UC Santa Barbara

UC Santa Barbara Electronic Theses and Dissertations

Title

Bayesian Quantification of Gene-Level Post Transcriptional Regulation Using High Throughput Multi Modal Measurements

Permalink

https://escholarship.org/uc/item/83x5217x

Author

Elcheikhali, Megan Kate

Publication Date

2023

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

University of California
Santa Barbara

Bayesian Quantification of Gene-Level Post Transcriptional Regulation Using High Throughput Multi Modal Measurements

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Statistics and Applied Probability

by

Megan Kate Elcheikhali

Committee in charge:

Professor Alexander Franks, Chair
Professor Wendy Meiring
Professor Sang-Yun Oh

June 2023

The Dissertation of Megan Kate Elcheikhali is approved.

Professor Wendy Meiring

Professor Sang-Yun Oh

Professor Alexander Franks, Committee Chair

June 2023

Bayesian Quantification of Gene-Level Post Transcriptional Regulation Using High Throughput Multi Modal Measurements

Copyright © 2023

by

Megan Kate Elcheikhali

This dissertation is dedicated to my family and loved ones. Mom and Dad, thank you for making it possible for me to focus on my education, teaching me to work hard to achieve my goals, and always cheering me on. Kelly, thank you for all of your care, encouragement, and friendship. Levi, thank you for joining me through this exciting and challenging journey, and for all that you do to support me in pursuing my degree. Erika, your kindness and brilliance are an inspiration, thank you so much for everything. I love all of you so much and could not ask to have more wonderful people in my life to look up to. To my aunties, uncles, Grandma, Grandpa, cousins, Mo and Mo, Chloe, Koenig family, Elliana and Pierce, I love you all and am so grateful to have you all in my life.

Acknowledgements

Thank you to all of the people who have made this accomplishment possible. Professor Alex Franks, thank you for all of your guidance and support in navigating this experience. Thank you for encouraging me to evaluate challenges in terms of the opportunities they create for learning and growth, and to think critically through every step of analysis. Thank you to my committee members, Professor Sang-Yan Oh and Professor Wendy Meiring for all your expertise and fantastic teaching. Thank you Professor Slavov and the entire Slavov lab for teaching me so much about this application area and for the terrific community. A very special thank you to all of my amazing friends and colleagues within the PSTAT department, including Matthew McDonough, Chao Zhang, Yi Zheng, Erika McPhillips, Ethan Marzban, Blythe King, Caroline Yovanovich, Myranda Flores, and many more.

Curriculum Vitæ

Megan Kate Elcheikhali

Education

2023	Ph.D. in Statistics and Applied Probability (Expected), University of California, Santa Barbara.
2020	M.A. in Applied Statistics, University of California, Santa Barbara.
2017	B.S. in Statistics, University of Georgia.

Publications

Abstract

Bayesian Quantification of Gene-Level Post Transcriptional Regulation Using High Throughput Multi Modal Measurements

by

Megan Kate Elcheikhali

Proteins represent a major component of nearly all biological processes, including metabolic reactions, cellular structure, and immune function. It is of interest to understand the utility of mRNA as proxy for protein abundance, as mRNA transcription represents a critical stage in the process of protein synthesis, and can be quantified with fewer technical challenges relative to measuring the proteome; however, its evaluation requires quantification of the role that post transcriptional regulation plays in protein synthesis, and the degree to which it influences relative protein abundance within individual genes. In this dissertation we explore the effect of post transcriptional regulation using high throughput mRNA and protein measurements, collected for individual tissues and cells, by estimating the relative protein-to-mRNA ratio (rPTR). Analysis of high throughput mRNA and protein measurements requires careful modeling of noise and an understanding of technical factors that contribute to measurement bias. We estimate rPTR using a Bayesian hierarchical model that directly models technical factors, relevant biological processes, and sampling variability. This model presents a useful approach to estimate rPTR at the gene level and identify genes and functional sets of genes that have extreme rPTR in measured tissues and cell types. Finally we evaluate the effect of gene-level technical factors, including batch effect and sequencing depth, and their influence of rPTR estimation by applying our Bayesian model to adjusted transcript counts using two technical-effect correction methods.

Contents

Curriculum Vitae **vi**

Abstract **vii**

1 Introduction **1**

2 Bayesian Inference of Post Transcriptional Regulation Across Tissues with Bulk mRNA and Protein Data **5**
2.1 Introduction 5
2.2 Methods 7
2.3 Results 19
2.4 Posterior Predictive Checking 30
2.5 Discussion 33

3 Bayesian Inference of Post Transcriptional Regulation Across Cell Types **35**
3.1 Introduction 35
3.2 Methods 38
3.3 Results 62
3.4 Posterior Predictive Checking 75
3.5 Discussion 80

4 Assessing Sensitivity to Choice of Technical-Effect Correction Methodology **82**
4.1 Introduction 82
4.2 Review of Batch and Technical Effect Correction Methods 84
4.3 Methods 89
4.4 Results 93
4.5 Posterior Predictive Checking 97
4.6 Leave-one-out cross validation 99
4.7 Discussion 100

5 Discussion **102**

A Appendix Chapter 2 **104**
A.1 Identifying tissues with extreme rPTR 104
A.2 Additional Plots 106

B Appendix Chapter 3 **109**
B.1 Model selection 109
B.2 Additional gene-level results 112
B.3 Additional results from GO Testing 115

C Appendix Chapter 4 **117**
C.1 Additional ComBat-seq plots 117
C.2 Additional SCTransform plots 122
C.3 Additional model comparison results 126

Bibliography **128**

Chapter 1

Introduction

Proteins play a crucial role in biological processes and are central to various cellular functions. From immune response to structural function and gene regulation, proteins are involved in nearly all aspects of molecular functions, and their dysregulation is linked to several diseases, including cancers, neurodegenerative diseases and genetic disorders [1]. Through understanding proteins and their synthesis, we can examine biological processes at the molecular level.

Protein synthesis is the process by which proteins are created from information contained in DNA. The first step of this process is transcription, where the information contained in a gene is transcribed into messenger RNA (mRNA) transcripts. Next the mRNA undergoes specific modifications, including post-transcriptional regulation, which influences the stability, location and accessibility of mRNA. After this point translation takes place, where transcripts are read by ribosomes in the cytoplasm and amino acids are assembled into polypeptide chains based on this information. After the polypeptide chain is released by the ribosome, protein folding ultimately yields the protein of interest, and chemical alterations are applied to the synthesized protein, called post translational modifications, which affect the chemical properties, activity and stability of the protein.

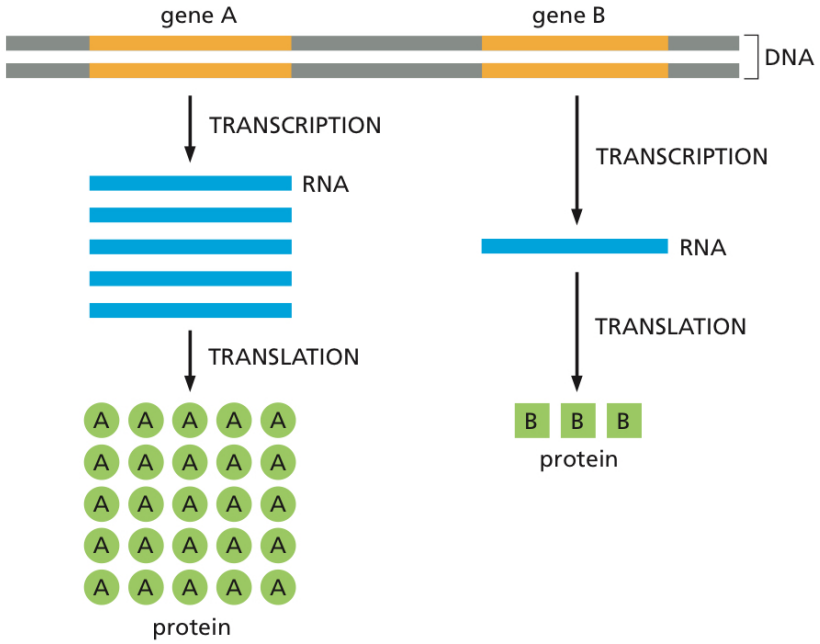

Figure 1.1: Depiction of protein synthesis as a product of transcription and translation from [1]. The amount of a particular protein that is synthesized corresponds to the amount of mRNA transcribed.

The quantification of gene products makes it possible to get a glimpse into biological processes in tissues and cells. mRNA sequencing involves counting transcripts associated with each gene in tissues or cells. mRNA sequencing was developed prior to the measurement of peptide intensity levels [2] [3], where peptide abundances are quantified instead of transcripts, and is less costly and easier to implement than peptide data collection [4]. This has resulted in mRNA often being used as a proxy to examine protein abundance [5] [6].

The utility of mRNA as a proxy for protein abundance depends on the degree of post transcriptional regulation in a given gene, as well as the capability of mRNA to reflect the variability of protein abundance within that gene across tissues or cells. By measuring both transcriptomics and proteomics data, we can quantify the effect of post

transcriptional regulation. This takes place by estimating the relative protein-to-mRNA ratio (rPTR), the ratio of relative mRNA and protein measurements to each of their respective gene-level averages, and the mRNA, protein correlation across tissues, cell types, and developmental stages. If a gene is regulated mostly transcriptionally, the rPTR will be close to 0, with little variability within a gene, and the mRNA, protein correlation across cell types or tissues will be high. If a gene is regulated mostly post-transcriptionally, rPTR will be extreme for some cell types or tissues within a gene, have high variability, and the resulting mRNA, protein correlation will be low.

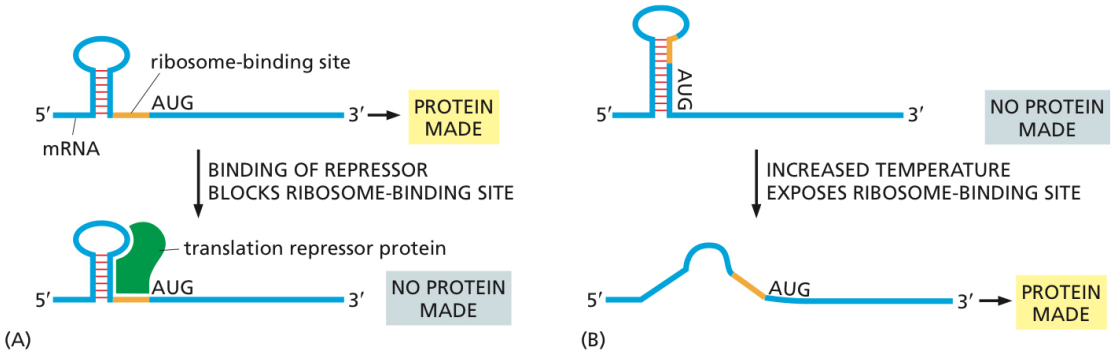

Figure 1.2: Examples of post transcriptional regulation being used to control protein synthesis from [1] In figure (A) protein translation is repressed as a result of the attachment of a sequence-specific binding protein that blocks a binding site necessary for translation. In figure (B) we can see that an increase in temperature exposes the ribosome binding site, making it possible for translation to be made possible.

In this dissertation, we quantify the effect of post transcriptional regulation by estimating rPTR for individual genes using bulk and single cell mRNA and protein data. We refer to modifications taking place after transcription and translation as "post transcriptional" regulation, as both can yield disparities between the amount of mRNA transcribed and protein synthesized. We use a Bayesian model that shares information across replicates to disentangle variability contributed by underlying biological processes and noise. In chapter 2, we apply this model to bulk transcriptomics and proteomics mea-

surements, and identify genes and functional sets of genes that may have elevated post transcriptional regulation. Bulk transcript sequencing and peptide quantification measures gene-level abundance of mRNA and protein, respectively, across all cells within a given tissue sample, without delineating between cells of different types and developmental stages. In chapter 3, we implement a version of this model adapted for single-cell mRNA and peptide measurements. Single cell measurements quantify the abundance of gene products at the cellular level, offering a glimpse into the transcriptome and proteome with unprecented resolution [7]. In chapter 4, we review approaches to correct for technical factors contributing bias and unwanted variability, including batch effect and sequencing depth. We apply two technical-effect correction methods, SCTransform from Seurat V2 [8] and ComBat-seq [9] to the single-cell model applied in chapter 3, and compare results of these fits to that of the model applied directly to observed data.

Chapter 2

Bayesian Inference of Post Transcriptional Regulation Across Tissues with Bulk mRNA and Protein Data

2.1 Introduction

Bulk mRNA and protein profiling allows for the quantification of gene products across cells in tissues for thousands of genes. During bulk mRNA sequencing, transcripts are counted for each gene across cells in tissue samples [10]. Protein intensity levels are measured separately using bulk proteomics, by breaking proteins into smaller fragments of peptides and measuring their abundance across cells in a given tissue sample [11]. As quantifying the proteome is newer and its measurements are more challenging to collect on a technical level, mRNA observations are often used as a proxy to examine cellular functions and how they're affected within healthy tissue, medical intervention or in the

event of disease [6]; however, several biological functions depend on the proteins later synthesized after mRNA transription. While the amount of a particular protein that is synthesized is often proportional to the amount of mRNA transcribed, post transcriptional regulation can often affect ultimate protein abundances. Some examples of post transcriptional regulation being carried out within cells include MicroRNA-Mediated Regulation, where small RNA molecules bind to mRNA such that expression is reduced or the mRNA begins to degrade, and RNA localization, the process where mRNA is transported to specific areas within the cell to control the synthesis of proteins at a spatial level [12] [13]. Additionally, post transcriptional regulation can affect the function of genes associated with disease. For example, dysregulation of MicroRNAs is associated with disease including cancer and neurodevelopmental disorders [12]. Thus, quantifying post transcriptional regulation at the gene level has implications with respect to improved understanding of the relationship between mRNA and protein and its analysis, and can reveal information about the process of post transcriptional regulation, a critical cellular process that is instrumental in this relationship.

Analyzing bulk measurements has specific technical challenges as a result of the technical and biological nuances involving data collection. This includes a large degree of measurement noise, as quantification of gene products takes place at just a "snap shot" in time for a collective set of cells [14], and this measurement noise can attenuate across-tissues correlations between mRNA and protein, making it more difficult to separate genes with elevated post transcriptional regulation from those that are simply measured with increased noise [15]. Additionally, missing data can restrict the view of the behavior of transcripts and proteins at the tissue level, a result of technical challenges associated with measurement and genes that are truly not expressed in a given sample [16].

In this chapter we take a Bayesian approach to the analysis in [17] to quantify the effect of post transcriptional regulation at the gene and tissue level using the relative

protein-to-mRNA ratio (rPTR). This approach offers direct modeling of noise, and allows us to leverage a hierarchical structure that is flexible enough to take into account heterogeneity associated with underlying gene product expression. Extreme rPTR, or rPTR with extreme variability across cell types, can reduce the across cell types correlation of mRNA and protein and potentially indicate increased post transcriptional regulation. This model utilizes technical replicates across two bulk mRNA and two bulk protein data sets to estimate the underlying $\log_2$ transcription rate and rPTR in each gene and tissue. This model disentangles variability contributed by underlying mRNA and protein abundance and sampling variability, and accounts for technical measurement biases associated with each data set.

2.2 Methods

2.2.1 Data Sources

All data sets used in this analysis are based on bulk mRNA and protein measurements. RNA-seq data are collected by lysing cells such that the contents in each cell is released, and mRNA is isolated [18] [19]. The mRNA is used to transcribe complementary DNA (cDNA) using reverse transcriptase, amplifying the molecules of mRNA. After some additional preocessing, the cDNA is sequenced to reflect mRNA measurements across all cells in a sample [10]. Protein measurements from [5] [20], are collected by a process similar to that of mRNA sequencing. Specific preparation is applied to lyse cells and release proteins. These proteins are digested using enzymes into smaller peptide fragments. Next, mass spectrometry takes place, wherein peptides are ionized and their mass-to-charge ratio and intensity are measured [11]. A total of nine tissues shared among these data sets are considered, where samples from multiple humans are

sampled for data collection. Tissues used in analysis are listed in table 2.1. As bulk approaches are capable of measuring a large subset of the transcriptome and proteome, measurements for over 6000 gene products are recorded. Resulting mRNA measurements are profiled transcript counts across cells within each tissue. For protein measurements, we have protein intensity levels profiled across cells associated with each tissue.

2.2.2 Preprocessing

To ensure comparison across genes, which may have drastically different average transcript and protein abundance across tissues, it is often necessary to perform some sort of scaling. In this analysis, we Z-transform $\log_2$ measurements across tissues in each gene, by subtracting the across-tissues mean and dividing by the standard deviation. While this preprocessing approach ensures that values are comparable across genes, sources of variability associated with underlying molecule abundance and noise can potentially be confounded. In figure 2.1 we can see Z-transformed observations for each tissue in one of the mRNA data sets [18]. To see transformed values for the remaining mRNA and protein data sets, please see the Appendix.

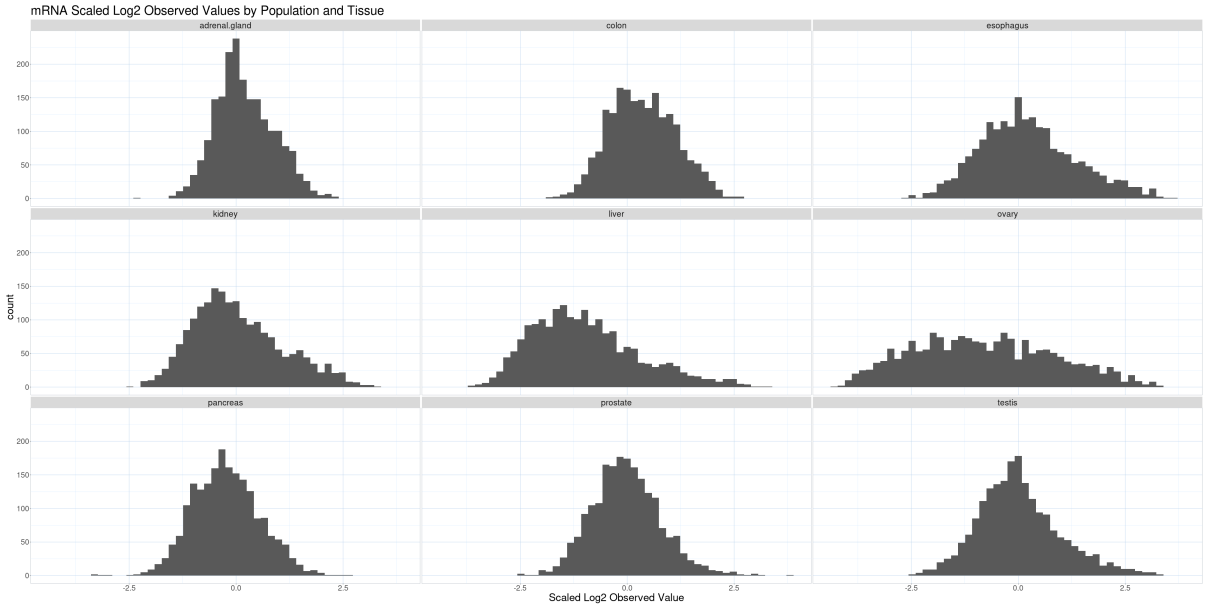

Figure 2.1: Z-transformed $\log_2$ transcript counts in mRNA data set [18] in each tissue considered in analysis.

2.2.3 Missing Data

Relative to single cell sequencing, bulk measurements have less missing data, as measurements are collected across cells in entire samples [21]. In tables 2.1 and 2.2 we can see the proportion of missing tissue, gene pairs in each mRNA and protein data set respectively, which can be missing as a result of various factors, such as sequencing errors, low expression levels, or inefficient capture of certain transcripts during library preparation [22]. While several imputation methods exist for bulk data, these methods vary with respect to treatment of different sources of variability [23] [24] and may not improve performance of downstream analysis [25].

Application of our Bayesian model assumes missingness at random (MAR), meaning that mechanisms guiding the probability that an observation is missing can be explained by data observed in our experiment [26]. This assumption can be somewhat unreasonable given the influence that abundance has on some of the different causes of missingness.

For computational simplicity and to make the MAR assumption more reasonable, we restrict analysis to genes that are less subject to missingness, with observed values in all 9 tissues in at least one mRNA and one protein data set. This allows for information within the same gene and across data sets to be shared.

Table 2.1: Proportion of missing observations in each mRNA data set and tissue. Observations are considered missing if a gene-tissue pair has 0 transcripts counted, but is observed in at least one other tissue.

Tissue	**mRNA Data Set 1**	**mRNA Data Set 2**
Adrenal Gland	0.015	0.021
Colon	0.011	0.005
Esophagus	0.020	0.011
Kidney	0.008	0.003
Liver	0.014	0.011
Ovary	0.039	0.017
Pancreas	0.018	0.018
Prostate	0.011	0.009
Testis	0.001	0.002

Table 2.2: Proportion of missing observations in each protein data set and tissue. Observations are considered missing if a gene-tissue pair has 0 transcripts counted, but each gene is observed in at least one other tissue.

Tissue	**Protein Data Set 1**	**Protein Data Set 2**
Adrenal Gland	0.002	0.006
Colon	0.003	0.009
Esophagus	0.013	0.008
Kidney	0.003	0.003
Liver	0.005	0.008
Ovary	0.001	0.003
Pancreas	0.003	0.008
Prostate	0.005	0.002
Testis	0.003	0.008

2.2.4 Bayesian inference of protein-to-mRNA ratios

To estimate rPTR for individual genes and tissues, we use a Bayesian model that estimates underlying expression-related parameters and takes into account high mea-

surement noise associated with high-throughput mRNA and protein data. We input Z-transformed observed abundances for each gene, tissue and data set. Let M_{mgt} represent Z-transformed $\log_2$ observed mRNA expression level in mRNA data set m, gene g and tissue t. Similarly, define P_{pgt} as the Z-transformed $\log_2$ protein abundance for gene g and tissue t in protein data set p. We assume that observations of the same modality in different data sets within the same tissue and gene are independent replicates. This assumption can be challenging considering that technical conditions are different for each of the data sets, as they are collected independently, although preprocessing performed to ensure that observations are relative to gene-level averages and have scaled across tissues variability makes this more feasible. Additionally, we assume that observed values within the same tissue and gene have shared underlying abundance, and conditional independence of genes given model parameters.

We estimate rPTR, $R_{gt} = \underset{p}{\mathrm{E}}[P_{pgt}] - \underset{m}{\mathrm{E}}[M_{mgt}]$ in each gene g and tissue t, allowing us to identify genes with extreme rPTR in at least one tissue, as well as explore the correlation of relative mRNA expression and protein abundance across tissues. If a gene has high post transcriptional regulation, its corresponding rPTR will be high in some tissues relative to the across-tissues average. As a result, the mRNA, protein correlation across tissues may be low compared to a gene that is regulated transcriptionally. We consider the likelihood and prior distributions for this model, as well as crucial details relating to the application of this model.

Likelihood

mRNA We model the Z-transformed $\log_2$ observed expression associated with each tissue for each gene and data set, M_{mgt} with likelihood,

$$M_{mgt} \sim \text{Normal}(\mu_{gt} + \gamma_{tm}, \tau^2_{\text{mRNA}}) \tag{2.1}$$

where μ_{gt} represents the underlying relative expression for gene g in tissue t. We account for systematic biases between data sets in each tissue using batch effect parameter γ_{tm}. Finally, τ^2_{mRNA} represents sampling variability associated with the mRNA data sets. As there is only one observation per gene-tissue-data set, we are unable to estimate sampling variability separately for each data set, but we accomplish this goal using higher resolution data sets in the following chapters.

Protein We model Z-transformed $\log_2$ protein intensities for each gene, tissue and protein population, P_{pgt}. For protein observations we use likelihood,

$$P_{pgt} \sim \text{Normal}(\mu_{gt} + R_{gt} + \kappa_{tp}, \tau^2_{\text{protein}}). \tag{2.2}$$

As μ_{gt} corresponds to relative mRNA expression, and we expect the amount of protein synthesized to be proportional to the amount of mRNA transcribed, R_{gt} represents the relative protein-to-mRNA ratio (rPTR) on the $\log_2$ scale. This is the estimand for our model, and reflects the degree of post transcriptional regulation relative to the gene-level average. If a gene is regulated mostly transcriptionally, R_{gt} will be close to 0 and will have little variability across tissues.

κ_{tp} performs the same function as that of γ_{tp} by aligning for each tissue and data set, and τ_{protein} represents sampling variability for protein observations, which is also shared across protein data sets.

Priors

We examine choice of prior distribution for each of our parameters. As observations are transformed to be relative to their respective gene-level averages, μ_{gt} and R_{gt} are assigned $\text{Normal}(0, \sigma^2_{\mu_g})$ and $\text{Normal}(0, \sigma^2_{R_g})$ priors, respectively. $\sigma^2_{\mu_g}$, the variance as-

sociated with μ_{gt} for gene g, is assigned a Normal$^+(0, 10)$ prior, and $\sigma^2_{R_g}$, the variance parameter for R_{gt} is assigned the same prior. Additionally, sampling variance parameters τ^2_{mrna} and τ^2_{protein} are assigned Normal$^+(0, 10)$ priors.

Model Details and Inference

In table 2.3, we can see model parameters, as well as their corresponding interpretations and prior distributions. Using these priors and likelihood, we can outline the full posterior distribution corresponding to this model. We assume independent priors for all parameters.

Table 2.3: Model parameters and corresponding prior distributions.

Parameter	Interpretation	Prior Distribution
μ_{gc}	Underlying $\log_2$ mRNA expression rate	Normal$(0, \sigma^2_{\mu_g})$
R_{gc}	Log relative protein-to-mRNA ratio (rPTR)	Normal$(0, \sigma^2_{R_g})$
$\sigma^2_{\mu_g}$	mRNA fold change variance	Normal$^+(0, 10)$
$\sigma^2_{R_g}$	rPTR variance	Normal$^+(0, 10)$
γ_{tm}	Batch effect correction (mRNA)	Normal$(0, 1)$
κ_{tp}	Batch effect correction (protein)	Normal$(0, 1)$
τ_{protein}	Protein sampling variance	Normal$^+(0, 10)$
τ_{mRNA}	mRNA sampling variance	Normal$^+(0, 10)$

Given the model likelihood and priors, we can write the full log likelihood for this model. Given all parameters $\vec{\theta} = [\mu, R, \sigma^2_\mu, \sigma^2_R, \gamma, \kappa, \tau_{\text{protein}}, \tau_{\text{mRNA}}]$ and representing the prior distribution as $p(\vec{\theta})$ the posterior distribution is proportional to,

$$\begin{aligned} p(\vec{\theta}|M, P) &\propto p(\vec{\theta})p(M|\vec{\theta})p(P|\vec{\theta}) \\ &\propto p(\vec{\theta}) \prod_{m=1}^{D_1} \prod_{g=1}^{G_m} \prod_{t=1}^{T_{mg}} p(M_{mgt}|\vec{\theta}) \prod_{p=1}^{D_2} \prod_{g=1}^{G_p} \prod_{t=1}^{T_{pg}} p(P_{pgt}|\vec{\theta}), \end{aligned} \tag{2.3}$$

with D_1 being the number of mRNA data sets and D_2 as the number of protein data

sets. G_m is the total number of genes for which there are transcripts present in mRNA data set m, G_p is the total number genes for which there are observed values in protein data set p, and T_{mg} and T_{pg} are number of tissues observed in each data set and gene. This results in log likelihood,

$$\begin{aligned} l(\vec{\theta}|M,P) &\propto l(\vec{\theta}) + \sum_{m=1}^{D_1}\sum_{g=1}^{G_m}\sum_{c=1}^{T_{mg}} \log(p(M|\vec{\theta})) + \sum_{p=1}^{D_2}\sum_{g=1}^{G_p}\sum_{t=1}^{T_{pg}} \log(p(P|\vec{\theta})) \\ &\propto l(\vec{\theta}) + \sum_{m=1}^{D_1}\sum_{g=1}^{G_m}\sum_{t=1}^{T_{mg}} \left[-\frac{1}{2}\log(2\pi\tau^2_{mRNA}) - \frac{(M_{mgt} - (a_g + \gamma_{pt} + \mu_{gt}))^2}{2\tau^2_{mRNA}} \right] \\ &+ \sum_{p=1}^{4}\sum_{g=1}^{G_p}\sum_{t=1}^{T_{pg}} \left[-\frac{1}{2}\log(2\pi\tau^2_{protein}) - \frac{(P_{pgt} - (b_g + \kappa_{pt} + \mu_{gt} + R_{gt}))^2}{2\tau^2_{protein}} \right]. \end{aligned} \tag{2.4}$$

We fit this model using Stan [27], a probabilistic programming language that performs fully Bayesian inference with No-U-Turn Sampling (NUTS) [28], resulting in posterior draws for all parameters in our model. This model was run using 2000 iterations and 5 chains, using default settings for all other sampling function arguments, including number of warm-up iterations. Sampling results in successful convergence diagnostics, such as maximum $\hat{R} < 1.01$ and 0 divergent transitions. Trace plots for R and μ in randomly selected genes and tissues are located in the Appendix.

2.2.5 Using across tissues mRNA, protein correlation and rPTR to identify genes with elevated post transcriptional regulation

Using posterior draws for μ_{gt} and R_{gt}, we can identify genes with extreme low correlation of μ_{gt} and $\mu_{gt} + R_{gt}$ across tissues, as well as genes with extreme R_{gt} in at least one tissue. As $\log_2$ observed values are Z-transformed before being inputted into the model,

they represent relative intensities and are centered at 0. Thus extreme R_{gt} indicates extreme rPTR and may suggest elevated post transcriptional regulation.

Identifying genes, tissues with nonzero rPTR

Using posterior samples of R, we can identify genes with rPTR significantly different from 0 by comparing R_{gt} for individual genes in each tissue to the across-genes average of R. As R is centered at zero for each gene, we compare R_{igt}, the ith posterior draw of R for gene g, tissue t, to $\bar{R}_{i-gt}$, the average R across all other genes within tissue t for posterior sample i. We define indicator,

$$\mathbb{1}_{ig} = \begin{cases} 1 & \text{if } \operatorname{sgn}(R_{igt}) == \operatorname{sgn}(R_{i-gt}) \\ 0 & \text{otherwise ,} \end{cases} \tag{2.5}$$

and we compute the posterior exclusion probability for gene g, tissue t as the average across draws i. This posterior probability p_{gt}, represents the posterior probability that R_{gt} has the same sign as that of all other genes. We expect that genes with extreme R_{igt} will consistently have a sign different from the average across all genes. We can then order genes according to this posterior probability, and compute the posterior expected proportion of false discoveries by taking the cumulative mean of p_{gt}. We consider a gene to be significant if its posterior expected false discovery proportion is below 0.01.

Across-tissues correlation of mRNA and protein

To compute across-tissues correlations, for each set of posterior draw i, we compute $\rho_{ig} = \operatorname{Cor}(\mu_{igt}, \mu_{igt} + R_{igt} | i, g)$. This is better than using raw data M_{mgt} and P_{pgt} because unwanted technical noise and sampling variability can attenuate the across-tissues correlation [17]. μ and R reflect underlying $\log_2$ mRNA expression and rPTR, respectively

and are estimated using observations for each gene and tissue in different data sets as independent replicates. To test whether our correlation across cell types is significantly lower than that of all other genes, we first define $\rho_{i-g} = \text{Cor}(\bar{\mu}_{i-gt}, \bar{\mu}_{i-gt} + \bar{R}_{i-gt} | i, g)$ to represent the correlation across tissue for the average of μ_{igt} and $\mu_{igt} + R_{igt}$ across all other genes for posterior sample i. we compute indicator $\mathbb{1}_{igt}$,

$$\mathbb{1}_{ig} = \begin{cases} 1 & \text{if } \rho_{ig} \geq \rho_{i-g} \\ 0 & \text{otherwise ,} \end{cases} \tag{2.6}$$

to represent the posterior indicator corresponding to ρ_{ig} being greater than the average across all other genes. We can average this indicator across posterior samples i to compute the posterior exclusion probability, p_g for each gene, and genes with a low across tissues correlation will have a low posterior probability. As these probabilities are defined as posterior exclusion probabilities, we can take the cumulative mean of the ordered vector of probabilities to compute the expected false discovery proportion. We consider genes significant if their false discovery proportion is below 0.01.

2.2.6 Functional gene set analysis

While testing each gene recovers a large set of significant genes, we can group genes together based on the biological processes with which they're involved to test with more statistical power. Gene ontology (GO) groups, are functional sets of genes that perform coordinated biological processes. These groups are predefined based on previous literature [29]. By testing at the GO-level, we're also able to gain insight into biological processes that may be particularly affected by post transcriptional regulation. We perform two types of tests with functional gene sets. First, we compare the mean among genes in each GO group and tissue to that of all other genes within the same tissue, and then we

test the correlation across tissues in each GO groups to that of all other genes.

Identifying groups, tissues with extreme mean rPTR relative to all other genes

To identify GO groups and tissues with extreme rPTR relative to genes belonging to all other GO groups, we directly test posterior samples of R_{gt}. Define R_{iGt} as the average R_{igt} across genes associated with gene set G for tissue t and posterior draw i. Next, let R_{i-Gt} be the average across genes not associated with GO group G. We define indicator $\mathbb{1}_{iGt}$ as,

$$\mathbb{1}_{iGt} = \begin{cases} 1 & \text{if } R_{iGt} \leq R_{i-Gt} \\ 0 & \text{otherwise .} \end{cases} \tag{2.7}$$

We average this indicator across posterior samples i to obtain $p_G^{'}$, the posterior probability that R_{Gt} is less than R_{-Gt}. Using $p_G = \min(p_G^{'}, 1 - p_G^{'})$, we have the posterior exclusion probability for R_{Gt} being more extreme than the average across all other genes, and we can compute the expected proportion of false discoveries as the cumulative mean of p_{Gt} ordered across groups. We consider significant groups as those that fall below the 1% threshold.

Across tissues correlation of mRNA and protein

In addition to identifying genes with extreme rPTR in individual GO groups and tissues, we can examine the correlation of mRNA and protein relative abundance parameters to identify groups with potentially high post transcriptional regulation. As we generally expect the amount of mRNA transcribed in a gene to be proportional to the amount of protein synthesized, groups with genes having extreme rPTR in one or more

tissues will have decreased $\mu, \mu + R$ correlations across tissues. For each GO group, let $\bar{\mu}_{iGt}$ and $\bar{\mu}_{iGt}+\bar{R}_{iGt}$ be the average μ_{igt} and $\mu_{igt}+R_{igt}$, respectively across genes associated with group G for tissue t and posterior sample i. Also define $\bar{\mu}_{i-Gt}$ and $\bar{\mu}_{i-Gt} + \bar{R}_{i-Gt}$ as the average for posterior sample i and tissue t across all genes not associated with GO group G. We define $\rho_{iG} = \text{Cor}(\bar{\mu}_{iGt}, \bar{\mu}_{iGt} + \bar{R}_{iGt}|i, G)$ as the mRNA, protein correlation across tissues in GO group G, and $\rho_{i-G} = \text{Cor}(\bar{\mu}_{i-Gt}, \bar{\mu}_{i-Gt} + \bar{R}_{i-Gt}|i, G)$ as the across tissues correlation corresponding to the average across all genes not associated with gene set G.

We use indicator,

$$\mathbb{1}_{iG} = \begin{cases} 1 & \text{if } \rho_{iG} \geq \rho_{i-G} \\ 0 & \text{otherwise ,} \end{cases} \tag{2.8}$$

to record whether the correlation of $\bar{\mu}_{iGt}$ and $\bar{\mu}_{iGt}+\bar{R}_{iGt}$ is greater than that of all other genes. We compute the average of this indicator across posterior samples i to recover the posterior probability that the correlation of mRNA and protein among genes in each group is higher than that of all other groups. As we are generally interested in gene sets with low correlation, groups with low posterior exclusion probability are significant, where significance is determined by a 1% FDR cutoff, as computed by ordering the posterior probabilities and computing the cumulative average.

2.2.7 rPTR within individual tissues

In addition to examining rPTR at the gene and gene-set levels, we consider tissues that are potentially associated with elevated post transcriptional regulation by testing the average of R_{gt} in each tissue t across genes g. For each posterior draw define $\bar{R}_{it}$ as the average of posterior draws R_{igt} across all genes g in tissue t for posterior draw

i. Next, define $\bar{R}_{i-t}$ as the average of posterior draws R_{ig-t} across all genes and tissues not being tissue t. To compute the posterior probability that $\bar{R}_t$, the mean $\bar{R}_{it}$ across posterior draws i, has the same sign as $\bar{R}_{-t}$, the average across all other tissues, we define indicator,

$$\mathbb{1}_{iGt} = \begin{cases} 1 & \text{if } \text{sgn}(\bar{R}_{it}) == \text{sgn}(\bar{R}_{i-t}) \\ 0 & \text{otherwise .} \end{cases} \tag{2.9}$$

to record instances where the signs of $\bar{R}_{it}$ and $\bar{R}_{i-t}$ are the same. By computing the average of this indicator across posterior draws i, we recover the posterior probability that $\bar{R}_t$ and $\bar{R}_{-t}$ have the same sign. Next, we compute the posterior expected proportion of false discoveries by computing the cumulative mean of this ordered probability across tissues, and consider significance based on a 0.01 false discovery rate cut off. Results for this test are shown in the Appendix.

