Statistical Hurdle Models for Single Cell Gene Expression: Differential Expression and Graphical Modeling

Andrew McDavid

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

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Raphael Gottardo, Chair

Mathias Drton, Chair

Noah Simon

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Statistical Hurdle Models for Single Cell Gene Expression:
Differential Expression and Graphical Modeling

Andrew McDavid

Co-Chairs of the Supervisory Committee:
Affiliate Professor Raphael Gottardo
Statistics

Professor Mathias Drton
Statistics

This dissertation describes a set of statistical methods developed for analysis of single cell gene expression. A characteristic of single cell expression is bimodal expression, in which two clusters of expression are present. In any given transcript, the null cluster corresponds to cells without detectable expression (hence a non-zero measurement reflects measurement error) while the signal cluster contains cells with a positive, detectable level of expression. Statistical models that accommodate this characteristic are considered.

- In Chapter 1, motivation and history of single cell gene expression is considered. Scientific and statistical questions addressable through single cell expression are discussed, and some statistical frameworks for bulk and single cell expression are described.

- In Chapter 2, I consider data generated from replicates of single cells and 100 cell aggregates that were assayed through single cell reverse-transcriptase qPCR (rt-qPCR). In rt-qPCR the null cluster manifests as bona-fide zeros, so expression is characterized by zero-inflation of otherwise continuous values. The average expression from single cells and 100-cell replicates is compared to develop quality control metrics that optimize

the single-cell, 100-cell concordance. A Hurdle model is proposed, which accounts for the fact that genes at the single-cell level can be *on* (and a continuous expression measure is recorded) or dichotomously *off* (and the recorded expression is zero). Based on this model, I derive a combined likelihood-ratio test for differential expression that incorporates both the discrete and continuous components. This chapter was originally published in McDavid et al. [2013].

- In Chapter 3, I consider application of the Hurdle model to single cell RNA sequencing (scRNAseq). In these technologies, the binary zero-inflation described found in rt-qPCR-based assays manifests itself as continuous, bimodal expression, motivating a clustering and thresholding procedure to assign expression to a cluster. The Hurdle model, extended and cast as a vector generalized linear model (vGLM), is provided as an `R` package named `MAST`. The cellular detection rate (CDR) is defined as the number of expressed genes found in a cell. It is identified as an important latent factor in single cell experiments, and is argued to measure size and efficiency variations among cells. Gene set enrichment analysis using the Hurdle model, and use of residuals defined through such models are discussed. Parts of this chapter were originally published in Finak et al. [2015], McDavid et al. [2014].

- In Chapter 4, the Hurdle model is generalized to model multivariate dependences between cells, permitting the parametrization of graphical models. A neighborhood selection-based method is proposed to leverage group-ℓ_1 penalized regression. Networks estimated on single-cell and multi-cell experiments are contrasted and found to be very distinct. In order to synthesize graphs estimated on transcriptome-scale data, a test for enrichment of connections between and within gene ontology categories is proposed.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables xi

Abbreviations xii

Chapter 1: Introduction 1
1.1 Biological motivation for single cell gene expression 1
1.2 Distributional properties of single cell gene expression 2
1.3 Previous methods for bulk and single cell expression 4
1.4 Statistical frameworks 10

Chapter 2: Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments 14
2.1 Single cell qPCR methods 14
2.2 A two-part Hurdle model 19
2.3 Quality control and filtering 20
2.4 Testing for differences between experimental groups 21
2.5 Application of filtering and Hurdle model in three datasets 26
2.6 Discussion 33

Chapter 3: Transcriptional change and heterogeneity in single-cell RNA sequencing data 39
3.1 Previous methods for single cell whole-transcriptome sequencing 39
3.2 Robust estimation of a vector regression model in scRNAseq 43
3.3 Adjusting for the cellular detection rate through vector regression 47
3.4 Single cell gene set enrichment analysis 52
3.5 Residual analysis in hurdle vGLMs 57

3.6 Data sets and biological protocols 61
3.7 Discussion 64

Chapter 4: Graphical models for zero-inflated single cell gene expression 66
4.1 Background 66
4.2 From single cells to cellular co-expression 68
4.3 Hurdle models 72
4.4 Neighborhood estimation via penalized regression 77
4.5 Simulations 81
4.6 T follicular helper cells 82
4.7 Mouse dendritic cells 85
4.8 Discussion 89
4.9 Supplemental methods 91

Bibliography 92

Appendix A: Supplementary derivations and figures 105
A.1 Empirical Bayes derivation of variance hyper parameters 105
A.2 Simulation exploring CDR effects 108
A.3 Supplementary figures 109

LIST OF FIGURES

Figure Number Page

1.1 Scatter plots and kernel density estimates from (a) Fluidigm single cell qPCR data (b) single cell RNA sequencing and (c) single cell RNA sequencing after thresholding. 3

2.1 Histogram and theoretical (normal) distribution of $(et_{ij}|v_{ij} = 1)$ for single cell (left, light gray) and hundred cell experiments (right, dark gray). Genes FASLG, IFN-γ, BIRC3 and CD69 are depicted. The frequency expression of each gene in the single cell experiments $\pi^{(1)}$ is printed above each histogram. The mean of the hundred cell and single cell experiments is indicated by a thick black line along the x-axis. 18

2.2 Normal quantile-quantile plots of z_{ij} for 49 genes, data set A. 22

2.3 The empirical cumulative distribution plot of Λ. The cumulative distribution of χ^2_2 is plotted in gray. 5% significance is indicated by a vertical line. The gene frequency π varies. The sample size $I = 100$. 25

2.4 The empirical cumulative distribution plot of Λ and of χ^2_2, considering departures from normality in the continuous component $\log \mathbf{u}$. $\log \mathbf{u}$ is now simulated from a t distribution with 4 degrees of freedom. 26

2.5 Concordance between hundred cell $y^{(100)}/100$ and $y^{(1)}$, the *in silico* average of single cell wells. The rows correspond to inclusion, case-wise deletion and inclusion and filtering of zeros $v_{ij} = 0$. Dark, thin lines show the initial location of a gene before filtering and connect to the location of the gene after filtering. The concordance correlation coefficient r_c and average weighted squared deviation $\overline{\text{WSS}}$ is printed. The dotted black line shows a loess fit through the data. In all cases, the expression values are transformed using a shifted log-transformation $(\log_2(y + 1))$. 28

2.6 Scatter plots of housekeeping genes GAPDH, POLR2A and other frequently expressed ($\pi > .95$) genes. Cells flagged for filtering are indicated in purple. A regression line of the form $et_y \sim et_x + \text{intercept}$, and its standard error is plotted using unfiltered cells. 35

2.7 Number of discoveries (genes × units) versus the false discovery rate, by treatment, data set A. The combined likelihood ratio test is compared to a Bernoulli or normal-theory only likelihood ratio test, as well as a t-test of the raw expression values (2^{et} scale), including zero measurements. 36

2.8 $-\log_{10} P$ of tests (genes × units) versus frequencies of expression π of the genes. The Bernoulli, normal-theory and combined likelihood ratio tests are plotted. * indicates test is different from the combined test at 5% significance in a Wilcoxon signed-rank test. 37

2.9 Heatmap of signed $\log_{10} p$ for selected genes (rows) and 16 individuals (columns). The color above each column indicates the antigen stimulation applied to the cells. Red and purple are two different CMV antigen pools; yellow and orange are two different HIV antigen pools. 38

3.1 The fraction of genes expressed, or cellular detection rate (CDR), is correlated with the first two principal components of variation in MAIT and DC data sets. 42

3.2 Single-cell expression ($\log_2$-TPM) of the top 100 genes identified as differentially expressed between cytokine (IL18, IL15, IL12) stimulated (purple) and non-stimulated (pink) MAIT cells using MAST (A). Partial residuals for up- and down- regulated genes are accumulated to yield an activation score (B), and this score is consistent with stimulation inducing heterogeneity compared to the unstimulated cells. 50

3.3 Module scores for individual cells for the top 9 enriched modules (A) and decomposed Z-scores (B) for single-cell gene set enrichment analysis in MAIT data set, using the blood transcription modules (BTM) database. The distribution of module scores suggests heterogeneity among individual cells with respect to different biological processes. Enrichment of modules in stimulated and non-stimulated cells is due to a combination of differences in the discrete (proportion) and continuous (mean conditional expression) of genes in modules. The combined Z-score reflects the enrichment due to differences in the continuous and discrete components. 56

3.4 Module scores (A) and decomposed Z-scores (B) for single-cell gene set enrichment analysis for LPS stimulated cells, mDC data set, using the mouse GO biological process database. The change in single-cell module scores over time for the nine most significantly enriched modules in response to LPS stimulation are shown in A. The *core antiviral*, *peaked inflammatory* and *sustained inflammatory* modules are among the top enriched modules, consistent with the original publication. Additionally we identify GO modules *cellular response to interferon-beta* and *response to virus*, which behave analogously to the core antiviral and sustained inflammatory modules. No GO analog for the *peaked inflammatory* module was detected. The majority of modules detected exhibit enrichment relative to the 1h time point (thus increasing with time). The "early marcher" cells identified in the original publication are highlighted here with triangles. We show the top 50 most significant modules (B). The combined Z-score summarizes the changes in the discrete and continuous components of expression. 58

3.5 Gene-gene correlation (Pearson's ρ) of model residuals in non-stimulated (A) and stimulated (B) cells, and principal components analysis biplot of model residuals (C) on both populations using the top 50 marginally differentially expressed genes. As marginal changes in the genes attributable to stimulation and CDR have been removed, clustering of subpopulations in (C) indicates co-expression of the indicated genes on a cellular basis. 60

3.6 Principal components analysis biplot of model residuals (A) and Gene-gene correlation (Pearson's R) of model residuals (B) by time point for LPS cells, mDC experiment using 20 genes with largest log-fold changes, given significant (FDR q <.01) marginal changes in expression. PC1 is correlated with change over time. The two "early marcher" cells are highlighted by an asterisk at the 1h time-point. Correlation structure in the residuals is increasingly evident over time and can be clearly observed at the 6h time-point compared to the earlier time-points. 62

4.1 Scatter plots of inverse cycle threshold ($40 - \text{Ct}$) measurements from a quantitative PCR (qPCR)-based single cell gene expression experiment. The cycle threshold (Ct) is the PCR cycle at which a predefined fluorescence threshold is crossed, so a larger inverse cycle threshold corresponds to greater log-expression [McDavid et al., 2013]. Measurements that failed to cross the threshold after 40 cycles are coded as 0. 70

4.2 Number of true positives vs false positives for simulated chain graphs under *dense* and *sparse dependence* with $m = 64$ nodes and $n = 100$ observations. The ribbon shows the simulation-induced standard errors about the average. The Anisometric and Isometric models use neighborhood selection with the multivariate Hurdle model (4.7) with group-ℓ_1 penalty based on the null-model Fisher information and identity matrix, respectively. The Gaussian, NPN and Logistic models use ℓ_1 penalized neighborhood selection under linear (Gaussian), Normal-score transformed linear (NPN) and logistic regressions. 82

4.3 Sensitivities $\left(\frac{\text{true positives}}{\text{total true}}\right)$ at 10% FDR as m, the number of nodes increases for a chain graph under fixed sparsity. See the caption of figure 4.2 for a description of the methods. 83

4.4 Networks of 35 edges estimated through neighborhood selection under the Hurdle, logistic, Gaussian model (single cells) and Gaussian model (10 cell aggregates) in T follicular helper cells. Brown hues indicate estimated negative dependences, while blue-green hues indicate positive dependences. The edge width and saturation are larger for stronger estimated dependences. 86

4.5 Networks estimated in LPS-treated mouse dendritic cells under Gaussian and Hurdle models. Hub genes are shown in red. Vertex colors indicate gene ontology membership. 88

4.6 Modules enriched at FDR $\leq 5\%$ using graphical geneset edge enrichment in mouse dendritic cells under Gaussian and Hurdle models. 90

A.1 Scatter plot of p-values for differential expression from adaptive and fixed thresholding on the A) MAIT and B) mDC data sets, demonstrating robustness to the thresholding method. Two selected genes from each data set, with large differences in p-values between fixed and adaptive thresholding in C) MAIT and D) mDC, are genes that exhibit substantial bimodality and our adaptive thresholding appears preferable. 110

A.2 Scatter plot of normalized (scaled to unit variance and zero mean) CDR (cellular detection rate) calculated from all genes vs. the CDR calculated from housekeeping genes , for stimulated A) and unstimulated B) MAIT cells. The estimated CDRs are linearly related within each condition. 111

A.3 Amount of variability, measured as percent of null model deviance, attributed to the CDR effect vs. the treatment effect, in each dataset. The CDR accounts for 5.2% of the variability in the MAIT and 4.8% of the variability in the mDC data sets for the average gene. Greater than 9% of the variability is attributed to over 10% of genes in both data sets. CDR contributes the most variability to the discrete component in both data sets and more so in the MAIT data than the mDC data. 112

A.4 Effect of CDR and confounding with treatment using different methods. A) ROC curve comparing the effect of controlling for CDR in the MAST model. The solid line is the median and the top and the bottom dashed line represents the 95 and 5 percent quantile. The result indicates that inclusion of CDR improves the performance when there is confounding between the CDR and stimulation and performs nearly the same when there is no confounding or when there is no CDR effect in the data generating model. B) Density plot of generated CDR values across cells using the three levels of confounding between the stimulation and the CDR effects. 113

A.5 Comparison of the empirical CDR (centered and scaled) and other correction methods, the cell by gene weights of Shalek et. al., and RUV and SVA. The CDR and Shalek et. al. weights are correlated, in fact generally just shifted by a constant (panel A, in a random subsample of genes, each in a different color), and the correlation coefficient is nearly unity (panel B). The location shift between the CDR and Shalek et. al. weights would be absorbed by the intercept term in the logistic regression. C) Scatterplots of CDR vs. the first and second SVA and RUV components. Treatment groups are shown in different colors. The first SVA and second RUV components are associated with CDR. D) In the mDC data, the first SVA and RUV components are correlated with CDR. 114

A.6 Gene Ontology Enrichment Analysis using the GOrilla online tools for the set of genes not detected as differentially expressed in the MAIT data set when the CDR is included in the MAST linear model. 115

A.7 False discoveries in genes (A) and modules (B) based on numeric permutation experiments for various methods. The unstimulated MAIT cells were permuted into two subsets, and were tested for differential expression under the Hurdle model (MAST), Limma, edgeR, and DEseq. In this scenario, any gene discovered is an *a priori* false discovery, so the number of false discoveries is plotted against the FDR-adjusted significance. We show the average values from ten permutations. 116

A.8 The distribution of $-\log$ p-values in permuted datasets is compared to its expected Exponential(1) distribution in (A) the hurdle model and (B) Normal-theory t-tests on the same data. In the smaller MAIT dataset ($N = 73$) the Hurdle is inflated in the tail of the test statistic, producing an additional .6 rejections per 1,000 tests at $\alpha = 10^{-3}$. The t-test is deflated, yielding .5 too few rejections per 1,000 tests at $\alpha = 10^{-3}$ 117

A.9 Proportion of immune-specific GO modules amongst all GO modules enriched in differentially expressed genes in the MAIT data set. Immune-specific GO modules were defined to be terms with experimental evidence codes within the Biological Process ontology that were descendants in the GO graph of the Immune System Process term. Differential expression of genes was determined at three increasing false discovery rate thresholds, and then GO enrichment in differentially expressed genes was tested using the hypergeometric distribution, calling significant enrichment at the 1% FDR level. Inclusion of the CDR in the model for differential expression increases the rate of detection of immune specific modules for the MAST and Limma methods. Among models that do not adjust for CDR, SCDE has highest specificity, but is dominated by MAST under CDR adjustment (SCDE cannot adjust for covariates, so was omitted from the CDR models). 118

A.10 Post-sort experiments via flow cytometry show that the sorted cell populations were over 90% pure MAITs (Figure A), and exhibited a change in cell size upon stimulation (Figure B) and that up to 44% of stimulated MAITs did not respond to cytokine stimulation (Figure C). 119

A.11 Gene set enrichment analysis of the mDC data set, LPS stimulated cells using the BTM (blood transcriptional modules) of Li et. al. Decreased expression for AP-1 transcriptional network genes is observed after LPS stimulation, consistent with previous findings in the literature [de Wit et al., 1996]. Type-1 interferon response and antiviral IFN modules are among the most significantly enriched and are consistent with the findings of the original publication [Shalek et al., 2014]. 120

A.12 Number of modules discovered plotted against FDR-adjusted significance of the module. MAST-based GSEA detects more modules than other methods. 121

A.13 Comparison of raw expression values ($\log_2$ TPM) and coefficients estimates (Unstimulated as reference) of modules identified as differentially expressed using MAST GSEA but not with CAMERA. Differences in the expression profile are evident, however CAMERA failed to detect them. A) Violin plots showing the expression of genes in the "T-cell surface signature" module. B) Model coefficient estimates for the genes in the 'T-cell surface signature module from GSEA, with 95% confidence intervals, from the discrete and continuous components of the model. C) Violin plots showing the expression of genes in the "chaperonin mediate protein folding" module. D) Model coefficient estimates for the genes in the chaperonin mediate protein folding module from GSEA, with 95% confidence intervals, from the discrete and continuous components of the model. 122

A.14 The six stimulated MAIT cells that did not exhibit an expression profile indicative of activation are shown in comparison to A) other stimulated MAITs and B) unstimulated MAITs. Differentially expressed genes between these six cells and the stimulated but activated and non-stimulated cells are shown, identified using MAST at a q-value of 15% and fold change threshold of at least 2. Panel C) shows PCA of the MAITs based on the differentially expressed genes. 13 selected genes with largest loadings discriminating between the three classes of cells are shown. 123

A.15 PCA of the model residuals of LPS stimulated cells using the genes in the core antiviral module identified in Shalek et al. [2014]. The two "outlier" cells evident at the 1h timepoint correspond to the "early marcher" precocious cells described previously. These results show that these cells exhibit coordinated co-expression of genes in the core antiviral signature at the single-cell level. 124

A.16 Co-expression plot for PAM (synthetic mimic of bacterial lipopeptides) stimulated cells of cells in the mDC data. Panel A in each figure shows principal component analysis (PCA) of the model residuals using the top 100 differentially expressed genes. Cells are faceted by time, which is correlated with the first principal component. Panel B shows heatmaps of the pairwise correlations between genes in the model residuals across cells at each timepoint. The order of genes in the heatmaps is based on clustering at the 6h timepoint. . 125

A.17 Co-expression plot for PIC (viral-like double-stranded RNA) stimulated cells of cells in the mDC data. Panel A in each figure shows principal component analysis (PCA) of the model residuals using the top 100 differentially expressed genes. Cells are faceted by time, which is correlated with the first principal component. Panel B shows heatmaps of the pairwise correlations between genes in the model residuals across cells at each timepoint. The order of genes in the heatmaps is based on clustering at the 6h timepoint. 126

LIST OF TABLES

Table Number Page

2.1 $\overline{\text{WSS}} - \min_{t_z, t_\zeta} \overline{\text{WSS}}$ values across data sets and filtering parameters. For each data set, the minimum $\overline{\text{WSS}}$ is subtracted so that cells that achieve that value contain zeros. 30

2.2 Effect of filtering, beyond the effect of excluding null wells, on control genes CD4 and CD8. Data set A is expected to be positive for CD8, negative for CD4. Data sets B and C are expected to be negative for CD8, positive for CD4. The percentage of cells filtered, the frequencies π before and after filtering and percentage change in π of these genes is printed. 31

4.1 Dissimilarities $\left(\frac{\text{Hamming Distance}}{\text{Number of edges}}\right)$ between networks of size 35 estimated through various methods. The Gaussian(10) model is a Gaussian model estimated on 10-cell replicates, while the Gaussian(raw) data is estimated on single cells without centering the data. The Hurdle and logistic models are described in the text. 85

A.1 Hyper parameter settings for CDR generation model. 109

A.2 Standard deviations of module scores for stimulated and non-stimulated MAIT cells 119

ABBREVIATIONS

CDR cellular detection rate.

CPM counts per million.

GLM generalized linear model.

GSEA gene set enrichment analysis.

MLE maximum likelihood estimate.

rt-qPCR reverse-transcriptase qPCR.

scRNAseq single cell RNA sequencing.

TPM transcripts per million.

UTR untranslated regions.

vGLM vector generalized linear model.

ACKNOWLEDGMENTS

I thank my PhD advisers and committee members for their mentorship, guidance and generosity with their time. The training you have given me will be an asset for the rest of my life. I am especially thankful to Raphael for taking the chance to fund me as a wait-listed student. None of this would have been possible without your support.

I learned so very much from the members of RGLAB these past five years. Thank you Greg Finak and Masanao for your help with MAST, and being a willing ear and critic for my harebrained programming and statistical schemes. Thanks to Renan, Leo, Phu and Mike for listening to me complain about `R`. Thanks to Greg Imholte, Sam and Amit for the engaging questions and conversations.

Erin, thank you for your patience and love during the long winters of my PhD. With you, the nights are shorter (possibly because I kept you awake until 2AM working in the other room). And Tatoosh, you have been a good cat with small, but non-zero probability.

I am particularly grateful for the support I received throughout the course of my PhD as a research assistant through R01 EB008400 from the National Institute of Biomedical Imaging and Bioengineering, US National Institutes of Health.

DEDICATION

To Mom and Dad–my first teachers–and all those who have followed.

Chapter 1

INTRODUCTION

1.1 Biological motivation for single cell gene expression

The ability to query the mRNA profile of an organism is one the experimental pillars of modern biology. Two complementary trends broadly characterize its evolution: its dimensionality has broadened, while the areas of application have deepened. Tiling arrays, and then sequencing, unlocked reference, and then *de novo* assembled transcriptomes and evolved from assuming a fixed transcript structure to allowing for exonic polymorphism and splicing variation. At the same time, experiments more richly capture additional organisms, tissue types and phenotypes.

The expansion of gene expression experiments in both cardinal directions has led to a more complete, less biased picture of the average state of an organism across time and space. A statistician is tempted to cast this as follows. Supposing that $\mathbf{Y}$ is a vector-valued gene expression measurement, and X is a covariate, then much of the progress in gene expression techniques has focused inference on

$$\mathrm{E}\big(\mathbf{Y}|X\big), \tag{1.1}$$

and expanding the cardinality of $\mathbf{Y}$ and X, as well as the precision of the estimate of $\mathrm{E}\big(\mathbf{Y}|X\big)$.

Compared to the bulk status quo, single cell gene expression experiments forge new frontier. Though measuring expression in single cells does both broaden and deepen gene expression application, it also adds a new rank: inference on heterogeneity. Cell-to-cell variation is both feature and goal in single cell gene expression. While measuring and accommodating cell-to-cell variation has long been a defining feature in flow cytometry, it is

unfamiliar ground in gene expression methods.

The study of variation can be expected to lead to several different forms of insight. The statistician imagines the study of cell-to-cell variation means inference on

$$\mathrm{Var}\big(Y_j|X\big), \tag{1.2}$$

for example, as per variance-components models. This would explain how unexplained variation in expression changes due to covariates. Or one might posit a latent variable Z, which identifies a sub-population of cells, or their spatial or temporal features. Then we desire inference on

$$\mathrm{E}\big(\mathbf{Y}|Z\big) \tag{1.3}$$

and $P(Z|\mathbf{Y})$, for example, as in factor analytic or finite mixture models. This would help identify heretofore unknown structure, and estimate how it affects gene expression. Lastly, we might desire inference on

$$\mathrm{Cov}\big(\mathbf{Y}\big), \tag{1.4}$$

which explains how the expression of one gene tends to covary with another gene, as for example is studied in graphical models. This might help us understand how genes can regulate each other, and ultimately answer causal questions.

1.2 Distributional properties of single cell gene expression

A characteristic of single cell expression is bimodal expression, in which two clusters of expression are present. The overt non-normality and bimodality of the data from these experiments suggest that modeling the bimodality might yield more efficient and complete answers to the previous statistical questions. Figure 1.1 demonstrates archetypal expression from reverse-transcriptase qPCR (rt-qPCR) and sequencing assays.

In any given transcript, the null cluster corresponds to cells without detectable expression (hence a non-zero measurement reflects error due to, for example, cross-hybridization or alignment ambiguity) while the signal cluster contains cells with a positive, detectable level

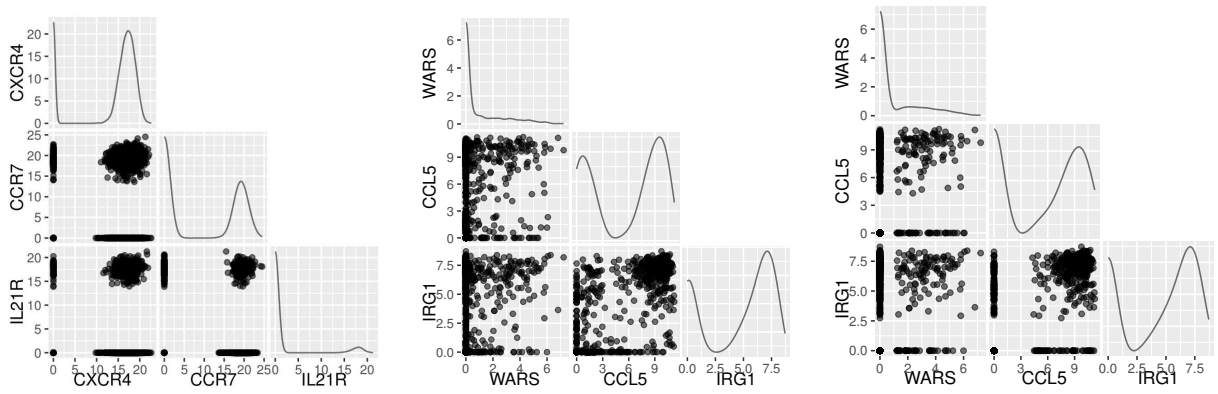

Figure 1.1: Scatter plots and kernel density estimates from (a) Fluidigm single cell qPCR data (b) single cell RNA sequencing and (c) single cell RNA sequencing after thresholding.

of expression. Here

$$V_{ij} = \begin{cases} 1 \text{ if } Y_{ij} \in \text{signal cluster}, \\ 0 \text{ else}, \end{cases}$$

will be used throughout this manuscript as a cluster indicator, where Y_{ij} gives expression of replicate i in gene j. This indicator may not always directly observed, and in some cases plug-in estimates of V_{ij} will be used. The manner in which bimodality and zero-inflation manifests depends on the technology used to measure expression. Micro-fluidic arrays (e.g. the Fluidigm Biomark) allow the use of targeted rt-qPCR to select for specific transcripts and efficiently pre-amplify them by combining rt-qPCR primers with single cell lysate in nanowells. The PCR thermocycler returns the cycle at which a fluorescence threshold is crossed; if after a fixed number of cycles the threshold is not crossed, then the value is reported as "undetected" which is mapped to a inverse cycle threshold of 0. Further details are available in chapter 2. This implies that continuous values, which are proportional to the log-expression, are inflated with zeros.

In contrast, in hybridization and sequencing-based assays, after log transforming normal-

ized hybridization events or aligned sequences, a null cluster may consist of small, continuous values. Often, but not always, two modes are present. Chapter 3 proposes methods for thresholding data to assign observations to null and signal clusters.

1.3 Previous methods for bulk and single cell expression

A cornucopia of methods have been proposed to analyze bulk gene expression. Although distributional differences between bulk and single cell data render some methods inappropriate for single cell data testing, many features translate usefully. Here I discuss several of the most wide-spread methods for testing for change in the conditional expression (1.1), variance component models (1.2), clustering and latent variable models (1.3) and covariance or graphical modeling (1.4). I also review proposals specifically tailored for single cell gene expression where available.

1.3.1 Adapting regression and ANOVA to test for bulk differential expression

The earliest specific proposals for bulk gene expression assumed that a vanilla linear model held for data represented as a matrix $\mathbf{y}$ with independent rows and columns which might represent normalized fluorescent intensities over an array of probes. The rows are "replicates" $i = 1, \ldots, n$ (which will be tenuously defined to include both *technical replicates* that are repeated measures of the same biological sample and *biological replicates* in which a new sample is derived) and the columns are genes or features $j = 1, \ldots, g$. Each replicate belongs to a class $c(i)$ denoting treatment or some other categorical covariate. A mundane, but nonetheless successful linear model for such data posits that each gene has a fixed intercept α_j and condition effect $\beta_{c(i)j}$ (which might be equal to zero) so that

$$Y_{ij} = \alpha_j + \beta_{c(i)j} + \epsilon_{ij}, \tag{1.5}$$

where $\mathrm{E}(\epsilon_{ij}) = 0$, i.e. the linear model is well-specified and $\mathrm{Cov}(\epsilon_i)$ is diagonal across genes j, and constant across replicates i. The expression vectors of each sample $\mathbf{Y}_i$ are typically

assumed to be independent; experiments with technical replicates nested within biological replicates would violate this assumption.

A pressing issue is that the number of replicates n is typically not large and the variance of each gene $\sigma_j^2 = \text{Var}(\epsilon_j)$ is typically similar, leading Smyth [2004] to propose `Limma`, a hierarchical model in which the σ_j^2 are assumed to come from a super-population. For example, the conjugate prior

$$\sigma_j^2 \sim \text{Inverse-Gamma}(a, b)$$

leads to a simple empirical Bayesian procedure in which the gene-specific variances can be integrated out and a and b chosen to maximize this marginal log-likelihood. Earlier, Tusher et al. [2001] had proposed the related idea of shrinking the standard deviation σ_j towards a global value σ_0, which can be defined in a variety of ways. In both cases, gene-level variances are moderated towards the *average* gene variance, with the amount of moderation determined adaptively.

In `Limma`, owing to the pseudo-observations available in the hierarchical model, *moderated* t-tests in this regime possess additional degrees of freedom. These additional degrees of freedom allow less stringent rejection regions for the same type I error rate α, since the shrinkage reduces the variance of the σ_j^2 estimates. A modification of the Smyth [2004] procedure is proposed in Chapter 3 of this dissertation and appears to lead to more stable inference.

Once sequencing-based **RNAseq** overtook array-based methods circa 2009, assuming homoscedasticity was less tenable, as integer-valued counts are naturally returned in sequencing. Quasipoisson and negative binomial generalized linear models replaced equation (1.5), again with some sort of de facto hierarchical model to share information on the dispersion parameter. In `edgeR`, Robinson et al. [2010] proposed a fairly direct adaptation of the `Limma` model. Anders and Huber [2010] proposed `Deseq`, which uses fitted values from local regression of gene $\times$ replicate estimates of the mean-variance relationship as plug-in estimates. Despite `Deseq` not explicitly treating variability about the fitted mean-variance

relationship and the distinct formulations compared to `edgeR`, both methods typically lead to similar inferences.

Later, it was observed that the linear link could still function (especially since near ubiquitous pre-modeling normalization destroys the integral nature of the data) as long as appropriate case-weights were available for weighted least squares [Law et al., 2014]. Pseudo-replication across genes allows a non-parametric estimate of the mean-variance relationship to derive these case-weights, and coherent use of similar moderated estimates as in `Limma`.

Auer and Doerge [2010], and Conesa et al. [2016] offer a comprehensive reviews of differential expression methods for RNAseq.

RNAseq challenges in library size and alignment

RNAseq presented other challenges besides the specification of the regression. Many protocols fragment each source molecule of mRNA (which consist canonically of one or more spliced exons and polyadenylated tail) into several pieces, zero or more of which will be amplified, sequenced and aligned. This one-to-many map suggests that at a minimum one might want to adjust for the transcript length (hence the number of fragments that might be produced) as a normalizing factor, or *offset* in a generalized linear model (GLM). On the other hand, an increase in the *sequencing depth* for a library (a blocking variable in which samples are often nested) produces an arbitrary increase in the fragments detected. This has been resolved through simple scalings such as **counts per million (CPM)**, which just divides each replicate by the total number of reads aligned in that replicate, or through more ornate normalizations.

Ambiguity in alignment of fragments to source transcripts provides another challenge. Alignment algorithms deterministically report one (or more) candidate alignments that locally optimize some alignment criteria. The candidates are later reduced to gene-level estimates. In some cases, the ambiguity is substantial, and ignoring it can lead to bias and lack of precision [Li and Dewey, 2011].

There is also a fundamental imprecision in the notion that each gene generates a single,

unique transcript. Some genes produce several isoforms from the same locus, but differ in the transcription start site, splicing, number of exons or untranslated regions (UTR). Trapnell et al. [2013] address complications and artifacts arising from differential isoform usage between samples, which can affect the effective length (and mappability) of the same locus in different samples.

In contrast to the vigorous methodological development for arrays and RNAseq, most early, and many current users of single cell differential expression have applied Normal-theory methods or ad-hoc adjustments for zero-inflation or bimodality. Section 2.1.1 reviews some early methods, many involving winzorization. Specialized proposals have included `SCDE` [Kharchenko et al., 2014] which uses a two-component mixture of Poisson and negative binomial distributions, `Monocle` [Trapnell et al., 2014] which straightforwardly applies Tobit regression, and Shalek et al. [2014] who adapt a Hurdle model.

1.3.2 Latent variation and clustering

In bulk arrays, many studies of latent variation have emphasized the detection of batch effects, owing to the pervasive (and often unavoidable) lack of randomization of treatment variables. **Surrogate Variable Analysis** [Leek, 2014] and **Remove Unwanted Variation** [Risso et al., 2014] have used principal component-like or factor analytic models to remove batch effects.

As it is hoped that single cell expression will help define and discover new subpopulations of cells, clustering and latent variation methods for single cell expression have received more rapid development. These methods have run the gamet from parametric, model-based to ad-hoc strategies for post processing. Pierson and Yau [2015] offer perhaps the most coherent and principled approach for accommodating zero-inflation. They propose a censoring variable V_{ij} that depends on a potentially unobserved level of background expression X_{ij}. Letting Y_{ij}

denote the observed level of expression in cell i, gene j then

$$Y_{ij} = \begin{cases} X_{ij} & V_{ij} = 1, \\ 0 & V_{ij} = 0, \end{cases} \tag{1.6}$$

where $P(v_{ij}|x_{ij}) = e^{-\lambda x_{ij}^2}$, and $\mathbf{X}_i$ follows a Normal distribution with low-rank (factor-analytic) covariance. Thus latent Gaussian variables are stochastically censored for smaller values of X_{ij}. In contra, Buettner et al. [2014] proposed a factor analytic Tobit model for dimensionality reduction of zero-inflated data in which the censoring occurs deterministically for X_{ij} sufficiently small. A non-zero inflated version of this model was later applied in Buettner et al. [2015] to estimate factors on subsets of genes with known cell cycle annotation to recover factors predictive of cell cycle.

Others have focused efforts on discovering temporal or spatial relationships between cells. `Monocle` [Trapnell et al., 2014] imposes a spanning tree on single cell RNA sequencing (scRNAseq) data after initial dimensionality reduction through independent components analysis. Once rooted, this tree yields "pseudotime" through the topological ordering induced by the tree. `Seurat` [Satija et al., 2015] takes a semi-supervised approach to inferring spatial information from disassociated single cells. Here, bulk, in-situ RNAseq experiments yielded an atlas of spatial expression in a set of landmark genes, providing the basis to train linear discriminants. These discriminants were used to predict the originating spatial location of the single cells using their gene expression measurements.

1.3.3 Variance components

Mixed models have found some use at modeling subject-to-subject or technical variability in bulk expression experiments, but unlike the case of unmodeled dispersion parameters, there have been few attempts to share information between genes and moderate estimates. van de Wiel et al. [2014] provide perhaps the most complete approach, fitting a hierarchical Bayesian model gene-wise to estimate the posterior likelihoods for dispersion and variance parameters. The marginal likelihood is generally not available in closed form, but an approximation is

available through integrated nested Laplace approximation, thus an EM-type algorithm is available to maximize the marginal posterior likelihood iteratively.

Few models have been proposed or applied specifically for single cell gene expression, although the above approach applies fairly directly, once a framework for zero-inflation is selected. Single cell gene expression designs often do have natural variance components, when multiple individuals, or subsets of cells from the same individual are repeatedly sampled, so development of moderated variance component models would be useful.

