Supplementary Information:

Cohort-scale differential expression analysis of single cell transcriptomics data using dreamlet

Gabriel E. Hoffman[1,2,3*], Donghoon Lee[1,2,3,4], Jaroslav Bendl[1,2,3,4], Prashant Fnu[1,2,3,4], Aram Hong[1,2,3,4], Clara Casey[1,2,3,4], Marcela Alvia[1,2,3,4], Zhiping Shao[1,2,3,4], Stathis Argyriou[1,2,3,4], Karen Therrien[1,2,3,4], Sanan Venkatesh[1,2,3,4], Georgios Voloudakis[1,2,3,4], Vahram Haroutunian[2,5,6], John Fullard[1,2,3,4], Panos Roussos[1,2,3,4,5,6*]

[1]Center for Disease Neurogenomics,
[2]Department of Psychiatry,
[3]Department of Genetics and Genomic Sciences,
[4]Friedman Brain Institute,
[5]Department of Neuroscience,
Icahn School of Medicine at Mount Sinai, New York, NY, USA
[6]Mental Illness Research, Education and Clinical Centers,
James J. Peters VA Medical Center, Bronx, New York

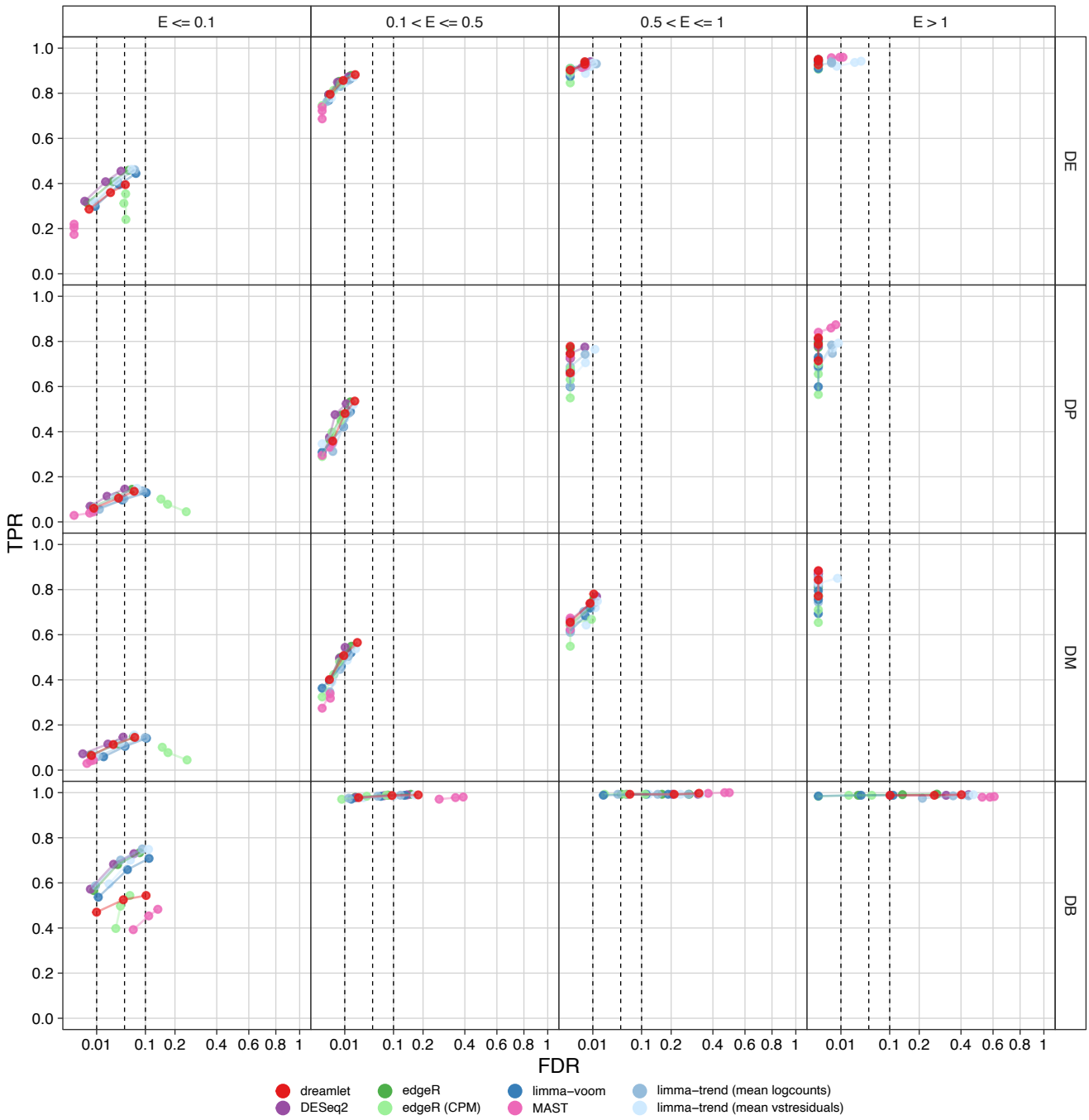
*Correspondence to: G.E.H (gabriel.hoffman@mssm.edu) and P.R. (panagiotis.roussos@mssm.edu)

