cell RNA-sequencing (scRNA-seq) experiments measure the RNA transcripts present in a single cell but are subject to many technical effects. Both experimental and analytical approaches have been proposed to adjust for technical effects and reduce biases in scRNA-seq data, including unique molecular identifiers (UMIs) and normalization procedures. We propose a measure of technical effects captured through the experiment, reads per UMI (rUMI). We further propose an estimator of the mean rUMI. We aim to estimate the unobserved proportion of molecules lost due to technical effects, i.e. 0 rUMI, by extrapolation. The estimated proportion of 0 rUMI can then serve as a measure of technical loss in an experiment. We observe the relationship of various transcript characteristics with rUMI based on a scRNA-seq dataset. Further, we generate a model to estimate rUMI based on biological characteristics of a gene and UMI. We discuss how patterns observed in our analysis can be incorporated into downstream normalization procedures, reducing some of the biases introduced from technical artifacts.

3.2 Introduction

Single-cell RNA-sequencing (scRNA-seq) provides the ability to record gene expression at the cellular level. This detailed level of measurement allows for finer understanding of a given tissue or sample. However, measurements are susceptible to a large number of technical effects. Some sources of technical effects, like amplification biases and sequencing errors, have well documented and understood effects [Dabney and Meyer, 2012; Benjamini and Speed, 2012; Aird *et al.*, 2011]; specifically, an imbalance of guanine and cytosine (GC) bases relative to adenine and thymine (AT) bases in a transcript results in inferior amplification. Other sources, like the introduction of

ambient mRNA or cell stress and death, have been proposed as sources of unwanted variation but are less understood [Bacher and Kendziorski, 2016; Stegle *et al.*, 2015].

Experimental and analytical methods designed to distinguish technical effects from biological ones have been proposed. Specifically, Unique Molecular Identifiers (UMIs) provide a barcode to tag a transcript captured from a cell. UMIs mark duplicated reads from the same transcript resulting from amplification, preventing artificial inflation of measured gene expression. In particular, UMIs reduce biases associated with transcript length and composition of GC bases in the transcript [Phipson *et al.*, 2017; Kivioja *et al.*, 2012; Islam *et al.*, 2014]. Analytically, Buettner *et al.* [2015] propose a method to identify genes that distinguish technical from biological variability. Chen and Zhou [2016] adjust for confounding effects, relying on the assumption of a linear model with single cell Partial Least Squares (scPLS). Methods to remove technical effects in other gene expression experiments include those proposed by Irizarry *et al.* [2003] for microarrays, Love *et al.* [2016] for bulk RNA-seq, and Risso *et al.* [2014] also for bulk RNA-seq.

One prominent feature in scRNA-seq data is a high proportion of zero counts, or dropouts. Zero counts occur when no transcripts are aligned to a given gene within a cell, resulting in a zero entry in the gene $\times$ cell digital gene expression matrix. Often, the proportion of zero counts is close to 90% [Bacher and Kendziorski, 2016; Grün *et al.*, 2014; Kim *et al.*, 2015; Dijk *et al.*, 2017]. The zero counts may be biologically accurate or a result of technical effects. Hicks *et al.* [2015] suspect that lowly expressed genes experience a larger proportion of zero counts, rather than zero counts occurring randomly, due to stronger technical variations that affect their measurements. Analytical approaches to account for the high proportion of zero counts include zero-inflated models proposed by Pierson and Yau [2015] and Risso *et al.* [2017].

In addition to the zero counts, experimental results indicate that scRNA-seq does not capture a complete representation of the tissue. Thus, in addition to high zero counts, the available data only provide a general measurement of the gene expression at the cellular level, suggesting additional normalization or analytical steps to refine the data. It is estimated that between $5 - 40\%$ of transcripts from a given cell are captured [Stegle *et al.*, 2015; Kolodziejczyk *et al.*, 2015]. Additionally, $10 - 60\%$ of cells are expected to be captured in droplet-based scRNA-seq protocols [Macosko *et al.*, 2015]. We then expect less than 25% of the transcripts in a sample to be captured in a droplet-based scRNA-seq dataset. For well-based scRNA-seq protocols, the number of cells captured are limited to the number of wells available; well-based protocols often capture fewer cells than droplet-based ones.

To assess batch effects in scRNA-seq, Tung *et al.* [2017] measured the conversion of reads to molecules based on UMIs at the cellular level, finding both biological and technical variation affecting the conversion. Thus, Tung *et al.* [2017] conclude that UMIs do not provide an unbiased estimate of gene expression. Specifically, Tung *et al.* [2017] perform scRNA-seq experiments to

measure both intra- and inter-individual differences by using multiple Fluidigm C1 plates for one individual. Technical effects are observed through different read to molecule conversion ratios across plates from the same individual, while biological effects are additionally present across individuals. We propose exploring a similar measure as Tung *et al.* [2017] at the transcript level, endeavoring to capture technical effects.

Our goals for this Chapter are: (1) to motivate a transcript-level reads per UMI measure as a way to capture technical effects, (2) to propose a robust estimator for a zero-truncated Poisson parameter, and (3) to summarize the relationship between biological characteristics of interest and rUMI, indicating what factors may be related to reliable detection of genes.

3.3 Methods

3.3.1 Reads per Unique Molecular Identifier

We calculate the number of reads per UMI (rUMI) as an approach to help understand some of the technical biases from the experiment. Typically, multiple reads associated with the same UMI are deduplicated before being summarized in the cell $\times$ gene expression matrix, partially removing amplification and sequencing biases. Information contained in duplicated reads from the same UMI may provide additional insights into the technical effects associated with the experiment. We anticipate that rUMI captures some of the technical biases present during both amplification, when multiple transcript copies are generated for a UMI, and sequencing, converting the transcripts into reads outputted by a sequencer. The rUMI measure summarizes technical effects at the transcript level, thereby capturing effects introduced by transcript and UMI characteristics, including the proportion of GC and length of a transcript. Compared to the cell-level conversion ratio of Tung *et al.* [2017], we can observe gene-level and UMI-level biases present using rUMI. Therefore, we measure the rUMI by counting the number of reads obtained after sequencing that are associated with that UMI; in essence, we enumerate the amount of deduplication that occurs as a result of the experimental inclusion of UMIs. Note that 0 rUMIs cannot be recorded.

While UMIs are intended to be unique, they are not perfectly unique. The length of the molecular barcode component of the UMI is often 8 base pairs; see Section 7.1.5 for more details. This provides 65,536 options for the molecular barcode; there is no way to ensure that every possible molecular barcode is used, or that there will be fewer than 65,536 transcripts captured in a given cell. Thus, some UMIs are repeated. We therefore consider the aligned gene as a secondary factor for distinguishing UMIs, with this characteristic helping to capture multiple transcripts being assigned to the same UMI. We discuss in Section 3.4.1 the plausibility of the UMI and gene fully distinguishing transcripts.

Zero counts are distinct from 0 rUMIs, although there is a connection between these situations. Specifically, zero counts can occur through one of three mechanisms. First, a gene may not be expressed in a given cell, resulting in a zero count for that given gene. Second, a gene may be expressed without any of its transcripts binding to a UMI. Thus, the transcripts are present but are never assigned a UMI. Finally, a gene may be expressed with at least one transcript assigned to a UMI. Every transcript assigned to a UMI for that gene must have 0 rUMI to result in a zero count in the gene expression matrix. Thus, a zero count indicates that any transcripts must have 0 rUMI, but 0 rUMI for a transcript does not ensure a zero count. While the rUMI estimator cannot distinguish between zeroes introduced from these three situations, it can estimate how often we expect 0 rUMI to occur.

3.3.2 Experimental Data

We obtained datasets generated and processed by the Hammoud lab at the University of Michigan [Green *et al.*, 2018]. Sertoli cells involved in spermatogenesis were isolated with an in-house drop-seq protocol (droplet-based cell isolation). UMIs were incorporated into the protocol, and the library was sequenced using an Illumina platform. Pre-processing steps were performed according to the pipeline published by the McCarroll Lab [Nemesh, 2015]. The typical pipeline was exited prior to the creation of the digital gene expression matrix and adjusted after alignment to record features about the barcodes, transcripts, and genes.

After filtering according to the standards set by Nemesh [2015], the gene expression for 1,521 cells was recorded for 17,237 genes (11,825,578 UMIs and 91.16% zeros). The 100 cells with the highest number of UMIs contribute 26.02% of the UMIs while comprising 6.57% of the cells in the sample (73.44% zeros across 15,215 genes). We filtered to the top 10 cells with the highest number of UMIs, resulting in 302,054 reads from 244,272 UMIs with an average of 1.24 rUMI (sd=0.54); note again that we can not record 0 rUMI. We consider 20 additional cells in Section 7.2.

3.3.3 Estimation of rUMI

Biologically, a zero-truncated Poisson distribution, that is a Poisson distribution modified so that $\mathbb{P}(X = 0) = 0$, is a reasonable model for the distribution of rUMIs; the probability mass function is shown in Equation 3.1. Other distributional choices are analytically and biologically plausible, but the Poisson model is often proposed by biologists [Robinson and Oshlack, 2010]. We discuss additional distributions and estimators resulting from these distributions in Chapter 4.

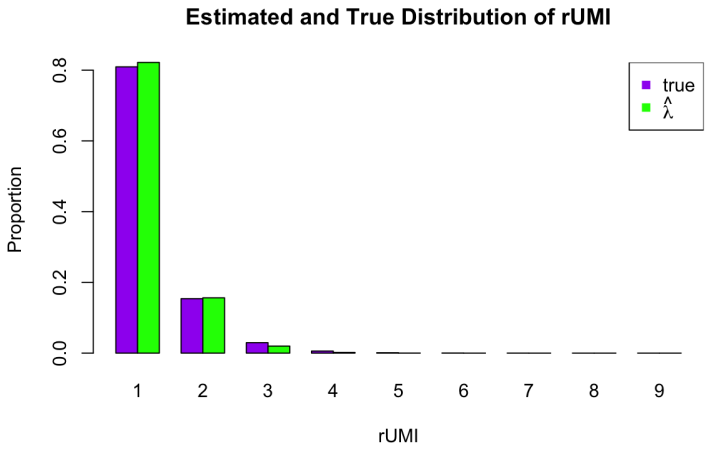

Figure 3.1: A comparison of the estimated distribution using $\hat{\lambda}$ from Equation 3.2 with the distribution of rUMI to be modeled.

$$\mathbb{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!(1 - e^{-\lambda})} \; ; \; x = 1, 2, \ldots, \infty. \tag{3.1}$$

We develop an estimator of the Poisson parameter λ, defined as

$$\hat{\lambda} = 2\frac{\sum \mathbb{I}(\text{rUMI} = 2)}{\sum \mathbb{I}(\text{rUMI} = 1)}. \tag{3.2}$$

When applied to the 244,272 UMIs from the 10 largest cells, $\hat{\lambda} = 0.38$. Figure 3.1 shows the estimated zero-truncated Poisson distributions based on the estimated $\hat{\lambda}$.

The estimator from Equation 3.2 provides that asymptotically in expectation $\hat{\lambda} \approx \lambda$.

$$\begin{aligned}
\mathbb{E}(\hat{\lambda}) &= \mathbb{E}(2\frac{\sum I(\text{rUMI}_s = 2)}{\sum I(\text{rUMI}_s = 1)}) \\
&\approx 2\frac{\mathbb{E}(\sum I(\text{rUMI} = 2))}{\mathbb{E}(\sum I(\text{rUMI} = 1))} \\
&= 2\frac{\mathbb{P}(\text{rUMI} = 2)}{\mathbb{P}(\text{rUMI} = 1)} \\
&= 2\frac{e^{-\lambda}\lambda^2/2!}{e^{-\lambda}\lambda/1!} \\
&= \lambda.
\end{aligned} \tag{3.3}$$

Further details can be found in Section 4.4.2.2.

We can further apply our estimator to subsets of the data. For example, we can estimate $\hat{\lambda}$ for

some group of the data i as

$$\hat{\lambda}_{,i} = 2\frac{\sum_{s\in S_i} I(\text{rUMI}_s = 2)}{\sum_{s\in S_i} I(\text{rUMI}_s = 1)}, \tag{3.4}$$

where S_i represents the members of group i. We create categorical groupings of our variable of interest to form S_i used in estimating $\hat{\lambda}_{,i}$. Redefining our variable in terms of groups provides flexibility for the form of the relationship between the variable of interest and rUMI without imposing any restrictions on the relationship. For quantitative variables, we use intervals that form (roughly) equally sized groups.

From the estimated $\hat{\lambda}$ values, we can also estimate the probability of the technical loss of a transcript based on the Poisson distribution, i.e. $\mathbb{P}(\text{rUMI} = 0) = e^{-\hat{\lambda}}$. We are particularly interested in observing how $\hat{\lambda}$ and its corresponding probabilities vary in relation to transcript characteristics.

3.4 Results

3.4.1 Data Characteristics

We consider the total number and standard deviation of UMIs recorded for each gene and for each cell in the 1,521 cells and 17,237 genes in Figure 3.2. Low measured expression is apparent for many of the cells and genes from Figure 3.2, with some genes and some cells accounting for a large proportion of the measured UMIs.

After filtering to the top 100 cells, we see that the relationship between the number of reads and number of UMIs appear fairly linear in Figure 3.3. There are some cells that experience higher overall rUMI than the general pattern. There is some variation of rUMI for genes but no obvious genes appear to be outliers from the right panel of Figure 3.3.

We perform our analyses for the top 10 cells based on total number of UMIs. Section 7.2 provides a similar analysis for 20 additional cells. Figure 3.4 displays the distribution of rUMI for the 244,272 UMIs. More than 80% of the UMIs are captured from a single read. Further evidence of shallow sequencing comes from $\hat{\lambda}$ evaluated on the full sample, which indicates that the probability of having 0 rUMI is 68.36%, or that more than 2 of every 3 transcripts that are assigned a UMI, i.e. bind to a bead, will not be included in the final set of reads.

As mentioned in Section 3.3.1, we applied the gene in addition to the UMI in identifying distinct transcripts. We suspect, however, that the long right tail in Figure 3.4 may be an artifact of UMIs not being perfectly unique. Specifically, it is possible that the 9 rUMI value in fact represents two (or more) distinct transcripts from the same gene that were assigned to identical UMIs. Thus,

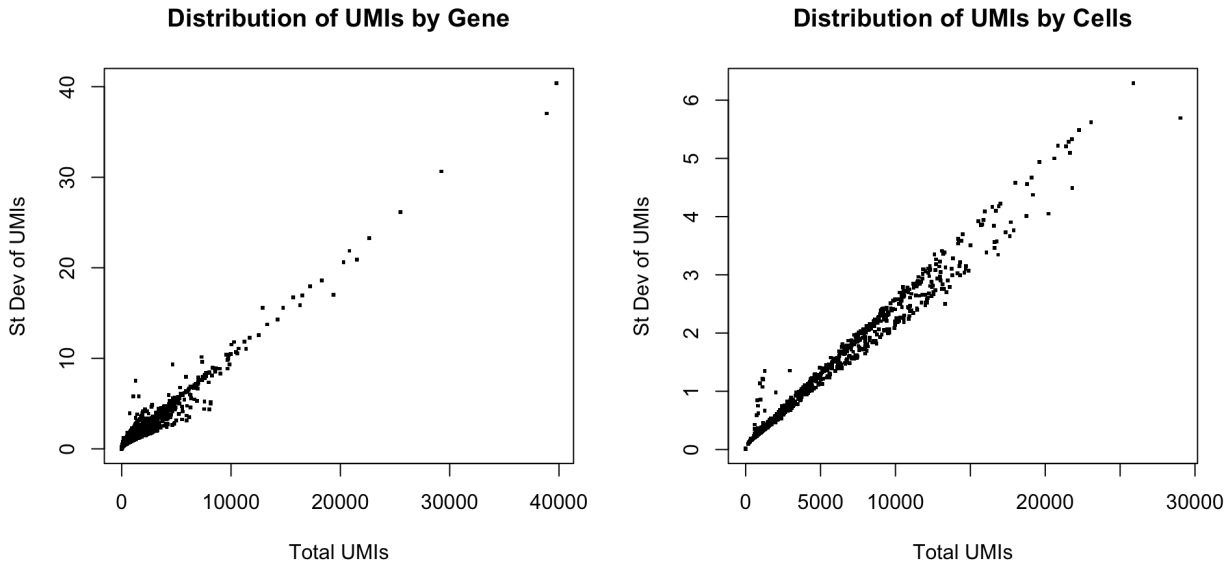

Figure 3.2: Scatterplots of the total number and standard deviation of UMIs from the full dataset of 17,237 genes (left panel) and 1,521 cells (right panel). A large proportion of genes and cells have small overall expression levels. There is a clear dependence between the total number of UMIs recorded and standard deviation.

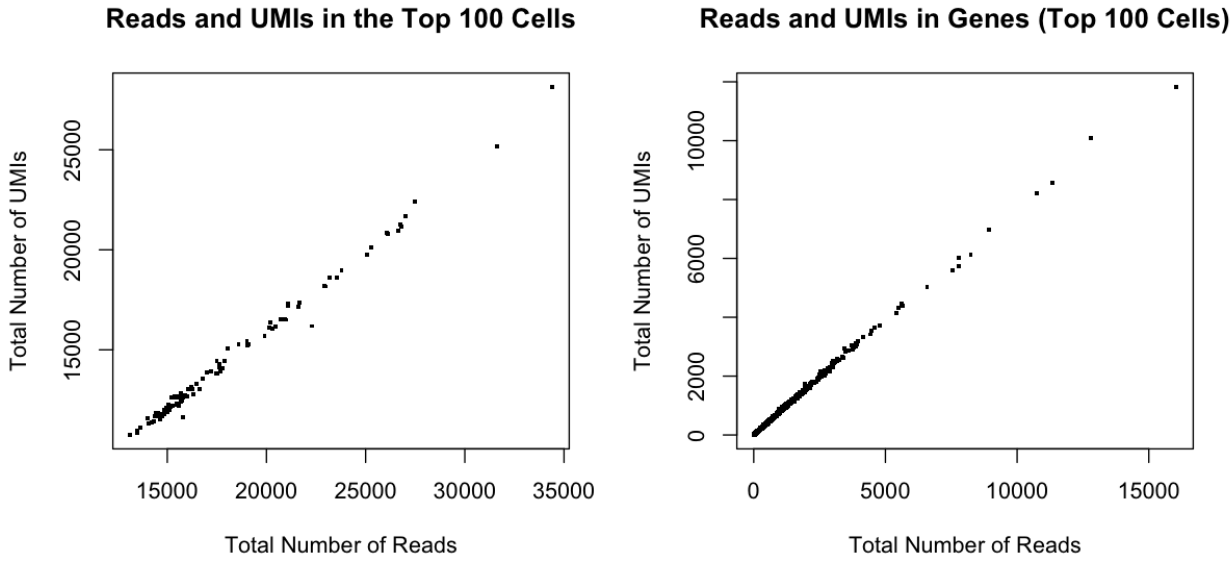

Figure 3.3: The relationship between the total number of reads and the total number of UMIs for the 100 cells with the highest number of UMIs. The left panel shows the relationship for the 100 cells, and the right panel shows the relationship for the genes.

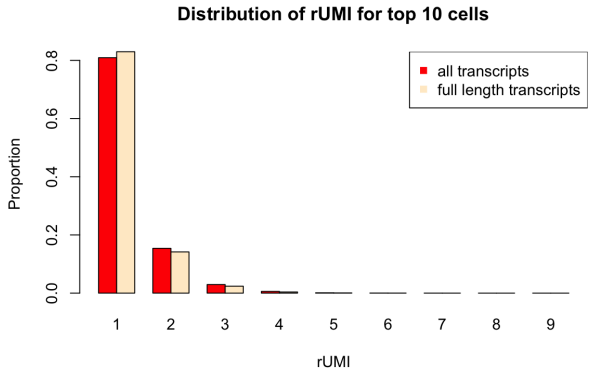

Figure 3.4: The distribution of rUMI for all transcripts and for full length transcripts from the top 10 cells. More than 80% of the UMIs were found in only 1 read, indicating shallow sequencing.

we consider the distribution of rUMIs without incorporating the gene as a distinguishing factor in Figure 3.5. We see the rUMI values now extend to 19, and that the proportion of 1 rUMIs is reduced to 62.6%. From this distribution, our overall estimates of $\hat{\lambda}$ changes to 0.74 from 0.38.

Additionally, we explore characteristics of the 20 largest rUMIs (7-9 rUMI). First, we examine the measured expression level of the genes associated with these 20 reads. 14 of the genes from the top 20 rUMIs are in the top 10% of measured gene expression. Genes with large measured expressions indicate that the transcripts are more abundant within a cell, and thus are more likely to have multiple transcripts assigned to identical UMIs. Additionally, we examine if molecular barcode characteristics might contribute to repeated UMIs, concluding that this is unlikely. Adding the gene as a separating characteristic resulted in a 30% increase in the number of molecules counted. Within the top 20 rUMIs, including the gene as a separating factor increased the number of molecules by 40%, a modest increase. While UMI characteristics could contribute to the large number of rUMIs measured, one would expect most of the top molecular barcodes to have multiple additional reads originating from each of the 10 cells. However, half of these molecular barcodes had 5 or fewer additional copies, most of which originated from different cells. Neither the high expression of a gene nor UMI characteristics seems to fully explain the large rUMI values observed.

3.4.2 Univariate Associations

3.4.2.1 Read Length

One of the first variables that we examine, transcript length, contains a high proportion of UMIs at the maximum length of 151 base pairs (bp), as seen in Figure 3.6. The maximum length of 151

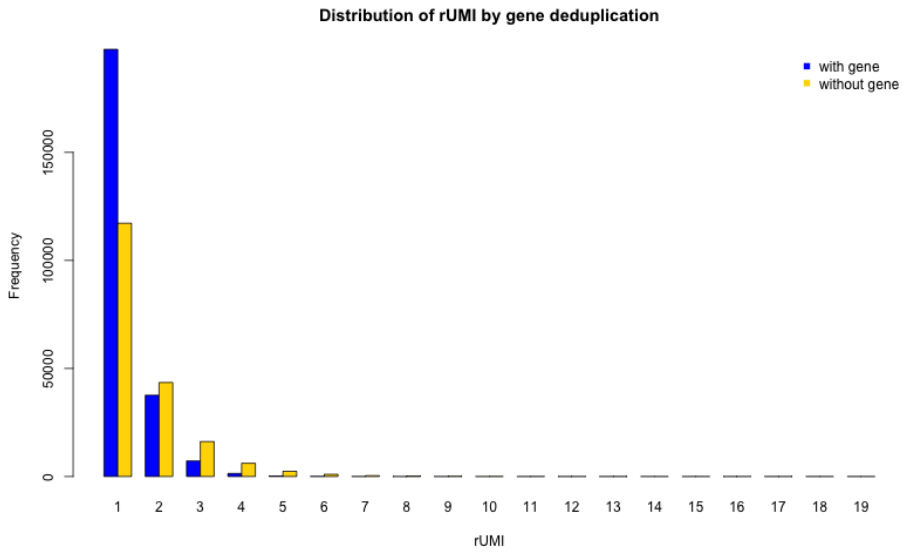

Figure 3.5: The distribution of rUMI with and without using gene to distinguish between molecules. UMIs with identical UMIs but distinct genes likely occur from different original molecules, even though the UMIs are identical.

bp was selected prior to sequencing based on experimental protocol. Thus, we perform two sets of analyses, one including all transcripts and another for full length transcripts, for the remaining variables. First, we examine the distribution of rUMI for the full length reads in the bottom panel of Figure 3.4, finding that an even larger proportion of UMIs are captured with only one rUMI than from all reads.

We examine how length is related to rUMI in the right panel of Figure 3.6. We see the largest estimate for mean rUMI occurs for the length interval that is not quite full length. We suspect that UMIs with multiple reads may consist of one (or more) full length reads and one (or more) reads that are close to full length. The average read length for the UMI would approach but not reach the full 151 bp, increasing the estimate for $\hat{\lambda}_{,i}$ for the UMIs with the second-longest average length.

3.4.2.2 Transcript Proportion of Guanine and Cytosine

We calculate the proportion of guanine and cytosine (GC content) for the full length of the transcript, for the first half of the transcript, and for the second half of the transcript. For UMIs with multiple reads, we summarize these measures with the mean of the proportion GC content for each read.

Figure 3.7 displays the estimates from Equation 3.2 for the average GC content of the full transcript. An AT/GC imbalance results in smaller $\hat{\lambda}_{,i}$ estimates, indicating that genes with an imbalance of bases may result in lower rUMI. The full length reads, however, exhibit a less clear

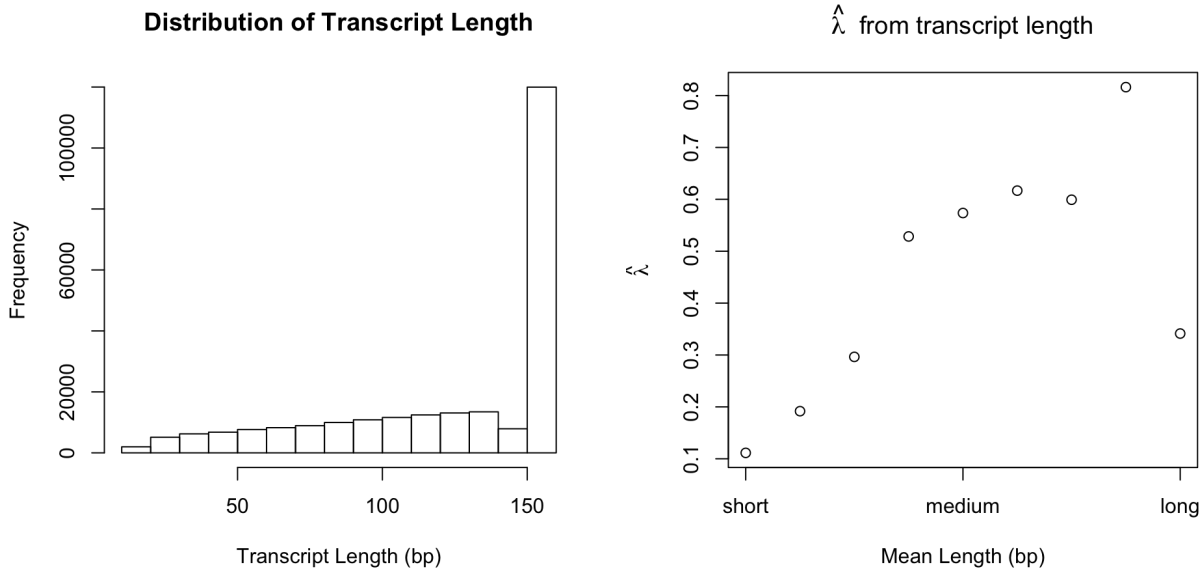

Figure 3.6: The average transcript length ranges from 13 to 151 bp, with almost half of the values consisting of full length reads (left panel). The scatterplot on the right displays how the estimated rUMI parameter is related to read length, with the largest estimate for reads that are close to but quite full length. The last category, long, represents all of the UMIs that are comprised of full length transcripts (roughly 50% of the UMIs). The estimates from the remaining categories were made with Equation 3.2 using roughly 6% of the UMIs each.

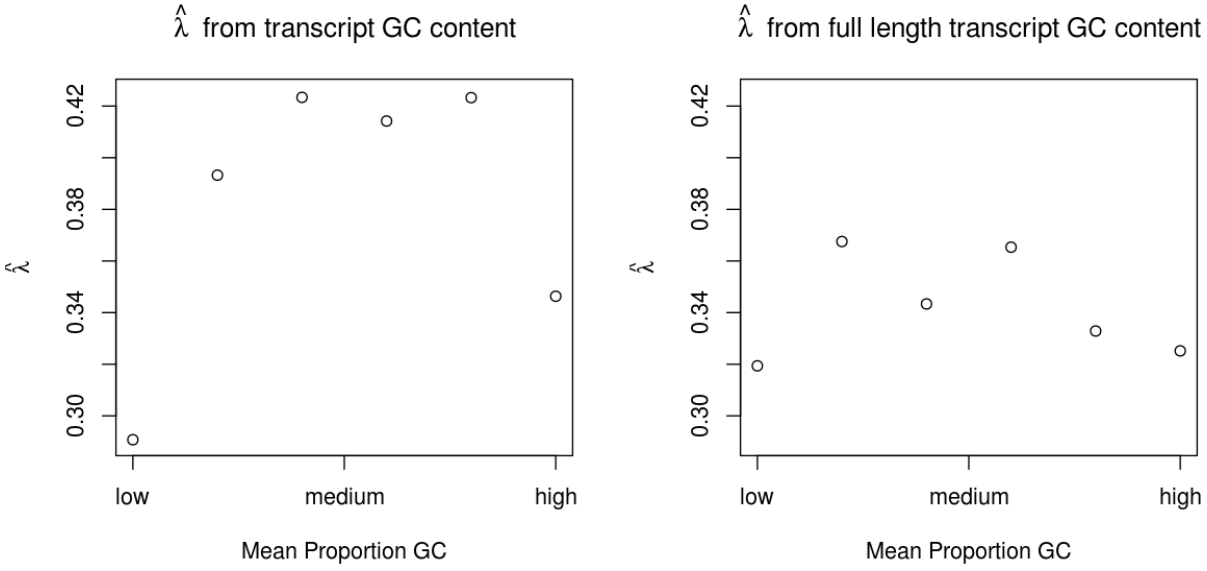

Figure 3.7: The estimated mean rUMI based on the GC content of the transcript is shown for all transcripts (left panel) and for full length transcripts (right panel) in the top 10 cells. rUMI appears roughly quadratically related to GC content for all transcripts, although this relationship does not hold for full length transcripts. AT/GC imbalances appear to result in smaller estimates of mean rUMIs. Full length transcripts tend to have smaller AT/GC imbalances and a slightly larger proportion of GC (Figure 3.8).

pattern. Note that the $\hat{\lambda}$ values estimated for the full length reads are smaller than those from all UMIs; this follows from the higher proportion of one rUMIs found in Figure 3.4 for the full length reads. In addition, the full length reads tend to have smaller AT/GC imbalances and higher proportions of GC (Figure 3.8). We suspect that these characteristics may explain the differences between the two panels of Figure 3.7; the larger number of bases with which the full length transcripts are normalized may also contribute to differences in both the GC characteristics and the relationship between GC and rUMI.

Second, we examine the relationship between rUMI and the GC content in the first half of a given transcript in Figure 3.9. Note that the GC content of the first half of a transcript contributes to the GC content of the entire transcript in Figure 3.7. Here, however, we observe a stronger quadratic relationship, with lower estimated rUMI occurring for an imbalance in AT/GC both for all transcripts and for full length transcripts, suggesting that the composition of the first bases of a transcript influences the technical biases that a UMI captures.

The relationship between rUMI and the proportion GC content in the second half of the transcript exhibited less clear patterns (Figure 7.1). Thus, it appears that the composition of the bases in the first half of the transcript is driving the relationship between the GC content of the full

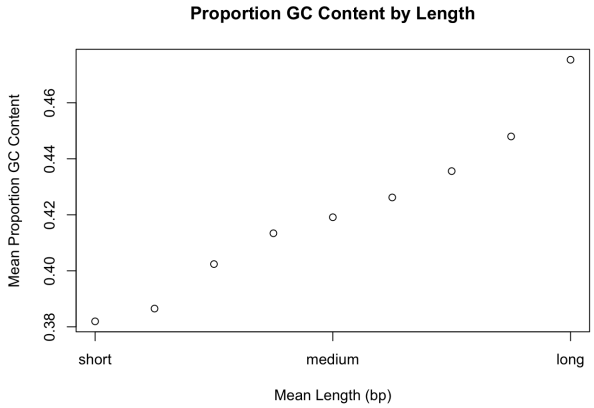

Figure 3.8: The relationship between average length of a UMI and the average proportion GC. The length categories are repeated from Figure 3.6. The full length UMIs have the highest proportion GC content and the smallest AT/GC imbalance.

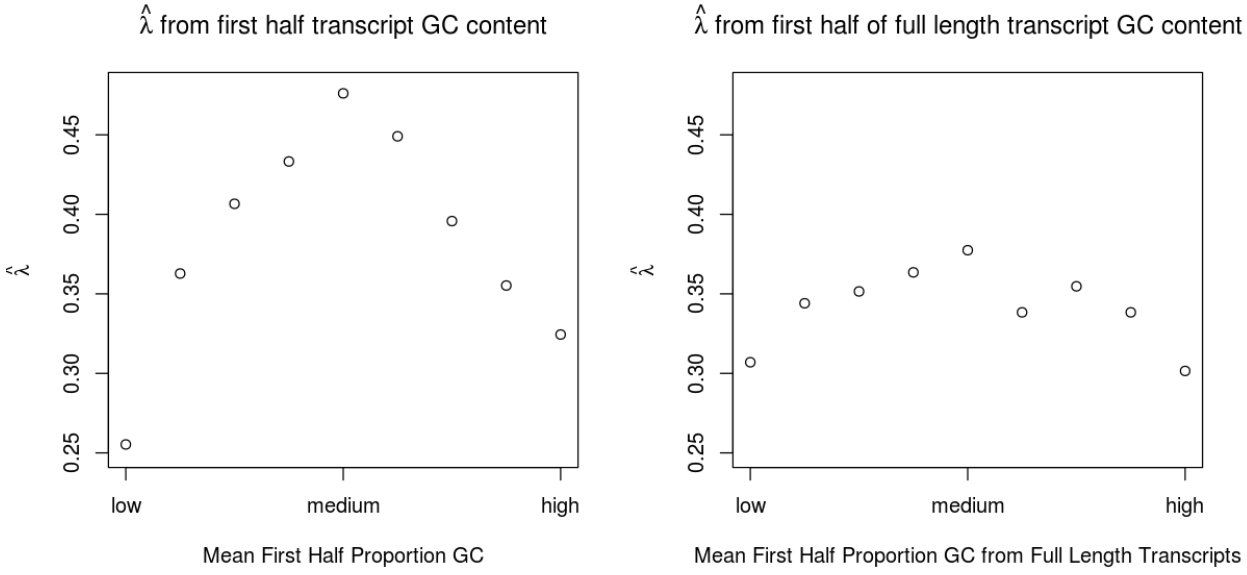

Figure 3.9: The estimated mean rUMI based on the GC content in the first half of the transcript is shown for all transcripts (left panel) and for full length transcripts (right panel) in the top 10 cells. A quadratic pattern is observed in both panels, with an AT/GC imbalance corresponding to a smaller average rUMI. Note that GC content in the first half of the transcript contributes to 50% of the GC content measured in Figure 3.7.

transcript and rUMI.

3.4.2.3 Two-somes

We further examine how the base pair composition of the transcript contributes to rUMI with measures of the specific sequence of the transcript. We consider a sliding window of length two for each transcript, recording the set of two base pairs in that window and defining this as a two-some. Note that with four bases, there are sixteen options for two-somes. We then calculate the proportion of a given two-some for the transcript. For UMIs with multiple reads, we summarize the proportion of a two-some with the mean proportion for each of the reads.

Figure 3.10 displays plots for the two-somes that exhibit the strongest relationships with rUMI; see Section 7.1.2 for additional figures and discussion. The AT, CT, TA, and TT two-somes show the strongest relationships with rUMI for all transcripts, while the AA and GC two-somes show the strongest relationships with rUMI for the full length transcripts. The lack of overlap between two-somes with strong patterns between all transcripts and full length transcripts does provide some hesitation as to the applicability of these results to all situations.

3.4.2.4 Other Variables

We consider the relationship between rUMI and other variables, with no especially interesting patterns observed, in Section 7.1. Other variables explored include the proportion of guanine and cytosine in the cellular barcode, the first base of the cellular barcode, the last base of the cellular barcode, the proportion of guanine and cytosine in the molecular barcode, the first base of the molecular barcode, and the last base of the molecular barcode. We selected these particular variables due to their known relationships with amplification biases, in the case of GC content, and their location before the transcript. We had suspected that features located at the beginning of a read may contribute to technical effects associated with amplification or sequencing biases. However, we observed no clear relationships between rUMI and these particular variables.

Additionally, we fit a generalized linear model to predict rUMIs based on a combination of variables considered here. See 7.3 for a discussion of this model and its findings.

