Supplementary Methods and Figures
dream: Powerful differential expression analysis for repeated measures designs

Gabriel E. Hoffman[1,2,3*] and Panos Roussos [1,2,3,4,5]

[1]Pamela Sklar Division of Psychiatric Genomics
[2]Icahn Institute for Data Science and Genomic Technology
[3]Department of Genetics and Genomic Sciences
[4]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, USA
[5]Mental Illness Research, Education, and Clinical Center (VISN 2 South), James J. Peters VA Medical Center, Bronx, New York, USA

[*]Contact: gabriel.hoffman@mssm.edu

Contents

1 Supplementary Methods

1.1 Estimating of the approximate degrees of freedom for the hypothesis test

For the linear model and generalized least squares model described above, the degrees of freedom of the hypothesis test is fixed based on the number of covariates and the sample size:

$$d = N - p \tag{1}$$

where N is the number of samples, and p is the number of covariates. For the linear mixed model, we can explicitly account for the fact that estimating the random effect changes the degrees of freedom of the distribution used to approximate the null distribution (Hoffman, 2013; Kenward and Roger, 1997; Giesbrecht and Burns, 1985; Kuznetsova *et al.*, 2017; Halekoh and Højsgaard, 2014). Thus let $d_{g,C,L}$ be the degrees of freedom of the hypothesis test for gene g for the specified set of covariates, C, and contrast matrix, L. We omit the statistical details here, but $d_{g,C,L}$ can be estimated from the model fit using the very fast Satterthwaite approximation (Giesbrecht and Burns, 1985; Kuznetsova *et al.*, 2017) (the default in dream) or the more accurate but computationally demanding Kenward-Roger approximation (Kenward and Roger, 1997; Halekoh and Højsgaard, 2014) used by dream-KR.

1.2 Modeling measurement error in RNA-seq counts

The limma package models measurement error in the RNA-seq counts by estimating precision weights for each observation (Law *et al.*, 2014). The `voom()` function does this by fitting a smooth function to the square root residual standard deviation regressed on the $\log_2$ counts. However, `voom()` can only model variables as fixed effects, and so cannot consider within-individual variation in estimating the precision weights. In `voomWithDreamWeights()`, we apply an identical procedure except that the residuals are computed by fitting a linear mixed model specified by the user. A dream analysis can use precision weights computed by either `voom()` or `voomWithDreamWeights()`.

1.3 Software

The dream method is available in the `dream()` function in the variancePartition (Hoffman and Schadt, 2016) package at `http://bioconductor.org/packages/variancePartition` from Bioconductor version ≥ 3.7.

1.4 Implementation

Linear mixed models are estimated using the `lme4` package (Bates *et al.*, 2015). Estimating the residual degrees of freedom is performed with either Satterthwaite approximation (Giesbrecht and Burns, 1985) in the `lmerTest` package (Kuznetsova *et al.*, 2017) or the Kenward-

Roger approximation (Kenward and Roger, 1997) in the `pbkrtest` package (Halekoh and Højsgaard, 2014). Parallel processing of thousands of genes on a multi-core computer is performed with `BiocParallel` (Morgan *et al.*, 2019). A naive implementation would copy the entire dataset to each thread so that memory usage would increase linearly with the number of threads used. However, this is problematic for large datasets and prevents the user from taking advantage of multi-core machines. Instead, `iterators` (Ooi and Weston, 2019) is used to stream chunks of data to each thread so that memory usage is almost constant regardless of the number of threads. This dramatically reduce memory usage. Visualization is performed with `ggplot2` (Wickham, 2009).

1.5 Simulating RNA-seq count data

In order to reproduce the characteristics of real biological datasets as closely as possible, we simulated gene expression data with

- 4 sources of expression variation with simulated magnitude designed to mimic real data
  - variance across individuals
  - variance across two disease classes
  - variance across two batches to simulate a batch effect
  - residual variance
- negative binomial error variance
- 20,738 protein coding genes from GENCODE v19
- $\sim$ 57 million total counts per sample

1.5.1 Further details

The true expression values were simulated from a linear mixed model with 4 components: 1) variance across individuals; 2) variance across two disease classes (i.e. cases versus controls); 3) variance across two batches; and 4) residual variance. All samples from a given individual have the same disease status, and analysis considers the cross-individual test of differential expression between cases and controls. Samples are randomly assigned to one of two batches. Including a batch component here models the expression heterogeneity across technical batches, but also across two brain regions, or tissues types as is common in real data. For each gene, the variance fractions for these 4 components were sampled from a beta distribution with parameters set to give a specified mean and variance described below. The residual variance was set so that the variance fractions summed to 1. If the randomly draw values sum to more than one, the residual variance is set to 0.05 and the other components are scaled accordingly. The variance component values were based on examining the variance fractions estimated with variancePartition across many datasets (Hoffman and Schadt, 2016; Hoffman *et al.*, 2017; Carcamo-Orive *et al.*, 2017; Girdhar *et al.*, 2018). In each simulation, 500 genes were randomly selected to be differentially expressed between cases and controls, and for all other genes the disease component was set to zero.

At the implementation level, simulating expression data is divided into three steps: 1) Simulate relative expression values, 2) convert these into multiplicative fold changes compared to a baseline, 3) simulate RNA-seq counts from negative binomial model given the expected multiplicative fold change values.

Let y_j be the vector of *relative* expression values for gene j across all individuals that varies from $-\infty$ to ∞. Let η_k correspond to the k^{th} component where $k \in$ (Individual, Disease, Batch, Noise) so that, for example, $\eta_{\text{Individual}}$ represents the vector of expected expression attributable to variance across individuals. Since each variable represented by k has a different number of levels, let $\tilde{\eta}$ be the standardized version of η with a mean of 0 and variance 1. The expression value of gene j is simulated according to

$$y_j \quad = \quad \tilde{\eta}_{\text{Individual}} + \tilde{\eta}_{\text{Disease}} + \tilde{\eta}_{\text{Batch}} + \tilde{\eta}_{\text{Noise}} \tag{2}$$

where each variance component η_k is drawn according to

$$\eta_k \quad = \quad X_k \beta_k \tag{3}$$

where X_k is the matrix of ANOVA coded indicator values for variable k. The number of columns in X_k equals the number of categories in variable k so that Disease has two categories, Batch has 2 categories and Individual has N categories. The vector β_k gives the expected value for each of the categories in variable k. These expected values are drawn from a normal distribution according to

$$\beta_k \quad = \quad \mathcal{N}(0, \sigma_k^2). \tag{4}$$

Since the expression values are simulated based on standardized variance components, σ_k^2 corresponds to the fraction of variance attributable to component k. Since fractions naturally fall between 0 and 1, they can be drawn from a beta distribution. The beta distribution is usually parameterized in terms of two shape parameters, α and β, but here we parameterize it with a mean and variance by matching the moments of the distribution. The simulated mean for each component corresponds to the expected variance fractions and were motivated based on variancePartition across many datasets (Hoffman and Schadt, 2016; Hoffman *et al.*, 2017; Carcamo-Orive *et al.*, 2017; Girdhar *et al.*, 2018). The simulated variance fractions are draw according to

$$\sigma^2_{\text{Individual}} \quad = \quad \text{Beta}(\text{mean} = 0.45, \text{var} = 0.03) \tag{5}$$

$$\sigma^2_{\text{Batch}} \quad = \quad \text{Beta}(\text{mean} = 0.20, \text{var} = 0.01). \tag{6}$$

For the 500 genes where disease has an effect on expression,

$$\sigma^2_{\text{Disease}} \quad = \quad \text{Beta}(\text{mean} = 0.30, \text{var} = 0.005), \tag{7}$$

and the value of $\sigma^2_{\text{Disease}}$ is subtracted from $\sigma^2_{\text{Individual}}$, since all samples from the same individual have the same disease status. For all other genes, $\sigma^2_{\text{Disease}} = 0$.

The random noise component is drawn so that the variance fractions sum to 1, but is set to a minimum of 0.05:

$$\sigma^2_{\text{Noise}} = \max\left(1 - \sigma^2_{\text{Individual}} - \sigma^2_{\text{Disease}} - \sigma^2_{\text{Batch}}, 0.05\right). \tag{8}$$

Given, y_j, the vector of relative expression values for gene j, convert them into multiplicative fold change values with a minimum of 1 according to

$$FC_j = y_j/2 - \min(y_j/2) + 1. \tag{9}$$

These simulated multiplicative fold change values are passed to polyester v1.14.0 (Frazee *et al.*, 2015) to produce counts with biologically realistic negative binomial error. The number of reads per gene is selected based on gene length using the 'mean model' from polyester. Otherwise, polyester defaults were used.

Expression values were simulated following above procedure for 20,738 protein coding genes from GENCODE v19. Approximately 57 million total reads counts were generated for each sample (mean: 57.3M, median: 55.0M, sd: 24.4M). Simulations using a range of values for each of these parameters did not change the conclusions.

