IMPROVED METHODS FOR THE ANALYSIS OF SINGLE-CELL RNA-SEQUENCING AND IMAGING DATA

Eric D. Van Buren

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2020

Approved by:

Yun Li

Di Wu

Yuchao Jiang

Michael Love

Kirk Wilhelmsen

ABSTRACT

Eric D. Van Buren: Improved Methods for the Analysis of single-cell RNA-Sequencing and Imaging Data
(Under the direction of Yun Li and Di Wu)

Two key challenges in the analysis of single cell RNA-seq (scRNA-seq) data are excess zeros due to "drop-out" events and substantial overdispersion due to stochastic and systematic differences. Association analysis of scRNA-seq data is further confronted with the possible dependency introduced by measuring multiple single cells from the same biological sample. To address these three challenges, the first chapter of this work proposes TWO-SIGMA: a TWO-component SInGle cell Model-based Association method for differential expression analysis of scRNA-seq data. The first component models the drop-out probability with a mixed-effects logistic regression, and the second component models the (conditional) mean read count with a mixed-effects negative binomial regression. Simulation studies and real data analysis show advantages in type-I error control and power enhancement over alternative approaches including MAST and a zero-inflated negative binomial model without random effects.

The second chapter of this dissertation expands the first to Gene set testing (GST). Here, we propose TWO-SIGMA-geneset to conduct competitive gene set testing, in which the genes in a given set are compared to the remaining collection of genes. Previous work has demonstrated that inter-gene correlation can substantially inflate type-I error. We provide an adjustment for inter-gene correlation, which is estimated using the residuals from the gene-level TWO-SIGMA model. Simulation studies show that type-I error is well controlled in a variety of representative scenarios, with or without inter-gene correlation present. Power is improved over state-of-the-art methods, including CAMERA, for a variety of scenarios consistent with real single-cell RNA-seq data.

Finally, the third chapter of this work studies chromosomal interactions at the single-cell level. First, we discuss the Hi-C technology for analyzing genome-wide chromosomal interactions. In particular, we focus on peak calling, in which the aim is to separate interactions between loci that are due to random chance from interactions that are not random. Second, we discuss state-of-the-art methods for single-cell imaging. We then show an example of a way to combine information from Hi-C and imaging data from Drosophilia embryos for peak calling using the Cauchy Combination Test. We conclude by discussing potential future research in this context.

ACKNOWLEDGEMENTS

I would like to thank my advisors, Yun Li and Di Wu, for their support and guidance over the past several years. Both have helped me tremendously in improving my understanding of statistics, genetics, and the research process, and I am grateful for their patience along the way. I would also like to thank my committee members Yuchao Jiang, Michael Love, and Kirk Wilhelmsen for their helpful suggestions. My time at UNC BIOS has strengthened me professionally and personally, and I appreciate the many friends I have made in my time in Chapel Hill.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DE	Differential Expression
GST	Gene Set Testing
HMM	Hidden Markov Model
HMRF	Hidden Markov Random Field
IGC	Inter-Gene Correlation
RNA-seq	RNA-sequencing
scRNA-seq	single-cell RNA-sequencing
TWO-SIGMA	TWO-component SInGle cell Model-based Association method
TWO-SIGMA-G	TWO-SIGMA for Gene set testing

CHAPTER 1: LITERATURE REVIEW

1.1 Sequencing Background

1.1.1 Bulk RNA-sequencing

High-throughput sequencing has allowed researchers to study the impact of the transcriptome on the molecular underpinnings of biological processes and diseases. Many such sequencing methods have been developed for a variety of purposes, but perhaps the most popular over the past decade has been bulk RNA-sequencing ("RNA-seq"). RNA-seq has become a fundamental tool in understanding of biological processes and genomic functions (Stark *et al.*, 2019; Conesa *et al.*, 2016). Initial steps in the pipeline are performed in a laboratory: RNA is extracted from a sample and messenger RNA (mRNA), which is usually of interest, is amplified to distinguish it from the less relevant ribosomal RNA. Sequencing follows and computational techniques are used to align read fragments to the genome, perform quality control, and produce a dataset consisting of read counts giving the amount of each gene that was observed. Samples consisting of thousands or tens of thousands of cells are pooled together, giving the aforementioned name "bulk" RNA-sequencing. Such pooling of the cells from a biological sample means that the expression levels studied constitute an average over cells that may have very different transcriptomic profiles due to many factors, such as cell type or cell cycle. Conclusions drawn using various computational and statistical techniques are then only generalizable to the aggregated population of cells.

1.1.2 Single-cell RNA-sequencing

Recently, single-cell RNA-sequencing ("scRNA-seq") technologies have been developed to allow sequencing of mRNA on each cell individually. In capturing transcriptomic variation at its fundamental level, this technology allows researchers to (i) study how cellular heterogeneity plays a role in disease etiology, (ii) discover new cell types, (iii) and make predictions of cellular development trajectories, among other items (Van den Berge *et al.*, 2018). Some of the most popular platforms for conducting single-cell sequencing include Drop-seq (Macosko *et al.*, 2015), Fluidigm C1 (https://www.fluidigm.com/products/c1-system), and the Chromium system from 10x Genomics (https://www.10xgenomics.com/solutions/single-cell/). Like RNA-seq, non-negative counts are produced. These counts tend to exhibit two major diverging characteristics from RNA-seq, however: larger variation and an excess of zero counts (Bacher and Kendziorski, 2016). This large variability is often imprecisely called "overdispersion"—more accurately the counts often exhibit excess variance relative to the Poisson distribution. It is well understood that only 10-30% of the mRNA expressed in a given cell will be captured successfully, and that distortions of the underlying truth likely occur as a consequence of this low capture rate (Hwang *et al.*, 2018; Yuan *et al.*, 2017; Haque *et al.*, 2017; Kelsey *et al.*, 2017; Buettner *et al.*, 2015). One such distortion is this presence of an excess number of zero read counts, relative to both what might be reasonably modeled using common discrete distributions and particularly as compared to bulk RNA-seq (Hicks *et al.*, 2017a; Bacher and Kendziorski, 2016; Chen *et al.*, 2019). One common approach in the literature to conceptualize zeros as being from two main sources: the first source is "biological," in that zero expression measurements occur due to stochastic biological factors (e.g. transcriptional bursting, cell cycle). The second source, commonly called "drop-out," is zeros that are mistakenly observed as a consequence of technical factors. (Wills *et al.*, 2013; Bacher and Kendziorski, 2016). These drop-out events tend to occur most often in genes with low mean expression, and also tend to differ in prevalence across different biological samples (Kharchenko *et al.*, 2014). This formulation of the source of observed zeros in scRNA-seq data will be discussed in more detail throughout this proposal.

One technical development of note is the development of Unique Molecular Identifiers (UMIs), which attach a unique barcode to each cell prior to sequencing (Chen *et al.*, 2018). UMIs can reduce the impact of biases created by repeated amplification of mRNA done before quantification, and may help to minimize the influence of these drop-out events in scRNA-seq data (Sena *et al.*, 2018; Townes *et al.*, 2019). Recent work has even suggested that if data was collected using UMI-based sequencing, droplet-based scRNA-seq data may not contain the excess zeros described above and further may contain very few drop-out events (Svensson, 2020). Nevertheless, there have been many methods developed which aim to borrow information and impute drop-out events based on the read counts of similar cells and/or genes prior to downstream analyses (Lin *et al.*, 2017; Gong *et al.*, 2018; Li and Li, 2018; Huang *et al.*, 2018; Tracy *et al.*, 2019). Although we do not focus on these methods extensively, they are important to consider and can be viewed as a supplement, rather than a competitor, to methods which use only observed expression values depending on the goals of the analysis.

The cellular detection rate (CDR) is defined in (Finak *et al.*, 2015) as the percentage of genes expressed over some background level of expression (often chosen to be zero). The CDR therefore has a biological interpretation as a cellular scaling factor and is a surrogate for both technical and biological variation. This confirms the conclusions of others that the CDR can explain a substantial proportion of observed expression variability and should be included in any association analysis of scRNA-seq data (Hicks *et al.*, 2017b). As such, including CDR as a covariate in a regression modelling framework using scRNA-seq data can help to control for unwanted technical variation and provide better estimates of effects of interest.

scRNA-seq datasets typically contain thousands of cells sequenced from a much smaller pool of individuals. This creates the potential for a repeated measures correlation that exists because multiple cells are sequenced from the same individual. This correlation will be discussed in more detail in the proposal section, however we find a gap in literature relating to methods that can specifically accommodate this correlation structure. At the gene level, such a within-subject correlation structure is additionally conflated with the gene-gene correlation that has been shown

to exist in scRNA-seq data (Buettner *et al.*, 2015). Consequences of such gene-gene correlation in tests of sets of genes in RNA-seq or scRNA-seq datasets will be discussed later in section 1.5 of this proposal.

1.2 Zero-Inflated Models for Count Data

Perhaps the most popular methods designed for data with an excess zeros are zero-inflated models, which have been applied to count data for nearly 30 years (Lambert, 1992; Greene, 1994). These models use zero-inflated distributions, which are a mixture of a point mass at zero and some second distribution, which can be discrete or continuous. For count-based sequencing studies, two of the most useful mixing distributions are Poisson and negative binomial. Following the notation used elsewhere (Greene, 1994; Hilbe, 2011), we can specify these two distributions in a nested fashion. Consider the following parameterization of the negative binomial probability mass function (p.m.f.) at a non-negative integer y:

$$\begin{aligned}\Pr(Y = y) &= f(y; \mu, \phi) \\ &= \frac{\Gamma\Big(y + \phi\Big)}{\Gamma(y+1)\Gamma\Big(\phi\Big)} \left(\frac{1}{1 + \frac{1}{\phi}\mu}\right)^{\phi} \left(\frac{\frac{1}{\phi}\mu}{1 + \frac{1}{\phi}\mu}\right)^{y}, \; y = 0, 1, 2, \ldots \end{aligned} \tag{1.1}$$

With this parameterization, E(Y) = μ and Var(Y) $= \mu + \frac{1}{\phi}\mu^2$, such that ϕ governs the amount of extra-Poisson variation and is thus called the "overdispersion" parameter ($\phi > 0$). This parameterization is appealing for interpretability because as $\frac{1}{\phi} \to 0^+$, the density above approaches the Poisson density with mean μ. Thus, the Poisson and negative binomial distributions are asymptotically nested (and nearly identical for large values of ϕ). One can test $H_0 : \frac{1}{\phi} = 0$ versus $H_a : \frac{1}{\phi} > 0$ using the likelihood ratio test, in which p-values come from a 50:50 mixture of χ^2_1 and a point mass at zero because the parameter $\frac{1}{\phi}$ is being tested on the boundary of the parameter space (Hilbe, 2011). Such a test is a useful way to increase model parsimony by using the Poisson distribution when the data does not suggest significant deviation from it.

Let p and μ be the probability of belonging to the degenerate zero component, and the mean read count conditional on being sampled from the non-zero component, respectively. The p.m.f. of one observation Y under the zero-inflated negative binomial (ZINB) distribution is given by:

$$\begin{aligned} \mathrm{P}(Y = 0) &= p + (1 - p) f(0; \mu, \phi) \\ \mathrm{P}(Y = y) &= (1 - p) f(y; \mu, \phi), y = 1, 2, 3, \ldots \end{aligned} \tag{1.2}$$

In this framework, observed zeros can therefore be thought of as being sampled from the (degenerate) point mass at zero with probability p or the negative binomial distribution f with probability $1 - p$. The ZINB distribution thus assumes that there are two sources of zeros in the data: the first source is the process that governs the "excess" zeros (called in the general literature "structural zeros") and the second source is from the negative binomial process (also called "sampling zeros"). Interpretations from the zero-inflated negative binomial model are quite natural for single-cell gene expression data because, as described earlier, it is reasonable to assume that some observed zeros are biological in origin while other zeros are drop-out events due to technical factors. Note that in such a mixture distribution the underlying source of any zero is an unknown, random variable. It is therefore not possible to use these models naively to determine whether a given zero is degenerate or random in nature, rather only to estimate a probability an observed zero is sampled from the point mass instead of the negative binomial distribution. Although semantic, the distinction regarding the source of zeros affects the interpretation of model coefficients and is important because these models are often misinterpreted by researchers as providing marginalized means when they actually provide means conditional on being in the second component (Preisser *et al.*, 2012; Todem *et al.*, 2016).

1.3 Generalized Linear Mixed Models

The generalized linear model (GLM) extends the standard linear regression model to distributions beyond the normal distribution with the use of various well-chosen link functions. These

models are used extensively in the analysis of genomic data and a variety of methods which employ GLMs will be discussed below. The generalized linear mixed model (GLMM) extends the GLM to include random effect terms, which allow model regression parameters to vary over, for example, different biological samples. Random effect terms are usually assumed to be distributed normally with mean zero and a variance that is referred to as a "variance component." Variance component(s) summarize the extent to which the model's intercept, for example, differs systematically across individuals. A large variance component suggests that individual heterogeneity in a given parameter exists, and a value of zero (or nearly zero) suggests that individuals do not deviate from the fixed effect parameter. GLMs assume that values from different samples are independent of one another. Including random effect terms via a GLMM relaxes this assumption, and helps to control for within-sample correlation when samples are not independent. The use of a GLMM can thereby improve mean parameter and standard error estimation to provide better control of type-I error (Fitzmaurice *et al.*, 2003).

1.3.1 Zero-Inflated Mixed Effects Models

A zero-inflated negative binomial mixed effects model (ZINBMEM) has previously been proposed for modelling count data (Min and Agresti, 2005). For the analysis of compositional microbiome data, a zero-inflated two-component mixed effects beta regression model has also been specified (Chen and Li, 2016). Both applications involve repeated measures data in which the number of repeated measures per individual is small relative to the number of individuals. This is in contrast to scRNA-seq datasets which include far more repeated measures (cells) than samples. Section 2.1 of this proposal will discuss how the performance of the ZINBMEM differs as the nature of the repeated measures data varies between these two extremes.

1.3.2 Different Ways to Account for Repeated Measures

In scRNA-seq datasets, which include a small number of biological samples relative to cells, it would be reasonable to alternatively consider controlling for biological sample as a fixed effect

rather than with a random effect. Most generally, there are two reasons to prefer incorporating a random effect to a fixed effect approach. First, it is often of interest to estimate a variance component that can apply to all samples in the population. This estimate can directly help to quantify the degree of heterogeneity in a population. Second, random effect terms explicitly help to control for the within-sample correlation, rather than providing only adjusted parameter estimates for included covariates. An alternative approach to accounting for such sample-level repeated measures using GLMMs would be to fit a marginal model with the generalized estimating equations (GEE) approach instead of a mixed effects model (Agresti, 2013). There are two main reasons to prefer incorporating random effect terms to a GEE approach: first, GEE-based models cannot be used for sample-level prediction, which can be a scientifically relevant task in the cellular context. Second, given that many scRNA-seq experiments are conducted over a small number of samples, it is likely that the empirical (sandwich) covariance estimate would underestimate the true standard errors and could thereby inflate type-I error (Agresti, 2013).

1.3.3 Evaluating the Need for Random Effect Terms

Evaluating the need for a random effect term in a GLMM typically involves a hypothesis test of whether the corresponding variance component(s) equal zero. Although Wald and score tests are possible, we will focus on the likelihood ratio test (LRT) to help with results discussed in section 2. Use of the LRT requires fitting the model under both the null and the alternative hypotheses, but is a preferred method to determine whether the random effect terms significantly improve model fit. In the linear mixed model, the asymptotic distribution of the LRT statistic for testing one variance component is a 50:50 mixture of χ^2_0 and χ^2_1 (Fitzmaurice *et al.*, 2003). P-values are then calculated as half of the usual p-value using the χ^2_1 distribution. This result holds in one-component GLMMs as well (Zhang and Lin, 2008), and in the case of testing one variance component in the ZINBMEM. To my knowledge the distribution when using the LRT to test two non-independent variance components in the ZINBMEM model has not been derived. One conservative choice is to use χ^2_2 as the reference distribution, although the true distribution is

likely a mixture of chi-square distributions as in other cases involving a GLMM (Zhang and Lin, 2008). This will be discussed more in section 2.1.

Other post-fitting options to compare models with and without random effects include information criteria like AIC and BIC or Wald tests of the variance components (Fitzmaurice *et al.*, 2003). Critically, all options discussed require fitting the "full" model including the random effect terms. As mentioned, the scRNA-seq application is distinct from typical repeated measures analyses in that the number of repeated measures (cells) typically far exceeds the number of samples. Such designs can entail more extensive computational time for each gene over scenarios involving a smaller number of repeated measures from a modest number of individuals. These computational burdens are especially relevant given that scRNA-seq data typically include thousands or tens of thousands of genes. In the proposal section we will discuss an *ad hoc* procedure that can bypass the need to fit the "full" model to identify the genes that are most likely to need the random effect terms in a computationally efficient manner.

1.4 Differential Expression Analysis

A task of primary interest in analyzing any high throughput sequencing that produces non-negative read counts, including RNA-seq and scRNA-seq, is to compare gene expression measurements across conditions of scientific or experimental interest. Identifying genes with different expression profiles between subjects with and without disease, for example, is a first step in explaining disease etiology and understanding phenotypic variation. This type of association is commonly referred to as *differential expression* (DE). Differential expression is a very general term, in that whether or not a gene is described as significantly differentially expressed depends on several factors, including the model used, its associated estimated measure (e.g. log fold change from a GLM), whether the data is represented on the original or a transformed scale, and the threshold used for judging significance.

1.4.1 Existing Methods for DE Analysis in Bulk RNA-seq

DE analysis is one of the most popular uses of RNA-seq data (Stark *et al.*, 2019). The first step in most methods developed for bulk RNA-seq data is typically some normalization of the read counts using one of many different methods. Quantile normalization was originally designed for microarray data, and uses observed quantiles to match distributions of observed counts across different sequencing runs (Smyth, 2005). The trimmed mean of M-values normalization method estimates scale factors between samples that can be used in DE pipelines without modifying the original data (Robinson and Oshlack, 2010). Other methods transform the count data into differing quantities such as Transcripts per Million (TPM) or FPKM (Fragments Per Kilobase per Million mapped reads) to account for technical biases including batch effects (Soneson and Delorenzi, 2013). The goal of all such transformations is to normalize the data across factors that could produce misleading results, such as transcript length, library size, or sequencing batch. We will revisit these transformations throughout the proposal. After such transformations, count-based distributions cannot be used.

Early methods designed for DE in microarray data, including voom and limma, consider the observed data (or some normalized and scaled version of it) to be from a log-normal distribution and use linear models for analysis (Law *et al.*, 2014; Smyth, 2005). While this distributional assumption may be suitable for the intensities measured in microarray experiments, it may be less realistic for RNA-seq data, which inherently measures a discrete count. Subsequent methods developed specifically for RNA-seq data most commonly use either the Poisson or negative binomial distribution to preserve the scale and properties of the data as much as possible. Perhaps the two most popular such methods are DESeq2 (encompassing its predecessor DESeq) and edgeR (Love *et al.*, 2014; Anders and Huber, 2010; Robinson *et al.*, 2009). Both methods employ negative binomial GLMs using a log link. Such models are advantageous for several reasons: (i) the negative binomial distribution captures the variability in expression observed with many biological replicates through its overdispersion parameter, (ii) the regression framework allows for the analysis of both simple and complex experiments, and (iii) the coefficients produced can

be interpreted naturally as log-fold changes. DESeq2 and edgeR differ in some default choices for filtering and outlier removal, and the normalization procedures used; DESeq2 normalizes data using the median of the ratio of the read count to the geometric mean read count across all genes while edgeR uses the TMM method described above. Many review papers have shown DESeq2 and edgeR to have excellent performance, which is particularly reassuring given that they seem to constitute the majority of DE analyses in RNA-seq data (Schurch *et al.*, 2016; Soneson and Delorenzi, 2013; Seyednasrollah *et al.*, 2013). As such, these negative binomial models constitute a well-supported starting point for DE methods for scRNA-seq data, and are sometimes used to benchmark DE methods designed for scRNA-seq data (Sekula *et al.*, 2019).

1.4.2 Existing Methods for DE Analysis in scRNA-seq Data

The excess zeros detailed in section 1.1.2, combined with the considerable biological and technical variation seen in scRNA-seq data, demand customized approaches for DE testing. As such, naively applying the methods described in section 1.4.1 is therefore considered insufficient (Bacher and Kendziorski, 2016) or at best equivalent to methods customized for scRNA-seq data (Soneson and Robinson, 2018). Many such methods have been developed for differential expression analysis in scRNA-seq data. Some early innovative Bayesian methods include SCDE, which utilizes a two-component negative binomial mixture method and scDD, which uses a Dirichlet mixture process (Kharchenko *et al.*, 2014; Korthauer *et al.*, 2016). Although both methods show strong performance in test cases, the former is limited to a two-group comparison, and the latter can only adjust for confounding covariates indirectly through a residualized analysis. DESingle employs a zero-inflated negative binomial distribution to detect differential expression (DE) in scRNA-seq data (Miao *et al.*, 2018). DESingle is also designed for DE detection with only a two-level grouping variable and does not employ a regression modeling framework to control for other covariates or account for within-sample correlation. Benchmarking papers typically limit themselves to two-group comparisons because most popular methods are limited to this case (Wang *et al.*, 2019; Soneson and Robinson, 2018).

ZINB-WaVE was designed for unsupervised settings in which a zero-inflated negative binomial model is used for dimension reduction (Risso *et al.*, 2018). One application of this method is to construct observation-level weights that can be incorporated into the generalized linear models mentioned earlier in the `DESeq2` or `edgeR` packages (Van den Berge *et al.*, 2018, 2017). The log transformations applied by these methods to the read counts have the benefit of reducing noise, which can mask true underlying biological signal, but may come at the expense of reduced interpretability when performing differential expression because the data is no longer perfectly in the gene space (Luecken and Theis, 2019). The ability to avoid transforming scRNA-seq data is particularly desirable given recent evidence which suggests that log transformation can distort many scRNA-seq datasets by specifically producing false variability and exaggerate the influence of zero counts (Townes *et al.*, 2019; Lun, 2018). Further recommendations suggest single-cell DE be analyzed using measured data while including relevant technical covariates such as the CDR or batch information as covariates in a regression model (Luecken and Theis, 2019).

MAST was introduced as a hurdle regression model for the analysis of scRNA-seq data (Finak *et al.*, 2015). Hurdle models are two-component models which mix a degenerate zero component with a distribution that is either continuous or left-truncated at zero for the positive expression component (Mullahy, 1986). Specifically, MAST models the log of TPM as normally distributed using a linear regression model, and separately models zero counts with a logistic regression model. As applied to sequencing data, the hurdle model thus does not allow zero expression measurements due to biological variation, and restricts zeros to derive from the degenerate zero distribution. This is in contrast to the zero-inflated negative binomial model mentioned above, which allows zeros to result from both components. Given the variability in the transcriptome that has been demonstrated to exist at the single-cell level, this restriction may sacrifice plausibility for some datasets. As discussed above, because the mean expression component involves a log transformation, there is further potential to distort the true biological signal present (Townes *et al.*, 2019). Although the ability to include random effect terms in either component of MAST is mentioned, they do not prioritize their inclusion in the model formulation and do

not evaluate the impact of random effects on the model's performance. MAST is considered to be one of the preferred methods for performing DE in scRNA-seq data given its regression modelling flexibility (Luecken and Theis, 2019).

Finally, a Bayesian hurdle model with cellular-level random effects was introduced for DE analysis of scRNA-seq data (Sekula *et al.*, 2019). This method is innovative in that it allows the cellular-level random effects to be either correlated with one-another via a compound symmetric structure or independent. Specifically, random effect terms are assumed correlated when the cells are assumed to come from the same latent subpopulation. These subpopulations are designed to represent unobserved groupings such as cell type and can be determined by a technique such as clustering before fitting the Bayesian hurdle model. When assumed independent, each cell has a random effect which is assumed to have mean zero and a common variance with other cells. This Bayesian hurdle model is somewhat restrictive, however, in that it does not allow for covariates other than CDR and a two-group treatment. Therefore, more complex experimental designs may not be fully analyzable using this method.

1.5 Gene Set Testing Background

Analyses of RNA-seq or scRNA-seq data typically begin at the gene level where, for example, one can generate test statistics summarizing the evidence of DE and rank all genes by the strength of such evidence. Most biological phenomena are understood to occur via interactions of many different genes (Barry *et al.*, 2005). Gene set analysis, also known as pathway analysis, therefore aims to put results in a broader biological context by studying expression changes in sets of genes. The purpose of these analyses is not to cluster genes into sets—this must be done *a priori*—but rather to put gene-level results into a biologically interpretable context. These analyses fundamentally require the constructed gene sets to represent meaningful biological pathways, and therefore require well curated and scientifically justified sets (Barry *et al.*, 2005). The Molecular Signatures Database (mSigDB) (Subramanian *et al.*, 2005; Liberzon *et al.*, 2015) provides the most extensive collection of gene sets aggregated from many different contributors and

categorized into eight major collections: hallmark (H), positional (c1), curated (c2), motif (c3), computational (c4), GO (c5), oncogenic (c6), and immunologic (c7). Sets are available for a variety of organisms, but of particular interest to this work are humans and mice. Typical sets consist of tens or hundreds of genes, but sizes can vary from two genes to over two thousand genes. There are several reasons for researchers to use set-level analyses as an addition to the gene-level analysis. First, because sets are constructed to represent shared annotations or functions, analyses can utilize previous biological knowledge in a useful manner. Second, in sharing strength across many genes, set-based analyses can improve statistical power and reduce the detection of spurious associations as compared to gene-level tests (Efron and Tibshirani, 2007; Gaynor *et al.*, 2019). Third, set-level analyses can increase reproducibility across experiments, which is often lower than desired due to the biological and technical variability present in RNA-seq and scRNA-seq data (Efron and Tibshirani, 2007; Gaynor *et al.*, 2019).

Broadly, set testing methods have two main steps. First, gene-level statistics are collected for all genes in the dataset. Second, the gene-level statistics are aggregated into some set-level (or "global") statistic which is used to calculate a set-level p-value. Two primary differences that exist between early gene set testing methods are the definition of the null hypothesis and calculation of the set-level p-value (Goeman and Buhlmann, 2007). It is now common to distinguish between "competitive" and "self-contained" tests and based on the null hypotheses of interest. Competitive tests compare the evidence of differential expression of a gene set to the evidence in some other reference set of genes; typically this reference set consists of either all other genes or a random sample of them. Self-contained tests, in contrast, compare the gene set to some fixed standard (usually the case of no DE) that does not incorporate the information from other genes. Competitive tests therefore compare the relative significance of gene sets to one another and thereby rank biological pathways by importance to the phenotype. Self-contained tests are most relevant for determining the significance of an individual biological pathway without making a relative comparison to other pathways (Wu and Smyth, 2012). Many early self-contained testing procedures are vulnerable to the scenario in which a larger test set leads to smaller p-values, even

in the cases where genes are chosen at random (Barry *et al.*, 2008; Khatri *et al.*, 2012). For this reason, combined with interpretability and method availability, competitive tests are far more common in the literature today (Wu and Smyth, 2012; Goeman and Buhlmann, 2007).

Most competitive methods permute either samples or genes to construct a sampling distribution and calculate set-level p-values (Goeman and Buhlmann, 2007). Inherently, permutation choice is linked to type of null hypothesis being tested. Sample permutation constructs a sampling distribution by randomly permuting sample-level variables, for example treatment arm, within the same set of genes. Such a within-set sampling distribution is useful to test a gene set in isolation, as in self-contained tests. Gene permutation, in contrast, constructs a sampling distribution by randomly allocating genes to the test set without permuting the samples. This is useful for a comparison between gene sets, as in competitive testing (Wu *et al.*, 2010; Wu and Smyth, 2012). Using such a gene-based permutation distribution to estimate p-values is equivalent to assuming independence between genes. This assumption is critically flawed and unrealistic, even in well-controlled microarray, RNA-seq, and scRNA-seq data (Wu and Smyth, 2012; Wu *et al.*, 2010; Goeman and Buhlmann, 2007; Finak *et al.*, 2015). It is particularly unreliable for gene set testing because sets are explicitly constructed to represent biological pathways, and it is natural to assume that genes in a given pathway are more correlated than a random collection of genes (Wu and Smyth, 2012). The presence of such an inter-gene correlation (IGC) can dramatically inflate type-I error or the false discovery rate (FDR) through inducing a correlation in the marginal gene-level statistics (Barry *et al.*, 2008; Gatti *et al.*, 2010; Wu and Smyth, 2012; Efron and Tibshirani, 2007). It is therefore essential that any method for gene set testing adequately account for the IGC to provide statistically rigorous set-level p-values.

1.6 Existing Methods for Gene Set Testing

Some of the earliest methods for gene set testing looked for over-representation of genes deemed significant in the test set using, for example, Fisher's exact test and a 2x2 table (Barry *et al.*, 2008). This approach is limited because it requires a hard cutoff to call a gene as DE or not

and is therefore highly sensitive to the cutoff choice. This kind of approach also does not use the strength of association, except to the extent that it passes the chosen threshold, and completely ignores any IGC present.

GSEA was developed as one of the first gene set testing methods developed for two-group comparisons of expression data (Subramanian *et al.*, 2005). The method has proven incredible popular, as seen by the fact that it has been citepd over 19,000 times. First, an enrichment score is calculated using the Kolmogorov-Smirnov statistic to represent the extent to which a given test set is over-represented at the extremes of the ranked list of genes. At the set-level, a hybrid of the competitive and self-contained null hypotheses is tested via sample permutation followed by comparison of the set-level statistic to those of other gene sets. One weakness of GSEA is that the null hypothesis being tested is not straightforward to precisely define given its hybrid nature. Larger gene sets can often be more significant, and other gene sets not being tested can influence results in ways that are counterintuitive (Damian and Gorfine, 2004; Tian *et al.*, 2005). Further work extended GSEA beyond reliance on the Kolmogorov-Smirnov statistic to include other more general test statistics, and to one-way ANOVA designs (Efron and Tibshirani, 2007; Oron *et al.*, 2008). A further extension, later implemented in a function called `sigPathway`, used a similar formulation as GSEA but used refined normalization for the set-level statistics before using gene permutation to estimate p-values (Tian *et al.*, 2005). Although an improvement to GSEA, some studies have shown that `sigPathway` can fail to preserve the FDR or type-I error rate (Tarca *et al.*, 2008; Wu and Smyth, 2012). SAFE was developed shortly after GSEA for a two-group comparison using an ordinary t-test or some related variant, the Wilcoxon test, or the F-test from ANOVA to get gene-level statistics (Barry *et al.*, 2005). Sample permutation was then used to estimate set-level p-values while accounting for IGC. Follow-up review studies have demonstrated that SAFE preserves type-I error and FDR but suffers from lower power than competing methods (Mathur *et al.*, 2018; Tarca *et al.*, 2008). All of the methods described here are limited to simple experiments in which test statistics relate to a categorical group comparison (two-group only except SAFE).

CAMERA was proposed as a method for competitive gene set testing method for microarray or RNA-seq data (Wu and Smyth, 2012). Gene-level statistics are first constructed using a linear model, with options including the usual t-statistic or a moderated t-statistic designed to provide more reliable estimates of the variance of the coefficients. The use of the linear model means that CAMERA can accommodate complex experimental designs beyond a two-group comparison. Set-level p-values are then computed using modifications of the t-test or Wilcoxon rank-sum test that allow for a common pairwise correlation in the test set. An efficiently computed estimate of the variance inflation factor is used to estimate this common pairwise correlation. Rather than use the raw data for this calculation, CAMERA uses the residuals from the linear model. Using the residuals means that the variation in gene expression explained by the covariates is removed, helping to give the most reliable estimate of the correlation between the gene-level statistics in the test set. By avoiding permutation, and unlike some early approaches, CAMERA provides a statistically valid, computationally efficient test of a precisely defined and fully specified null hypothesis (Goeman and Buhlmann, 2007). The hypothesis corresponds to a test that the average absolute value of each coefficient in the test set is larger in magnitude as compared to the reference set.

MAST, which was developed for scRNA-seq DE analysis and discussed in section 1.4.2, has an extension to allow for competitive gene set testing comparing a test set to its complement (Finak *et al.*, 2015). This extension is quite flexible given the regression-based framework employed by MAST. Once the gene-level statistics are collected, the bootstrap is used to estimate the inter-gene correlation of the regression coefficients. Set-level tests are conducted using the Z-test and computed separately for the two components of the hurdle model. The use of the bootstrap is computationally intensive and subject to variability based on the number of bootstrap samples used. Only single regression coefficients are testable using MAST's competitive gene set procedure.

Finally, I will briefly discuss several methods that have been recently developed for gene set analysis with a different emphasis to the set-level DE detection we discussed above. A deep

learning framework was proposed for gene set inference (Lukassen *et al.*, 2019). The method is a broad tool which can simultaneously integrate gene set inference and batch effect correction. One limitation of this deep learning framework is that rigorous statistical testing of gene sets, similar to that of CAMERA, for example, is not readily available. The PAGODA method was developed as an extension of the SCDE method discussed in section 1.4.2 to search for gene sets that exhibit coordinated overdisperion (Fan *et al.*, 2016). It can utilize both previously annotated gene sets or *de novo* gene sets. Similarly, the SCENIC method was developed for single-cell regulatory network construction (Aibar *et al.*, 2017). The primary focus of these methods is to analyze variance patterns in search of transcriptional heterogeneity and not detecting sets of genes that show systematic differential expression.