2.3 Results

2.3.1 Increased across tissues correlation relative to empirical values

Using our Bayesian model to estimate underlying relative mRNA expression and protein abundance, we can recover across-tissues correlations that are more consistent with a central concept of molecular biology, that mRNA must first be transcribed to synthesize protein [30]. As mRNA is transcribed as a part of the pipeline of protein synthesis, we expect most across-tissues correlations to be non-negative. By estimating underlying relative mRNA expression and protein abundance using replicates from multiple data sets, inferred correlations are generally higher, and fewer genes have negative across tissues

correlations. This is in part due to disentangling sampling variability and bias associated with empirical correlations. Figure 2.2 displays inferred and empirical across tissues correlations. The median model-inferred mRNA, protein correlation across tissues across genes is higher than that of empirical correlations, as correlations for most genes have increased.

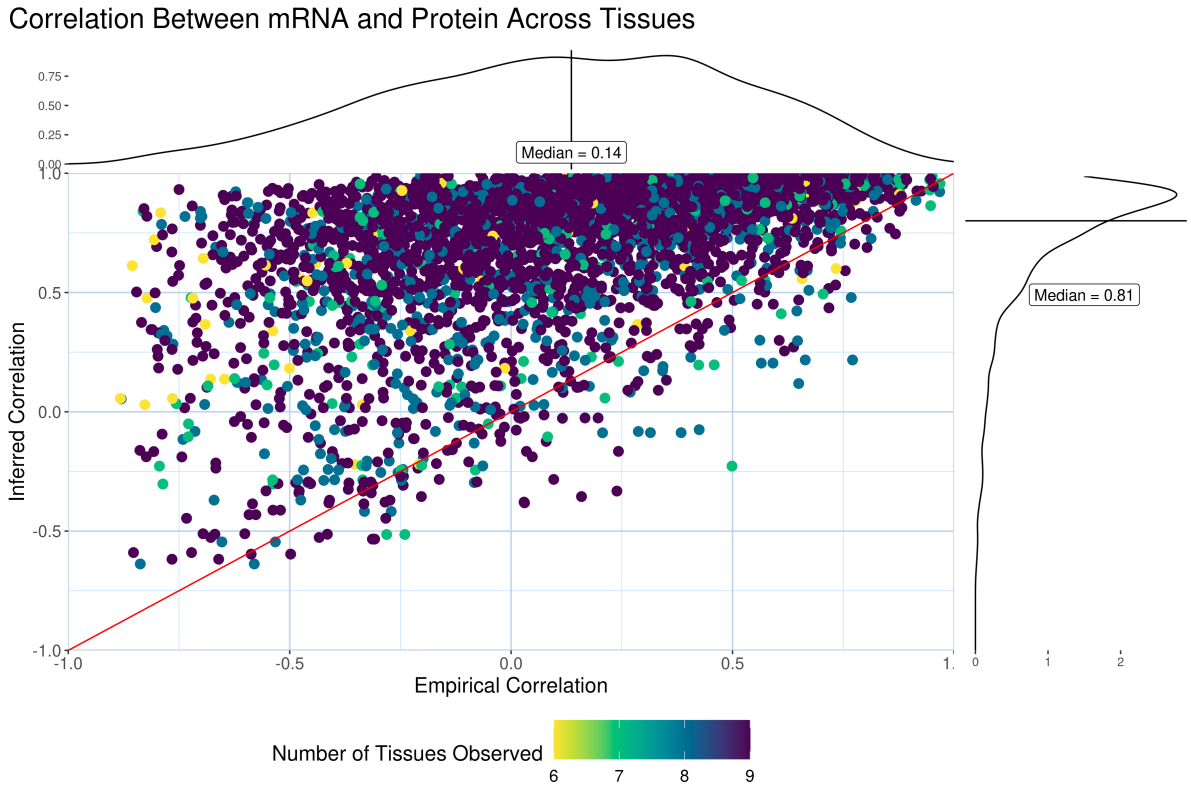

Figure 2.2: Empirical and inferred across tissues correlations using both mRNA data sets and [5]. Inferred correlations are computed using posterior means of μ and $\mu + R$ for each gene and tissue based on the implemented Bayesian model. Observed values are based on Z-transformed $\log_2$ observations. Please see Appendix for an additional plot using the remaining protein data set, [20].

Across tissues correlation of mRNA and protein

Using inferred mRNA, protein correlations, we can identify potentially post transcriptionally regulated genes by testing the across tissues correlation of μ and $\mu + R$.

If R_{gt} has high variability across tissues t, this correlation will be lower than that of transcriptionally regulated genes. For each gene, we compute the posterior probability that the across tissues correlation is higher than that of all other genes, so as to test for genes with extreme low correlation. Additional information on testing procedure can be found in Methods. In table 2.4, we can see sumary information regarding the results of this test.

Table 2.4: Number of significant test results for across tissues correlation of mRNA and protein, computed as the across tissues correlation of μ and $\mu + R$, at 1% and 0.1% expected false discovery proportion.

Significant at 1% **FDR**	**Significant at** 0.1% **FDR**
83	38

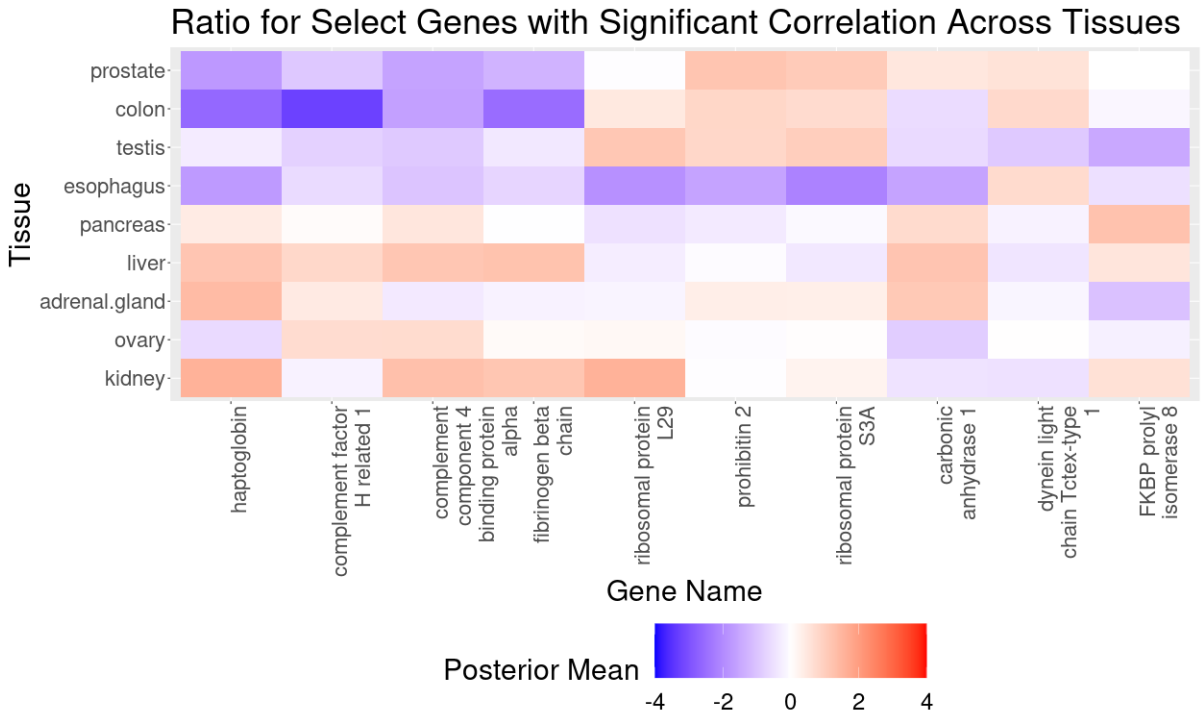

Figure 2.3: Posterior mean of R among select genes with significant low across tissues correlation, shown for each tissue.

In figure 2.3, we can see posterior means for R among select significant genes from this test. Included among significant genes is haptoglobin, a protein made by liver that

binds to dead hemoglobin to be later removed from red blood cells and the bloodstream. Dyfunction of haptoglobin is associated with several diseases and is known to undergo several post transcriptional modifications [31]. Additionally, ribosomal protein S3A, a protein directly related to the synthesis of protein, has significant low correlation [32].

Identifying genes, tissues with nonzero rPTR

In addition to examining post transcriptional regulation using across tissues mRNA, protein correlations, we can simply use our posterior samples to identify genes and tissues with nonzero rPTR. As values are modeled on a $\log_2$ scale and relative to their respective gene-level averages, genes with nonzero rPTR potentially have increased post transcriptional regulation. We test R_{gt} against the average across all other genes by computing the posterior probability that the sign of R_{gt} is the same as that of all other genes across posterior draws. Additional discussion on how this test is implemented is available in Methods. Figure 2.4 displays posterior means of mRNA and protein for each gene in each tissue, with purple displaying genes with nonzero rPTR. We can see that the colon, prostate, esophagus and testis have the highest number of genes with significant rPTR, and some of these tissues, including colon and esophagus, have genes with significant rPTR that is consistently positive or negative. Results for this test are summarized in table 2.5, where the number of significant genes in each tissue at a 1% and 0.1% expected proportion of false discoveries are displayed.

Additionally in figure 2.5 we have posterior means corresponding to select significant genes. Several genes associated with complement factors are significant, including complement factor H related-1, which also had significant low correlation across tissues. The complement system involves plasma proteins that induce inflammatory reactions to fight infection, and its dysregulation is linked with a range of complement deficiency diseases and increased risk for several bacterial infections [33].

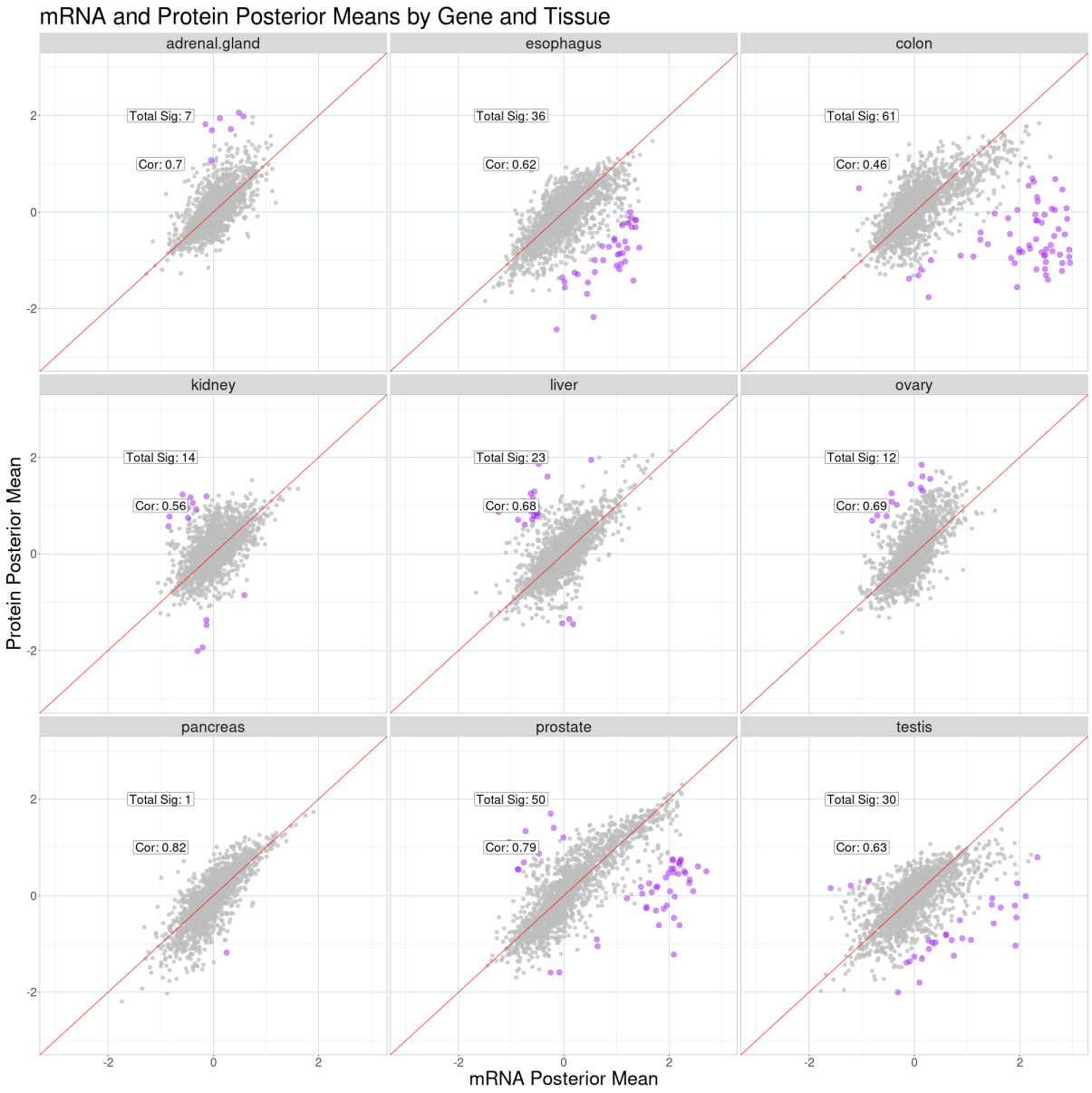

Figure 2.4: Posterior mean of mRNA and protein in each tissue. Gray points correspond to genes that are not significant at 1% FDR cut off, while purple points are significant. Each facet is labeled with the number of signficant genes as well as the mRNA, protein correlation across genes in each tissue.

Table 2.5: Number of significant test results for gene-level comparison of mean of R to all other genes, shown at a 1% and 0.1% FDR cut off for each cell type.

Tissue	**Significant at 1% FDR**	**Significant at 0.1% FDR**
Adrenal Gland	7	0
Esophagus	61	43
Colon	36	10
Kidney	14	1
Liver	23	7
Ovary	12	0
Pancreas	1	0
Prostate	50	15
Testis	30	4

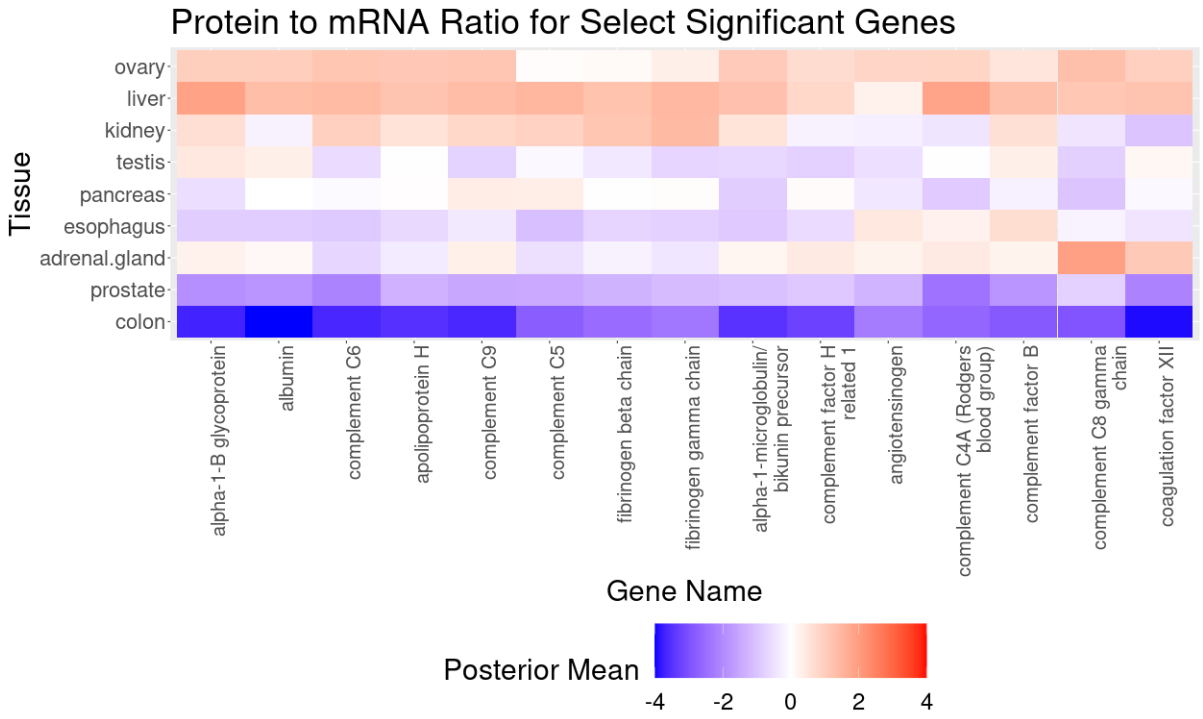

Figure 2.5: Posterior mean of R among select genes with significant extreme rPTR in at least one tissue.

2.3.2 Functional gene set analysis

By grouping together genes based on shared biological functions, we can identify GO groups with extreme rPTR. This approach has the advantage of more statistical power relative to gene-level testing, and its results directly relate to essential biological processes. We test R for individual tissues and gene sets to identify GO groups with extreme rPTR. Additionally we test the across tissues correlation by GO group to compare within-set correlations to that of all other genes.

Identifying functional gene sets, tissues with extreme average rPTR relative to all other genes

To identify functional gene sets with extreme rPTR in each tissue, we compare the mean R_{gt} across genes associated with each GO group to that of all other genes. Groups with consistently higher or lower R relative to the average across all other genes have extreme rPTR. Please see Methods for additional testing details. In table 2.6 we can see the number of significant GO groups at 1% and 0.01% FDR cutoffs for each tissue. We can see that tissues with a large number of significant genes during gene-level testing also have the largest set of significant groups, including the colon, esophagus, and prostate.

Table 2.6: Number of significant test results for group-level comparison of mean of R to all other genes, shown at a 1% and 0.1% FDR cut off for each cell type.

Tissue	**Significant at** 1% **FDR**	**Significant at** 0.1% **FDR**
Adrenal Gland	61	2
Colon	847	409
Esophagus	245	89
Kidney	183	45
Liver	214	75
Ovary	238	62
Pancreas	23	1
Prostate	322	126
Testis	186	40

We examine the posterior mean among genes associated with each GO group for a select set of significant groups in each tissue. As tests are run separately for each tissue, we review results only for select tissues. Please see results for the remaining tissues in the Appendix.

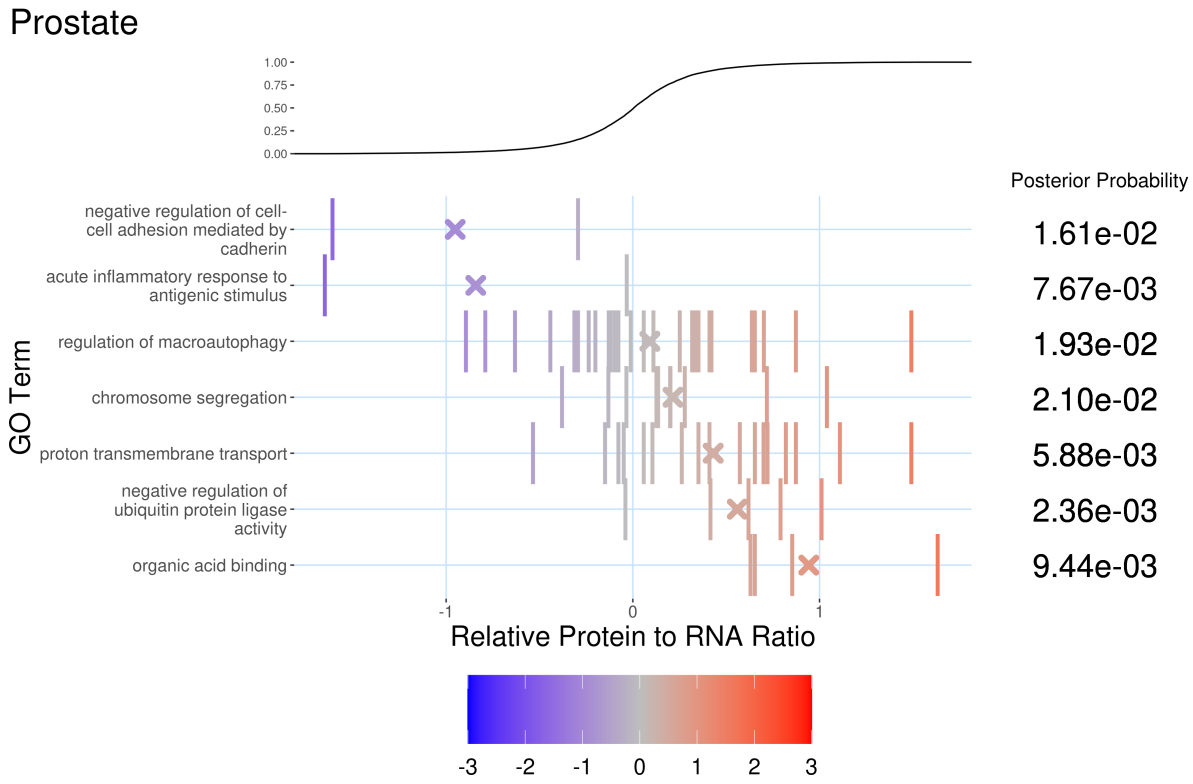

Figure 2.6: Posterior mean of R among genes in select significant GO groups for prostate tissue. Tick marks represent posterior mean for each gene. x marks represent average across genes in each group, and color corresponds to value. The cumulative distribution function of R across all genes is shown at the top panel, and the side panel shows posterior probabilities corresponding to this test.

Included among significant groups in figure 2.6 are acute inflammatory response to antigenic stimulus, which consists of genes that trigger an immune response. Previous literature identifies post transcriptional controls as a crucial component of immune system function [34]. Additionally, negative regulation of ubiquitin protein ligase activity has significant rPTR. Ubiquitination is a post translational modification involving the

attachment of the ubiquitin protein, affecting downstream protein stability and activity [35].

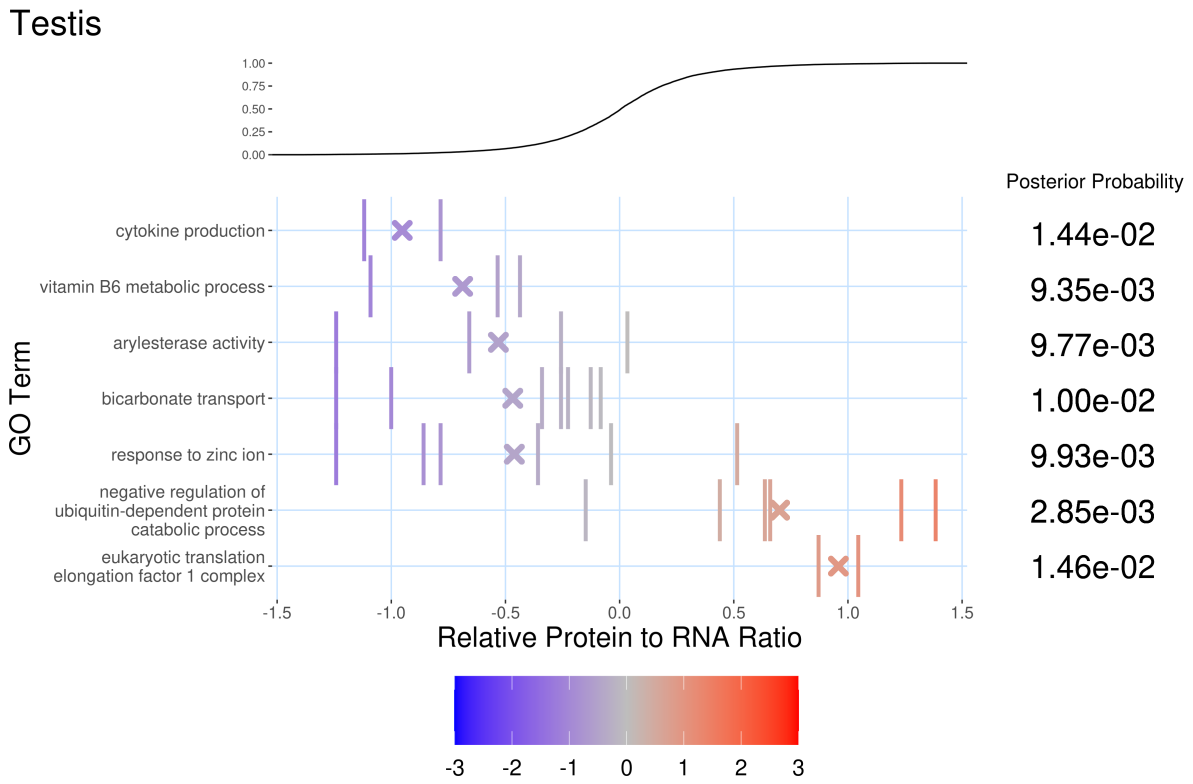

Figure 2.7: Posterior mean of R among genes in select significant GO groups for testes tissue. Tick marks represent posterior mean for each gene. x marks represent average across genes in each group, and color corresponds to value. The cumulative distribution function of R across all genes is shown at the top panel, and the side panel shows posterior probabilities corresponding to this test.

Among significant groups for testis tissue, we can see another ubiquitin-related functional gene set, as well as a group relating to response to zinc. Zinc finger proteins are a large gene set closely involved in protein synthesis including performing transcriptional activation and protein folding and assembly [36], and zinc is present in these proteins as a structural cofactor or "helper molecule".

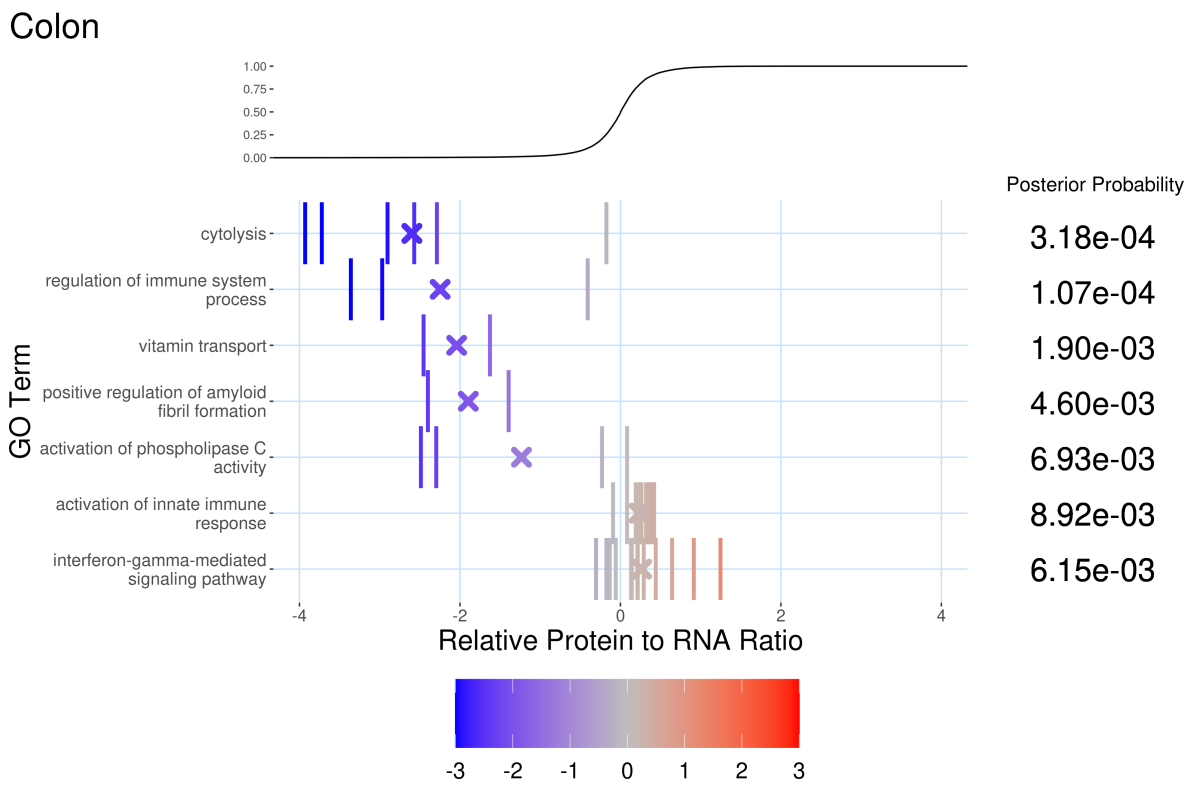

Figure 2.8: Posterior mean of R among genes in select significant GO groups for colon tissue. Tick marks represent posterior mean for each gene. x marks represent average across genes in each group, and color corresponds to value. The cumulative distribution function of R across all genes is shown at the top panel, and the side panel shows posterior probabilities corresponding to this test.

As we can see among significant groups, there are several groups relating to immune function that have significant rPTR. Positive regulation of amyloid fibril formulation also has significant low rPTR. While amyloid formulation is a normal biological process, its dysregulation, especially paired with an imbalance of formulation and clearence has been linked to several diseases including Alzheimer's and Parkinson's diseases. The role of post transcriptional modifications in amyloid function is of great interest [37].

Across tissues correlation of mRNA and protein

In addition to considering groups with extreme rPTR, we test the across tissues mRNA, protein correlation by comparing samples of μ_{gt} and $\mu_{gt} + R_{gt}$ in each GO group to that of all other genes. We compute the across tissues correlation of the group-level averages of μ_{igt} and $\mu_{igt} + R_{igt}$ where i represents posterior sample draw in each GO group to that of all other genes. Please see Methods for additional details on group-level testing. In table 2.7 we can see the number of significant groups at a 1% and 0.1% cutoff. As the posterior distributions for correlation are somewhat large, few groups are considered significant.

Table 2.7: Number of significant test results for group-level comparison across tissues correlation of mRNA and protein, μ and $\mu + R$.

Significant at 1% **FDR**	**Significant at** 0.1% **FDR**
10	2

In figure 2.9, we can see the gene-level across tissues correlation for the posterior mean among genes among each group. It appears that most groups with significant low correlation have within-gene mRNA, protein correlations that span a wide range. Included among significant groups are at least one group relating to complement factors. We saw several significant complement-related genes as a result of gene level testing, so it is expected that a group consisting of these genes have significant low correlation. Additionally, two gene sets relating to fibronolysis are significant. Fibronolysis is the normal biological process that prevents the growth of blood clots, and as such its activity is limited to specific tissues [38].

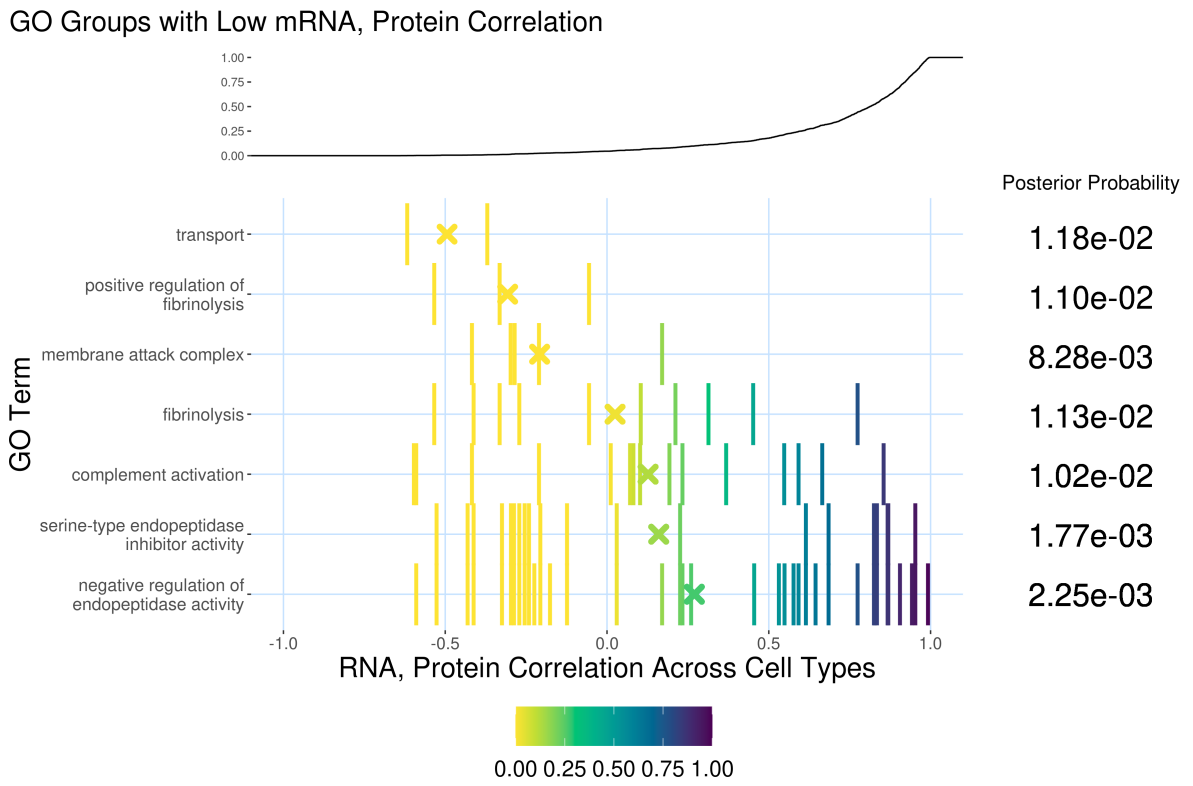

Figure 2.9: Across tissues correlation of posterior mean of mRNA and protein within each gene for select groups with significant low across tissues correlations. Each tick mark represents an across tissues correlation, with color representing value. Small x marks indicate the group-level average correlation. Additionally we can see posterior probabilities corresponding to each group and the empirical cumulative distribution function of the mRNA, protein across tissues correlation for all genes.

2.4 Posterior Predictive Checking

Using posterior predictive checks (PPCs), we are able to compare values simulated under the model asserted to that of observed values, using the observed values directly or test statistics of interest. A model that is performing acceptably should generally have posterior predictive draws that resemble observed values. We begin by examining the resemblance of observed values to posterior predictive draws directly, and then move on to compare test statistics essential for evaluating model performance with respect to

this type of data. In figure 2.8 we can see the proportion of Z-transformed $\log_2$ observed mRNA intensities falling within posterior predictive draws for each cell type and data set. While both posterior predictive draws for both modalities seem to under cover observed values, protein data sets are associated with generally worse coverage than that of mRNA.

Table 2.8: Proportion of genes with observed values in 95% quantile interval for posterior predictive draws in each tissue.

Tissue	**mRNA 1**	**mRNA 2**	**Protein 1**	**Protein 2**
Adrenal Gland	0.905	0.860	0.741	0.785
Colon	0.879	0.913	0.760	0.812
Esophagus	0.812	0.899	0.602	0.833
Kidney	0.867	0.845	0.770	0.762
Liver	0.663	0.855	0.723	0.812
Ovary	0.563	0.794	0.761	0.726
Pancreas	0.904	0.784	0.582	0.755
Prostate	0.902	0.931	0.787	0.726
Testis	0.854	0.819	0.825	0.722

Next, we compare the across tissues mRNA, protein correlation among posterior predictive draws to that of observed values. As observed values have across tissues correlation attenuated by sampling variability and unwanted sources of noise, we expect the correlation to be generally lower than that of underlying mRNA expression and protein abundance parameters; however, posterior predictive draws should accurately recover the effect of these sources of noise to reflect the correlation seen in observed values.

In table 2.9 we can see coverage information regarding posterior predictive and observed across tissues correlations. Posterior predictive correlations generally undercover that of observed values, similar to results in table 2.8. This may be a result of the way that sampling variability is modeled, as variance of posterior predictive draws do not reflect the variability of that of observed values across tissues and data sets. In figure

2.10 we can compare observed and posterior predictive variances across genes in each tissue for protein data sets. Note that this plot is available for mRNA data sets in the Appendix.