1.6 Performance metrics

In each simulation described above, 500 genes were simulated with non-zero coefficients. These genes were considered as 'positives', and all remaining genes were considered as 'negatives'. All performance metrics follow from standard definitions based on a p-value cutoff C:

False positives (FP): genes in the 'negative' class having a p-value $< C$

True positives (TP): genes in the 'positive' class having a p-value $< C$

False negatives (FN): genes in the 'positive' class having a p-value $> C$

True negatives (TN): genes in the 'negative' class having a p-value $> C$

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP+FP} \\ \text{Recall} &= \frac{TP}{TP+FN} \end{aligned}$$

Precision-Recall (PR) curves and area under the PR curve (AUPR) were implemented using the `PRROC` package (Grau *et al.*, 2015).

Confidence intervals were computed using the asymptotic approximation of the binomial model implemented in R package `binom` (Dorai-Raj, 2015). Letting $\hat{p}$ be the estimated proportion, the confidence interval is $\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where z is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution and $\alpha = 0.05$. Since AUPR can be interpreted as a binomial proportion Boyd *et al.* (2013), confidence intervals for AUPR were computed the same way.

1.7 Null simulations using real RNA-seq data

Since simulated counts cannot fully reproduce the biological, technical and random variability of real data, we used counts from 317 RNA-seq samples of induced pluripotent stem cells from 101 individuals available from GEO at GSE79636 (Carcamo-Orive *et al.*, 2017). Subsets of the data were generated using $N \in (5, 10, 20, 40)$ individuals and $R \in (2, 3)$ replicates per individual. In each simulation, the first $\sqrt{N} - 1$ principal components of the gene expression were included as covariates to account for batch effects (Leek *et al.*, 2010). A continuous variable to be the focus of the differential expression analysis was simulated for each sample. For this purpose, a normally distributed variable independent of the gene expression data was simulated with 99% of the variance across individuals and 1% of the variance within individuals. (We note that simulating binary values for this variable gives similar results.) Since this phenotype is independent of the gene expression, we can evaluate the control of type I error from the 12 differential expression analysis. If the hypothesis tests for each gene are statistically independent, then an accurate statistical test will give p-values that are uniformly distributed under the null. We note that in real data, co-expression between genes (Langfelder and Horvath, 2008) can cause a deviation from uniformity even under the null.

For each (N, R) pair, 5 independent simulation were performed each using $14,634$ expressed genes. The false positive rate was evaluated as the fraction of genes with $p < 0.05$. The relationship between the false positive rate and expression magnitude was evaluated by fitting a logistic regression model where the response is the binary variable indicating if the gene is a false positive at $p < 0.05$ and the predictor is the $\log_2$ counts per million.

1.8 Data analysis

Data for Timothy syndrome was downloaded from GEO at GSE25542 (Pasca *et al.*, 2011). Data for childhood onset schizophrenia was downloaded from `https://www.synapse.org/#!Synapse:syn9907463` (Hoffman *et al.*, 2017). Post mortem brain RNA-seq data from Alzheimer's and controls was downloaded from `https://www.synapse.org/#!Synapse:syn3159438` (Wang *et al.*, 2018). Analysis was performed on individuals from European ancestry that were assayed in each of 4 brain regions (Brodmann areas 10, 22, 36 and 4), had ApoE genotype data, had Braak stage information, and were either controls or definite AD patients (i.e. possible and probable cases were excluded). Differential expression analysis corrected for batch, sex, RIN, rRNA rate, post mortem interval, mapping rate and ApoE genotype.

Data from GENESIPS was obtained from GEO at GSE79636 (Carcamo-Orive *et al.*, 2017). Data from Warren *et al.* (2017) was obtained from GEO at GSE90749. Data from Mariani *et al.* (2015) was obtained from recount2 (Collado-Torres *et al.*, 2017) at SRP047194. Enrichment analysis was performed with cameraPR (Wu and Smyth, 2012) in the limma package (Ritchie *et al.*, 2015). In order to avoid using arbitrary cutoffs to identify differentially expressed genes, gene set enrichments were evaluated by applying cameraPR to the differential expression test statistics from each analysis. The fraction of expression variation explainable by cis regulatory variants was obtained from Gamazon *et al.* (2015) and Huckins *et al.* (2019).

Code for simulations and reproducible analysis, figures, and statistics from differential expression and enrichment analyses are available at `https://github.com/GabrielHoffman/dream_analysis`.

1.9 A note on shrinkage by combining gene-level results

In functional genomics, shrinkage/regularization is widely used to borrow information across multiple genes and Smyth's `limma` is a prime example. The `eBayes` function in `limma` uses an empirical Bayes approach to shrink the observed residual variances towards a common value. Note that this approach pushes the gene-level value *towards* a common value, instead of assigning a single summary value to all genes. This has worked extremely well in practice.

The `duplicateCorrelation` function in `limma` estimates the contribution of replicates by estimating a variance component for each gene, τ_g^2, and then summarizing the gene-level values with a single genome-wide value, τ^2. Setting a single genome-wide value is a very strong shrinkage, and with sufficient sample size using the estimated gene-level value (as `dream` does) performs improves performance.

This raises two related questions: 1) Can the gene-level variance component estimate, τ_g^2, be shrunk towards a central value in a compromise between using the estimated gene-level value and a single genome-wide summary? 2) Can Smyth's empirical Bayes shrinkage of the residual variance be applied to the linear mixed model?

Yu *et al.* (2019) has recently developed a framework to combine shrinkage of both the variance component and the residual variance term in order to create a new moderated t-statistic. Yu *et al.* (2019) refers to this approach as fully moderated t-statistics (FMT). For each gene, the shrunken variance component and shrunken residual variance are then combined to approximate the degrees of freedom of the t-statistic under the null.

The FMT methods did not accurately control the false positive rate in our simulations, and no shrinkage is used by dream with default settings or with the KR method. However, the FMT method is available for the sake of reproducibility in the function `variancePartition:::eBayesFMT()` which can be run on the result of `dream` just as `eBayes` is used by `limma`. This function is adapted from the code kindly provided provided by Yu *et al.* (2019).

We note that an earlier version (before v1.15.5 from September 5, 2019) of `dream` in the `variancePartition` package included an empirical Bayes step. However, after additional testing, we removed this step so that only gene-level estimates are used in the linear mixed model. We note that the elevated false positive rate that Yu *et al.* (2019) observed in simulations using `dream` are due this this issue and has been resolved.

Running the `eBayes` function on a linear mixed model fit with `dream` in version ≥ v1.15.5 gives the following warning:

```
Warning message:
In eBayes(fit) :
  Empircal Bayes moderated test is no longer supported for dream analysis
Returning original results for use downstream
```

2 Supplementary Figures

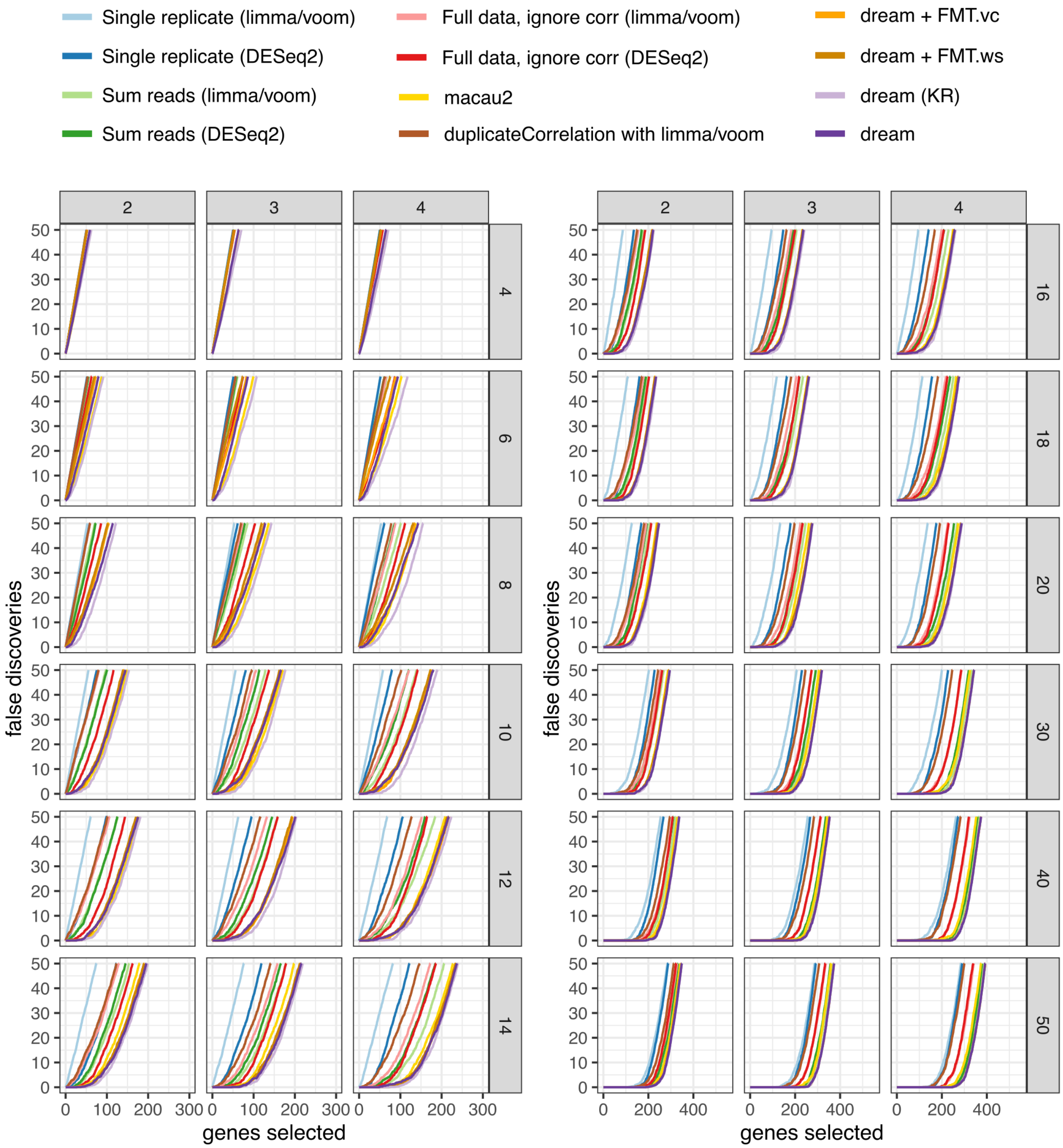

Figure S 1: **False discovery rates for multiple simulation conditions.** False discoveries plotted against the number of genes called differentially expressed by each method. Results are shown for between 4 and 50 individuals (rows) and 2 to 4 replicates (columns). For each combination, 50 simulated datasets were analyzed.

