A COMPARATIVE MULTIPLE SIMULATION STUDY FOR PARAMETRIC AND NONPARAMETRIC METHODS IN THE IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Daniel Grant Palmer

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Program:
Applied Statistics

June 2021

Fargo, North Dakota

North Dakota State University
Graduate School

Title

A COMPARATIVE MULTIPLE SIMULATION STUDY FOR PARAMETRIC AND NONPARAMETRIC METHODS IN THE IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES

By

Daniel Grant Palmer

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Megan Orr
Chair

Dr. Timothy Greives

Dr. Bong-Jin Choi

Approved:

June 28, 2021
Date

Dr. Rhonda Magel
Department Chair

ABSTRACT

RNA-seq data simulated from a negative binomial distribution, sampled without replacement, or modified from read counts were analyzed to compare differential gene expression analysis methods in terms of false discovery rate control and power. The goals of the study were to determine optimal sample sizes/proportions of differential expression needed to adequately control false discovery rate and which differential gene expression methods performed best with the given simulation methods.

Parametric tools like edgeR and limma-voom tended to be conservative when controlling false discovery rate from a negative binomial distribution as the proportion of differential expression increased. For the nonparametric simulation methods, many differential gene expression methods did not adequately control false discovery rate and results varied greatly when different reference data sets were used for simulations.

ACKNOWLEDGMENTS

I want to thank my parents, Glen and Renee Palmer, for always supporting my career goals. This is my love letter to you guys. There is no way I would have made it through my undergraduate and graduate degrees without your unwavering love and encouragement.

Thank you, God, for imbuing me with skills and talents that I am utilizing to bring You glory. I hope I continue to grow in my faith.

I also would like to thank my advisor, Dr. Megan Orr, for helping me to decide on this research topic and setting up weekly progress meetings to make sure I was on track. Thank you, Dr. Tim Grieves and Dr. Bong-Jin Choi for agreeing to be on my committee.

I want to extend a special thanks to Professor Kathryn Lemm and Coach Jim Clark for the support through my undergraduate degree at the University of Jamestown. It was a pleasure to get to know you two over the years, and I appreciate the occasion holiday text from Coach Clark.

Thank you to all the friends and acquaintances I met over the years. You all have no doubt shared a role in shaping me into the person I am today.

I wish you all nothing but the best in your future endeavors.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

CPM .....Count per million

DDE .....Declared differentially expressed

DE .....Differentially expressed

DGE .....Differential gene expression

EE.....Equivalently expressed

FDR.....False discovery rate

GEO .....Gene Expression Omnibus

(G)LM .....(Generalized) linear model

NB .....Negative binomial

NCBI .....National Center for Biotechnology Information

RNA-seq .....Ribonucleic acid sequencing

SE.....Standard error

SRSWOR .....Simple random sampling without replacement

T1D .....Type I diabetes

LIST OF APPENDIX TABLES

CHAPTER 1. INTRODUCTION

RNA sequencing (RNA-seq) is a popular and relatively new method of identifying potentially differentially expressed (DE) genes in a dataset as opposed to the more antiquated microarray technology. Identifying DE genes is done through differential gene expression (DGE) analysis. Gene expression data generated from microarray technology are generally assumed to be continuous and normally distributed, implying that linear models can be used to analyze the data. RNA-seq technologies, on the other hand, generate expression levels that are counts, implying that generalized linear models (GLMs) should be used to analyze the data. Nonparametric methods can also be used to analyze both microarray and RNA-seq data. There are multiple types of RNA-seq data, including bulk RNA-seq, microRNA-seq, and single-cell RNA-seq. For clarity, any reference to RNA-seq in this disquisition will refer to bulk RNA-seq.

Before analysis can be performed on RNA-seq data, the reads must be aligned (Frazee, Jaffe, Kirchner, and Leek 2015) based on if the samples are single or paired-end. After alignment, the data are formatted into a read count matrix with columns represented by the samples (experimental units) and rows represented by the genes.

RNA-seq read counts were originally analyzed using Poisson GLMs because of the discrete nature of the data (Wu, Wang, and Wu 2014; Anders and Huber 2010). The support of the read counts has a lower bound of zero reads and no upper bound, which is consistent with Poisson random variables. Mathematically speaking, this GLM would work in practice, but previous research tells us that this model only accounts for the technical variation in the data (Frazee, Jaffe, Kirchner, and Leek 2015; Robinson, Mccarthy, and Smyth 2010; McCarthy, Chen, and Smyth 2012). Another component needs to be accounted for in our model: biological variation. Biological variation is associated with the individual replicates in the treatments and is

not accounted for in the Poisson GLM because the variance is greater than the mean (Robinson, Mccarthy, and Smyth 2010; Anders and Huber 2010; McCarthy, Chen, and Smyth 2012). The negative binomial (NB) model accounts for this overdispersion of the Poisson model.

Multiple simulations run under the assumption that RNA-seq read counts are NB distributed have been proposed in previous research (Wu, Wang, and Wu 2014) to determine appropriate sample sizes and compare differential expression analysis methods. Some researchers argue that parametric methods of simulations fail to replicate the complex structure of real RNA-seq data, and thus, should not be used to determine appropriate sample size or compare DGE methods (Benidt and Nettleton 2015; Gerard 2020). The debate of parametric and non-parametric simulations for RNA-seq was the motivation for this paper.

The goal of this study was to determine appropriate combinations of sample sizes and proportions of equivalent expression regarding each DGE method for both the parametric simulation method and the nonparametric simulation methods. Proportions of equivalently expressed (EE) genes cannot be prespecified in practice, but a simulation study allows us to control this particular setting of a study. Powers for DGE analysis methods that adequately controlled false discovery rate (FDR) were also of interest.

This paper is organized as follows: (1) chapter II contains a literature review of all the methods reviewed by the researcher (implemented or not), (2) chapter III is an outline of the procedures followed in each implemented method, (3) chapter IV presents the results found in the study, and (4) chapter V discusses the results while providing thoughts and recommendations for future research as well as limitations of the study.

CHAPTER 2. LITERATURE REVIEW

To our knowledge, little research has been done on directly comparing previous RNA-seq simulation methods. Most papers (e.g. seqgendiff and SimSeq) propose their own simulation methods and discuss the findings in the context of their proposed method (Benidt and Nettleton 2015; Gerard 2020). More research has been done for DGE analysis methods. Schurch et al. (2016), a popular example, made recommendations for DGE analysis tools to use for small and large sample sizes alike in yeast data.

Previous research led to R (R Core Team 2020) being a popular tool of choice for RNA-seq both for its open-source capabilities and the high popularity of Bioconductor (Huber et al. 2015, Gentleman et al. 2004), a software for molecular biology and bioinformatics.

Simulation Methods

Parametric

PROPER (Wu, Wang, and Wu 2014), which stands for PROspective Power Evaluation for RNASeq, is a parametric simulation method that simulates data from the NB distribution based on the parameters of a count matrix from a real RNA-seq dataset. This method starts by simulating gene-wise baseline expression levels and dispersions from the provided dataset. Effect sizes can also be set. Some DGE analysis methods are built into the package, but predefined DGE methods are used to have more control over specific steps and genes filtered. It also has nice graphical and tabular visualizations for power and sample size estimation.

Another method that was examined in the literature search was powsimR (Vieth, Ziegenhain, Parekh, Enard, and Hellmann 2017). Like PROPER, it is a parametric simulation method that simulates counts from a NB distribution. It can also be used for single-cell RNA-seq data simulation. This method also starts by simulating gene-wise baseline expression levels, but

it differs from PROPER in that it estimates dispersion parameters through fitting a mean-dispersion spline. Also, unlike PROPER, the powsimR method can directly account for batch data as part of the simulation. DGE analysis methods are also built into the package.

Semiparametric

A semiparametric method for simulating RNA-seq data is seqgendiff (Gerard 2020). It takes a nonparametric approach to the systematic components of the model but adds a signal to the data represented by an error term that follows a prespecified distribution. The read counts are modified by binomial thinning. Additionally, data from more complex models can be simulated by providing a correlation matrix or additional surrogate variables, but the simplest procedure for the two-group model can also be used.

Nonparametric

A nonparametric simulation method that is appropriate for reference data sets with large sample sizes in two or more treatment groups is SimSeq (Benidt and Nettleton 2015). SimSeq simulates RNA-seq reads by "subsampling columns from a large source RNA-seq dataset and then swapping individual read counts within genes adjusted by a correction factor to create differential expression" (Benidt and Nettleton 2015).

Another simulation method explored is a simple resampling method. This method randomly samples columns of equal sample sizes for each treatment group from the original read count matrix. According to the procedure outlined in Schurch et al. (2016), genes declared to be DE for the original read count matrix (the "true" DE genes) are identified using a DGE analysis method. Then, for each resampled data set, DGE analysis is again performed to identify genes declared differentially expressed (DDE). Performance of the DGE methods is performed by

comparing the list of true DE genes to the lists of DDE genes from the resampled data sets. More information on this can be found in the "Hypotheses of Interest" section.

Parametric or Nonparametric

Polyester (Frazee, Jaffe, Kirchner, and Leek 2015) is another simulation method explored. It is a unique simulation method in that it starts at the sequencing level to account for read alignment and counting. To start, the researcher downloads SRA (Sequence Read Archive) files for each sample from the NCBI SRA. Next, the SRA files are converted to FASTA files (a special type of FASTQ file) for input to the package. The SRA Run Selector is used to examine the samples to determine if they are single or paired-end prior to converting the files. With the FASTA files, the researcher can start simulating count matrices. Polyester has a built-in NB model for simulating reads or a user-defined count model that can be specified. The library sizes can also be modified by a certain factor.

Differential Gene Expression Analysis Methods

Parametric

Many parametric and nonparametric DGE analysis methods have been proposed for RNA-seq data. One of the first methods was edgeR (Robinson, Mccarthy, and Smyth 2010). edgeR is a parametric DGE analysis method that assumes the count data follows a NB distribution. edgeR starts preprocessing the data by filtering out lowly expressed genes before running analysis. The normalization factors are calculated, and the dispersion parameters are also calculated. A GLM fit or exact test is conducted for each gene to obtain a p-value.

Another DGE analysis method is DESeq2 (Love, Huber, and Anders 2014). This tool assumes the data follows a NB distribution and is the successor of DESeq (Anders and Huber 2010). It makes improvements in estimating dispersion parameters, introduces a regularized log

transformation option, and is overall less conservative in controlling type I error. The DESeq2 procedure begins by calculating size factors and dispersion parameters. Then, the dispersions are shrunk by an empirical Bayes' method. Gene-wise GLMs are fit, and Wald statistics are calculated and used to calculate a p-value for each gene. DESeq2 includes filter options, including independent filtering and Cook's distance filtering.

Another commonly used DGE analysis method is limma-voom. limma (Ritchie et al. 2015; McCarthy, Chen, and Smyth 2012) is a parametric DGE analysis method that was originally proposed for microarray data. Data modeled with limma is assumed to follow a log-normal distribution. The voom (Law, Chen, Shi, and Smyth 2014) method in limma is one of the more common methods used for RNA-seq data. The read count data is first transformed into log-count per million units. Gene-wise linear models (LMs) are fit to the data and precision weights are calculated. Gene-wise LMs are refit to the data, and dispersion parameters are shrunk with an empirical Bayes' method.

Another DGE analysis method is baySeq (Hardcastle and Kelly 2010), which also utilizes a parametric model for DGE analysis. It is similar to edgeR in that it can estimate parameters with quasi-likelihood estimation. As initial inputs, the user provides a list of a vector assuming the null hypothesis of no differential expression among replicates and a vector assuming differential expression between replicates. baySeq also uses an empirical Bayes approach to calculating the dispersion parameters. The prior distribution for read counts is assumed to be a NB distribution, and the posteriors are calculated with either maximum likelihood estimation or quasi-likelihood estimation.

Nonparametric

A nonparametric DGE analysis method is NOISeq (Tarazona, Garcia-Alcalde, Dopazo, Ferrer, and Conesa 2011; Tarazona et al. 2015). NOISeq is used for gene expression experiments with two treatments. The function used for biological replicates utilizes an empirical Bayes method for differential expression. This approach assumes there are two distinct populations: (1) genes with non-varying expression between two conditions of an experiment and (2) genes with changing expression between two experimental conditions. First, Z-statistics are calculated from the groups by permuting samples from the read count matrices and corresponding posterior probabilities of differential expression are computed and filtered by a cutoff.

CHAPTER 3. METHODS

Data

Two RNA-seq read count datasets were examined to compare simulation methods and DGE analysis methods. The primary dataset of interest measured expression levels of kidney genes. The participants ($N = 144$) were paired into tumor ($N_1 = 72$) and non-tumor ($N_2 = 72$) groups. There were 20531 genes total for each participant. The data for this experiment was queried from the SimSeq package.

The second dataset used for this paper measured expression levels of whole blood genes. The participants ($N = 82$) were comprised of patients diagnosed with type I diabetes (T1D) ($N_1 = 39$) and healthy volunteers ($N_2 = 43$). There were 16785 genes examined for each patient/volunteer. The data for this experiment was queried from the NCBI (National Center for Biotechnology Information) GEO (Gene Expression Omnibus) (Edgar, Domrachev, and Lash 2002). The accession for the data was GSE123658.