1.7 Hi-C and Imaging Methods for 3D Structure Recovery

Analyzing spatial interactions within chromatin can help researchers understand the interactions between genomic loci that are close in three-dimensional (3D) distance but may be farther in linear (genomic) distance. These interactions can be used to reconstruct 3D structures, learn about gene regulation and epigenetic signatures, and compartmentalize genomic regions into active or closed chromatin regions (Varoquaux *et al.*, 2014; Oluwadare *et al.*, 2019; Hu *et al.*, 2013). Pairs of regions/loci that interact frequently are thought of as "close" to one another in distance in the 3D space, although there is not always a clear relationship between these two ideas (Lajoie *et al.*, 2015; Fudenberg and Imakaev, 2017). One of the earliest techniques to understand 3D chromosomal structure was a microscopic method called multicolor fluorescent in situ hybridization (FISH). As an imaging based approach, FISH can directly record spatial distance in single-cells to reconstruct 3D structure. FISH relies on prior specification to identify regions to study and has difficulty detecting *cis* interactions that occur within 100 kilobases due to low-resolution, low-throughput data (Oluwadare *et al.*, 2019; Hu *et al.*, 2013). Furthermore, a genome-wide analysis is currently not feasible using FISH (Fudenberg and Imakaev, 2017).

Early sequencing-based techniques, referred to as chromosome confirmation capture (CCC) methods, included 3C, 4C, and 5C. All of these innovative techniques have practical restrictions on either the genomic loci analyzed or the size of the region in which to investigate any chromatin-chromatin interaction (Lajoie *et al.*, 2015). "Hi-C" is a more recent technique that uses high-throughput sequencing to simultaneously reveal genome-wide pairwise interactions between loci. Both CCC methods and Hi-C are used to identify topicologically associated domains (TADs), which are megabase-long self-interacting regions of the genome that tend to exhibit more intra-region interactions than inter-region interactions (Mateo *et al.*, 2019; Pal *et al.*, 2019; Oluwadare *et al.*, 2019). TADs are considered a fundamental piece of proper gene regulatory functioning, and are highly reproducible across different Hi-C experiments (Finn *et al.*, 2019). It has thus been hypothesized that TADs comprise micro-environments within which promoters interact with enhancers (Spielmann *et al.*, 2018; Lajoie *et al.*, 2015). Disruption of TAD boundaries has been associated with many negative health outcomes such as cancer, limb deformities, and adult-onset degradation of brain white matter (Spielmann *et al.*, 2018; Krumm and Duan, 2019). It is therefore of scientific interest to understand and discover TADs.

Typically, data from a Hi-C experiment is represented as a symmetric matrix (often called the "contact matrix") which gives the number of interactions (or contacts) between loci. Differing loci are usually binned into fixed-sized genomic intervals ranging from 40 kB to 1 MB to reduce sparsity and increase coverage, with some recommendations suggesting that all bins should have at least 1,000 reads (Varoquaux *et al.*, 2014; Lajoie *et al.*, 2015). The bin size can vary depending on the objectives of the analysis, which can vary from genome-wide searches to studies of specific small-scale regions. There are many systematic biases that impact the quality and mappability of Hi-C data, included restriction enzyme cutting frequencies, GC content, and sequence uniqueness (Yaffe and Tanay, 2011; Hu *et al.*, 2012).

In Hi-C, interactions are measured over some aggregated population of cells. Thus, it is impossible to distinguish interaction pairs that only occur simultaneously in a given cell, are mutually exclusive in an individual cell, or are somewhere in between (Lajoie *et al.*, 2015). Se-

quencing and analysis pipelines have been developed to extend Hi-C to single-cells to the study of cell-to-cell variability in chromatin structure and spatial genome organization (Nagano *et al.*, 2013). As compared to bulk approaches, however, single-cell Hi-C data suffers from low resolution (as measured by a low number of total reads), with state-of-the-art single-cell approaches typically covering only 5-10% of the genome (Mateo *et al.*, 2019; Zhou *et al.*, 2019). When reported as a two-dimensional contact matrix, this means that less than 1% of all possible contacts can be captured at the single-cell level. Methods specifically designed to analyze single-cell Hi-C data are somewhat limited, although recent work showed promise in clustering single-cell Hi-C data to form constituent cell types for further downstream analyses (Zhou *et al.*, 2019).

Evidence from imaging studies, Hi-C studies, and integrative analyses of both suggests that cell-to-cell variability in spatial genome organization is substantial across space and time, even in cell populations that would appear to be functionally homogeneous (Zhou *et al.*, 2019; Ramani *et al.*, 2020; Finn *et al.*, 2019; Mateo *et al.*, 2019). Therefore, understanding this variability requires techniques that can provide data at the cellular level. Optical reconstruction of chromatin architecture (ORCA) was developed as one option for identifying new TADs with Hi-C while simultaneously using the high resolution and tissue imaging information information provided by FISH (Mateo *et al.*, 2019). In using microscopy, the chromatin structure of a few thousand single cells can be simultaneously analyzed to investigate chromatin structure and detect interactions between enhancers and promoters at a fine resolution. Cell types can be determined through the simultaneous measurement of mRNA and nascent transcription, and thousands of cells can be processed quickly.

Previous work in Hi-C has demonstrated the counterintuitive result that, for some loci, physical distance and contact frequency may not always be inversely related as expected (Fudenberg and Imakaev, 2017). This work concluded that contact frequency and spatial distance should be considered separately in integrative analyses of Hi-C and FISH data. Further integrative analyses of physical distance and interaction frequency suggest that genomic distance is not sufficient to explain all interactions observed (Finn *et al.*, 2019). This analysis suggested that gene dense re-

gions seem to display a larger amount of cell-to-cell heterogeneity in interaction frequencies than gene poor regions.

CHAPTER 2: TWO-SIGMA: A TWO-COMPONENT SINGLE CELL MODEL-BASED ASSOCIATION METHOD FOR SINGLE-CELL RNA-SEQ DATA

2.1 TWO-SIGMA for DE Analysis in scRNA-seq Data

Let $i = 1, ..., n$ index biological sample (individuals) and $j = 1, ..., n_i$ index the single cells associated with each biological sample. Following the notation in section 1.2, we will use a zero-inflated negative binomial GLMM to model the probability of dropout p_{ij} and the conditional mean read count μ_{ij}. Our proposed TWO-component SInGle cell Model-based Association method (TWO-SIGMA) for single-cell RNA-seq data is given by:

$$\begin{aligned} &\text{logit}(p_{ij}) = \boldsymbol{z_{ij}^T}\boldsymbol{\alpha} + a_i, a_i \sim N(0, \sigma_a^2) \\ &\log(\mu_{ij}) = \boldsymbol{x_{ij}^T}\boldsymbol{\beta} + b_i, b_i \sim N(0, \sigma_b^2), \text{ assume } a_i \perp\!\!\!\perp b_i \end{aligned} \tag{2.1}$$

The model is fit for each gene individually, so all parameters are gene-specific. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are fixed effect coefficient vectors and the corresponding vectors of sample-level and/or cell-level covariates $\boldsymbol{z_{ij}}$ and $\boldsymbol{x_{ij}}$ can be different. a_i and b_i are sample-specific intercepts (discussed more in the next section). Prediction of sample-specific intercepts and estimation of the variance components σ_a^2 and σ_b^2 allow us to investigate heterogeneity among individuals, and tests of whether the variance components equal zero allow us to separately (or jointly) evaluate the need for random effects. Separate variance components are estimated because the different link functions in the two components correspond to linear predictors with different scales. Including the random intercept terms also helps control for any within-sample correlation, providing more accurate estimates and standard errors of fixed effect parameters.

As part of our `twosigma` R package, we employ the `glmmTMB` package (Brooks *et al.*, 2017) to fit the model specified in equation (2). This package is well-suited to fit generalized linear mixed models (GLMMs) because the user can easily specify an arbitrarily complex model composed of fixed and random effects. The marginal likelihood $\mathrm{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \phi, \sigma_a^2, \sigma_b^2)$ of the TWOSIGMA model is given by

$$\prod_{i=1}^{n} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} \Bigg([\mathrm{P}(Y_{ij} = 0)]^{\mathrm{I}(y_{ij}=0)} [\mathrm{P}(Y_{ij} = y_{ij})]^{\mathrm{I}(y_{ij}>0)} \times g(a_i, b_i \mid \sigma_a^2, \sigma_b^2) \mathrm{d}a_i \mathrm{d}b_i \Bigg)$$

where $g(a_i, b_i \mid \sigma_a^2, \sigma_b^2)$ is the product of two normal densities (assuming $a_i \perp\!\!\!\perp b_i$), and $P(Y_{ij})$ is as in equation (1.2) substituting p_{ij} and μ_{ij} for p and μ, respectively.

Because no analytic solutions to this integral are available, the marginal likelihood must be approximated to obtain parameter estimates. Models that include many random effects can be fit efficiently using the implementation in the `twosigma` package because the Laplace approximation is used to integrate out random effects and automatic differentiation is used to compute gradients (Skaug and Fournier, 2006). It can be shown that the Laplace approximation is equivalent to using Gaussian quadrature with one quadrature point (Fitzmaurice *et al.*, 2003). Although estimates can be biased, the Laplace approximation often performs well for count response variables (Diggle *et al.*, 2002). For further comments on situations in which the Laplace approximation performs suitably well in practical applications, including the analysis of count data, see (Breslow and Clayton, 1993). Finally, others have demonstrated that the Laplace approximation works quite well in non-linear mixed-effects models (Pinheiro and Bates, 1995). This framework also does not require balanced data, as is sometimes assumed for mixed-effects models; for instance, balanced data are implicitly included in the setup of (Chen and Li, 2016).

The likelihood ratio test can be used to test several types of hypotheses in equation (2.1). As mentioned in sections 1.2 and 1.3.3, respectively, one can use the LRT in testing $\phi = 0$ and either of the variance components marginally (either $H_0 : \sigma_a = 0$ or $H_0 : \sigma_b = 0$) with p-values coming

from an equal mixture of the χ^2_0 and χ^2_1 distributions. A joint test of the variance components given by $H_0 : \sigma_a = \sigma_b = 0$ is conservative if taking p-values from the χ^2_2 distribution.

To summarize, TWO-SIGMA can control for different covariates in each component, incorporate random effects to accommodate within-sample dependency, analyze unbalanced data, and allows for zero-inflated and overdispersed counts. The regression modelling framework controls for additional covariates and provides the ability to examine any DE hypothesis that can be expressed as a contrast of regression coefficients. The implementation of the model strikes a balance between computational accuracy and efficiency, even as the number of random effects (number of samples in the context of the scRNA-seq data) or the number of single cells per sample increases.

2.1.1 Performance of TWO-SIGMA

To evaluate the performance of TWO-SIGMA, we simulated data in a variety of scenarios. Although many methods exist for DE in scRNA-seq data, as described in section 1.4.2, we chose to focus our comparison to MAST because, like TWO-SIGMA, it is designed using a regression modeling framework that is suitable for complex designs beyond a two-group comparison and can incorporate multiple cell-level and subject-specific covariates. We also compare to a ZINB model without random effects to highlight the impact random effect terms can have on model performance. Simulated additional covariates included age and the CDR (discussed in section 1.1.2) to mimic our real data analysis. Values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ were designed to represent realistic parameter values observed in our pancreatic data analysis. Models were evaluated using the likelihood ratio test on the joint null hypothesis that a binary disease status indicator is not associated with expression through either drop-out probability or the conditional mean, $\mathrm{H}_0 : \alpha_1 = \beta_1 = 0$. We consider two different ways of simulating data: one in which the number of samples far exceeds the number of cells per sample, as is typical in most repeated measures contexts, and the other in which the number of cells far exceeds the number of samples, as is the case in scRNA-seq data.

In each scenario we simulated 10,000 genes, each with 50,000 cells, and used 0.05 as the nominal significance rate to evaluate type-I error and power.

Table 2.1: Type-I error evaluations in simulated data: Shows type-I error using the LRT to test the joint null hypothesis of a simulated binary disease status indicator, $H_0 : \alpha_1 = 0, \beta_1 = 0$ versus $H_a : \alpha_1 \neq 0$ or $\beta_1 \neq 0$, with a significance level of 0.05. "T-S" refers to TWO-SIGMA, ZINB refers to a zero-inflated negative binomial model without random effects and MAST refers to the model described in Finak *et al.* (2015). 10,000 genes were simulated.

				Case 1 in Supplement:			Case 3 in Supplement:			Case 4 in Supplement:		
	Sim Params			50 Cells per 1000 Ind.			500 Cells per 100 Ind.			2000 Cells per 25 Ind.		
	ϕ	σ_a	σ_b	T-S	ZINB	MAST	T-S	ZINB	MAST	T-S	ZINB	MAST
No R.E.	10	0	0	**0.049**	0.051	0.089	**0.042**	0.050	0.090	**0.041**	0.052	0.090
	2	0	0	**0.048**	0.051	0.080	**0.038**	0.044	0.079	**0.041**	0.052	0.086
	1	0	0	**0.048**	0.052	0.081	**0.044**	0.051	0.087	**0.042**	0.051	0.090
Small R.E.	10	0.1	0.1	**0.051**	0.132	0.144	**0.056**	0.534	0.313	**0.077**	0.795	0.487
	2	0.1	0.1	**0.051**	0.078	0.089	**0.057**	0.323	0.176	**0.072**	0.643	0.361
	1	0.1	0.1	**0.049**	0.066	0.095	**0.053**	0.224	0.174	**0.075**	0.548	0.361
Large R.E.	10	0.5	0.5	**0.051**	0.621	0.290	**0.055**	0.941	0.716	**0.076**	0.984	0.875
	2	0.5	0.5	**0.053**	0.505	0.275	**0.056**	0.909	0.685	**0.076**	0.974	0.857
	1	0.5	0.5	**0.050**	0.404	0.247	**0.053**	0.873	0.649	**0.074**	0.964	0.827

Table 2.1 shows results from simulations in which the true values of the overdispersion parameter ϕ and the variance components σ_a, σ_b vary. Type-I error is well-controlled for TWO-SIGMA in the scenarios involving more individuals than cells. When the number of cells increases, type-I errors from TWO-SIGMA are slightly inflated over the nominal rate of 5%, but consistently remain superior to the results from the ZINB model or MAST in the presence of non-zero variance components. For example, the last row of table 2.1 shows that, when $\phi = 1$ and $\sigma_a = \sigma_b = 0.5$, type-I error for TWO-SIGMA increases from 0.05 to 0.053 to 0.074 as the number of individuals decreases from 1000 to 100 to 25. In contrast, the ZINB model and MAST have inflated type-I errors in every scenario that increase to nearly 1 as the number of individuals decreases. This is not surprising because both of the latter methods cannot account for any within-sample dependency structure among the single cells from the same sample. Ignoring the dependency introduced by even a moderate random effect size can thus have a drastic impact on the type-I error. When true variance components are zero, both TWO-SIGMA and the ZINB model preserve type-I error while MAST consistently has higher type-I error, as seen in the first three rows of table 2.1. Coverage of confidence intervals for α, β, and ϕ always approaches the

nominal level (see tables A.1-A.4 in section A.1.1 of the appendix). The reason for the slightly inflated type-I error for TWO-SIGMA observed in the scenario with 25 individuals is worth mentioning briefly. The smaller number of individuals (25) provides less information to estimate the sample-specific variance components σ_a and σ_b and fewer unique values of the simulated binary disease status indicator. The slightly lower coverage for variance components in the last 6 sets of case 4 in the appendix (table A.4) is one illustration of the (relative) difficulty in getting precise variance component estimates. TWO-SIGMA outperforms MAST or the ZINB model in preserving type-I error and estimating parameters under a variety of sample size breakdowns and with a variety of true parameter values. Figures A.3 and A.4 in section A.1.1 of the appendix confirm that type-I error is preserved across more stringent significance thresholds for representative scenarios.

2.1.1.1 Power Improvement under a variety of scenarios

Because the ZINB model and MAST both have heavily inflated type-I errors in many cases, using raw (or "apparent") power does not provide a fair comparison for these two methods. For each method and each simulation setting under the null, we therefore calculate the empirical significance threshold, defined as the test statistic value at the quantile associated with 1 minus the significance level. A percentage of statistics equal to the nominal significance level will then be larger than this threshold. For various alternative hypotheses, we calculate "true" power for MAST and the ZINB model by using the empirical significance threshold from the corresponding setting under the null as the rejection threshold instead of a usual theoretical threshold (e.g. 5.9915 from χ^2_2 at the .05 level). Figure 2.1 plots raw power for TWO-SIGMA and true power for MAST in the ZINB model in the following four scenarios: effect in both components, in either the same or opposite directions, and effects in only one of the two components. In the first three scenarios, MAST consistently has the lowest power, while TWO-SIGMA and the ZINB model have very similar power in the first two scenarios, beginning at around 20% and increasing to nearly 100%. The ZINB model has higher power than TWO-SIGMA in the third scenario but the

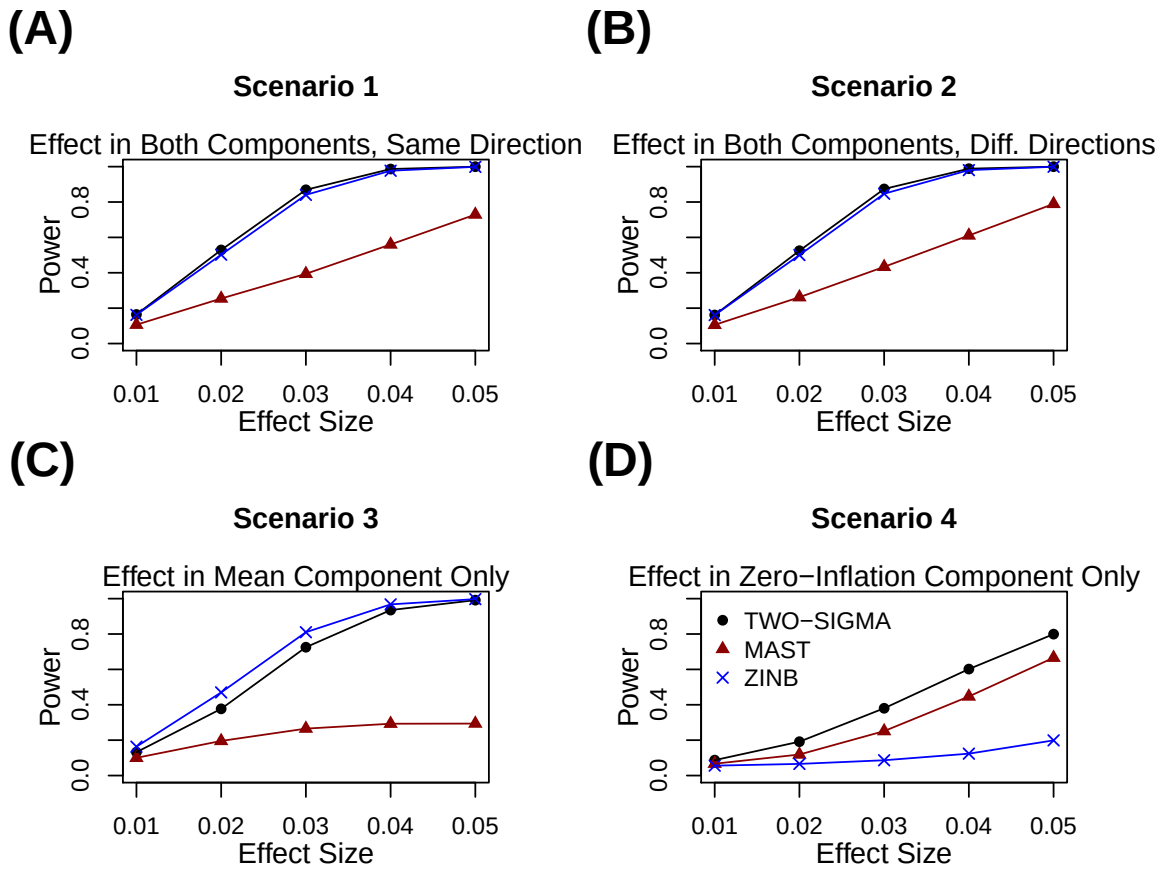

Figure 2.1: Power evaluations in simulated data: Shows the power to test $H_0 : \alpha_1 = \beta_1 = 0$ by varying the effect size with 500 cells from each of 100 individuals. Values of ϕ, σ_a, and σ_b were all set to 0.1 to mimic the "Small R.E." section of table 2.1 and 10,000 genes were simulated. Because of the type-I error inflation from the ZINB model and MAST seen in table 2.1, true power was calculated and plotted for these methods using the empirical significance threshold from the corresponding setting under the null. TWO-SIGMA can bypass the need for computationally expensive resampling procedures needed to generate true power because it preserves the type-I error as seen in table 2.1. See the discussion in section A.1.2 of the appendix for more details about computing true power and discussion regarding power trends across the different methods.

lowest power in the fourth scenario. In simulation, computing the empirical significance thresholds and true power is straightforward and computationally included when evaluating type-I error. In real data settings, however, computationally intensive resampling approaches are needed for reliable estimates of the empirical significance thresholds. Because TWO-SIGMA preserves type-I error, we can rely on raw power and can therefore bypass the need for any resampling approach for valid inference. This is a key advantage and shows that TWO-SIGMA is more robust and flexible than the ZINB model while both preserving the type-I error and retaining high power without any additional computation. When the effect is only in the zero-inflation component, power is lower for all methods than in the first three scenarios. Such effects present only in the zero-inflation component are known to be more difficult to detect, as seen in (Chen and Li, 2016). For full power results, including more detailed comparisons to the ZINB model with additional discussion, see section A.1.2 of the appendix.

2.1.2 ad hoc approach

One primary methodological contribution of TWO-SIGMA for scRNA-seq data analysis is the inclusion of random effect terms in each of the two components, which is a well-established technique to account for within-sample correlation. As mentioned in section 1.3, ignoring random effects in TWO-SIGMA is equivalent to assuming that cells from the same sample/individual are independent. This assumption can lead to underestimated standard errors and thus inflated type-I errors. An example is given in table 2.3 in the real data analysis section.

We utilize the following *ad hoc* approach to determine whether random effects are needed: using a one-way ANOVA, we regress the Pearson residuals from a zero-inflated negative binomial regression model without random effects on the sample label and take the p-value from the overall ANOVA F test. This p-value serves as a rudimentary measure of whether the residuals tend to differ across samples. If they do, this is evidence that residuals are not exchangeable across samples. The full TWO-SIGMA specification including random effects will then be fit to more formally evaluate the need for random effect terms. In contrast, when the residuals show

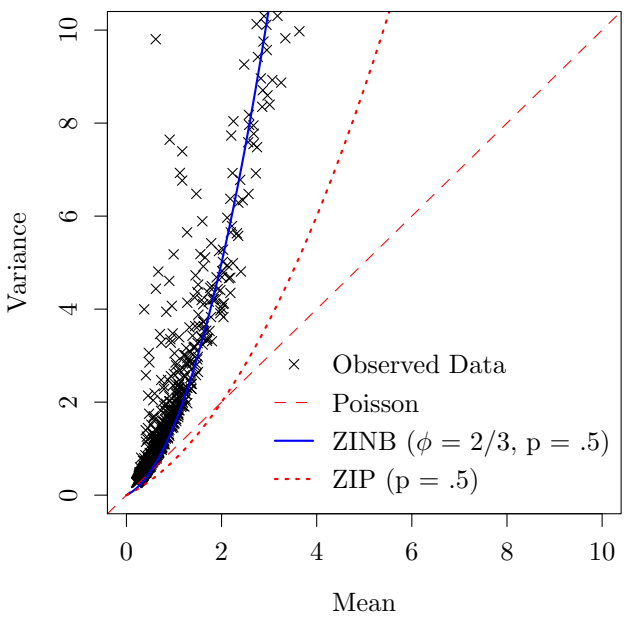

Figure 2.2: Presence of overdispersion in real data: Shows the need of a non-linear mean-variance relationship in the pancreatic islet data. Each point represents the mean-variance relationship for one gene. In the legend ϕ represents the overdispersion parameter of the negative binomial distribution and p represents the drop-out probability.

no tendency of differing across samples, we do not have evidence to believe that they are structured/clustered within samples and thus will not fit the full model with random effects. Through simulations we found that this procedure is very effective in identifying the need for random effects. Results from applying this proposed method to a real dataset of pancreatic islet cells are given in the data analysis section. In simulations, computation runtime was the longest for models attempting to fit random effects when variance components were truly zero (see tables A.1–A.4 in the appendix). Therefore, as discussed more in section A.1 of the appendix, the *ad hoc* method can dramatically reduce overall computation time over many genes in addition to increasing model parsimony where most appropriate.

2.1.3 Pancreas real data analysis

For illustrative purposes we applied TWO-SIGMA to a dataset of pancreatic islet cells isolated from nine individuals (see (Fang *et al.*, 2019) for full details on the data processing and generation steps). To focus on the most informative cells and genes, we applied rather aggressive

filtering of the data to keep the top 2,000 genes by number of transcripts observed and only keep cells with more than 1000 transcripts across these genes. After merging across all nine individuals, we were left with 1,290 genes and 10,269 single cells of which we used only the 7,774 for which cell type information was available based on the expression of signature genes. Here we focus our attention on alpha and beta cells, which compose the majority (55% and 34%, respectively) of the cells in our dataset. Type-II diabetes (T2D) status and age were used as sample-specific covariates in all analyses. The CDR was computed for each cell and included in all analyses, which were further stratified by cell type.

Figure 2.2 plots the relationship of mean versus variance for the 1,290 genes we used in our analysis. It shows that the Poisson and zero-inflated Poisson models cannot adequately account for the overdispersion observed in many genes. In contrast, TWO-SIGMA can accommodate these mean-variance pairs in a quadratic relationship via the overdispersion parameter ϕ. Because we have only nine individuals, we chose to focus on analyses excluding the zero-inflation random effect terms a_i to improve convergence and overall model fit. Some genes still showed convergence issues. This is partly indicative of a misspecified or overparameterized model and partly due to the small number of cells and samples in the dataset. As a general guideline, users with concerns or limited computational resources begin including random effects in the mean component only, and scale upwards to include random effects in the zero-inflation component if performance is satisfactory.

Table 2.2 shows the proportions of genes showing statistically significant results at the .05 level for three types of hypothesis tests: the joint test of significance for the binary disease indicator $H_0 : \alpha_1 = \beta_1 = 0$, the test of the mean model variance component $H_0 : \sigma_b = 0$, and the test for the presence of overdispersion $H_0 : \frac{1}{\phi} = 0$. For example, when fitting the TWO-SIGMA model without the zero-inflation variance component to alpha cells, 73.8% of genes had statistically significant variance components in the mean model. Most genes showed the need for a random effect term or the negative binomial distribution (or both).

Table 2.2: Rejection summaries from the pancreas data: Shows the proportion of genes in the pancreatic islet data with rejected nulls for various hypotheses related to T2D. The TWO-SIGMA model as specified in equation (2.1) was fit with no zero-inflation variance component (no ZIVC).

Hypothesis	Alpha Cells No ZIVC	Beta Cells No ZIVC
Overall Disease Status	0.161	0.111
Overall R.E. Test	0.738	0.724
NB vs. Poisson	0.627	0.555

2.1.3.1 Impact of ignoring within-sample correlation

Table 2.3: Influence of failing to include needed random effects: Gives mean component estimates for gene *RPS29* with (top panel) and without (bottom panel) random effects.

Effect	Estimate	Std. Error	z value	p-value
Intercept	0.521	0.207	2.515	0.012
T2D	-0.349	0.292	-1.197	0.231
age	-0.284	0.256	-1.109	0.267
CDR	0.394	0.011	36.284	$<.001$
σ_b	0.490			
Intercept	0.833	0.021	40.094	$<.001$
T2D	-0.605	0.032	-19.090	$<.001$
age	-0.057	0.017	-3.324	$<.001$
CDR	0.390	0.015	26.611	$<.001$

Models for genes that mistakenly exclude the b_i random effect term often show highly significant results for covariates; this significance can disappear when including the random intercept term—possibly indicative of a false positive due to failing to account for within-sample correlation. For example, gene *RPS29* demonstrates this pattern in alpha cells. Table 2.3 shows that failing to include random effects—and thereby assuming independence of all single cells—can lead to vastly underestimated standard errors. T2D status and age change from highly significant to insignificant when including a random intercept term. The standard error for the coefficient of T2D increases by a factor of 9 from 0.032 to 0.292, and the magnitude of the point estimate is halved from -0.605 to -0.349. Individual covariates such as T2D can thus exhibit dramatically increased type-I error when random effects are incorrectly ignored. In contrast, the coefficient and associated standard error for the cellular detection rate (CDR) are nearly identical in the two

models. This result is expected given that CDR is a cell-level covariate and shows that including sample-specific random effects leads to very minor changes in the estimation of any covariates that are not sample-specific. Our emphasis in this section is not to draw conclusions about any association between *RPS29* and T2D, but rather to illustrate that ignoring random effects has the potential to alter scientific conclusions.

We also used alpha cells to test the overall effect of T2D using both TWO-SIGMA to MAST. Table 2.4 shows that MAST rejects in many more instances than TWO-SIGMA. Of the 273 genes that were rejected with MAST but not with TWO-SIGMA, 234 have statistically significant variance components in TWO-SIGMA. This further illustrates the possibility that fixed effect coefficients can be mistakenly deemed significant in the presence of within-sample correlation.

2.1.3.2 Cell-type specific genes often show a need for random effect inclusion

Table 2.4: Agreement between TWO-SIGMA and MAST: Shows the agreement in rejecting the omnibus null hypothesis of an association between T2D status and gene expression in alpha cells using a Bonferroni adjusted significance level of 5×10^{-5}.

TWO-SIGMA	MAST	
	No Reject	Reject
No Reject	1013	273
Reject	1	3

We matched 234 and 120 genes in our data that were identified in previous studies as cell-type specific in alpha or beta cells, respectively. ((Lawlor *et al.*, 2017), supplementary table 10). After stratifying the data by cell type and removing genes with more than 90% or less than 10% zeros, we fit TWO-SIGMA (excluding a_i as mentioned previously) to the remaining 222 alpha cell-specific and 111 beta cell-specific genes to alpha cells and beta cells, respectively. Of these, 93 alpha cell-specific genes and 85 beta cell-specific genes had statistically significant variance components σ_b. This suggests that non-negligible between-sample variation—not attributable to cell-type—is present for these cell-type specific genes. As discussed in (Lawlor *et al.*, 2017), cell-type specific expression profiles are often of primary interest to study (dis)function at the cellular

level and reveal novel approaches to treat and manage diseases such as T2D. Thus, it is critical to have reliable inference for these genes. As seen in the previous section, incorrectly excluding random effect terms can provide very misleading results and can thereby misdirect attempts to understand disease etiology at a cellular level.

2.1.3.3 The *ad hoc* method successfully separates genes that need random effects

Finally, we used all 1,290 genes from the islet dataset to demonstrate the usefulness of the *ad hoc* method to determine the need for the random effects terms b_i. Figure 2.3 shows that likelihood ratio statistics from formal testing of b_i are consistently larger for genes selected by the procedure than those not selected. This pattern suggests that the *ad hoc* procedure described earlier can effectively identify genes that will exhibit non-zero variance components in real data.

2.1.4 Discussion of TWO-SIGMA

TWO-SIGMA builds on the well-established literature in both zero-inflated models and generalized linear mixed models. It keeps the data on the original scale while simultaneously allowing for zero-inflation, overdispersion, and random effects to account for within-sample correlation. As compared to existing methods, its flexibility is demonstrated both in the use of random effect terms and the ability to test any hypothesis of DE that can be expressed as a contrast of regression coefficients while controlling for multiple sample-level and individual-level covariates.

Incorrectly excluding random effects and assuming independence of cells can lead to underestimated standard errors of fixed effects and can therefore increase the type-I error of hypothesis tests relating to fixed effects parameters. See table 2.3 for an example. If the random effect terms do not contribute to the model fit, as judged by a statistical test or practical significance, they can be removed easily within the general framework of TWO-SIGMA. Random intercepts can also be useful even when they are not of direct interest: they often capture the effects of omitted

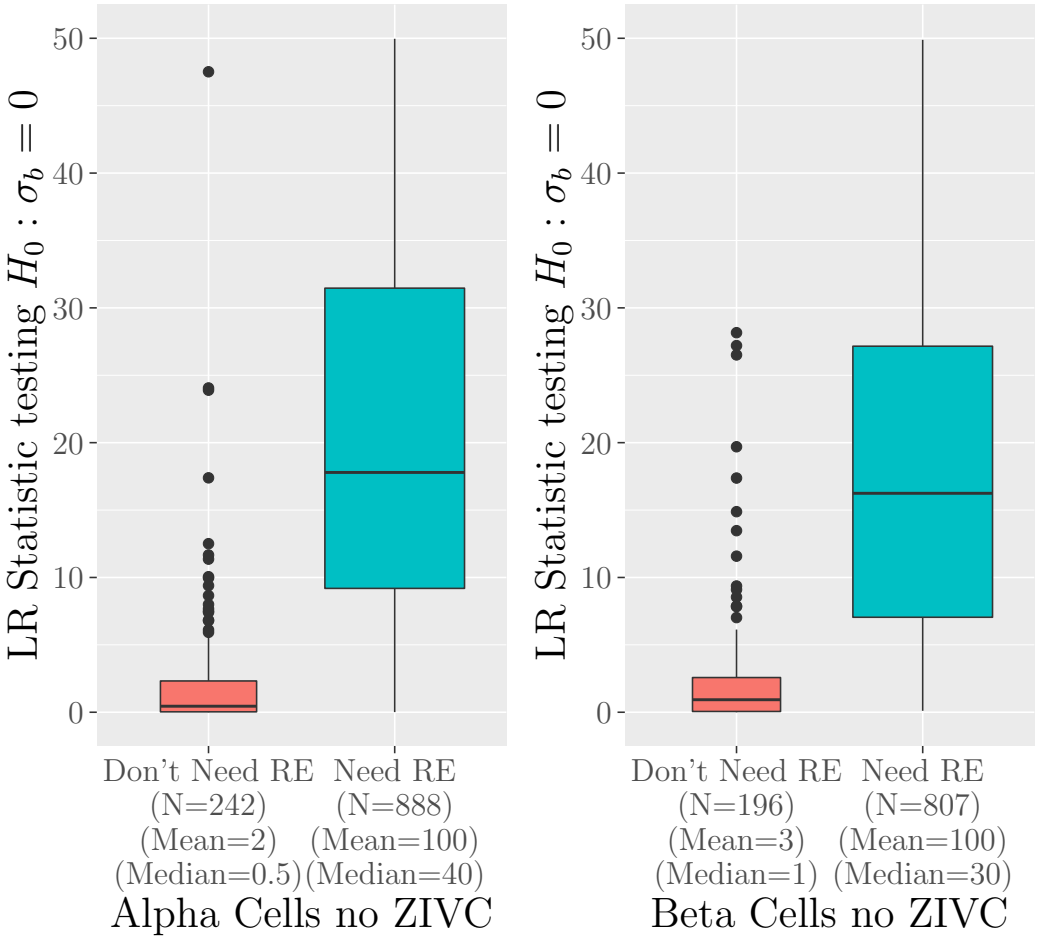

Figure 2.3: Ability of the ad hoc method to identify genes in need of random effects: Shows boxplots of the LR statistics from the joint test of the need for random effects, $H_0 : \sigma_a = \sigma_b = 0$, using TWO-SIGMA. Genes that our *ad hoc* procedure suggests need random effects ("Need RE") and genes the procedure suggests do not ("Don't Need RE") are compared. Both panels were created using TWO-SIGMA as specified in equation (2.1) but with no zero-inflation variance component (no ZIVC).

sample-specific covariates, and can limit the bias of fixed effect coefficients caused by misspecification. For example, if cell-type information is missing, and varies between individuals, a random intercept term can limit the resulting bias and p-value inflation observed in fixed effects parameters. Our *ad hoc* method proves to be a useful tool to both (i) select genes that could benefit from including random effect terms and (ii) reduce overall computation time by suggesting genes that do not need to be fit including random effect terms.