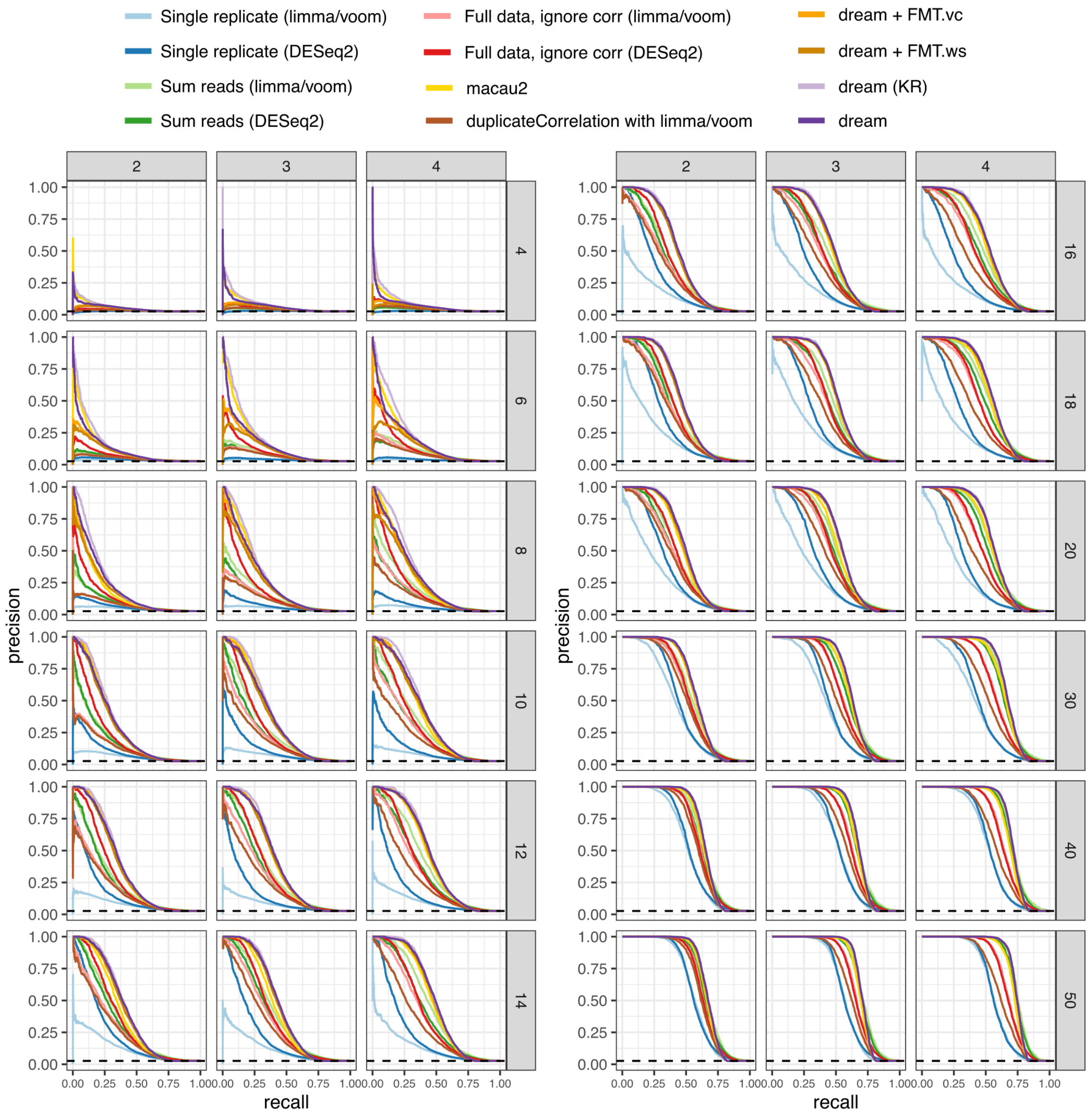

Figure S 2: **Precision-recall curves for multiple simulation conditions.** Plots shows performance in identifying true differentially expressed genes. Dashed lines indicate performance of a random classifier. Results are shown for between 4 and 50 individuals (rows) and 2 to 4 replicates (columns). For each combination, 50 simulated datasets were analyzed.

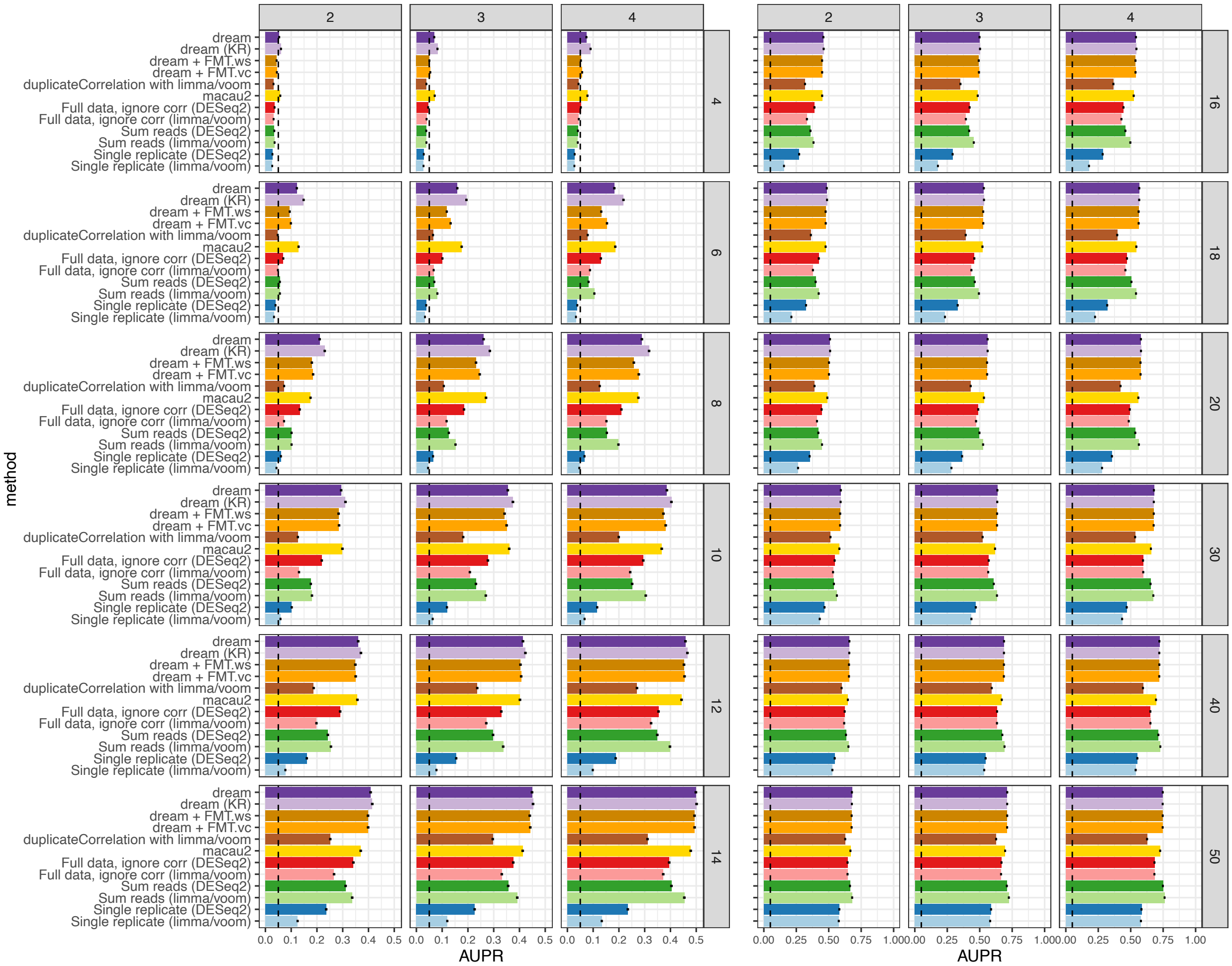

Figure S 3: **Area under the precision-recall (AUPR) for multiple simulation conditions.** Dashed line indicates AUPR of a random classifier. Error bars indicate 95% confidence intervals. Results are shown for between 4 and 50 individuals (rows) and 2 to 4 replicates (columns). For each combination, 50 simulated datasets were analyzed.