Settings

For each simulation method, a subset of samples (columns) was randomly selected from the original dataset. Denoting the subsetted samples as n such that $n \subset N_1$ and $n \subset N_2$, three different sample size settings were examined: $n = 3$, $n = 6$, and $n = 10$. Letting p denote the proportion of EE genes (rows) for each simulation, four different proportion settings were examined: $p = 0.5$, $p = 0.7$, and $p = 0.9$, and $p = 1$. There were 10000 genes simulated for each simulation. At each simulation setting, 100 total simulations were run. Table 1 below gives a visual representation of the simulation settings.

Table 1. Simulation Settings.

Sample Size (n)	**Proportion EE Genes (p)**	**Simulation Method**	**DGE Method**
3	0.5	PROPER	edgeR
6	0.7	seqgendiff	DESeq2
10	0.9	SimSeq	limma-voom
	1	Resampling	NOISeq

Every possible combination of sample size, proportion of EE genes, simulation method, and DGE analysis method was run.

Filtering

In agreement with the SimSeq paper (Benidt and Nettleton 2015), genes were kept that had a mean read count of ten or greater and a lower bound of two for nonzero read counts for the SimSeq, seqgendiff, and resampling methods. From the total number of genes in the original read count matrices, $10000p$ EE genes were randomly sampled from the total number of EE genes, and consequently $10000(1 - p)$ DE genes were randomly sampled from the total number of DE genes.

For the PROPER NB simulation method, genes from the RNA-seq count data were removed if they yielded all zero read counts.

Design Matrix

Some preliminary analysis showed that there was a significant batch effect in this experiment. The batch information was obtained from the NCBI GEO website with the GEOquery (Davis and Meltzer 2007) package in R. The methods that did not account for a batch effect used a design matrix X

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \quad (1)$$

where the column vector of 1s corresponds to the intercept term, and the second column vector corresponds to the indicator variable (0/1) of T1D. The methods that did account for the batch effect were represented by a different design matrix X such that

$$X = \begin{bmatrix} 1 & 0 & b_{13}^{2} & b_{14}^{3} \\ 1 & 0 & b_{23}^{2} & b_{24}^{3} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & b_{(n-1)3}^{2} & b_{(n-1)4}^{3} \\ 1 & 1 & b_{n3}^{2} & b_{n4}^{3} \end{bmatrix} \quad (2)$$

where

b_{i3}^{2} represents the column vector corresponding to the indicator variable of Batch 2 and,

b_{i4}^{3} represents the column vector corresponding to the indicator variable of Batch 3.

Rows where $(b_{i3}^{2}, b_{i4}^{3}) = (0, 0)$ represented the effect of Batch 1.

Simulation Methods

PROPER

The first simulation method, PROPER (Wu, Wang, and Wu 2014), had three main steps to simulate data from a read count matrix: (1) estimation of the simulation parameters from the original read count matrix, (2) determination of the options of the simulation using the returned objects, and (3) running the simulations with the given options. The first step computed the sequencing depth for each sample, the log baseline expression rate per gene, and the log dispersion for each gene.

To run the simulations, y_{gk} represented the observed read count of the g^{th} gene and the k^{th} replicate. It was assumed that the corresponding random variable

$$Y_{gk} \sim Negative\ Binomial(s_k \mu_g, \phi_g) \quad (3)$$

where

s_k is the normalization factor for the k^{th} replicate,

μ_g represents the baseline expression level (mean) for the g^{th} gene, and

ϕ_g represents the dispersion parameter for the g^{th} gene.

This step could have been skipped entirely if the researcher decided to estimate the simulation parameters with the built-in datasets (i.e., Cheung, Gilad, Bottomly, MAQC). The second step retained the previous simulation options as well as user input for number of genes simulated, the percentage of genes DE, the log (base 2) fold change of the DE genes, and a random seed number used to reproduce the results if needed. The function returned the read counts, the treatment vector (indicator for T1D) for the selected samples, an index of selected DE genes, and the previous simulation options.

seqgendiff

Two main steps were performed to simulate data from the seqgendiff package: (1) selection of the samples and (2) addition of a signal following a specific parametric model to the data through binomial thinning.

A simple random sample without replacement (SRSWOR) approach was implemented in the first step. The N_1 samples (tumor group and T1D patients) were selected first and then the N_2 samples (non-tumor group and healthy volunteers) were selected afterward for the datasets. All genes from the original count matrices were retained. It was worth noting that for the sampled read counts the matrix

$$\mathbf{Y}_{g\,x\,k} \sim Poisson(2^{\theta gk}) \quad (4)$$

so that

$$\theta = \mu \mathbf{1}_k^T + \Omega \quad (5)$$

where

k is the number of replicates,

μ is the column vector of intercept terms for the genes,

$\mathbf{1}_k$ is the column vector of 1s, and

$\Omega \in \mathbb{R}_{g\ x\ k}$ is the matrix accounting for variation in the model.

A signal (Ω) was then added to the distribution in order to better account for the real variability in RNA-seq data through binomial thinning. Other inputs to the function included the proportion of genes that were null and the proportion of genes in group 1. The remaining function inputs were left to their defaults.

Suppose now there is a design matrix with p covariates

$$X \in \mathbb{R}_{k\ x\ p}$$

and a square permutation matrix

$$\Pi \in \mathbb{R}_{k\ x\ k}$$

Then the resultant matrix of read counts is modeled as

$$\tilde{\mathbf{Y}}_{g\ x\ k} \sim Poisson\left(2^{\tilde{\theta}gk}\right) \quad (6)$$

so that

$$\tilde{\theta} = \tilde{\mu} \mathbf{1}_k^T + bx^T \Pi^T + \Omega \quad (7)$$

where

$\tilde{\mu}$ is the new column vector of intercept terms for the genes,

b is the coefficients vector corresponding to the design matrix X, and

x corresponds to the appropriate row vector of the design matrix X.

The coefficients vector, b, was used to identify DE and EE genes. This vector of the selected genes can be notated as

$$b = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_g \end{bmatrix} \tag{8}$$

where

c_i for $i = 1, 2, \ldots, g$ represent constants.

For any $c_i = 0$, the genes were deemed EE, otherwise they were DE.

The binomial thinning part of these simulations was not readily apparent to the researcher from these above equations, so the following steps are used to explain how to go from equation 4 to equation 6. The binomial conditional distribution is applicable to both RNA-seq data with just technical variation (Poisson) and also biological variation (NB). For data with technical variation, we assume $\mathbf{Y} \sim Poisson(\mu)$. If the conditioning on $\mathbf{Y}$ is such that $\widetilde{\mathbf{Y}} \mid \mathbf{Y} \sim Binomial(\boldsymbol{y}, p)$, then $\widetilde{\mathbf{Y}} \sim Poisson(\mu p)$ (Gerard 2020), resulting in the same distribution of the original random variable. For data with biological variation, we assume $\mathbf{Y} \sim NB(\mu, \phi)$. If the conditional distribution is binomial such that $\widetilde{\mathbf{Y}} \mid \mathbf{Y} \sim Binomial(\boldsymbol{y}, p)$, then $\widetilde{\mathbf{Y}} \sim NB(\mu p, \phi)$.

SimSeq

The third simulation method, SimSeq, had a ten-step process used to simulate read counts (Benidt and Nettleton 2015):

1. The set of all genes in the count matrix was denoted as G, with g being an individual gene in G. Gene-wise ($\forall g \in G$) p-values were computed with a DE test using the Wilcoxon Signed Rank test.
2. Gene-wise local false discovery rates (FDRs) were calculated for each p-value.

3. Probability weights for each gene were calculated by taking the difference of 1 and the local FDR.
4. $\mathbf{G_1}$ genes were randomly selected without replacement to be DE from G according to the probability weights computed above and called G_1.
5. $\mathbf{G_0}$ genes were randomly selected without replacement from G to be EE genes and called G_0. A new set of genes was defined as $G^* = G_0 \cup G_1$ which encompassed all DE and EE genes selected.
6. $\mathbf{Y}$ was denoted as the matrix of read counts initially provided. A sample $\boldsymbol{y}$ was randomly selected without replacement from the first treatment group and subsetted to the set of genes G^*. This column was denoted as $\boldsymbol{x_1}$ and was assigned to the treatment group 1 of simulations.
7. A sample was randomly selected without replacement from both treatment groups in $\mathbf{Y}$ with columns $\mathbf{Y_1}$ and $\mathbf{Y_2}$. The appropriate normalization factors were denoted as s_1 and s_2 for these columns. The 0.75 quantile of the k^{th} replicate was selected for normalization.
8. The set of genes G^* were subsetted for $\mathbf{Y_1}$ and $\mathbf{Y_2}$.
9. $\boldsymbol{x_2}$ was created $\forall g \in G^*$ by

$$\boldsymbol{x_{2g}} = \begin{cases} \boldsymbol{y_{1g}} \; if \; g \in G_0 \\ \left\lfloor \boldsymbol{y_{2g}} \left(\frac{s_1}{s_2} \right) + 0.5 \right\rfloor \; if \; g \in G_1 \end{cases} \tag{9}$$

   where $\lfloor \cdot \rfloor$ is the floor operator to round the quantity down to the nearest integer.
10. The sixth through ninth steps were repeated n times.

Resampling

The resampling method used the SRSWOR method outlined in the seqgendiff procedure. Denoting $\mathbf{Y}$ as the read count matrix, n samples were selected from each treatment group and the

submatrices were denoted as $\mathbf{Y_1}$ and $\mathbf{Y_2}$. The new read count matrix was $\mathbf{Y}^* = [\mathbf{Y_1} \quad \mathbf{Y_2}]$. Unfortunately, resampling did not explicitly define the DE genes used in the simulations. In this case, the approach used was such that the DGE analysis methods were run on the full dataset of "clean replicates" and the subsetted samples outlined in the Schurch et al. (2016) paper. The DDE genes from the subsetted samples were compared to the DE genes from the full dataset.

Differential Gene Expression Analysis Methods

edgeR

edgeR (Robinson, Mccarthy, and Smyth 2010) is one of the more popular methods used for DGE analysis. The first step in the edgeR process was to create a DGEList object to hold all the appropriate information. The next step was to calculate the normalization factors using the trimmed mean of M-values method (Robinson and Oshlack 2010). There were three main steps in this process: (1) calculation of the M-values [gene-wise log (base 2) fold changes] of the k^{th} replicate and the r^{th} reference sample so that

$$M_{gk}^{r} = log_2\left(\frac{y_{gk}/n_k}{y_{gr}/n_r}\right) \tag{10}$$

and calculate the A-values (absolute expression levels) of the k^{th} replicate and the r^{th} reference sample so that

$$A_{gk}^{r} = \frac{1}{2}\left[log_2\left(y_{gk}/n_k\right) + log_2\left(y_{gr}/n_r\right)\right] \tag{11}$$

(2) trimming the M-values and the A-values (defaults are 30% for M-values and 5% for A-values), and (3) calculation of the weighted mean of the trimmed M-values for the k^{th} replicate so that

$$log_2(TMM_k^r) = \frac{w_{gk}^r M_{gk}^r}{w_{gk}^r} \tag{12}$$

where

$$w_{gk}^{r} = \frac{n_k - y_{gk}}{n_k y_{gk}} + \frac{n_r - y_{gr}}{n_r y_{gr}}.$$

The next step in the edgeR procedure was to calculate the dispersion parameters (ϕ_g) by shrinking them toward a common dispersion. This used an empirical Bayes weighted likelihood given by

$$WL(\phi_g) = l_g(\phi_g) + \alpha l_C(\phi_g) \tag{13}$$

where

$l_g(\cdot)$ is the log-likelihood function for gene g,

$l_C(\phi_g) = \sum_{g=1}^{m} l_g(\phi_g)$ is the gene-wise common log-likelihood and prior distribution of ϕ_g, and

α is the weight given to the common likelihood (prior precision).

A likelihood-ratio test was performed for each gene and p-values were calculated. More information is given in the "Hypotheses of Interest" section.

DESeq2

DESeq2 (Love, Huber, and Anders 2014) is another popular method for DGE analysis used by researchers. The first step was to create a DESeq object. The next step was to call the DESeq function, which used three processes internally: (1) estimation of the normalization factors, (2) estimation of the dispersion parameters, (3) fitting a NB GLM and computation of the Wald statistics. Estimating the normalization factors differed from edgeR in that it used the median-of-ratios method defined below

$$\widehat{s_k} = median_g \frac{y_{gk}}{\left(\prod_{v=1}^{n} y_{gv}\right)^{1/n}} \tag{14}$$

where

$\left(\Pi_{\nu=1}^{n} y_{g\nu}\right)^{1/n}$ is the geometric mean of the read counts for gene g and the normalization factor, and

$median_g(\cdot)$ is the median function for gene g.

Next, let the Fisher Information Matrix be denoted by

$$\mathbb{I}_g = X^T W_g X \tag{15}$$

where X is the design matrix, and W_g is the diagonal weights matrix for gene g. The dispersion parameters for DESeq2 were estimated using the Cox Reid-adjusted profile likelihood which is given by

$$APL_g(\phi_g) = l(\phi_g; y_g, \widehat{\beta_g}) - \frac{1}{2} ln\left[det(\mathbb{I}_g)\right] \tag{16}$$

where

$l(\cdot)$ is the log-likelihood function,

ϕ_g is the dispersion parameter for gene g,

y_g is the read count for gene g, and

$\widehat{\beta_g}$ is the estimated coefficient vector for gene g.