Contents

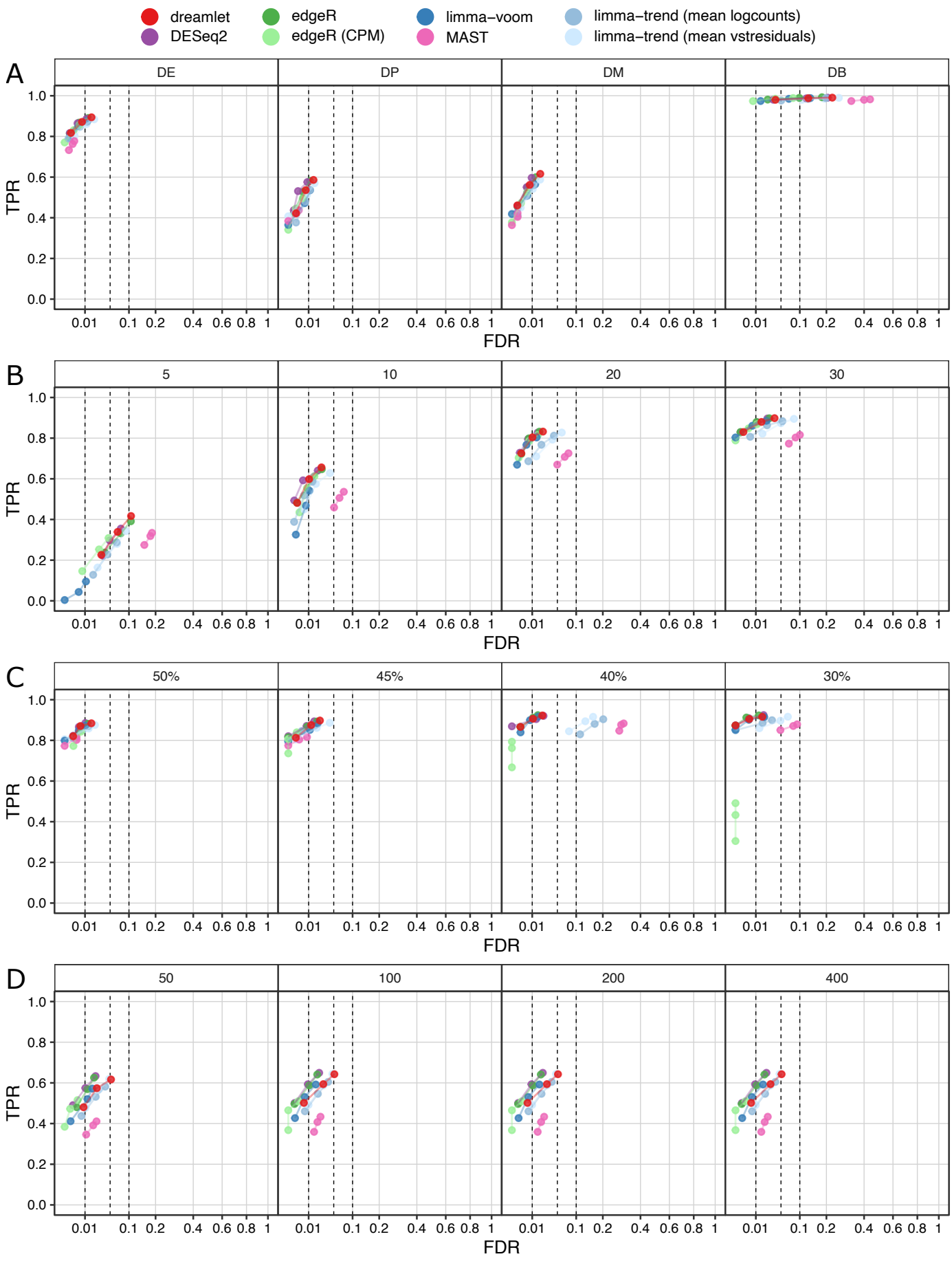
1 Supplementary Figures **2**

2 Supplementary Methods **11**

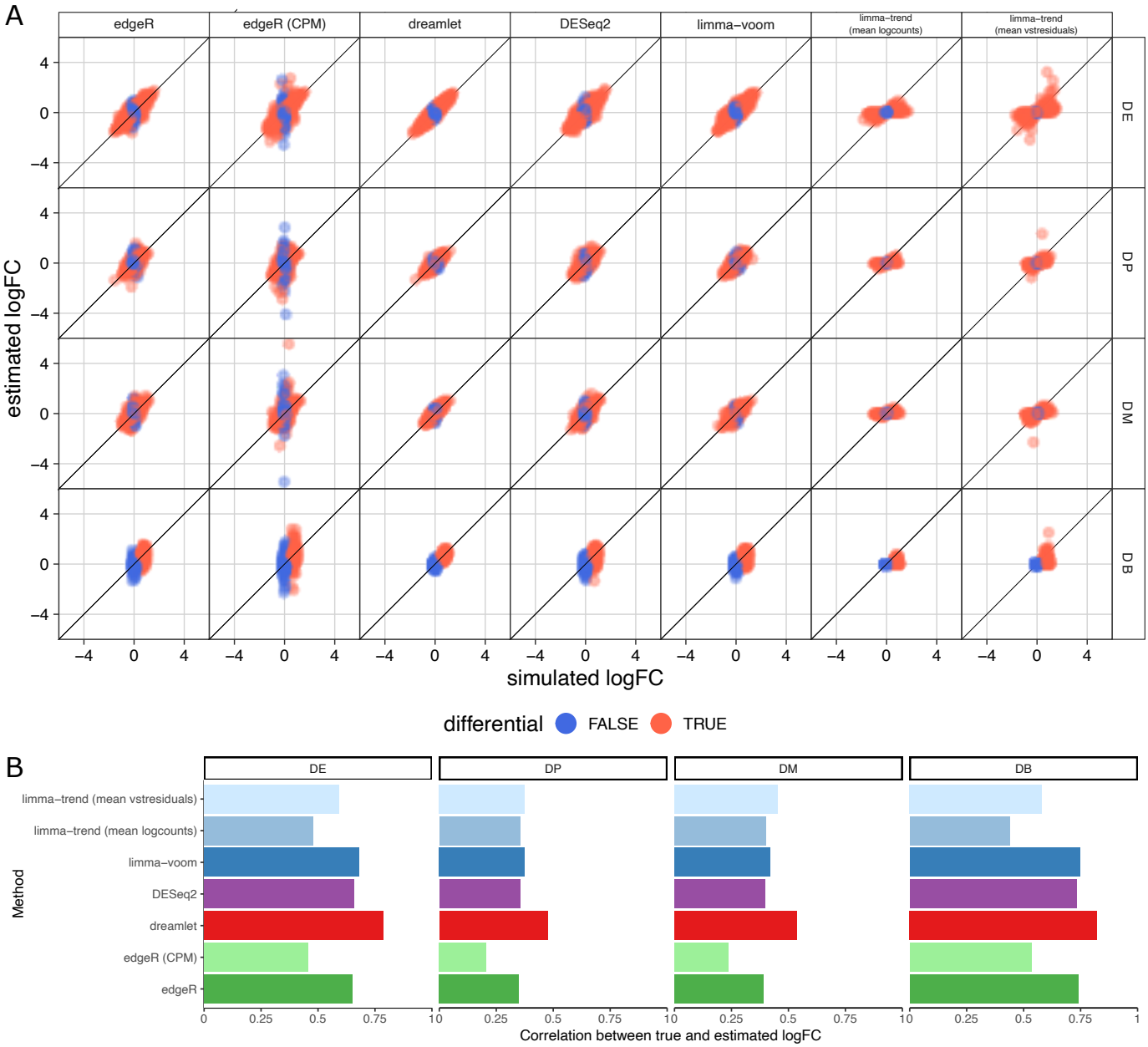
1 Supplementary Figures



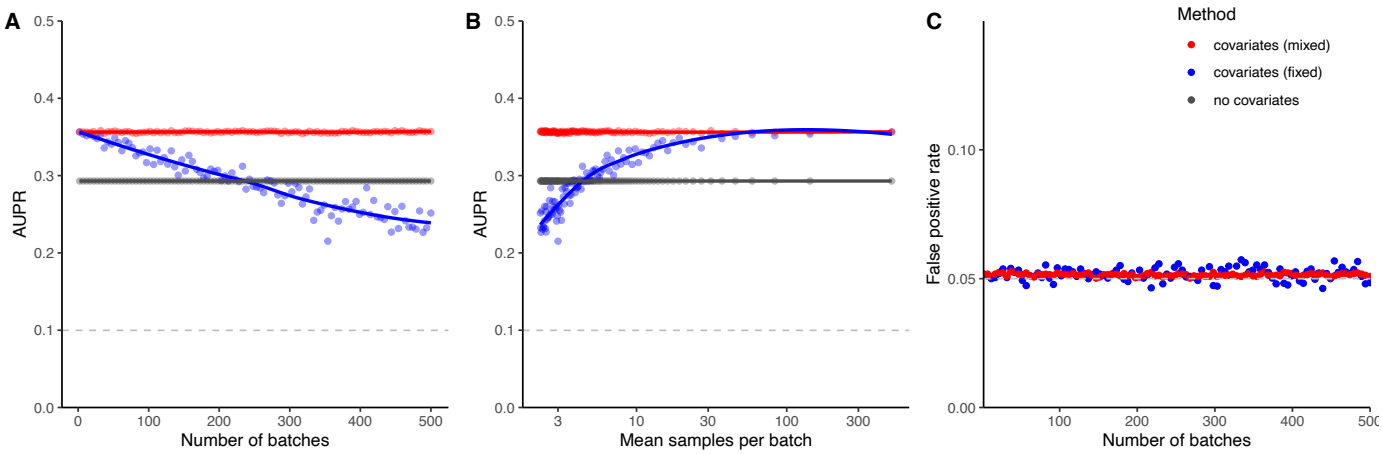
Supplemental Figure 1: Performance of differential expression analysis across expression levels. True positive rate (TPR) and false discovery rate (FDR) are shown for 4 simulation conditions (DE, DP, DM, DB). Performance is stratified based on mean difference in simulated gene expression between case and controls subjects. Vertical dashed lines indicate 1%, 5%, and 10% FDR, and points for each differential expression method are shown for each of these target values. Changes in the mean expression (DE), changes in the proportions of low and high expression-state components (DP), differential modality (DM), or changes in both proportions and modality (DB)


dreamlet
DESeq2
edgeR
edgeR (CPM)
limma−voom
MAST
limma−trend (mean logcounts)
limma−trend (mean vstresiduals)
A
DE
DP
DM
DB
B
5
10
20
30
C
50%
45%
40%
30%
D
50
100
200
400
TPR
FDR
0.0
0.2
0.4
0.6
0.8
1.0
0.01
0.1
0.2
0.4
0.6
0.8
1