3.5 Discussion and Conclusions

In this chapter, we have motivated the use of rUMI as a measure of technical effects in scRNA-seq data. Specifically, we describe some of the technical effects that rUMI could reasonably capture. We introduce a zero-truncated Poisson distribution to model the rUMI distribution. We propose an estimator for the λ parameter from a zero-truncated Poisson distribution. We continue developing

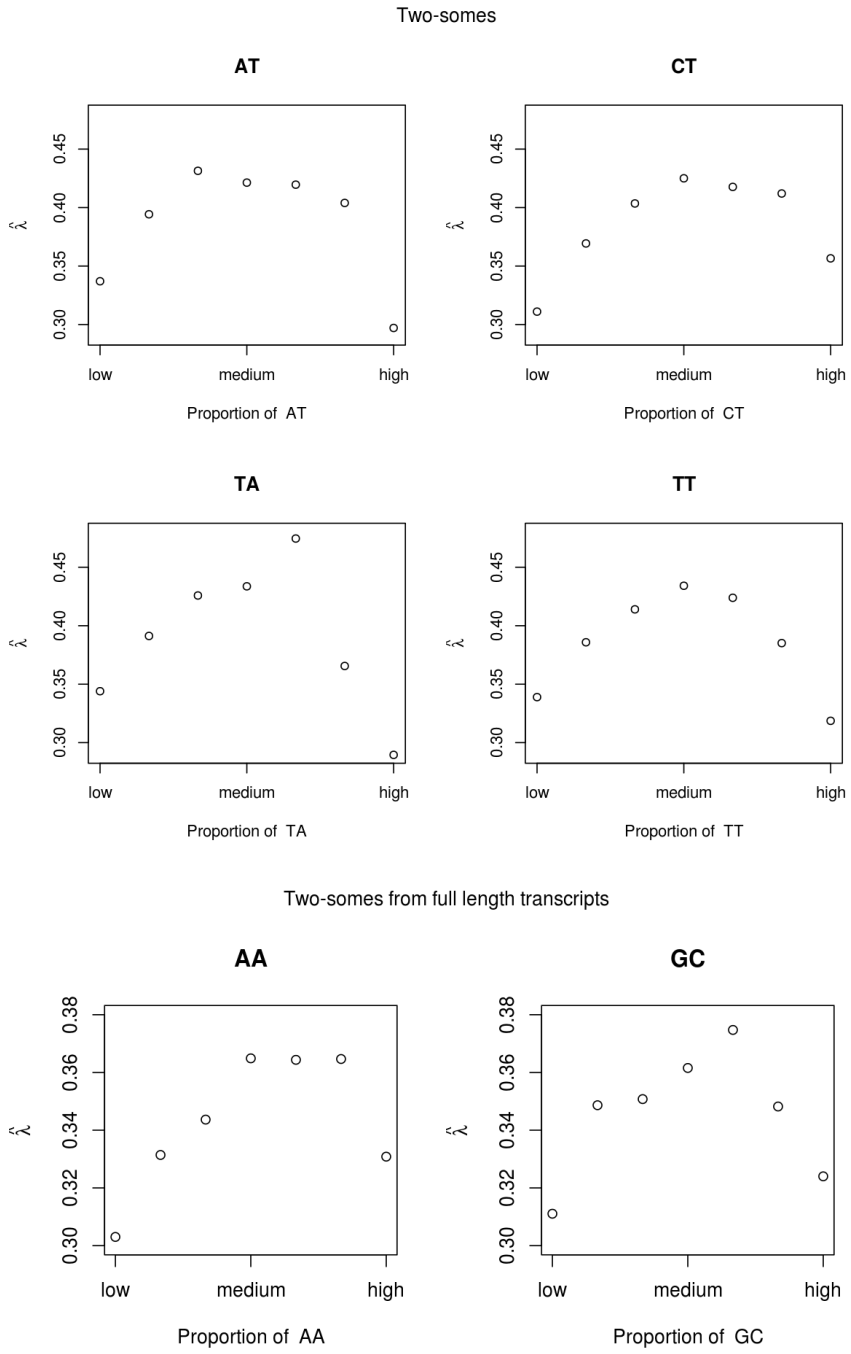

Figure 3.10: The estimated mean rUMI is shown based on the average proportion of the specified two-somes in transcripts. The two-somes in the top two rows (AT, CT, TA, and TT) are calculated from all transcripts in the top 10 cells, while the two-somes in the bottom row (AA and GC) are calculated from the full length transcripts in the top 10 cells. Average levels of these two-somes have higher average rUMI, while enrichment or depletion of these two-somes results in lower average rUMI.

this estimators and discuss additional estimators in Chapter 4. We have explored how biological gene and barcode characteristics are related to rUMI, suggesting that the GC content of a transcript is especially informative in estimating rUMI.

The rUMI measure summarizes the amount of deduplication that has occurred for a given transcript. Additionally, rUMI captures technical biases associated with specific genes and transcripts. Specifically, since rUMI measures the number of copies of a single transcript, it also can also suggest characteristics stemming from the gene or the UMI that indicate a transcript is more likely to be read. Thus, we can identify technical biases that increase the likelihood of a given transcript being read, and, conversely, infer characteristics that reduce the likelihood of a transcript being read.

Further work is needed to fully address how we anticipate our findings could be incorporated into later downstream analyses. We envision that these results could provide meaningful insights in multiple ways.

The estimated $\hat{\lambda}$ for rUMI indicates how reliably detectable a given gene is, with higher estimated rUMI measures indicating that the UMI is more likely to be captured in scRNA-seq data. Genes with specific characteristics, like a proposed range of proportion GC in the first 75 bp identified from Figure 3.9, could be identified as genes that might be especially reliably detectable. We may need to adapt gene-level characteristics from the transcript-level ones used in our current approach by removing introns or assuming that the first 75 bp of a gene are preserved in order to identify these genes.

The methods that we have applied to this scRNA-seq data could be applied to other scRNA-seq data with UMIs. We anticipate that our results may be generalizable, although some characteristics may capture technical effects differently based on the experiment. Estimating $\hat{\lambda}$ for rUMI and modeling rUMI based on the experimental results to be analyzed could further inform how reliably detectable a given gene is within an experiment.

A list of reliably detectable genes or a model to identify reliably detectable genes could be applied to later analyses and normalizations in a variety of ways. First, we could evaluate zeros occurring in scRNA-seq data based on how reliably detectable the gene is. We may be able to infer whether zeros from a given gene are systematic, semi-systematic, or stochastic based on features of the gene. Additionally, reliably detectable genes may be selected as genes that experience fewer technical biases and can serve as features for clustering or cell type identification. Finally, we proposed in Chapter 2 that the cytosolic ribosomal genes appear to be proportionally stable. We may be able to use the characteristics of reliably detectable genes to further refine the list of proportionally stably expressed genes, as an additional desirable characteristic of stably expressed genes is reliable detection.

Conversely to the methods described above, we could apply these results to identify genes that

are particularly unreliably detected. We may choose to remove unreliably detected genes from later analyses or adjust their measurements in some way. By recognizing that the gene expression is unreliable, we may be able to remove some of the technical effects from the data. Through this extrapolation of unreliable detection, we may also be able to identify zeros in scRNA-seq data that we believe are biologically unexpressed or that we suspect are zero due to some technical effects. This information could be incorporated into later analyses. Specifically, data imputation and normalization methods have been proposed both for the zero counts of scRNA-seq data [Huang *et al.*, 2017] and for all genes using pools of cells [Dijk *et al.*, 2017; Lun *et al.*, 2016].

Before applying our reliably detectable characteristics to many datasets, we recommend assessing the generalizability of our results in additional datasets. We have applied the analysis discussed here to 20 additional cells from the same experiment, observing similar patterns regardless of cell size; see Section 7.2. We could apply these same methods to cells or reads from any scRNA-seq experiments with UMIs, providing robust support for the generalizability of the results presented here. One can imagine tissue, platform, or experimental protocol differences that change the exact technical effects captured by the rUMI measure. If the same general results are found in multiple datasets, then one set of genes may be proposed as reliably detectable. If different results occur, then it may be advisable to repeat the analysis to identify characteristics specific to the given experiment.

Different experimental conditions could occur for a number of reasons. The experiment itself may apply different protocols to different tissue types, depending on the experimental purpose. Often, one of the experimental conditions the scientist has to balance is the cost and the amount of new information. Sequencing costs are often one of the most expensive, and thus, scientists try to reduce cost and redundant information in the form of duplicated reads. It is often desirable to have an rUMI with many low values, indicating that little redundant information was purchased. Statistically, however, the low rUMI values indicate low sampling of the transcripts, and thus less certainty about conclusions. Experimenters strive to have low experimental waste (small values for rUMIs) while obtaining the maximum amount of information.

One possible byproduct of low experimental waste is the high proportion of zero counts in scRNA-seq data. Zero counts indicate that no copies of a gene were measured in a given cell but were measured in a different cell. These zero counts are not necessarily problematic and may arise from different causes, including the gene not being expressed in the cell, the transcripts not being assigned to a UMI, or the transcripts assigned to a UMI not being read, i.e. 0 rUMIs. The first scenario is not problematic but is biologically accurate. The second scenario may arise due to technical biases but information to evaluate these biases is not available currently. The third scenario, however, can be addressed with the rUMI measure described here. Specifically, genes that have characteristics of lower rUMIs, like an imbalance of AT and GC bases, may have less

reliable detection and thus less trustworthy zero counts. These factors may be uniformly applied to all transcripts from a given gene, or differentially applied based on other factors. More work needs to be undertaken to evaluate how helpful the rUMI measure could be clarify the zero counts, as we suspect that a large portion of zero counts result from the first or second scenarios.

Our hope is that our measure of reliable detection using rUMIs in conjunction with the proportional stability of the cytosolic ribosomal genes will result in a procedure that can be used to normalize scRNA-seq data in a more systematic and intentional manner.

Chapter 4

Zero-truncated Distribution Models of Reads per UMI

4.1 Abstract

Single-cell RNA-sequencing (scRNA-seq) experiments measure the RNA transcripts present in a single cell but are subject to many technical effects. We proposed reads per unique molecular identifier (rUMI) as a way to measure technical effects in a scRNA-seq experiment in Chapter 3. One prominent feature of the rUMI statistic is that it cannot take zero values. In this chapter, we extend the statistical motivation and methods for rUMI estimation and explore characteristics of plausible zero-truncated distributions that the rUMIs could follow. Specifically, we examine three different special case models for the rUMIs. We define five estimators: the maximum likelihood estimator for the three data generating models (zero-truncated Binomial, zero-truncated Poisson, and zero-truncated Negative Binomial), a moment-based estimator from a zero-truncated Poisson distribution, and a robust estimator from a zero-truncated Poisson distribution. We then examine various features of these estimators, including their performance under different data generating schemes. Finally, we apply these estimators to data obtained from a scRNA-seq experiment. We specifically examine the proportion GC in the transcript as described in Chapter 3 and the expression level of the gene as discussed by Hicks *et al.* [2018] in relation to the estimated mean rUMI for each of the five estimators.

4.2 Introduction

Single-cell RNA-sequencing (scRNA-seq) measures gene expression at the cellular level. Various factors have been proposed as affecting the measurements, including both technical and biological features [Hicks *et al.*, 2018]. Understanding the technical effects of a scRNA-seq experiment could improve data normalization, informing the analytical processes applied to data.

We previously introduced a measure of some of the technical effects associated with scRNA-seq experiments, taking advantage of information recorded from an experimental innovation, Unique

Molecular Identifiers (UMIs). UMIs are barcodes that tag transcripts, marking duplicate reads resulting from amplification. UMIs provide a basis for deduplication of reads and reduce biases in the measurements, including those resulting from transcript length and proportion of GC bases [Phipson *et al.*, 2017; Kivioja *et al.*, 2012; Islam *et al.*, 2014]. We propose that additional information can be gleaned from the deduplication step and suggested calculating the reads per UMI (rUMI) to measure technical effects in Chapter 3. We suspect that rUMI may summarize some of the technical effects due to amplification and sequencing. Indeed, Tung *et al.* [2017] explore the conversion of reads to molecules at the cellular level to assess sources of variation in scRNA-seq experiments, finding evidence of both biological and technical sources of variation. In contrast, we explore the reads per UMI at the molecular level.

With rUMIs measured from a given experiment, we can estimate features of rUMIs, including the proportion of zero rUMIs, based on various distributional assumptions. The proportion of zero rUMIs specifically estimates the proportion of the population of molecules that attach to beads that are captured in the experimental sample. Stegle *et al.* [2015] and Kolodziejczyk *et al.* [2015] estimate that the proportion of molecules captured in the experiment could be $60 - 95\%$. In fact, experimenters are pleased with a distribution of rUMIs that have many low values, as sequencing is expensive and duplicate reads provide redundant experimental information. Statistically, however, that same high proportion of zero rUMIs indicates less confidence in conclusions, as they are drawn from a smaller portion of the population. Biases in detection of genes can also affect the conclusions drawn from scRNA-seq experiments. Particularly, Hicks *et al.* [2018] suggest that genes that are lowly expressed are more likely to result in zero rUMI measures than genes that are highly expressed within a cell, i.e. that the expression level of a gene affects the probability that a molecule of that gene is captured in the reads.

The distribution of observed rUMIs is composed of integer values of one or more, with a maximum possible value determined by amplification. Because we cannot observe 0 rUMIs, we examine zero-truncated (ZT) distributions, or distributions that do not allow values of zero. Typical observed rUMI distributions have a large proportion of values of one and two, with a long right tail of higher values. Based on these characteristics, ZTBinomial, ZTPoisson, and ZTNegative Binomial distributions are reasonable to propose. The Binomial distribution is appropriate based on the underlying scientific processes of amplification and sequencing. The Poisson distribution is often suggested as an approximation to the Binomial distribution, while the Negative Binomial distribution relaxes the assumption from the Poisson distribution that the mean and variance are equivalent.

In this chapter, we further explore the rUMI statistic used to measure technical effects proposed in Chapter 3 and possible underlying ZT distributions. In particular, our goals are to: (1) develop estimators for the unobserved proportion of zero rUMIs in a given sample, (2) evaluate

our estimators for their theoretical properties, and (3) apply our estimators to simulated and real data.

We particularly aim to predict the proportion of zero rUMIs from characteristics of the molecule. With this information in hand, we are better able to estimate the sampled proportion of transcripts and to inform normalizations applied to scRNA-seq data. We suspect that transcript characteristics may result in different proportions of 0 rUMIs. Thus, our estimators should be capable of being applied to the entire sample of rUMIs or to subsets of the sample based on specific molecular characteristics. This chapter focuses on the development and properties of the estimators.

4.3 Statistical Framework

4.3.1 Notation

Let M denote the number of molecules, i.e. transcripts, within a sample. Let N_i for $i \in 1, \ldots, M$ denote the number of copies of the transcript following amplification within that same sample. Further, let p_i denote the probability for a given copy of molecule i to be captured in the sequencing reads.[1] Then, we can consider X_{ij} the indicator that the j^{th} copy of molecule i is captured, with $\mathbb{P}(X_{ij} = 1) = p_i$. We further denote the number of reads for molecule i as $R_i = \sum_{j=1}^{N_i} X_{ij}$.

The R_i for $i \in 1, \ldots, M$ is then the true distribution of rUMI for this given sample. However, we only observe R_i when it takes a non-zero value. We denote Z_i as an indicator of molecule i being captured in the reads, and C_i as the observed R_i value when $Z_i = 1$. When $Z_i = 0$, C_i is undefined. Further, let $n = \sum_{i=1}^{M} Z_i$, or the sample size of the observed C_i.

We are interested in estimating the probability that a molecule is not captured in the reads, $\pi_i = \mathbb{P}(R_i = 0)$, which is equivalent to $1 - \mathbb{E}(Z_i)$. We will denote the true probability $\mathbb{P}(R_k = 0 \mid k \in K)$ as $\pi_{,K}$ and any estimators as $\hat{\pi}_{,K}$ where K is an index for some group of molecules, providing the flexibility for estimators to be evaluated for different groups of molecules. The estimators that we develop can be applied to all of the observed reads in a scRNA-seq experiment or to subsets of the reads, capturing the relationship between attributes of molecules and the probability that a molecule will be included in the final reads, i.e. $\mathbb{P}(Z_i = 1)$. In other words, we use features of our observed rUMI distribution (or a subset of it) to estimate the proportion of molecules that are unobserved in the reads.

[1] Since all N_i copies of molecule i are identical, barring any amplification errors, it is assumed that p_i is equivalent for all N_i copies.

4.3.2 Special Cases

Special Case 1: Binomial Distribution Suppose that $p_i = p_0 \; \forall \; i$, i.e. that the probability of capturing any molecule is the same for all molecules. In addition, suppose that $N_i = N_0 \; \forall \; i$, i.e. that amplification is equally efficient for all molecules. Additionally, we assume that N_0 is finite. If we let $\lambda_0 = N_0 p_0$, then $R_i \sim$ Binomial $\left(N_0, p_0 = \frac{\lambda_0}{N_0}\right)$ and C_i a ZTBinomial one. Thus, for $c \in 1, \ldots, N_0$,

$$\mathbb{P}(C_i = c) = \binom{N_0}{c} \frac{\lambda_0^c (N_0 - \lambda_0)^{N_0 - c}}{N_0^{N_0} - (N_0 - \lambda_0)^{N_0}}. \tag{4.1}$$

Our parameter of interest for the binomial distribution is $\pi = \left(1 - \frac{\lambda_0}{N_0}\right)^{N_0}$.

Special Case 2: Poisson Distribution Again, suppose that the probability that any given molecule is captured is equivalent for all molecules, i.e. $p_i = p_0 \; \forall \; i$, and that amplification is equally efficient for all molecules, i.e. $N_i = N_0 \; \forall \; i$. As $N_0 \to \infty$ and $p_0 \to 0$, letting $N_0 p_0 \to \lambda$, then $R_i \sim$ Poisson(λ), and $C_i \sim$ ZTPoisson(λ). Thus, for $c \in 1, 2, 3, \ldots$,

$$\mathbb{P}(C_i = c) = \frac{e^{-\lambda} \lambda^c}{c!(1 - e^{-\lambda})}. \tag{4.2}$$

In this scenario, $\pi = e^{-\lambda}$.

Special Case 3: Negative Binomial Distribution Suppose that $N_i = N_0 \; \forall \; i$ again, but that p_i is no longer identical but drawn from a Gamma$\left(\alpha, \beta = \frac{\alpha N_0}{\lambda_0}\right)$ distribution. The p_i distribution is equivalent to $\lambda_i \sim$ Gamma $\left(\alpha, \frac{\alpha}{\lambda_0}\right)$ when $\lambda_i = N_0 p_i$. For a given λ_i, suppose $N_0 \to \infty$, $p_i \to 0$, and $N_0 p_i \to \lambda_i$. Then, $R_i \sim$ Negative Binomial $\left(\alpha, \frac{\lambda_0}{\alpha + \lambda_0}\right)$ and $C_i \sim$ ZTNB with the same parameters. Thus, for $c \in 1, 2, 3, \ldots$,

$$\mathbb{P}(C_i = c) = \binom{c + \alpha - 1}{c} \frac{(\alpha)^\alpha \lambda_0^c}{(\alpha + \lambda_0)^c [(\alpha + \lambda_0)^\alpha - (\alpha)^\alpha]}. \tag{4.3}$$

Our parameter of interest for the negative binomial distribution is $\pi = \left(\frac{\alpha}{\alpha + \lambda_0}\right)^\alpha$.

Above, we have proposed three distributions that could plausibly motivate the distribution of rUMIs. Moments for these three distributions can be found in Section 8.1.

4.4 Estimators

4.4.1 Approach for Extrapolation

Our primary goal is estimating the unobserved probability that a molecule is not captured, i.e. π_i, based on the rUMI values that we do observe. We are attempting to estimate an unobserved quantity that falls outside the range of observed values, resulting in an extrapolation problem with all of its related challenges.

We propose above three plausible forms of the probability density function for the distribution of our observed C. Our approach is to identify estimators by imposing assumptions that imply a distributional form to the data generating process (e.g., Binomial, Poisson, or Negative Binomial). We then can take advantage of the assumed form to predict $\pi_{,K}$ from the existing information.

The distributional form of the probability density function is needed to calculate $\hat{\pi}_{,K}$. However, while we can assess the fit of our estimated probability density function to the observed distribution of C, this information in fact does not provide an assessment of the accuracy of the estimated probability of 0. We are extrapolating about the proportion of rUMIs that are 0, a value that falls outside of the observable range of rUMIs.

Because we are extrapolating, we anticipate that the rUMIs measured with small values will be more informative in predicting the probability of 0 rUMIs. Specifically, the small values of rUMI, i.e. 1s and 2s, are closer to 0 rUMIs than larger values, i.e. 6s and 7s, and therefore are plausibly better able to predict 0 rUMIs. The overall fit of the assumed distribution to the observed distribution provides evidence of a good representation of the observed data but does not indicate an accurate estimate of the probability of 0 rUMIs. In fact, a good fit of our estimated distribution to the observed distribution could also indicate overfitting to the data. Therefore, we anticipate well-fitting local probability estimates to be more informative of the accuracy of our estimator compared to an overall goodness of fit.

4.4.2 Estimator Definitions

We are primarily interested in estimating $\pi_{,K} \coloneqq \mathbb{P}(R_k = 0 \mid k \in K)$, the probability that a given transcript is not included in the final sample, i.e. 0 rUMI, for some group K of molecules. Secondarily, we can also estimate $\lambda_{,K}$, which is equivalent to the mean rUMI based on the distribution of rUMIs including the unobserved zeros proposed in each of the Special Cases in Section 4.3.2. Below, we provide estimators for $\lambda_{,K}$; we examine the performance of their corresponding probability estimates $\hat{\pi}_{,K}$ in Section 4.4.3 as described in Section 4.3.2.

4.4.2.1 Estimator 1: Moments Estimator

Under Special Case 2, $C \sim \text{ZTPoisson}(\lambda)$. Ordinarily, for a method of moments estimator, we would solve for λ after equating the sample mean with $\mathbb{E}(C)$ in Table 8.2. However, this equation does not provide a closed form estimator for λ. Therefore, we consider the first two moments together. Note that $\mathbb{E}(C^2) = \mathbb{E}(C)(1+\lambda)$, or equivalently that $\text{Var}(C) = \mathbb{E}(C)[1+\lambda-\mathbb{E}(C)]$. Using the first two moments from Table 8.2, our moments-based estimator is

$$\hat{\lambda}_1 = \frac{\hat{\sigma}^2}{\bar{x}} + \bar{x} - 1 = \frac{n\sum c_i^2}{(n-1)\sum c_i} - \frac{\sum c_i}{n(n-1)} - 1. \tag{4.4}$$

4.4.2.2 Estimator 2: Robust Estimator

We define $\hat{\lambda}_2$ motivated from the probabilities of the ZTPoisson distribution. $\hat{\lambda}_2$ is defined as

$$\hat{\lambda}_2 = 2\frac{\sum \mathbb{I}(C_i = 2)}{\sum \mathbb{I}(C_i = 1)}. \tag{4.5}$$

In other words, $\hat{\lambda}_2$ is 2 times the ratio of the number of rUMIs equal to two to the number of rUMIs equal to one. Taking the expectations of the indicators and assuming a ZTPoisson distribution simplifies $\hat{\lambda}_2$ to λ and provides the motivation for this estimator.

We describe $\hat{\lambda}_2$ as a robust estimator, because it is the only estimator that does not involve the sample moments. Specifically, $\hat{\lambda}_2$ is not influenced by large observed rUMI values through the mean in the same way as the other four estimators.

4.4.2.3 Estimator 3: Poisson Maximum Likelihood Estimator

The maximum likelihood estimator, $\hat{\lambda}_3$, from the ZTPoisson distribution is

$$\hat{\lambda}_3 = \left\{ \lambda : \bar{c} = \frac{\lambda}{1-e^{-\lambda}} \right\}. \tag{4.6}$$

There is no closed form solution for the MLE of the ZTPoisson distribution. Note that the maximum likelihood estimator $\hat{\lambda}_3$ is equivalent to the standard method of moments estimator and is calculated only using the sample mean, unlike $\hat{\lambda}_1$, which is calculated from the sample mean and the sample variance.

Note that if the sample consists entirely of 1 rUMI, then we set $\hat{\lambda}_3 = 0$ and $\hat{\pi}_3 = 1$.

4.4.2.4 Estimator 4: Binomial Maximum Likelihood Estimator

We next consider the maximum likelihood estimator from Special Case 1, the ZTBinomial distribution. Again, there is no closed form solution for the estimator, and $\hat{\lambda}_4$ is

$$\hat{\lambda}_4 = \left\{ \lambda : \bar{c} = \frac{\lambda}{1 - (1 - \frac{\lambda}{N_0})^{N_0}} \right\}. \tag{4.7}$$

Since N_0 is an additional parameter, we consider the possible values of N_0 when identifying the maximum likelihood. We scan over integer values for N_0 ranging from the maximum of the sample to 4,096 (2^{12}, selected based on the number of amplification rounds) or another large value, calculating the $\hat{\lambda}_4$ and corresponding likelihood function for each N_0. We then select the pair with the maximum likelihood value and use that to calculate the probability of zero rUMIs, $\hat{\pi}_4 = \left(1 - \frac{\hat{\lambda}_4}{\hat{N}_0}\right)^{\hat{N}_0}$. Note that in terms of λ, this is again the standard method of moments estimator but for the ZTBinomial distribution and assuming that N_0 is known.

Note that if the sample consists entirely of 1 rUMI, we set $\hat{\lambda}_4 = 0$, and $\hat{\pi}_4 = 1$.

4.4.2.5 Estimator 5: Negative Binomial Maximum Likelihood Estimator

We finally consider the maximum likelihood estimator from Special Case 3, the ZTNegative Binomial distribution. Again, there is no closed form solution for the estimator, and $\hat{\lambda}_5$ is

$$\hat{\lambda}_5 = \left\{ \lambda : \bar{c} = \frac{\lambda}{1 - (\frac{\alpha}{\alpha+\lambda})^{\alpha}} \right\}. \tag{4.8}$$

We optimize the likelihood function over values of α ranging from 1×10^{-9} to 200 and the corresponding $\hat{\lambda}_5$ estimates. From the optimal $\hat{\lambda}_5$ and $\hat{\alpha}$ pair, we calculate the probability of zero rUMIs, $\hat{\pi}_5 = \left(\frac{\hat{\alpha}}{\hat{\alpha}+\hat{\lambda}_5}\right)^{\hat{\alpha}}$. Note that $\hat{\lambda}_5$ is again the standard method of moments estimator but for the ZTNegative Binomial distribution assuming α is known.

Note that if the sample consists entirely of 1 rUMI, we set $\hat{\lambda}_5 = 0$, and $\hat{\pi}_5 = 1$.

4.4.3 Estimator Properties

4.4.3.1 Estimator 1: Moments Estimator

Our first estimator is given in Equation 4.4. For this estimator, we derive the asymptotic bias and variance using Taylor expansions in Section 8.2.1.

From these results, asymptotically in sample size, i.e. as $n \to \infty$,

$$\mathbb{E}(\hat{\lambda}_1) \to \frac{\mathbb{E}(C^2)}{\mathbb{E}(C)} - 1, \tag{4.9}$$

and

$$\text{Var}(\sqrt{n}\hat{\lambda}_1) \to \frac{\mathbb{E}(C^4) - \mathbb{E}(C^2)^2}{\mathbb{E}(C)^2} - \frac{2\mathbb{E}(C^2)(\mathbb{E}(C^3) - \mathbb{E}(C^2)\mathbb{E}(C))}{\mathbb{E}(C)^3} + \frac{\mathbb{E}(C^2)^2(\mathbb{E}(C^2) - \mathbb{E}(C)^2)}{\mathbb{E}(C)^4}. \tag{4.10}$$

The moments of our three special cases in Tables 8.1 to 8.3 can be combined with Equations 4.9 and 4.10 to calculate the asymptotic biases and variances of the $\hat{\lambda}_1$ in terms of the parameters. Under a ZTPoisson distribution (Special Case 2), $\hat{\lambda}_1$ is asymptotically unbiased and $\text{Var}(\sqrt{n}\hat{\lambda}_1) \to (\lambda + 2)(1 - e^{-\lambda})$. The asymptotic bias is $-\frac{\lambda}{N_0}$ with data drawn from a ZTBinomial distribution (Special Case 1) and is $\frac{\lambda}{\alpha}$ with data from a ZTNegative Binomial distribution (Special Case 3).

If the true model is a ZTPoisson distribution, the MSE for $\hat{\lambda}_1$ will approach 0 as the sample size grows. However, if the true model is either ZTBinomial or ZTNegative Binomial, the asymptotic biases will prevent the MSE from approaching 0.

4.4.3.2 Estimator 2: Robust Estimator

The formula for the robust estimator contains two variables. Observe that the variables $\sum \mathbb{I}(C_i = 1) \sim \text{Binomial}(n, \mathbb{P}(C = 1))$ and $\sum \mathbb{I}(C_i = 2) \sim \text{Binomial}(n, \mathbb{P}(C = 2))$ where $n = \sum Z_i$. With this observation, we derive asymptotic characterics of the estimator in Section 8.2.2.

Asymptotically,

$$\mathbb{E}(\hat{\lambda}_2) \to 2\frac{\mathbb{P}(C = 2)}{\mathbb{P}(C = 1)} \tag{4.11}$$

and

$$\text{Var}(\sqrt{n}\hat{\lambda}_2) \to 4\mathbb{P}(C = 2)\left(\frac{1 - 2\mathbb{P}(C = 2)}{\mathbb{P}(C = 1)^2} + 2\mathbb{P}(C = 2) + \frac{\mathbb{P}(C = 2)}{\mathbb{P}(C = 1)^3}\right). \tag{4.12}$$

We can solve for the asymptotic bias and variance by substituting values for $\mathbb{P}(C = 1)$ and $\mathbb{P}(C = 2)$ into the above formulas for various sampling schemes.

When $C \sim$ ZTPois(λ), $\mathbb{P}(C = 1) = \frac{e^{-\lambda}\lambda}{1-e^{-\lambda}}$ and $\mathbb{P}(C = 2) = \frac{e^{-\lambda}\lambda^2}{2(1-e^{-\lambda})}$. Therefore, $\hat{\lambda}_2$ is asymptotically unbiased and has an asymptotic variance of

$$\text{Var}(\sqrt{n}\hat{\lambda}_2) \to \frac{1}{e^{-\lambda}(1 - e^{-\lambda})^2}\left(2 - 6e^{-\lambda} - 2\lambda^2 e^{-\lambda} + 6e^{-2\lambda} + 4\lambda^2 e^{-2\lambda} - 2e^{-3\lambda} - 2\lambda^2 e^{-3\lambda} + 2\lambda^4 e^{-3\lambda} + \lambda - 3\lambda e^{-\lambda} + 3\lambda e^{-2\lambda} - \lambda e^{-3\lambda}\right).$$

Note again that the MSE for $\hat{\lambda}_2$ under a ZTPoisson distribution will approach 0 as the sample size increases. Additionally, under a ZTPoisson distribution, ARE$(\hat{\lambda}_1, \hat{\lambda}_2) \rightarrow e^{\lambda}$ as $\lambda \rightarrow \infty$ (Figure 4.1). Thus, $\hat{\lambda}_1$ and $\hat{\lambda}_2$ perform similarly for small λ, but $\hat{\lambda}_1$ is more efficient for larger λ.

The asymptotic biases for $\hat{\lambda}_2$ are $\frac{\lambda(\lambda-1)}{N_0-\lambda}$ when $C \sim$ ZTBinom $\left(N_0, \frac{\lambda}{N_0}\right)$ and $\frac{\lambda(1-\lambda)}{\lambda+\alpha}$ when C is drawn from a ZTNegative Binomial distribution. A description of the asymptotic variances can be found in Section 8.2.2 and can be generated by combining Equation 4.12 and probabilities from Equations 4.1 and 4.3.

4.4.3.3 Estimator 3: Poisson Maximum Likelihood Estimator

$\hat{\lambda}_3$ is the maximum likelihood estimator (mle) from the ZTPoisson distribution. We can obtain the asymptotic distribution of the mle from the Fisher Information for the correctly specified distribution. Specifically, when $C \sim$ ZTPois,

$$\sqrt{n}(\hat{\lambda}_3 - \lambda) \xrightarrow{d} \mathrm{N}\left(0, \frac{\lambda(e^{\lambda}-1)(1-e^{-\lambda})}{e^{\lambda}-1-\lambda}\right).$$

Note that although $\hat{\lambda}_3$ approaches the variance specified by the Cramér-Rao Lower Bound, it will not attain it.

The asymptotic relative efficiencies for combinations of $\hat{\lambda}_1$, $\hat{\lambda}_2$, and $\hat{\lambda}_3$ are shown in Figure 4.1. When data are drawn from a ZTPoisson distribution, the robust estimator $\hat{\lambda}_2$ has the largest variance, while the maximum likelihood estimator $\hat{\lambda}_3$ has the smallest.

The asymptotic bias and variance for the ZTBinomial and ZTNegative Binomial cannot be calculated. Estimates from simulations are discussed in 4.5.1.

4.4.3.4 Estimator 4: Binomial Maximum Likelihood Estimator

If we assume that N_0 is fixed and known in the ZTBinomial distribution, we can obtain the asymptotic distribution of $\hat{\lambda}_4$ when C is drawn from a ZTBinomial distribution. Specifically,

$$\sqrt{n}(\hat{\lambda}_4 - \lambda) \xrightarrow{d} \mathrm{N}\left(0, \frac{\lambda(N_0-\lambda)^2(N_0^{N_0}-(N_0-\lambda)^{N_0})^2}{N_0^{N_0+1}[(N_0-\lambda)(N_0^{N_0}-(N_0-\lambda)^{N_0})-\lambda N_0(N_0-\lambda)^{N_0}]}\right).$$

We cannot calculate the asymptotic biases and variances for the ZTPoisson or ZTNegative Binomial cases directly; we discuss estimates from simulations in Section 4.5.1. We additionally discuss estimates from simulations where N_0 is unknown and data are drawn from a ZTBinomial distribution in Section 4.5.1.

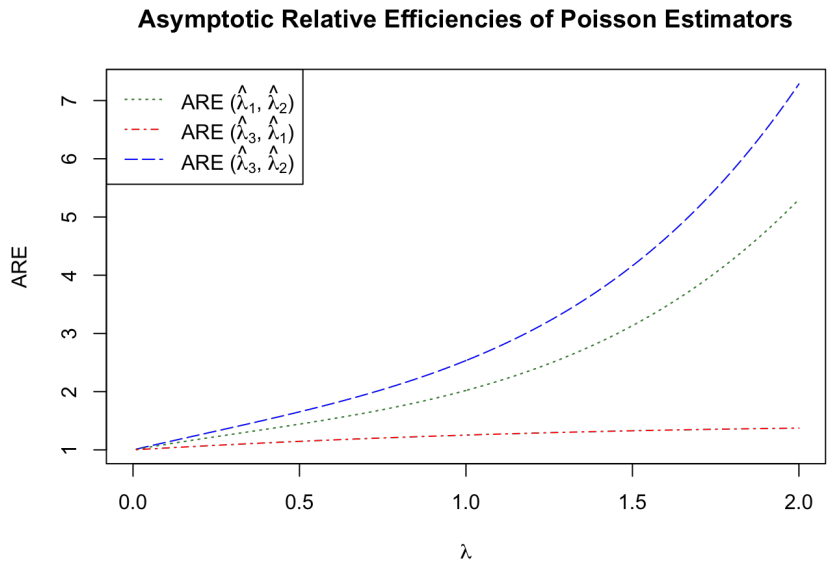

Figure 4.1: The asymptotic relative efficiencies of the ZTPoisson-based estimators $\hat{\lambda}_1$ (moment-based estimator), $\hat{\lambda}_2$ (robust estimator), and $\hat{\lambda}_3$ (maximum likelihood estimator).

4.4.3.5 Estimator 5: Negative Binomial Maximum Likelihood Estimator

If we assume that α is fixed and known in the ZTNegative Binomial distribution, we can obtain the asymptotic distribution of $\hat{\lambda}_5$ when C is drawn from a ZTNegative Binomial distribution. Specifically,

$$\sqrt{n}(\hat{\lambda}_5 - \lambda) \xrightarrow{d} \mathrm{N}\left(0, \frac{\lambda[(\alpha+\lambda)^\alpha - \alpha^\alpha]^2}{\alpha(\alpha+\lambda)^{\alpha-1}[(\alpha+\lambda)^\alpha - \alpha^\alpha] - \lambda\alpha^{\alpha+2}(\alpha+\lambda)^{\alpha-2}}\right).$$

We discuss the estimated asymptotic biases and variances from simulations for the ZTPoisson and ZTBinomial cases and for the ZTNegative Binomial case when α is unknown in Section 4.5.1.