The dispersion parameters were again shrunk toward a common value, but the difference between this process and the one used in edgeR was that the dispersions were shrunk to a common prior (Wu, Wang, and Wu 2012). The log-fold changes (LFCs) were shrunk in a similar fashion to the dispersion parameters. The gene-wise LFCs were tested for significance with Wald tests and p-values were generated. More information is given in the "Hypotheses of Interest" section.

limma-voom

limma (Ritchie et al. 2015; McCarthy, Chen, and Smyth 2012) is another popular method used for DGE analysis. To use limma with RNA-seq data, an edgeR DGEList object was created and normalization factors from edgeR were applied. The voom (Law, Chen, Shi, and Smyth 2014) method in limma log-transformed (base 2) the count data so the units were in counts per million (CPM). There was an offset of 0.5 to account for read counts of zero. The documentation defined the formula converting read counts to log-CPM as

$$y_{gk} = log_2\left(\frac{r_{gk} + 0.5}{\sum_{g=1}^{G} r_{gk} + 1.0} x 10^6\right) \tag{17}$$

where

r_{gk} is the read count for gene g and sample k, and

$\sum_{g=1}^{G} r_{gk}$ is the sum of all read counts (mapped reads) for gene g and sample k.

Linear models (LMs) were then fit to the data in a similar way as what is done for microarray data. The gene-wise LMs were fit using ordinary least squares where the expected value of the log-CPM was represented as

$$E(y_g) = X\beta_g \tag{18}$$

Gene-wise mean log-CPMs were computed as

$$\tilde{r} = \overline{y_g} + log_2(\tilde{R}) - log_2(10^6) \tag{19}$$

where

$\tilde{R}$ is the geometric mean of the library sizes plus 1.

A locally weighted regression (LOWESS) curve (notated as $lo(\cdot)$) was fit to create a smooth mean-variance trend. The fitted values from the LMs ($\widehat{\mu_{gl}}$) can be written as

$$\widehat{\lambda_{gk}} = \widehat{\mu_{gk}} + log_2\left(\sum_{g=1}^{G} r_{gk} + 1\right) - log_2(10^6) \tag{20}$$

and the precision weights are calculated as

$$w_{gk} = lo\left(\widehat{\lambda_{gk}}\right)^{-4} \tag{21}$$

and these weights are associated with the read counts, y_{gk}. Gene-wise LMs were fit to the data and empirical Bayes' dispersions were shrunk to a common value. The corresponding p-values were calculated. More information is given in the "Hypotheses of Interest" section.

NOISeq

For the nonparametric DGE analysis method, NOISeq (Tarazona, Garcia-Alcalde, Dopazo, Ferrer, and Conesa 2011; Tarazona et al. 2015), there were three main steps to identify DE genes: (1) computation of the Z-statistics, (2) estimation of the Z-scores, and (3) computation of the posterior probabilities of differential expression.

A Z-statistic was proposed such that

$$Z = \frac{M + D}{2} \tag{22}$$

where M is the log-ratio (base 2) of the two conditions (treatment and control) and the difference between the two conditions, D. Next, the biological replicates were accounted for by recomputing the M and D values (denoted as M^* and D^*, respectively). The formulas are

$$M^* = \frac{\mathrm{M}}{\mathrm{a}_0 + \sqrt{SE(\mathrm{M})}} \tag{23}$$

and

$$D^* = \frac{D}{a_0 + \sqrt{SE(D)}} \tag{24}$$

where $SE(D)$ and $SE(M)$ are the standard errors of the D and M statistics, respectively, and a_0 is the user-inputted quantile of the values of $\sqrt{SE(M)}$ and $\sqrt{SE(D)}$. The new DE statistic was then computed as

$$Z = \frac{M^* + D^*}{2} \tag{25}$$

To calculate the Z-scores for the null genes (Z_0), the read count matrix was partitioned into two subsets of the original: $\mathbf{Y_1}$ (control) and $\mathbf{Y_2}$ (treatment), both of dimension g genes and n samples. The sample names of these matrices were permuted r (default is 50) times, and the Z-statistics outlined by equations 22-25 were calculated again. Matrices of g rows and r columns were generated with the Z_0 statistics being the pooled results. Different approaches were used in dealing with small and large sample sizes. For samples with less than 5 replicates, a k-means clustering algorithm was used to compute M and D values for determining DE genes by borrowing information across genes. Within the k clusters of genes, the sample names were permuted in a similar fashion and Z_0 statistics were calculated.

The distribution of the test statistic Z can be represented by the probability density function

$$f(z) = p_0 f_0(z) + p_1 f_1(z) \tag{26}$$

where

p_0 is the probability of a gene having similar expression levels between groups,

f_0 is the density corresponding to p_0,

p_1 is the probability of a gene having dissimilar expression levels between groups, and

f_1 is the density corresponding to p_1.

For any gene g, Bayes' rule can be used to compute the probability of differential expression ($p_1(z)$) by

$$p_1(z) = \frac{p_1 f_1(z)}{f(z)} = 1 - \frac{p_0 f_0(z)}{f(z)} \tag{27}$$

which has significance in calculating gene-wise FDRs shown below.

Other Methods

Simulation methods and DGE analysis methods not listed here in the "Methods" sections were not used in this study due to issues with at least one of the datasets or with memory issues in the R computing environment.

Hypotheses of Interest

Gene-wise hypothesis tests were performed for the two-group setting to identify DE genes. The hypotheses of interest were

$$H_0: \mu_1 = \mu_2 \tag{28}$$

$$H_1: \mu_1 \neq \mu_2$$

where rejecting H_0 declared the gene DE. For 10000 genes, 10000 hypotheses were run. This introduced the common multiple testing problem of type I error. In table 2, a graphical representation of type I (and type II) error is shown to illustrate the outcomes based on the decision rule of these hypotheses.

Table 2. Confusion Matrix for Decision Rule on Hypothesis.

	Fail to reject H_0	Reject H_0
H_0 is true	Correct Decision (True Non-Discovery)	Type I Error (False Discovery)
H_0 is false	Type II Error (False Non-Discovery)	Correct Decision (True Discovery)

In gene expression, true positives are known as true discoveries (S), false positives are known as false discoveries (V), true negatives are known as true non-discoveries (U), false

negatives are known as false non-discoveries (T), total positives are known as total discoveries (R), and total negatives are known as total non-discoveries (W). These quantities are considered random variables and are displayed in table 3.

Table 3. Confusion Matrix for Decision Rule on Hypothesis in the Context of Differential Gene Expression.

	Fail to reject H_0	Reject H_0 (Gene is DDE)	Total
H_0 is true (Gene is EE)	U	V	$\mathbf{G_0}$
H_0 is false (Gene is DE)	T	S	$\mathbf{G_1}$
Total	W	R	$\mathbf{G}$

Source: Adapted from Benjamini and Hochberg (1995).

The total column in table 3 represents the total number of EE genes ($\mathbf{G_0}$), the total number of DE genes ($\mathbf{G_1}$), and total number of genes ($\mathbf{G}$) in a simulated dataset. These quantities are fixed and were determined a priori to the experiment.

False Discovery Rate

Controlling the FDR is imperative as a natural result of the multiple testing conundrum in gene expression experiments where thousands of hypothesis tests are performed simultaneously. The observed FDR is defined as

$$Q = \frac{V}{R} \text{ if } R > 0 \tag{29}$$

$$Q = 0 \text{ otherwise}$$

(Benjamini and Hochberg 1995) which was calculated for each simulation.

Each R software package used either a q-value or Benjamini-Hochberg correction (adjusted p-value) to calculate the gene-wise FDR for each hypothesis test. edgeR, DESeq2, and

limma use the Benjamini-Hochberg method. Benjamini and Hochberg defined the theoretical FDR as the expected value of the observed FDR (Benjamini and Hochberg 1995) which is denoted as

$$FDR = E(Q) \tag{30}$$

.

Gene-wise FDRs were controlled at a significance level of $\alpha = 0.1$, and the genes with an estimated FDR less than α were deemed DE. NOISeq (Tarazona, Garcia-Alcalde, Dopazo, Ferrer, and Conesa 2011; Tarazona et al. 2015) used a value defined in the literature as

$$p_1(z) = 1 - FDR \tag{31}$$

where FDR is the theoretical FDR for declaring genes DE.

Controlling the theoretical FDR using the Benjamini and Hochberg (1995) procedure at a significance level, α, is actually controlling the FDR such that

$$E(Q) \leq \alpha \left(\frac{\mathbf{G_0}}{\mathbf{G}}\right) \leq \alpha \tag{32}$$

for multiple hypotheses. Thus, in our results, it was better to be more conservative than α.

Power

Assuming the FDR was adequately controlled at a predefined level, running simulations like these allowed the researcher to calculate the power of each hypothesis test post hoc. Power is generally represented in the general form

$$Power = 1 - P(Type\ II\ error) \tag{33}$$

and in this case, it is the difference of 1 and the false non-discovery rate (FNDR)

$$Power = 1 - FNDR \tag{34}$$

where

$$FNDR = \frac{T}{T+S}$$

(Schurch et al. 2016) in practice.

For $p = 1$, a different criterium must be used. From table 3, it is shown that $T + S = \mathbf{G_1} = 0$ DE genes. It follows that the power simplifies to

$$Power = 1 - \frac{T}{T+S}$$

$$Power\ |_{T=0,S=0} = 1 - \frac{0}{0}$$

so that the power of the test is always undefined. As such, the power calculations were not included in the tables or figures for simulation settings with $p = 1$.

Software

Version 4.0.2 of R (R Core Team 2020) was used to run the code for the experiment. Most R packages were available through Bioconductor (Huber et al. 2015, Gentleman et al. 2004) and are listed as follows: GEOquery (Davis and Meltzer 2007), PROPER (Wu, Wang, and Wu 2014), seqgendiff (Gerard 2020), SimSeq (Benidt and Nettleton 2015), edgeR (Robinson, Mccarthy, and Smyth 2010), DESeq2 (Love, Huber, and Anders 2014), limma (Ritchie et al. 2015; McCarthy, Chen, and Smyth 2012) with voom (Law, Chen, Shi, and Smyth 2014), NOISeq (Tarazona, Garcia-Alcalde, Dopazo, Ferrer, and Conesa 2011; Tarazona et al. 2015), and kableExtra (Zhu 2021).

CHAPTER 4. RESULTS

As stated above, observed power and observed FDR were calculated using both the Schurch et al. (2016) method (resampling) and the "true" DE genes in the simulated datasets. The appendix displays two different types of tables for each simulation method: (1) results corresponding to the mean observed FDR and (2) results corresponding to the mean observed FDR and the mean power. Tables were generated for every possible combination of simulation method, DGE analysis method, and proportion of EE genes. The first type of table is only used for ($p = 1$) because the power is undefined at this setting (see the "Power" subsection in the "Methods" section). The first column of each table shows the sample size for each combination and the DGE analysis tool used. The second column displays the mean power of the 100 simulations. The third column shows the standard errors of the power estimates. The third and fourth columns display the mean observed FDRs and the standard errors of their estimates, respectively.

Kidney Data

PROPER ($p = 1$)

Results for this setting are listed in table 4. The mean observed FDR tended to decrease with the exception of DESeq2 as the sample size increased. The mean observed FDRs were close to 1 for edgeR, DESeq2, and NOISeq. The standard errors of the estimates increased as the sample size increased for edgeR and NOISeq while they decreased for limma-voom. These estimates stayed the same for DESeq2. None of the observed FDRs were controlled adequately at an appropriate level, α. The limma-voom method yielded the lowest observed FDRs for all sample sizes with 0.54 ($n = 3$), 0.25 ($n = 6$), and 0.16 ($n = 10$). According to the table,

limma-voom may be the preferred DGE method for large sample sizes of the PROPER simulation method because it controls FDR the best.

PROPER ($p = 0.9$)

The results for this simulation setting are listed in table 5. As the sample size increased, the powers increased, and the standard error of the powers decreased for the DGE analysis tools. Observed FDRs tended to decrease for increasing sample sizes. The voom method from limma controlled FDR the best for each sample size with 0.1318 ($n = 3$), 0.0861 ($n = 6$), and 0.0837 ($n = 10$), respectively. The corresponding mean powers for these observed FDRs are 0.1306, 0.3304, and 0.4250. NOISeq controlled FDR adequately for $n = 10$ with an observed FDR of 0.0845, and the corresponding power for this FDR is 0.0193. For $n = 6$, NOISeq yielded an observed FDR of 0.0936 and power of 0.0082. The DGE analysis method edgeR controlled FDR at 0.1815 ($n = 6$) and 0.1353 ($n = 10$). The corresponding mean powers were 0.3677 and 0.4661. Standard errors of the observed FDRs tended to decrease for edgeR and voom with an increase in sample size, but they increased for DESeq2 and NOISeq. The DGE method limma-voom is the best choice for this setting because it controls FDR well and yields the highest powers for all sample sizes.