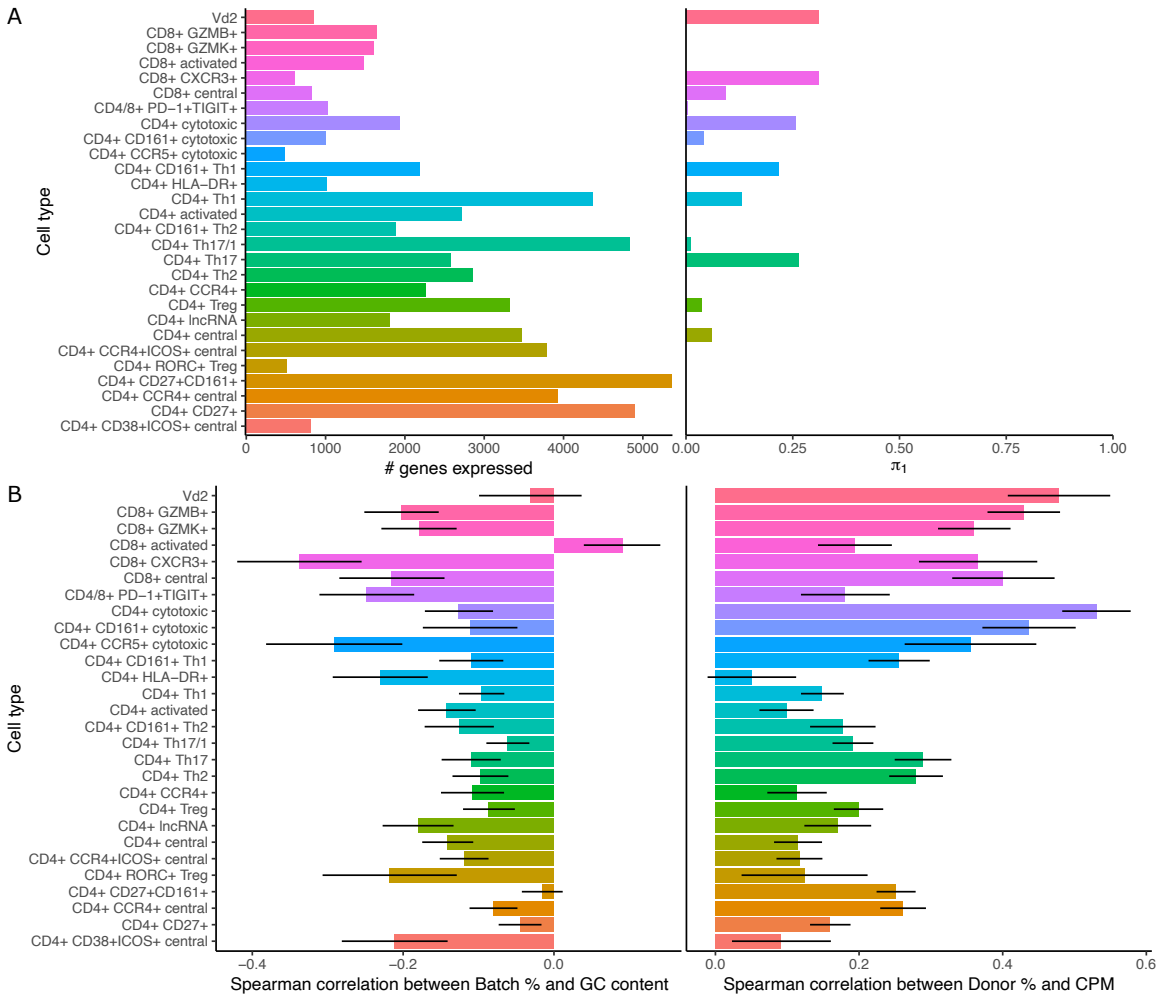
Supplemental Figure 2: Performance of differential expression analysis stratified by multiple variables. True positive rate (TPR) and false discovery rate (FDR) are shown for each differential expression method. Vertical dashed lines indicate 1%, 5%, and 10% FDR, and points for each method are shown for each of these target values. Performance is show stratified by **A**) simulation condition, **B**) number of subjects, **C**) balance between case and control groups (i.e. 45% indicates percentage of subjects in one class so 55% are in the other), and **D**) number of cells observed per subject.



