# SIEVE: One-stop differential expression, variability, and skewness analyses using RNA-Seq data

Hongxiang Li[1,2,3] and Tsung Fei Khang[3,4*]

[1]Department of Mathematics, Yunnan Normal University, Kunming 650500, Yunnan, P.R. China.
[2]Yunnan Key Laboratory of Modern Analytical Mathematics and Applications, Kunming 650500, Yunnan, P.R. China.
[3]Institute of Mathematical Sciences, Faculty of Science, Universiti Malaya, 50603 Kuala Lumpur, Malaysia.
[4]Universiti Malaya Centre for Data Analytics, Universiti Malaya, 50603 Kuala Lumpur, Malaysia.

*Corresponding author(s). E-mail(s): tfkhang@um.edu.my;

**Abstract**

**Motivation:** RNA-Seq data analysis is commonly biased towards detecting differentially expressed genes and insufficiently conveys the complexity of gene expression changes between biological conditions. This bias arises because discrete models of RNA-Seq count data cannot fully characterize the mean, variance, and skewness of gene expression distribution using independent model parameters. A unified framework that simultaneously tests for differential expression, variability, and skewness is needed to realize the full potential of RNA-Seq data analysis in a systems biology context.

**Results:** We present SIEVE, a statistical methodology that provides the desired unified framework. SIEVE embraces a compositional data analysis framework that transforms discrete RNA-Seq counts to a continuous form with a distribution that is well-fitted by a skew-normal distribution. Simulation results show that SIEVE controls the false discovery rate and probability of Type II error better than existing methods for differential expression analysis. Analysis of the Mayo RNA-Seq dataset for Alzheimer's disease using SIEVE reveals that a gene

set with significant expression difference in mean, standard deviation and skewness between the control and the Alzheimer's disease group strongly predicts a subject's disease state. Furthermore, functional enrichment analysis shows that relying solely on differentially expressed genes detects only a segment of a much broader spectrum of biological aspects associated with Alzheimer's disease. The latter aspects can only be revealed using genes that show differential variability and skewness. Thus, SIEVE enables fresh perspectives for understanding the intricate changes in gene expression that occur in complex diseases.

**Availability:** The SIEVE R package and source codes are available at https://github.com/Divo-Lee/SIEVE.

**Author Summary:** In RNA-Seq data analysis, the majority of current tools focus only on finding with significant differences in mean expression levels between groups, but other variations in gene expression patterns, such as the variance and even the skewness, are potentially important. We developed SIEVE, a new method that enables the simultaneous detection of genes that have significant differences in mean, variance and skewness between two biological conditions. The key innovation is a data transformation that renders the analysis tractable, making it possible to detect more complex patterns in RNA-Seq data. Applied to a large Alzheimer's Disease dataset, we showed that SIEVE confirmed many previously known findings, as well as uncovered novel genes and biological pathways that were missed by traditional methods. SIEVE should be useful for researchers who study complex conditions by enabling a much richer class of biologically relevant genes to be detected and studied than was previously possible.

# 1 Introduction

The goal of gene expression analysis is to uncover genes with patterns of expression variation that are significantly associated with variation of some biological states. Three such patterns are of potential biological interest: (i) differential expression (DE); (ii) differential variability (DV); (iii) differential skewness (DS). These correspond to shifts in the mean, variance, and skewness of gene expression levels respectively when comparing two biological conditions.

By far, the statistical detection of genes that are significantly up or down-regulated (i.e. DE genes) forms the main use of RNA-Seq data. While finding significant DE genes is important, not all genes that are associated with change in biological states show change in mean expression level. For example, a gene's