Table 2.9: Proportion of genes with observed across-tissue correlations in 95% quantile interval for posterior predictive across-tissues correlation. For each population, proportions are averaged with results from both data sets of the opposite modality.

	mRNA 1	**mRNA 2**	**Protein 1**	**Protein 2**
Proportion	0.855	0.902	0.801	0.824

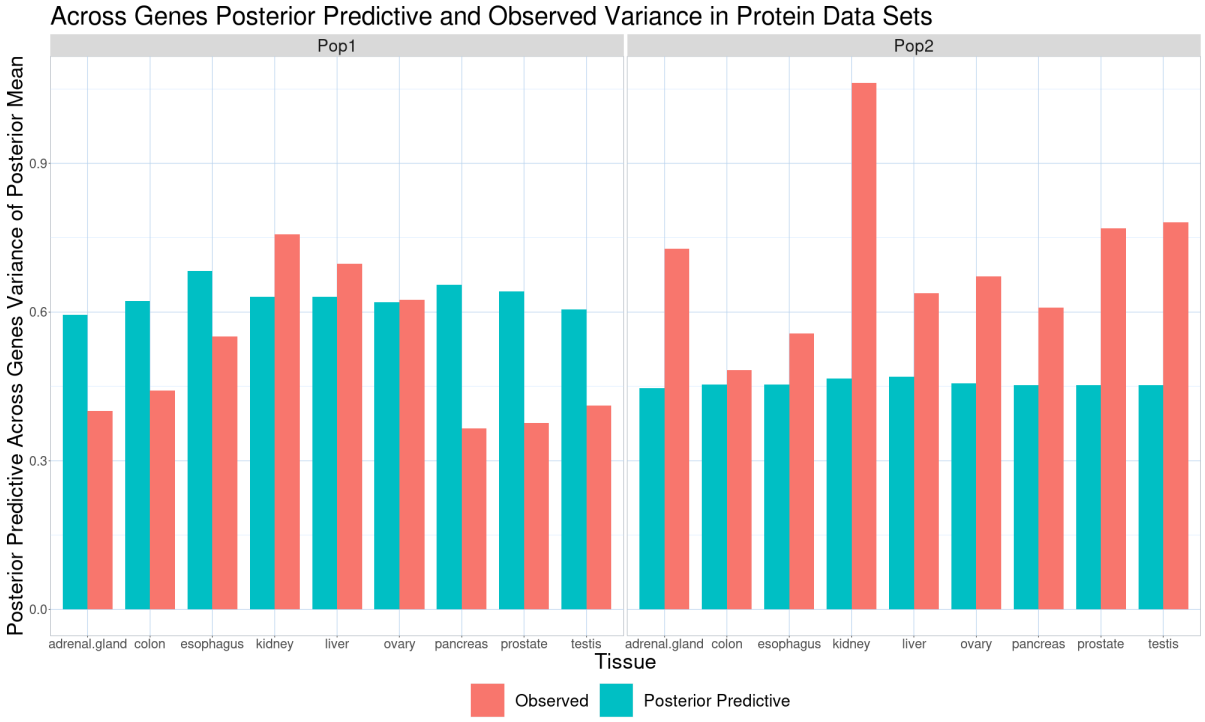

Figure 2.10: Posterior predictive and observed variance across genes in each tissue and data set. Posterior predictive variances are based on the posterior mean of posterior predictive draws for each gene, tissue and protein data set.

2.5 Discussion

Using a Bayesian model that leverages replicates across data sets and separates different sources of variation corresponding to sampling variability, variability of interest associated with biological processes, and systematic technical effects, we can better examine the relationship between mRNA and protein across tissues for individual genes. Posterior samples of μ_{gt} and $\mu_{gt} + R_{gt}$, representing $\log_2$ relative mRNA expression and protein abundance in each and tissue, reveal genes and functional sets of genes with potentially high rPTR, as rPTR and its variability across tissues reveals information about post trancriptional regulation at the gene and cell type levels. We test these parameters by comparing R_{gt} to zero for genes and sets of genes within individual tissues, as well as the across tissues mRNA, protein correlation for individual genes and GO groups.

Results from this analysis, and in particular functional gene set testing, find several significant ontology groups that are also found to be significant in previous analysis [17]—specifically, gene sets relating to translational initiation, cell adhesion, stress fiber, and ribosome function, among others.

One major challenge associated with this analysis is the resolution associated with bulk data analysis. Rather than measuring transcript and protein levels in individual genes, bulk data collection profiles values across cells belonging to the same tissue. Thus, measurements may reflect the most abundant cell types in each tissue, or may really reflect no cell type at all [7]. Important variability of interest relating to the functions of individual cells is unable to be examined. Additionally, only one observation is provided per gene-tissue-data set triplet, presenting limited opportunity to take advantage of replicates. This makes the equal-variance assumption used in this model necessary, but seems to fail to account for noise differences in each of the mRNA data sets. Another area of improvement for this model is its approach with respect to preprocessing and

scaling with and between modalities. By applying scaling directly to observed values, rather than underlying biological expression parameters, different sources of variability may be confounded.

Single cell mRNA and protein data, where observations are measured at the individual cell type and gene level, allow for greater resolution to estimate post transcriptional regulation. At the level of thousands of cells, transcript levels and peptide measurements are collected, and information is shared across cells of the same type. This allows for better estimation of noise variability at the individual data set level, and observations better reflect the biological systems under which data is collected. In the following chapter we evaluate application of this Bayesian model to single-cell data in post mortem human testes.

Chapter 3

Bayesian Inference of Post Transcriptional Regulation Across Cell Types

3.1 Introduction

Single cell RNA sequencing allows researchers to explore gene expression profiles with much higher resolution than is available with bulk RNA sequencing. Where bulk sequencing measures average expression across cells in a sample, without delineation between cells of different types, single cell RNA sequencing counts transcripts corresponding to each gene in individual cells. This creates the opportunity to explore biological processes with consideration of each unique cell type and developmental state, sources of important variability that are otherwise confounded in the analysis of bulk transcript data [7]. Single cell sequencing makes it possible to identify to analyze rare cell types, states, and subpopulations. While scRNA-seq offers unprecedented resolution of gene expression profiles, its analysis has major challenges, including technical artifacts, high

missingness, measurement noise, and high dimensionality [7], [39].

Single cell RNA sequencing data have often been used as a proxy to study proteins, which are involved in a wide range biological processes [40]. While gene expression and protein intensity levels are generally correlated at the mean-level, post transcriptional regulation can yield disagreement between these values [17]. Post transcriptional regulation can affect protein synthesis via processes such as splicing, RNA editing, and control of RNA degradation, and takes place after RNA has been transcribed [1]. Thus, it is of interest to quantify the effect of post transcriptional regulation at the gene level in order to better understand this relationship and its effect on biological systems. Additionally, quantification of the effect of post transcriptional regulation on protein synthesis can allow for evaluation of the fitness of scRNA-seq data as a proxy for single cell proteomics data.

Advancements in single-cell technology have enabled the measurement of peptide intensities at the cellular level with mass spectrometry and multiplexing based methods [41]. After required preparation, including isolating single cells, lysing to release proteins from cells, and digestion to reduce proteins to individual peptides, mass spectrometry involves ionizing and separating peptides based on their respective mass to charge ratios. During fragmentation, ions are separated into small fragments or precursors. Next precursors are measured within mass to charge ratio windows to identify corresponding peptides. To ensure that peptides have measurements on the same scale, each intensity measurement is divided by the median across all cells. As peptides are the building blocks of proteins, intensity levels are captured for individual peptides and may be aggregated across peptides associated with the same protein or treated as individual replicates [42]. This enables the identification of rare cell types and subpopulations and analysis of the proteome without using RNA as a proxy.

Joint analysis of transcript and protein data is critical because it provides a window

into post transcriptional regulation at the gene level, as it becomes possible to quantify the relative protein-to-RNA ratio (rPTR) as the ratio of $\log_2$ protein intensities and mRNA expression for each gene and cell type relative to the across cell types average. If a gene is regulated mostly transcriptionally, the relative intensity of mRNA and protein will be similar, and its ratio will have little variability across cell types. Conversely, if a gene is regulated mostly post transcriptionally, rPTR will have higher variability across cell types and the correlation for mRNA and protein across cell types will be low [43].

Estimating post transcriptional regulation cannot be adequately accomplished with empirical data alone, as observed differences between mRNA and protein can be mistakenly attributed to biological causes, when it is in fact a result of high measurement noise, sampling variability, and differences in dynamic range between the two modalities. As a result, the true correlation across tissues between mRNA and protein is attenuated, and rPTR is overestimated [17]. Careful analysis is required to successfully overcome these challenges. In this chapter, we present and evaluate our Bayesian approach to estimate rPTR, which leverages information contained across modalities and data sets. This model uses a hierarchical structure to denoise the data and estimate underlying biological mRNA expression and protein abundance for both mRNA and protein and ultimately compute rPTR. This method is applied to two single cell mRNA data sets and four single cell protein data sets [44]. Analysis of testes data has the unique advantage of exposure to cells of several types and stages within the spermatogenesis process, wherein sperm is created, providing insight into the various mechanisms that regulate protein synthesis during spermatogenesis.

3.2 Methods

3.2.1 Data sources

Data in this chapter are from postmortem human testes from healthy individuals. Tissue is sequenced at the single-cell level separately for both mRNA and protein. With the exception of joint measurements of transcripts and limited cell surface proteins [45], technology has not yet been developed to simultaneously sequence genes and peptides within the same cell. Thus, the transcriptome and proteome from the same tissue samples are measured separately. Cells in this analysis have Nano ProteOmic sample Preparation (nPOP) [46] applied, which involves lysing cells in water with a freezing and heat cycle as a part of preparing cells for data collection.

Protein data are generated using the SCoPE2 pipeline (Single Cell ProtEomics by Mass Spectrometry 2), a mass-spectrometry based pipeline that employs an isobaric carrier to enhance peptide sequence identification [41]. As "top-down" proteomics, where intensity-levels are measured for entire proteins, has not yet been developed, measurements are collected at the peptide-level, and each peptide is identified and matched to its corresponding predefined proteins. SCoPE2 enables the quantification of thousands of proteins per cell with high efficiency, representing a major advancement compared to classical antibody-based approaches, typically limited to quantifying only 50-100 proteins per cell. Furthermore, SCoPE2 has been shown to have high accuracy and reproducibility, making it a valuable tool for single cell proteomics research [41]. This pipeline also utilizes Dart-ID and MaxQuant softwares [47] [42], which measure essential peptide-level features, such as retention time [47]. This is used to compute a posterior error probability (PEP) associated with each peptide identification and subsequent false discovery rates (FDRs) for each peptide. Using MaxQuant, it is possible to compute precursor ion fractions (PIFs), the degree to which co-isolation of multiple precursors takes place prior

to fragmentation, which represents a measure of purity. Data from this experiment consists of intensities for each peptide in each cell, and peptides are later grouped together according to their corresponding proteins.

Single-cell RNA sequencing (scRNA-seq) enables the measurement of gene expression in individual cells, and requires careful preparation involving cell isolation and lysis, mRNA capture, and library preparation. During these steps, individual cells are isolated and mRNA within these cells are released. mRNA is then reverse transcribed into complementary DNA (cDNA), which is synthetic DNA transcribed from mRNA using the reverse transcriptase enzyme [48]. The cDNA is amplified and later sequenced using high-throughput sequencing technologies, such as clampFISH [49]. These technologies use techniques such as barcoding to retain the cell label associated with each sequenced unique molecule, resulting in counts for each gene in each cell. As scRNA-seq enables the measurement of gene expression in individual cells, researchers can identify cellular heterogeneity and subpopulations that might be missed by traditional bulk RNA sequencing methods. scRNA-seq also presents challenges as a result of unwanted technical variability missingness, and thus requires careful analysis.

This data collection process has resulted in ~ 4000 and ~ 2800 cells for each of the mRNA data sets, and between 759 and 1547 cells in each of the protein data sets. There are about 3100 shared genes between the two modalities.

3.2.2 Preprocessing, Normalization and Scaling

To mitigate the presence of low-quality cells or cells with poorly measured features, filtering is first applied to both the transcript and peptide-level data sets. For mRNA data sets, we remove cells with an extreme number of transcript counts or features, which can be a result of technical artifacts including "doublets," where two cells are

accidentally sequenced as one. Cells with elevated mitochondrial percentage (which can indicate technical artifacts or damaged cells) are also removed. This filtering step is essential to ensure that low quality cells don't contribute unwanted noise to analysis.

Filtering is also applied to peptide-level data, leveraging analysis performed by Dart-ID and MaxQuant. This includes the removal of peptides with FDR $> 1\%$, PIF < 0.8, as well as contaminant peptides identified by MaxQuant. Cell-level filtering is also applied, including comparing the coefficient of variation (CV) across proteins with more than five peptides within each measured cell to that of a control sample. Cells with low CV relative to the control sample are filtered out.

Preprocessing required to input observed values into the Bayesian model is different from that required to perform cell type identification, which will be described for each cell type assignment method. For mRNA data sets, the sum of unique molecular identifier (UMI) counts among cells assigned to each type are used for each gene in each data set separately. Let M_{mgc} be this sum in each data set $m \in \{1, 2\}$ for each gene g and cell type c. Included with each sum is a record of the total number of transcript counts across all genes and cells assigned to cell type c in data set m.

For the protein data sets we use sufficient statistics representing the average $\log_2$ intensity for each peptide across cells assigned to each cell type, consisting of the sum and sum of squared values. Let P_{pg_kc} represent the average $\log_2$ intensity across cells assigned to type c for peptide k associated with gene g in protein data set p. Included with sufficient statistics $\sum_{k \in g} P_{pg_kc}$ and $\sum_{k \in g} P^2_{pg_kc}$ is the number of peptides k corresponding to protein g in data set p.

3.2.3 Missing Data

The management of missing data is a crucial aspect of transcriptomics and proteomics-related data analysis. Among transcriptomics data, missing observations can be a result of two main causes. First, "true" or "structural" zeros occur because a particular gene is not being expressed in a given cell at the time of isolation. Alternatively, "dropouts" occur when a gene is being expressed, but at an abundance so low that it is not detected. This is a result of limitations with respect to sequencing technology and sample prep [50]. Thus, the total set of missing observations in scRNA-seq data contains a subset of observations that are not actually missing.

Proteomics data also involves missingness as a result of several mechanisms, with instances including when a protein is not analyzed or identified, when a protein has very low abundance such that it is not detected, and when a protein truly has zero abundance [51]. A protein is not analyzed when its corresponding ions are not sampled for measurement at all, as a limited subset of the total ions present in a sample will actually be quantified. Whether or not an ion will be sampled depends on choice of measurement technique. This includes data dependent acquisition, where only a fraction of peptides present will later be fragmented (often based on the abundance of peptide) data independent acquisition, where all peptides within predetermined sets of isolation windows move forward to fragmentation, and targeted acquisition, where, for example, measurements are collected for a predefined set of proteins [52]. Targeted acquisition is sometimes also data-dependent [53]. With several of these approaches ion detection may be dependent on abundance of such ion [54]. With some data-independent acquisition measurement techniques, all detectable ions within a window may be identified, but not all peptides will be identified, creating a similar challenge where probability of peptide identification is affected by the abundance of the peptide itself among other factors

[52]. Additionally, missingness can occur when an ion is selected for sampling, but is below the limit of detection during measurement. This detection limit is a direct result of technology-based aspects including retention time, and detailed knowledge of these factors and selected measurement technique makes it possible to identify which proteins are missing as a result of this cause, and model them accordingly [51]. Finally, observations may be missing because a protein is truly not present in a measured sample. For example, proteins specific to functions within particular cell types, such as hormone-related proteins, may truly be not present in a given cell. Thus, it is crucial not to impute such values [51]. Data in this analysis are generated using a data-dependent acquisition technique, and observations below the limit of detection are not specifically labeled.

These various sources of missingness result in a high level of missing observations at the cell-level. Tables 3.1 and 3.2 display the proportion of missing observations across all cells and genes in each of the data sets used in this analysis. As both tables depict, the majority of gene-cell gene products are missing.

Table 3.1: Proportion of missing observations for each cell type within each gene $\times$ cell mRNA data matrix. Note that a gene is considered missing in a given gene, cell pair if no transcripts are counted corresponding to that gene in the particular cell, but is measured in at least one other cell in the data set. Therefore, the subset of coding genes not present in any cell within a data set is not reflected.

Cell Type	**mRNA Pop1**	**mRNA Pop2**
Peri-Tubular Myoid (PTM)	0.907	0.913
Spermatogonia (SPG)	0.838	0.854
Leydig Cells (LC)	0.924	0.928
Endothelial Cells (EC)	0.908	0.914
Macrophage (M)	0.935	0.938
Spermatocytes (SPC)	0.873	0.883
Spermatid (St)	0.912	0.919

Table 3.2: Proportion of missing observations at the protein, cell level for each protein data set and cell type. Missingness is defined using the same approach as 3.1. Additionally, note that missingness for protein data sets are displayed at the protein-level as opposed to the peptide-level. Thus, peptides may not be measured within a particular cell but the corresponding protein will not be missing if another peptide is present for the same protein.

Cell Type	**Prot Pop1**	**Prot Pop2**	**Prot Pop3**	**Prot Pop4**
Peri-Tubular Myoid (PTM)	0.774	0.789	0.786	0.675
Spermatogonia (SPG)	0.781	0.783	0.789	0.688
Leydig Cells (LC)	0.780	0.776	0.784	0.676
Endothelial Cells (EC)	0.781	0.777	0.788	0.679
Spermatocytes (SPC)	0.783	0.782	0.786	0.685
Spermatid (St)	0.789	0.791	0.782	0.694

As single-cell data has much greater resolution than previous measurement techniques, grouping cells together according to each of their respective cell types is an essential tool in mitigating missingness. For example, below we can see missingness at the cell-type level, after aggregating across cells for each of the mRNA data sets. Note that, after aggregating across cells assigned to each cell type among the protein data sets, every protein is observed at least once in every cell type; however, there are large disparities across data sets with respect to the total set of genes observed. In table 3.4, we can see the proportion of proteins observed across all data sets observed in each data set individually.

Table 3.3: Proportion of genes present in each data set that are present in all other protein data sets.

Cell Type	**mRNA Pop1**	**mRNA Pop2**
Peri-Tubular Myoid (PTM)	0.145	0.222
Spermatogonia (SPG)	0.061	0.0377
Leydig Cells (LC)	0.162	0.227
Endothelial Cells (EC)	0.227	0.296
Macrophage (M)	0.348	0.356
Spermatocytes (SPC)	0.216	0.471
Spermatid (St)	0.378	0.474

Table 3.4: Proportion of genes in each cell type with missing observations across all cells assigned to each type for each Protein data set.

	Prot Pop1	**Prot Pop2**	**Prot Pop3**	**Prot Pop4**
Proportion of Observed Genes	0.493	0.456	0.421	0.360

The different mechanisms behind missingness in single cell data may contribute to both missingness at random and missingness not at random or non-ignorable missingness [55]. Observations are missing not at random (MNAR) when the mechanism driving missingness depends on the missing observations. This presents a major challenge in analysis. While some imputation methods exist to stratify missingness due to dropouts and true zeros [56] [57], imputation methods that do not acknowledge biological zeros [23] or preserve true zeros, but very conservatively perform imputation [24] may fail to properly account for these different sources of missingness. As improvement in performance for downstream analysis has not been shown [25], imputation should be considered with serious caution.

To manage missingness within our model, which assumes missingness at random (MAR) and is not performed on imputed data, we restrict the subset of genes in analysis to those where the MAR assumption is more reasonable, as information can be shared across cell-gene pairs. In particular, we consider only genes with observed values in three or more cell types in at least one mRNA and one protein data set. As values inputted into the model are computed across cells assigned to the same type, genes that are lowly abundant across multiple cells in a type will have low averages. While zeros in gene, cell pairs are not included in across-cell averages, total transcript counts include counts corresponding to all other observed genes in each cell.

3.2.4 Cell Type Identification

Cell type assignment is a crucial component of analysis for which there is no available "ground truth" in most cases. Cells of different types have different transcriptomic and proteomic profiles, as each cell type is responsible for carrying out specific functions. Thus, disease and treatment can affect different cell types differently, genes can be expressed at high rates in some cell types but not others, and functional sets of genes may behave distinctly depending on cell type identity. Thus awareness of cell type identity is a critical component of single-cell data analysis.

Cells present in the human testes data set mostly fall into the following types [58]:

- Peritubular Myoid (PTM) cells, which are smooth muscle cells surrounding seminiferous tubules, the structures in the testes where sperm are produced.
- Endothelial Cells (EC), which line blood vessels within the testes. Endothelial cells are essential in preventing harmful materials from entering seminiferous tubules and damaging sperm.
- Leydig Cells (LC), which are responsible for producing testosterone and are located between seminiferous tubules.
- Spermatogonia (SPG) cells, which are the germ cells that later produce mature sperm, located inside seminiferous tubules. Spermatagonia are the precursor cells that initiate spermatogenesis, the process of sperm cell development.
- Spermatocyte (SPC) cells, which are intermediate cells in the spermatogenesis process that undergo division and differentiation to later create genetically diverse sperm cells.

- Spermatids (St) are the final stage of sperm cell development before becoming mature sperm. Some spergmatogonia undergo differentiation and several rounds of meiosis to later become spermatids.

A variety of approaches have been used to identify cell type using mRNA or protein data, [8] [59], and more recently multimodal approaches at cell type assignment have emerged [60], [61], [62]. First, we review the cell type assignment approach from Seurat V2, one the most popular approaches to assign cell types specifically with transcriptomics data [8], followed by techniques used for assigning cell type labels to multimodal data, including LIGER [60] and scDMVAE [61].

Seurat [8], a package by Satija Lab, was initially created in 2015 for the purpose of analyzing transcriptomics data, and has developed an array of methods to analyze multi-omics data, including proteins and chromatin accessibility [62]. Cell type assignment using Seurat V2 includes linear and nonlinear dimension reduction, clustering, and typically predefined marker genes to perform cell type identification. After filtering is complete, each cell is normalized across genes by total expression count and multiplied by a scaling factor. At this point, highly variable genes are selected, which have high cell-to-cell variability and can help mitigate the high degree of noise. Then each gene is scaled across cells such that each gene has mean 0 variance 1. At a high level, using filtered, scaled, and normalized scRNA-seq counts for each gene and cell, linear dimension reduction such as principal component analysis is used to reduce the dimensionality of each cell across highly variable features (genes). Next, a graph-based clustering approach identifies clusters of cells based on relationships of shared nearest neighbors to each cell. In order to accomplish this clustering, cells are embedded to a graph structure such as K-nearest neighbors (KNN) on the euclidean distance of principal components corresponding to each cell. This establishes edges between cells corresponding to princi-

pal component similarity. In the clustering step, modularity optimization techniques are utilized to group cells together. Finally, cell types must be assigned to each cluster. This can be accomplished by identifying marker genes, which are highly expressed in specific clusters compared to other clusters. Marker genes can be selected using previous literature, separate data sets or data collection methods. Cells, as well as their corresponding cluster assignments, are often plotted in a low-dimensional representation using nonlinear dimension reduction, such as Uniform Manifold Approximation and Projection (UMAP) [63]. This technique approximates the manifold of the high-dimensional data with a graph structure, and then optimizes the embedding of the graph into a low-dimensional space that preserves both the global and local structure of the data.

Next we consider some multimodal cell type assignment approaches to assign cell type labels to samples with multi-omic single cell measurements.

scDMVAE [61] performs multimodal cell type assignment by first projecting cells in the mRNA and protein data sets to shared latent space. During this step, attention scores are computed for each cell in each modality, which represents a weighted average of cells in the opposite modality by similarity [64]. As this method is parametric and uses a Zero-Inflated Negative Binoimal model for mRNA counts, log normalization or scaling is not required; however, highly variable features are selected. For protein observations, $\log_2$ normalized and scaled peptide-intensities are used for each. By default, cosine similarity is used to evaluate the relationship between cells. Next, the weighting strategy from Seurat V3 [62] is employed, wherein attention scores are used as distances, and KNN is applied. Cells with neighbors belonging to a different modality are then anchors, and these anchors are scored using a metric inspired by Shared Nearest Neighbors (SNN) [65]. As the final step of the integration portion, anchor cells and all remaining cells are connected using the scores associated with each anchor, as well as the distances from each cell to its nearest anchors. Finally, cell type labels are transferred from cells in the

mRNA data sets to that of the protein data sets by computing predicted label scores corresponding to each cell type for each cell as a weighted average of labels of reference assigned cells based on the distance to anchor cells within the same modality, anchor stores, and attention scores across modalities.

LIGER [60] also consists of an integration step, which uses integrative Non-negative Matrix Factorization (iNMF) [66] to construct a shared low-dimensional space that accounts for both similarities and differences between modalities. This approach requires normalization, the selection of highly variable features, and scaling, similarly to the approach used in Seurat. Unsupervised clustering can then be applied specifically to mRNA data, or to the joint space. Next, marker genes are identified with differential expression to identify the cell type that corresponds to each cluster. In this case, as labels are assigned based on the cells from the mRNA data sets, label transfer takes place for each cell in the protein data set by majority vote among the k-nearest cells from the mRNA data set.

Figure 3.2 depicts UMAP reductions for each protein population, as well as their corresponding cell type assignments. Figure 3.1 displays this same dimension reduction for the mRNA data sets. We can see clustering corresponding to cell type in each cluster. Note that one cell type, Macrophages (M), were identified in the mRNA data sets but not in protein data sets. Few cells among the mRNA data set were assigned to this cell type, so this can be a result of general low abundance of this cell type among the samples tested. As cell types are only analyzed if they are present in both mRNA and protein data sets, the macrophage cell type is not included in analysis.

Using these cell type assignments, each cell is modeled based on its identified type, and transcripts across cells belonging to the same type represent biological replicates, to be used for underlying relative gene product abundance estimation.

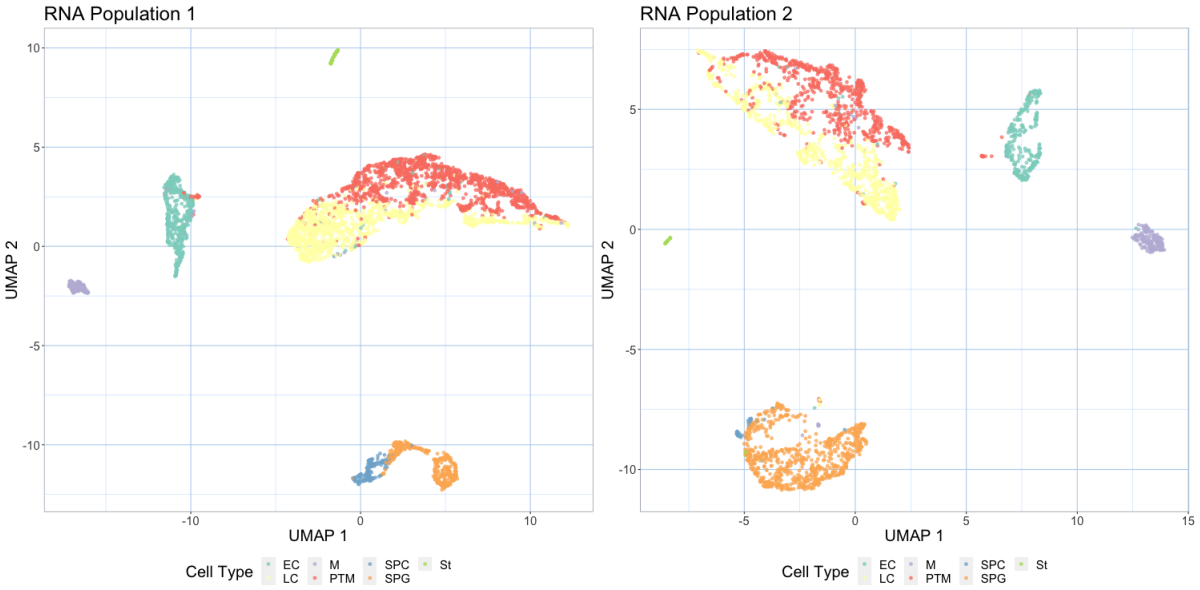

Figure 3.1: UMAP Dimension Reduction of mRNA Data sets. Point color corresponds to cell type assignment in each panel. We expect to see points form clusters corresponding to cell type.

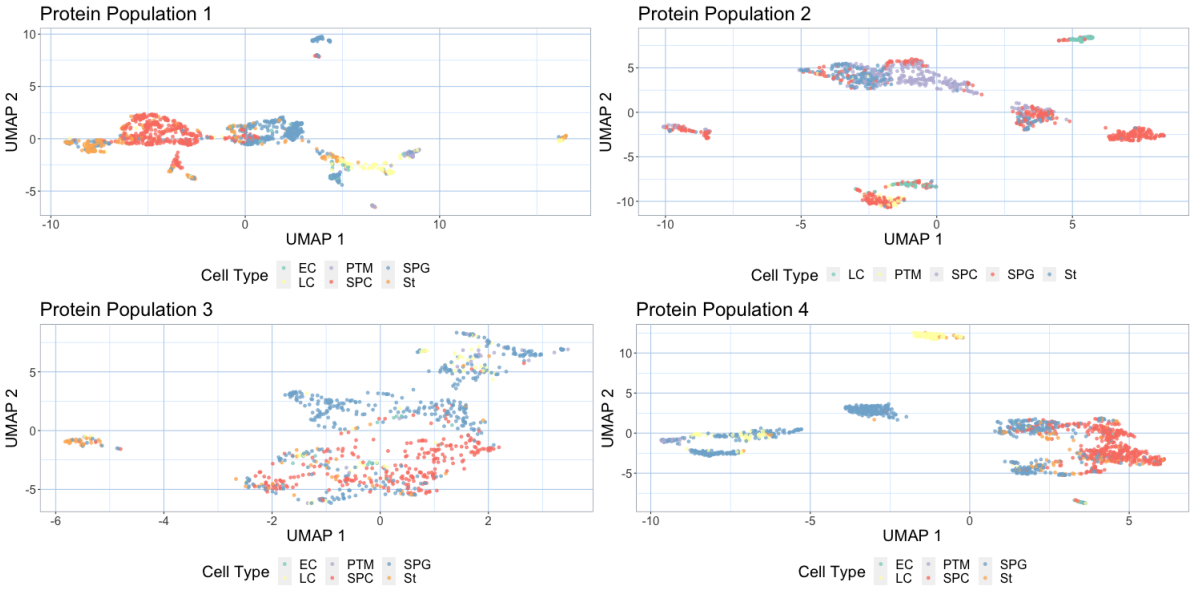

Figure 3.2: UMAP Dimension Reduction of Protein Data sets, with cell type reflecting cell type assignment.

3.2.5 Bayesian Inference of protein to mRNA ratios

We introduce our Bayesian Model to estimate rPTR using observed transcript and protein data. Let M_{mgc} represent the sum of transcript counts across cells assigned to cell type c for gene g and mRNA data set m. In this particular case, $m = \{1, 2\}$. Additionally, let P_{pg_kc} represent the average $\log_2$ peptide intensities for protein data set $p = \{1, 2, 3, 4\}$ across cells assigned to cell type c for peptide k associated with gene g.

Assumptions included for this model are that cells are correctly associated with their corresponding cell type, which depends on performance of type identification, and that peptides are properly identified during the data collection process. Additionally, we assume conditional independence of genes given model parameters. Another critical assumption for this model is that M_{mgc} and P_{pg_kc} are independent replicates corresponding to some shared underlying biological process. In particular, transcript counts for gene g among cells assigned to cell type c have some shared relative rate of $\log_2$ transcription, μ_{gc}, and peptide-level averages have shared underlying protein intensity $\mu_{gc} + R_{gc}$, where R_{gc} is defined as the $\log_2$ relative ratio of protein to mRNA. If a gene is not post transcriptionally regulated, the amount of protein synthesized will be directly proportional to the amount of mRNA transcribed; therefore, R_{gc}, the $\log_2$ relative protein-to-mRNA ratio, will reflect the degree of post transcriptional regulation affecting gene g within cell type c, relative to the average transcription or protein synthesis for gene g. While computing rPTR based on observed values directly is subject to high measurement noise and missingness, our model enables us to utilize replicates across data sets, obtaining more robust estimates of rPTR and estimating the effect of distinct sources of variability– that associated with biological processes, unwanted technical noise, and random error.

We explore the likelihood and prior distributions for this model, as well as specification of the corresponding posterior distribution.

Likelihood

mRNA We model the sum of transcript counts across cells assigned to each cell type for each individual gene, M_{mgc} using the following likelihood,

$$M_{mgc} \sim \text{Negative Binomial}^*(n_{mc}2^{(\mu_{gc})}a_g\gamma_{mc}, \phi_m), \tag{3.1}$$

where n_{mc} represents the total transcript count across cells assigned to cell type c. To ensure that $\log_2$ transcription rate, μ_{gc}, is relative to its gene-level average, we use parameter a_g to represent the average transcript rate in gene g across all cell types. It is also critical to account for batch effect, an unwanted source of variability that can yield systematic difference between different mRNA data sets. As cells of different types are often different sizes, we allow batch parameter γ_{mc} to vary by both mRNA data set and cell type. Finally we define ϕ_m, which represents overdispersion for each mRNA data set m. This enables the model to account for the high noise associated with transcript count, as the equal mean and variance assumption often associated with the Poisson likelihood may be otherwise unreasonable due to non-biological sources of variability and additional biological-source related variability, including transcriptional bursts, where transcription occurs in "pulses" [67].

Consider a Negative Binomial(r, p) random variable with traditional parameterization. Under this alternative parameterization, Negative Binomial$^*(\mu^*, \phi^*)$, $r = \frac{\mu^*(1+\phi^*)}{\phi^*}$ and $p = \frac{\phi^*}{1+\phi^*}$. While several different likelihoods were considered to model transcript counts, the Negative Binomial likelihood was ultimately chosen based on model performance relative to Poisson and Normal likelihoods. Please see the Appendix for a detailed discussion on the model comparison process.

Protein We use sufficient statistics to represent P_{pg_kc}, the average log peptide-level intensity across cells belonging to each cell type for proteins corresponding to each gene g. We model peptide-level averages with likelihood,

$$P_{pg_kc} \sim \text{Normal}(s_g(\mu_{gc} + R_{gc}) + b_g + \kappa_{pc}, \tau_p^2). \tag{3.2}$$

This models each peptide as an independent replicate of the protein to which it forms.

To consider biological process parameters μ and R relative to the average protein-level intensity associated with each gene, we use b_g to represent gene-level average protein intensities. Similarly to the batch-effect parameter utilized in the mRNA likelihood, κ_{pc} represents technical data set, cell type effect.

Analysis of single cell transcript count and peptide intensity data often involves scaling the variance of observed values, for example, by dividing by an observed standard deviation or the total abundance of a protein across cells or cell types [68] [69]. This ensures reasonable comparisons between genes, as intensity can vary widely both across genes and with respect to modalities; however, this approach may also confound different sources of variability, including variability contributed by biological processes of interest and sampling variability. Thus, we ensure comparison between genes, and correspondence across modalities without confounding these variability sources by scaling $\log_2$ relative protein abundance parameter $\mu_{gc} + R_{gc}$ directly with s_g.

Finally, we use τ_p, which varies according to protein data set p, to model sampling variability for each of the peptide-level averages. This allows for the separation of unwanted variability and biological variability of interest associated with μ_{gc} and R_{gc}, and accounts for different sampling variability in different data sets.

Prior Distributions

We model biological process parameters μ_{gc} and R_{gc} with normal priors having mean 0, as μ_{gc} and R_{gc} are modeled respective to gene-level means and the presence of post transcriptional regulation is not assumed in advance. We model variance of μ_{gc} and R_{gc} on a per-gene basis to allow for flexibility in modeling μ and R. $\sigma^2_{\mu_g}$ represents variance for μ_{gc}, and varies by gene g. We implement a hierarchical structure for $\sigma^2_{\mu_g}$, as $\sigma^2_{\mu_g}$ is assigned a Gamma(α_μ, β_μ) hyperprior. The variance associated with R_{gc}, $\sigma^2_{R_g}$, has the same hierarchical structure, $\sigma^2_{R_g} \sim$ Gamma(α_R, β_R). Note that $\alpha_\mu, \beta_\mu, \alpha_R, \beta_R$ are all assigned Gamma priors, with shape and rate determined by considering domain expertise, convergence statistics, and leave-one-out cross validation.

The scale parameter, s_g, is modeled with the same hierarchical structure as the biological process variance parameters. This is to avoid mistakenly associating genes with disagreement in scaling across modalities with having high post transcriptional regulation. Each $\frac{1}{s_g}$ is assigned a Uniform$(0, 10)$ prior, shared across genes. Setting the prior on $\frac{1}{s}$ as opposed to s directly helps to prevent a disproportionate amount of prior mass being set on extreme scales.

To account for technical differences between data sets, which can potentially affect cells differently depending on factors such as size [70], we use batch effect parameters γ_{mc} and κ_{pc} for mRNA and protein, respectively. γ_{mc} is given a Gamma(α_m, α_m) prior, where $\alpha_m \sim$ Gamma$(2, 1)$. $\kappa_{pc} \sim$ Normal$(0, 0.25)$, as $\log_2$ peptide intensities are modeled using a normal likelihood.

We assume that sampling variability is constant across genes and cell types within each protein data set with parameter $\tau^2_p \sim$ Cauchy$^+(0, 1)$. As we model mRNA using the Negative Binomial distribution with alternative parameterization, ϕ_m models overdispersion for each mRNA data set and $\text{Var}(M_{mgc}) = E[M_{mgc}] + \frac{E[M_{mgc}]^2}{\phi_m}$. Rather than setting

a generic prior on ϕ_m directly, which may inadvertently put more prior mass on fits with higher overdispersion, we place a Cauchy$^+(0, 10)$ prior on $\frac{1}{\phi_m}$.

Additional Model Details

To clearly reflect the structural aspects of our model, figure 3.3 displays a plate diagram of parameters present in the likelihood. μ_p, while not an explicit model parameter, indicates how underlying $\log_2$ relative protein intensity is formed as the sum of transcription rate μ_{gc} and rPTR R_{gc}.

