Compare RNA-Seq Differential Expression Analysis Methods

by

Xiyuan Sun

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Statistics

Program of Study Committee:

Jarad Niemi, Major Professor

Danniel Nettleton

Peng Liu

Iowa State University

Ames, Iowa

2019

DEDICATION

I would like to dedicate this thesis to my parents.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

This research was built on Niemi et al's approach (Niemi et al., 2015). Their research was supported by National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health and joint National Science Foundation / NIGMS Mathematical Biology Program under award number R01GM109458.

Abstract

We simulated RNA-Seq count data based on parameters estimated from a maize RNA-Seq dataset (Paschold et al., 2012). We comperehensively compared six differential expression (DE) analysis methods (eBayes, edgeR, DESeq2, DESeq, sSeq, and EBSeq) and evaluated their performance by receiver operator characteristic (ROC) curves and areas under the curve (AUC). eBayes tends to give the best performance in terms of AUC. We observed the following patterns: (1) the difference among methods shrinks as proportion of DE genes (pDiff) increases; (2) the number of genes (nGenes) doesn't affect the methods performance in terms of AUC values; (3) all methods perform better when the number of samples increases. Supplementary materials accompanying this paper is on github at https://github.com/xiyuansun/kellycc.

CHAPTER 1. Overview

1.1 Introduction

A gene is regarded as differentially expressed when the expected count reads of this gene corresponding to one condition differs from that of another condition. Finding genes that are differntially expressed between conditions is an integral part of understanding the molecular basis of phenotypic variation. High-throughout sequencing of cDNA (RNA-Seq) has been used to quantify the abundance of mRNA corresponding to different genes(Soneson and Delorenzi, 2013).

RNA-Seq is a new approach to transcriptome analysis based on next-generation sequencing technology. RNA-Seq data are a set of short RNA reads that are often summarized as discrete counts. The negative binomial distribution has become widely used to analyze RNA-Seq data which allows more flexibility in assigning between-sample variation(Ching et al., 2014). We analyzed simulated datasets with a defined total number of genes, total number of samples, and proportion of DE genes. We compared eBayes(Niemi et al., 2015) method to the alternative methods implemented by existing R packages, such as `edgeR, DESeq, DESeq2, sSeq, EBSeq`. All methods are available within the R framework and take a matrix of counts as input, i.e., the number of reads mapping to each genomic feature of interest in each of a number of samples. We evaluated the methods based on simulated datasets, as demonstrated in RNA-Seq Data section.

Ji et al. (Ji et al., 2014) introduced an approach to assess gene expression heterosis using microarray data under the assumption that these data are continuous. They built a normal hierarchical model for microarray measurements of transcript abundance that allows borrowing of information across genes to estimate means and variances. They introduced an empirical

Bayes framework that first estimates model hyperparameters, then estimates the posterior distribution for gene-specific parameters conditional on those hyperparameters, and finally computes heterosis probabilities based on integrals of regions under this posterior. Building on the work of Ji et al. with the normal data model, (Niemi et al., 2015) constructed a hierarchical model based on a negative binomial data model. They utilized an empirical Bayes approach to obtain estimates of the hyperparameters and the posterior distributions for the gene-specific parameters conditional on those hyperparameters. In this creative component report, we applied Niemi's empirical Bayes method in the differential gene expression analysis context, and call it `eBayes` method.

Five other methods for differential expression analysis of RNA-Seq data were also evaluated in this study. All of them work on the count data directly: `edgeR`(Robinson et al., 2010), `DESeq`(Anders and Huber, 2010), `DESeq2`(Love et al., 2014), `sSeq`(Yu et al., 2013), `EBSeq`(Leng et al., 2013). More detailed descriptions of the methods can be found in the Method section and in the respective original publication.

The six methods were evaluated based on simulated datasets, where we could control the settings and the true differential expression status of each gene. Details regarding the different simulation scenarios can be found in the Simulation section. We explored each method's ability to rank truly DE genes ahead of non-DE genes. This was evaluated in terms of the area under a Receiver Operating Characteristic (ROC) curve (AUC).

The remainder of the report proceeds as follows.Chapter 2.1-2.2 presents the hierarchical model, an empirical Bayes method of estimating the parameters, and the calculation of posterior probabilities of DE. Chapter 3 presents a simulation study based on a maize experiment and compares Niemi et al's approach to alternative methods. Chapter 4 summarizes the result with a facetted AUC plot and includes the discussion part for future research direction.

1.2 RNA-Seq Data

RNA-Seq is a next generation sequencing (NGS) procedure of the entire transcriptome by which one can measure the expression of several features, such as gene expression. The number of reads mapped to a given gene is considered to be the estimate of the expression level of that

feature using the technology(Marioni et al., 2008).

The end-product of a RNA-seq experiment id a sequence of read counts, typically a matrix with rows representing genes and columns representing samples from different gene varieties, as in Table 1.1. In this example, there are $V = 2$ gene varieties: $B73, Mo17$, 4 replicates of each variety. The genes shown above the double horizontal line are part of the genes with differential expression between the two varieties. The genes shown below the double horizontal line are examples of the non-DE genes.

	B73_1	B73_2	B73_3	B73_4	Mo17_1	Mo17_2	Mo17_3	Mo17_4
AC148152.3_FG001	3	4	6	0	8	17	18	20
AC148152.3_FG008	3	3	4	1	31	40	45	49
AC152495.1_FG002	33	46	18	13	4	0	2	6
AC152495.1_FG017	41	44	16	13	2	2	2	0
AC184130.4_FG012	24	47	18	21	110	144	121	96
AC184133.3_FG001	0	1	1	0	14	13	4	9
AC148152.3_FG005	2323	1533	1932	1945	2070	1582	2196	1882
AC148167.6_FG001	672	598	728	713	743	655	821	824
AC149475.2_FG002	459	438	451	483	467	448	634	532
AC149475.2_FG003	1184	976	1131	1206	891	743	1288	1107
AC149475.2_FG005	551	535	360	353	550	524	492	440
AC149475.2_FG007	245	214	169	159	297	262	210	302

Table 1.1 Maize RNA-Seq Data (Paschold,2012)

Our interest is in the detection of differentially expressed genes between the two varieties, i.e., genes for which read count distributions differ between varieties.

CHAPTER 2. Method

2.1 Estimating the difference between read counts for a given gene

To detemine whether the read count differences between different conditions for a given gene are greater than expected by chance, differential gene expression (DGE) tools must find a way to estimate that difference (Dündar et al., 2015).The two basic tasks of all DGE tools are: (1) Estimate the magnitude of differential expression between two or more conditions based on read counts from replicated samples, i.e., calculate the fold change of read counts, taking into account the differences in sequencing depth and variability; (2) Estimate the significance of the difference and correct for multiple testing.

2.2 Standard Setup of Negative Binomial Model in Generalized Linear Model Framework

To most easily explain the following DE analysis methods, we followed closedly the notation used in McCarthy et al.'s paper(McCarthy et al., 2012).

Let the Y_{gij} be the read cound in replicate j of condition i for gene g. Assume y_{gij} follows a NB distribution with mean μ_{gij} and gene-wise dispersion ϕ_g, denoted by Equation (2.1)

$$Y_{gij} \overset{ind}{\sim} NB(\mu_{gij}, \phi_g) \tag{2.1}$$

Gene g's varince equals $\mu_{gij}(1 + \phi_g \cdot \mu_{gij})$, while the dispersion ϕ_g is the square of the biological coefficient of variation (McCarthy et al., 2012).

In the generalized linear model (GLM) setting, the mean response, μ_{gij}, is linked to a linear predictor with the natural logorithm link according to Equation (2.2)

$$\log(\mu_{gij}) = x_i^T \beta_{\mathbf{g}} + \log(N_{ij}) \tag{2.2}$$

where x_i is the design matrix containing the covariates (e.g., experimental conditions, batch effects, etc.), $\beta_{\mathbf{g}} = (\beta_{g1}, \beta_{g2})$ is a vector of regression parameters (a subset of which are of interest for differential expression inference) and N_{ij} is the normalized library size for replicate j of condition i. In the Empirical Bayes Method Part, we denote the terms of primary scientific interest as Equation (2.3)

$$\lambda_{gi} = x_i^T \beta_{\mathbf{g}} \tag{2.3}$$

and the noamalization factors as Equation (2.4)

$$\gamma_{ij} = \log(N_{ij}) \tag{2.4}$$

Different DE analysis methods adopted different algorithms to estimate the regression parameters and gene-wise dispersion parameters. More detials could be checked in the Alternative Methods section.

2.3 Empirical Bayes identification of gene differential expression from RNA-seq read counts

To use RNA-seq counts to identify genes displaing differential expression (DE), we built a hierarchical model to borrow information across gene-variety means and across gene-specific overdispersion parameters, estimate the hyperparameters using an empirical Bayes procedure, and calculate empirical Bayes posterior probabilities for DE.

