TWO-SIGMA-G: A New Competitive Gene Set Testing Framework for scRNA-seq Data Accounting for Inter-Gene and Cell-Cell Correlation

Eric Van Buren [1], Ming Hu [2], Liang Cheng[3,4,5] John Wrobel[3], Kirk Wilhelmsen [6], Lishan Su[3,4,7], Yun Li [8,9,10*], Di Wu [8,11*]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health

[2]Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation

[3] Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill

[4] Department of Microbiology and Immunology, The University of North Carolina at Chapel Hill

[5] Frontier Science Center for Immunology and Metabolism, Medical Research Institute, Wuhan University

[6]Departments of Genetics and Neurology, Renaissance Computing Institute, University of North Carolina at Chapel Hill

[7] Departments of Pharmacology, Microbiology & Immunology University of Maryland School of Medicine

[8]Department of Biostatistics, The University of North Carolina at Chapel Hill

[9]Department of Genetics, The University of North Carolina at Chapel Hill

[10]Department of Computer Science, The University of North Carolina at Chapel Hill

[11]Division of Oral and Craniofacial Health Sciences, Adams School of Dentistry, The University of North Carolina at Chapel Hill

* To whom correspondence should be addressed: Email: did@email.unc.edu;
yun_li@med.unc.edu

January 24, 2021

Funding: This work was supported by the National Institute of Health [R01GM105785 and U54HD079124 to YL, R01HL129132 to YL and EVB, UM1HG011585 to M.H., R03DE028983 to DW], the National Cancer Institute [R35CA197449 to EVB], and the University of North Carolina Computational Medicine Program [to DW and LS].

Abstract

We propose TWO-SIGMA-G, a competitive gene set test designed for scRNA-seq data. TWO-SIGMA-G uses the mixed-effects regression modelling approach of our previously published TWO-SIGMA to test for differential expression at the gene-level. This regression-based approach can analyze complex designs while accommodating zero-inflated and overdispersed counts and within-sample cell-cell correlation. TWO-SIGMA-G uses a novel approach to adjust for inter-gene-correlation (IGC) at the set-level, which can inflate type-I error when ignored. Simulations demonstrate that TWO-SIGMA-G preserves type-I error and increases power in the presence of IGC compared to other methods designed for bulk and single-cell RNA-seq data. Application to two real datasets of HIV infection in mice and Alzheimer's disease progression in humans reveal biologically meaningful results. TWO-SIGMA-G is available at https://github.com/edvanburen/twosigma.

1 Background

In the past decade and a half, gene set tests utilizing pre-constructed gene sets [1] have been used to contextualize gene-level differential expression analyses and identify both important pathways and real biological mechanisms [12, 16, 23, 11] by connecting expression datasets in microarray data and bulk RNA-seq data [25, 33]. Gene set tests are used to determine whether predefined sets of genes exhibit differential expression (DE). These tests improve statistical power and reduce spurious associations as compared to gene-level DE testing [5, 9]. This further increases reproducibility across experiments, which is often lower than desired due to biological and technical variability in the data [5, 9]. Therefore, set-level tests constitute an essential step in performing differential expression (DE) based analyses for bulk RNA-seq data.

It is now common to distinguish between the two primary types of gene set tests, "competitive" and "self-contained," based on the null hypotheses of interest [10]. Self-contained tests compare a candidate gene set to a fixed standard (usually the case of no DE) and do not incorporate information from genes not in the set [32]. Competitive tests compare the evidence of differential expression of a gene set to the evidence in a reference set of genes [14, 28, 5, 22]. Competitive tests use a battery of gene sets to rank which sets are the most important for a given phenotype. In contrast, self-contained tests are commonly used to compare the similarity of gene expression patterns between different data sets. Because of interpretability and method availability, competitive tests are far more common in the literature today [33, 10].

The first step of gene set tests is to produce gene-level test statistics for DE. Then, these statistics are aggregated into a set-level summary statistic and corresponding p-value. Competitive tests usually permute genes or use parametric assumptions to construct a distribution of these summary statistics under the null hypothesis, often assuming that genes are independent of each other [33, 32, 10, 7]. Previous studies show that methods which assume independence often suffer from inflated type-I error [33, 10, 7]. This is because genes within a given gene set tend to have a positive inter-gene correlation (IGC). Ignoring this IGC underestimates the variance of set-level summary statistics and can dramatically inflate type-I error through inducing a correlation in the marginal gene-level statistics [2, 8, 33, 5]. Therefore, it is essential that any competitive gene set test adequately accounts for IGC to provide statistically rigorous set-level p-values.

Several gene set testing methods have been developed for bulk RNA-seq and microarray data, including GSEA [25] and related extension `sigPathway`, CAMERA [33], and PAGE [14]. GSEA has proven incredibly popular, as seen by its over 19,000 citations. However, one weakness of GSEA is that the null hypothesis being tested is not straightforward to precisely define because it is a hybrid of the self-contained and competitive null hypothesis. Larger gene sets can often be more significant, even if additional genes represent noise, and other gene sets not being tested can influence results in counterintuitive ways [4, 28]. `sigPathway` and PAGE sometimes suffer from inflated type-I error [27, 33]. CAMERA does not suffer from any of these disadvantages, and is described in more detail in the Methods section. GSEA, `sigPathway`, CAMERA, and PAGE all rely on assumptions chosen to represent the features of bulk RNA-seq data. Such assumptions may be unreasonable in scRNA-seq data, which often exhibits zero-inflated and overdispersed counts and the possible within-sample correlation between cells from the same sample [30]. Misspecified gene-level models can lead to misleading set-level inference, taking the form of inflated type-I error or reduced power. Thus, there is a need for methodological advancements to tailor gene set testing frameworks to scRNA-seq data, and a need to evaluate the ability of methods designed for bulk data to provide statistically valid results when applied to scRNA-seq data.

We are aware of two existing methods explicitly created for competitive gene set testing in scRNA-seq data: iDEA [18] and an extension of MAST [7]. iDEA jointly conducts gene-level DE testing using zingeR [31] and uses a Bayesian approach to produce set-level p-values. iDEA does not explicitly adjust for IGC, however, and may not detect the scenario in which the same proportion of genes are significant in the test and reference set but the magnitude of the association differs. MAST fits a log-normal hurdle model at the gene-level and uses a Z-test with a computationally-intensive bootstrapping procedure that was not studied in great detail to produce set-level p-values. We discuss iDEA and MAST in more detail in the Methods section. BAGSE was proposed as an improvement to GSEA designed to quantify the level of enrichment while preserving both type-I error and the power of GSEA [13]. As in iDEA, BAGSE utilizes gene-level estimates that can come from methods designed for DE analysis in scRNA-seq data. However, we do not classify BAGSE as a competitive gene set testing option for scRNA-seq because it tests the hybrid null hypothesis of GSEA and not the competitive null hypothesis mentioned above.

This paper develops TWO-SIGMA-G, a set-level framework for DE testing in scRNA-seq data with competitive null hypothesis. Our approach utilizes the flexible mixed-effects zero-inflated negative binomial regression model of TWO-SIGMA [30] to produce gene-level statistics. Using TWO-SIGMA, the type-I error is preserved in the presence of cell-cell correlation, and, as discussed below, many choices are available for gene-level statistics. Using a regression-based framework means that complex experiments with additional sample-level and cell-level covariates can be analyzed. IGC is estimated using an innovative residual-based approach and explicitly adjusted for at the set-level. We demonstrate TWO-SIGMA-G outperforms existing competitive gene set tests methods using extensive simulation scenarios under the null and alternative hypotheses. Application of TWO-SIGMA-G to an HIV-related humanized mouse scRNA-seq dataset and an Alzheimer's Disease human brain scRNA-seq dataset reveal exciting biological findings.

2 Results

2.1 Estimation of Inter-Gene Correlation

Before specifying our new gene set testing method, we first propose a novel strategy to estimate IGC between pairs of genes from their respective gene-level DE regression models. Cell-level covariates such as the cellular detection rate (CDR), which measures the percentage of genes expressed in a cell, have been previously demonstrated to be highly influential to observed expression levels [7]. Subject-specific covariates, such as disease status or race, can further create an additional correlation structure in the raw data. Therefore, using the raw data to estimate IGC can overestimate the correlation that remains between gene-level statistics, which come from regression models that directly adjust for these other covariates. Thus, the use of residuals to estimate IGC can better represent the remaining correlation of the gene-level statistics under the null.

We 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 test set genes. Given these residual matrices, we can compute the pairwise $(s \times s)$ correlation matrix $\boldsymbol{C}_i$, which contains s choose two unique non-diagonal elements. These elements give the pairwise correlations between the residuals of two different genes in the test set. We average these values to produce one average pairwise correlation $\hat{\rho}_i$ per individual. Finally, we estimate the overall correlation with the average of these values such that $\hat{\rho} = \sum_{i=1}^{n} \hat{\rho}_i / n$.

Therefore, our IGC procedure builds off of the advantages of a residual-based approach in removing the correlation from sample-level and cell-level covariates. We further use individual-level calculations to help mitigate the impacts of the large individual heterogeneity often seen in scRNA-seq datasets. In simulations, we found that this IGC estimate preserves type-I error in a conservative manner while still producing improved power in a variety of realistic scenarios. The estimate of the IGC is virtually free computationally in that the model is not refit via permutation or bootstrapping.