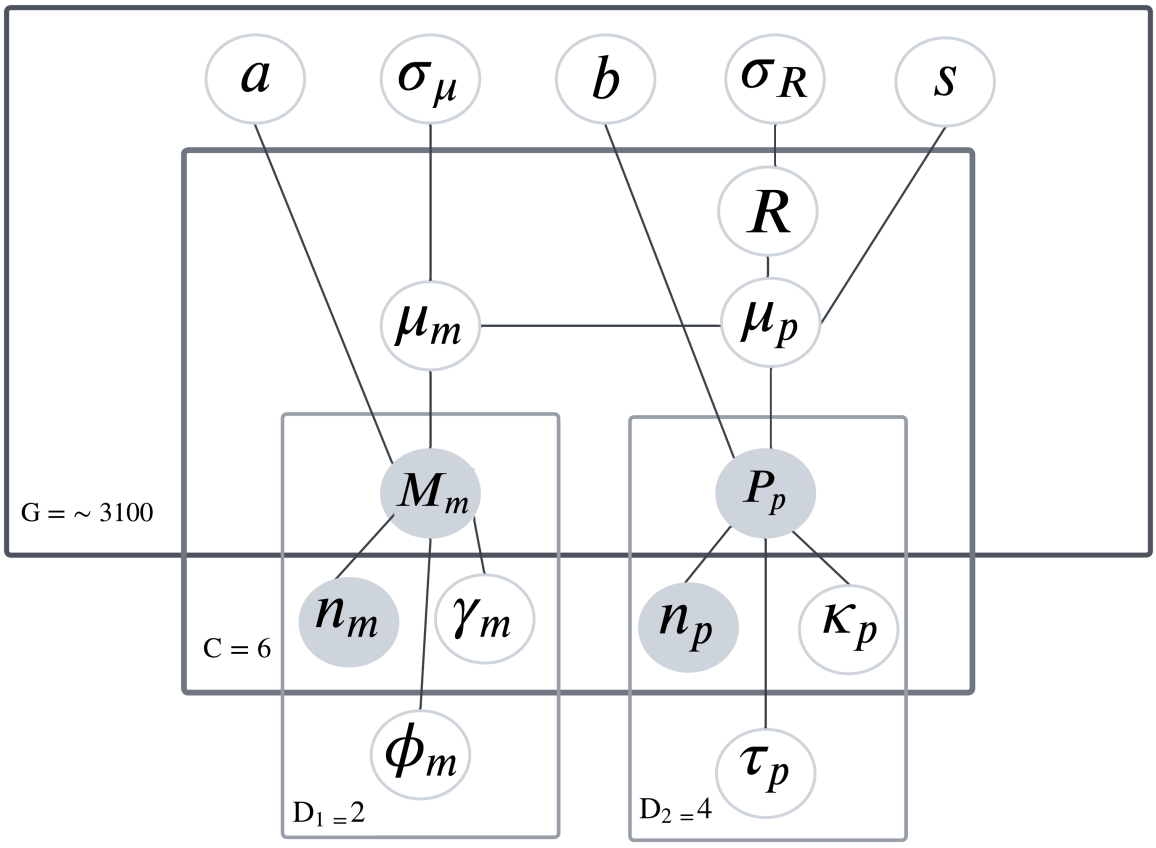

Figure 3.3: Plate Diagram of Bayes Model. Closed circles represent observed values, and open circle represent unknown parameter. Overlapping rectangles represent model structure, and lines denote relationships between parameters.

In table 3.5, we can see a full specification of parameters present in the model as well

as their corresponding prior distribution, as outlined in the Priors section. We assume independent priors for all parameters.

Table 3.5: Model parameters and corresponding prior distributions.

Parameter	Interpretation	Prior Distribution
μ_{gc}	Underlying $\log_2$ transcription rate	Normal$(0, \sigma^2_{\mu_g})$
R_{gc}	Log relative protein-to-mRNA ratio (rPTR)	Normal$(0, \sigma^2_{R_g})$
$\sigma^2_{\mu_g}$	mRNA fold change variance	Gamma(α_μ, β_μ)
$\sigma^2_{R_g}$	rPTR variance	Gamma(α_R, β_R)
a_g	mRNA centering parameter	Gamma$(3, 4)$
b_g	Protein centering parameter	Normal$(0, 0.5)$
s_g	Protein scaling parameter	$1/s_g \sim$ Uniform$(0, 10)$
γ_{mc}	Batch effect correction (mRNA)	Gamma$(\alpha_\gamma, \alpha_\gamma)$
κ_{pc}	Batch effect correction (protein)	Normal$(0, 0.25)$
τ_p	Peptide noise variance	Cauchy$^+(0, 1)$
ϕ_m	mRNA overdispersion	Cauchy$^+(0, 10)$
n_{mc}	Transcript counts	—
n_{pgc}	Number of peptides associated with protein	—

Posterior Distribution Given the model likelihood and priors, we can write the full log likelihood for this model. Given all parameters $\vec{\theta} = [\mu, R, \sigma^2_\mu, \sigma^2_R, \gamma, \kappa, a, b, s, \phi, \tau]$ the posterior distribution is proportional to,

$$\begin{aligned} p(\vec{\theta}|M, P) &\propto p(\vec{\theta})p(M|\vec{\theta})p(P|\vec{\theta}) \\ &\propto p(\vec{\theta}) \prod_{m=1}^{D_1} \prod_{g=1}^{G_m} \prod_{c=1}^{C_{mg}} p(M_{mgc}|\vec{\theta}) \prod_{p=1}^{D_2} \prod_{g=1}^{G_p} \prod_{c=1}^{C_{pg}} p(P_{pgc}|\vec{\theta}), \end{aligned} \tag{3.3}$$

where D_1 is the total number of mRNA data sets and D_2 is the total number of protein data sets. G_m is the total number of genes for which there are transcripts present in mRNA data set m, G_p is the total number genes for which there are peptides measured in protein data set p, and C_{mg} and C_{pg} are number of cell types observed in each respective data set and gene.

The log likelihood is,

$$
\begin{aligned}
l(\vec{\theta}|M,P) &\propto l(\vec{\theta}) + \sum_{m=1}^{D_1}\sum_{g=1}^{G_m}\sum_{c=1}^{C_{mg}} \log(p(M|\vec{\theta})) + \sum_{p=1}^{D_2}\sum_{g=1}^{G_p}\sum_{c=1}^{C_{pg}} \log(p(P|\vec{\theta})) \\
&\propto l(\vec{\theta}) + \sum_{m=1}^{D_1}\sum_{g=1}^{G_m}\sum_{c=1}^{C_{mg}} \Bigg[\log\binom{M_{mgc}+\phi_m-1}{M_{mgc}} \\
&+ M_{mgc}\log\left(\frac{n_{mc}a_g\gamma_{mc}2^{\mu_{gc}}}{n_{mc}a_g\gamma_{mc}2^{\mu_{gc}}+\phi_m}\right) + \phi_m\log\left(\frac{\phi_m}{n_{mc}a_g\gamma_{mc}2^{\mu_{gc}}+\phi_m}\right)\Bigg] \\
&+ \sum_{p=1}^{D_2}\sum_{g=1}^{G_p}\sum_{c=1}^{C_{pg}} \left[-\frac{1}{2}\log(2\pi\tau_p^2) - \frac{(P_{pgc}-(b_g+\kappa_{pc}+s_g(\mu_{gc}+R_{gc})))^2}{2\tau_p^2}\right].
\end{aligned}
\tag{3.4}
$$

Note that the representation of the mRNA likelihood is using the alternative parameterization, as described in the Likelihood section.

Inference Using Stan

We fit this model using the probabilistic programming language Stan [27], which uses No-U-Turn Sampling (NUTS) to perform fully Bayesian inference with a flexible probabilistic framework. NUTS, an extension of Hamiltonian Monte Carlo (HMC), automatically adjusts number of steps and step size at each iteration to draw samples from the posterior distribution more efficiently, as tuning required for step size and number of steps in HMC is not needed using NUTS [28]. We use 3000 iterations for sampling and 5 chains, with 2000 warm-up iterations. While maximum $\hat{R}$ is higher than that of the bulk model, it is still acceptable at 1.1. There are no convergence-related warnings, such as divergent transitions or saturation of maximum tree depth. Trace plots for μ and R in randomly selected genes and cell types are located in the Appendix.

3.2.6 Across cell type correlations and identifying candidate genes for high PTR

To understand which genes are associated with elevated post transcriptional regulation, we explore the posterior distribution of R_{gc} within and across cell types c, as this parameters recovers rPTR. We can identify genes with potentially high post transcriptional regulation by comparing R_{gc} to 0, as both μ_{gc} and $\mu_{gc} + R_{gc}$ are centered at 0 and represent log intensities relative to each respective gene-level mean. Thus, extreme R_{gc} represents rPTR of interest to reflect elevated post transcriptional regulation. Additionally, we consider the correlation of μ_{gc} and $\mu_{gc} + R_{gc}$ across cell types c so as to identify genes with low mRNA, protein correlation. If a cell type or set of cell types within a gene have a high degree of post transcriptional regulation, this correlation may be reduced, as the amount of mRNA transcribed is not proportional to the amount of protein synthesized for such cell types.

Identifying genes, cell types with nonzero rPTR

Using posterior samples of R, we can identify genes with R significantly different from zero. Consider posterior sample R_{igc}, representing draw i from the posterior distribution of R for gene g and cell type c. Additionally, consider mean $\bar{R}_{i-gc}$, representing the average posterior draws across all genes other than g for set of samples i within cell type c. We can compute the probability that the sign of R_{gc} is different from that of $\bar{R}_{gc}$ first using,

$$\mathbb{1}_{igc} = \begin{cases} 1 & \text{if } \text{sgn}(R_{igc}) == \text{sgn}(\bar{R}_{i-gc}) \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

Finally, p_{gc} is computed using the average of $\mathbb{1}_{igc}$ across posterior samples i, rep-

resenting the posterior probability that R_{gc} has the same sign. Using these posterior samples, we can also compute false discovery probabilities, so as to compute a Bayesian false discovery rate [71]. Let $\bar{R}_{gc}$ represent the posterior mean of R_{gc}. The posterior exclusion probability is computed as $\underset{i}{\text{mean}}(\text{sgn}(R_{igc}) \neq \text{sgn}(\bar{R}_{i-gc}))$. By computing the cumulative mean across genes g ordered by posterior exclusion probability, we can recover the expected proportion of false discoveries by number of significant genes for each cell type c.

mRNA, protein correlation across cell types

In addition to identifying genes with R significantly different from zero, we can identify genes with across cell types correlation of mRNA and protein significantly different from that of all other genes.

We compute the correlation between μ_{gc} and $\mu_{gc} + R_{gc}$ across cell types for each gene as opposed to using the empirical correlation of M_{mgc} and P_{pgc}, as this correlation can vary widely across genes as a result of both post transcriptional regulation and sampling variability. Due to a high degree of missingness on the cell and cell-type levels, empirical estimates of the correlation across cell types for individual genes are prone to noise and attenuates true correlation. By implementing the Bayesian model, we can obtain more reliable posterior estimates by utilizing observations from replicates in multiple data sets, as well as entire posterior distributions to use for further analysis.

Let $\rho_{ig} = \text{Cor}(\mu_{igc}, \mu_{igc} + R_{igc} | i, g)$ represent the correlation of posterior draws corresponding to mRNA and protein, respectively, across cell types for gene g and set of posterior samples i. Additionally, let $\rho_{i-g} = \text{Cor}(\bar{\mu}_{i-gc}, \bar{\mu}_{i-gc} + \bar{R}_{i-gc} | i, g)$ be the correlation across cell types of average of μ and $\mu + R$ across all genes other than g for cell type c and set of posterior samples. Next, we compute $\mathbb{1}_{ig}$ as,

$$\mathbb{1}_{igc} = \begin{cases} 1 & \text{if } \rho_{ig} >= \rho_{i-g} \\ 0 & \text{otherwise ,} \end{cases} \tag{3.6}$$

and compute the posterior probability that the correlation of mRNA and protein across cell types for gene g is greater than that of other genes as $p_{gc} = \underset{i}{\text{mean}}(\mathbb{1}_{igc})$. We consider posterior probabilities with low correlation as significant. We can similarly compute false positive probabilities for each gene using the cumulative mean of the posterior exclusion probability.

3.2.7 Functional gene set analysis

While exploring rPTR at the gene level can provide valuable insights into post transcriptional regulation across cell types, it is also of interest to evaluate rPTR within functional sets of genes, involved in carrying out unified biological processes. This can provide increased statistical robustness, as we consider groups of genes instead of individual genes. Additionally, we may expect that genes that work together to carry out unified functions may have similar behavior with respect to post transcriptional regulation. We leverage draws from the posterior distributions of μ and R and can run several types of tests.

Identifying groups, cell types with extreme rPTR relative to all other genes

We can compare the posterior samples of R_{gc} among genes in each ontology group to posterior samples among all other genes for each cell type by computing the posterior probability of obtaining a sample mean across all genes as extreme as that across genes belonging to each GO group and cell type. For each set of posterior samples i, GO group

G, and cell type c, consider mean,

$$\bar{R}_{iGc} = \frac{\sum_{g \in G} R_{igc}}{n_G},$$

where n_G is the number of genes observed in ontology group G.

Next, consider posterior samples R_{i-Gc}, representing draws of R within cell type c corresponding to all genes not associated with GO group G. Let there be n_{-G} such genes, and compute average,

$$\bar{R}_{i-Gc} = \frac{\sum_{g \notin G} R_{igc}}{n_{-G}}.$$

Next define indicator $\mathbb{1}_{iGc}$ as,

$$\mathbb{1}_{iGc} = \begin{cases} 1 & \text{if } \bar{R}_{iGc} \leq \bar{R}_{i-Gc} \\ 0 & \text{otherwise ,} \end{cases} \tag{3.7}$$

By computing $p^{'}_{Gc} = \underset{i}{\text{mean}}(\mathbb{1}_{iGc})$ and $\min(p^{'}_{Gc}, 1 - p^{'}_{Gc})$, we can obtain the posterior exclusion probability corresponding to R_{Gc} being more extreme than the average across all other groups. We use the cumulative mean of the ordered posterior exclusion probabilities to compute the posterior expected proportion of false discoveries.

Across cell types mRNA, protein correlation

Rather than considering rPTR with respect to individual cell types, we can examine the correlation of mRNA and protein across cell types. If a gene, or genes within an ontology group, are regulated mostly transcriptionally, $\mu_{gc}, \mu_{gc} + R_{gc}$ will have high correlation across cell types. Alternatively, if that gene is regulated mostly post transcriptionally mRNA and protein will have low correlation across cell types. Thus, we focus the at-

tention of this test to genes with low correlation across cell types. To determine which ontology groups have low correlation, consider correlations,

$$\rho_{iG} = \text{Cor}(\bar{\mu}_{iG}, \bar{\mu}_{iG} + \bar{R}_{iG}) = \text{Cor}\left(\frac{\sum_{g\in G}\mu_{igc}}{n_G}, \frac{\sum_{g\in G}\mu_{igc} + R_{igc}}{n_G}|i, g, n_g\right),$$

and,

$$\rho_{i-G} = \text{Cor}(\bar{\mu}_{i-G}, \bar{\mu}_{i-G} + \bar{R}_{i-G}) = \text{Cor}\left(\frac{\sum_{g\notin G}\mu_{igc}}{n_G^C}, \frac{\sum_{g\in G}\mu_{igc} + R_{igc}}{n_G^C}\right),$$

where n_G^C is the number of genes not assigned to GO group G. Define indicator $\mathbb{1}_{iG}$ as,

$$\mathbb{1}_{iG} = \begin{cases} 1 & \text{if } \rho_{iG} >= \rho_{i-G} \\ 0 & \text{otherwise ,} \end{cases} \tag{3.8}$$

We compute $p_G = \underset{i}{\text{mean}}(\mathbb{1}_{iG})$ across posterior samples as the posterior probability that the correlation across cell types among genes in group G is greater than that of all other genes. Additionally, we can use the cumulative mean of this probability across groups to compute the false discovery rate corresponding to each functional gene set.

Testing Scale Parameter

The final type of functional gene set testing involves testing the scale parameter across gene sets. This test intends to examine the degree to which significant GO groups from the two previous tests are affected by difference in scales between modalities. Specifically, we are interested in groups with an extreme scale parameter relative to genes not in each group. Let $\bar{t}_{iG}$ be the sample mean of $\log(s_{ig})$ across genes associated with GO group

G for set of posterior samples i. Similarly, let $\bar{t}_{i-G}$ be the average of $\log(s_g)$ across an equal number of randomly sampled genes not belonging to group G posterior samples i. Define indicator $\mathbb{1}_{iG}$ as,

$$\mathbb{1}_{iG} = \begin{cases} 1 & \text{if } \bar{t}_{iG} <= \bar{t}_{i-G} \\ 0 & \text{otherwise .} \end{cases} \tag{3.9}$$

Next, we define $p_G^{'} = \underset{i}{\text{mean}}(\mathbb{1}_{iG})$ and $p_G = \min(p_G^{'}, 1 - p_G^{'})$, which represents the posterior probability that the sample mean of the log scale parameter across genes associated with group G is more extreme than that of all other genes. Using this posterior exclusion probability, we can compute the false discovery rate across groups as a function of the number of significant GO groups as the cumulative mean of p_G in increasing order.

3.3 Results

3.3.1 Increased mRNA, protein correlations across cell types

Application of the Bayesian model shows a marked improvement of correlations between mRNA and protein across cell types for individual genes, in that correlations are generally higher across cell types and there are fewer genes with negative correlation across cell type. As the model leverages information across data sets and estimates variability contributed by biological processes, technical noise and sampling variability, estimates of underlying gene expression and protein intensities are on the same scale as one another and are less noisy than that computed with empirical values. As figure 3.4 displays, correlations computed on inferred parameters are generally higher than that of empirical estimates, and there is a clear reduction in genes with estimated negative correlation between mRNA and protein, which is unrealistic to expect from a biological

perspective and is likely a result of noise. Figure 3.4 displays results for one protein population. Results for the remaining protein data sets can be found in the Appendix.

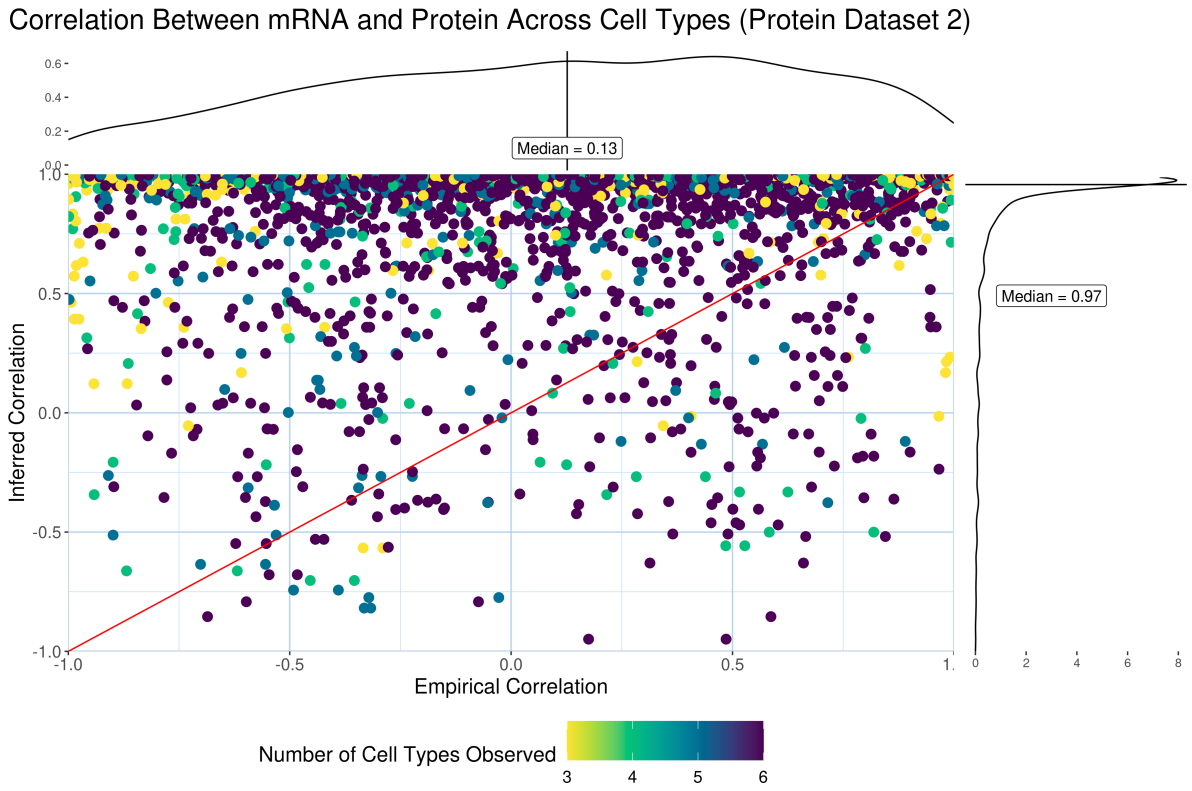

Figure 3.4: mRNA, Protein Correlation Across Cell Types. Correlations across cell types using the posterior mean of each gene, cell type are plotted against empirical across cell type correlations using protein data set 2. Point color represents the number of cell types observed for each gene, and a red diagonal line displays $y = x$. Opposite each axis label is a marginal density plot of corresponding correlations, with vertical lines denoting the labeled median value.

3.3.2 Identifying genes with potentially high post transcriptional regulation

Identifying genes, cell types with nonzero rPTR

To identify genes with nonzero rPTR, we use posterior samples of R_{gc} from the Bayesian model to identify which genes, if any, consistently have a sign different than the

average across all other cell types. As R_{gc} is defined relative to the average $\log_2$ intensity in gene g, R_{gc} of 0 can indicate an equal ratio of protein intensity and transcription rate, suggesting that the gene may not be post transcriptionally regulated.

In figure 3.5 we can see posterior means for mRNA expression and protein abundance parameters μ_{gc} and $\mu_{gc} + R_{gc}$, with colored points to indicate genes significant at a 1% false discovery rate in each cell type. Information on how FDRs are computed is provided in Methods.

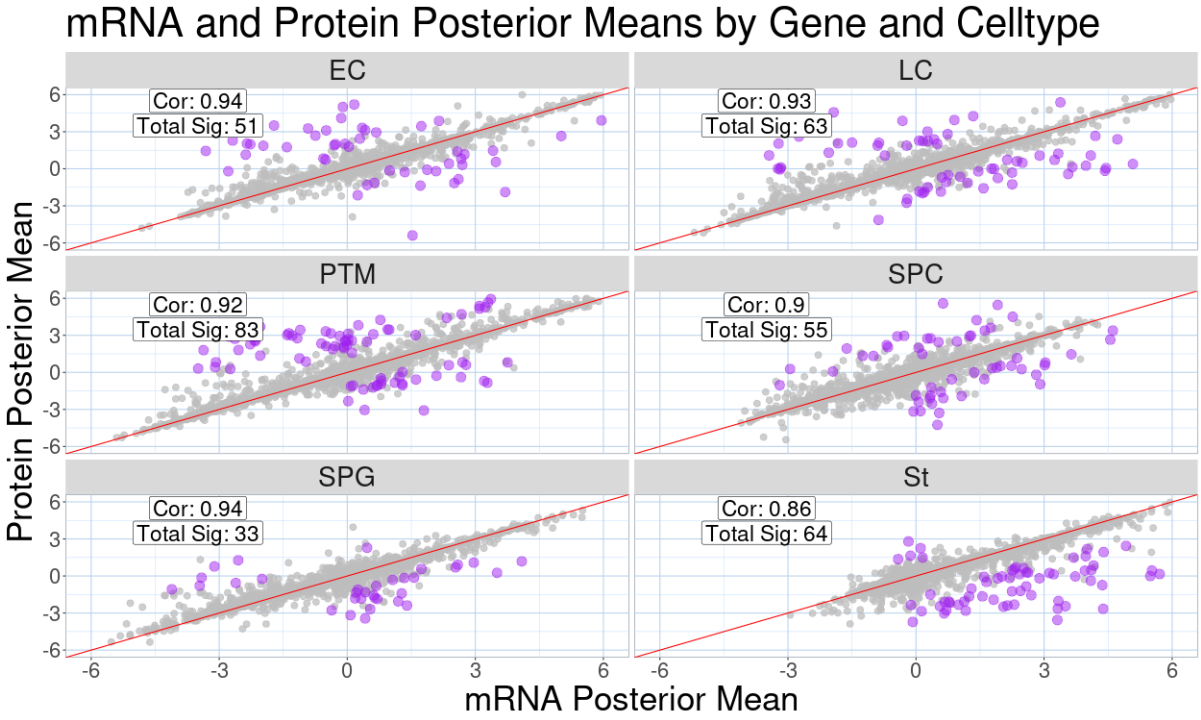

Figure 3.5: Posterior Means for $\log_2$ relative mRNA expression, and Protein abundance per gene and cell type. Protein posterior mean, $\mu_{gc} + R_{gc}$, is plotted by mRNA posterior mean, μ_{gc}. Purple points denote genes significantly different from zero at a 1% FDR. We can see the number of these points labeled as "Total Sig", as well as the correlation across genes of these posterior averages.

For a select list of genes, figure 3.6 displays a heatmap with the posterior mean of R for each gene and cell type. Included among significant genes is a gene relating to zinc finger protein, which is directly involved with protein folding and transcriptional

activation, among other essential cellular functions [36]. In addition, at least two genes relating to myosin are significant, which are filaments of muscle cells involved in motion of other cells. Both actin and myosin are essential in facilitating spermatogenesis [72].

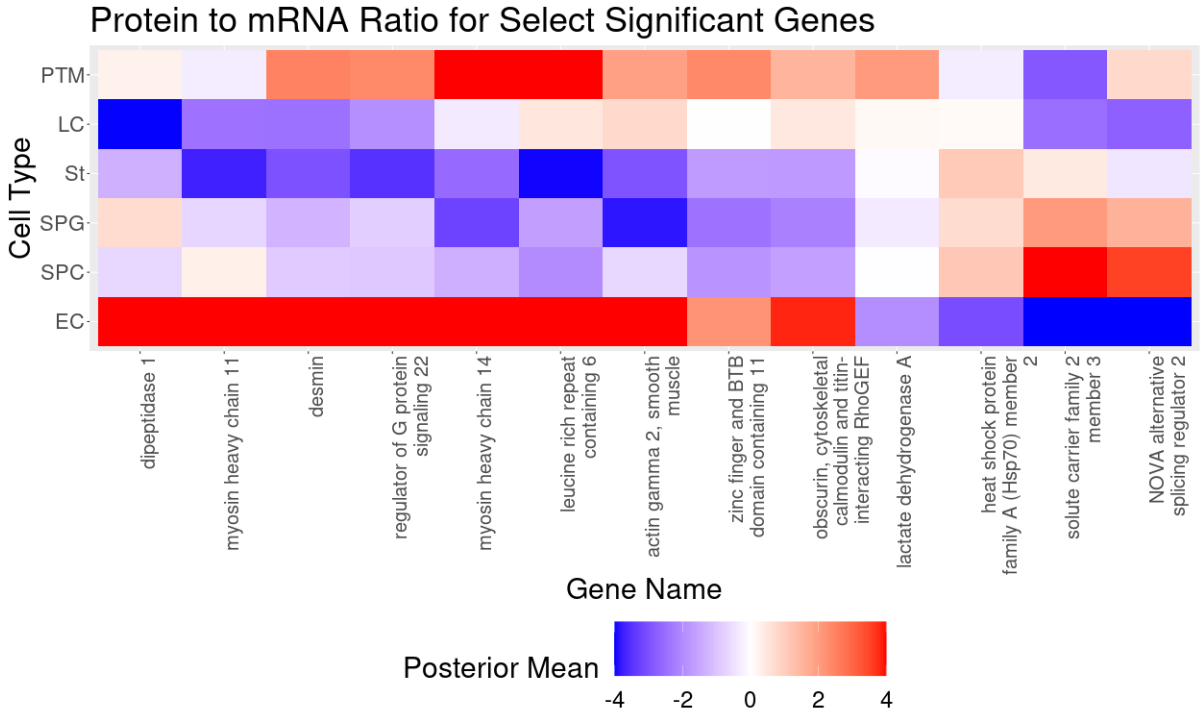

Figure 3.6: Posterior Mean of Ratio for select genes with nonzero rPTR in at least one cell type. Genes with the lowest posterior exclusion probability are shown in this plot, and color reflects the average posterior mean of R among each cell type and genes in each group.

With posterior samples of R we are able to compute false discovery statistics, such as the number of significant genes with a 1% false discovery rate (FDR) by cell type. Table 3.6 and figure B.2 in the Appendix present some summary information relating to false discovery rates.

Table 3.6: Significant Genes by Cell Type. Number of significant genes is shown for each cell type, at 1% and 0.1% false discovery rate.

Cell Type	**Significant Genes at** 1% **FDR**	**Significant Genes at** 0.1% **FDR**
EC	51	22
LC	63	14
PTM	83	25
SPC	55	9
SPG	33	9
St	64	31

Across cell types mRNA, protein correlation

In addition to examining genes based on whether R_{gc} is significantly different from zero, we can evaluate R based on the mRNA, protein correlation within each gene across cell types. Genes with low $\mu_{gc}, \mu_{gc} + R_{gc}$ across cell types c may have increased post transcriptional regulation, as a disproportionate amount of protein is synthesized compared to mRNA for one or more cell types. Thus, this test identifies genes with low correlation by considering significance based on a 1% FDR using the posterior probability that the mRNA, protein correlation across cell types is larger than that of all other genes.

The total number of genes with significant low correlation are shown in table 3.7. Additionally, figure 3.7 shows a heatmap of the posterior mean of R_{gc} for select genes with significant test results.

Table 3.7: Significant genes for correlation test. Displayed for 1% and 0.1% false discovery rate.

Significant Genes at 1% **FDR**	**Significant Genes at** 0.1% **FDR**
231	142

Included among significant genes are several genes relating to enzyme production and function, including alpha l iduronidase and ubiquitin conjugating enzyme. Additionally, one gene relating to zinc finger protein has significant low correlation across cell types.

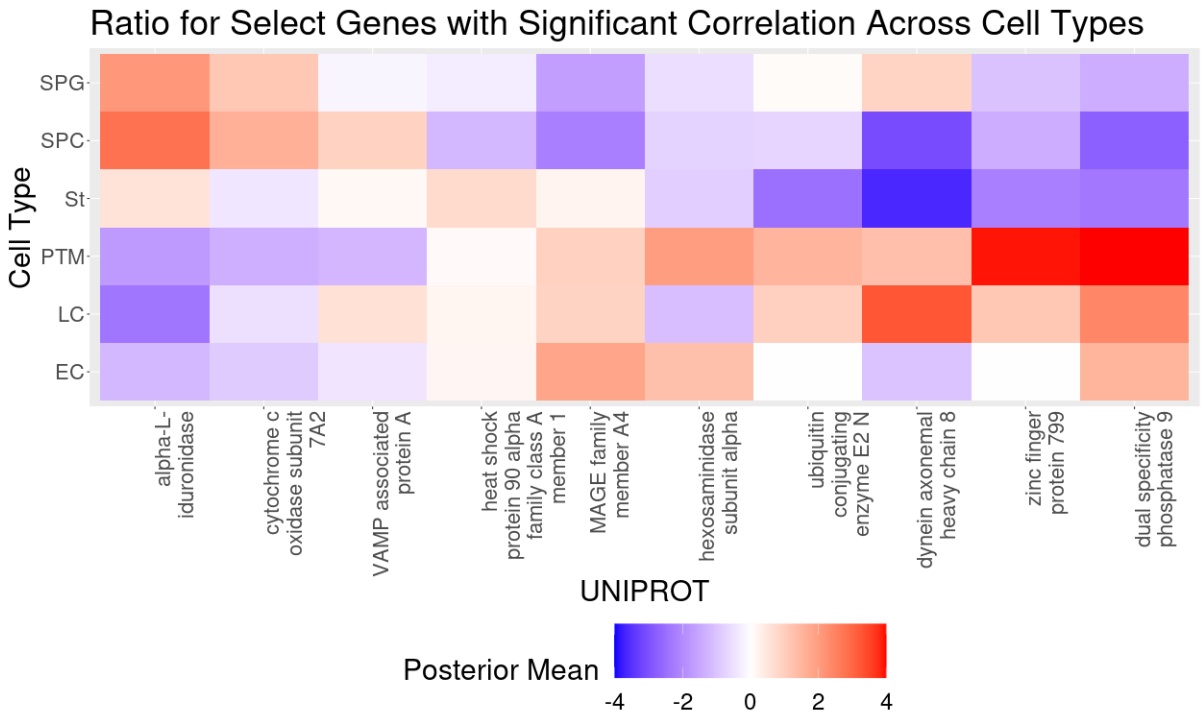

Figure 3.7: Average of posterior mean of R among genes with significant low mRNA, protein correlation across cell types. Averages are computed using the mean of posterior means across genes belonging to each group for each cell type.

3.3.3 Testing protein to mRNA ratio within functional gene sets

Considering sets of functionally related genes, which work together to perform coordinated biological processes, can help uncover biological functions that are affected by post transcriptional regulation. Additionally, this approach offers statistical power that otherwise may not be present by testing each gene individually. We can approach this testing by examining R_{gc} within each gene ontology (GO) group and cell type, as well as the correlation of μ_{gc} and $\mu_{gc} + R_{gc}$ across cell types for each GO group.

Identifying gene sets with extreme rPTR relative to all other genes

First, we test the distribution of R_{gc} for each functional gene set individually for each cell type, by computing the probability that the average R_{gc} among genes in each GO group is as extreme as that of all other genes. Please see methods for additional details on the mechanics of this test. Below a brief summary is displayed of the number of significant GO groups in each cell type.

Table 3.8: Significant GO Groups by Cell Type. Displayed at 1% and 0.1% false discovery rate for each cell type.

Cell Type	**Significant Groups at** 1% **FDR**	**Significant Groups at** 0.1% **FDR**
EC	127	45
LC	163	40
PTM	206	56
SPC	132	27
SPG	91	19
St	189	58

Below we can see posterior means for genes belonging to select significant groups, as well as the distribution of posterior means of R for all genes. As this test is run separately for each cell type, we will examine results for select cell types. Please see the Appendix for results corresponding to remaining cell types.

In figure 3.8, we have posterior means for the genes associated with each significant functional gene set for peritubular myoid (PTM) cells. These cells are smooth muscle cells that surround the seminiferous tubules in the testis. There appear to be significant groups with both high and low R relative to the average across all genes. We can see the empirical cumulative distribution function corresponding to all genes above the tick marks, as well as the posterior probability corresponding to each group.

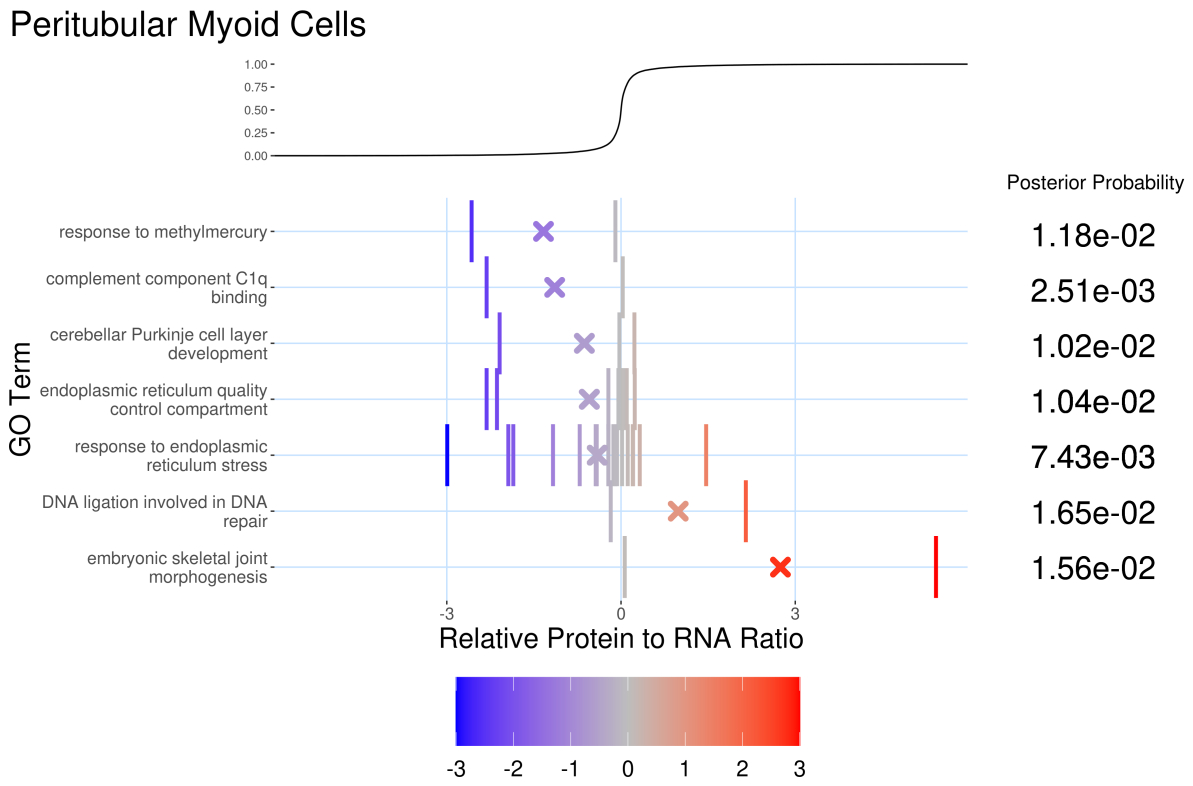

Figure 3.8: Genes in Significant Groups for Cell Type PTM. Each tick represents the posterior mean of R_{gc} for each gene assigned to groups on the y-axis. Color also represents posterior mean, and an x point represents the group-level mean across genes. The posterior exclusion probability corresponding to each group is displayed, as well as the cumulative distribution function of R across all genes.