2.3.1 Hierarchical model for RNA-seq counts

Let Y_{gij} be the count for gene $g = 1, 2, ..., G$, variety $i = 1, 2$, and replicate $j = 1, 2, 3, ..., n_i$.

We assume Y_{gij} follows Equation (2.1) with a different parameterization as $\mu_{gij} = \exp(x_i^T \beta_g +$ $\log(N_{ij}), \lambda_{gi} = x_i^T \beta_g, \gamma_{ij} = \log(N_{ij}), \phi_g = \exp(\psi_g)$, and $\log(N_{ij})$ are normalization factors that account for differencees in the thoroughness of sequencing from sample to sample.

Following (Ji et al., 2014), we reparameterize the gene-variety mean structure into the genespecific average β_{g1} and half-variety difference β_{g2} as shown in Equation (2.5). For our differential expression study where number of varieties is 2, let $i = 1, 2$ indicate the two varieties. The reparameterization is

$$\beta_{g1} = \frac{\lambda_{g1} + \lambda_{g2}}{2}, \beta_{g2} = \frac{\lambda_{g1} - \lambda_{g2}}{2} \tag{2.5}$$

We assume a hierarchical model for the gene-specific mean parameters and overdispersion parameters with the variety averages, half-variety averages, and overdispersion parameters follow normal distributions as in Equation (2.6)

$$\beta_{g1} \overset{ind}{\sim} N(\eta_{\beta_1}, \sigma^2_{\beta_1}), \beta_{g2} \overset{ind}{\sim} N(\eta_{\beta_2}, \sigma^2_{\beta_2}), \psi_g \overset{ind}{\sim} N(\eta_\psi, \sigma^2_\psi) \tag{2.6}$$

We assume a priori independence among the variety averages, half-variety averages, and overdispersion parameters.

2.3.2 Empirical Bayes Method

We categorized the parameters of the model in Section 2.2.1 into gene-specific parameters $\theta = (\theta_{\mathbf{1}}, ..., \theta_{\mathbf{G}})$ where $\theta_{\mathbf{g}} = (\beta_{g1}, \beta_{g2}, \psi_g)$, normalization factors $\gamma = (\gamma_{11}, ..., \gamma_{Vn_V})$, and hyperparameters $\pi = (\eta, \sigma)$ where $\eta = (\eta_{\beta_1}, \eta_{\beta_2}, \eta_\psi)$ and $\sigma = (\sigma_{\beta_1}, \sigma_{\beta_2}, \sigma_\psi)$. We obtained estimates for the hyperparameters and then based gene-specific inference on the posterior conditional on these estimates (Niemi et al., 2015).

To obtain normalization factors $\hat{\gamma}$, we followed Niemi et al's approch (Niemi et al., 2015) using the weighted trimmed mean of M values (TMM)(Robinson and Oshlack, 2010). We used `edgeR` to obtain genewise dispersion estimates, $\hat{\psi}_g$, through the adjusted profile likelihood (APL) introduced byCox and Reid (Cox and Reid, 1987), and the generalized linear model methods to obtain estimates for the remaining gene-specific parameters $(\hat{\beta}_{g1}, \hat{\beta}_{g2})$(Robinson and Oshlack, 2010). Using $\hat{\theta}_g = (\hat{\beta}_{g1}, \hat{\beta}_{g2}, \hat{\psi}_g)$, I estimate hyperparameters for the location and scale parameters in the hierarchical model using a central method of moments approach, shown in Equation (2.7) and Equation (2.8). For $\hat{\eta}_\psi, \hat{\sigma}_\psi$, the formula is similar to Equation (2.7).

$$\hat{\eta}_{\beta_1} = \sum_{g=1}^{G} \hat{\beta}_{g1}/G, \hat{\sigma}^2_{\beta_1} = \sum_{g=1}^{G} (\hat{\beta}_{g1} - \hat{\eta}_{\beta_1})^2/(G-1) \tag{2.7}$$

$$\hat{\eta}_{\beta_2} = \sum_{g=1}^{G} \hat{\beta}_{g2}/G, \hat{\sigma}^2_{\beta_2} = \sum_{g=1}^{G} (\hat{\beta}_{g1} - \hat{\eta}_{\beta_1})^2/(G-1) \tag{2.8}$$

Conditional on the estimated normalization factors $\hat{\gamma}$ and hyperparameters $\hat{\pi}$, we perform a Bayesian analysis to re-estimate the gene-specific parameters and describe their uncertainty (Niemi et al., 2015). Equation 2.9 shows that conditional on $\hat{\gamma}$ and $\hat{\pi}$, the gene-specific parameters are independent and therefore conditional posterior inference across the genes can be parallelized.

$$p(\theta|y,\hat{\pi},\hat{\gamma}) \propto$$
$$\prod_{g=1}^{G}\left[\prod_{i=1}^{2}\prod_{j=1}^{n_i} NB(y_{gij};\hat{\mu}_{gij}=\exp(\lambda_{gi}+\hat{\gamma}_{ij}),\phi_g=\exp(\psi_g))N(\beta_{g1};\hat{\eta}_{\beta_1},\hat{\sigma}^2_{\beta_1})p(\beta_{g2};\hat{\eta}_{\beta_2},\hat{\sigma}_{\beta_2})N(\psi_g;\hat{\eta}_{\psi},\hat{\sigma}^2_{\psi})\right] \tag{2.9}$$

To perform the conditional posterior inference on the gene-specific parameters, we used the statistical software `Stan` (Team et al., 2014) executed through the `RStan` interface (Team et al., 2016). Stan implements a Hamiltonion Monte Carlo (Neal et al., 2011) to obtain samples from the posterior in Equation 2.9. We used the default NUTS sampler(Annis et al., 2017).

We ran four simultaneous chains with random initial starting values for 1000 burn-in (and tuning) iterations followed by another 1000 iterations retaining every fourth sample (to reduce storage space) for inference. We monitored convergence using the potential scale reduction factor and effective sample size (ESS) for all gene-wise parameters $\theta_{\mathbf{g}} = (\beta_{g1}, \beta_{g2}, \psi_g)$ (Gelman et al., 1992). According to Niemi et al's approach(Niemi et al., 2015), we reran the chains with double the iterations for both burn-in and inference if the minimum ESS was less than 1000. We continued the restarting and doubling until we obtained minimum ESS greater than 1000 for all parameters.

2.3.3 Gene expression differentiation

In the maize context that motivates this work, we are interested in differential expression (DE). For a specific gene g, non-DE occurs when expected expression in the second variety is the same as the expected expression of first variety, i.e., $\mu_{g1} = \mu_{g2}$, or equivalently, $\beta_{g2} = 0$. I evaluate measurements based on empirical Bayes estimates of their posterior probabilities, e.g.,

$$P(DE_g|y,\hat{\pi},\hat{\gamma}) = \min(P(\beta_{g2} < 0|y,\hat{\pi},\hat{\gamma}), P(\beta_{g2} > 0|y,\hat{\pi},\hat{\gamma})) \tag{2.10}$$

$P(\beta_{g2} < 0|y,\hat{\pi},\hat{\gamma}) \approx \frac{1}{M}\sum_{m=1}^{M} I(\beta_{g2}^{(m)} < 0)$, $P(\beta_{g2} > 0|y,\hat{\pi},\hat{\gamma}) \approx \frac{1}{M}\sum_{m=1}^{M} I(\beta_{g2}^{(m)} > 0)$ where $\beta_{g1}^{(m)}, \beta_{g2}^{(m)}$ is the m^{th} MCMC sample from the empirical Bayes posterior.

We do not evaluate $\beta_{g2} = 0$ since we rather treat β_{g2} as continuous.

We based our DE tag decisions on the estimates of the posterior probabilities shown in Equation (2.10). We constructed a ranked list of genes according to the minimum of $P(\beta_{g2} < 0|y,\hat{\pi},\hat{\gamma})$ and $P(\beta_{g2} > 0|y,\hat{\pi},\hat{\gamma})$. Geneticists can use this list to prioritize future experiments to understand the molecular genetic mechanisms for differential expression (Niemi et al., 2015).

We will use the term `eBayes` to refer to the approach defined in Sections 2.1 - 2.2 and we are assuming normal distribution for half-variety differences.

2.4 Alternative Methods

We compared the `eBayes` method to five alternative methods. To follow the recent progress in the RNA-Seq DE area, we selected two widely used methods, `edgeR, DESeq`, and three other newly released DE analysis packages `DESeq2, EBSeq, and sSeq`. For each method, we attempted to provide a measure of the strength of DE for each gene such that small values of this measure indicate support for DE.