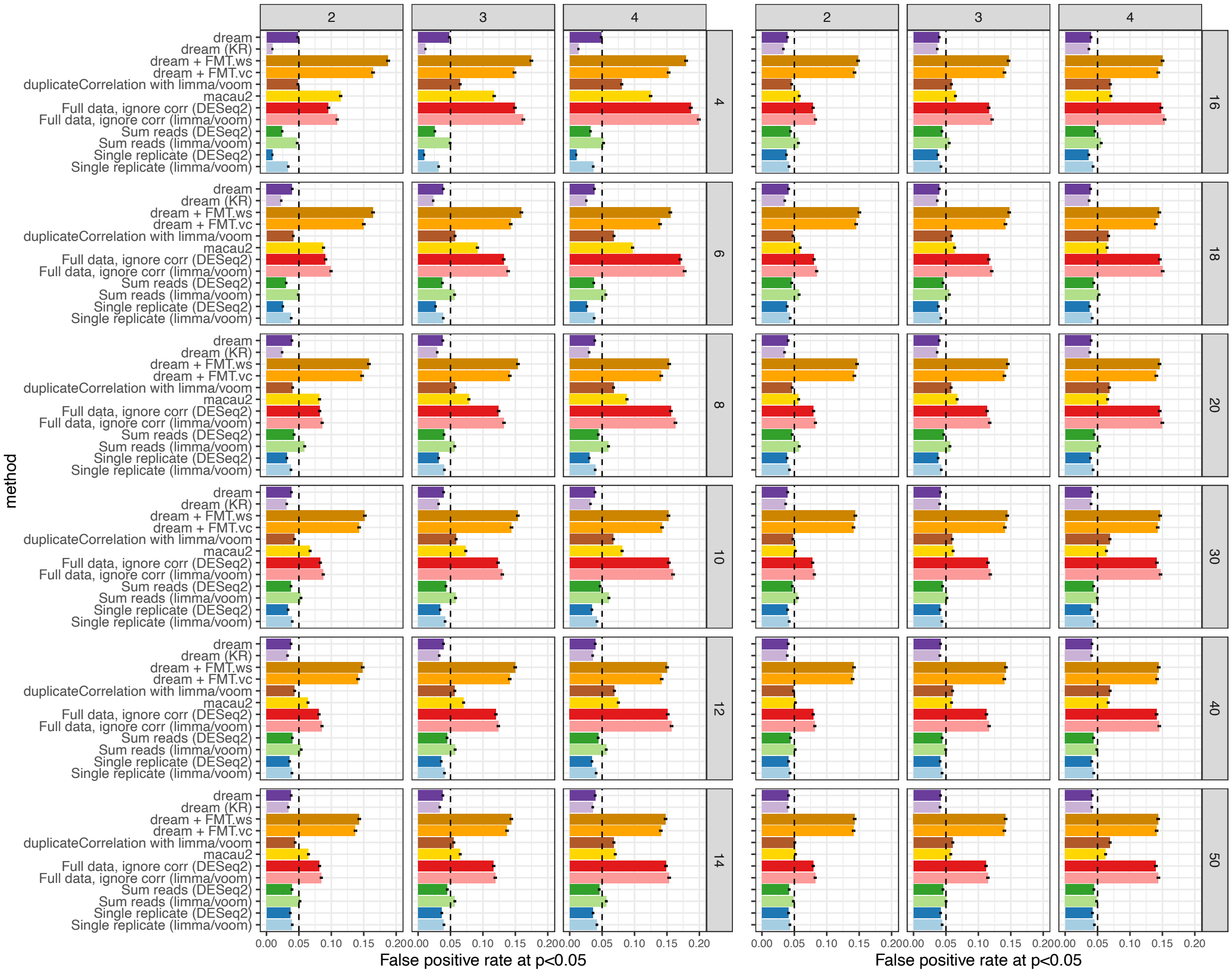

Figure S 4: **False positive rate for multiple simulation conditions.** False positive rate at p < 0.05 evaluated under a null model were no genes are differentially expressed illustrates calibration of type I error from each method. As indicated by the dashed line, a well calibrated method should give p-values < 0.05 for 5% of tests under a null model. Results are shown for number of individuals between 4 and 50 (rows), and replicates between 2 and 4 (columns).

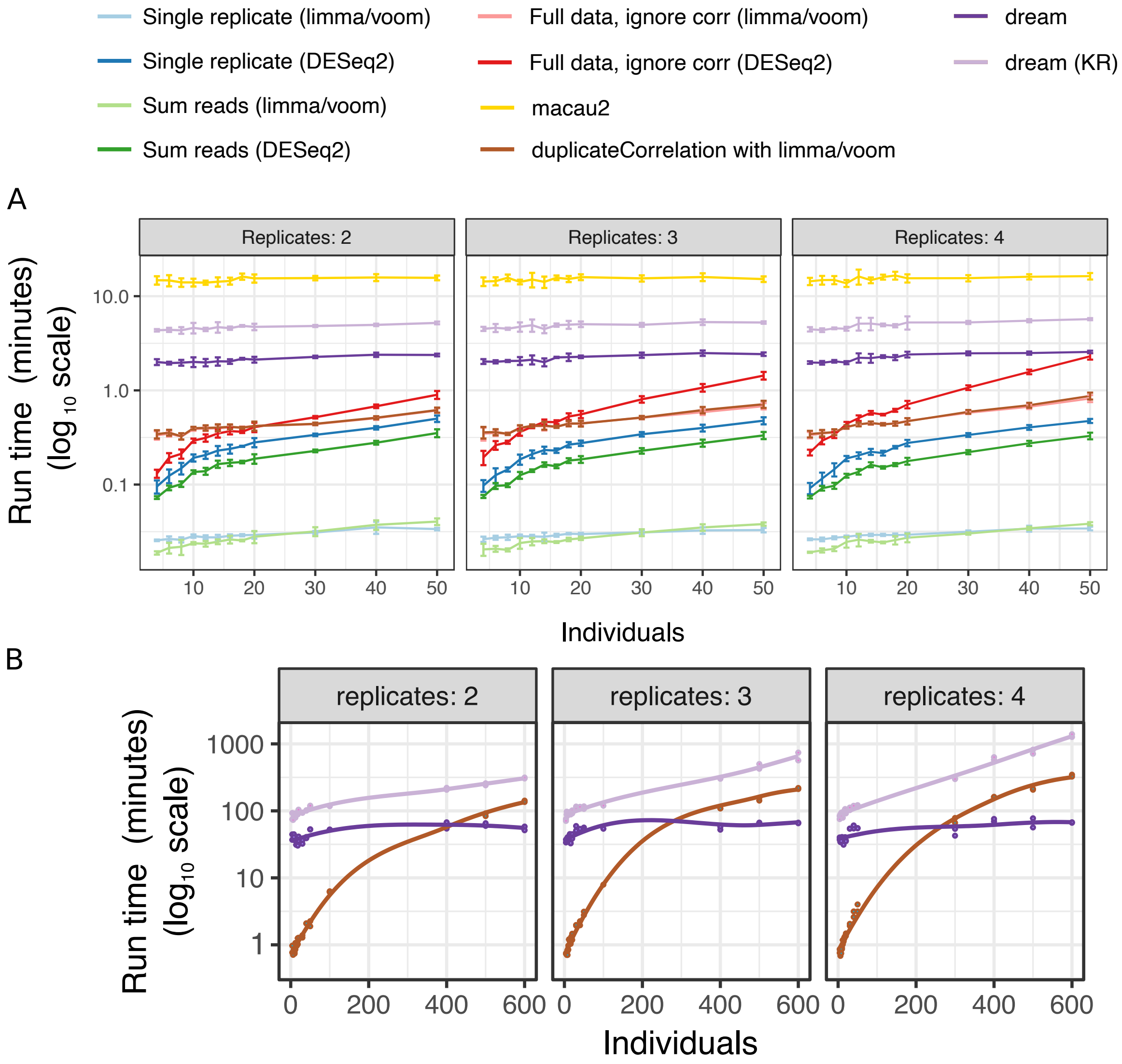

Figure S 5: **Run time comparison.** Run time for was evaluated on the simulated datasets on a 12 core Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz. **A**) Run time in minutes for each method in the simulation study presented in the previous figures. Averages and standard errors are shown. Analysis was performed using 5 cores. **B**) Run time in minutes for 3 methods on larger simulated datasets using 12 cores. Each combination of individuals, replicates, methods and threads was evaluated on 2 simulated datasets. Lines show loess smoothing. The formula used was: ∼ `Disease + (1|Individual)`.