previously tightly-controlled expression range can be lost, resulting in expression over a much wider range (Bahar et al, 2006; Cheung et al, 2003). Pritchard et al (2001) reported that genes associated with modulation of the immune system, stress and hormonal regulation have high gene expression variability. Furthermore, gene expression variability often shows consistent differences between individuals with and without certain disease states (Ho et al, 2008; Mar et al, 2011; Gorlov et al, 2012). In cancer biology, cancer tissues show higher variability of gene expression compared with normal tissues (Gorlov et al, 2012; Corrada Bravo et al, 2012). In all these cases, the focus is on genes that show differential variability (i.e. DV genes). Although some methods of detecting DV genes are available, they are not integrated with the framework of commonly used DE tests. Differences in the skewness of gene expression levels have also been suggested to affect disease pathogenesis (Casellas and Varona, 2012; Marko and Weil, 2012; Church et al, 2019; Mar, 2019). Skewness, which takes higher moments into account, characterizes the shape of the distribution of gene expression levels. It can potentially capture more nuanced patterns of variation compared to the mean and the variance. Church et al (2019) reported elevated positive skewness in the expression of genes within immune-related pathways when comparing cancer against control subjects. Unfortunately, no methods are currently available for detecting the class of genes that show differetial skewness (i.e. DS genes).

Fundamentally, the absence of a unified framework in RNA-Seq data analysis that allows simultaneous testing of DE, DV, and DS has led researchers to focus predominantly on DE tests, thus unintentionally biasing findings. Such biases are the inevitable outcome of letting tool availability dictate analysis, rather than the biological question. The first step toward correcting this bias is the development of a unified framework.

Conventional discrete count models have been effective in developing useful statistical tests of DE, but the lack of model parameters that explicitly characterize variance or skewness makes the development of DV and DS tests difficult. For context, we first provide an overview of DE and DV tests in RNA-Seq data analysis. The discrete negative binomial (NB) model underlies popular DE methods such as edgeR (Robinson and Smyth, 2007; McCarthy et al, 2012) and DESeq2 (Love et al, 2014). For the NB model, the variance $\sigma^2$ is modeled as a linear ($\sigma^2 = \phi\mu$) or quadratic ($\sigma^2 = \mu + \phi\mu^2$) function of the population mean $\mu$, with $\phi$ as the dispersion parameter. These two parametrizations

lead to the negative binomial 1 (NB1) and 2 (NB2) models, respectively. However, since mean and variance are confounded, it may be difficult to use these models effectively for analysis of gene expression variability (Geiler-Samerotte et al, 2013). Other models such as the Generalized Poisson (Srivastava and Chen, 2010) and the Poisson-Tweedie distribution (Esnaola et al, 2013) have also been used. Alternatively, the voom algorithm (Law et al, 2014) assumes a Gaussian distribution for the logarithm of RNA-Seq counts, and repurposes the limma pipeline (Smyth, 2005) used in microarrays for DE analysis. Gauthier et al (2020) proposed dearseq, a distribution-agnostic method that uses a test statistic based on the variance component score. Nonparametric methods (e.g. NOISeq (Tarazona et al, 2015, 2011) and SAMSeq (Li and Tibshirani, 2013)) are available, but they are less commonly used compared to DESeq2 or edgeR. Nevertheless, when sample size is sufficiently large, the Wilcoxon rank-sum test has been shown to have better false discovery rate (FDR) control compared to edgeR, DESeq2 and dearseq (Li et al, 2022).

In contrast, there are fewer DV tests available, and they are also less widely used. Currently, five DV methods are available: (i) DiffVar (Phipson and Oshlack, 2014), an empirical Bayes method that uses the limma framework (Smyth, 2005); (ii) MDSeq (Ran and Daye, 2017), which uses an NB1 generalized linear model to test DE and DV separately; (iii) the generalized additive model for location, scale and shape (GAMLSS; Rigby and Stasinopoulos, 2005), which tests the effects of biological factors on a gene's Poisson and non-Poisson variation using the NB2 model (de Jong et al, 2019); (iv) DiffDist (Roberts et al, 2022), a hierarchical Bayesian model based on the NB2 distribution; (v) clrDV (Li and Khang, 2023), our recent contribution that uses the compositional data framework in the present work but restricted to DV test.

The non-normality of gene expression distribution, particularly as characterized by skewness, may be potentially crucial under certain biological contexts (Mar, 2019). It can shed light on gene functionalities that cannot be discovered using conventional analysis of difference in mean - and, less commonly, differences in variability (Hasegawa et al, 2015; Ho et al, 2008; Mar et al, 2011). The relevance of skewness of distributions of gene expression as measured using microarray data was documented in several studies (Casellas and Varona, 2012; Hardin and Wilson, 2009; Marko and Weil, 2012; Thomas et al, 2010). More recently, Church et al (2019) highlighted skewness as an effective indicator of biological heterogeneity in gene pathways associated with specific

cancer cohorts. Despite these studies, no methods are currently available for testing DS.