Several authors proposed differential expression methods based on the negative binomial distribution, motivated by observation that real RNA-Seq data sets typically exhibited greater variability than could be modeled via the Poisson distribution(Lorenz et al., 2014).

Robinson and Smyth (Robinson and Smyth, 2007a) assumed a negative binomial distribution for the read counts for all genes with a common dispersion parameter, i.e., $Y_{gij} \overset{ind}{\sim} NB(\mu_{gij}, \phi)$, where $\mu_{gij} = N_{ij}\exp(\lambda_{gi})$, N_{ij} is the normalized library size for sample j in population i, and $\exp(\lambda_{gi})$ is the relative abundance parameter for gene g in population i, which is assumed to be the same to the replicate samples within a population. The dispersion parameter ϕ is estimated by maximizing the conditional likelihood given the sum of the counts in each poopulation. Quantile adjusted conditional maximum likelihood (qCML) is applied if the li-

brary sizes are not equal within each population. The null hypothesis for the test of differential expression is the equality of the relative abindance parameters, $H_0 : \lambda_{g1} = \lambda_{g2}, g = 1, 2, ..., G$. The authors suggested an exact NB test based on the same quantile adjustment used in estimating the dispersion parameter, and a p-value calculated as the probability of observing counts greater than those observed(Lorenz et al., 2014). But the assumption of dispersion parameter ϕ common to all genes is often implausible. The authors extended their NB approach and suggested using gene-specific dispersion parameter ϕ_g(Robinson and Smyth, 2007b). The exact test with empirical Bayes adjustment was better at detecting DE genes and was better able to control false discovery rates when gene-specific overdispersion was introduced(Lorenz et al., 2014). So they extended the standard NB approach by estimating gene-specific dispersion parameter via empirial Bayes weighted likelihood estimation, in which gene-specific dispersion parameter estimates were shrunk toward a common dispersion. Their method was implemented in R package called `edgeR`. It moderates the dispersion per gene toward a local estimate with genes of similar expression.

Anders and Huber (Anders and Huber, 2010) noted that dispersion often varies with expected read count, and suggested an extended NB model in which the variances of the read count are defined a nonparametric function of their expectations, as $Y_{gij} \overset{ind}{\sim} NB(\mu_{gij}, \phi_\mu)$, where $\mu_{gij} = N_{ij} \exp(\lambda_{gi})$. Then $Var(Y_{gij}) = \mu_{gij}(1 + \phi_\mu \mu_{gij})$. They employ a gamma-family generalized linear local regression to model the mean-dispersion relationship. The nulll hypothesis in the test of differential expression is $H_0 : \lambda_{g1} = \lambda_{g2}$, which is tested by an exact test similar to Robinson and Smyth's. Their method was implemented in an R package called `DESeq`. It detects and corrects dispersion estimates that are too low through modeling of the dependence of the dispersion on the average expression strength over all samples. `DESeq` (by default) estimates dispersion by pooling all samples together, fitting them to a parametric distribution and taking the maximum.

Love and Huber (Love et al., 2014) then proposed another method for differential analysis of count data, using shrinkage estimation for dispersions and fold change to improve stability and interpretability of the estimates based on Anders and Huber's. They noticed the limitation of the most common approach in the comparative analysis of transcriptomics data. The noisiness

of LFC estimates for genes with low counts would complicate the ranking by fold change. So they developed a statistical framework to facilitate gene ranking and visualization based on stable estimation of effect sizes (LFCs), as well as testing of differential expression with respect to user-defined thresholds of biological significance. They first perform ordinary GLM fits to obtain MLEs for the LFCs and then fit a zero-centered normal distribution to the observed distribution of MLEs over all genes. This distribution is used as a prior on LFCs in a second round of GLM fits, and the MAP estimates are kept as the final estimates of LFC. A standard error for each estimate is derived from the posterior's curvature at its maximum. These shrunken LFCs and their standard errors are used in the Wald tests for differential expression. Their method was implemented by `DESeq2`. It uses a Wald test: the shrinken estimate of LFC is divided by its standard error, resulting in a z-statistic, which is compared to a standard normal distribution. `DESeq2` is a new update to `DESeq`, and it uses shrinkage estimation for dispersion: the first round of dispersion-mean relationship is obtained by MLE, and this fit is then used as a prior to estimate the maximum a posteriori estimate for dispersion in the second round.

`edgeR, DESeq` and `DESeq2` differ in how the dispersion is estimated. The default normalization method is also different among `edgeR`, `DESeq` and `DESeq2`. `edgeR` uses trimmed-mean-of-M-values (TMM), while `DESeq, DESeq2` use a relative log expression approach.

Yu and Huber used the method of moment estimates for the dispersion and shrank them towards an estimated target, which minimizes the average squared difference between the shrinkage estimates an the initial estimates. They estimate dispersion by pooling all the samples using the method of moments, and then shrinking the gene-wise estimates through minimizing the mean-square error. They also used exact test for the DE analysis. The model has little practical difference from the model in Anders and Huber's. Yu and Huber use the Hansen's generalized shrinkage estimator $\hat{\phi}_g$ in conjunction with the NB distribution to test genes for differential expression. They follow `edgeR, DESeq` by testing $H_0 : \mu_{g1} = \mu_{g2}$ per gene with the exact test. Under H_0, the p-values are calculated with respect to $Y_{gij} \sim NB(s_{ij}\mu_{gi}, \phi_{gi})$ and are adjusted to control the false discovery rate(Yu et al., 2013). s_{ij} is the size factor. It can be thought of the representative ratio of counts in the library to the geometric mean of the counts

in all the libraries. Their method was implemented in `sSeq`.

Leng developed an empirical Bayes model for identifying DE genes and isoforms. This method was implemented in `EBSeq`. It provides posterior probabilities as the evidence in favor of DE. Estimates of the gene-specific means and variances are obtained via method-of-moments, and the hyperparameters are obtained via the expectation-maximization (EM) algorithm(Leng et al., 2013). `EBSeq` estimates the posterior likelihoods of differential expression by the aid of empirical Bayesian methods. To account for the different sequencing depths, a median normalization procedure similar to `DESeq` is used.

CHAPTER 3. Simulation Study

3.1 Simulation Study based on a maize experiment

We built a simulation framework that aims to reflect the reality of RNA-seq data.

3.1.1 Parameter Estimation

To assess the efficacy of `eBayes` method to identify DE genes, we used a maize dataset with varieties $B73$ and $Mo17$ (Paschold et al., 2012) to determine realistic parameter values for our simulation study. Chapter 2 describes the maize dataset in detail.

We removed the genes with zero counts in all conditions, as well as genes whose maximum counts are less than 5 as recommended (Rau et al., 2013).The description of parameters for the real RNA-seq dataset is summarized in Table 3.1.

Number of Trimmed Genes	27619
Number of Samples	8
Median expression ($\log 2$ counts per million)	3.92
Median dispersion	0.03
Median $\log 2$ fold change (LFC) of genes	0.39
Median library size (sum of total counts, $\log 10$)	6.99

Table 3.1 Description of estimated parameters

We estimated parameters from the maize RNA-seq data (Paschold et al., 2012), and fit it by GLM with negative binomial distribution. The genewise dispersions for negative binomial GLMs were estimated using Cox-Reid Adjusted Profile Likelihood(McCarthy et al., 2012). This method modifies the maximum likelihood estimate of dispersion by accounting for the experimental design through Fisher's information matrix in the log-likelihood function(McCarthy et al., 2012).

In summary, the library size (reads mapped to the transcriptome) is $\log 10$ mean of 6.99, the normalized median gene expressions $\log 2$ counts per million (CPM) is 3.92, and the median LFCs of DE genes is 0.39.

3.1.2 Model

In our simulated data, we used a generalized linear model (GLM) with negative binomial distribution. For the dataset of two groups, the counts for a particular gene g in group i replicate j were modeled by Equation (2.1), where ϕ_g is the genewise dispersion calculated by the CR-APL method, and the expected value μ_{gi} is a function of the library size of group i replicate j as Equation 2.2, where μ_{gi} is the expected counts of gene g in group i, N_{ij} is the normalized library size for group i replicate j, β_g is the vector of coefficients for the two experimental conditions (two groups), x_i is a vector of length 2 indicating whether replicate j belongs to group one or group two in the experiment(Ching et al., 2014).