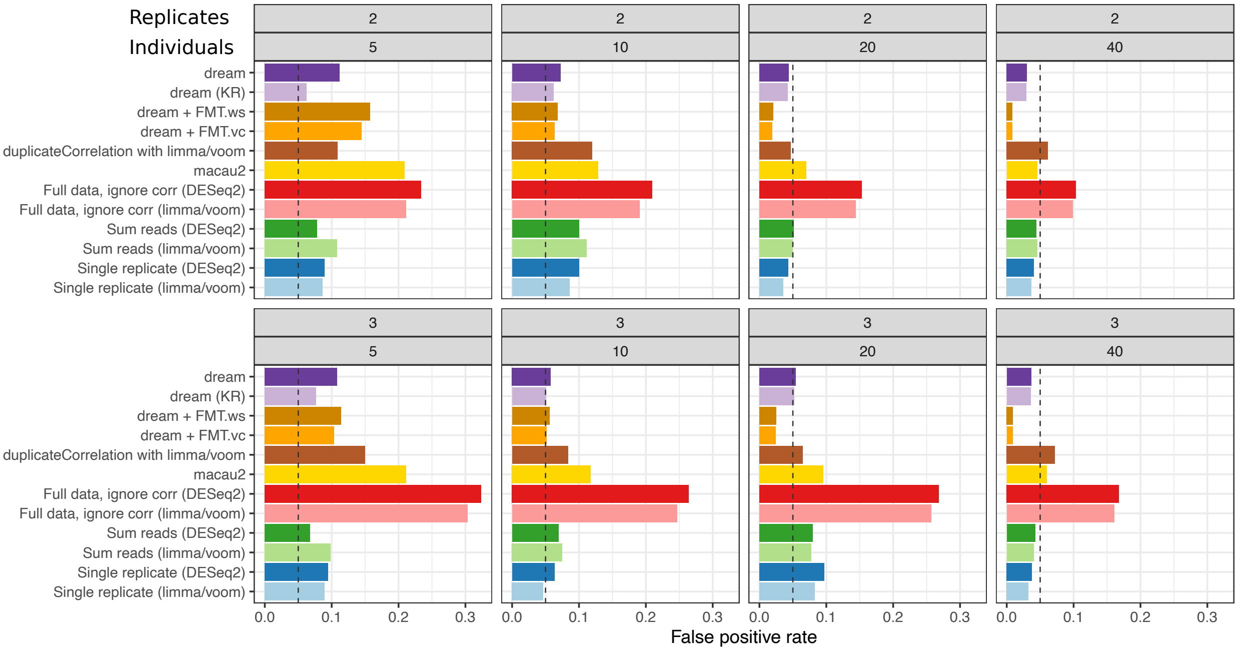

Figure S 6: **False positive rates from null simulations using real data.** False positive rate at p < 0.05 evaluated under a null model were no genes are differentially expressed based on real RNA-seq data from (Carcamo-Orive *et al.*, 2017). The dashed lines indicate the target 5% false positive rate.

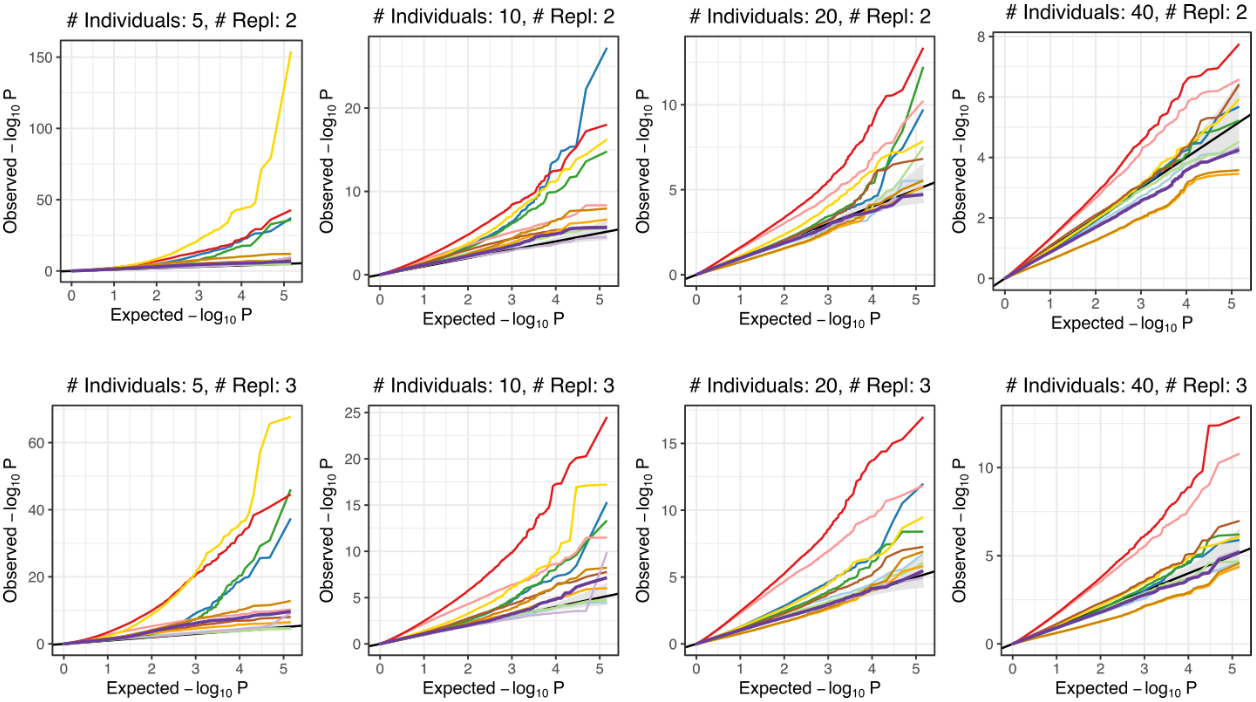

Figure S 7: **QQ plots from null simulations using real data.** QQ plots under a null model were no genes are differentially expressed based on real RNA-seq data from (Carcamo-Orive *et al.*, 2017). Colors as same as in previous figure.

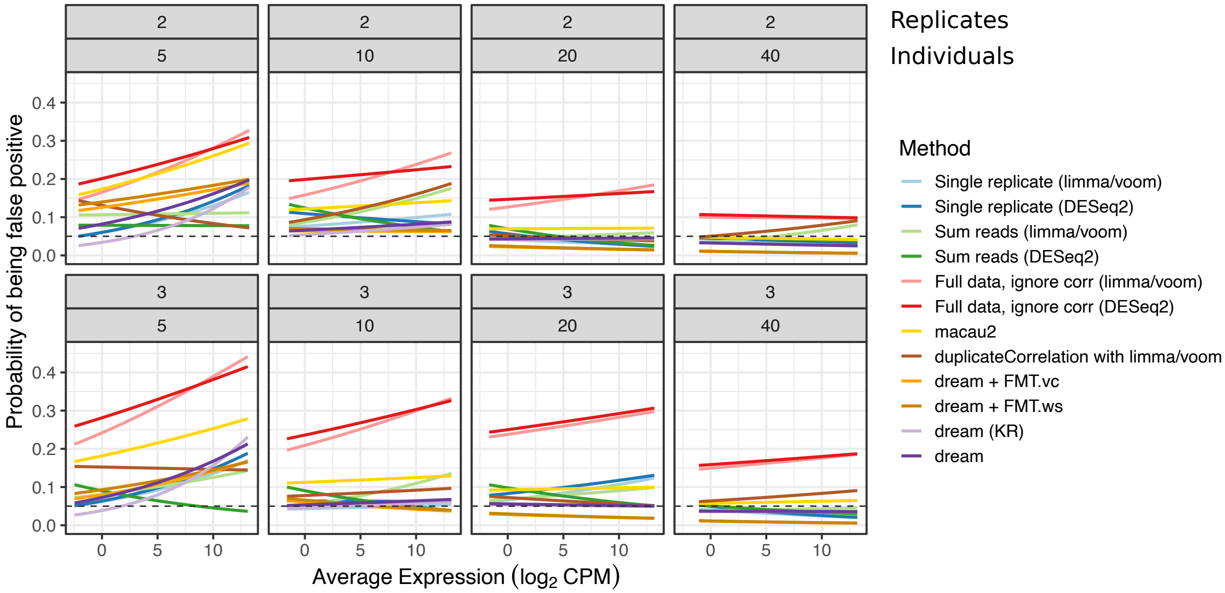

Figure S 8: **Effect of expression magnitude on false positive rate**. The relationship between expression and the probability of being a false positive by each of 12 methods was evaluated from null simulations based on real RNA-seq data from (Carcamo-Orive *et al.*, 2017). For each simulation condition and each differential expression method, the probability of each gene being a false positive was modeled as a logistic function of the the expression magnitude. Curves show fit of logistic regression. The dashed lines indicate the target 5% false positive rate.