A unified framework that enables the simultaneous testing of mean, variance and skewness requires a model for which these parameters are explicit and independent of one another. This suggests transformation of RNA-Seq count data into continuous data, followed by their appropriate modeling using a continuous distribution. There is good evidence that this approach is promising. Quinn et al (2018) adopted a compositional data framework for RNA-Seq data and repurposed the ALDEx2 algorithm (Fernandes et al, 2013, 2014) - originally developed for differential abundance testing in microbiome data - for detecting DE genes. ALDEx2 incorporates a Dirichlet process to produce centered log-ratio (CLR) transformation (Aitchison, 1986) of count data. The transformed data is assumed to be normally distributed, and a Welch t-test or Wilcoxon test is then applied to test for differential abundance between two populations. Interestingly, ALDEx2 has been shown to have comparable or superior statistical properties against edgeR and DESeq2 (Quinn et al, 2018). Although the Welch t-test is adequate for testing equality of mean due to its robustness against non-normality (Rasch et al, 2007), without explicitly modeling the CLR-transformed data, it is unclear how tests concerning other aspects such as equality of variances or skewness can be done.

Toward the goal of erecting a unified framework for RNA-Seq data analysis, it is necessary to use a model that is both mathematically tractable and provides a good empirical fit the transformed RNA-Seq data. To this end, we present SIEVE (one-*S*top d*I*fferential *E*xpression, *V*ariability, and sk*E*wness analyses using RNA-Seq data), our solution for the simultaneous testing of differential expression, variability, and skewness of genes using RNA-Seq data. SIEVE adopts a compositional data analysis approach to modeling discrete RNA-Seq count data, applies Aitchison's CLR transformation (Aitchison, 1986) to convert them into continuous form, and uses a skew-normal distribution to model them. We previously described this framework in the context of developing a DV test, which we named clrDV (Li and Khang, 2023). With SIEVE, genes with significant variation in skewness can now be detected for the first time, alongside those with significant variation in mean and standard deviation.

# 2 Materials and Methods

## 2.1 The compositional data framework and the centered log-ratio transformation

Let $X_{gi}$ be the read count of gene $g$ and sample $i$, where $g = 1, 2, \ldots, G$ and $i = 1, 2, \ldots, n$. The $i$th sample data becomes compositional if we normalize the counts by dividing them with the sum of gene counts. Application of the CLR transformation converts the variable space from the simplex to the Euclidean space, for which standard mathematical and statistical methods operate in. The CLR-transformation of the count of the $i$th gene is defined as the logarithm of the gene count divided by the geometric mean of gene counts across all $G$ genes. To avoid multiplication or division by 0, a pseudo-value of 0.5 replaces any zero gene counts.

## 2.2 Modeling Centered Log-ratio-transformed Data Using the Skew-normal Distribution

Denote the CLR-transformed count from gene $g$ in sample $i$ by $Y_{gi}$. We shall assume a skew-normal distribution with centered parameters (CP) for the random variable $Y_{gi}$. Notationally, we write $Y_{gi} \sim \mathrm{SN_C}(\mu_g, \sigma_g, \gamma_g)$, where $\mu_g$ is the mean, $\sigma_g$ is the standard deviation, and $\gamma_g$ is the skewness parameter (see Supplementary Materials S1 for mathematical details). The parameter vector $\boldsymbol{\theta}_g^{(C)} = (\mu_g, \sigma_g, \gamma_g)$ has parameter space $\mathbb{R} \times \mathbb{R}^+ \times (-k, k)$, where $k = \sqrt{2}(4-\pi)/(\pi-2)^{3/2} \approx 0.9953$. The skew-normal distribution simplifies to the normal distribution with mean $\mu_g$ and variance $\sigma_g^2$ when $\gamma_g = 0$. Given two populations and sufficiently large sample sizes, tests of equality of mean, standard deviation and skewness can be done using the Wald statistic:

$$Z_g = \frac{\hat{\theta}_{g,2} - \hat{\theta}_{g,1}}{\sqrt{\mathrm{Var}(\hat{\theta}_{g,2}) + \mathrm{Var}(\hat{\theta}_{g,1})}},$$

where $g = 1, 2, \ldots, G$ index the genes, and $j = 1, 2$ index the groups. Denote the maximum likelihood estimator (MLE) or the maximum penalized likelihood estimator (MPLE; Azzalini and Arellano-Valle, 2013; Azzalini and Capitanio, 2014) as $\hat{\theta}_{g,j}$. Then, for testing DE, $\hat{\theta}_{g,j} = \hat{\mu}_{g,j}$; for testing DV, $\hat{\theta}_{g,j} = \hat{\sigma}_{g,j}$; for testing DS, $\hat{\theta}_{g,j} = \hat{\gamma}_{g,j}$. For estimating $\mathrm{Var}(\hat{\theta}_{g,j})$, the inverse of

the Fisher information matrix of $\boldsymbol{\theta}_{g,j}^{(C)}$ is used. When sample size is sufficiently large, $Z_g$ converges to the standard normal distribution. The Benjamini-Yekutieli (BY) procedure (Benjamini and Yekutieli, 2001), which allows for arbitrary dependence between the tested hypotheses, is used to control the FDR and adjust $p$-values for multiple hypothesis testing.

Under the SIEVE framework, eight gene classes are possible (see also Fig. S1 in the Supplementary Materials). These range from genes that do not show these properties, to genes that simultaneously show them (Table 1).

**Table 1**: Eight gene classes identifiable using SIEVE. Abbreviations: DE = differential expression; DV = differential variability; DS = differential skewness; 1 = True; 0 = False.

| Gene class | DE | DV | DS | Label |
|---|---|---|---|---|
| non-DE, non-DV and non-DS | 0 | 0 | 0 | 000 |
| pure DE | 1 | 0 | 0 | 100 |
| pure DV | 0 | 1 | 0 | 101 |
| pure DS | 0 | 0 | 1 | 001 |
| DE, DV, non-DS (DEV) | 1 | 1 | 0 | 110 |
| DE, non-DV, DS (DES) | 1 | 0 | 1 | 101 |
| non-DE, DV, DS (DVS) | 0 | 1 | 1 | 011 |
| DE, DV, DS (DEVS) | 1 | 1 | 1 | 111 |

## 2.3 Simulation Design for Evaluating Performance of Differential Expression Tests

To compare the performance of SIEVE with existing DE methods, RNA-Seq count data were simulated from an NB2 model with model parameters estimated from real datasets using `polyester` R (Frazee et al, 2015). Three RNA-Seq datasets with reasonably large sample sizes were used to facilitate reliable estimation of the standard deviation and skewness parameters.

The first dataset (GEO accession number: GSE123658) contains whole blood RNA-Seq data from 39 Type 1 diabetes patients and 43 healthy donors (Leal Valentim et al, 2020), with 16,785 genes. The second dataset (GEO accession number: GSE150318) consists of longitudinal RNA-Seq data from 114 short-lived killfish *Nothobranchius furzeri*. For each fish, samples were taken at 10 and 20 weeks of age. A total of 26,739 genes are available (Kelmer Sacramento et al, 2020). The third dataset (GEO accession number: GSE179250) comprises 29,245 genes from 192 human liver samples (Zhou et al,

2022). Hereafter, we refer to these three datasets as the Valentim, Kelmer, and Zhou datasets.

For use in the simulation study, only genes that had average count-per-million (CPM) above 0.5 and zero counts in not more than 85% of the samples were admitted. After filtering, the Valentim, Kelmer and Zhou datasets consisted of 12,283, 16,670, and 17,862 genes, respectively.

Assuming the distribution of the RNA-Seq count data follows the NB2 model, we randomly sampled 5,000 of the filtered genes and then estimated the parameters of the NB2 model using samples from one group (Type 1 diabetes group in the Valentim dataset ($n = 39$); 10-weeks group in the Kelmer dataset ($n = 114$); all liver samples in the Zhou dataset ($n = 192$)). After this, we generated 200 biological replicates from the estimated NB2 models. The goodness-of-fit of the skew-normal model on the distribution of the CLR-transformed simulated count data was checked using the Kolmogorov-Smirnov (KS) test.