3.1.3 Simulation Scenario

To evaluate the performance of `eBayes` method and the alternative methods across a variety of reasonable scenarios, we created several options to make up each scenario. We set up the simulation scenarios as the following Table 3.2:

sc	nGenes	nSamples	pDiff(%)
1	10000	8	10
2	10000	8	30
3	10000	8	1
4	10000	4	10
5	10000	4	30
6	10000	4	1
7	10000	16	10
8	10000	16	30
9	10000	16	1
10	1000	8	10
11	1000	8	30
12	1000	8	1
13	1000	4	10
14	1000	4	30
15	1000	4	1
16	1000	16	10
17	1000	16	30
18	1000	16	1

Table 3.2 Simulation Scenario Table

We set up a simulation framework with parameters based on the joint distribution of mean μ_{gi} and gene-wise dispersion estimates ϕ_g from the maize RNA-seq data (Paschold et al., 2012).

We generated true NB model parameters, μ_{gi} and ϕ_g, using the joint distribution of estimates $\hat{\mu}_{gi}$ and $\hat{\phi}_g$, estimated using `edgeR` from the real dataset (Paschold et al., 2012). The derived from real data parameters were used to simulate the counts, from a NB distribution.

For a particular gene in the simulated dataset, the number of true DE tag was determined via the proportion of differential gene expression (pDiff) in the simulation scenario setup. I randomly selected $nGenes \times (1 - pDiff)$ genes and assign the expected variety count means

the same as $\mu_{g1} = \mu_{g2} = 1/(nSamples)\sum_{j=1}^{nRep}\left[N_{1j}\exp(X_1^T\beta_{g1}) + N_{2j}\exp(X_2^T\beta_{g2})\right]$ for the selected non-DE genes, where $\beta_{gi}, i = 1, 2$ was the regression coefficient estimates shown in equation (2.3) got from the parameter estmation based on the real RNA-seq dataset.

We simulated nGenes $\in (10000, 1000)$ total number of genes with negative binomial distribution counts. To simulate data with realistic moments, the mean and dispersions were drawn from the joint distribution of means and gene-wise dispersion estimates from the real maize data (Paschold et al., 2012). These simulated datasets were of varying total sample size nSamples $\in (4, 8, 16)$, and the samples were split into two equal-sized groups. pDiff $\in (10\%, 30\%, 1\%)$ of genes are true DE genes. For each scenario, we simulated 5 datasets with different seeds.

CHAPTER 4. Result

4.1 Methods Comparison Result

For the methods in Chapter 2, we sorted genes according to the computed measure of the strength of evidence for DE, which were the adjusted p-values for alternative methods and the statistics shown in Equation (2.10) for the `eBayes` method. From these sorted lists, we calculated area under ROC curve (AUC) values to evaluate the ability of these methods to distinguish genes with DE.

To evaluate the true positive rate (TPR) and false positive rate (FPR) together, we generated receiver operator characteristic (ROC) curves based on the DE analysis results of the simulated datasets in each simulation scenario. It is a plot of the true positive rate against the false positive rate for the different possible cutpoints of a test (DE test). The slope of the tagent line at a cutpoint gives the likelihood ratio for that value of the test. Figure 4.1 is an example of the ROC curves we generated based on scenario 1 all simulated datasets. More scenario simulation ROC curves are in the Appendix.

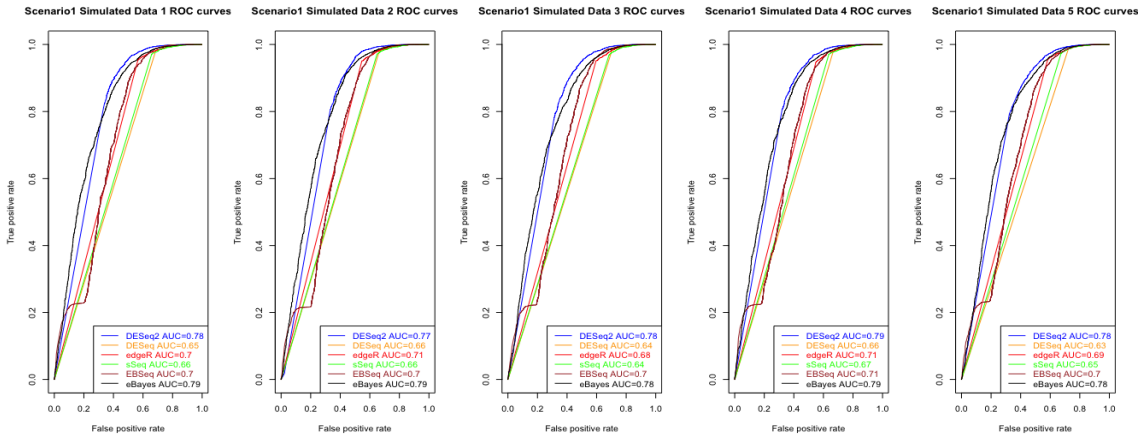

Figure 4.1 ROC curves of Simulated Datasets with $nGenes = 10000, nSamples = 8, pDiff = 10\%$

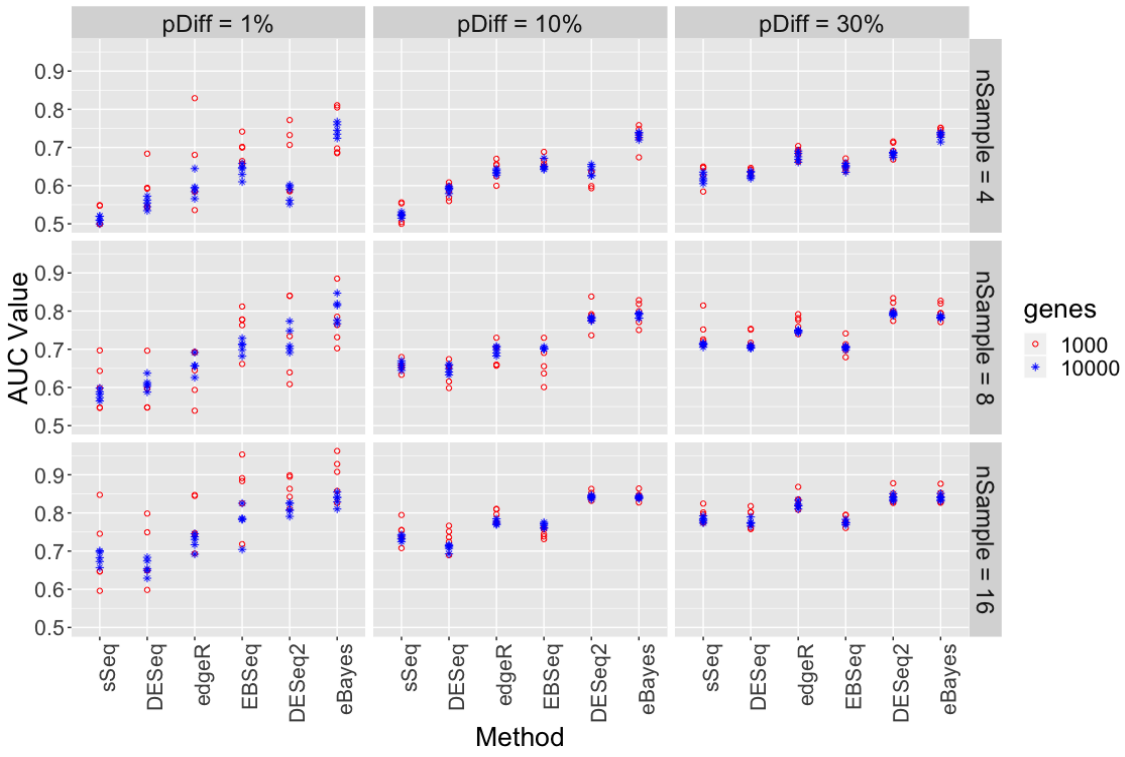

Figure 4.2 AUC Plot of Simulated Datasets across Six Methods, Facetted by proportion of DE (pDiff) and number of samples (nSample), Grouped by total number of Genes (nGenes)

We also evaluated the different methods with another performance metric: area under the curve (AUC) of ROC curves. The accuracy of the DE test depends on how well the test seperates the DE and non-DE groups. Accuracy is measured by the area under the ROC curve. Figure 4.2 provides the area under the ROC curve (AUC) across the 5 simulations for each of the scenario defined in 3.2. We facetted the plots by number of samples (nSample) and differential gene expression proportion (pDiff), grouped by different level of total number of genes.

Similar to the single ROC curve, the `eBayes` method appears to outperform the other methods in terms of AUC, when number of samples and true DE proportion are small. With four or eight replicates per variety, there does not appear to be much of a difference between `eBayes` and `DESeq2`, but as the number of replicates decreases, the `eBayes` approach appears to improve relative to `DESeq2` in terms of AUC values.

We noticed that the difference among methods increased as proportion of true DE genes

(pDiff) decreased. pDiff shrank means fewer genes were taged as true DE genes. `eBayes` seems a better DE genes analysis tool handling smaller pDiff. We also found that all methods perform better when the number of samples increases, while the number of genes (nGenes) doesn't affect the methods performance obviously in terms of AUC values.