Effort on variance component models for single cell expression have focused on injudicious additive partitioning of observed gene-level variances into "technical" versus "biological" components. Brennecke et al. [2013] propose using mRNAs that have been spiked in at known concentrations to bound the technical variability (the variance that be present in hypothetical repeated sampling of an individual cell.) Genes with sample coefficients of variability that exceed thresholds found through a parametric fit of the mean-variance function in the spike-in genes are declared to have significant biological variability, and are perhaps exclusively used in downstream analysis.

This idea received further treatment in Vallejos et al. [2015] who proposed a hierarchical model to explain extra-Poisson variation in sequencing counts through a GLM including random cell-level and gene-level intercepts, the former rendered estimable through spike-in genes. Since single cell data is characterized by replicates of *single cells*, the benefit of decomposing technical variability and biological single cell variability is unclear, apart from forming the basis of a feature selection heuristic. It could be used in efforts to develop biological protocols that minimize technical variability. Paradoxically, the spike-in RNAs that render the technical variability estimable may actually increase it, as they complicate the protocol.

1.3.4 Covariation

Attempts at studying (1.4) in bulk data have used both shrinkage and parametric restrictions focused on model selection. Schäfer and Strimmer [2005] propose a shrinkage procedure for

general covariance matrices, which can be of intrinsic interest for hierarchical clustering, for instance.

However, when $\mathbf{Y}$ is multivariate Gaussian with a covariance matrix Σ, its precision matrix Σ^{-1} is perhaps of even greater interest, since zeros in Σ^{-1} correspond to conditional independences between genes. These conditional independences can be used to encode a graph, in which nodes are genes and edges are identified through a symmetric adjacency matrix $A = (a_{ij})$ with $a_{ij} = 1$ if $[\Sigma^{-1}]_{ij} \neq 0$ and 0 otherwise. These graphs provide a concise visual summary of the relationship between genes.

Several methods have been proposed to estimate the pattern of zeros in Σ^{-1}. The graphical Gaussian lasso [Friedman et al., 2008] proposes an ℓ_1 penalization of the joint Gaussian log-likelihood, while neighborhood selection [Meinshausen and Bühlmann, 2006] uses an ℓ_1 penalized node-conditional Gaussian log-likelihood. Other innovations and variations of these procedures are discussed in chapter 4.

In single cell gene expression, few investigators have considered measures of statistical independence or covariation. `Seurat` [Satija et al., 2015] exploits the correlation structure of expression for single imputation of a subset of "landmark" genes through an ℓ_1-penalized linear regression. The authors argue that the imputed estimate has lower variance and more stability than the observed values. Beyond this, I am unaware of other attempts to estimate or utilize dependence structures in single cell gene expression.

1.4 Statistical frameworks

Extant statistical methods provide ways to tackle many of the questions discussed above. Here some of these methods are discussed in greater depth.

1.4.1 Censoring and sampling models

Two common approaches exist classically to accommodate zero inflation: censoring models and mixture models. In a censoring model, a latent, univariate variable u follows a

distribution $P(u)$, and we observe

$$Y = \begin{cases} g(U) \text{ if } h(U) = 1, \\ 0 \text{ else,} \end{cases} \tag{1.7}$$

where $g(U)$ is generally a simple transformation of U, e.g., the identity function and $h(U)$ is an indicator function. The **Tobit** model lets $P(u)$ be a univariate Normal distribution, $g(U) = U$, and $h(U) = 1_{U>a}$ be an indicator that U exceeds some threshold. Thus U is censored below some threshold and only a "0" is recorded.

In a mixture model, zeros arise from a function that does not deterministically depend on U. Upon generalizing to a pair of variables (U, V), formulation (1.7) encompasses many of these models as well. Zero-inflated discrete models, such as the zero-mixed Poisson, can be written as follows. Let

$$U \sim \text{Poisson}(\lambda),$$
$$V \sim \text{Bernoulli(p)}$$

be independent random variables, $g(U, V) = U$ and $h(U, V) = V$. Zeros arise both naturally through Poisson variation, as well as from U. When U follows the *zero-truncated* Poisson distribution supported on positive integers, $P(u) = \frac{\lambda^u e^{-\lambda}}{u!\left(1-e^{-\lambda}\right)}$, *Hurdle models* are generated. Now, zeros only arise through the Bernoulli process on V. A non-zero is recorded only if the Bernoulli hurdle is passed.

This model is adopted in chapter 2 and 3 of this dissertation by setting $U \sim \text{Normal}(\mu, \sigma^2)$ and $V \sim \text{Bernoulli}(p)$, $g(U, V) = U$, $h(U, V) = V$, and is referred to as a (Normal) Hurdle model. It is extended to multivariate, zero-inflated variables in Chapter 4. A convenient feature of this model is that V is identified by the event $U = 0$, since that event occurs with measure zero otherwise under the Normal distribution. (This convenience does not occur generally when U follows some discrete distribution.)

An interesting generalization of the Normal Hurdle is the Heckit, or type II Tobit model. In these, (U, V) are bi-variate Normal and $g(U, V) = U$, $h(U, V) = 1_{V>a}$ so that (U, V) is

censored in one coordinate given a threshold is crossed in the other. Identifiability requires constraints on the mean and variance of the bi-variate Normal distribution. Amemiya [1984] and Toomet and Henningsen [2008] review these generalized Tobit models extensively.

1.4.2 Vector GLMs

It is natural to extend the Normal Hurdle to a regression setting by letting the parameters be functions of covariates:

$$U_i \sim \text{Normal}(\mathbf{x}_i\beta, \sigma^2), \tag{1.8}$$

$$V_i \sim \text{Bernoulli}(\text{expit}(\mathbf{x}_i\beta')), \tag{1.9}$$

where $\mathbf{x}_i, \beta, \beta' \in \mathbb{R}^p$ are vectors of covariates, and parameters respectively, and $\text{expit}(z) = e^z/(1+e^z)$. The density of $Y = UV$ is

$$f(y;\beta,\beta') = \exp\left\{1_{y\neq 0}\left[\mathbf{x}\beta' - \frac{(y-\mathbf{x}\beta)^2}{2\sigma^2} - 1/2\log\left(2\pi\sigma^2\right)\right] - \log\left(1+e^{\mathbf{x}\beta'}\right)\right\} \tag{1.10}$$

$$= \exp\left\{1_{y\neq 0}\left[\eta' - \frac{(y-\eta)^2}{2\sigma^2} - 1/2\log\left(2\pi\sigma^2\right)\right] - \log\left(1+e^{\eta'}\right)\right\} \tag{1.11}$$

This is an example of a **vector generalized linear model (vGLM)** [Yee, 2015] in which a base distribution depending on several parameters is extended by setting them to be linear functions (predictors) $\eta = \mathbf{x}\beta, \eta' = \mathbf{x}\beta'$ of covariates. The vGLM specification of expanding model parameters in terms of linear predictors η and η' is deliberately vague compared to the relative formalism in GLM, which links the linear predictor to the first moment of an exponential family base distribution. However, the vGLM specification allows great flexibility.

Hurdle model properties

The log-likelihood $\log f(\mathbf{y};\beta,\beta')$ of model (1.10) of a sequence of data $\mathbf{y} = y_1, \ldots, y_n$ separates as a sum of the Gaussian log-likelihood and the logistic log-likelihood. Consequently, it has

score

$$\frac{\partial \log f}{\partial \beta} = \sum_{i:y_i \neq 0} \frac{1}{\sigma^2} \mathbf{x}_i^T (y_i - \mathbf{x}_i^T \beta),$$
$$\frac{\partial \log f}{\partial \sigma} = \sum_{i:y_i \neq 0} \frac{(y_i - \mathbf{x}_i^T \beta)^2}{\sigma^3} - \frac{1}{\sigma},$$
$$\frac{\partial \log f}{\partial \beta'} = \sum_i \mathbf{x}_i \left(1_{y_i \neq 0} - \text{expit}(\mathbf{x}_i^T \beta')\right).$$

The Fisher information is block-diagonal with (β, σ) and β' blocks consisting of the typical Fisher information of the Gaussian linear and logistic models, respectively.

1.4.3 Graphical models

Graphical models provide a way to parametrize the conditional independences present in some high-dimensional distribution through a graph. They have been used extensively as a model for gene expression. Further background will be delayed until Chapter 4.

Chapter 2

DATA EXPLORATION, QUALITY CONTROL AND TESTING IN SINGLE-CELL QPCR-BASED GENE EXPRESSION EXPERIMENTS

2.1 Single cell qPCR methods

The development of fluorescence-based flow cytometry (FCM) revolutionized single-cell analysis. Although populations of cells sorted by flow cytometry using surface markers may appear monolithic, mRNA expression of specific genes within these cells can be heterogeneous [Dalerba et al., 2011] and could further discriminate cell subsets. On the other hand, classical gene expression experiments (microarrays, RNA-seq, qPCR) richly characterize a cellular population, but at the cost of reporting a summation of expression from many individual cells. Recent advances in microfluidic technology now permit performing thousands of PCRs in a single device, enabling gene expression measurements at the single-cell level across hundreds of cells and genes [Kalisky and Quake, 2011]. This provides a technology that probes the stochastic nature of biochemical processes, resulting in relatively large cell-to-cell expression variability.

This heterogeneity may carry important information: thus single cell expression data should not be analyzed in the same fashion as population-level data. At the scale of a single cell, biological variability (the object of interest) and technical variability (a nuisance factor) are often of the same magnitude, making it difficult to distinguish between the two. No expression (*i.e.* the gene is off) may be detected in individual cells due to real biological effects, resulting in zero-inflation of otherwise continuous measures. These features of single-cell data require special attention during analysis.

Here we focus on the reverse-transcriptase qPCR (rt-qPCR)-based Fluidigm (San Fran-

cisco, CA) single-cell gene expression assay, which provides simultaneous measurements of up to 96 genes on mRNA sources as minute as a single cell. In traditional rt-qPCR, despite careful measurement of input cDNA concentrations, differences in starting material below the limit of detection require correction for reliable results Vandesompele et al. [2002]. Subtraction of internal control genes, or averages thereof is typically used (*e.g.*, the Δ-Ct method), and results are often reported in fold increase per cell [Schmittgen and Livak, 2008]. In array-based gene expression, differences in hybridization and washing of non-specific DNA between chips require additional correction.

Such normalization schemes are not directly applicable in single-cell gene expression experiments, nor is it obvious that they are needed. For single cells, the individual cell is the atomic unit of normalization and the amount of starting material naturally measured in number of cells per reaction. Even if one attempted direct application of traditional normalization approaches, the dichotomous nature of single-cell expression hinders their use.

Nonetheless, it is important to test for and address any technical biases. We present a filtering approach for removing outlying measurements at the single-cell level that accounts for the dichotomous nature of the data. Using concordance measures derived from three data sets where gene expression was measured at the single-cell and hundred-cell levels, we show that classical rt-qPCR type normalization is not necessary with single-cell multiplexed PCR data and that our filtering step removes technical artifacts that most severely impact quantitation.

A typical goal of gene expression experiments is to search for differential expression across groups. The zero-inflation of expression in Fluidigm introduces problems for testing differential representation of cell subsets characterized by expression patterns, as well. Traditional tests of differential expression such as the t-test or other approaches based on normality are likely inappropriate for zero-inflated data [Smyth, 2004, Gottardo et al., 2006]. Approaches to this problem have varied. Powell et al. [2012] used a winsorized z-transformation of the expression values, then treated them as continuous. Glotzbach et al. [2011] used the non-parametric, Kolmorgov-Smirnov test for differences in distribution to find differentially

expressed genes, after winsorizing. Flatz et al. [2011] dichotomized the expression and worked with the binary trait. Of these authors, only Flatz et al. [2011] and Glotzbach et al. [2011] made use of formal tests of differential expression. However, as we will see later, both the continuous and discrete parts of the measurements are informative for differential expression and should be used. A parametric test allows directions of difference to be assessed.

Here we propose a discrete/continuous model for single-cell expression data based on a mixture of a point mass at zero and a log-normal distribution. Using this model, we derive a likelihood ratio test that can simultaneously test for changes in mean expression (conditional on the gene being expressed) and in the percentage of expressed cells.

2.1.1 Data sets and notations

We use three Fluidigm single-cell gene expression data sets described below. We offer a brief overview of the assay technology used for our data. Desired cells (*e.g.* antigen-specific CD8+T cells) are selected and lysed, and a cDNA library is generated through rt-qPCR. A short (c. 15 cycle), multiplexed pre-amplification selects and enriches for the desired genes. These products are loaded onto the Fluidigm chip and gene-specific primers are added for single-cell gene expression quantitation. For the data presented here we used a 96×96 format plate, *i.e.* 96 genes across 96 cells. The design of the chip generates each combination of the 96 genes and 96 enriched cDNA libraries producing 9216 separate PCR reactions. After each cycle, the fluorescence is read. The cycle (or interpolated fraction thereof) at which the fluorescence crosses a pre-determined threshold is recorded, defined as the "*ct*" value. For all data sets considered here, primers were chosen to have $> 90\%$ amplification efficiency.

Data set A: Twenty-eight 96×96 format plates of CMV- or HIV-specific CD8+ single cell T cells were isolated from 16 individuals. The donors' cells were stimulated with one of four tetramers. Cells were sorted immediately after tetramer incubation ("unstimulated") or after 3 hours of exposure ("stimulated"). Approximately 90 individual cells were measured for each patient-stimulation combination ("unit").

Data set B: Ten subjects were considered, and approximately 180 activated CD4+ memory

T cells were sorted per subject, with each subject crossed between two arrays.

Data set C: Two subjects were considered. Fluorescent staining of CD4+ T cells allowed cytometric sorting into CD154+/- sub-populations. Approximately 40 cells were sorted per sub-population per subject across three arrays.

Additionally, for each individual and treatment within each data set, aggregates of 100 cells (*i.e.*, 100 cells per well on the array) were isolated and assayed by Fluidigm technology. The expression measured in these 100-cell aggregates, after dividing by 100, provides a "biological" average of expression per–cell, and can be compared to an *in silico* average of the single-cell measurements. The *concordance* between these two averages serves as a measure of experimental fidelity [Lin, 1989].

Notations: The standard assumptions of qPCR-based assays apply to the Fluidigm technology, namely that the cycle threshold (ct) is inversely proportional to the log of fluorescence. The fluorescence is directly proportional to the starting concentration of mRNA [Higuchi et al., 1992, Karlen et al., 2007]. The Fluidigm instrument returns the cycle threshold (ct), however, we find it more useful to work with the complement of ct, which we define as the *expression threshold* (et)

$$et = c_{\max} - ct,$$

where $c_{\max}$ is the maximum number of cycles used, 40 in our case. Assuming all reactions are in the exponential amplification phase, this quantity should be directly proportional to the log-abundance of mRNA, plus an intercept term corresponding to the number of cycles it takes for the minimally-detectable quantity of mRNA to cross threshold. If the fluorescence does not cross the threshold after 40 cycles, then the Fluidigm instrument records a value of N/A, and we say that the gene is *not detected*. As we will see in the results section, detected genes typically have a value of ct much less than $c_{\max}$ suggesting that undetected genes might be regarded as unexpressed genes. This assumption is supported by the idea that transcription of mRNA is thought to occur in bursts of activity [Levsky et al., 2002, Kaufmann and van Oudenaarden, 2007], followed by quiescence. Other authors have noted this feature in single cell expression studies as well [Glotzbach et al., 2011]. When looking at

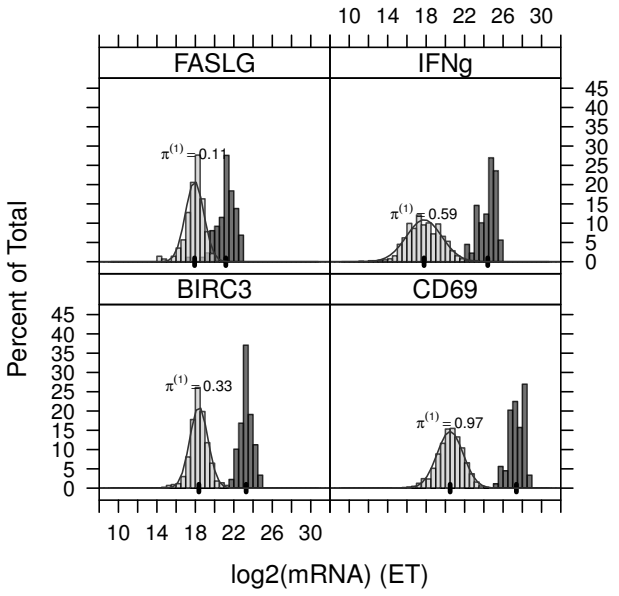

Figure 2.1: Histogram and theoretical (normal) distribution of $(et_{ij}|v_{ij} = 1)$ for single cell (left, light gray) and hundred cell experiments (right, dark gray). Genes FASLG, IFN-γ, BIRC3 and CD69 are depicted. The frequency expression of each gene in the single cell experiments $\pi^{(1)}$ is printed above each histogram. The mean of the hundred cell and single cell experiments is indicated by a thick black line along the x-axis.

the concordance of the single-cell and hundred-cell experiments, this assumption is reasonable and leads to better concordance than omitting the N/A values. As a consequence, we treat the undetected genes as unexpressed genes, and we set the corresponding *et* value to $-\infty$, so that the mRNA abundance is zero (*i.e.* $2^{et} = 0$).

For a fixed sample or experimental unit, let us denote by et_{ij} the expression threshold of *well* i and *gene* j, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. This results in a matrix of $\log_2$ based expression values, $\mathbf{ET} = (et_{ij})$, just as in array-based gene expression. Similarly, we will denote by $\mathbf{Y} = (y_{ij})$ the matrix of untransformed expression values where $y_{ij} = 2^{et_{ij}}$. Usually a well contains one cell but the Fluidigm technology can be used with multiple cells per well to quantify the gene expression of a mixture of cells. As a consequence, we prefer to use the term "well" instead of "cell". In the three data sets used here, wells will contain either one or hundred cells. Finally, several biological units are typically measured in an experiment, and in this case we will use an extra index k to refer to the biological units.

2.2 A two-part Hurdle model

As described previously, for a given cell, a gene can be defined as *on* (*i.e.* a positive *et* value is recorded) or as *off* (*i.e.* the gene is undetected and $y_{ij} = 0$). To simplify our model, we will denote by $v_{ij} = \mathbf{1}[y_{ij} > 0]$ the indicator variable equal to one if the gene j is expressed in well i and zero otherwise. Following classical statistical conventions, we use upper cases to denote the random variables, and lower cases to denote the values taken by these random variables. Using these notations, we introduce the following model of single-cell expression

$$(Y_{ij}|V_{ij} = 1) \sim \text{logNormal}(\mu_j, \sigma_j^2), \tag{2.1}$$

$$(Y_{ij}|V_{ij} = 0) \sim \delta_0, \tag{2.2}$$

$$V_{ij} \sim \text{Be}(\pi_j), \tag{2.3}$$

where δ_0 denotes a point mass at zero, μ_j and σ_j^2 are the $\log_2$-based mean and variance expression level parameters conditional on the gene being expressed (*i.e.* $V_{ij} = 1$), and π_j is the frequency of expression of gene j across all cells. In the data sets considered here, the

frequency of expression greatly varies across genes from 0 to .99 with a median value of π_j around .1. Note that assuming a log-Normal model for $(Y_{ij}|V_{ij} = 1)$ is equivalent to modeling $(\mathbf{ET}_{ij}|V_{ij} = 1)$ as normally distributed. The empirical distribution of the data (Figure 2.1 and 2.2 motivates our selection of a log-normal distribution and follows observations of previous authors [Bengtsson et al., 2005].

Thus in a particular gene, three parameters characterize the expression distribution: μ_j, σ_j, the mean and standard deviation of the $et_{ij}|V_{ij} = 1$, and π_j, the Bernoulli probability of expression.

2.3 Quality control and filtering

The Fluidigm assay is sensitive, and due to the exponential amplification of starting mRNA, even minute contamination can render a measurement unreliable. Similarly, variation in cell preparation can have significant impact on the resulting experiment and data, such as unintentional empty wells, which would distort estimates of π_j. This suggests identifying, and possibly removing outliers before conducting further analysis. We examine both the discrete component v_{ij} and the continuous component $(et_{ij}|v_{ij} = 1)$ in screening for outliers. We define the robust z-transformed positive expression value as

$$z_{ij} \equiv \frac{et_{ij} - \text{median}_i(et_{ij})}{k \cdot \text{MAD}_i(et_{ij})},$$

where the median and median absolute deviation are calculated, for a given gene, over expressed cells (*i.e.* $v_{ij} = 1$), and $k = 1.48$ is a scaling constant that gives the standard deviation in terms of the MAD for the normal distribution. Next, let $f_i = \text{asin}\sqrt{\bar{v}_{i\cdot}}$ be the Bernoulli variance-stabilizing transformation of the proportion of genes expressed in well i. Then we define the robust z-transformed fraction as

$$\zeta_i \equiv \frac{f_i - \text{median}_i(f_i)}{k \cdot \text{MAD}_i(f_i)},$$

where the median, MAD and k are as defined previously. This leads to the following steps for filtering:

1. Remove *null* cells with no detected genes, *i.e.* $V_{ij} = 0$, for all j.

2. Pick threshold for z filtering (t_z); threshold for ζ filtering (t_ζ).

3. Calculate z_{ij} and ζ_i

4. Remove wells in which genes have $|z| > t_z$ OR $|\zeta| > t_\zeta$.

Step 1 removes wells where no cells were loaded, and thus all measured expression values are null. It is important to perform this step first to prevent break-down in the median and MAD estimates for the ζ's in experiments with many amplification or flow cytometry failures. Finally, step 4 removes unreliable wells that either have an extreme proportion of expression or extreme cell×gene expression values. The thresholds t_z and t_ζ control the tolerance to outliers, so typical advice for outlier thresholding applies. Biological replicates, such as the hundred cell replicates described in Data set and Notations, permit the assessment of intra-class deviance and then the thresholds can be selected to minimize this deviance. Using this approach, we find that picking $t_z = 9, t_\zeta = 9$ works well for the data sets we consider here, see section 2.5.

2.4 Testing for differences between experimental groups

One typical goal of gene expression analysis is to test for difference in expression patterns between experimental units. Here, we focus on testing differential gene expression between two paired-biological units, *e.g.* before and after stimulation. Our framework should be generalizable to other types of situations, see Section 2.6. The classical test for changes in mean for samples with continuous measurements is the t-test. Conversely, if only a change in π were of interest, then a contingency table test (Chi-square, Fisher's Exact or Bernoulli likelihood ratio) is appropriate. However, in our case, we would like to test for a change in μ and π simultaneously, since both could be biologically relevant. Formally, we wish to test

$$H_0 : \pi_0 = \pi_1 \quad \text{and} \quad \mu_0 = \mu_1$$

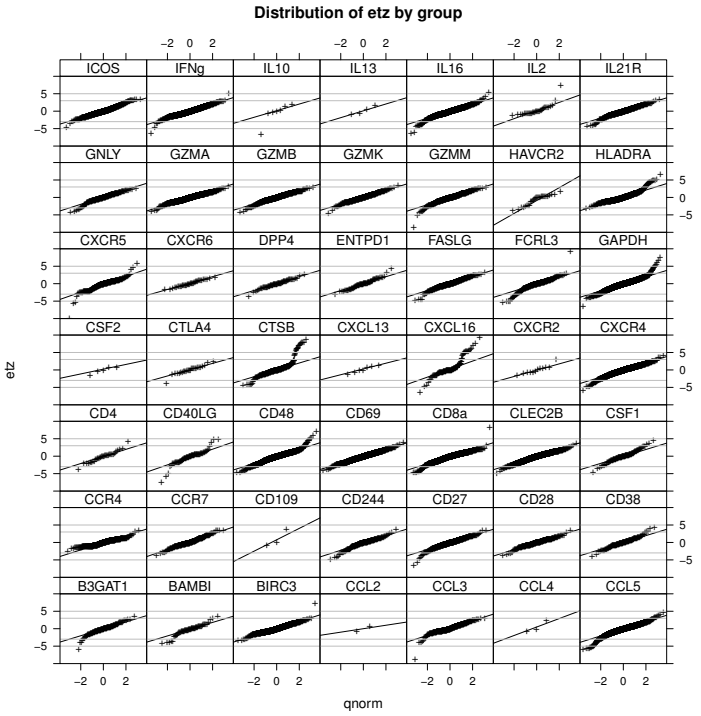

Figure 2.2: Normal quantile-quantile plots of z_{ij} for 49 genes, data set A.

versus the alternative

$$H_a : \pi_0 \neq \pi_1 \quad \text{and} \quad \mu_0 \neq \mu_1.$$

This can be accomplished using a likelihood ratio test that would simultaneously test for differences in means or proportions of expression.

Suppose that I wells are assayed in each unit, though the unbalanced case ($I_0 \neq I_1$) would be treated similarly with obvious changes of notation. Based on (2.1), the likelihood function for one gene across two biological units, omitting the gene index j for clarity, is given by

$$\mathrm{L}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{v}) = \prod_k \pi_k^{n_k}(1-\pi_k)^{I-n_k} \prod_{i \in S_k} g(y_{ik}|\mu_k, \sigma^2), \tag{2.4}$$

where $\mathbf{y}$ and $\mathbf{v}$ are the vectors of observations for the gene across the two groups, $\boldsymbol{\theta} = \{\mu_k, \sigma^2, \pi_k; k = 0, 1\}$ is the vector of unknown parameters, S_k is the set of cells expressing the gene in group k (*i.e.* $S_k = \{i : v_{ik} = 1\}$) , $n_k = \sum_i v_{ik}$ is the number of cells expressing the gene in group k, and g is the density function of the log-normal distribution with parameters μ_k and σ^2. The likelihood ratio test (LRT) statistic $\Lambda(\mathbf{y}, \mathbf{v})$ is then defined as the ratio of the null and alternative likelihoods obtained by replacing the unknown parameters with their null and alternative maximum likelihood estimates.

Derivation of LRT

The likelihood ratio test is defined as

$$\Lambda(\mathbf{y}, \mathbf{v}) = \frac{\sup_{\boldsymbol{\theta} \in H_0} \mathrm{L}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{v})}{\sup_{\boldsymbol{\theta} \in H_A} \mathrm{L}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{v})}, \tag{2.5}$$

where the likelihood is given by equation (2.4). Using the following change of variable, $et_{ik} = \log y_{ik}$, the likelihood function can be written as

$$\mathrm{L}(\boldsymbol{\theta}|\mathbf{et}, \mathbf{v}) = \prod_k \pi_k^{n_k}(1-\pi_k)^{I-n_k} \prod_{i \in S_k} \mathrm{N}(et_{ik}|\mu_k, \sigma^2), \tag{2.6}$$

where $\mathrm{N}(\cdot|\mu, \sigma^2)$ is the density function of a normal distribution with mean μ and variance σ^2. It follows that the likelihood ratio test can be written as

$$
\begin{aligned}
\Lambda(\mathbf{et}, \mathbf{v}) &= \frac{\sup_{\boldsymbol{\theta} \in H_0} \mathrm{L}(\boldsymbol{\theta}|\mathbf{et}, \mathbf{v})}{\sup_{\boldsymbol{\theta} \in H_A} \mathrm{L}(\boldsymbol{\theta}|\mathbf{et}, \mathbf{v})}, \\
&= \frac{\sup_{\{\pi_0, \mu_0, \sigma^2\}} \pi_0^{n_0+n_1}(1-\pi_0)^{2I-n_0-n_1} \prod_k \prod_{i \in S_k} \mathrm{N}(et_{ik}|\mu_0, \sigma^2)}{\sup_{\{\pi_0, \mu_0, \sigma^2, \pi_1, \mu_1\}} \prod_k \pi_k^{n_k}(1-\pi_k)^{I-n_k} \prod_{i \in S_k} \mathrm{N}(et_{ik}|\mu_k, \sigma^2)}, \\
&= \frac{\sup_{\pi_0} \pi_0^{n_0+n_1}(1-\pi_0)^{2I-n_0-n_1}}{\sup_{\{\pi_0, \pi_1\}} \prod_k \pi_k^{n_k}(1-\pi_k)^{I-n_k}} \cdot \frac{\sup_{\{\mu_0, \sigma^2\}} \prod_k \prod_{i \in S_k} \mathrm{N}(et_{ik}|\mu_0, \sigma^2)}{\sup_{\{\mu_0, \mu_1, \sigma^2\}} \prod_k \prod_{i \in S_k} \mathrm{N}(et_{ik}|\mu_k, \sigma^2)}, \\
&= \Lambda_b(\mathbf{v}) \cdot \Lambda_n(\mathbf{et}^+),
\end{aligned}
$$

where Λ_b is a binomial LRT, Λ_n is a normal LRT and $\mathbf{et}^+$ is the set of positive *et* values. Thus our combined LRT can be computed as the product of a binomial and a normal LRT statistic, both of which can easily be derived using classical statistical theory.

An interesting observation is that the likelihood function given by (2.4) is the product of the Bernoulli likelihood for all cells and the log-normal likelihood for the expressed cells. It follows that the log-LRT statistic decomposes as a sum of a Bernoulli log-LRT test statistic and a log-normal log-LRT test statistic, since each component can be maximized independently. It thus combines the two sources of information in a natural way, and this decomposition allows post-hoc assessment of which of the component(s) drive the detected difference by simply comparing the magnitude of the two log-LRTs. In Section 2.5 we will show that our combined LRT test is more powerful than the Bernoulli or log-normal tests alone.

Validity of asymptotic approximation

Applying classical asymptotic results about LRTs [Wilks, 1938], $-2\log\Lambda(\mathbf{y}, \mathbf{v})$ converges to a χ^2 distribution with two degrees of freedom under H_0. Some care is warranted in invoking this asymptotic result, since even for large I, the sample size for the log-normal LRT will be πI. In simulation, in Figures 2.3 and 2.4 we find that the χ^2 convergence is adequate for $\pi I > 8$, even under departures from normality.

Bivariate samples of $(\mathbf{y}, \mathbf{v})$ are simulated from hypothetical genes with $\pi = .02, .04, .08, .16, .32$

and sample size $I = 100$ in which there is no difference means and proportions between classes. The standard deviation is 1.3, which is the median empirical standard deviation across data sets. There is some inflation of the null distribution for smaller effective sample sizes ($\pi I < 8$), chiefly in the form of additional skewness that would result in the size of the test being higher than the nominal significance level. However, one is not forced to rely on the asymptotic distribution of Λ for assessing significance. When the effective sample size of the continuous component, πI is too small, *e.g.* $\pi I < 8$, the null distribution can be approximated using permutations Ge et al. [2003]. This proviso applies similarly for purpose of power calculations hence one may wish to conduct these through simulation.

Departures from normality

Figure 2.4 depicts the same scenario as in Figure 2.3, however now $\log \mathbf{u}$ is simulated from a t-distribution with 4 degrees of freedom, and scaled by 1.3, the standard deviation used in the prior simulation. This allows the assessment of the level of the test when the continuous distribution does not follow a normal distribution. As expected, since the Λ_n depends on the mean of $\mathbf{et}^+$, the central limit theorem results in this mean converging in distribution to a normal distribution, hence the test is somewhat robust to departures from normality. The null distributions are very close to the ones obtained under the normal assumption.

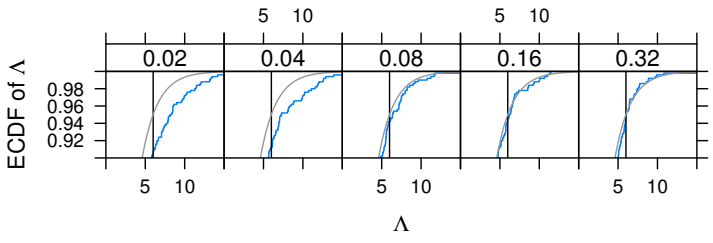

Figure 2.3: The empirical cumulative distribution plot of Λ. The cumulative distribution of χ^2_2 is plotted in gray. 5% significance is indicated by a vertical line. The gene frequency π varies. The sample size $I = 100$.

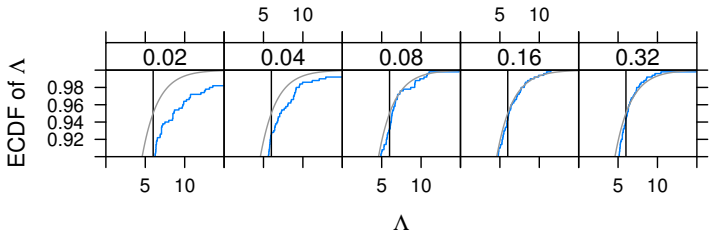

Figure 2.4: The empirical cumulative distribution plot of Λ and of χ^2_2, considering departures from normality in the continuous component $\log \mathbf{u}$. $\log \mathbf{u}$ is now simulated from a t distribution with 4 degrees of freedom.

2.5 Application of filtering and Hurdle model in three datasets

2.5.1 Distributional assumptions

In Figure 2.1, we observe good agreement between the empirical distributions of positive *et* values and their postulated normal distribution for four genes in data set A. This confirms that a log-normal model for the positive expression level, $y_{ij}|v_{ij} = 1$, is appropriate. Even cells in the lowest quantiles of *et* (and lowest quantiles of expression) still have expression far away from the bound at 0, suggesting that undetected genes represent cells with null or negligible RNA abundance. It is also noteworthy that the difference between the means (shown as a heavy, vertical line) of the hundred cell replicates and single cell replicates is approximately $\log_2(100 \cdot \pi_j^{(1)})$ cycles, where $\pi_j^{(1)}$ is the expression frequency of gene j in the single-cell experiments. As such, in genes with $\pi_j^{(1)} \ll 1$, such as FASLG, this difference between means is smaller than genes with $\pi_j^{(1)} \approx 1$. As we will see the next section, inclusion of the unexpressed cells ($v_{ij} = 0$) is important to accurately relate the expression level of the single-cell experiments to the hundred-cell experiments.

2.5.2 Concordance between hundred-cell and single-cell experiments

The 100-cell aggregates (see section 2.1.1, *data sets and notation*) allows us to assess the accuracy and reliability of our single-cell experiments by comparing this *in-vitro* 100-cell expression to an *in-silico* estimate obtained by averaging the expression of 100 single-cell measurements. The *in silico* average of signal in a gene j and unit k from 100 single-cell wells is $y_{jk}^{(1)} = \sum_{i=1}^{100} y_{ijk}/100$ where y_{ijk} is the expression measurement of gene j in cell i and unit k. We compare this to the *in-vitro* "average" of signal from a 100 cell aggregate. In this case, we just use the expression of a gene-unit and divide by the number of cells (100).

The concordance here is assessed both visually by plotting $\log_2(y_{jk}^{(1)} + 1)$ *vs.* $\log_2(y_{jk}^{(100)} + 1)$ (Figure 2.5) and by calculating the concordance correlation coefficient (r_c) between the two variables, which is often used to quantify reproducibility [Lin, 1989]. The shifted log transformation allows visualization of both the discrete and continuous components while being on the *et* scale.

We first use this concordance experiment to test whether wells that do not cross the fluorescence threshold after $c_{\max}$ should be treated as exact zeros or missing values. If we suppose that $v_{ij} = 0$ implies an assay failure and the measurement should be discarded, we would simply compute the single cell average over expressed cells, *i.e.* $y_j^{(1)} = \sum_i y_{ij} v_{ij} / \sum_i v_{ij}$. Figure 2.5 demonstrates good concordance between the hundred-cell and single-cell experiments when the undetected genes are treated as zeros. However, this is not the case when the zeros are treated as missing values.