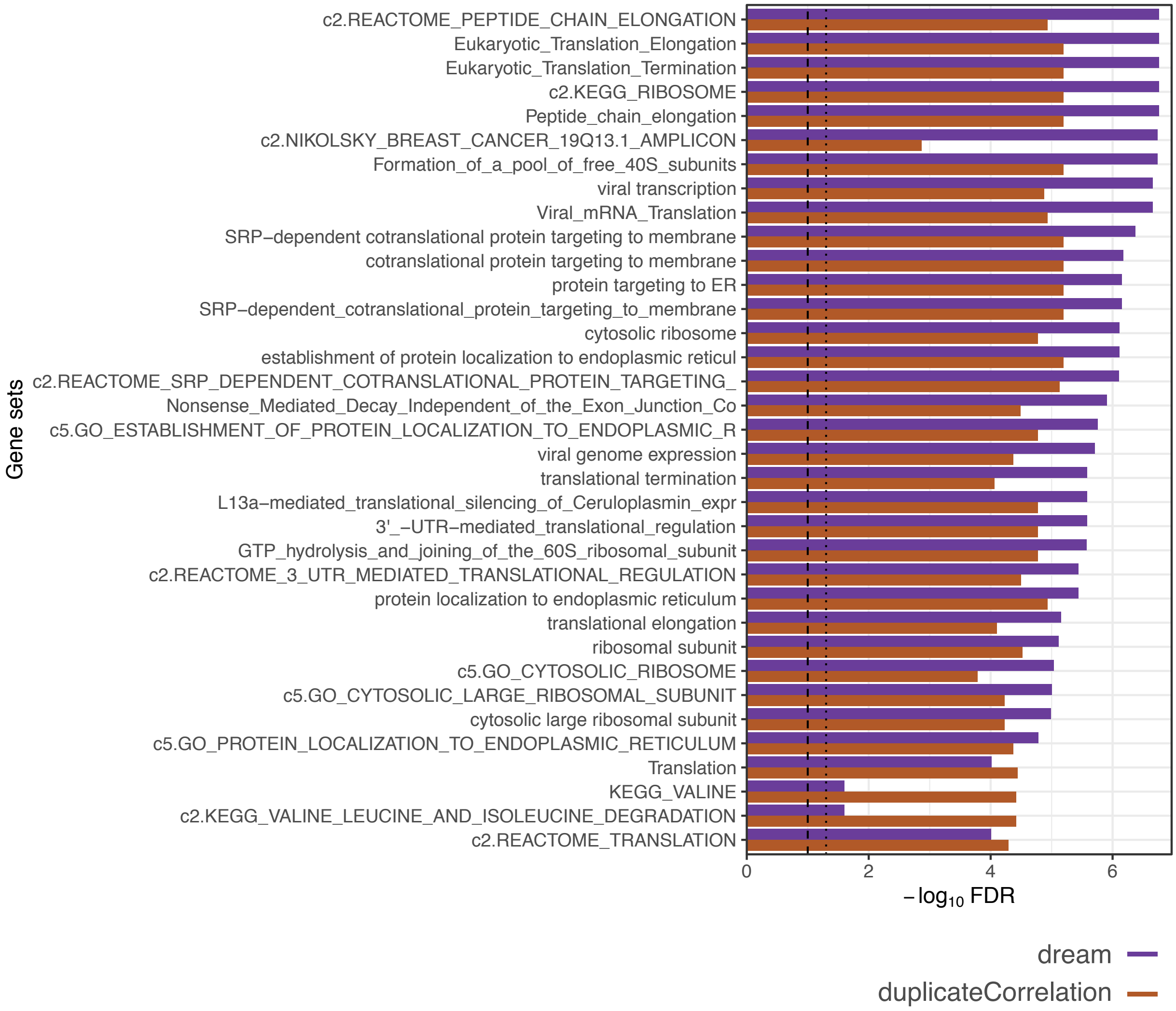

Figure S 9: **Gene set enrichment FDR for top 30 genesets from differential expression analysis of Braak stage.** Enrichment FDRs were computed using t-statistics from dream and duplicateCorrelation.

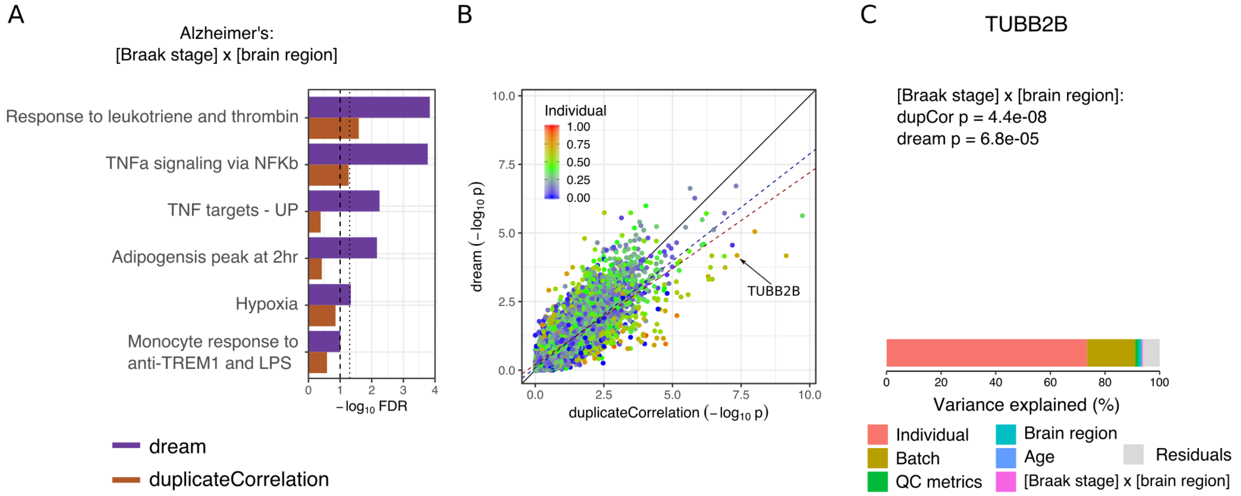

Figure S 10: **Application to transcriptome data from Alzheimer's disease. A**) Gene set enrichment FDR for genes associated with test of Braak stage by brain region term with 4 coefficients. Results are shown for dream and duplicateCorrelation. Lines with broad and narrow dashes indicate 10% and 5% FDR cutoff, respectively. **B**) Comparison of $-\log_{10}$ p-values from applying dream and duplicateCorrelation. Each point is a gene, and is colored by the fraction of expression variation explained by variance across individuals. Black solid line indicates a slope of 1. Dashed line indicates the best fit line for the 20% of genes with the highest (red) and lowest (blue) expression variation explained by variance across individuals. **C**) Results for TUBB2B. Box plot is omitted because it is identical to Figure 3C. Bar plot of variance decomposition for TUBB2B shows that 73.4% of variance is explained by expression variance across individuals. Since this value is much larger than the genome-wide mean, duplicateCorrelation under-corrects for the repeated measures.

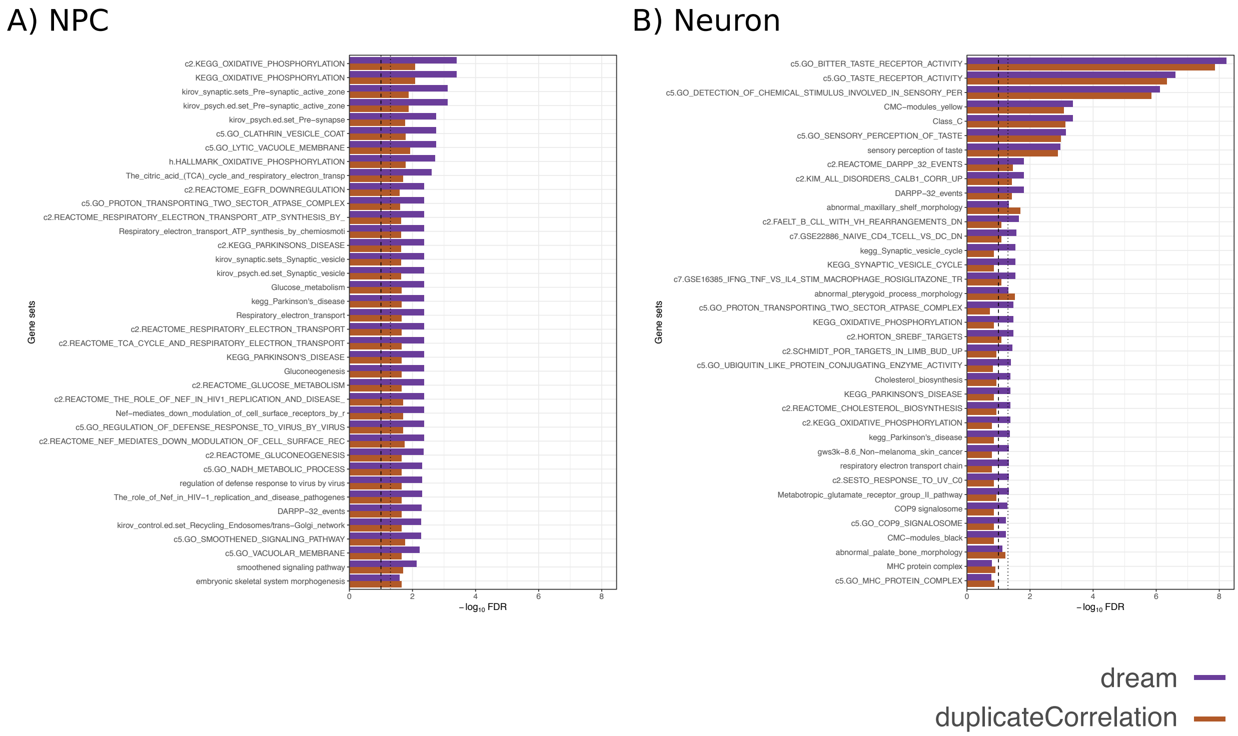

Figure S 11: **Gene set enrichment FDR for top 30 genesets from differential expression analysis of childhood onset schizophrenia.** Enrichment FDRs were computed using t-statistics from dream and duplicateCorrelation analysis of iPSC-derived **A**) neural progenitor cells (NPCs) and **B**) neurons.