Because we expect *a priori* that zero-inflation will occur in scRNA-seq, it is beneficial to include a component dedicated to it. The zero-inflation component in TWO-SIGMA is flexible

in that it allows for a different set of covariates from the mean model, or no covariates at all. For example, one might be interested in using zero-inflation only to improve mean parameter estimation. In this scenario, a constant probability of drop-out could be assumed via an intercept-only regression model. This would prevent coefficient estimates in the mean model from being overly shrunk towards zero, as would occur if drop-out was not accounted for, but would also limit the total number of parameters estimated and maximize model parsimony. Even if the data are not truly generated from a zero-inflated process, or if drop-out is viewed as a "nuisance," using the two component model in equation (2.1) can be a convenient choice to improve model fit and fixed effect parameter estimation. As a point of comparison, MAST requires covariates in each component to be identical and therefore could not fit such an intercept-only zero-inflation component.

Finally, our experience suggests that variance component estimates are often much smaller in the zero-inflation component than in the mean component. Therefore, as we did in the real data analysis, it might be a pragmatic choice to exclude random effects from the zero-inflation component of TWO-SIGMA. A key strength of TWO-SIGMA is the flexibility to easily customize the model within the general framework either *a priori* or via iterative removal based on statistical hypothesis tests of features such as random effects, overdispersion, or the drop-out component.

CHAPTER 3: TWO-SIGMA-G: A TWO-COMPONENT SINGLE CELL MODEL-BASED ASSOCIATION METHOD FOR SINGLE-CELL RNA-SEQ GENESET TESTING

3.1 TWO-SIGMA-G

We further extend our TWO-SIGMA method to competitive gene set testing via TWO-SIGMA-Geneset (TWO-SIGMA-G). First, *gene-level* statistics are collected for test and reference set genes. In a similar vein to CAMERA, we choose to compare the test and reference set statistics using the two-sample Wilcoxon rank-sum test. In using ranks, we can provide robustness against the influence of very large gene-level statistics.

As mentioned, IGC can inflate type-I error in competitive gene set testing (Wu and Smyth, 2012). Assuming a common pairwise correlation ρ in the test set of size m_1 and no correlation in the reference set of size m_2, it can be shown that the variance of the two-group Wilcoxon rank-sum statistic is given by:

$$\frac{m_1 m_2}{2\pi}\left(\sin^{-1}1 + (m_2-1)\sin^{-1}\frac{1}{2} + (m_1-1)(m_2-1)\sin^{-1}\frac{\rho}{2} + (m_1-1)\sin^{-1}\frac{\rho+1}{2}\right)$$

Using this variance formula, we can compute p-values using the usual normal approximation to the Wilcoxon rank-sum statistic. The reference set used in TWO-SIGMA-G can be chosen in one of two ways: either using a random sample of other genes of size m_1, or as the collection of all genes not in the test set under consideration. Previous studies have cautioned that set size can inflate the type-I error of some gene set testing procedures (Damian and Gorfine, 2004; Tian *et al.*, 2005). For larger gene sets, which are likely of more interest scientifically, the difference between these two approaches for choosing a reference set diminishes. We will discuss the choice

of reference set in more detail in section 3.2 and in the context of a real data example in section 3.3.

Cell-level covariates such as the cellular detection rate (CDR) have been previously demonstrated to be highly influential to observed expression levels (Finak *et al.*, 2015). Subject-specific covariates, such as disease status or race, can further create an additional correlation structure in the raw data. Gene-level DE statistics will be adjusted for the effects of these covariates When they are included in a regression model. Using the raw data to estimate IGC can therefore overestimate the correlation that remains between the gene-level statistics, and can thereby lead to conservative set-level inference. Residuals, in contrast, remove covariate effects and can better represent the correlation in the gene-level statistics.

We therefore estimate the inter-gene correlation of a given gene set using the residuals from the TWO-SIGMA model as follows: Define the $(n_i \times 1)$ vector of residuals for gene s from individual i as $\boldsymbol{r_{is}} = \boldsymbol{Y_{is}} - \widehat{\boldsymbol{Y}}_{\boldsymbol{is}}$. Then, by individual, construct the $n_i \times s$ matrix $\boldsymbol{R_i} = \{\boldsymbol{r_{is}}\}$ consisting of the residuals for all genes in the test set. Given these residual matrices, we can compute the $(s \times s)$ correlation matrix $\boldsymbol{C}_i$, which contains s choose 2 unique non-diagonal elements. These elements give the pairwise correlation between residuals of two different genes in the test set. We average these (s choose 2) values to produce one average pairwise correlation $\hat{\rho}_i$ per individual. Finally, our overall correlation estimate is taken as the average of these values such that $\hat{\rho} = \sum_{i=1}^{n} \hat{\rho}_i / n$. We found in simulations that this IGC estimate preserves type-I error in a conservative manner while still producing improved power in a variety of realistic scenarios. Our IGC procedure therefore builds off of the advantages of a residualized approach while further using individual-level calculations to help mitigate the impacts of the large individual heterogeneity often seen in scRNA-seq datasets. Additionally, use of the residuals also removes the correlation explained by sample-specific random effects terms if included at the gene-level.

As compared to other methods, TWO-SIGMA-G has several key advantages in applicability and interpretability. First, it is explicitly tailored to scRNA-seq data at the gene-level in that it can flexibly and optionally include zero-inflation, overdispersion, and within-subject random effect

terms to account for with-subject correlation. Second, the use of a regression modeling framework at the gene-level enables the analysis of complex designs including multiple confounding covariates as will be displayed further in the real data analysis of section 3.3. Third, as in CAMERA, estimating the IGC using residuals allows us to remove the effects of included sample-level and cell-level covariates. Our approach adds another step in first stratifying by individual in the IGC calculation then averaging over individuals to produce an overall estimate. Finally, like CAMERA, the estimate of the IGC is virtually free computationally in that the model is not refit via permutation or bootstrapping.

3.2 Performance of TWO-SIGMA-G

Extensive simulation studies were conducted to evaluate the performance of TWO-SIGMA-G. A total of six different settings were constructed which varied both the amount of inter-gene correlation induced and the presence of other covariates in the model, which can create additional complex correlation structures between cells and genes. For each setting, we aggregated over ten biological replicates consisting of differing cell populations to minimize the impact of initial cell population on results. Settings were repeated using reference sets of size 30 and 100. See section B1 in appendix for more details regarding the simulation procedure and the settings used. TWO-SIGMA-G was compared to two other methods for competitive testing: CAMERA, the leading method for bulk RNA-seq and thus for competitive testing, and the procedure in MAST, which is one of the most popular packages for scRNA-seq data analysis. Methods designed for self-contained testing, a hybrid of self-contained and competitive testing, or other aspects of gene set testing, such as ROAST (Wu *et al.*, 2010), GSEA (Subramanian *et al.*, 2005), `sigPathway` (Tian *et al.*, 2005), and PAGODA (Fan *et al.*, 2016) were not included because they are testing fundamentally different null hypotheses. We had difficulties obtaining reliable p-values from iDEA for the main simulations. We believe this is because our simulations were calibrated using gene-level statistics summarizing evidence from both components, while iDEA uses only the effect size from the mean component. TWO-SIGMA-G, CAMERA, and MAST all utilize the

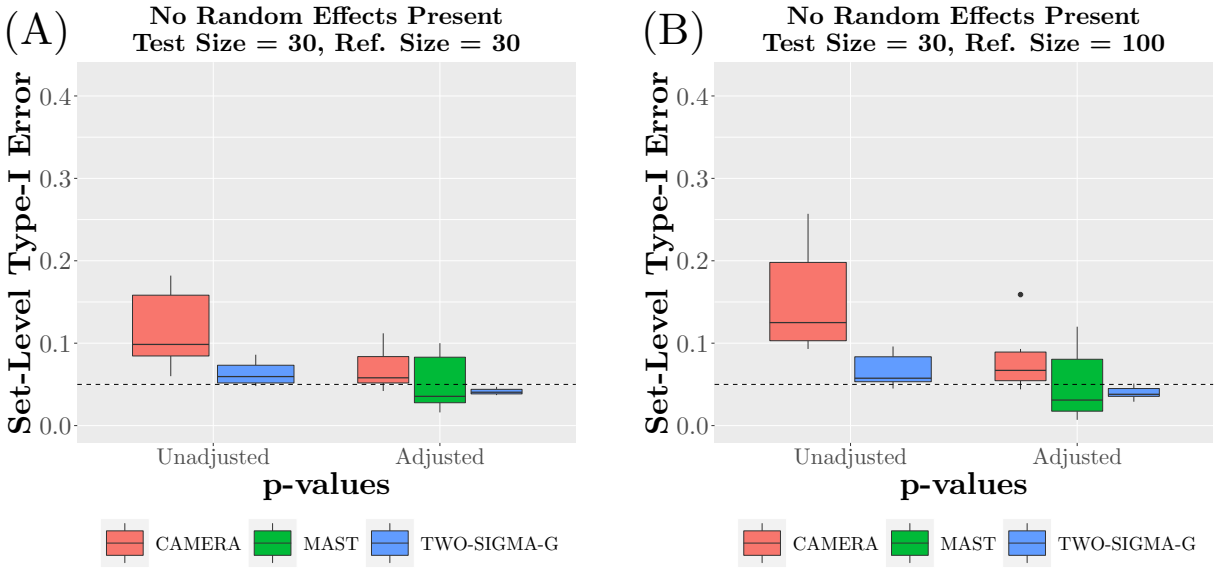

Figure 3.1: Shows the set-level type-I error of CAMERA, MAST, and TWO-SIGMA-G for genes simulated with IGC. Two scenarios are presented: using reference set sizes of 100 and 30. Within each scenario, both unadjusted and adjusted set-level p-values are plotted (unadjusted p-values are unavailable for MAST). Each boxplot aggregates 6 different settings which vary both the magnitude of the average inter-gene correlation in the test set and the nature of the correlation structure via the introduction of other individual-level covariates. Such settings are meant to represent the diversity seen in real data sets to paint an accurate picture of testing properties over a wide range of gene sets. Each of the 6 settings is further composed of 10 replicates which vary only random seed to mimic the impact of a different starting pool of cells from which genes were simulated. See appendix section B1 for more details regarding the simulation procedure.

raw data and as such can capture general set-level enrichment coming from expression changes in zero proportion or mean value. iDEA does not use the raw data and uses only the summary statistics from the mean component, however. Thus, in Figure B5 of the appendix, we provide a meaningful comparison to iDEA using data simulated to emphasize the mean component effect, and show that TWO-SIGMA-G shows comparable power with gains in some scenarios.

3.2.1 Set-Level Type-I Error Control

Figure 3.1 shows the type-I error performance of TWO-SIGMA-G, CAMERA, and MAST across the six simulated settings. Unadjusted p-values demonstrate the implications of ignored IGC—type-I error is consistently inflated without adjustment. After p-value adjustment, both TWO-SIGMA-G and MAST tend to preserve type-I error at the 5% level. In contrast, CAMERA suffers from inflated type-I error after IGC adjustment. Differences between the three methods

are likely partially due to a combination of factors that lead to a misspecified model for the features of scRNA-seq data. First, CAMERA and MAST use a log transformation of the data, which may distort true signals, particularly in the presence of many zero counts (Townes *et al.*, 2019; Lun, 2018). Second, unlike TWO-SIGMA-G and MAST, CAMERA does not separately model the excess zeros in the data and may underestimate parameters relating to mean expression as a result. Performance of MAST and CAMERA tends to be the worst in scenarios involving additional covariates which create more complex correlation structures. The procedure used in TWO-SIGMA-G to estimate and adjust for IGC is well-calibrated and produces valid set-level inference.

Additional analyses showed that the type-I error from TWO-SIGMA-G is preserved or approximately preserved in the presence of non-zero gene-level random effect terms, whether or not they are included in the gene-level models (Appendix Figure B1). For both CAMERA and MAST, however, type-I error tends to be inflated on average and the variance in the type-I error across the six settings tends to increase in the presence of gene-level random effects. Performance for these methods tends to once again be the worst in settings involving the largest inter-gene correlation. For both methods, however, this inflation is much lower in magnitude than at the gene-level (Van Buren *et al.*, 2020). This highlights an advantage of competitive testing: because it makes a relative comparison to a reference set of genes, it is partially robust to the consequences of a systematic, gene-level misspecification. The real data analysis in section 3.3 further shows large agreement in set-level results from TWO-SIGMA-G regardless of random effect inclusion in the gene-level model.

Results seen in figure 3.1 and Appendix Figure B1 suggest,that the null distribution of all three methods are nearly identical with a larger reference set size. In the interest of being conservative, however, we will evaluate performance using results in which the test and reference sets are of equal size.

Figure S2 in the appendix shows the type-I error at alternative set-level null hypotheses, in which an equal percentage of genes in the test and reference sets are DE (with the same gene-

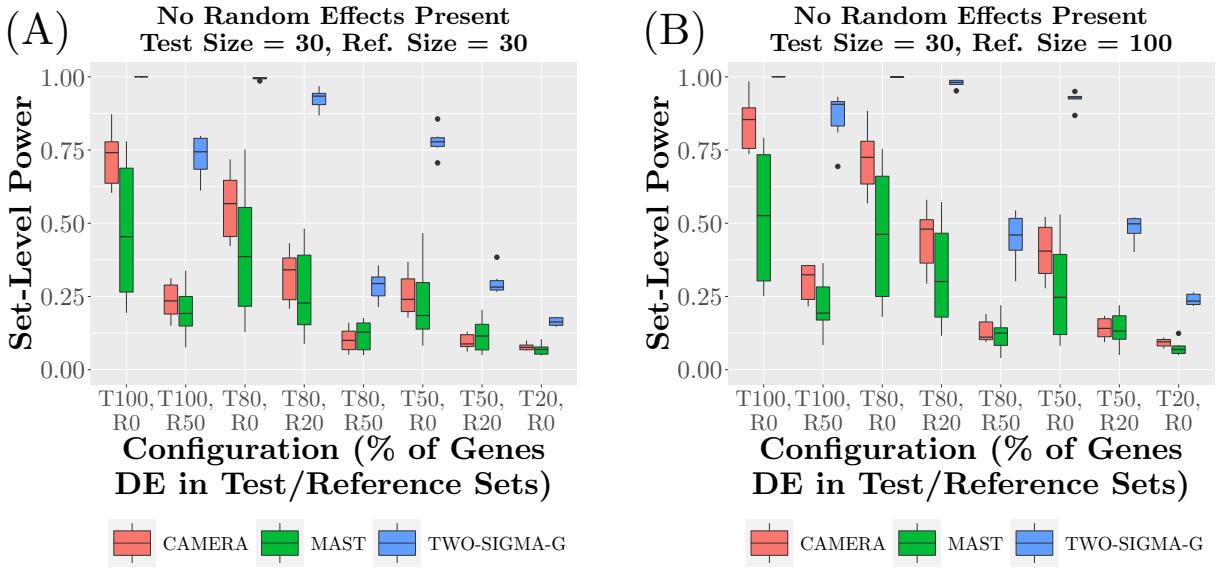

Figure 3.2: Shows the set-level power of TWO-SIGMA-G and CAMERA for genes simulated with IGC. Two scenarios are presented: using reference set sizes of 100 and 30. Within each scenario, the percentage of genes that are differentially expressed (with the same effect size) in the test and reference set is varied. For example, "T80,R50" corresponds to the configuration under the alternative hypothesis in which 80% of test set genes are DE and 50% of reference set genes are DE. Each boxplot aggregates 6 different settings which vary both the magnitude of the average inter-gene correlation in the test set and the nature of the correlation structure via the introduction of other individual-level covariates. Such settings are meant to represent the diversity seen in real data sets to paint an accurate picture of testing properties over a wide range of gene sets. Each of the 6 settings is further composed of 10 replicates which vary only random seed to mimic the impact of a different starting pool of cells from which genes were simulated. See section B1 of the appendix for more details regarding the simulation procedure.

level effect size). Generally, TWO-SIGMA-G and MAST tend to both become more conservative as more DE genes are introduced, both without gene-level random effects and when they are mistakenly absent. The type-I error of MAST tends to become inflated once the background percentage of DE genes increases, particularly when gene-level random effects are mistakenly excluded.

3.2.2 Set-Level Power Improvement

Figure 3.2 shows the power of CAMERA, MAST, and TWO-SIGMA-G on simulated data. Different configurations are presented which involves a differing proportion of DE genes (with the same effect size) in the test and reference set. For example, "T100,R50" corresponds to the configuration in which 100% of genes in the test set are DE and 50% of genes in the reference set are DE. Scenarios which combine DE and non-DE genes in both the test and reference set are the

most informative to study because it is unlikely in real data to have a completely null reference set and/or a completely alternative test set. Our simulations demonstrate several interesting findings. First, TWO-SIGMA-G is consistently the most powerful method. Second, power depends primarily on the proportion difference in DE between the test and reference set, and less on the precise composition of the test and reference sets. For example, the "T80,R50" and "T50,R20" configurations have the same difference in percentage of DE genes, and similar power profiles for all methods. Third, using a reference set of size 100 tends to improve power for both methods and particularly for TWO-SIGMA-G. As discussed in the previous section, this power increase does not seem to be a consequence of an increase in type-I error. This provides some evidence in favor of using a larger reference set in lieu of a balanced reference set. We will revisit this topic in the real data analysis of section 3.3. Figure B3 in the appendix shows power when truly present gene-level random effects are included or mistakenly excluded from the gene-level model. In either case, power is only slightly reduced versus the case without gene-level random effects. Thus, if interested primarily in set-level inference, gene-level random effect terms and the associated increase in computational cost may not be necessary for valid and powerful inference. If gene-level inference is simultaneously of interest, however, this power loss may be acceptable to prevent the massive type-I error inflation that has been shown to occur at the gene-level when random effects are mistakenly absent (Van Buren *et al.*, 2020). Figure B4 in the appendix varies the magnitude of DE, with half of genes having twice the effect size of the other half. Set-level power is improved, as expected, but the relative positions of each configuration remain as in figure 3.2, suggesting that power results presented are applicable to alternative DE breakdowns.

3.3 Real Data Analysis

We demonstrate our TWO-SIGMA-G method on two real datasets. The first is a dataset of 15,351 single-cells collected from 6 donor mice, three of which were infected HIV and three were given a mock treatment. As with other UMI-based scRNA-seq data, we found that this data was not consistent with zero-inflation, and thus we fit the TWO-SIGMA model without the zero-

inflation component at the gene-level. Cells were typed, and for our analysis we will consider the 11 cell types which had cells both with and without HIV. Because the primary interest is in comparisons between HIV and mock cells within a cell type, we categorize the remaining 14,354 cells into one of 2*11 = 22 mutually exclusive groups. Many cell types are quite rare, and thus an ANCOVA model additionally adjusting for CDR was fit as a way to improve estimation for the very rare cell types. TWO-SIGMA-G is ideal for this analysis because gene-level statistics can come from a test of such an arbitrary contrast matrix. Gene-level statistics for each cell type are Z-statistics contrasting the mean values in observed expression between the two treatment groups within a cell type. Gene sets were taken from the Molecular Signatures Database (Subramanian *et al.*, 2005; Liberzon *et al.*, 2015) version 7, c2 collection, accessed via the `msigdf` R package (https://github.com/ToledoEM/msigdf). After filtering to keep sets with at least two genes present in the data, a total 4,630 genes and 5,011 sets were analyzed. Each gene set has an associated set-level p-value for each of the 11 cell types. Figure 3.3 shows a heatmap of average set-level log fold-changes and associated p-values from the HIV dataset. Hierarchical clustering groups the gene sets into to main groups: the rightmost group represents sets that are biologically expected responses to virus introduction. These sets tend to display consistent effect sizes across all 11 cell types.

The second dataset we analyzed consists of 80,660 single cells sequenced from 48 human donors Mathys *et al.* (2019). Half of these donors have clinically diagnosed Alzheimer's disease (AD), categorized into early and late stages (12 individuals each), and the other half are control patients without an Alzheimer's diagnosis. Our geneset analysis was conducted analogously to the HIV dataset: a one-component ANCOVA model was fit including cell-type and AD status jointly, with age at death, sex, and the CDR used as an additional covariates. Once again, the gene-level statistics were contrasts of the difference in expression between AD stages within cell-type. For the analysis, we were left with six cell types: excitatory neurons (Ex), inhibitory neurons (In), microglia (Mic), astrocytes (Ast), oligodendrocytes (Oli), and oligodendrocyte pro-

genitor cells (Opc). In total, filtering genes expressed in at least 10% of all cells and restricting to these cell types left us with 6,048 genes and 70,634 cells for analysis.

Results were generated comparing AD patients (early or late stage) to control, early stage AD patients to control, late stage AD patients to control, and late stage AD patients to early stage AD patients. Appendix B shows the complete results, but here we focus on the comparison between early stage AD patients and controls, seen in figure 3.4. Many of the top sets are downregulated in all cell types and involved in cellular respiration, such as "MOOTHA_VOXPHOS" and "KEGG_OXIDATIVE_PHOSPHORYLATION." Previous studies have suggested that dysfunction in mitochondrial functioning, particularly in cellular respiration as caused by oxidative damage, is among the earlist events in Alzheimer's disease (Nunomura *et al.*, 2001). As figure 3.4 demonstrates, we replicate this finding with particularly strong and robust downregulation seen in respiration related pathways in neuronal cells. Previous differential expression analyses of AD patients have suggested that the trend of decreased expression in genes associated with cellular respiration may reverse over time (Nunomura *et al.*, 2001; Manczak *et al.*, 2004). We observe this pattern in our set-level analyses as well: sets of genes functionally annotated to be related to cellular respiration are significantly downregulated comparing early stage AD patients to control, significantly upregulated comparing late stage to early stage AD patients, and upregulated but not among the most significant sets comparing late stage AD patients to control. Thus, it would seem that as the disease progresses, the initial downregulation in these gene sets is reversed over time, possibly due to cellular degeneration and an increasing demand for energy in remaining cells (Nunomura *et al.*, 2001).

Cell-type specific heterogeneity is also revealed by our analysis. For example, microglia cells exhibit stronger and more significant downregulation of pathways involved in immune response, such as " while showing less of an impact in pathways related to cellular respiration. Given the nature of microglia cells as the immune defenders, this is not a surprising finding. In different contexts, the ability to identify cell-type specific variability can reveal previously unknown functional differences.

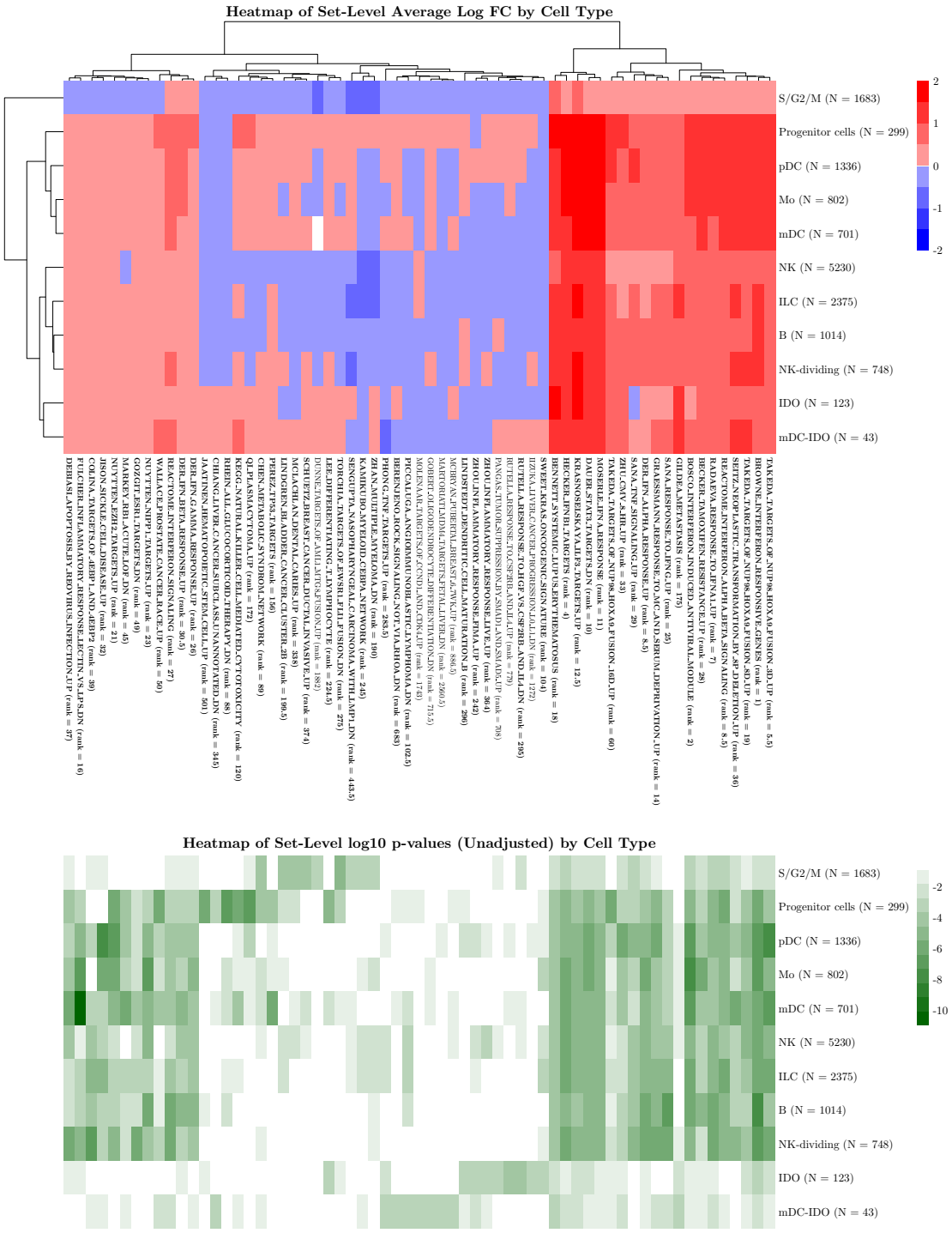

Figure 3.3: Shows cell-type specific variation in set-level significance for the HIV data. Sets which are significant after FDR-adjustment are bolded.

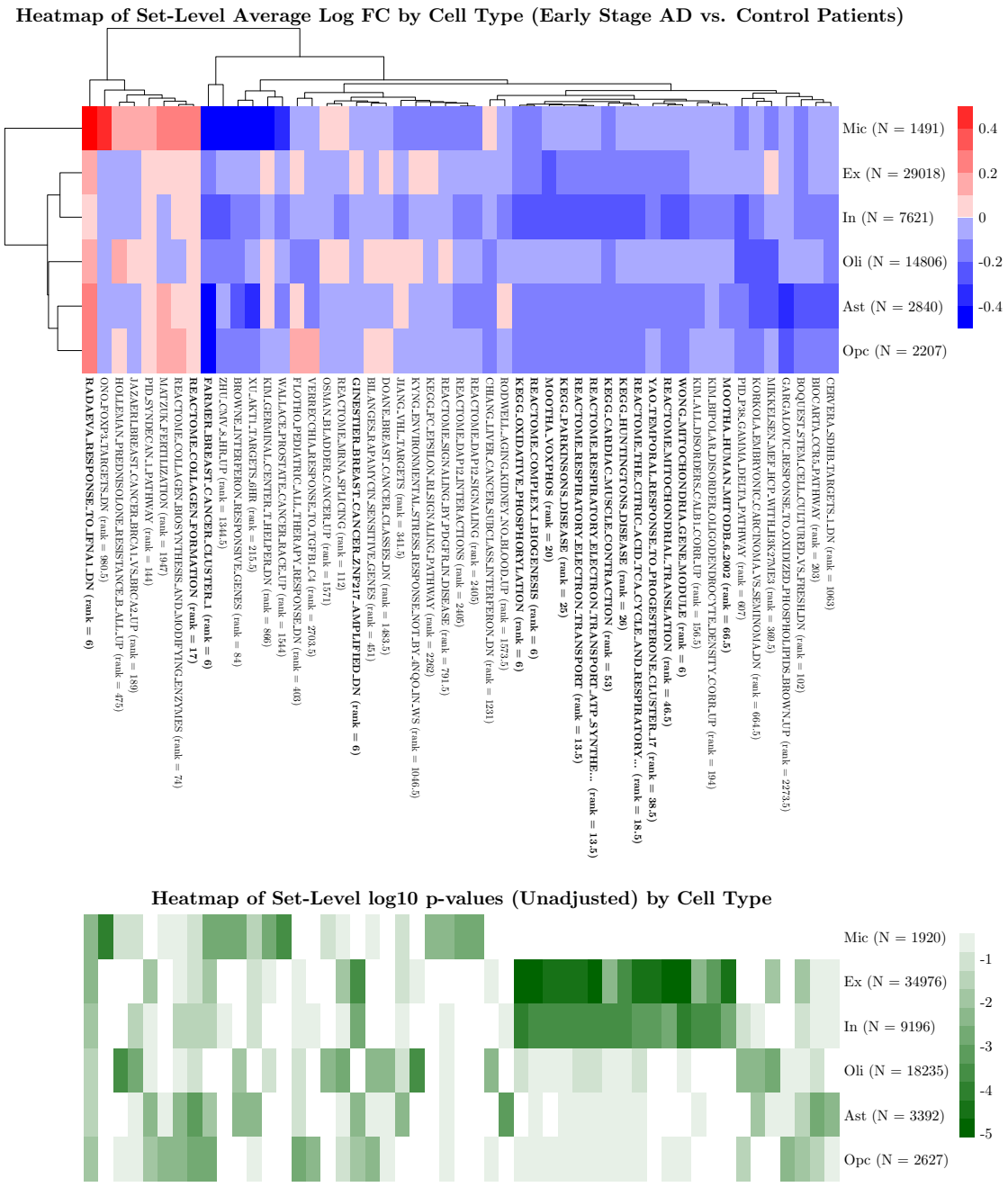

Figure 3.4: Shows cell-type specific variation in set-level significance for the Alzheimer's data. Sets which are significant after FDR-adjustment are bolded.

CHAPTER 4: INTEGRATIVE ANALYSIS OF HI-C AND SINGLE-CELL IMAGING DATA

4.1 Exploratory Analysis of Hi-C and single-cell Imaging Data

As discussed in Section 1.7, there is substantial evidence which suggests that the physical distance between two genetic loci will be correlated with the contact frequency in three dimensions (3D) between the two regions. Loci closer in physical distance would naturally be assumed to be closer in 3D distance, however recent evidence has been suggested physical distance and 3D distance cannot be assumed as perfect substitutes for one another. Therefore, it would be advantageous to collect both measures of interactions separately. As part of an introductory investigation into the prospects of integrating single-cell imaging data and Hi-C data, we first analyze the two data sources separately. Single-cell imaging data originally published in (Mateo *et al.*, 2019) was collected from the common fruit fly (Drosophila). The data structure consists of cells sequenced from eight embryos for a total of 646 cells. For each cell, 70 probes were placed every 10 kilobases (kb) over a 700kb region (from position 12200000 to 12900000 on chromosome 3R), and microscopic imaging captured the relative 3D positions of these probes. For technical reasons, the coordinates of some probes will be missing. Therefore, we first perform linear interpolation to fill in coordinates of missing probes. Such an approach naturally involves some loss of precision, and could be refined by supervised methods in future work.