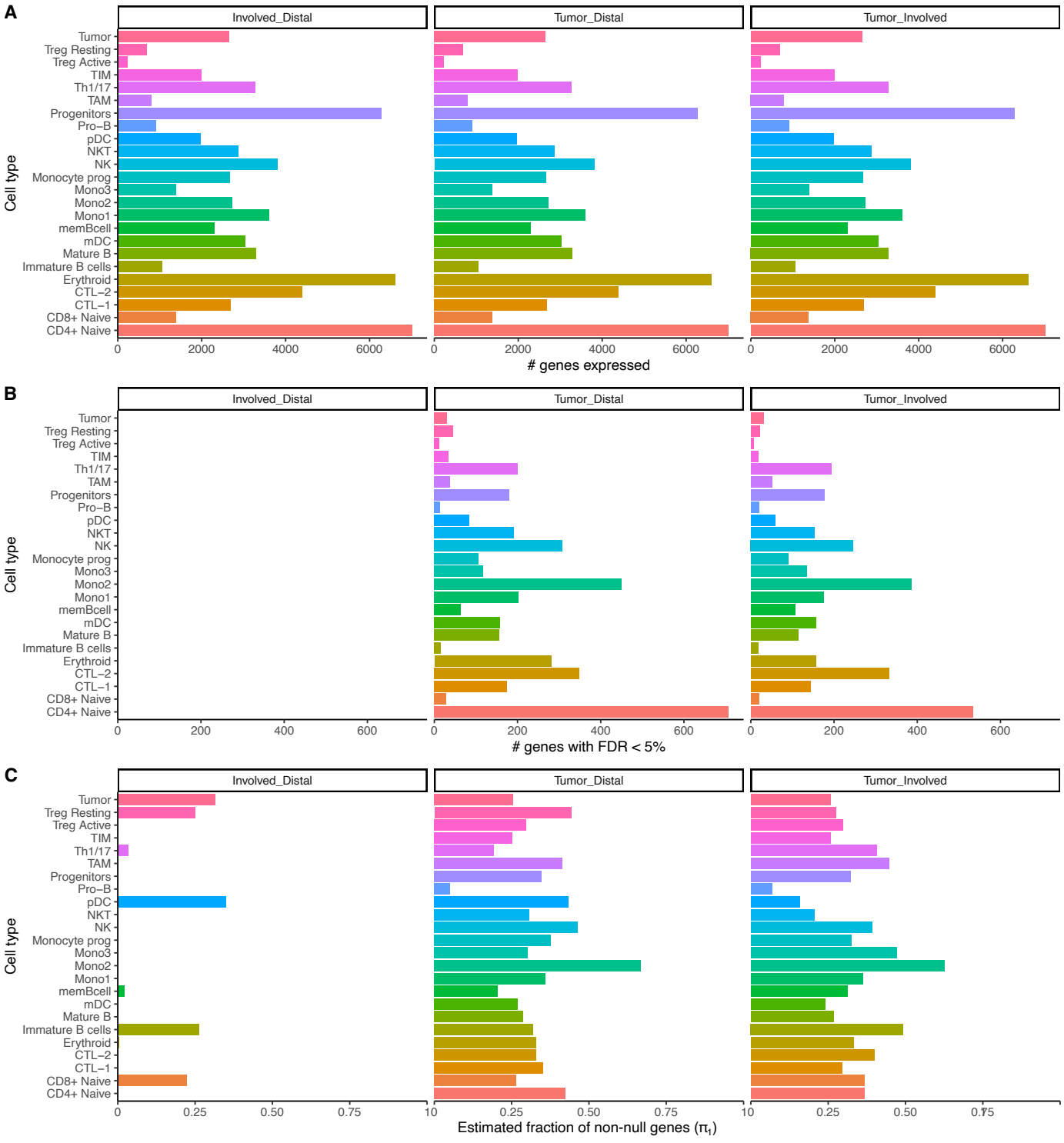
Supplemental Figure 3: Estimation of simulated log fold changes. **A**) Plot of simulated and estimated log fold changes values for each method under 4 simulation conditions (DE, DP, DM, DB). Genes with simulated effect size of zero are show in blue, those with non-zero effect size are shown in orange. **B**) Pearson correlation between simulated and estimated log fold changes for each method and simulation condition using results from (**A**).



Supplemental Figure 4: Performance of mixed models on simulated data. Simulations were performed for $n = 1000$ samples, and an increasing number of batches where the effect of each batch is drawn from a standard normal distribution. For each number of batches, 5000 genes were simulated and 500 were associated with a phenotype variable that was normally distributed. For each simulation, regression models accounting for batch using a fixed or mixed effect model, or no covariate correction were evaluated. The area under the precision recall curve (AUPR) and false positive rate (FPR) at 5% was computed for each simulation, and a lowess smooth is plotted. **A**) AUPR for each model versus the number of batches in the simulation. **B**). The same results as in **(A)** shown as a function of mean number of samples per batch on a log scale. **C**) FPR as a function of the number of batches. Both fixed and mixed effects models control the false positive rate, but the mixed model shows substantially reduced variance around the 5% target compared to the fixed effect model. Omitting covariates give inflated FPR beyond the y-axis shown here.



Supplemental Figure 5: Summary of differential expression results from tuberculosis infection. **A)** Number of genes passing expression cutoff for each cell population (left), and Storey's π_1 estimate of the fraction of genes rejecting the null hypothesis of no association. **B)** Correlation between the fraction of expression variation for each gene explained by experimental batch and the GC content of each gene (left), and correlation between the fraction of expression variation explained by donor and expression magnitude in counts per million (right).

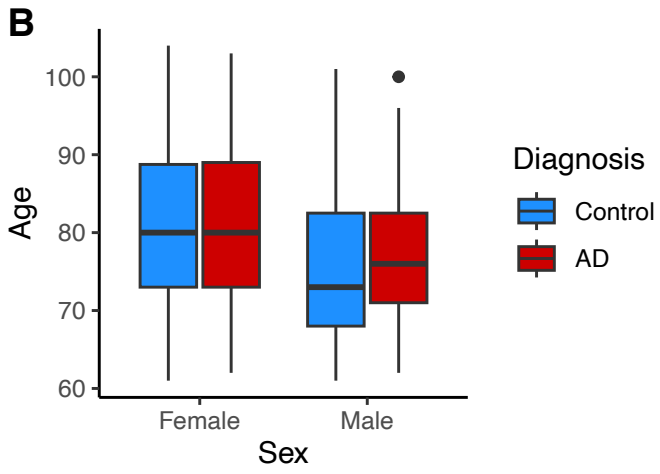


Supplemental Figure 6: Summary of differential expression results bone cancer metastases . **A)** Number of genes passing expression cutoff for comparison between tumor, involved tissue and distal tissue for each cell type. **B)** Number of differentially expressed genes at FDR 5%. Note that no genes passed the cutoff for the "Involved vs Distal" comparison. **C)** Storey's π_1 estimate of the fraction of genes rejecting the null hypothesis of no association.