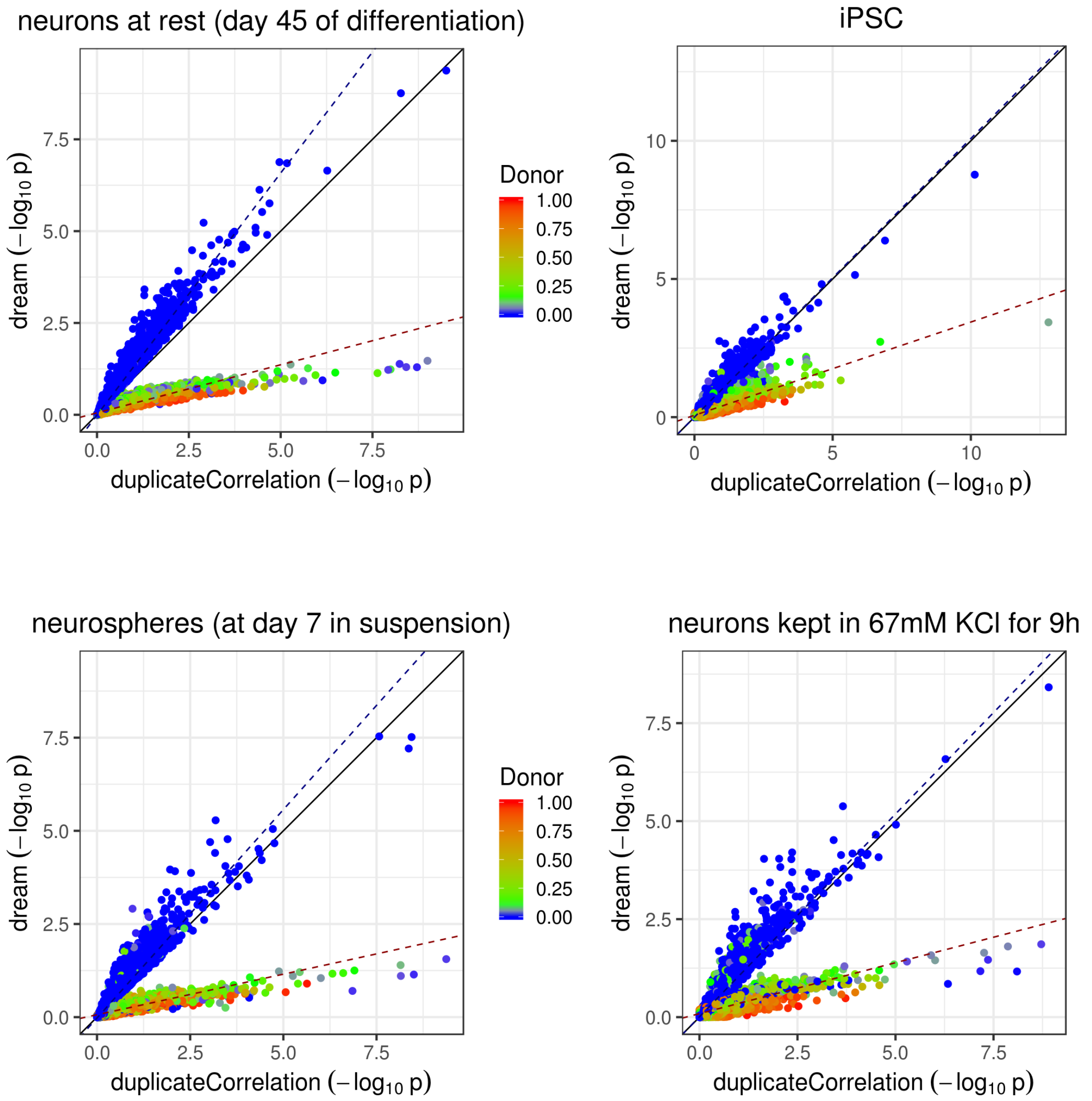

Figure S 12: **Differential expression analysis of Timothy syndrome compared to controls in four cell types or conditions.** Comparison of $-\log_{10}$ p-values from applying dream and duplicateCorrelation analyze case/control differences. Each point is a gene, and is colored by the fraction of expression variation explained by variance across individuals. Black solid line indicates a slope of 1. Dashed line indicates the best fit line for the 20% of genes with the highest (red) and lowest (blue) expression variation explained by variance across individuals.

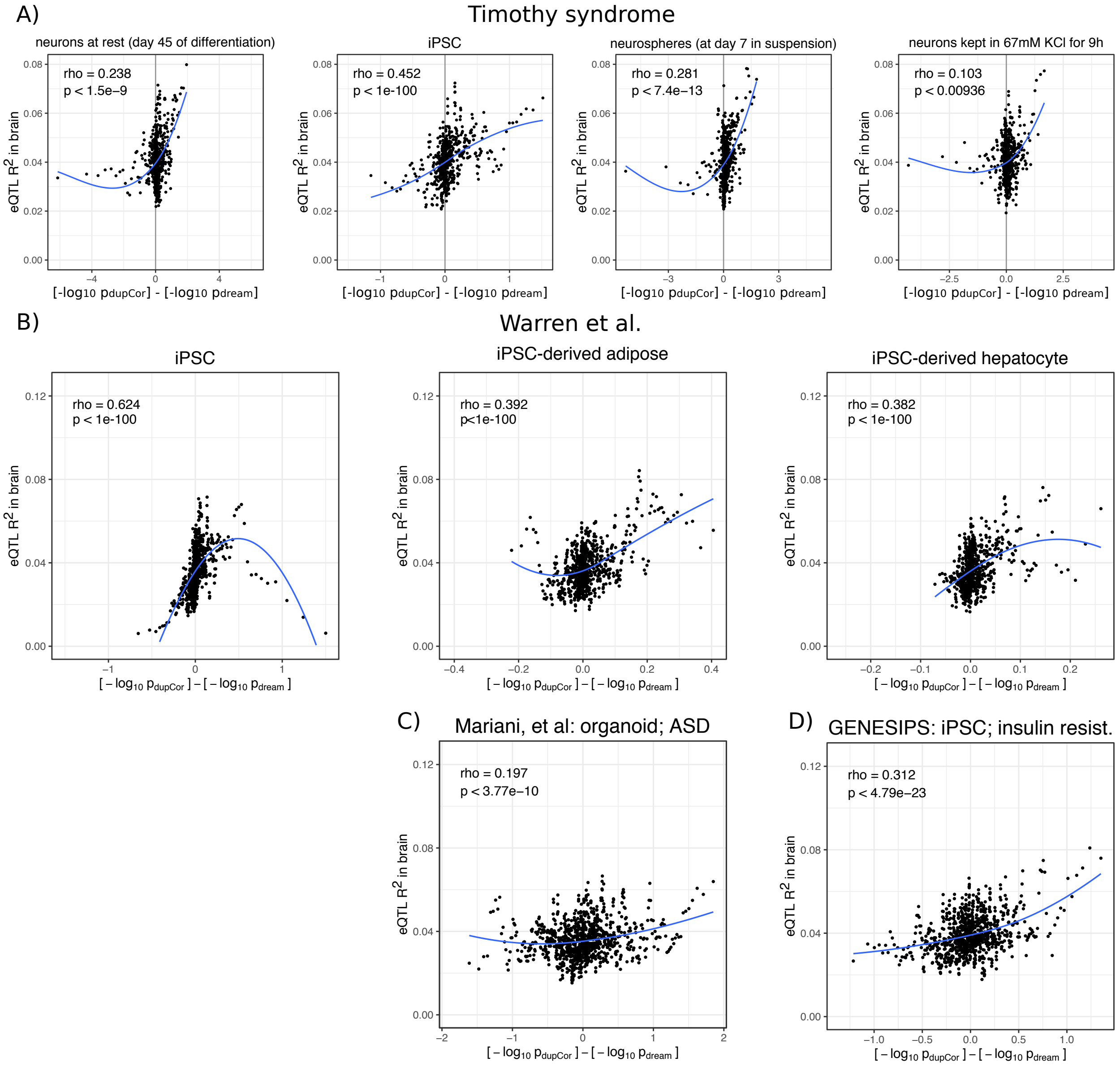

Figure S 13: **Relationship between differential expression results and genetic regulation.** For each gene the fraction of expression variation explainable by cis-eQTLs is compared to the difference in $-\log_{10}$ p-value from duplicateCorrelation and dream differential expression analysis. Due to the large number of genes, a sliding window analysis of 100 genes with an overlap of 20 was used to summarize the results. For each window, the average fraction of expression variation explainable by cis-eQTLs (i.e. eQTL R^2) in brains from the CommonMind Consortium (Fromer *et al.*, 2016) and average difference in $-\log_{10}$ p-values from the two methods are reported when differential expression analysis is performed on **A**) Timothy Syndrome in 4 cell types from Pasca *et al.* (2011) **B**) the SNP rs12740374 in 3 cell types from Warren *et al.* (2017), **C**) Autism Spectrum in organoids from Mariani *et al.* (2015), and **D**) insulin resistance in iPSC from Carcamo-Orive *et al.* (2017). Spearman rho correlations and p-values are shown along with loess curve.