Note that, included among these significant groups, are at least two groups with function relating to the endoplasmic reticulum. This organelle is essential to protein folding, and is the location where post transcriptional modifications often take place [73].

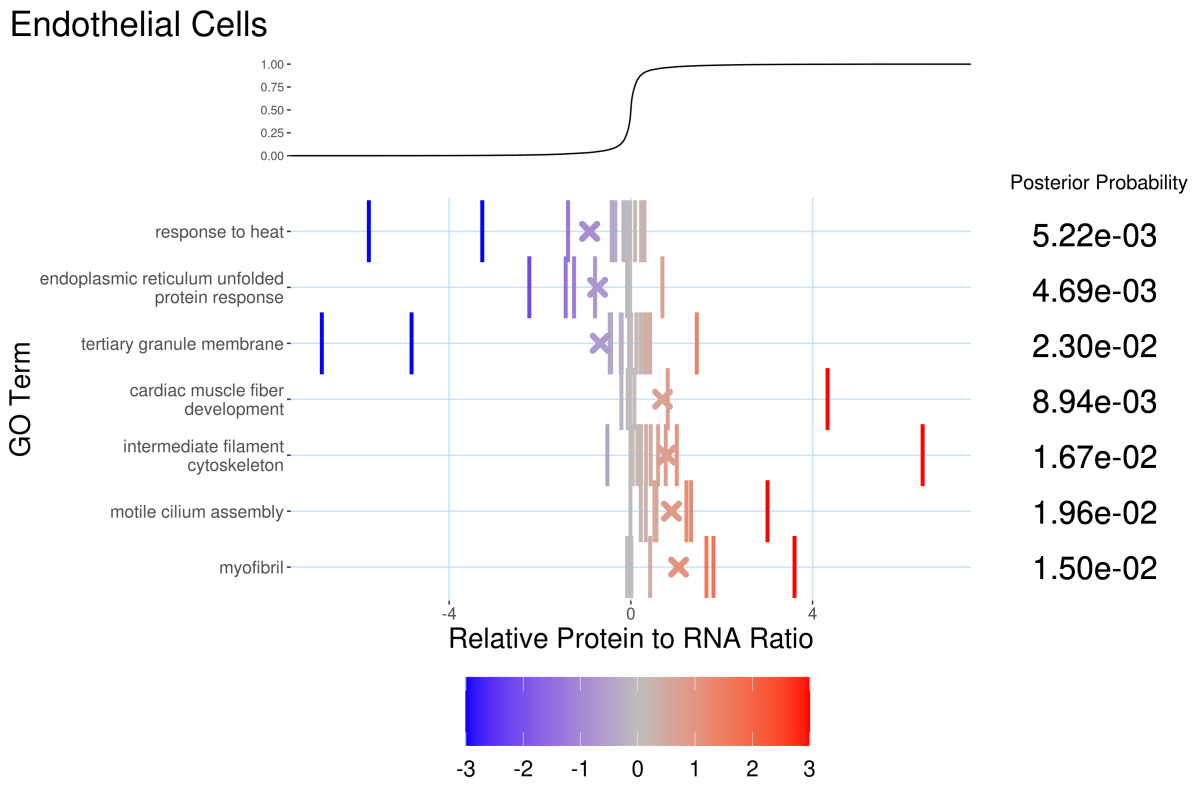

Figure 3.9: Genes in Significant Groups for Cell Type EC. The posterior mean of R_{gc} is shown for each gene as a member of each of the selected significant GO groups. An x labels each group-level mean. Color also represents posterior mean of R, and posterior probability for each group is depicted on the right. The upper panel also shows the cumulative distribution function of R across genes.

In figure 3.9, we can see posterior means corresponding to each gene belonging to significant GO groups for endothelial cells, EC. Endothelial cells line blood vessels and are essential to the regulation of exchanges between the bloodstream and surrounding tissues. For cell type EC, more significant groups have significant positive R than negative R. In addition to GO groups relating to the endoplasmic reticulum, as present among PTM results, a group relating to motile cillium assembly is present. Motile cillia are present in spermatozoa and are involved in the movement of fluid. A defect in motile cillia results in a set of diseases known as motile ciliopathies [74]. Additionally, several groups relating to protein function are significant, including myofibril and intermediate

filament cytoskeleton.

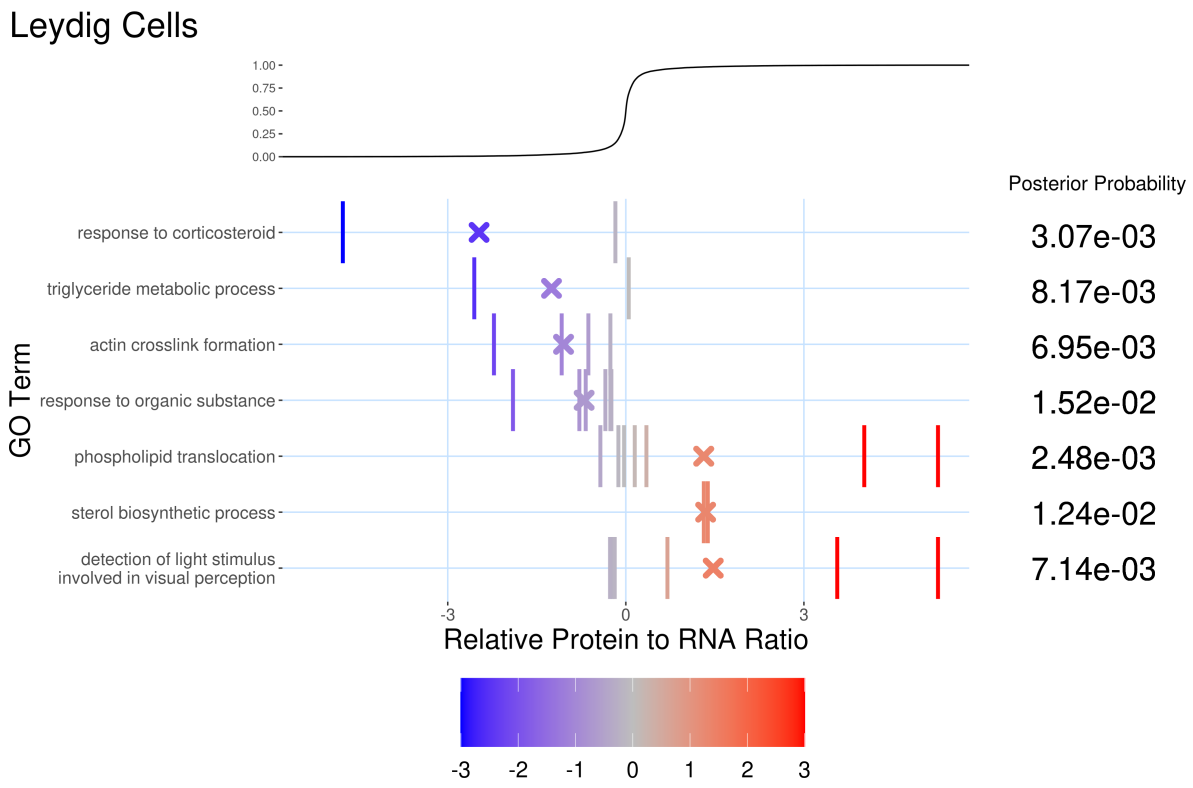

Figure 3.10: Genes in Significant Groups for Cell Type LC. The posterior mean of R_{gc} is shown for each gene as a member of each of the selected significant GO groups. For each GO group, x labels the average posterior mean across genes belonging to each group. Color also represents posterior mean of R, and posterior probability for each group is depicted on the right. The upper panel also shows the cumulative distribution function of R across genes.

Finally we have posterior means for significant GO groups among cells assigned to the Leydig Cell Type (LC) in figure 3.10. These cells are particularly important for sperm production and the maintenance of sexual characteristics, as they are directly responsible for producing testosterone. Included among significant groups are gene sets relating to hormone production, including response to corticosteroid and sterol biosynthetic process, both types of steroids.

We compare significant groups from this testing to groups with significant rPTR

from testis tissue in the bulk analysis. Several groups, including groups associated with ubiquitin processes, translation, protein binding, and ribosomal function, were significant in both analyses.

Across cell types correlation of mRNA, protein

In addition to testing R by cell type, we test the correlation between mRNA and protein, $\mu_{gc}, \mu_{gc} + R_{gc}$, across cell types c for the average R_{gc} across genes belonging to each GO group. This allows us to uncover functional sets with low correlation across cell types, which may indicate elevated post transcriptional regulation and can subsequently affect the biological process associated with genes in each group. Please see methods for additional details on test procedure. Table 3.9 displays summary information about significant groups.

Table 3.9: Significant GO Groups for Correlation Test at 1% and 0.1% false discovery rate.

Significant at 1% **FDR**	**Significant at** $0.1\% FDR$
8	2

In figure 3.11, we can see the mRNA, protein correlation corresponding to posterior means for genes in select significant functional gene sets with low protein, mRNA correlation. Additionally, we can observe the posterior probability corresponding to each of these groups and the cumulative distribution function for this correlation across genes. We can see that groups relating to phospholipids are significant, which were present in 3.10. Interestingly, several groups relating to protein synthesis are present as well, including peptide loading and SUMO transferase activity. Proteins in the small ubiquitin-related modifier (SUMO) pathway directly regulate cellular processes including post transcriptional regulation [75].

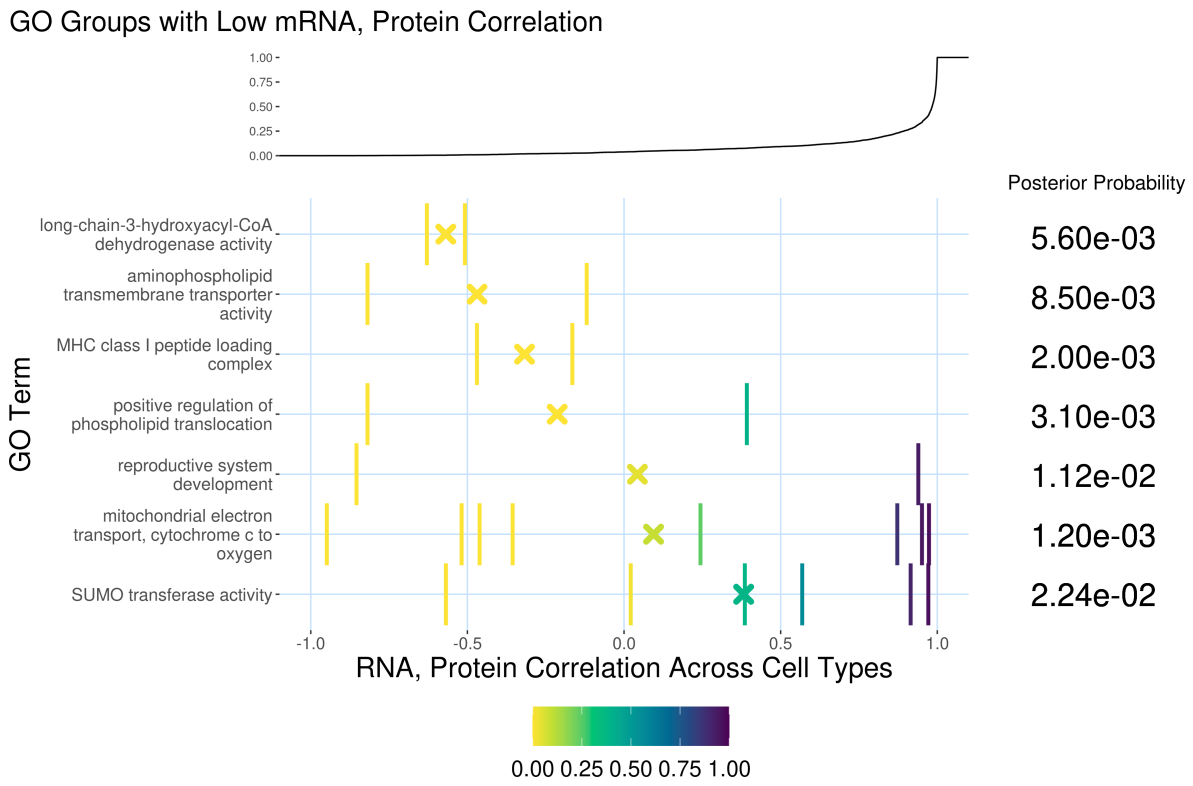

Figure 3.11: Genes in Groups with Significant Low Correlation. Each tick represents the correlation across cell types of the posterior mean of μ and $\mu + R$ among genes in each group. An x corresponds to the group-level mean of the scale parameter. Color also represents correlation across cell types. Additionally we can see the cumulative distribution of the across cell types correlation across genes, as well as the posterior exclusion probability for each group.

Testing the distribution of the scale parameter for each GO Group

Since the scaling parameter is responsible for ensuring that μ_{gc} and $\mu_{gc} + R_{gc}$ are on comparable scales between different modalities, we test the log scale parameter among genes in each ontology group to recover any groups that may have extreme scale relative to other groups and significant results in other tests. Ideally, it will be the case that genes or groups of genes are not significant simply because of poor scaling between mRNA and protein. Figure B.4 in the Appendix displays the posterior mean of the log scale parameter across genes. In table 3.10, we can see a summary of significant groups

corresponding to this test. We consider significant groups as those with a low posterior probability that the mean log scale in each group is more extreme than that of a randomly sampled group of equal size. Using this posterior exclusion probability, we can recover groups with extreme log scale at 1% FDR.

Table 3.10: Significant GO Groups for Log Scale Test and 1% and 0.1% false discovery rate.

Significant at 1% **FDR**	**Significant at** 0.1% **FDR**
3	0

In figure 3.12, we can see the posterior mean for the scale parameter among genes corresponding to select significant groups. While one group has an extreme low log scale, two groups have extreme high scales.

As the intention of testing the scale parameter is to identify which groups may have significant results for other tests as a result of scale difference between modalities, we compare the list of groups with significant scale parameter to the significant groups from other GO tests. The platelet aggregation group, in addition to having a significant scale, is significant for and extreme R_{gc} in all cell types. Meanwhile, the GO group SRP-dependent cotranslational protein targeting to membrane is significant only among spermatagonia. None of the groups with significant scale parameters also have significant across cell type mRNA, protein correlation.

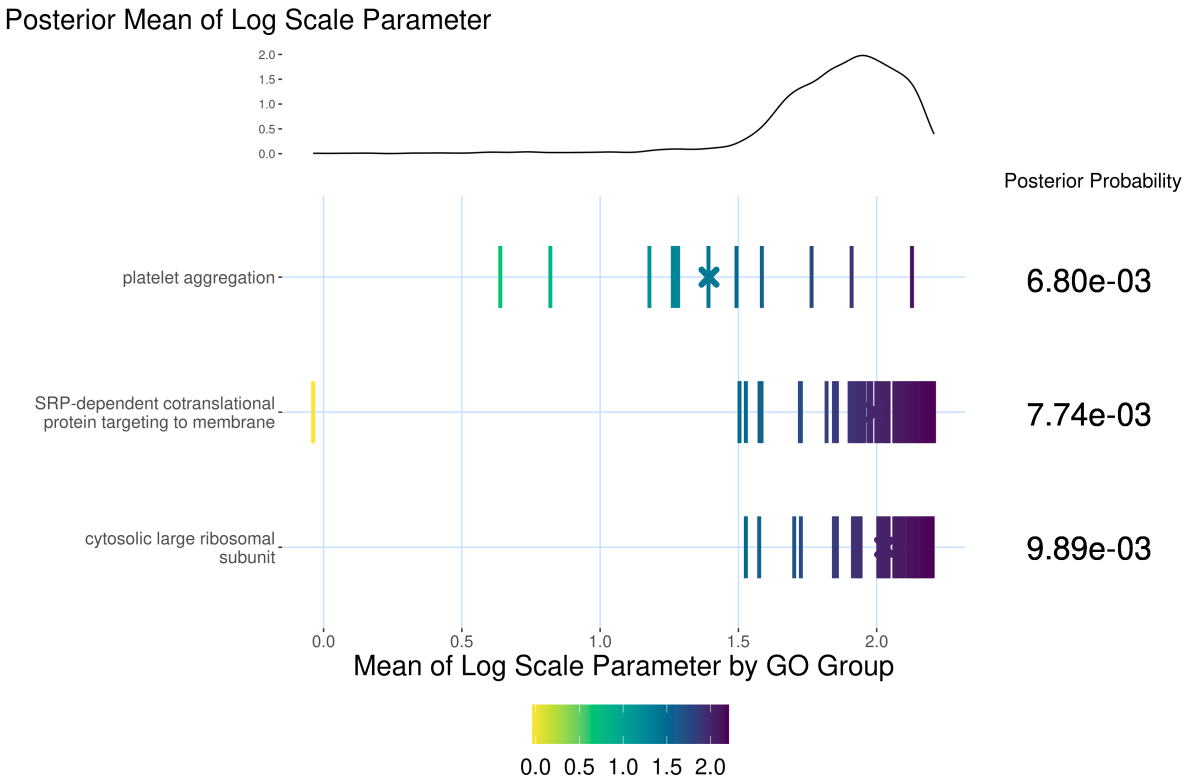

Figure 3.12: Genes in Groups with Significant Scale. Each tick represents the posterior mean of the scale parameter among genes in each group. Color represents posterior mean of scale parameter as well, with "x" marks for each group-level mean. The posterior probability for each group is shown on the right. The probability density function for the posterior mean of scale is shown as well.

3.4 Posterior Predictive Checking

To examine model fitness relative to observed data, we use posterior predictive checks (PPCs). PPCs are a technique for assessing the goodness-of-fit of a Bayesian model by comparing observed values and test statistics corresponding to those observed values to the distribution of such test statistics based on the data generated by the posterior predictive distribution. The posterior predictive distribution is obtained by simulating data using the same likelihood and model parameters as that assigned to the observed data. The simulated data is then compared to the observed data to evaluate the fit of

the model.

Tables 3.11 and 3.12 denote the proportion of genes with observed averages within the 95% posterior interval corresponding to posterior predictive simulated values by cell type and data set. High coverage of observed values is maintained for all cell types and populations, indicating that the model is successfully replicating observed values.

Table 3.11: Proportion of observed mRNA gene, cell type averages within 95% posterior predictive interval. Using upper and lower quantiles based on posterior predictive draws, the proportion of observed mRNA data set values within these bounds are presented.

Cell Type	mRNA Data set 1	mRNA Data set 2
EC	0.992	0.997
LC	0.989	0.995
PTM	0.992	0.996
SPC	0.994	0.996
SPG	0.988	0.997
St	0.983	0.997

Table 3.12: Proportion of observed protein gene, cell type averages within 95% posterior predictive interval. Using upper and lower quantiles based on posterior predictive draws, the proportion of observed protein data set values within these bounds are presented.

Cell Type	Protein Pop1	Protein Pop2	Protein Pop3	Protein Pop4
EC	0.980	0.971	0.891	0.982
LC	0.952	0.978	0.966	0.953
PTM	0.917	0.961	0.967	0.957
SPC	0.984	0.981	0.985	0.988
SPG	0.979	0.991	0.983	0.991
St	0.989	0.977	0.743	0.982

Beyond simply comparing whether the observed values appear reasonable, we can also compare statistics of interest between the observed and simulated values, such as variance or correlation of the average log transcript counts and protein intensity across cell types.

Figure 3.13 displays the observed variance across cell types of average relative transcript counts for each gene individually, as well as 95% posterior intervals for the same quantity computed on simulated values. The variances of simulated values successfully resemble that of the of the observed values, with both high coverage of posterior intervals and correlation with the posterior mean. This indicates that the model is adequately capturing the range of variances seen in the observed data.

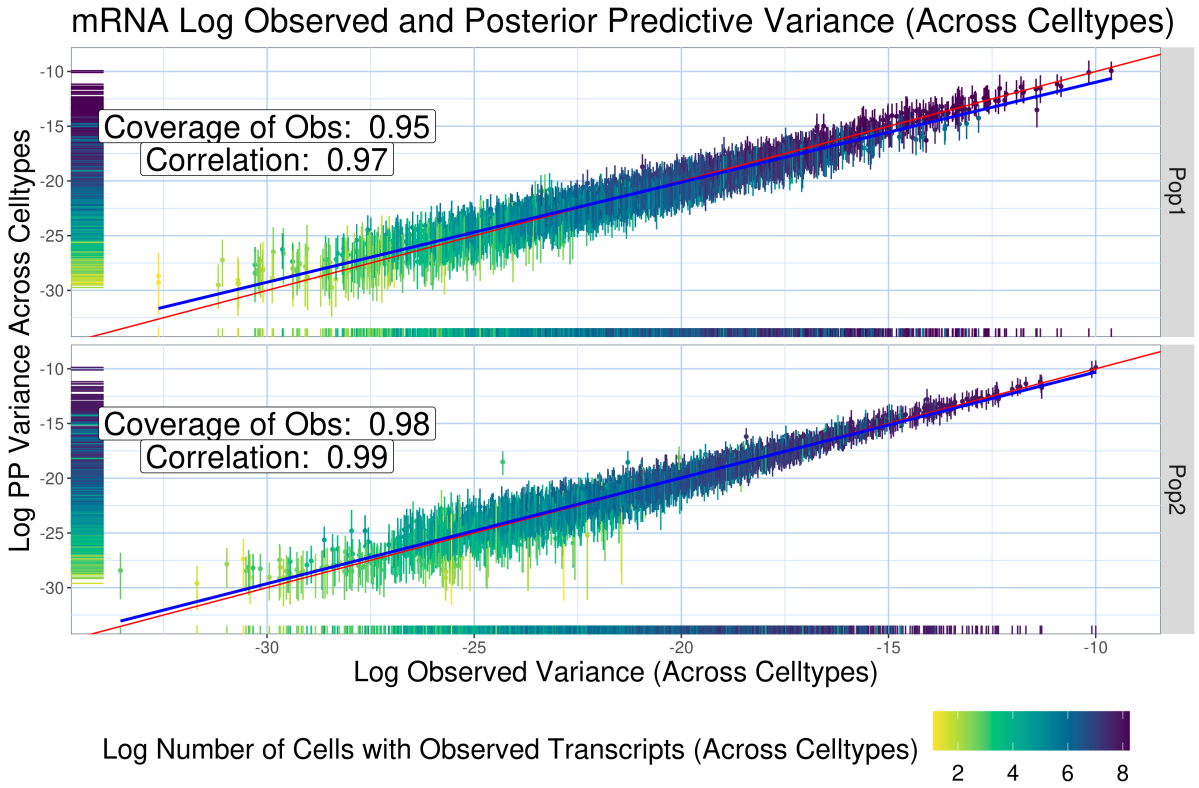

Figure 3.13: Observed and Posterior Predictive mRNA Variance Across Cell Types. The log variance of observed averages across cell types is represented on the x-axis, as well as the 95% interval for the across cell types variance of posterior predictive draws. Color represents number of observed cells, and a red line represents $y = x$. The coverage of observed values by the posterior predictive variance, as well as the correlation of observed variances and median posterior predictive variance are labeled.

We can examine this same quantity for the protein-level average across peptides in figure 3.14. We see a strong positive correlation between observed and posterior predictive

variances across cell types, with slight under coverage affecting proteins with only one peptide observed in particular.

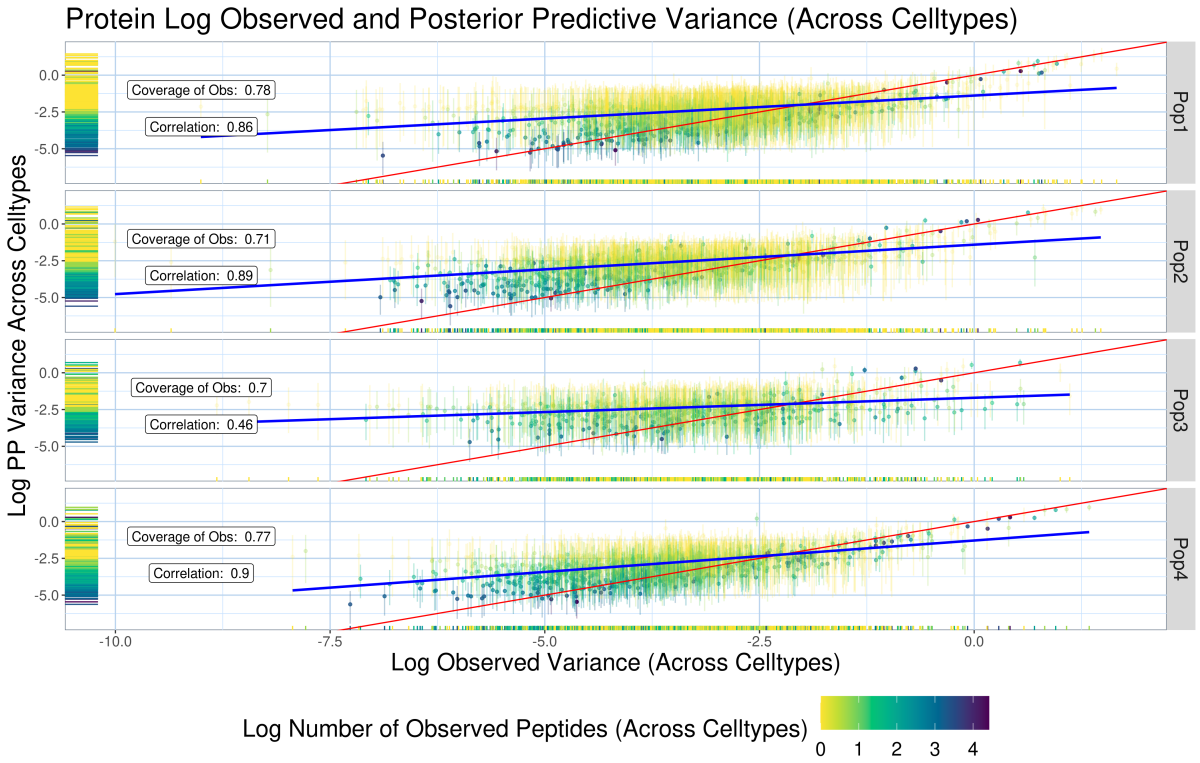

Figure 3.14: Observed and Posterior Predictive Protein Variance Across Cell Types. The log variance of observed averages across cell types is represented on the x-axis, as well as the 95% interval for the across cell types variance of posterior predictive draws. Color represents number of observed cells, and a red line represents $y = x$. The coverage of observed values by the posterior predictive variance, as well as the correlation of observed variances and median posterior predictive variance are labeled.

Additionally, we can compare posterior predictive correlations between mRNA and protein to that of the observed values to examine whether the model can replicate observed correlations. As figure 3.15 shows, the posterior predictive distribution of correlation across cell types generally encapsulates that of the observed data for most populations. While parameters corresponding to mRNA and protein $\log_2$ relative abundance, μ_{gc} and $\mu_{gc} + R_{gc}$ are recovering higher correlations across cell types, on average, the posterior predictive samples are reflecting sampling variability well enough to recover

observed correlations.

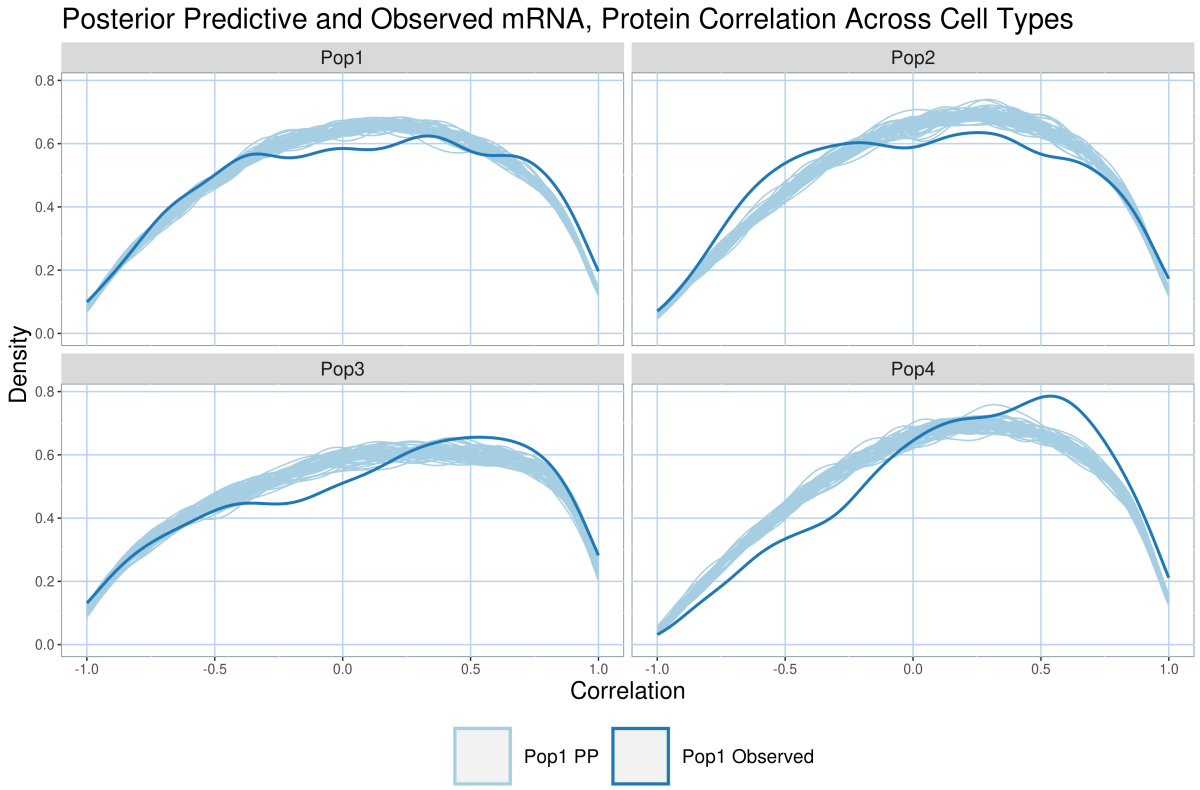

Figure 3.15: Observed and Posterior Predictive mRNA, Protein Correlation Across Cell Types. Density plots are displayed for observed correlation of mRNA, protein across cell types for all four protein populations and mRNA data set 1. Additionally we can see posterior predictive correlations across cell types for randomly selected posterior predictive draws.

To evaluate the relationship between correlation across cell types for observed values and that of posterior predictive draws, figure 3.13 depicts the proportion of observed correlations within the 95% interval for correlation across cell types among posterior draws. There is high coverage of observed correlations across both protein and mRNA data sets.

Table 3.13: Proportion of observed mRNA, protein correlation across cell types within 0.025 and 0.975 quantiles of posterior predictive correlation across cell types.

Protein Data set	mRNA Pop 1	mRNA Pop 2
Pop 1	0.954	0.952
Pop 2	0.944	0.95
Pop 3	0.92	0.913
Pop 4	0.962	0.962

3.5 Discussion

Using a Bayesian hierarchical model, we were able to estimate relative underlying mRNA and protein abundance, and ultimately quantify the effect of post transcriptional regulation at the gene-level by estimating rPTR. In addition to examining rPTR for individual genes and cell types, we consider sets of functionally related genes to identify pathways potentially affected by post transcriptional regulation. Through the exploration of posterior samples, we have identified functional gene sets with low mRNA, protein correlation across cell types and extreme rPTR within individual cell types.

This model is able leverage data from multiple data sets to better mitigate serious model challenges, including high noise and technical variability. While these sources of variation typically attenuate mRNA, protein correlation across cell types, inferred mRNA, protein correlations from our model are markedly higher than that of empirical estimates. Additionally, examination of posterior predictive draws indicates that the model well-captures variation present in observed data.

One potential improvement for this model is a more strategic approach at missing data, in particular the MNAR qualities of single cell RNA-seq and protein data. While protein technologies exist to identify whether an observation is missing at random as the result of a peptide not being identified, and peptides not identified as a result of being below the limit of detection and thus MNAR, transcriptomics data currently includes

no mechanism to indicate missingness source. Thus, some observations are "true zeros," as a gene is not being expressed in a particular cell, and other observations are zero as a result of transcript abundance being too low for measurement equipment to capture. Thus, improved approaches at modeling missingness is going to require careful modeling of mechanisms underneath missingness.

Another improvement of interest for this model is management of technical effects, in particular batch effect and effect of sequencing depth. Sequencing depth, the number of molecules sequenced in each cell, can confound biological and technical sources of variability [76]. While our model incorporates the total number of transcript counts across cells assigned to each type, n_{mc}, it does not account for this artifact at the cellular level and assumes that its effect is constant across cells assigned to the same type. Batch effect can also be accounted for in a way that more comprehensively accounts for its effects at the gene level [9]. In the following chapter, we evaluate the effect of two technical factor correction methods [76] [9] on model fit, to probe our model to areas of improvement with respect to this area, and better understand the effect of these technical factors on observed values.

Chapter 4

Assessing Sensitivity to Choice of Technical-Effect Correction Methodology

4.1 Introduction

While single-cell and high throughput gene product quantification offers unprecedented resolution to examine the transcriptome and proteome, technical factors can confound biological heterogeneity of interest. This includes batch effect, which can emerge as a result of collecting measurements for cells in separate batches, as quantifying transcript or peptide abundance for an entire tissue sample at a single-cell resolution is not reasonable for a large number of cells. This yields systemic biases among cells in each batch [77]. While our Bayesian model includes a batch-effect correction method at the cell type and population levels, biases associated with batch can even affect counts at the gene-level [9], which our model would fail to account for.

Sequencing depth, the coverage of the transcriptome or proteome that the measure-

ment collection process is able to capture successfully, is another technical factor that can potentially confound quantification of underlying biological processes. If a cell is sequenced with high depth, it means that a large number of features (gene products) are measured in that cell. Choices made during the data collection and sample preparation processes, including measurement instrument, prep reagent and read length, can affect downstream sequencing depth. As a result, the normalized expression in a cell-gene pair is correlated with sequencing depth such that genes expressed among cells with high sequencing depth appear to have higher expression relative to those with lower sequencing depth. Sequencing depth can also confound gene-level variance, as cells with higher sequencing depth show higher gene-level variance than that among cells with lower sequencing depth [8]. While our model uses total transcript count across cells assigned to each cell type to attempt to control for this technical factor, it does not capture the effect of sequencing depth at the gene or individual cell levels, which may be necessary to properly control for this effect.

In this chapter, we review and apply popular technical-effect correction methods, including ComBat [78], ComBat-seq [9], and SCTransform [8] to each of the mRNA data sets to examine the influence that technical factors have on results. We compare gene-level and group-level results from the Bayesian model applied to observed data to that of the corrected data to better understand the effect that technical factors and their correction have on model fit. Additionally, we use approximated leave-one-out cross validation to compare the model applied to different technical-effect correction methods directly.

4.2 Review of Batch and Technical Effect Correction Methods

4.2.1 ComBat

We review ComBat [78], the first batch correction method developed of this group. ComBat was initially developed to address batch effect in bulk data, but was later applied to other data types, including scRNA-seq [79]. This method uses an empirical Bayes (EB) framework that pools information across genes in each batch to accomplish "shrinkage" of per-gene batch-effect estimates toward their across-genes average. In general, empirical Bayes refers to the estimation of a prior distribution of interest using data [80]. One of the most crucial assumptions to apply ComBat is normally distributed errors. Additionally, it is assumed that batch effects represent systematic, non-biologically motivated shifts in the per-gene mean and variance of all samples within a batch. The three main steps involved in ComBat batch effect correction are data standardization, batch effect parameter estimation with EB, and data adjustment.

To ensure that each gene will have similar mean and variance across tissues or cells, standardization is first performed. Let y_{ijg} represent observed transcript count for batch i, gene g and cell or tissue j (depending on whether data is measured within individual cells or across cells in a sampled tissue). Next, consider regression $y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijg}$ with errors $\epsilon_{ijg} \sim N(0, \sigma_g^2)$. α_g represents overall gene-level expression, X is a known matrix corresponding to sample conditions, and β_g is the vector of corresponding regression coefficients. Parameters γ_{ig} and δ_{ig} represent additive and multiplicative batch effects for each batch and gene. Using ordinary least squares estimates for these parameters, observations are scaled as, $Z_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g}{\hat{\sigma}_g}$.