CHAPTER 5. Discussion

`eBayes` method is based on obtaining estimate for hyperparameters followed by MCMC to estimate gene-specific parameters. The empirical Bayes posteriors were used to estimate posterior probabilities of DE . Through a simulation study, we demonstrated that this method outperformed alternative methods in most simulation scenarios with higher AUC values. More samples (nSamples) would improve all methods' performance given the same proportion of DE genes (pDiff). DE analysis methods work better on count data with more replicates per condition and higher true DE genes proportion. This is not surprising considering that the focus of most methods is to model the variability in gene expression measurements and therefore increasing the number of replicates strengthen the estimate. The true DE genes proportion (pDiff) affects the outperformance of `eBayes`. When pDiff is smaller, `eBayes` performs much better than other methods in terms of AUC values, which means eBayes method has some advantages handling smaller true DE proportion scenario. We observed that the ROC curves for `sSeq, DESeq` were almost straight lines in some senarios, which might be associated with equal cost of misclassifying DE and misclassifying non-DE cases. We got rid of such issues when we used unadjusted p-value as the statistics of `sSeq, DESeq`. But using unadjusted p-values ended up inherently including a large number of false positives given cutoff of 0.05 for p-values. Soneson (Soneson and Delorenzi, 2013) also mentioned that some methods including `DESeq` exhibit an excess of large p-values, which has been attributed to the use of exact tests based on discrete probability distributions(Robles et al., 2012).

For future researches, we would recommend adding more methods to the comparison, such as other Bayesian methods implemented by `baySeq, ShrinkSeq`, nonparametric methods implemented by `NOISeq, SAMseq` , and fully Bayesian method using `fbseq`. The fully Bayesian method would involve some heavy computation. It would be easier with access to high perfor-

mance clusters equipped with gpu nodes. People could also try more complicated experimental designs, i.e., including more experimental conditions or adding the flow cell effects. `eBayes` method might be improved by refining the hierarchical model for the gene-specific parameter distribution and the hyperparameters estimation.

BIBLIOGRAPHY

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010.

Jeffrey Annis, Brent J Miller, and Thomas J Palmeri. Bayesian inference with stan: A tutorial on adding custom distributions. *Behavior research methods*, 49(3):863–886, 2017.

Travers Ching, Sijia Huang, and Lana X Garmire. Power analysis and sample size estimation for rna-seq differential expression. *Rna*, 20(11):1684–1696, 2014.

David Roxbee Cox and Nancy Reid. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(1):1–18, 1987.

Friederike Dündar, Luce Skrabanek, and Paul Zumbo. Introduction to differential gene expression analysis using rna-seq. *Applied Bioinformatics Core/Weill Cornell Medical College*, pages 1–67, 2015.

Andrew Gelman, Donald B Rubin, et al. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

Tieming Ji, Peng Liu, and Dan Nettleton. Estimation and testing of gene expression heterosis. *Journal of agricultural, biological, and environmental statistics*, 19(3):319–337, 2014.

Ning Leng, John A Dawson, James A Thomson, Victor Ruotti, Anna I Rissman, Bart MG Smits, Jill D Haag, Michael N Gould, Ron M Stewart, and Christina Kendziorski. Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, 29(8):1035–1043, 2013.

Douglas J Lorenz, Ryan S Gill, Ritendranath Mitra, and Susmita Datta. Using rna-seq data to detect differentially expressed genes. In *Statistical analysis of next generation sequencing data*, pages 25–49. Springer, 2014.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.

Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research*, 40(10):4288–4297, 2012.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Jarad Niemi, Eric Mittman, Will Landau, and Dan Nettleton. Empirical bayes analysis of rna-seq data for detection of gene expression heterosis. *Journal of agricultural, biological, and environmental statistics*, 20(4):614–628, 2015.

Anja Paschold, Yi Jia, Caroline Marcon, Steve Lund, Nick B Larson, Cheng-Ting Yeh, Stephan Ossowski, Christa Lanz, Dan Nettleton, Patrick S Schnable, et al. Complementation contributes to transcriptome complexity in maize (zea mays l.) hybrids relative to their inbred parents. *Genome research*, 22(12):2445–2454, 2012.

Andrea Rau, Mélina Gallopin, Gilles Celeux, and Florence Jaffrézic. Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics*, 29(17): 2146–2152, 2013.

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

Mark D Robinson and Gordon K Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007a.

Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–332, 2007b.

Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1): 139–140, 2010.

José A Robles, Sumaira E Qureshi, Stuart J Stephen, Susan R Wilson, Conrad J Burden, and Jennifer M Taylor. Efficient experimental design and analysis strategies for the detection of differential expression using rna-sequencing. *BMC genomics*, 13(1):484, 2012.

Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. *BMC bioinformatics*, 14(1):91, 2013.

Stan Developent Team et al. Rstan: the r interface to stan. *R package version*, 2(1), 2016.

Stan Development Team et al. Stan: A c++ library for probability and sampling. *Online: http://mc-stan. org*, 2014.

Danni Yu, Wolfgang Huber, and Olga Vitek. sseq: A simple and shrinkage approach of differential expression analysis for rna-seq experiments. 2013.

.1 Appendix

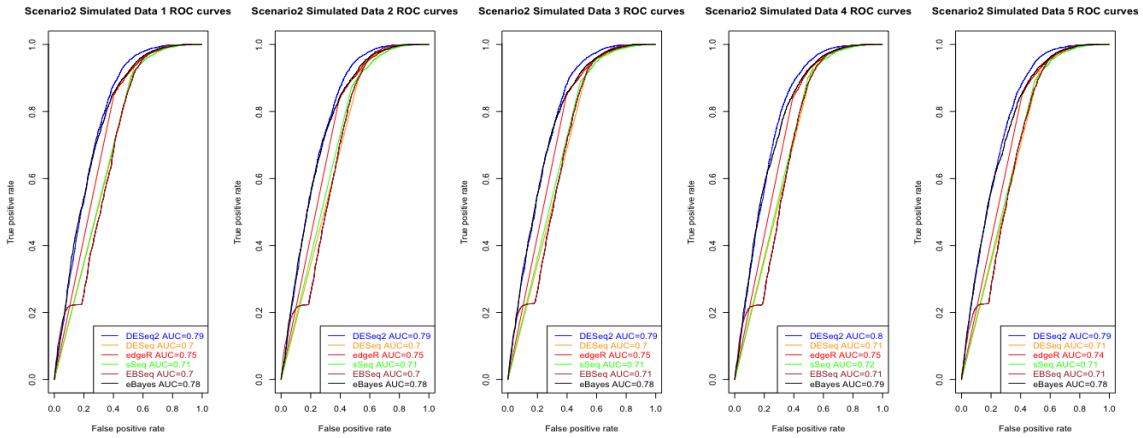

Figure 1 ROC curves of Simulated Datasets with $nGenes = 10000, nSamples = 8, pDiff = 30\%$

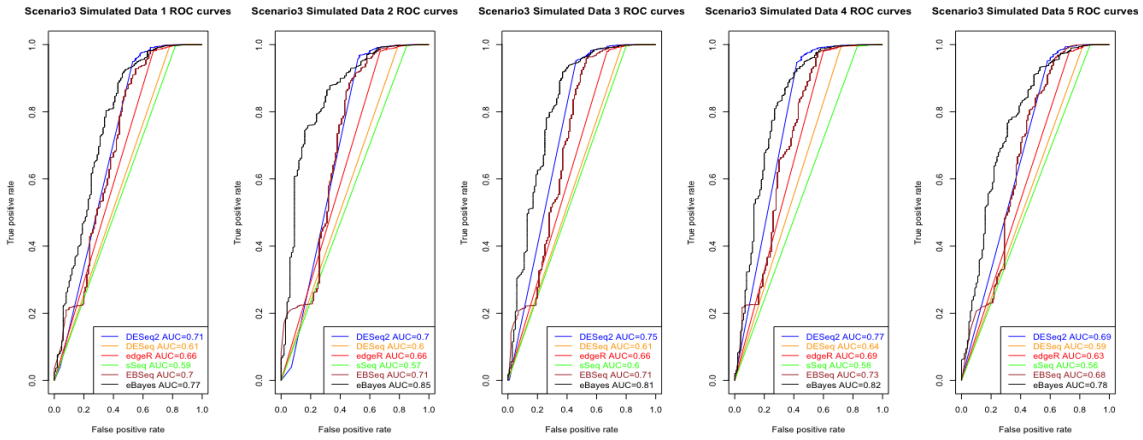

Figure 2 ROC curves of Simulated Datasets with $nGenes = 10000, nSamples = 8, pDiff = 1\%$