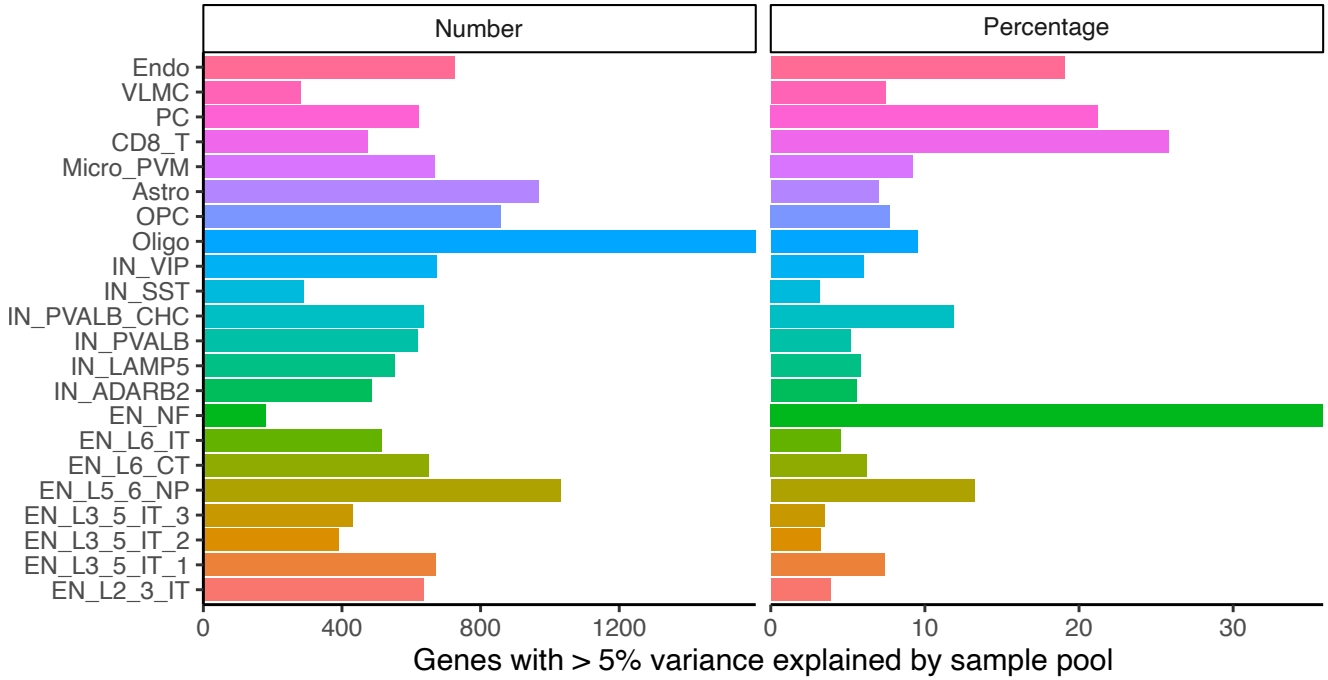
A

	Female	Male
Control	70	79
AD	71	79

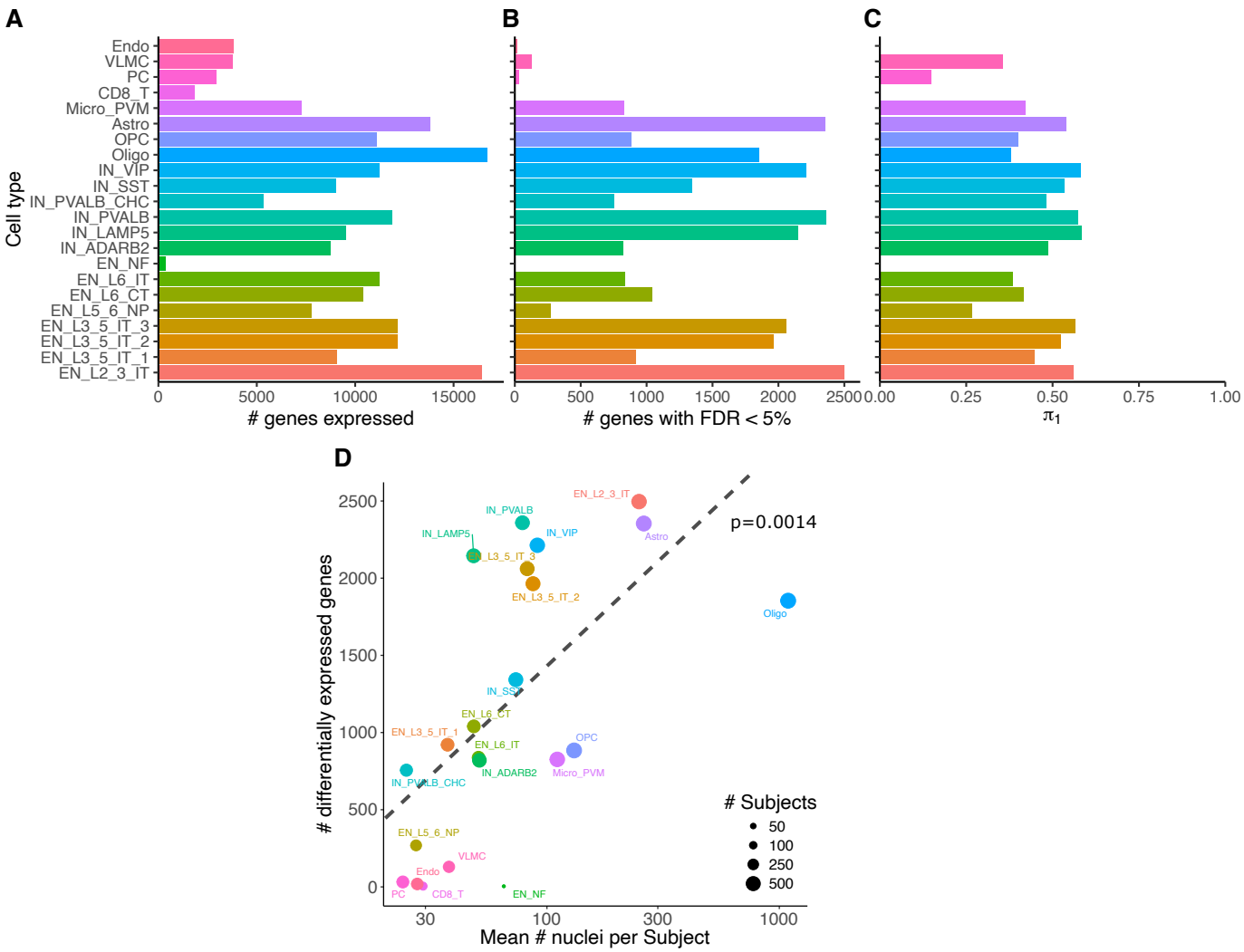
B



Supplemental Figure 7: Summary of Alzheimer's disease postmortem brain cohort A) Sample size stratified by sex and disease status **B)** Age distribution stratified by sex and disease status.