With an interpolated and complete dataset, for two probes i and j in the same cell, we can compute the pairwise Euclidean distances d_{ij} between their 3D coordinates (x_i, y_i, z_i) and (x_j, y_j, z_j) as follows:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

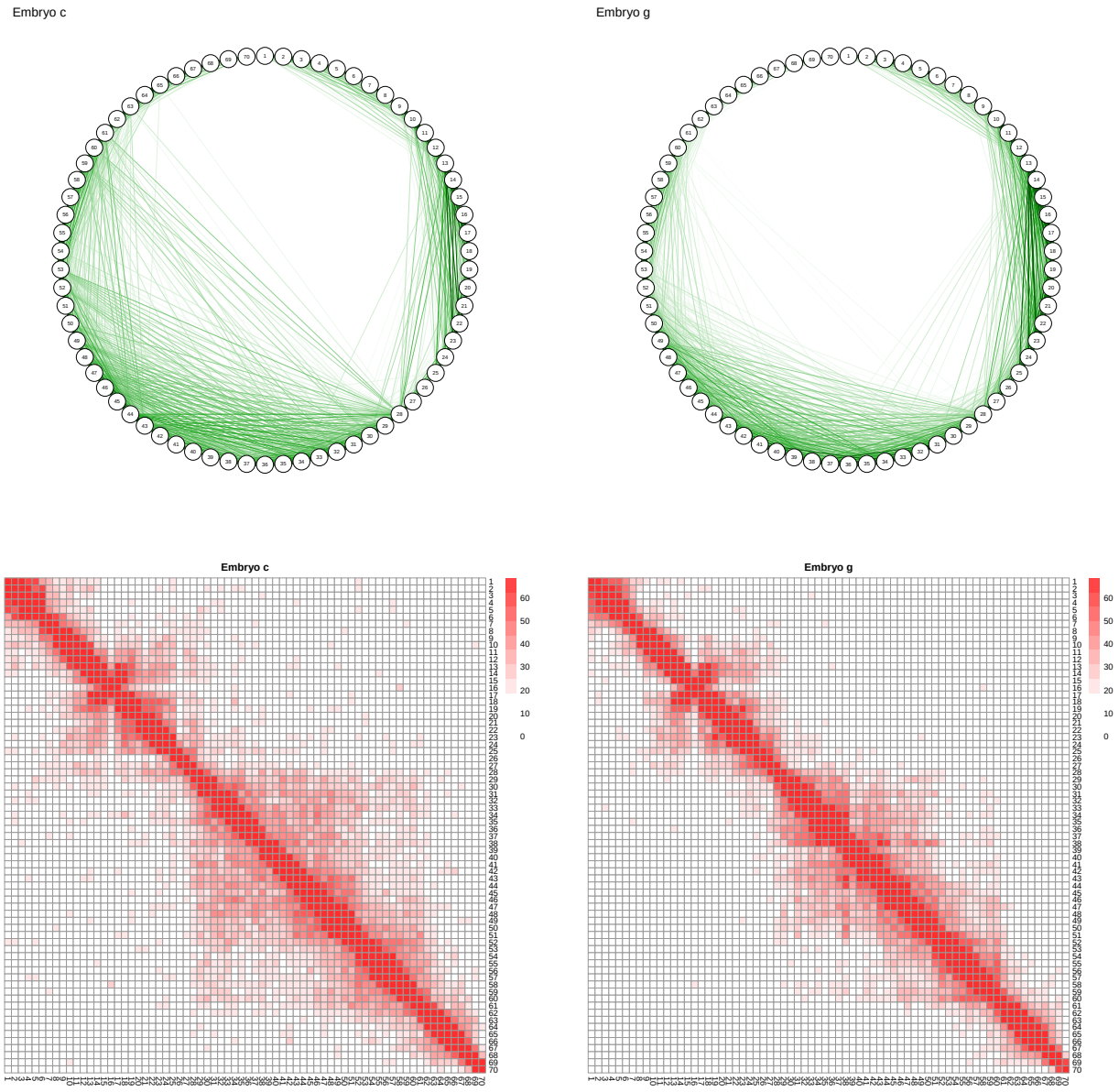

Figure 4.1: Visualizes the average distances between distant genetic loci. Larger edges in the top row correspond to a closer 3D distance, and two representative embryos from the single-cell imaging data are plotted. The bottom row shows heatmaps created from classifying a contact based on a distance within 150nm.

First, we can directly analyze such distances to investigate regions that are relatively far in genomic distance but relatively close in 3D distance. Figure 4.1 plots the average distances between probes in two representative embryos. Proximal probes are the closest in distance, but other patterns emerge: such as the connections involving probes 13 and 28 and other non-adjacent probes. Such plots can represent the distance data in a quasi-continuous way, but are still aggregating all

cells from a given embryo. The heatmaps below show similar patterns, but also reveal several possible Topologically Associated Domains (TAD)s within the data.

As a first step in combining the imaging and Hi-C data, we can categorize a contact based on an observed distance between probes of 150nm or less, as was done in Mateo *et al.* (2019). Doing so, we can produce analogous heatmaps as seen in the bottom row of 4.1. This is in a sense a loss of information, because the distance matters only to the extent of its relation to the threshold of 150nm. Using this approach, however, we can combine the information between the imaging and Hi-C data using the following procedure:

1. Compute the pairwise distances between each probe combination in each cell
2. Categorize contacts based on the Euclidean distance (as specified above) being within 150nm
3. Add these counts over all cells to get total contacts
4. Standardize Hi-C contacts and distance-based contacts to contacts per million
5. Add the two sources of contacts and multiply by one million

Figure 4.2 displays corresponding heatmaps for the Hi-C data and the combined Hi-C and imaging data. In general, the signal presented by both data sources is similar, but some differences do exist in certain regions. In the next section we will focus on the task of peak calling. Evaluating the sensitivity of possible TADs, for example, would be an interesting application that we do not explore here.

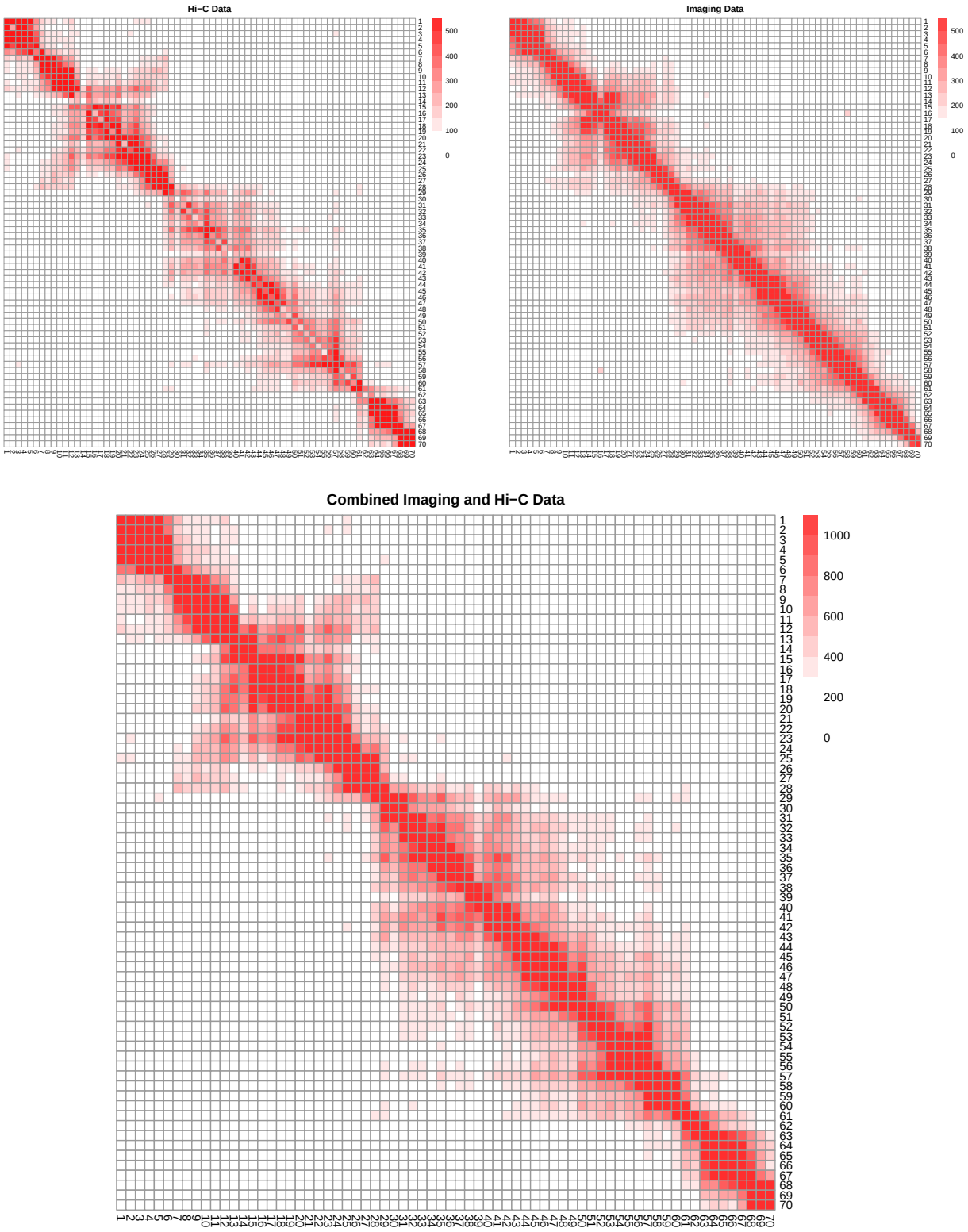

Figure 4.2: Heatmaps of contacts between probes. Larger edges correspond to a closer 3D distance, and two representative embryos from the single-cell imaging data are plotted.

4.2 Integrating Hi-C and Imaging Data for Peak Calling

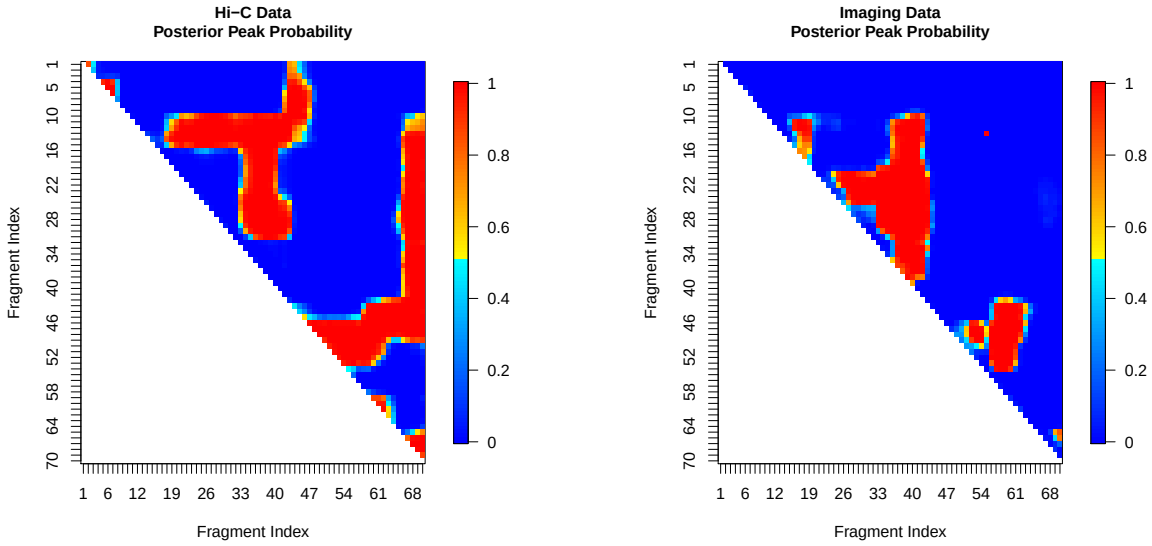

Figure 4.3: Shows estimated posterior peak probabilities from the Bayesian HMRF Model of Xu *et al.* (2015) for the Hi-C and imaging data.

One of the most commonly performed tasks with Hi-C data is peak calling. The goal of these analyses is to determine which interactions observed are due to random chance, and which interactions are not random. Although many methods exist for peak calling, we choose to focus on a Bayesian method utilizing Hidden Markov Random Fields (HMRF) first proposed in Xu *et al.* (2015). HMRFs are an extension of Hidden Markov Models (HMMs) to two dimensions by creating an underlying Markov random field. The Markov property explicitly accounts for the dependency between neighboring genetic loci (in genetic distance). Unlike many competitor methods, information is thus borrowed from neighboring regions to improve peak calling. As input, the Bayesian HMRF takes in the observed contact counts, and an estimate of expected interaction counts. Although there are a number of ways to get these expected counts, here we use a modification of the Fit-Hi-C package Ay *et al.* (2014). As output, the method produces a posterior peak probability estimate for each pairwise interaction combination. Figure 4.3 shows the output of the Bayesian HMRF caller for the same 70kb region separately in the Hi-C and imaging data. The left panel in particular shows a broad peak near the end of the region studied. The low

expected number of interaction counts seems to be the primary driver of this pattern. Our next goal is to combine the estimated posterior probabilities from the Hi-C and imaging data to produce consensus peak probability estimates. This task must be undertaken with caution because the peak probability estimates from the two data sources are necessarily correlated. The nature of this dependency is unknown. Therefore, to combine the probabilities, we use the Cauchy Combination test, which was recently articulated to combine dependent p-values Liu and Xie (2020). Although we are combining probabilities and not strictly p-values, the method itself can still be applied. The CCT statistic is given by:

$$T(p_1, ..., p_j) = \sum_{i=1}^{j} \omega_j \text{tan}\{\pi(.5 - p_i)\}$$

where $\sum_j \omega_j = 1$. As discussed more in Liu and Xie (2020), this quantity can be approximated using the standard Cauchy distribution.

Let p_i be the peak probability from the imaging data for a given loci, and p_h be the peak probability from the Hi-C data for a given loci. Then, we can calculate the combined posterior peak probability as follows:

$$p_c = 1 - F_C(T(p_h, p_i)) = .5 - \frac{1}{\pi}\text{arctan}\{T(p_h, p_i)\}$$

where F_C is the c.d.f. of the standard Cauchy distribution. Figure 4.4 shows p_c as applied to the integration of the Hi-C and imaging data. As expected, it generally paints a kind of "average" of the two heatmaps in figure 4.3, but should be more robust to the dependency between these two heatmaps. Follow-up analysis can aim to improve the biological interpretation of these called peaks.

Future work in this context can focus on a number of possible improvements. First, potential alternatives to producing expected count estimates for input to the Bayesian HMRF can be explored. These may reduce the seemingly strange pattern observed in the left panel of figure 4.3. Second, the distance from the imaging data could be applied in a continuous manner, rather than

used only to create a binarized interaction count. Such approaches could include a joint model that aims to use all observed distances from every cell to define interactions. Third, such integration of different datasets could be applied to other common tasks involving Hi-C data, such as TAD calling or 3D structure modelling. Finally, and perhaps most importantly, the ultimate usefulness in imaging methods will be in an ability to refine prediction in genomic regions that are not imaged. For example, the null distribution of interactions or peak posterior probabilities could be refined using the imaging data where it is available. Then, adjustments to regions outside of those imaged could be made to leverage the imaging data to its fullest potential.

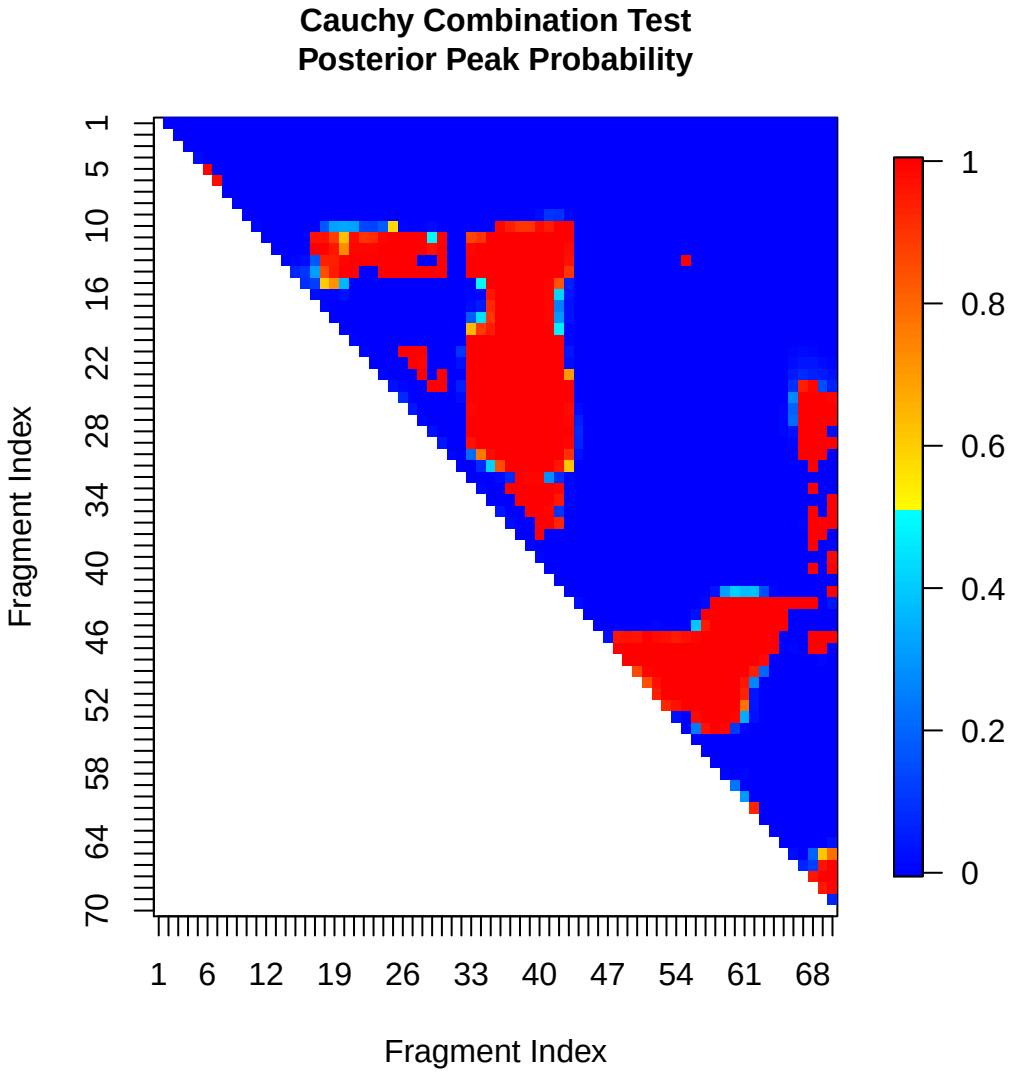

Figure 4.4: Shows the estimated posterior peak probability from the Bayesian HMRF Model of Xu *et al.* (2015) for the Hi-C and imaging data.

CHAPTER 5: CONCLUSION

In this work, we discuss topics related to the analysis of single-cell genomic data. In the first chapter, we discuss a general framework for DE testing of single-cell RNA-sequencing data. This is one of the most common, and most important, statistical tasks performed on scRNA-seq data. Considerable attention from researchers has lead to the development of a number of methods for DE analysis. Here, we introduce TWO-SIGMA, which builds on these methods by accommodating some novel features and some aspects of previously developed work. TWO-SIGMA is a general two-component mixed-effects zero-inflated negative binomial regression framework. As such, it can accommodate zero-inflated and overdispersed counts while keeping the data on its original scale. Random effect terms can be used to adjust standard error estimates for the within-sample correlation that may exist by the nature of scRNA-seq experimental designs, which sequence many cells from a much smaller number of donors. The regression modelling framework can control for different and arbitrary covariates in each component and test general DE beyond a two-group comparison as we demonstrated in this dissertation. Users can also remove the zero-inflation component or reduce the negative binomial distribution to the Poisson distribution if supported by the data. Our results suggest quite strongly that the presence of even a moderate within-sample correlation can radically inflate type-I error in some genes.

The second chapter of this work extends TWO-SIGMA to geneset testing with TWO-SIGMA-G. At the gene-level, TWO-SIGMA-G retains the advantages of TWO-SIGMA. At the set-level, TWO-SIGMA-G proposes a competitive framework that adjusts for the inter-gene correlation expected within pathways given their construction to represent genes with biologically similar functions. Our adjustment is necessary to protect set-level type-I error, as shown in simulations,

and is computationally efficient because no bootstrap or permutation is required. TWO-SIGMA-G is one of the first competitive testing methods for scRNA-seq data, and shows great performance as compared to other methods designed for geneset testing in either bulk or single-cell RNA-sequencing data. The real data analysis demonstrates the ability of TWO-SIGMA-G to test complex gene-level hypotheses and revealed some novel findings regarding gene expression profiles in Alzheimer's patients.

Finally, the third chapter of this dissertation begins to bridge the gap between single-cell imaging data and Hi-C data for understanding the 3D structure of the genome. The ultimate goal of the research began here is to fully leverage the cell-type heterogeneity and the resolution provided by imaging methods to contribute to a refined understanding of the 3D genome, and how its architecture can vary in different cells. Hopefully, the kind of imaging data can be improved to cover larger genomic regions, and leveraged effectively to also improve statistical inference in genomic regions that are not imaged.

APPENDIX A: ADDITIONAL RESULTS FOR CHAPTER 2

A.1 More Results from the TWO-SIGMA *ad hoc* procedure

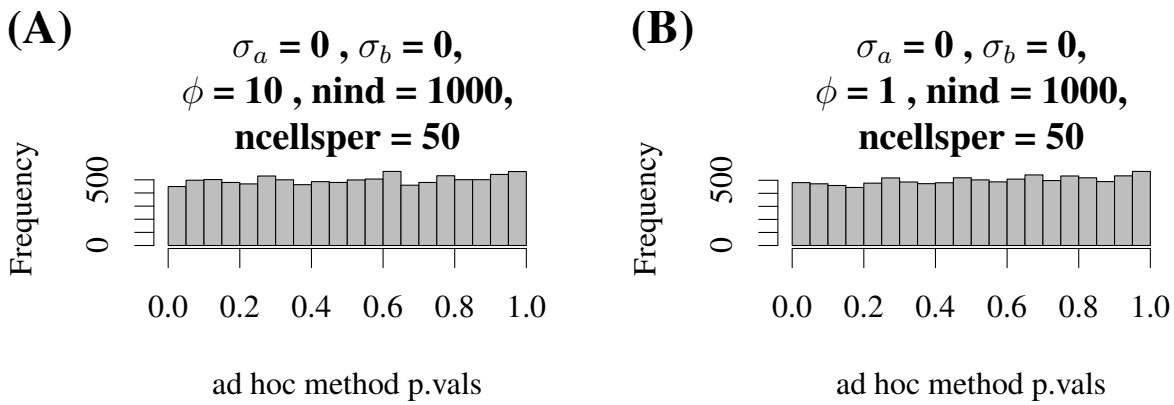

Figure A.1: *ad hoc* procedure with zero variance components: Shows the distribution of p-values from the *ad hoc* method described in the main text when variance components are zero under some representative scenarios.

For each of the simulation results discussed in the main text and summarized in the tables below, we also calculated the p-values from the *ad hoc* method for determining if random effects are needed. Figures A.1 and A.2 show histograms of p-values from representative scenarios when variance components are zero and non-zero, respectively. When variance components are zero, p-values are close to uniformly distributed, meaning that most genes will not be flagged as in need of random effects. When variance components are non-zero, the method produces small p-values which can successfully flag the need to include random effects using a p-value cutoff threshold of, for example, 0.05 or 0.10.

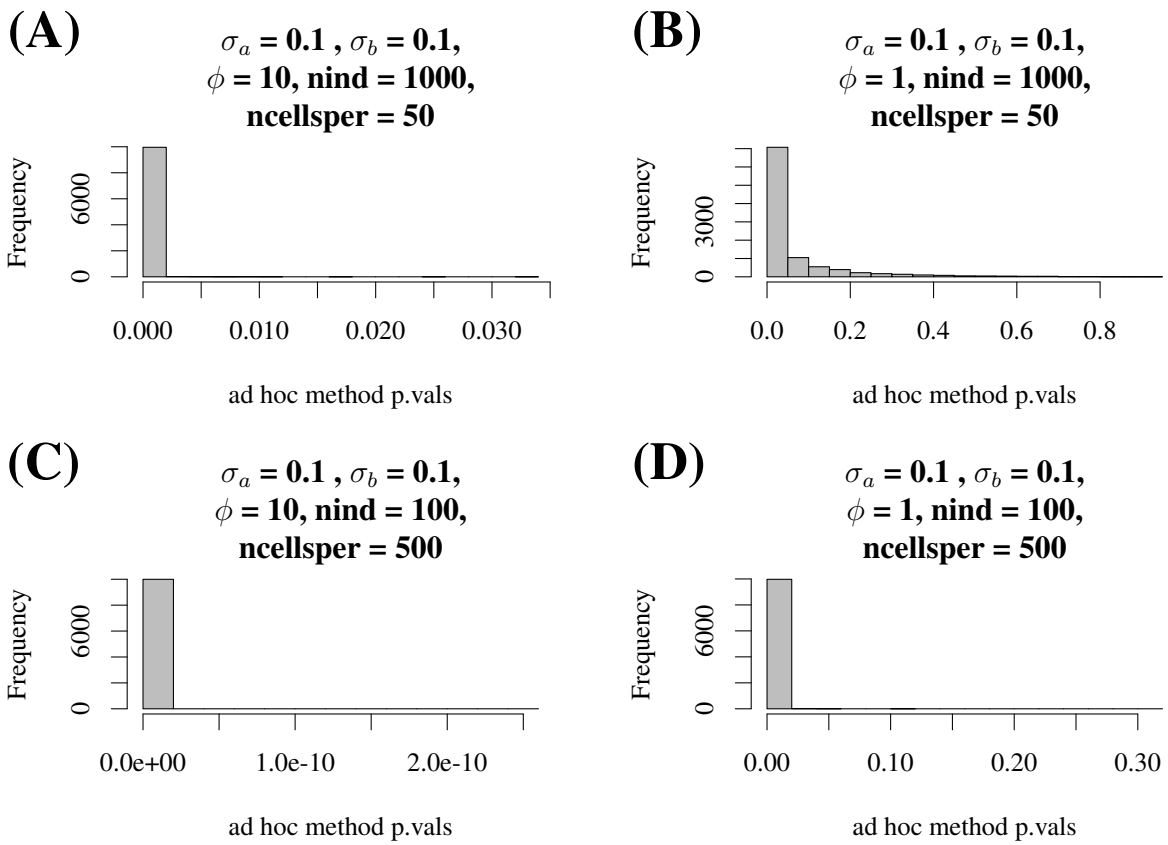

Figure A.2: *ad hoc* procedure with non-zero variance components: Shows the distribution of p-values from the *ad hoc* method described in section 3 of the main text when variance components are non-zero under some representative scenarios.

A.1.1 Extended TWO-SIGMA Type-I Error Simulation Results

Tables A.1–A.4 give more detailed results for the type-I error simulations at the 0.05 level. Each setting has results from three models: TWO-SIGMA, the zero-inflated negative binomial model (without random effects) "ZINB," and MAST Finak *et al.* (2015). Parameters α_1 and β_1 correspond to the coefficients on a binary disease status indicator, and are set to 0 under the null and are non-zero under some alternative hypothesis. Other coefficients in both components include an intercept (α_0, β_0), coefficients from simulated age values (α_2, β_2) and coefficients from simulated CDR values (α_3, β_3). Parameter values were designed to mimic realistic values observed in the pancreas data analysis. "LRT" refers to the likelihood ratio statistic (on 2 d.f.), and the combined χ^2 statistic is defined as the sum of the squared z-statistics from each of the two coefficients related to the binary disease status indicator. Coverage is given for parameters α_1 and β_1, ϕ, and σ_a and σ_b. Note that confidence intervals for the variance components are computed on the log scale and exponentiated. Therefore, the intervals will not contain zero and thus coverage for σ_a and σ_b when equal to zero is not entirely meaningful. Finally, note that the average time column includes the average over all genes of the time needed for two runs of TWO-SIGMA, MAST, and the ZINB model (each with and without the coefficients corresponding to the binary disease status indicator) and to simulate the data as well as process the results for the entire replication. These times are therefore most informative to compare within a table, and are not included to compare across the various methods—fitting a random effect term will always entail a large computational burden. Within a given table, differences in the running time are largely due to TWO-SIGMA run time. More discussion of run time is given in the paragraph below.

For example, consider table A.1. The first column "N/N Max" gives the number of genes that converged compared to the total number of genes simulated for each scenario, and as mentioned above the last column gives total runtime for all computations in a given simulation replication. One highly consistent trend is that the convergence percentage is lower and running time higher when variance components were zero. This is because the marginal likelihood for TWO-SIGMA is evaluated more times when the true values of σ_a and σ_b are on the boundary space of zero.

Comparing the first and last runtimes in table A.1 shows that there can be nearly a 50% increase in run-time when true variance components are zero yet random effects are included in the model. This underscores the usefulness of our *ad hoc* procedure to avoid fitting random effects where they are unnecessary. In table 1, Type-I error for TWO-SIGMA is well-preserved for any variance component value but becomes increasingly inflated for the ZINB model and MAST when variance components are non-zero. Furthermore, coverage from TWO-SIGMA for all parameters shown remains near the nominal level of 95%. These results hold well for all four cases varying the breakdown of the total of 50,000 cells between number of individuals and number of cells per individual, with a slight inflation of type-I error seen in table A.4 and discussed in the main text. Figures A.3 and A.4 show the observed type-I error across more stringent significance levels for representative simulation scenarios. These figures do not show systematically different patterns than seen at the .05 level in tables A.1–A.4 but can highlight the substantial type-I error seen for MAST and the ZINB model in some situations.