4.5 Simulations

4.5.1 Estimator Performances

We simulate data generated according to each of the three special cases described in Section 4.3.2. For each generated sample, we apply our five estimators and record the values of $\hat{\lambda}$ and $\hat{\pi}$. The following tables summarize the results from these simulations.

Tables 4.1 and 8.4 provide the MSEs for $\hat{\pi}$ and $\hat{\lambda}$ respectively for each of the five estimators

Distribution			MSE $\times 10^3$				
λ	Family	π	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	$\hat{\pi}_5$
0.1	ZTP	0.905	0.001	0.001	0.001	0.010	0.011
	ZTB	0.904	0.580	0.497	0.552	0.013	0.568
	ZTNB	0.906	0.550	0.430	0.510	0.511	0.024
0.4	ZTP	0.670	0.001	0.002	0.001	0.006	0.008
	ZTB	0.656	7.168	3.640	5.942	0.002	6.130
	ZTNB	0.683	5.850	2.326	4.510	4.520	0.016
1.5	ZTP	0.223	0.001	0.003	0.000	0.001	0.005
	ZTB	0.153	29.603	0.161	16.018	0.000	16.591
	ZTNB	0.280	15.975	0.582	9.549	9.569	0.003

Table 4.1: Mean squared error (multiplied by 10^3) for the estimated probability of 0 from three data generating processes. 500 samples of size 250,000 were collected. π represents the true probability of zero for each of the three distributions. $N = 4$ for the ZTB distribution, and $\alpha = 4$ for the ZTNB distribution.

and from each of the three data generating processes. We consider three different true λ values of 0.1, 0.4, and 1.5. We set $N_0 = 4$ for the ZTBinomial and $\alpha = 4$ for the ZTNegative Binomial case. We simulate a sample of size 250,000 from each of the zero truncated models as specified, taking 500 repeated samples. Note that the true π in Table 4.1 varies for the different data generating functions.

The estimators derived from the correctly specified model often perform the best in terms of the MSE. When the true distribution of observed rUMIs is ZTPoisson, $\hat{\pi}_2$ performs worse than $\hat{\pi}_1$ and $\hat{\pi}_3$, the other two estimators derived from the ZTPoisson distribution. When the true distribution is ZTBinomial, the estimator derived from the ZTBinomial distribution, $\hat{\pi}_4$, performs the best, followed by $\hat{\pi}_2$. The best performing estimator when the true distribution is ZTNegative Binomial is $\hat{\pi}_5$, the estimator derived from the ZTNegative Binomial distribution, again followed by $\hat{\pi}_2$.

We examine how adjusting the additional parameter in the ZTBinomial distribution (N) and in the ZTNegative Binomial distribution (α) affects the five estimators. We continue generating samples of size 250,000 as before, but we simulate 100 samples instead of 500. We set $\lambda = 0.4$ for these simulations.

Table 4.2 indicates that $\hat{\pi}_4$ seems to perform best when N is small followed by $\hat{\pi}_2$. As N increases, the ZTBinomial distribution begins to more closely match the ZTPoisson distribution. When $N \leq 64$, the MSE of $\hat{\pi}_4$ is smaller than the other estimators; when $N = 256$, however, $\hat{\pi}_3$ and $\hat{\pi}_1$, estimators derived from the ZTPoisson distribution, have surpassed the estimator from the correctly specified distribution, likely due to a close resemblance of the ZTBinomial to the ZTPoisson distribution at this point.

Table 4.3 presents MSE estimates for $\hat{\pi}$ based on data generated from ZTNegative Binomial

N	π	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	$\hat{\pi}_5$
2	0.640	31.918	19.235	27.831	0.002	28.126
4	0.656	7.147	3.622	5.922	0.002	6.110
8	0.663	1.706	0.807	1.385	0.013	1.485
16	0.667	0.414	0.191	0.334	0.010	0.386
64	0.669	0.027	0.013	0.022	0.009	0.037
256	0.670	0.003	0.002	0.002	0.007	0.010
1024	0.670	0.002	0.002	0.002	0.007	0.010

Table 4.2: Mean squared error (multiplied by 10^3) for the estimated probability of 0 from ZT-Binomial data with varying N. 100 samples of size 250,000 were collected.

α	π	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	$\hat{\pi}_5$
1	0.714	70.248	22.353	50.771	50.814	0.016
5	0.681	3.833	1.560	2.974	2.982	0.011
20	0.673	0.254	0.110	0.200	0.202	0.016
50	0.671	0.041	0.019	0.033	0.033	0.010
70	0.671	0.023	0.013	0.019	0.021	0.010
100	0.671	0.012	0.008	0.010	0.015	0.006

Table 4.3: Mean squared error (multiplied by 10^3) for the estimated probability of 0 from ZTNegative Binomial data with varying α. 100 samples of size 250,000 were collected.

distributions with varying α. The estimator from the correctly specified model, $\hat{\pi}_5$ performs best for the α considered in this table followed by the robust $\hat{\pi}_2$. The advantages of both of these estimators decrease as α increases and the ZTNegative Binomial distribution more closely approximates the ZTPoisson one.

Figures 8.1 to 8.7 display the estimated asymptotic biases and variances for our five estimators under the three different data generating functions along with the theoretical results obtained in Section 4.4.3. In particular, we note that the MSEs only approach 0 when the estimator is unbiased. Additionally, we note that the estimation of the second parameter for $\hat{\lambda}_4$ and $\hat{\lambda}_5$ do affect the asymptotic biases and variances and quite substantially.

Our simulations indicate that the robust estimator $\hat{\pi}_2$ performs well across most situations, although another estimator can typically outperform it. Generally, the performance of $\hat{\pi}_2$ does not rely on a specific underlying distribution or set of parameters to perform well. The performance of the other estimators, however, do rely on the underlying distribution. The corresponding tables for the $\hat{\lambda}$ estimates can be found in Tables 8.4 to 8.6.

Distribution			MSE $\times 10^3$				
Collision	Family	π	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	$\hat{\pi}_5$
0.1%	ZTP	0.670	0.002	0.002	0.001	0.007	0.007
	ZTB	0.656	4.469	2.563	3.779	0.941	4.287
	ZTNB	0.683	5.964	2.380	4.597	4.608	0.014
0.5%	ZTP	0.670	0.010	0.006	0.008	0.010	0.008
	ZTB	0.656	4.510	2.422	3.776	0.382	3.932
	ZTNB	0.683	6.322	2.464	4.839	4.850	0.022
1%	ZTP	0.670	0.029	0.015	0.023	0.025	0.011
	ZTB	0.656	4.444	2.331	3.703	0.145	3.858
	ZTNB	0.683	6.813	2.663	5.207	5.218	0.017
2%	ZTP	0.670	0.111	0.054	0.088	0.089	0.008
	ZTB	0.656	4.448	2.354	3.719	0.129	3.875
	ZTNB	0.683	7.791	2.962	5.889	5.901	0.029

Table 4.4: Mean squared error (multiplied by 10^3) for the estimated probability of 0 from data with varying levels of collisions. 100 samples of size 250,000 were collected. True generating parameters are $\lambda = 0.4$, $N = 4$, and $\alpha = 4$.

4.5.2 Modifications to Special Cases

The special cases described in Section 4.3.2 present idealized versions of data generating functions. However, we suspect that various perturbations to these special cases may exist. We simulate data with various modifications and perturbations, evaluating the five estimators in each of these situations.

4.5.2.1 Collisions

We define a collision as the situation where two distinct molecules attach to identical UMIs.

Note that there are two classes of collisions: observable and unobservable, depending on whether the transcripts originate from different or the same gene, respectively. From the data generated by Green *et al.* [2018], we can identify observable collisions as two (or more) molecules aligned to different genes with the same UMI.

We observe in Chapter 3 that UMIs are not in fact perfectly unique using scRNA-seq drop-seq data [Green *et al.*, 2018]. Therefore, we use the gene as an additional characteristic to distinguish two distinct molecules that attach to identical UMIs in the data in an observable collision. In fact, it is not surprising that there are some observable collisions, as there are 65,536 distinct molecular barcodes and 244,272 molecules in the top 10 cells.

We simulate data assuming that unobserved collisions occur for some portion of the UMIs. Our data generating process is modified from that described in Section 4.5.1. First, we simu-

late our non-collision data from a ZTPoisson (λ), ZTBinomial $(N_0, \frac{\lambda}{N_0})$, or ZTNegative Binomial $(\alpha, \frac{\lambda}{\alpha+\lambda})$ distribution, as before. We simulate our collision data from a ZTPoisson (2λ), ZTBinomial $(2N_0, \frac{\lambda}{2N_0})$, or ZTNegative Binomial $(\alpha, \frac{2\lambda}{\alpha+2\lambda})$ distribution, respectively. We take random samples of size 250,000 from the non-zero values simulated from the non-collision and collision data, mixed at the specified collision rate. We simulate 100 samples of 250,000 UMIs in total. We set $\lambda = 0.4$, $N_0 = 4$, and $\alpha = 4$.

Table 4.4 displays the simulation results for data including unobserved collisions occurring at different rates, with Table 8.7 displaying the analogous results for the respective $\hat{\lambda}$ estimators. The true π is calculated for the non-collision data. The results from Table 8.7 are similar to those described in Section 4.5.1. Specifically, the estimator derived from the correct data generating function typically has the best performance of the five estimators considered; the robust estimator $\hat{\pi}_2$ typically follows this estimator in performance. As the collision rate increases, the ZTPoisson distribution resembles the ZTNegative Binomial distribution. $\hat{\pi}_5$ based on the ZTNegative Binomial distribution then has the best performance followed by $\hat{\pi}_2$. Generally, the advantage of the robust $\hat{\pi}_2$ increases as the collision rate increases.

We chose our simulated collision rates from observed collision rates in scRNA-seq data [Green *et al.*, 2018]. Specifically, we first estimated the probability that two randomly selected molecules share a UMI. From this estimated probability, we calculated the number of effective UMIs as 37,452. Specifically, the probability that two randomly selected molecules share a UMI from our observed data is 2.67×10^{-5}. If our UMIs were perfectly random, this corresponds to 37,452 possible UMIs.

The unobserved collision rate depends on the expression level of individual genes. Therefore, we increased the observed expression level of a gene from our data three-fold to more closely approximate the true expression level of the cell before the estimated loss of molecules prior to assignment of UMIs (Table 4.6). We simulated the assignment of molecules to one of the effective UMIs based on our expression level of a gene. We estimate that unobserved collisions occur for approximately 0.1% of the UMIs in the data collected by Green *et al.* [2018]. Almost all of our simulated collisions had two molecules assigned to one UMI, while a handful had three molecules assigned to the same UMI. Table 8.7 is the analogous table to Table 4.4 for $\hat{\lambda}$.

4.5.2.2 Corruption

We noted in Chapter 3 that the observed distribution of rUMIs had a long right tail, extending to 9 rUMI. This characteristic in particular led us to develop the robust estimator $\hat{\lambda}_2$ that would not be affected by the right skew. Therefore, we were interested in representing this skewness in our simulated data. We first considered modifying our simulated data with collisions with our primary collision rate estimated from real data. From the simulated samples of rUMIs with collisions in

Table 4.4, 1 of the 400 ZTPoisson samples had 9 rUMIs and 47 of the 400 ZTNegative Binomial samples had 9 or 10 rUMIs. We therefore considered a direct perturbation of the data generating procedure.

We artificially require 25 of the 250,000 simulated molecules to have rUMIs of 6-10 with respective probabilities of 5%, 35%, 20%, 35%, and 5%. We refer to the artificially high rUMIs as corrupt rUMIs. For the remaining 249,975 molecules, we simulate them according to a ZTPoisson, ZTBinomial, or ZTNegative Binomial distribution with $\lambda = 0.4$, $N_0 = 16$, and $\alpha = 4$. For this simulation, we increased N_0 from 4 to 16 to reduce the separation between the simulated rUMIs and the corrupt rUMIs for the ZTBinomial data.

Table 4.5 shows that $\hat{\pi}_2$ is one of the most consistently well performing estimators, although it is not always optimal. Table 8.8 represents an analogous table to Table 4.5 for the respective $\hat{\lambda}$ estimates.

Distribution			MSE $\times 10^3$				
λ	Family	π	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	$\hat{\pi}_5$
0.1	ZTP	0.905	0.024	0.001	0.002	0.002	2.888
	ZTB	0.905	0.002	0.030	0.021	0.021	3.290
	ZTNB	0.906	0.791	0.431	0.565	0.566	1.851
0.4	ZTP	0.670	0.011	0.002	0.002	0.002	0.231
	ZTB	0.667	0.299	0.191	0.304	0.282	0.391
	ZTNB	0.683	6.245	2.322	4.604	4.615	0.090
1.5	ZTP	0.223	0.001	0.003	0.000	0.001	0.005
	ZTB	0.207	1.402	0.027	0.800	0.003	0.930
	ZTNB	0.280	16.056	0.581	9.579	9.600	0.003

Table 4.5: Mean squared error (multiplied by 10^3) for the estimated probability of 0 from three data generating processes with corruption ($N_0 = 16$ and $\alpha = 4$). 500 samples of size 250,000 were collected.

4.6 Sertoli Cell Drop-seq Data

We apply our five estimators to the data described in Chapter 3. Table 4.6 provides a summary of the results. The bottom two rows of Table 4.6 calculate the five estimators without taking into account the observable collisions. In other words, the top two rows use the gene as a deduplication factor, while the bottom two rows only use the cellular and molecular barcode to deduplicate the UMIs. The observable collisions change the estimates quite substantially. Unobserved collisions could additionally affect the estimates.

We noted in Chapter 3 that there was a relationship between $\hat{\lambda}$ and the GC content of a transcript

	Estimator 1	Estimator 2	Estimator 3	Estimator 4	Estimator 5
$\hat{\lambda}$	0.4757	0.3804	0.4408	0.4409 (4096)	0.2228 (0.882)
$\hat{\pi}$	0.6215	0.6836	0.6435	0.6434	0.8198
$\hat{\lambda}$	1.2788	0.7424	1.0468	1.0470 (4096)	0.5173 (0.695)
$\hat{\pi}$	0.2784	0.4760	0.3511	0.3509	0.6793

Table 4.6: The estimated $\hat{\lambda}$ and associated probabilities from the Sertoli cell sample. The top two rows are calculated from the sample using gene as a deduplication factor while the bottom two rows do not use gene. For estimators 4 and 5, the $\hat{N}_0$ and $\hat{\alpha}$ associated with the MLE are provided in parentheses along with the $\hat{\lambda}$ value.

(Figure 3.7). We have provided an extension of this figure in the left panel of Figure 4.2, where we estimate the $\hat{\lambda}$ for each of our five estimators. The numbers in the figure represent the estimator used to calculate the given estimate; in other words, the 1 in the figure corresponds to $\hat{\lambda}_1$, the 2 to $\hat{\lambda}_2$, and so on. The size of the estimators appear consistent with the full sample estimates in Table 4.6. We observe a similar shape for each of the five estimators, indicating that there does appear to be a relationship between the rate of 0 rUMIs and the GC content of a transcript.

The $\hat{\lambda}$ estimates in Figure 4.2 are different for most of the estimators; in fact, the associated $\hat{\pi}$ vary from 59.0% to 88.2% (Figure 8.10); see Figures 8.8, 8.9, 8.11, and 8.12 for analogous estimates calculated for twenty additional Sertoli cells. The robust estimator is often near the middle in terms of the estimates. We also note that $\hat{\lambda}_3$ and $\hat{\lambda}_4$ are very near each other; this is in part due to the large $\hat{N}$, which makes the estimated Binomial distribution very similar to the Poisson distribution.

Because $\hat{\pi}$ is extrapolated, it is difficult to know which estimator provides the best estimate. From simulations, we have observed that $\hat{\lambda}_2$ tends to have a low MSE under different data generating processes. Additionally, $\hat{\lambda}_2$ is not the most extreme estimate of the five options. However, the scientific understanding is the same regardless of the estimator selected. An imbalance of AT and GC in the transcript has lower mean rUMI than transcripts that are more balanced.

4.7 Hicks (2018) Revisited

Hicks *et al.* [2018] describe technical variability that occurs in scRNA-seq data by studying multiple datasets. As part of their investigation, they examine how scRNA-seq data compare to bulk RNA-seq data performed on the same sample of cells. Ignoring any technical effects, one would expect the average of the gene expression across the cells in scRNA-seq data would closely approximate the measured expression from bulk RNA-seq data. Reproducing work performed by Shalek *et al.* [2013], they generate a graph comparing the average of logs of the single-cell and

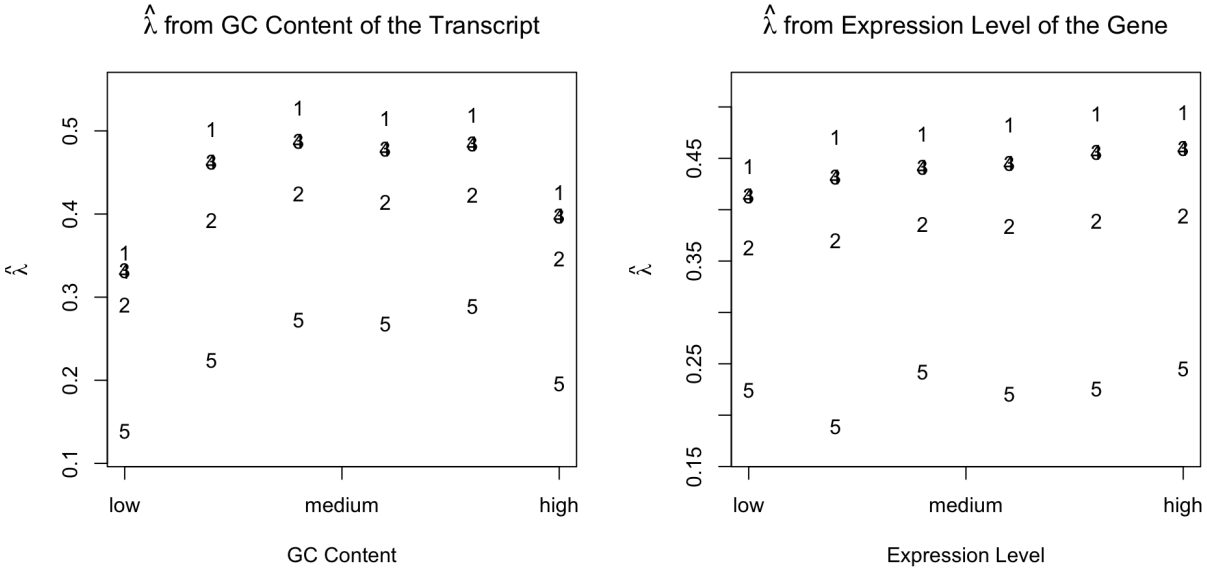

Figure 4.2: The estimated $\hat{\lambda}$ values for each of the five estimators. The left panel shows an extension of Figure 3.7, estimating λ based on the GC content of the transcript. The right panel estimates λ based on the expression level of the gene.

bulk profiles with the difference between the logs of the single-cell and bulk profiles (Figure 2 and Figure S11). They then fit a smoothed line to the data by binning the values on the x-axis and calculating the mean difference for a given bin, identifying that lowly expressed genes have larger measurements in bulk data than in their single-cell counterparts.

Hicks *et al.* [2018] in essence argue that genes that are lowly expressed are more likely to experience 0 rUMI than genes that are highly expressed. In other words, Hicks *et al.* [2018] indicate that the probability that a given molecule is captured is not equally likely for all molecules but depends on the expression level of that gene. We use a different approach to assess this statement; we estimate the λ parameter after subsetting the molecules into categories based on the expression level of their associated gene. The right panel of Figure 4.2 displays the estimates based on the expression level of the gene. We note a general trend that as the expression level of a cell increases, so does the $\hat{\lambda}$ regardless of the estimator. The corresponding $\hat{\pi}$ estimates from these $\hat{\lambda}$ differ by less than 4.5%; in fact, the $\hat{\pi}_2$ estimates differ by 2.1%. For comparison, the $\hat{\pi}$ based on the GC content have a range of at least 9% for all estimators.

Our results appear to be somewhat discordant with those obtained by Hicks *et al.* [2018]. Therefore, further investigation is warranted to understand these two results jointly.

4.8 Discussion

In this chapter, we have developed estimators for rUMI, furthering the work introduced in Chapter 3. We describe the characteristics of the distribution of rUMIs, including that rUMIs take positive integer values. We provide motivation and details surrounding three plausible distributions for rUMIs, the ZTBinomial, ZTPoisson, and ZTNegative Binomial distributions. Additionally, we propose five estimators derived from these three distributions: maximum likelihood estimators for each of the three distributions, a closed form methods-based estimator for the ZTPoisson distribution, and a robust estimator for the ZTPoisson distribution. We study asymptotic properties of these estimators both theoretically where obtainable and through simulations to understand the biases and uncertainties associated with the estimators.

No one estimator has universally ideal properties. In particular, the estimators derived from the true data generating function appear to perform best based on the MSE. However, the robust $\hat{\lambda}_2$ typically is optimal after the estimator from the correctly specified model. We examine perturbations to the data generating processes, including collisions and corruption. Collisions occur when two or more transcripts are assigned to the same UMI; these collisions are unobservable when the transcripts originate from the same gene. Corruption modifies a given distribution by sampling a portion of observations from high values for rUMI that are unlikely to be sampled from the data generating distribution by chance. When high levels of unobserved collisions or corruption are present, $\hat{\lambda}_2$ tends to perform better than most of the other estimators.

We apply our estimator to scRNA-seq data in two ways. First, we examine the relationship of the five different estimators with the GC content of the transcript. The five estimators identify the same pattern between the mean rUMI and the GC content of the transcript; that is, we see that an imbalance in GC and AT in a transcript results in smaller estimated mean rUMIs. The values for the estimates do differ between the five estimators. Additionally, we note that Hicks *et al.* [2018] claim that genes that are more lowly expressed result in smaller mean rUMIs based on joint single-cell and bulk RNA sequencing experiments. We apply that claim to our data, estimating the mean rUMI based on the expression level of the gene. We note a modest increase in mean rUMI for more highly expressed genes, although our observed increase is not consistent with the size of the effect noted by Hicks *et al.* [2018]. This inconsistency warrants further investigation.

We note that our five estimators do in fact provide different values for the estimates in Table 4.6 and in Figure 4.2. The variability of the five estimators seems inconsistent with the MSEs calculated in Tables 8.4 to 8.8. Specifically, the difference between $\hat{\lambda}_1$ and $\hat{\lambda}_5$ seems quite large relative to the observed MSEs. Thus, it appears that our simulated data does not represent the observed distribution of rUMIs well, as the MSEs from our simulated data are not consistent with those that would result from our observed estimates in Figure 4.2. An area for further exploration

is to identify differences between our simulated data and the observed distribution or to consider additional perturbations to simulated data in order to generate MSEs that appear consistent with the variability in our observed estimators.

Overall, we were interested in estimating the true proportion of zero rUMIs. This extrapolation problem can be informed by the observed rUMI distribution but cannot be assessed purely based on the observed data. As discussed in Chapter 3, we are interested in estimating the technical biases present in an experiment through the rUMI measure. The proportion of zero rUMI serves as a way to quantify some of the technical biases in a scRNA-seq experiment. Specifically, we would like to predict the proportion of molecules not included in our reads based on various molecule and gene characteristics. We have demonstrated that we can estimate the unobservable proportion of zero rUMI using five different estimators. These estimators can be applied to different datasets to infer the reliable detection of certain sets of genes.

Chapter 5

Conclusions

In this work, we have presented two approaches to address biases that are present in scRNA-seq data. Both of these two approaches work to identify biases with characteristics that can be leveraged in normalization procedures to help data more accurately represent what we expect to be the true biological gene expression at the cellular level.

In Chapter 2, we first define different notion of stably expressed genes at the cellular level. In particular, we define these notions based on the true biological nature of the cell rather than the measurements obtained from scRNA-seq experiments. Absolute stability indicate that a gene is expressed with approximately the same number of transcripts in each cell, while proportional stability and stability in concentration indicate that a gene is expressed approximately constantly with respect to either the total RNA content of the cell or the cell volume, respectively. Genes exhibiting different notions of stability are capable of capturing different types of technical effects. The different notions of stability can define different types of differential expression. However, scRNA-seq data is subjected to many technical effects, and identifying stably expressed genes can be challenging. Specifically, different normalizations can obscure or highlight different types of stably expressed genes. Therefore, we examine multiple scRNA-seq datasets to identify genes that appear to be generally stably expressed, organizing our analysis by studying sets of genes that are associated with cellular structures. We motivate our analysis by proposing a data generating model for the measured expression based on the true expression and technical factors and interpret our results through this model. We find that the set of cytosolic ribosomal genes is enriched with proportionally stable genes, based on both scRNA-seq and bulk GTEx data, indicating the potential for their use in downstream normalization procedures. We also generate a database of gene information that can be used to further evaluate and refine sets of stably expressed genes based on scRNA-seq and GTEx data.

In Chapter 3, we approach scRNA-seq data by considering the number of reads sequenced for each unique molecular identifier (UMI). The rUMI measure can be applied to scRNA-seq data in two primary ways: as a way to estimate the unobserved probability that a given molecule will have 0 rUMIs, i.e. be excluded from the reads due to technical loss, and as a way to identify genes

that are reliably detected, i.e. have a large estimated rUMI value. We begin by proposing a zero-truncated Poisson distribution to model the rUMI distribution and an estimator for the associated λ. We create categories for a variable, calculating our $\hat{\lambda}$ for each category and observing its relationship with rUMI. We find that the proportion of guanine and cytosine in the first half of the transcript is particularly informative in estimating mean rUMI, with an imbalance resulting in less reliable detection. We consider additional characteristics, including transcript length, two-somes, base composition of the cellular barcode, and base composition of the molecular barcode, with less informative results.

In Chapter 4, we extend and strengthen the theoretical framework and results introduced in Chapter 3. Specifically, we motivate potential rUMI distributions with assumptions about the underlying processes that contribute to the sample rUMIs, suggesting three possible underlying zero-truncated distributions. We then propose five estimators both for the distributional parameters and for the probability of 0 rUMIs resulting from the parameters. We aim to extrapolate the true proportion of 0 rUMIs due to technical loss from our observed rUMI distribution both generally and as a function of characteristics of the molecules. We study the properties of our five estimators under different data generating functions and perturbations. The robust estimator derived from the ZTPoisson distribution and calculated as 2 times the ratio of the number of 2 rUMIs to the number of 1 rUMIs performs well under all data generating processes although is rarely the optimal choice in any given situation based on its MSE. Finally, we estimate the probability of technical loss based on the observed expression level of the gene with our five estimators, finding that there does appear to be a marginal increase in technical loss of molecules among lowly expressed genes.

The removal of biases from scRNA-seq data is challenging. However, we have generated a list of proportionally stably expressed genes that can assist in normalizing the data to remove technical effects associated with cell size. In addition, we have identified characteristics that are indicative of reliable detection, and conversely can be estimated as indicative of unreliable detection. These two findings could be used for normalization separately or could be combined to generate a list of genes that appear to be both proportionally stably expressed and reliably detected, two characteristics that are both favorable for genes to be used in normalizations.

While these results are encouraging, and indeed illuminating for scRNA-seq experiments, there are additional avenues of research stemming from these biologically motivated results.

First, we have proposed a set of proportionally stable genes, but we have thus far struggled to identify any genes that appear absolutely stable in scRNA-seq data. We attempted to leverage the stable "expression" of external spike-ins as described in Chapter 2 but found that the spike-ins appeared more stable than endogenous measurements. We interpret this as indicative of stronger technical effects present in endogenous genes compared to the external spike-ins, i.e. a more variable t than u from Model 2.1. Thus, we would like to consider leveraging other endogenous

genes that have been suggested as biologically plausible to be absolutely stably expressed. In addition, we would like to further refine the cellular structures that we have considered. Some of the structures are quite large, and subsetting them into smaller categories may provide more specific gene sets. These provide two preliminary approaches as a starting point for identifying absolutely stably expressed genes.

Second, we have proposed an approach to quantify reliable detection of genes with rUMI by estimating the proportion of 0 rUMIs. How reliably detectable a gene is could inform our interpretation of gene expression measurements, indicating which measurements are more accurate. We would like to further explore the rUMI measure in additional scRNA-seq protocols, experiments, and tissue types to assess the generalizability of our given estimators. The approach could be applied to any scRNA-seq data with UMIs, but similar results across multiple experiments would suggest that similar technical effects are present across all scRNA-seq experiments. Additionally, we proposed frameworks for the true distribution of rUMIs. We would like to continue exploring the fit of various distributions with different perturbations to the true data to model it more accurately, as well as identifying the conditions that are optimal for each of the estimators.

Finally, we would like to explore exactly how best to incorporate cytosolic ribosomal genes and other stably and reliably expressed genes into normalization procedures. We would like to present a method to remove some technical biases from scRNA-seq data. While we have made preliminary advances towards this end goal by defining sets of genes that can be applied to existing normalization methods, we would like to understand how well these procedures work in adjusting scRNA-seq data for technical effects.

Chapter 6

Supplement: Stably Expressed Genes

6.1 Additional Single-Cell Figures

We provide in this section figures from scRNA-seq data that are referenced in Chapter 2 (Figures 6.1 to 6.6). These figures provide additional details to support conclusions drawn in Chapter 2.

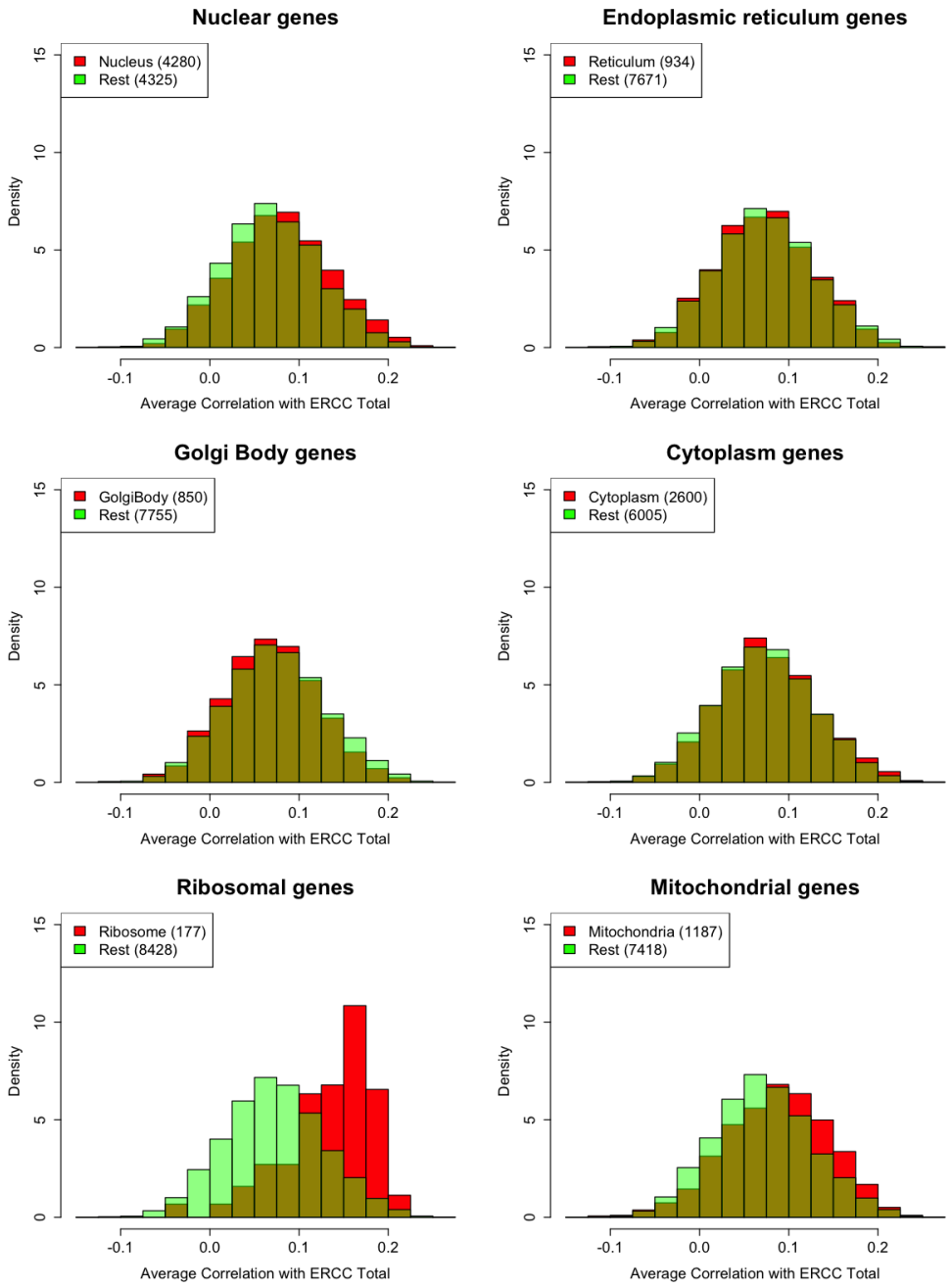

Figure 6.1: Histograms of the average correlations of each gene with the ERCC total over the six C1 datasets, with comparisons between the gene set of interest and the remaining genes. Note that the x-axis ranges from -0.15 to 0.275.