2.5.3 Filtering outlying cells

In addition to the concordance measure r_c, we use another goodness-of-fit measure to optimize our filtering parameters t_z, t_ζ defined by,

$$\overline{\mathrm{WSS}} = \sum_{j,k} n_{jk} \left(\log_2(y_{jk}^{(1)} + 1) - \log_2(y_{jk}^{(100)} + 1) \right)^2 / JK, \tag{2.7}$$

where $n_{jk} = \sum_i v_{ijk}$ is the number of positive wells for gene j in unit k in the single-cell experiments. For a particular gene and unit, the $\overline{\mathrm{WSS}}$ decreases as we lower the filtering

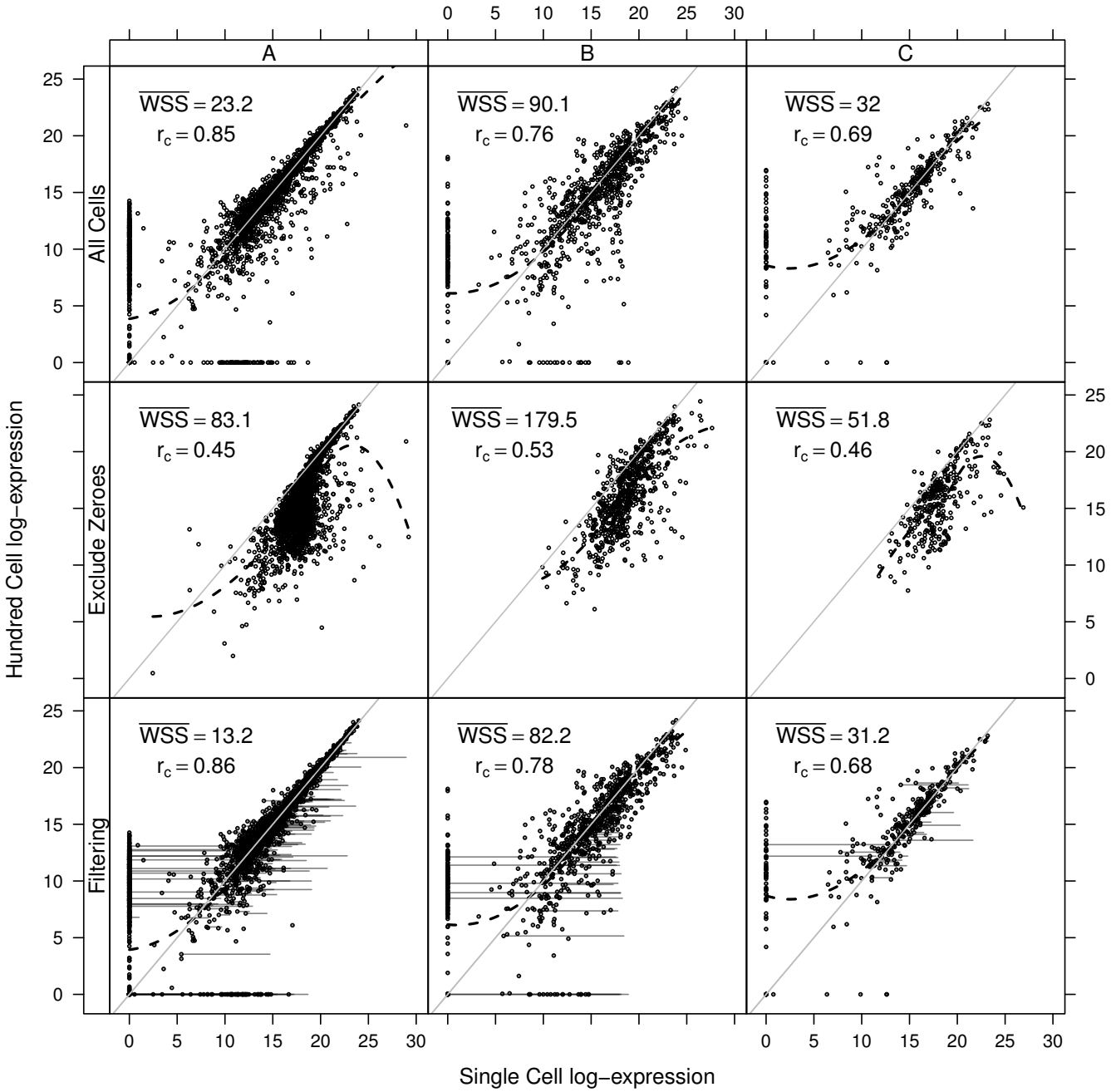

Figure 2.5: Concordance between hundred cell $y^{(100)}/100$ and $y^{(1)}$, the *in silico* average of single cell wells. The rows correspond to inclusion, case-wise deletion and inclusion and filtering of zeros $v_{ij} = 0$. Dark, thin lines show the initial location of a gene before filtering and connect to the location of the gene after filtering. The concordance correlation coefficient r_c and average weighted squared deviation $\overline{\text{WSS}}$ is printed. The dotted black line shows a loess fit through the data. In all cases, the expression values are transformed using a shifted log-transformation ($\log_2(y+1)$).

threshold and extreme values are filtered. Eventually, so many cells are removed that there is zero expression (and a large deviance) for the *in-silico* estimate. Thus we wish to find a set of values for the filtering parameters that would lead to the lowest $\overline{\text{WSS}}$ measure across the three data sets used here. The addition of the scaling factor n_{jk} gives higher weight to combinations with more *ex ante* positive observations, so that the contribution to the sum of squares would be smaller in gene×unit combinations that have fewer expressed cells. The factor n_{jk} can also be interpreted as the scaling factor for the variance of the mean over positive observations. Finally, the $\overline{\text{WSS}}$ is computed on the $\log_2(y+1)$ scale to reduce the effect of extreme outliers.

When hundred-cell aggregates are available, one can optimize the filter parameters t_z, t_ζ by minimizing the $\overline{\text{WSS}}$ over possible combinations. In our case, we found that setting $t_z = 9, t_\zeta = 9$ achieves the best reduction in $\overline{\text{WSS}}$ across the three data-sets explored here (Table 2.1). Using these values, our filtering criteria moderately improve the concordance between the single-cell and hundred-cell experiments in two of the data sets but dramatically improve (decrease) the weighted sum of squares. This improvement is evident graphically, since the per-unit averages of *et* of multiple genes move towards the diagonal.

Beside improving $\overline{\text{WSS}}$ and generally improving r_c, we explore the effect of filtering on detection of control genes in the (Table).

Filtering parameter optimization

We determine appropriate values of the continuous parameter t_z and the expression proportion parameter t_ζ by searching the grid $t_z, t_\zeta \in [3, 5, \ldots, 9]$.For each value in the grid, the weighted residual sum of squares $\overline{\text{WSS}}$ is calculated. The minimizing values vary somewhat on the data set, so we based our recommendation of $t_z = t_\zeta = 9$ by choosing values that minimize the maximum residual $\overline{\text{WSS}}$ across data sets.

		t_z			
Data set	t_ζ	3	5	7	9
A	3	6.44	2.95	2.95	2.83
A	5	5.26	1.17	0.11	0.00
A	7	5.26	0.01	0.40	1.02
A	9	5.21	0.00	0.43	1.64
B	3	8.85	5.33	5.56	6.28
B	5	3.18	0.00	0.04	0.69
B	7	3.18	0.00	0.04	0.69
B	9	3.18	0.00	0.04	0.69
C	3	27.99	16.88	8.56	8.56
C	5	27.98	8.90	7.08	6.13
C	7	25.84	6.41	4.55	0.00
C	9	25.84	6.41	4.55	0.00

Table 2.1: $\overline{\text{WSS}} - \min_{t_z,t_\zeta} \overline{\text{WSS}}$ values across data sets and filtering parameters. For each data set, the minimum $\overline{\text{WSS}}$ is subtracted so that cells that achieve that value contain zeros.

Effect of filtering on control genes

The flow cytometric sorting of the data sets we consider allows examination of how filtering affects the quality of the sorted cells. In data sets B and C, the cells were putatively sorted to be CD4+, and CD8- using surface markers. In data set A, the cells were sorted to be CD8+, CD4-. The flow sorting is based on protein presence, while the gene expression is based on mRNA, so it's expected that there will be differences between the two that reflect biological differences between transcription and translation.

Nonetheless, with the proviso that it would be surprising for there to be 100% concordance

between translated protein and transcribed mRNA, these genes still may serve limited roles as positive and negative controls. As seen in Table 2.2, in two out of the three data sets (A and C), the percentage of unexpected transcript decreases substantially after filtering. In B, little change is noted in either positive control or negative control genes. Since this unexpected transcript could reflect contamination or assay failure, and the filtering is agnostic to the presence or absence of any particular gene, this provides additional evidence beyond $\overline{\text{WSS}}$ that filtering may improve sample quality. The insubstantial changes in the positive control genes (CD8 for A, CD4 for B and C) likely reflects the overall rareness of filtering ($< .5\%$) in any of the data sets.

In this calculation, the reference group was taken to be all wells with detected expression in at least one gene, hence already includes step one of the filtering algorithm described in section 2.3 of the main text. If we take the reference group to be the raw measurements, then substantial increases in the expression of positive control genes are noted in B and C (comparison not shown).

Data set	Filtered	Pre-CD4	Post-CD4	ΔCD4	Pre-CD8	Post-CD8	ΔCD8
A	0.3%	0.013	0.011	-16.9%	0.879	0.879	0.0%
B	0.2%	0.424	0.423	-0.0%	0.002	0.002	0.2%
C	0.3%	0.484	0.485	0.3%	0.016	0.013	-16.4%

Table 2.2: Effect of filtering, beyond the effect of excluding null wells, on control genes CD4 and CD8. Data set A is expected to be positive for CD8, negative for CD4. Data sets B and C are expected to be negative for CD8, positive for CD4. The percentage of cells filtered, the frequencies π before and after filtering and percentage change in π of these genes is printed.

2.5.4 Normalization and housekeeping genes

Other authors have noted that "the gene transcript number is ideally standardized to the number of cells" [Vandesompele et al., 2002], which is the case with gene expression from sorted cells. So it is not entirely a surprise that we find little evidence for housekeeping genes providing useful normalization here. For a housekeeper to have good validity, it should have high cross correlation with other housekeeping genes. This is not the case for housekeepers GAPDH and POLR2A, which in data set A, in linear regression have an $R^2 = .027$. In Figure 2.6, we observe in scatter plots of housekeepers' *et* that the correlation drops even further (to an $R^2 = .017$) after filtering outlying cells (see previous section). Since the correlation between housekeepers is present primarily in cells we suspect suffered from technical error, we find little utility in normalization schemes. In fact, the use of housekeeping genes for normalization could even result in masking cellular artifacts that should be filtered.

2.5.5 An efficient test of differential expression for single-cells

In data set A, approximately 90 cells in each of 16 subjects were measured in unstimulated and stimulated states (see 2.1.1). This permits conducting a test for each gene in each subject for differences in π and μ, as described in section 2.4. We plot the number of discoveries at various false discovery rates (FDR) in Figure 2.7. The combined likelihood test produces the greatest number of discoveries over a wide range of FDR. For example, at an FDR of 1%, our combined test could detect more than 20 additional gene×unit changes across the four stimulations.

Another feature of the combined LRT is its robustness to background gene frequency π_j. Of course, if $\pi_j \approx 0$ on average, then any test will be underpowered to detect group differences. But using only the continuous components amounts to "throwing away" data for genes with intermediate π_j. And similarly, using only the dichotomous component results in a test insensitive to differences in μ_j in frequently expressed genes. This robustness to the π_j spectrum is shown in Figure 2.8 in which $-\log_{10}$ p-values are shown for the Bernoulli,

normal and combined LRTs versus frequency of π_j.

A total of 65 genes were detected at an FDR of 1% in at least one individual. We define $p^\star = -\text{sign}(\mu_1 - \mu_0) \cdot \log_{10} p$ as the negative $\log_{10}$ p-value times an indicator variable which is positive when stimulated groups have greater expression, and negative otherwise. Figure 2.9 plots a heatmap of signed $\log_{10}$ p-values. The selected genes are in clustered rows; the 16 individuals are arranged in columns by stimulation. The color above each column indicates which of the four antigen stimulations the individual received. From this, it is clear that genes cluster into up-regulated and down-regulated modules and that there is much individual variability in response. Some genes appear to have stronger responses to particular antigens, such as the response to CMV (red and purple columns) in FASLG and CLEC2B.

2.6 Discussion

Current approaches for analysis of single-cell assays have incompletely utilized the salient features of the experiment, and the resulting inference can be sub-optimal. In this chapter, we have presented a framework for data exploration, quality control and testing for differential expression using single-cell data. Our comparison of 100-cell and single-cell measurements shows that undetected genes in an assay should be treated as effective "zeros." Both the discrete, zero-inflated portion and continuous portion of single-cell expression data are meaningful for detecting outliers. Moreover, differences in either could be of biological interest, so it is desirable to combine evidence from both to detect changes in expression. Our likelihood ratio test allows just that.

Although we have suggested default parameters for the filtering of outliers, informed from several data sets, our defaults are likely conservative. They are 3-4 times larger than the most substantial difference in expression between experimental groups we observed. Acquiring forms of ground-truth besides "bulk" experiments (in our case, 100-cell aggregates) could allow forming tighter bounds. As in any outlier-based filtering procedure, it is desirable to tune for the problem at hand. The thresholds t_z and t_ζ should be different when eliminating

potential technical error is of greatest concern than when one is most interested in searching for biological heterogeneity.

We have used the χ^2 asymptotic distribution of the LRT to compute p-values and assess significance. This approximation is relatively accurate and robust to the distributional form of Y when the expected number of expressed cells is greater than 8. Otherwise, approximating the null distribution using permutations as in Ge et al. [2003] is more appropriate.

Further work, incorporating a mixed-effects model to our likelihood ratio test, could extend its applicability. The test outlined in this chapter may not be appropriate in cases where traits of interest are not blocked within individuals (*e.g.*, comparing between phenotypes like HIV+ *vs.* HIV-). In this case, one wishes to identify gene expression changes across groups, in spite of high individual-to-individual heterogeneity. By modeling the mean and proportion of expression as common across groups and adding specific random effects for between-individual variability, our model could be extended to address such experimental questions as well.

Single-cell gene expressions assays have already been shown to be useful in multiple studies and will become even more routine once sequencing at the single-cell level becomes practical [Varadarajan et al., 2011, Ramskold et al., 2012]. As a consequence, the development of effective statistical methods to analyze such data is becoming increasingly important. This chapter offers a coherent framework for researchers using this nascent technology.

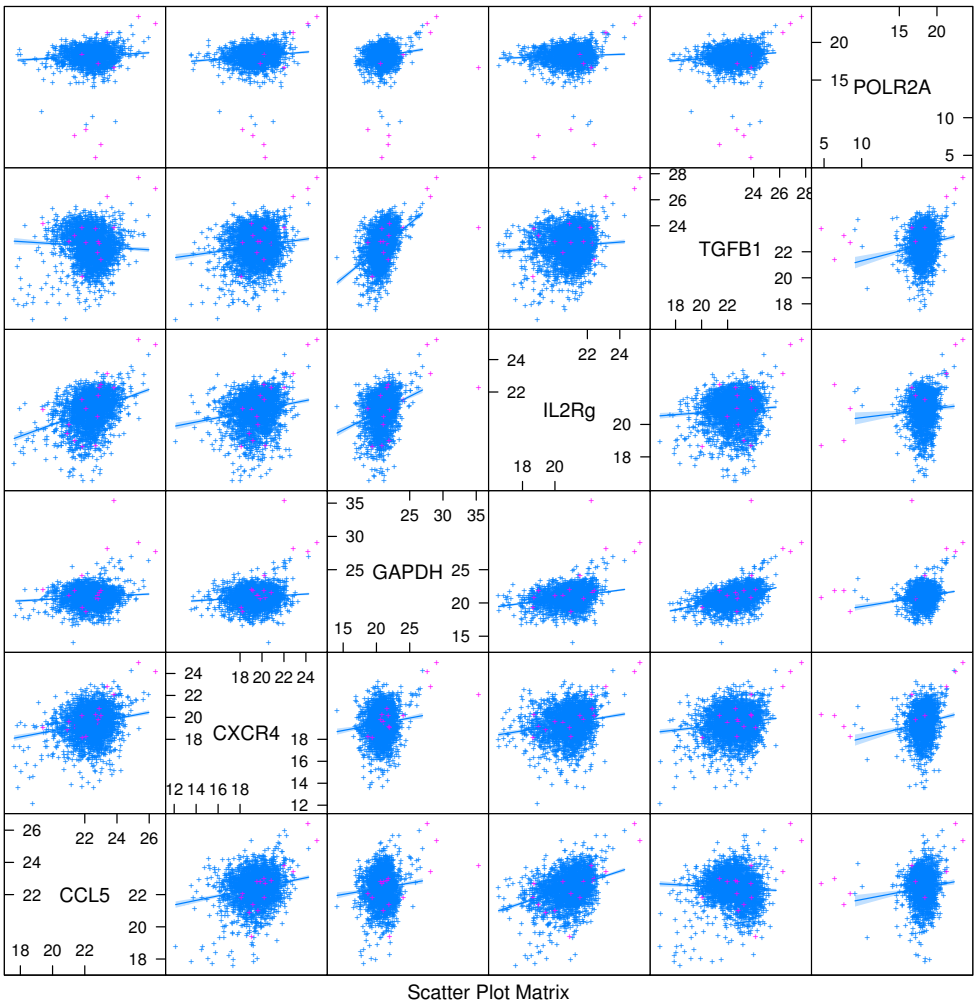

Figure 2.6: Scatter plots of housekeeping genes GAPDH, POLR2A and other frequently expressed ($\pi > .95$) genes. Cells flagged for filtering are indicated in purple. A regression line of the form $et_y \sim et_x +$ intercept, and its standard error is plotted using unfiltered cells.

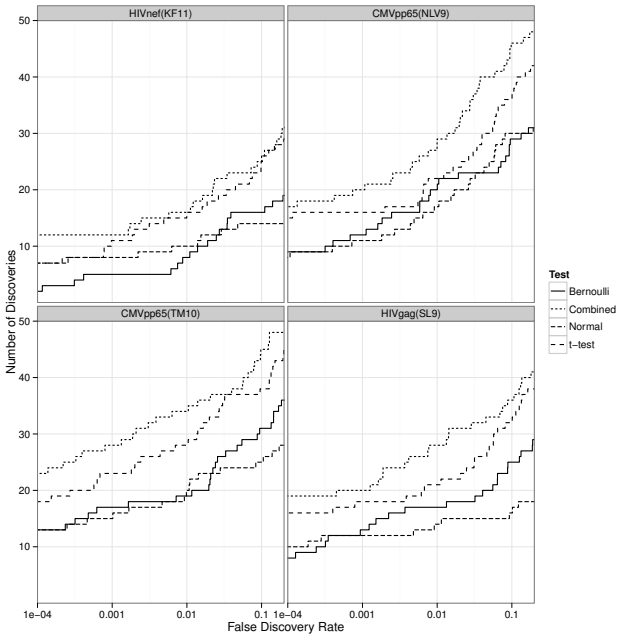

Figure 2.7: Number of discoveries (genes × units) versus the false discovery rate, by treatment, data set A. The combined likelihood ratio test is compared to a Bernoulli or normal-theory only likelihood ratio test, as well as a t-test of the raw expression values (2^{et} scale), including zero measurements.

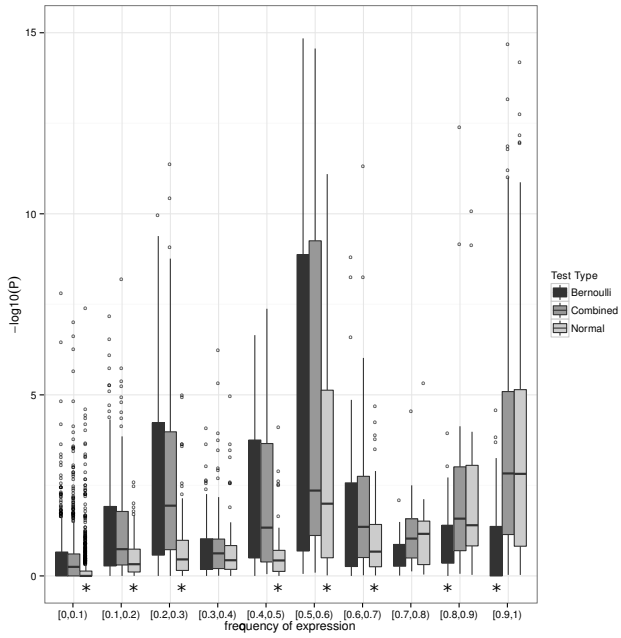

Figure 2.8: $-\log_{10} P$ of tests (genes $\times$ units) versus frequencies of expression π of the genes. The Bernoulli, normal-theory and combined likelihood ratio tests are plotted. * indicates test is different from the combined test at 5% significance in a Wilcoxon signed-rank test.

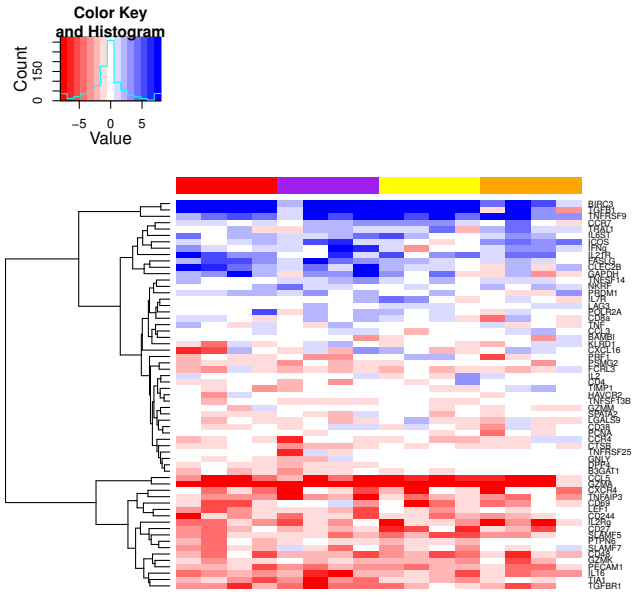

Figure 2.9: Heatmap of signed $\log_{10} p$ for selected genes (rows) and 16 individuals (columns). The color above each column indicates the antigen stimulation applied to the cells. Red and purple are two different CMV antigen pools; yellow and orange are two different HIV antigen pools.

Chapter 3

TRANSCRIPTIONAL CHANGE AND HETEROGENEITY IN SINGLE-CELL RNA SEQUENCING DATA

3.1 Previous methods for single cell whole-transcriptome sequencing

Whole transcriptome expression profiling of single cells via RNA sequencing (scRNA-seq) is the logical apex to single cell gene expression experiments. In contrast to transcriptomic experiments on mRNA derived from bulk samples, this technology provides powerful multi-parametric measurements of gene co-expression at the single-cell level. However, the development of equally potent analytic tools has trailed the rapid advances in the biochemistry and molecular biology, and several challenges need to be addressed to fully leverage the information in single-cell expression profiles.

First, single-cell expression has repeatedly been shown to exhibit a characteristic bimodal expression pattern, wherein the expression of otherwise abundant genes is either strongly positive, or undetected within individual cells. This is due in part to low starting quantities of RNA such that many genes will be below the threshold of detection, but there is also a biological component to this variation (termed extrinsic noise in the literature) that is conflated with the technical variability [Elowitz et al., 2002, Raj et al., 2008, Sanchez and Golding, 2013]. We and other groups [McDavid et al., 2013, Shalek et al., 2014, Kharchenko et al., 2014, Trapnell et al., 2014] have shown that the proportion of cells with detectable expression reflects both technical factors and biological differences between samples. Results from synthetic biology also support the notion that bimodality can arise from the stochastic nature of gene expression [Kaufmann and van Oudenaarden, 2007, Marinov et al., 2014].

Secondly, measuring single cell gene expression might seem to obviate the need to normalize for starting RNA quantities, but recent work shows that cells scale transcript copy

number with cell volume (a factor that affects gene expression globally) to maintain a constant mRNA concentration and thus constant biochemical reaction rates [Marguerat et al., 2012, Padovan-Merhar et al., 2015]. In scRNA-seq, cells of varying volume, and hence mRNA copy number, are diluted to an approximately fixed reaction volume leading to differences in detection rates of various mRNA species that are driven by the initial cell volumes. Technical assay variability (e.g. mRNA quality, pre-amplification efficiency) and extrinsic biological factors (e.g. nuisance biological variability due to cell size) that globally affect transcription remain, and can significantly influence expression level measurements. Our approach easily allows for estimation and control of the *cellular detection rate* (CDR) while simultaneously estimating treatment effects.

Previously, Kharchenko et al. [2014] developed a so-called three-component mixture model to test for differential gene expression while accounting for bimodal expression. Their approach is limited to two-class comparisons and cannot adjust for important biological covariates such as multiple treatment groups and technical factors such as batch or time information, limiting its utility in more complex experimental designs. On the other hand, several methods have been proposed for modeling bulk RNA-seq data that permit sophisticated modeling through linear [Law et al., 2014] or generalized linear models [Robinson et al., 2010, Anders and Huber, 2010] but these models have not yet been adapted to single-cell data as they do not properly account for the observed bimodality in expression levels. This is particularly important when adjusting for covariates that might affect the expression rates. As we will demonstrate later, such model mis-specification can significantly affect sensitivity and specificity when detecting differentially expressed genes and gene-sets.

Here, we propose a Hurdle model tailored to the analysis of scRNA-seq data, providing a mechanism to address the challenges noted above. It is a two-part generalized linear model that simultaneously models the rate of expression over background of various transcripts, and the positive expression mean. Leveraging the established theory for generalized linear modeling allows us to accommodate complex experimental designs while controlling for covariates (including technical factors) in both the discrete and continuous parts of the model.

We introduce the *cellular detection rate (CDR)*: the fraction of genes that are detectably expressed in each cell, which, as discussed above, acts as a proxy for both technical (e.g. dropout, amplification efficiency, etc.) and biological factors (e.g. cell volume and other extrinsic factors other than treatment of interest) that globally influence gene expression. As a result it represents an important source of variability in scRNA-seq data that needs to be modeled (Figure 3.1). Our approach of modeling the CDR as a covariate, offers an alternative to the weight correction of Shalek et al. [2014] that does not depend on the use of control genes and allows us to jointly estimate nuisance and treatment effects. Our framework permits the analysis of complex experiments, such as repeated single cell measurements under various treatments and/or longitudinal sampling of single cells from multiple subjects with a variety of background characteristics (e.g. gender, age, etc.) as it is easily extended to accommodate random effects. These features are especially important when sampling single cells since there are multiple sources of variance (e.g. cell-to-cell variance within a subject, and subject-to-subject variance). These type of experiments/designs will become routine in future single-cell studies such as for clinical trials where single-cell assays will be performed on large cohorts with complex designs.

In our Hurdle model, differences between treatment groups are summarized with pairs of regression coefficients whose sampling distributions are available through bootstrap or asymptotic expressions, enabling us to perform complementary differential gene expression and gene set enrichment analyses (GSEA). We use an empirical Bayesian framework to regularize model parameters, which helps improve inference for genes with sparse expression, much like what has been done for bulk gene expression [Smyth, 2004]. Our GSEA approach accounts for gene-gene correlations, which is important for proper control of type I errors [Wu and Smyth, 2012]. This GSEA framework is particularly useful for synthesizing observed gene-level differences into statements about pathways or modules. Finally, our model yields *single cell residuals* that can be manipulated to interrogate cellular heterogeneity and gene-gene correlations across cells and conditions. We have named our approach MAST for **Model-based Analysis of Single-cell Transcriptomics**.

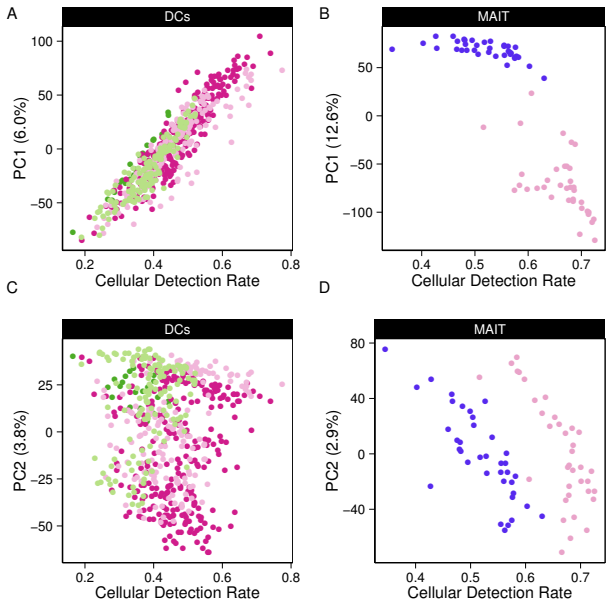

Figure 3.1: The fraction of genes expressed, or cellular detection rate (CDR), is correlated with the first two principal components of variation in MAIT and DC data sets.

We illustrate the method on two data sets. We first apply our approach to an experiment comparing primary human non-stimulated and cytokine-activated Mucosal-Associated Invariant T (MAIT) cells. MAST identifies novel expression signatures of activation, and the single-cell residuals produced by the model highlights a population of MAIT cells showing partial activation but no induction of effector function. We then illustrate the application of MAST to a previously-published complex experiment studying temporal changes in murine bone marrow-derived dendritic cells subjected to LPS stimulation. We both recapitulate the findings of the original publication and describe additional coordinated gene expression changes at the single-cell level across time in LPS (lipopolysaccharide) stimulated mDC (myeloid dendritic cells).

3.2 Robust estimation of a vector regression model in scRNAseq

Our MAST framework models the $\log_2$(transcripts per million (TPM)+1) single-cell gene expression matrix using a two-part generalized linear model. One component of MAST models the discrete expression rate of each gene across cells, while the other component models the conditional continuous expression level (conditional on the gene being expressed). We address several obstacles to the use of this model in practice in transcriptomic data sets:

- The residual variance of each gene can be difficult to estimate for infrequently expressed genes, yet many genes are expected to have similar residual variances. Accurate estimation of this quantity is important for testing for differential expression. We propose a hierarchical, empirical Bayes model to share information between genes on their residual variances.

- The maximum likelihood estimate (MLE) may not exist when a covariate can perfectly explain the discrete expression rate of a gene. Paradoxically, these *linearly separable* genes are of greatest interest, since their expression patterns are so predictable. We solve this by adopting a boundary-avoiding prior which results in a finite (penalized) MLE.

Given an experimental design that provides covariates, the cell expression rate is modeled using logistic regression. The expression level is modeled as Gaussian, conditional on the presence of positive expression.

Given normalized, possibly thresholded (see 3.2.1), scRNA-seq expression $Y = [y_{ij}]$, the rate of expression and the level of expression for the expressed cells are modeled conditionally independent for each gene j. Define the indicator $V = [v_{ij}]$ indicating whether gene j is expressed in cell i, i.e. $v_{ij} = 0$ if $y_{ij} = 0$ and $v_{ij} = 1$ if $y_{ij} > 0$. We fit logistic regression models for the discrete variable V and Gaussian linear model for the continuous variable ($Y|V = 1$) independently, as follows,

$$\text{logit}\left(Pr\left(V_{ij}=1\right)\right)=X_i\beta_j^D,$$

$$Pr\left(Y_{ij}=y|V_{ij}=1\right)=\text{Normal}\left(X_i\beta_j^C,\sigma_j^2\right).$$

Shrinkage of the continuous variance

As the number of expressed cells varies from gene to gene, so does the amount of information available to estimate the residual variance of the gene. On the other hand, many genes can be expected to have similar variances. To accommodate this feature of the assay, we shrink the gene-specific variances estimates to a global estimate of the variance using an empirical Bayes method. Let τ_j^2 be the precision (1/variance) for $Y_j|V_j=1$ in gene j. We suppose $\tau_j^2 \sim Gamma\left(\alpha,\beta\right)$, find the joint likelihood (across genes) and integrate out the gene-specific inverse variances. Then maximum likelihood is used to estimate α and β. Due to conjugacy, these parameters are interpretable providing 2α pseudo-observations with precision β/α. This leads to a simple procedure where the shrunken gene-specific precision is a convex combination of its MLE and the common precision. This approach accounts for the fact that the number of cells expressing a gene varies from gene to gene. Genes with fewer expressed cells end up with proportionally stronger shrinkage, as the ratio of pseudo observations to actual observations is greater. Further details are available in section A.1.

Bayesian logistic regression for discrete component

In logistic regression, when the binary outcome can be perfectly predicted by a covariate (or linear combination of covariates), then "linear separation" is said to be present, and parameter estimates will diverge towards $\pm\infty$ while the Fisher information becomes singular. (In contrast, if even a single cell were to violate this linear separation, then the Fisher Information would be invertible.) Yet cases with linear separation are of particular interest, since a gene that so sharply changes by condition is noteworthy. To accommodate this scenario, we apply a Bayesian logistic regression procedure available in the `bayesglm` function in the

`R` package `arm`. A Cauchy distribution prior centered at zero for the regression coefficients results in maximum a posteriori (MAP) estimates nearly identical to the MLE when linear separation is not present. Under linear separation the Bayesian MAP estimate is finite, with non-singular Hessian about the MAP (providing an estimate of the statistical precision, akin to the Fisher information.) Favorable small-sample frequentist properties have also been described in [Gelman et al., 2008].

Testing for differential expression

Because V_j and Y_j are defined conditionally independent for each gene, tests with asymptotic χ^2 null distributions, such as the likelihood ratio or Wald tests can be summed and remain asymptotically χ^2, with the degrees of freedom of the component tests added. For the continuous part, we use the shrunken variance estimates derived through our empirical Bayes approach described above. The test results across genes can be combined and adjusted for multiplicity using the false discovery rate (FDR) adjustment [Benjamini and Hochberg, 1995]. In this chapter, we declare a gene differentially expressed if the FDR adjusted p-value is less than 0.01 and the estimated fold-change is greater than 1.5 (on $\log_2$ scale).

3.2.1 Estimating thresholds of expression

In previous studies, small, but non-zero expression values were thresholded using an arbitrary fixed threshold [Shalek et al., 2014]. These conservative fixed thresholds do not allow any variation between genes for differing levels of background noise as suggested by Kharchenko et al. [2014]. In order to adaptively determine the level of background noise, we propose a thresholding routine that shares information across genes. We, and others, have observed that threshold for background expression may depend on the total expression in a gene, consistent with varying levels of contamination with homologous genomic DNA, or differing levels of mappability for various transcripts. Let m_j denote median expression across cells in gene j. The G genes are partitioned into K bins based on m_j such the median of bin k is

greater than bin $k+1$, and so forth. This binning allows for thresholding that varies with the expression level of a gene.

For each bin we apply kernel density estimation and determine if the distribution is bimodal, then apply peak finding to estimate the threshold t_k as the minimum density point between the two major peaks in the bin. For bins that are not bimodal, their thresholds are set as follows:

1. Determine a threshold $t*$ in an examplar bin $k*$ with two peaks that reliably represent signal and background modes. We use the 75th percentile bin amongst bins where two peaks are found, though other criteria (formal tests for bimodality) could be used.

2. Find adjusted thresholds t'_k: For $k < k*$, set $t'_k = \min(t*, t_k)$. For $k > k*$, set $t'_k = \max(t*, t_k)$.

This ensures that the thresholds are monotonically increasing and shares information across bins to impute thresholds for bins where the distribution of the data was not bimodal. This function is implemented in the `thresholdSCRNACountMatrix` function of the MAST package.

3.2.2 Module Scores

In order to assess the degree to which each cell exhibits enrichment for each gene module, we use quantities available through our model to define module "scores," which are defined as the observed expression corrected for cellular detection rate (CDR) effect, analogous to those defined by Shalek et al. The score s_{ij} for cell i and gene j is defined as the observed expression corrected for the CDR effect: $s_{ij} = y_{ij} - \widetilde{y}_{ij}$ where $\widetilde{y}_{ij}$ is the expected expression, given all covariates except the treatment effects, from the fitted model. In our two part model, $\widetilde{y}_{ij} = \hat{v}_{ij}\hat{y}_{ij}$ where $\hat{v}_{ij}$ and $\hat{y}_{ij}$ are the predicted values from the discrete and continuous components of our hurdle model.

This can be interpreted as correcting the observed expression of gene j in cell i by subtracting the conditional expectation of nuisance effects. A gene module score for cell i is

the average of the scores for the genes contained in the module, i.e. $\sum_{\{j \in module\}} s_{ij}/|module|$

3.3 Adjusting for the cellular detection rate through vector regression

The CDR, the proportion of genes expressed in a single cellis an important source of variability in the data sets we explore here where

$$\text{CDR}_i = 1/G \sum_{j=1}^{G} v_{ij}.$$

This is supported through principal component analysis (PCA) (Figure 3.1). It is highly correlated with the second principal component (PC, Pearson's rho=0.76 grouped, 0.91 stimulated, 0.97 non-stimulated) in the MAIT dataset and with the first PC (rho=0.92 grouped, 0.97 non-stimulated, 0.92 LPS, 0.89 PAM, 0.92 PIC) in the mDC dataset. As we observe larger CDR variability within treatment groups than across groups, it is likely that the CDR is a nuisance factor. This is further supported by the fact that the CDR calculated using control (e.g. housekeeping) genes is highly correlated with the CDR calculated over all genes (Figure A.2).