In the next step, batch effect parameters are estimated using parametric empirical pri-

ors. ComBat assumes that standardized data is normally distributed, $Z_{ijg} \sim N(\gamma_{ig}, \delta_{ig}^2)$. The following priors are used for batch effect parameters, $\gamma_{ig} \sim N(Y_i, \tau_i^2)$ and $\delta_{ig}^2 \sim$ Inverse Gamma(λ_i, θ_i). Hyperparameters are estimated using the standardized data and method of moments. With these priors and the assumed normal likelihood for standardized observations, final batch-effect parameter estimates are computed as posterior means for γ_{ig} and δ_{ig}.

In the final step, data is adjusted to account for estimated batch effects. This amounts to reverse-transforming standardized observations using parameter estimates. Let γ_{ijg}^* represent batch-adjusted data. $\gamma_{ijg}^* = \frac{\hat{\sigma}_g}{\hat{\delta}_{ig}}(Z_{ijg} - \hat{\gamma}_{ig}) + \hat{\alpha}_g + X\hat{\beta}_g$.

ComBat, a robust and flexible EB framework to estimate and correct for the effect of batch, accounts for batch at the gene-level for each observation. This method results in batch-adjusted data, but unfortunately does not take on the original count structure of RNA-seq data, as a normality assumption is required. Additionally, as a mean-variance dependence in RNA-seq count data with over-dispersion has been identified [81], the application of this method intended for bulk analysis to single-cell data may not be appropriate. Next, we consider, ComBat-seq, an adapatation of ComBat to be applied to single-cell RNA-seq data.

4.2.2 ComBat-seq

ComBat-seq [9] is a batch correction method developed after ComBat for the specific purpose of being applied to RNA-seq data. Where ComBat models observed values as a regression with normally distributed errors and does not account for overdispersion, ComBat-seq applies Negative Binomial regression to reflect gene-specific overdispersion and the count structure of observed data.

ComBat-seq begins with a following gene-wise regression model. Let y_{gij} be transcript

counts in for gene g in column (cell if applied to single-cell observations) j in batch i. Assuming $y_{gij} \sim \text{NB}(\mu_{gij}, \phi_{gi})$, with μ_{gij} and ϕ_{gi} representing mean and dispersion parameters, respectively, the regression model is as follows,

$$\log(\mu_{gij}) = \alpha_g + X_j\beta_g + \gamma_{gi} + \log(N_j)$$
$$\text{var}(y_{gij}) = \mu_{gij} + \phi_{gi} + \mu_{gij}^2.$$

where α_g represents log expected counts, X_j is a known biological covariate, such as biological condition, and β_g is the covariate corresponding to X_j. Batch effect parameters are reflected by γ_{gi} and ϕ_{gi} within the mean and dispersion parameters, respectively. N_j corresponds to library size, or the total counts across genes in cell j. The purpose of this regression is to obtain estimates for batch effect parameters as well as the expected counts, $\hat{\mu}_{gij}$. Estimates are computed using Fisher scoring.

Data adjustment is performed next, using the assumption that adjusted data follows a "batch-free" negative binomial distribution. Let y^*_{gj} represent the adjusted data for gene g and cell j, with negative binomial parameters μ^*_{gj} and ϕ^*_g computed as,

$$\log(\mu^*_{gj}) = \log\hat{\mu}_{gij} - \hat{\gamma}_{gi}\phi^*_g = \frac{1}{\text{N}_{\text{batch}}}\sum_i \hat{\phi}_{gi}. \tag{4.1}$$

Adjusted data is generated as the closest quantile on the batch-free distribution to that of the original observed counts, y_{gij}. Zeros are preserved in this adjustment.

The preservation of the integer nature of count data and modeling of library size, an important strategy to account for difference in cell size as total number of transcripts counted in a cell, are important advantages of ComBat-seq. In the event that there are no actual differences between batches, ComBat-seq may inadvertently yield increased false positive rates. As performing modification on the data carries the risk of influenc-

ing estimation of underlying gene products, ComBat-seq and batch effect correction in general should be carefully considered.

4.2.3 SCTransform V2

While ComBat-seq is generally focused on batch correction, but includes a covariate to account for total transcript counts in a cell, Seurat's SCTransform is a normalization and variance stabilization technique that uses total transcript counts as a proxy for sequencing depth, modeled at the gene level and allows for an across-genes covariate corresponding to batch. SCTransform is specifically intended for single-cell RNA-seq data, with a primary goal of removing the influence of technical effect while preserving variability contributed by underlying mRNA expression. As SCTransform performs negative binomial regression, observations are modeled on the count scale. SCTransform is carried out in three major steps, including independent gene-level negative binomial regression, parameter regularization, and finally computing Pearson residuals based on observed UMI counts to obtain normalized values. To obtain adjusted counts instead of normalized values, Pearson residuals can be back-transformed into counts to represent the adjusted count matrix.

We will first review SCTransform without considering batch effect correction. After discussing the method with total transcript counts as the only covariate, we will review an extension of this model that can be applied for batch effect correction.

The first step of SCTransform is performing individual gene-level negative binomial regressions with total count as a proxy for sequencing depth. Let y_g represent the vector of transcript counts assigned to gene g, across all cells. For each gene we have,

$$\log(E(y_g)) = \beta_0 + \beta_1 + \log_{10}(m),$$

where m is a vector of total transcript counts across genes in each cell. Specifically for each entry j of m, $m_j = \sum_i x_{ij}$. β_1 represents the coefficient corresponding to this covariate and β_0 is the intercept term for each gene. The overdispersion parameter ϕ is estimated for each gene using the observed variance and mean-variance relationship for the corresponding negative binomial parameterization, mean μ and variance $\mu + \frac{\mu^2}{\phi}$.

The next step is to perform parameter regularization to avoid overfitting associated with running the model at the gene-level. For each parameter independently, kernel regression is applied to smooth parameter estimates based on gene-level averages, using a normal kernel, bandwidth determined by R function bw.SJ [82], and a default bandwidth adjustment factor of 3. The bandwidth selected by bw.SJ is multiplied by this bandwidth adjustment factor, and this value determines the neighborhood of points influencing weighted averages.

Using the regularized parameter estimates, UMI counts are converted into Pearson residuals to represent normalized values. Let z_{gj} represent Pearson residuals in gene g and cell j. The following transformation is used,

$$
\begin{aligned}
z_{gj} &= \frac{x_{gj} - \mu_{gj}}{\sigma_{gj}}, \\
\mu_{gj} &= \exp(\beta_{0_g} + \beta_{1_g} \log_{10} m_j), \\
\sigma_{gj} &= \sqrt{\mu_{gj} + \frac{\mu_{gj}^2}{\phi_g}},
\end{aligned}
$$

to obtain these residuals. The reverse of this transformation can be applied to obtain the adjusted counts matrix.

Now, consider accounting for additional nuisance parameters, such as batch. No regularization is peformed to account for parameters for these covariates. Instead, the method is applied using sequencing depth as the only covariate, and an additional round

of negative binomial regression is run, with the previously regularized sequencing-depth parameters and a parameter corresponding to batch effect included in the model. Pearson residuals are computed using this model including nuisance parameter coefficients.

4.3 Methods

In this section, we discuss application of ComBat-seq and SCTransform separately to the scRNA-seq data sets analyzed in the previous chapter. As the original ComBat approach requires the normality assumption, it is not suitable to use in this case. After performing technical-effect correction, we run the Bayesian model and examine the effect of these methods on results. Note that, while protein data is still included in the models, technical effect correction methods are applied to the RNA-seq data specfically.

We review implementation of ComBat-seq and SCTransform, as well as its application to the Bayesian model. Additionally we review gene and gene-set level testing and leave-one-out cross validation.

4.3.1 ComBat-seq corrected data

We apply ComBat-seq batch correction to each of the single cell mRNA testes data sets. Let matrix M_{g,c_1+c_2} represent a matrix consisting of UMI counts, where each row corresponds to a gene and each column represents a cell. The first c_1 columns of M are cells from mRNA data set 1, and columns c_1+1 through c_2 are measurements from cells in mRNA data set 2. Additionally define b as a vector of length c_1+c_2, where each entry denotes the data set from which that cell was observed.

The sva package from the bioconductor repository contains the ComBat_seq function, which allows us to run the ComBat seq method on M within R [83] [84]. The resulting output is matrix M^{ComBat} with the same dimensions as M, representing adjusted counts

based on ComBat seq. We subset M^{ComBat} by cell, to form M_1^{ComBat}, batch corrected counts for mRNA data set 1, and M_2^{ComBat}, batch corrected counts for mRNA data set 2.

Next we use adjusted counts M_1^{ComBat} and M_2^{ComBat}, along with peptide-level data sets P_1, P_2, P_3, P_4 to apply the same Bayesian model as that in the previous chapter. This model is run using 3000 iterations in stan with 2000 warm-up iterations. The largest $\hat{R}$ is 1.07; however, some iterations saturate maximum tree depth. As there are no other convergence warnings we can safely ignore this warning; however, this can indicate that the model can potentially be run more efficiently. Trace plots are shown for this fit in the Appendix.

4.3.2 SCTransform corrected data

Next we apply SCTransform normalization and variance stabilization to each of the single-cell mRNA testes data sets. While batch can be included as a covariate to be used during a second round of negative binomial regression, its modeled at a per-data set resolution, with the assumption that the effect of batch is uniform across cell types. Since our Bayesian model would therefore account for batch at a greater resolution, as batch is modeled for each data set and cell type, we perform SCTransform on each of the mRNA data sets independently without including batch as a covariate.

To apply SCTransform, we first create a Seurat object, a convenient data object that contains all corresponding UMI counts, transformed values, and meta data reflecting cell type assignments and any other cell-level information. Original counts are stored in the "RNA" assay of these Seurat objects. Next we call the SCTransform function on the Seurat object for each of the mRNA data sets, specifying "V2" regularization to refer to the correct version. Resulting adjusted counts are stored in the "SCT" assay, which we

pull out for application to the Bayesian model.

Let M_1^{SCT} and M_2^{SCT} represent adjusted count matrices for mRNA data sets 1 and 2, respectively. We input these outputted matrices directly into the original Bayesian model pipeline, along with the peptide-level data sets. Sampling is run using the same function arguments as that of the ComBat-seq fit, with maximum $\hat{R}$ of 1.08 and no convergence-related warnings. Trace plots can be see in the Appendix.

4.3.3 Gene-level testing

We repeat analysis from the previous chapter for posterior draws from the ComBat-seq and SCTransform preprocessed fits to explore the degree to which these results are influenced by the effect of technical factors.

We use the same testing procedure for posterior draws from the separately preprocessed data sets to compare R_{gc} in each gene g and cell type c to all other genes in each cell type. This recovers the posterior probability that R_{gc} has the same sign as that of all other genes. We recover significant genes based on a 1% FDR cutoff.

To test gene-level mRNA, protein correlations, we use the same testing approach as the previous chapter, recovering the posterior probability that the across cell types correlation for each gene is higher than that of the averages across all other genes. Significance is determined based on a 1% expected proportion of false discoveries.

For each gene-level test, we intersect the list of significant genes corresponding to each technical-effect correction method with significant genes from the previous chapter to examine the agreement between these approaches.

4.3.4 Group-level testing

We replicate GO-level testing for posterior draws corresponding to each of the technical-effect correction methods. To identify groups with extreme rPTR in each cell type, we again compare the average of R_{gc} across genes in each functional gene set to that of all other genes for each cell type. This recovers the posterior probability that the rPTR in each gene, cell type is as extreme as the average across all other genes. Using this posterior exclusion probability, we can compute the posterior expected proportion of false discoveries and identify significant groups at a 1% cutoff.

Similarly, we can repeat the analysis in the previous chapter to recover groups with low mRNA, protein correlation. We compute the average R_{gc} across genes in each functional gene set for each cell type, and then compute the correlation of these averages across cell types. This value is compared to the average across genes not assigned to that gene set. We determine significance based on a 1% FDR cutoff.

4.3.5 Leave-one-out cross validation

To examine how technical-effect correction method affects model results, we approximate leave-one-out cross validation (LOO-CV) for each of the resulting fitted models using Pareto smoothed important sampling (PSIS). LOO-CV involves fitting the model to all data points but one (cycling through such that each data point is left out once), and then predicting the corresponding left-out point using the fitted model. This results in an estimate to reflect the predictive accuracy for the model. Rather than fitting and refitting the model for each point, which would be computationally expensive, we use the PSIS approximation. PSIS uses importance sampling to assign weights to each posterior draw, smoothed using the Pareto distribution. Next, the posterior likelihood is adjusted based on these weights to estimate the posterior distribution corresponding to leaving out

each point, resulting in an estimation of the model's out of sample predictive accuracy. We can implement this cross-validation approach using the loo package in R [85].

To compare different models, or in this case the same model applied to observations that are preprocessed differently, we compare the expected log pointwise predictive density (elpd), a representation of each model's out-of-sample predictive accuracy, across models. Additionally, the estimated shape parameter of the Pareto distribution, k, and whether finite mean and variance exist based on this k summarise the percentage of points for which the estimation is "good". This comparison can be generated using the compare_loo function within the loo package.

4.4 Results

4.4.1 Gene-level comparison of original results to ComBat-seq and SCTransform adjusted fits

To better understand how each of these technical-effect correction methods affect the resulting fit, we repeat gene and group-level testing and compare results to that of the previous chapter. We directly compare results of gene-level and group-level tests.

In figures C.1 and C.8, we can see across cell types correlations corresponding to the posterior mean of μ_{gc} and $\mu_{gc} + R_{gc}$ in each gene and cell type corresponding to ComBat-seq and SCTransform adjusted counts. There appears to be only a marginal reduction in the median inferred correlation for ComBat-seq application, and inferred correlations are generally higher than that of empirical values, just as we obtained with the previous fit.

For each gene, we test the correlation across cell types to that of all other genes, recovering significant genes with low correlation. We compare results of this test to that

run on the original data. While 233 genes had significant low correlation for the previous fit, the ComBat-seq prepared data have more significant genes at 289, with 230 of these genes being significant in both tests. 224 genes have significant low correlation based on the SCTransform corrected data, and 209 of these genes are also significant in the previous fit.

In addition to comparing across cell-types correlations, we compare gene-level testing for nonzero rPTR in individual cell types. In figures C.3 C.10 we have posterior means for μ and $\mu + R$ in each gene and cell type based on the ComBat-seq and SCTransform associated fits, respectively. Across both correction methods, we see an increased number of significant genes in each cell type at a 1% FDR cutoff.

In table 4.1 we have the proportion of significant genes in this test that were also significant in the previous test. The majority of significant genes based on the each of the technical effect correction methods are also significant in the original model. When we compare ComBat-seq and SCTransform-corrected results directly, we see strong overlap between significant genes, indicating that results are also consistent across these approaches.

Table 4.1: Comparison of genes with significant rPTR from fit applied to ComBat-seq and SCTransform prepared observed data and fit applied to observed data directly. We can see the proportion of significant genes that are also significant for the previous chapter.

Correction Method	**EC**	**LC**	**PTM**	**SPC**	**SPG**	**ST**
ComBat-seq	0.839	0.866	0.960	0.652	0.833	0.723
SCTransform	0.810	0.894	0.975	0.629	0.914	0.716

4.4.2 Group-level comparison of original results to ComBat-seq and SCTransform adjusted fits

In addition to comparing gene-level results, we explore the relationship between significant gene sets using the ComBat-seq and SCTransform adjusted count data sets and that of the uncorrected data. First, we consider groups with extreme rPTR, as applied to both fits. Please see Methods for additional details on how this test is run. In table 4.2 we can see summary information for GO testing applied to both transformed data sets, including the number of significant gene sets and proportion of significant gene sets that are also significant for the previous fit. Across all cell types, there is an increase in the number of significant groups. Additionally, there's a large degree of overlap among significant groups between the fit on uncorrected data and that applied to corrected data sets.

Table 4.2: Comparison of genes with significant rPTR from fits applied to ComBat-seq and SCTransform prepared observed data and fit applied to observed data directly. For each cell type we can see the proportion of significant genes that are also significant in the previous chapter.

Cell Type	**Number of Significant Groups**		**Proportion Previously Significant**	
	ComBat-seq	**SCTransform**	**ComBat-seq**	**SCTransform**
EC	132	129	0.886	0.837
LC	178	195	0.910	0.790
PTM	232	237	0.831	0.819
SPC	164	160	0.835	0.819
SPG	106	119	0.849	0.663
St	195	178	0.872	0.860

Next, we compare groups with low across cell types correlation of μ and $\mu + R$ among the ComBat-seq and SCTransform associated fits and the original one. In table 4.3 we have the total number of groups with low correlation and the corresponding proportion that were also significant based on the previous fit. With the ComBat-seq prepared counts, nearly twice as many groups have significant low correlation relative

to the previous fit, as 15 groups are significant instead of 8, and all 8 of the previously significant groups are significant for this test. One fewer group is found to have significant low correlation with the SCTransform-prepared data, as 7 groups are significant, and 6 of these 7 groups were also significant when the same test was run on fit from the original data.

Table 4.3: Comparison of genes with significant low across cell types mRNA, protein correlation from fit applied to ComBat-seq and SCTransform prepared observed data sets and fit applied to observed data directly. For each correction method we can see the proportion of significant genes that are also significant for the previous chapter.

Method	**Number of Significant Groups**	**Proportion Previously Significant**
ComBat-seq	15	0.533
SCTransform	7	0.857

In table 4.4, we can compare results of the scale parameter GO testing applied to this fit to that of the fit applied to the uncorrected data. Among ComBat-seq associated results, 2 groups have significant scale, compared to 3 in the previous chapter, and only one of these significant groups were previously significant. For the fit applied to SCTransform corrected counts, there are twice as many groups with significant scale parameter as there were when the scale parameter test as applied to the original fit, with 6 groups having significant scale parameter. All 3 of the groups with significant groups applied to the previous fit are significant for the test applied to this fit.

Table 4.4: Comparison of groups with significant scale parameter from fit applied to ComBat-seq and SCTransform corrected counts and fit applied to observed data directly. For each cell type we can see the proportion of significant groups that are also significant for the previous chapter.

Method	**Number of Significant Groups**	**Proportion Previously Significant**
ComBat-seq	2	0.5
SCTransform	6	0.5

4.5 Posterior Predictive Checking

In this section we both evaluate fitness of the model applied to each of the technical-effect correction methods, and compare these results to the fit applied to the uncorrected data in the previous chapter. We examine posterior predictive draws and their test statistics relative to the corrected values that are inputted into the model. While the influence that these changes have over posterior predictive draws for the protein data sets are of interest, they are less directly affected because batch and sequencing depth correction are applied to the mRNA data sets. Please see the appendix for posterior predictive check figures relating to the protein data sets.

In table 4.5, we can see the proportion of corrected gene, cell type averages in each of the mRNA data sets that fall within the 95% posterior predictive interval for each technical effect correction method. ComBat-seq results indicate that for mRNA data set 1, we see a marginal reduction in the proportion of genes within the interval, and a marginal increase in coverage for mRNA population 2. With SCTransform results we can see only marginal changes in coverage. Across both cell types, populations, and correction methods, however, the coverage of corrected values remains high.

Table 4.5: Comparison of mRNA gene proportion within 95% posterior predictive interval for ComBat-seq and SCTransform adjusted data. The values represent the proportion of mRNA data set averages within the interval for each cell type and data set.

Cell Type	**mRNA Data Set 1**		**mRNA Data Set 2**	
	ComBat-seq	**SCTransform**	**ComBat-seq**	**SCTransform**
EC	0.986	0.993	0.997	0.996
LC	0.983	0.990	0.996	0.995
PTM	0.986	0.993	0.997	0.997
SPC	0.994	0.995	0.996	0.995
SPG	0.979	0.989	0.999	0.997
St	0.971	0.978	0.998	0.997

Next, we explore the effect of ComBat-seq correction on the posterior predictive

across cell types variance. In figures C.4 C.11, we have log observed variances of mean ComBat-seq and SCTransform corrected counts across cell types, respectively, as well as the corresponding posterior predictive interval. In the ComBat-seq associated fit, we see an increase in the coverage and correlation corresponding to mRNA Pop2, which may suggest that the adjustments from ComBat-seq mitigate some technical effects such that the model is better able to replicate batch-corrected values. There appears to be no difference in the variance coverage for the SCTransform corrected data. This is surprising, considering the potential gene-level effect that sequencing depth has on observed values and that the model currently models sequencing depth at the cell type and not cell level.

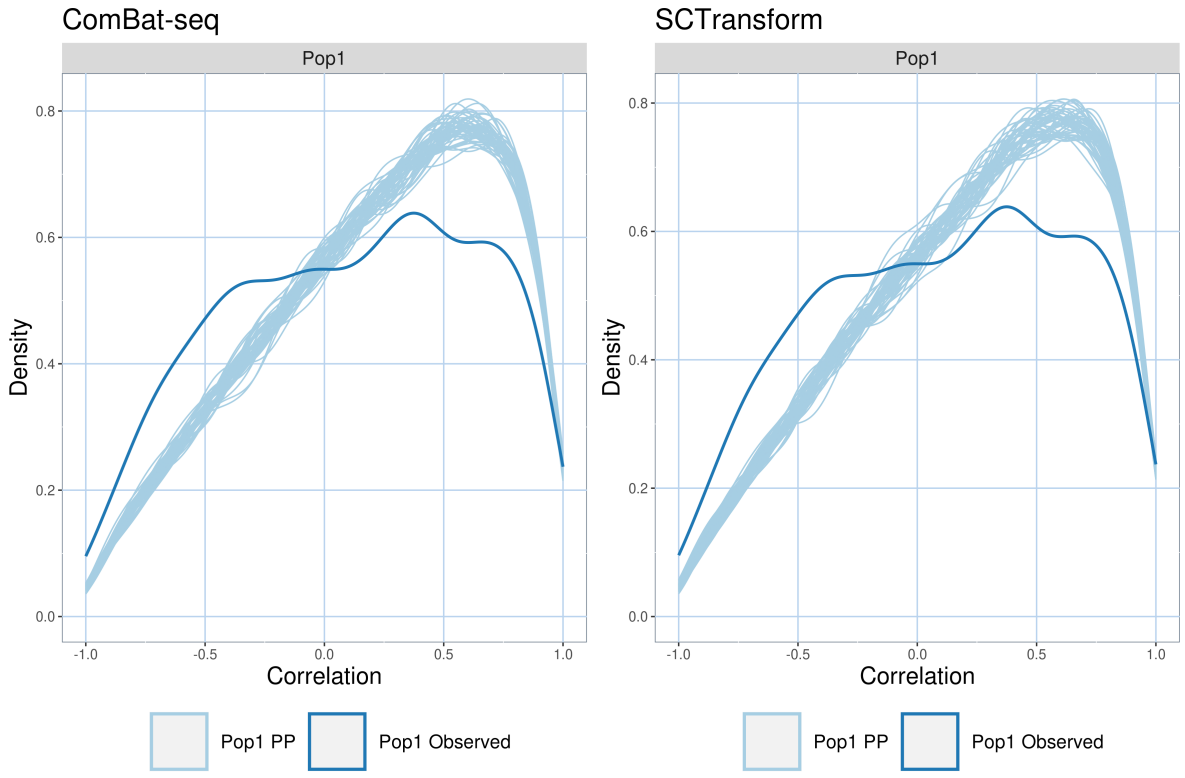

Figure 4.1: ComBat-Seq and SCTranform processed and posterior predictive mRNA, Protein Correlation Across Cell Types. Density plots are displayed for observed correlation of mRNA, protein across cell types for protein data set 1 and mRNA data set 1. Additionally we can see posterior predictive correlations across cell types for randomly selected posterior predictive draws.

Finally we examine the effect of ComBat-seq adjustment on posterior predictive mRNA, protein correlations across cell types. In figure 4.1, we can see posterior predictive correlations of mRNA and protein for both ComBat-seq and SCTransform corrected counts. Observed correlations are based on the adjusted counts associated with each of the correction methods. First, we note that the across cell types correlation with the ComBat-seq adjustment seems to be reduced for some genes in protein data sets 1 and 2 in particular relative to the uncorrected data. Interestingly, it appears that posterior predictive correlations do not show enough variability across protein data sets, failing to capture the lower correlation associated with some genes, especially those that have observed correlation below -0.5. Note that these plots reflect mRNA data set 1 and protein data set 1, and posterior predictive correlations for the remaining data sets are available in the Appendix.

4.6 Leave-one-out cross validation

Next, we quantitatively compare performance of these models by approximating leave-one-out cross validation (LOO-CV) using Pareto smoothed importance sampling (PSIS). The expected log pointwise predictive density (elpd) will reflect the expected predictive accuracy, and the resulting model of choice will have the highest expected predictive accuracy comparatively. While LOO-CV is not an appropriate method to compare models run on different data sets, it is appropriate in this case because the data sets on which the model is fit are the same, and is simply transformed differently in each case before being applied to the same model. In table 4.6, we can see summary information for to LOO-PSIS application corresponding to the mRNA likelihood. Interestingly, despite better performance associated with some of the other methods with respect to across cell types variance and mRNA, protein correlation, the LOO-PSIS results indicate that

SCTransform preprocessing yields the most accurate predictive distribution. While the standard error indicates that there is no "overlap" between elpd differences, there is also a fairly large proportion of points that do not have "good" fit based on Pareto diagnostics, suggesting that we should accept these results with caution, and can potentially run LOO directly on points with poor diagnostics. A summary of LOO-PSIS results is also available for protein in the Appendix.

Table 4.6: LOO-PSIS Comparison for mRNA Observations. We can see the difference in elpd for each model, standard error, and the proportion of points labeled to have "good" fit based on Pareto diagnostics.

Model	elpd diff	se diff	good prop
SCTransform	$0.00e+00$	$0.00e+00$	51.2
ComBat-seq	-2417.420	97.106	48.2
Original	-5203.899	56.695	51.0

4.7 Discussion

In this chapter we reviewed technical-effect correction methods ComBat, ComBat-seq, and SCTransform, and evaluated the performance of applying our Bayesian model to ComBat-seq and SCTransform corrected count matrices relative to the fit obtained by applying the model to observed counts directly.

Important aspects of ComBat-seq and SCTransform include the modeling of technical effects at the gene and cellular levels as opposed to aggregating across cells assigned to the same type or measured in the same data set, as both sequencing depth and batch can potentially influence gene-level variability [9] [76]. Additionally, these methods leverage the count structure associated with scRNA-seq data, and result in corrected count values consistent with this structure. This allows for the application of downstream methods that use negative binomial or Poisson assumptions, such as differential expression or rPTR estimation.

Application of these correction methods resulted in marginal changes to most posterior predictive test statistics of interest, but overestimated the across cell types mRNA, protein correlation within posterior predictive values relative to corrected counts for some genes. Application of these correction methods both resulted in an increase in genes and functional gene sets associated with extreme rPTR compared to the original fit. Approximations for leave-one-out-cross validation expected log pointwise predictive density indicate that SCTransform preprocessing provides the best predictive accuracy relative to ComBat-seq preprocessing and applying the existing model to observed data directly.

By examining the influence of technical factors and their correction methods on results, we can better identify potential model changes for improved handling of technical factors. Improved approaches at taking unwanted technical effects into account may include expanded modeling of batch effect at the gene and data set levels within the model, as opposed to continuing to model batch by cell type and data set. Additionally, accounting for sequencing depth at a higher resolution or including a coefficient so that its effect can vary by gene may better quantify technical effects within the model. These techniques would avoid the potential "double dipping" that can arise as a result of performing one round of estimation on the entire set of data to carry out technical effect correction, and then another round of estimation corresponding to the Bayesian model. This technique takes into account the potential for false discovery independently at each stage of analysis but fails to account for their shared influence on false discovery probability across stages. This can result in failing to achieve statistical guarantees relating to Type 1 error and power [86]. By jointly quantifying the effect of technical effect parameters, this issue can be avoided. Finally, shifting focus to model technical factors within protein data, such as batch, at the gene level, as well as modeling of missingness mechanisms that may contribute to missingness not at random represent valuable improvements to this model that can be made in the future.

Chapter 5

Discussion

In this dissertation, we have quantified the effect of post transcriptional regulation by developing a Bayesian model that estimates relative protein-to-mRNA ratios (rPTR) within individual genes. This model has been applied to bulk measurements of proteins and mRNA, which reflect gene product abundance across cells in the same tissue for each gene, as well as single-cell peptide and mRNA measurements, quantifying gene product abundance for each gene within indiviudal cells. Using replicates corresponding to the same gene and cell type or tissue across data sets, we are able to quantify sampling variability and the effect of unwanted technical factors, allowing for the estimation of underlying relative $\log_2$ mRNA intensity and protein-to-mRNA ratios. Using posterior draws from this model, we identify genes and sets of functionally related genes that have extreme rPTR, and thus potentially have elevated post transcriptional regulation. In addition to quantifying rPTR using observed transcript counts and peptide intensities, we explored the effect of technical factors, including batch and sequencing depth, to better understand its influence on modeling results. By applying the model to adjusted transcript counts using two technical-effect correction methods, we examine the effect of gene-level correction and explore opportunities for model improvement in this area.

Future developments that can improve estimation of rPTR include gene-level quantification of batch and sequencing depth within the model, which would leverage the benefits of technical-effect correction techniques while achieving statistical guarantees relating to false positive probability and power, as well as modeling of technical effects within protein observations. Another meaningful improvement is modeling missing data to better reflects mechanisms contributing to missingness and suitable for MNAR assumptions.

Appendix A

Appendix Chapter 2

A.1 Identifying tissues with extreme rPTR

In addition to considering R at the gene and gene-set levels, we test R by tissue to identify tissues with extreme rPTR. This can help to identify tissues that consistently have positive or negative rPTR. For each posterior draw, we compute the average R_{gt} across genes g and compare it to the average R_{gt} across all other tissues. Please see Methods for additional details on this test. In table A.1, the results of this tissue-level testing are displayed. Interestingly, tissues with a large set of genes with significant rPTR did not have significant tissue-level results. Instead, tissues with few significant genes are significant at the tissue level, and average differences are fairly small. Additionally, it appears that most across-tissue averages are slightly negative.

Table A.1: Number of significant test results for tissue-level comparison of mean of R to all other tissues.

Tissue	**Posterior probability**	**Average R**	**Average R Other**
Ovary	0.0004	0.0444	-0.0188
Kidney	0.0012	0.0409	-0.0184
Liver	0.0032	0.0374	-0.0179
Adrenal Gland	0.0112	0.0331	-0.0174
Esophagus	0.7380	-0.0812	-0.0031
Colon	0.7668	-0.0776	-0.0036
Testis	0.9308	-0.0482	-0.0072
Pancreas	0.9436	-0.0241	-0.0103
Prostate	0.9544	-0.0308	-0.0094

A.2 Additional Plots

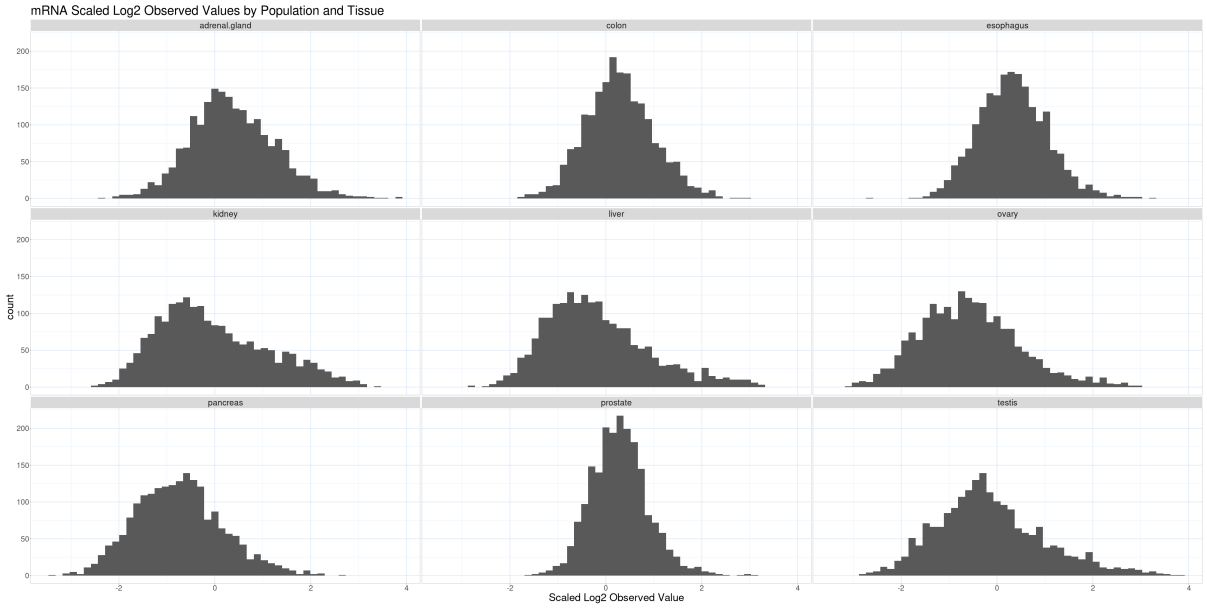

Figure A.1: Z-transformed $\log_2$ transcript counts in mRNA data set [19] in each tissue considered in analysis.

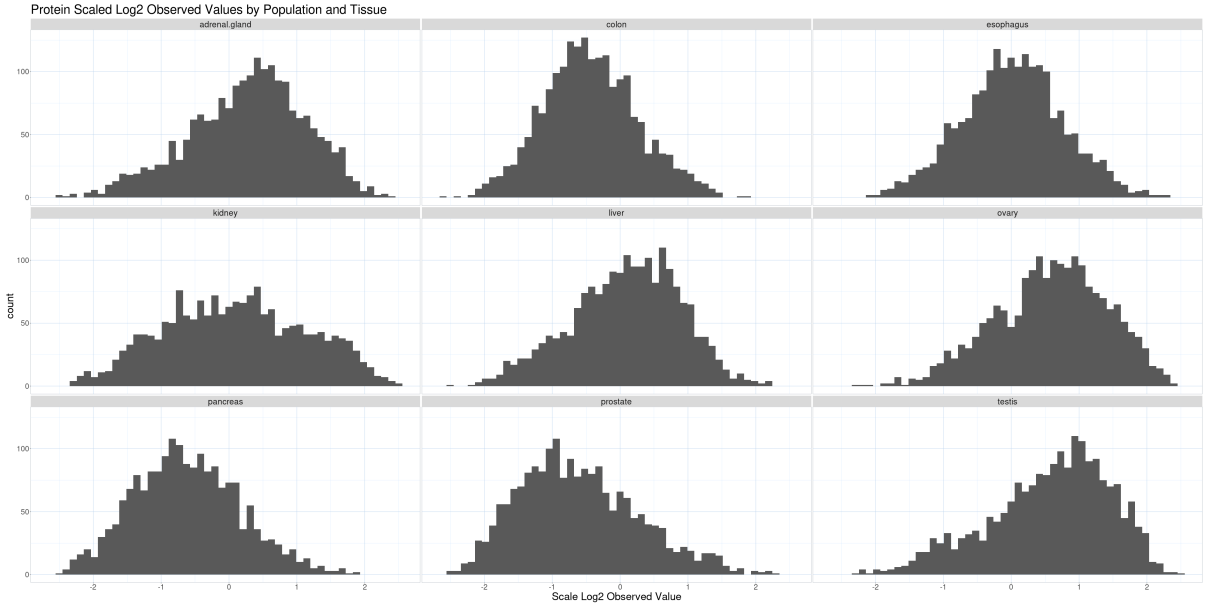

Figure A.2: Z-transformed $\log_2$ transcript counts in Protein data set [5] in each tissue considered in analysis.

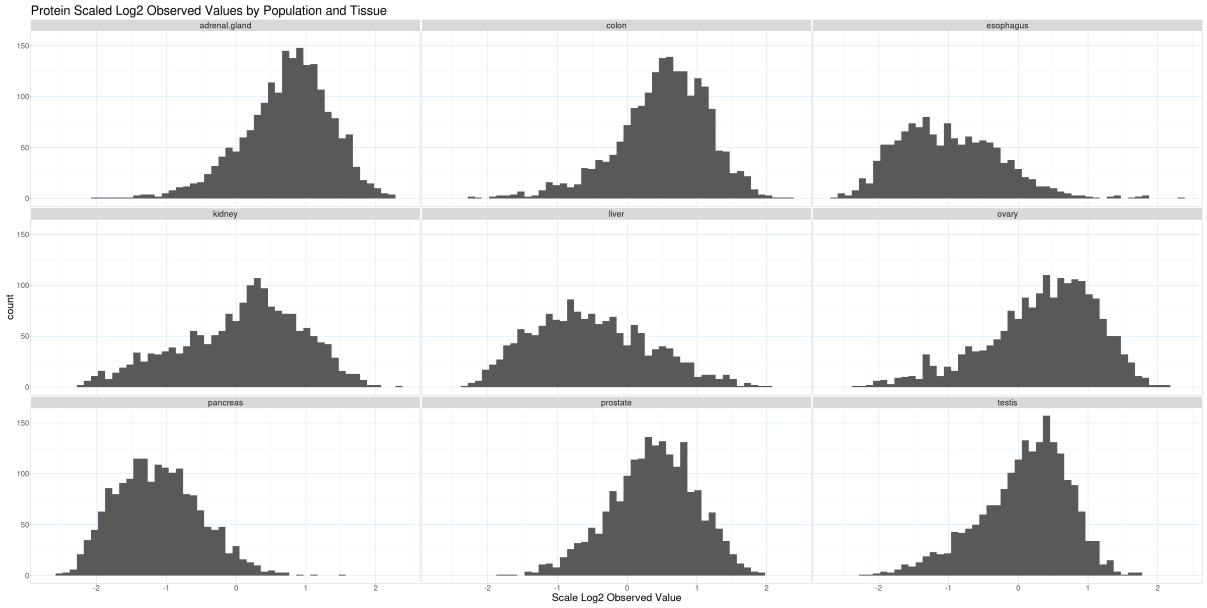

Figure A.3: Z-transformed $\log_2$ transcript counts in Protein data set [20] in each tissue considered in analysis.