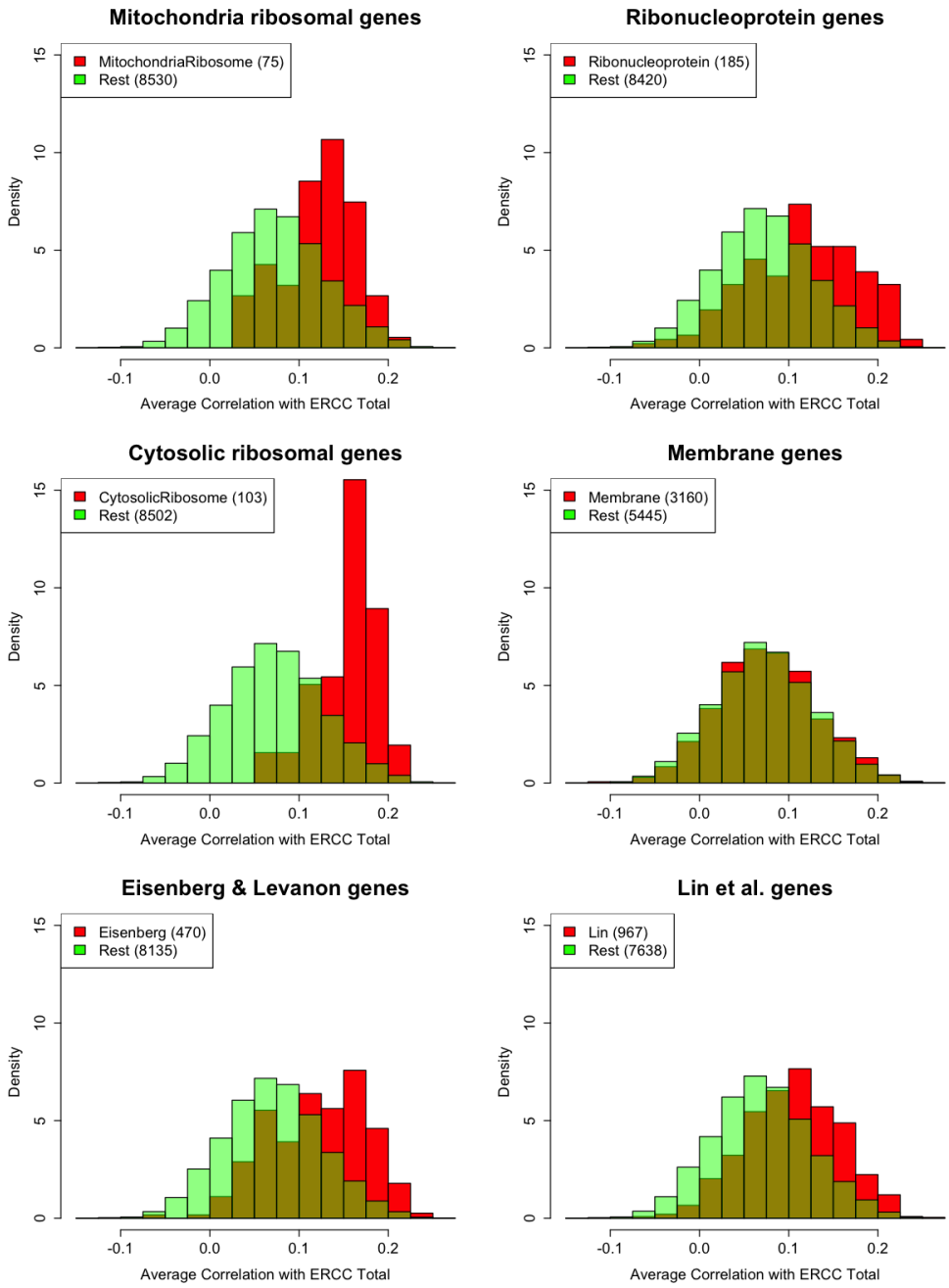

Figure 6.2: Continuation of Figure 6.1

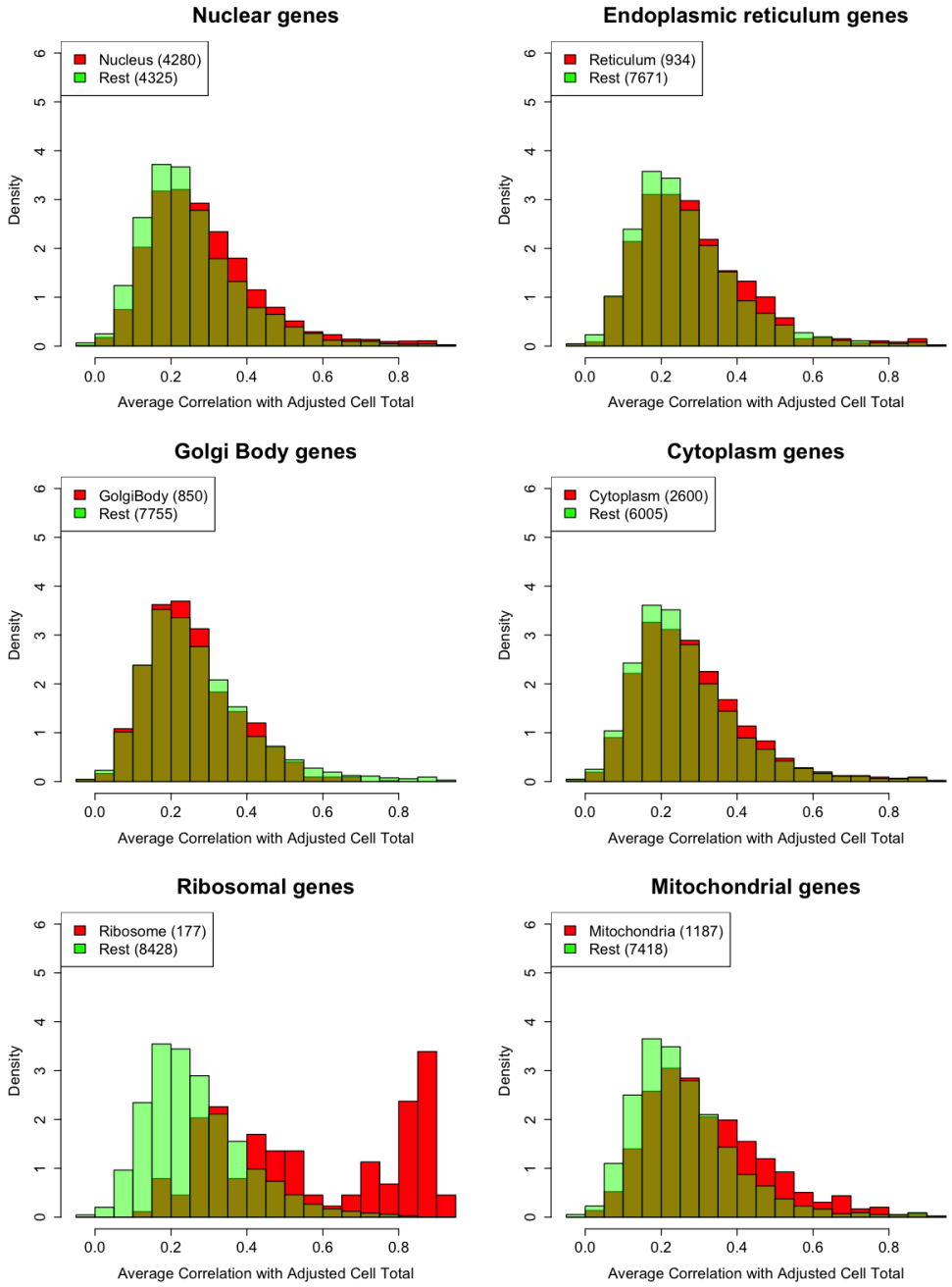

Figure 6.3: Histograms of the average correlations of each gene with the adjusted cell total over the six C1 datasets, with comparisons between the set of genes of interest and the remaining genes. Note that, unlike Figures 6.1 and 6.2, the x-axis ranges from -0.05 to 0.95.

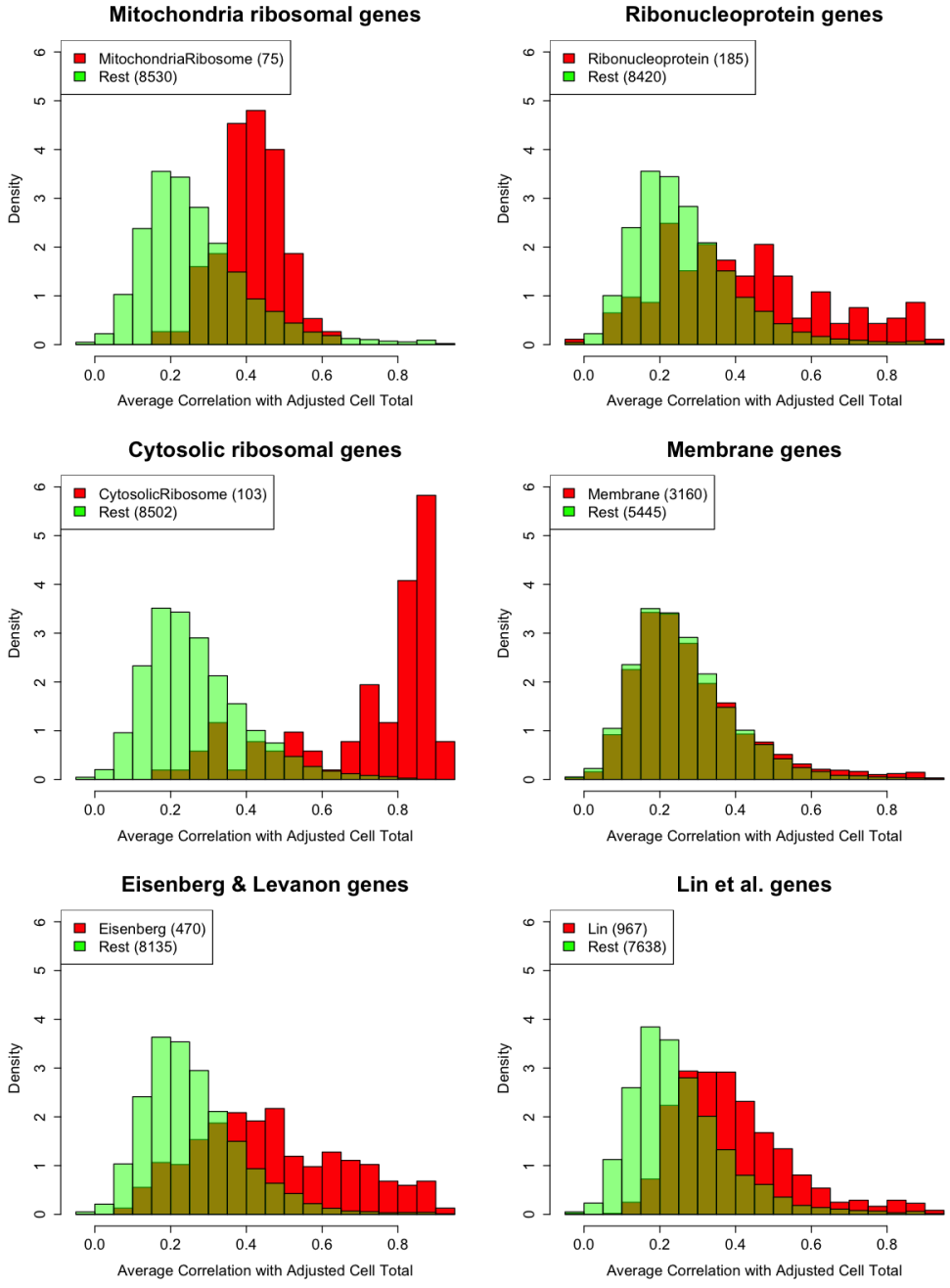

Figure 6.4: Continuation of Figure 6.3.

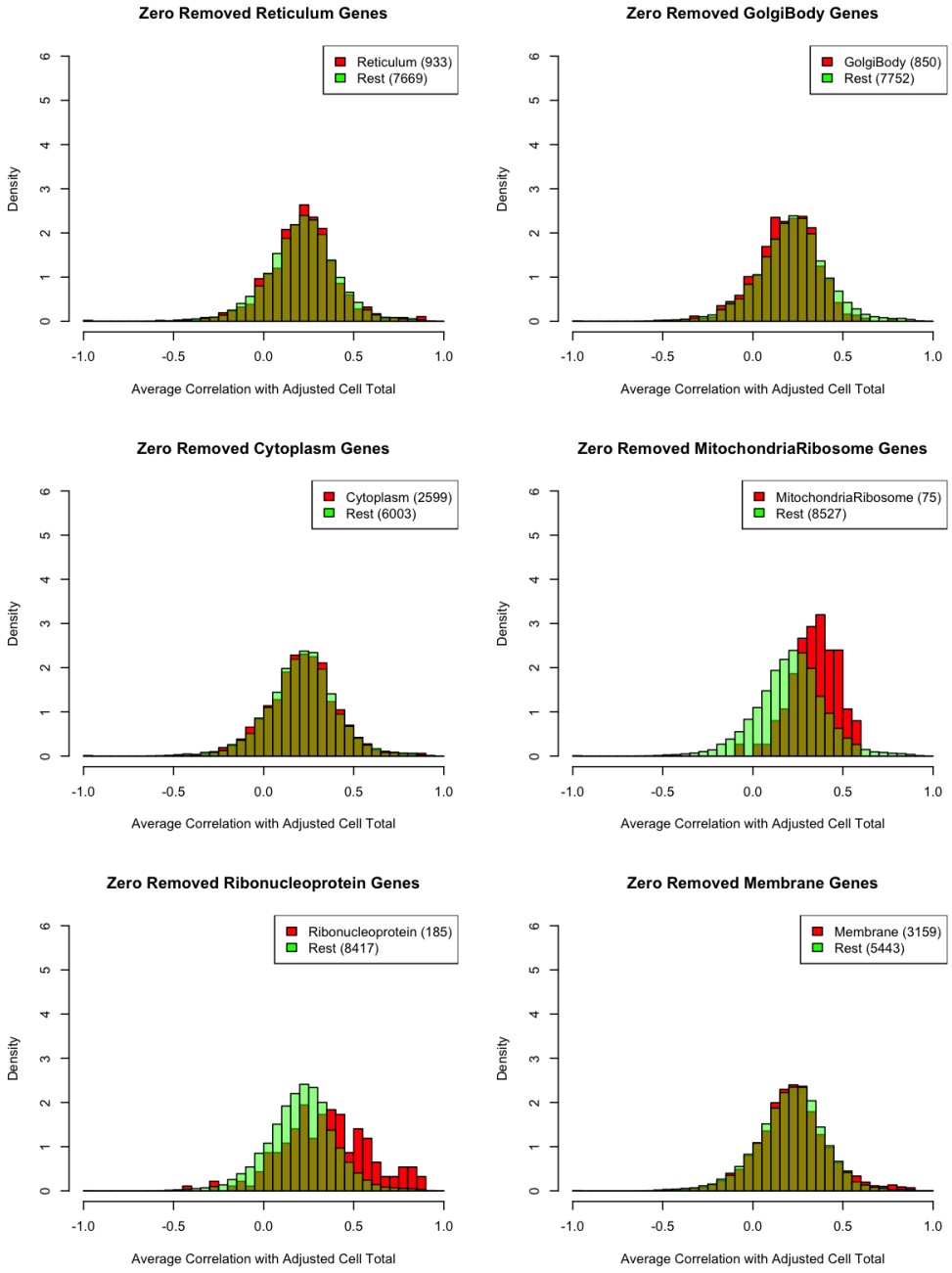

Figure 6.5: Continuation of Figure 2.9.

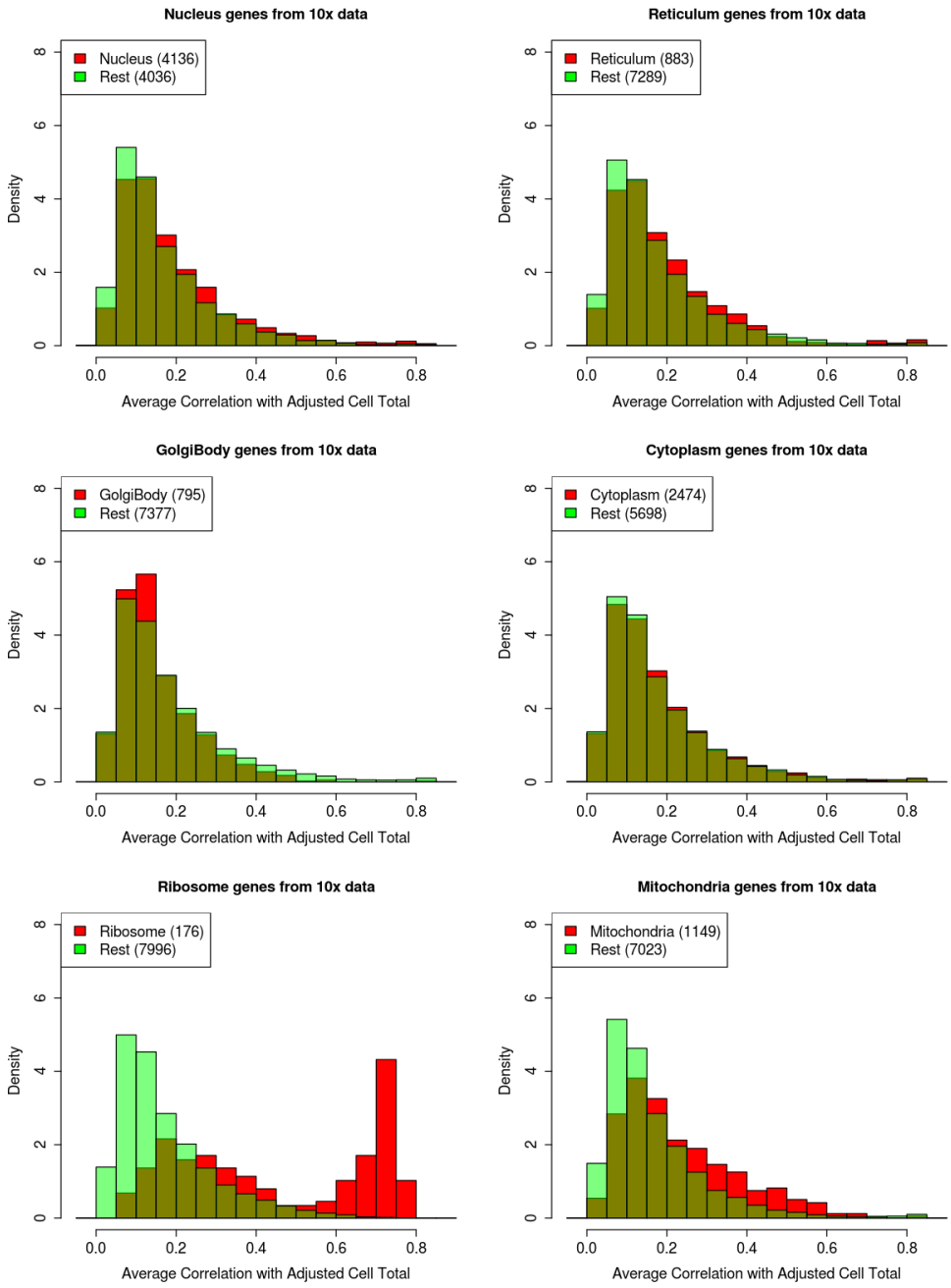

Figure 6.6: Continuation of Figure 2.12.

6.2 Gene Summaries

6.2.1 Structural Annotation Dictionary

Our structural annotation dictionary combines a list of human genes from Ensembl's Biomart with the Gene Ontology Consortium's (GO) information on genes gathered with `mygene` in R [Wu *et al.*, 2013; Xin *et al.*, 2016]. Gene names, not Ensembl identifiers, were used.

The query through `mygene` returns the top matches from GO for a given gene name. We retained only the top match. Any genes not identified from GO were removed. The biological process, molecular function, and cellular component were returned from GO. Each of these fields could have no values or multiple values.

We considered the 1,629 unique cellular components returned for mapping the genes into ten broad cell structure (or location) categories. The cellular categories that were mapped to the cell structures were identified with a combination of text matching and review. First, terms associated with the structure of interest were identified with the text term in the second column of Table 6.1. The cellular components that matched that text string were reviewed for appropriate fit. Cellular component categories that did not seem to fit were removed with the text string(s) specified in the third column of Table 6.1. The cell structures we consider are: nucleus, endoplasmic reticulum, Golgi apparatus, cytoplasm, membrane, mitochondria, ribosome, mitochondrial ribosome, ribonucleoprotein complex, and cytosolic ribosome.

The mapping was not one-to-one. In fact, each cellular component could be mapped to no structure, one structure, or multiple structures. Some categories were nested, with the cytosolic ribosome being a subset of the ribosome. In addition, each gene had zero to 41 cellular components as annotated by GO, with 75% of genes having four or fewer cellular components.

Table 6.1: Structural Annotation Dictionary Criteria

Structure	Matching Term	Removal Term(s)
Nucleus	nucl	mitochond, nucleotide, ribonucleoprotein, male, endonuclease, ribonuclease
Endoplasmic reticulum	reticu	sarcoplasmic, cortical
Golgi bodies	Golgi	vesicle
Cytoplasm	cytop	side of, axon, nuclear pore
Membrane	membrane, cytoplasmic vesicle membrane, azurophil granule membrane	mitochond, Golgi, nucl, endoplasmic, organ, secretory, lysosom, endosome, ruffle, photoreceptor, vacuol, platelet, synaptic, peroxisomal, sperm, vesic, acrosomal, granule, basement, melanosome, coat, neuron, phago, sarcoplasmic, muscle, omegasome, tubular, spine, attack, junctional, transport, ER, chain, bounded, succinate, lamellar
Ribosome	ribosom	mitochond
Mitochondria	mitochond	ribosom
Mitochondria ribosome	mitochond, ribosom	
Ribonucleoprotein complex	nucl, ribonucleoprotein	
Cytosolic ribosome	ribosom, cytosolic	

6.2.2 Cytosolic Ribosomal Genes

The 103 cytosolic ribosome genes from the Fluidigm C1 scRNA-seq datasets are:

COA1, DDX3X, EIF2A, EIF2AK4, EIF2D, EIF4G1, FXR2, GEMIN5, LARP 4, MRPL1, MRPL4, MRPS18A, MRPS18C, MARPS5, NAA10, NHP2, NUFIP1, RPL10, RPL10A, RPL11, RPL12, RPL13, RPL13A, RPL14, RPL15, RPL17, RPL18, RPL18A, RPL19, RPL21, RPL22, RPL22L1, RPL23, RPL23A, RPL24, RPL26, RPL26L1, RPL27, RPL27A, RPL28, RPL29, RPL3, RPL30, RPL31, RPL32, RPL34, RPL35, RPL35A, RPL36, RPL36A, RPL36AL, RPL37, RPL37A, RPL38, RPL39, RPL39L, RPL4, RPL41, RPL6, RPL7, RPL7A, RPL7L1, RPL8, RPL9, RPLP0, RPLP1, RPLP2, RPS10, RPS11, RPS12, RPS13, RPS14, RPS15, RPS15A, RPS16, RPS18, RPS19, RPS2, RPS20, RPS21, RPS23, RPS24, RPS25, RPS26, RPS27, RPS27A, RPS27L, RPS28, RPS29, RPS3, RPS3A, RPS4X, RPS5, RPS6, RPS7, RPS8, RPS9, RPSA, RSL1D1, RSL24D1, SURF6, UBA52, ZNF622

The 89 cytosolic ribosome genes from the GTEx dataset are:

APOD, DDX3X, EIF2A, EIF2AK4, FXR2, GEMIN5, HBA1, HBA2, MCTS1, MRPL4, MRPS18A, MRPS18C, MRPS5, NAA11, NHP2, NUFIP1, PPARGC1A, RPL10A, RPL10L, RPL11, RPL12, RPL13, RPL13AP3, RPL14, RPL17, RPL18, RPL18A, RPL19, RPL21, RPL22, RPL23, RPL23A, RPL24, RPL26, RPL26L1, RPL27, RPL27A, RPL28, RPL29, RPL30, RPL31, RPL32, RPL36, RPL36A, RPL36AL, RPL37, RPL37A, RPL38, RPL39, RPL39L, RPL39P5, RPL3L, RPL41, RPL7, RPLP1, RPLP2, RPS10, RPS10P5, RPS11, RPS12, RPS13, RPS14, RPS15, RPS16, RPS17, RPS18, RPS19, RPS20, RPS23, RPS25, RPS27, RPS27A, RPS27L, RPS28, RPS29, RPS3A, RPS4X, RPS5, RPS6, RPS7, RPS8, RPS9, RPSA, RSL1D1, RSL24D1, SNU13, SURF6, UBA52, ZNF622

6.3 Stability Measures

Here, we provide mathematical formulations for the stability measures described in Sections 2.4.4 and 2.4.5.

Let i represent a gene, j a dataset, k a cell, $\mathcal{G}$ some set of genes, and $c_{i,j,k}$ the expression recorded for gene i in cell k in dataset j. Let $s_{\mathcal{G},j}$ be a vector representing the sum of the set of genes $\mathcal{G}$; in other words, $s_{\mathcal{G},j,k} = \sum_{i \in \mathcal{G}} c_{i,j,k}$ is the kth entry in the vector $s_{\mathcal{G},j}$. Conversely, let $s_{-\mathcal{G},j}$ be a vector representing the sum of the genes not contained in the set of genes $\mathcal{G}$, or $s_{-\mathcal{G},j,k} = \sum_{i \notin \mathcal{G}} c_{i,j,k}$ is the kth entry of the vector $s_{-\mathcal{G},j}$.

The correlations that serve as a measure of absolute stability for a given gene i and a given dataset j are calculated by $r_{\text{ERCC},i,j} = \text{cor}(\log(c_{i,j} + 1), \log(s_{\mathcal{E},j}))$ where $\mathcal{E}$ is the set of ERCC spike-ins. The overall measure of absolute stability for gene i is $r_{\text{ERCC},i} = \frac{1}{6} \sum_{j=1}^{6} r_{\text{ERCC},i,j}$.

The correlations that serve as a measure of proportional stability for a given gene i, a given dataset j, and a given cell structure with gene set $\mathcal{S}$ are defined as $r_{\mathcal{S},i,j} = \text{cor}(\log(c_{i,j}+1), \log(s_{-\mathcal{S},j}))$. The overall measure of stability for gene i is $r_{\mathcal{S},i} = \frac{1}{6} \sum_{j=1}^{6} r_{\mathcal{S},i,j}$.

We calculated an additional measure of proportional stability that was comparable across all genes. This uses an unadjusted cell total, unlike the measure of proportional stability defined above that uses a cell total adjusting for a given structure. s_j is a vector where $s_{j,k} = \sum_{i=1}^{I} c_{i,j,k}$; that is, we sum over all of the genes for a given cell. The measure for a given dataset is $r_{i,j} = \text{cor}(\log(c_{i,j} + 1), \log(s_j))$ and is summarized over the 6 datasets by $r_i = \frac{1}{6} \sum_{j=1}^{6} r_{i,j}$.

6.4 Absolute Stability from Individual ERCC spike-ins

We examine the ERCC measurements separately. Within an experiment, ERCC spike-ins are introduced in the same manner and are subjected to the same technical protocols. However, they are intentionally designed to have different characteristics, providing the ability to capture different technical effects. In our analysis, we summed the ERCC spike-in measurements to capture an overall technical effect.

We examine the correlations between individual ERCC spike-ins to assess the similarity of the technical effects captured by the spike-ins. Figure 6.7 displays the strong correlations among the spike-ins in GSE77288 compared to the remaining datasets. The pattern of expressions of the ERCC spike-ins on the log scale are summarized in Figure 6.8. Higher expressions for GSE84686 are evident, as are some differences in the expression profiles across the six datasets.

Recognizing that the individual ERCC spike-ins may be capturing a variety of technical effects, and that highly expressed ERCCs may dominate the ERCC total used in our analysis, we replace the sum of the spike-in measurements with the median, a robust measure (Figures 6.9 and 6.10). The correlations are calculated in the same manner as described in Chapter 2 with the median of the ERCC measurements replacing the ERCC total, resulting in a correlation between the log of the gene expression + 1 and the log of the median of the ERCC measurements + 1. The calculated correlations are all less than 0.25 in absolute value. This suggests that regardless of the quantity that summarizes the ERCC measurements, the correlations with gene expression will remain small. It appears that there are limits to how well the ERCC measurements capture the technical effects with respect to biological gene expressions, reducing their effectiveness in addressing questions of absolute stability.

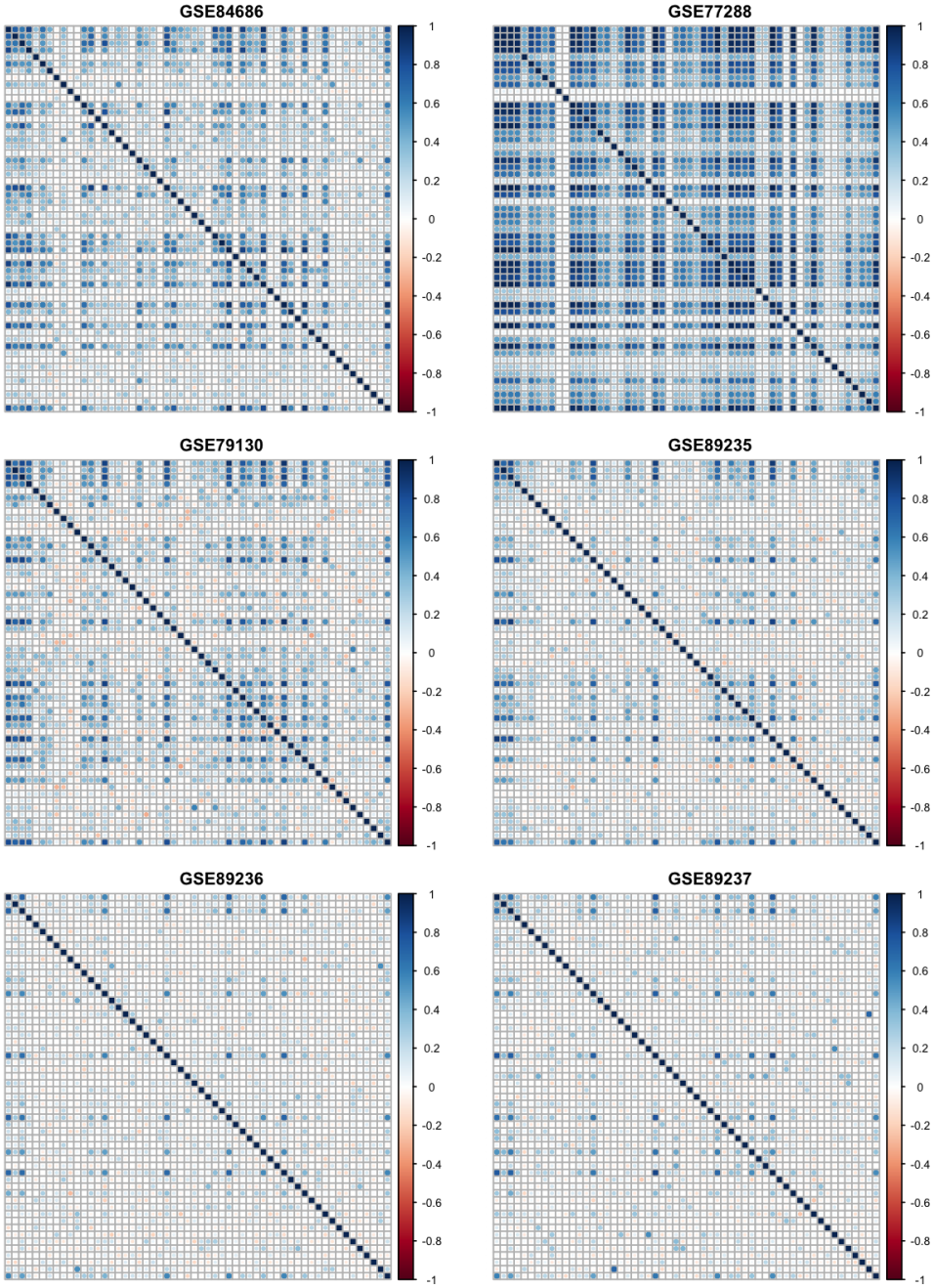

Figure 6.7: The correlations of ERCC spike-in measurements within a dataset. The same set of 56 ERCCs that were measured in each of the six Fluidigm C1 datasets are included in this figure.

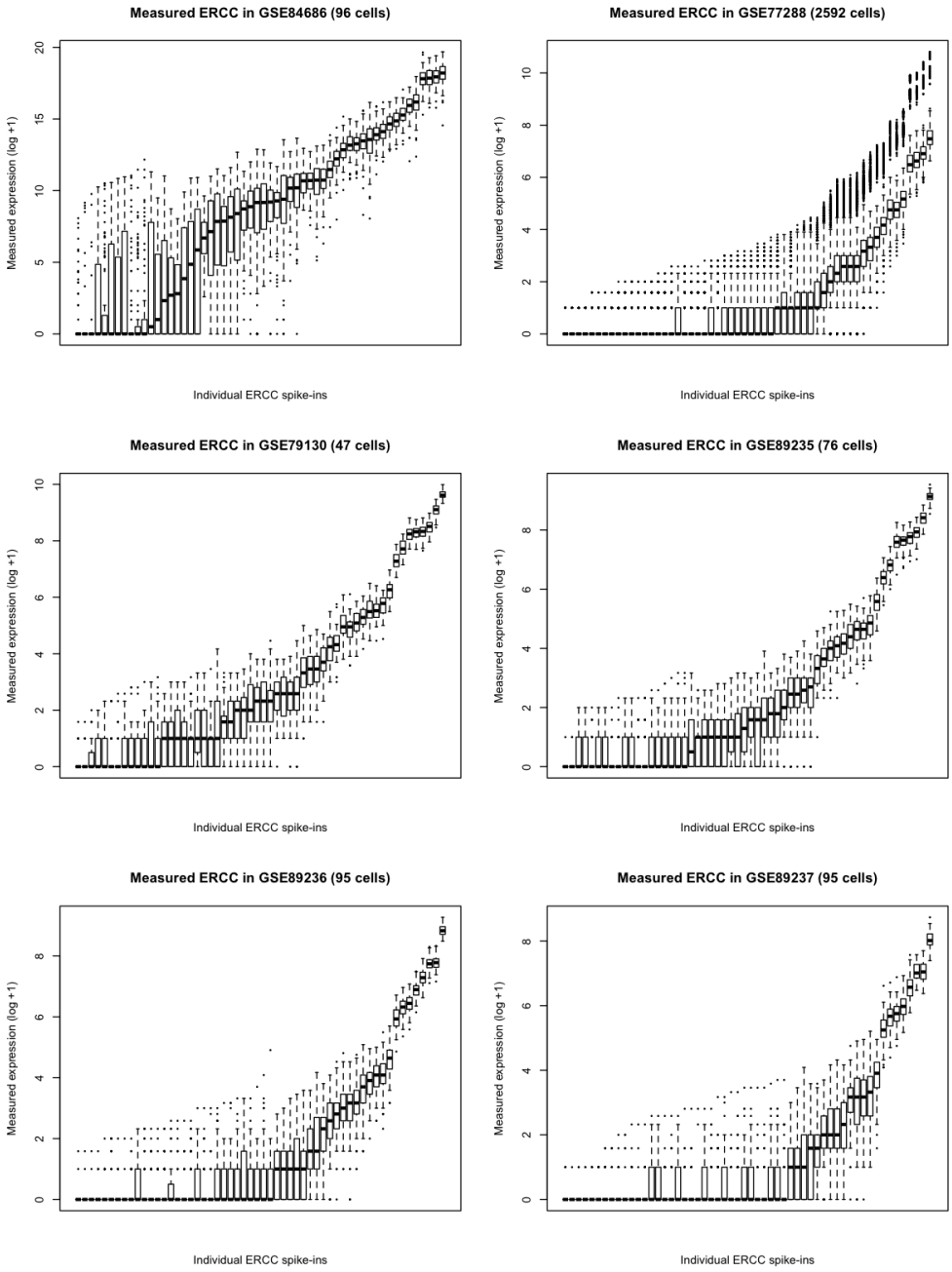

Figure 6.8: The expression levels of ERCC spike-in measurements within each dataset. The same set of 56 ERCCs that were measured in each of the six Fluidigm C1 datasets are included in each of these plots.

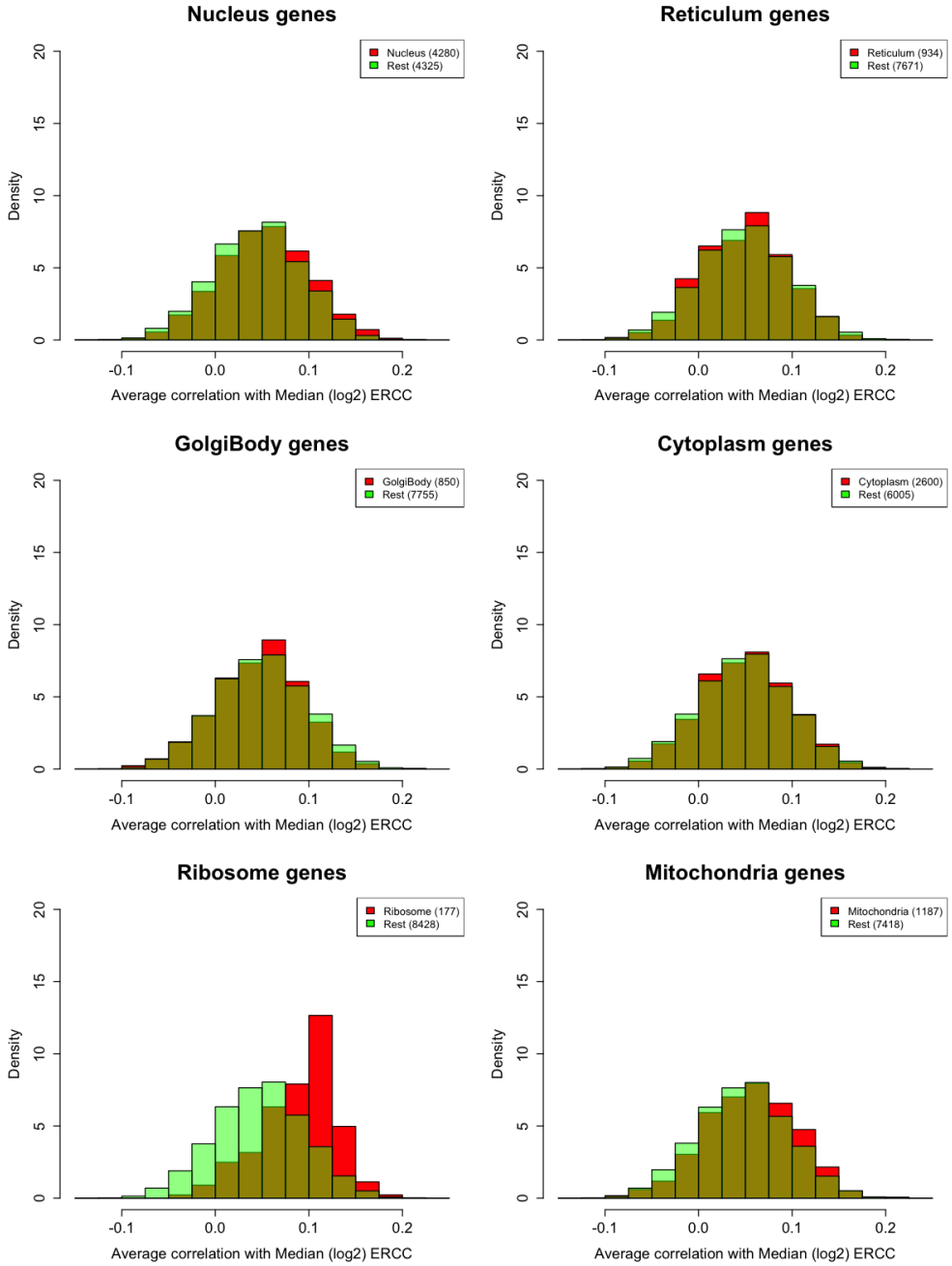

Figure 6.9: Histograms of the average correlations of each gene over the six Fluidigm C1 datasets considered, with comparisons between the set of genes of interest and the remaining genes. This figure is similar to Figure 6.1, but with the median of the log-transformed ERCC spike-in measurements replacing the sum.