We thus conjecture that CDR is a proxy for unobserved nuisance factors that should be explicitly modeled. In particular, we suggest that the CDR captures variation in global transcription rate due to difference in cell size (among other factors) [Padovan-Merhar et al., 2015], as well as technical variation due to factors such as cell viability and efficiency in first strand synthesis. Fortunately, MAST easily accommodates covariates, such as the CDR, and more importantly allows joint, additive modeling of them with other biological variables of interest, with the effect of each covariate decomposed into its discrete and continuous parts. Applying an analysis of deviance with the MAST hurdle model, we quantified the amount of variability that could be attributed to CDR. The CDR accounts for 5.2% of the deviance in the MAIT data set and 4.8% in the mDC data set for the average gene, and often times much more than that: it comprises more than 9% of the deviance in over 10% of genes in both data sets, particularly for the discrete component of the model (Figure A.3). It should also be noted that the CDR deviance estimates for many of the genes are comparable (if

not greater) to the treatment deviance estimates. It is possible that the CDR and treatment effects could be partially confounded, for example, treated cells could become larger in volume. We explored the effect of confounding between the CDR and treatment effects on the MAST false positive rate in the presence and absence of CDR control in the MAST model (Figure A.4). Controlling for CDR improves the sensitivity and specificity of MAST in the presence of confounding, and doesn't negatively impact its performance either in the absence of confounding or in the absence of a CDR effect.

That CDR predicts expression levels contradicts the model of independent expression between genes, since the level of expression (averaged across many genes) would not affect the level in any given gene were expression independent. It is especially important to adjust for it when testing for co-expression between genes, or the apparent correlation between genes is greatly inflated (see **Residual analysis identifies networks of co-expressed genes implicated in MAIT cell activation**)

Comparison to other approaches

Finally, we have investigated the relationship between our approach and the weight correction of Shalek et al. [2014] and other technical bias correction approaches like RUV [Risso et al., 2014] and SVA [Leek, 2014] (Figure A.5). The CDR has a strong linear relationship to the weights of Shalek et al. [2014], as well as with the first component of SVA and second component of RUV. Thus, use of the CDR as a covariate can be seen as a statistically rigorous way to correct for the dropout biases of Shalek et al, without the need to use control genes, and more importantly with the ability to control for these while estimating treatment effects. Although CDR is correlated to the latent components found via RUV or SVA in the data sets we consider here, CDR has the advantage of biological interpretability as a cellular scaling factor.

3.3.1 Single-cell sequencing identifies a transcriptional profile of MAIT cell activation

We applied MAST to our MAIT dataset to identify genes up- or down-regulated by cytokine stimulation while accounting for variation in the CDR. We detected 291 differentially expressed genes , as opposed to 1413 when excluding CDR. To determine whether this was due to a change in ranking or a simply a shift in significance, we compared the overlap between the top n genes in both models (varying n from 100 to 1413), and found that, on average, 35% (range 32% - 38%) of genes are excluded when CDR is modeled, suggesting that inclusion of this variable allows global changes in expression, manifest in the CDR, to be decomposed from local changes in expression. This is supported by gene ontology enrichment analysis (Figure A.6) of these CDR-specific genes (n=539), where we see no enrichment for modules associated with treatment of interest These CDR-specific GO terms (e.g. involvement of regulation of RNA stability and protein folding) may hint at biology underlying differences in the CDR that are not necessarily associated with treatment.

In order to assess the type-I error rate of our approach, we also applied MAST to identify differentially expressed genes across random splits of the MAIT cells. As expected, MAST did not detect any significant differences (Figure A.7), whereas DEseq and edgeR, designed for bulk RNA-seq, detected large number of differentially expressed genes even at stringent nominal FDR. SCDE, a single-cell RNA-seq specific method, also had higher false discovery rates than MAST. Permutation analysis demonstrated that the null distribution of the MAST test statistic was well calibrated (Figure A.8).

We examined the GO enrichment of genes detected by limma, edgeR, DESeq, or SCDE but not MAST and found that these sets generally lacked significant enrichment for modules related to the treatment of interest. MAST with CDR control also detected enrichment of immune-specific GO terms at a higher rate than other methods (Figure A.9). MAST's testing framework has better sensitivity and specificity than these approaches. Among models that do not adjust for CDR, SCDE performs best, but trails MAST and limma, which can adjust for CDR.

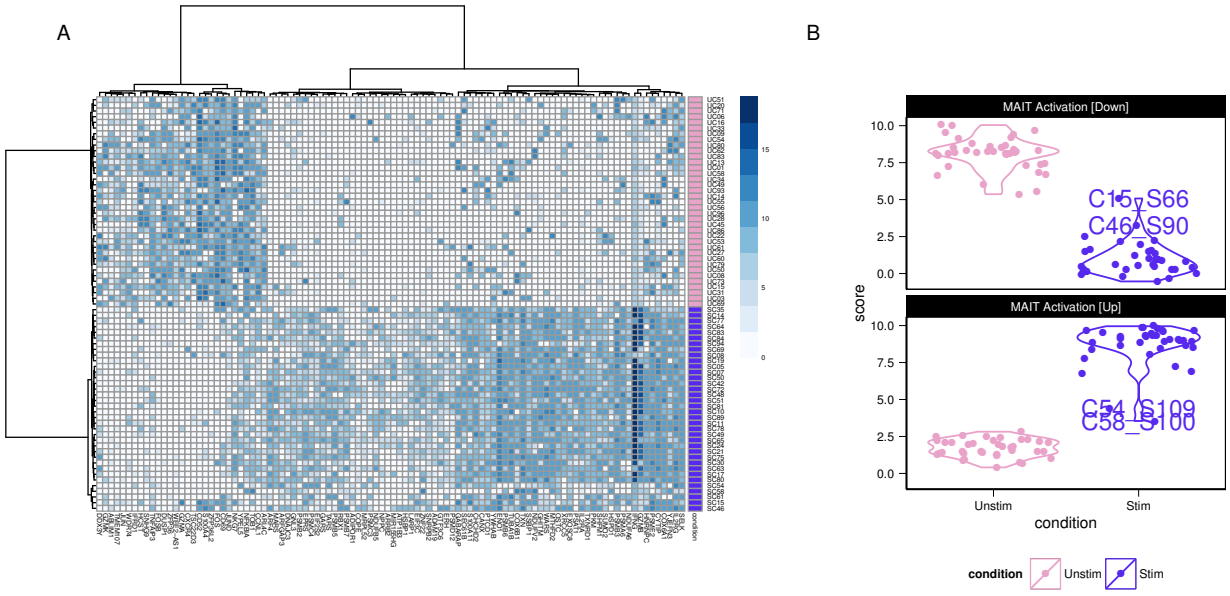

Figure 3.2: Single-cell expression ($\log_2$-TPM) of the top 100 genes identified as differentially expressed between cytokine (IL18, IL15, IL12) stimulated (purple) and non-stimulated (pink) MAIT cells using MAST (A). Partial residuals for up- and down- regulated genes are accumulated to yield an activation score (B), and this score is consistent with stimulation inducing heterogeneity compared to the unstimulated cells.

Figure 3.2 shows the single-cell expression ($\log_2$-TPM) of the top 100 genes identified as differentially expressed between cytokine (IL18, IL15, IL12) stimulated (purple) and non-stimulated (pink) MAIT cells using MAST. Following stimulation with IL12/15/18, we observe increased expression in proteins with effector function including Interferon$-\gamma$ (IFNG), granzyme-B (GZMB) as has been reported in Natural Killer (NK), Natural Killer T-cells (NKT) and memory T cells, and a concomitant downregulation of the AP-1 transcription factor network. CD69 is an early and only transient marker of activation that can be induced by stimulation of the T cell receptor or by cytokine signals. Its downregulation at the mRNA level after 24h is likely preceding subsequent protein-level downregulation [Chu et al., 2013, Tyznik et al., 2014, Smeltz, 2007].

We used these lists of up- and down-regulated genes to define a MAIT activation score that differentiates between stimulated and non-stimulated MAITs as shown in Figure 3.2. This yields a score for each cell, based on the model fit and adjusting for nuisance factors (see Methods), defined as the expected expression level across genes in a module. The score differentiates stimulated and non-stimulated cells, and demonstrates that the stimulated MAIT population is more heterogeneous in its expression phenotype. In particular, a few stimulated MAIT cells (SC08, SC54, SC48, SC15, SC46, and SC61 in Figure 3.2) exhibit low expression of IFN-γ response genes, suggesting these cells did not fully activate despite stimulation. Post-sort experiments via FCM show that the sorted populations were over 99% pure MAITs (Figure A.10B), and exhibited a change in cell size upon stimulation (Figure A.10B), and that up to 44% of stimulated MAITs didn't express IFN-γ or GZMB following cytokine stimulation (Figure A.10A). The non-responding cells in the RNA-seq experiment likely correspond to these non-responding cells from the flow cytometry experiment, and the observed frequencies of these cells in the RNA-seq and flow populations are consistent with each other (Pr(observing 6 or fewer non-responding cells) = 0.16 under binomial sampling). We discuss this heterogeneity in a further section. Importantly, the lists of up- and down-regulated genes can be used to define gene sets for gene set enrichment analysis in order to identify transcriptional changes related to MAIT activation in bulk experiments.

3.3.2 Temporal expression patterns of mouse dendritic cell maturation

Shalek et al analyzed murine bone-marrow derived dendritic cells simulated using three pathogenic components over the course of six hours and estimated the proportion of cells that expressed a gene and the expression level of expressing cells. We compared results from applying our model to those obtained by Shalek et al when analyzing their lipopolysaccharide (LPS) stimulated cells. As with the MAIT analysis, we used MAST adjusting for the CDR. MAST identified a total of 1359 differentially expressed genes (1996 omitting the CDR), and the CDR accounted for 5.2% of the model deviance in the average gene.

The most significantly elevated genes at 6h include CCL5, CD40, IL12B, and Interferon-inducible (IFIT) gene family members, while down-regulation was observed for EGR1 and EGR2, transcription factors that are known to negatively regulate dendritic cell immunogenicity [Miah et al., 2013].

3.4 Single cell gene set enrichment analysis

In many cases, it is desirable to synthesize the differential expression signature of a treatment into its effects on various pathways or gene sets. A competitive gene set enrichment analysis (GSEA) compares the average model coefficient in the *test* set (gene set of interest) to the average model coefficient in the *null* set (everything else) with a Z-test and is useful to compare gene sets to a background. We adapt the hurdle vector generalized linear model (vGLM) to provide such competitive tests. The approach is similar to that used by CAMERA [Wu and Smyth, 2012] for bulk experiments in its accounting for inter-gene correlation that is known to inflate the false significance (type-I error) in permutation-based GSEA protocols, although it differs in that it uses the sampling variance of each model coefficient to estimate the variance of the average coefficient, whereas CAMERA uses the empirical variance of the model coefficients.

Fix a gene module, ie, a collection of gene indices. Let $j = 1, \ldots, G_0, \ldots, G$ index the G genes measured, with $G - G_0$ genes in the *test set* (set of interest) and G_0 genes in the *null*

set (outside the set of interest). We assume that the following hurdle linear models

$$E\left(\mathbf{Y_j}|\mathbf{Y_j}>0\right) = x\beta_j + Z\eta_j,$$
$$\text{logit } E(\mathbf{Y_j}>0) = x\beta'_j + Z\eta'_j$$

have been fit for $j = 1,\ldots,G$. Here x is a simple covariate of interest (scalar in each observation) while Z is all other covariates (potentially a vector in each observation). The competitive gene set enrichment test considers the average coefficient of interest in the test and null sets:

$$\hat{\theta} = \frac{1}{G-G_0}\sum_{j=G_0+1}^{G}\hat{\beta}_j,$$
$$\hat{\theta}' = \frac{1}{G-G_0}\sum_{j=G_0+1}^{G}\hat{\beta}'_j,$$
$$\hat{\theta}_0 = \frac{1}{G_0}\sum_{j=1}^{G_0}\hat{\beta}_j,$$
$$\hat{\theta}'_0 = \frac{1}{G_0}\sum_{j=1}^{G_0}\hat{\beta}'_j,$$

and forms the test statistics

$$Z = \frac{\hat{\theta}-\hat{\theta}_0}{\sqrt{\widehat{\text{Var}}(\hat{\theta})+\widehat{\text{Var}}(\hat{\theta}_0)}},$$

with Z' formed analogously. The goal is to estimate $\text{Var}(\hat{\theta})$ and $\text{Var}(\hat{\theta}_0)$. Since, for example,

$$\begin{aligned}\text{Var}(\hat{\theta}_0) &= \text{Var}\left[\frac{1}{G_0}\sum_{j=1}^{G_0}\hat{\beta}_j\right]\\ &= \frac{1}{G_0^2}\left[\sum_{j=1}^{G_0}\text{Var}\,\hat{\beta}_j + 2\sum_{1\le j<h\le G_0}\text{Cov}\left(\hat{\beta}_j,\hat{\beta}_h\right)\right],\end{aligned}$$

it suffices to find some estimate of the genewise covariance matrix $\Sigma \in \mathbb{R}^{G\times G}$ for $\hat{\beta}$.

We chose to accomplish this by bootstrap. Repeat R times: sample cells with replacement and generate an expression matrix Y^*, and refit the hurdle linear model providing coefficients

$\hat{\beta}^*$ (and $\hat{\beta}'^*$). Collect the the bootstrapped coefficients in matrix $\hat{\boldsymbol{\beta}}^* \in \mathbb{R}^{R\times G}$. Estimate Σ via the sample covariance on $\hat{\boldsymbol{\beta}}^*$. Z scores are formed and calculated equivalently for the logistic regression coefficients. GSEA tests are done separately on the two components of the hurdle model and the results from the two components are combined using the Stouffer's method, which favors consensus in the two components.

Implementation Notes

We find that the bootstrapped covariances converge rather quickly, and $R = 100$ typically more than suffices. An adjustment to Z to account for Monte Carlo variation in the bootstrap estimate $\hat{\Sigma}$ is also available by comparing Z to a t-distribution with degrees of freedom determined through Welch's approximation on R effective observations. This modest requirement can be relaxed for exploratory analysis by assuming independence across genes and using model-based (asymptotic) estimates.

Note that the full covariance matrix estimate $\hat{\Sigma}$ never need be explicitly formed (since it is potentially memory intensive). Rather we accumulate the sum over the $(G - G_0)(G - G_0 + 1)/2$ inner products on the genes in the test set, to yield $\hat{\mathrm{Var}}(\theta)$. A working estimate of $\hat{\mathrm{Var}}(\theta_0)$ is updated by adding and subtracting only the covariances that have changed as G_0 changes between modules.

Stouffer's method for combining Z-scores is used to form the composite $\bar{Z} = (Z + Z')/\sqrt{2}$.

3.4.1 GSEA highlights pathways implicated in MAIT cell activation.

We used MAST to perform GSEA in the MAIT data using the blood transcriptional modules of Li et al. [2014]. The cell-level scores for the top 9 enriched modules (Figure 3.3A) continue to show significant heterogeneity in the stimulated and non-stimulated cells, particularly for modules related to T-cell signaling, protein folding, proteasome function, and the AP-1 transcription factor network. Although the standard deviations of the module scores were greater for stimulated than non-stimulated cells in 7 of the top 9 enriched modules (Table A.2), the

magnitude of variability for stimulated and non stimulated cells was fairly similar. Enrichment in stimulated cells (green) and non-stimulated cells (pink) is displayed for each module for the discrete and continuous components of the model (Figure 3.3B, see Methods), as well as a Z-score combining the discrete and continuous parts. The enrichment in the T-cell signaling module is driven by the increased expression of IFNg, GZMB, IL2RA, IL2RB, and TNFRSF9, 5 of the 6 genes in the module. Stimulated cells also exhibit increased energy usage, translation and protein synthesis, while down-regulating genes involved in cell cycle growth and arrest (and other cell cycle related modules). The down-regulation of cell cycle growth inhibition genes indicates that IL-12/15/18 signals are sufficient to prepare MAIT cells for cell proliferation. Interestingly, we observe down-regulation of mRNA transcripts from genes in the AP-1 transcription factor network. This has been previously described in dendritic cells in response to LPS stimulation [de Wit et al., 1996] and, indeed, we observe this effect in the mDC data set analyzed here (Figure A.11).

Our GSEA approach is more powerful than existing methods for bulk RNA-seq data (Figure A.12), and we discover significantly enriched modules with clear patterns of stimulation-induced changes that other methods omit (Figure A.13). Two such modules include the "T-cell surface signature" and "chaperonin mediated protein folding, whose component genes show elevated expression in response to stimulation (Figure A.13). These additional discoveries are not solely due to greater permissiveness in MAST. We applied MAST to identify differentially expressed gene sets across random partitions of the non-stimulated cells, to examine its false discovery rate. As expected, MAST did not detect any significant differences, which suggests that it has good type I error control (Figure A.7).

3.4.2 GSEA of mouse bone marrow-derived dendritic cells

In the mDC data set, besides finding signatures consistent with the modules from Shalek et. al. (Figure 3.4A), we identify modules that show similar annotation and overlap significantly with the *core antiviral* and *sustained inflammatory* signatures, including several modules linked to type 1 interferon response and antiviral signatures (Figure 3.4B). The "cellular

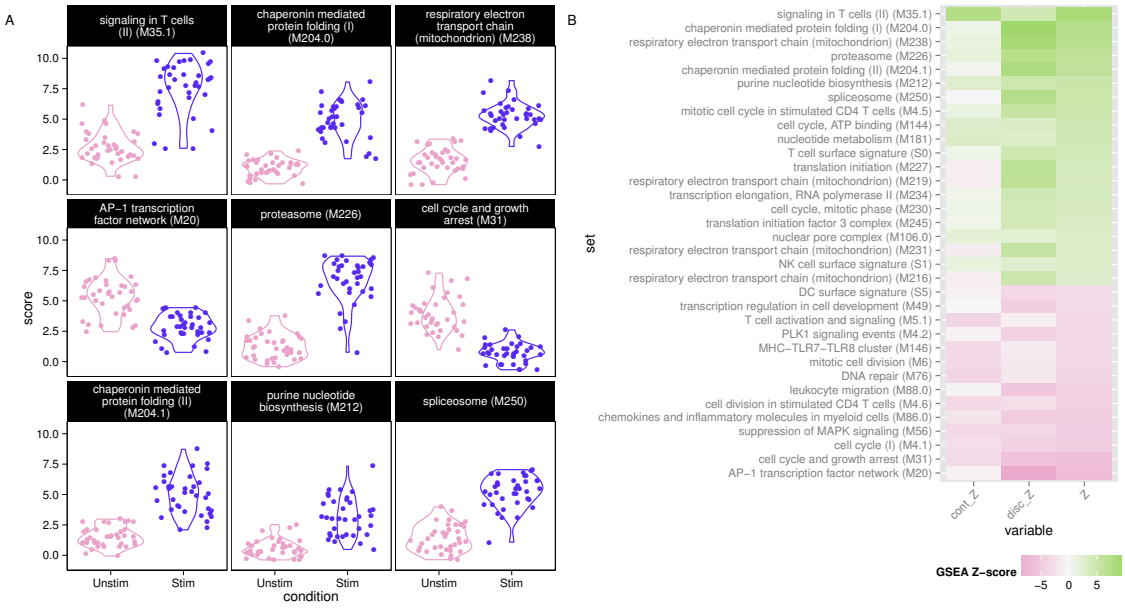

Figure 3.3: Module scores for individual cells for the top 9 enriched modules (A) and decomposed Z-scores (B) for single-cell gene set enrichment analysis in MAIT data set, using the blood transcription modules (BTM) database. The distribution of module scores suggests heterogeneity among individual cells with respect to different biological processes. Enrichment of modules in stimulated and non-stimulated cells is due to a combination of differences in the discrete (proportion) and continuous (mean conditional expression) of genes in modules. The combined Z-score reflects the enrichment due to differences in the continuous and discrete components.

response to interferon-beta" signature ($n = 22$) overlaps with the original core antiviral signature ($n = 99$) by 13 genes; *response* and *defense response to virus* signatures overlap with the core antiviral signature by 17/43 and 22/74 genes, respectively, suggesting the core antiviral signature captures elements of these known signatures. The *chemokine* (n=16) and *cytokine activity* (n=51) modules overlap with the sustained inflammatory (n = 95) module by 5 and 12 genes, respectively. All overlaps are significant at $p < 10^{-8}$ under hypergeometric sampling.

Our modeling approach identifies the two "early marcher" cells in the core antiviral module (marked with triangles on Figure 3.4A) corresponding to the same cells highlighted in Figure 4b of Shalek et al. Other modules exhibiting significant time-dependent trends include a module of genes involved in the AP-1 transcription factor network that is down-regulated (Figure A.11), a finding which has been previously shown in human monocytes following LPS stimulation [de Wit et al., 1996] . As with the MAITs, under monte carlo permutation of time-labels, no significant modules were found when using FDR control.

3.5 Residual analysis in hurdle vGLMs

Much of the heterogeneity between the non-responding and responding stimulated cells remains even after removal of marginal (gene level) stimulation effects. Since, MAST models the expected expression value for each cell, we can compute residuals adjusted for known sources of variability. The residuals can be compared across genes to characterize cellular heterogeneity and correlation.

The hurdle model, in general, provides two residuals: one for the discrete component and one for the continuous component. The notion of a residual in a GLM is not uniquely defined. We choose standardized deviance residuals. They are calculated for the discrete and continuous component separately, and then we combine the residuals by averaging them. If a cell is unexpressed, then its residual is missing and it is omitted from the average, thus the approach assumes Y is missing completely at random, given V.

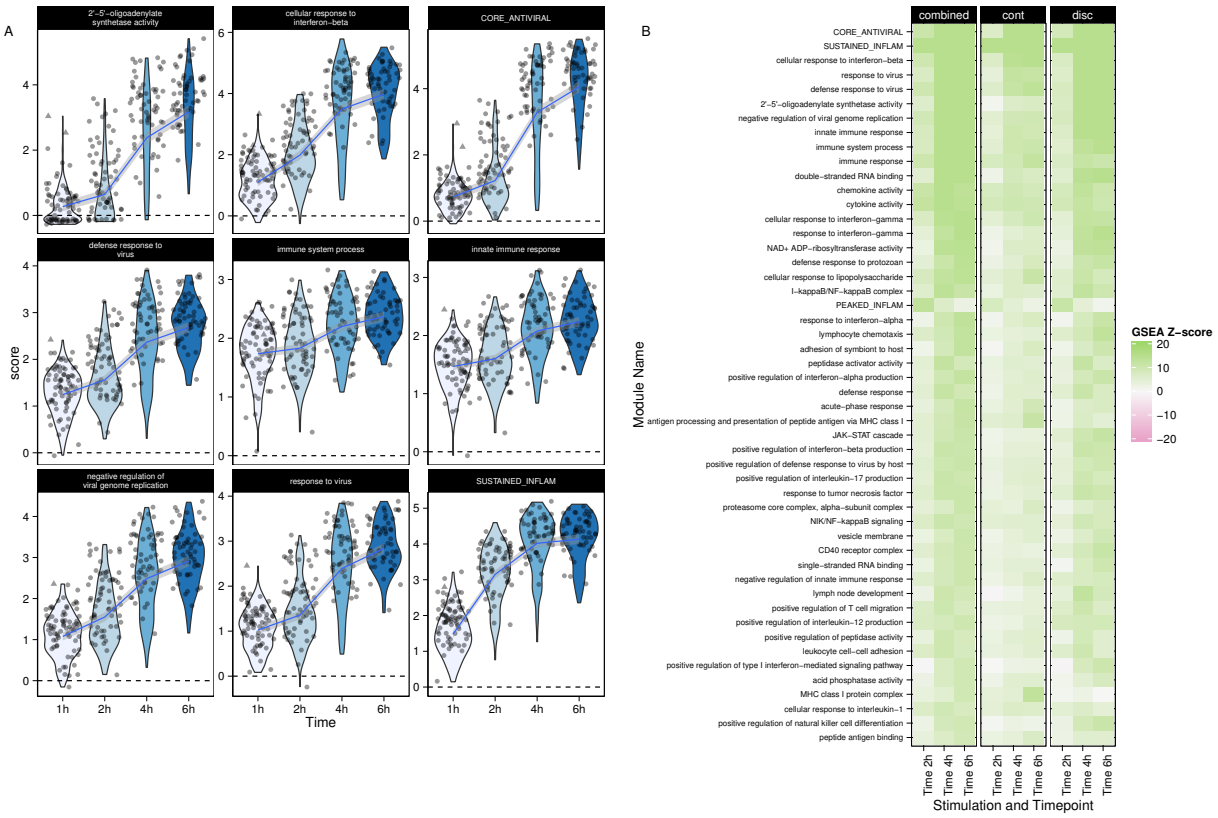

Figure 3.4: Module scores (A) and decomposed Z-scores (B) for single-cell gene set enrichment analysis for LPS stimulated cells, mDC data set, using the mouse GO biological process database. The change in single-cell module scores over time for the nine most significantly enriched modules in response to LPS stimulation are shown in A. The *core antiviral*, *peaked inflammatory* and *sustained inflammatory* modules are among the top enriched modules, consistent with the original publication. Additionally we identify GO modules *cellular response to interferon-beta* and *response to virus*, which behave analogously to the core antiviral and sustained inflammatory modules. No GO analog for the *peaked inflammatory* module was detected. The majority of modules detected exhibit enrichment relative to the 1h time point (thus increasing with time). The "early marcher" cells identified in the original publication are highlighted here with triangles. We show the top 50 most significant modules (B). The combined Z-score summarizes the changes in the discrete and continuous components of expression.

Deviance Residuals

For a given gene, and model component (discrete or continuous) the residual deviance D is -2 times the maximized log-likelihood, (centered so that $D = 0$ when every observation has its own mean parameter). The deviance can be written as a sum of -2 times the log-likelihood of each observation, or

$$D = \sum_{i=1}^{N} d_i.$$

The deviance residual is defined as

$$r_i = \text{sign}(y_i - \hat{y}_i)\sqrt{d_i},$$

and the standardized deviance residual given by $r'_d = \frac{r_d}{\sqrt{1-h_i}}$, where h_i is the leverage associated with observation i.

The combined deviance residual for cell i is the average of the standardized discrete and continuous deviance residuals for the cell if both are present, otherwise it is only the standardized discrete residual.

3.5.1 Identifying co-expressed genes implicated in MAIT cell activation.

We observe co-expression in the residuals from stimulated cells that is not evident in the non-stimulated group (Figure A.14A,B). Since the residuals have removed any marginal changes due to stimulation in each gene, the average residual in the two groups is comparable. The co-expression observed, meanwhile, is due to individual cells expressing these genes *dependently*, where pairs of genes appear together more often than expected under a model of independent expression.

Two clusters of co-expressed genes stand out in the residuals of the stimulated cells (Figure 3.5). These clusters show coordinated, early up-regulation of GZMB and IFN-γ in response to stimulation in MAIT cells and a concomitant decrease in CD69 expression, an early and transient activation marker. PCA of the model residuals highlights the non-responsive stimulated MAIT cells.

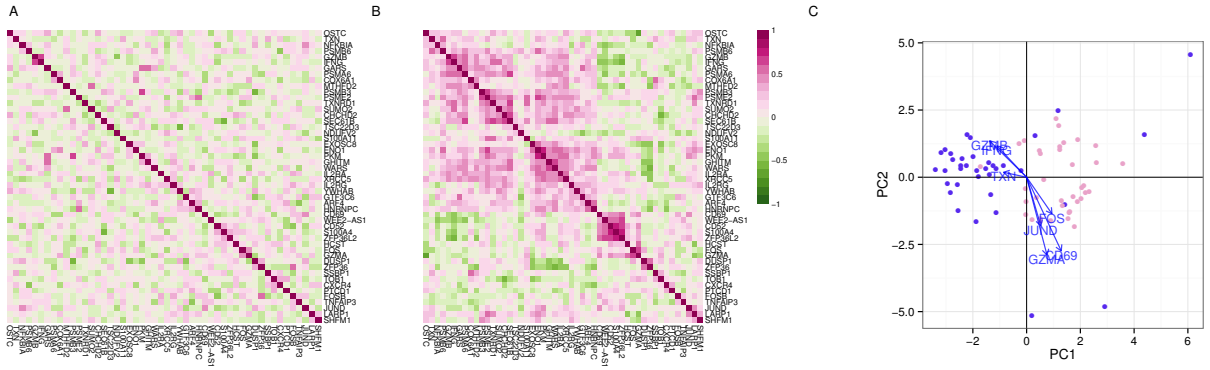

Figure 3.5: Gene-gene correlation (Pearson's ρ) of model residuals in non-stimulated (A) and stimulated (B) cells, and principal components analysis biplot of model residuals (C) on both populations using the top 50 marginally differentially expressed genes. As marginal changes in the genes attributable to stimulation and CDR have been removed, clustering of subpopulations in (C) indicates co-expression of the indicated genes on a cellular basis.

Accounting for the CDR reduces the background correlation observed between genes. When the CDR is included in the model, the number of differentially expressed genes with significant correlations across cells (FDR adjusted p-value $< 1\%$) decreases from 73 to 61 in the stimulated cells, and from 808 to 15 in non-stimulated cells. This shows that adjusting for CDR is also important for co-expression analyses as it reduces background co-expression attributable to cell volume, which otherwise results in dense, un-interpretable patterns of correlation.

MAIT non-responding stimulated cells

The hurdle model expression residuals identify six MAIT cells that do not have a typical activated expression profile in response to stimulation (Figures 3.2 and 3.3). The proportion of these cells detected in the single cell RNA sequencing (scRNAseq) experiment is consistent with what was detected in the flow cytometry experiment. The cells exhibit lower levels of

IFN-γ and GZMB than activated cells (Figure A.14), but also exhibit decreased expression of AP-1 component genes Fos and FosB, consistent with other stimulated cells (Figure A.14B). They do not produce IFN-γ or GZMB upon to cytokine stimulation and exhibit expression profiles intermediate to non-stimulated and stimulated cells (Figure A.14C).

3.5.2 Residual analysis of mouse bone marrow-derived dendritic cells identifies sets of co-expressed genes.

We also explored stimulation-driven correlation patterns. Principal component analysis (Figure 3.6A) of the model residuals demonstrates a clear time trend associated with PC1, as cells increase co-expression of interferon-activated genes. After removing the marginal stimulation and adjusting for the CDR, we observe correlation between chemokines CCL5, TNF receptor CD40, and interferon-inducible (IFIT) genes (Figure 3.6B). A principal finding of the original publication was the identification of a subset of cells that exhibited an early temporal response to LPS stimulation. Recapitulating the original results here, when we examine the PCA of the residuals using the genes in the core antiviral module, we can identify the "early marcher" cells at the 1h time-point (Figure A.15). The co-expression plot for other stimulations can be found in the supplementary material (Figures A.16-A.17).

3.6 Data sets and biological protocols

3.6.1 MAIT cell isolation and stimulation

Cryopreserved PBMC were thawed and stained with Aqua Live/Dead Fixable Dead Cell Stain and the following antibodies: CD3, CD8, CD4, CD161, Vα7.2, CD56 and CD16. $CD8^+$ MAIT cells were sorted as live $CD3^+CD8^+$ $CD4^-CD161^{hi}V\alpha7.2^+$ cells and purity was confirmed by post-sort FACS analysis. Sorted MAIT cells were divided into aliquots and immediately processed on a C1 Fluidigm machine or treated with a combination of IL-12 (eBioscience), IL-15 (eBioscience), and IL-18 (MBL) at 100ng/mL for 24 hours followed by C1 processing.

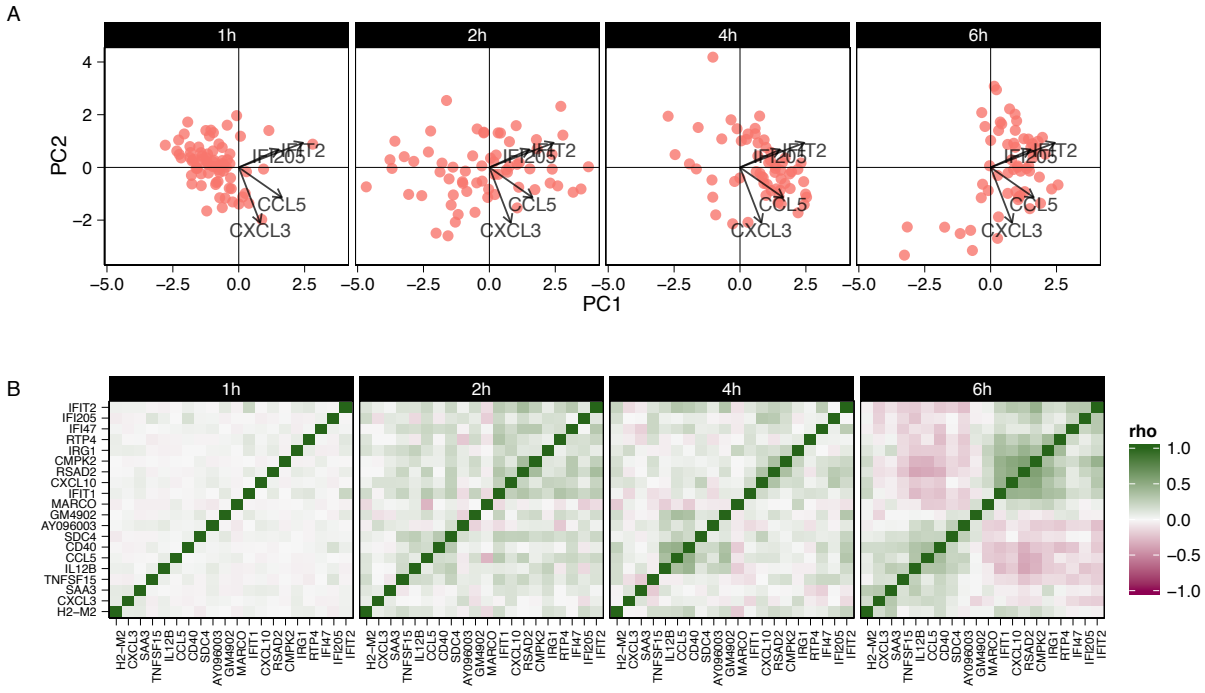

Figure 3.6: Principal components analysis biplot of model residuals (A) and Gene-gene correlation (Pearson's R) of model residuals (B) by time point for LPS cells, mDC experiment using 20 genes with largest log-fold changes, given significant (FDR q <.01) marginal changes in expression. PC1 is correlated with change over time. The two "early marcher" cells are highlighted by an asterisk at the 1h time-point. Correlation structure in the residuals is increasingly evident over time and can be clearly observed at the 6h time-point compared to the earlier time-points.

Data for the MAIT study were derived from a single donor who provided written informed consent for immune response exploratory analyses. The study was approved by the relevant institutional review boards.

C1 processing, Sequencing, and Alignment

After flow sorting, single cells were captured on the Fluidigm™ C1 Single-Cell Auto Prep System (C1), lysed on chip and subjected to reverse transcription and cDNA amplification using the SMARTer® Ultra™ Low Input RNA Kit for C1 System (Clontech). Sequencing libraries were prepared using the Nextera XT DNA Library Preparation Kit (Illumina) according to C1 protocols (Fluidigm). Barcoded libraries were pooled and quantified using a Qubit® Fluorometer (Life Technologies). Single-read sequencing of the pooled libraries was carried out either on a HiScanSQ or a HiSeq2500 sequencer (Illumina) with 100-base reads, using TruSeq v3 Cluster and SBS kits (Illumina) with a target depth of >2.5M reads. Sequences were aligned to the UCSC Human genome assembly version 19 and gene expression levels quantified using RSEM [Li and Dewey, 2011] and TPM values were loaded into R [Gentleman et al., 2004] for analyses. See supplement for more details on data processing procedures.

Our thresholding approach does not adversely affect detection of differentially expressed genes (Figure A.1) and serves to make the continuous expression (Et > 0) more Normal.