Case 1: 1000 individuals, 50 single-cells each, 0.05 level

		LRT	Combined χ^2	95 % CI Coverage					Simulation Parameters					Avg. Time
N / N Max	Model	Type-I Error	Type-I Error	α_1	β_1	ϕ	σ_a	σ_b	$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$	$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$	ϕ	σ_a	σ_b	(min)
	TWO-SIGMA	**0.049**	**0.049**	0.951	0.953	0.954	—	—						
9195/10000	ZINB	0.051	0.051	0.950	0.951	0.953	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0	0	31.7
	MAST	0.089	0.020	0.950	0.997	—	—	—						
	TWO-SIGMA	**0.048**	**0.047**	0.953	0.951	0.952	—	—						
9612/10000	ZINB	0.051	0.049	0.951	0.949	0.953	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	2	0	0	33.5
	MAST	0.080	0.032	0.950	0.978	—	—	—						
	TWO-SIGMA	**0.048**	**0.050**	0.954	0.953	0.949	—	—						
9464/10000	ZINB	0.052	0.053	0.952	0.951	0.949	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	1	0	0	32.2
	MAST	0.081	0.042	0.953	0.966	—	—	—						
	TWO-SIGMA	**0.051**	**0.051**	0.949	0.952	0.952	0.936	0.948						
10000/10000	ZINB	0.132	0.131	0.941	0.853	0.001	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	30.7
	MAST	0.144	0.059	0.942	0.950	—	—	—						
	TWO-SIGMA	**0.051**	**0.051**	0.950	0.949	0.951	0.936	0.963						
10000/10000	ZINB	0.078	0.078	0.945	0.918	0.666	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	2	0.1	0.1	30.7
	MAST	0.089	0.048	0.944	0.960	—	—	—						
	TWO-SIGMA	**0.049**	**0.049**	0.948	0.948	0.950	0.935	0.954						
10000/10000	ZINB	0.066	0.066	0.941	0.930	0.869	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	1	0.1	0.1	30.1
	MAST	0.095	0.049	0.941	0.960	—	—	—						
	TWO-SIGMA	**0.051**	**0.051**	0.947	0.952	0.946	0.944	0.949						
10000/10000	ZINB	0.621	0.621	0.776	0.417	0	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.5	0.5	20.9
	MAST	0.290	0.494	0.778	0.556	—	—	—						
	TWO-SIGMA	**0.053**	**0.053**	0.948	0.948	0.947	0.947	0.946						
9999/10000	ZINB	0.505	0.503	0.794	0.539	0	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	2	0.5	0.5	18.0
	MAST	0.275	0.400	0.792	0.658	—	—	—						
	TWO-SIGMA	**0.050**	**0.050**	0.950	0.954	0.949	0.950	0.948						
10000/10000	ZINB	0.404	0.398	0.817	0.639	0	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	1	0.5	0.5	17.0
	MAST	0.247	0.301	0.810	0.759	—	—	—						

Table A.1: Type-I Error using LRT to test $H_0 : \alpha_1 = 0, \beta_1 = 0$ with a significance level of 0.05

Case 2: 500 individuals, 100 single-cells each, 0.05 level

		LRT	Combined χ^2	95 % CI Coverage					Simulation Parameters					Avg. Time
N / N Max	Model	Type-I Error	Type-I Error	α_1	β_1	ϕ	σ_a	σ_b	$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$	$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$	ϕ	σ_a	σ_b	(min)
	TWO-SIGMA	**0.044**	**0.044**	0.955	0.952	0.950	—	—						
8895/10000	ZINB	0.047	0.047	0.953	0.950	0.952	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0	0	31.8
	MAST	0.086	0.020	0.953	0.996	—	—	—						
	TWO-SIGMA	**0.042**	**0.043**	0.952	0.957	0.946	—	—						
9285/10000	ZINB	0.045	0.046	0.949	0.955	0.949	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	2	0	0	32.7
	MAST	0.084	0.030	0.949	0.981	—	—	—						
	TWO-SIGMA	**0.046**	**0.045**	0.952	0.952	0.952	—	—						
9502/10000	ZINB	0.049	0.049	0.951	0.949	0.951	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	1	0	0	30.6
	MAST	0.085	0.039	0.950	0.969	—	—	—						
	TWO-SIGMA	**0.052**	**0.052**	0.942	0.951	0.950	0.955	0.948						
10000/10000	ZINB	0.217	0.218	0.926	0.754	0.002	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	23.8
	MAST	0.187	0.099	0.926	0.899	—	—	—						
	TWO-SIGMA	**0.050**	**0.052**	0.948	0.953	0.950	0.948	0.958						
10000/10000	ZINB	0.106	0.105	0.935	0.885	0.670	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	2	0.1	0.1	25.3
	MAST	0.104	0.065	0.934	0.944	—	—	—						
	TWO-SIGMA	**0.052**	**0.052**	0.950	0.950	0.949	0.946	0.967						
10000/10000	ZINB	0.081	0.082	0.939	0.914	0.866	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	1	0.1	0.1	22.9
	MAST	0.107	0.063	0.938	0.945	—	—	—						
	TWO-SIGMA	**0.051**	**0.052**	0.950	0.949	0.946	0.948	0.947						
10000/10000	ZINB	0.753	0.753	0.667	0.307	0	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.5	0.5	17.8
	MAST	0.419	0.673	0.670	0.414	—	—	—						
	TWO-SIGMA	**0.054**	**0.055**	0.951	0.942	0.953	0.949	0.950						
10000/10000	ZINB	0.663	0.663	0.691	0.414	0	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	2	0.5	0.5	18.0
	MAST	0.382	0.579	0.685	0.516	—	—	—						
	TWO-SIGMA	**0.054**	**0.054**	0.944	0.948	0.947	0.950	0.942						
10000/10000	ZINB	0.583	0.577	0.705	0.509	0	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	1	0.5	0.5	15.5
	MAST	0.366	0.494	0.696	0.628	—	—	—						

Table A.2: Type-I Error using LRT to test $H_0 : \alpha_1 = 0, \beta_1 = 0$ with a significance level of 0.05

Case 3: 100 individuals, 500 single-cells each, 0.05 level

		LRT	Combined χ^2	95 % CI Coverage					Simulation Parameters					Avg. Time
N / N Max	Model	Type-I Error	Type-I Error	α_1	β_1	ϕ	σ_a	σ_b	$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$	$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$	ϕ	σ_a	σ_b	(min)
	TWO-SIGMA	**0.042**	**0.044**	0.954	0.953	0.951	—	—						
8773/10000	ZINB	0.050	0.050	0.950	0.950	0.950	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0	0	32.1
	MAST	0.090	0.021	0.950	0.995	—	—	—						
	TWO-SIGMA	**0.038**	**0.038**	0.960	0.957	0.953	—	—						
8901/10000	ZINB	0.044	0.043	0.955	0.954	0.953	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	2	0	0	32.1
	MAST	0.079	0.028	0.955	0.977	—	—	—						
	TWO-SIGMA	**0.044**	**0.045**	0.956	0.954	0.951	—	—						
9199/10000	ZINB	0.051	0.050	0.952	0.949	0.951	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	1	0	0	31.3
	MAST	0.087	0.038	0.950	0.969	—	—	—						
	TWO-SIGMA	**0.056**	**0.059**	0.938	0.947	0.951	0.979	0.938						
9999/10000	ZINB	0.534	0.533	0.869	0.465	0.007	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	25.7
	MAST	0.313	0.376	0.869	0.634	—	—	—						
	TWO-SIGMA	**0.057**	**0.060**	0.940	0.942	0.952	0.978	0.943						
10000/10000	ZINB	0.323	0.320	0.877	0.685	0.673	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	2	0.1	0.1	25.9
	MAST	0.176	0.226	0.872	0.791	—	—	—						
	TWO-SIGMA	**0.053**	**0.058**	0.939	0.947	0.950	0.977	0.955						
10000/10000	ZINB	0.224	0.219	0.887	0.789	0.883	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	1	0.1	0.1	25.6
	MAST	0.174	0.169	0.882	0.860	—	—	—						
	TWO-SIGMA	**0.055**	**0.058**	0.945	0.942	0.951	0.935	0.936						
10000/10000	ZINB	0.941	0.941	0.367	0.142	0	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.5	0.5	21.4
	MAST	0.716	0.914	0.367	0.193	—	—	—						
	TWO-SIGMA	**0.056**	**0.060**	0.940	0.945	0.950	0.936	0.934						
10000/10000	ZINB	0.909	0.909	0.386	0.196	0	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	2	0.5	0.5	20.4
	MAST	0.685	0.884	0.383	0.254	—	—	—						
	TWO-SIGMA	**0.053**	**0.056**	0.943	0.947	0.952	0.939	0.934						
10000/10000	ZINB	0.873	0.872	0.412	0.256	0	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	1	0.5	0.5	20.0
	MAST	0.649	0.839	0.400	0.324	—	—	—						

Table A.3: Type-I Error using LRT to test $H_0 : \alpha_1 = 0, \beta_1 = 0$ with a significance level of 0.05

Case 4: 25 individuals, 2000 single-cells each, 0.05 level

		LRT	Combined χ^2	95 % CI Coverage					Simulation Parameters					Avg. Time
N / N Max	Model	Type-I Error	Type-I Error	α_1	β_1	ϕ	σ_a	σ_b	$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$	$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$	ϕ	σ_a	σ_b	(min)
	TWO-SIGMA	**0.041**	**0.045**	0.953	0.951	0.950	—	—						
8698/10000	ZINB	0.052	0.052	0.947	0.946	0.950	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0	0	28.9
	MAST	0.090	0.021	0.947	0.995	—	—	—						
	TWO-SIGMA	**0.041**	**0.046**	0.955	0.954	0.952	—	—						
8585/10000	ZINB	0.052	0.053	0.949	0.949	0.952	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	2	0	0	28.3
	MAST	0.086	0.034	0.948	0.976	—	—	—						
	TWO-SIGMA	**0.042**	**0.044**	0.954	0.954	0.946	—	—						
8763/10000	ZINB	0.051	0.050	0.949	0.949	0.946	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	1	0	0	27.8
	MAST	0.090	0.041	0.949	0.966	—	—	—						
	TWO-SIGMA	**0.076**	**0.088**	0.920	0.923	0.949	0.980	0.909						
9544/10000	ZINB	0.817	0.817	0.689	0.235	0.056	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	22.3
	MAST	0.497	0.720	0.689	0.354	—	—	—						
	TWO-SIGMA	**0.072**	**0.087**	0.926	0.923	0.946	0.994	0.896						
9999/10000	ZINB	0.643	0.642	0.708	0.424	0.719	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	2	0.1	0.1	26.4
	MAST	0.361	0.562	0.704	0.527	—	—	—						
	TWO-SIGMA	**0.075**	**0.094**	0.922	0.923	0.949	0.992	0.906						
10000/10000	ZINB	0.548	0.542	0.733	0.541	0.880	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	1	0.1	0.1	26.3
	MAST	0.361	0.467	0.718	0.637	—	—	—						
	TWO-SIGMA	**0.076**	**0.094**	0.920	0.920	0.951	0.888	0.888						
10000/10000	ZINB	0.984	0.984	0.194	0.070	0	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.5	0.5	22.7
	MAST	0.875	0.979	0.195	0.098	—	—	—						
	TWO-SIGMA	**0.076**	**0.092**	0.925	0.922	0.949	0.886	0.882						
10000/10000	ZINB	0.974	0.975	0.202	0.101	0	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	2	0.5	0.5	22.0
	MAST	0.857	0.966	0.197	0.132	—	—	—						
	TWO-SIGMA	**0.074**	**0.089**	0.923	0.922	0.950	0.891	0.880						
10000/10000	ZINB	0.964	0.963	0.218	0.135	0	—	—	(1, **0**, -0.5, -2)	(2, **0**, -0.1, 0.6)	1	0.5	0.5	22.5
	MAST	0.827	0.953	0.213	0.174	—	—	—						

Table A.4: Type-I Error using LRT to test $H_0 : \alpha_1 = 0, \beta_1 = 0$ with a significance level of 0.05

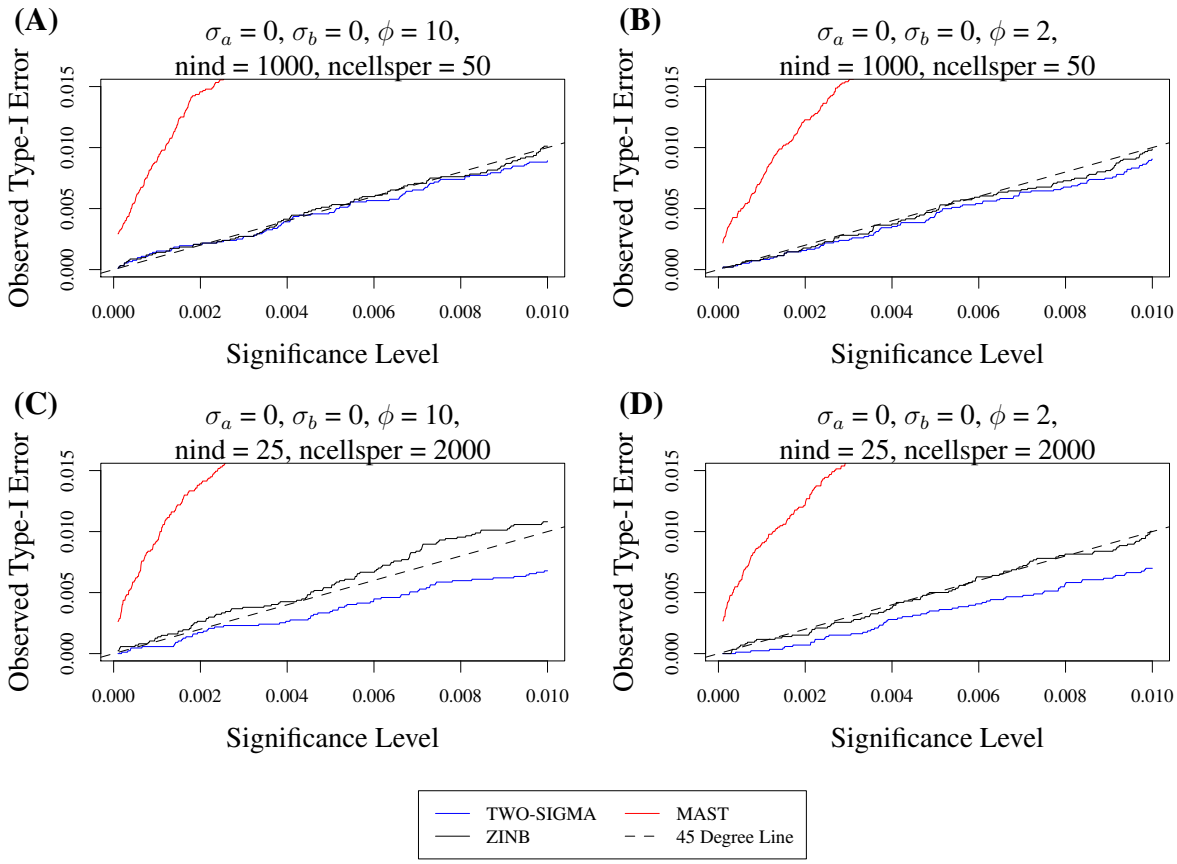

Figure A.3: Type-I error across different significance levels: Shows the observed type-I error across various nominal significance levels.

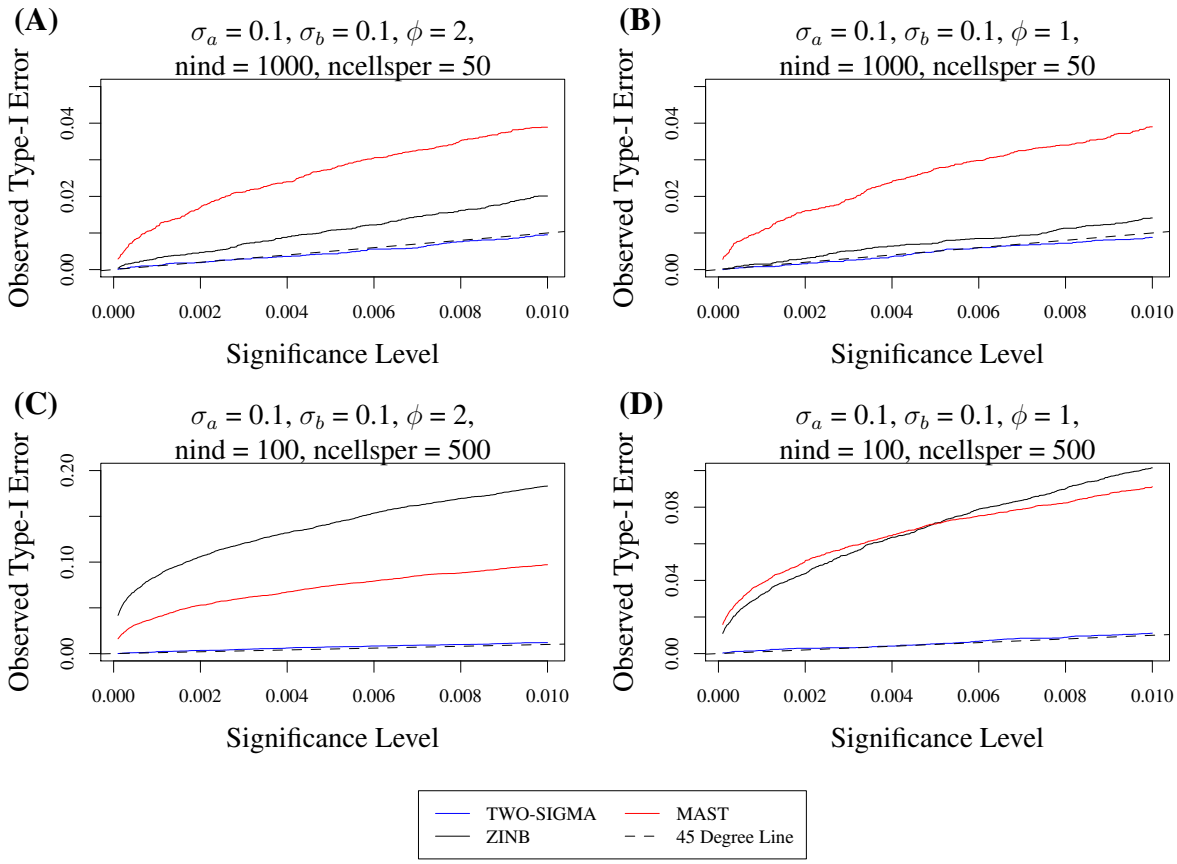

Figure A.4: Type-I error across different significance levels: Shows the observed type-I error across various nominal significance levels.

A.1.2 Extended TWO-SIGMA Power Results

Simulations under the same framework were also performed for non-zero values of α_1 and β_1 (both as defined in the previous section) to evaluate the power of TWO-SIGMA in testing $H_0 : \alpha_1 = 0, \beta_1 = 0$. As seen in Table 1 of the main text and tables A.1–A.4, MAST and the ZINB model can suffer from vastly inflated type-I error. Thus, the observed (or "apparent") power does not always provide a fair comparison to TWO-SIGMA. For each of the three methods we therefore calculated empirical significance thresholds for all null simulation settings. These are cutoffs such that the percentage of statistics larger than the threshold equals the significance level. "True" power is then calculated by rejecting the null if the test statistic is larger than the empirical significance threshold from the corresponding simulation setting under the null. In simulation settings this does not add computation, but in real data setting this procedure involves additional computation and is therefore not preferred.

Because the type-I error for TWO-SIGMA is approximately preserved in all four sample size cases, true power is nearly identical to apparent power for TWO-SIGMA. We therefore found it unnecessary to use true power for TWO-SIGMA in figures A.5-A.7 shown here and figure 2 in the main text. In contrast, true power can be very different than apparent power for both the ZINB model and MAST given their inflated type-I errors. For example, one simulation setting shows that the apparent power of MAST is 0.375 which the true power for this scenario is only 0.194 (see the third rows of table A.5 and A.9). Although not presented, this discrepancy between apparent and true power would be even more pronounced if the simulated data here were based on larger values of the variance components σ_a and σ_b because type-I errors are more inflated for larger variance components (see tables A.1-A.4).

One general observation from tables A.9 to A.12 and figure 2 of the main text is that the ZINB model retains very high true power in both sample size settings and across all four effect scenarios. For smaller values of α_0 the ZINB model can sometimes have higher true power than TWO-SIGMA. As the dropout proportion increases (via increasing α_0), TWO-SIGMA tends to eventually have higher power. TWO-SIGMA does not require the use of computationally expen-

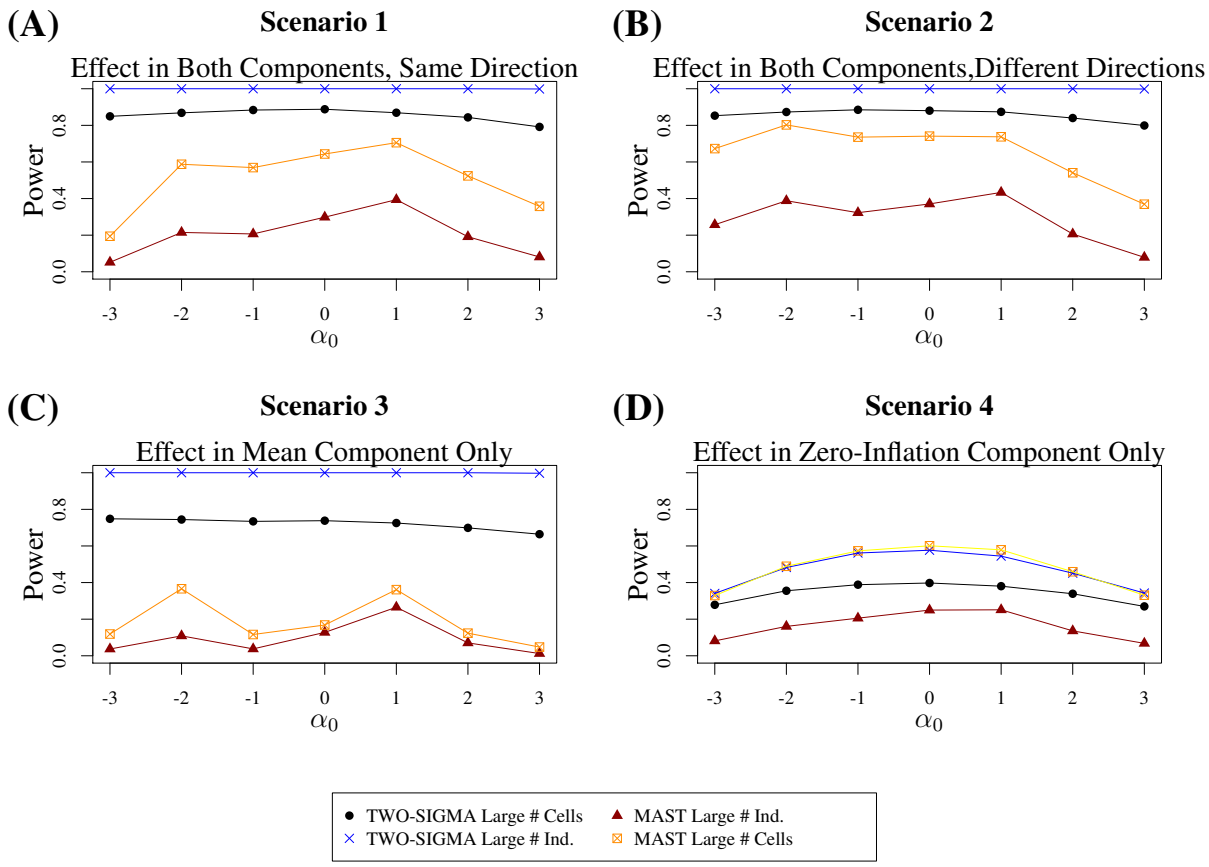

Figure A.5: Power evaluations in simulated data: Shows the power to test $H_0 : \alpha_1 = \beta_1 = 0$ by varying the intercept α_0 to control the drop-out proportion in four setups: TWO-SIGMA and MAST with 50 cells from each of 1000 individuals or 500 cells from each of 100 individuals. Values of ϕ, σ_a, and σ_b were all set to 0.1 and an effect size of 0.03 was used. Larger values of α_0 correspond to more drop-out in the data. 10,000 genes were simulated. Because of the type-I error inflation from MAST seen in tables A.1–A.4, true power was calculated and plotted using the empirical significance threshold from the corresponding setting under the null. TWO-SIGMA retains higher power in the first three scenarios and half of the fourth scenario without the need to use true power. See section A.1.2 for more details about computing true power and discussion regarding power trends across all three methods.

sive resampling procedures for valid inference, giving it a key advantage over the ZINB model, which is furthermore not articulated explicitly as a DE method for scRNA-seq data, but included to contrast the impact of R.E.s on model performance.

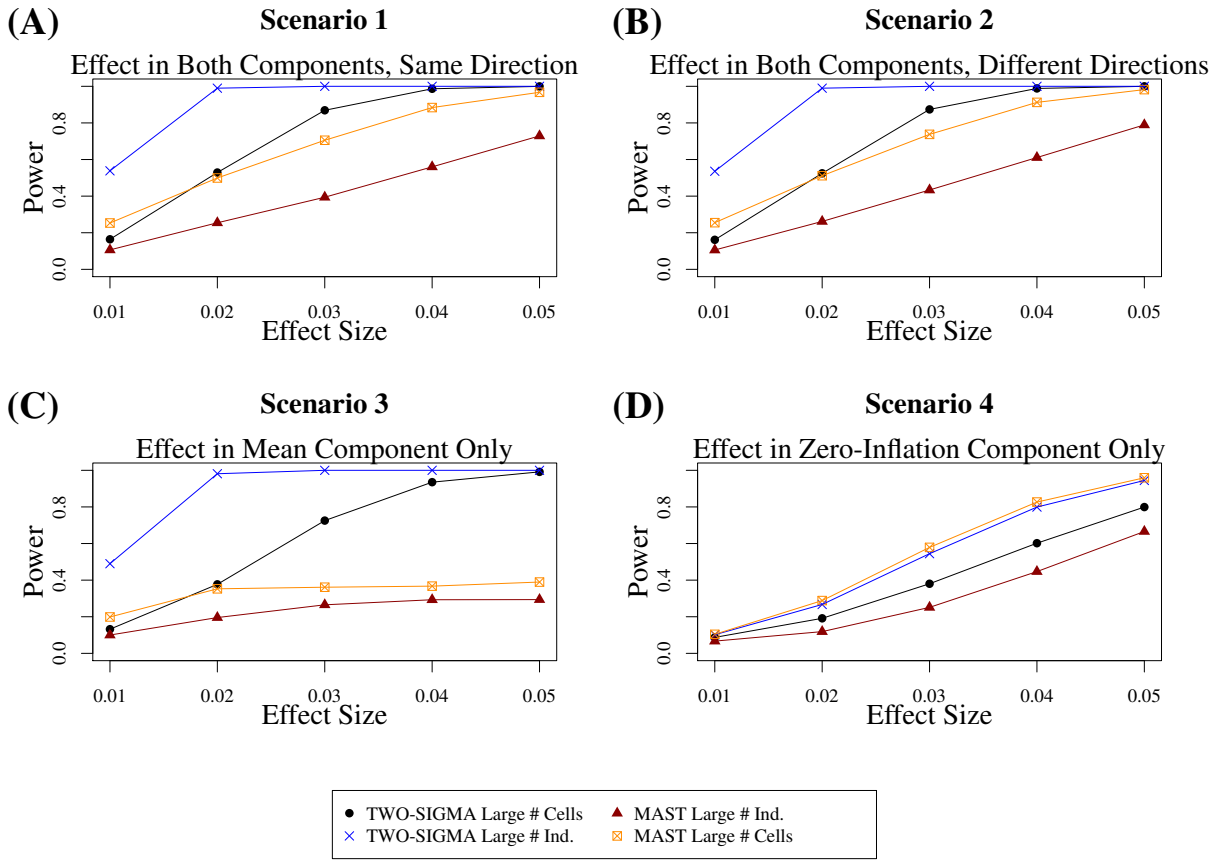

Figure A.6: Power evaluations in simulated data: Shows the power to test $H_0 : \alpha_1 = \beta_1 = 0$ by varying the effect size in two sample size setups: 50 cells from each of 1000 individuals or 500 cells from each of 100 individuals. Values of ϕ, σ_a, and σ_b were all set to 0.1 and 10,000 genes were simulated. Because of the type-I error inflation from the ZINB model and MAST seen in tables A.1–A.4, true power was calculated and plotted using the empirical significance threshold from the corresponding setting under the null for both of these methods. TWO-SIGMA retains higher power in the first three scenarios and half of the fourth scenario without the need to use true power. See the discussion at the beginning of section A.1.2 for more details about computing true power and discussion regarding power trends across all differing methods.

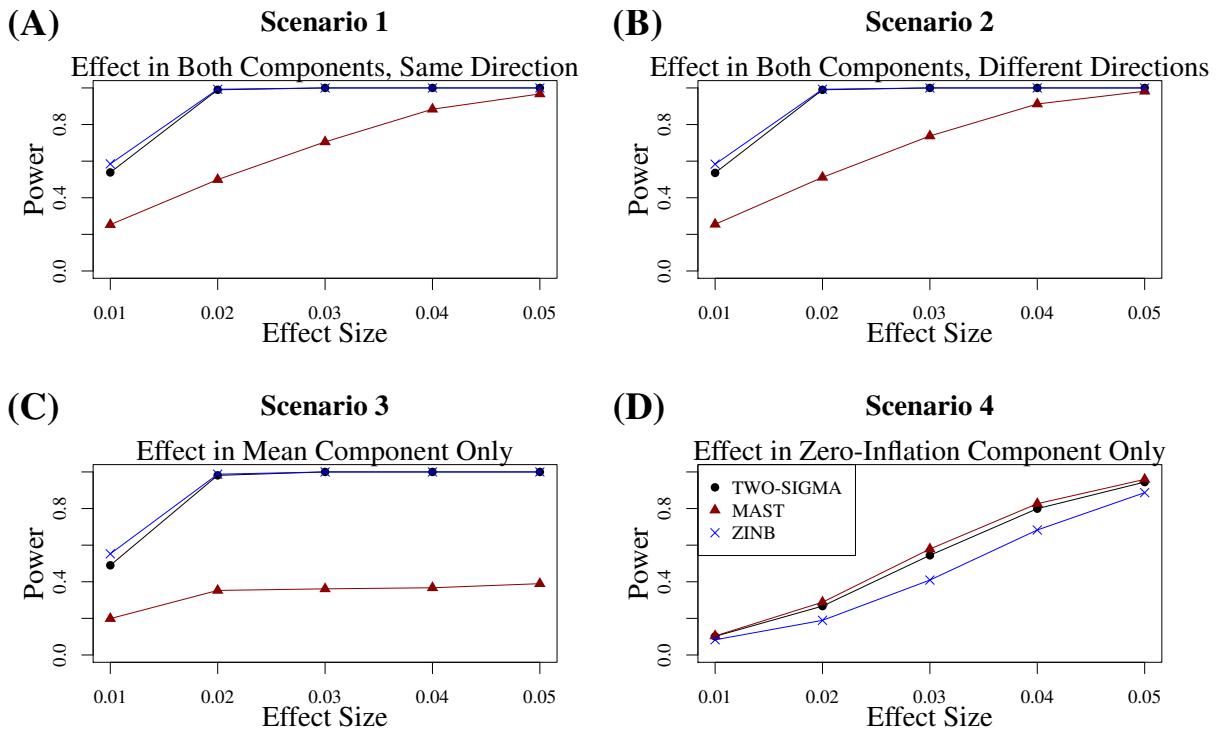

Figure A.7: Power evaluations in simulated data: Shows the power to test $H_0 : \alpha_1 = \beta_1 = 0$ by varying the effect size with 50 cells from each of 1000 individuals. Values of ϕ, σ_a, and σ_b were all set to 0.1 and 10,000 genes were simulated. Because of the type-I error inflation from the ZINB model and MAST seen in tables A.1–A.4, true power was calculated and plotted using the empirical significance threshold from the corresponding setting under the null for these two methods. In the first three scenarios, MAST consistently has lower true power while TWO-SIGMA and the ZINB model typically have very similar true power. When the effect is only in the zero-inflation component, power is lower for all methods at all effect sizes. Using TWO-SIGMA can bypass the need for computationally expensive resampling procedures needed to generate true power. See the discussion at the beginning of section A.1.2 for more details about computing true power and discussion regarding power trends across all differing methods.

A.1.2.1 Results using "Apparent" Power for MAST and ZINB model

Case 1: 1000 individuals, 50 single-cells each, 0.05 level

Power Scenarios 1 & 2 from Figure 1 of main text

Effects in both components in either the same or different directions

		LRT	Combined χ^2	Simulation Parameters					Avg. Time
N / N Max	Model	P(Reject H_0)	P(Reject H_0)	$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$	$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$	ϕ	σ_a	σ_b	(min)
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-3, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	27.9
	MAST	0.375	1.000						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-2, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	27.8
	MAST	0.738	1.000						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-1, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	27.0
	MAST	0.767	1.000						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(0, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	29.4
	MAST	0.816	1.000						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(1, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	29.4
	MAST	0.834	1.000						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(2, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	27.5
	MAST	0.720	0.999						
	TWO-SIGMA	**0.999**	**0.999**						
10000/10000	ZINB	1.000	1.000	(3, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	27.2
	MAST	0.572	0.986						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-3, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	27.4
	MAST	0.843	1.000						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-2, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	27.6
	MAST	0.913	1.000						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-1, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	27.7
	MAST	0.884	1.000						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(0, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	29.5
	MAST	0.879	1.000						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(1, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	28.2
	MAST	0.863	1.000						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(2, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	27.0
	MAST	0.734	0.999						
	TWO-SIGMA	**0.999**	**0.999**						
10000/10000	ZINB	0.999	0.999	(3, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	26.2
	MAST	0.579	0.985						

Table A.5: Apparent Power using LRT to test $H_0 : \alpha_1 = 0, \beta_1 = 0$ with a significance level of 0.05

Case 1: 1000 individuals, 50 single-cells each, 0.05 level

Power Scenarios 3 & 4 from Figure 1 of main text

Effects in one component at a time

		LRT	Combined χ^2	Simulation Parameters					Avg. Time
N / N Max	Model	P(Reject H_0)	P(Reject H_0)	$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$	$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$	ϕ	σ_a	σ_b	(min)
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-3, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	26.6
	MAST	0.262	1.000						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-2, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	27.3
	MAST	0.461	1.000						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-1, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	26.9
	MAST	0.246	1.000						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(0, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	29.8
	MAST	0.280	1.000						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(1, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	30.2
	MAST	0.420	1.000						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(2, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	29.5
	MAST	0.211	0.998						
	TWO-SIGMA	**0.997**	**0.997**						
10000/10000	ZINB	0.999	0.999	(3, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	29.7
	MAST	0.134	0.966						
	TWO-SIGMA	**0.341**	**0.343**						
10000/10000	ZINB	0.495	0.495	(-3, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	26.6
	MAST	0.546	0.368						
	TWO-SIGMA	**0.482**	**0.484**						
10000/10000	ZINB	0.611	0.611	(-2, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	26.4
	MAST	0.703	0.513						
	TWO-SIGMA	**0.561**	**0.563**						
9999/10000	ZINB	0.668	0.669	(-1, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	29.3
	MAST	0.768	0.588						
	TWO-SIGMA	**0.577**	**0.577**						
10000/10000	ZINB	0.674	0.676	(0, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	30.9
	MAST	0.786	0.606						
	TWO-SIGMA	**0.544**	**0.546**						
10000/10000	ZINB	0.624	0.624	(1, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	31.0
	MAST	0.768	0.566						
	TWO-SIGMA	**0.451**	**0.452**						
10000/10000	ZINB	0.526	0.526	(2, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	28.6
	MAST	0.670	0.449						
	TWO-SIGMA	**0.343**	**0.343**						
10000/10000	ZINB	0.394	0.393	(3, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	28.4
	MAST	0.539	0.318						

Table A.6: Apparent Power using LRT to test $H_0 : \alpha_1 = 0, \beta_1 = 0$ with a significance level of 0.05

Case 2: 100 individuals, 500 single-cells each, 0.05 level

Power Scenarios 1 & 2 from Figure 1 of main text

Effects in both components in either the same or different directions

		LRT	Combined χ^2	Simulation Parameters					Avg. Time
N / N Max	Model	P(Reject H_0)	P(Reject H_0)	$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$	$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$	ϕ	σ_a	σ_b	(min)
	TWO-SIGMA	**0.849**	**0.858**						
10000/10000	ZINB	0.995	0.995	(-3, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	21.6
	MAST	0.441	0.988						
	TWO-SIGMA	**0.868**	**0.873**						
10000/10000	ZINB	0.997	0.997	(-2, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	23.9
	MAST	0.712	0.992						
	TWO-SIGMA	**0.884**	**0.890**						
10000/10000	ZINB	0.996	0.996	(-1, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	22.6
	MAST	0.739	0.991						
	TWO-SIGMA	**0.888**	**0.893**						
10000/10000	ZINB	0.996	0.996	(0, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	23.7
	MAST	0.784	0.990						
	TWO-SIGMA	**0.869**	**0.875**						
10000/10000	ZINB	0.993	0.993	(1, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	26.3
	MAST	0.790	0.985						
	TWO-SIGMA	**0.844**	**0.850**						
10000/10000	ZINB	0.986	0.986	(2, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	25.6
	MAST	0.672	0.964						
	TWO-SIGMA	**0.792**	**0.799**						
10000/10000	ZINB	0.969	0.969	(3, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	26.1
	MAST	0.564	0.914						
	TWO-SIGMA	**0.853**	**0.858**						
9996/10000	ZINB	0.996	0.996	(-3, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	22.1
	MAST	0.802	0.991						
	TWO-SIGMA	**0.873**	**0.876**						
10000/10000	ZINB	0.997	0.997	(-2, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	23.8
	MAST	0.869	0.992						
	TWO-SIGMA	**0.885**	**0.889**						
9999/10000	ZINB	0.997	0.997	(-1, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	22.7
	MAST	0.839	0.992						
	TWO-SIGMA	**0.880**	**0.885**						
10000/10000	ZINB	0.996	0.996	(0, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	23.3
	MAST	0.839	0.990						
	TWO-SIGMA	**0.874**	**0.878**						
9999/10000	ZINB	0.993	0.993	(1, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	25.8
	MAST	0.824	0.985						
	TWO-SIGMA	**0.840**	**0.846**						
10000/10000	ZINB	0.989	0.989	(2, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	25.9
	MAST	0.709	0.965						
	TWO-SIGMA	**0.799**	**0.806**						
10000/10000	ZINB	0.974	0.974	(3, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	25.6
	MAST	0.573	0.919						