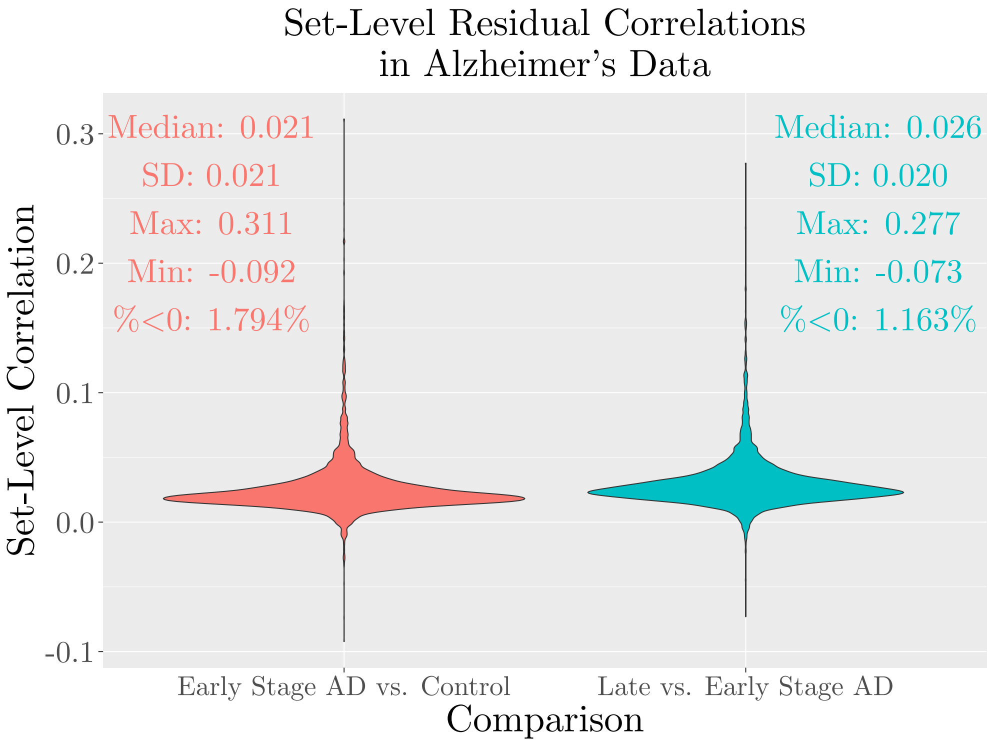

Figure 1: Shows the set-level IGC estimates from TWO-SIGMA-G's residual-based approach for the two Alzheimer's dataset comparisons (see the real data analysis section). Most sets demonstrate a substantially positive correlation after regressing out sample-level and cell-level covariates. Sets plotted are taken from the c2 collection of the Molecular Signatures Database, and negative estimated correlations are set to zero to compute the TWO-SIGMA-G p-value in a conservative manner.

Figure 1 shows the average set-level IGC estimates from TWO-SIGMA-G for each of the two comparisons in the Alzheimer's dataset described more in the real data analysis section. A non-negligible correlation exists in the residual space for both comparisons, with half of the sets having a correlation larger than 0.02. For each comparison, over 98% of the estimated pairwise correlations are positive. Ignoring this remaining correlation would therefore make inflated type-I error a possibility.

2.2 TWO-SIGMA-G for Set-Level Testing

We extend our TWO-SIGMA method [30] to competitive gene set testing via TWO-SIGMA-Geneset (TWO-SIGMA-G). The Methods section contains full details regarding TWO-SIGMA. Briefly, TWO-SIGMA uses a zero-inflated negative binomial regression model to test for DE at the gene level in scRNA-seq data. It is flexible and can be customized in several different ways. First, the zero-inflation component can be removed from the model entirely (as was done in the real data analysis section), leaving a standard negative binomial regression model. Second, the model can additionally include random effect terms to account for cell-cell correlation within the same sample and limit type-I error inflation in gene-level DE inference. Finally, many gene-level statistics measuring the evidence of DE can be used for set-level testing. For example, the likelihood ratio or z statistic corresponding to a treatment effect are commonly used gene-level statistics. We discuss uses for other, more complex, gene-level statistics based on custom contrasts of regression parameters in the real data analysis section. TWO-SIGMA-G employs the Wilcoxon rank-sum test to compare the statistics of genes in the test set to the statistics of genes in the reference set, and therefore uses the sum of the ranks in the test set as the set-level summary statistic. In using the ranks, TWO-SIGMA-G provides robustness against the influence of very large gene-level statistics.

Traditionally, the Wilcoxon rank-sum test assumes that observations within a group are independent. However, as mentioned, IGC is expected given the construction of gene sets as harmonious biological pathways, and can inflate type-I error if ignored [33]. To create a gene set testing method designed for single-cell data, we therefore utilize a modified version of the rank-sum test. This modification allows for correlated gene-level statistics in the test set [2], similar to the approach of CAMERA for bulk RNA-seq [33]. We assume a pairwise correlation ρ between gene-level statistics in the test set of size m_1 and no correlation in the reference set of size m_2. With these assumptions, variance of the two-group Wilcoxon rank-sum statistic is:

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

A positive ρ increases the variance as compared to a value of zero. Therefore, ignoring a positive ρ leads to an underestimated variance and inflated type-I error as a result. As discussed in the previous section, we estimate ρ using a residual-based approach. Using this modified variance, and the known mean of rank-sum statistics under the null, set-level p-values are computed analytically using a standard normal approximation [33]. 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.

In addition to producing set-level significance, TWO-SIGMA-G also identifies the directionality of sets as up or down-regulated. Whether or not a zero-inflation component is included in gene-level models, directionality is produced by averaging gene-level log fold-change estimates in the test set to produce a set-level effect size and taking the sign of the result. These effect sizes are demonstrated further in the real data analysis section.

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 account for zero-inflation, overdispersion, and within-subject random effect terms to account for within-subject cell-cell 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 demonstrated further in the real data analysis section. Third, estimating IGC using residuals after regressing out sample-level and cell-level covariates provides estimates of IGC that more closely reflect the remaining correlation of the gene-level statistics. Simulation studies show that TWO-SIGMA-G preserves type-I error and improves power over other methods. Applications to two real datasets demonstrate the TWO-SIGMA-G's ability to produce meaningful, cell-type-specific findings that can elucidate differentiation in pathway expression profiles.

2.3 TWO-SIGMA-G preserves type-I error in the presence of inter-gene correlation

Figure 2 shows the set-level type-I error performance of TWO-SIGMA-G, CAMERA, and MAST across various simulation scenarios (see Supplementary Figures S1 and S2 for results using smaller significance thresholds). Panel (A) shows that all three methods correctly hold the type-I error rate when genes are simulated independently and no gene-level within-sample random effects exist. In contrast, panels (B), (C), and (D) show that type-I error is consistently inflated when IGC is present and ignored. After *p*-value adjustment using the estimated average IGC, both TWO-SIGMA-G and MAST tend to preserve type-I error at the 5% level in the presence of IGC. In contrast, CAMERA suffers from inflation of type-I error after IGC adjustment. Differences between the three methods are likely 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 [29, 17]. 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. The procedure used in TWO-SIGMA-G to estimate and adjust for IGC is well-calibrated and produces valid set-level inference.

Panels (C) and (D) of figure 2 show that the type-I error from TWO-SIGMA-G is preserved or approximately preserved when gene-level random effect terms are truly present and either correctly included ("present") or incorrectly excluded ("incorrectly absent") from the fitted gene-level model. 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. For both methods, however, this type-I inflation is much lower in magnitude than can exist at the gene-level [30]. This highlights