3.6.2 Time-series stimulation of mouse bone-marrow derived dendritic cells (mDC)

Processed RNA-seq data (transcripts-per-million, TPM) were downloaded from GEO under accession number GSE41265. Alignment, pre-processing and filtering steps have been previously described [Shalek et al., 2014]. Low quality cells were filtered as described.

3.6.3 Availability of Supporting Data

MAST is available as an R package (`http://www.github.com/RGLab/MAST`, doi: 10.5281/zenodo.18539), released under the GPL license. All data and results presented in this chapter– including code to reproduce the results – are available at: (`http://github.com/RGLab/MASTdata/archive/v1.0.1.tar.gz`, doi: 10.5281/zenodo.19041). Raw data files have been submitted to NCBI's sequence read archive under project accession SRP059458.

3.7 Discussion

We have presented MAST, a flexible statistical framework for the analysis of scRNA-seq data. MAST is suitable for supervised analyses about differential expression of genes and gene-modules, as well as unsupervised analyses of model residuals, to generate hypotheses regarding co-expression of genes. MAST accounts for the bimodality of single-cell data by jointly modeling rates of expression (discrete) and positive mean expression (continuous) values. Information from the discrete and continuous parts is combined to perform inference about changes in expression levels using gene or gene-set based statistics. Because our approach uses a generalized linear framework, it can be used to jointly estimate nuisance variation from biological and technical sources, as well as biological effects of interest. In particular, we have shown that it is important to control for the proportion of genes detected in each cell, which we refer to as the cellular detection rate (CDR), as this factor can single-handedly explain 13% of the variability in the 90% percentile gene. Adjusting for CDR at least partially controls for differences in abundance due to cell size and other extrinsic biological and technical effects. Using several scRNA-seq datasets, we showed that our approach provides a statistically rigorous improvement to methods proposed by other groups in this context [Shalek et al., 2014]. Although MAST has greatest efficiency when the continuous (log)-expression is Normally distributed transformations (such as the Box-Cox) could also be applied if the non-zero continuous measurements are skewed.

As discussed by Padovan-Merhar et al. [2015], care must be taken when interpreting

experiments where the system shows global changes in CDR across treatment groups. The question is essentially ontological: is the CDR a mediator of the treatment effect (is it caused by the treatment and intermediate to expression of the gene of interest), or does it confound the treatment effect (does it happen to co-occur with treatment). Regardless, the CDR-adjusted treatment estimates are interpreted as the change in expression due to treatment, if CDR were held constant between the two conditions.

Two other alternative uses of the CDR are of note. It is also possible to use CDR as a precision variable (an uncorrelated secondary cause) by centering the CDR within each treatment groups, which makes the CDR measurement orthogonal to treatment. This would implicitly assume that the observed changes are treatment induced, while still modeling the heterogeneity in cell volume within each treatment group. An alternative approach would be to estimate the CDR coefficient using a set of control genes assumed to be treatment invariant, such as housekeeping or ERCC spike-ins [Brennecke et al., 2013, Buettner et al., 2015] and including it as an offset to the linear predictors in the regression. As noted by Hicks et al. [2015], the optimal approach to handle confounding between technical and biological effects on the CDR is to design experiments with biological replicates across multiple batches. Finally, we note that while the methodology presented here was developed using scRNA-seq data sets, it appears applicable to other single-cell gene expression platforms where bimodal, conditionally Normal expression patterns are seen such as single cell RNA seq with unique molecular identifiers.

Chapter 4

GRAPHICAL MODELS FOR ZERO-INFLATED SINGLE CELL GENE EXPRESSION

4.1 Background

Graphical models have synthesized high-dimensional biological experiments into understandable, canonical forms [Dobra et al., 2004, Markowetz and Spang, 2007]. Although inferring causal relationships between genes is perhaps the ultimate goal of such analysis, causal models are difficult to estimate with observational data, while experimental manipulation of specific genes has remained costly, and largely inimitable to high-throughput biology. On the other hand, undirected models parametrize the conditional independences present between variables with a graph consisting of a set of vertices $\mathcal{V}$ and a set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. These Markov random fields are estimable on *iid* observational data, and provide descriptions of the statistical predictors of each gene. Each gene is optimally predicted using only its neighbors.

Improved descriptions of conditional dependence in gene expression experiments would help answer a variety of scientific questions. They might provide new insights on–or at least falsify–models of gene regulation, since statistical dependence is expected, given causal dependence. In immunology, polyfunctional immune cells, which simultaneously express multiple cytokines, have been identified as useful predictors of vaccine response [Precopio et al., 2007]. Simultaneous expression or *co-expression* of cellular surface markers has been used to define cellular phenotypes [Lin et al., 2015], so expanding the "dictionary" of co-expression may allow phenotypic refinement.

Several different approaches have been described to estimate the structure of parametric, undirected graphical models. Fully parametric joint models have assumed a Gaussian distribution [Yuan and Lin, 2007], or that the marginal distributions are monotone transformations

thereof [Liu et al., 2009]. Pseudo-likelihood based models, which posit only the conditional distribution of each gene without necessarily guaranteeing joint compatibility [Arnold and Press, 1989, Meinshausen and Bühlmann, 2006, Chen et al., 2015] are more flexible, and have also allowed high-dimensional generalized additive models [Voorman et al., 2014], where the conditional expectation is a smooth, additive function. Many of these methods have seen profitable application to gene expression experiments on bulk aggregates of cells assayed through microarrays or RNA sequencing.

Microfluidic and molecular barcoding advances have enabled the measurement of the minute quantities of mRNA present in single cells. A characteristic of single cell expression is zero-inflation of otherwise continuous measurements, in which measurements are either strongly positive, or undetectable. These experiments have the potential to provide unique resolution of gene co-expression, but the distributions are inadequately modeled with a Gaussian distribution, and in any case the properties of the zero-inflation may be of intrinsic interest [Kim and Marioni, 2013]. To accommodate these features, we propose a joint probability density function $f(\mathbf{y})$ of the form

$$\log f(\mathbf{y}) = \mathbf{v_y}^T \mathbf{G} \mathbf{v_y} + \mathbf{v_y}^T \mathbf{H} \mathbf{y} - \frac{1}{2} \mathbf{y}^T \mathbf{K} \mathbf{y} - C(\mathbf{G}, \mathbf{H}, \mathbf{K}), \tag{4.1}$$

in which both binary and continuous versions of gene expression are sufficient statistics, and interactions thereof are parametrized. Here $\mathbf{y}$ represents an m-vector of gene expression, $[\mathbf{v_y}]_i = I_{y_i \neq 0}$ is the element-wise non-zero indicator function, and $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{K}$ are matrices of interaction parameters. This model can be shown equivalent to a finite mixture model of singular Gaussian distributions. The neighborhood of each gene can be estimated using an **anisometric** group-ℓ_1 penalized conditional likelihood of the form

$$\text{argmax}_\theta \log f_{[b|A]}(y; \theta) - \lambda \sum_{a \in A} \sqrt{\theta_a^T \mathbf{H}_{aa} \theta_a},$$

where $\log f_{[b|A]}(y; \theta)$ is the conditional likelihood of $y_b | \mathbf{y}_A$, θ is the concatenation of rows and columns of interaction matrices, and $\lambda \geq 0$ is a tuning parameter. Typically the group-ℓ_1 penalty [Yuan and Lin, 2006] takes $\mathbf{H}_{aa} = I$. We propose to use the observed Fisher information in block a under a null model $\theta_a = 0$ for all $a \in A$.

Section 2 of this chapter discusses the parameter targeted in single cell gene expression experiments, and why it is not accessible from traditional gene expression experiments. Section 3 describes the parametric Hurdle model for single cell gene expression and estimation of graphical models using neighborhood selection via penalized regression. Section 4 provides a simulation study. Section 5 illustrates the method in a data set in which selected gene profiles were available for both single- and several-cell aggregates, while section 6 applies the method to a high-dimensional data set. The proposed method yields substantial improvements in simulation, and uncovers distinct networks compared to existing approaches.

4.2 From single cells to cellular co-expression

A typical cell contains 1-50 picograms of total RNA, of which perhaps 5% is assayable messenger RNA encoding for proteins (the remainder is structural tRNA and rRNA) [Livesey, 2003]. Protocols for bulk gene expression experiments, such as for Illumina TrueSeq, may call for 100 nanograms of total mRNA, hence require the equivalent of 80,000 cells' worth of mRNA. On the one hand, this biological "summation" over thousands of cells is expected to yield sharper inference on the mean expression level of each gene. However, this comes at the cost of distorting the conditional dependences present between genes.

Consider $\mathbf{Y}_i$, an *iid* sequence of random vectors in $\mathbb{R}^p$ representing the copy number of p transcripts present in single cells $i = 1, \ldots, n$. Now suppose the n cells are aggregated and the total expression is measured using some linear quantification that reports values proportional to the input counts of mRNA. Then the sum of expression is observed in *bulk* experiments is

$$\mathbf{Z} = \sum_{i}^{n} \mathbf{Y}_i.$$

Although most bulk experiments are designed to test for differences in mean expression due to experimental treatments and lack extensive replication within a condition, *stochastic profiling* [Janes et al., 2010] experiments have provided *iid* replicates of $\mathbf{Z}$ to suitable for estimating higher order moments. But when the distribution of $\mathbf{Y}_i$ obeys some conditional

independence relationships, in general the distribution of $\mathbf{Z}$ does not obey these same relationships. For example, take $p = 3$ and suppose that $\mathbf{Y}_i$ are *iid* samples from a tri-variate distribution $[Y_1, Y_2, Y_3]$ on $\{0, 1\}^3$. Suppose the probability mass function (PMF) factors as the chain graph $p(y_1, y_2, y_3) = p(y_1)p(y_2|y_1)p(y_3|y_2)$. Then $p(y_3, y_2|y_1) = p(y_2|y_1)p(y_3|y_2)$, which is equivalent to saying that each 2×2 probability table $p(y_3, y_2|y_1 = j)$, $j = 0, 1$ has non-negative rank one. Yet even summing over $n = 2$ cells, the PMF of $\mathbf{Z} = \mathbf{Y_1} + \mathbf{Y_2}$ will not generally factor as such, as one may exhibit a 3×3 probability table for $p(z_3, z_2|z_1)$ that is not a non-negative rank one matrix.

The infamous case in which graphical structure commutes under convolution is when the $\mathbf{Y}_i$ are multivariate Normal. But as argued in the next section, single cell gene expression is generally bimodal and zero-inflated, so not plausibly described by a multivariate Normal distribution. Even though for large enough n the distribution of the bulk experiment $\mathbf{Z}$ might approach multivariate (log-)normality, the networks estimated in bulk data will not reflect the conditional independences that hold in single cell data.

In the limit of n large, $\mathrm{Cov}(\mathbf{Z})$ converges to the population covariance of $\mathrm{Cov}(\mathbf{Y})$. In some models, and some graphs, such as tree-structured Ising models, the independence structure of $\mathbf{Y}$ can be at least partially identified from $\mathrm{Cov}(\mathbf{Y})^{-1}$, but generally there is no easy connection between $\mathrm{Cov}(\mathbf{Y})$ and the conditional independence structure of $\mathbf{Y}$ [Loh and Wainwright, 2013].

4.2.1 Single cell expression

A distinctive feature of single cell gene expression data–across methods and platforms–is the bimodality of expression values [McDavid et al., 2013, Finak et al., 2015, Marinov et al., 2014, Shalek et al., 2014]. Genes can be on (and a positive expression measure is recorded) or off–or below a limit of detection–and the recorded expression is zero or negligible. The cause of the distributional bimodality remains unresolved. It has been argued that it represents censoring of expression below a substantial limit of detection, yet comparison of *in silico* signal summation from many single cells, to the signal measured in biological sums of cells

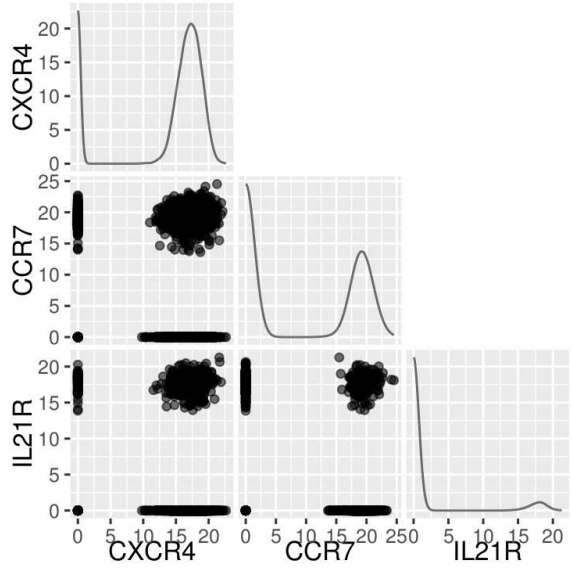

Figure 4.1: Scatter plots of inverse cycle threshold ($40 - \text{Ct}$) measurements from a quantitative PCR (qPCR)-based single cell gene expression experiment. The cycle threshold (Ct) is the PCR cycle at which a predefined fluorescence threshold is crossed, so a larger inverse cycle threshold corresponds to greater log-expression [McDavid et al., 2013]. Measurements that failed to cross the threshold after 40 cycles are coded as 0.

suggest that the limit of detection is essentially zero.

Moreover, the empirical distribution of the log-transformed counts appears rather different, than would be expected from censoring: the distribution of the log-transformed, positive values is generally symmetric. Yet the presence of bimodality in technically replicated experiments ("Pool/split" experiments) implicates technical factors as a cause [Marinov et al., 2014].

4.2.2 Markov graphs

An Markov random field encodes the conditional independence between components of a random vector $\mathbf{Y}$ through a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E} \subseteq \mathcal{V} \times \mathcal{V})$. The set $\mathcal{V}$ contains vertices indexing $\mathbf{Y}$, while $\mathcal{E}$ is the edge set. This graph is undirected in the sense that if $(a, b) \in \mathcal{E}$ then so is (b, a). Associated with $\mathcal{G}$ is a probability distribution $P(\mathbf{Y})$. The graph specifies minimal forms of conditional independence on $P(\mathbf{Y})$.

Of primary use in this chapter will be when $P(\mathbf{Y})$ is compatible with $\mathcal{G}$ under the **local Markov property**. A distribution $P(\mathbf{Y})$ satisfies this property with respect to $\mathcal{G}$ when for all $a \in \mathcal{V}$, $Y_a \perp Y_{\mathcal{V}\setminus \mathrm{cl}(a)} | Y_{\mathrm{ne}(a)}$, where $\mathrm{ne}(a)$ denotes the neighborhood (adjacent vertices) of a and $\mathrm{cl}(a) = \mathrm{ne}(a) \cup a$ is the closure. When $P(\mathbf{Y})$ has a positive and continuous density f with respect to some product measure, then the local Markov property is equivalent to the pairwise and global Markov properties. The Hammersley-Clifford theorem then states that any density–and only these densities–that are compatible with $\mathcal{G}$ will factor as a product of potential functions that depend only on the cliques in $\mathcal{G}$; see for example Lauritzen [1996, ch. 3].

The above implies that given some family that can parametrize all conditional independences that $\mathcal{G}$ encodes, if for each $a \in \mathcal{V}$, we infer the node-wise local conditional independences $Y_a \perp Y_{\mathcal{V}\setminus \mathrm{cl}(a)} | Y_{\mathrm{ne}(a)}$, this suffices to identify all conditional independences that hold in the joint distribution of $\mathbf{Y}$. Moreover, if S is a separating set of genes a and b in $\mathcal{G}$, then knowing the expression of a can provide no information on the expression of b, given S.

4.3 Hurdle models

Univariate Hurdle models arise as modifications of a density through excision of points in the support, generally at the origin. Suppose F is a measure on $\mathbb{R}$ admitting a density f with respect to some dominating measure λ. Then for any Borel set A, the **zero-modified** Hurdle measure F_0 on F is defined as

$$F_0(A) \equiv pF(A \setminus \{0\})/F(\{0\}^c) + (1-p)\delta_0(A),$$

where δ_0 is a point mass at 0. When λ is Lebesgue measure, as in the case in this work, $F(\{0\}) = 0$. This implies a density with respect to the sum-measure $\lambda + \delta_0$ of

$$f_0(y) = pf(y)v_y + (1-p)(1-v_y),$$

where $v_y = I_{y \neq 0}$ is the indicator function for non-zero values of y.

4.3.1 Hurdle exponential families

When F belongs to an exponential family, F_0 also belongs to an exponential family with modified sufficient statistics. In particular,when $f(y)$ is the Normal density with mean ξ and precision τ^2 then the Hurdle modification at zero has density

$$f_0(y) = \exp\left\{v_y\left[1/2\log\left(\tau^2/(2\pi)\right) + \log p/(1-p) - \xi^2\tau^2/2\right] + y\xi\tau^2 - y^2\tau^2/2 + \log(1-p)\right\} \quad (4.2)$$

with respect to the measure $\lambda + \delta_0$. This implies that f_0 is a member of an exponential family with sufficient statistics $v_y, y, -y^2/2$ and natural parameters

$$\begin{aligned} g &= 1/2\log\left(\tau^2/(2\pi)\right) + \log p/(1-p) - \xi^2\tau^2/2, \\ h &= \xi\tau^2, \\ k &= \tau^2, \end{aligned}$$

or inversely the original parameters in terms of the natural parameters are:

$$\begin{aligned}(\operatorname{Var} Y|v_y)^{-1} = &\quad \tau^2 = k,\\ \operatorname{E} Y|v_y = &\quad \xi = h/k,\\ \operatorname{logit} \operatorname{E} v_y = &\quad \log p/(1-p) = g - 1/2 \log(k/2\pi) + h^2/(2k).\end{aligned} \tag{4.3}$$

4.3.2 Multivariate Hurdle models

Based on figure 4.1, a plausible multivariate model puts positive mass on every one of the 2^m coordinate subspaces, including the origin and the entire Euclidean space $\mathbb{R}^m$. It is easiest to construct this model conditionally by first defining $\mathbf{V} = [V_1, \ldots, V_m]^T \equiv [I_{y_1 \neq 0}, \ldots, I_{y_m \neq 0}]^T$ to be the vector of non-zero indicator functions on $\mathbf{Y}$, where the dependence on $\mathbf{Y}$ will be suppressed from here on. Under *any* distribution on $\mathbf{Y}$, $\mathbf{V}$ is most generally a collection of Bernoulli variables distributed according to some 2^m probability table.

A random vector $\mathbf{Y}$ has singular Normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ [Rao, 1973] with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ with rank $r < m$ if $\mathbf{U}$ is a $m \times m - r$ matrix such that $\mathbf{U}^T\boldsymbol{\Sigma} = 0$ and the following holds: a) $\mathbf{U}^T\mathbf{Y} = \mathbf{U}^T\boldsymbol{\mu}$ almost everywhere; and b) $\mathbf{Y}$ has a density $\frac{(2\pi)^{-r/2}}{(\det{}^+ \boldsymbol{\Sigma})^{1/2}} \exp\{-(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^-(\mathbf{y}-\boldsymbol{\mu})/2\}$, restricted to the hyperplane $\mathbf{U}^T\mathbf{Y} = \mathbf{U}^T\boldsymbol{\mu}$. Here $\det^+$ is the pseudo-determinant (product of non-zero eigenvalues) and $\boldsymbol{\Sigma}^-$ is a pseudo-inverse, such as the Moore-Penrose inverse. In the case that $\boldsymbol{\Sigma}$ is zero outside a positive-definite submatrix of size $r \times r$, $\mathbf{U}$ can be chosen to be a diagonal selection matrix consisting of zeros and ones, and $\mathbf{Y}$ has a density with respect to the measure $\lambda^r \times \delta_0^{m-r}$, which is the case treated subsequently.

In detail, given $\mathbf{V}$, let $r = |\mathbf{V}|$ and $\mathcal{I} \equiv \operatorname{diag} \mathbf{V}$ be the diagonal matrix with (i,i) entry equal to V_i, thus selecting the the non-zero entries of $\mathbf{V}$, then

$$\mathbf{Y}|\mathbf{V} = \mathbf{v} \sim \mathcal{N}\left(\mathcal{I}\boldsymbol{\mu}(\mathbf{v}), \mathcal{I}\boldsymbol{\Sigma}(\mathbf{v})\mathcal{I}\right),$$

where $\boldsymbol{\mu}(\mathbf{v})$ and $\boldsymbol{\Sigma}(\mathbf{v})$ depend on $\mathbf{v}$ arbitrarily, but are only identifiable along the obvious subspaces. This is the multivariate analog of excision of mass that defines the univariate

Hurdle model. This construction is also equivalent to $\mathbf{Y}|\mathbf{V}$ being distributed as

$$\mathbf{Y}|\mathbf{V} = \mathbf{v} \sim \mathcal{N}\left(\mathbf{K}_{\mathbf{v}}^{-}\mathbf{h}_{\mathbf{v}}, \mathbf{K}_{\mathbf{v}}^{-}\right), \tag{4.4}$$

where $\mathbf{K}_{\mathbf{v}}^{-} = (\mathcal{I}\mathbf{K}(\mathbf{v})\mathcal{I})^{-}$ is the pseudo-inverse of a precision matrix with arbitrary dependence on $\mathbf{v}$ subject to the constraint that its rows and columns are filled with zeros whenever a coordinate of $\mathbf{v}$ is zero, and $\mathbf{h}_{\mathbf{v}} = \mathcal{I}\mathbf{h}(\mathbf{v})$ is a column vector that may depend on $\mathbf{v}$ arbitrarily after being suitably zeroed out with its selection matrix.

The joint density $P(\mathbf{Y}, \mathbf{v}(\mathbf{Y})) = P(\mathbf{Y})$ can be decomposed as $P(\mathbf{Y}|\mathbf{V})P(\mathbf{V})$:

$$\begin{aligned} p(\mathbf{y}) &= p(\mathbf{v}) \times (2\pi)^{-r/2}\left(\det{}^{+}\mathbf{K}_{\mathbf{v}}\right)^{1/2} \exp\left\{-\left(\mathbf{y}-\mathbf{h}_{\mathbf{v}}\right)^{T}\mathbf{K}_{\mathbf{v}}\left(\mathbf{y}-\mathbf{h}_{\mathbf{v}}\right)/2\right\} \\ &= p(\mathbf{v}) \times (2\pi)^{-r/2}\left(\det{}^{+}\mathbf{K}_{\mathbf{v}}\right)^{1/2} \exp\left\{-\mathbf{h}_{\mathbf{v}}^{\mathbf{T}}\mathbf{K}_{\mathbf{v}}\mathbf{h}_{\mathbf{v}}/2\right\} \\ &\quad \times \exp\left\{-\frac{1}{2}\mathbf{y}^{T}\mathbf{K}_{\mathbf{v}}\mathbf{y} + \mathbf{h}_{\mathbf{v}}^{\mathbf{T}}\mathbf{y}\right\} \\ &= \exp\left\{g(\mathbf{v}) - \mathbf{y}^{T}\mathbf{K}_{\mathbf{v}}\mathbf{y}/2 + \mathbf{h}_{\mathbf{v}}^{\mathbf{T}}\mathbf{y}\right\}, \end{aligned} \tag{4.5}$$

with respect to the m-fold product of $\lambda + \delta_0$, the sum of Lebesgue measure and point mass at 0.

As arbitrary functions of $\mathbf{v}$, $g(\mathbf{v})$, $\mathbf{h}_{\mathbf{v}}$ and $\mathbf{K}_{\mathbf{v}}$ are generically $k \leq m$-order polynomials in $\mathbf{v}$, but for the sake of parsimony one may restrict k strictly less than m. When the polynomial order is 2, 1, 0 for $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{K}$ respectively, then the quadratic statistics $\mathbf{y}\mathbf{y}^{T}$, $\mathbf{v}\mathbf{y}^{T}$, $\mathbf{v}\mathbf{v}^{T}$ are minimally sufficient and the model simplifies to

$$p(\mathbf{y}) = \exp\left\{\mathbf{v}^{T}\mathbf{G}\mathbf{v} + \mathbf{v}^{T}\mathbf{H}\mathbf{y} - \frac{1}{2}\mathbf{y}^{T}\mathbf{K}\mathbf{y} - C\right\}, \tag{4.6}$$

where $C = \log\sum_{\mathbf{v}\in\{0,1\}^m}\exp\left[\mathbf{v}^{T}\mathbf{G}\mathbf{v} + \mathbf{h}^{T}\left(\mathcal{I}\mathbf{K}\mathcal{I}\right)^{-}\mathbf{h}/2\right]\left(\det{}^{+}\mathcal{I}\mathbf{K}\mathcal{I}\right)^{1/2}(2\pi)^{-r/2}$. It follows immediately that model (4.6) is a mixture of singular Gaussian distributions:

Proposition 1. *Model (4.6) is a 2^m mixture of Normal distributions, each singular along a different coordinate axes. Given $\mathbf{v}$, the conditional expectation is $\left(\mathcal{I}\mathbf{K}\mathcal{I}\right)^{-}\mathbf{H}\mathbf{v}$ and the conditional variance is $\left(\mathcal{I}\mathbf{K}\mathcal{I}\right)^{-}$.*

4.3.3 Conditional distributions identify interaction parameters

The normalizing constant in equation (4.6) is a difficult sum of 2^m terms, as is true in the pairwise Ising model. However, the conditional likelihood of each coordinate $y_b|\mathbf{y}_A$ is tractable, and identifies parameters from a given row/column of the interaction matrices.

This is seen through examination of the conditional distributions. Recalling section 4.2.2, $\mathcal{V}$ denotes the vertex set. Fixing a coordinate b and its complement $A = \mathcal{V} \setminus b$, and making the simplifications $v_i y_i = y_i$ and $v_i^2 = v_i$, the kernel of the distribution in (4.6) as a function of y_b is

$$\log f_{[b|A]}(y) = v_b g_{bb} + 2v_b^T \mathbf{g_{bA}} \mathbf{v}_A + v_b \mathbf{h_{bA}} \mathbf{y}_A + y_b h_{bb} + y_B \mathbf{h_{Ab}}^T \mathbf{v}_A \\ - \frac{1}{2}\left[y_B^2 k_{bb} + 2y_b \mathbf{k_{bA}} \mathbf{y}_A\right] - C_{[b|A]}.$$

Factoring by the sufficient statistics over y_b,

$$\log f_{[b|A]}(y) = v_b \underbrace{\left[g_{bb} + 2\mathbf{g}_{bA}\mathbf{v}_A + h_{bA}\mathbf{y}_A\right]}_{g_{[b|A]}} + y_b^T \underbrace{\left[h_{bb} + \mathbf{h}_{Ab}^T \mathbf{v}_A - \mathbf{k}_{bA}\mathbf{y}_A\right]}_{h_{[b|A]}} \tag{4.7}$$

$$- \frac{1}{2} y_B^2 \underbrace{k_{bb}}_{k_{[b|A]}} - C_{[b|A]}$$

$$= v_b \eta_i^{(1)} + y_b \eta_i^{(2)} - \frac{1}{2} y_b^2 k_{bb} - C_{[b|A]}. \tag{4.8}$$

Thus the conditional distribution is just an example of the univariate Hurdle of equation (4.2) with natural parameters $g_{[b|A]}$, $h_{[b|A]}$ and $k_{[b|A]}$, which also serve as linear predictors that depend on a design matrix constructed from $\mathbf{y}_A$ and $\mathbf{v}_A$.

Concretely, we can write

$$g_{[b|A]} = Z_0^T \theta_{g0} + \sum_{a \in A} X_a \theta_{ga},$$

where Z_0 in this case is taken to be 1, but generally could include a vector of covariates, $X_a = [v_a, y_a]$ and $\theta_{ga} = [g_{ba}, h_{ba}]$. The linear predictor for $h_{[b|A]}$ can be written analogously. We can solve for the natural parameters to yield parameters recognizable from the univariate

case using equation 4.3. Then

$$\begin{aligned}
\tau^2_{b|A]} &= k_{bb}, \\
\xi_{[b|A]} &= \frac{1}{k_{bb}} \left[h_{bb} - \mathbf{k_{Ab}}^T \mathbf{y}_A + \mathbf{h_{Ab}}^T \mathbf{v}_A \right], \\
[\log p/(1-p)]_{[b|A]} &= [g_{bb} + 2\mathbf{g_{bA}} v_a + \mathbf{h_{bA}} \mathbf{y}_A] \\
&\qquad \underbrace{-1/2 \log\left(k_{bb}/2\pi\right) + \left[h_{bb} - \mathbf{k_{Ab}}^T \mathbf{y}_A + \mathbf{h_{Ab}}^T \mathbf{v}_A \right]^2 / \left(2k_{bb}\right)}_{C_{[b|A]}},
\end{aligned}$$

so $\log p/(1-p)$ depends on y_A both through a linear term $2g_{bA}\mathbf{v}_A + h_{bA}\mathbf{y}_A$ as well as a quadratic term (derived from the normalizing constant for the Gaussian). Integrating with respect to $\delta_0 + \lambda$ confirms directly that the normalizing constant is indeed given by $C_{[b|A]}$.

The conditional distribution in (4.7) defines a vector generalized linear model, since the univariate family (4.2) is parametrized by three natural parameters, g, h and k, the first two of which are modeled as a linear function of covariates.

4.3.4 Other work on mixed graphical models

The notation used here follows Lauritzen [1996], who describes conditional Gaussian (CG) models with *inhomogeneous, non-singular* precision $\mathbf{K}(\mathbf{v})$ that can depend on the discrete set of covariates in arbitrary, positive-definite fashion. The formulation here is both a special case, and an extension of Lauritzen's inhomogeneous CG models, since it imposes the inhomogeneity (conditional singularity, in fact) in a structured fashion.

Several authors have described algorithms to infer the structure of specializations of Lauritzen's CG models. Lee and Hastie [2013] and Cheng et al. [2013] describe ℓ_1-penalized, pseudo-likelihood algorithms to estimate structure in specializations of the inhomogeneous, pairwise case (in which the variance of continuous variables, given both discrete and continuous neighbors are homoscedastic).

Other authors have described classes of graphical models with mixed node conditional distributions. Chen et al. [2015], and Yang et al. [2014] describe exponential-family graphical models for which the conditional distributions follow a generalized linear model with

neighbors entering additively. While this chapter was in preparation, Tansey et al. [2015] proposed *vector space graphical models* that include the multivariate Hurdle as a special case, estimated through sparse group-lasso penalized neighborhood selection. Their isometric group-lasso does not account for heterogeneity in the scaling of predictors in the conditional distributions. The anisometric group-lasso proposed in the following section yields substantial benefits in finite samples.

Other work considering dependence measures in mixed distributions include Olkin and Tate [1961], Cox and Wermuth [1992].

4.4 Neighborhood estimation via penalized regression

Equation (4.7) shows that the conditional distributions of each node, given the rest, identify rows and columns of the interaction matrices. If m is fixed and $n \to \infty$, maximum likelihood, methods of moments or Bayesian estimators will concentrate around the true values and hypothesis testing might reveal the neighborhood of a node. Although in single cell experiments, the number of cell replicates, n, is larger than in many bulk mRNA experiments, it is still generally the case that the number of genes, m measures in the thousands, while n measures in the hundreds (though emerging technologies may change this), so estimators and consistency guarantees under fixed m asymptotics are not applicable.

On the other hand, under scenarios in which $n, m \to \infty$ while $n > Cd^{\phi}(\log m)^{\psi}$, where C, ϕ and ψ are constants that depend on the model and d is the maximum vertex degree, penalized regression has been shown to consistently identify signed sparsity patterns in precision matrices in multivariate Normal models [Meinshausen and Bühlmann, 2006], in interaction matrices for Ising (auto-logistic) graphical models [Ravikumar et al., 2010] and exponential family graphical models [Yang et al., 2014]. This motivates the application of node-wise penalized regression for the problem at hand.

4.4.1 Penalty and computation

By inspection of (4.7) for $y_b \perp y_a | y_{\mathcal{V}\setminus\{a,b\}}$ the four parameters $[g_{ba}, h_{ba}, h_{ab}, k_{ba}] = \theta_a$ must simultaneously vanish. Penalizing the conditional log-likelihood $\log f_{[b|A]}(\mathbf{y})$ with the grouped ℓ_1 penalty $P_\lambda(\theta) = \lambda \sum_{a \in A} \sqrt{\theta_a^T \theta_a}$ can lead to an optimum that is sparse in parameter blocks responsible for vertex a. This penalty is equivalent to placing a sequence of independent, multivariate Laplace (multivariate exponential power distribution [Eltoft et al., 2006]) priors on blocks of θ and reporting the MAP. It is well-known that this results in both shrinkage and variable selection.

Viewed as a prior, the default group-lasso penalty implicitly assumes that each variable in each block has a similar effect size. This may be reasonable, provided they are measured in comparable units. For example if covariate X_1 is measured in meters, while covariate X_2 in centimeters, then the distribution of effect sizes for X_2 would be 1000-times more dispersed than the distribution for X_1, revealing a kind of expected scale-equivariance. In penalized GLMs, this is typically enforced "at run time" by ensuring covariates are on comparable scales, or $Z-$scoring each column of the design if no intrinsic scale exists.

In the case of a vector regression, terms from linear predictor $g_{[b|A]}$ and linear predictor $h_{[b|A]}$ end up together in blocks, and these coefficients are not necessarily comparable, as one specifies log-odds of $E(V_b|V_a = 0)$ while the other specifies conditional expectations of $E(Y_a|Y_b)$. Re-scaling is not an option, since the same design matrix $X_a = [V_a, Y_a]$ is used in each linear predictor, and even if it were, only reparametrization through isometric transformations produces the same solution (in terms of fitted values) as has been noted for the group-lasso in linear regression [Simon and Tibshirani, 2012]. But replacing the isometric ℓ_2 norm in the sum so that the penalty is

$$P_{\mathbf{H},\lambda}(\theta) = \lambda \sum_{a \in A} \sqrt{\theta_a^T \mathbf{H}_{aa} \theta_a}, \tag{4.9}$$

where $\mathbf{H} \equiv \mathrm{diag}\,(\mathbf{H}_{aa})$ is a block-diagonal, positive-definite matrix allows terms from the linear predictors to have different scales of penalty. It also accounts for correlation between

components of θ_a, since columns of the design are correlated due to both v_a and y_a appearing as predictors.

If prior information existed, the matrix $\mathbf{H}$ could be chosen accordingly, with interpretation as a multivariate Laplace prior. Absent prior information, setting $\mathbf{H}$ equal to the Fisher information under a null model $\theta_a = 0$ for all a results in variable selection approximately equal to conducting score tests, with exact equivalence holding under a null hypothesis of $\theta_a = 0$ for all a, as is shown in the following proposition:

Proposition 2. *Suppose that the inverse information* $\mathbf{H}^{-1}$ *is also block-diagonal, where* $\mathbf{H} = [\frac{\partial^2 \log f_{[b|A]}(\mathbf{y})}{\partial\theta_i\theta_j}]$ *is the conditional information. This holds, for example, when* $\mathbf{H}$ *is block-diagonal. The scaled* ℓ_1 *penalty is equivalent to a score test of the null hypothesis that* $\theta = 0$ *vs the alternative that a pre-specified subvector* $\theta_a \neq 0$.

Proof: Let $c = \mathcal{V} \setminus \{a, b\}$ and suppose that $\theta_c = 0$. From the KKT conditions, $\theta_a = 0$ is an optimum if and only if

$$\nabla_a^T H_{aa}^{-1} \nabla_a < \lambda^2,$$

where $\nabla_a = \frac{\partial \log f_{[b|A]}(\mathbf{y})}{\partial\theta_a}$ is the a-subvector of the conditional log-likelihood gradient. Taking λ^2 to be an appropriate quantile from a χ^2-distribution with $\dim(H_{aa})$ degrees of freedom results yields a score test.