Table A.7: Apparent Power using LRT to test $H_0 : \alpha_1 = 0, \beta_1 = 0$ with a significance level of 0.05

Case 2: 100 individuals, 500 single-cells each, 0.05 level

Power Scenarios 3 & 4 from Figure 1 of main text

Effects in one component at a time

		LRT	Combined χ^2	Simulation Parameters					Avg. Time
N / N Max	Model	P(Reject H_0)	P(Reject H_0)	$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$	$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$	ϕ	σ_a	σ_b	(min)
9999/10000	TWO-SIGMA	**0.748**	**0.757**	(-3, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	21.9
	ZINB	0.992	0.992						
	MAST	0.341	0.982						
10000/10000	TWO-SIGMA	**0.744**	**0.754**	(-2, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	22.7
	ZINB	0.993	0.993						
	MAST	0.478	0.981						
10000/10000	TWO-SIGMA	**0.734**	**0.744**	(-1, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	21.8
	ZINB	0.991	0.991						
	MAST	0.347	0.978						
10000/10000	TWO-SIGMA	**0.738**	**0.746**	(0, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	22.1
	ZINB	0.991	0.990						
	MAST	0.369	0.975						
10000/10000	TWO-SIGMA	**0.725**	**0.735**	(1, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	25.4
	ZINB	0.984	0.984						
	MAST	0.444	0.962						
10000/10000	TWO-SIGMA	**0.699**	**0.711**	(2, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	24.9
	ZINB	0.974	0.974						
	MAST	0.282	0.923						
10000/10000	TWO-SIGMA	**0.664**	**0.677**	(3, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	26.6
	ZINB	0.949	0.949						
	MAST	0.187	0.851						
9999/10000	TWO-SIGMA	**0.279**	**0.289**	(-3, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	22.7
	ZINB	0.770	0.769						
	MAST	0.558	0.654						
9999/10000	TWO-SIGMA	**0.355**	**0.367**	(-2, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	21.1
	ZINB	0.819	0.820						
	MAST	0.688	0.729						
10000/10000	TWO-SIGMA	**0.388**	**0.400**	(-1, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	19.7
	ZINB	0.834	0.834						
	MAST	0.732	0.756						
10000/10000	TWO-SIGMA	**0.398**	**0.408**	(0, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	20.6
	ZINB	0.823	0.823						
	MAST	0.755	0.744						
9899/10000	TWO-SIGMA	**0.380**	**0.391**	(1, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	25.3
	ZINB	0.784	0.784						
	MAST	0.754	0.706						
10000/10000	TWO-SIGMA	**0.339**	**0.349**	(2, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	26.4
	ZINB	0.718	0.719						
	MAST	0.651	0.601						
10000/10000	TWO-SIGMA	**0.270**	**0.280**	(3, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	26.1
	ZINB	0.588	0.588						
	MAST	0.533	0.446						

Table A.8: Apparent Power using LRT to test $H_0 : \alpha_1 = 0, \beta_1 = 0$ with a significance level of 0.05

A.1.2.2 Results using "True" Power for MAST and the ZINB model

Case 1: 1000 individuals, 50 single-cells each, 0.05 level

Power Scenarios 1 & 2 from Figure 1 of main text

Effects in both components in either the same or different directions

		LRT	Combined χ^2	Simulation Parameters					Avg. Time
N / N Max	Model	P(Reject H_0)	P(Reject H_0)	$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$	$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$	ϕ	σ_a	σ_b	(min)
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-3, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	27.9
	MAST	0.194	0.352						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-2, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	27.8
	MAST	0.588	0.721						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-1, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	27.0
	MAST	0.569	0.746						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(0, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	29.4
	MAST	0.644	0.802						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(1, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	29.4
	MAST	0.705	0.822						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(2, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	27.5
	MAST	0.524	0.700						
	TWO-SIGMA	**0.999**	**0.999**						
10000/10000	ZINB	0.998	0.998	(3, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	27.2
	MAST	0.358	0.547						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-3, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	27.4
	MAST	0.673	0.829						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-2, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	27.6
	MAST	0.803	0.904						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-1, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	27.7
	MAST	0.736	0.872						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(0, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	29.5
	MAST	0.741	0.867						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(1, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	28.2
	MAST	0.738	0.852						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(2, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	27.0
	MAST	0.541	0.715						
	TWO-SIGMA	**0.998**	**0.998**						
10000/10000	ZINB	0.997	0.997	(3, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	26.2
	MAST	0.369	0.557						

Table A.9: True Power using LRT to test $H_0 : \alpha_1 = 0, \beta_1 = 0$ with a significance level of 0.05

Case 1: 1000 individuals, 50 single-cells each, 0.05 level

Power Scenarios 3 & 4 from Figure 1 of main text

Effects in one component at a time

		LRT	Combined χ^2	Simulation Parameters					Avg. Time
N / N Max	Model	P(Reject H_0)	P(Reject H_0)	$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$	$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$	ϕ	σ_a	σ_b	(min)
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-3, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	26.6
	MAST	0.119	0.242						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-2, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	27.3
	MAST	0.366	0.450						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(-1, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	26.9
	MAST	0.116	0.229						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(0, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	29.8
	MAST	0.168	0.263						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(1, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	30.2
	MAST	0.361	0.411						
	TWO-SIGMA	**1.000**	**1.000**						
10000/10000	ZINB	1.000	1.000	(2, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	29.5
	MAST	0.124	0.200						
	TWO-SIGMA	**0.997**	**0.997**						
10000/10000	ZINB	0.996	0.996	(3, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	29.7
	MAST	0.048	0.120						
	TWO-SIGMA	**0.336**	**0.335**						
10000/10000	ZINB	0.295	0.294	(-3, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	26.6
	MAST	0.329	0.522						
	TWO-SIGMA	**0.476**	**0.474**						
10000/10000	ZINB	0.401	0.400	(-2, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	26.4
	MAST	0.489	0.684						
	TWO-SIGMA	**0.554**	**0.553**						
9999/10000	ZINB	0.459	0.459	(-1, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	29.3
	MAST	0.573	0.747						
	TWO-SIGMA	**0.570**	**0.568**						
10000/10000	ZINB	0.465	0.464	(0, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	30.9
	MAST	0.601	0.769						
	TWO-SIGMA	**0.539**	**0.537**						
10000/10000	ZINB	0.409	0.408	(1, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	31.0
	MAST	0.579	0.750						
	TWO-SIGMA	**0.445**	**0.444**						
10000/10000	ZINB	0.306	0.305	(2, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	28.6
	MAST	0.459	0.649						
	TWO-SIGMA	**0.337**	**0.335**						
10000/10000	ZINB	0.205	0.204	(3, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	28.4
	MAST	0.331	0.518						

Table A.10: True Power using LRT to test $H_0 : \alpha_1 = 0, \beta_1 = 0$ with a significance level of 0.05

Case 2: 100 individuals, 500 single-cells each, 0.05 level

Power Scenarios 1 & 2 from Figure 1 of main text

Effects in both components in either the same or different directions

		LRT	Combined χ^2	Simulation Parameters					Avg. Time
N / N Max	Model	P(Reject H_0)	P(Reject H_0)	$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$	$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$	ϕ	σ_a	σ_b	(min)
	TWO-SIGMA	**0.837**	**0.830**						
10000/10000	ZINB	0.937	0.937	(-3, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	21.6
	MAST	0.052	0.085						
	TWO-SIGMA	**0.858**	**0.852**						
10000/10000	ZINB	0.935	0.934	(-2, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	23.9
	MAST	0.215	0.298						
	TWO-SIGMA	**0.872**	**0.866**						
10000/10000	ZINB	0.926	0.926	(-1, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	22.6
	MAST	0.207	0.290						
	TWO-SIGMA	**0.876**	**0.870**						
10000/10000	ZINB	0.899	0.899	(0, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	23.7
	MAST	0.298	0.373						
	TWO-SIGMA	**0.859**	**0.852**						
10000/10000	ZINB	0.840	0.840	(1, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	26.3
	MAST	0.394	0.456						
	TWO-SIGMA	**0.831**	**0.822**						
10000/10000	ZINB	0.729	0.728	(2, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	25.6
	MAST	0.191	0.252						
	TWO-SIGMA	**0.778**	**0.768**						
10000/10000	ZINB	0.502	0.500	(3, **0.03**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	26.1
	MAST	0.081	0.130						
	TWO-SIGMA	**0.842**	**0.835**						
9996/10000	ZINB	0.943	0.943	(-3, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	22.1
	MAST	0.257	0.350						
	TWO-SIGMA	**0.862**	**0.855**						
10000/10000	ZINB	0.937	0.937	(-2, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	23.8
	MAST	0.388	0.483						
	TWO-SIGMA	**0.875**	**0.868**						
9999/10000	ZINB	0.927	0.926	(-1, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	22.7
	MAST	0.323	0.417						
	TWO-SIGMA	**0.870**	**0.864**						
10000/10000	ZINB	0.901	0.900	(0, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	23.3
	MAST	0.371	0.452						
	TWO-SIGMA	**0.862**	**0.855**						
9999/10000	ZINB	0.847	0.846	(1, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	25.8
	MAST	0.434	0.503						
	TWO-SIGMA	**0.827**	**0.819**						
10000/10000	ZINB	0.717	0.716	(2, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	25.9
	MAST	0.206	0.278						
	TWO-SIGMA	**0.785**	**0.775**						
10000/10000	ZINB	0.503	0.500	(3, **0.03**, -0.5, -2)	(2, **-0.03**, -0.1, 0.6)	10	0.1	0.1	25.6
	MAST	0.079	0.132						

Table A.11: True Power using LRT to test $H_0 : \alpha_1 = 0, \beta_1 = 0$ with a significance level of 0.05

Case 2: 100 individuals, 500 single-cells each, 0.05 level

Power Scenarios 3 & 4 from Figure 1 of main text

Effects in one component at a time

		LRT	Combined χ^2	Simulation Parameters					Avg. Time
N / N Max	Model	P(Reject H_0)	P(Reject H_0)	$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$	$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$	ϕ	σ_a	σ_b	(min)
	TWO-SIGMA	**0.730**	**0.719**						
9999/10000	ZINB	0.932	0.932	(-3, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	21.9
	MAST	0.037	0.058						
	TWO-SIGMA	**0.725**	**0.716**						
10000/10000	ZINB	0.925	0.924	(-2, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	22.7
	MAST	0.109	0.152						
	TWO-SIGMA	**0.715**	**0.704**						
10000/10000	ZINB	0.907	0.906	(-1, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	21.8
	MAST	0.038	0.059						
	TWO-SIGMA	**0.723**	**0.713**						
10000/10000	ZINB	0.870	0.870	(0, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	22.1
	MAST	0.128	0.142						
	TWO-SIGMA	**0.710**	**0.699**						
10000/10000	ZINB	0.810	0.809	(1, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	25.4
	MAST	0.265	0.278						
	TWO-SIGMA	**0.680**	**0.669**						
10000/10000	ZINB	0.675	0.673	(2, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	24.9
	MAST	0.070	0.083						
	TWO-SIGMA	**0.645**	**0.632**						
10000/10000	ZINB	0.452	0.451	(3, **0**, -0.5, -2)	(2, **0.03**, -0.1, 0.6)	10	0.1	0.1	26.6
	MAST	0.012	0.020						
	TWO-SIGMA	**0.263**	**0.252**						
9999/10000	ZINB	0.184	0.184	(-3, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	22.7
	MAST	0.081	0.129						
	TWO-SIGMA	**0.337**	**0.324**						
9999/10000	ZINB	0.184	0.184	(-2, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	21.1
	MAST	0.161	0.238						
	TWO-SIGMA	**0.368**	**0.356**						
10000/10000	ZINB	0.162	0.160	(-1, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	19.7
	MAST	0.206	0.290						
	TWO-SIGMA	**0.375**	**0.364**						
10000/10000	ZINB	0.128	0.126	(0, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	20.6
	MAST	0.250	0.335						
	TWO-SIGMA	**0.358**	**0.346**						
9899/10000	ZINB	0.086	0.085	(1, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	25.3
	MAST	0.251	0.336						
	TWO-SIGMA	**0.321**	**0.310**						
10000/10000	ZINB	0.042	0.042	(2, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	26.4
	MAST	0.136	0.206						
	TWO-SIGMA	**0.253**	**0.242**						
10000/10000	ZINB	0.011	0.011	(3, **0.03**, -0.5, -2)	(2, **0**, -0.1, 0.6)	10	0.1	0.1	26.1
	MAST	0.068	0.115						

Table A.12: True Power using LRT to test $H_0 : \alpha_1 = 0, \beta_1 = 0$ with a significance level of 0.05

APPENDIX A: ADDITIONAL RESULTS FOR CHAPTER 3

B.1 Gene Set Simulation Details

1. Simulate set of independent ("original") genes

   (a) Simulate covariates and random effects (if present) to create cell population

   (b) Randomly sample (or set to zero to exclude) parameter values for additional covariates to include in the model

   - Random sampling creates variability in read counts
   - Intercepts fixed to ensure drop-out percentages and data scale comparable

   (c) Simulate Y_{ij} from the ZINB distribution

   (d) Repeat 1,000 times without RE and 300 times with RE

   - Cell population same in each scenario, genes differ due to differing parameters and randomness
   - Need to make sure there are enough unique data values to limit spurious correlation

2. Generate correlated gene sets of size 30

   (a) For each "original" gene, call it Y_{input}, add noise from NB distn using pre-specified, fixed parameters a_1, μ_{perm}, ϕ_{perm} to create 29 correlated genes Y_{out}:

$$Y_{out} = \text{round}(a_1 * Y_{input} + a_2 * NB(\mu_{perm}, \phi_{perm}))$$

   - Added noise has the same distribution for each scenario
   - Weight noise by a_2 to control the amount of correlation (larger a_2 means lower correlation) and change mean patterns across scenarios

- If gene is under the alternative, add additional noise $a_3 * NB(\mu_{perm}, \phi_{perm})$ to preserve signal (a_3 taken as 0.15 in "mixed" alternatives and 0.1 otherwise)

(b) Randomly set some non-zero counts to zero to keep the proportion of zeros the same in correlated and original gene

- Ensures that proportion of zeros alone does not drive significant results

3. Gene set testing procedure

   (a) Randomly choose correlated test set (including "original" gene) and reference set

   (b) Compute *gene-level* statistics using TWO-SIGMA

   (c) Use modified Wilcoxon rank-sum procedure adjusting for inter-gene correlation (IGC), estimated as in section 2.2 of the main text

4. Vary magnitude of IGC in 2(a) by drawing a_2 randomly or treating as various fixed values as in table B.1

5. Repeat 1-4 using 10 different random seeds and aggregate results to minimize the impact of random seed

Table B.1: Shows the six different settings used to simulate data for gene set simulations. "O.C." refers to the presence of other covariates besides treatment in the true model, which can serve to create complex gene-gene correlation structures.

a_2 low	a_2 high	O.C.
3	3	No
5	5	No
10	10	No
3	3	Yes
5	5	Yes
10	10	Yes

B.2 Additional Power and Type-I Error Results

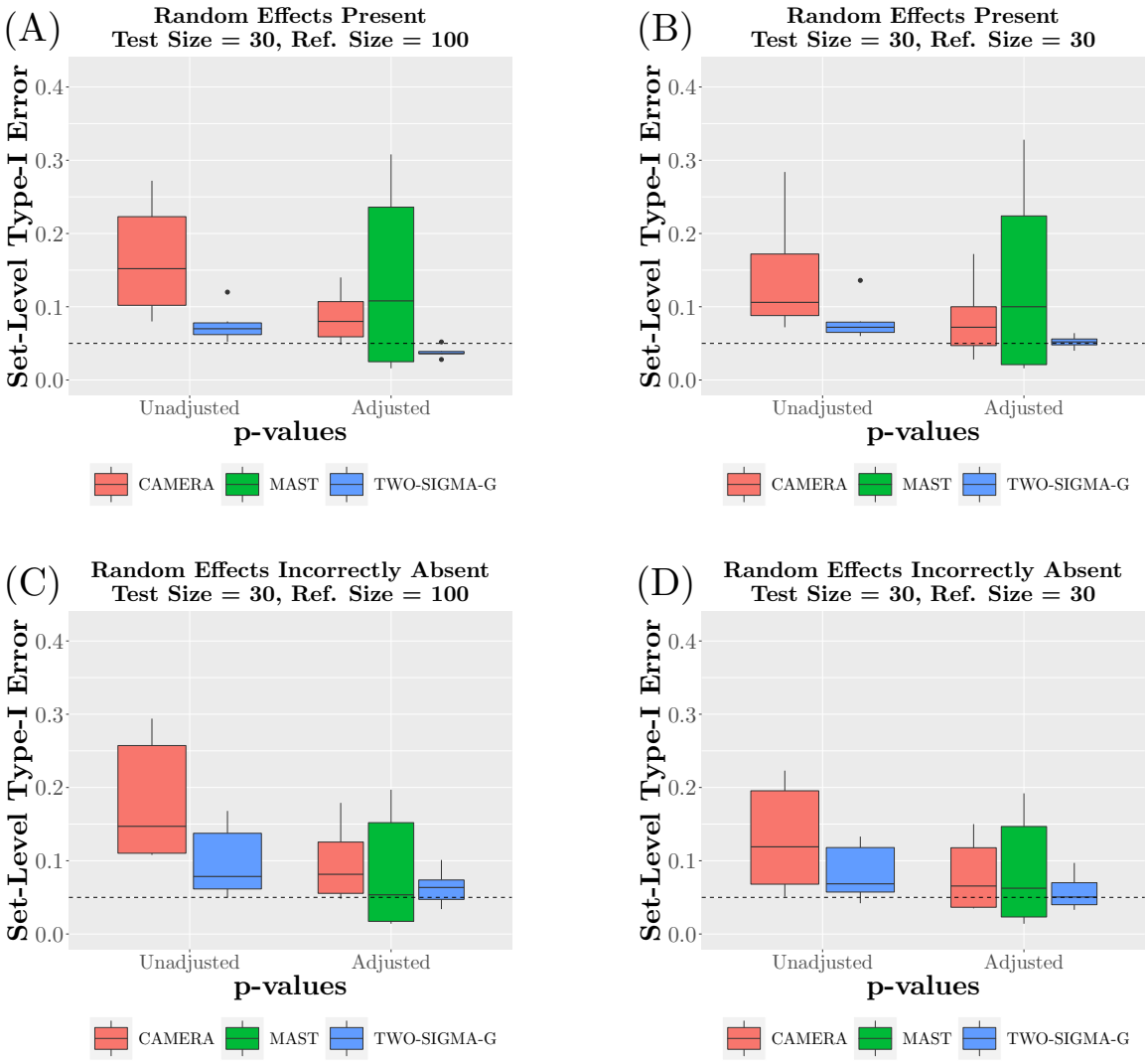

Figure B.1: Shows type-I error performance for CAMERA, MAST, and TWO-SIGMA-G when gene-level random effects are truly present and either incorrectly absent or correctly included in MAST and TWO-SIGMA-G gene-level models. Generally, there appears to be a limited need to incur the increased computational cost of fitting gene-level random effects if interested primarily in set-level inference. Note that CAMERA does not have the ability to fit random effects at the gene-level. Each boxplot aggregates 6 different settings which vary both the magnitude of the average inter-gene correlation in the test set and the nature of the correlation structure via the introduction of other individual-level covariates. Such settings are meant to represent the diversity seen in real data sets to paint an accurate picture of testing properties over a wide range of gene sets. Each of the 6 settings is further composed of 10 replicates which vary only random seed to mimic the impact of a different starting pool of cells from which genes were simulated. See section S1 for more details regarding the simulation procedure.

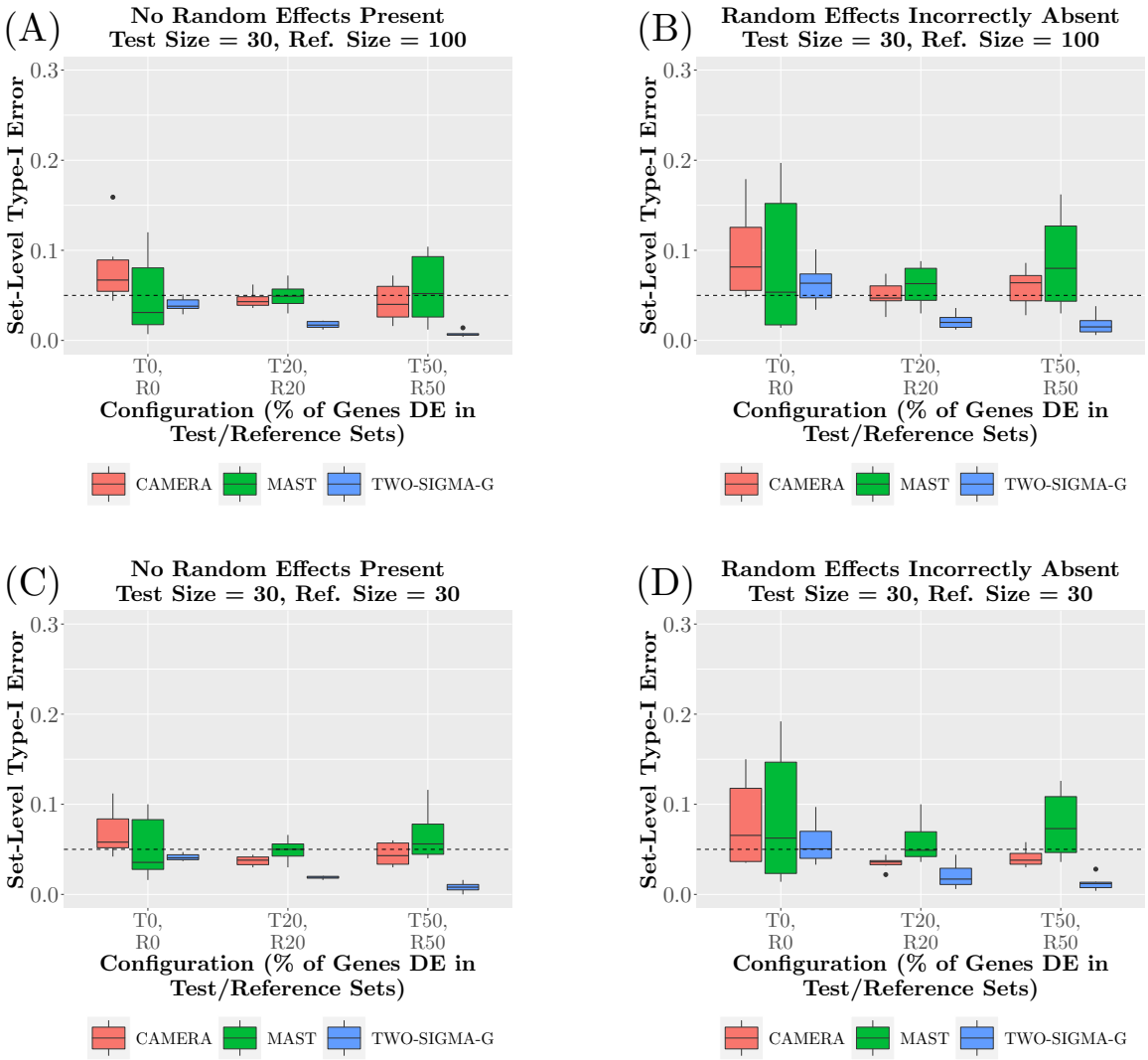

Figure B.2: Shows the type-I error of TWO-SIGMA-G, CAMERA, and MAST for various set-level null hypotheses when gene-level random effects are not present or are incorrectly absent in MAST and TWO-SIGMA-G gene-level models. Generally, TWO-SIGMA-G becomes more conservative, MAST becomes anti-conservative, and CAMERA's performance varies as the proportion of DE genes increases. Each boxplot aggregates 6 different settings which vary both the magnitude of the average inter-gene correlation in the test set and the nature of the correlation structure via the introduction of other individual-level covariates. Such settings are meant to represent the diversity seen in real data sets to paint an accurate picture of testing properties over a wide range of gene sets. Each of the 6 settings is further composed of 10 replicates which vary only random seed to mimic the impact of a different starting pool of cells from which genes were simulated. See section S1 for more details regarding the simulation procedure.

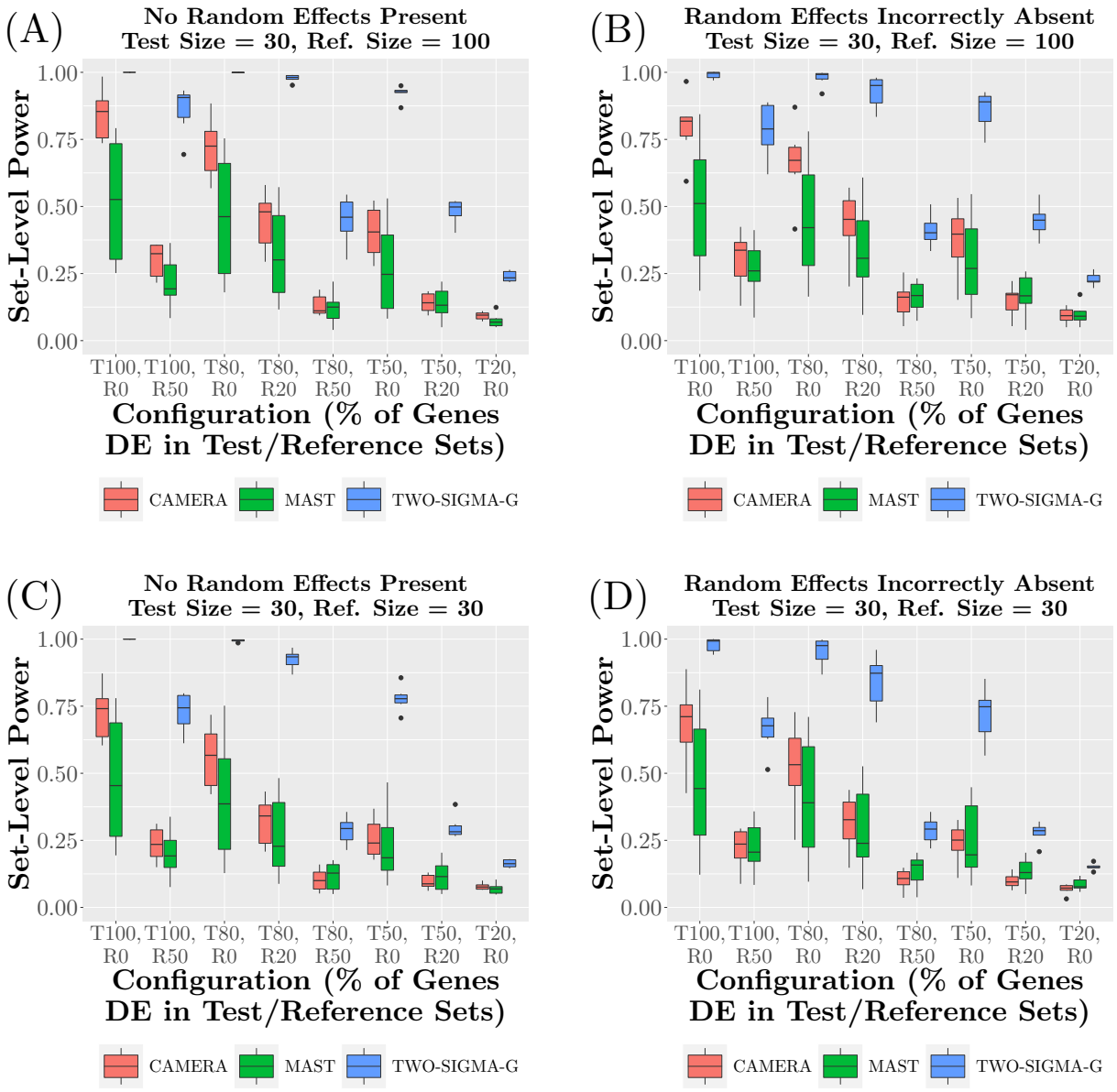

Figure B.3: Shows the power of TWO-SIGMA-G and CAMERA when random effect terms are excluded or incorrectly absent from the gene-level TWO-SIGMA model. Two scenarios are presented: using reference set sizes of 100 and 30. Within each scenario, the percentage of genes that are differentially expressed in the test and reference set is varied. For example, "T80,R50" corresponds to the configuration under the alternative hypothesis in which 80% of test set genes are DE and 50% of reference set genes are DE. Within the test set, the amount of DE is mixed: with 50% of genes having twice as large of an effect size as the other half. Each boxplot aggregates 6 different settings which vary both the magnitude of the average inter-gene correlation in the test set and the nature of the correlation structure via the introduction of other individual-level covariates. Such settings are meant to represent the diversity seen in real data sets to paint an accurate picture of testing properties over a wide range of gene sets. Each of the 6 settings is further composed of 10 replicates which vary only random seed to mimic the impact of a different starting pool of cells from which genes were simulated. See section S1 for more details regarding the simulation procedure.

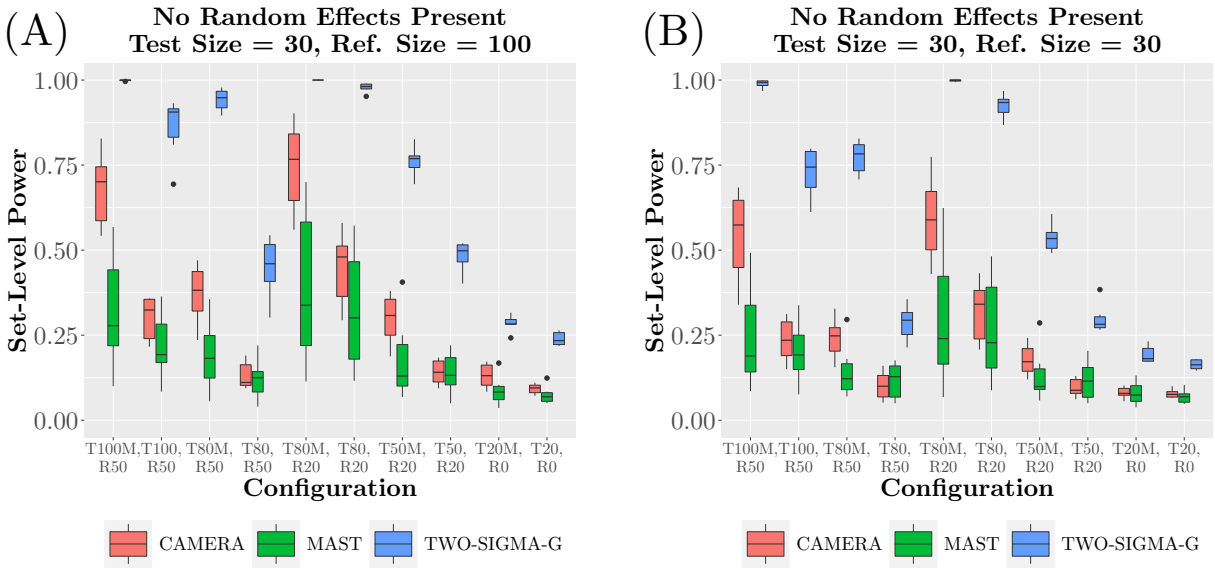

Figure B.4: Shows the power of TWO-SIGMA-G and CAMERA using different DE magnitudes for genes simulated with IGC. Four scenarios are presented: using reference set sizes of 100 and 30, both with and without random effects truly present at the gene-level. Within each scenario, the percentage of genes that are differentially expressed in the test and reference set is varied. For example, "T80,R50" corresponds to the configuration under the alternative hypothesis in which 80% of test set genes are DE and 50% of reference set genes are DE. Within the test set, the amount of DE is mixed: with 50% of genes having twice as large of an effect size as the other half. Each boxplot aggregates 6 different settings which vary both the magnitude of the average inter-gene correlation in the test set and the nature of the correlation structure via the introduction of other individual-level covariates. Such settings are meant to represent the diversity seen in real data sets to paint an accurate picture of testing properties over a wide range of gene sets. Each of the 6 settings is further composed of 10 replicates which vary only random seed to mimic the impact of a different starting pool of cells from which genes were simulated. See section S1 for more details regarding the simulation procedure.