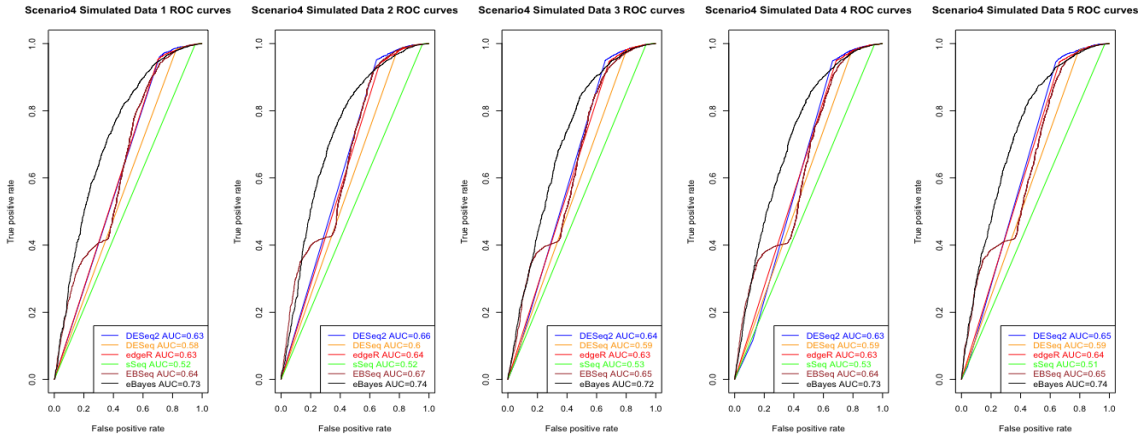

Figure 3 ROC curves of Simulated Datasets with $nGenes = 10000, nSamples = 4, pDiff = 10\%$

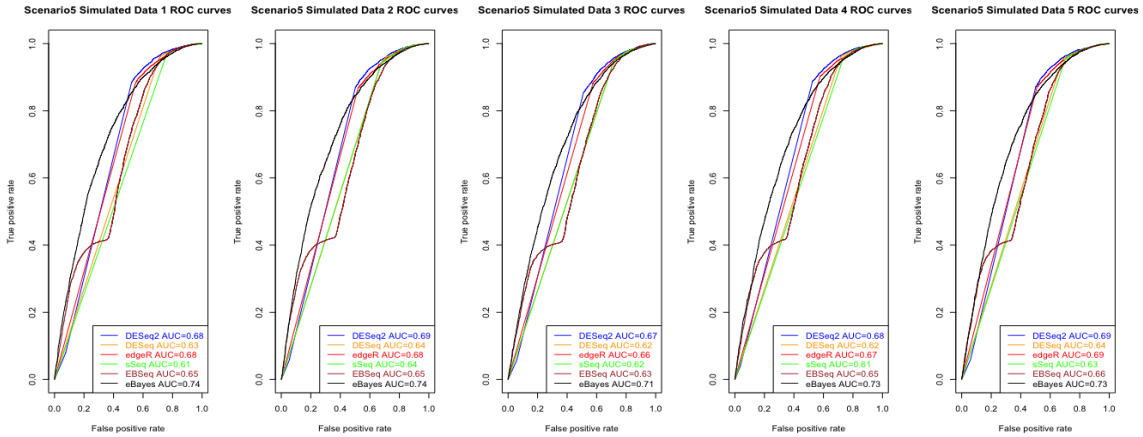

Figure 4 ROC curves of Simulated Datasets with $nGenes = 10000, nSamples = 4, pDiff = 30\%$

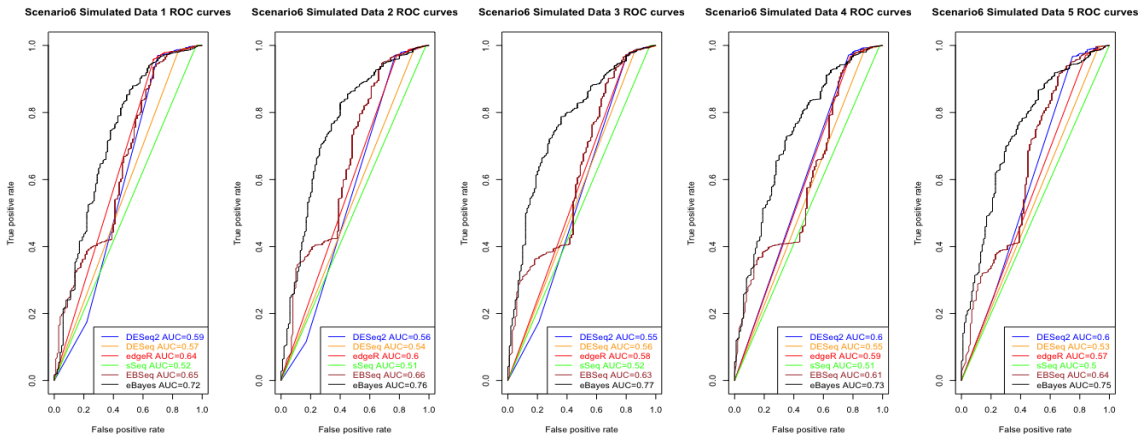

Figure 5 ROC curves of Simulated Datasets with $nGenes = 10000, nSamples = 4, pDiff = 1\%$

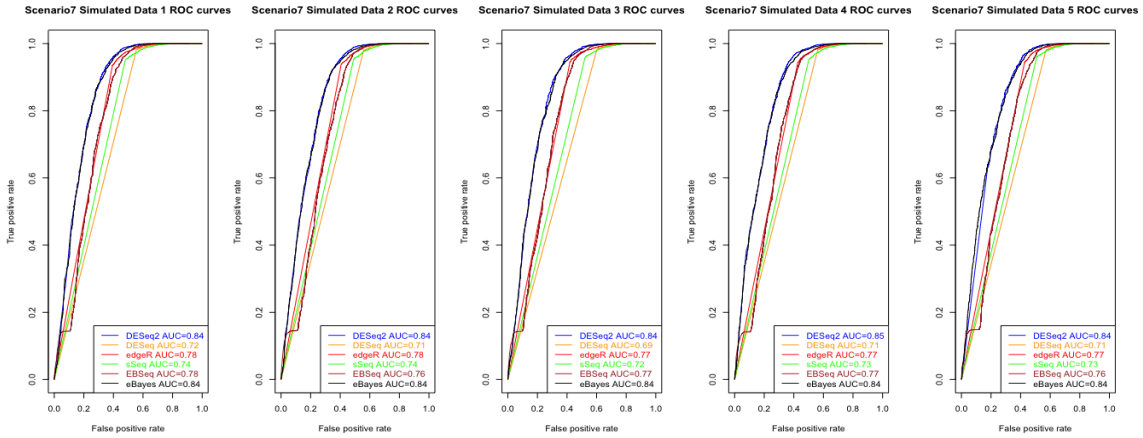

Figure 6 ROC curves of Simulated Datasets with $nGenes = 10000, nSamples = 16, pDiff = 10\%$

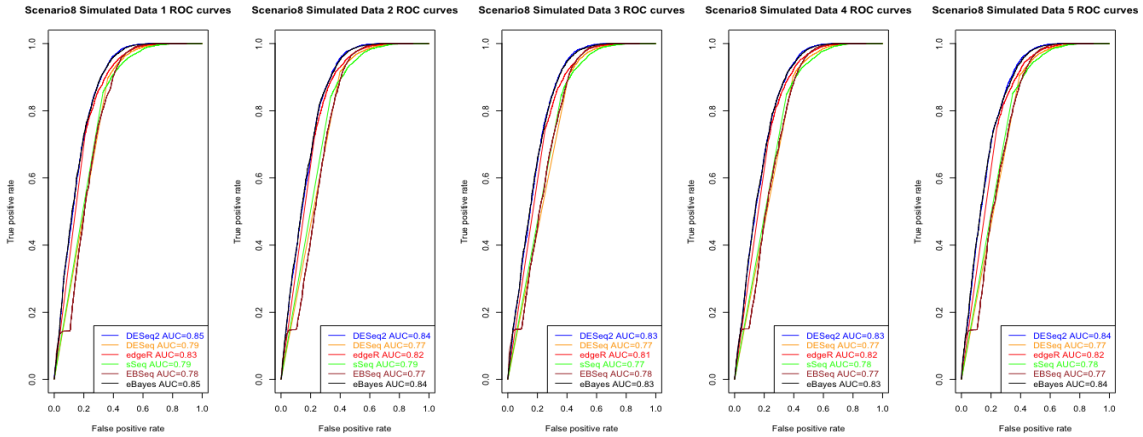

Figure 7 ROC curves of Simulated Datasets with $nGenes = 10000, nSamples = 16, pDiff = 30\%$