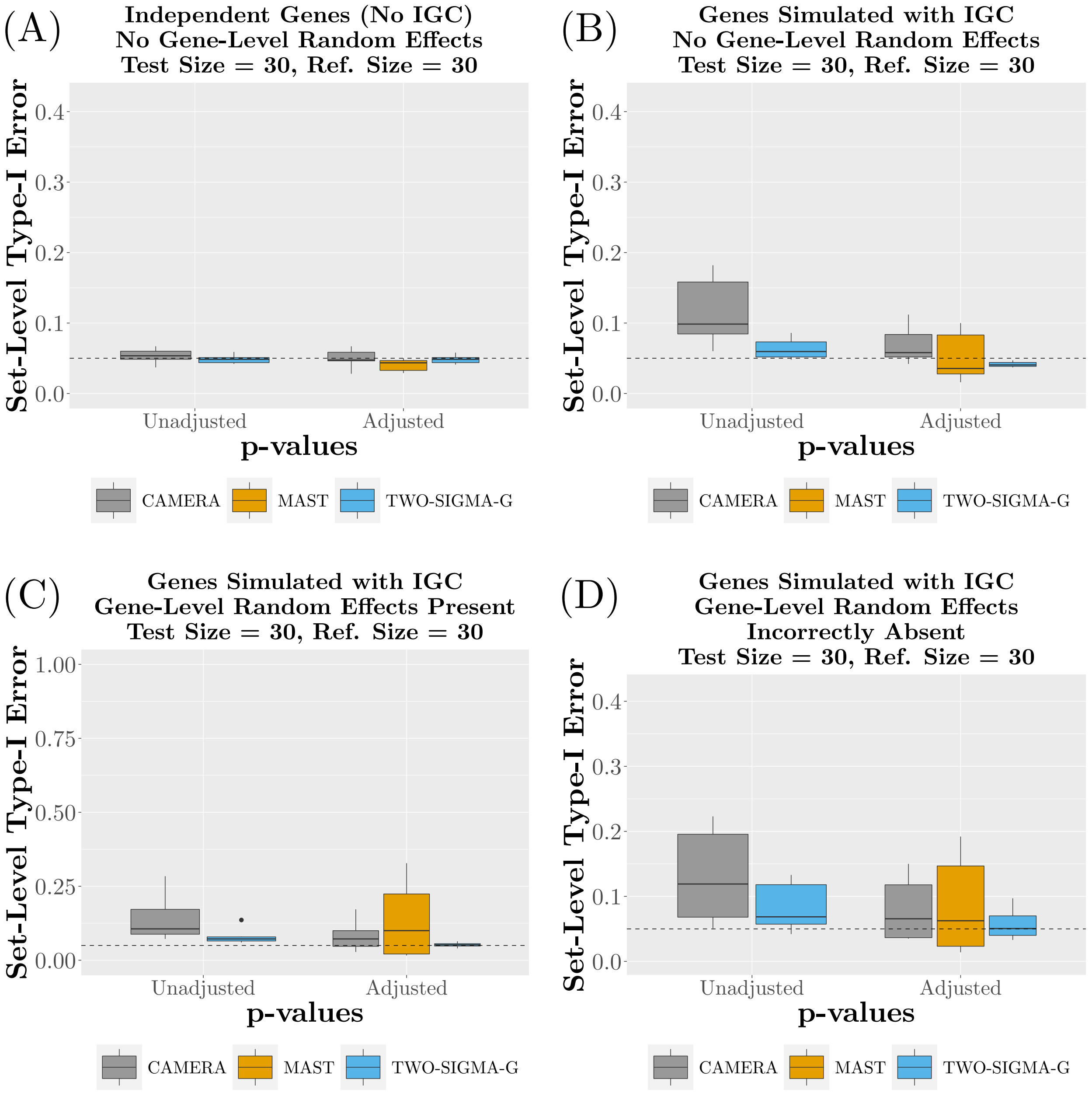

Figure 2: Shows type-I error performance for CAMERA, MAST, and TWO-SIGMA-G using a reference set size of 30 genes. Each panel varies the existence of IGC between genes in the test set and the presence of gene-level random effect terms in the gene-level model (CAMERA never includes gene-level random effect terms). Within each panel, both unadjusted and adjusted set-level p-values are plotted (unadjusted p-values are unavailable for MAST). Each boxplot aggregates six different settings which vary both the magnitude of the average inter-gene correlation (where applicable) in the test set and the nature of the correlation structure via the introduction of other individual-level covariates. Such settings are intended 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 six 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 the Methods section for more details regarding the simulation procedure.

an advantage of competitive gene set 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 section further shows a large agreement in set-level inference from TWO-SIGMA-G regardless of random effect inclusion in the gene-level model.

We additionally found that the null distributions of all three methods are nearly identical with a larger reference set size (Supplementary Figure S3). However, in the interest of being conservative, we will evaluate performance using results in which the test and reference sets are of equal size.

We also evaluated type-I error at various set-level null hypotheses, in which an equal but non-zero percentage of genes in the test and reference sets are DE (with the same gene-level effect size, Supplementary Figure S4). For example, the scenario in which 20% of genes are DE in both the test and reference sets is one set-level competitive null hypothesis. Generally, TWO-SIGMA-G becomes more conservative, MAST becomes anti-conservative, and CAMERA's performance varies as the proportion of DE genes increases. MAST's type-I error tends to become inflated once the background percentage of DE genes increases, particularly when gene-level random effects are mistakenly excluded from the gene-level model.

As discussed in the methods section, we had difficulty obtaining reliable p-values for iDEA and PAGE using our main simulation structure. We found that TWO-SIGMA provided improved type-I error control as compared to these methods using a modified simulation framework (Supplementary Figure S5).

2.4 TWO-SIGMA-G improves power over alternative approaches

Figure 3 shows the power of CAMERA, MAST, and TWO-SIGMA-G on simulated data, and demonstrates that TWO-SIGMA-G is consistently the most powerful method. Different configurations are presented, involving 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 that 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. Results suggest that 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 in all four panels of figure 3. We found that using a reference set size of 100 tends to improve power for all methods and particularly for TWO-SIGMA-G (Supplementary Figure S6). This power increase does not seem to be a consequence of an increase in type-I error (Supplementary Figure S3). This provides some evidence in favor of using a larger

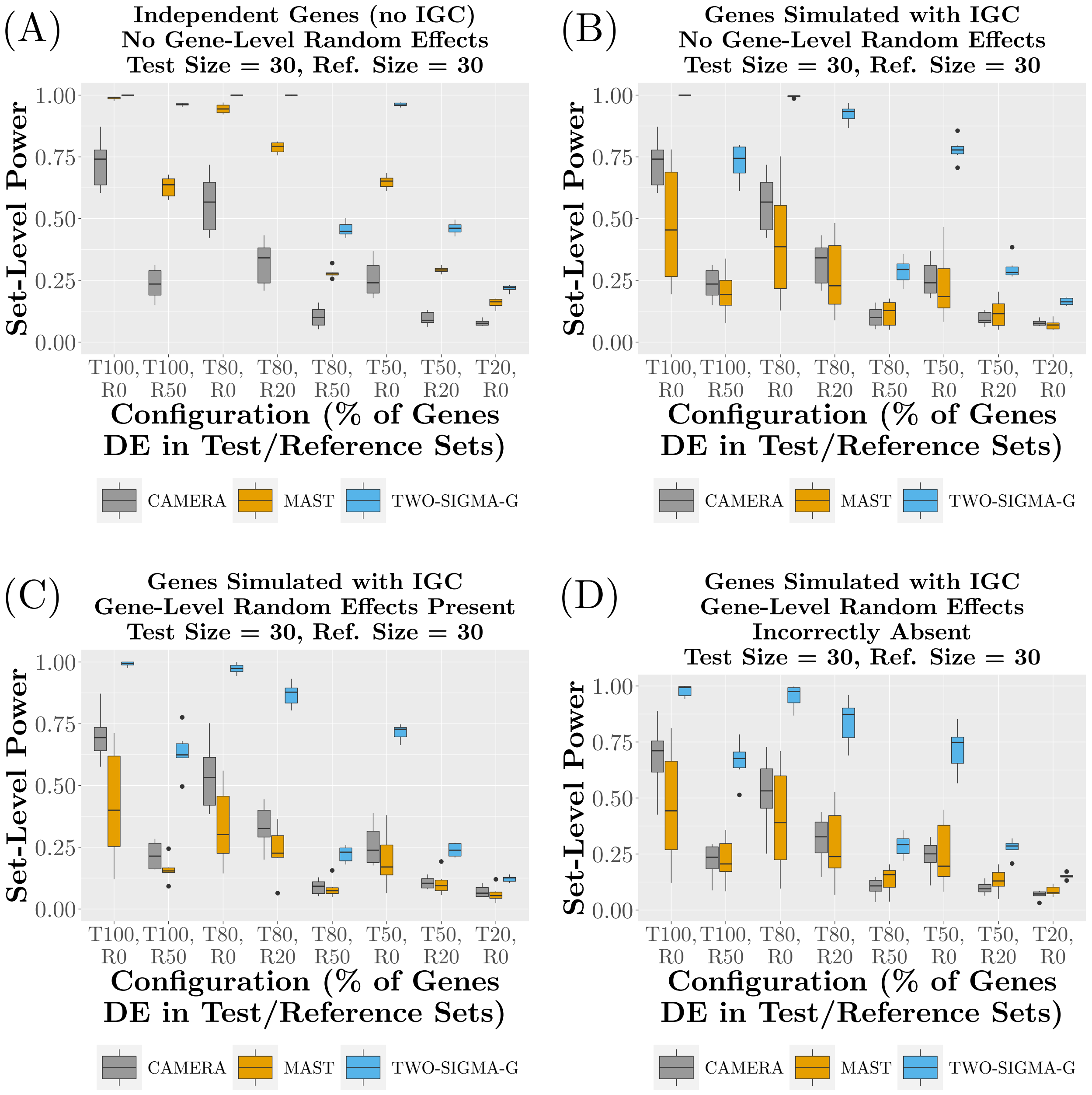

Figure 3: Shows the set-level power of CAMERA, MAST, and TWO-SIGMA-G using a reference set size of 30 genes. Each panel varies the existence of IGC between genes in the test set and the presence of gene-level random effect terms in the gene-level model (CAMERA never includes gene-level random effect terms). Scenarios along the x-axis of each panel vary the percentage of differentially expressed genes (with the same effect size) in the test and reference sets. 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 six different settings that 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 six 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 the Methods section for more details regarding the simulation procedure.

reference set in lieu of a balanced reference set.

Panels (A) and (B) of figure 3 summarize results from genes that do not contain gene-level random effect terms. In contrast, panels (C) and (D) of figure 3 show power when gene-level random effect terms are truly non-zero and either correctly included ("present") or incorrectly excluded ("incorrectly absent") from the fitted 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, the increased computational cost from gene-level random effect terms may not be necessary for valid and powerful inference. The set-level 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 if gene-level inference is of interest[30].