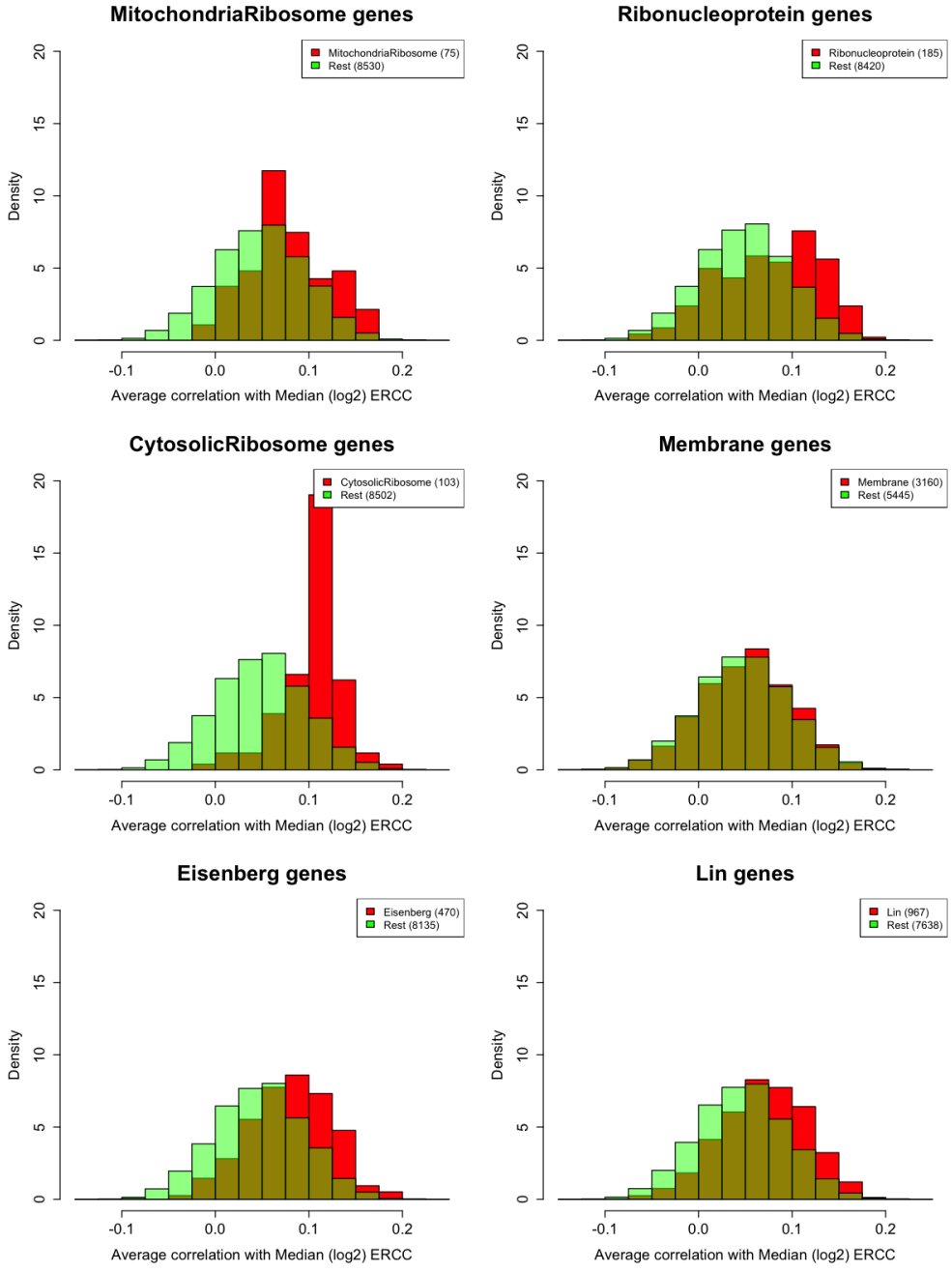

Figure 6.10: Continuation of Figure 6.9.

6.5 GTEx Features

6.5.1 Difference in Expression Profiles of Genes

In Chapter 2 we use SVD plots of the GTEx data to investigate the stability of different gene sets. We interpret lack of clustering by cell type to indicate that gene expression is stable across cell types. However, another possible explanation for lack of clustering by cell type would be the existence of variation due to some other, stronger factor that dominates the SVD plot. To ensure that the lack of separation is indeed due to stability of expression, we examine the IQR of gene expression as a way to measure stability directly. In particular, we calculate, for each gene and for each tissue type, the IQR of log-RPKM-adjusted expression. We can then compare the distributions of these IQRs for the cytosolic ribosomal genes, the genes from Eisenberg and Levanon [2003], the genes from Lin *et al.* [2019a], and other highly expressed genes. The IQRs provide a sense of how stably expressed the given set of genes are within the tissue type, with smaller values indicating more stable expression (Figures 6.11 to 6.16).

In general, it appears that the cytosolic ribosomal genes have small IQRs within tissues, indicating that their differences between subtissue types are small. This does not discount the potential for cytosolic ribosomal genes to have differences between tissue types; any tissue-specific dependencies suggest caution to their later implementation in analyses that contain multiple tissue types.

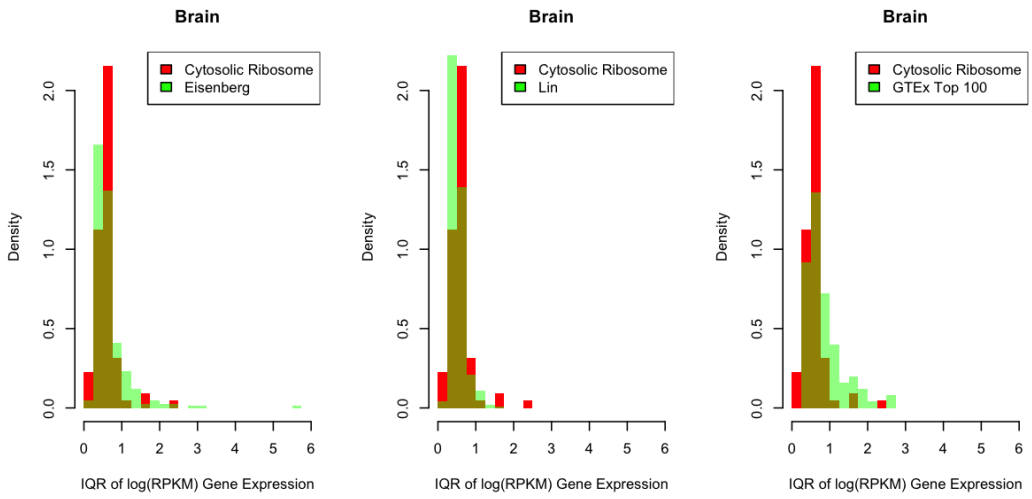

Figure 6.11: The distribution in IQR of genes within the brain in the GTEx data.

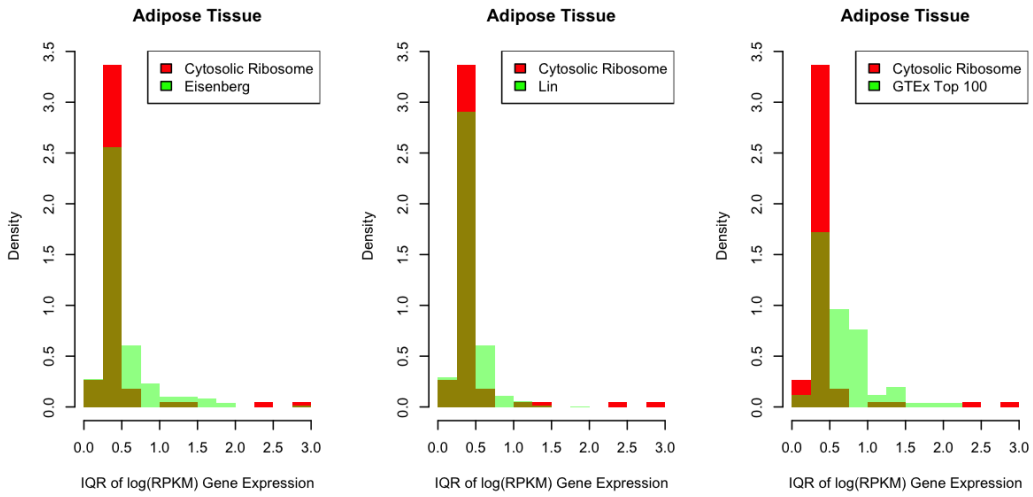

Figure 6.12: The IQRs for the expression of genes within adipose tissue samples.

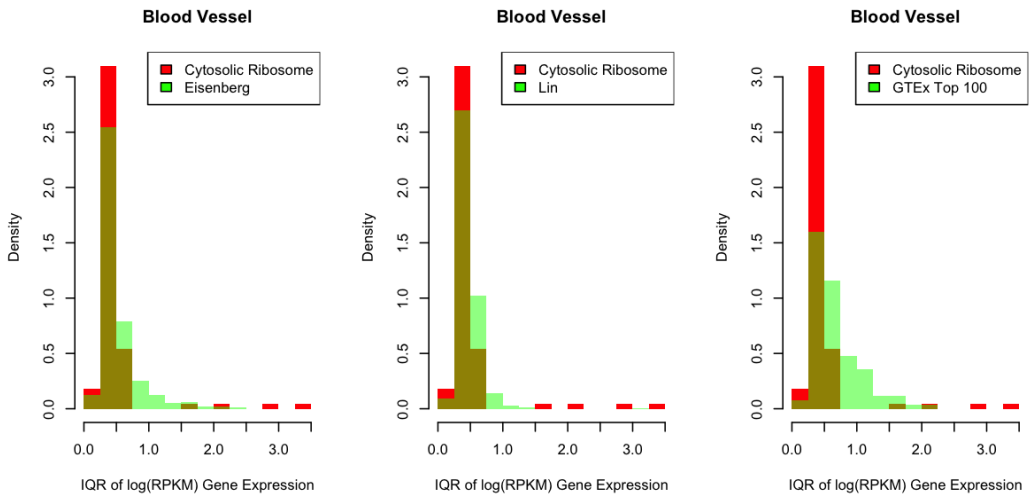

Figure 6.13: The IQRs for the expression of genes within blood vessel samples.

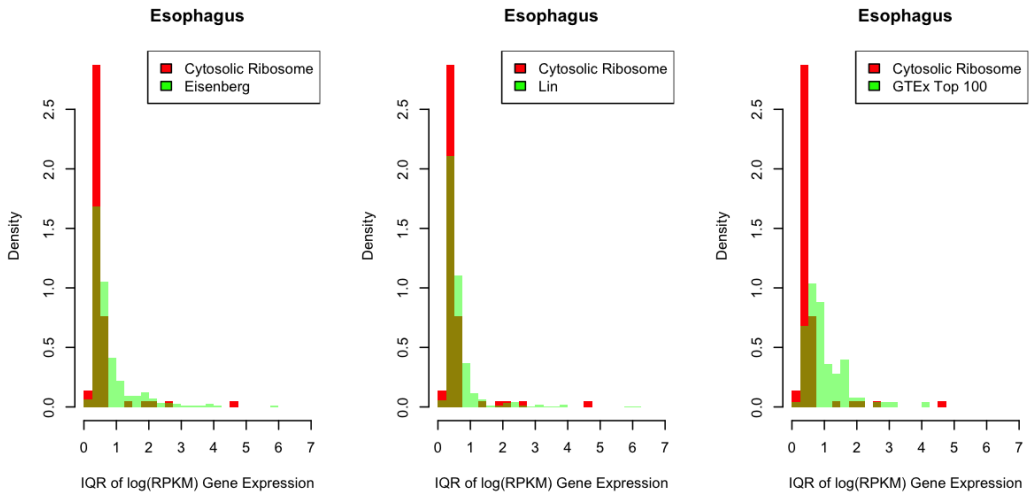

Figure 6.14: The IQRs for the expression of genes within esophagus samples.

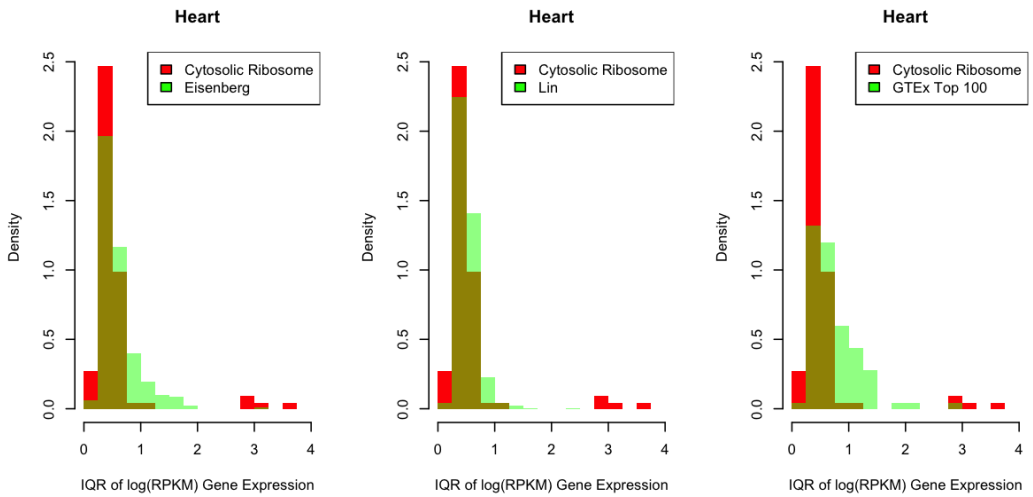

Figure 6.15: The IQRs for the expression of genes within heart samples.

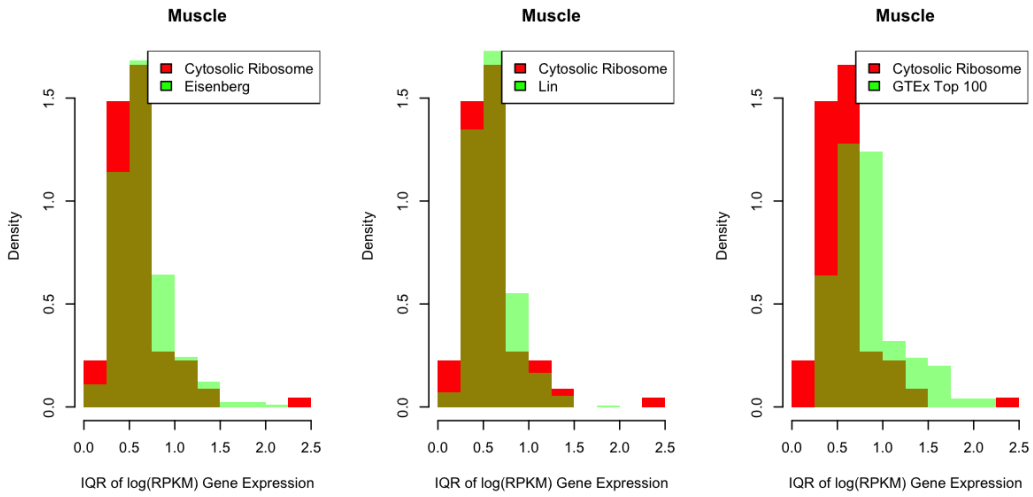

Figure 6.16: The IQRs for the expression of genes within muscle samples.

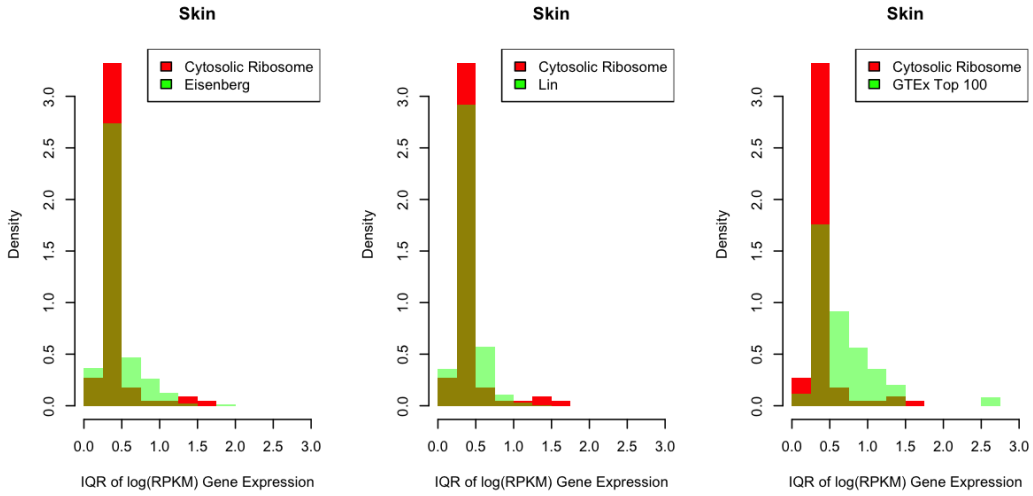

Figure 6.17: The IQRs for the expression of genes within skin samples.

6.6 Supplementary Table Description

We created a table that contains information for human genes, including structural annotations and numerical summaries from our single-cell and GTEx analyses. The first ten columns contain indicators for whether the gene was classified as a member of each of our structures; these names are denoted as the full name of the structure followed by "_indicator". The following two columns contain indicators of membership of the gene sets published by Eisenberg and Levanon [2003] and by Lin *et al.* [2019a] and denoted with the author names.

Summary statistics, including the mean, standard deviation, and correlation, for each single cell dataset are calculated for the genes expressed in all six scRNA-seq datasets. The mean and standard deviation of the log + 1 transformed expressions for each of the genes is calculated for each of the datasets; these variables are denoted with the five digit GEO accession number (GSE#####) followed by either a "mean" or "sd" suffix to denote the summary measure. The correlation with the unadjusted cell total is also included for each dataset, as well as the mean unadjusted cell total across the six datasets. Our analysis in Chapter 2 primarily looks at correlations with adjusted cell totals; we provide the mean correlation across the six datasets for each gene with each of the adjusted cell totals. These variables have the name "sc_cormean_less" followed by the name of the structure removed during the cell total calculation.

Prior to analysis, the GTEx data is normalized according to an adjusted RPKM transformation. The total number of reads for a given sample is calculated and divided by 1,000,000. Each individual entry is then transformed with log ([entry + 1]/ total samples per million). Then, the mean and standard deviation of each gene are calculated for each of the tissue types contained within the data. F-statistics and their respective degrees of freedom are calculated for each of the tissue types that contain subtissue types; these F-statistics are a measure of how differentially expressed the gene is between the subtissue types. Note that genes that are stably expressed will have low F-statistics. These variables are denoted with GTEx_tissue type_statistic, where the "tissue type" and "statistic" are replaced with the appropriate designation; the options for the statistic are mean, standard deviation, F statistic, or degrees of freedom. Note that we have removed the "Transformed fibroblasts" from the Skin tissue type, as these were prepared differently than the other samples.

Chapter 7

Supplement: Modeling Biases of Reads per UMI

7.1 Univariate Associations

7.1.1 Transcript Proportion of Guanine and Cytosine

We consider the proportion of GC content for various parts of the transcript in Section 3.4.2.2. Figure 3.7 displays the estimated rUMI parameter based on the mean proportion GC in the transcript, and Figure 3.9 displays the estimated rUMI parameter based on the mean proportion GC in the first half of the transcript.

We also considered the relationship between rUMI and the mean proportion GC content in the second half of the transcript (Figure 7.1). We see that the variation of rUMI observed amongst the full length transcripts is less than all transcripts. However, the shape of the relationship for the full length transcripts appears quadratic, while the shape for all transcripts is indistinct.

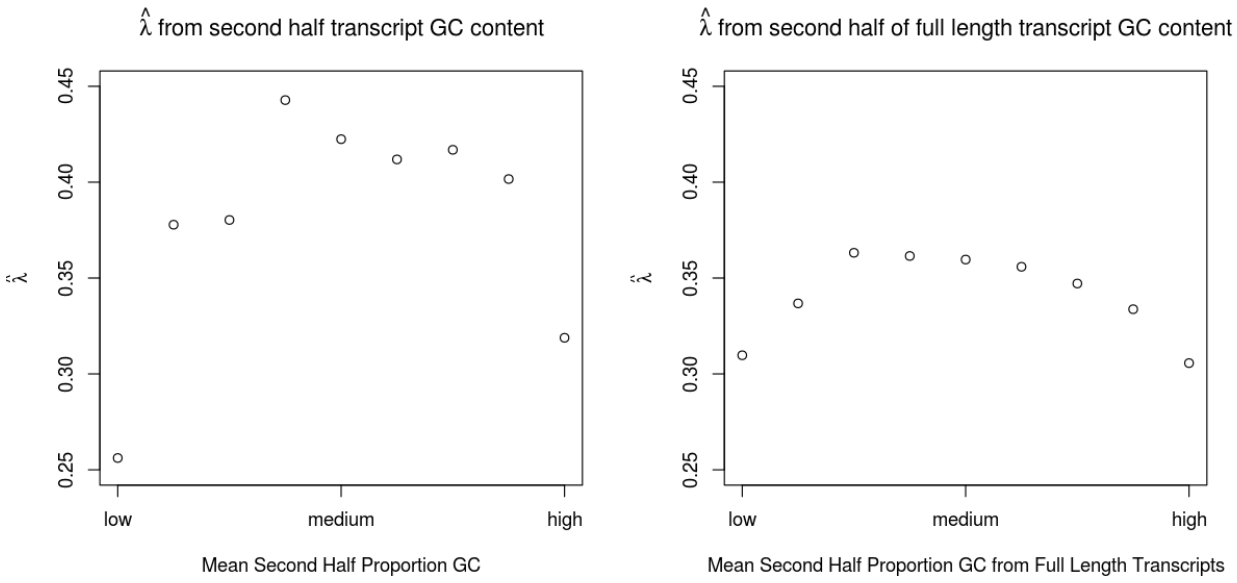

Figure 7.1: The estimated mean rUMI based on the GC content in the second half of the transcript is shown for all transcripts (left panel) and for full length transcripts (right panel) in the top 10 cells. An AT/GC imbalance corresponds to a smaller average rUMI. The shape from the full length transcripts appears quadratic, while the shape from all transcripts is indistinct. Note that together with the proportion GC in Figure 3.9, the proportion GC in the full transcript can be calculated.

7.1.2 Two-Somes

Two-somes are calculated to measure the detailed composition of a transcript. We define two-somes as every possible permutation of two bases that could appear within a read. Note that with four bases, there are sixteen possible two-somes. We calculate the two-somes by applying a sliding window of length two to the read, recording the number of times each of the sixteen possible two-somes occur; the number of two-somes for a read are one less than the length of the read. Thus, we can calculate the proportion of a given two-some for a read with the length and the number of two-somes.

Figures 7.2 and 7.3 display the relationship between the proportion of each of the twosomes with the estimated mean rUMI. Note that these figures all have the same scale on the y-axis, so the distributions are comparable in terms of estimated mean rUMI. We see that the strongest relationships with rUMI occur with AT, CT, TA, and TT two-somes. Note that, with the exception of CT, all of these two-somes consist of adenine and thymine bases and are from the same side of the AT/GC divide. The remaining two-somes exhibited weaker relationships with rUMI.

We also consider the two-somes for the full length transcripts. Figures 7.4 and 7.5 are similar to Figures 7.2 and 7.3 but calculated on only the full length transcripts. The full length transcripts in general have smaller mean rUMIs, as described in Chapter 3 and evidenced by the reduced scale of the y-axis. In addition, the patterns observed for full length transcripts generally appear weaker than those for all transcripts. The AA and GC two-somes seem to have the strongest relationship with rUMI. Again, the AA and GC two-somes occur on the same side of the AT/GC divide.

Note that the two-somes with the strongest relationships with rUMI do not seem to be preserved across Figures 7.2 and 7.3 to Figures 7.4 and 7.5. Thus, we are hesitant to conclude that two-some results are generalizable.

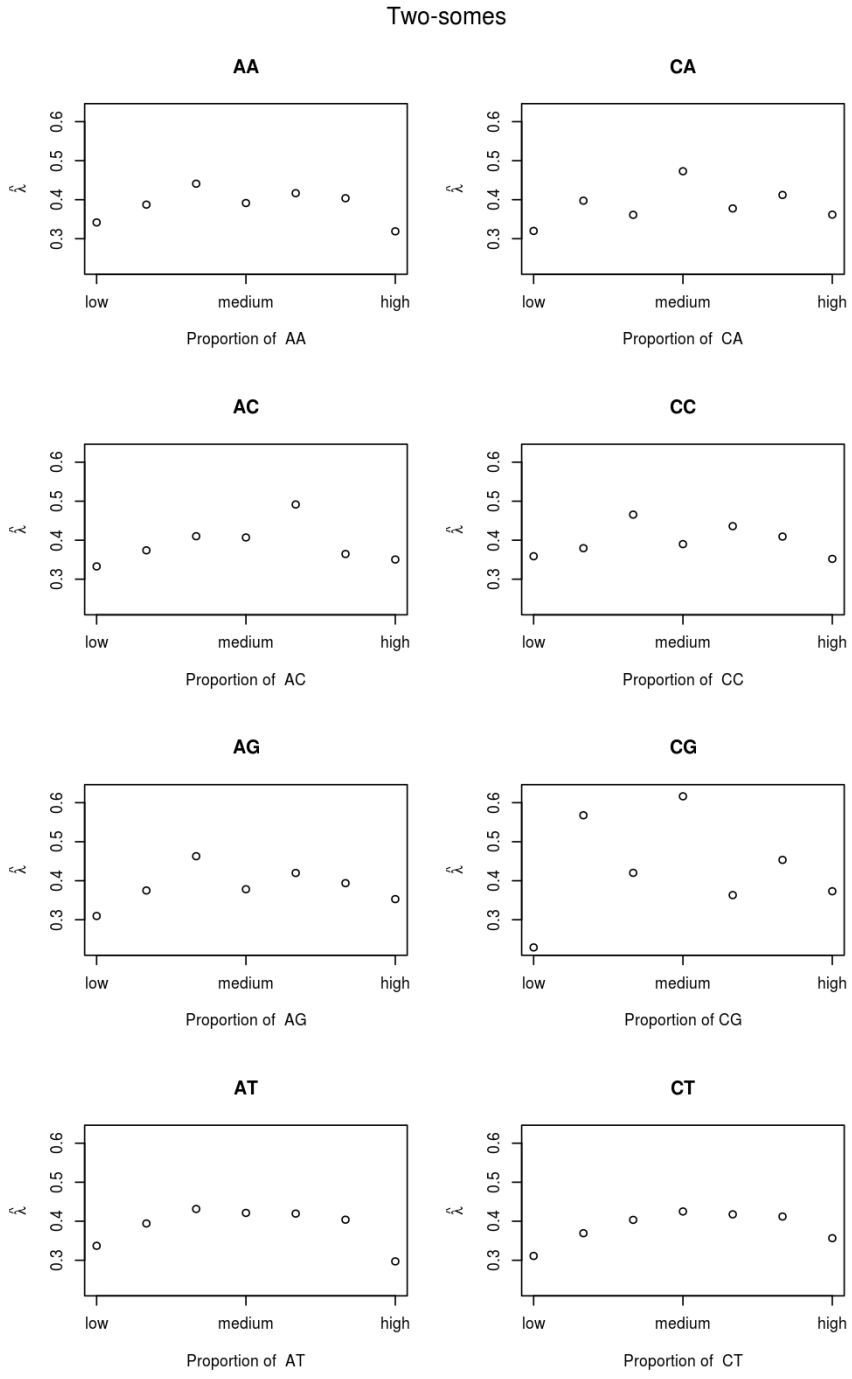

Figure 7.2: The relationship between two-somes calculated from all transcripts and $\hat{\lambda}$ for rUMI. Note that the scale of the y-axis is preserved across all plots. The clearest relationships seem to be visible for the AT, CT, TA, and TT two-somes.

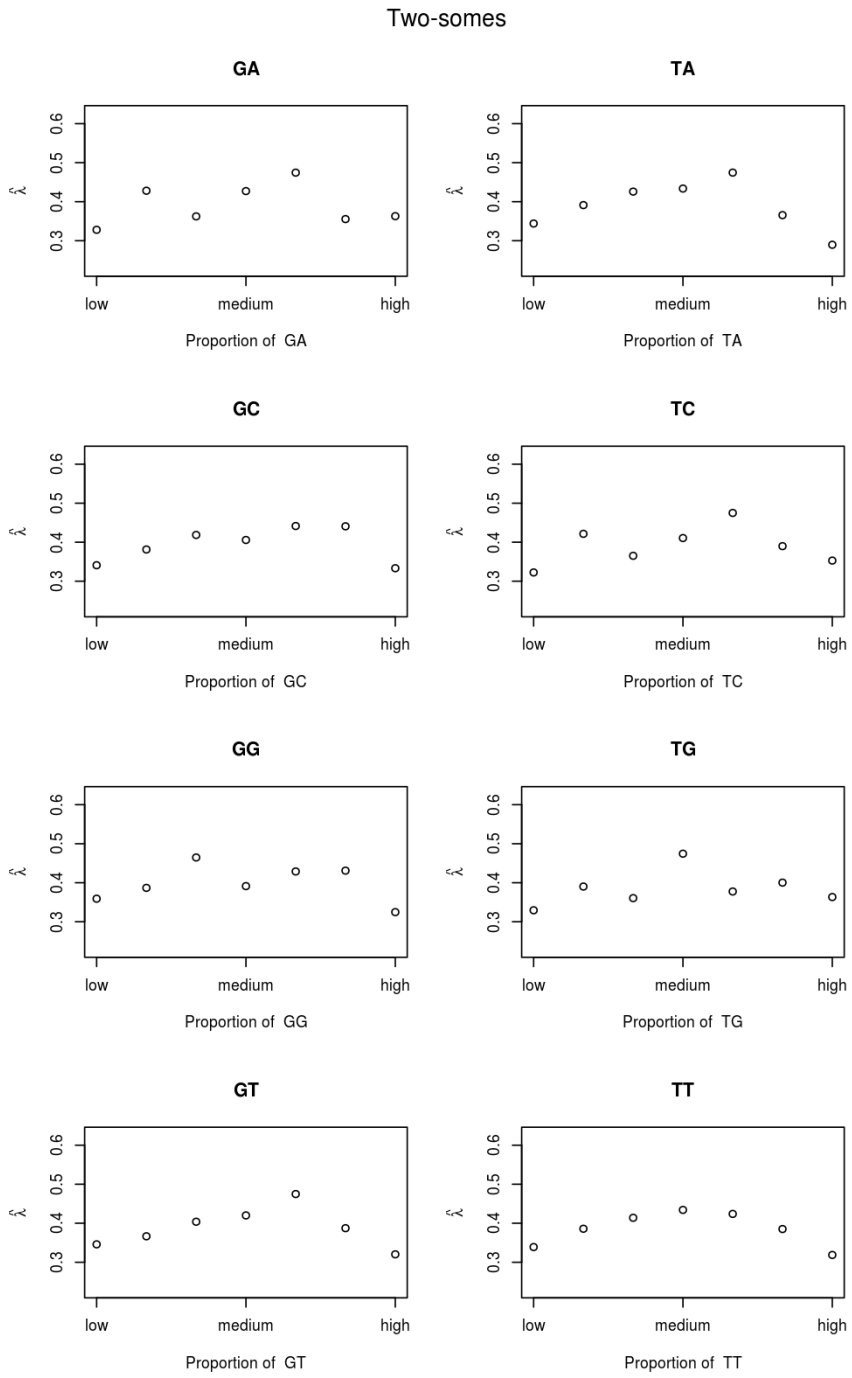

Figure 7.3: Continuation of Figure 7.2.

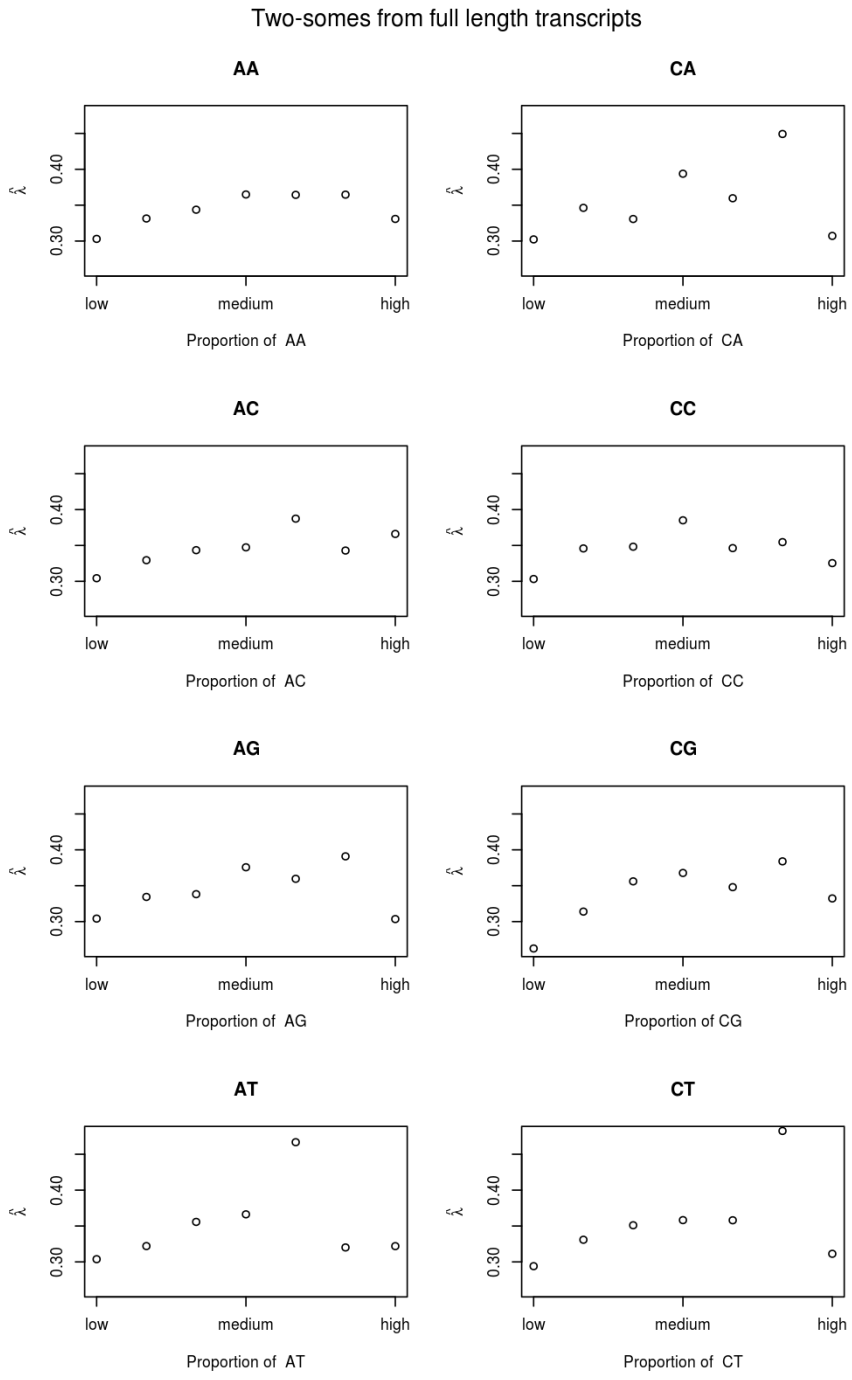

Figure 7.4: The relationship between two-somes calculated from full length transcripts and $\hat{\lambda}$ for rUMI. Compared to Figure 7.2, we have removed those UMIs that do not have full length transcripts. Note that the scale of the y-axis is smaller than in Figure 7.2 but is preserved across all plots. The clearest relationships seem to be visible for the AA and GC two-somes.