This leads to algorithm 1. The covariates $\mathbf{Z}$ might just be an intercept column, but generally could be any cell-level covariate deemed relevant, in which case the estimated model has the interpretation of being a *conditional Markov random field*. The optimization step in line 3 is solved using any Newton-like algorithm (eg BFGS) as the object is smooth and concave, while proximal gradient ascent [Parikh and Boyd, 2014] solves line 6, as the objective is a sum of a concave, smooth function and a structured concave function. In particular, we exploit the fact that while the proximal operator

$$\text{prox}_\gamma(x) = \text{argmax}_u \frac{1}{\gamma}\|x - u\|_2^2 + \sum_{a \in A} \sqrt{u_a^T H_{aa} u_a}$$

is not available in the familiar form of a soft-thresholding operator as in the isometric group-lasso, the proximal operator of the *anisometric* group-lasso can be efficiently found via a line search and a few 4×4 matrix multiplications after one-time pre-calculation of the singular value decomposition of H_{aa} [Foygel and Drton, 2010]. Throughout the inner-loop, warm starts are exploited for $\hat{\theta}$ as λ varies. Active set heuristics using the strong rules of Tibshirani et al. [2012] yield computational gains for sparse solutions with large m. In the accompanying software, the algorithm is written in a combination of `R` and `C++`.

Data: Expression matrix $\mathbf{Y} \in \mathbb{R}^{n \times m}$, covariates $\mathbf{Z} \in \mathbb{R}^{n \times q}$, penalty path $\mathbf{\Lambda}$

Result: Neighborhoods $ne(i, \lambda)$, $1 \leq i \leq m$, $\lambda \in \mathbf{\Lambda}$

for $i = 1 \dots m$ **do**
  $\mathbf{X} \leftarrow \left[\mathbf{Z}, V_1, Y_1, \dots, V_{i-1}, Y_{i-1}, V_{i+1}, Y_{i+1}, \dots, V_m, Y_m\right]$;
  $\bar{\theta}_0 \leftarrow \text{argmax}_{\theta_0, \theta=0} \log f([\theta_0, \theta], \mathbf{X})$;
  $\mathbf{H} \leftarrow \text{Hessian}\big(\log f\left([\bar{\theta}_0, 0], \mathbf{X}\right)\big)$;
  for $\lambda \in \mathbf{\Lambda}$ **do**
    $[\hat{\theta}_0, \hat{\theta}] \leftarrow \text{argmax}_{\theta_0, \theta} \log f([\theta_0, \theta], \mathbf{X}) - P_{\lambda, \mathbf{H}}(\theta)$;
    $ne(i, \lambda) \leftarrow \text{ProcessNeighborhood}(\hat{\theta})$;
  end
end

Algorithm 1: Neighborhood selection

4.4.2 Convergence and model selection consistency

Tansey et al. [2015] provide model selection consistency guarantees under assumptions involving both the sample information matrix and the joint log-partition function. The log-partition function in model (4.6) is computationally intractable for even moderate m (although it could be numerically approximated). Thus these assumptions are generally difficult to verify, even in simulation.

Total selection consistency may require astronomically large samples; for example in the course of running these simulations, it fails to occur even once in 30 simulations for $m = 128, n = 10000$. The simulations in Tansey et al. [2015] also do not achieve total selection consistency even for samples of $n = 25000$. However, at realistic sample sizes imperfect recovery, in which some number of false edges are included in a portion of the true set, is feasible, and is explored further in section 4.5. Applying a positive-definite penalty matrix rather than the isometric penalty offers drastically improved rates of imperfect recovery.

4.5 Simulations

We consider a series of simulations under two sets of parametric alternatives. Observations are generated through Gibbs sampling from model (4.6), with 2000 iteration burn-in, 10% down-sampling. Down-sampled iterations exhibited only mild auto-correlation.

In the first study, the dependence structure is "dense:" if any of $g_{ij}, h_{ij}, h_{ji}, k_{ij} \neq 0$, then all of them are non-zero. In the second study, only g_{ij} is non-zero, and the Hurdle model is a superset of the true model, which is Ising/logistic. In both cases the edge density is fixed at 1.5% and the underlying graph is a chain, thus the maximum degree is 2. The number of observations $n = 100$ is fixed and the dimension varies from $m = 16$ to $m = 128$, with 30 replicates run.

Five methods were examined to test graph structure inference: neighborhood selection under the Hurdle model using 1) isometric and 2) anisometric penalties, neighborhood selection using ℓ_1-penalized 3) logistic regression, 4) linear regression and 5) Gaussian copula (NPN) models [Liu et al., 2009]. The logistic model is fit using the R package `glmnet`. Models 4 and 5 were fit using the R package `huge`, with "truncation" transformation function. Neighborhoods are stitched together using an "or" rule, i.e. vertices a and b are adjacent if either $b \in \text{ne}(a)$ or $b \in \text{ne}(a)$.

In each method, as a tuning parameter decreases, more edges are selected. Figure 4.2 shows a representative ROC curve (true positives vs false positives) for $n = 100$ and $m = 64$.

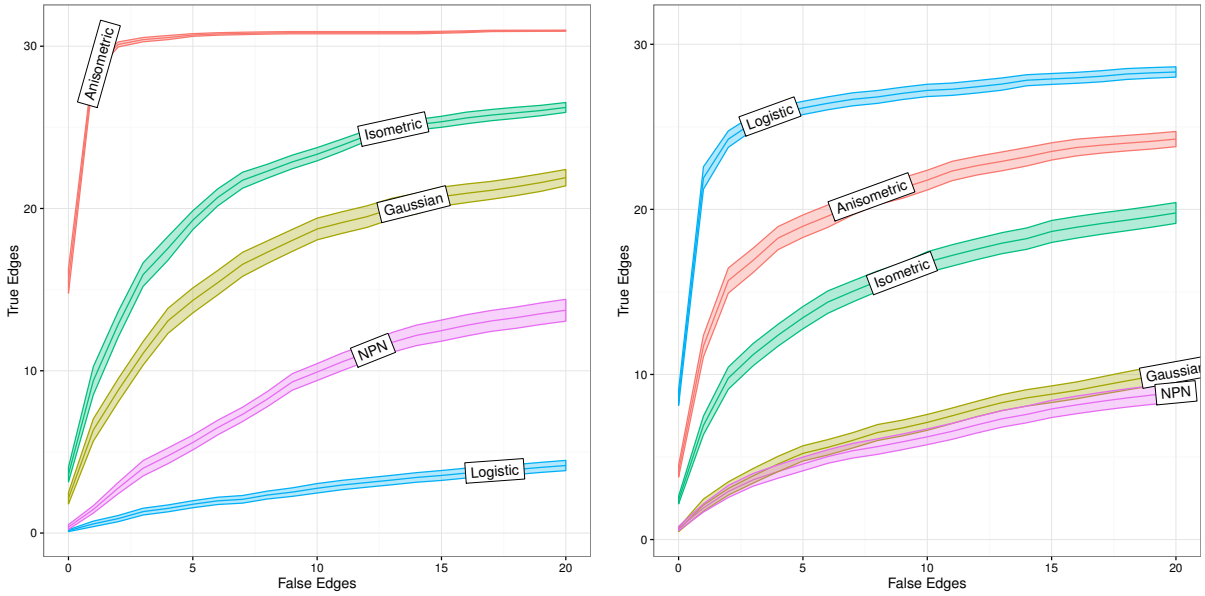

Figure 4.2: Number of true positives vs false positives for simulated chain graphs under *dense* and *sparse dependence* with $m = 64$ nodes and $n = 100$ observations. The ribbon shows the simulation-induced standard errors about the average. The Anisometric and Isometric models use neighborhood selection with the multivariate Hurdle model (4.7) with group-ℓ_1 penalty based on the null-model Fisher information and identity matrix, respectively. The Gaussian, NPN and Logistic models use ℓ_1 penalized neighborhood selection under linear (Gaussian), Normal-score transformed linear (NPN) and logistic regressions.

In figure 4.3, the maximum sensitivity $\left(\frac{\text{true positives}}{\text{total true}}\right)$ under the oracle value of the tuning parameter that admits fewer than 10% false discoveries $\left(\frac{\text{false positives}}{\text{total positives}}\right)$ is shown.

4.6 T follicular helper cells

T follicular helper (Tfh) cells are a class of $CD4^+$ lymphocytes resident in the germinal centers of lymph nodes. B-cells that actively secrete antibodies require Tfh cell co-stimulation to activate these secretion properties [Ma et al., 2012]. HIV is known to induce changes in Tfh cells, increasing numbers of Tfh cells in germinal centers, while reducing numbers of Tfh-

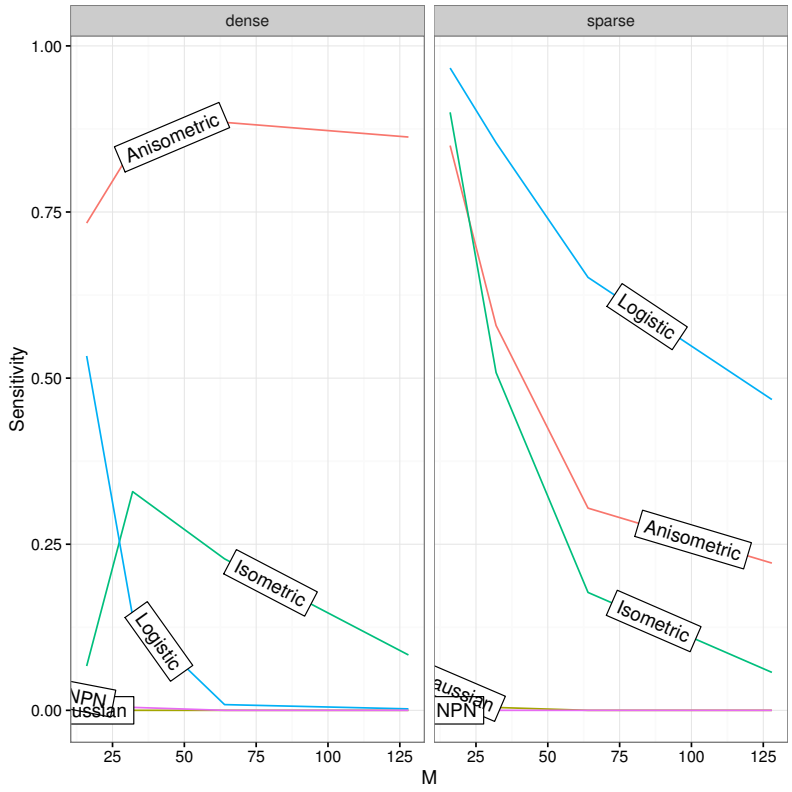

Figure 4.3: Sensitivities $\left(\frac{\text{true positives}}{\text{total true}}\right)$ at 10% FDR as m, the number of nodes increases for a chain graph under fixed sparsity. See the caption of figure 4.2 for a description of the methods.

like cells in peripheral blood [Onabajo and Mattapallil, 2013]. We consider co-expression networks in Tfh cells measured in eight donors with recent HIV diagnosis and naive to anti-retroviral therapy. 65 genes were selected for profiling via qPCR on the basis of their role in Tfh signaling and differentiation, generally with sparse expression across single cells (median expression 28%, 90th percentile, 82%). 573 single cell, and 61 10-cell replicates were taken.

Figure 4.4 shows networks of 35 edges (network size chosen through stability selection, see following) estimated using Hurdle, Gaussian (with centered data, see section 4.9) and Logistic, and Gaussian model using 10-cell aggregates. Normalized Hamming distances between the four methods and the Gaussian model fit on the "raw", uncentered data are reported in table 4.1. The Hurdle and (centered) Gaussian models are most similar, while the logistic and Gaussian 10-cell network are quite distinct. The `Gaussian(raw)` model on untransformed data is similar to the logistic model. The zeros exert substantial leverage on the regression compared to the variation among the non-zero values, so this is expected.

In the Gaussian and Hurdle networks, the gene CD3ϵ, the backbone of T cell receptor(TCR) signal transduction by which T cells receive antigens presented by other non-immune cells, has a strong, positive association with CD4, which is a co-receptor necessary for activation in helper T-cells. It is also strongly connected to IL2Rg, a subunit common to several different interleukin receptors. CXCR4 and CD28 are also highly connected hubs in both graphs. Of particular note, CXCR4 is the receptor by which T-tropic HIV isolates are able to infect $CD4^+$ T-cells, so co-expression of STAT5B/CD95/CXCR4/FYN would tend to make a Tfh cell more susceptible to this path of infection. IL21, a cytokine inducing cell division and proliferation in its ligands, has strong positive association with CTLA4, an inhibitory immune system checkpoint, and NFATC1 (Nuclear factor of activated T-cells), a transcription factor triggered during an immune response. Lck and Fyn, which both play a key role in TCR signaling, interact via CD27, part of the TNF-receptor superfamily.

On the other hand, in the Hurdle model, CXCL13 is negatively associated with CCR4 and CTLA4. In the Gaussian model, the CXCL13-CTLA4 edge has inverted sign, and the CCR4 edge is absent. In fact, no edges indicating negative associations are present in the

	Gaussian(10)	Hurdle	logistic	Gaussian(raw)
Gaussian	1.00	**0.27**	0.95	.87
Gaussian(10)	-	1.00	1.00	1.00
Hurdle	-	-	0.96	.87
logistic	-	-	-	.31

Table 4.1: Dissimilarities $\left(\frac{\text{Hamming Distance}}{\text{Number of edges}}\right)$ between networks of size 35 estimated through various methods. The Gaussian(10) model is a Gaussian model estimated on 10-cell replicates, while the Gaussian(raw) data is estimated on single cells without centering the data. The Hurdle and logistic models are described in the text.

Gaussian model. The logistic is wildly different, with CXCL13 as a hub, and predominately consists of negative connections.

Shah and Samworth [2013] propose a form of antithetic stability selection, and when employed here the richness of the networks varies from 2 (Gaussian10), 17 (Logistic), 26 (Hurdle) and 52 edges (Gaussian). In order to compare networks with equivalent richness, we therefore examine a compromise of 35 edges in the four methods.

4.7 Mouse dendritic cells

In an experiment originally described in Shalek et al. [2014], bone marrow-derived dendritic cells, from *mus musculus* were exposed to lipopolysaccharide (LPS), a toxic compound secreted and structurally utilized by gram-negative bacteria. Cells were sampled after 0, 1, 2, 4, and 6 hours. Here we consider the transcription networks estimated using 4431 transcripts expressed in at least 20% of 65 cells sampled 2 hours after LPS exposure. Rather than attempting to perform model selection on this limited sample size, we consider highly sparse ($< .01\%$ sparsity) networks of 700 edges, chosen to provide tractable visualization and illustration of the method.

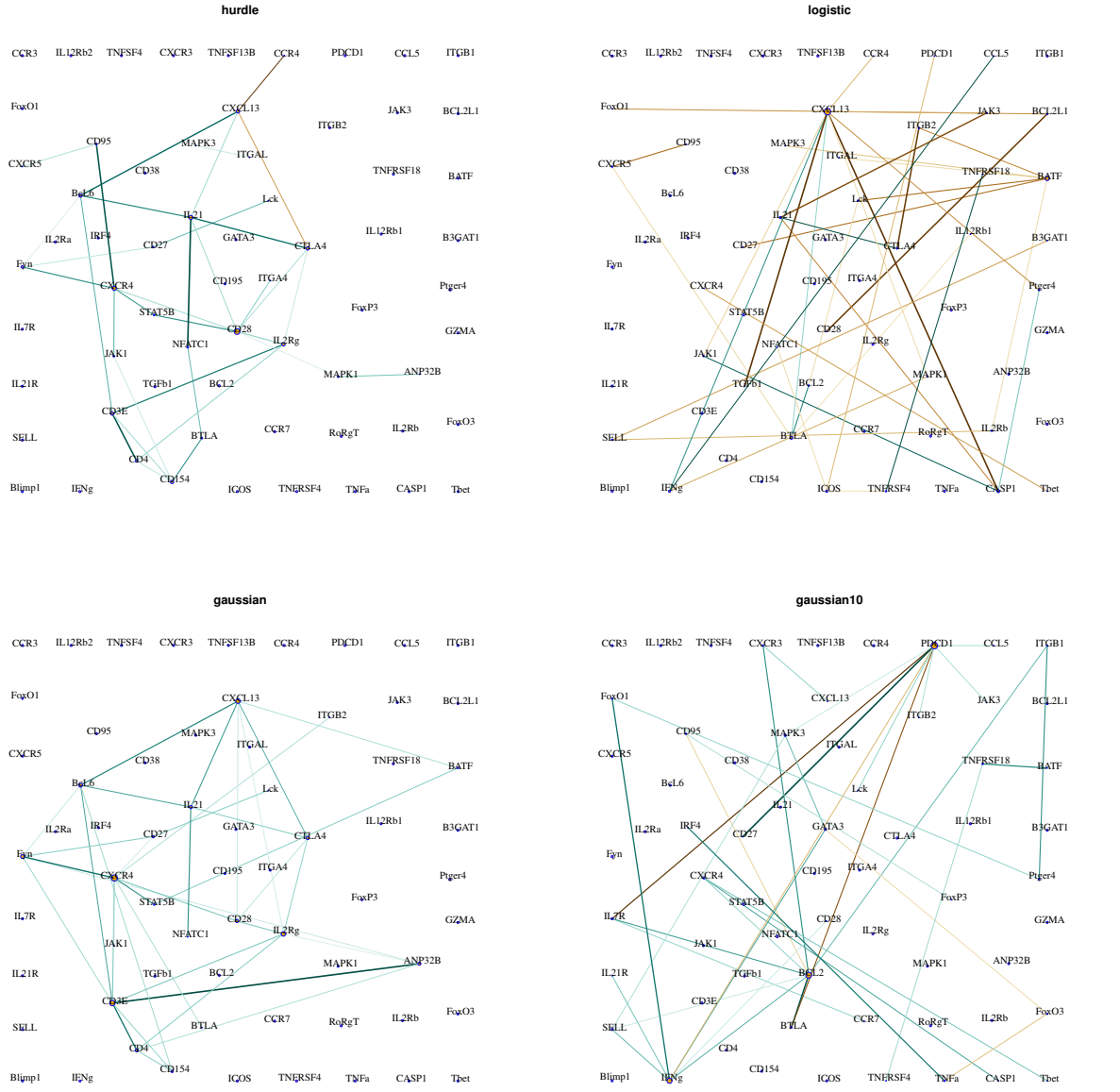

Figure 4.4: Networks of 35 edges estimated through neighborhood selection under the Hurdle, logistic, Gaussian model (single cells) and Gaussian model (10 cell aggregates) in T follicular helper cells. Brown hues indicate estimated negative dependences, while blue-green hues indicate positive dependences. The edge width and saturation are larger for stronger estimated dependences.

In a Gaussian model, the network is star-shaped, with MX1, CCL17, TAX1BP3 and CCL3 as hubs all with degrees ≥ 15, though none are directly inter-connected (figure 4.5). In all, 2.5% of non-isolated vertices contribute 50% of the edges in the network. With the exception of TAX1BP3, these hub genes are all immune-signaling related.

In the Hurdle model, the graph is chain-shaped, with the maximum degree being 12: 7% of nodes provide 50% of the edges. The strongest hub, MGL2 (also known as CD301b), has been recently described to be involved in uptake and presentation of glycosylated antigens, such as LPS, by dendritic cells [Denda-Nagai et al., 2010]. A sub-connected set of genes coding for MHC-II antigen presentation (H2-AB1, H2-EB1, H2-AA) is the densest sub-component, and interconnected to MGL2 as well as FABP5. Increased expression of FABP5 has been shown to increase expression of cytokines IL-7 and IL-18, hence is also involved in immune cell stimulation [Adachi et al., 2012]. Many of the neighbors of MGL2,H2-AB1, H2-EB1, H2-AA and FABP5 are neighbors of the hub genes in the Gaussian graph, whereas MX1, CCL17 and CCL3 are sparsely connected in the Hurdle network. TAX1BP3 is absent.

4.7.1 Graphical geneset edge enrichment

We consider how well the 700 edges recapitulate known relationships between genes using the Gene Ontology (GO) annotations. The Gene Ontology Consortium [2015] provides a directed, acyclic graph of ontologies to which genes may be annotated if they have been shown experimentally or computationally to be involved in a biological process, component or function. In the GO annotations for mouse, most genes are members of several categories (median= 14). We test for enrichment between and within categories under a hypergeometric model, in which each pair (i, j) of (possibly non-disjoint) GO categories induces a *coloring* $(i, j) \to c$ of vertices, coloring any vertex belonging to either i or j. Evidence that this color is interconnected, given 700 total edges, n_c of which are between c-colored vertices, is evaluated using the hypergeometric tail probability $t(c) = P(N_c > n_n)$ in an urn of $4431^2/2$ possible edges, of which m_c could possibly join c-colored vertices. The distribution of the smallest $k < 200$ order statistics $t_{(k)}$ across the $\sim$ 16 million dependent tests possible on $3987^2/2$

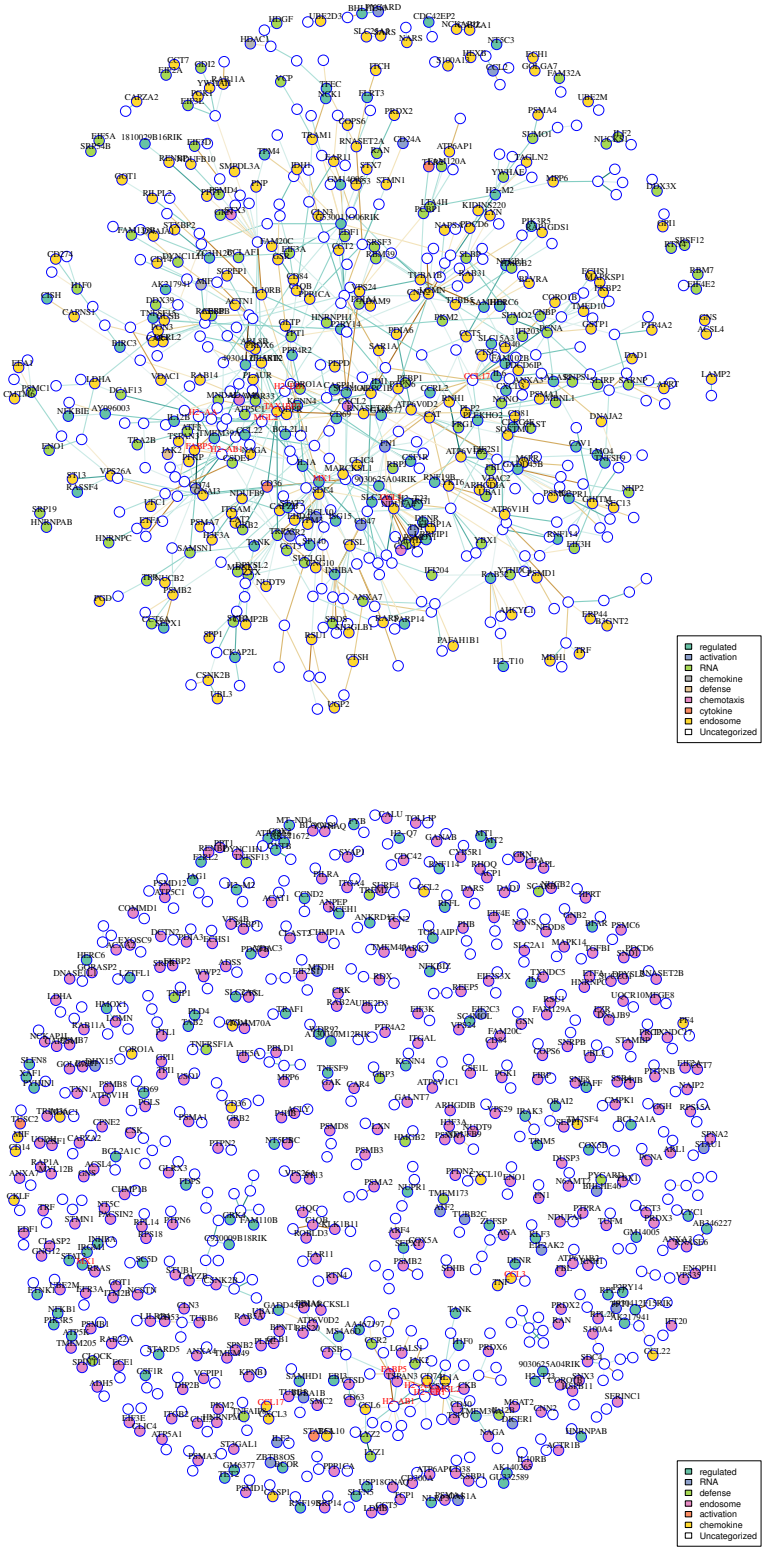

Figure 4.5: Networks estimated in LPS-treated mouse dendritic cells under Gaussian and Hurdle models. Hub genes are shown in red. Vertex colors indicate gene ontology membership.

pairs of categories is found under a Erdos-Renyi random graph model, yielding Monte Carlo p-values $p(t_{(k)})$. A color is declared significant if $p(t_c) < .05$ and all colors ranked above it are also significant at .05 (using the ordering on $t(c)$), thus controlling the FDR at less than 5%.

In the Gaussian model, more than 100 pairs of categories (colors) are significantly enriched at an FDR of less than 5%, however in these 100 pairs, only 7 correspond to intra-category enrichment (figure 4.6). These are: response to salt stress, potassium channel regulator activity, extracellular space, extracellular exosome and three manually curated genesets containing genes with significant time-course differential expression in the original experiment. In the Hurdle model, 14/60 significantly enriched pairs form intra-connections, including defense response to Gram-negative bacteria, and cell-cell adhesion and several modules involving extracellular secretion via the Golgi apparatus. Also of particular note, genes annotated to the activation of innate immune response are directly connected to RNA PolII transcription factors, a connection absent in the Gaussian model. This suggests that using the more parametrically appropriate Hurdle model manages to identify transcription factor-induced expression changes in these regulated genes, a direct method by which one gene would induce expression changes in another. No significant enrichment was found in the logistic model.

4.8 Discussion

Graphical models estimated from single cell data are distinct from networks estimating from bulk data, or even repeated stochastic samples. In simulations, the Hurdle model with anisometric penalty has much greater sensitivity compared to available methods, while in the two data sets here, it yields substantially different network estimates compared to Gaussian and Logistic models on these zero-inflated data. When enrichment of gene ontology categories is considered between vertices in transcriptome-wide data, the enrichment uncovered with the Hurdle model is consistent with identifying direct effects of transcription factors on genes undergoing dynamic regulation due to LPS exposure.

Although measuring transcriptome-wide data allows conditional estimation of direct ef-

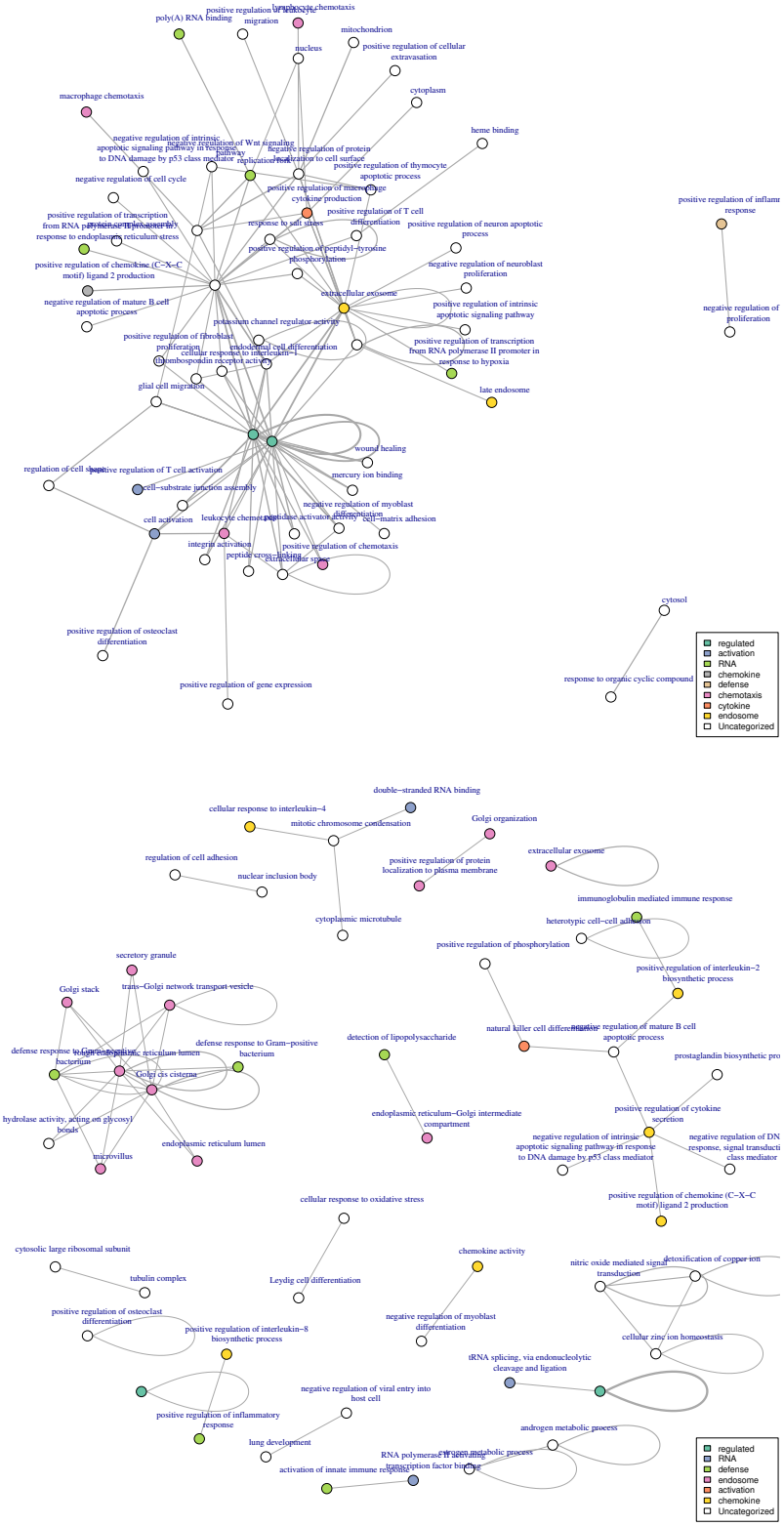

Figure 4.6: Modules enriched at FDR $\leq$ 5% using graphical geneset edge enrichment in mouse dendritic cells under Gaussian and Hurdle models.

fects between genes, non-mRNA factors may also greatly affect gene expression. In this sense, important variables have still been marginalized over, and in the case of the Tfh data, indeed, most of the transcriptome has been marginalized over. Thus, co-expression under sparsity assumptions is most helpful as a screening procedure. Methods to adapt graphical model variable selection to clustering and/or factor analytic models would extend its usefulness greatly, and allow greater biological insight with these data sets.

4.9 Supplemental methods

In all models and data sets, the cellular detection rate $\sum_j I_{y_{ij}>0}$ [Finak et al., 2015] was used as an unpenalized covariate in $\mathbf{Z}$ as described in algorithm 1. In the Tfh data, a separate intercept was fit for donor, as well. For the Gaussian and Hurdle models, positive values were conditionally centered

$$\tilde{y}_{ij} = \begin{cases} 0 & v_{ij} = 0 \\ y_{ij} - \bar{y}_j^+ & \text{else} \end{cases}$$

where $\bar{y}_j^+$ is the average in a gene over positive values, to make V_j and Y_j marginally orthogonal, which enhanced optimization convergence, and reduced the leverage of zeros in the Gaussian model.

The mDC data set was thresholded as described previously [Finak et al., 2015], and filtered for low-expression and cluster-disrupted cells.

BIBLIOGRAPHY

Yasuhiro Adachi, Sumie Hiramatsu, Nobuko Tokuda, Kazem Sharifi, Majid Ebrahimi, Ariful Islam, Yoshiteru Kagawa, Linda Koshy Vaidyan, Tomoo Sawada, Kimikazu Hamano, and Yuji Owada. Fatty acid-binding protein 4 (FABP4) and FABP5 modulate cytokine production in the mouse thymic epithelial cells. *Histochemistry and Cell Biology*, 138(3): 397–406, 2012.

Takeshi Amemiya. Tobit models: A survey. *Journal of Econometrics*, 24(81):3–61, 1984.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, jan 2010.

Barry Arnold and James Press. Compatible Conditional Distributions. *Journal of the American Statistical Association*, 84(405), 1989.

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

Martin Bengtsson, Anders Ståhlberg, Patrik Rorsman, and Mikael Kubista. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Research*, 15(10):1388–1392, oct 2005.

Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B Methodological*, 57(1):289–300, 1995.

Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C

Marioni, and Marcus G Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, advance on, sep 2013.

Florian Buettner, Victoria Moignard, B Göttgens, and FJ Theis. Probabilistic PCA of censored data: accounting for uncertainties in the visualisation of high-throughput single-cell qPCR data. *Bioinformatics*, pages 1–9, 2014.

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

Shizhe Chen, Daniela M. Witten, and Ali Shojaie. Selection and estimation for mixed graphical models. *Biometrika*, 102(1):47–64, nov 2015.

Jie Cheng, Elizaveta Levina, and Ji Zhu. High-dimensional Mixed Graphical Models. *arXiv preprint arXiv:1304.2810*, apr 2013.

Talyn Chu, Aaron J Tyznik, Sarah Roepke, Amy M Berkley, Amanda Woodward-Davis, Laura Pattacini, Michael J Bevan, Dietmar Zehn, and Martin Prlic. Bystander-activated memory CD8 T cells control early pathogen load in an innate-like, NKG2D-dependent manner. *Cell Reports*, 3(3):701–708, 2013.

Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):13, 2016.

DR Cox and N Wermuth. Response models for mixed binary and quantitative variables. *Biometrika*, 79(3):441–461, 1992.

Piero Dalerba, Tomer Kalisky, Debashis Sahoo, Pradeep Rajendran, Michael Rothenberg, Anne Leyrat, Sopheak Sim, Jennifer Okamoto, Darius Johnston, Dalong Qian, Maider

Zabala, Janet Bueno, Norma Neff, Jianbin Wang, Andrew Shelton, Brendan Visser, Shigeo Hisamori, Yohei Shimono, Marc van de Wetering, Hans Clevers, Michael Clarke, and Stephen Quake. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnology*, 29(12):1120–1127, nov 2011.

H de Wit, D Hoogstraten, R M Halie, and E Vellenga. Interferon-gamma modulates the lipopolysaccharide-induced expression of AP-1 and NF-kappa B at the mRNA and protein level in human monocytes. *Experimental Hematology*, 24(2):228–235, feb 1996.

Kaori Denda-Nagai, Satoshi Aida, Kengo Saba, Kiwamu Suzuki, Saya Moriyama, Sarawut Oo-puthinan, Makoto Tsuiji, Akiko Morikawa, Yosuke Kumamoto, Daisuke Sugiura, Akihiko Kudo, Yoshihiro Akimoto, Hayato Kawakami, Nicolai V. Bovin, and Tatsuro Irimura. Distribution and function of macrophage galactose-type C-type lectin 2 (MGL2/CD301b): Efficient uptake and presentation of glycosylated antigens by dendritic cells. *Journal of Biological Chemistry*, 285(25):19193–19204, 2010.

Adrian Dobra, Chris Hans, Beatrix Jones, Joseph R. Nevins, Guang Yao, and Mike West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90:196–212, 2004.

Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–6, aug 2002.

Torbjørn Eltoft, Taesu Kim, and Te Won Lee. On the multivariate Laplace distribution. *IEEE Signal Processing Letters*, 13(5):300–303, 2006.

Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana McElrath, Martin Prlic, Peter S. Linsley, and Raphael Gottardo. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16(1):278, 2015.

Lukas Flatz, Rahul Roychoudhuri, Mitsuo Honda, Abdelali Filali-Mouhim, Jean-Philippe Goulet, Nadia Kettaf, Min Lin, Mario Roederer, Elias Haddad, Rafick Sékaly, and Gary Nabel. Single-cell gene-expression profiling reveals qualitatively distinct CD8 T cells elicited by different gene-based vaccines. *Proceedings of the National Academy of Sciences*, mar 2011.

Rina Foygel and Mathias Drton. Exact block-wise optimization in group lasso and sparse group lasso for linear regression. *Arxiv preprint arXiv:1010.3320*, pages 1–19, 2010.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, jul 2008.

Youngchao Ge, Sandrine Dudoit, and Terence Speed. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77, jun 2003.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. *Annals of Applied Statistics*, 2(4):1360–1383, 2008.

Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

Jason Glotzbach, Michael Januszyk, Ivan Vial, Victor Wong, Alexander Gelbard, Tomer Kalisky, Hariharan Thangarajah, Michael Longaker, Stephen Quake, Gilbert Chu, and Geoffrey Gurtner. An Information Theoretic, Microfluidic-Based Single Cell Analysis Permits Identification of Subpopulations among Putatively Homogeneous Stem Cells. *PloS one*, 6(6):e21211, jun 2011.