Using the FDR and the probability of Type II error ($\beta$) as performance metrics, we compared SIEVE against other nine other DE methods: edgeR, DESeq2, voom, DSS, tweeDEseq, NOISeq, ALDEx2, dearseq, and the Wilcoxon rank-sum test. edgeR, DESeq2 and voom were selected because of their popularity as DE tests. Then, to comprehensively survey the landscape of DE tests using a diversity of approaches, we included non-parametric methods such as NOISeq and the Wilcoxon rank-sum test, as well as DSS, dearseq and tweeDEseq, which are based on ideas of dispersion shrinkage, variance component score, and the Poisson-Tweedie model, respectively. RNA-Seq data were similarly simulated using the same three datasets with `polyester` R (Frazee et al, 2015). For each of the three datasets, a total of 2000 filtered genes were randomly selected, and their mean ($\mu$) and size ($1/\phi$) parameters under assumption of an NB2 model were estimated. The samples used were the 39 Type I diabetes patients for the Valentim dataset, the 114 fish at 10 weeks of age for the Kelmer dataset, and all 192 livers for the Zhou dataset.

In a two-group comparison setting, both groups have identical NB2 model parameters under the null hypothesis of equal mean. Hence, to simulate DE genes, we spiked 10% of the genes to be differentially expressed by multiplying or dividing their estimated mean parameter with a random value uniformly chosen between 2 to 4. Biological replicates of size 30, 50, and 100 per group were then generated from the NB2 model, with 30 instances simulated for all three datasets.

We defined the region of desirable performance as where both FDR and $\beta$ are below 0.05. Then, we computed the percentage of simulated instances falling within this region ($\theta$). For SIEVE and ALDEx2, we applied the CLR transformation to the raw counts; for the other eight methods, we normalized the raw counts using the trimmed mean of M values (TMM; Robinson and Oshlack, 2010). We identified genes with adjusted $p$-value below 0.05 as DE genes, and recorded the computing time of each method.

To evaluate performance similarity of DE methods, we used a cluster heatmap with the Ward clustering algortihm and between-sample distances measured using the Euclidean distance. We computed the mean and standard deviation of FDR and $\beta$ (which are used as features) for the 30 simulated instances across each of the three sample size scenarios, and subjected them to min-max normalization before constructing the cluster heatmap.

Finally, to evaluate the effect of using pseudo-values that are smaller than 0.5, we conducted a sensitivity analysis using two smaller pseudo-values (0.01 and 0.1)

## 2.4 Empirical Validation

To ascertain the value of mining DE, DV, and DS genes using SIEVE, we applied SIEVE to the Mayo RNASeq dataset (Allen et al, 2016) to assess its capability for detecting genes that are contextually meaningful in Alzheimer's disease (AD). This dataset consists of $64,253$ genes and 278 samples from patients who are normal ($n = 80$), have Alzheimer's disease ($n = 84$), progressive supranuclear palsy ($n = 84$) and pathologic aging ($n = 30$). Here, we compared the AD group against the control group. We chose the Mayo RNA-Seq dataset for two main reasons. Firstly, the study has a rigorous experimental design and quality control for ensuring data fidelity. Secondly, the sample size per group is sufficiently large (50 or more per group) to allow the standard deviation and skewness parameters of the skew-normal model to be reliably estimated (Azzalini and Capitanio, 2014; Azzalini, 2022). Thirdly, analysis of an AD dataset would enable us to leverage on the substantial literature in this well-researched field to cross-check the biological significance of genes that are detected using SIEVE. The KS test was used to check the fit of the skew-normal distribution on the CLR-transformed count data.

During the data preprocessing step, we excluded genes that have average CPM $< 0.5$ and genes with zero counts in 85% or more samples. Furthermore,

we removed samples with missing value labels. At the end, we retained a total of 18,664 genes, with 78 control and 82 AD samples.