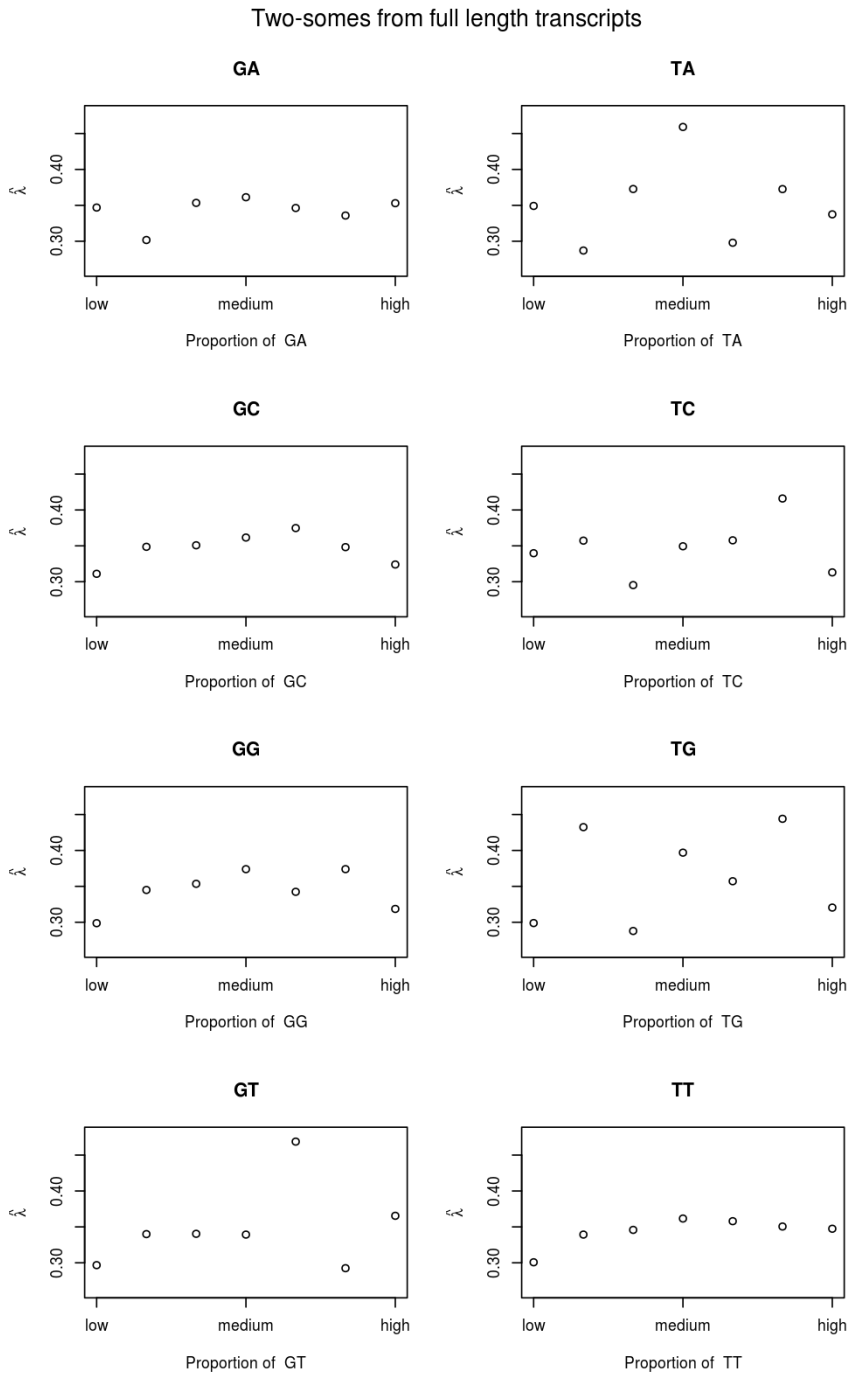

Figure 7.5: Continuation of Figure 7.4.

7.1.3 Cells

We calculated the $\hat{\lambda}$ for rUMI for each of the ten cells examined (Figure 7.6). There are variations in the estimated mean rUMI for each cell, but the amount of variation is fairly small. The general pattern of mean rUMI is preserved across all transcripts and full length transcripts. The first and third largest cells do seem to have the smallest mean rUMI values.

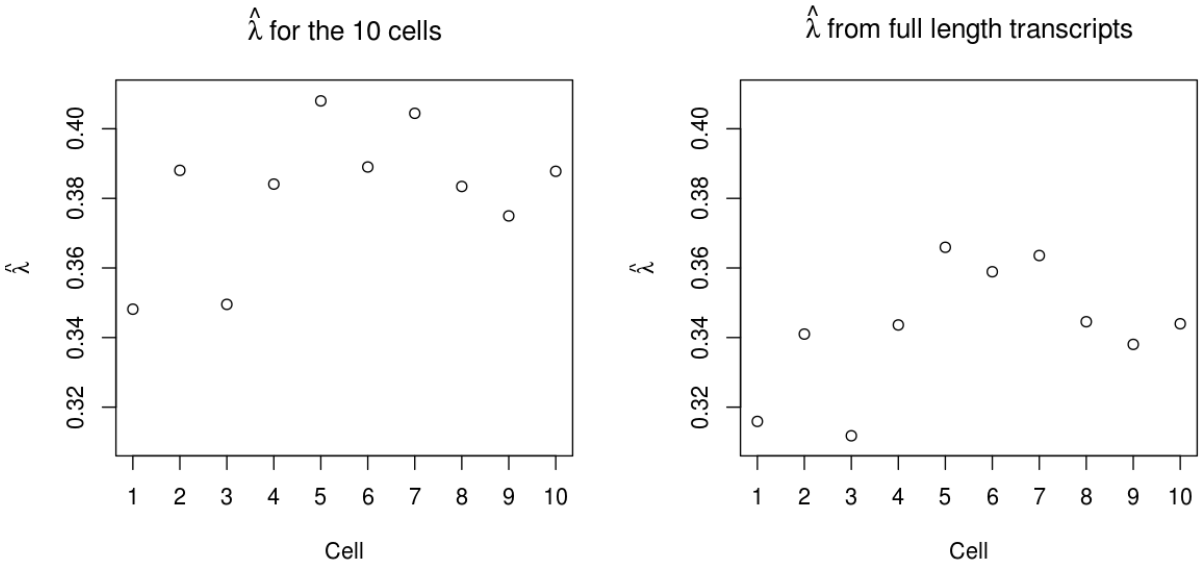

Figure 7.6: The relationship between each of the cells and the estimated mean rUMI for all transcripts (left panel) and for full length transcripts (right panel).

7.1.4 Cellular Barcode

For this experiment, the UMI consists of 20 bp. Each cell has a 12 bp barcode to identify transcripts originating from that cell. The cellular barcode allows researchers the ability to record the gene expression at the cellular level. As the UMI is located near the beginning of the genetic material during amplification and sequencing, we hypothesize that the cellular barcode may affect the rUMI. Therefore, we consider various features of the cellular barcode and their relationships with the estimated mean rUMI. Specifically, we examine the proportion GC in the cellular barcode, the first base, and the last base of the barcode.

Figure 7.7 displays the relationship between the proportion GC content of the cellular barcode with the estimated mean rUMI. We analyze ten cells, i.e. ten cellular barcodes, which reduces the values that the proportion GC content in the cellular barcode takes. The ten cells in Figure 7.6 are summarized in three categories in Figure 7.7. No meaningful differences in mean rUMI are calculated in Figure 7.7.

Figure 7.8 shows the relationship between the first and last base of the cellular barcode with the estimated mean rUMI. Similar to Figure 7.7, the ten cells from Figure 7.6 are summarized in four categories in Figure 7.8. Again, no meaningful differences in mean rUMI are calculated based on the first or last base of the cellular barcode.

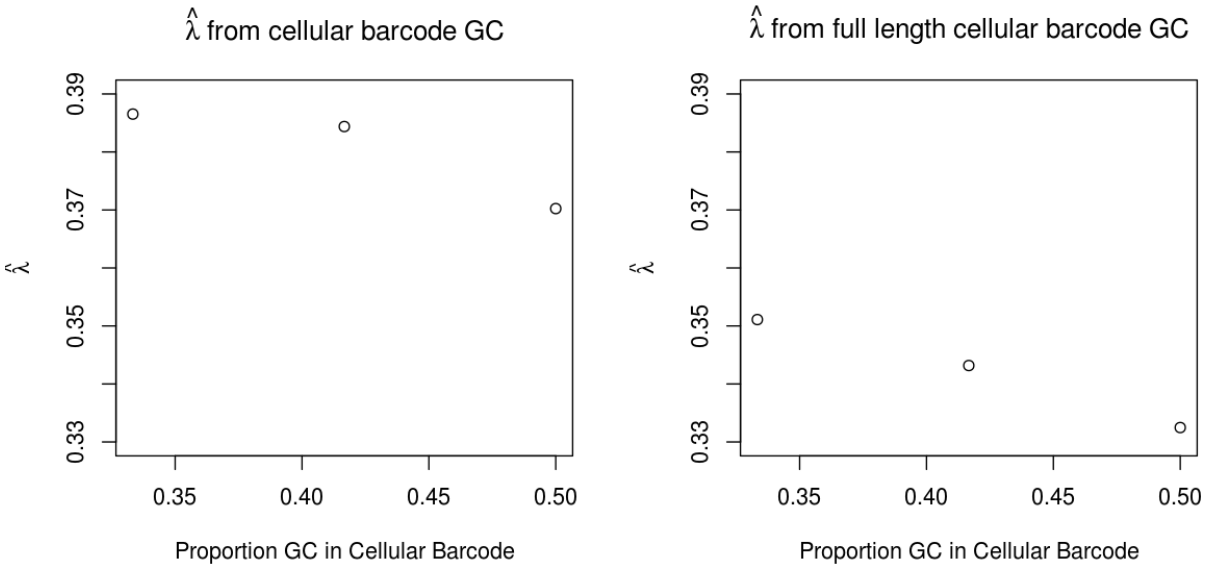

Figure 7.7: The relationship between the proportion GC in the cellular barcode and the estimated mean rUMI for all transcripts (left panel) and full length transcripts (right panel). With ten cellular barcodes, we calculate three values for the proportion GC. The ten cells displayed in Figure 7.6 are reduced to three categories. The estimated mean rUMI is similar for each of the groups considered here.

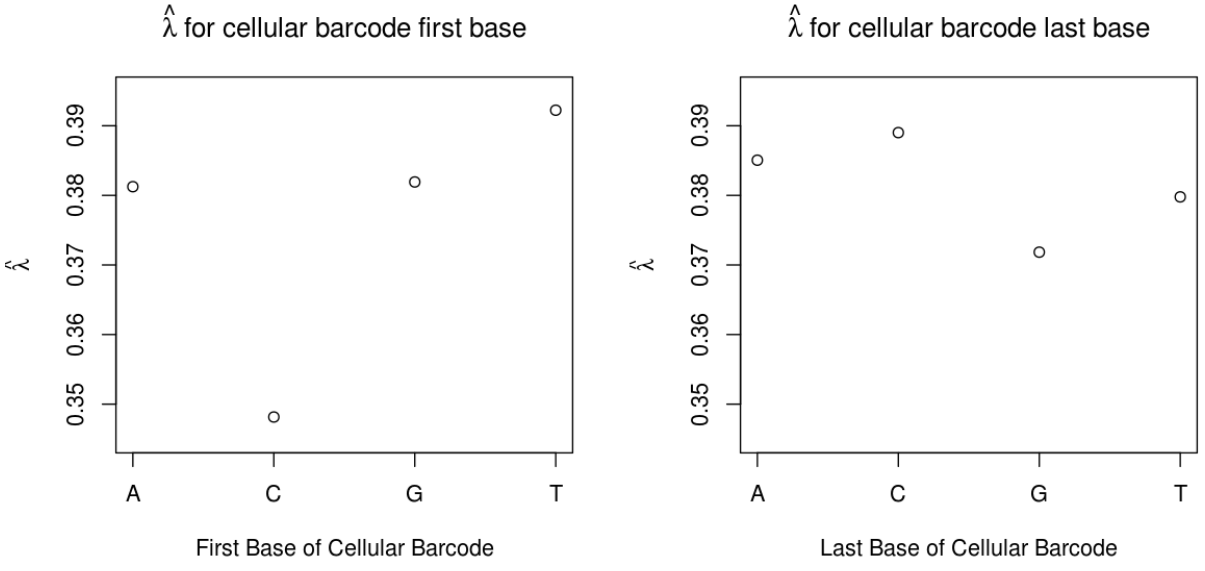

Figure 7.8: The relationship between the estimated mean rUMI and the first base in the cellular barcode (left panel) and the last base in the cellular barcode (right panel). The ten cells displayed in Figure 7.6 are reduced to four categories for this figure. The estimated mean rUMI is similar for each of the groups considered here.

7.1.5 Molecular Barcode

The remaining 8 bp of the UMI compose a molecular barcode. The molecular barcode allows researchers to follow a transcript through sequencing, identifying reads originating from the same transcript and motivating deduplication of reads. In addition, we could not measure rUMI without the molecular barcode.

Similar to the cellular barcode, we hypothesize that features of the molecular barcode may be informative in estimating rUMI. Therefore, we examine the proportion GC (Figure 7.9) and the first and last base (Figure 7.10) in the molecular barcode. Figure 7.9 indicates that molecular barcodes consisting only of A and T result in lower mean rUMI; however, only 366 UMIs contain only A and T compared to 2,986 UMIs containing only G and C (the next smallest group). Thus, the uneven (and small) sample size for proportion GC in the molecular barcode may be driving the small estimated mean rUMI for the AT-only molecular barcodes. No other meaningful differences in estimated mean rUMI are observed in Figures 7.9 and 7.10.

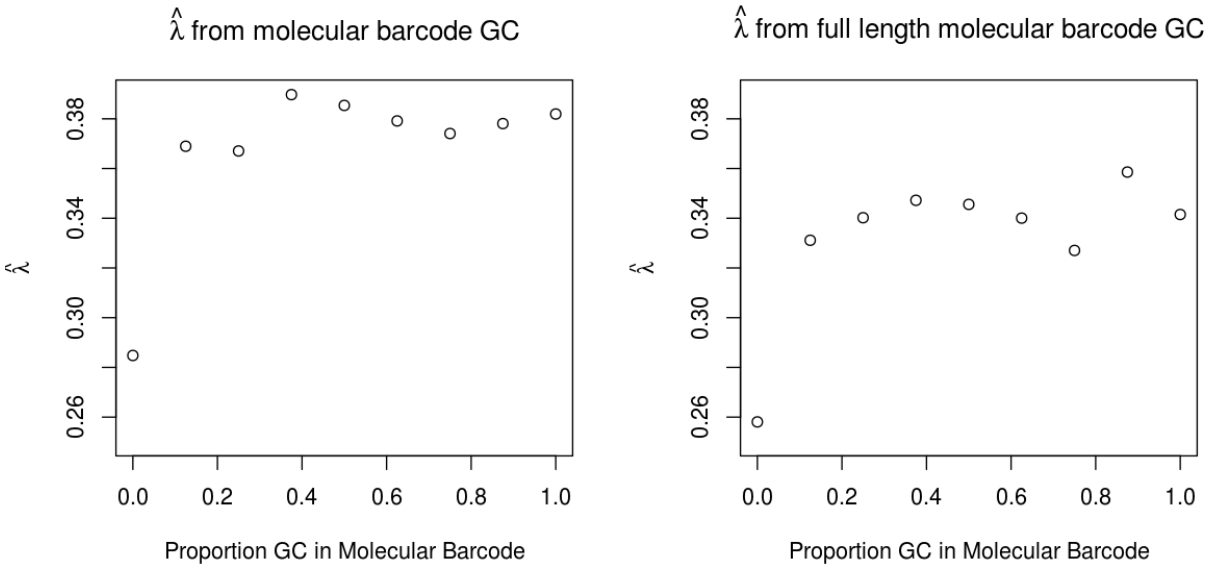

Figure 7.9: The relationship between the proportion GC in the molecular barcode and the estimated mean rUMI for all transcripts (left panel) and full length transcripts (right panel). We are limited to nine values for the proportion of GC in the eight bases of the molecular barcode. The estimated mean rUMI is similar for each of the groups considered here except for those molecular barcodes that consist of only adenine and thymine.

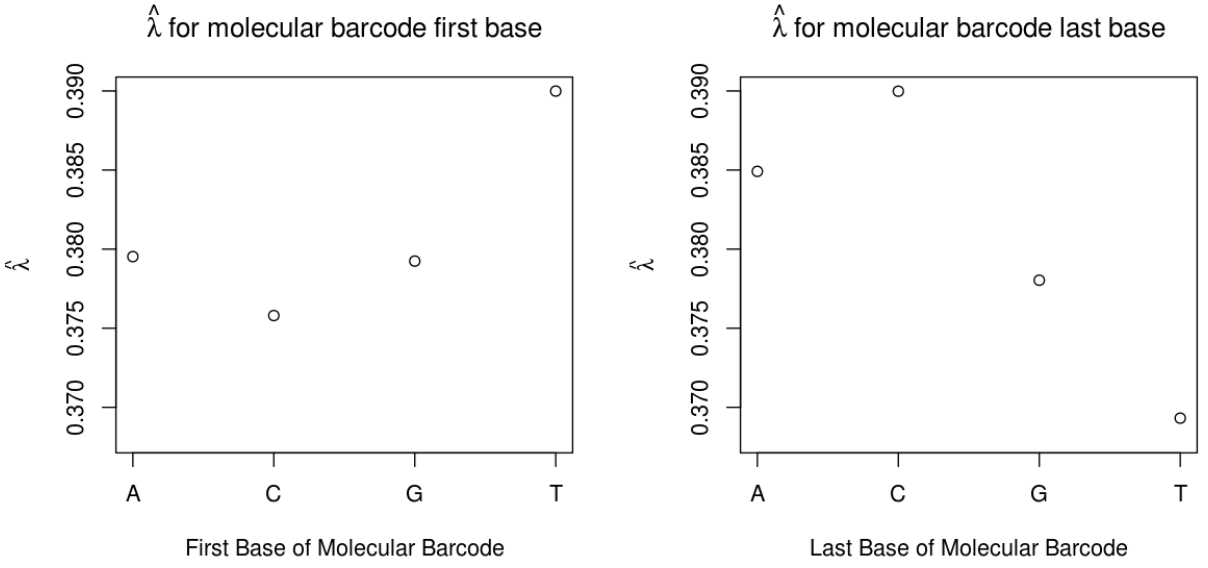

Figure 7.10: The relationship between the estimated mean rUMI and the first base in the molecular barcode (left panel) and the last base in the molecular barcode (right panel). The four bases provide the four categories for this figure. The estimated mean rUMI is similar for each of the groups considered here.

7.2 Results for Twenty Additional Cells

7.2.1 Data Characteristics

In addition to the 10 cells with the most UMIs, we considered 20 additional cells. We chose 10 cells that had close to the median level of UMIs (1,369-1,383 UMIs/cell), and we chose 10 additional cells that had gene expressions near the 80th percentile (6,951-7,174 UMIs/cell); we call these groups of cells median and medium cells, respectively. We repeated a similar analysis on these cells that we did on the 10 highly expressed cells. We present analogous figures here, as evidence that the patterns we observe with the 10 highly expressed cells are typically preserved between cells of different gene expression levels.

From the 10 medium cells, there were 91,943 reads stemming from 74,853 UMIs. The median cells had 15,344 UMIs deduplicated from 18,691 reads. Figure 7.11 shows the distribution of rUMIs for each of the sets of 10 cells with the medium cells on the left and the median cells on the right. Note that, again, over 80% of the UMIs have a single read. The medium cells have a maximum of 7 rUMI, while for the median cells 6 rUMI is the maximum. This is reduced from 9 rUMI in the highly expressed cells, although fewer UMIs contribute to these histograms.

The medium cells have a mean rUMI of 1.24 (sd=0.53, $\hat{\lambda} = 0.37$ based on all of the UMIs). The mean of the median cells is 1.22 rUMI (sd=0.51; $\hat{\lambda} = 0.36$ based on all of the UMIs).

We again consider how the gene aids in deduplication of the UMIs, as discussed in Section 3.4.1 and specifically in Figure 3.5. Here, we see that the gene is not as helpful in distinguishing distinct UMIs. As there are fewer UMIs stemming from a given cell compared to the top 10 cells, it is reasonable to have fewer overlapping UMIs. The maximum rUMI does increase to 10 and 7 rUMI from 7 and 6 rUMI for the medium and median cells, respectively.

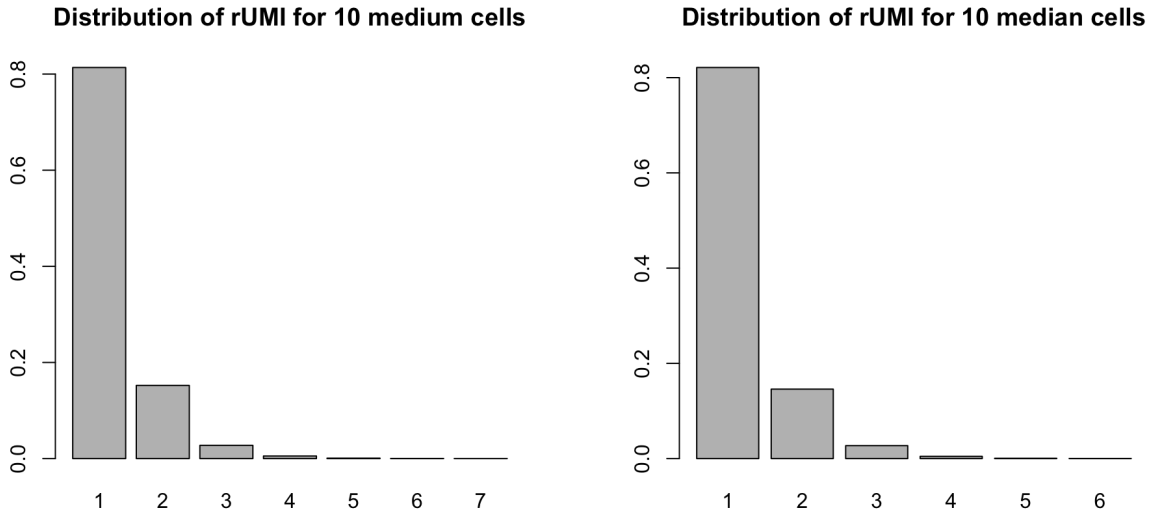

Figure 7.11: Histograms of rUMI for the 10 medium (left panel) and 10 median (right panel) cells. Compare to Figure 3.4.

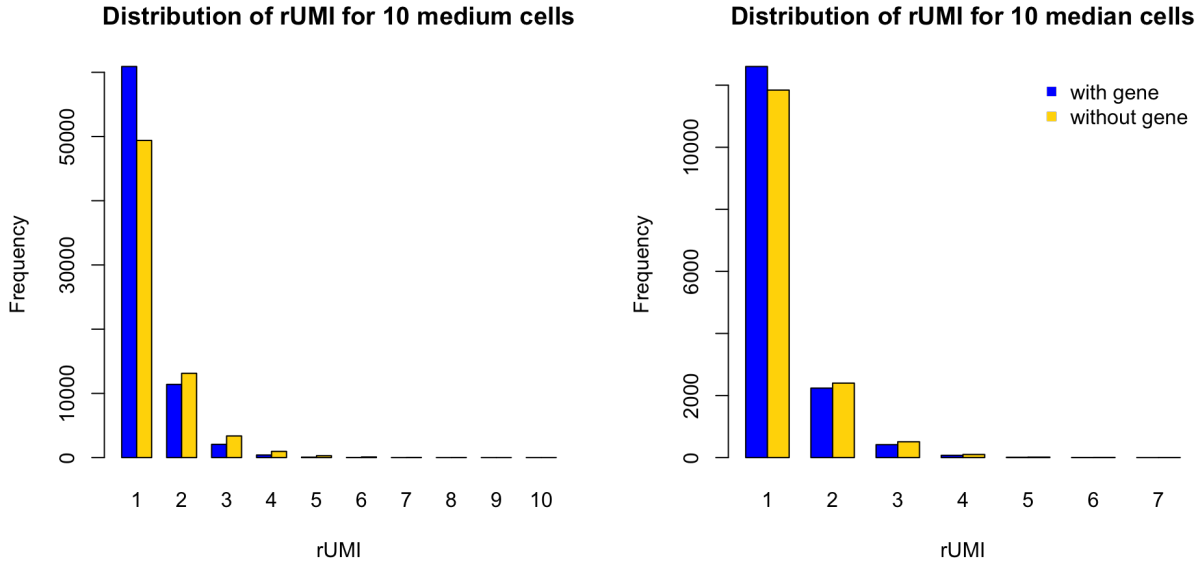

Figure 7.12: Histograms of rUMI for the 10 medium (left panel) and 10 median (right panel) cells with (in blue) and without (in yellow) using gene as an additional deduplication characteristic. Compare to Figure 3.5.

7.2.2 Univariate Associations

We recreate Figures 3.6 to 3.9, 7.1, 7.6, and 7.9 for the twenty additional cells in Figures 7.13 to 7.20. The results are largely similar to those found in the corresponding Figures from the 10 cells with the highest expression. It appears that the patterns noted for the 10 cells with high expression are not a product of the cell size but are experienced by many cells. Note that fewer UMIs are represented in the medium and median cells than in the 10 cells with high expression shown earlier.

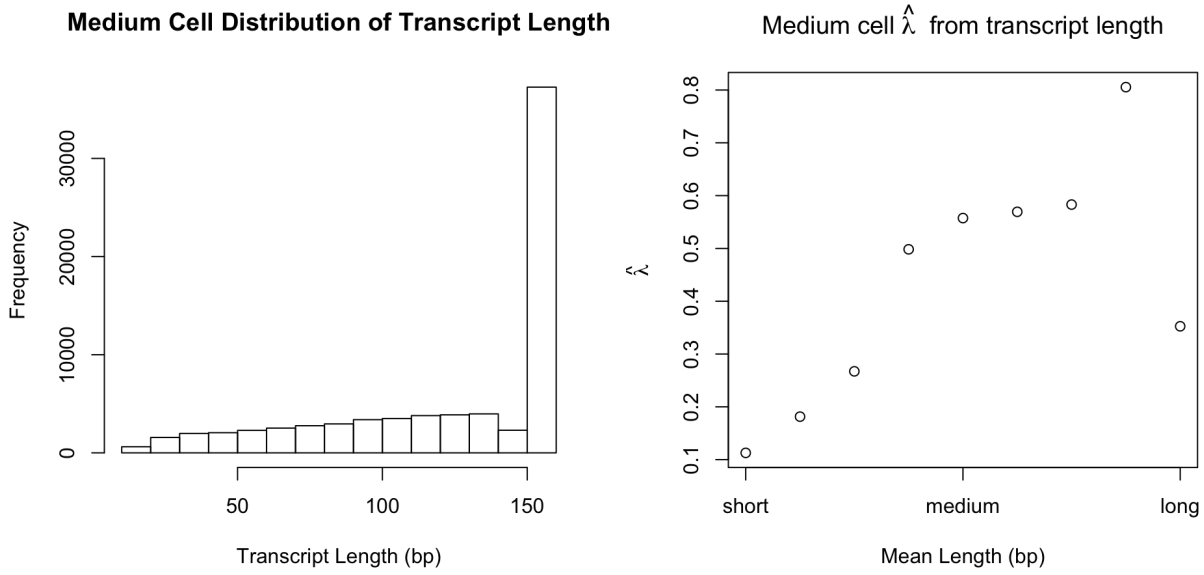

Figure 7.13: Histogram of length (bp) of the transcripts (left panel) and estimated $\hat{\lambda}$ based on length (right panel) for the medium cells. Note that roughly 50% of the transcripts are full length (151 bp) and represented by the one $\hat{\lambda}$ estimate for long transcripts. Compare to Figure 3.6.

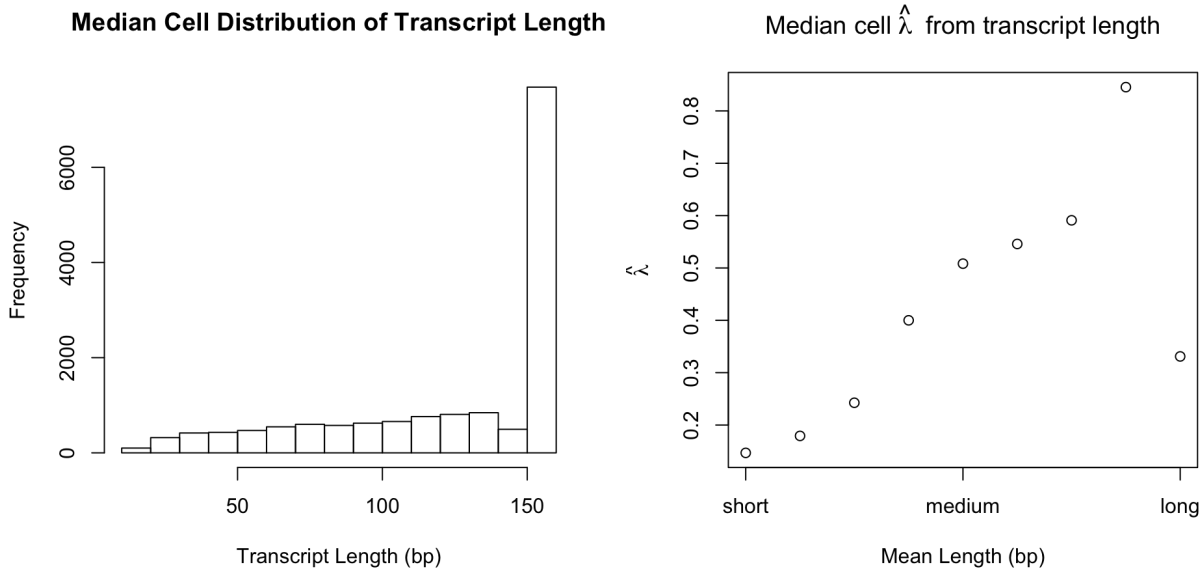

Figure 7.14: Histogram of length (bp) of the transcripts (left panel) and estimated $\hat{\lambda}$ based on length (right panel) for the median cells. Note that roughly 50% of the transcripts are full length (151 bp) and represented by the one $\hat{\lambda}$ estimate for long transcripts. Compare to Figure 3.6.

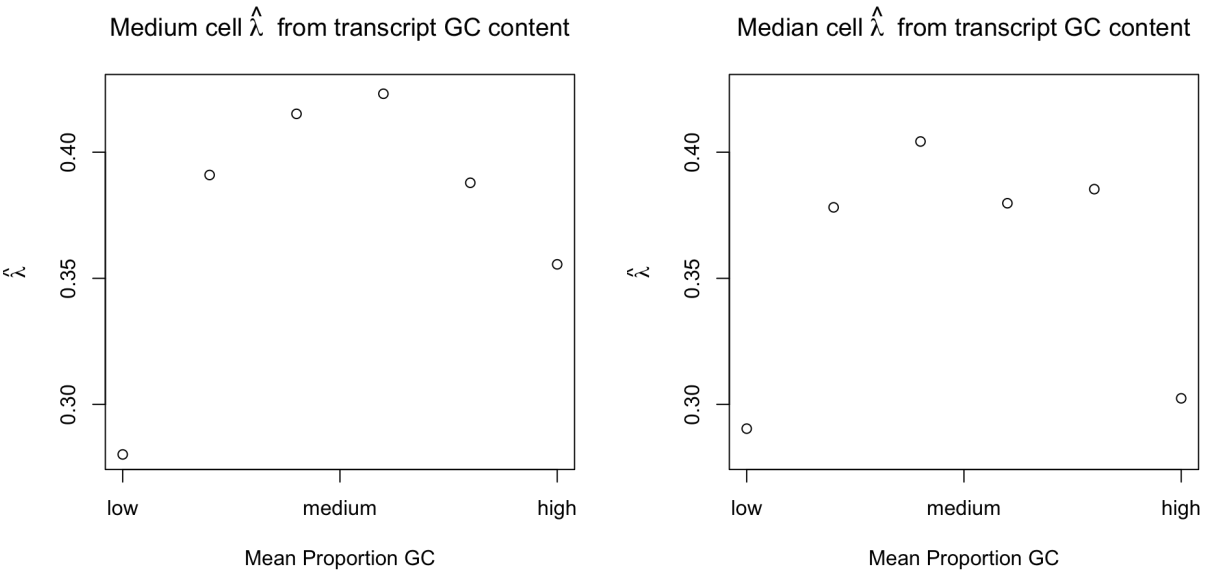

Figure 7.15: Estimated $\hat{\lambda}$ based on the mean proportion GC content of the transcript for the medium (left panel) and median (right panel) cells. Compare to Figure 3.7.

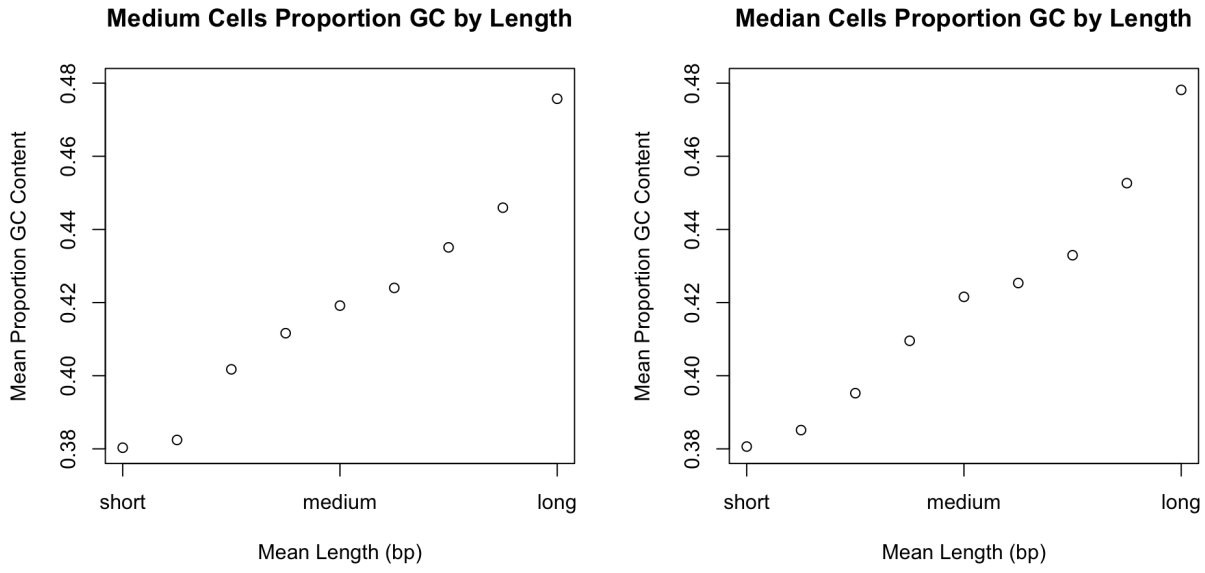

Figure 7.16: Scatterplot of the relationship between the mean length (bp) and the mean proportion GC content of the transcripts for the medium (left panel) and the median (right panel) cells. Compare to Figure 3.8.