Raphael Gottardo, Adrian Raftery, Ka Yee Yeung, and Roger Bumgarner. Bayesian Robust Inference for Differential Gene Expression in Microarrays with Multiple Samples. *Biometrics*, 62(1):10–18, mar 2006.

Stephanie C Hicks, Mingxiang Teng, and Rafael A Irizarry. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *bioRxiv 025528*, 2015.

R Higuchi, G Dollinger, P S Walsh, and R Griffith. Simultaneous amplification and detection of specific DNA sequences. *Biotechnology*, 10(4):413–417, apr 1992.

Kevin a Janes, Chun-Chao Wang, Karin J Holmberg, Kristin Cabral, and Joan S Brugge. Identifying single-cell molecular programs by stochastic profiling. *Nature Methods*, 7(4): 311–317, 2010.

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

Yann Karlen, Alan McNair, Sébastien Perseguers, Christian Mazza, and Nicolas Mermod. Statistical significance of quantitative PCR. *BMC Bioinformatics*, 8(1):131, apr 2007.

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

Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):1–5, may 2014.

Jong Kyoung Kim and John C Marioni. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biology*, 14(1):R7, 2013.

Steffen Lauritzen. *Graphical Models*. Oxford University Press, Oxford, 1st edition, 1996.

Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2), 2014.

J Lee and T Hastie. Structure Learning of Mixed Graphical Models. In *AISTATS 16*, volume 31, pages 388–396, Scottsdale, AZ, USA, 2013.

Jeffrey T. Leek. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research*, 42(21):e161–e161, dec 2014.

Jeffrey Levsky, Shailesh Shenoy, Rossanna Pezo, and Robert Singer. Single-Cell Gene Expression Profiling. *Science*, 297(5582):836–840, aug 2002.

Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, 2011.

Shuzhao Li, Nadine Rouphael, Sai Duraisingham, Sandra Romero-Steiner, Scott Presnell, Carl Davis, Daniel S Schmidt, Scott E Johnson, Andrea Milton, Gowrisankar Rajam, Sudhir Kasturi, George M Carlone, Charlie Quinn, Damien Chaussabel, A Karolina Palucka, Mark J Mulligan, Rafi Ahmed, David S Stephens, Helder I Nakaya, and Bali Pulendran. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nature Immunology*, 15(2):195–204, 2014.

Lawrence Lin. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, 45(1):255–268, mar 1989.

Lin Lin, Greg Finak, Kevin Ushey, Chetan Seshadri, Thomas R Hawn, Nicole Frahm, Thomas J Scriba, Hassan Mahomed, Willem Hanekom, Pierre-Alexandre Bart, Giuseppe Pantaleo, Georgia D Tomaras, Supachai Rerks-Ngarm, Jaranit Kaewkungwal, Sorachai Nitayaphan, Punnee Pitisuttithum, Nelson L Michael, Jerome H Kim, Merlin L Robb, Robert J O'Connell, Nicos Karasavvas, Peter Gilbert, Stephen C De Rosa, M Juliana McElrath, and Raphael Gottardo. COMPASS identifies T-cell subsets correlated with clinical outcomes. *Nature Biotechnology*, 33(6):610–616, 2015.

Han Liu, John Lafferty, and J Wainwright. The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. *Journal of Machine Learning Research*, 10(10): 2295–2328, 2009.

F. J. Livesey. Strategies for microarray analysis of limiting amounts of RNA. *Briefings in Functional Genomics and Proteomics*, 2(1):31–36, 2003.

PL Loh and MJ Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *Annals of Statistics*, 41(6):3022–3049, 2013.

Cindy S. Ma, Elissa K. Deenick, Marcel Batten, and Stuart G. Tangye. The origins, function, and regulation of T follicular helper cells. *Journal of Experimental Medicine*, 209(7):1241–1253, 2012.

Samuel Marguerat, Alexander Schmidt, Sandra Codlin, Wei Chen, Ruedi Aebersold, and Jürg Bähler. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell*, 151(3):671–683, 2012.

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

Florian Markowetz and Rainer Spang. Inferring cellular networks: a review. *BMC Bioinformatics*, 8(Suppl 6):S5, 2007.

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

Andrew McDavid, Lucas Dennis, Patrick Danaher, Greg Finak, Michael Krouse, Alice Wang, Philippa Webster, Joseph Beechem, and Raphael Gottardo. Modeling Bi-modality Im-

proves Characterization of Cell Cycle on Gene Expression in Single Cells. *PLoS Computational Biology*, 10(7):e1003696, feb 2014.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, jun 2006.

Mohammad Alam Miah, Se Eun Byeon, Md Selim Ahmed, Cheol-Hee Yoon, Sang-Jun Ha, and Yong-Soo Bae. Egr2 induced during DC development acts as an intrinsic negative regulator of DC immunogenicity. *European Journal of Immunology*, 43(9):2484–2496, sep 2013.

I Olkin and RF Tate. Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics*, 1961.

Olusegun O. Onabajo and Joseph J. Mattapallil. Expansion or Depletion of T Follicular Helper cells During HIV Infection: Consequences for B cell Responses. *Current HIV Research*, 11(8):595–600, 2013.

Olivia Padovan-Merhar, Gautham P Nair, Andrew G Biaesch, Andreas Mayer, Steven Scarfone, Shawn W Foley, Angela R Wu, L Stirling Churchman, Abhyudai Singh, and Arjun Raj. Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Molecular Cell*, 58(2): 339–352, 2015.

Neal Parikh and Stephen Boyd. Proximal Algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2014.

Emma Pierson and Christopher Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1):241, 2015.

Ashley Powell, AmirAli Talasaz, Haiyu Zhang, Marc Coram, Anupama Reddy, Glenn Deng, Melinda Telli, Ranjana Advani, Robert Carlson, Joseph Mollick, Shruti Sheth, Allison

Kurian, James Ford, Frank Stockdale, Stephen Quake, Fabian Pease, Michael Mindrinos, Gyan Bhanot, Shanaz Dairkee, Ronald Davis, and Stefanie Jeffrey. Single Cell Profiling of Circulating Tumor Cells: Transcriptional Heterogeneity and Diversity from Breast Cancer Cell Lines. *PLoS One*, 7(5):e33788, may 2012.

Melissa L Precopio, Michael R Betts, Janie Parrino, David A Price, Emma Gostick, David R Ambrozak, Tedi E Asher, Daniel C Douek, Alexandre Harari, Giuseppe Pantaleo, Robert Bailer, Barney S Graham, Mario Roederer, and Richard A Koup. Immunization with vaccinia virus induces polyfunctional and phenotypically distinctive CD8(+) T cell responses. *Journal of Experimental Medicine*, 204(6):1405–1416, 2007.

Arjun Raj, Patrick van den Bogaard, Scott a Rifkin, Alexander van Oudenaarden, and Sanjay Tyagi. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods*, 5(10):877–879, 2008.

Daniel Ramskold, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid Faridani, Gregory Daniels, Irina Khrebtukova, Jeanne Loring, Louise Laurent, Gary Schroth, and Rickard Sandberg. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, 30(8):777–782, jul 2012.

C. Radhakirshna Rao. *Linear Statistical Inference and its Applications*. John Wiley & Sons, Ltd., 2nd edition, 1973.

Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using 1 -regularized logistic regression. *The Annals of Statistics*, 38(3): 1287–1319, jun 2010.

Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9): 896–902, aug 2014.

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

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

Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, 2015.

Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4, 2005.

Thomas Schmittgen and Kenneth Livak. Analyzing real-time PCR data by the comparative CT method. *Nat. Protocols*, 3(6):1101–1108, jun 2008.

Rajen D. Shah and Richard J. Samworth. Variable selection with error control: Another look at stability selection. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 75(1):55–80, 2013.

Alex K. Shalek, Rahul Satija, Joe Shuga, John J. Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S. Gertner, Jellert T. Gaublomme, Nir Yosef, Schraga Schwartz, Brian Fowler, Suzanne Weaver, Jing Wang, Xiaohui Wang, Ruihua Ding, Raktima Raychowdhury, Nir Friedman, Nir Hacohen, Hongkun Park, Andrew P. May, and Aviv Regev. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):263–269, jun 2014.

Noah Simon and Robert Tibshirani. Standardization and the Group Lasso Penalty. *Statistica Sinica*, 22(3):983–1001, 2012.

Ronald B Smeltz. Profound enhancement of the IL-12/IL-18 pathway of IFN-gamma secretion in human CD8+ memory T cell subsets via IL-15. *Journal of Immunology*, 178(8): 4786–4792, 2007.

Gordon Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), jan 2004.

Wesley Tansey, Oscar Hernan Madrid Padilla, Arun Sai Suggala, and Pradeep Ravikumar. Vector-Space Markov Random Fields via Exponential Families. *Proceedings of The 32nd International Conference on Machine Learning*, 37:684–692, 2015.

The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 2015.

Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 74(2): 245–266, 2012.

O Toomet and A Henningsen. Sample selection models in R: Package sampleSelection. *Journal of Statistical Software*, 27(7):1–23, 2008.

Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, 31(1):46–53, 2013.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, mar 2014.

V G Tusher, R Tibshirani, and G Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–21, 2001.

Aaron J Tyznik, Shilpi Verma, Qiao Wang, Mitchell Kronenberg, and Chris A Benedict. Distinct requirements for activation of NKT and NK cells during viral infection. *Journal of Immunology*, 192(8):3676–3685, 2014.

Catalina A. Vallejos, John C. Marioni, and Sylvia Richardson. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLoS Computational Biology*, 11(6):1–18, 2015.

Mark a van de Wiel, Maarten Neerincx, Tineke E Buffart, Daoud Sie, and Henk M W Verheul. ShrinkBayes: a versatile R-package for analysis of count-based sequencing data in complex study designs. *BMC Bioinformatics*, 15:116, 2014.

Jo Vandesompele, Katleen De Preter, Filip Pattyn, Bruce Poppe, Nadine Van Roy, Anne De Paepe, and Frank Speleman. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology*, 3(7): 1–12, jun 2002.

Navin Varadarajan, Boris Julg, Yvonne Yamanaka, Huabiao Chen, Adebola Ogunniyi, Elizabeth McAndrew, Lindsay Porter, Alicja Piechocka-Trocha, Brenna Hill, Daniel Douek, Florencia Pereyra, Bruce Walker, and Christopher Love. A high-throughput single-cell analysis of human CD8 T cell functions reveals discordance for cytokine secretion and cytolysis. *The Journal of Clinical Investigation*, 121(11):4322–4331, nov 2011.

Arend Voorman, Ali Shojaie, and Daniela Witten. Graph estimation with joint additive models. *Biometrika*, 101(1):85–101, 2014.

S S Wilks. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, mar 1938.

Di Wu and Gordon K Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17):e133, sep 2012.

Eunho Yang, Y Baker, P Ravikumar, G Allen, and Z Liu. Mixed Graphical Models via Exponential Families. In *AISTATS 17*, volume 33, Reykjavik, Iceland, 2014.

Thomas W Yee. *Vector Generalized Linear and Additive Models*. Springer Series in Statistics. Springer New York, New York, NY, 1 edition, 2015. ISBN 978-1-4939-2817-0.

M Yuan and Y Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94:19–35, 2007.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 68(1):49–67, 2006.

Appendix A

SUPPLEMENTARY DERIVATIONS AND FIGURES

A.1 Empirical Bayes derivation of variance hyper parameters

Suppose there are genes $j = 1, \ldots, G$. Assume that the precision (inverse variance) for the continuous component of gene j is distributed

$$\tau_j | \alpha_0, \beta_0 \sim \text{Gamma-rate}(\alpha_0, \beta_0)$$

and that $i \neq j \Rightarrow \tau_i \perp \tau_j | \alpha_0, \beta_0$. Thus $\tau_j | \alpha_0, \beta_0$ has density

$$f(\tau_j | \alpha_0, \beta_0) = \tau_j^{\alpha_0 - 1} e^{-\tau_j \beta_0} \beta_0^{\alpha_0} / \Gamma(\alpha_0).$$

Assume that n_j cells have non-zero expression vector Y_j in gene j under the linear model $\mathrm{E}[Y_j | X] = X\eta$, with $\dim(\eta) = p$, so that

$$Y_j | \tau_j, \eta \sim \mathcal{N}(X\eta, \tau_j).$$

This implies that $R_j = \sum (y_i - \hat{\eta} X_i)^2$ is sufficient for τ_j and that statistic has scale chi-square distribution with $n_j - p$ degrees of freedom, or equivalently, a gamma-rate distribution with shape $\alpha_j = (n_j - p)/2$ and rate $\beta_j = \tau_j/2$. Here $\hat{\eta}$ is the typical OLS estimator.

The joint distribution of $\tau_j, R_j | \alpha_0, \beta_0$ has density

$$f(R_j, \tau_j | \alpha_0, \beta_0) = \tau_j^{\alpha' - 1} \exp\left(-\tau_j \beta'\right) \beta_0^{\alpha_0} / \Gamma(\alpha_0) R_j^{(n_j - p)/2 - 1} \left(1/2\right)^{(n_j - p)/2} / \Gamma\left((n_j - p)/2\right).$$

for $\alpha' = \alpha_0 + (n_j - p)/2$ and $\beta' = \beta_0 + R_j/2$. In terms of τ_j this density has the kernel of a

gamma distribution, with aforementioned parameters, so that marginalizing out τ_j yields

$$\begin{aligned} f(R_j|\alpha_0,\beta_0) &= \frac{\Gamma(\alpha')\beta_0^{\alpha_0}}{\Gamma(\alpha_0)\beta'^{\alpha'}}\frac{(1/2)^{(n_j-p)/2}}{\Gamma\left((n_j-p)/2\right)}R_j^{(n_j-p)/2-1} \\ &= \frac{\Gamma((n_j-p)/2+\alpha_0)\beta_0^{\alpha_0}}{\Gamma(\alpha_0)\left(\beta_0+R_j/2\right)^{(n_j-p)/2+\alpha_0}}\frac{(1/2)^{(n_j-p)/2}}{\Gamma\left((n_j-p)/2\right)}R_j^{(n_j-p)/2-1} \\ &= \frac{\Gamma((n_j-p)/2+\alpha_0)\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\frac{(1/2)^{(n_j-p)/2}}{\Gamma\left((n_j-p)/2\right)}\frac{R_j^{(n_j-p)/2-1}}{\beta_0^{(n_j-p)/2+\alpha_0}\left(1+R_j/(2\beta_0)\right)^{(n_j-p)/2+\alpha_0}} \\ &= \frac{(1/2)^{(n_j-p)/2}}{\mathcal{B}\left((n_j-p)/2,\alpha_0\right)\beta_0^{(n_j-p)/2}}\frac{R_j^{(n_j-p)/2-1}}{\left(1+R_j/(2\beta_0)\right)^{(n_j-p)/2+\alpha_0}} \end{aligned} \tag{A.1}$$

where $\mathcal{B}$ is the beta function. This is recognized as the kernel of a (scale) F-distribution. Since $1+R_j$ is raised to the $-\left((n_j-p)/2+\alpha_0\right)$, R_j is raised to the $(n_j-p)/2-1$ power, and R_j is divided by $2\beta_0$, we identify the parameters of the scale-F distribution d_1, d_2, σ as

$$\begin{aligned} \frac{d_1+d_2}{2} &= [(n_j-p)/2+\alpha_0] \\ \frac{d_1}{2}-1 &= (n_j-p)/2-1 \\ \frac{d_1}{d_2\sigma} &= 1/(2\beta_0). \end{aligned}$$

Solving this system gives

$$\begin{aligned} d_1 &= n_j-p \\ d_2 &= 2\alpha_0 \\ \sigma &= \frac{\beta_0(n_j-p)}{\alpha_0}. \end{aligned}$$

Working backwards from a $F\left(n_j-p, 2\alpha_0, \frac{\beta_0(n_j-p)}{\alpha_0}\right)$ distribution, we would have that

$$f(R_j) = \frac{1}{\mathcal{B}\left(\frac{(n_j-p)}{2},\alpha_0\right)}\left(\frac{(n_j-p)}{2\alpha_0}\right)^{(n_j-p)/2}\left[\frac{R_j\alpha_0}{\beta_0(n_j-p)}\right]^{(n_j-p)/2-1}\left[1+\frac{R_j}{(2\beta_0)}\right]^{-(n_j-p+2\alpha_0)/2}\frac{\alpha_0}{\beta_0(n_j-p)},$$

which after some algebra verifies to be equivalent to equation A.1.

Maximum Likelihood Estimators

Equation A.1 can be used as the basis for maximum likelihood estimation of α_0, β_0. Dropping constants that do not depend on the parameters, the log-likelihood and score functions have

the form

$$\begin{aligned}\mathcal{L}(\alpha_0,\beta_0) &= -\log \mathcal{B}\left((n_j-p)/2,\alpha_0\right) - \frac{n_j-p}{2}\log\beta_0 - \log\left(1+R_j/(2\beta_0)\right)\left((n_j-p)/2+\alpha_0\right)\\ \mathcal{L}_{\alpha_0} &= \psi\left((n_j-p)/2+\alpha_0\right) - \psi\left(\alpha_0\right) - \log\left(1+R_j/(2\beta_0)\right)\\ \mathcal{L}_{\beta_0} &= \frac{\alpha_0 R_j - (n_j-p)\beta_0}{R_j\beta_0 + 2\beta_0^2},\end{aligned}$$

where ψ is the digamma function $\frac{d\Gamma(x)}{dx}$. This likelihood may be maximized numerically, e.g., using the *optim* function in R.

Posterior MLE for τ_j

Given estimates α_0, β_0 derived by MLE, then the posterior distribution of τ_j is Gamma-rate with parameters $\alpha' = \alpha_0 + (n_j - p)/2$ and $\beta' = \beta_0 + R_j/2$. The log-likelihood and score for τ_j is

$$\begin{aligned}\mathcal{L}(\tau_j) &= (\alpha'-1)\log\tau_j - \tau_j\beta'\\ \mathcal{L}_{\tau_j} &= \frac{\alpha'-1}{\tau_j} - \beta'\end{aligned}$$

which implies that

$$\hat{\tau_j} = \frac{\alpha'-1}{\beta'}$$

which has an interpretation in terms of pseudo-observations as follows

$$\begin{aligned}1/\hat{\tau_j} &= \frac{R_j/2+\beta_0}{\alpha_0+(n_j-p)/2} = \frac{R_j}{n_j-p}\frac{n_j-p}{2\alpha_0+n_j-p} + \frac{\beta_0}{\alpha_0}\frac{2\alpha_0}{2\alpha_0+n_j-p}\\ &= \left(\tau_j^{-1}\right)^{\text{MLE}}\lambda + 1/\tau_0(1-\lambda)\end{aligned}$$

noting that $\left(\tau_j^{-1}\right)^{\text{MLE}} = \frac{R_j}{n_j-p}$ would be the typical MLE of the variance τ_j^{-1} and that $\tau_0 = \alpha_0/\beta_0$ would be the MLE of τ using only the prior information. This final formulation of the shrunken precision as a convex combination of the MLE and the global value τ_0 is used in practice.

A.2 Simulation exploring CDR effects

In order to assess the effect of the including/excluding CDR effects when modeling single-cell gene expression data, we simulated $\log_2$ TPM expression matrix with 2500 genes where 100 genes are differentially expressed for sample size of 100 in each of two stimulation conditions. We tested four scenarios: one with no CDR effect in the simulated data generating process; and three others with varying levels of confounding between CDR and stimulation effect. The four scenarios are described in Table A.1. The parameters in the data generating model were chosen to mimic the the observed features of the MAIT experiment, as described below.

The results based on 100 replication is summarized by the ROC curves in Figure A.4 showing the importance of controlling for CDR when there is a CDR effect in the data generating process. This is especially important when there is confounding between the stimulation and the CDR, as ignoring the CDR effect would typically inflate the type I error rate. At the same time our results also indicates the robustness of our proposed model for including CDR even when there is no CDR effect in the data generating process, promoting the inclusion of CDR as a default model.

A.2.1 Data generating protocol

We set the sample size $N = 200$, the number of genes J=2500 and defined the stimulation indicator s as 0 for first 100 cells and 1 for the last 100. Given stimulation indicator s we generated the data accordingly:

$$\begin{aligned}
\tau_j^2 &\sim Gamma(a_0, b_0) \\
\text{CDR}_i &\sim (1 - s_i)\text{Beta}(a_u, b_u) + s_i\text{Beta}(a_s, b_s) \\
z_{ij}|\text{CDR}_i &\sim \text{Bernoulli}(\text{logit}^{-1}(\mu_j^d + \alpha_j^d s_i + \beta_j^d \text{CDR}_i)) \\
y_{ij}|z_{ij}, \text{CDR}_i &\sim z_{ij}\text{N}(\mu_j^c + \alpha_j^c s_i + \beta_j^c \text{CDR}_i, 1/\tau_j^2)
\end{aligned}$$

The coefficients μ, α, and β were generated from a Normal distribution with hyper parameters based on the distribution of these quantities observed in the MAIT experiment.

We also set $\alpha_j = 0$ for $j > 100$, since only the first 100 genes are differentially expressed (i.e. have a non-zero treatment effect). The precision hyperparameters a_0 and b_0 are set to the point estimates found in the MAIT experiment. The code with all the simulation details can be found in `AdditionalAnalyses.Rmd`.

Table A.1: Hyper parameter settings for CDR generation model.

	strong confounding	moderate confounding	no confounding
a_u	4	6	8
b_s	16	14	12
a_s	12	10	8
b_s	8	10	12

A.3 Supplementary figures

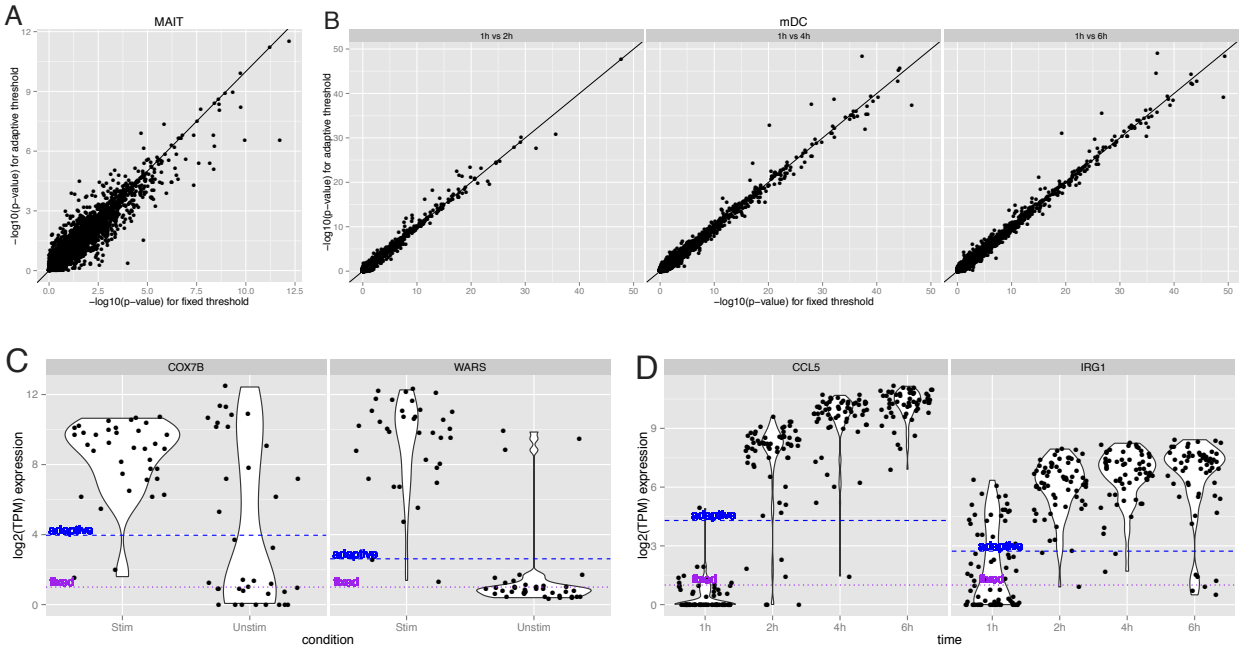

Figure A.1: Scatter plot of p-values for differential expression from adaptive and fixed thresholding on the A) MAIT and B) mDC data sets, demonstrating robustness to the thresholding method. Two selected genes from each data set, with large differences in p-values between fixed and adaptive thresholding in C) MAIT and D) mDC, are genes that exhibit substantial bimodality and our adaptive thresholding appears preferable.

Figure A.2: Scatter plot of normalized (scaled to unit variance and zero mean) CDR (cellular detection rate) calculated from all genes vs. the CDR calculated from housekeeping genes , for stimulated A) and unstimulated B) MAIT cells. The estimated CDRs are linearly related within each condition.

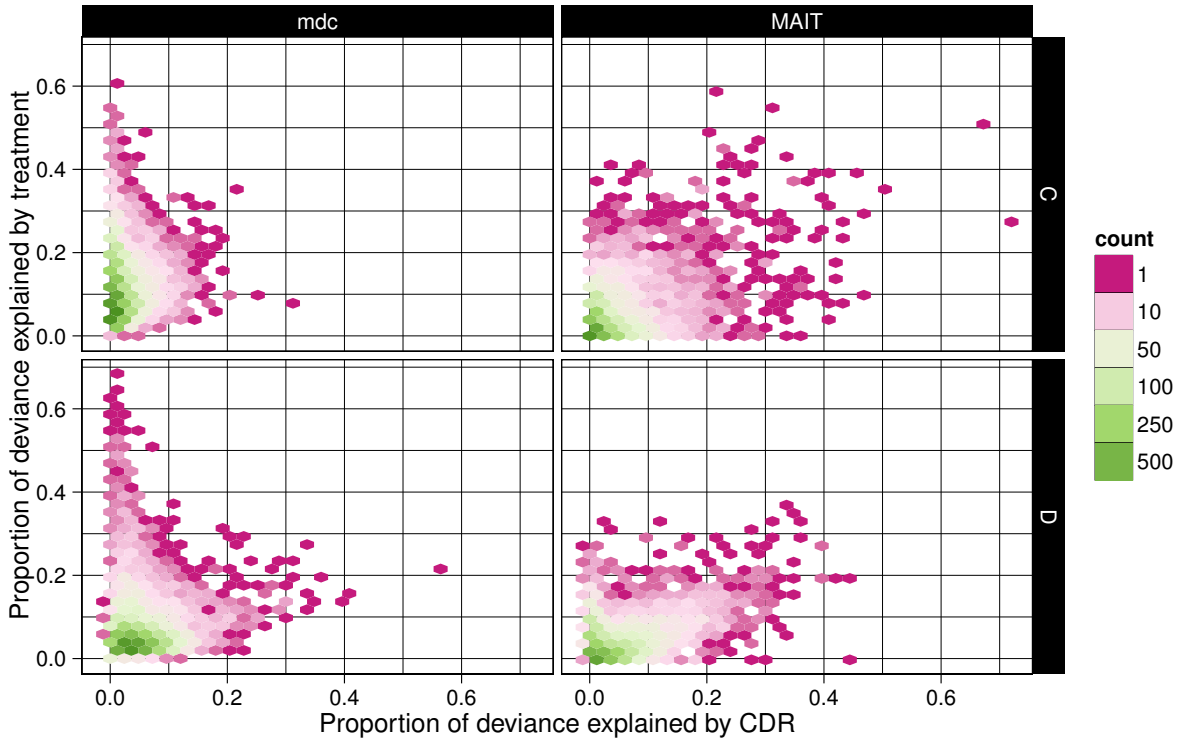

Figure A.3: Amount of variability, measured as percent of null model deviance, attributed to the CDR effect vs. the treatment effect, in each dataset. The CDR accounts for 5.2% of the variability in the MAIT and 4.8% of the variability in the mDC data sets for the average gene. Greater than 9% of the variability is attributed to over 10% of genes in both data sets. CDR contributes the most variability to the discrete component in both data sets and more so in the MAIT data than the mDC data.

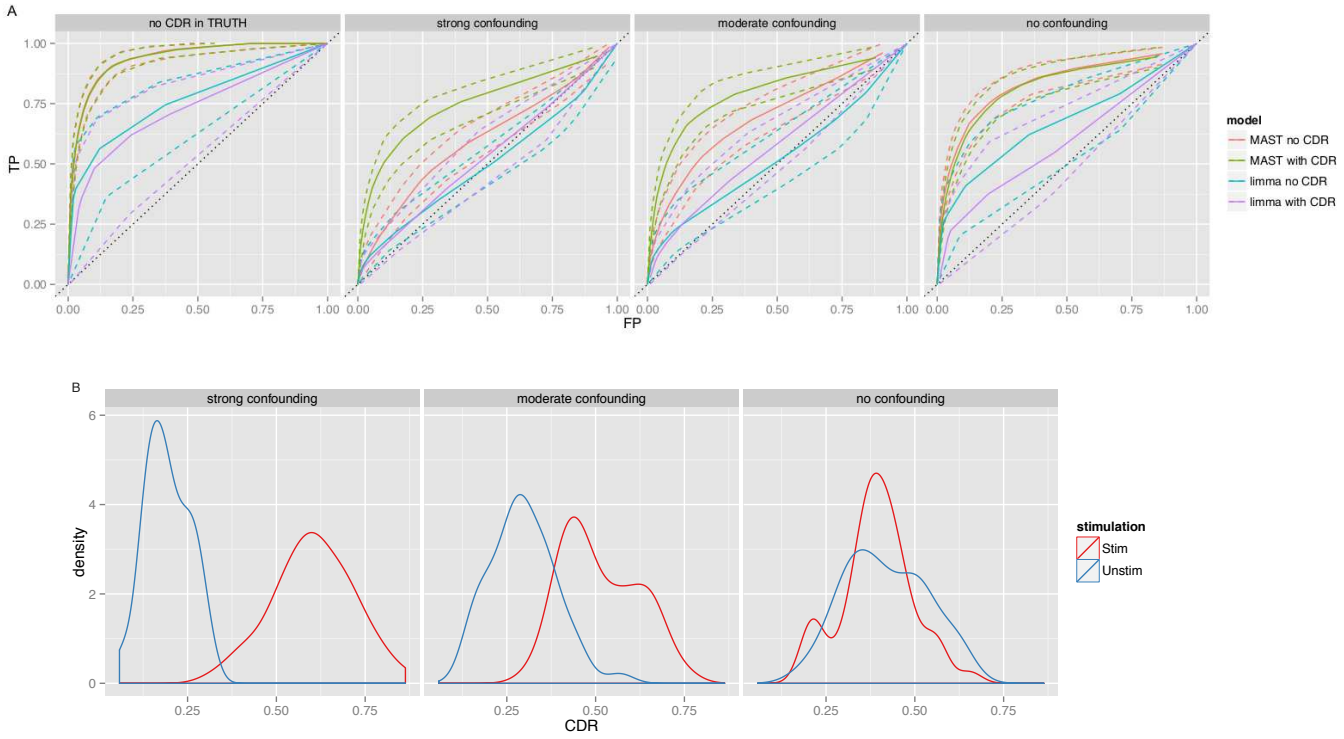

Figure A.4: Effect of CDR and confounding with treatment using different methods. A) ROC curve comparing the effect of controlling for CDR in the MAST model. The solid line is the median and the top and the bottom dashed line represents the 95 and 5 percent quantile. The result indicates that inclusion of CDR improves the performance when there is confounding between the CDR and stimulation and performs nearly the same when there is no confounding or when there is no CDR effect in the data generating model. B) Density plot of generated CDR values across cells using the three levels of confounding between the stimulation and the CDR effects.

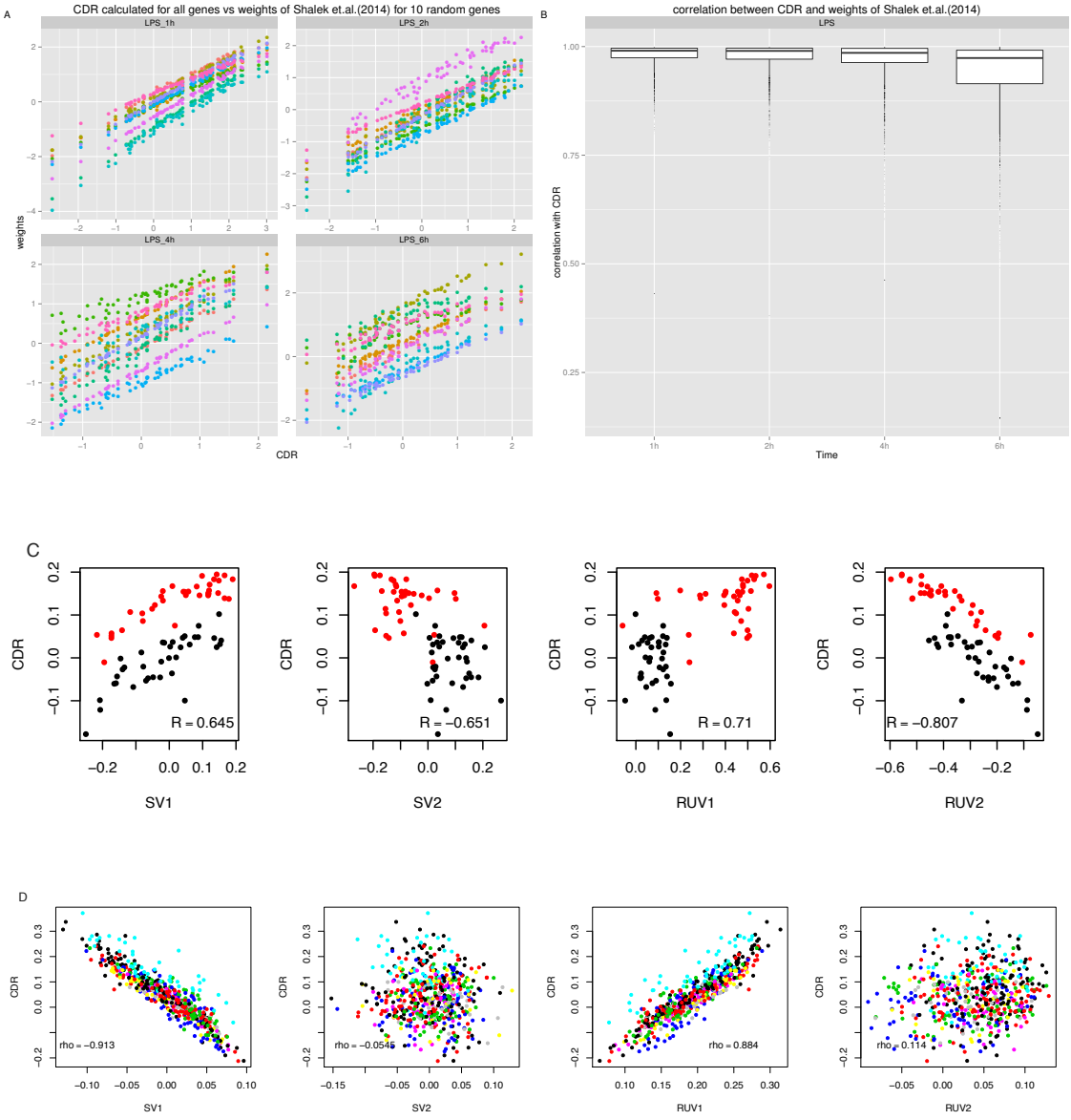

Figure A.5: Comparison of the empirical CDR (centered and scaled) and other correction methods, the cell by gene weights of Shalek et. al., and RUV and SVA. The CDR and Shalek et. al. weights are correlated, in fact generally just shifted by a constant (panel A, in a random subsample of genes, each in a different color), and the correlation coefficient is nearly unity (panel B). The location shift between the CDR and Shalek et. al. weights would be absorbed by the intercept term in the logistic regression. C) Scatterplots of CDR vs. the first and second SVA and RUV components. Treatment groups are shown in different colors. The first SVA and second RUV components are associated with CDR. D) In the mDC data, the first SVA and RUV components are correlated with CDR.