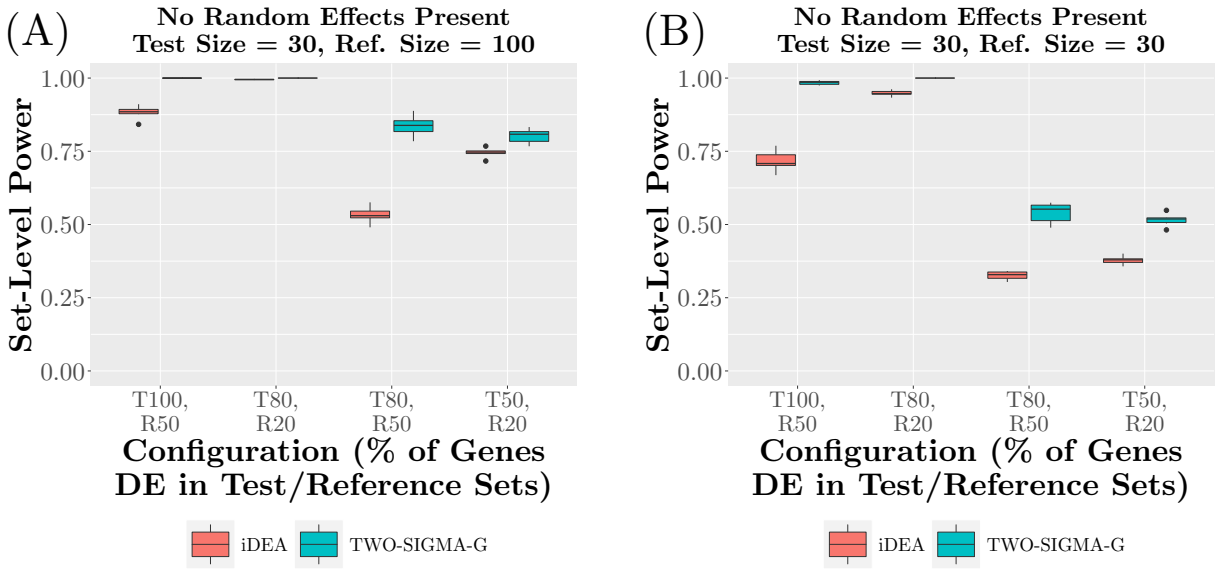

Figure B.5: Shows the power of TWO-SIGMA-G and iDEA using different DE magnitudes for genes simulated with IGC. Two scenarios are presented: using reference set sizes of 100 and 30. Within each scenario, the percentage of genes that are differentially expressed in the test and reference set is varied. For example, "T80,R50" corresponds to the configuration under the alternative hypothesis in which 80% of test set genes are DE and 50% of reference set genes are DE. Because iDEA performed poorly in scenarios involving "R0", they were excluded. Each boxplot aggregates 6 different settings which vary both the magnitude of the average inter-gene correlation in the test set and the nature of the correlation structure via the introduction of other individual-level covariates. Such settings are meant to represent the diversity seen in real data sets to paint an accurate picture of testing properties over a wide range of gene sets. Each of the 6 settings is further composed of 10 replicates which vary only random seed to mimic the impact of a different starting pool of cells from which genes were simulated. See section S1 for more details regarding the simulation procedure.

B.3 Additional Real Data Figures

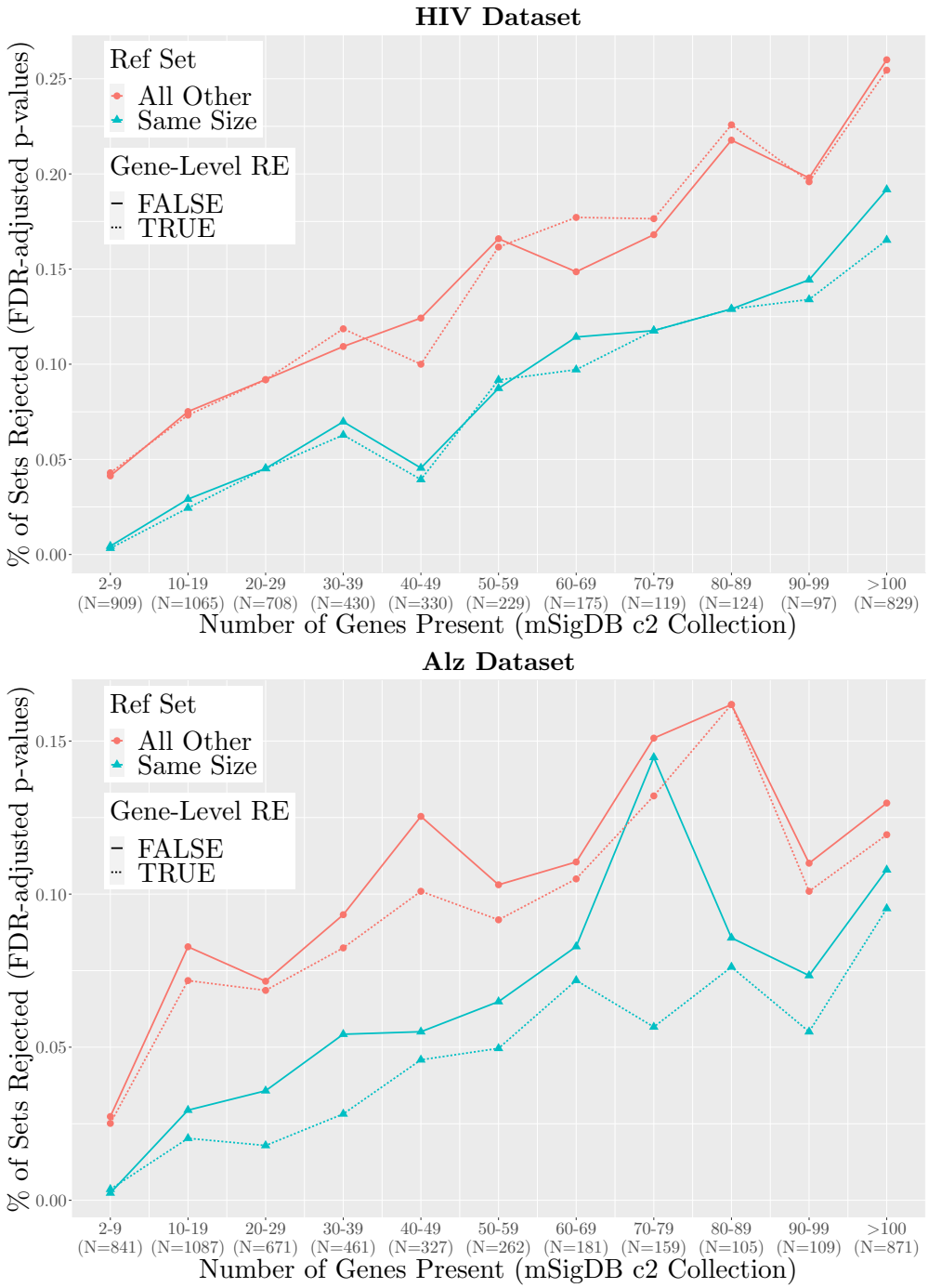

Figure B.6: Shows the percentage of sets rejected using Fisher's-method p-values adjusted to control FDR in four settings varying the choice of reference set between the complement set of genes ("All Other") or a random reference of the same size as the test set ("Same Size"), and with and without random effects present at the gene-level. The presence of gene-level random effects in the model does not greatly affect the percentage of sets rejected in either the HIV dataset (top) or the Alzheimer's dataset (bottom).

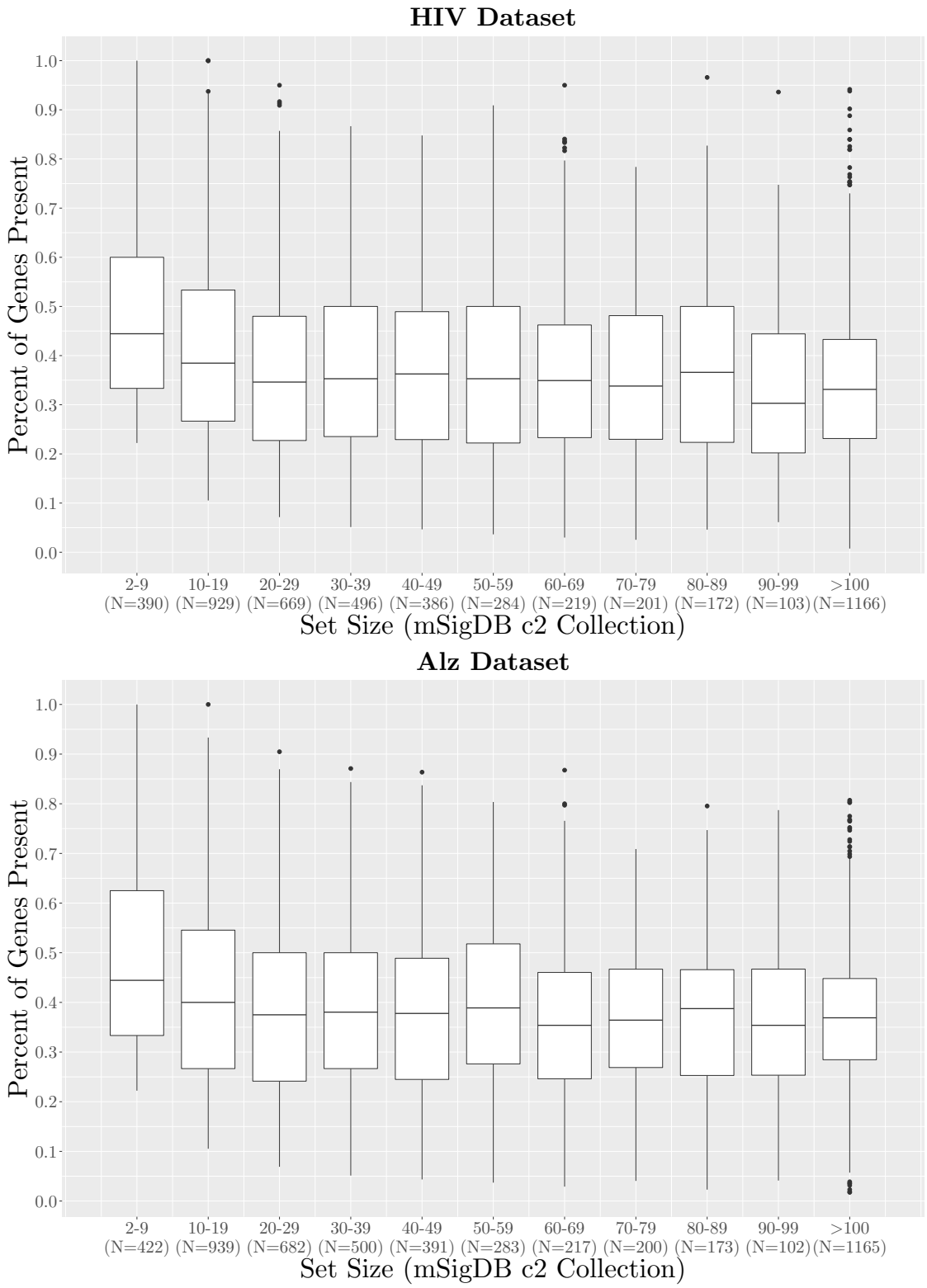

Figure B.7: Shows how the percentage of genes present varies by set size in the HIV dataset (top) and the Alzheimer's dataset(bottom).

B.4 Comparing Early Stage AD Patients to Control

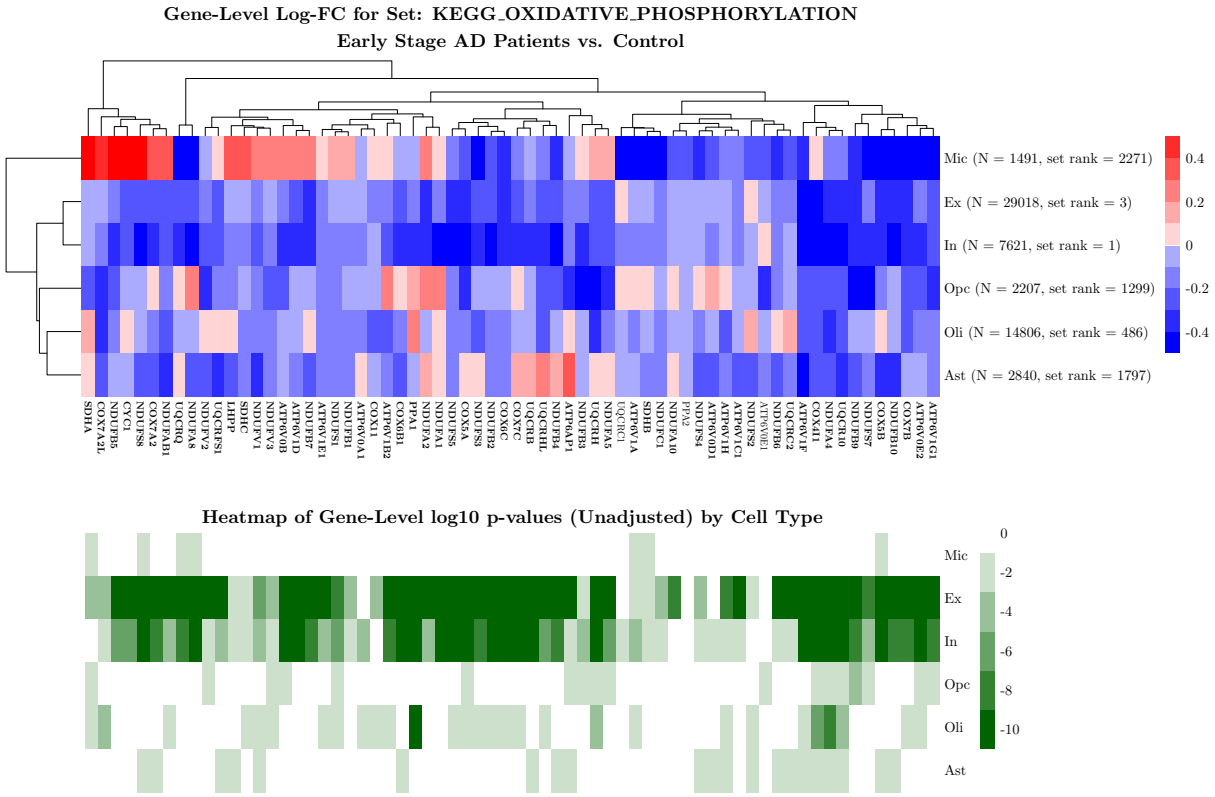

Figure B.8: Shows cell-type specific variation in gene-level significance for genes in the KEGG_OXIDATIVE_PHOSPHORYLATION pathway comparing early stage AD patients to controls. Gene names that are bolded are signifcant over all cell types after FDR-adjustment of the Fisher's method p-value.

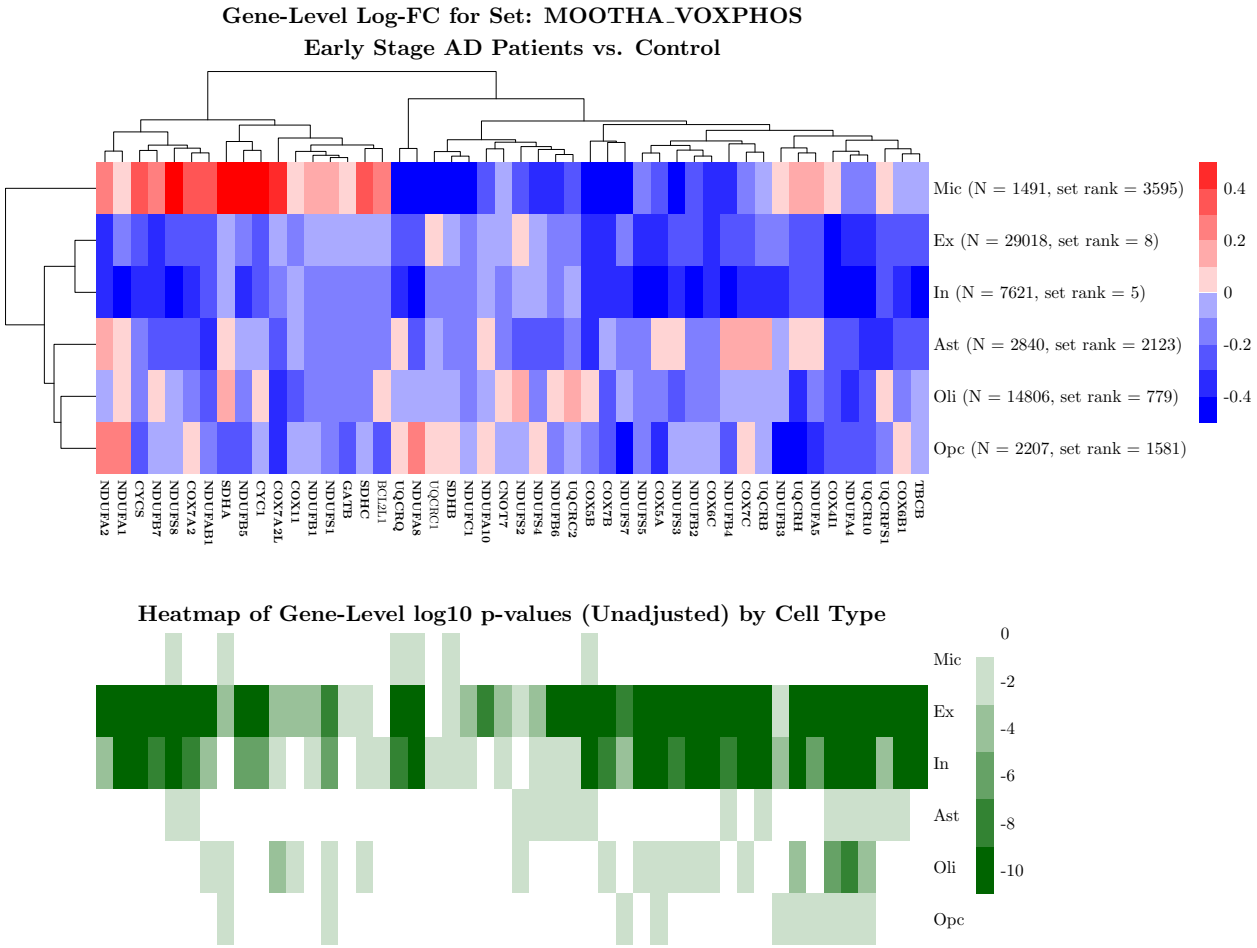

Figure B.9: Shows cell-type specific variation in gene-level significance for genes in the MOOTHA_VOXPHOS pathway comparing early stage AD patients to controls. Gene names that are bolded are signifcant over all cell types after FDR-adjustment of the Fisher's method p-value.

B.5 Comparing Late to Early Stage AD Patients

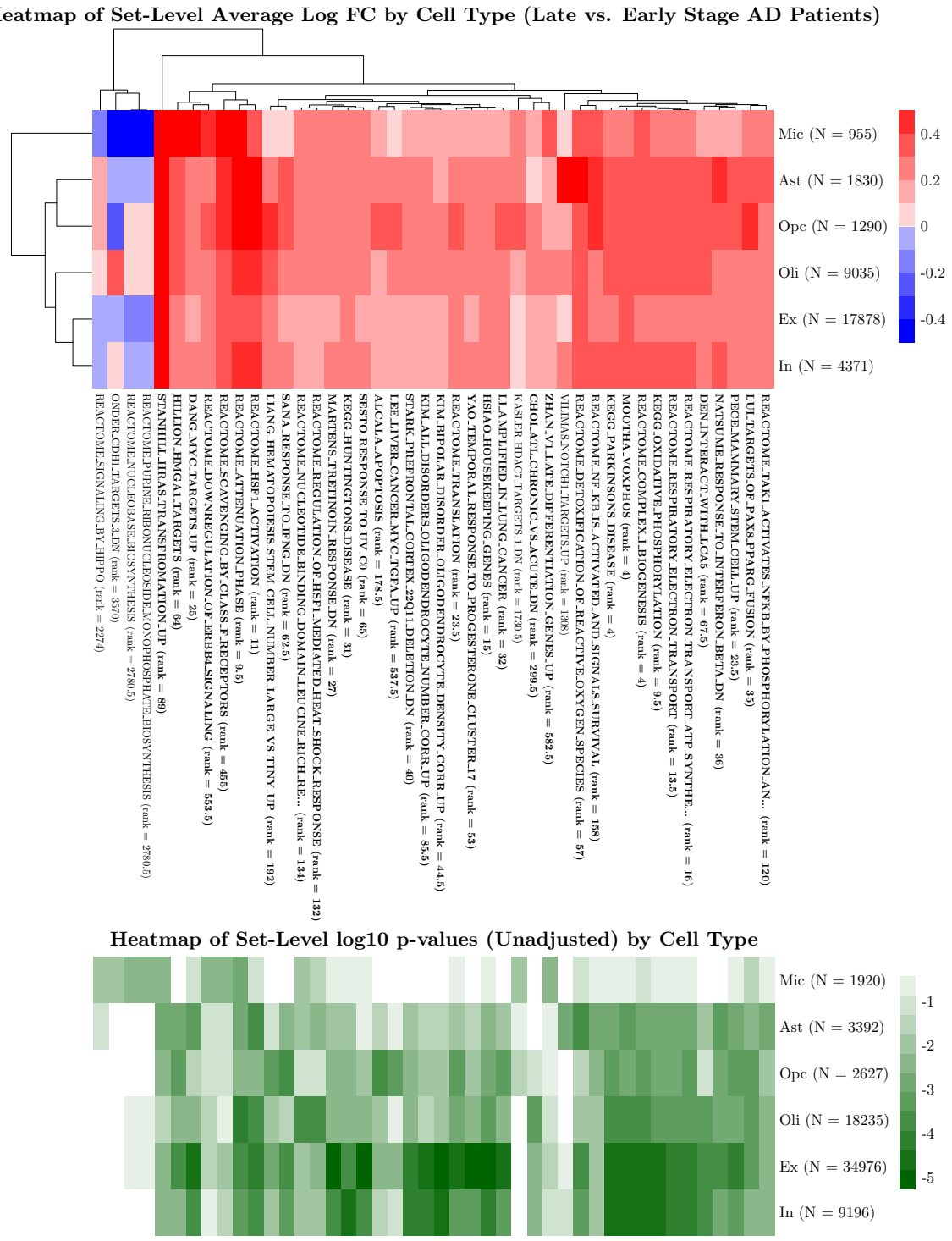

Figure B.10: Heatmap of the most significant gene sets (and their corresponding p-values) comparing late state AD patients to early stage AD patients by cell type. Sets plotted are among the top 10 in significance for at least once cell type. Sets in bold are significant over all cell types after FDR-adjustment of the Fisher's method p-value.

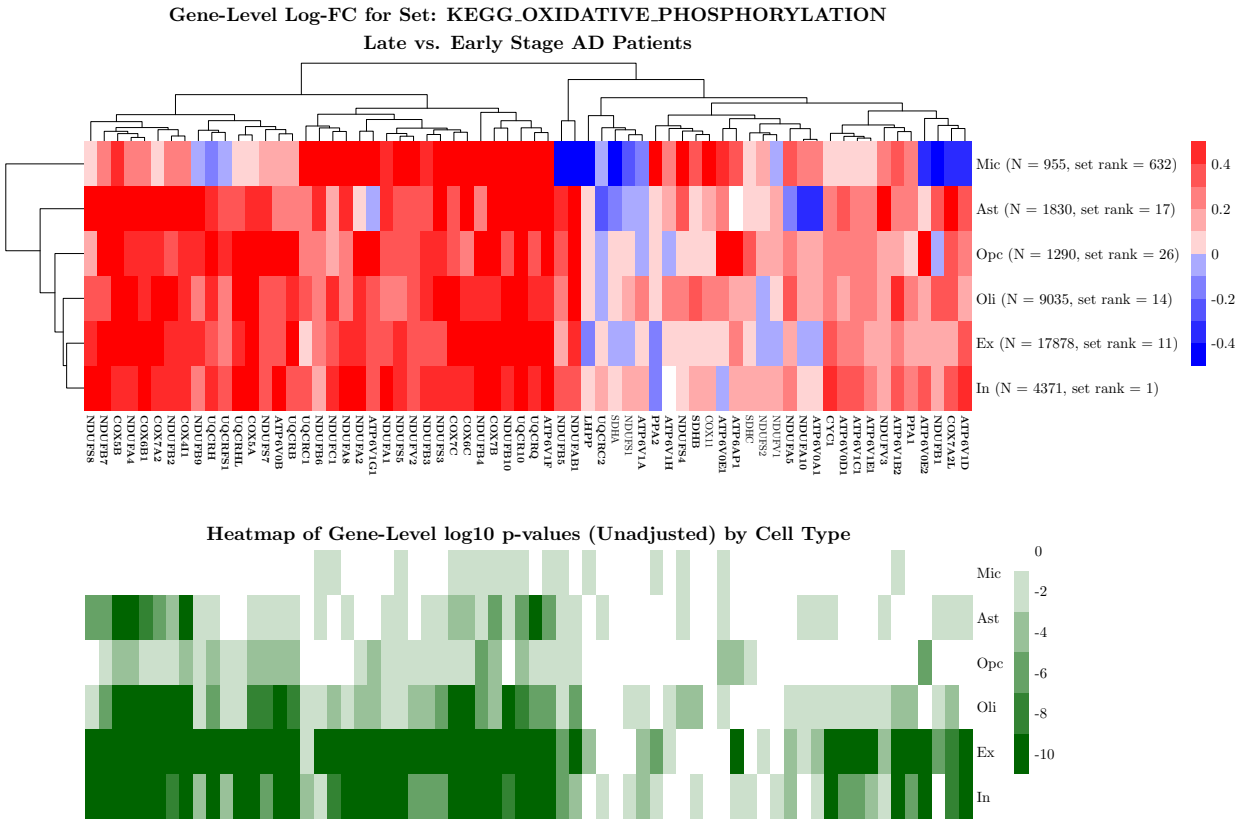

Figure B.11: Shows cell-type specific variation in gene-level significance for genes in the KEGG_OXIDATIVE_PHOSPHORYLATION pathway comparing late stage AD patients to early stage AD patients. Gene names that are bolded are signifcant over all cell types after FDR-adjustment of the Fisher's method p-value.

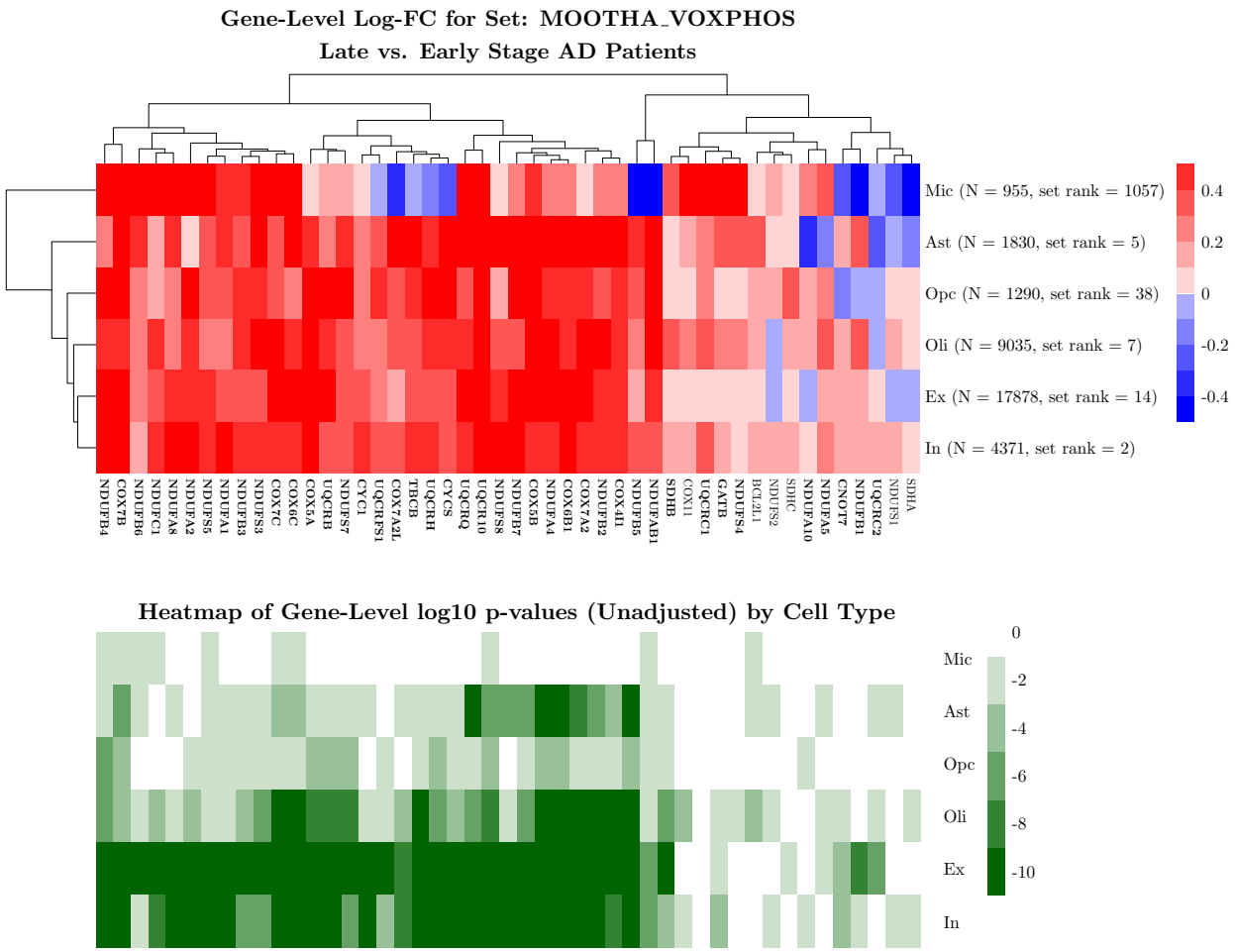

Figure B.12: Shows cell-type specific variation in gene-level significance for genes in the MOOTHA_VOXPHOS pathway comparing late stage AD patients to early stage AD patients. Gene names that are bolded are signifcant over all cell types after FDR-adjustment of the Fisher's method p-value.

B.6 Comparing Late Stage AD Patients to Control

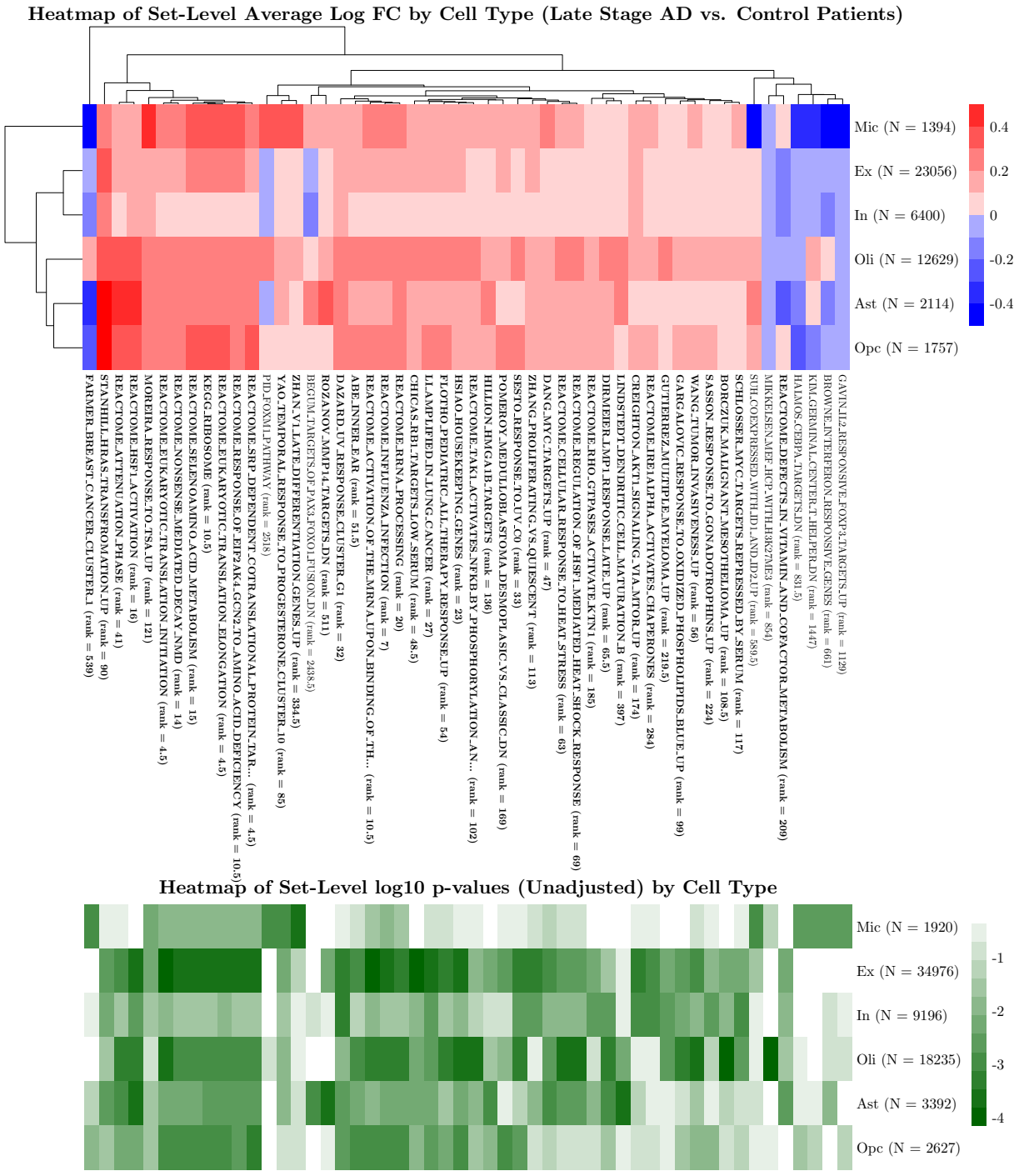

Figure B.13: Heatmap of the most significant gene sets (and their corresponding p-values) comparing late state AD patients to controls by cell type. Sets plotted are among the top 10 in significance for at least once cell type. Sets in bold are significant over all cell types after FDR-adjustment of the Fisher's method p-value.