PROPER ($p = 0.7$)

Table 6 lists the results for this simulation setting. As the sample size increased, mean power increased for the DGE analysis methods. Standard errors of these estimates tended to decrease or stay the same with an increase in sample size. The voom method controlled FDR adequately for all sample sizes with observed FDRs of 0.0722 ($n = 3$), 0.0623 ($n = 6$), and 0.0602 ($n = 10$). The corresponding powers are 0.2290, 0.4018, and 0.4842. NOISeq had observed FDRs of 0.0796 ($n = 6$) and 0.0701 ($n = 10$). It yielded corresponding powers of

0.0127 and 0.0624, respectively. The DGE analysis method edgeR controlled FDR adequately with observed FDRs of 0.0944 ($n = 6$) and 0.0736 ($n = 10$). Corresponding powers for edgeR were 0.4306 and 0.5243, respectively. The edgeR method observed FDR at $n = 3$ was 0.1594, and its corresponding power was 0.2715. DESeq2 observed FDRs were 0.1601 ($n = 6$) and 0.1212 ($n = 10$). Their corresponding powers were 0.5005 and 0.5578. Standard errors of the observed FDRs tended to decrease or stay the same with an increase in sample size. The DGE method limma-voom is recommended for small sample sizes at this setting because of adequate FDR control, and edgeR is recommended for moderate/large sample sizes at this setting as it yields larger powers than voom with adequate FDR control.

PROPER ($p = 0.5$)

Table 7 shows the results of this simulation setting. Mean power increased as sample size increased. Standard errors of the powers decreased or stayed the same as sample size increased. Observed FDRs tended to roughly stay the same or slightly decrease as sample size increased. The limma-voom method yielded the lowest observed FDRs of 0.0439 ($n = 3$), 0.0409 ($n = 6$), and 0.0406 ($n = 10$). The corresponding powers for these observed FDRs are 0.2907, 0.4478, and 0.5273. The DGE analysis method edgeR controlled FDR adequately with observed FDRs of 0.0794 ($n = 3$), 0.0507 ($n = 6$), and 0.0452 ($n = 10$). Corresponding powers for edgeR are 0.3138, 0.4756, and 0.5691. DESeq2 controlled FDR adequately with observed FDRs of 0.0848 ($n = 6$) and 0.0675 ($n = 10$). Corresponding powers for DESeq2 are 0.5360 and 0.5927. At $n = 3$, the observed FDR for DESeq2 was 0.1175 with a corresponding power of 0.4392. The observed FDRs for NOISeq are 0.0601 ($n = 6$) and 0.0459 ($n = 10$). Corresponding powers are 0.0426 and 0.2024. At $n = 3$, the observed FDR for NOISeq is 0.2020, and its corresponding power is 0.5426. Standard errors of the observed FDRs tended to

decrease with sample size increases. edgeR is recommended for small sample sizes because it yielded the highest power and controlled FDR adequately for this simulation setting. DESeq2 is recommended for moderate/large sample sizes because it yielded the largest powers and controlled FDR well.

seqgendiff

The simulation settings for seqgendiff are found in tables 8, 9, 10, and 11. No DGE analysis method adequately controlled FDR at any sample size or proportion of equivalent expression. Standard errors of the observed FDRs decreased or stayed the same as sample size increased.

SimSeq ($p = 1$)

Results for this simulation setting are listed in table 12. Observed FDRs had a decreasing trend as sample size increased for each DGE analysis tool. The voom method controlled FDR adequately with observed FDRs of 0.04 ($n = 3$), 0.05 ($n = 6$), and 0.03 ($n = 10$). The edgeR and NOISeq methods' standard errors of the observed FDRs increased as sample size increased, but the DESeq2 and limma-voom methods' standard errors stayed roughly the same. limma-voom is the only recommended DGE method for adequate FDR control for this setting.

SimSeq ($p = 0.9$)

The simulation results are displayed in table 13 for this setting. Mean power increased with increased sample size for the DGE analysis methods. Standard errors of these statistics increased as sample size increased with the exception of NOISeq. Observed FDRs tended to decrease with an increase in sample size. The limma-voom method yielded the smallest observed FDRs of 0.0466 ($n = 3$), 0.0725 ($n = 6$), and 0.0583 ($n = 10$). Corresponding powers for the voom FDRs are 0.0111, 0.1321, and 0.2934. NOISeq controlled FDR adequately at $n = 10$

with an observed FDR of 0.0742 and a corresponding power of 0.0420. At $n = 6$, NOISeq yielded an observed FDR of 0.1225 and a corresponding power of 0.0197. Standard errors of these statistics tended to decrease as the sample size increased. limma-voom is the DGE analysis method recommended for all sample sizes for this setting because of its adequate FDR control.

SimSeq ($p = 0.7$)

Table 14 lists the simulation results for this setting. Mean power increased with increased sample size for each DGE analysis tool. Standard errors of the powers varied between the DGE analysis methods as sample size increased. The limma-voom method yielded the smallest observed FDRs of 0.0412 ($n = 3$), 0.0618 ($n = 6$), and 0.0560 ($n = 10$). The corresponding powers for these observed FDRs are 0.0386, 0.2389, and 0.4062. For NOISeq, FDR was controlled adequately with observed FDRs of 0.0705 ($n = 6$) and 0.0647 ($n = 10$). Corresponding powers for NOISeq are 0.0471 and 0.0950. The observed FDR at $n = 10$ is 0.1548 for DESeq2, and its corresponding power is 0.4747. The observed FDR at $n = 6$ is 0.1880 for edgeR, and its corresponding power is 0.2964. The observed FDR at $n = 10$ is 0.1736 for edgeR, and its corresponding power is 0.4268. Standard errors of the observed FDRs tended to decrease as sample size increased. The DGE method limma-voom is recommended for all sample sizes at this setting because it yields the highest powers for methods that control FDR adequately.

SimSeq ($p = 0.5$)

Simulation results for this setting are displayed in table 15. Power increased for the DGE analysis methods as the sample size increased. Standard errors of the powers decreased or stayed the same as sample size increased. Observed FDRs tended to decrease as sample size increased. The DGE analysis method limma-voom yielded the smallest observed FDRs of 0.0466 ($n = 3$), 0.0516 ($n = 6$), and 0.0408 ($n = 10$). Corresponding powers for voom are 0.0738, 0.2841, and

0.4633, respectively. FDRs were controlled adequately also for NOISeq with observed FDRs of 0.0539 ($n = 6$) and 0.0463 ($n = 10$). Corresponding powers for these sample sizes are 0.0816 and 0.1510, respectively. The edgeR method yielded observed FDRs of 0.1426 ($n = 3$), 0.1165 ($n = 6$), and 0.1002 ($n = 10$). Corresponding powers for edgeR observed FDRs are 0.1697, 0.3220, and 0.4671. DESeq2 had observed FDRs of 0.1688 ($n = 3$), 0.1358 ($n = 6$), and 0.0926 ($n = 10$). Corresponding powers for these observed FDRs are 0.2216, 0.3830, and 0.5127, respectively. For lower to moderate sample sizes, limma-voom is recommended for this setting because it controls FDR adequately. DESeq2 is recommended for large sample sizes with this setting because it yields a larger power than limma-voom and controls FDR.

Resampling

The simulation results for this setting are in table 16. NOISeq controlled FDR adequately with observed FDRs of 0.0321 ($n = 6$) and 0.0304 ($n = 10$). Corresponding powers for these observed FDRs are 0.1525 and 0.2328. At $n = 3$, NOISeq yielded an observed FDR of 0.1695, and its corresponding power is 0.3733. Standard errors of the observed FDRs decreased or stayed the same as sample size increased. NOISeq is recommended for the resampling method because it is the only DGE method that controls FDR adequately.

Boxplots

The PROPER power box-and-whisker plots are shown in figure 1, and the observed FDR box-and-whisker plots are shown in figure 2. Parametric DGE methods yielded a low spread of data regardless of the proportion of EE genes, p, and displayed a consistent upward trend for the power. NOISeq yielded a high spread for the power boxplots for low replicates but a smaller observed FDR spread. It yielded a lower spread for moderate to high sample sizes. The

parametric methods yielded a low spread for observed FDRs regardless of proportion of EE genes while NOISeq was more variable at high proportions of equivalent expression.

The seqgendiff power box-and-whisker plots are shown in figure 1, and the observed FDR box-and-whisker plots are shown in figure 2. Parametric and nonparametric DGE methods alike yielded a large spread for the power boxplots and increasing proportions of equivalent expression did not seem to increase power. Sample size, however, did increase power. The observed FDRs were consistently large for all DGE methods, but they had a small spread. They also decreased as p decreased.

The SimSeq power box-and-whisker plots are shown in figure 1, and the observed FDR box-and-whisker plots are shown in figure 2. Parametric and nonparametric DGE methods alike yielded a large spread for the power boxplots and increasing proportions of equivalent expression along with sample size increased power. For lower proportions of EE genes, the parametric methods seemed to be more robust and perform even better than NOISeq with potential model assumption violations. The limma-voom method was more robust regardless of sample size or proportion of equivalent expression.

The resampling power box-and-whisker plots are shown in figure 1, and the observed FDR box-and-whisker plots are shown in figure 2. NOISeq was the only method to control FDR well, and it yielded decent power.

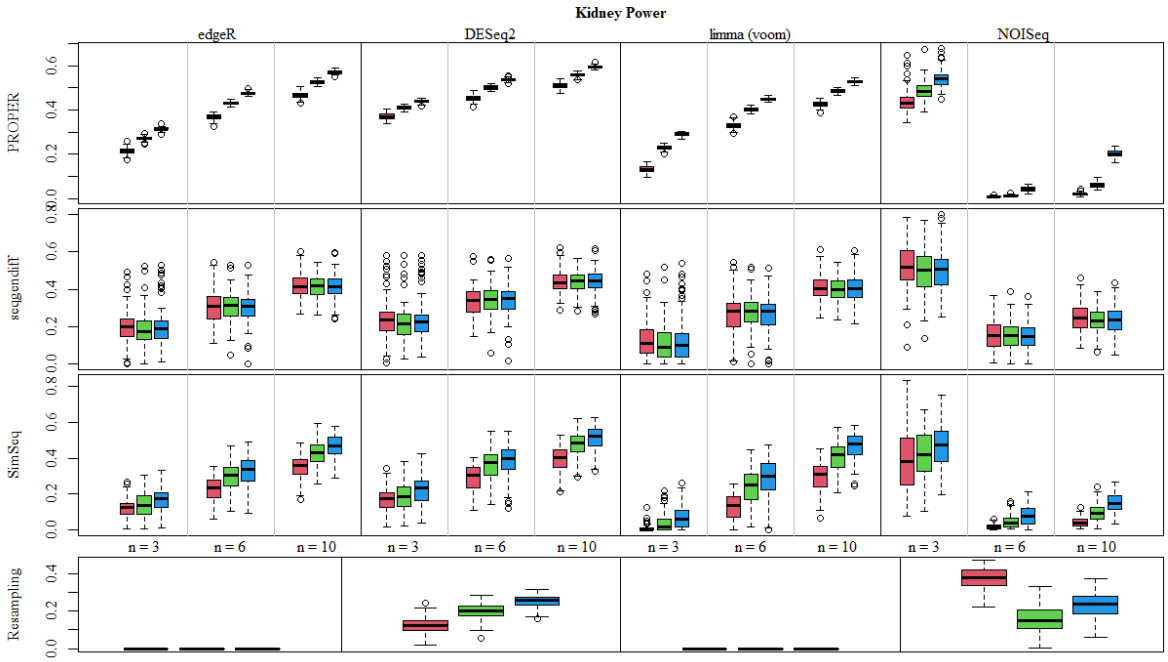

Figure 1. Observed Powers for Kidney Data (Figure by author). Box-and-whisker plots of power by sample size for DGE analysis methods. From left to right for each DGE analysis method and sample size combination: pink ($p = 0.9$), green ($p = 0.7$), blue ($p = 0.5$) for PROPER, seqgendiff, and SimSeq. From left to right for each DGE analysis method: pink ($n = 3$), green ($n = 6$), and blue ($n = 10$) for the resampling simulation method.