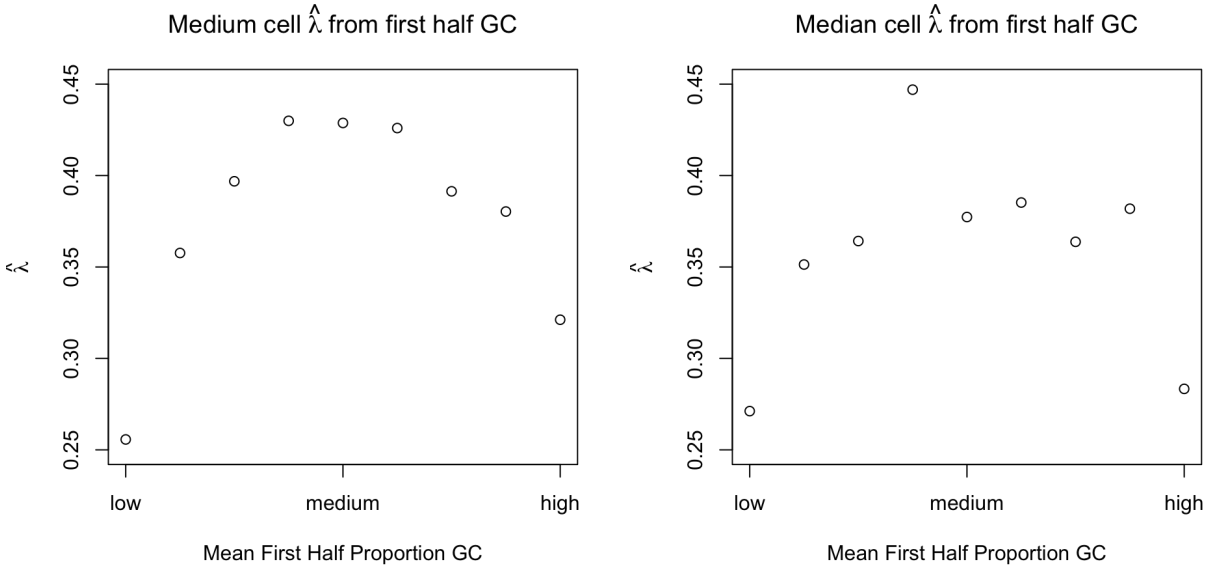

Figure 7.17: Estimated $\hat{\lambda}$ based on the mean proportion GC content in the first half of the transcript for the medium (left panel) and the median (right panel) cells. Compare to Figure 3.9.

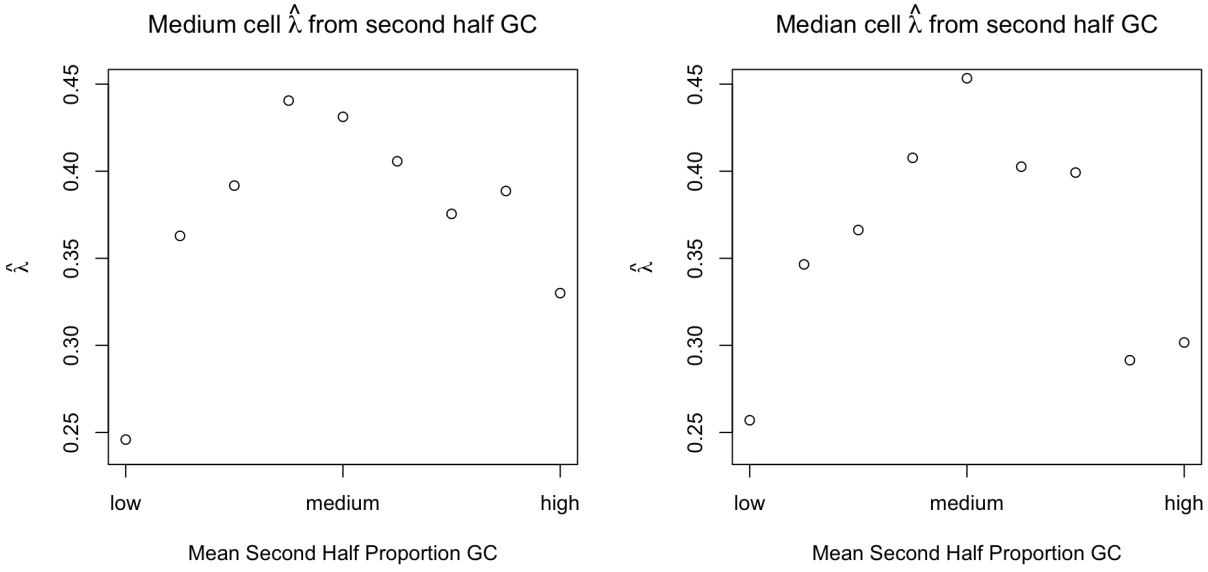

Figure 7.18: Estimated $\hat{\lambda}$ based on the mean proportion GC content in the second half of the transcript for the medium (left panel) and the median (right panel) cells. Compare to Figure 7.1.

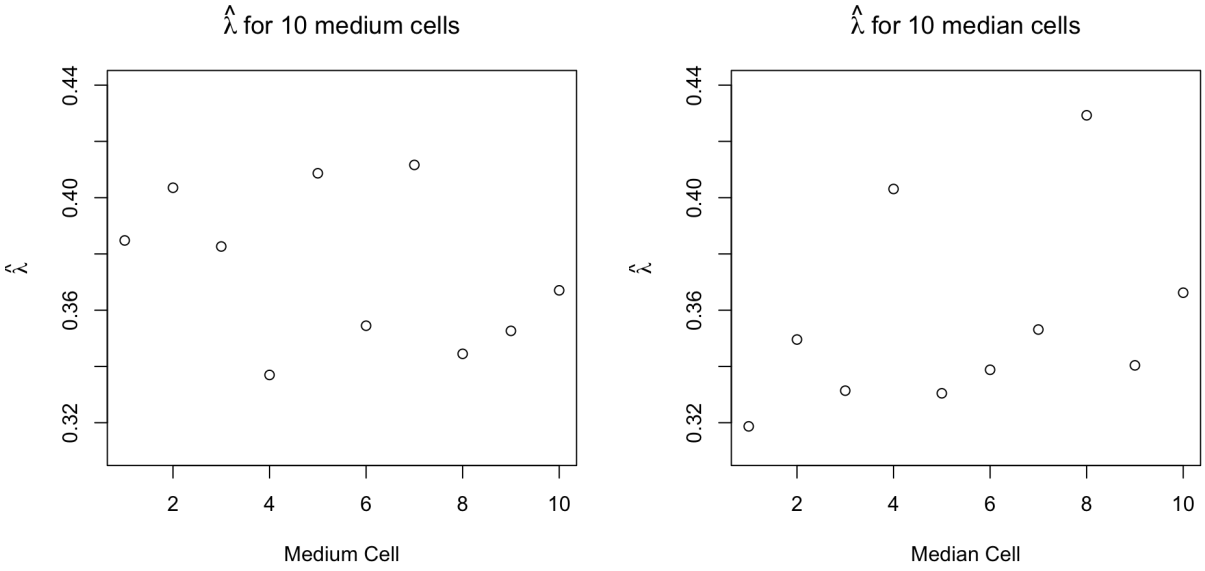

Figure 7.19: Estimated $\hat{\lambda}$ for each of the individual cells included in the 10 medium (left panel) and 10 median (right panel) cells. Compare to Figure 7.6.

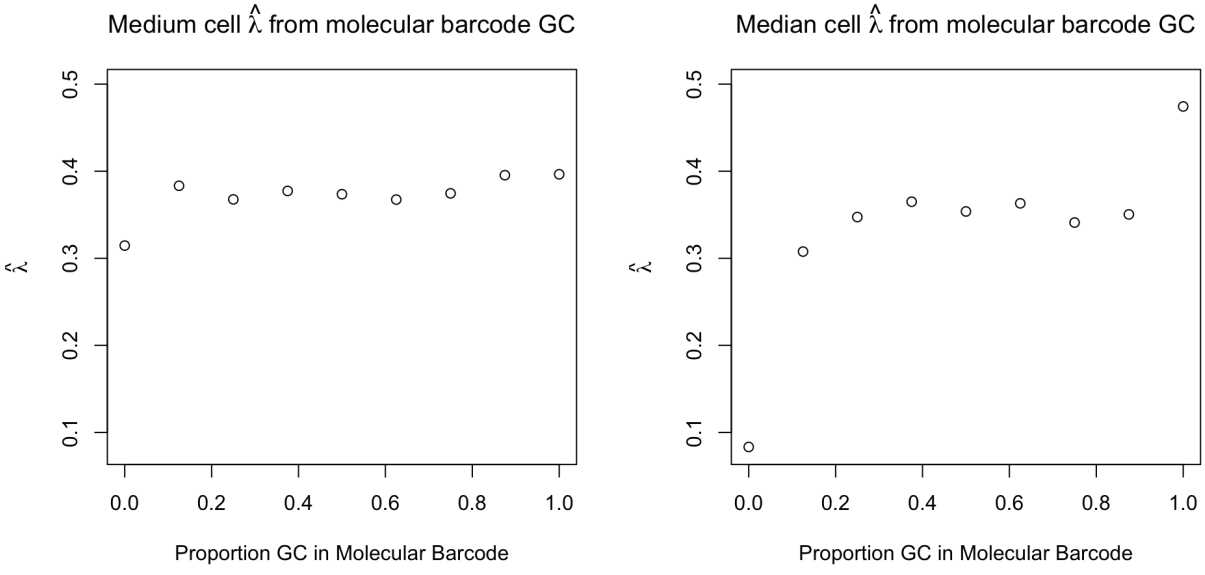

Figure 7.20: Estimated $\hat{\lambda}$ based on the proportion GC content in the molecular barcode from the medium (left panel) and the median (right panel) cells. Compare to Figure 7.9.

7.3 Model for rUMIs

We filter to the UMIs with full length reads for model fitting purposes, resulting in a mean rUMI of 1.20 (sd = 0.50). Our estimated $\hat{\lambda}$ from Equation 3.2 is reduced to 0.34 from 0.38 for all transcripts.

We predict rUMI with a general linear model from a zero-truncated Poisson family, based on biological and analytical motivation. Recall that 0 rUMI cannot be recorded. We generated analogous linear models, as well. Based on our findings in Section 3.4.2, we consider full sequence proportion GC content, first half of sequence proportion GC content, AA two-somes, and GC two-somes as predictors. Due to collinearity between the proportion GC content in the full sequence and first half of the sequence, we selected the full sequence GC content based on interpretability. The two-somes also exhibit collinearity, so we selected only the two-somes with the strongest relationships observed in Figures 7.4 and 7.5.

The fitted model for rUMI is:

$$\hat{\mu}_W(T, U, V) = \exp(-2.20 + 4.85T - 5.54T^2 + 2.72U - 13.35U^2 + 3.60V - 20.15V^2), \quad (7.1)$$

where $W =$ rUMI, $T =$ average proportion GC content of the UMI, $U =$ average proportion of AA two-somes of the UMI, and $V =$ average proportion of GC two-somes of the UMI. The analogous linear model has an $R^2 = 0.0011$.

With over 80% UMIs recorded with 1 read (bottom panel of Figure 3.4), the fitted values for this model are close to 1 (Figure 7.21). The average fitted rUMI increases from 1.205 to 1.216 for UMIs with 1 read compared to multiple reads.

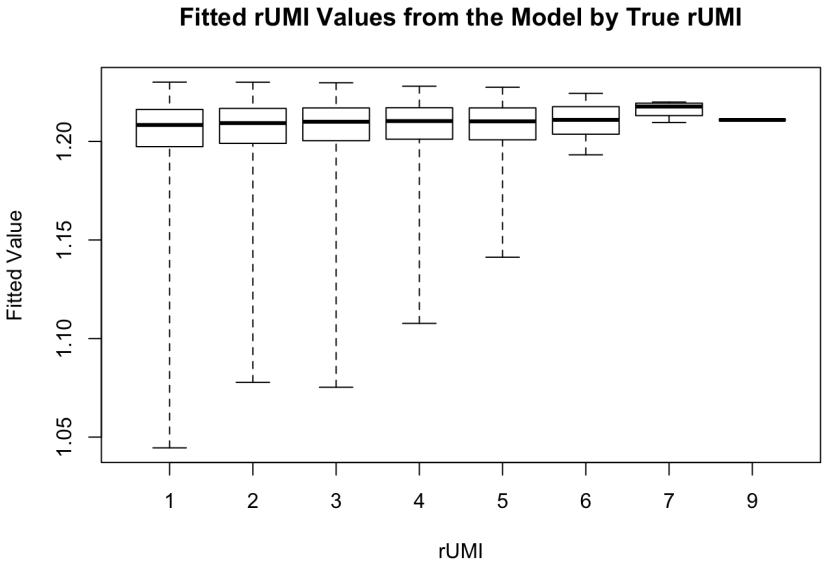

Figure 7.21: Side-by-side boxplots of the fitted values based on Model 7.1 are shown for each rUMI level. As rUMI increases, the fitted values generally increase. Note that the whiskers extend to the maximum and minimum fitted values for each level. Outliers (defined as being beyond 1.5 times the IQR from Q1 or Q3) are present on the lower end for each of 1 to 5 rUMIs.

Chapter 8

Supplement: Zero-truncated Distribution Models

8.1 Distribution Moments

Special Case 1: Binomial Distribution For the ZTBinomial distribution, moments are as follows:

Moment	Expression
$\mathbb{E}(C)$	$\frac{1}{1-\left(\frac{N_0-\lambda}{N_0}\right)^{N_0}}\lambda$
$\mathbb{E}(C^2)$	$\frac{1}{1-\left(\frac{N_0-\lambda}{N_0}\right)^{N_0}}\left(\lambda^2-\frac{\lambda^2}{N_0}+\lambda\right)$
$\mathbb{E}(C^3)$	$\frac{1}{1-\left(\frac{N_0-\lambda}{N_0}\right)^{N_0}}\left((N_0-1)(N_0-2)\frac{\lambda^3}{N_0^2}+3(N_0-1)\frac{\lambda^2}{N_0}+\lambda\right)$
$\mathbb{E}(C^4)$	$\frac{1}{1-\left(\frac{N_0-\lambda}{N_0}\right)^{N_0}}\left((N_0-1)(N_0-2)(N_0-3)\frac{\lambda^4}{N_0^3} +6(N_0-1)(N_0-2)\frac{\lambda^3}{N_0^2}+7(N_0-1)\frac{\lambda^2}{N_0}+\lambda\right)$

Table 8.1: Moments of the Zero Truncated Binomial Distribution

Special Case 2: Poisson Distribution For the ZTPoisson distribution, moments are as follows:

Moment	Expression
$\mathbb{E}(C)$	$\frac{\lambda}{1-e^{-\lambda}}$
$\mathbb{E}(C^2)$	$\frac{\lambda+\lambda^2}{1-e^{-\lambda}}$
$\mathbb{E}(C^3)$	$\frac{\lambda^3+3\lambda^2+\lambda}{1-e^{-\lambda}}$
$\mathbb{E}(C^4)$	$\frac{\lambda^4+6\lambda^3+7\lambda^2+\lambda}{1-e^{-\lambda}}$

Table 8.2: Moments of the Zero Truncated Poisson Distribution

Special Case 3: Negative Binomial Distribution For the ZTNegative Binomial distribution, moments are as follows:

Moment	Expression
$\mathbb{E}(C)$	$\frac{1}{1-\left(\frac{\alpha}{\alpha+\lambda_0}\right)^{\alpha}}\lambda_0$
$\mathbb{E}(C^2)$	$\frac{1}{1-\left(\frac{\alpha}{\alpha+\lambda_0}\right)^{\alpha}}\left(\lambda_0^2+\frac{\lambda_0^2}{\alpha}+\lambda_0\right)$
$\mathbb{E}(C^3)$	$\frac{1}{1-\left(\frac{\alpha}{\alpha+\lambda_0}\right)^{\alpha}}\left(\lambda_0+3(\alpha+1)\frac{\lambda_0^2}{\alpha}+(\alpha+1)(\alpha+2)\frac{\lambda_0^3}{\alpha^2}\right)$
$\mathbb{E}(C^4)$	$\frac{1}{1-\left(\frac{\alpha}{\alpha+\lambda_0}\right)^{\alpha}}\left(\lambda_0+7(\alpha+1)\frac{\lambda_0^2}{\alpha}+6(\alpha+1)(\alpha+2)\frac{\lambda_0^3}{\alpha^2}+(\alpha+1)(\alpha+2)(\alpha+3)\frac{\lambda_0^4}{\alpha^3}\right)$

Table 8.3: Moments of the Zero Truncated Negative Binomial Distribution

8.2 Estimator Properties

We begin by deriving a general form of a multivariate Taylor expansion for fractions. We use this throughout Sections 8.2.1 and 8.2.2.

Suppose that we have some general fraction of the form $\frac{X_1}{X_2}$. Then, we can take a multivariate Taylor expansion about the means of the two variables (μ_1, μ_2).

This first-degree Taylor approximation is

$$\frac{X_1}{X_2} \approx \frac{\mu_1}{\mu_2} + \frac{X_1 - \mu_1}{\mu_2} - \frac{\mu_1(X_2 - \mu_2)}{\mu_2^2}. \tag{8.1}$$

The second-degree Taylor approximation is

$$\frac{X_1}{X_2} \approx \frac{\mu_1}{\mu_2} + \frac{X_1 - \mu_1}{\mu_2} - \frac{\mu_1(X_2 - \mu_2)}{\mu_2^2} - \frac{(X_1 - \mu_1)(X_2 - \mu_2)}{\mu_2^2} + \frac{\mu_1(X_2 - \mu_2)^2}{\mu_2^3}. \tag{8.2}$$

From Equation 8.2, we see that

$$\begin{aligned}
\mathbb{E}\left(\frac{X_1}{X_2}\right) &\approx \mathbb{E}\left(\frac{\mu_1}{\mu_2} + \frac{X_1 - \mu_1}{\mu_2} - \frac{\mu_1(X_2 - \mu_2)}{\mu_2^2} - \frac{(X_1 - \mu_1)(X_2 - \mu_2)}{\mu_2^2} + \frac{\mu_1(X_2 - \mu_2)^2}{\mu_2^3}\right) \\
&= \frac{\mu_1}{\mu_2} - \frac{\text{Cov}(X_1, X_2)}{\mu_2^2} + \frac{\mu_1}{\mu_2^3}\text{Var}(X_2) \\
&= \frac{\mathbb{E}(X_1)}{\mathbb{E}(X_2)} - \frac{\text{Cov}(X_1, X_2)}{\mathbb{E}(X_2)^2} + \frac{\mathbb{E}(X_1)}{\mathbb{E}(X_2)^3}\text{Var}(X_2).
\end{aligned} \tag{8.3}$$

From Equation 8.1, we observe that

$$\begin{aligned}
\text{Var}\left(\frac{X_1}{X_2}\right) &\approx \text{Var}\left(\frac{\mu_1}{\mu_2} + \frac{X_1 - \mu_1}{\mu_2} - \frac{\mu_1(X_2 - \mu_2)}{\mu_2^2}\right) \\
&= \frac{\text{Var}(X_1)}{\mu_2^2} + \frac{\mu_1^2}{\mu_2^4}\text{Var}(X_2) - \frac{2\mu_1}{\mu_2^3}\text{Cov}(X_1, X_2) \\
&= \frac{\text{Var}(X_1)}{\mathbb{E}(X_2)^2} + \frac{\mathbb{E}(X_1)^2}{\mathbb{E}(X_2)^4}\text{Var}(X_2) - \frac{2\mathbb{E}(X_1)}{\mathbb{E}(X_2)^3}\text{Cov}(X_1, X_2).
\end{aligned} \tag{8.4}$$

8.2.1 Estimator 1

Estimator 1, as defined in Section 4.4.2.1, is a moments-based estimator. For the derivations in this Section, we use the right-side version of Estimator 1 in Equation 4.4.

We take a multivariate Taylor expansion of $\frac{\sum C_i^2/n}{\sum C_i/n}$ around the means of $\left(\frac{\sum C_i^2}{n}, \frac{\sum C_i}{n}\right)$. Combining the Taylor expansion from Equation 8.3 with the other components of Equation 4.4, we derive that

$$\mathbb{E}(\hat{\lambda}_1) \approx \frac{n}{n-1}\left(\frac{\mathbb{E}(C^2)}{\mathbb{E}(C)} - \frac{\mathbb{E}(C^3) - \mathbb{E}(C^2)\mathbb{E}(C)}{n\mathbb{E}(C)^2} + \frac{\mathbb{E}(C^2)(\mathbb{E}(C^2) - \mathbb{E}(C)^2)}{n\mathbb{E}(C)^3}\right) - \frac{\mathbb{E}(C)}{n-1} - 1.$$

This simplifies to

$$\mathbb{E}(\hat{\lambda}_1) \to \frac{\mathbb{E}(C^2)}{\mathbb{E}(C)} - 1.$$

The variance of Estimator 1 is found using the multivariate Taylor expansion around the means of $\left(\frac{\sum C_i^2}{n}, \frac{\sum C_i}{n}\right)$. Using Equation 8.4, we see that

$$\text{Var}(\sqrt{n}\hat{\lambda}_1) \to \frac{\mathbb{E}(C^4) - \mathbb{E}(C^2)^2}{\mathbb{E}(C)^2} - \frac{2\mathbb{E}(C^2)(\mathbb{E}(C^3) - \mathbb{E}(C^2)\mathbb{E}(C))}{\mathbb{E}(C)^3} + \frac{\mathbb{E}(C^2)^2(\mathbb{E}(C^2) - \mathbb{E}(C)^2)}{\mathbb{E}(C)^4}.$$

Note that $\text{Cov}\left(\sum C_i^2, \sum C_i\right) = n\mathbb{E}(C_i^3) - n\mathbb{E}(C_i^2)\mathbb{E}(C_i)$.

We can apply the moments from Tables 8.1 to 8.3 to the Equations above to calculate the asymptotic biases and variances from different data generating functions.

Figure 8.1 shows simulated asymptotic biases and variances of $\hat{\lambda}_1$ with different data generating procedures. The dotted lines represent the theoretical asymptotic biases and variances described here, while the solid lines represent the simulated biases and variances for the given sample size.

We can additionally provide details about the asymptotic means and variances for $\hat{\pi}_1$. We take a Taylor approximation of $e^{-\lambda}$ about $\mathbb{E}(\hat{\lambda}_1)$. Note that $\mathbb{E}(\hat{\lambda}_1)$ will change depending on the distribution that C follows.

For the ZTPoisson distribution, $\hat{\pi}_1$ remains asymptotically unbiased, and the asymptotic variance is multiplied by $e^{-2\lambda}$. When $C \sim$ ZTBin, the asymptotic bias is $e^{-\lambda}(e^{\frac{\lambda}{N_0}} - 1)$, and the asymptotic variance is multiplied by $e^{-2\lambda\left(1-\frac{1}{N_0}\right)}$. When the data are drawn from a ZTNegative Binomial distribution, the asymptotic bias of $\hat{\pi}_1$ is $e^{-\lambda}(e^{-\frac{\lambda}{\alpha}} - 1)$, and the asymptotic variance is the asymptotic variance of $\hat{\lambda}_1$ when $C \sim$ ZTNegBin multiplied by $e^{-2\lambda\left(1+\frac{1}{\alpha}\right)}$.

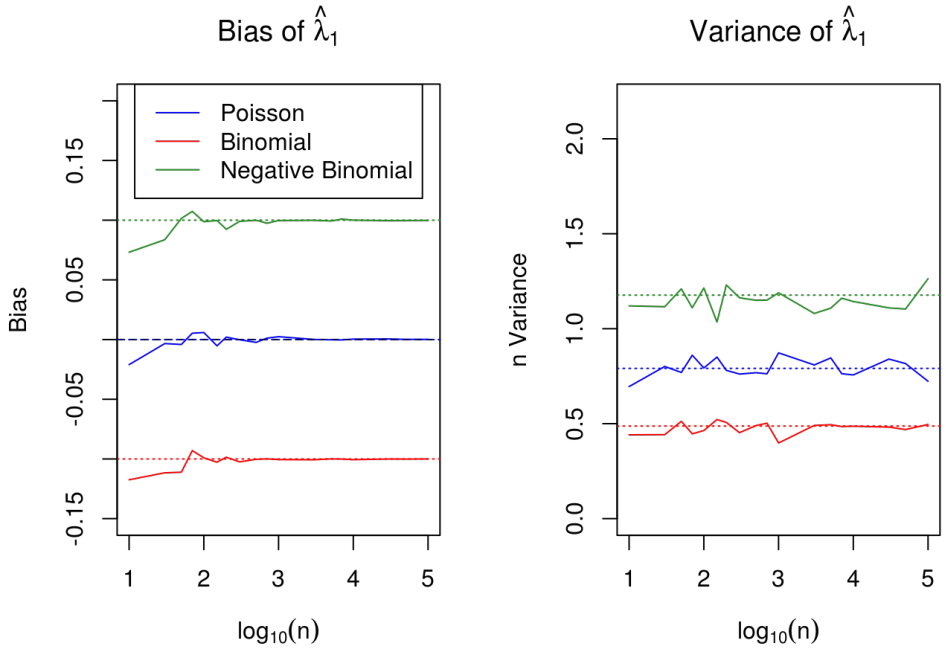

Figure 8.1: Simulated biases and variances for $\hat{\lambda}_1$ for data drawn from ZTPoisson, ZTBinomial, and ZTNegative Binomial distributions with $\lambda = 0.4$, $N_0 = 4$, and $\alpha = 4$. The dotted lines show the theoretical results. 500 samples of 18 different sizes ranging from 10 to 100,000 were taken.

8.2.2 Estimator 2

Estimator 2, as defined in Section 4.4.2.2, is a robust estimator. Note that the two variables used in the estimator, $\sum \mathbb{I}(C_i = 1)$ and $\sum \mathbb{I}(C_i = 2)$, both follow a Binomial distribution with the same n trials and probabilities of success of $\mathbb{P}(C_i = 1)$ and $\mathbb{P}(C_i = 2)$, respectively.

We take a multivariate Taylor expansion of $\hat{\lambda}_2$ around the means of $(n\mathbb{P}(C_i = 2), n\mathbb{P}(C_i = 1))$. From Equation 8.3 and the Binomial distributions of the numerator and denominator,

$$\mathbb{E}(\hat{\lambda}_2) \approx \frac{2\mathbb{P}(C = 2)}{\mathbb{P}(C = 1)} \left(1 + \frac{1}{n\mathbb{P}(C = 1)}\right).$$

This simplifies to

$$\mathbb{E}(\hat{\lambda}_2) \to \frac{2\mathbb{P}(C = 2)}{\mathbb{P}(C = 1)}.$$

From Equation 8.4, the asymptotic variance of Estimator 2 is

$$\text{Var}(\sqrt{n}\hat{\lambda}_2) \to 4\mathbb{P}(C = 2) \left(\frac{1 - 2\mathbb{P}(C = 2)}{\mathbb{P}(C = 1)^2} + 2\mathbb{P}(C = 2) + \frac{\mathbb{P}(C = 2)}{\mathbb{P}(C = 1)^3}\right).$$

Note that together, $\sum \mathbb{I}(C_i = 1)$ and $\sum \mathbb{I}(C_i = 2)$ follow a Multinomial distribution. Therefore, $\text{Cov}\left[\sum \mathbb{I}(C_i = 1), \sum \mathbb{I}(C_i = 2)\right] = -n\mathbb{P}(C_i = 1)\mathbb{P}(C_i = 2)$.

We can apply the moments from Tables 8.1 to 8.3 to the general equations above to calculate the asymptotic biases and variances from different data generating functions.

For example, when $C \sim \text{ZTBinom}\left(N_0, \frac{\lambda}{N_0}\right)$, the asymptotic variance of $\sqrt{n}\hat{\lambda}_2$ is

$$4\frac{N_0 - 1}{N_0(N_0 - \lambda)^4} \left[N_0^4(1 + \lambda - \lambda^2 + 2\lambda^4) + N_0^3\lambda(-5 - 2\lambda + 2\lambda^2 - \lambda^3) \right.$$
$$\left. + N_0^2\lambda^2(9 + \lambda - \lambda^2) + N_0\lambda^3(-7 - \lambda) + 2\lambda^4\right].$$

Figure 8.2 shows simulated asymptotic biases and variances of $\hat{\lambda}_2$ with different data generating procedures. The solid lines represent the simulated biases and variances, while the dotted lines are the theoretical biases and variances described here.

We can also calculate the asymptotic means and variances for $\hat{\pi}_2$. To do this, we take a Taylor approximation of $e^{-\lambda}$ about $\mathbb{E}(\hat{\lambda}_2)$. Note that $\mathbb{E}(\hat{\lambda}_2)$ changes depending on the distribution of the data.

When $C \sim \text{ZTPois}(\lambda)$, $\hat{\pi}_2$ remains asymptotically unbiased and the asymptotic variance of $\hat{\lambda}_2$ is multiplied by $e^{-2\lambda}$. If instead the true distribution of C is ZTBinomial, then the asymptotic bias is $e^{-\lambda}(e^{\frac{\lambda(1-\lambda)}{N_0-\lambda}} - 1)$ and the asymptotic variance of $\hat{\lambda}_2$ is multiplied by $e^{-2\lambda\left(1+\frac{\lambda-1}{N_0-\lambda}\right)}$. If $C \sim \text{ZTNegBin}\left(\alpha, \frac{\lambda}{\alpha+\lambda}\right)$, then the asymptotic bias of $\hat{\pi}_2$ is $e^{-\lambda}(e^{\frac{\lambda(\lambda-1)}{\lambda+\alpha}} - 1)$. The asymptotic variance

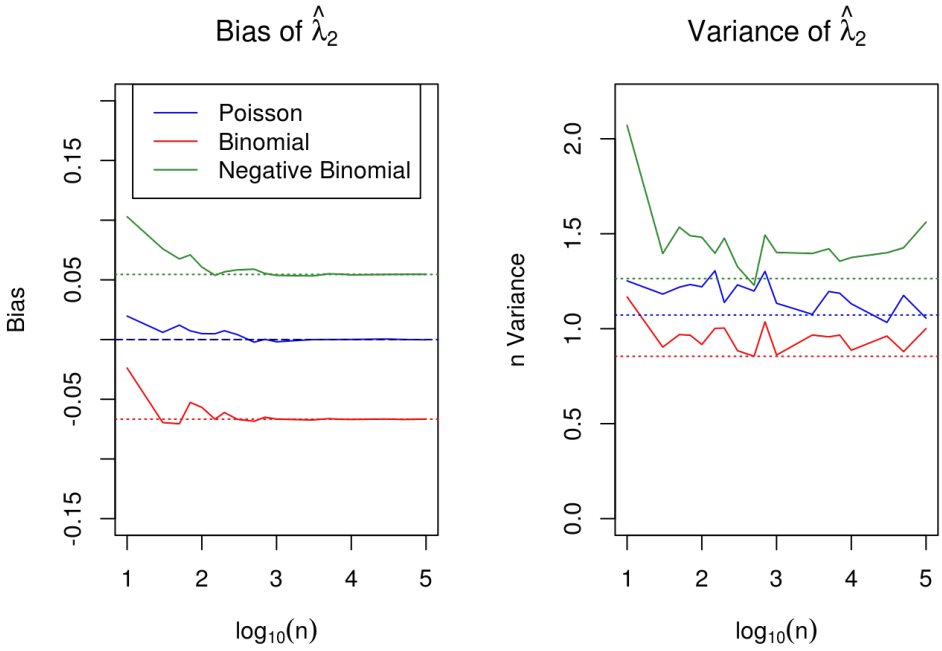

Figure 8.2: Simulated biases and variances for $\hat{\lambda}_2$ for data drawn from ZTPoisson, ZTBinomial, and ZTNegative Binomial distributions with $\lambda = 0.4$, $N_0 = 4$, and $\alpha = 4$. The dotted lines show the theoretical results. 500 samples of 18 different sizes ranging from 10 to 100,000 were taken.

of $\hat{\lambda}_2$ is multiplied by $e^{-2\lambda\left(1+\frac{1-\lambda}{\lambda+\alpha}\right)}$ to calculate the asymptotic variance of $\hat{\pi}_2$.

8.2.3 Estimator 3

The asymptotic bias and variance for $\hat{\pi}_3$ when $C \sim \text{ZTPois}(\lambda)$ can be estimated from the asymptotic distribution for $\hat{\lambda}_3$ in Section 4.4.3.3. By the Delta Method,

$$\sqrt{n}(\hat{\pi}_3 - \pi) \xrightarrow{d} \text{N}\left(0, \frac{\lambda e^{-\lambda}(1 - e^{-\lambda})^2}{e^{\lambda} - 1 - \lambda}\right).$$

Figure 8.3 shows simulated asymptotic biases and variances of $\hat{\lambda}_3$ with different data generating procedures. Note that the theoretical bias and variance (dotted lines) could only be calculated when the data are ZTPoisson. The simulated biases and variances are shown as solid lines.

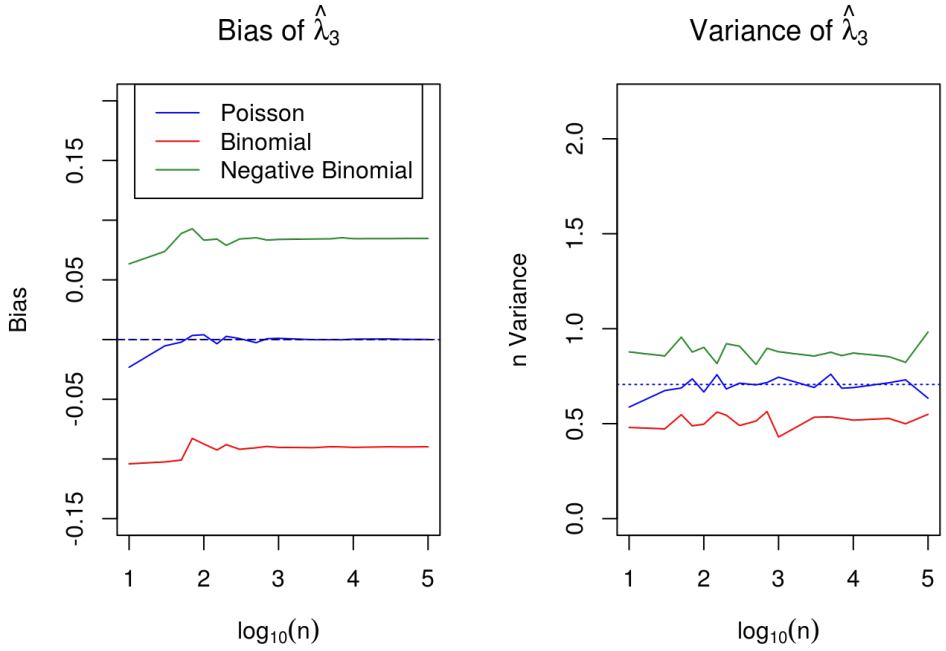

Figure 8.3: Simulated biases and variances for $\hat{\lambda}_3$ for data drawn from ZTPoisson, ZTBinomial, and ZTNegative Binomial distributions with $\lambda = 0.4$, $N_0 = 4$, and $\alpha = 4$. The dotted lines show the theoretical results. 500 samples of 18 different sizes ranging from 10 to 100,000 were taken.

8.2.4 Estimator 4

The asymptotic bias and variance for $\hat{\pi}_4$ can be estimated from the asymptotic distribution for $\hat{\lambda}_4$ in Section 4.4.3.4 if we assume that N_0 is known beforehand. By the Delta Method,

$$\sqrt{n}(\hat{\pi}_4 - \pi) \xrightarrow{d} \mathrm{N}\left(0, \frac{\lambda(N_0-\lambda)^{2N_0}(N_0^{N_0} - (N_0-\lambda)^{N_0})^2}{N_0^{3N_0-1}[(N_0-\lambda)(N_0^{N_0} - (N_0-\lambda)^{N_0}) - \lambda N_0(N_0-\lambda)^{N_0}]}\right).$$

Figure 8.4 shows simulated asymptotic biases and variances (solid lines) of $\hat{\lambda}_4$ with different data generating procedures. Note that the theoretical bias and variance (dotted lines) are calculated only when the data are ZTBinomial and assuming a known N_0. This assumption is violated when estimating $\hat{\lambda}_4$, as $\hat{N}_0$ is also estimated.

Figure 8.5 shows simulated asymptotic biases and variances (solid lines) of $\hat{\lambda}_4$ if we assume that N_0 is known to be 4 with different data generating procedures. The difference between Figure 8.5 and Figure 8.4 shows how estimating the N_0 changes the asymptotic biases and variances from those calculated when N_0 is known.