When the magnitude of gene-level DE is varied, such that half of genes have twice the effect size of the other half, we found that set-level power is improved (Supplementary Figure S7). The relative positions of each configuration remain as in figure 3, suggesting that power results in figure 3 apply to alternative DE breakdowns. For example, whether or not genes in the test have varying DE magnitudes, the "T80, R20" scenarios have improved power over the "T100,R50" scenarios. The relative rankings of the three compared methods also remains when the magnitude of gene-level DE is varied.

As above, we also used a modified simulation framework to compare to iDEA and PAGE. Results are very similar to our main simulations: TWO-SIGMA-G improves power over both methods, and differing configurations with the same difference in DE percentage between the test and reference sets tend to have very similar power profiles (Supplementary Figure S8).

2.5 Analysis of HIV data reveals biologically expected findings

First, we analyze a dataset of 11,630 single-cells collected from of 4 humanized donor mice, two of which were infected with HIV and two were given a mock treatment [3]. A total of 3,549 genes and 4,772 gene sets were analyzed. Table 1 shows the number of differentially expressed genes and sets in all cell types comparing HIV to a mock treatment. Having more DE genes does not always correspond to more DE gene sets. For example, erythroid cells have the second largest number of DE genes, but rank eighth in terms of the number of DE gene sets. This result is expected using TWO-SIGMA-G because, as a competitive test, it focuses on the relative signal of gene sets as compared to a background reference set of genes. The lack of a clear relationship between the number of DE genes and gene sets was also reflected when analyzing the overlap in significance between genes and gene sets among the four most prevalent cell types as seen in panels (B) and (C) of figure 4.

Figure 4 shows more detailed cell-type-specific results comparing HIV to the mock treatment. Panel

	Genes		Sets	
Cell Type	Up	Down	Up	Down
NK (N = 4249)	489	531	123	131
Erythroid (N = 2085)	413	523	56	94
ILC (N = 1421)	235	198	130	91
B (N = 1205)	168	273	95	133
mDC (N = 1088)	350	346	149	67
pDC (N = 821)	214	194	130	37
Progenitor (N = 555)	93	127	92	65
Macrophages (N = 126)	41	40	92	77
Mast (N = 80)	16	8	57	44

Table 1: Shows the number of differentially expressed genes (using TWO-SIGMA) and gene sets (using TWO-SIGMA-G) after FDR-adjustment for the HIV dataset. Gene-level p-values were adjusted using the Benjamini-Hochberg method, and significance was determined by comparing these adjusted p-values to the 5% significance threshold. Gene sets tested passed FDR-adjustment of the Fisher's method p-value combining the nine cell-type specific p-values, and marginal significance was judged as compared to the 5% significance threshold.

(A) presents heatmaps showing cell-type-specific average log fold changes (FC) and corresponding p-values for gene sets among the ten most significant in at least one of the nine cell-types. Sets related to virus introduction and interferon release are expected to be consistently upregulated and highly significant at both the set-level (as seen in a representative gene set in Supplementary figure S9) and the gene-level [24]. The significance of these sets is found both when combining p-values into a consensus FDR-adjusted p-value using Fisher's method and within cell types other than erythroid cells, albeit with differing strength of significance. Given the known functionality of erythroid cells as oxygen carriers in contrast to the immune function of the other cell types, this result is expected. Rather, it demonstrates that TWO-SIGMA-G can recover expected biological findings using cell-type-specific analyses and quantify differing strengths of association even among sets that may not exhibit large cell-type-specific heterogeneity. Panels (B) and (C) of figure 4 show the overlap in FDR-adjusted DE genes and gene sets, respectively, among the four most prevalent cell types. For example, there are 41 gene sets that are significant over all nine cell types analyzed that are also significant in each of NK, Erythroid, ILC, and B cells. These Venn diagrams show that our analysis reveals a large degree of cell-type specific heterogeneity at the gene level and the set level.

2.6 Analysis of Alzheimer's data reveals cell-type-specific heterogeneity in set-level expression

The second real data analysis is designed to demonstrate TWO-SIGMA-G using a more complex application. Specifically, we use the scRNA-seq data of [20] (see Methods section for more details) to analyze changes in gene expression as Alzheimer's Disease (AD) progresses. The data provides gene expression across three

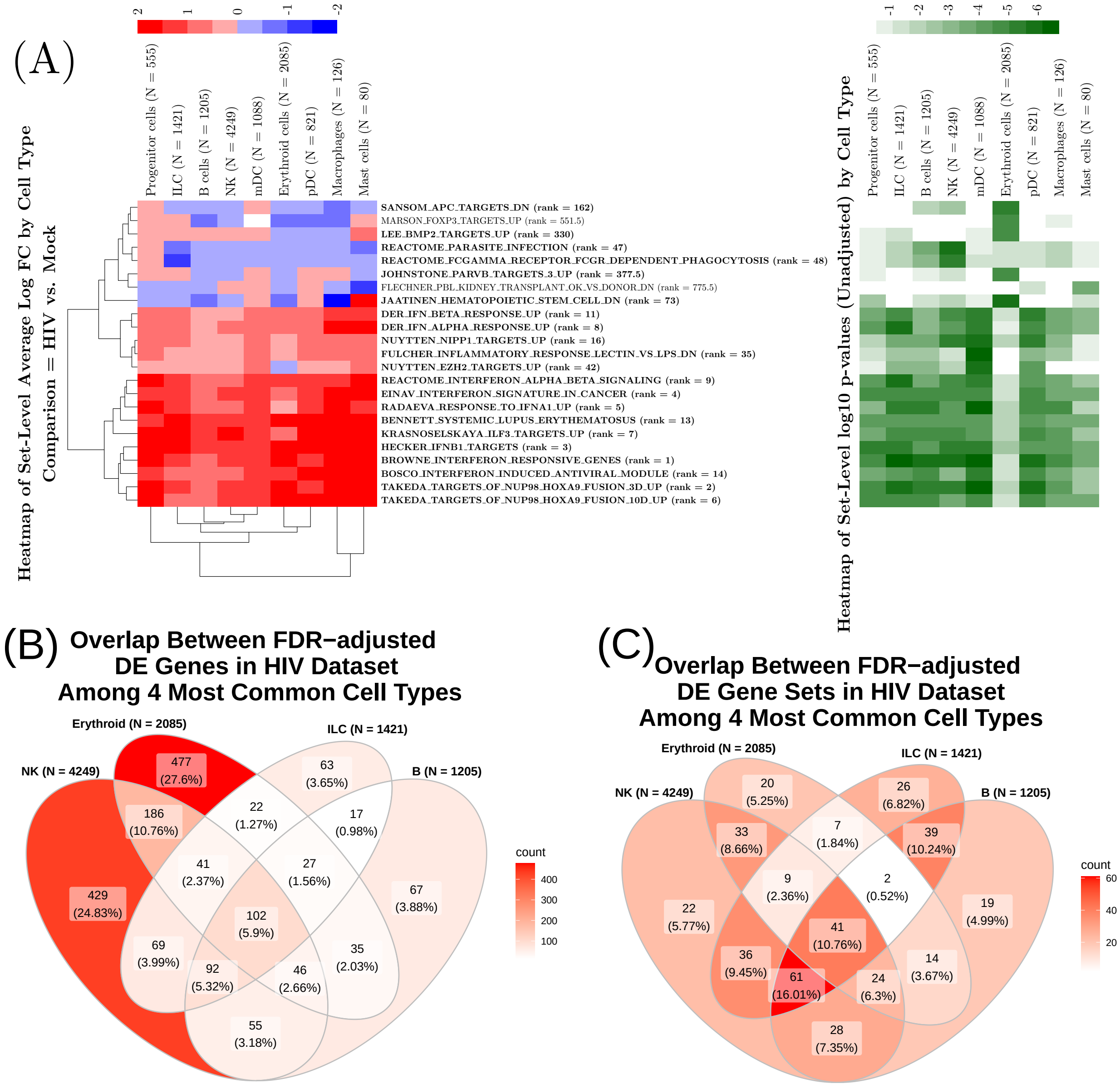

Figure 4: Results from analysis of the HIV dataset. (A) Cell-type specific variation in average set-level log fold-change (left) and significance (right). Sets plotted are among the top 10 in significance for at least one cell type. Sets in bold are significant at the 5% level over all cell types after FDR-adjustment of the Fisher's method p-value, and the rank of the Fisher's p-value among all sets is provided next to the set name. (B) Overlap between FDR-adjusted DE genes (5% significance level) among the four most prevalent cell types. (C) Overlap between FDR-adjusted DE gene sets (determined as in panel (A)) among the four most prevalent cell types.

distinct pathology groups: control (AD free), early-stage AD progression, and late-stage AD progression. We focus on two relevant comparisons: late vs. early-stage AD, and early-stage AD vs. control. A total of 6,048 genes and 5,074 gene sets were analyzed.

	Early Stage AD vs. Control				Late vs. Early Stage AD			
	Genes		Sets		Genes		Sets	
Cell Type	Up	Down	Up	Down	Up	Down	Up	Down
Excitatory Neuron (N = 29018)	1055	1893	5	26	1619	2645	410	7
Oligodendrocyte (N = 14806)	339	619	3	5	1443	74	297	1
Inhibitory Neuron (N = 7621)	58	1781	4	29	1483	761	374	7
Astrocyte (N = 2840)	61	311	4	4	265	44	245	5
Oligodendrocyte Progenitor (N = 2207)	13	64	4	3	256	3	251	3
Microglia (N = 1491)	27	20	5	6	14	0	42	2

Table 2: Shows the number of differentially expressed genes (using TWO-SIGMA) and gene sets (using TWO-SIGMA-G) after FDR-adjustment for both comparisons of the Alzheimer's dataset. Gene-level p-values were adjusted using the Benjamini-Hochberg method, and significance was determined by comparing these adjusted p-values to the 5% significance threshold. Gene sets tested passed FDR-adjustment of the Fisher's method p-value combining the six cell-type-specific p-values, and marginal significance was judged as compared to the 5% significance threshold.