Supplemental Figure 8: Summary of batch effect across sample pools The number (left) and percentage (right) of expressed genes in each cell type where $> 5\%$ of variance is explained by variation across sample pools.



Supplemental Figure 9: Summary of Alzheimer's disease differential expression analysis. **A)** Number of genes passing expression cutoff for each cell type. **B)** Number of differentially expressed genes at 5% FDR. **C)** Storey's π_1 estimate of the fraction of genes rejecting the null hypothesis of no association. **D)** Number of differentially expressed genes for each cell cluster increases with the number of nuclei observed per subject. Circle size indicates the number of subjects with at least 10 nuclei observed for the cluster. P-value from linear regression is shown.

2 Supplementary Methods

For small sample sizes, parameter estimates can have high sampling variance. In seminal work, Smyth (Smyth, 2004) developed an empirical Bayes approach that borrows information across genes to estimate the residual variance. The widely used limma package (Ritchie et al., 2015) fits a linear model for each gene, performs the empirical Bayes step, and then computes a moderated t-statistic with a modified null distribution. In the case of a linear model, Smyth's empirical Bayes method uses a conjugate prior on the residual variances and assumes they are drawn from an inverse gamma distribution (i.e. precisions are drawn from a scaled chi-squared distribution) with parameters estimated from the data. A key value in this calculation is the residual degrees of freedom. In the case of a linear model with n samples and p covariates (including the intercept), the residual degrees of freedom (df_r) is $n-p$. However, the case of a linear mixed model used here is more complicated. In this case, we show that the residual variance estimates follow a distribution given by a weighted mixture of n chi-squared random variables, where the weights depend on both the data and the estimated model parameters. We match the expected value of this mixture distribution using a single chi-square and use its degrees of freedom to approximate the df_r of the linear mixed model. Importantly, this method is exact in the case of a linear model, is approximate for linear mixed models with any number of random effects, and the approximation improves with the sample size.

Residual degrees of freedom

In the case of the linear model with

$$\begin{aligned} y &= X\beta + \varepsilon & (1)\\ \varepsilon &\sim \mathcal{N}(0, \sigma_\varepsilon^2) & (2) \end{aligned}$$

n samples and p predictors (including intercept), the hat matrix $H = X(X^TX)^{-1}X^T$ transforms observed and fitted response values according to $\hat{y} = Hy$ (Hastie and Tibshirani, 1990). Similarly, the residuals are

$$\begin{aligned} r &= y - \hat{y} & (3)\\ &= y - Hy & (4)\\ &= (I - H)y. & (5) \end{aligned}$$

The residual sum of squares is therefore

$$r^Tr = y^T(I-H)^T(I-H)y. \quad (6)$$

For any vector of normally distributed values $z \sim \mathcal{N}(0,1)$ and a positive semi-definite matrix A, the quadratic form is distributed according to a mixture of chi-squared distributions weighted according to the eigen-values of A. Formally, $z^TAz \sim \sum_i \lambda_i \chi_1^2$, where λ_i values are the eigen-values of A.

Since the expected value of the sum of weighted χ^2 random variables is the sum of the weights, then $E[z^T Az] = \sum_i \lambda_i = \text{tr}(A)$.

For the linear model, $A = (I - H)^T(I - H)$ and has $n - p$ eigen-values with value 1, and the rest are 0. Therefore, $\text{tr}(A) = p$.

In general,

$$r^T r/\sigma_e^2 \sim \sum_i \lambda_i \chi_1^2 \tag{7}$$

where the weights λ_i are defined by H. The unbiased estimate of the residual variance is thus $E[\sigma_e^2] = r^T r/(n - \text{tr}(A))$.

In the case of a linear model, this reduces to the standard theory from linear modeling

$$r^T r/\sigma_e^2 \sim \chi_{n-p}^2. \tag{8}$$

For the linear mixed model, we can use a single χ^2 distribution to approximate the weighted mixture of χ^2. Setting the degrees of freedom to $\text{tr}(A)$ matches the mean of the mixture distribution, while being exact in the case of the linear model. The value of $\text{tr}(A)$ is computed based on

$$\begin{aligned} \text{tr}(A) &= \text{tr}\left[(I - H)^T(I - H)\right] &(9)\\ &= n - 2\text{tr}(H) + \text{tr}(HH^T). &(10) \end{aligned}$$