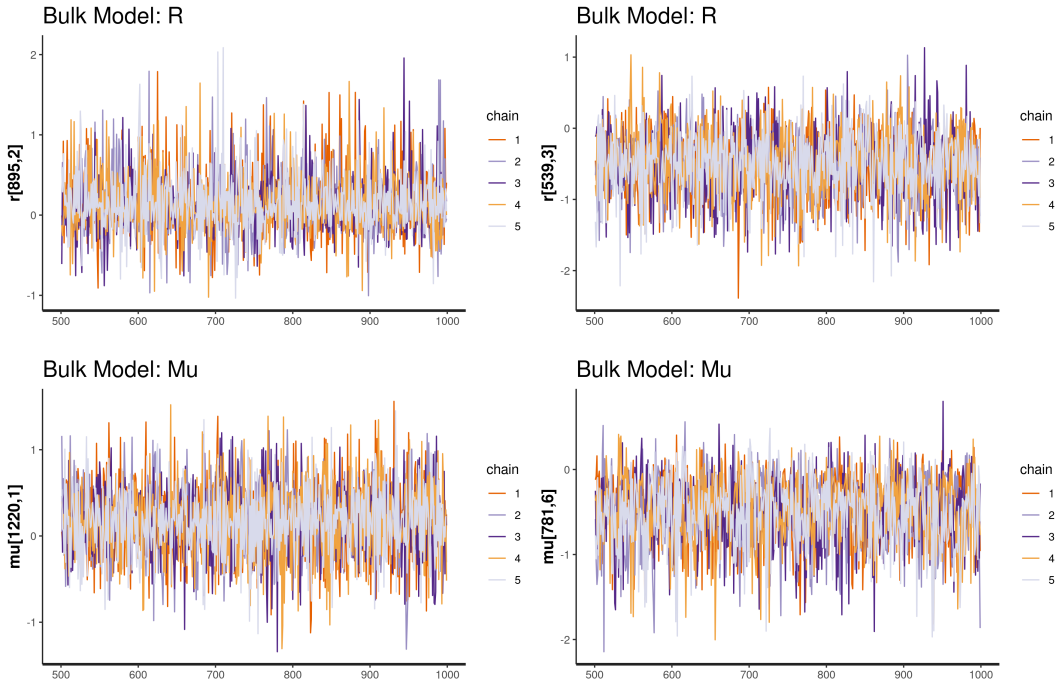

Figure A.4: Trace plots for R and μ parameters in randomly selected genes and tissues.

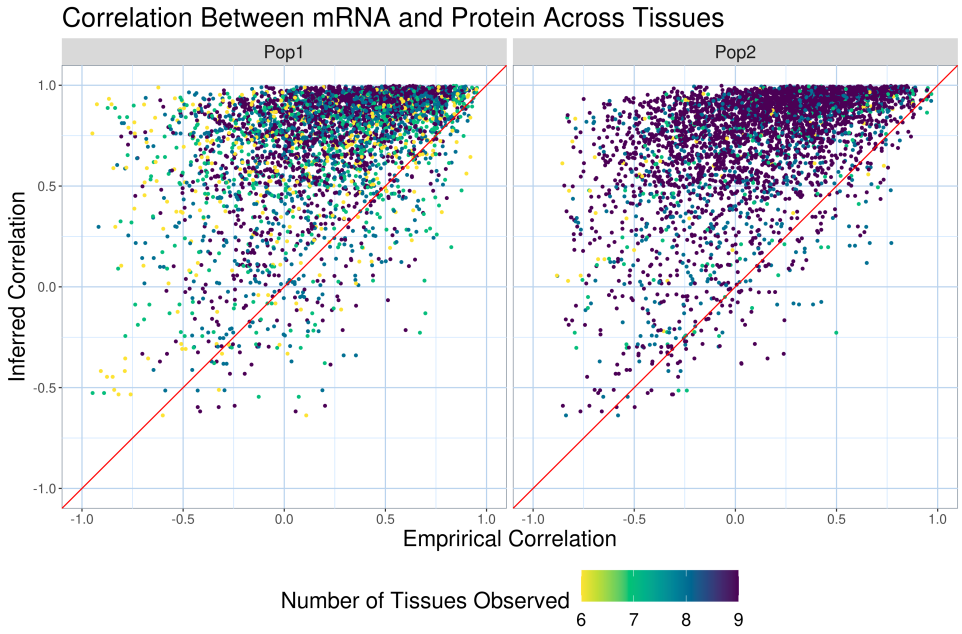

Figure A.5: Empirical and inferred across tissues correlations using both mRNA data sets and both protein data sets. Inferred correlations are computed using posterior means of μ and $\mu + R$ for each gene and tissue based on the implemented Bayesian model. Observed values are based on Z-transformed $\log_2$ observations.

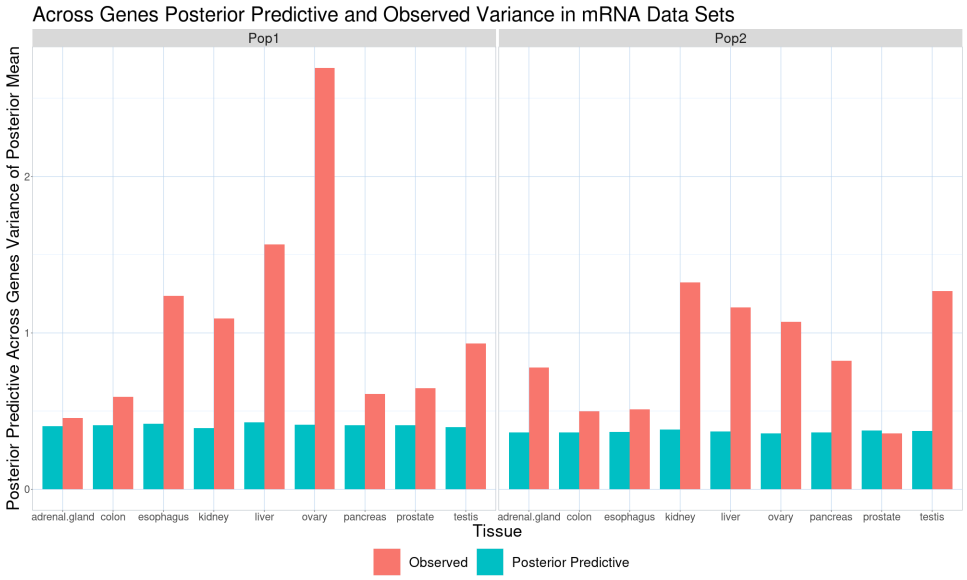

Figure A.6: mRNA observed and posterior predictive across-genes correlation for each tissue and data set. Posterior predictive variances are based on posterior means of posterior predictive draws for each data set, tissue and gene.

Appendix B

Appendix Chapter 3

B.1 Model selection

While the negative binomial likelihood takes advantage of the overdispersion present in transcript counts, we opted to consider several likelihood options and compare them. The Poisson model, also applicable to UMI counts across cells for each gene, is another common choice that assumes equal mean and variance [67]. Additionally the normal model is considered, using average log counts across cells corresponding to each cell type by gene and data set. Note that all of these likelihood choices apply to mRNA count data, as protein-level measurements correspond to intensity and are thus not reasonable for the negative binomial or Poisson likelihoods.

We explore choice of likelihood using leave-one-out cross validation, applied to protein and mRNA likelihoods separately. While likelihood choice only applies to M_{mgc}, mRNA and protein are tied through μ, and thus likelihood choice can effect both modalities. More information on leave-one-out cross validation is available in this section.

B.1.1 Poisson

As opposed to using a Negative Binomial likelihood to model transcript counts, we can use a Poisson likelihood, which assumes equal mean and variance.

$$M_{mgc} \sim \text{Poisson}(n_{mc} 2^{\mu_{gc}} a_g \gamma_{mc}).$$

This assumption can be restrictive considering the large degree of measurement noise, but models counts directly as opposed to normalized or scaled counts. This model retains batch corrections and centering and scaling at the mean level; however, sampling variability is not modeled on the per data set-cell type basis as is in the negative binomial model.

B.1.2 Normal

Modeling transcripts using a Normal likelihood requires that we model the mean $\log_2$ transcript counts, instead of working with the nonnegative transcript counts directly. Define M^*_{mgc}, the average $\log_2$ transcript count for gene g, cell type c for mRNA data set m. We use likelihood,

$$M^*_{mgc} \sim \text{Normal}(\mu_{gc} + a_g + \gamma_{mc}, \tau^{*2}_m),$$

where the priors for parameters a and γ are adjusted to resemble that of the Normal likelihood used for protein. τ^{*2}_m represents the sampling variance for transcript count averages, which is also assigned a half-Cauchy prior. While normal likelihoods are a frequent approach for average $\log_2$ counts [39], it fails to reflect the count structure of transcript data.

In order to determine which choice of likelihood is most reasonable with respect to

Table B.1: LOO-PSIS Comparison for mRNA Observations

Distribution	elpd diff	se diff
Negative binomial	$0.000000e+00$	$0.000000e+00$
Poisson	$-1.671164e+05$	$6.978348e+03$
Normal	$-3.211972e+11$	$1.551923e+09$

model performance, we use Pareto smoothed importance sampling (PSIS) to approximate leave-one-out cross validation (LOO-CV) prediction errors [85]. LOO-CV involves fitting the model to all but one of the data points, and then using the fitted model to predict the response for the left-out data point. The process is repeated for each data point in the data set, resulting in an estimate of the model's predictive accuracy. PSIS is a less computationally complex approach to LOO and is based on importance sampling, as weights are assigned to each posterior draw, effectively adjusting the posterior as though that draw was not observed [85]. The PSIS method uses a Pareto distribution to smooth the importance weights, resulting in a more stable and accurate estimate of the model's out-of-sample predictive accuracy. This comparison is implemented using the loo package in R [85], which returns a set of metrics including the expected log pointwise predictive density (elpd) as a reflection of the model's out-of-sample predictive accuracy. Table B.1 displays elpd comparisons for the mRNA likelihood on each of the candidate models.

As B.1 shows, the Negative binomial model has the best expected log predictive density relative to the Poisson and Normal models. In B.2, we can also see how choice of mRNA likelihood affects protein expected log predictive density. While the Poisson model has the best elpd comparatively, the standard error indicates that this difference is negligible.

Table B.2: LOO-PSIS Comparison for Protein Observations

Distribution	elpd diff	se diff
Poisson	0.000	0.000
Negative binomial	−0.506	11.689
Normal	−259.237	26.053

B.2 Additional gene-level results

Next, we review plots relating to our Bayesian model at the gene-level, including false discovery rates for gene testing and posterior means for the scale parameter.

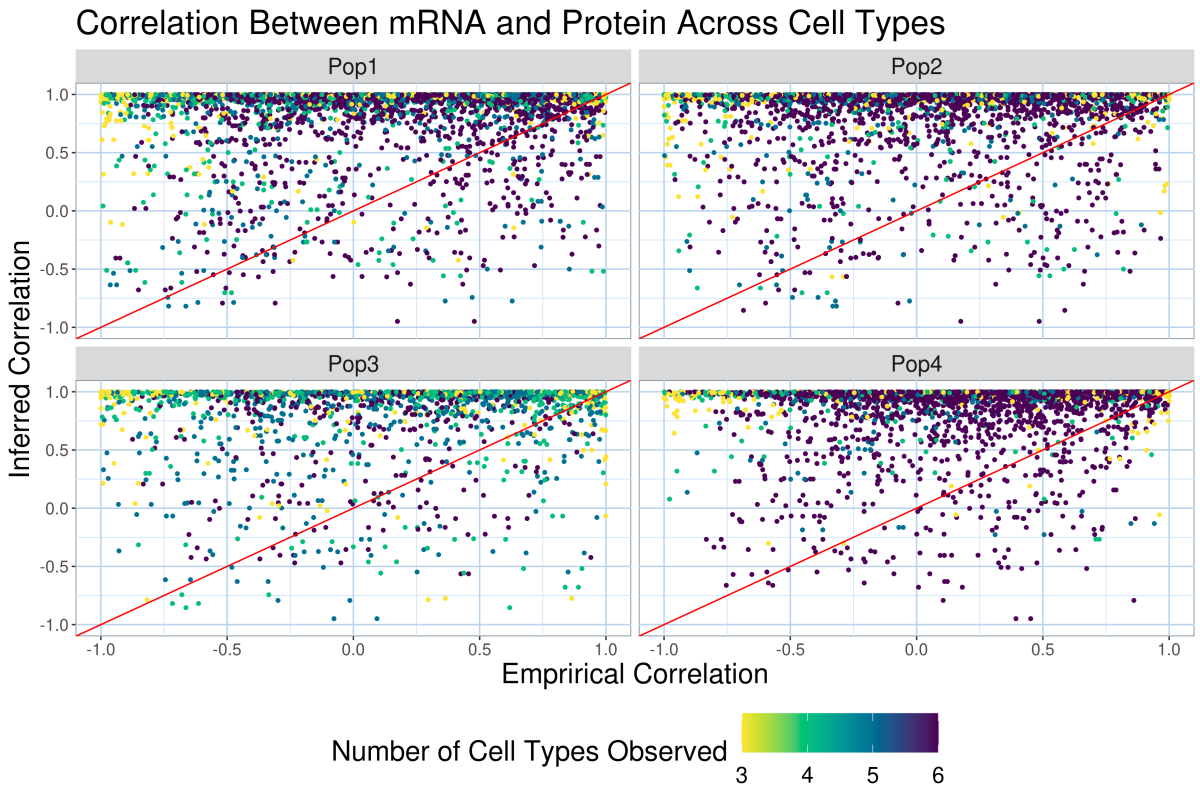

Figure B.1: mRNA, Protein Across Cell Type Correlations for all protein populations. On the y-axis is the inferred correlation between mRNA and protein using μ_{gc} and $\mu_{gc} + R_{gc}$ across cell types c. On the x-axis is the observed correlation between each faceted protein population and both of the mRNA populations. A red line depicts $y = x$.

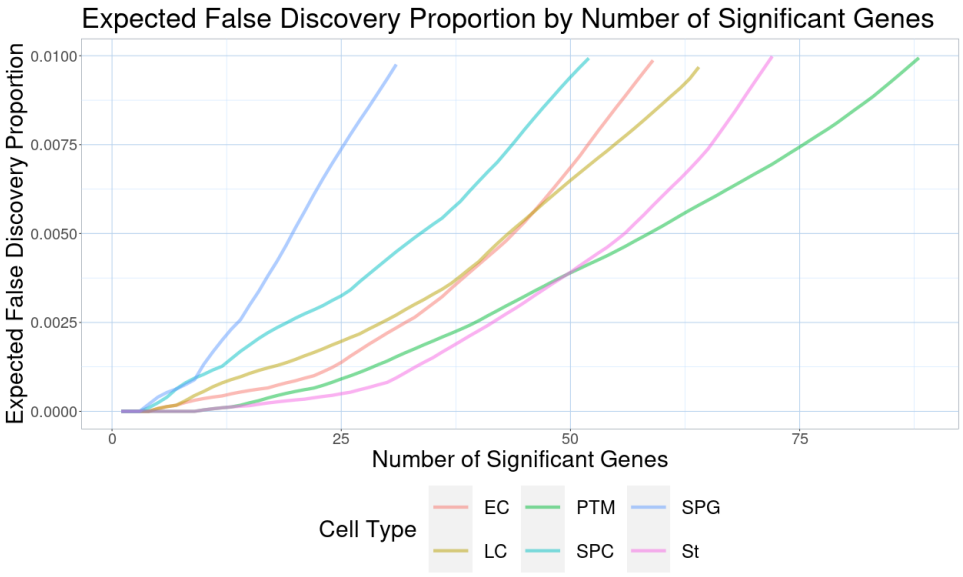

Figure B.2: Expected Proportion of False Discoveries by Number of Significant Genes. For each cell type, the expected false discovery proportion is computed as the cumulative mean of ordered posterior exclusion probabilities. Each cell type has a 1% cutoff for FDR.

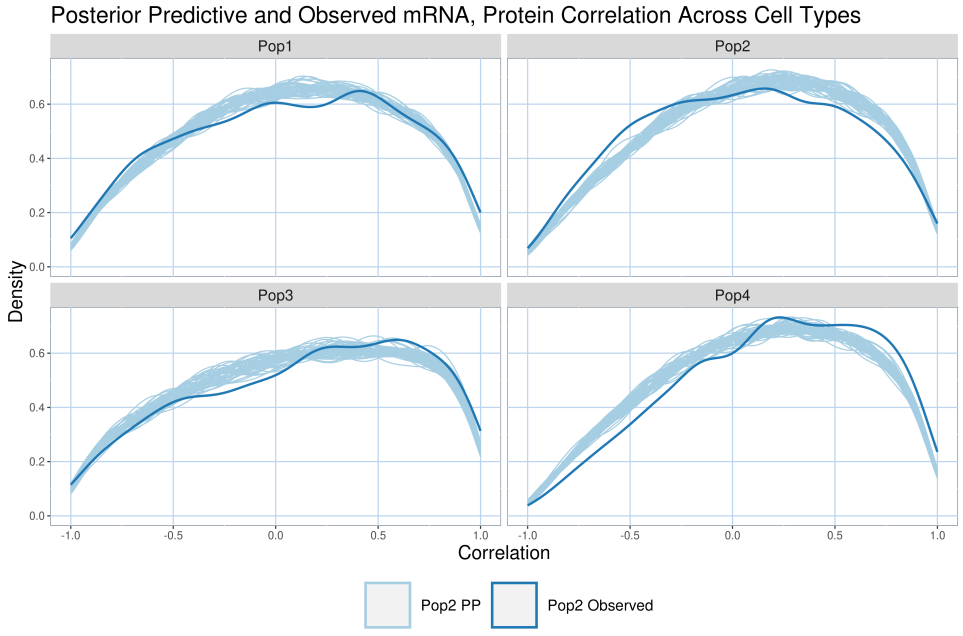

Figure B.3: Observed and Posterior Predictive mRNA, Protein Correlation Across Cell Types for mRNA population 2. Light blue represents correlations for posterior predictive draws, and dark blue represents correlations for observed values.

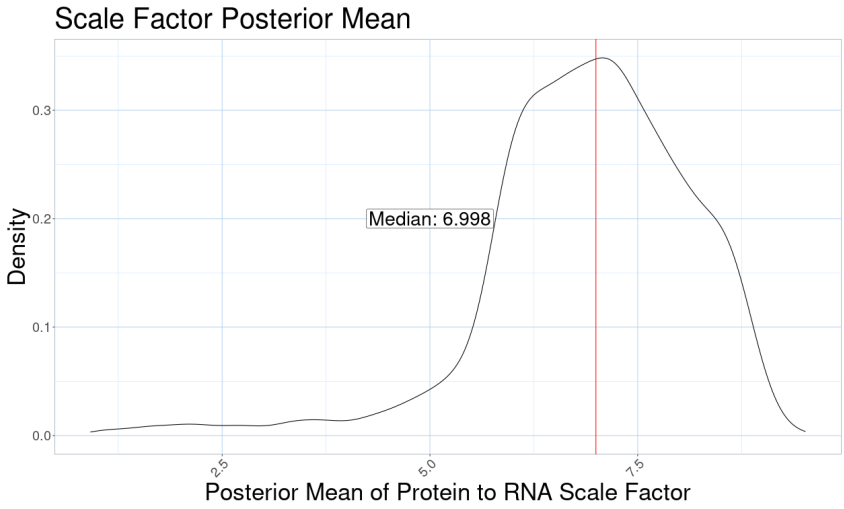

Figure B.4: Posterior Mean of Protein Scale Parameter. A red line indicates the median across genes.

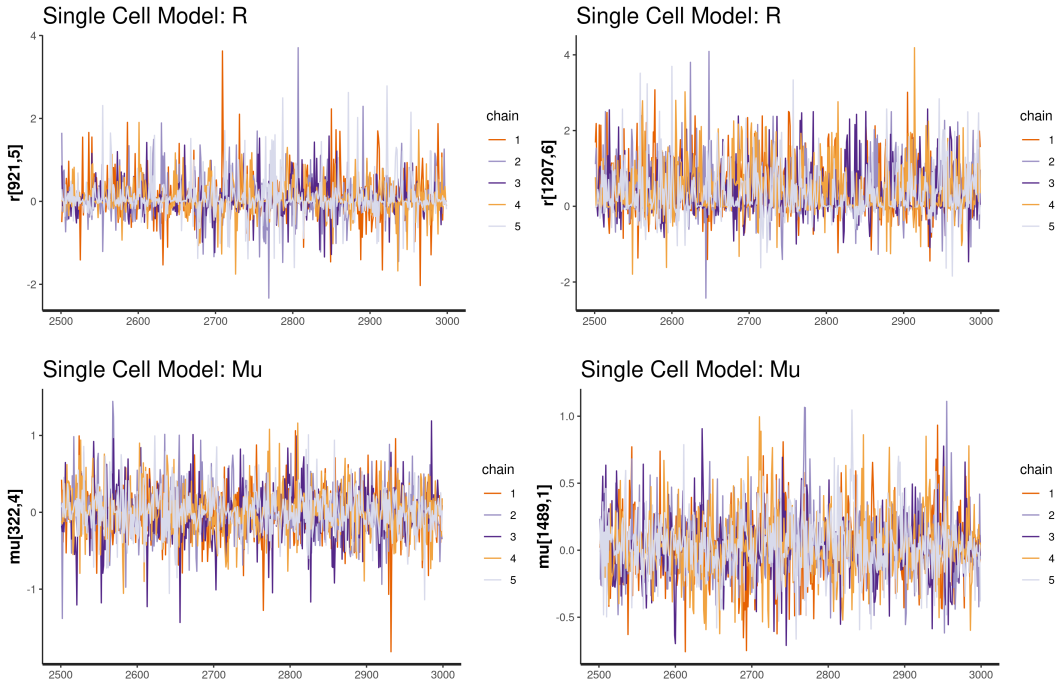

Figure B.5: Trace plots for R and μ parameters in randomly selected genes and tissues.

B.3 Additional results from GO Testing

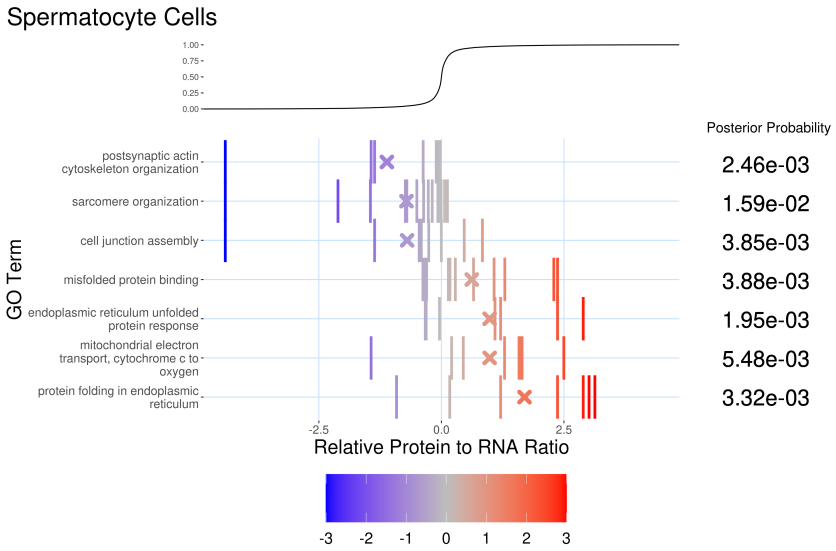

Figure B.6: Genes in Significant Groups for Cell Type SPC. Each tick represents the posterior mean of R_{gc} for genes associated with each GO group. x marks represent group-level averages, and color corresponds to posterior mean of each gene. Beside the tick marks we can see posterior probabilities for this test, as well as the cumulative distribution function for the posterior mean of R across all genes.

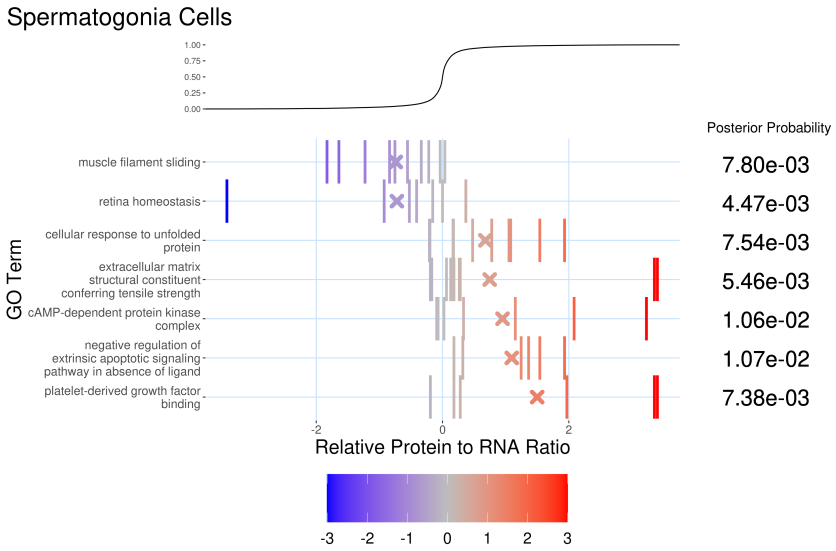

Figure B.7: Genes in Significant Groups for Cell Type SPG. Each tick represents the posterior mean of R_{gc} for genes associated with each GO group. x marks represent group-level averages, and color corresponds to posterior mean of each gene. Beside the tick marks we can see posterior probabilities for this test, as well as the cumulative distribution function for the posterior mean of R across all genes.

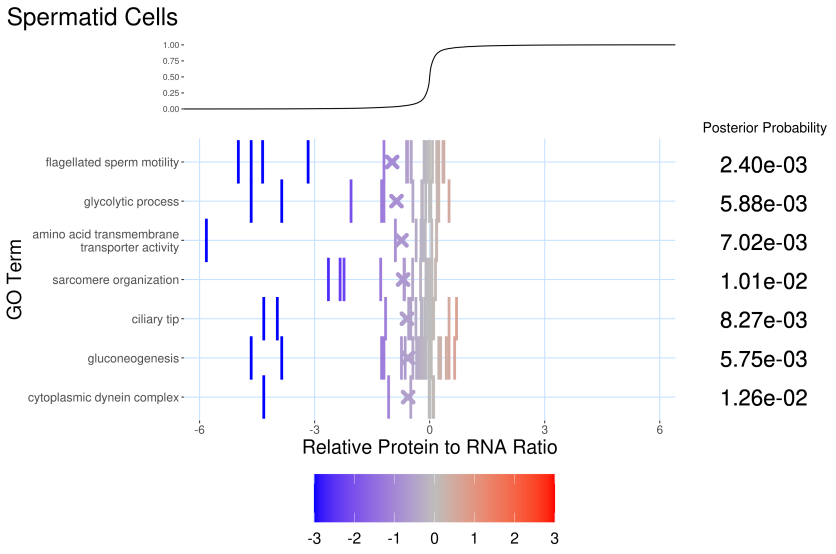

Figure B.8: Genes in Significant Groups for Cell Type St. Each tick represents the posterior mean of R_{gc} for genes associated with each GO group. x marks represent group-level averages, and color corresponds to posterior mean of each gene. Beside the tick marks we can see posterior probabilities for this test, as well as the cumulative distribution function for the posterior mean of R across all genes.

Appendix C

Appendix Chapter 4

C.1 Additional ComBat-seq plots

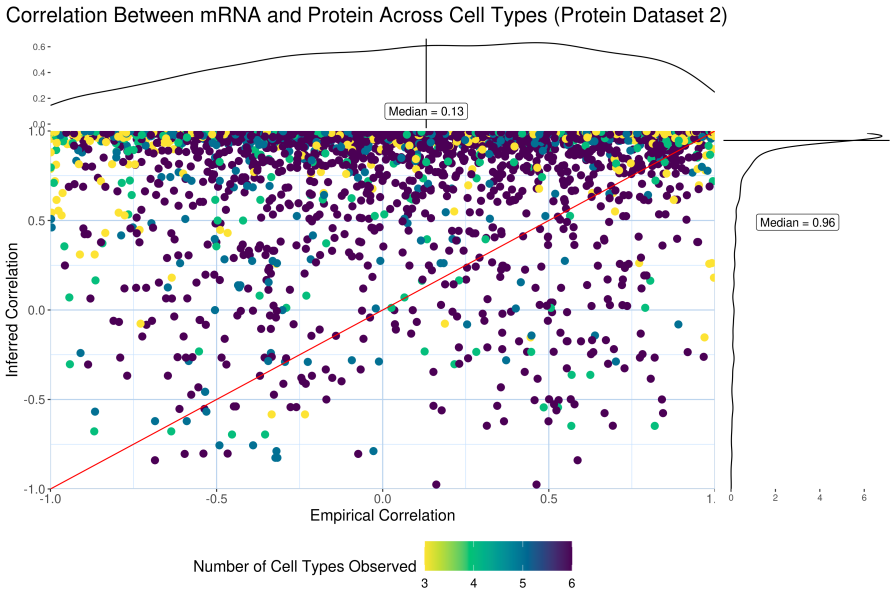

Figure C.1: mRNA, Protein Correlation Across Cell Types applied to ComBat-seq-processed data. Correlations across cell types using the posterior mean of each gene, cell type are plotted against empirical across cell type correlations using protein data set 2. Point color represents the number of cell types observed for each gene, and a red diagonal line is at $y = x$. Opposite each axis label is a marginal density plot of corresponding correlations, with vertical lines denoting the labeled median value.

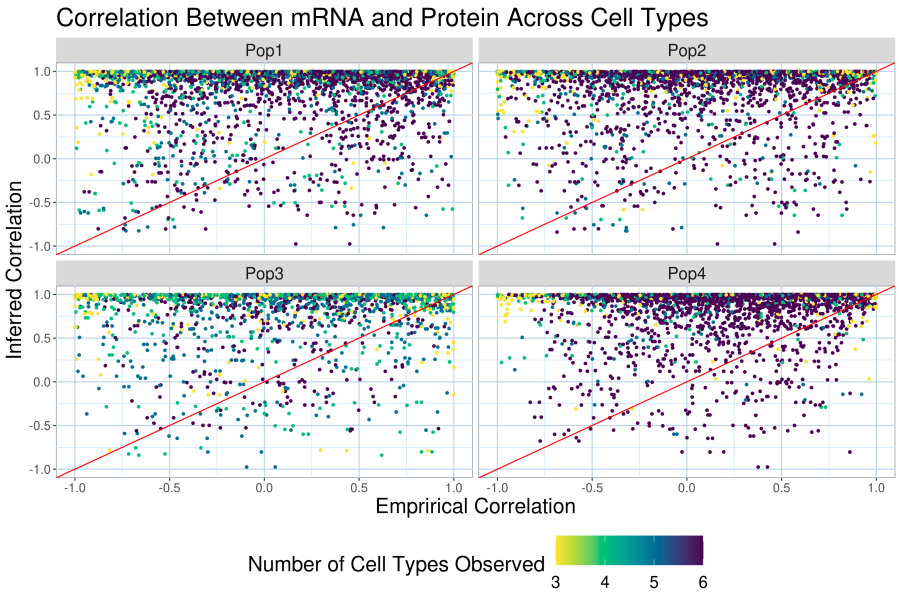

Figure C.2: mRNA, protein across cell type correlations for all protein and mRNA data sets using ComBat-seq corrected observations. y-axis is inferred correlation between mRNA and protein using μ_{gc} and $\mu_{gc} + R_{gc}$ across cell types c. x-axis is the empirical correlation between each faceted protein population and both of the mRNA populations. A red line depicts $y = x$.

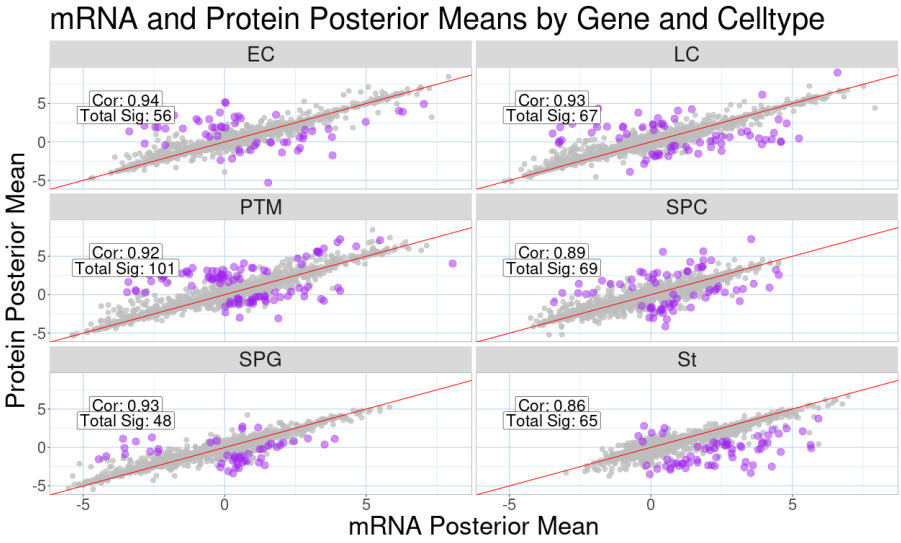

Figure C.3: Posterior Means for $\log_2$ relative mRNA transcription and protein abundance per gene and cell type applied to ComBat-seq-processed model. Protein posterior mean, $\mu_{gc} + R_{gc}$, is plotted by mRNA posterior mean, μ_{gc}. Purple points denote genes significantly different from zero at a 1% FDR.

Table C.1: Proportion of ComBat-seq corrected protein gene, cell type averages within 95% posterior predictive interval. Using upper and lower quantiles based on posterior predictive draws, the proportion of ComBat-seq adjusted protein data set values within these bounds are presented.

Cell Type	Protein Pop1	Protein Pop2	Protein Pop3	Protein Pop4
EC	0.979	0.970	0.891	0.983
LC	0.952	0.979	0.967	0.955
PTM	0.920	0.959	0.967	0.957
SPC	0.984	0.981	0.984	0.987
SPG	0.978	0.991	0.983	0.993
St	0.989	0.976	0.747	0.978

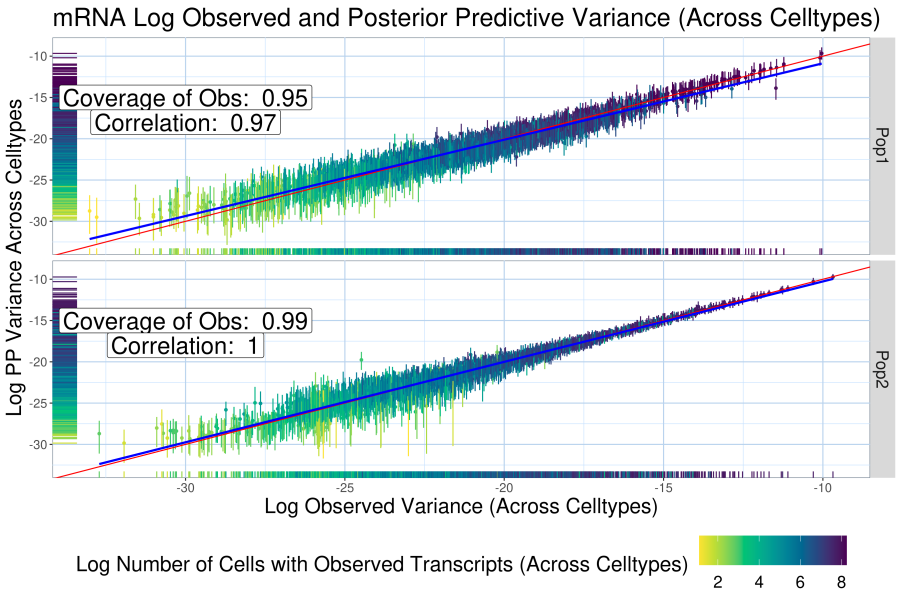

Figure C.4: Observed and Posterior Predictive mRNA Variance Across Cell Types. The log variance of ComBat-seq corrected averages across cell types is represented on the x-axis, as well as the 95% interval for the across cell types variance of posterior predictive draws. Color represents number of observed cells, and a red line represents $y = x$. The coverage of observed values by the posterior predictive variance, as well as the correlation of observed variances and median posterior predictive variance are labeled for each mRNA data set.