Figure 5 shows cell-typ- specific results comparing early stage AD patients to control. 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 [21]. As panel (A) of figure 5 demonstrates, we replicate this finding with particularly robust downregulation seen in pathways related to cellular respiration, such as "KEGG_OXIDATIVE_PHOSPHORYLATION" and "MOOTHA_VOXPHOS" (see Supplementary Figures S10 and S11 for more detailed gene-level results for these sets). The neuronal cell types demonstrate highly consistent statistical significance in these cases. The Venn diagrams in panels (B) and (C) of figure 5 show that, after FDR correction, the most significant gene sets are shared between only the two neuronal cell types. This suggests that pathway changes in the early stages of Alzheimer's disease are most identifiable in these cell types. Table 2 shows breakdowns of the totals by direction of differential expression, and additionally includes all cell types.

Previous differential expression analyses in AD patients have further suggested that this trend of decreased expression in genes associated with cellular respiration may reverse as the disease progresses [21, 19]. To investigate this possibility, figure 6 shows cell-type-specific results comparing late-stage AD to early-stage AD. Heatmaps in panel (A) show that most of the top gene sets are now consistently and significantly upregulated, and furthermore many of these sets are also seen as highly significant but downregulated in panel (A) of figure 6. Thus, the initial downregulation in gene sets related to cellular respiration is reversed over time, possibly due to cellular degeneration and an increasing demand for energy in remaining cells [21] (see Supplementary figures S12 and S13 for gene-level information for the sets discussed above). Interestingly, this observed upregulation and the possible increase in demand for energy is highly significant in the neuronal cells, as in the previous comparison, but also highly significant in astrocytes, oligodendrocytes, and oligodendrocyte progenitor cells. Panel (C) of figure 6 shows that, unlike in the previous comparison, there is a large degree of overlap between the top four cell types among all DE gene sets. When comparing late-stage AD to control (Supplementary figures S14-S16), there is a slight upregulation in respiration related gene sets, although such sets are not among the most significant sets for any cell type. Thus, our analyses suggest that there is a systematic decrease in the expression of genes related to cellular respiration in the neuronal cells of early-stage AD patients. This decrease is reversed and even slightly over-compensated for when comparing late-stage patients to early stage patients, both in neuronal cells and in other cell types. Without the breakdowns of AD patients into the early and late stages, this pattern is obscured (Supplementary Figures S17-S19). Table 2 shows that the number of differentially expressed sets and genes increases dramatically in the late vs. early-stage AD comparison over the early stage AD to control comparison. This further reinforces the idea that massive changes in gene and pathway expression profiles occur in late-stage AD patients.

Our analysis explicitly reveals other cell-type type-specific heterogeneity. For example, microglia cells tend to have a unique set-level effect size profile, as demonstrated by the hierarchical clustering in the left heatmap of figure 5. This uniqueness also extends to significance. In comparing early-stage patients to control, microglia cells exhibit stronger significance in pathways involved in immune response, such as "RADAEVA_RESPONSE_TO_IFNA1_DN" or "BROWNE_INTERFERON_RESPONSIVE_ GENES," while showing less or no significance in previously mentioned pathways related to cellular respiration. Given the role of microglia cells in immune response, these results are not surprising. For a general application, however, TWO-SIGMA-G can help researchers to investigate cell-type-specific heterogeneity using the approach used here. The ability to test complex gene-level hypotheses as contrasts of regression parameters increases the diversity of cell-type-specific hypotheses that can be explored.

3 Discussion

We propose TWO-SIGMA-G, a novel method designed for competitive gene set testing using scRNA-seq data. At the gene-level, we employ our previously developed TWO-SIGMA method to test for DE. TWO-SIGMA is a flexible regression modelling framework that can fit both one-component and two-component negative binomial regression models to allow for overdispersed and zero-inflated counts. Additional covariates can be included in each of the two components, and sample-specific random effect terms can be included to account for within-sample correlation. The gene-level hypothesis is not limited to a binary or categorical group comparison, but rather can be a general contrast of regression parameters, as demonstrated in the real data analyses. This flexibility allows the testing of complex hypotheses and the analysis of complex experimental designs. At the set-level, we adjust for IGC, which has been demonstrated to inflate type-I error if mistakenly ignored. Using gene-level residuals to estimate IGC, we produce set-level p-values that preserve type-I error and improve power over alternative approaches.

The ability of TWO-SIGMA-G to include random effect terms at the gene-level provides a distinguishing factor from many methods for gene set analysis. Such random effects can improve inference at the gene-level substantially for some genes [30]. However, if only interested in set-level inference, our simulations suggest that statistical inference remains valid when excluding gene-level random effects and reducing computational burden as a result. When gene-level inference is of interest, it is likely desirable to fall back on including random effect terms into the regression modelling framework. However, we suggest that inference in real data analyses is likely not influenced greatly at the set-level by the presence or absence of gene-level random effects (Supplementary Figure S20).

TWO-SIGMA-G is implemented in the `twosigma` R package (https://github.com/edvanburen/twosigma), which is computationally efficient and allows for parallelization. To benchmark computational performance, we ran a modified version of our HIV data analysis, testing for a treatment effect of HIV pooled over all cell types. This modification to a one degree of freedom hypothesis allows us to test identical hypotheses in TWO-SIGMA-G, MAST, and CAMERA to provide a fairer comparison of computation. Using three computing cores on a MacBook Pro laptop, the methods had the following respective runtimes: 33.2 minutes for TWO-SIGMA-G, 33.5 minutes for MAST (25 bootstrap replications), and 5 seconds for CAMERA. TWO-SIGMA-G shows slightly improved yet nearly identical computational performance to MAST in the presence of the other advantages for performing gene set testing in scRNA-seq data described throughout this paper.

Unlike bulk RNA-seq data, many genes are often uncaptured or fail to survive filtering in scRNA-seq data. In gene set analysis, we must therefore assume that a gene set can be represented by the genes that exist in the dataset. In both the HIV and Alzheimer's datasets, we typically have around 40% representation regardless of set size after gene filtering (Supplementary Figure S21). Given the biologically meaningful and interpretable results we presented, the absence of these genes does not seem to threaten the ability of scRNA-seq gene set analyses to contribute new biological insights.