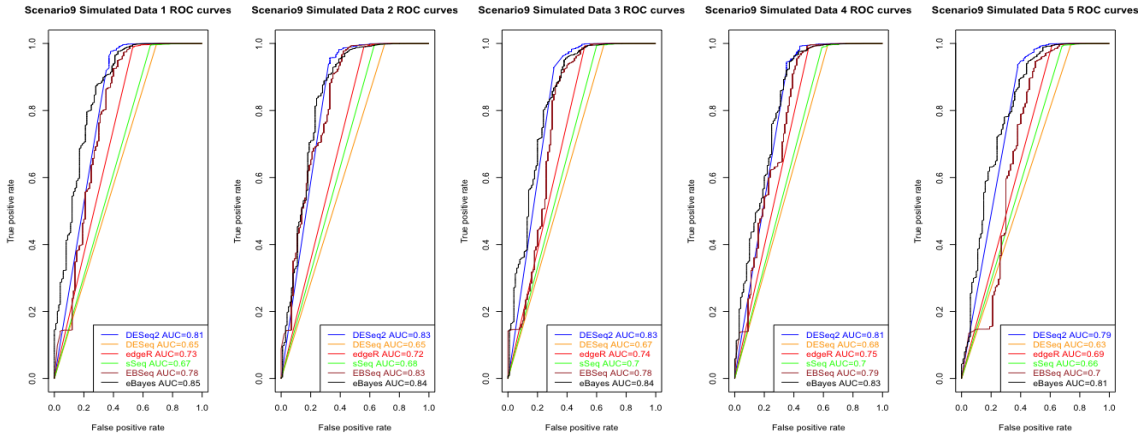

Figure 8 ROC curves of Simulated Datasets with $nGenes = 10000, nSamples = 16, pDiff = 1\%$

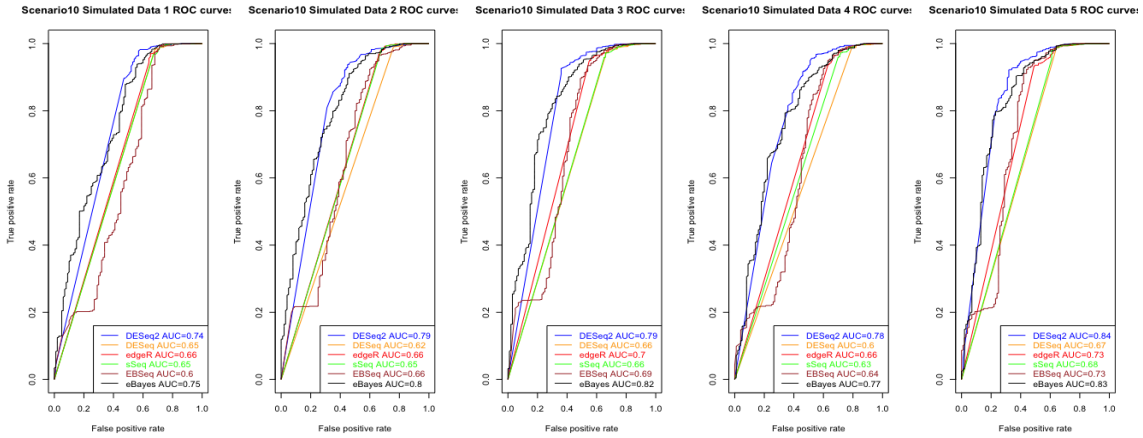

Figure 9 ROC curves of Simulated Datasets with $nGenes = 1000, nSamples = 8, pDiff = 10\%$

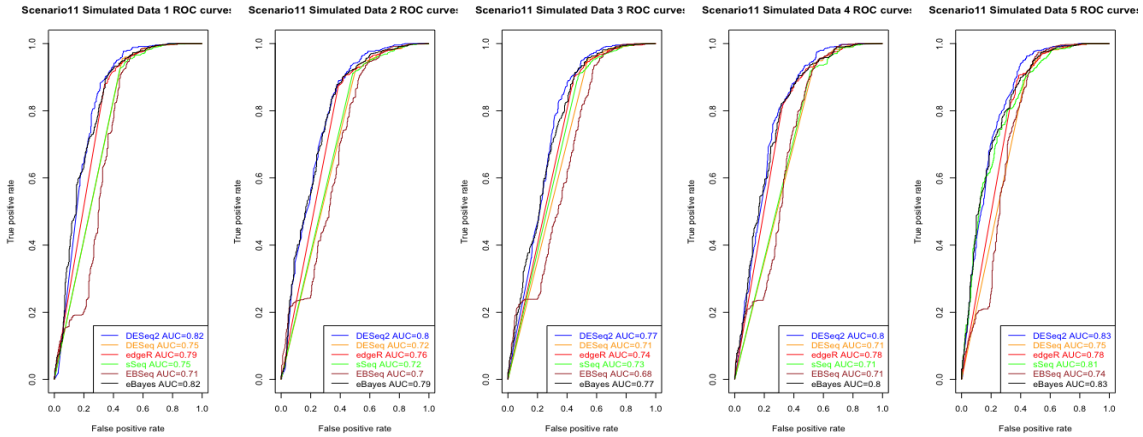

Figure 10 ROC curves of Simulated Datasets with $nGenes = 1000, nSamples = 8, pDiff = 30\%$

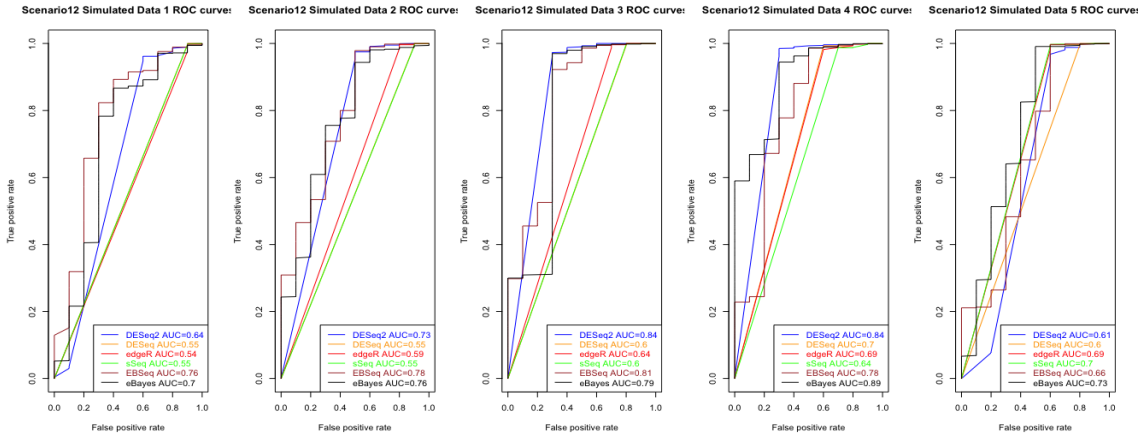

Figure 11 ROC curves of Simulated Datasets with $nGenes = 1000, nSamples = 8, pDiff = 1\%$

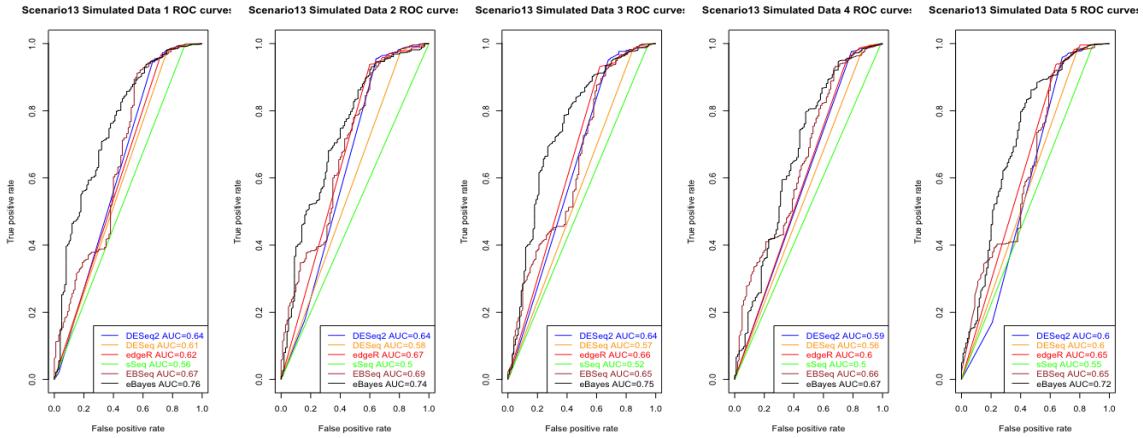

Figure 12 ROC curves of Simulated Datasets with $nGenes = 1000, nSamples = 4, pDiff = 10\%$