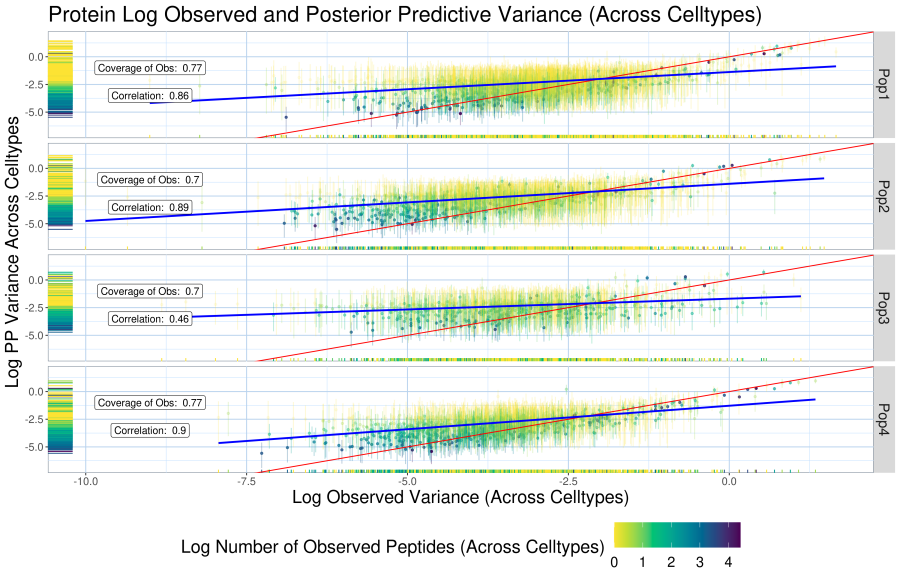

Figure C.5: Observed and Posterior Predictive Protein Variance Across Cell Types. The log variance of ComBat-seq corrected averages across cell types is represented on the x-axis, as well as the 95% interval for the across cell types variance of posterior predictive draws. Color represents number of observed cells, and a red line represents $y = x$.

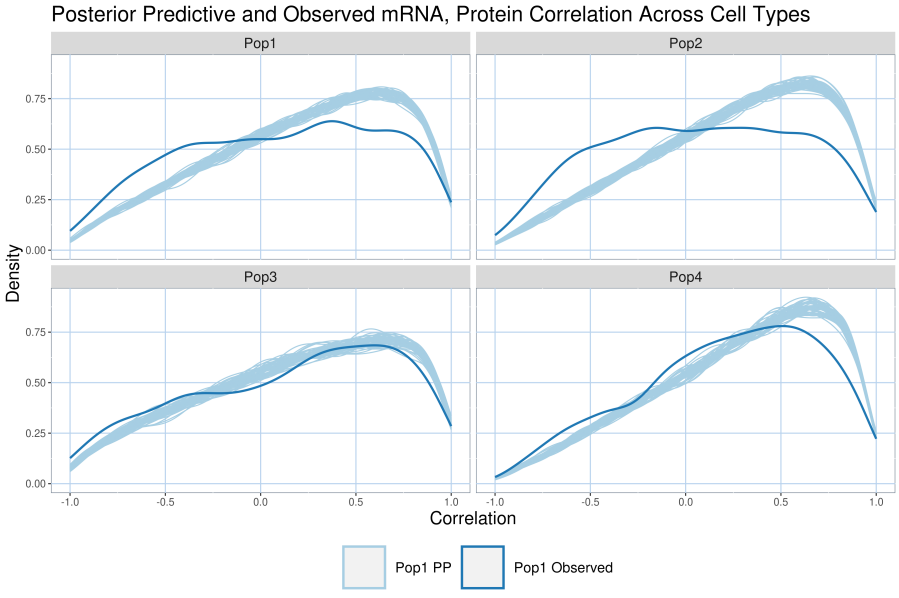

Figure C.6: ComBat-Seq processed and Posterior Predictive mRNA, Protein Correlation Across Cell Types. Density plots are displayed for observed correlation of mRNA, protein across cell types for all four protein populations and mRNA data set 1.

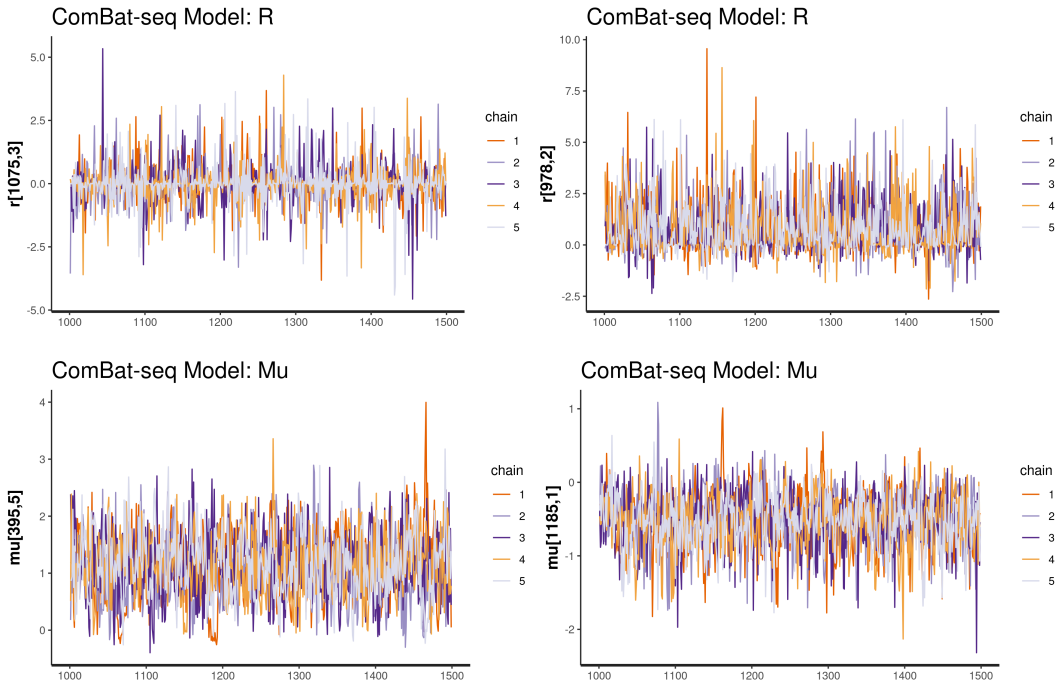

Figure C.7: Trace plots for R and μ parameters in randomly selected genes and tissues using model run on ComBat-seq adjusted counts.

C.2 Additional SCTransform plots

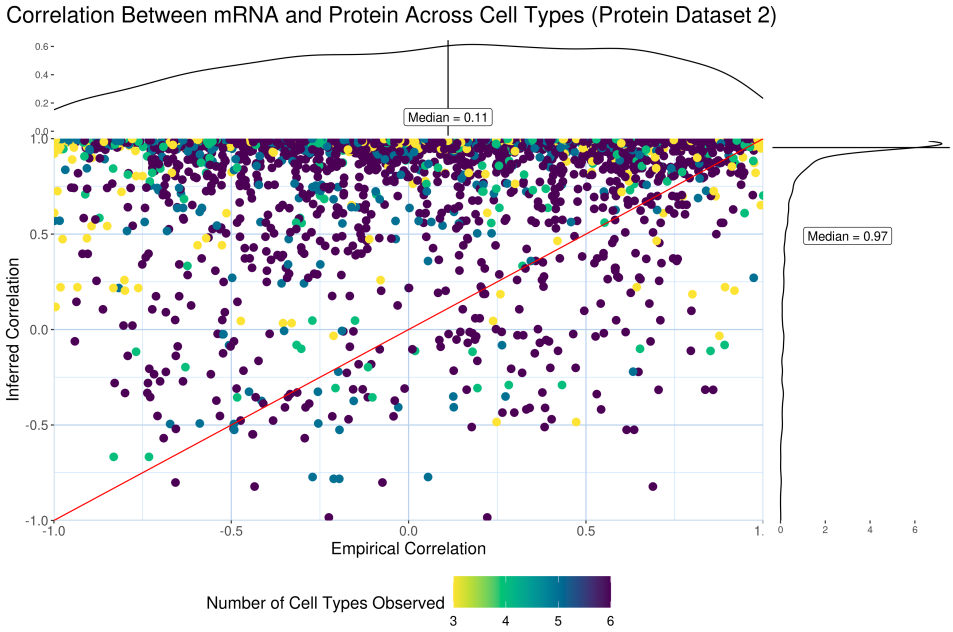

Figure C.8: mRNA, Protein Correlation Across Cell Types applied to SCTransform-processed data. Correlations across cell types using the posterior mean of each gene, cell type are plotted against empirical SCTransform-corrected across cell type correlations using protein data set 2. Density plots are shown at the margins for empirical and inferred correlation across genes.

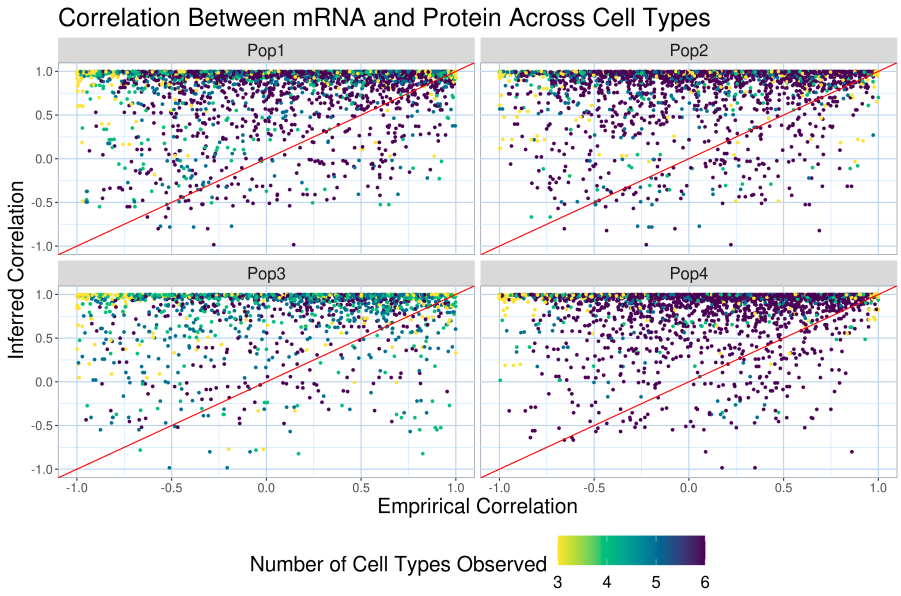

Figure C.9: mRNA, protein across cell type correlations for all protein and mRNA data sets using SCTransform adjusted counts. y-axis is the inferred correlation between mRNA and protein using μ_{gc} and $\mu_{gc} + R_{gc}$ across cell types c. x-axis is the empirical correlation between each faceted protein population and both of the mRNA populations.

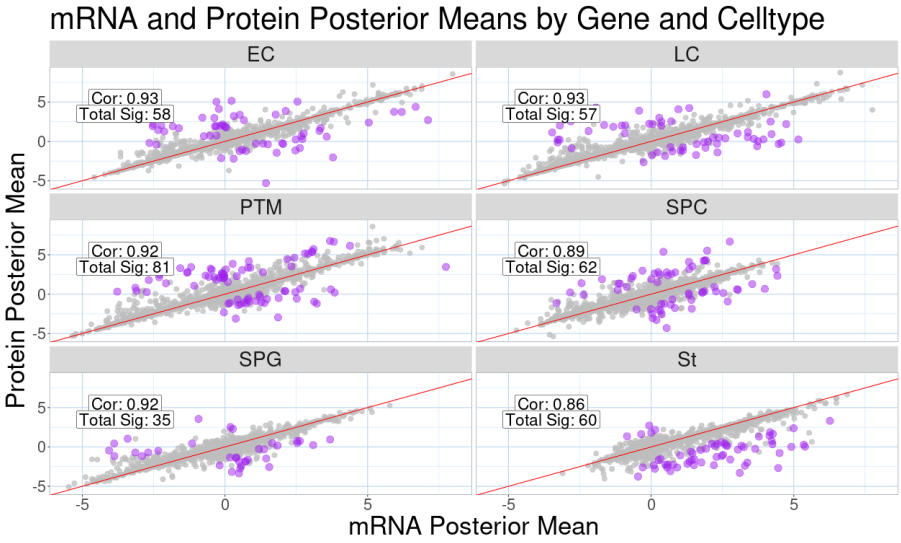

Figure C.10: Posterior Means for mRNA expression, Protein abundance per Gene and Cell Type applied to SCTransform-processed model. Protein posterior mean, $\mu_{gc}+R_{gc}$, is plotted by mRNA posterior mean, μ_{gc}. Purple points denote genes significantly different from zero at a 1% FDR. We can see the number of these points labeled as "Total Sig", as well as the correlation across genes of these posterior averages.

Table C.2: Proportion of SCTranform adjusted protein gene, cell type averages within 95% posterior predictive interval. Using upper and lower quantiles based on posterior predictive draws, the proportion of adjusted protein observed values within these bounds are presented.

Cell Type	Protein Pop1	Protein Pop2	Protein Pop3	Protein Pop4
EC	0.979	0.971	0.899	0.984
LC	0.954	0.977	0.965	0.951
PTM	0.920	0.959	0.968	0.956
SPC	0.982	0.983	0.985	0.987
SPG	0.976	0.990	0.982	0.993
St	0.988	0.978	0.742	0.982

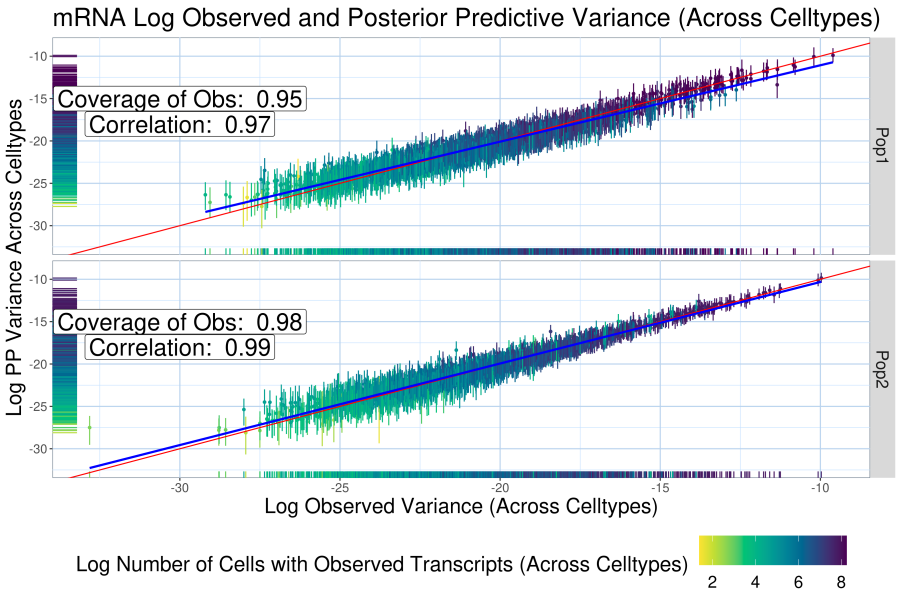

Figure C.11: Observed and Posterior Predictive mRNA Variance Across Cell Types. The log variance of SCTransform corrected averages across cell types is represented on the x-axis, as well as the 95% interval for the across cell types variance of posterior predictive draws.

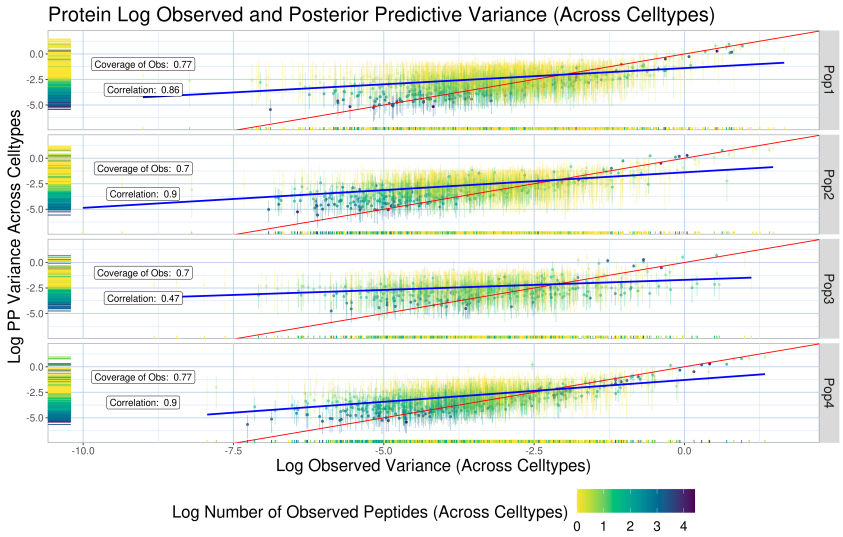

Figure C.12: Observed and Posterior Predictive Protein Variance Across Cell Types. The log variance of SCTransform corrected averages across cell types is represented on the x-axis, as well as the 95% interval for the across cell types variance of posterior predictive draws. Color represents number of observed cells, and a red line represents $y = x$.

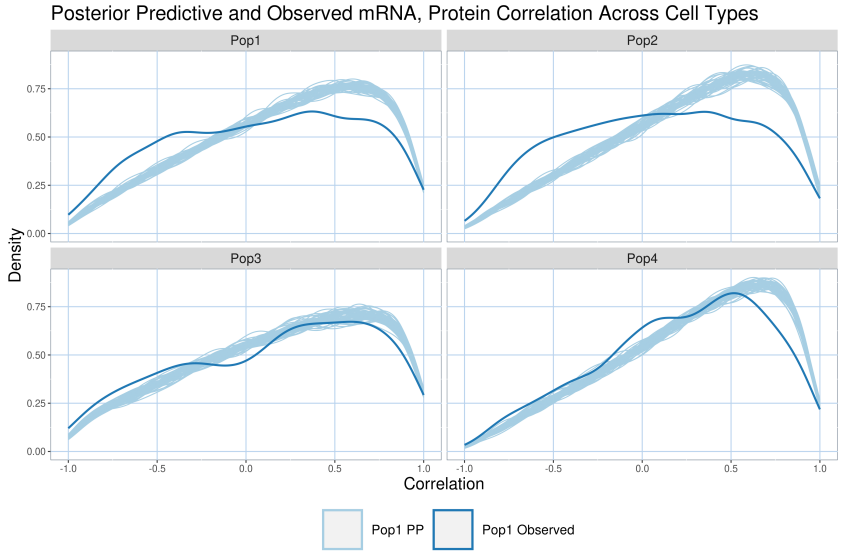

Figure C.13: SCTransform processed and Posterior Predictive mRNA, Protein Correlation Across Cell Types. Density plots are displayed for observed correlation of mRNA, protein across cell types for all four protein populations and mRNA data set 1.

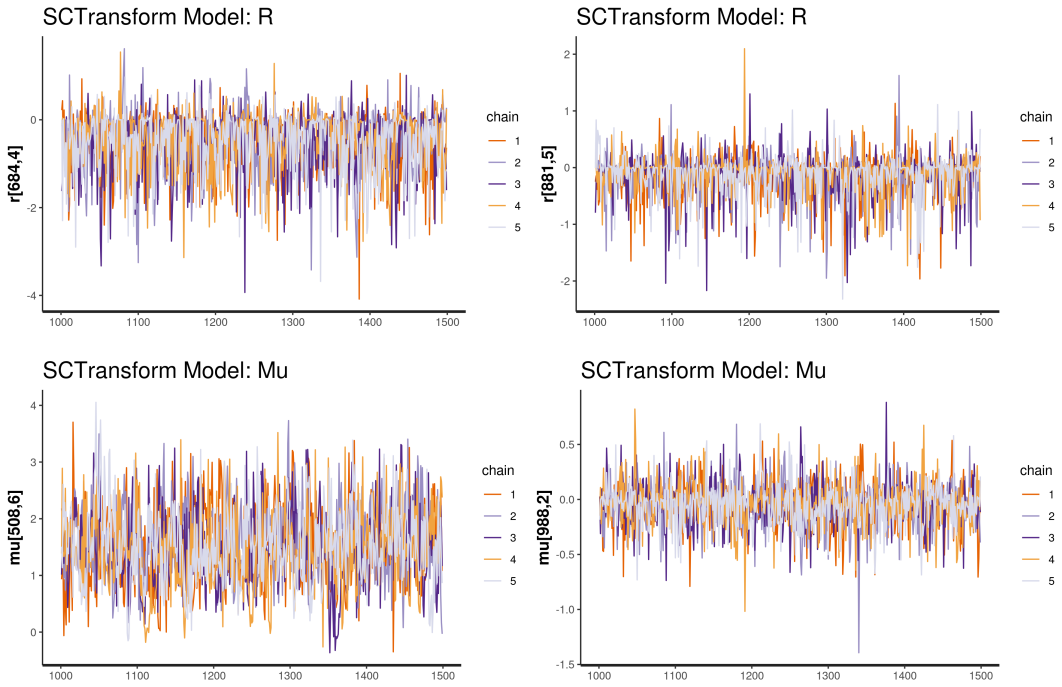

Figure C.14: Trace plots for R and μ parameters in randomly selected genes and tissues for model applied to SCTransform-corrected counts.

C.3 Additional model comparison results

In addition to examining influence of preprocessing choice on model fit for mRNA points, we can see how choice of mRNA preprocessing indirectly influences protein fit. In table C.3, we can see LOO-PSIS summary information for the protein fit. There is an increase in points associated with a "good" fit relative to the mRNA observations, and applying the model to observed data directly has the best LOO-PSIS results relative to SCTransform and ComBat-seq correction. This can further indicate that by making direct improvements to the model to account for batch and sequencing depth at the gene level may yield better results than applying the model to corrected data.

Table C.3: LOO-PSIS Comparison for protein Observations. We can see the difference in elpd for each model, standard error, and the proportion of points labeled to have "good" fit based on Pareto diagnostics.

Model	**elpd diff**	**se diff**	**good prop**
Original	$0.00e+00$	$0.00e+00$	77.7
SCTransform	-106.472	19.664	77.4
ComBat-seq	-170.944	15.792	77.0

Bibliography

[1] Alberts B, Bray D, Hopkin K, Johnson A, Lewis J, Raff M, Roberts K, Walter P., *Essential Cell Biology.* W. W. Norton & Company., fifth ed., 2019.

[2] Schena M, Shalon D, Davis R W, Brown P O., *Quantitative monitoring of gene expression patterns with a complementary dna microarray.*, *Science.* **270** (1995) 467–470.

[3] Aebersold R, Mann M., *Mass spectrometry-based proteomics.*, *Nature.* **422** (2003) 198–207.

[4] Wang H, Mann M., *Advances and challenges in mass spectrometry-based proteomics data analysis.*, *Journal of Proteomics.* (2020), no. 220.

[5] Wilhelm M, Schlegl J, Hahne H, Gholami A, Lieberenz M, et al., *Mass-spectrometry-based draft of the human proteome.*, *Nature.* **509** (2014) 582–587.

[6] Edfors F, Danielsson F, Hallström B M, Käll L, Lundberg E, Pontén F, Uhlén M., *Gene-specific correlation of rna and protein levels in human cells and tissues.*, *Molecular Systems Biology.* **12** (2016) 883.

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

[8] Butler A, Hoffman P, Smibert P., *Integrating single-cell transcriptomic data across different conditions, technologies, and species.*, *Nat Biotechnol.* **36** (2018) 411–420.

[9] Zhang Y, Parmigiani G, Johnson W E., *ComBat-seq: batch effect adjustment for RNA-seq count data.*, *NAR Genomics and Bioinformatics.* **2** (09, 2020) [https://academic.oup.com/nargab/article-pdf/2/3/lqaa078/34054992/lqaa078.pdf].

[10] Wang Z, Gerstein M, Snyder M, *Rna-seq: a revolutionary tool for transcriptomics.*, *Nature Reviews Genetics.* **10** (2009) 57–63.

[11] Aebersold R, Mann M., *Mass-spectrometric exploration of proteome structure and function.*, *Nature.* **537** (2016) 347–355.

[12] Ha M, Kim V N., *Regulation of microrna biogenesis.*, *Nature Reviews Molecular Cell Biology.* **15** (2014) 509–524.

[13] Martin K C, Ephrussi A., *mrna localization: gene expression in the spatial dimension.*, *Cell.* **136** (2009) 719–730.

[14] van den Berge K, Roux de Bézieux H, Street K, Saelens W, Cannoodt R, Saeys Y, Dudoit S., *Trajectory-based differential expression analysis for single-cell sequencing data.*, *Nature Communications.* **11** (2020) 1–12.

[15] Csárdi G, Franks A, Choi D S, Airoldi E M, Drummond D A., *Accounting for experimental noise reveals that mrna levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast.*, *PLoS Genet.* **11** (2015).

[16] Li J, Tibshirani R., *Finding consistent patterns: A nonparametric approach for identifying differential expression in rna-seq data.*, *Statistical Methods in Medical Research.* **22** (2013) 519–536.

[17] Franks A, Airoldi E, Slavov N., *Post-transcriptional regulation across human tissues.*, *PLOS Computational Biology.* **13** (2017).

[18] Djebali S, Davis C A, Merkel A, Dobin A, Lassmann T, Mortazavi A., *Landscape of transcription in human cells.*, *Nature.* **489** (2012) 101–108.

[19] Fagerberg L, Hallstrom B M, Oksvold P, Kampf C, Djureinovic D, Odeberg J., *Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics.*, *Molecular & Cellular Proteomics.* **13** (2014) 397–406.

[20] Kim M S, Pinto S M, Getnet D, Nirujogi R S, Manda S S, Chaerkady R, et al., *A draft map of the human proteome.*, *Nature.* **509** (2014) 575–581.

[21] Ilicic T, McCarthy D J, *Bulk rna-seq for investigating the transcriptome of single cells.*, *Briefings in Functional Genomics.* **16** (2016) 59–62.

[22] Rau A, Marot G, Jaffrézic F., *Differential meta-analysis of rna-seq data from multiple studies.*, *BMC Bioinformatics.* **15** (2014) 91.

[23] van Dijk D et al., *Recovering gene interactions from single-cell data using data diffusion.*, *Cell.* **174** (2018) 716–729.

[24] Huang M et al., *Saver: gene expression recovery for single-cell rna sequencing.*, *Nat Methods.* **15** (2018), no. 539.

[25] Hou W, Ji Z, Ji H et al., *A systematic evaluation of single-cell rna-sequencing imputation methods.*, *Genome Biol.* **21** (2020), no. 218.

[26] Bhaskaran K, Smeeth L., *What is the difference between missing completely at random and missing at random?*, *Int J Epidemiol.* **43** (2014), no. 4.

[27] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A., *Stan: A probabilistic programming language.*, *Journal of statistical software.* **76** (2017), no. 1.

[28] G. A. Hoffman M D, *The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo.*, *Journal of Machine Learning Research.* **15** (2014) 1351–1381.

[29] Pagès H, Carlson M, Falcon S, Li N, *Annotationdbi: Manipulation of sqlite-based annotations in bioconductor. r package version 1.62.1.*, 2023.

[30] Crick F H, *Central dogma of molecular biology.*, *Nature.* **227** (1970) 561–563.

[31] Naryzny S N, Legina O K., *Haptoglobin as a biomarker.*, *Biochem.* **15** (2021) 184–198.

[32] Naora H, Takai I, Adachi M, Naora H., *Altered Cellular Responses by Varying Expression of a Ribosomal Protein Gene: Sequential Coordination of Enhancement and Suppression of Ribosomal Protein S3a Gene Expression Induces Apoptosis.*, *Journal of Cell Biology.* **141** (05, 1998) 741–753.

[33] Janeway C A Jr, Travers P, Walport M, et al., *The complement system and innate immunity*, *Immunobiology: The Immune System in Health and Disease.* **5** (2001).

[34] Carpenter S, Ricci E, Mercier B et al., *Post-transcriptional regulation of gene expression in innate immunity.*, *Nat Rev Immunol.* **14** (2014) 361–376.

[35] Song L, Luo Z Q., *Post-translational regulation of ubiquitin signaling.*, *Journal of Cell Biology.* **218** (04, 2019) 1776–1786.

[36] Laity J H, Lee B M, Wright P E., *Zinc finger proteins: new insights into structural and functional diversity.*, *Current opinion in structural biology.* **11** (2001) 39–46.

[37] Kluve-Beckerman B, Smith J T, Ivancic C, Benson M D., *Post-translational modification of amyloid a protein in patients with aa amyloidosis.*, *Amyloid.* **29** (2022), no. 1 50–57, [https://doi.org/10.1080/13506129.2021.1997985]. PMID: 34787027.

[38] Chapin JC , Hajjar K A., *Fibrinolysis and the control of blood coagulation.*, *Blood Rev.* **29** (2015) 17–24.

[39] Luecken M D, Theis F J., *Current best practices in single cell RNA-seq analysis: a tutorial.*, *Molecular Systems Biology.* **15** (2019).

[40] Haque A, Engel J, Teichmann S A, *A practical guide to single-cell rna-sequencing for biomedical research and clinical applications.*, *Genome Med.* **9** (2017) 75.

[41] Petelski A, Emmott E, Leduc A, Huffman G, Specht H, Perlman D, Slavov N., *Multiplexed single-cell proteomics using scope2.*, *Nature protocols.* **16** (2021), no. 12.

[42] Cox J, Mann M., *Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.*, *Nat Biotechnol.* **26** (2008) 1367–72.

[43] Daran-Lapujade P, Rossell S, M. van Gulik W, Luttik M, J L de Groot M, Slijper M, J R Heck A, Daran JM, de Winde JH, Westerhoff HV, Pronk JT, Bakker BM., *The fluxes through glycolytic enzymes in ¡i¿saccharomyces cerevisiae are predominantly regulated at posttranscriptional levels.*, *Proceedings of the National Academy of Sciences.* **104** (2007), no. 40 15753–15758.

[44] M. LLC., "Chan Zuckerberg Initiative mapping the transcriptome and proteome of human testis in 3d.."

[45] Stoeckius, M., Hafemeister, C., Stephenson, W. et al., *Simultaneous epitope and transcriptome measurement in single cells.*, *Nat Methods.* **14** (2017) 865–868.

[46] Leduc A, Huffman GR, Slavov N., *Droplet sample preparation for single-cell proteomics applied to the cell cycle.*, *bioRxiv* (2021) [https://www.biorxiv.org/content/early/2021/04/26/2021.04.24.441211.full.pdf].

[47] Chen A, Franks A, Slavov N., *Dart-id increases single-cell proteome coverage.*, *PLoS Com- put Biol.* (2019).

[48] Shchelochkov OA., *Cdna (copy dna).*, 2023. May 4 2023.

[49] Tavakoli S, Liu Y, Potts J, Rouhanifard S H., *Click chemistry-based amplification and detection of endogenous rna and dna molecules in situ using clampfish probes.*, *Methods in enzymology.* **641** (2020) 459–476.

[50] Stephanie C H, Townes DW, Teng M, Irizarry R A., *Missing data and technical variability in single-cell rna-sequencing experiments.*, *bioRxiv* (2017) [https://www.biorxiv.org/content/early/2017/05/08/025528.full.pdf].

[51] Slavov N, *Types of missing data in proteomics.*, 2021.

[52] Guan S, Taylor PP, Han Z, Moran M F, Ma B., *Data dependent–independent acquisition (ddia) proteomics.*, *Journal of Proteome Research.* **19** (2020), no. 8 3230–3237, [https://doi.org/10.1021/acs.jproteome.0c00186]. PMID: 32539411.

[53] Borràs E, Sabidó E., *What is targeted proteomics? a concise revision of targeted acquisition and targeted data analysis in mass spectrometry.*, *PROTEOMICS.* **17** (2017), no. 17-18 1700180, [https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.201700180].

[54] Slavov N, Franks A., *Non-ignorable missing data.*, 2021.

[55] Mack C, Su Z, Westreich D., *Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide.* Agency for Healthcare Research and Quality (US)., 3 ed., 2018.

[56] Linderman G C, Zhao J, Roulis M et al., *Zero-preserving imputation of single-cell rna-seq data.*, *Nat Commun.* **13** (2022), no. 192.

[57] Li W V, Li J J., *An accurate and robust imputation method scimpute for single-cell rna-seq data.*, *Nat. Commun.* **9** (2018), no. 997.

[58] Gilbert S F, *Developmental Biology.*, vol. 6. Sinauer Associates., 2000.

[59] Cao J et. al., *The single-cell transcriptional landscape of mammalian organogenesis.*, *Nature.* **566** (2019) 496–502.

[60] Joshua W, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko E., *Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity.*, *Cell.* **177** (2019) 1873–1887.

[61] Zheng Y, *Single Cell Multi-modal Analysis Using scDMVAE with an Emphasis on SCoPE2 Technology.* PhD thesis, University of California Santa Barbara., 2023.

[62] Tim S, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck W III, Hao Y, Stoeckius M, Smibert P, Satija R., *Comprehensive integration of single-cell data.*, *Cell.* **177** (2019) 1888–1902.

[63] McInnes L, Healy J, Melville J., *Umap: Uniform manifold approximation and projection for dimension reduction.*, .

[64] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I., *Attention is all you need.*, *Advances in neural information processing systems.* **30** (2017).

[65] Jarvis R, and Patrick, Edward A., *Clustering using a similarity measure based on shared near neighbors.*, *IEEE Transactions on computers.* **100** (1973), no. 11 1025–1034.

[66] Yang Z, Michailidis G., *A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data.*, *Bioinformatics.* **32** (2016), no. 1 1–8.

[67] Choudhary S, Satija R., *Comparison and evaluation of statistical error models for scrna-seq.*, *Genome Biol.* **23** (2022), no. 27.

[68] Wiśniewski J R Mann M., *A proteomics approach to the protein normalization problem: Selection of unvarying proteins for ms-based proteomics and western blotting.*, *Journal of Proteome Research.* **15** (2016), no. 7 2321–2326. PMID: 27297043.

[69] Mulè M P, Martins A J, Tsang J S, *Normalizing and denoising protein expression data from droplet-based single cell profiling.*, *Nat Commun.* **13** (2022), no. 2099.

[70] Lun L, Aaron T, Bach K, Marioni J C. *Genome Biology.* **17** (2016).

[71] R. D., *Understanding the bayesian approach to false discovery rates.*, 2022. May 4 2023.

[72] Sun X, Kovacs T, Hu Y J, Yang W X., *The role of actin and myosin during spermatogenesis. molecular biology reports.*, .

[73] Ellgaard L, McCaul N, Chatsisvili A, Braakman I., *Co- and post-translational protein folding in the er.*, *Traffic.* **17** (2016), no. 6 615–638, [https://onlinelibrary.wiley.com/doi/pdf/10.1111/tra.12392].

[74] Ringers C, Olstad E W, Jurisch-Yaksi N., *The role of motile cilia in the development and physiology of the nervous system.*, *Philosophical Transactions of the Royal Society B: Biological Sciences.* **375** (2020), no. 1792 20190156, [https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2019.0156].

[75] Gareau J R, Lima C D., *The sumo pathway: emerging mechanisms that shape specificity, conjugation and recognition.*, *Nat Rev Mol Cell Biol.* **11** (2010) 861–71.

[76] Hafemeister C, Satija R., *Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression.*, *Genome Biol.* **20** (2019), no. 296.

[77] Lütge A, Zyprych-Walczak J, Brykczynska Kunzmann U, Crowell H L, Calini D, Malhotra D, Soneson C, Robinson M D., *Cellmixs: quantifying and visualizing batch effects in single-cell rna-seq data.*, *Life Sci Alliance.* **4** (Mar, 2021).

[78] Zhang Y, Jenkins D F, Manimaran S et al., *Alternative empirical bayes models for adjusting for batch effects in genomic studies.*, *BMC Bioinformatics* **19** (2018), no. 262.

[79] Campbell R K, Fukui D J et al., *Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors.*, *Nature Biotechnology.* **36** (2015) 421–427.

[80] van de Wiel M A, Te Beest D E, Münch M M., *Learning from a lot: Empirical bayes for high-dimensional model-based prediction.*, *Scand Stat Theory Appl.* **46** (2019) 2–25.

[81] Law C W, Chen Y, Shi W, Smyth G K., *voom: precision weights unlock linear model analysis tools for rna-seq read counts.*, *Genome Biol.* **15** (2014).

[82] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[83] Leek J T, Johnson W E, Parker H S, Fertig E J, Jaffe A E, Zhang Y, Storey J D, Torres L C., *sva: Surrogate variable analysis. r package version 3.48.0.*, 2023.

[84] Gentleman R C, Carey V J, Bates D M et al., *Bioconductor: open software development for computational biology and bioinformatics.*, 2004.

[85] Vehtari A, Gabry J, Magnusson M, Yao Y, Bürkner P C, Paananen T, Gelman A., *loo: Efficient leave-one-out cross-validation and waic for bayesian models.*, 2023. R package version 2.6.0.

[86] A. Neufeld, L. L. Gao, J. Popp, A. Battle, and D. Witten, *Inference after latent variable estimation for single-cell rna sequencing data*, 2022.