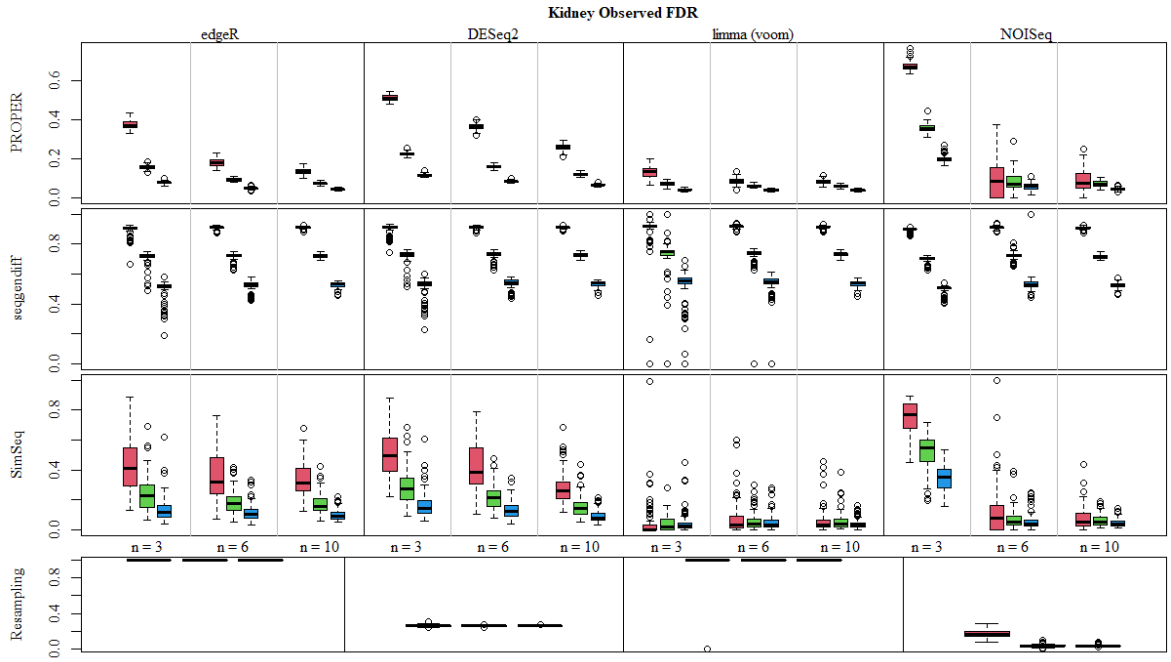

Figure 2. Observed FDRs for Kidney Data (Figure by author). Box-and-whisker plots of observed FDRs by sample size for DGE analysis methods. From left to right for each DGE analysis method and sample size combination: pink ($p = 0.9$), green ($p = 0.7$), blue ($p = 0.5$) for PROPER, seqgendiff, and SimSeq. From left to right for each DGE analysis method: pink ($n = 3$), green ($n = 6$), and blue ($n = 10$) for the resampling simulation method.

Original Read Count Data

To calculate observed powers and FDRs with the resampling method, each DGE analysis method needed to be run on the full dataset, which is referred to in Schurch et al. (2016) as the set of "clean" replicates. For the kidney data, DESeq2 declared 15161 genes to be DE. edgeR yielded 15120 "true" DE genes. limma with the voom method had 14818 DE genes. NOISeq yielded 11706 DE genes.

Whole Blood Data

PROPER ($p = 1$)

The results for this setting are listed in table 17. Observed FDRs tended to decrease with an increase in sample size. The DGE analysis method limma-voom yielded the lowest overall

observed FDR of 0.14 at $n = 10$. Standard errors of the observed FDRs increased as sample size increased for edgeR and NOISeq, but they decreased or stayed the same for the DESeq2 and voom methods. The voom method may be the most appropriate method to use for large samples at this setting, according to the table, because of its FDR control.

PROPER ($p = 0.9$)

The results for this setting are listed in table 18. Power tended to increase as sample size increased for the DGE analysis methods. Standard errors of these estimates increased with increased sample size for edgeR and limma-voom, but they decreased with increased sample size for DESeq2 and NOISeq. Observed FDRs tended to decrease with an increase in sample size. NOISeq yielded the lowest observed FDRs of 0.0729 ($n = 6$) and 0.0559 ($n = 10$). Corresponding powers for these observed FDRs are 0.0439 and 0.1143. The voom method controlled FDRs adequately with observed FDRs of 0.0880 ($n = 3$), 0.0768 ($n = 6$), and 0.0800 ($n = 10$). Corresponding powers for these estimates are 0.2416, 0.4355, and 0.5256, respectively. The DGE analysis method edgeR controlled FDR adequately at $n = 10$ with an observed FDR of 0.0985, and its corresponding power was 0.5510. At $n = 6$, edgeR yielded an observed FDR of 0.1282 and a power of 0.4591. DESeq2 had an observed FDR of 0.1632 at $n = 10$ and a power of 0.5642. Standard errors of the observed FDRs decreased as sample size increased. limma-voom is recommended for small/moderate sample sizes because it controls FDR adequately at this setting. edgeR is recommended for large sample sizes because it yields the largest power with adequate FDR control.

PROPER ($p = 0.7$)

The results for this setting are shown in table 19. As the sample size increased, mean power increased for the DGE analysis methods. Standard errors of these estimates tended to

decrease with an increase in sample size. NOISeq controlled FDR adequately with observed FDRs of 0.0491 ($n = 6$) and 0.0339 ($n = 10$). The corresponding powers are 0.1251 and 0.2556. The limma-voom method controlled FDR adequately with observed FDRs of 0.0645 ($n = 3$), 0.0589 ($n = 6$), and 0.0571 ($n = 10$). The corresponding powers for the observed FDRs are 0.3659, 0.5166, and 0.5908, respectively. The DGE analysis method edgeR controlled FDR adequately with observed FDRs of 0.0677 ($n = 6$) and 0.0565 ($n = 10$). Corresponding powers for edgeR were 0.5296 and 0.6128, respectively. The edgeR method observed FDR at $n = 3$ was 0.1018. The corresponding power for this FDR was 0.3784. DESeq2 controlled FDR adequately for $n = 10$ with an observed FDR of 0.0829 and a corresponding power of 0.6156. Observed FDRs for other sample sizes were 0.1419 ($n = 3$) and 0.1019 ($n = 6$). These corresponding powers were 0.4413 and 0.5541. Standard errors of the observed FDRs tended to decrease with an increase in sample size. The DGE method limma-voom is recommended for small sample sizes because it controls FDR well for this setting. edgeR is recommended for moderate sample sizes because it yields the largest power for methods that control FDR adequately. DESeq2 is recommended for large sample sizes because it yields the highest power for methods that control FDR well.

PROPER ($p = 0.5$)

The results of this simulation setting are found in table 20. Mean power increased as sample size increased. Standard errors of the powers decreased or stayed the same as sample size increased. Observed FDRs tended to roughly stay the same or slightly decrease as sample size increased. NOISeq yielded the lowest observed FDRs of 0.0304 ($n = 6$) and 0.0223 ($n = 10$). The corresponding powers for these observed FDRs are 0.2631 and 0.3672. The voom method controlled FDR adequately with observed FDRs of 0.0377 ($n = 3$), 0.0371 ($n = 6$), 0.0393 ($n =$

10). The corresponding powers are 0.4300, 0.5628, and 0.6340. The DGE analysis method edgeR also controlled FDR adequately with observed FDRs of 0.0533 ($n = 3$), 0.0398 ($n = 6$), and 0.0362 ($n = 10$). Corresponding powers for edgeR are 0.4353, 0.5805, and 0.6559. DESeq2 controlled FDR adequately with observed FDRs of 0.0714 ($n = 3$), 0.0541 ($n = 6$), and 0.0482 ($n = 10$). Corresponding powers for DESeq2 are 0.4810, 0.5911, and 0.6497. Standard errors of the observed FDRs tended to decrease with sample size increases. DESeq2 is recommended for small/moderate sample sizes because it yielded the largest powers for methods that adequately controlled FDR. edgeR is recommended for large sample sizes because it yielded the highest power for methods that controlled FDR well.

seqgendiff ($p = 1$, $p = 0.9$)

The results for these simulation settings were listed in tables 21 and 22. No DGE analysis method adequately controlled FDR at any sample size. Observed FDRs tended to decrease as sample size increased. Standard errors of the observed FDRs increased or stayed the same as sample size increased.

seqgendiff ($p = 0.7$)

The results for this simulation setting are in table 23. No DGE analysis method adequately controlled FDR at any sample size. The limma-voom method yielded the lowest observed FDRs of 0.1029 ($n = 3$), 0.1103 ($n = 6$), and 0.1365 ($n = 10$). With the exception of DESeq2, standard errors of the observed FDRs increased as sample size increased.

seqgendiff ($p = 0.5$)

The results for this setting are shown in table 24. The voom method controlled FDR adequately with observed FDRs of 0.0397 ($n = 3$) and 0.0990 ($n = 6$). The corresponding powers for these observed FDRs are 0.0375 and 0.0079. At $n = 10$, the limma-voom observed

FDR was 0.1123, and its corresponding power is 0.0068. With the exception of DESeq2, standard errors of the observed FDRs increased as sample size increased. limma-voom is recommended for small/moderate sample sizes because it was the only DGE method that controlled FDR adequately.

SimSeq ($p = 1$)

Results for this simulation setting are listed in table 25. Observed FDRs had an overall decreasing trend as sample size increased for each DGE analysis tool. The voom method controlled FDR adequately for $n = 6$ with an observed FDR of 0.04. The observed FDRs for other sample sizes of limma-voom were 0.18 ($n = 3$) and 0.10 ($n = 10$). The edgeR, NOISeq, and DESeq2 methods' standard errors of the observed FDRs increased as sample size increased, but the limma-voom method's standard errors stayed roughly the same. The voom method is recommended for moderate/large sample sizes at this setting because FDR is controlled adequately.

SimSeq ($p = 0.9$)

The simulation results are displayed in table 26 for this setting. Mean power increased with increased sample size for the DGE analysis methods. Standard errors of these statistics increased as sample size increased with the exception of NOISeq. With the exception of the voom method, observed FDRs tended to increase with an increase in sample size. The limma-voom method controlled FDR adequately with observed FDRs of 0.0577 ($n = 6$) and 0.0379 ($n = 10$). Corresponding powers for the voom FDRs are 0.0026 and 0.0031. limma-voom is recommended for moderate/large sample sizes because it controlled FDR well at this simulation setting.

SimSeq ($p = 0.7$)

Table 27 lists the simulation results for this setting. Mean power increased with increased sample size for each DGE analysis tool. With the exception of NOISeq, standard errors of the powers increased for the DGE analysis methods as sample size increased. Observed FDRs tended to decrease as the sample size increased. The limma-voom method controlled FDR adequately with observed FDRs of 0.0257 ($n = 6$) and 0.0632 ($n = 10$). The corresponding powers for these observed FDRs are 0.0034 and 0.0093. At $n = 3$, the voom method yielded an observed FDR of 0.1117 and a corresponding power of 0.0001. The observed FDR at $n = 10$ for NOISeq was 0.1592 and yielded a corresponding power of 0.0039. For $n = 10$, DESeq2 yielded an observed FDR of 0.1900 and a corresponding power of 0.0251. limma-voom is recommended for moderate/large sample sizes because it was the only DGE method that controlled FDR adequately.

SimSeq ($p = 0.5$)

Simulation results for this setting are displayed in table 28. Power increased for the DGE analysis methods as the sample size increased. With the exception of NOISeq, standard errors of the powers increased as sample size increased. Observed FDRs tended to decrease as sample size increased. The DGE analysis method limma-voom controlled FDR adequately with observed FDRs of 0.0096 ($n = 6$) and 0.0279 ($n = 10$). Corresponding powers for voom are 0.0020 and 0.0157, respectively. At $n = 3$, limma-voom yielded an observed FDR of 0.1332 and a corresponding power of 0.0012. Observed FDRs for NOISeq were 0.1613 ($n = 6$) and 0.0985 ($n = 10$) with corresponding powers of 0.0029 and 0.0064, respectively. The DESeq2 observed FDR was 0.1251 for $n = 10$, and its corresponding power is 0.0298. With the exception of NOISeq, standard errors of the observed FDRs decrease with increased sample size. The DGE

method limma-voom is recommended for moderate/large sample sizes because it controls FDR adequately and yields the largest powers.

Resampling

The simulation results for this setting are in table 29. As sample size increased, mean power increased. As sample size increased, observed FDRs tended to decrease. NOISeq yielded an observed FDR of 0.0994 at $n = 10$ with a corresponding power of 0.0273. NOISeq is recommended for large sample sizes for this simulation setting for adequate FDR control.

Boxplots

The PROPER power box-and-whisker plots are shown in figure 3, and the observed FDR box-and-whisker plots are shown in figure 4. Parametric DGE methods yielded a low spread of data regardless of the proportion of EE genes, p, and displayed a consistent upward trend for the power, again. NOISeq yielded a high spread for the power boxplots for low replicates, also. It also displayed an interesting stair step pattern for the power in contrast to the kidney data. There was a smaller spread for NOISeq regarding high proportions of equivalent expression for observed FDRs.

The seqgendiff power box-and-whisker plots are shown in figure 3, and the observed FDR box-and-whisker plots are shown in figure 4. Powers were quite variable regardless of sample size or proportion of equivalent expression, p. The only DGE analysis method that showed some semblance of consistency for power in terms of increasing sample size and decreasing p was DESeq2. Observed FDRs exhibited a downward trend for the DGE methods but did not appear to be controlled well by any method, although, DESeq2 showed better consistency.

The SimSeq power box-and-whisker plots are shown in figure 3, and the observed FDR box-and-whisker plots are shown in figure 4. All DGE analysis methods were consistent in producing powers that increased with both increased sample size and decreased proportion of equivalent expression. The observed FDRs were variable, but they were more consistent in being controlled better at lower proportions of equivalent expression.

The resampling power box-and-whisker plots are shown in figure 3, and the observed FDR box-and-whisker plots are shown in figure 4. NOISeq was the only DGE analysis method to control FDR well at some sample size. It performed fairly well for larger sample sizes.