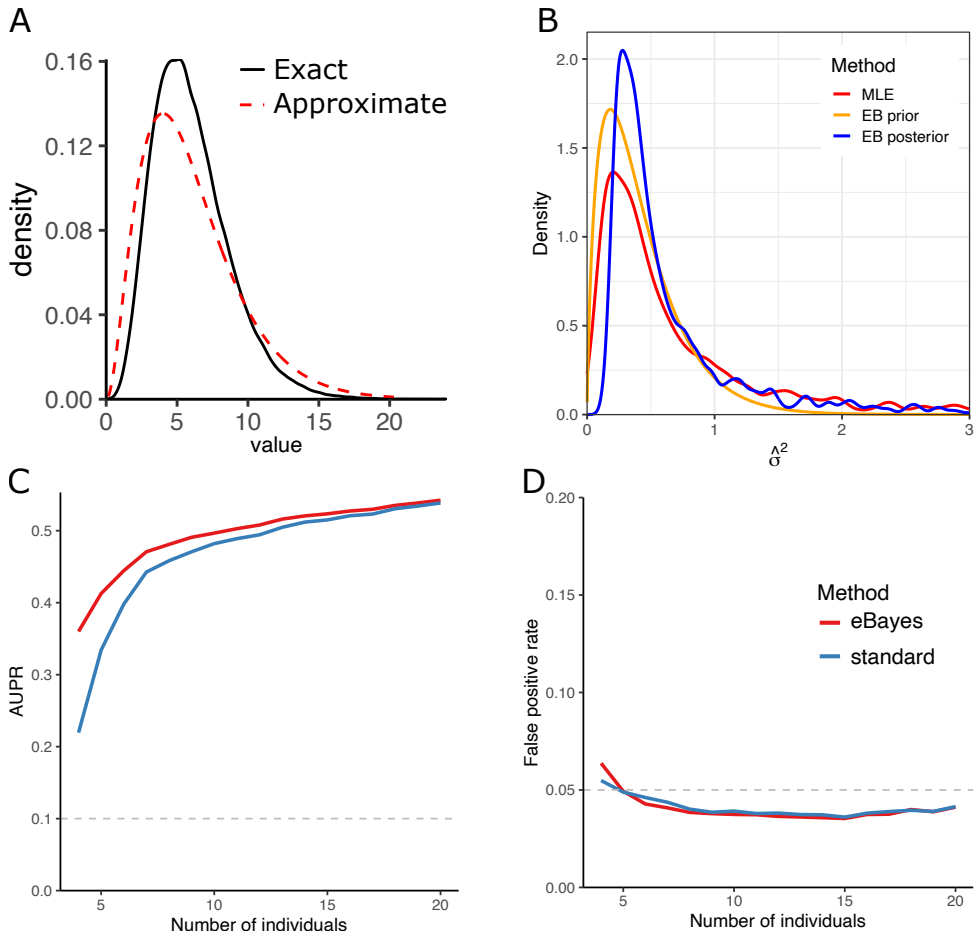
Mixture of chi-squares

Let x be a weighted mixture of k χ_1^2 values where $\lambda_i \in [0, 1]$ is the weight of the i^{th} component so that $x \sim \sum_{i=1}^k \lambda_i \chi_1^2$. This distribution has k parameters and an expectation $E[x] = \sum_{i=1}^k \lambda_i$. When all λ_i are 0 or 1, x is χ_ν^2 distributed with $\nu = \sum_{i=1}^k \lambda_i$. In this case, matching the mean of the mixture with a single chi-squared distribution is exact. For arbitrary non-negative values of λ_i, the distribution of x does not reduce to this simple form, but setting $\nu = \sum_{i=1}^k \lambda_i$ approximates the mixture distribution by matching its mean. We examine the chi-square approximation of the weighted sum of chi-square random variables for increasing values of k (**Supplementary Figure 10A**).

While Smyth uses the df_r for both the EB step and the null of the moderated t-statistic, here we used the approximate df_r just for EB. We use the Satterthwaite or KR method for the null distribution. The df_r of a linear mixed model fit by `lme4::lmer()` can be computed by `variancePartition::rdf.merMod()`.

Simulations

The EB step shrinks the MLE estimates toward the global mean to reduce the impact of genes with extremely high or low residual variance inflating the false positive rate (**Supplementary Figure 10B**). Simulations of $p = 20k$ genes, a discrete covariate with two categories modeled as a random effect, and varying sample sizes indicates that the EB post processing of results from a linear mixed models increases statistical performance (**Supplementary Figure 10C,D**).



Supplemental Figure 10: Empirical Bayes shrinkage. **A)** Approximation of *scaled* chisq distribution with a chisq distribution. Here, 100,000 values were sampled from a weighted mixture of k χ^2_1 variables with $\lambda_i = 0.5$ and the kernel density is plotted in black for 10. The χ^2_ν approximation matching the mean is shown in red. **B)** Empirical Bayes (EB) shrinkage of residual variance from simulated dataset of 20k genes with residual precision drawn from $\Gamma(4, 4)$. The EB posterior estimates (blue) shrink the MLE estimates (red) toward the global mean. **C)** Area under the precision-recall curve for simulated data for the standard method (blue) and EB shrinkage (red) for increasing sample size. **D)** False positive rate for these simulations.

References

Amezquita, R. A., Lun, A. T., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., Waldron, L., Pagès, H., Smith, M. L., Huber, W., Morgan, M., Gottardo, R., and Hicks, S. C. (2020). Orchestrating single-cell analysis with Bioconductor. *Nature Methods* **17,** 137–145.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall / CRC, New York.

Hoffman, G. E. and Roussos, P. (2021). Dream: Powerful differential expression analysis for repeated measures designs. *Bioinformatics* **37,** 192–201.

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,** R29.

Lun, A. T. L., Pagès, H., and Smith, M. L. (2018). beachmat: A Bioconductor C++ API for accessing high-throughput biological data from a variety of R matrix types. *PLOS Computational Biology* **14,** e1006135.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888–1902.e21.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19,** 15.

Zappia, L. and Lun, A. (2021). zellkonverter: Conversion Between scRNA-seq Objects. In *https://github.com/theislab/zellkonverter*.