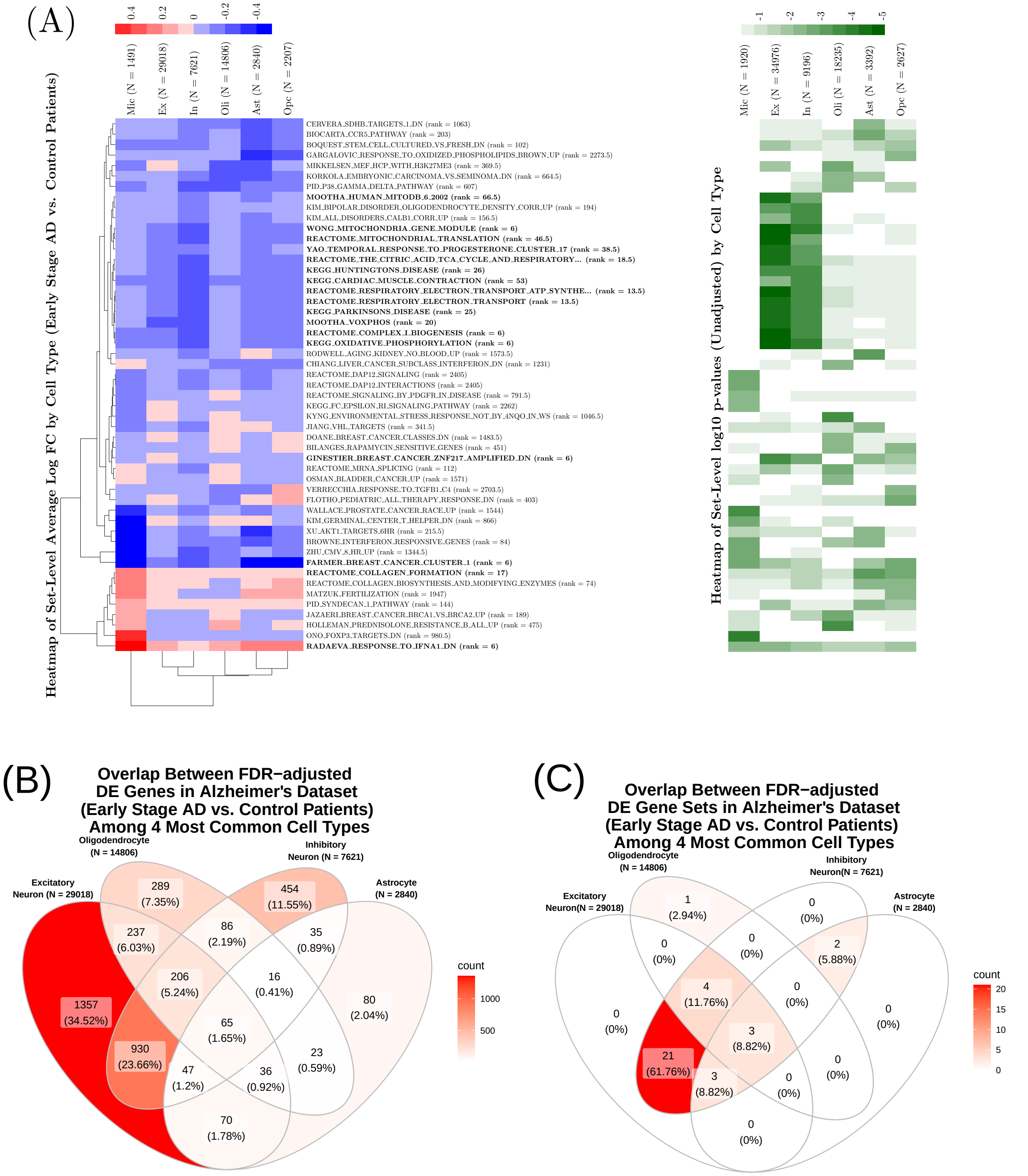

Figure 5: Results from Alzheimer's dataset comparing Early Stage AD to Control. (A) Cell-type-specific variation in average set-level log fold-change (left) and significance (right). Gene sets plotted are among the top 10 in significance for at least one cell type. Sets in bold are significant at the 5% level over all cell types after FDR-adjustment of the Fisher's method p-value, and the rank of the Fisher's p-value among all sets is provided next to the set name. (B) Overlap between FDR-adjusted DE genes (5% significance level) among the four most prevalent cell types. (C) Overlap between FDR-adjusted DE gene sets (determined as in panel (A)) among the four most prevalent cell types.

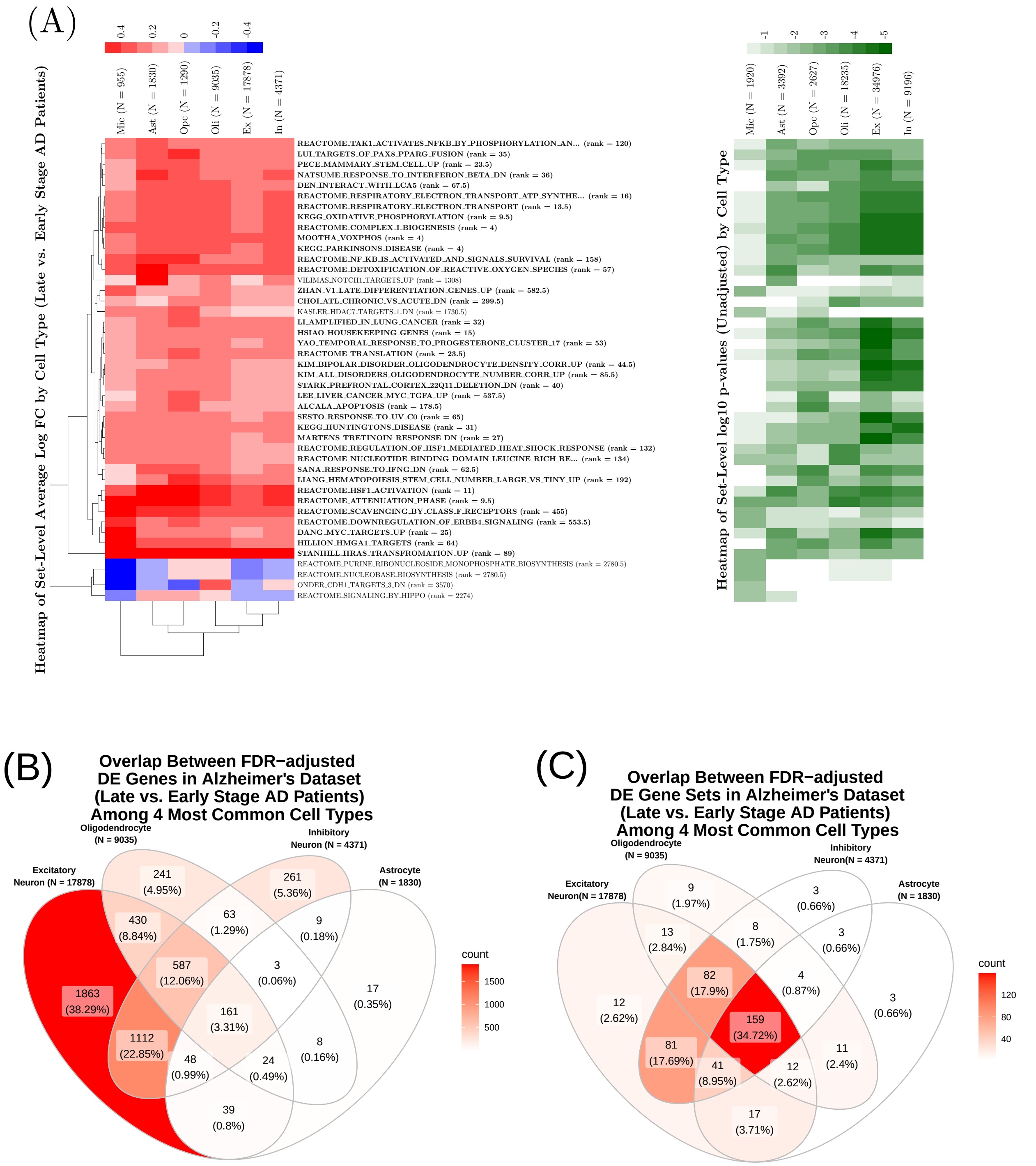

Figure 6: Results from Alzheimer's dataset comparing Late to Early Stage AD. (A) Cell-type-specific variation in average set-level log fold-change (left) and significance (right). Gene sets plotted are among the top 10 in significance for at least one cell type. Sets in bold are significant at the 5% level over all cell types after FDR-adjustment of the Fisher's method p-value, and the rank of the Fisher's p-value among all sets is provided next to the set name. (B) Overlap between FDR-adjusted DE genes (5% significance level) among the four most prevalent cell types. (C) Overlap between FDR-adjusted DE gene sets (determined as in panel (A)) among the four most prevalent cell types.

4 Methods

4.1 Compared Methods

We compared TWO-SIGMA-G to four other methods for competitive gene set testing in R version 3.6.3: the competitive testing procedure of MAST (version 1.13.5, accessed via the gseaAfterBoot function), CAMERA (accessed via version 3.42 of the limma package), iDEA (version 1.0.1), and PAGE (accessed via the PGSEA function of version 1.60 of the PGSEA) package. Log fold-change values from TWO-SIGMA were used as input for both iDEA and PAGE, and for MAST, 25 bootstrap replicates were used. In all other cases, default options were used for each method.

PAGE was developed as an extension of GSEA, and compares the log fold-change values in the test set to those from the complement set of genes using a one-sample z-test, where the sample mean and variance are estimated using all genes. PAGE does not adjust for IGC, but is computationally quite efficient as compared to GSEA. It may fail to preserve type-I error in some cases, however.

CAMERA was proposed as a competitive gene set test for microarray or RNA-seq data [33]. Gene-level statistics are first constructed using a linear model, meaning 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. Rather than using the raw data to estimate the IGC, CAMERA uses the residuals from the linear model. Use of the residuals means that the variation in gene expression explained by the covariates is removed, giving 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 [10]. 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, has an extension to allow for competitive gene set testing comparing a test set to its complement set of genes [7]. This extension is quite flexible given the log-normal hurdle regression framework employed by MAST. Once the gene-level statistics are collected, a bootstrap procedure 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 performance of the method does not seem to have been studied in great depth, and recent evidence has suggested that log transforming scRNA-seq data may distort true signal [29, 17].

iDEA was developed as an integrative method for both DE and gene set enrichment analysis [18]. The method takes gene-level DE summary statistics and gene sets as input. For each gene set, the method produces a posterior probability of DE for each gene and a set-level p-value. As a competitive test, the set-level p-value compares the gene-level odds of DE in the test set to the reference set. Because it focuses on the posterior probability of DE at the gene level, iDEA may not capture a gene set where enrichment of the test and reference sets is similar in proportion but the effect sizes themselves are systematically larger in the test set.

Both MAST and iDEA use the complement set of genes as the reference set. Previous studies have, however, cautioned that set size may inflate the type-I error of some gene set testing procedures [4, 28]. For larger gene sets, which are likely of more interest scientifically, the difference between these two approaches for choosing a reference set diminishes.