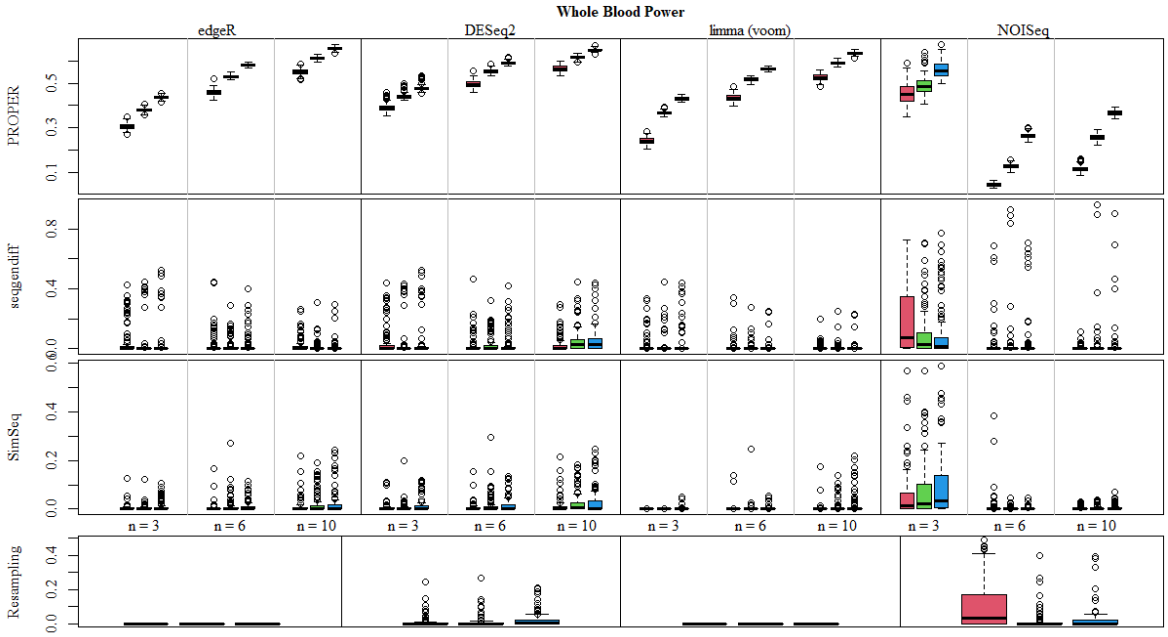

Figure 3. Observed Powers for Whole Blood Data (Figure by author). Box-and-whisker plots of power by sample size for DGE analysis methods. From left to right for each DGE analysis method and sample size combination: pink ($p = 0.9$), green ($p = 0.7$), blue ($p = 0.5$) for PROPER, seqgendiff, and SimSeq. From left to right for each DGE analysis method: pink ($n = 3$), green ($n = 6$), and blue ($n = 10$) for the resampling simulation method.

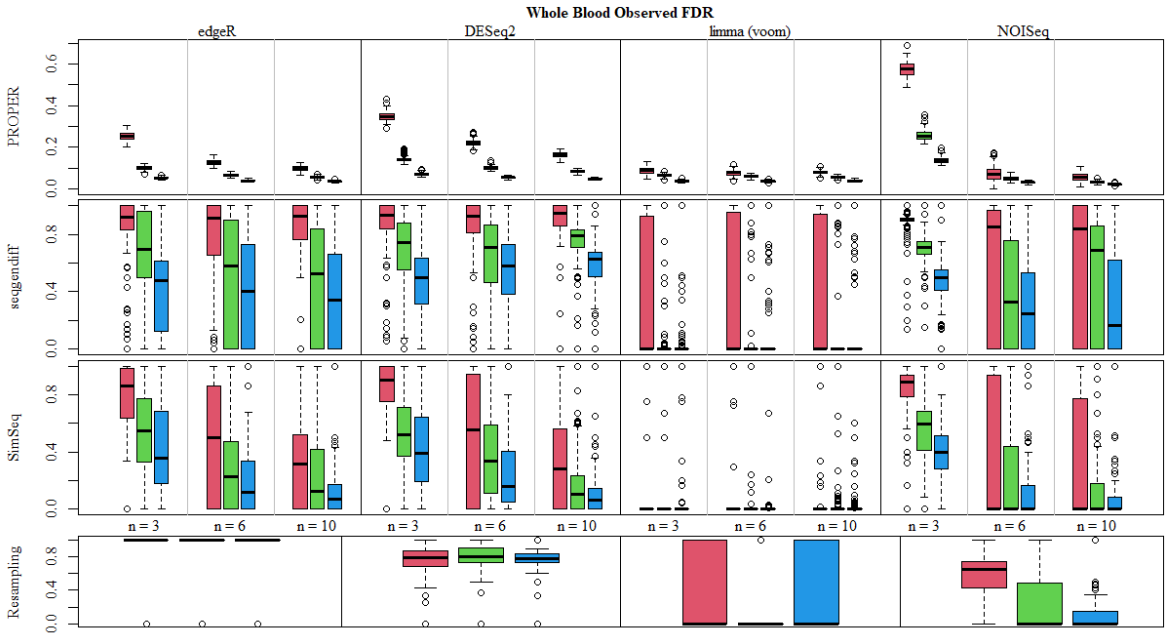

Figure 4. Observed FDRs for Whole Blood Data (Figure by author). Box-and-whisker plots of observed FDRs by sample size for DGE analysis methods. From left to right for each DGE analysis method and sample size combination: pink ($p = 0.9$), green ($p = 0.7$), blue ($p = 0.5$) for PROPER, seqgendiff, and SimSeq. From left to right for each DGE analysis method: pink ($n = 3$), green ($n = 6$), and blue ($n = 10$) for the resampling simulation method.

Original Read Count Data

For the whole blood data, DESeq2 declared 3924 genes to be DE. NOISeq yielded 3009 "true" DE genes. edgeR had 1153 DE genes. limma-voom yielded 1151 DE genes.

CHAPTER 5. DISCUSSION

Hypotheses

In general, it was hypothesized that when sample size was increased that power would also increase. It was also hypothesized that lowering the proportion of EE genes would increase the power of the test and the FDR would decrease.

For the data simulated with PROPER, it was expected that DGE analysis tools following a NB parametric model (edgeR and DESeq2) would perform the best in terms of DE and power analysis. These methods would also produce the lowest FDRs, and no guesses were made on the rankings of said methods. Previous research tells us that the observed FDR of parametric methods (edgeR, DESeq2, limma-voom) is "typically lower" than the theoretical FDR (Benidt and Nettleton 2015). NOISeq was expected to produce the lowest powers.

For the data simulated with SimSeq and seqgendiff, it was expected that there would be a lot of variability within the individual datasets themselves, so there would naturally be more variance in the observed FDR and power estimations. NOISeq was expected to perform fairly well regardless of sample size due to the nonparametric assumptions this method uses. The researcher was not sure in regards to violations of model assumptions how robust the DGE analysis parametric methods were.

Recommendations

The recommendations for simulation settings are given for datasets with similar compositions to the originals and similar to the datasets simulated. Results may vary for other researchers.

PROPER

For datasets in which gene expressions follow a NB distribution, using parametric DGE analysis methods seem to perform as expected. The edgeR method seems to be a little too liberal in estimating FDR for the smaller sample sizes and larger proportions of equivalent expression. For moderate to larger sample sizes, the FDR control is quite consistent across proportions, p. Relative to limma-voom, edgeR yields better power with only a small disadvantage in FDR control as the proportion of EE genes is low. When the proportion of EE genes is high, however, the disparity is more pronounced, and it might be better to opt toward using the voom method instead.

The DESeq2 method provides a greater power than both the edgeR or limma-voom methods alike, but it is too liberal in its control of FDR for most settings. For larger sample sizes and lower proportions of EE genes, it is expected that it overtakes edgeR and limma-voom in both FDR control and higher power.

The limma-voom method is quite consistent in controlling FDR adequately at all proportions of EE genes (with more liberal estimates for $p = 1$, but still fairly conservative generally) due to the lower number of DDE genes identified (see the "Original Read Count Data" subsections in the "Results" section). Using the voom method is a safe bet for controlling FDR at all proportions, p, with a small sacrifice in power.

NOISeq does a decent job of controlling FDR at moderate and large sample sizes regardless of the proportion of EE genes (p), but the power is greatly sacrificed due to the method's lack of distributional assumptions. The k-means clustering method of genes for low sample sizes yields a large variance for all proportion of EE genes settings, but based on the

decreasing observed FDRs, it was expected to be controlled adequately for some $p < 0.5$ and be the superior method for low replicates.

seqgendiff

It is difficult to make specific recommendations for the seqgendiff simulation methods both due to the varied results in simulated datasets between the types of genes examined (kidney and whole blood) and the unpredictability of how the DGE analysis methods will perform for data of different structures. For the kidney data, observed FDRs were much too liberal and not controlled well, but they were quite consistent regardless of sample size and DGE analysis method. For the whole blood data, unusually, mean powers decreased with sample size. The limma-voom method seemed to control FDR fairly well for most proportions, p, while the mean observed FDRs for the other DGE analysis methods were too liberal with FDR control (although somewhat better than the kidney data). Surprisingly, it did not appear that NOISeq would perform better any better than a parametric DGE analysis method for the nonparametric simulated seqgendiff data. In general, it would be wise to use seqgendiff for low proportions of EE genes for good FDR control. Other gene-filtering methods that are specific to the original dataset used may need to be implemented to yield more consistent results.

SimSeq

More inference can be made on datasets simulated with the nonparametric SimSeq method. Based on adequate FDR control and high power (for the kidney data), it can be said that the limma-voom method of DGE analysis is quite robust to the potential violation of NB model assumptions while edgeR and DESeq2 require larger sample sizes and/or smaller proportions of EE genes. Again, it was surprising to see that NOISeq did not perform as well as expected for the SimSeq simulated data relative to the parametric DGE analysis methods. As with seqgendiff,

further experimentation could be done on gene-filtering approaches in a more general sense or more specific to the particular RNA-seq dataset.

Resampling

NOISeq performed the best for all sample sizes as it adequately controlled FDR relative to the other parametric DGE analysis methods.

Future Recommendations

The choice for the number of simulations was arbitrary and could have just as easily been increased to 1000 or 10000. Other simulation methods such as those listed in the "Literature Review" section that were not implemented could be explored in other studies. Different datasets could also be explored.

Limitations

Batch Effect

Unfortunately, some batch effects could not be accounted for with the DGE analysis tools. There were three main reasons for this: (1) the design matrix was not full rank, (2) the simulation method did not return column sample names, and (3) the DGE analysis method did not accept a design matrix as input. For small/moderate sample sizes ($n = 3$ and sometimes $n = 6$), samples were selected such that at least two of the columns of the design matrix were linearly dependent. For each simulation out of 100, the design matrix needed to be full rank for the DGE analysis methods to use it. Multiple runs of a simulation method at a particular setting could ensure that all design matrices were full rank, but this approach is time-consuming. Thus, if at least one design matrix had linearly dependent columns, the researcher used the matrix given in equation 1 rather than rerun the simulations.

Simulation methods like PROPER and SimSeq did not return sample names of the samples selected. If l is denoted as the number of samples selected to generate the count matrix, the number of samples used in the SimSeq package was $l = 3n$. This implied that at each sample size setting there were 9, 18, and 30 replicates, respectively, actually used in the final read counts matrix. This discrepancy between the n samples selected and the l samples used did not allow us to have conformable arguments for the model matrix.

NOISeq did not take input for a design matrix, so the batch effect was not accounted for with this DGE analysis method.

DE Genes

For the resampling simulation method, DE genes were not predefined in the simulated datasets. As a result, it could not be ascertained whether the identification of DE genes was dependent on the DGE analysis method, dependent on the random sampling of genes outlined in the "Filtering" section, or dependent on both DGE analysis and gene sampling. The same could be said for the other simulation methods (especially the nonparametric ones) because the filtering strategies seemed to impact the identification of DE genes more than was anticipated.

As mentioned in the "Introduction" section, the proportion of EE genes cannot be prespecified in practice. Hence, experimentation was conducted with multiple proportions to see if there were significant differences across simulation settings. Observed FDRs and powers that yield consistent results for differing proportions of equivalent expression (holding everything else constant) are the most helpful in making general recommendations on methods to use for researchers.

Conclusion

It is difficult to say whether parametric simulation methods following the Poisson and NB distributions truly reflect the structure of most RNA-seq data in contrast to the nonparametric simulation methods. Either way, the parametric simulation methods are most widely used in practice.

As expected, parametric DGE analysis methods performed best with the parametric simulation model assumptions met. The limma-voom method may be the preferred DGE analysis tool for both parametric and nonparametric simulation methods because it seems to be consistent and robust to model assumption violations as opposed to other DGE methods.

After running the simulations, results differed considerably by the type of RNA-seq data in terms of simulation method and DGE analysis method. While it cannot be determined how generalizable these particular results are to kidney and whole blood data, respectively, there are differences between datasets that are unaccounted for, especially for nonparametric simulation methods.

REFERENCES

Anders, Simon, and Wolfgang Huber. "Differential Expression Analysis for Sequence Count Data." *Genome Biology* 11, no. R106 (October 27, 2010). doi:10.1038/npre.2010.4282.2.

Benidt, Sam, and Dan Nettleton. "SimSeq: A Nonparametric Approach to Simulation of RNA-sequence Datasets." *Bioinformatics* 31, no. 13 (July 1, 2015): 2131-140. doi:10.1093/bioinformatics/btv124.