References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, **67**(1).

Boyd, K., Eng, K. H., and Page, C. D. (2013). Area under the precision-recall curve: Point estimates and confidence intervals. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 451–466. Springer-Verlag Berlin Heidelberg.

Carcamo-Orive, I., Hoffman, G. E., Cundiff, P., Beckmann, N. D., D'Souza, S. L., Knowles, J. W., Patel, A., Papatsenko, D., Abbasi, F., Reaven, G. M., Whalen, S., Lee, P., Shahbazi, M., Henrion, M. Y., Zhu, K., Wang, S., Roussos, P., Schadt, E. E., Pandey, G., Chang, R., Quertermous, T., and Lemischka, I. (2017). Analysis of Transcriptional Variability in a Large Human iPSC Library Reveals Genetic and Non-genetic Determinants of Heterogeneity. *Cell Stem Cell*, **20**(4), 518–532.e9.

Collado-Torres, L., Nellore, A., Kammers, K., Ellis, S. E., Taub, M. A., Hansen, K. D., Jaffe, A. E., Langmead, B., and Leek, J. T. (2017). Reproducible RNA-seq analysis using recount2. *Nature Biotechnology*, **35**(4), 319–321.

Dorai-Raj, S. (2015). Binomial confidence intervals for several parameterizations. *R package version 1.1-1 https://cran.r-project.org/package=binom*.

Frazee, A. C., Jaffe, A. E., Langmead, B., and Leek, J. T. (2015). Polyester: Simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, **31**(17), 2778–2784.

Fromer, M., Roussos, P., Sieberts, S. K., Johnson, J. S., Kavanagh, D. H., Perumal, T. M., Ruderfer, D. M., Oh, E. C., Topol, A., Shah, H. R., Klei, L. L., Kramer, R., Pinto, D., Gümüş, Z. H., Cicek, A. E., Dang, K. K., Browne, A., Lu, C., Xie, L., Readhead, B., Stahl, E. A., Xiao, J., Parvizi, M., Hamamsy, T., Fullard, J. F., Wang, Y.-C., Mahajan, M. C., Derry, J. M. J., Dudley, J. T., Hemby, S. E., Logsdon, B. A., Talbot, K., Raj, T., Bennett, D. A., De Jager, P. L., Zhu, J., Zhang, B., Sullivan, P. F., Chess, A., Purcell, S. M., Shinobu, L. A., Mangravite, L. M., Toyoshiba, H., Gur, R. E., Hahn, C.-G., Lewis, D. A., Haroutunian, V., Peters, M. A., Lipska, B. K., Buxbaum, J. D., Schadt, E. E., Hirai, K., Roeder, K., Brennand, K. J., Katsanis, N., Domenici, E., Devlin, B., and Sklar, P. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience*, **19**(11), 1442–1453.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., and Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, **47**(9), 1091–1098.

Giesbrecht, F. G. and Burns, J. C. (1985). Two-Stage Analysis Based on a Mixed Model: Large-Sample Asymptotic Theory and Small-Sample Simulation Results. *Biometrics*, **41**(2), 477.

Girdhar, K., Hoffman, G. E., Jiang, Y., Brown, L., Kundakovic, M., Hauberg, M. E., Francoeur, N. J., Wang, Y.-c., Shah, H., Kavanagh, D. H., Zharovsky, E., Jacobov, R., Wiseman, J. R., Park, R., Johnson, J. S., Kassim, B. S., Sloofman, L., Mattei, E., Weng, Z., Sieberts, S. K., Peters, M. A., Harris, B. T., Lipska, B. K., Sklar, P., Roussos, P., and Akbarian, S. (2018). Cell-specific histone modification maps in the human frontal lobe link schizophrenia risk to the neuronal epigenome. *Nature Neuroscience*, **21**(8), 1126–1136.

Grau, J., Grosse, I., and Keilwagen, J. (2015). PRROC: Computing and visualizing Precision-recall and receiver operating characteristic curves in R. *Bioinformatics*, **31**(15), 2595–2597.

Halekoh, U. and Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models - The R Package pbkrtest. *Journal of Statistical Software*, **59**(9), 3–4.

Hoffman, G. E. (2013). Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. *PLoS ONE*, **8**(10), e75707.

Hoffman, G. E. and Schadt, E. E. (2016). variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics*, **17**(1), 483.

Hoffman, G. E., Hartley, B. J., Flaherty, E., Ladran, I., Gochman, P., Ruderfer, D. M., Stahl, E. A., Rapoport, J., Sklar, P., and Brennand, K. J. (2017). Transcriptional signatures of schizophrenia in hiPSC-derived NPCs and neurons are concordant with post-mortem adult brains. *Nature Communications*, **8**(1), 2225.

Huckins, L. M., Dobbyn, A., Ruderfer, D. M., Hoffman, G., Wang, W., Pardiñas, A. F., Rajagopal, V. M., Als, T. D., T. Nguyen, H., Girdhar, K., Boocock, J., Roussos, P., Fromer, M., Kramer, R., Domenici, E., Gamazon, E. R., Purcell, S., Demontis, D., Børglum, A. D., Walters, J. T. R., O'Donovan, M. C., Sullivan, P., Owen, M. J., Devlin, B., Sieberts, S. K., Cox, N. J., Im, H. K., Sklar, P., and Stahl, E. A. (2019). Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. *Nature Genetics*, **51**(4), 659–674.

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53**(3), 983–97.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, **82**(13).

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, **9**, 559.

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.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. a. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews. Genetics*, **11**(10), 733–9.

Mariani, J., Coppola, G., Zhang, P., Abyzov, A., Provini, L., Tomasini, L., Amenduni, M., Szekely, A., Palejev, D., Wilson, M., Gerstein, M., Grigorenko, E. L., Chawarska, K., Pelphrey, K. A., Howe, J. R., and Vaccarino, F. M. (2015). FOXG1-Dependent Dysregulation of GABA/Glutamate Neuron Differentiation in Autism Spectrum Disorders. *Cell*, **162**(2), 375–390.

Morgan, M., Obenchain, V., Lang, M., Thompson, R., and Turaga, N. (2019). BiocParallel: Bioconductor facilities for parallel evaluation.

Ooi, H. and Weston, S. (2019). iterators: Provides Iterator Construct. *R package version 1.0.12 https://CRAN.R-project.org/package=iterators*.

Pasca, S. P., Portmann, T., Voineagu, I., Yazawa, M., Shcheglovitov, A., Paşca, A. M., Cord, B., Palmer, T. D., Chikahisa, S., Nishino, S., Bernstein, J. A., Hallmayer, J., Geschwind, D. H., and Dolmetsch, R. E. (2011). Using iPSC-derived neurons to uncover cellular phenotypes associated with Timothy syndrome. *Nature medicine*, **17**(12), 1657–62.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, **43**(7), e47.

Wang, M., Beckmann, N. D., Roussos, P., Wang, E., Zhou, X., Wang, Q., Ming, C., Neff, R., Ma, W., Fullard, J. F., Hauberg, M. E., Bendl, J., Peters, M. A., Logsdon, B., Wang, P., Mahajan, M., Mangravite, L. M., Dammer, E. B., Duong, D. M., Lah, J. J., Seyfried, N. T., Levey, A. I., Buxbaum, J. D., Ehrlich, M., Gandy, S., Katsel, P., Haroutunian, V., Schadt, E., and Zhang, B. (2018). The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. *Scientific data*, **5**, 180185.

Warren, C. R., O'Sullivan, J. F., Friesen, M., Becker, C. E., Zhang, X., Liu, P., Wakabayashi, Y., Morningstar, J. E., Shi, X., Choi, J., Xia, F., Peters, D. T., Florido, M. H. C., Tsankov, A. M., Duberow, E., Comisar, L., Shay, J., Jiang, X., Meissner, A., Musunuru, K., Kathiresan, S., Daheron, L., Zhu, J., Gerszten, R. E., Deo, R. C., Vasan, R. S., O'Donnell, C. J., and Cowan, C. A. (2017). Induced Pluripotent Stem Cell Differentiation Enables Functional Validation of GWAS Variants in Metabolic Disease. *Cell Stem Cell*, **20**(4), 547–557.e7.

Wickham, H. (2009). *ggplot2*. Springer New York, New York, NY.

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

Yu, L., Zhang, J., Brock, G., and Fernandez, S. (2019). Fully moderated t-statistic in linear modeling of mixed effects for differential expression analysis. *BMC Bioinformatics*, **20**(Suppl 24), 1–9.