4.2 TWO-SIGMA for Gene-Level DE Statistics

TWO-SIGMA fits a zero-inflated mixed-effects negative binomial regression model and simultaneously models, for cell j of individual i, the probability of dropout p_{ij} and the negative binomial mean μ_{ij} as follows:

$$\begin{aligned}
&\text{logit}(p_{ij}) = \boldsymbol{z_{ij}^T}\boldsymbol{\alpha} + a_i, a_i \sim N(0, \sigma_a^2) \\
&\log(\mu_{ij}) = \boldsymbol{x_{ij}^T}\boldsymbol{\beta} + b_i, b_i \sim N(0, \sigma_b^2), \text{ assume } a_i \perp\!\!\!\perp b_i
\end{aligned} \tag{1}$$

$\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are fixed effect coefficient vectors and the corresponding vectors of covariates $\boldsymbol{z_{ij}}$ and $\boldsymbol{x_{ij}}$ can be different. a_i and b_i are sample-specific random intercept terms. Including these terms helps control for any within-sample correlation, providing more accurate estimates and standard errors of fixed effect parameters. In this context, gene-level statistics will correspond to tests of estimable contrasts of regression parameters in the TWO-SIGMA model seen in equation (1). Examples could include a likelihood ratio statistic of a treatment effect, or an ANOVA style pairwise comparison between treatment groups within cell-type as we demonstrate in the real data analysis of this paper. If an effect of interest is present in both the mean and zero-inflation component, options for a joint test include the likelihood ratio test, Stouffer's method to combine Z-scores, or using simply using the test statistic from the mean model. To summarize, TWO-SIGMA can control for additional covariates in both components, incorporate random effects to accommodate within-sample dependency, analyze unbalanced data, test general DE hypotheses beyond a two-group comparison, and allow for zero-inflated and overdispersed counts at the gene-level. The zero-inflation component can be removed in its entirety if desired, as is done in the real data analyses.

4.3 Gene Set Simulation Procedure

To simulate correlated gene sets, with varying magnitudes of correlation designed to represent the diversity seen in real data, we employ the following simulation procedure:

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. Vary magnitude of IGC in 2(a) by taking a_2 from various values in table 3

Table 3: 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	O.C.
3	No
5	No
10	No
3	Yes
5	Yes
10	Yes

4. Repeat 1-4 using 10 different random seeds (mimic different cell populations)

Genes were simulated using the zero-inflated negative binomial distribution to be under the gene-level null or, for simplicity, a single gene-level alternative (this is relaxed in supplementary Figure S7). Gene sets possessing a compound symmetric correlation structure were generated by a procedure described fully in supplementary section S1. Briefly, a total of six different settings were constructed to vary both the amount of inter-gene correlation in the test set and the presence of other covariates in the gene-level model. The presence of such other covariates can create additional complex correlation structures between cells and genes as discussed in section 2.2. For each setting, we aggregated over ten biological replicates consisting of different cell populations to minimize the impact of the initial cell population on results. At the set-level, sets can be constructed to be under various null or alternative hypotheses by varying the proportion of genes that are under the *gene-level* null or alternative in both the test sets and reference sets. Settings were repeated using reference sets of size 30 and 100 to evaluate the impact of reference set size on set-level inference. This simulation strategy introduces a small positive correlation which varies from 0 to about 0.05 depending on simulation scenario and the computational method used to estimate the correlation. Our aim is not to evaluate correlation estimates directly, but rather to introduce small but positive gene-gene correlations and evaluate the ability of various competitive gene set testing methods based on their set-level performance after adjusting for inter-gene correlation.

In the main simulations TWO-SIGMA-G was compared to two other methods for competitive testing using regression modelling approaches: CAMERA [33], the leading method for bulk RNA-seq and thus for competitive testing, and the procedure in MAST [7], which is one of the most popular packages for scRNA-seq data analysis. We had difficulties obtaining reliable p-values from iDEA [18] and PAGE [14] for the main simulations. We believe this is because our main simulations were calibrated using gene-level statistics which summarize evidence from both the mean and zero-inflation components, while iDEA and PAGE use only the effect size (and standard error in iDEA) from the mean component. TWO-SIGMA-G, CAMERA, and MAST all utilize the raw data and as such can capture general set-level enrichment coming from expression changes in zero proportion or mean value. In contrast, iDEA and PAGE do not use the raw data. To provide a meaningful and fair comparison to both iDEA and PAGE, we simulated correlated genes emphasizing signal in the mean component. Type-I error results for these methods are shown in Supplementary Figure S2 and provide similar conclusions to the above. Methods designed for self-contained testing, a hybrid of self-contained and competitive testing, or other aspects of gene set testing, such as ROAST [32], GSEA [25], `sigPathway` [28], PAGODA [6], and BAGSE [13] were not included because they are testing fundamentally different null hypotheses.

4.4 HIV Dataset

Our first dataset consists of single-cells collected from humanized mice [3]. Given our focus is at the set level, we filtered genes to keep those with zero proportion no higher than the mean percentage, leaving the most relevant and highly expressed 3,549 genes. A total of nine cell types are present in the data: natural killer (NK) cells, erythroid cells, innate lymphoid cells (ILC), B cells, myeloid dendritic cells (mDC), plasmacytoid dendritic cells (pDC), progenitor cells, macrophages, and mast cells. The read counts are then treated as the outcome of interest; as with other UMI-based scRNA-seq count data, we found that this data was not consistent with zero-inflation [26], and thus we fit the TWO-SIGMA model without the zero-inflation component at the gene-level. Because the primary interest is in comparisons between HIV and mock cells within cell-type, we categorize cells into one of 2*9 = 18 mutually exclusive groups. An ANCOVA model additionally adjusting for CDR was fit as a way to test for cell-type specific differences in expression levels comparing HIV to mock. TWO-SIGMA-G is ideal for this analysis because gene-level statistics can come from a test of such an arbitrary contrast matrix. These gene-level statistics are, for each cell-type, Wald Z-statistics contrasting the mean values in observed expression between the two groups within a cell-type. Gene sets were taken from the Molecular Signatures Database (mSigDB) [25, 15] 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 our data, a total of 4,772 sets with at least two genes present in our data were analyzed. More detailed breakdowns showing the percentage of genes available by set size are available in Supplementary Figure S19.

4.5 Alzheimer's Dataset

Our second dataset consists of 70,634 single cells from human donors [20]. We did not remove cells beyond what was done in the original manuscript. Given our focus is at the set level, however, we chose to filter the original 17,926 genes to the 6,048 most highly expressed genes by removing genes unexpressed in at least 90% of cells. The read counts are once again treated as the outcome of interest in a model without a zero-inflation component. A total of 48 individual donors are present, categorized into three pathology groups: 24 individuals are control patients free of a diagnosis of AD, 12 were diagnosed with early stage AD, and 12 were diagnosed with late stage AD. The 70,634 single cells from the six most common cell-types were analyzed: astrocytes (Ast), excitatory neurons (Ex), inhibitory neurons (In), microglia (Mic), oligodendrocytes (Oli), and oligodendrocyte progenitor cells (Opc). The existence of the pathology groups allows us to explore cell-type specific variability in gene expression as AD progresses into early and late stages of disease severity. Our geneset analysis was conducted similarly to above: 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. In total, 5,074 sets with at least two genes present from the MsigDB c2 collection were analyzed. More detailed breakdowns showing the percentage of genes available by set size are available in Supplementary Figure S19.

5 Availability of data and materials

Both datasets analyzed in this manuscript are publicly available. The HIV dataset is available at the Gene Expression Omnibus under accession GSE148796. The Alzheimer's dataset is available upon completion of a data usage agreement at the Rush Alzheimer's Disease Center (RADC) Research Resource Sharing Hub (https://www.radc.rush.edu/docs/omics.htm) under "snRNA-seq PFC." TWO-SIGMA-G is implemented in the function `twosigmag` in the `twosigma` R package, which is freely available on GitHub at https://github.com/edvanburen/twosigma.

6 Declarations

Competing Interests: No authors have a competing interest.

References

[1] William T. Barry, Andrew B. Nobel, and Fred A. Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–1949, 01 2005.

[2] William T. Barry, Andrew B. Nobel, and Fred A. Wright. A statistical framework for testing functional categories in microarray data. *Ann. Appl. Stat.*, 2(1):286–315, 03 2008.

[3] Liang Cheng, Haisheng Yu, John A. Wrobel, Guangming Li, Peng Liu, Zhiyuan Hu, Xiao-Ning Xu, and Lishan Su. Identification of pathogenic trail-expressing innate immune cells during hiv-1 infection in humanized mice by scrna-seq. *JCI Insight*, 5(11), 6 2020.

[4] Doris Damian and Malka Gorfine. Statistical concerns about the gsea procedure. *Nature Genetics*, 36(7):663–663, 2004.

[5] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *Ann. Appl. Stat.*, 1(1):107–129, 06 2007.

[6] Jean Fan, Neeraj Salathia, Rui Liu, Gwendolyn E Kaeser, Yun C Yung, Joseph L Herman, Fiona Kaper, Jian-Bing Fan, Kun Zhang, Jerold Chun, and Peter V Kharchenko. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature Methods*, 13(3):241–244, 2016.

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

[8] Daniel M. Gatti, William T. Barry, Andrew B. Nobel, Ivan Rusyn, and Fred A. Wright. Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics*, 11(1):574, 2010.

[9] Sheila M Gaynor, Ryan Sun, Xihong Lin, and John Quackenbush. Identification of differentially expressed gene sets using the Generalized Berk–Jones statistic. *Bioinformatics*, 35(22):4568–4576, 05 2019.

[10] Jelle J. Goeman and Peter Buhlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 02 2007.

[11] Prakash K. Gupta, Jernej Godec, David Wolski, Emily Adland, Kathleen Yates, Kristen E. Pauken, Cormac Cosgrove, Carola Ledderose, Wolfgang G. Junger, Simon C. Robson, E. John Wherry, Galit Alter, Philip J. R. Goulder, Paul Klenerman, Arlene H. Sharpe, Georg M. Lauer, and W. Nicholas Haining. Cd39 expression identifies terminally exhausted cd8+ t cells. *PLOS Pathogens*, 11(10):1–21, 10 2015.