Benjamini, Yoav, and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57, no. 1 (1995): 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

Davis, Sean, and Paul S. Meltzer. "GEOquery: A Bridge between the Gene Expression Omnibus (GEO) and BioConductor." *Bioinformatics* 23, no. 14 (July 15, 2007): 1846-847. doi:10.1093/bioinformatics/btm254.

Edgar, Ron, Michael Domrachev, and Alex E. Lash. "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository." *Nucleic Acids Research* 30, no. 1 (January 01, 2002): 207-10. doi:10.1093/nar/30.1.207.

Frazee, Alyssa C., Andrew E. Jaffe, Ben Langmead, and Jeffrey T. Leek. "Polyester: Simulating RNA-seq Datasets with Differential Transcript Expression." *Bioinformatics* 31, no. 17 (September 1, 2015): 2778-784. doi:10.1093/bioinformatics/btv272.

Gentleman, Robert C., Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J. Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean YH Yang, and Jianhua Zhang. "Bioconductor: Open Software Development for Computational Biology and Bioinformatics." *Genome Biology* 5, no. R80 (September 15, 2004). doi:10.1186/gb-2004-5-10-r80.

Gerard, David. "Data-based RNA-seq Simulations by Binomial Thinning." *BMC Bioinformatics* 21, no. 206 (May 24, 2020). doi:10.1186/s12859-020-3450-9.

Hardcastle, Thomas J., and Krystyna A. Kelly. "BaySeq: Empirical Bayesian Methods for Identifying Differential Expression in Sequence Count Data." *BMC Bioinformatics* 11, no. 422 (August 10, 2010). doi:10.1186/1471-2105-11-422.

Huber, Wolfgang, Vincent J. Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S. Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, Raphael Gottardo, Florian Hahne, Kasper D. Hansen, Rafael A. Irizarry, Michael Lawrence,

Michael I. Love, James Macdonald, Valerie Obenchain, Andrzej K. Oleś, Hervé Pagès, Alejandro Reyes, Paul Shannon, Gordon K. Smyth, Dan Tenenbaum, Levi Waldron, and Martin Morgan. "Orchestrating High-throughput Genomic Analysis with Bioconductor." *Nature Methods* 12 (January 29, 2015): 115-21. doi:10.1038/nmeth.3252.

Law, Charity W., Yunshun Chen, Wei Shi, and Gordon K. Smyth. "Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-seq Read Counts." *Genome Biology* 15, no. R29 (February 3, 2014). doi:10.1186/gb-2014-15-2-r29.

Love, Michael I., Wolfgang Huber, and Simon Anders. "Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2." *Genome Biology* 15, no. 12 (December 5, 2014). doi:10.1186/s13059-014-0550-8.

Mccarthy, Davis J., Yunshun Chen, and Gordon K. Smyth. "Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation." *Nucleic Acids Research* 40, no. 10 (May 1, 2012): 4288-297. doi:10.1093/nar/gks042.

R Core Team. R: A Language and Environment for Statistical Computing. Computer software. 2020. http://www.R-project.org/.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. "Limma Powers Differential Expression Analyses for RNA-sequencing and Microarray Studies." *Nucleic Acids Research* 43, no. 7 (April 20, 2015): E47. doi:10.1093/nar/gkv007.

Robinson, Mark D., and Alicia Oshlack. "A Scaling Normalization Method for Differential Expression Analysis of RNA-seq Data." *Genome Biology* 11, no. 3 (March 2, 2010). doi:10.1186/gb-2010-11-3-r25.

Robinson, Mark D., Davis J. Mccarthy, and Gordon K. Smyth. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26, no. 1 (January 1, 2010): 139-40. doi:10.1093/bioinformatics/btp616.

Schurch, Nicholas J., Pietá Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Gharbi, Gordon G. Simpson, Tom Owen-Hughes, Mark Blaxter, and Geoffrey J. Barton. "How Many Biological Replicates Are Needed in an RNA-seq Experiment and Which Differential Expression Tool Should You Use?" *RNA* 22 (February 17, 2016): 839-51. doi:10.1261/rna.053959.115.

Tarazona, Sonia, Pedro Furió-Tarí, David Turrà, Antonio Di Pietro, María José Nueda, Alberto Ferrer, and Ana Conesa. "Data Quality Aware Analysis of Differential Expression in RNA-seq with NOISeq R/Bioc Package." *Nucleic Acids Research* 43, no. 21 (December 2, 2015): E140. doi:10.1093/nar/gkv711.

Tarazona, Sonia, Fernando García-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. "Differential Expression in RNA-seq: A Matter of Depth." *Genome Research* 21 (September 8, 2011): 2213-223. doi:10.1101/gr.124321.111.

Vieth, Beate, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. "PowsimR: Power Analysis for Bulk and Single Cell RNA-seq Experiments." *Bioinformatics* 33, no. 21 (November 1, 2017): 3486-488. doi:10.1093/bioinformatics/btx435.

Wu, Hao, Chi Wang, and Zhijin Wu. "A New Shrinkage Estimator for Dispersion Improves Differential Expression Detection in RNA-seq Data." *Biostatistics* 14, no. 2 (September 22, 2012): 232-43. doi:10.1093/biostatistics/kxs033.

Wu, Hao, Chi Wang, and Zhijin Wu. "PROPER: Comprehensive Power Evaluation for Differential Expression Using RNA-seq." *Bioinformatics* 31, no. 2 (October 1, 2014): 233-41. doi:10.1093/bioinformatics/btu640.

Zhu, Hao. KableExtra: Construct Complex Table with Kable and Pipe Syntax. Computer software. KableExtra: Construct Complex Table with Knitr::kable() Pipe. May 25, 2021. https://github.com/haozhu233/kableExtra.

APPENDIX. TABLES

Table A1. PROPER Observed FDR ($p = 1$) for Kidney Data.

	Mean Observed FDR	SE Observed FDR
edgeR		
n = 3	1.00	0.0000
n = 6	1.00	0.0000
n = 10	0.85	0.0359
DESeq2		
n = 3	1.00	0.0000
n = 6	1.00	0.0000
n = 10	1.00	0.0000
limma (voom)		
n = 3	0.54	0.0501
n = 6	0.25	0.0435
n = 10	0.16	0.0368
NOISeq		
n = 3	1.00	0.0000
n = 6	0.63	0.0485
n = 10	0.71	0.0456

Table A2. PROPER Power and Observed FDR ($p = 0.9$) for Kidney Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.2140	0.0014	0.3760	0.0023
n = 6	0.3677	0.0013	0.1815	0.0019
n = 10	0.4661	0.0014	0.1353	0.0017
DESeq2				
n = 3	0.3692	0.0015	0.5133	0.0014
n = 6	0.4536	0.0013	0.3661	0.0016
n = 10	0.5096	0.0014	0.2610	0.0018
limma (voom)				
n = 3	0.1306	0.0016	0.1318	0.0029
n = 6	0.3304	0.0014	0.0861	0.0016
n = 10	0.4250	0.0014	0.0837	0.0012
NOISeq				
n = 3	0.4403	0.0052	0.6764	0.0021
n = 6	0.0082	0.0003	0.0936	0.0102
n = 10	0.0193	0.0006	0.0845	0.0058

Table A3. PROPER Power and Observed FDR ($p = 0.7$) for Kidney Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.2715	0.0008	0.1594	0.0011
n = 6	0.4306	0.0007	0.0944	0.0008
n = 10	0.5243	0.0008	0.0736	0.0007
DESeq2				
n = 3	0.4090	0.0008	0.2285	0.0009
n = 6	0.5005	0.0009	0.1601	0.0008
n = 10	0.5578	0.0008	0.1212	0.0007
limma (voom)				
n = 3	0.2290	0.0010	0.0722	0.0009
n = 6	0.4018	0.0009	0.0623	0.0007
n = 10	0.4842	0.0008	0.0602	0.0007
NOISeq				
n = 3	0.4860	0.0042	0.3596	0.0020
n = 6	0.0127	0.0004	0.0796	0.0048
n = 10	0.0624	0.0013	0.0701	0.0016

Table A4. PROPER Power and Observed FDR ($p = 0.5$) for Kidney Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.3138	0.0007	0.0794	0.0006
n = 6	0.4756	0.0006	0.0507	0.0005
n = 10	0.5691	0.0007	0.0452	0.0004
DESeq2				
n = 3	0.4392	0.0008	0.1175	0.0006
n = 6	0.5360	0.0006	0.0848	0.0005
n = 10	0.5927	0.0006	0.0675	0.0005
limma (voom)				
n = 3	0.2907	0.0008	0.0439	0.0005
n = 6	0.4478	0.0006	0.0409	0.0004
n = 10	0.5273	0.0007	0.0406	0.0004
NOISeq				
n = 3	0.5426	0.0041	0.2020	0.0017
n = 6	0.0426	0.0010	0.0601	0.0017
n = 10	0.2024	0.0015	0.0459	0.0007

Table A5. seqgendiff Observed FDR ($p = 1$) for Kidney Data.

	Mean Observed FDR	SE Observed FDR
edgeR		
n = 3	1.00	0.0000
n = 6	1.00	0.0000
n = 10	1.00	0.0000
DESeq2		
n = 3	1.00	0.0000
n = 6	1.00	0.0000
n = 10	1.00	0.0000
limma (voom)		
n = 3	0.92	0.0273
n = 6	1.00	0.0000
n = 10	1.00	0.0000
NOISeq		
n = 3	1.00	0.0000
n = 6	1.00	0.0000
n = 10	1.00	0.0000

Table A6. seqgendiff Power and Observed FDR ($p = 0.9$) for Kidney Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.2021	0.0098	0.8985	0.0035
n = 6	0.3028	0.0083	0.9096	0.0009
n = 10	0.4191	0.0067	0.9106	0.0008
DESeq2				
n = 3	0.2411	0.0106	0.9043	0.0030
n = 6	0.3370	0.0082	0.9131	0.0009
n = 10	0.4414	0.0064	0.9128	0.0007
limma (voom)				
n = 3	0.1296	0.0105	0.8411	0.0247
n = 6	0.2664	0.0095	0.9172	0.0010
n = 10	0.4093	0.0071	0.9134	0.0008
NOISeq				
n = 3	0.5201	0.0119	0.8990	0.0011
n = 6	0.1550	0.0083	0.9108	0.0010
n = 10	0.2494	0.0074	0.9082	0.0008

Table A7. seqgendiff Power and Observed FDR ($p = 0.7$) for Kidney Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.1769	0.0088	0.7108	0.0042
n = 6	0.3079	0.0080	0.7211	0.0023
n = 10	0.4137	0.0062	0.7229	0.0015
DESeq2				
n = 3	0.2168	0.0095	0.7223	0.0041
n = 6	0.3454	0.0080	0.7312	0.0023
n = 10	0.4388	0.0059	0.7290	0.0015
limma (voom)				
n = 3	0.1055	0.0094	0.6613	0.0245
n = 6	0.2780	0.0089	0.7291	0.0078
n = 10	0.4023	0.0065	0.7309	0.0017
NOISeq				
n = 3	0.4890	0.0115	0.7026	0.0015
n = 6	0.1518	0.0075	0.7241	0.0023
n = 10	0.2371	0.0070	0.7178	0.0014

Table A8. seqgendiff Power and Observed FDR ($p = 0.5$) for Kidney Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.1982	0.0102	0.5018	0.0065
n = 6	0.3038	0.0080	0.5264	0.0028
n = 10	0.4129	0.0067	0.5279	0.0018
DESeq2				
n = 3	0.2399	0.0110	0.5160	0.0064
n = 6	0.3416	0.0083	0.5394	0.0029
n = 10	0.4382	0.0066	0.5359	0.0019
limma (voom)				
n = 3	0.1251	0.0117	0.5051	0.0148
n = 6	0.2710	0.0090	0.5400	0.0064
n = 10	0.3994	0.0073	0.5380	0.0021
NOISeq				
n = 3	0.5022	0.0114	0.5001	0.0025
n = 6	0.1494	0.0070	0.5340	0.0054
n = 10	0.2356	0.0074	0.5236	0.0018

Table A9. SimSeq Observed FDR ($p = 1$) for Kidney Data.