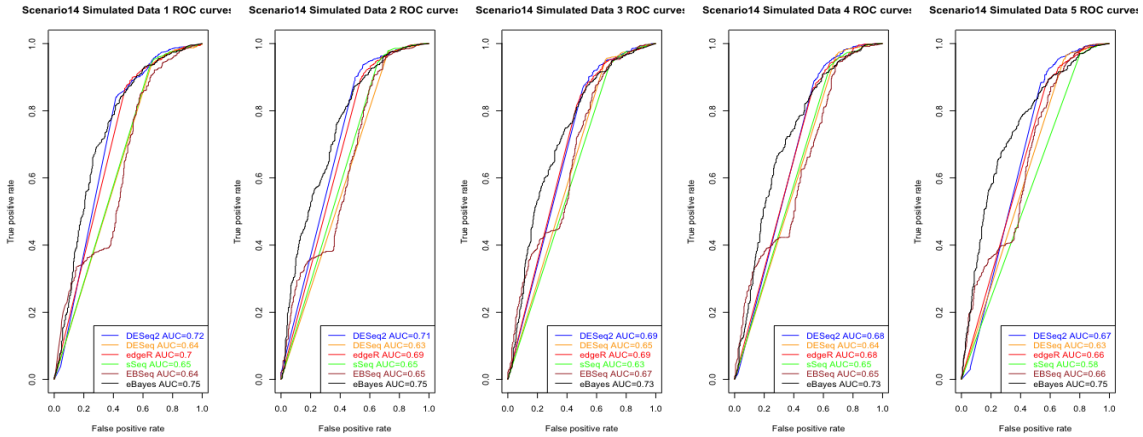

Figure 13 ROC curves of Simulated Datasets with $nGenes = 1000, nSamples = 4, pDiff = 30\%$

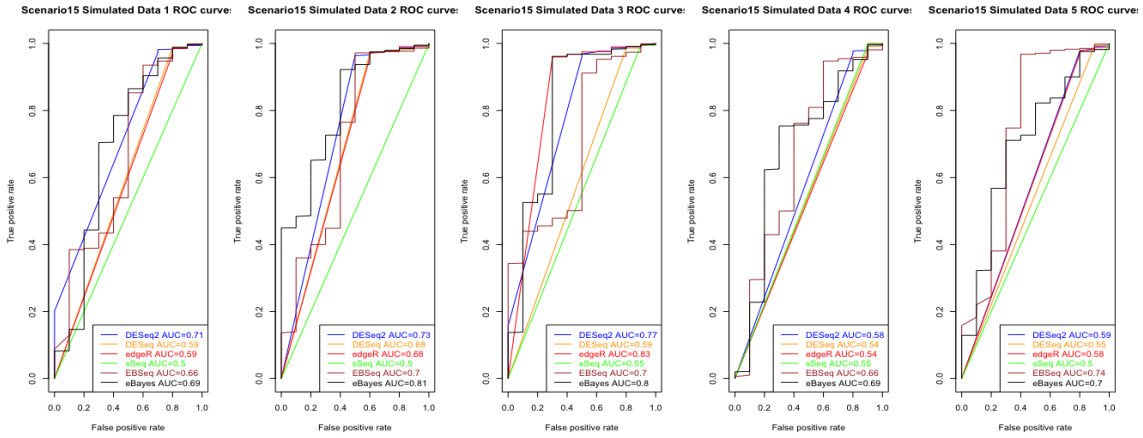

Figure 14 ROC curves of Simulated Datasets with $nGenes = 1000, nSamples = 4, pDiff = 1\%$

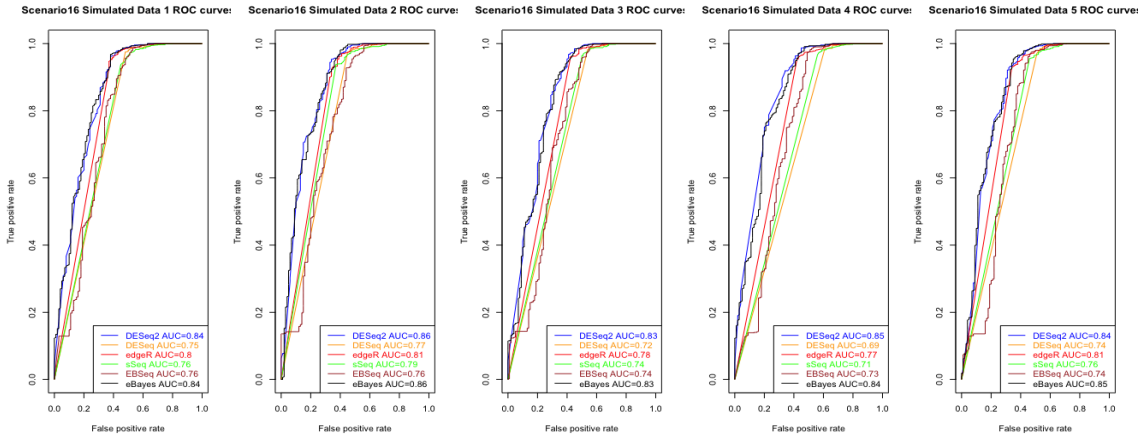

Figure 15 ROC curves of Simulated Datasets with $nGenes = 1000, nSamples = 16, pDiff = 10\%$

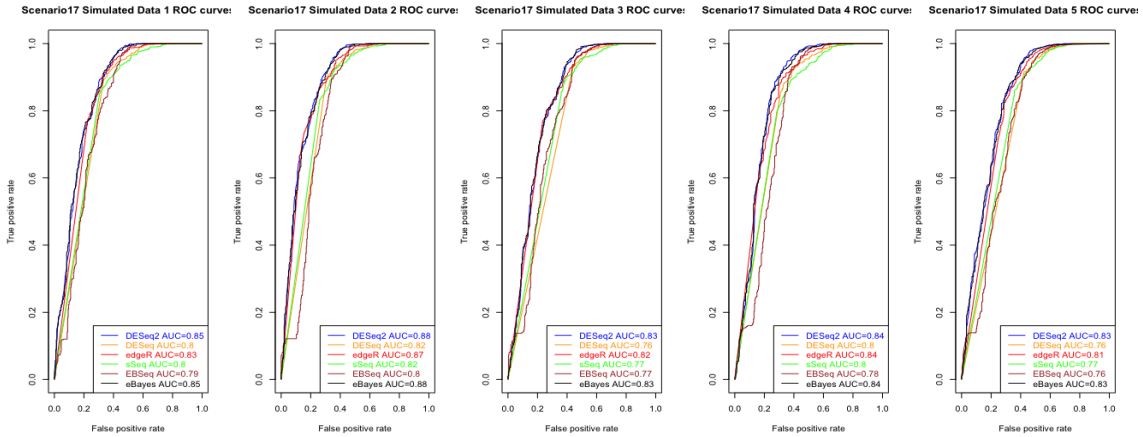

Figure 16 ROC curves of Simulated Datasets with $nGenes = 1000, nSamples = 16, pDiff = 30\%$

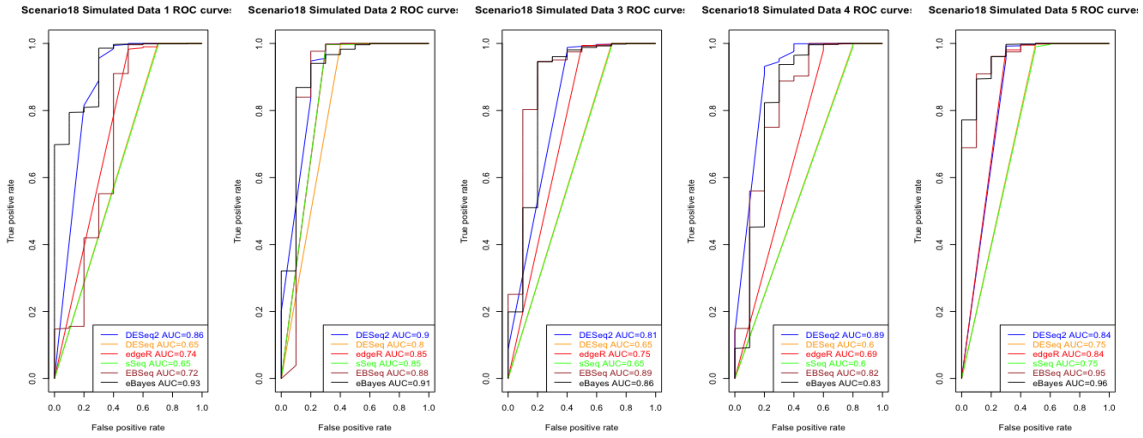

Figure 17 ROC curves of Simulated Datasets with $nGenes = 1000, nSamples = 16, pDiff = 1\%$