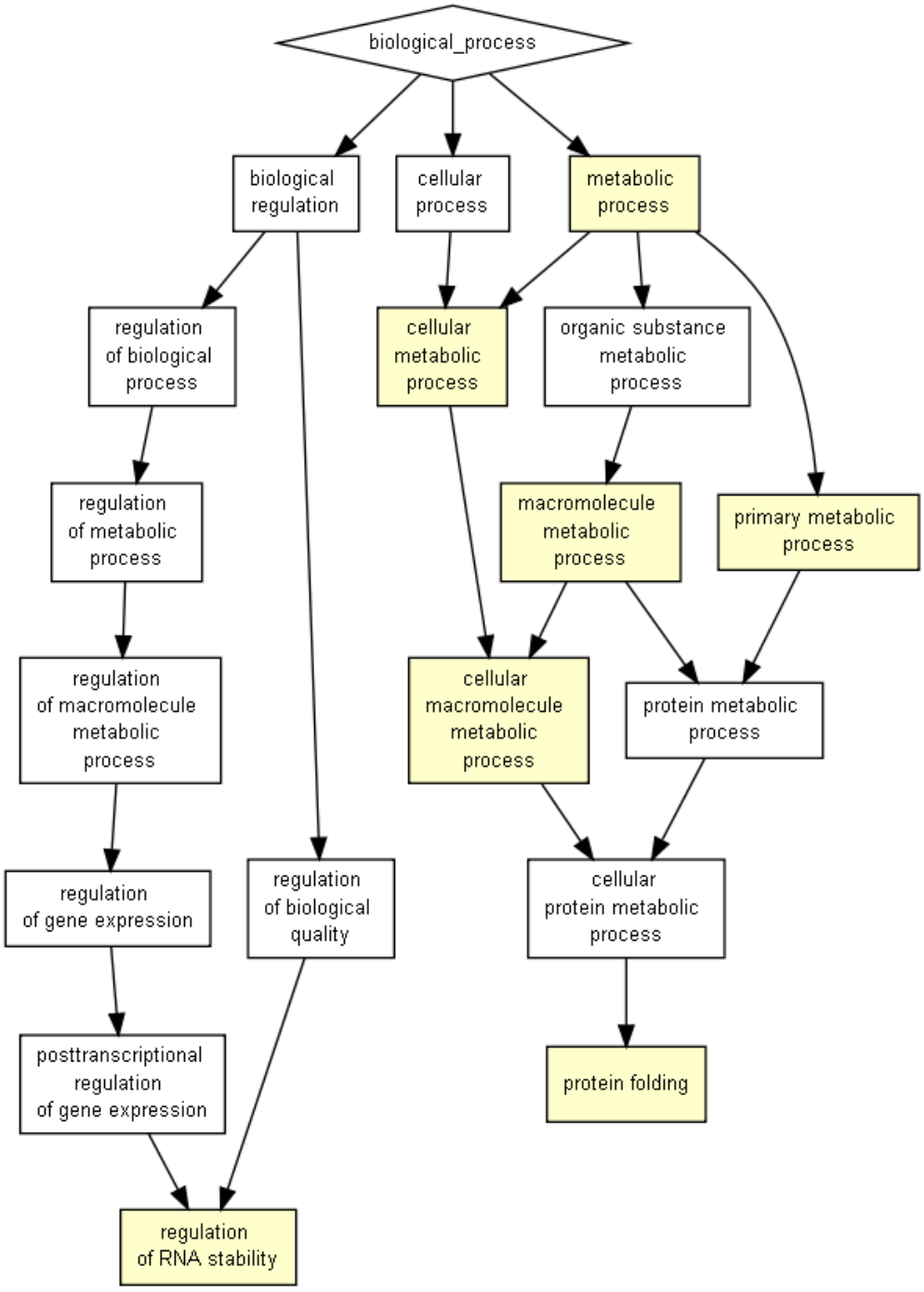

Figure A.6: Gene Ontology Enrichment Analysis using the GOrilla online tools for the set of genes not detected as differentially expressed in the MAIT data set when the CDR is included in the MAST linear model.

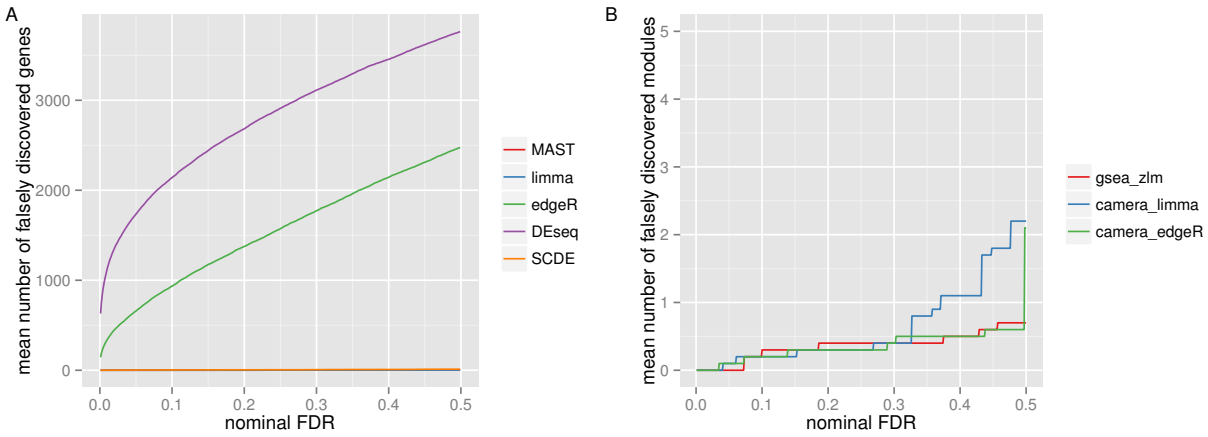

Figure A.7: False discoveries in genes (A) and modules (B) based on numeric permutation experiments for various methods. The unstimulated MAIT cells were permuted into two subsets, and were tested for differential expression under the Hurdle model (MAST), Limma, edgeR, and DEseq. In this scenario, any gene discovered is an *a priori* false discovery, so the number of false discoveries is plotted against the FDR-adjusted significance. We show the average values from ten permutations.

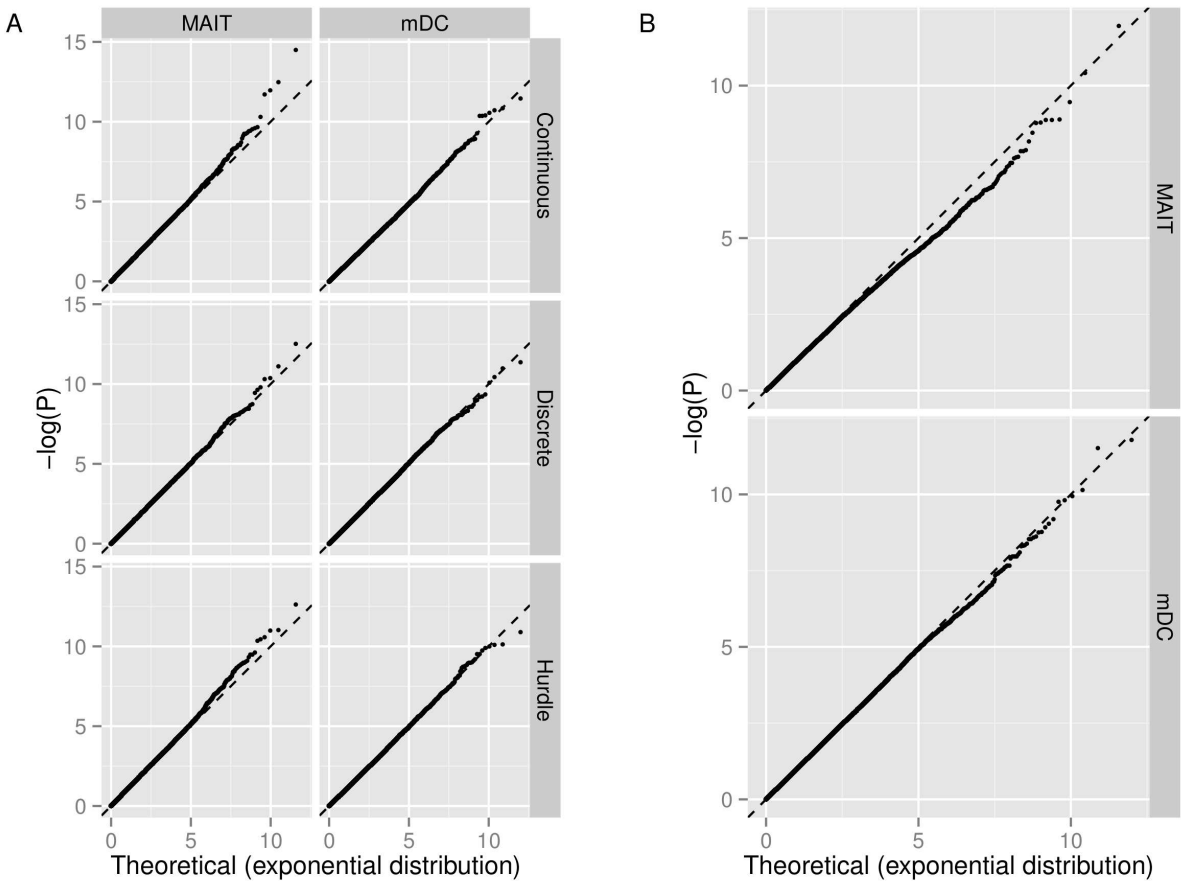

Figure A.8: The distribution of $-\log$ p-values in permuted datasets is compared to its expected Exponential(1) distribution in (A) the hurdle model and (B) Normal-theory t-tests on the same data. In the smaller MAIT dataset ($N = 73$) the Hurdle is inflated in the tail of the test statistic, producing an additional .6 rejections per 1,000 tests at $\alpha = 10^{-3}$. The t-test is deflated, yielding .5 too few rejections per 1,000 tests at $\alpha = 10^{-3}$

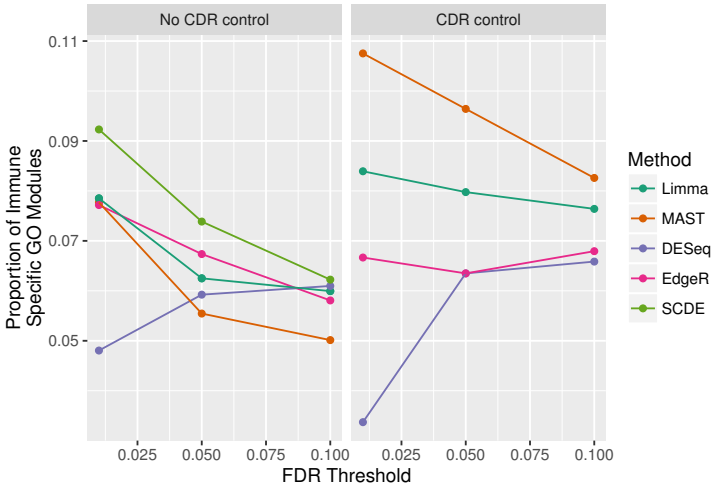

Figure A.9: Proportion of immune-specific GO modules amongst all GO modules enriched in differentially expressed genes in the MAIT data set. Immune-specific GO modules were defined to be terms with experimental evidence codes within the Biological Process ontology that were descendants in the GO graph of the Immune System Process term. Differential expression of genes was determined at three increasing false discovery rate thresholds, and then GO enrichment in differentially expressed genes was tested using the hypergeometric distribution, calling significant enrichment at the 1% FDR level. Inclusion of the CDR in the model for differential expression increases the rate of detection of immune specific modules for the MAST and Limma methods. Among models that do not adjust for CDR, SCDE has highest specificity, but is dominated by MAST under CDR adjustment (SCDE cannot adjust for covariates, so was omitted from the CDR models).

set	Unstim	Stim
signaling in T cells (II) (M35.1)	1.23	2.10
chaperonin mediated protein folding (I) (M204.0)	0.67	1.48
respiratory electron transport chain (mitochondrion) (M238)	0.99	1.04
AP-1 transcription factor network (M20)	1.54	1.00
proteasome (M226)	1.00	1.80
cell cycle and growth arrest (M31)	1.42	0.78
chaperonin mediated protein folding (II) (M204.1)	0.81	1.60
purine nucleotide biosynthesis (M212)	0.70	1.57
spliceosome (M250)	1.10	1.30

Table A.2: Standard deviations of module scores for stimulated and non-stimulated MAIT cells

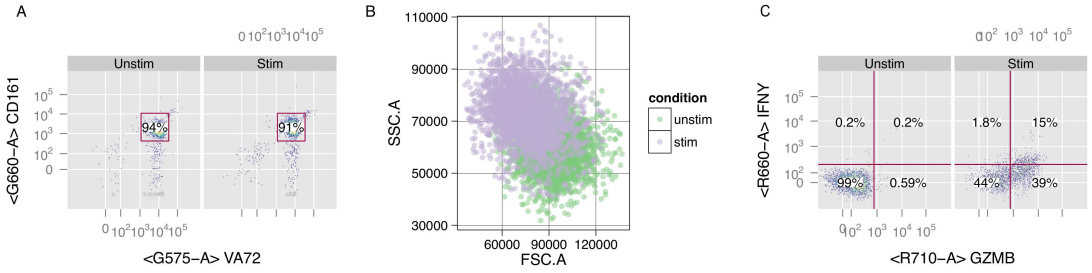

Figure A.10: Post-sort experiments via flow cytometry show that the sorted cell populations were over 90% pure MAITs (Figure A), and exhibited a change in cell size upon stimulation (Figure B) and that up to 44% of stimulated MAITs did not respond to cytokine stimulation (Figure C).

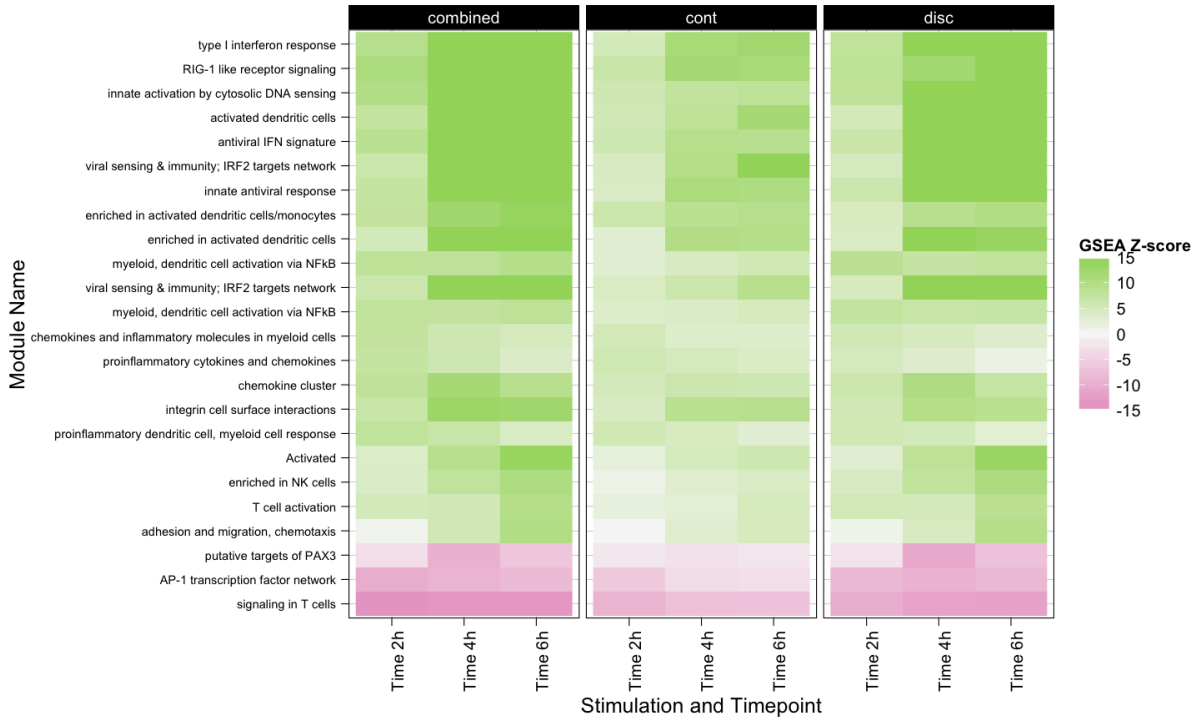

Figure A.11: Gene set enrichment analysis of the mDC data set, LPS stimulated cells using the BTM (blood transcriptional modules) of Li et. al. Decreased expression for AP-1 transcriptional network genes is observed after LPS stimulation, consistent with previous findings in the literature [de Wit et al., 1996]. Type-1 interferon response and antiviral IFN modules are among the most significantly enriched and are consistent with the findings of the original publication [Shalek et al., 2014].

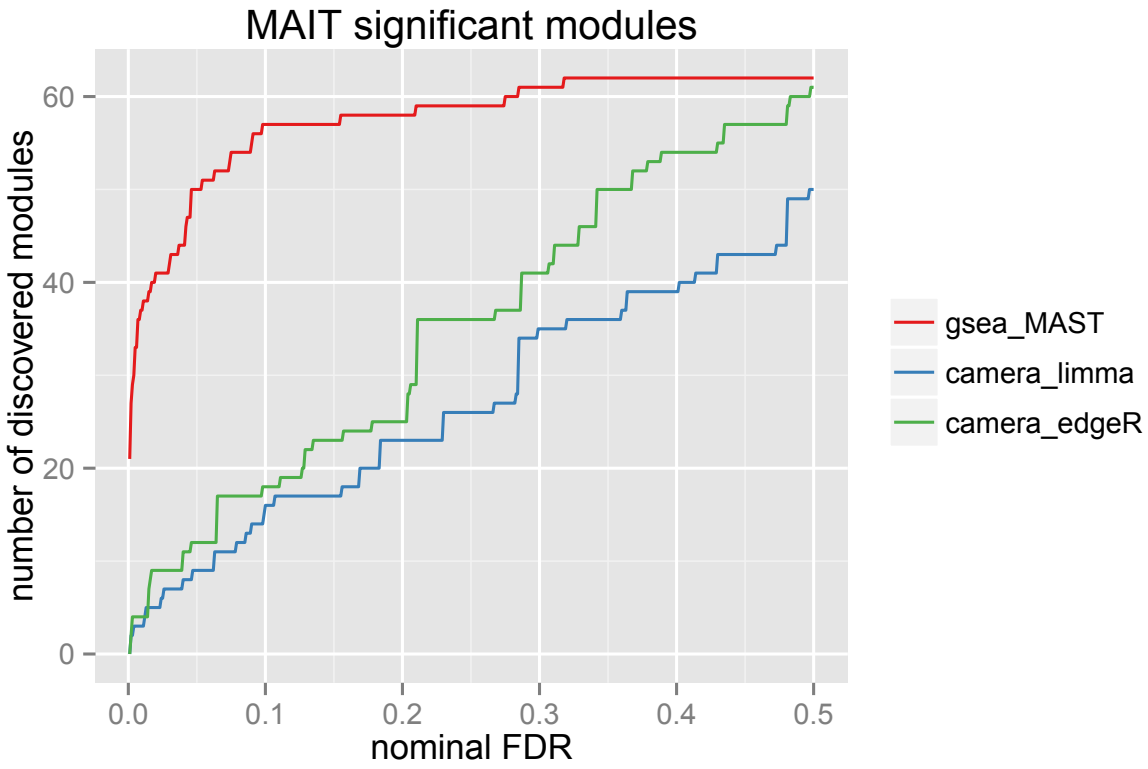

Figure A.12: Number of modules discovered plotted against FDR-adjusted significance of the module. MAST-based GSEA detects more modules than other methods.

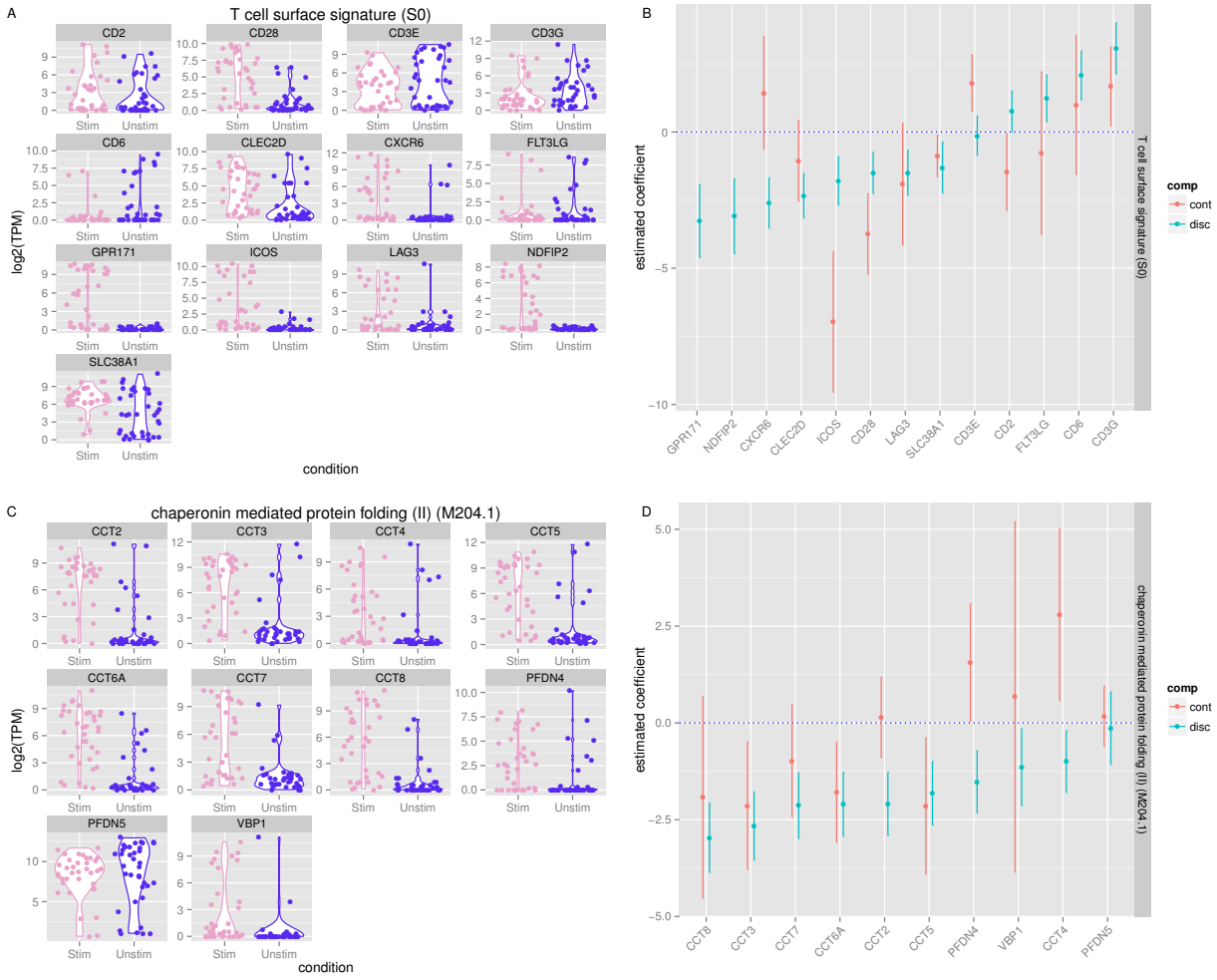

Figure A.13: Comparison of raw expression values ($\log_2$ TPM) and coefficients estimates (Unstimulated as reference) of modules identified as differentially expressed using MAST GSEA but not with CAMERA. Differences in the expression profile are evident, however CAMERA failed to detect them. A) Violin plots showing the expression of genes in the "T-cell surface signature" module. B) Model coefficient estimates for the genes in the 'T-cell surface signature module from GSEA, with 95% confidence intervals, from the discrete and continuous components of the model. C) Violin plots showing the expression of genes in the "chaperonin mediate protein folding" module. D) Model coefficient estimates for the genes in the chaperonin mediate protein folding module from GSEA, with 95% confidence intervals, from the discrete and continuous components of the model.

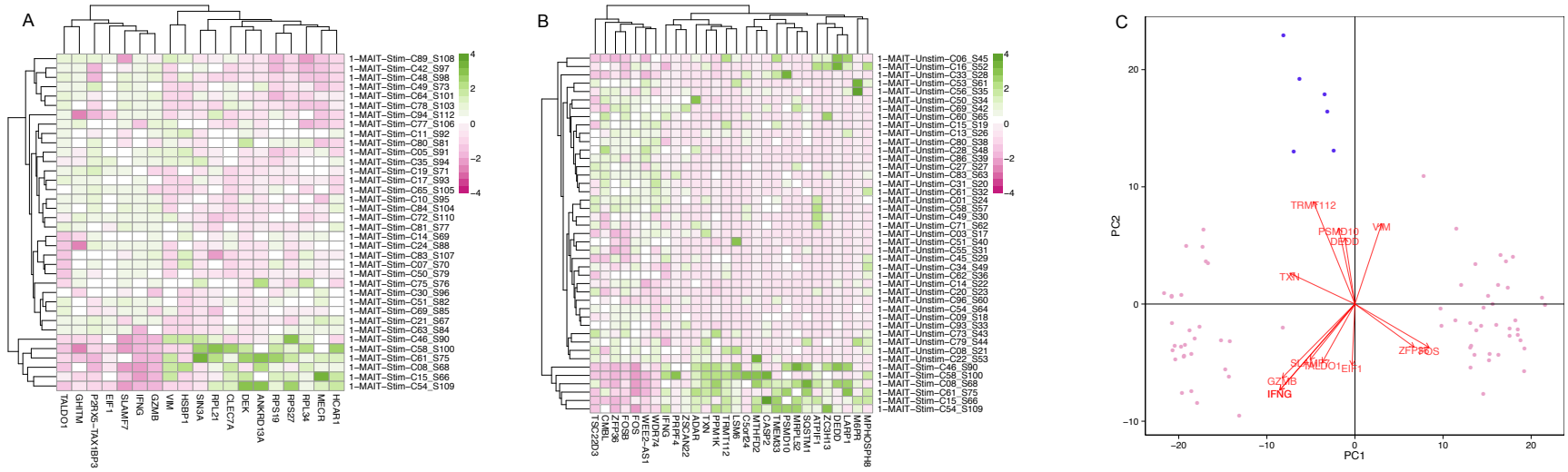

Figure A.14: The six stimulated MAIT cells that did not exhibit an expression profile indicative of activation are shown in comparison to A) other stimulated MAITs and B) unstimulated MAITs. Differentially expressed genes between these six cells and the stimulated but activated and non-stimulated cells are shown, identified using MAST at a q-value of 15% and fold change threshold of at least 2. Panel C) shows PCA of the MAITs based on the differentially expressed genes. 13 selected genes with largest loadings discriminating between the three classes of cells are shown.

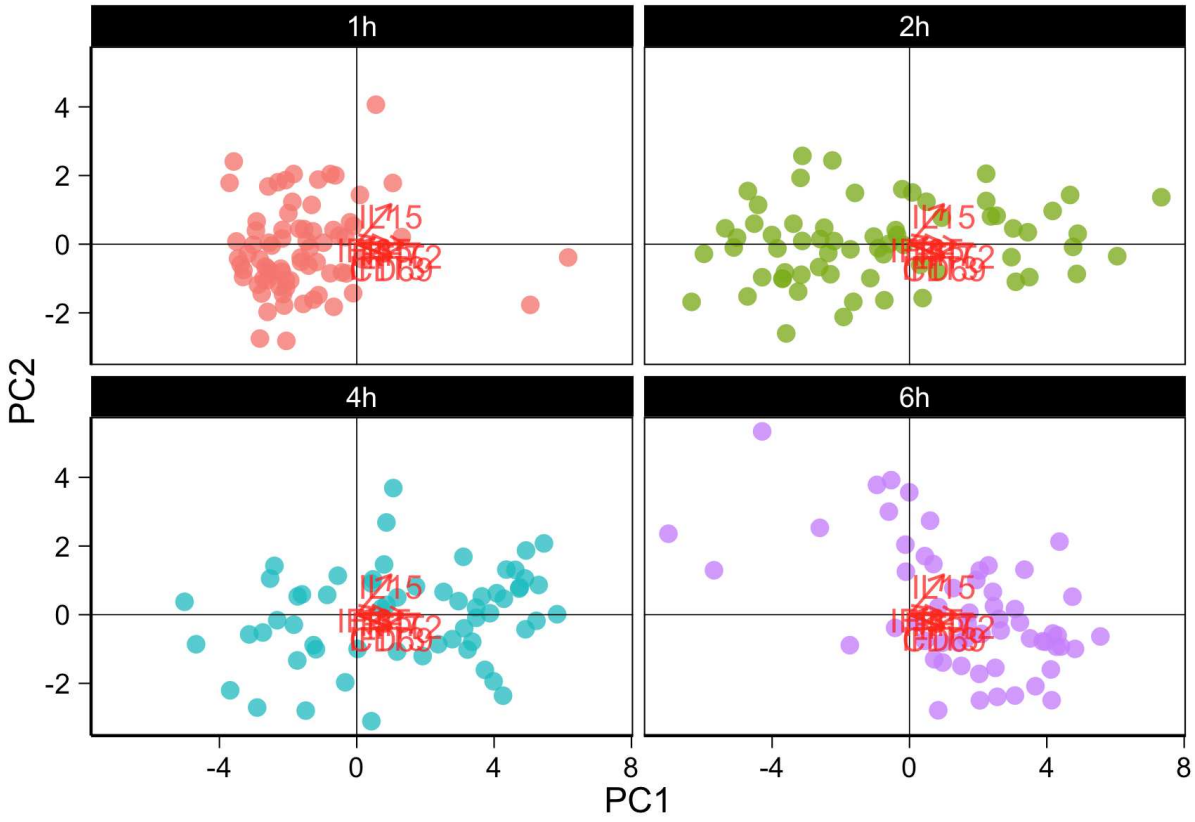

Figure A.15: PCA of the model residuals of LPS stimulated cells using the genes in the core antiviral module identified in Shalek et al. [2014]. The two "outlier" cells evident at the 1h timepoint correspond to the "early marcher" precocious cells described previously. These results show that these cells exhibit coordinated co-expression of genes in the core antiviral signature at the single-cell level.

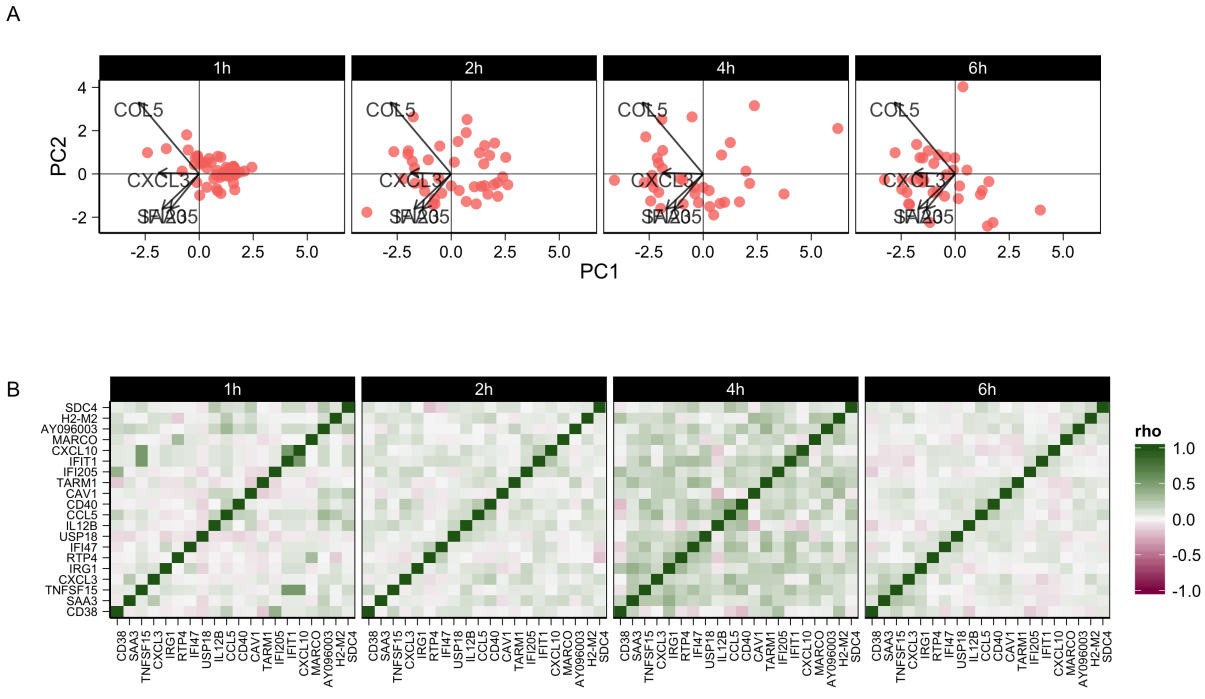

Figure A.16: Co-expression plot for PAM (synthetic mimic of bacterial lipopeptides) stimulated cells of cells in the mDC data. Panel A in each figure shows principal component analysis (PCA) of the model residuals using the top 100 differentially expressed genes. Cells are faceted by time, which is correlated with the first principal component. Panel B shows heatmaps of the pairwise correlations between genes in the model residuals across cells at each timepoint. The order of genes in the heatmaps is based on clustering at the 6h timepoint.

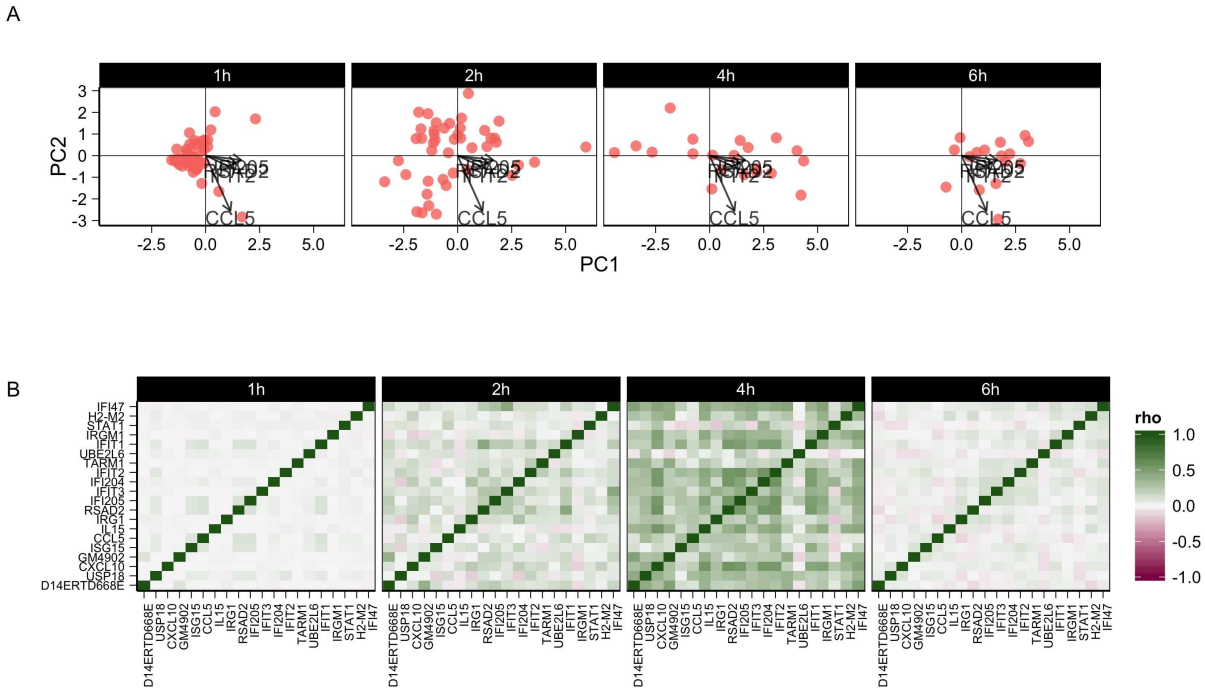

Figure A.17: Co-expression plot for PIC (viral-like double-stranded RNA) stimulated cells of cells in the mDC data. Panel A in each figure shows principal component analysis (PCA) of the model residuals using the top 100 differentially expressed genes. Cells are faceted by time, which is correlated with the first principal component. Panel B shows heatmaps of the pairwise correlations between genes in the model residuals across cells at each timepoint. The order of genes in the heatmaps is based on clustering at the 6h timepoint.