	Mean Observed FDR	SE Observed FDR
edgeR		
n = 3	0.98	0.0141
n = 6	0.92	0.0273
n = 10	0.96	0.0197
DESeq2		
n = 3	1.00	0.0000
n = 6	1.00	0.0000
n = 10	1.00	0.0000
limma (voom)		
n = 3	0.04	0.0197
n = 6	0.05	0.0219
n = 10	0.03	0.0171
NOISeq		
n = 3	1.00	0.0000
n = 6	0.67	0.0473
n = 10	0.65	0.0479

Table A10. SimSeq Power and Observed FDR ($p = 0.9$) for Kidney Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.1215	0.0050	0.4250	0.0164
n = 6	0.2251	0.0066	0.3528	0.0153
n = 10	0.3485	0.0066	0.3358	0.0108
DESeq2				
n = 3	0.1705	0.0061	0.5100	0.0144
n = 6	0.2901	0.0071	0.4197	0.0152
n = 10	0.3955	0.0071	0.2801	0.0107
limma (voom)				
n = 3	0.0111	0.0019	0.0466	0.0124
n = 6	0.1321	0.0071	0.0725	0.0104
n = 10	0.2934	0.0078	0.0583	0.0077
NOISeq				
n = 3	0.3878	0.0163	0.7552	0.0103
n = 6	0.0197	0.0014	0.1225	0.0159
n = 10	0.0420	0.0026	0.0742	0.0071

Table A11. SimSeq Power and Observed FDR ($p = 0.7$) for Kidney Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.1450	0.0068	0.2389	0.0112
n = 6	0.2964	0.0078	0.1880	0.0075
n = 10	0.4268	0.0068	0.1736	0.0065
DESeq2				
n = 3	0.1939	0.0079	0.2887	0.0114
n = 6	0.3617	0.0083	0.2204	0.0085
n = 10	0.4747	0.0071	0.1548	0.0071
limma (voom)				
n = 3	0.0386	0.0049	0.0412	0.0052
n = 6	0.2389	0.0099	0.0618	0.0062
n = 10	0.4062	0.0078	0.0560	0.0056
NOISeq				
n = 3	0.4239	0.0138	0.5208	0.0117
n = 6	0.0471	0.0034	0.0705	0.0064
n = 10	0.0950	0.0048	0.0647	0.0039

Table A12. SimSeq Power and Observed FDR ($p = 0.5$) for Kidney Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.1697	0.0066	0.1426	0.0086
n = 6	0.3220	0.0091	0.1165	0.0056
n = 10	0.4671	0.0065	0.1002	0.0035
DESeq2				
n = 3	0.2216	0.0079	0.1688	0.0086
n = 6	0.3830	0.0094	0.1358	0.0061
n = 10	0.5127	0.0065	0.0926	0.0037
limma (voom)				
n = 3	0.0738	0.0066	0.0466	0.0074
n = 6	0.2841	0.0114	0.0516	0.0053
n = 10	0.4633	0.0075	0.0408	0.0030
NOISeq				
n = 3	0.4697	0.0117	0.3354	0.0082
n = 6	0.0816	0.0053	0.0539	0.0050
n = 10	0.1510	0.0052	0.0463	0.0026

Table A13. Resampling Power and Observed FDR for Kidney Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.0000	0.0000	1.0000	0.0000
n = 6	0.0000	0.0000	1.0000	0.0000
n = 10	0.0000	0.0000	1.0000	0.0000
DESeq2				
n = 3	0.1212	0.0044	0.2603	0.0009
n = 6	0.1976	0.0041	0.2599	0.0006
n = 10	0.2511	0.0031	0.2585	0.0004
limma (voom)				
n = 3	0.0000	0.0000	0.9400	0.0239
n = 6	0.0000	0.0000	1.0000	0.0000
n = 10	0.0000	0.0000	1.0000	0.0000
NOISeq				
n = 3	0.3733	0.0057	0.1695	0.0043
n = 6	0.1525	0.0070	0.0321	0.0013
n = 10	0.2328	0.0062	0.0304	0.0009

Table A14. PROPER Observed FDR ($p = 1$) for Whole Blood Data.

	Mean Observed FDR	SE Observed FDR
edgeR		
n = 3	1.00	0.0000
n = 6	0.95	0.0219
n = 10	0.70	0.0461
DESeq2		
n = 3	1.00	0.0000
n = 6	1.00	0.0000
n = 10	1.00	0.0000
limma (voom)		
n = 3	0.33	0.0473
n = 6	0.35	0.0479
n = 10	0.14	0.0349
NOISeq		
n = 3	1.00	0.0000
n = 6	0.82	0.0386
n = 10	0.87	0.0338

Table A15. PROPER Power and Observed FDR ($p = 0.9$) for Whole Blood Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.3059	0.0013	0.2537	0.0022
n = 6	0.4591	0.0015	0.1282	0.0015
n = 10	0.5510	0.0015	0.0985	0.0012
DESeq2				
n = 3	0.3887	0.0020	0.3504	0.0025
n = 6	0.4957	0.0017	0.2222	0.0017
n = 10	0.5642	0.0015	0.1632	0.0014
limma (voom)				
n = 3	0.2416	0.0014	0.0880	0.0017
n = 6	0.4355	0.0015	0.0768	0.0014
n = 10	0.5256	0.0015	0.0800	0.0011
NOISeq				
n = 3	0.4504	0.0055	0.5755	0.0037
n = 6	0.0439	0.0008	0.0729	0.0038
n = 10	0.1143	0.0014	0.0559	0.0020

Table A16. PROPER Power and Observed FDR ($p = 0.7$) for Whole Blood Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.3784	0.0009	0.1018	0.0009
n = 6	0.5296	0.0008	0.0677	0.0006
n = 10	0.6128	0.0007	0.0565	0.0006
DESeq2				
n = 3	0.4413	0.0015	0.1419	0.0014
n = 6	0.5541	0.0011	0.1019	0.0009
n = 10	0.6156	0.0008	0.0829	0.0007
limma (voom)				
n = 3	0.3659	0.0008	0.0645	0.0007
n = 6	0.5166	0.0008	0.0589	0.0006
n = 10	0.5908	0.0008	0.0571	0.0006
NOISeq				
n = 3	0.4911	0.0045	0.2566	0.0027
n = 6	0.1251	0.0013	0.0491	0.0011
n = 10	0.2556	0.0015	0.0339	0.0007

Table A17. PROPER Power and Observed FDR ($p = 0.5$) for Whole Blood Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.4353	0.0007	0.0533	0.0005
n = 6	0.5805	0.0006	0.0398	0.0004
n = 10	0.6559	0.0006	0.0362	0.0004
DESeq2				
n = 3	0.4810	0.0016	0.0714	0.0008
n = 6	0.5911	0.0008	0.0541	0.0004
n = 10	0.6497	0.0007	0.0482	0.0004
limma (voom)				
n = 3	0.4300	0.0007	0.0377	0.0004
n = 6	0.5628	0.0006	0.0371	0.0004
n = 10	0.6340	0.0007	0.0393	0.0004
NOISeq				
n = 3	0.5607	0.0038	0.1386	0.0016
n = 6	0.2631	0.0014	0.0304	0.0005
n = 10	0.3672	0.0011	0.0223	0.0004

Table A18. seqgendiff Observed FDR ($p = 1$) for Whole Blood Data.

	Mean Observed FDR	SE Observed FDR
edgeR		
n = 3	0.98	0.0141
n = 6	0.86	0.0349
n = 10	0.80	0.0402
DESeq2		
n = 3	0.98	0.0141
n = 6	0.90	0.0302
n = 10	0.99	0.0100
limma (voom)		
n = 3	0.37	0.0485
n = 6	0.31	0.0465
n = 10	0.33	0.0473
NOISeq		
n = 3	0.97	0.0171
n = 6	0.64	0.0482
n = 10	0.70	0.0461

Table A19. seqgendiff Power and Observed FDR ($p = 0.9$) for Whole Blood Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.0454	0.0099	0.8064	0.0295
n = 6	0.0224	0.0071	0.7296	0.0375
n = 10	0.0202	0.0048	0.7586	0.0355
DESeq2				
n = 3	0.0502	0.0105	0.8459	0.0226
n = 6	0.0267	0.0077	0.7950	0.0313
n = 10	0.0254	0.0053	0.8181	0.0312
limma (voom)				
n = 3	0.0237	0.0066	0.3104	0.0441
n = 6	0.0107	0.0048	0.3056	0.0444
n = 10	0.0056	0.0023	0.3012	0.0447
NOISeq				
n = 3	0.1827	0.0210	0.8667	0.0146
n = 6	0.0310	0.0114	0.5319	0.0461
n = 10	0.0075	0.0020	0.5541	0.0465

Table A20. seqgendiff Power and Observed FDR ($p = 0.7$) for Whole Blood Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.0318	0.0104	0.6304	0.0331
n = 6	0.0205	0.0049	0.4956	0.0401
n = 10	0.0103	0.0037	0.4951	0.0405
DESeq2				
n = 3	0.0340	0.0105	0.6727	0.0295
n = 6	0.0284	0.0060	0.6253	0.0330
n = 10	0.0531	0.0080	0.7518	0.0159
limma (voom)				
n = 3	0.0195	0.0066	0.1029	0.0276
n = 6	0.0073	0.0034	0.1103	0.0289
n = 10	0.0059	0.0029	0.1365	0.0320
NOISeq				
n = 3	0.1012	0.0176	0.7067	0.0141
n = 6	0.0430	0.0176	0.3973	0.0405
n = 10	0.0291	0.0136	0.4875	0.0427

Table A21. seqgendiff Power and Observed FDR ($p = 0.5$) for Whole Blood Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.0412	0.0121	0.4238	0.0308
n = 6	0.0228	0.0066	0.4020	0.0367
n = 10	0.0105	0.0044	0.3639	0.0358
DESeq2				
n = 3	0.0451	0.0124	0.4820	0.0276
n = 6	0.0297	0.0074	0.5269	0.0296
n = 10	0.0593	0.0092	0.5900	0.0188
limma (voom)				
n = 3	0.0375	0.0101	0.0397	0.0138
n = 6	0.0079	0.0039	0.0990	0.0253
n = 10	0.0068	0.0035	0.1123	0.0271
NOISeq				
n = 3	0.1073	0.0188	0.4659	0.0175
n = 6	0.0448	0.0150	0.3302	0.0352
n = 10	0.0294	0.0127	0.3244	0.0347

Table A22. SimSeq Observed FDR ($p = 1$) for Whole Blood Data.

	Mean Observed FDR	SE Observed FDR
edgeR		
n = 3	0.94	0.0239
n = 6	0.63	0.0485
n = 10	0.46	0.0501
DESeq2		
n = 3	0.97	0.0171
n = 6	0.73	0.0446
n = 10	0.44	0.0499
limma (voom)		
n = 3	0.18	0.0386
n = 6	0.04	0.0197
n = 10	0.10	0.0302
NOISeq		
n = 3	0.98	0.0141
n = 6	0.63	0.0485
n = 10	0.64	0.0482

Table A23. SimSeq Power and Observed FDR ($p = 0.9$) for Whole Blood Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.0049	0.0014	0.7471	0.0301
n = 6	0.0049	0.0019	0.4624	0.0411
n = 10	0.0112	0.0030	0.3525	0.0339
DESeq2				
n = 3	0.0066	0.0019	0.8377	0.0203
n = 6	0.0061	0.0019	0.5371	0.0392
n = 10	0.0131	0.0033	0.3409	0.0332
limma (voom)				
n = 3	0.0000	0.0000	0.2325	0.0419
n = 6	0.0026	0.0018	0.0577	0.0221
n = 10	0.0031	0.0019	0.0379	0.0169
NOISeq				
n = 3	0.0578	0.0102	0.8109	0.0229
n = 6	0.0103	0.0048	0.3820	0.0453
n = 10	0.0037	0.0008	0.3376	0.0420

Table A24. SimSeq Power and Observed FDR ($p = 0.7$) for Whole Blood Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.0053	0.0014	0.5434	0.0294
n = 6	0.0122	0.0034	0.2977	0.0310
n = 10	0.0210	0.0040	0.2439	0.0280
DESeq2				
n = 3	0.0077	0.0021	0.5334	0.0232
n = 6	0.0157	0.0038	0.3630	0.0312
n = 10	0.0251	0.0043	0.1900	0.0243
limma (voom)				
n = 3	0.0001	0.0000	0.1117	0.0301
n = 6	0.0034	0.0025	0.0257	0.0143
n = 10	0.0093	0.0027	0.0632	0.0212
NOISeq				
n = 3	0.0751	0.0113	0.5409	0.0221
n = 6	0.0022	0.0006	0.2136	0.0334
n = 10	0.0039	0.0008	0.1592	0.0297

Table A25. SimSeq Power and Observed FDR ($p = 0.5$) for Whole Blood Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.0105	0.0022	0.4296	0.0301
n = 6	0.0108	0.0022	0.2193	0.0277
n = 10	0.0252	0.0051	0.1287	0.0174
DESeq2				
n = 3	0.0148	0.0029	0.4215	0.0274
n = 6	0.0154	0.0029	0.2710	0.0284
n = 10	0.0298	0.0054	0.1251	0.0183
limma (voom)				
n = 3	0.0012	0.0007	0.1332	0.0327
n = 6	0.0020	0.0009	0.0096	0.0070
n = 10	0.0157	0.0043	0.0279	0.0094
NOISeq				
n = 3	0.0948	0.0126	0.4077	0.0208
n = 6	0.0029	0.0007	0.1613	0.0315
n = 10	0.0064	0.0012	0.0985	0.0233

Table A26. Resampling Power and Observed FDR for Whole Blood Data.

	Mean Power	SE Power	Mean Observed FDR	SE Observed FDR
edgeR				
n = 3	0.0000	0.0000	0.9600	0.0197
n = 6	0.0000	0.0000	0.8700	0.0338
n = 10	0.0000	0.0000	0.9100	0.0288
DESeq2				
n = 3	0.0111	0.0033	0.7486	0.0214
n = 6	0.0130	0.0037	0.7772	0.0234
n = 10	0.0228	0.0044	0.7379	0.0218
limma (voom)				
n = 3	0.0000	0.0000	0.3200	0.0469
n = 6	0.0000	0.0000	0.1800	0.0386
n = 10	0.0000	0.0000	0.3300	0.0473
NOISeq				
n = 3	0.1080	0.0144	0.5724	0.0219
n = 6	0.0217	0.0063	0.2533	0.0338
n = 10	0.0273	0.0068	0.0994	0.0166