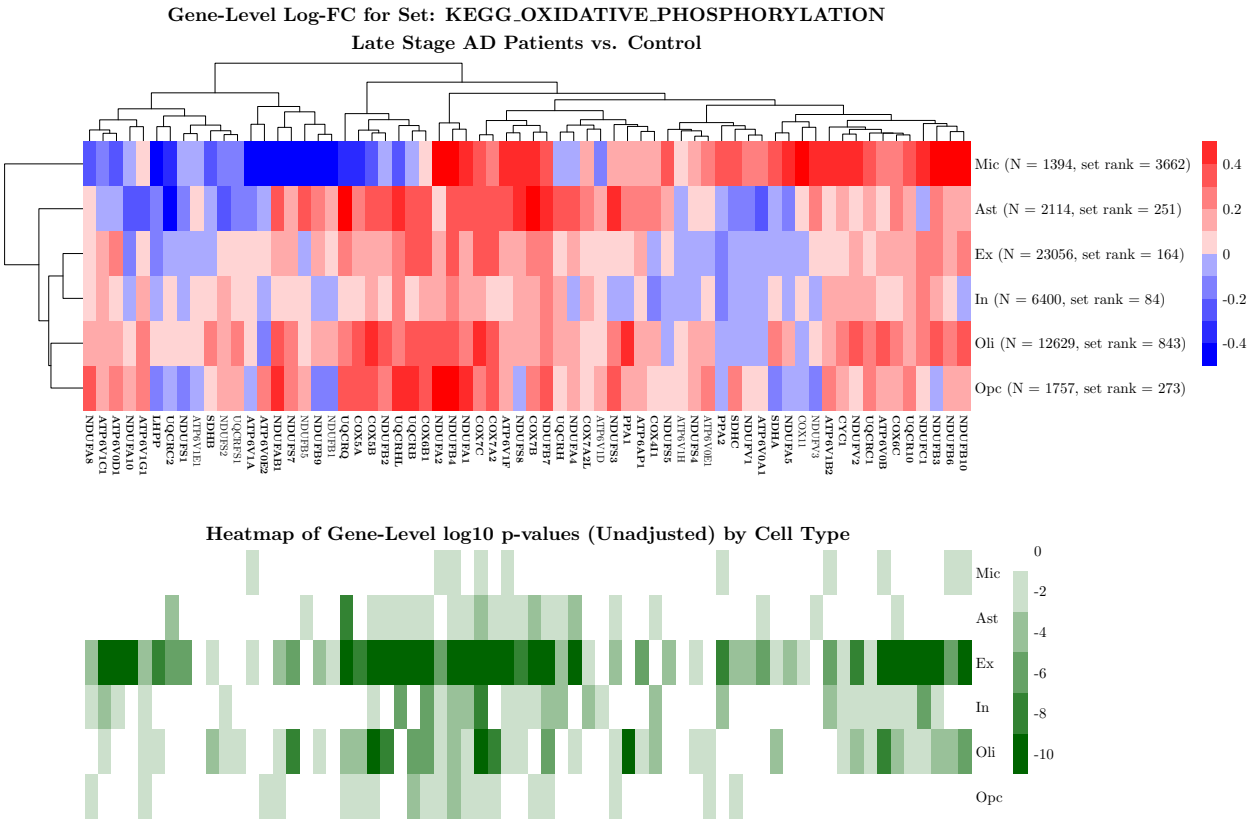

Figure B.14: Shows cell-type specific variation in gene-level significance for genes in the KEGG_OXIDATIVE_PHOSPHORYLATION pathway comparing late stage AD patients to controls. Gene names that are bolded are signifcant over all cell types after FDR-adjustment of the Fisher's method p-value.

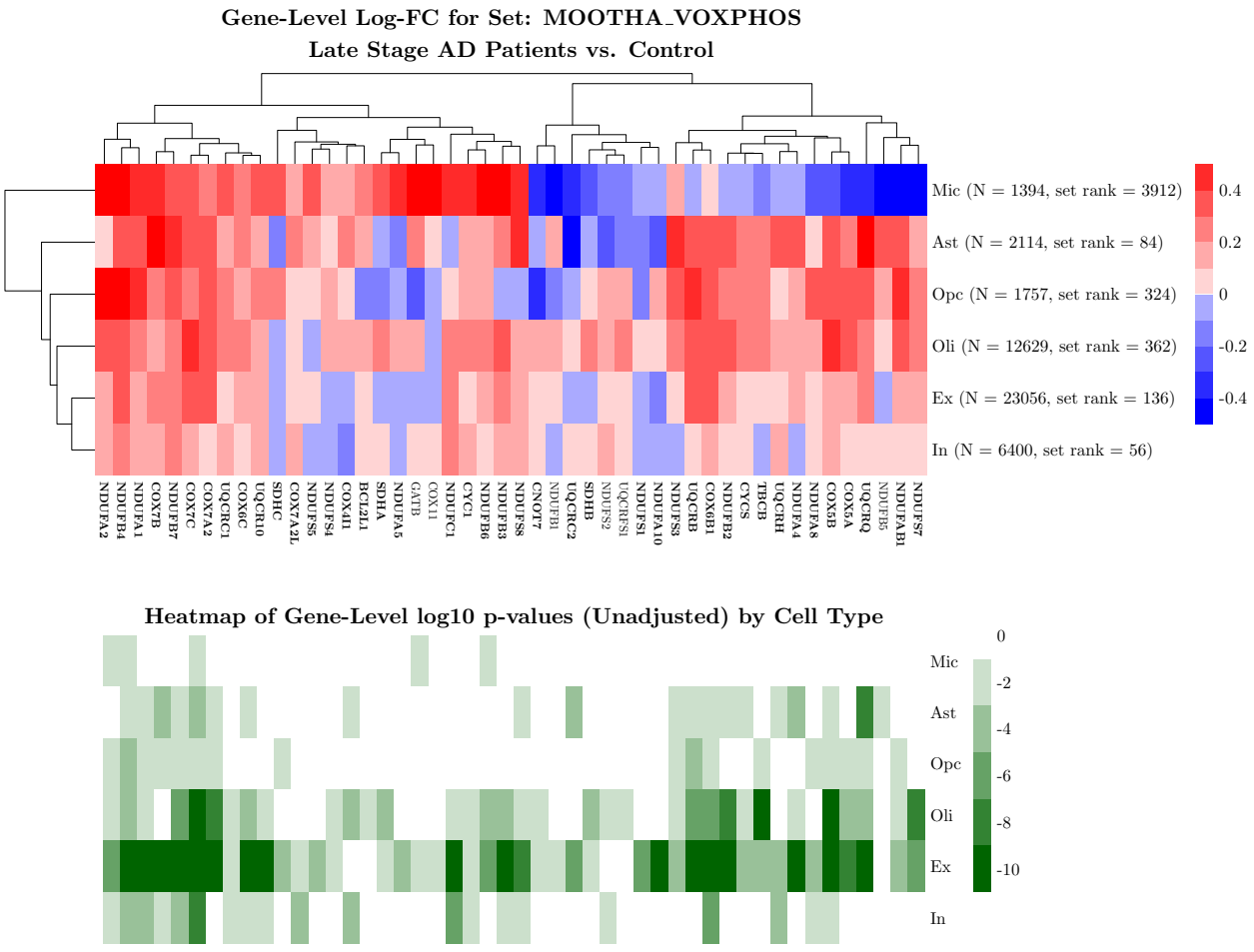

Figure B.15: Shows cell-type specific variation in gene-level significance for genes in the MOOTHA_VOXPHOS pathway comparing late stage AD patients to controls. Gene names that are bolded are signifcant over all cell types after FDR-adjustment of the Fisher's method p-value.

B.7 Comparing AD Patients (Early and Late Stage) to Control

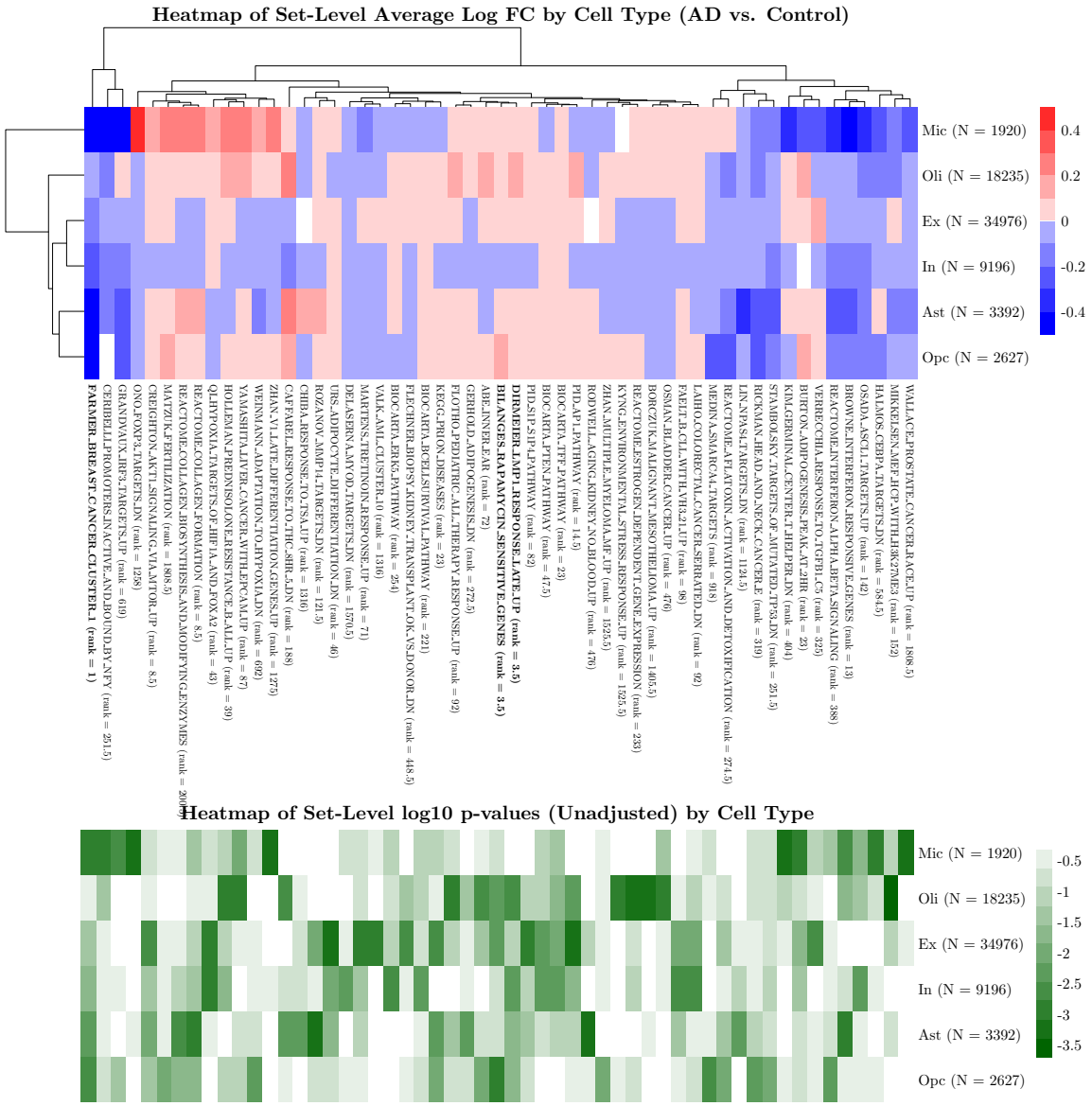

Figure B.16: Heatmap of the most significant gene sets (and their corresponding p-values) comparing AD patients (early and late stage) to controls by cell type. Sets plotted are among the top 10 in significance for at least once cell type. Sets in bold are significant over all cell types after FDR-adjustment of the Fisher's method p-value.

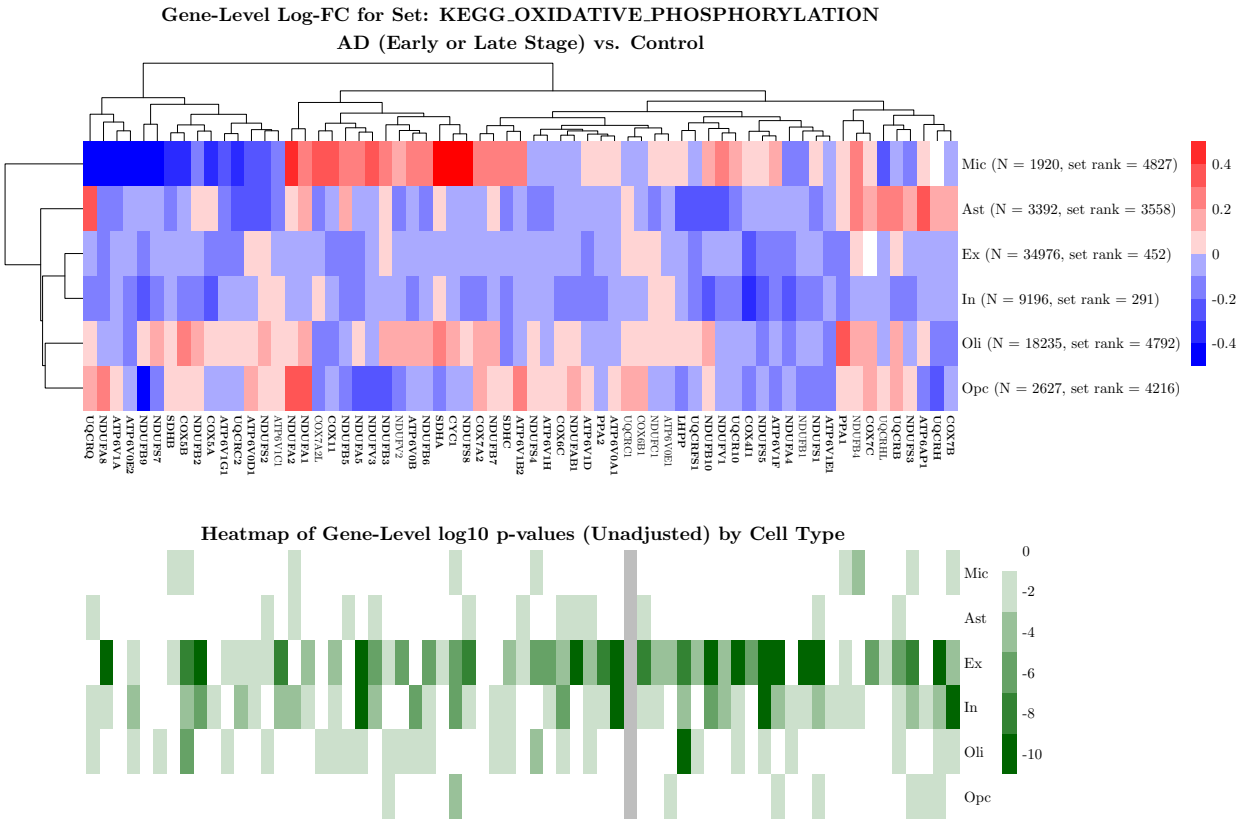

Figure B.17: Shows cell-type specific variation in gene-level significance for genes in the KEGG_OXIDATIVE_PHOSPHORYLATION pathway comparing AD patients (early and late stage) to controls. Gene names that are bolded are signifcant over all cell types after FDR-adjustment of the Fisher's method p-value.

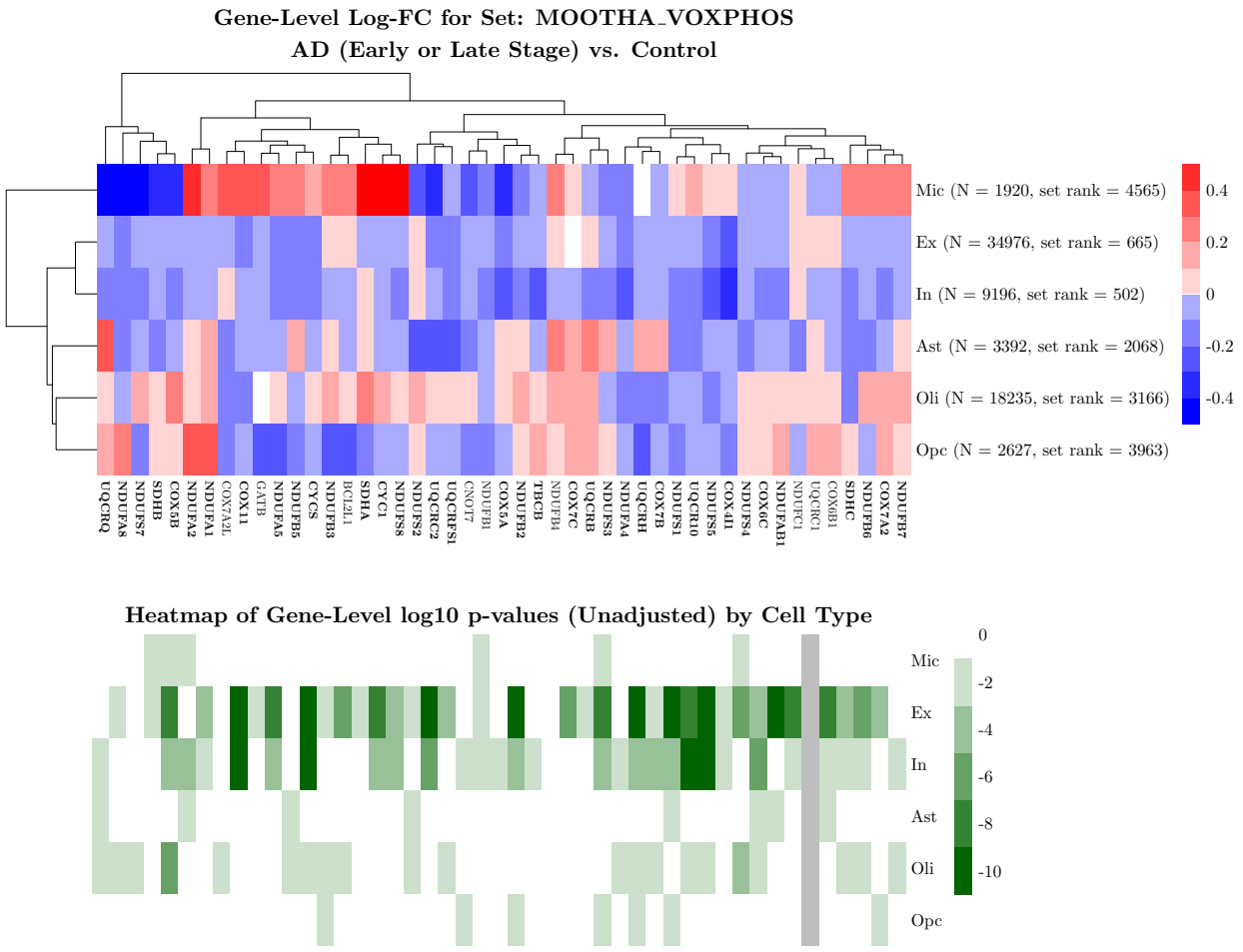

Figure B.18: Shows cell-type specific variation in gene-level significance for genes in the MOOTHA_VOXPHOS pathway comparing AD patients (early and late stage) to controls. Gene names that are bolded are signifcant over all cell types after FDR-adjustment of the Fisher's method p-value.

BIBLIOGRAPHY

Agresti, A. (2013). *Categorical Data Analysis, Third Edition*. John Wiley & Sons, Inc., Hoboken, NJ.

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z. K., Wouters, J., and Aerts, S. (2017). Scenic: single-cell regulatory network inference and clustering. *Nature Methods*, **14**(11), 1083–1086.

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

Ay, F., Bailey, T. L., and Noble, W. S. (2014). Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts. *Genome Research*, **24**(6), 999–1011.

Bacher, R. and Kendziorski, C. (2016). Design and computational analysis of single-cell rna-sequencing experiments. *Genome Biology*, **17**(1), 63.

Barry, W. T., Nobel, A. B., and Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**(9), 1943–1949.

Barry, W. T., Nobel, A. B., and Wright, F. A. (2008). A statistical framework for testing functional categories in microarray data. *Ann. Appl. Stat.*, **2**(1), 286–315.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**(421), 9–25.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., and Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, **9**(2), 378–400.

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

Chen, E. Z. and Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, **32**(17), 2611–2617.

Chen, G., Ning, B., and Shi, T. (2019). Single-cell rna-seq technologies and related computational data analysis. *Frontiers in Genetics*, **10**, 317.

Chen, W., Li, Y., Easton, J., Finkelstein, D., Wu, G., and Chen, X. (2018). Umi-count modeling and differential expression analysis for single-cell rna sequencing. *Genome Biology*, **19**(1), 70.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M., Gaffney, D. J., Elo, L. L., Zhang, X., and Mortazavi, A. (2016). A survey of best practices for rna-seq data analysis. *Genome Biology*, **17**(1), 13.

Damian, D. and Gorfine, M. (2004). Statistical concerns about the gsea procedure. *Nature Genetics*, **36**(7), 663–663.

Diggle, P. J. *et al.* (2002). *Analysis of Longitudinal Data*. Oxford University Press, Oxford, UK.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**(1), 107–129.

Fan, J., Salathia, N., Liu, R., Kaeser, G. E., Yung, Y. C., Herman, J. L., Kaper, F., Fan, J.-B., Zhang, K., Chun, J., and Kharchenko, P. V. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature Methods*, **13**(3), 241–244.

Fang, Z., Weng, C., Li, H., Tao, R., Mai, W., Liu, X., Lu, L., Lai, S., Duan, Q., Alvarez, C., Arvan, P., Wynshaw-Boris, A., Li, Y., Pei, Y., Jin, F., and Li, Y. (2019). Single-cell heterogeneity analysis and crispr screen identify key beta cell-specific disease genes. *Cell Reports*, **26**(11), 3132 – 3144.e7.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., Linsley, P. S., and Gottardo, R. (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome Biology*, **16**(1), 278.

Finn, E. H., Pegoraro, G., Brando, H. B., Valton, A.-L., Oomen, M. E., Dekker, J., Mirny, L., and Misteli, T. (2019). Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell*, **176**(6), 1502 – 1515.e10.

Fitzmaurice, G. M. *et al.* (2003). *Applied Longitudinal Analysis, Second Edition*. John Wiley & Sons, Inc., Hoboken, NJ.

Fudenberg, G. and Imakaev, M. (2017). Fish-ing for captured contacts: towards reconciling fish and 3c. *Nature Methods*, **14**(7), 673–678.

Gatti, D. M., Barry, W. T., Nobel, A. B., Rusyn, I., and Wright, F. A. (2010). Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics*, **11**(1), 574.

Gaynor, S. M., Sun, R., Lin, X., and Quackenbush, J. (2019). Identification of differentially expressed gene sets using the Generalized BerkJones statistic. *Bioinformatics*, **35**(22), 4568–4576.

Goeman, J. J. and Buhlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**(8), 980–987.

Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N., and Garry, D. J. (2018). Drimpute: imputing dropout events in single cell rna sequencing data. *BMC Bioinformatics*, **19**(1), 220.

Greene, W. H. (1994). Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. Working Papers 94-10, New York University, Leonard N. Stern School of Business, Department of Economics.

Haque, A., Engel, J., Teichmann, S. A., and Lönnberg, T. (2017). A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome Medicine*, **9**(1), 75.

Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2017a). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, **19**(4), 562–578.

Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2017b). Missing data and technical variability in single-cell rna-sequencing experiments. *bioRxiv*.

Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press, 2 edition.

Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., and Liu, J. S. (2012). HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, **28**(23), 3131–3133.

Hu, M., Deng, K., Qin, Z., Dixon, J., Selvaraj, S., Fang, J., Ren, B., and Liu, J. S. (2013). Bayesian inference of spatial organizations of chromosomes. *PLOS Computational Biology*, **9**(1), 1–14.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., and Zhang, N. R. (2018). Saver: gene expression recovery for single-cell rna sequencing. *Nature Methods*, **15**(7), 539–542.

Hwang, B., Lee, J. H., and Bang, D. (2018). Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, **50**(8), 96.

Kelsey, G., Stegle, O., and Reik, W. (2017). Single-cell epigenomics: Recording the past and predicting the future. *Science*, **358**(6359), 69–75.

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

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLOS Computational Biology*, **8**(2), 1–10.

Korthauer, K. D., Chu, L.-F., Newton, M. A., Li, Y., Thomson, J., Stewart, R., and Kendziorski, C. (2016). A statistical approach for identifying differential distributions in single-cell rna-seq experiments. *Genome Biology*, **17**(1), 222.

Krumm, A. and Duan, Z. (2019). Understanding the 3d genome: Emerging impacts on human disease. *Seminars in Cell and Developmental Biology*, **90**, 62 – 77. 3D Genome and Diseases.

Lajoie, B. R., Dekker, J., and Kaplan, N. (2015). The hitchhiker's guide to hi-c analysis: Practical guidelines. *Methods*, **72**, 65 – 75. (Epi)Genomics approaches and their applications.

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

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, **15**(2), R29.

Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., Kycia, I., Robson, P., and Stitzel, M. L. (2017). Single-cell transcriptomes identify human islet cell signatures and reveal cell-typespecific expression changes in type 2 diabetes. *Genome Research*, **27**(2), 208–222.

Li, W. V. and Li, J. J. (2018). An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature Communications*, **9**(1), 997.

Liberzon, A., Birger, C., Thorvaldsdttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Systems*, **1**(6), 417 – 425.

Lin, P., Troup, M., and Ho, J. W. K. (2017). Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome Biology*, **18**(1), 59.

Liu, Y. and Xie, J. (2020). Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, **115**(529), 393–402.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, **15**(12), 550.

Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular Systems Biology*, **15**(6), e8746.

Lukassen, S., Ten, F. W., Eils, R., and Conrad, C. (2019). Gene set inference from single-cell sequencing data using a hybrid of matrix factorization and variational autoencoders. *bioRxiv*.

Lun, A. (2018). Overcoming systematic errors caused by log-transformation of normalized single-cell rna sequencing data. *bioRxiv*.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., and McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**(5), 1202 – 1214.

Manczak, M., Park, B. S., Jung, Y., and Reddy, P. H. (2004). Differential expression of oxidative phosphorylation genes in patients with alzheimer's disease. *NeuroMolecular Medicine*, **5**(2), 147–162.

Mateo, L. J., Murphy, S. E., Hafner, A., Cinquini, I. S., Walker, C. A., and Boettiger, A. N. (2019). Visualizing dna folding and rna in embryos at single-cell resolution. *Nature*, **568**(7750), 49–54.

Mathur, R., Rotroff, D., Ma, J., Shojaie, A., and Motsinger-Reif, A. (2018). Gene set analysis methods: a systematic comparison. *BioData Mining*, **11**(1), 8.

Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., Menon, M., He, L., Abdurrob, F., Jiang, X., Martorell, A. J., Ransohoff, R. M., Hafler, B. P., Bennett, D. A., Kellis, M., and Tsai, L.-H. (2019). Single-cell transcriptomic analysis of alzheimer's disease. *Nature*, **570**(7761), 332–337.

Miao, Z., Deng, K., Wang, X., and Zhang, X. (2018). Desingle for detecting three types of differential expression in single-cell rna-seq data. *Bioinformatics*, **34**(18), 3223–3224.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, **5**, 1–19.

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

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

Nunomura, A., Perry, G., Aliev, G., Hirai, K., Takeda, A., Balraj, E. K., Jones, P. K., Ghanbari, H., Wataya, T., Shimohama, S., Chiba, S., Atwood, C. S., Petersen, R. B., and Smith, M. A. (2001). Oxidative Damage Is the Earliest Event in Alzheimer Disease. *Journal of Neuropathology & Experimental Neurology*, **60**(8), 759–767.

Oluwadare, O., Highsmith, M., and Cheng, J. (2019). An overview of methods for reconstructing 3-d chromosome and genome structures from hi-c data. *Biological Procedures Online*, **21**(1), 7.

Oron, A. P., Jiang, Z., and Gentleman, R. (2008). Gene set enrichment analysis using linear models and diagnostics. *Bioinformatics*, **24**(22), 2586–2591.

Pal, K., Forcato, M., and Ferrari, F. (2019). Hi-c analysis: from data generation to integration. *Biophysical Reviews*, **11**(1), 67–78.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, **4**(1), 12–35.

Preisser, J. S., Stamm, J. W., Long, D. L., and Kincade, M. E. (2012). Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. *Caries Research*, **46**(4), 413–423.

Ramani, V., Deng, X., Qiu, R., Lee, C., Disteche, C. M., Noble, W. S., Shendure, J., and Duan, Z. (2020). Sci-hi-c: A single-cell hi-c method for mapping 3d genome organization in large number of single cells. *Methods*, **170**, 61 – 68. Methods for Mapping Three-Dimensional Genome Architecture.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell rna-seq data. *Nature Communications*, **9**(1), 284.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, **11**(3), R25.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

Schurch, N. J., Schofield, P., Gierliski, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., and Barton, G. J. (2016). How many biological replicates are needed in an rna-seq experiment and which differential expression tool should you use? *RNA*, **22**(6), 839–851.

Sekula, M., Gaskins, J., and Datta, S. (2019). Detection of differentially expressed genes in discrete single-cell rna sequencing data using a hurdle model with correlated random effects. *Biometrics*, **75**(4), 1051–1062.

Sena, J. A., Galotto, G., Devitt, N. P., Connick, M. C., Jacobi, J. L., Umale, P. E., Vidali, L., and Bell, C. J. (2018). Unique molecular identifiers reveal a novel sequencing artefact with implications for rna-seq based gene expression analysis. *Scientific Reports*, **8**(1), 13121.

Seyednasrollah, F., Laiho, A., and Elo, L. L. (2013). Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*, **16**(1), 59–70.

Skaug, H. J. and Fournier, D. A. (2006). Automatic approximation of the marginal likelihood in non-gaussian hierarchical models. *Computational Statistics & Data Analysis*, **51**(2), 699 – 709.

Smyth, G. K. (2005). *limma: Linear Models for Microarray Data*, pages 397–420. Springer New York, New York, NY.

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics*, **14**(1), 91.

Soneson, C. and Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, **15**(4), 255–261.

Spielmann, M., Lupiáñez, D. G., and Mundlos, S. (2018). Structural variation in the 3d genome. *Nature Reviews Genetics*, **19**(7), 453–467.

Stark, R., Grzelak, M., and Hadfield, J. (2019). Rna sequencing: the teenage years. *Nature Reviews Genetics*, **20**(11), 631–656.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**(43), 15545–15550.

Svensson, V. (2020). Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*.

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., Kim, C. J., Kusanovic, J. P., and Romero, R. (2008). A novel signaling pathway impact analysis. *Bioinformatics*, **25**(1), 75–82.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences*, **102**(38), 13544–13549.

Todem, D., Kim, K., and Hsu, W.-W. (2016). Marginal mean models for zero-inflated count data. *Biometrics*, **72**(3), 986–994.

Townes, F. W., Hicks, S. C., Aryee, M. J., and Irizarry, R. A. (2019). Feature selection and dimension reduction for single cell rna-seq based on a multinomial model. *bioRxiv*.

Tracy, S., Yuan, G.-C., and Dries, R. (2019). Rescue: imputing dropout events in single-cell rna-sequencing data. *BMC Bioinformatics*, **20**(1), 388.

Van Buren, E., Hu, M., Weng, C., Jin, F., Li, Y., Wu, D., and Li, Y. (2020). Two-sigma: a novel two-component single cell model-based association method for single-cell rna-seq data. *bioRxiv*.

Van den Berge, K., Soneson, C., Love, M. I., Robinson, M. D., and Clement, L. (2017). zinger: unlocking rna-seq tools for zero-inflation and single cell applications. *bioRxiv*.

Van den Berge, K., Perraudeau, F., Soneson, C., Love, M. I., Risso, D., Vert, J.-P., Robinson, M. D., Dudoit, S., and Clement, L. (2018). Observation weights unlock bulk rna-seq tools for zero inflation and single-cell applications. *Genome Biology*, **19**(1), 24.

Varoquaux, N., Ay, F., Noble, W. S., and Vert, J.-P. (2014). A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*, **30**(12), i26–i33.

Wang, T., Li, B., Nelson, C. E., and Nabavi, S. (2019). Comparative analysis of differential gene expression analysis tools for single-cell rna sequencing data. *BMC Bioinformatics*, **20**(1), 40.

Wills, Q. F., Livak, K. J., Tipping, A. J., Enver, T., Goldson, A. J., Sexton, D. W., and Holmes, C. (2013). Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature Biotechnology*, **31**(8), 748–752.

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

Wu, D., Lim, E., Vaillant, F., Asselin-Labat, M.-L., Visvader, J. E., and Smyth, G. K. (2010). ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, **26**(17), 2176–2182.

Xu, Z., Zhang, G., Jin, F., Chen, M., Furey, T. S., Sullivan, P. F., Qin, Z., Hu, M., and Li, Y. (2015). A hidden Markov random field-based Bayesian method for the detection of long-range chromosomal interactions in Hi-C data. *Bioinformatics*, **32**(5), 650–656.

Yaffe, E. and Tanay, A. (2011). Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, **43**(11), 1059–1065.

Yuan, G.-C., Cai, L., Elowitz, M., Enver, T., Fan, G., Guo, G., Irizarry, R., Kharchenko, P., Kim, J., Orkin, S., Quackenbush, J., Saadatpour, A., Schroeder, T., Shivdasani, R., and Tirosh, I. (2017). Challenges and emerging directions in single-cell analysis. *Genome Biology*, **18**(1), 84.

Zhang, D. and Lin, X. (2008). *Variance Component Testing in Generalized Linear Mixed Models for Longitudinal/Clustered Data and other Related Topics*, pages 19–36. Springer New York, New York, NY.

Zhou, J., Ma, J., Chen, Y., Cheng, C., Bao, B., Peng, J., Sejnowski, T. J., Dixon, J. R., and Ecker, J. R. (2019). Robust single-cell hi-c clustering by convolution- and random-walk–based imputation. *Proceedings of the National Academy of Sciences*, **116**(28), 14011–14018.