[12] Pleun Hombrink, Christina Helbig, Ronald A Backer, Berber Piet, Anna E Oja, Regina Stark, Giso Brasser, Aldo Jongejan, RenéE Jonkers, Benjamin Nota, Onur Basak, Hans C Clevers, Perry D Mo-

erland, Derk Amsen, and RenéA W van Lier. Programs for the persistence, vigilance and control of human cd8+ lung-resident memory t cells. *Nature Immunology*, 17(12):1467–1478, 2016.

[13] Abhay Hukku, Corbin Quick, Francesca Luca, Roger Pique-Regi, and Xiaoquan Wen. BAGSE: a Bayesian hierarchical model approach for gene set enrichment analysis. *Bioinformatics*, 36(6):1689–1695, 11 2019.

[14] Seon-Young Kim and David J. Volsky. Page: Parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6(1):144, 2005.

[15] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417 – 425, 2015.

[16] Elgene Lim, François Vaillant, Di Wu, Natasha C Forrest, Bhupinder Pal, Adam H Hart, Marie-Liesse Asselin-Labat, David E Gyorki, Teresa Ward, Audrey Partanen, Frank Feleppa, Lily I Huschtscha, Heather J Thorne, Stephen B Fox, Max Yan, Juliet D French, Melissa A Brown, Gordon K Smyth, Jane E Visvader, Geoffrey J Lindeman, and kConFab. Aberrant luminal progenitors as the candidate target population for basal tumor development in brca1 mutation carriers. *Nature Medicine*, 15(8):907–913, 2009.

[17] Aaron Lun. Overcoming systematic errors caused by log-transformation of normalized single-cell rna sequencing data. *bioRxiv*, 2018.

[18] Ying Ma, Shiquan Sun, Xuequn Shang, Evan T. Keller, Mengjie Chen, and Xiang Zhou. Integrative differential expression and gene set enrichment analysis using summary statistics for scrna-seq studies. *Nature Communications*, 11(1):1585, 2020.

[19] Maria Manczak, Byung S. Park, Youngsin Jung, and P. Hemachandra Reddy. Differential expression of oxidative phosphorylation genes in patients with alzheimer's disease. *NeuroMolecular Medicine*, 5(2):147–162, 2004.

[20] Hansruedi Mathys, Jose Davila-Velderrain, Zhuyu Peng, Fan Gao, Shahin Mohammadi, Jennie Z. Young, Madhvi Menon, Liang He, Fatema Abdurrob, Xueqiao Jiang, Anthony J. Martorell, Richard M. Ransohoff, Brian P. Hafler, David A. Bennett, Manolis Kellis, and Li-Huei Tsai. Single-cell transcriptomic analysis of alzheimer's disease. *Nature*, 570(7761):332–337, 2019.

[21] Akihiko Nunomura, George Perry, Gjumrakch Aliev, Keisuke Hirai, Atsushi Takeda, Elizabeth K. Balraj, Paul K. Jones, Hossein Ghanbari, Takafumi Wataya, Shun Shimohama, Shigeru Chiba, Craig S.

Atwood, Robert B. Petersen, and Mark A. Smith. Oxidative Damage Is the Earliest Event in Alzheimer Disease. *Journal of Neuropathology & Experimental Neurology*, 60(8):759–767, 08 2001.

[22] Assaf P. Oron, Zhen Jiang, and Robert Gentleman. Gene set enrichment analysis using linear models and diagnostics. *Bioinformatics*, 24(22):2586–2591, 09 2008.

[23] Dalila Pinto, Alistair T. Pagnamenta, Lambertus Klei, Richard Anney, Daniele Merico, Regina Regan, Judith Conroy, Tiago R. Magalhaes, Catarina Correia, Brett S. Abrahams, Joana Almeida, Elena Bacchelli, Gary D. Bader, Anthony J. Bailey, Gillian Baird, Agatino Battaglia, Tom Berney, Nadia Bolshakova, Sven Bölte, Patrick F. Bolton, Thomas Bourgeron, Sean Brennan, Jessica Brian, Susan E. Bryson, Andrew R. Carson, Guillermo Casallo, Jillian Casey, Brian H. Y. Chung, Lynne Cochrane, Christina Corsello, Emily L. Crawford, Andrew Crossett, Cheryl Cytrynbaum, Geraldine Dawson, Maretha de Jonge, Richard Delorme, Irene Drmic, Eftichia Duketis, Frederico Duque, Annette Estes, Penny Farrar, Bridget A. Fernandez, Susan E. Folstein, Eric Fombonne, Christine M. Freitag, John Gilbert, Christopher Gillberg, Joseph T. Glessner, Jeremy Goldberg, Andrew Green, Jonathan Green, Stephen J. Guter, Hakon Hakonarson, Elizabeth A. Heron, Matthew Hill, Richard Holt, Jennifer L. Howe, Gillian Hughes, Vanessa Hus, Roberta Igliozzi, Cecilia Kim, Sabine M. Klauck, Alexander Kolevzon, Olena Korvatska, Vlad Kustanovich, Clara M. Lajonchere, Janine A. Lamb, Magdalena Laskawiec, Marion Leboyer, Ann Le Couteur, Bennett L. Leventhal, Anath C. Lionel, Xiao-Qing Liu, Catherine Lord, Linda Lotspeich, Sabata C. Lund, Elena Maestrini, William Mahoney, Carine Mantoulan, Christian R. Marshall, Helen McConachie, Christopher J. McDougle, Jane McGrath, William M. McMahon, Alison Merikangas, Ohsuke Migita, Nancy J. Minshew, Ghazala K. Mirza, Jeff Munson, Stanley F. Nelson, Carolyn Noakes, Abdul Noor, Gudrun Nygren, Guiomar Oliveira, Katerina Papanikolaou, Jeremy R. Parr, Barbara Parrini, Tara Paton, Andrew Pickles, Marion Pilorge, Joseph Piven, Chris P. Ponting, David J. Posey, Annemarie Poustka, Fritz Poustka, Aparna Prasad, Jiannis Ragoussis, Katy Renshaw, Jessica Rickaby, Wendy Roberts, Kathryn Roeder, Bernadette Roge, Michael L. Rutter, Laura J. Bierut, John P. Rice, Jeff Salt, Katherine Sansom, Daisuke Sato, Ricardo Segurado, Ana F. Sequeira, Lili Senman, Naisha Shah, Val C. Sheffield, Latha Soorya, Inês Sousa, Olaf Stein, Nuala Sykes, Vera Stoppioni, Christina Strawbridge, Raffaella Tancredi, Katherine Tansey, Bhooma Thiruvahindrapduram, Ann P. Thompson, Susanne Thomson, Ana Tryfon, John Tsiantis, Herman Van Engeland, John B. Vincent, Fred Volkmar, Simon Wallace, Kai Wang, Zhouzhi Wang, Thomas H. Wassink, Caleb Webber, Rosanna Weksberg, Kirsty Wing, Kerstin Wittemeyer, Shawn Wood, Jing Wu, Brian L. Yaspan, Danielle Zurawiecki, Lonnie Zwaigenbaum, Joseph D. Buxbaum, Rita M. Cantor, Edwin H. Cook, Hilary Coon, Michael L. Cuccaro, Bernie Devlin, Sean Ennis, Louise Gallagher, Daniel H. Geschwind,

Michael Gill, Jonathan L. Haines, Joachim Hallmayer, Judith Miller, Anthony P. Monaco, John I. Nurnberger Jr, Andrew D. Paterson, Margaret A. Pericak-Vance, Gerard D. Schellenberg, Peter Szatmari, Astrid M. Vicente, Veronica J. Vieland, Ellen M. Wijsman, Stephen W. Scherer, James S. Sutcliffe, and Catalina Betancur. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466(7304):368–372, 2010.

[24] Andrew Soper, Izumi Kimura, Shumpei Nagaoka, Yoriyuki Konno, Keisuke Yamamoto, Yoshio Koyanagi, and Kei Sato. Type i interferon responses by hiv-1 infection: Association with disease progression and control. *Frontiers in Immunology*, 8:1823, 2018.

[25] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. 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, 2005.

[26] Valentine Svensson. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 2020.

[27] Adi Laurentiu Tarca, Sorin Draghici, Purvesh Khatri, Sonia S. Hassan, Pooja Mittal, Jung-sun Kim, Chong Jai Kim, Juan Pedro Kusanovic, and Roberto Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, 11 2008.

[28] Lu Tian, Steven A. Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S. Kohane, and Peter J. Park. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences*, 102(38):13544–13549, 2005.

[29] F. William Townes, Stephanie C. Hicks, Martin J. Aryee, and Rafael A. Irizarry. Feature selection and dimension reduction for single cell rna-seq based on a multinomial model. *bioRxiv*, 2019.

[30] Eric Van Buren, Ming Hu, Chen Weng, Fulai Jin, Yan Li, Di Wu, and Yun Li. Two-sigma: A novel two-component single cell model-based association method for single-cell rna-seq data. *Genetic Epidemiology*, n/a(n/a).

[31] Koen Van den Berge, Charlotte Soneson, Michael I. Love, Mark D. Robinson, and Lieven Clement. zinger: unlocking rna-seq tools for zero-inflation and single cell applications. *bioRxiv*, 2017.

[32] Di Wu, Elgene Lim, François Vaillant, Marie-Liesse Asselin-Labat, Jane E. Visvader, and Gordon K. Smyth. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–2182, 07 2010.

[33] Di Wu and Gordon K. Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17):e133–e133, 05 2012.