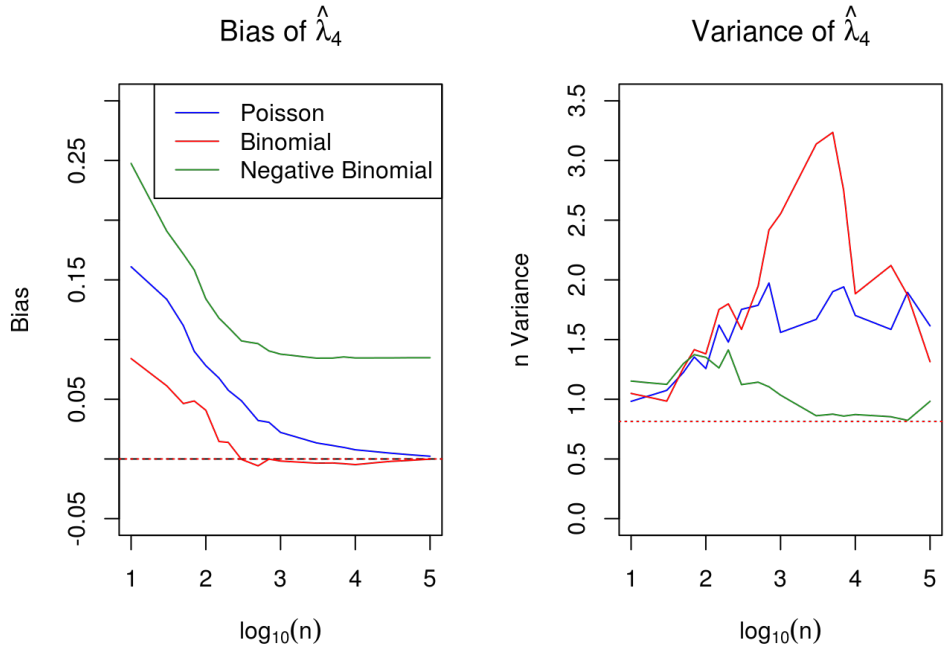

Figure 8.4: Simulated biases and variances for $\hat{\lambda}_4$ for data drawn from ZTPoisson, ZTBinomial, and ZTNegative Binomial distributions with $\lambda = 0.4$, $N_0 = 4$, and $\alpha = 4$. The dotted lines show the theoretical results. 500 samples of 18 different sizes ranging from 10 to 100,000 were taken.

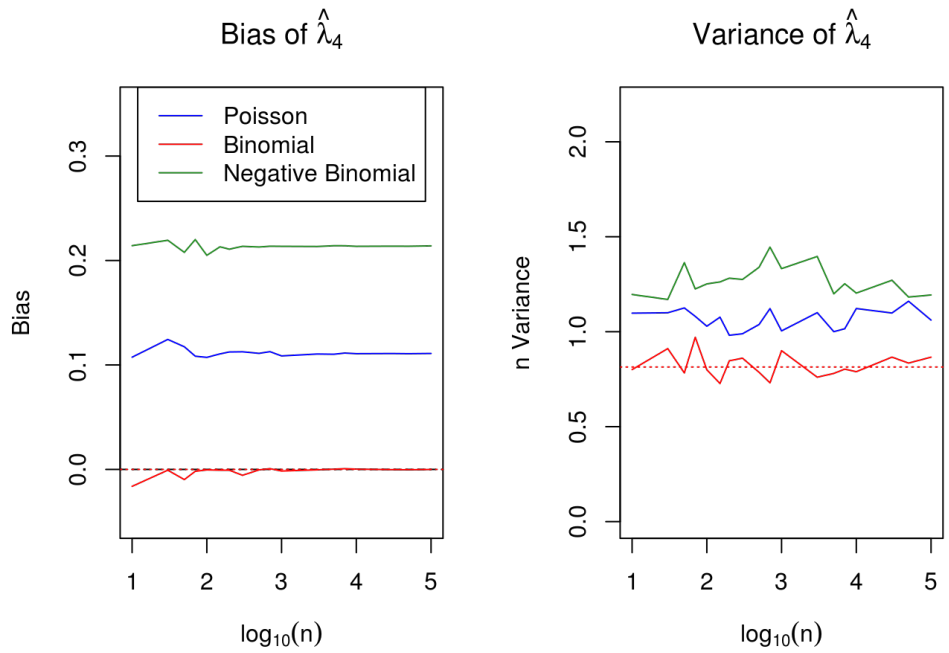

Figure 8.5: Simulated biases and variances for $\hat{\lambda}_4$ if we assume $N_0 = 4$ is known for data drawn from ZTPoisson, ZTBinomial, and ZTNegative Binomial distributions with $\lambda = 0.4$, $N_0 = 4$, and $\alpha = 4$. The dotted lines show the theoretical results. 500 samples of 18 different sizes ranging from 10 to 100,000 were taken.

8.2.5 Estimator 5

The asymptotic bias and variance for $\hat{\pi}_5$ can be estimated from the asymptotic distribution for $\hat{\lambda}_5$ in Section 4.4.3.5 if we assume that α is known beforehand. By the Delta Method,

$$\sqrt{n}(\hat{\pi}_5 - \pi) \xrightarrow{d} \mathrm{N}\left(0, \frac{\lambda\alpha^{2\alpha+1}[(\alpha+\lambda)^{\alpha} - \alpha^{\alpha}]^2}{(\alpha+\lambda)^{3\alpha+1}[(\alpha+\lambda)^{\alpha} - \alpha^{\alpha}] - \lambda\alpha^{\alpha+1}(\alpha+\lambda)^{3\alpha}}\right).$$

Figure 8.6 shows simulated asymptotic biases and variances (solid lines) of $\hat{\lambda}_5$ with different data generating procedures. Note that the theoretical bias and variance (dotted lines) are only calculated when the data are ZTNegative Binomial and assume a known α. We violate the assumption of a known α when estimating $\hat{\lambda}_5$ from simulated data, which contributes to the simulated variance exceeding the theoretical one.

Figure 8.7 shows simulated asymptotic biases and variances (solid lines) of $\hat{\lambda}_5$ assuming that α is known to be 4 with different data generating procedures. The difference between Figure 8.7 and Figure 8.6 shows how estimating the α changes the asymptotic biases and variances from those calculated when α is known.

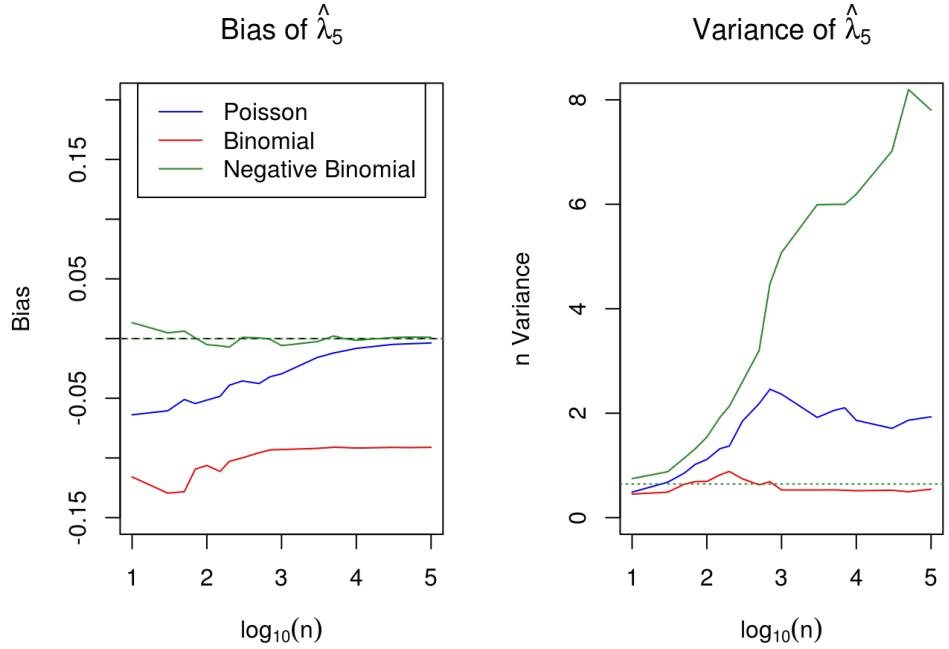

Figure 8.6: Simulated biases and variances for $\hat{\lambda}_5$ for data drawn from ZTPoisson, ZTBinomial, and ZTNegative Binomial distributions with $\lambda = 0.4$, $N_0 = 4$, and $\alpha = 4$. The dotted lines show the theoretical results. 500 samples of 18 different sizes ranging from 10 to 100,000 were taken.

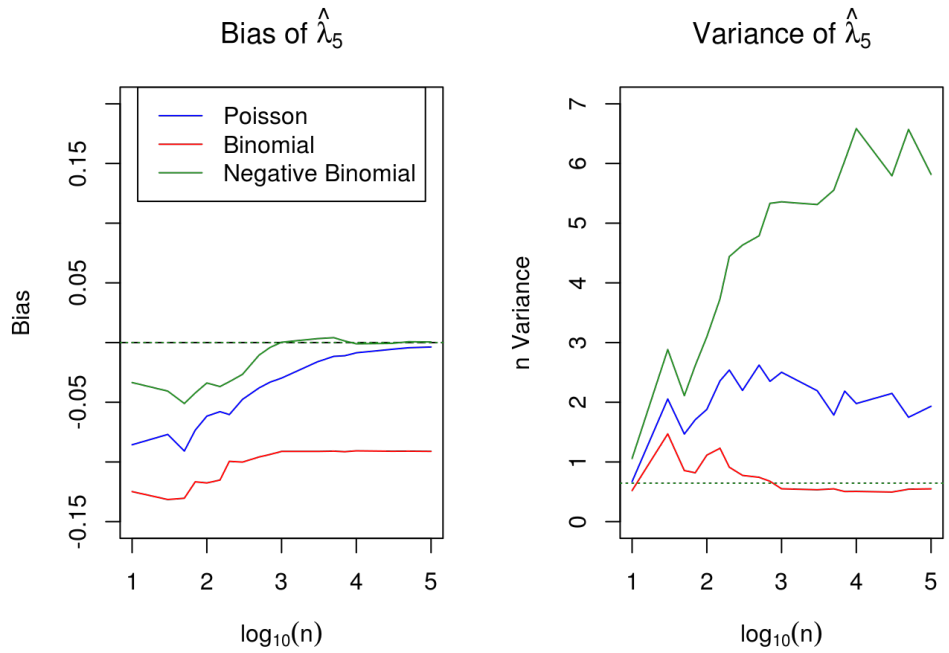

Figure 8.7: Simulated biases and variances for $\hat{\lambda}_5$ if we assume $\alpha = 4$ is known for data drawn from ZTPoisson, ZTBinomial, and ZTNegative Binomial distributions with $\lambda = 0.4$, $N_0 = 4$, and $\alpha = 4$. The dotted lines show the theoretical results. 500 samples of 18 different sizes ranging from 10 to 100,000 were taken.

8.3 Simulation Results

In Section 4.5, we provide Tables 4.1 to 4.3 to summarize the MSEs of the estimated probabilities that a given molecule was not recorded, i.e. $\hat{\pi}$. Tables 8.4 to 8.6 provide the corresponding MSE values for the $\hat{\lambda}$ estimators.

Additionally, we provide Tables 4.4 and 4.5 to demonstrate the MSEs of the estimated probabilities under perturbations of the general data generating functions. Tables 8.7 and 8.8 provide the corresponding MSEs for the $\hat{\lambda}$ estimators.

Distribution		MSE $\times 10^3$				
λ	Family	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$
0.1	ZTP	0.001	0.001	0.001	0.011	0.012
	ZTB	0.627	0.534	0.595	0.015	0.612
	ZTNB	0.626	0.481	0.577	0.578	0.027
0.4	ZTP	0.003	0.005	0.003	0.010	0.013
	ZTB	9.985	4.426	8.036	0.003	8.290
	ZTNB	10.002	2.970	7.182	7.199	0.026
1.5	ZTP	0.012	0.053	0.009	0.011	0.027
	ZTB	140.604	90.326	50.178	0.005	51.972
	ZTNB	140.577	18.593	41.423	41.511	0.022

Table 8.4: Mean squared error (multiplied by 10^3) for the five estimators from three data generating processes. 500 samples of size 250,000 were collected.$N = 4$ for the ZTB distribution, and $\alpha = 4$ for the ZTNB distribution.

N	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$
2	39.963	22.457	34.356	0.004	34.719
4	9.952	4.399	8.005	0.005	8.258
8	2.505	0.998	1.956	0.020	2.098
16	0.625	0.239	0.482	0.016	0.557
64	0.043	0.018	0.032	0.013	0.055
256	0.005	0.005	0.004	0.011	0.016
1024	0.004	0.005	0.003	0.011	0.016

Table 8.5: Mean squared error (multiplied by 10^3) for the five estimators from ZTBinomial data with varying N. 100 samples of size 250,000 were collected.

α	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$
1	160.162	29.355	99.525	99.611	0.032
5	6.430	1.989	4.669	4.682	0.018
20	0.402	0.141	0.299	0.302	0.025
50	0.064	0.026	0.049	0.050	0.016
70	0.037	0.019	0.029	0.033	0.017
100	0.020	0.012	0.016	0.023	0.010

Table 8.6: Mean squared error (multiplied by 10^3) for the five estimators from ZTNegative Binomial data with varying α. 100 samples of size 250,000 were collected.

Distribution		MSE $\times 10^3$				
Collision	Family	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$
0.1%	ZTP	0.004	0.005	0.003	0.012	0.010
	ZTB	5.743	2.877	4.674	1.261	5.346
	ZTNB	10.249	3.065	7.361	7.378	0.023
0.5%	ZTP	0.022	0.013	0.018	0.022	0.012
	ZTB	5.786	2.616	4.646	0.504	4.849
	ZTNB	11.027	3.217	7.861	7.879	0.035
1%	ZTP	0.065	0.033	0.051	0.055	0.017
	ZTB	5.677	2.465	4.528	0.177	4.729
	ZTNB	12.110	3.578	8.630	8.649	0.025
2%	ZTP	0.252	0.121	0.199	0.201	0.019
	ZTB	5.681	2.497	4.550	0.136	4.752
	ZTNB	14.309	4.134	10.086	10.107	0.028

Table 8.7: Mean squared error (multiplied by 10^3) for the five estimators from data with varying levels of collisions. 100 samples of size 250,000 were collected. True generating parameters are $\lambda = 0.4$, $N = 4$, and $\alpha = 4$.

Distribution		MSE $\times 10^3$				
λ	Family	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$
0.1	ZTP	0.029	0.001	0.003	0.003	3.185
	ZTB	0.002	0.032	0.022	0.022	3.609
	ZTNB	0.919	0.482	0.643	0.645	2.089
0.4	ZTP	0.024	0.005	0.004	0.005	0.332
	ZTB	0.422	0.238	0.430	0.399	0.554
	ZTNB	10.859	2.962	7.376	7.393	0.140
1.5	ZTP	0.016	0.051	0.010	0.011	0.022
	ZTB	8.323	2.705	2.815	0.016	3.277
	ZTNB	142.149	18.627	41.774	41.862	0.022

Table 8.8: Mean squared error (multiplied by 10^3) for the five estimators from three data generating processes with corruption ($N_0 = 16$ and $\alpha = 4$). 500 samples of size 250,000 were collected.

8.4 Application to Sertoli Cells

We apply our five $\hat{\lambda}$ estimators to the additional 20 cells (10 medium and 10 median) described in Section 7.2. Figures 8.8 and 8.9 are analogous to Figure 4.2 for the medium and median cells, respectively. Note that the medium and median cells appear to have similar relationships as the top 10 cells between the $\hat{\lambda}$ estimates of rUMI and the proportion GC or expression level of a gene. We also plot the corresponding $\hat{\pi}$ estimates in Figures 8.10 to 8.12.

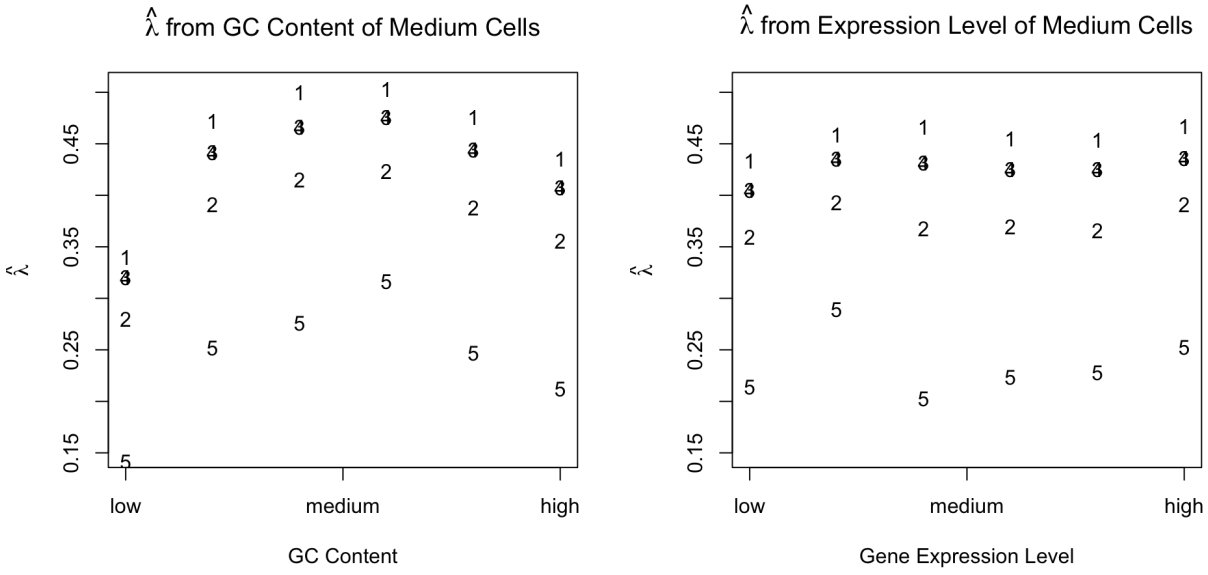

Figure 8.8: Estimated $\hat{\lambda}$ based on the proportion GC content in the transcript (left panel) and expression level of the gene as calculated from the 10 cells (right panel) for the medium cells. Compare to Figure 4.2.

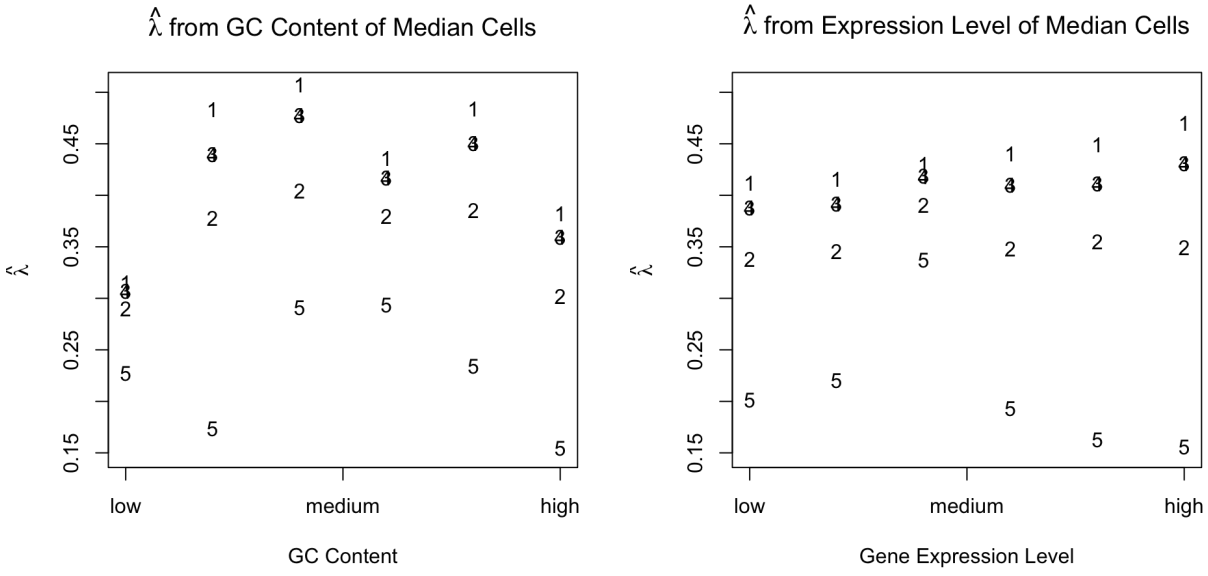

Figure 8.9: Estimated $\hat{\lambda}$ based on the proportion GC content in the transcript (left panel) and expression level of the gene as calculated from the 10 cells (right panel) for the median cells. Compare to Figure 4.2.

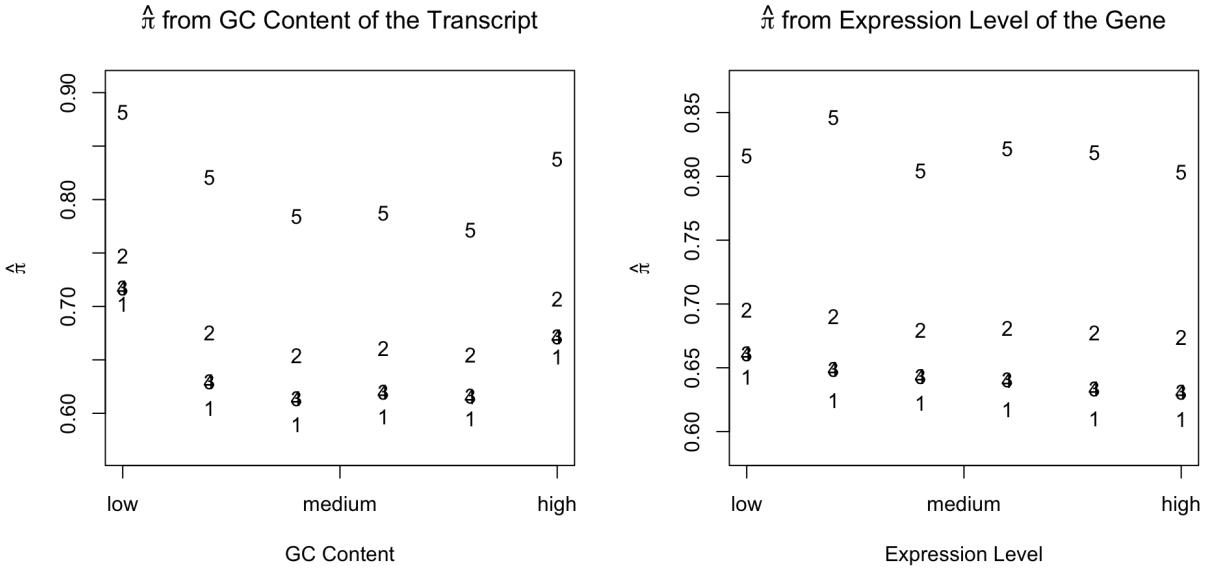

Figure 8.10: Estimated $\hat{\pi}$ corresponding to the $\hat{\lambda}$ estimates in Figure 4.2 based on the proportion GC content in the transcript (left panel) and expression level of the gene as calculated from the 10 cells (right panel) for the highly expressed cells.

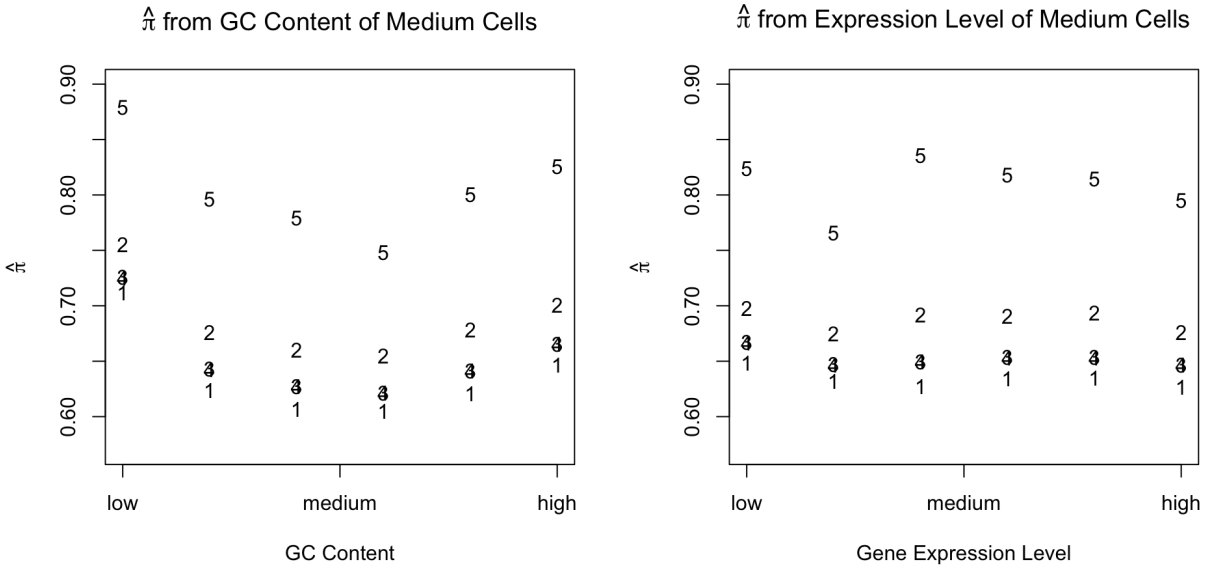

Figure 8.11: Estimated $\hat{\pi}$ corresponding to the $\hat{\lambda}$ estimates in Figure 8.8 based on the proportion GC content in the transcript (left panel) and expression level of the gene as calculated from the 10 cells (right panel) for the medium cells.

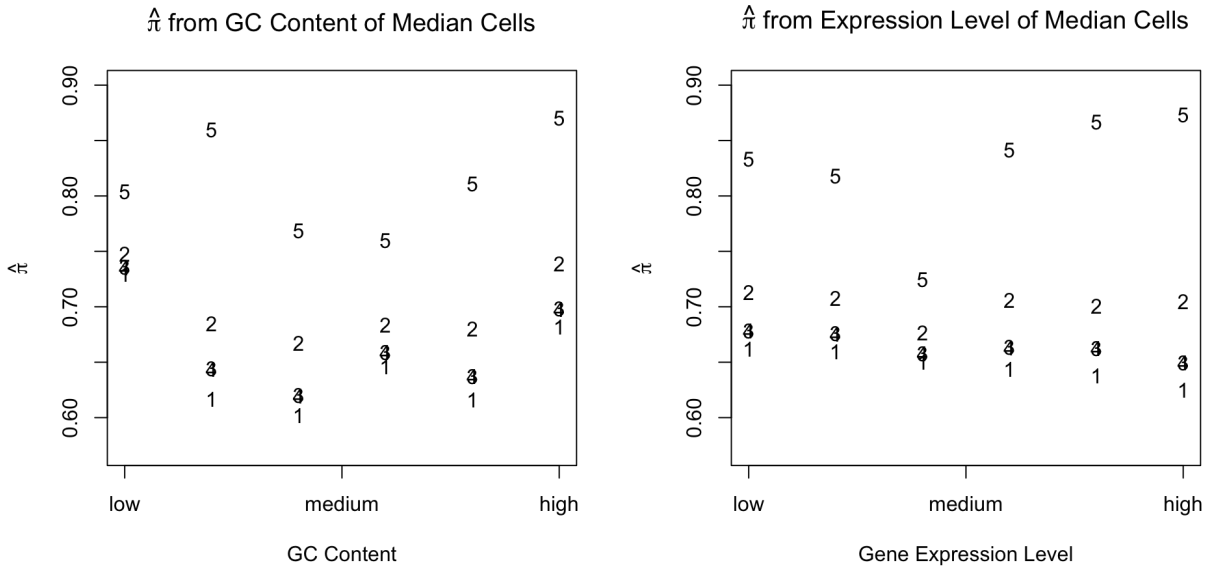

Figure 8.12: Estimated $\hat{\pi}$ corresponding to the $\hat{\lambda}$ estimates in Figure 8.9 based on the proportion GC content in the transcript (left panel) and expression level of the gene as calculated from the 10 cells (right panel) for the median cells.

Bibliography

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, **12**(2).

Arguel, M.-J., LeBrigand, K., Paquet, A., RuizGarcía, S., Zaragosi, L.-E., Barbry, P., and Waldmann, R. (2017). A cost effective 5‘ selective single cell transcriptome profiling approach with improved UMI design. *Nucleic Acids Research*, **45**(7), e48–e48.

Bacher, R. and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology*, **17**(63).

Baker, S. C., Bauer, S. R., Beyer, R. P., Brenton, J. D., Bromley, B., Burrill, J., Causton, H., Conley, M. P., Elespuru, R., Fero, M., Foy, C., Fuscoe, J., Gao, X., Gerhold, D. L., Gilles, P., Goodsaid, F., Guo, X., Hackett, J., Hockett, R. D., Ikonomi, P., Irizarry, R. A., Kawasaki, E. S., Kaysser-Kranich, T., Kerr, K., Kiser, G., Koch, W. H., Lee, K. Y., Liu, C., Liu, Z. L., Lucas, A., Manohar, C. F., Miyada, G., Modrusan, Z., Parkes, H., Puri, R. K., Reid, L., Ryder, T. B., Salit, M., Samaha, R. R., Scherf, U., Sendera, T. J., Setterquist, R. A., Shi, L., Shippy, R., Soriano, J. V., Wagar, E. A., Warrington, J. A., Williams, M., Wilmer, F., Wilson, M., Wolber, P. K., Wu, X., and Zadro, R. (2005). The External RNA Controls Consortium: a progress report. *Nature Methods*, **2**(10), 731–734.

Barber, R. D., Harmer, D. W., Coleman, R. A., and Clark, B. J. (2005). Gapdh as a housekeeping gene: analysis of gapdh mrna expression in a panel of 72 human tissues. *Physiological genomics*.

Benjamini, Y. and Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic acids research*, **40**(10), 1–14.

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., and Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, **10**(11), 1093–1095.

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

Carithers, L. J., Ardlie, K., Barcus, M., Branton, P. A., Britton, A., Buia, S. A., Compton, C. C., DeLuca, D. S., Peter-Demchok, J., Gelfand, E. T., Guan, P., Korzeniewski, G. E., Lockhart, N. C., Rabiner, C. A., Rao, A. K., Robinson, K. L., Roche, N. V., Sawyer, S. J., Segrè, A. V.,

Shive, C. E., Smith, A. M., Sobin, L. H., Undale, A. H., Valentino, K. M., Vaught, J., Young, T. R., and Moore, H. M. (2015). A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking*, **13**(5), 311–319.

Chen, M. and Zhou, X. (2016). Controlling for confounding effects in single cell RNA sequencing studies using both control and target genes. *bioRxiv*.

Chen, M. and Zhou, X. (2017). Controlling for Confounding Effects in Single Cell RNA Sequencing Studies Using both Control and Target Genes. *Scientific Reports*, **7**(1), 13587.

Dabney, J. and Meyer, M. (2012). Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques*, **52**(2).

Das, A., Rouault-Pierre, K., Kamdar, S., Gomez-Tourino, I., Wood, K., Donaldson, I., Mein, C. A., Bonnet, D., Hayday, A. C., and Gibbons, D. L. (2017). Adaptive from Innate: Human IFN-γ^+ CD4$^+$ T Cells Can Arise Directly from CXCL8-Producing Recent Thymic Emigrants in Babies and Adults. *The Journal of Immunology*, **199**(5), 1696–1705.

De Jonge, H. J., Fehrmann, R. S., de Bont, E. S., Hofstra, R. M., Gerbens, F., Kamps, W. A., de Vries, E. G., van der Zee, A. G., te Meerman, G. J., and ter Elst, A. (2007). Evidence based selection of housekeeping genes. *PloS one*, **2**(9), e898.

Dijk, D. V., Nainys, J., Sharma, R., Kaithail, P., and Carr, A. J. (2017). MAGIC : A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *bioRxiv*.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, **30**(1), 207–10.

Eisenberg, E. and Levanon, E. Y. (2003). Human housekeeping genes are compact. *Trends in Genetics*, **19**(7), 362–365.

Eisenberg, E. and Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends in Genetics*, **29**(10), 569–574.

Fukaya, T., Lim, B., and Levine, M. (2016). Enhancer Control of Transcriptional Bursting. *Cell*, **166**(2), 358–368.

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

Gagnon-Bartsch, J. A., Jacob, L., and Speed, T. P. (2013). Removing unwanted variation from high dimensional data with negative controls.

Green, C. D., Ma, Q., Manske, G. L., Shami, A. N., Zheng, X., Marini, S., Moritz, L., Sultan, C., Gurczynski, S. J., Moore, B. B., *et al.* (2018). A comprehensive roadmap of murine spermatogenesis defined by single-cell rna-seq. *Developmental cell*, **46**(5), 651–667.

Grün, D. and vanOudenaarden, A. (2015). Design and Analysis of Single-Cell Sequencing Experiments. *Cell*, **163**(4), 799–810.

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nature Methods*, **11**(6), 637–640.

Hicks, S. C., Teng, M., and Irizarry, R. A. (2015). On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *bioRxiv*.

Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, **19**(4), 562–578.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J., Raj, A., Li, M., and Zhang, N. R. (2017). Gene Expression Recovery For Single Cell RNA Sequencing. *bioRxiv*.

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., and Teichmann, S. A. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, **17**(1), 29.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2), 249–264.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex rna-seq. *Genome research*, **21**(7), 1160–1167.

Islam, S., Zeisel, A., Joost, S., Manno, G. L., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, **11**(2), 163–166.

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, pages 1543–1551.

Jiang, Y., Zhang, N. R., and Li, M. (2017). SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biology*, **18**(1), 74.

Kim, J. K., Kolodziejczyk, A. A., Illicic, T., Teichmann, S. A., and Marioni, J. C. (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nature Communications*, **6**.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, **9**(1), 72–74.

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

Lin, Y., Ghazanfar, S., Strbenac, D., Wang, A., Patrick, E., Lin, D. M., Speed, T., Yang, J. Y., and Yang, P. (2019a). Evaluating stably expressed genes in single cells. *GigaScience*, **8**(9), giz106.

Lin, Y., Ghazanfar, S., Wang, K. Y., Gagnon-Bartsch, J. A., Lo, K. K., Su, X., Han, Z.-G., Ormerod, J. T., Speed, T. P., Yang, P., *et al.* (2019b). scmerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell rna-seq datasets. *Proceedings of the National Academy of Sciences*, **116**(20), 9775–9784.

Love, M. I., Hogenesch, J. B., and Irizarry, R. A. (2016). Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature Biotechnology*, **34**(12), 1287–1291.

Lun, A. T. L. and Marioni, J. C. (2017). Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics (Oxford, England)*, **18**(3), 451–464.

Lun, A. T. L., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, **17**(1), 75.

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

Nemesh, J. (2015). Drop-seq Core Computational Protocol.

Oyolu, C., Zakharia, F., and Baker, J. (2012). Distinguishing human cell types based on housekeeping gene signatures. *Stem Cells*, **30**(3), 580–584.

Phipson, B., Zappia, L., and Oshlack, A. (2017). Gene length and detection bias in single cell RNA sequencing protocols. *F1000Research*, **6**, 595.

Pierson, E. and Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, **16**.

Pine, P. S., Munro, S. A., Parsons, J. R., McDaniel, J., Lucas, A. B., Lozach, J., Myers, T. G., Su, Q., Jacobs-Helber, S. M., and Salit, M. (2016). Evaluation of the External RNA Controls Consortium (ERCC) reference material using a modified Latin square design. *BMC Biotechnology*, **16**(1), 54.

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, **32**(9), 896–902.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2017). ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data. *bioRxiv*.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**.

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., *et al.* (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, **498**(7453), 236.

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature reviews. Genetics*, **16**, 133–145.

Suter, D. M., Molina, N., Gatfield, D., Schneider, K., Schibler, U., and Naef, F. (2011). Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. *Science*, **332**(6028), 472–474.

Thorrez, L., Van Deun, K., Tranchevent, L.-C., Van Lommel, L., Engelen, K., Marchal, K., Moreau, Y., Van Mechelen, I., and Schuit, F. (2008). Using ribosomal protein genes as reference: a tale of caution. *PloS one*, **3**(3), e1854.

Tung, P.-Y., Blischak, J. D., Hsiao, C. J., Knowles, D. A., Burnett, J. E., Pritchard, J. K., and Gilad, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. *Scientific reports*, **7**, 39921.

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell rna sequencing data: challenges and opportunities. *Nature methods*, **14**(6), 565.

Wang, J., Zhao, Q., Hastie, T., and Owen, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *Ann. Statist.*, **45**(5), 1863–1894.

Wu, C., MacLeod, I., and Su, A. I. (2013). BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Research*, **41**(D1), D561–D565.

Xin, J., Mark, A., Afrasiabi, C., Tsueng, G., Juchler, M., Gopal, N., Stupp, G. S., Putman, T. E., Ainscough, B. J., Griffith, O. L., Torkamani, A., Whetzel, P. L., Mungall, C. J., Mooney, S. D., Su, A. I., and Wu, C. (2016). High-performance web services for querying gene and variant annotation. *Genome Biology*, **17**(1), 91.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., *et al.* (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications*, **8**, 14049.