For comparison of DE methods, we selected those that showed reasonably good performance in the simulation study. Comparison of DV methods have already been reported previously by us in Li and Khang (2023). Volcano plots were used to inspect joint patterns of biological and statistical significance. Biological significance was measured using the difference of mean for DE genes, $\log_2$ of the ratio of standard deviation for DV genes, and the difference of skewness for DS genes. Thus, zero corresponds to genes with no significant biological difference between the two groups. Significant DE, DV and DS genes were flagged using the significance score method (Xiao et al, 2014; see Supplementary Materials S2). Sets of DE genes detected using the different methods were examined using Venn diagrams. Finally, the joint distributions of estimated mean, standard deviation, and skewness parameters for the control and the AD group were visualized using contour heatmaps.

Subsets of the genes detected using SIEVE that are strongly predictive of the AD state were identified using the Generalized, Unbiased Interaction Detection and Estimation (GUIDE, Version 40.3; Loh, 2022) classification and regression tree algorithm (Loh, 2009). To assess the informativeness of the set of significant DE, DV, and DS genes (the top 10 genes on the left and the right tails of the distribution of biological significance scores) for predicting the AD state, we estimated the out-of-bag generalization error using the GUIDE random forests model with default hyperparameters (equal priors, unit misclassification costs, univariate split highest priority, no interaction splits, number of trees = 1000, SE parameter = 0.25). For benchmarking, AD state classification using GUIDE was done using a random subset of 60 putatively uninformative genes that are non-DE, non-DV, and non-DS with BY-adjusted $p$-value of 1.

To check whether the lists of DE, DV, and DS genes detected using SIEVE are contextually meaningful, we conducted gene set functional enrichment analysis using WebGestalt (Liao et al, 2019; Zhang et al, 2005). We explored enriched biological processes in the gene ontology (GO) functional database using over representation analysis of the following five gene lists: (i) high confidence DE genes obtained the intersection of sets of DE genes detected using SIEVE, edgeR, DESeq2, voom, tweeDEseq, ALDEx2, DSS and Wilcoxon; (ii) non-DE, DV and/or DS genes; (iii) pure DV genes; (iv) pure DS genes; and

(v) the union of DE, DV, and DS genes. The human genome reference set was chosen. Only the top 20 GO terms were included, and affinity propagation was used for redundancy reduction. GO hierarchies were checked using WebGestalt's GOView application (Liao et al, 2019).

### 2.5 Computing Tools and Environment

Computational works for DE, DV and DS tests were done using a MacBook Pro (macOS Monterey version 12.5) with 32GB RAM, an M1 pro chip and a 10-core CPU. The R (version 4.2.1; R Core Team, 2022) computing environment was used. Supplementary Materials S6 gives the complete list of R packages used. For converting ENSEMBL gene ID to gene symbol, we used the application programming interface of www.biotools.fr (Saurin, 2022).

## 3 Results

### 3.1 Simulation Study

The skew-normal distribution fits the CLR-transformed count data of genes in the three datasets well, with 99.2%, 98.9% and 98.7% of genes having KS-test $p$-values above 0.05 (Fig. 1). Graphical summaries of the simulation results are given in Fig. 2, 3, and 4. For brevity, details of the summary statistics of the performance metrics are given only for the Valentim dataset (Table 2). For the other two simulations, see Tables S1 and S2 in the Supplementary Materials.

Overall, SIEVE, ALDEx2 and NOISeq prioritise FDR control over $\beta$. In all three simulations, these three methods, along with edgeR, are the top four methods with lower mean FDR. Specifically, NOISeq generally has the lowest mean FDR, followed by SIEVE, edgeR, and ALDEx2 in the simulated Valentim and Zhou datasets. In the simulated Kelmer dataset, NOISeq again has the lowest FDR, followed by SIEVE, ALDEx2, and edgeR. However, we note that NOISeq also has the highest mean $\beta$ among all DE methods compared. ALDEx2 has relatively higher mean $\beta$, followed by SIEVE and edgeR. In all three simulations, edgeR has mean FDR that is relatively constant at about 0.03 (range: 0.026 to 0.042) across the three sample size scenarios. Furthermore, its mean $\beta$ is at most about 0.08 at $n = 30$, and almost 0 as $n$ increases to 100.

In contrast, DESeq2, voom, tweeDEseq, DSS, dearseq, and the Wilcoxon rank-sum test control FDR at around 0.05, with generally greater variation than edgeR's. As expected, all methods have decreasing $\beta$ as sample size increases.

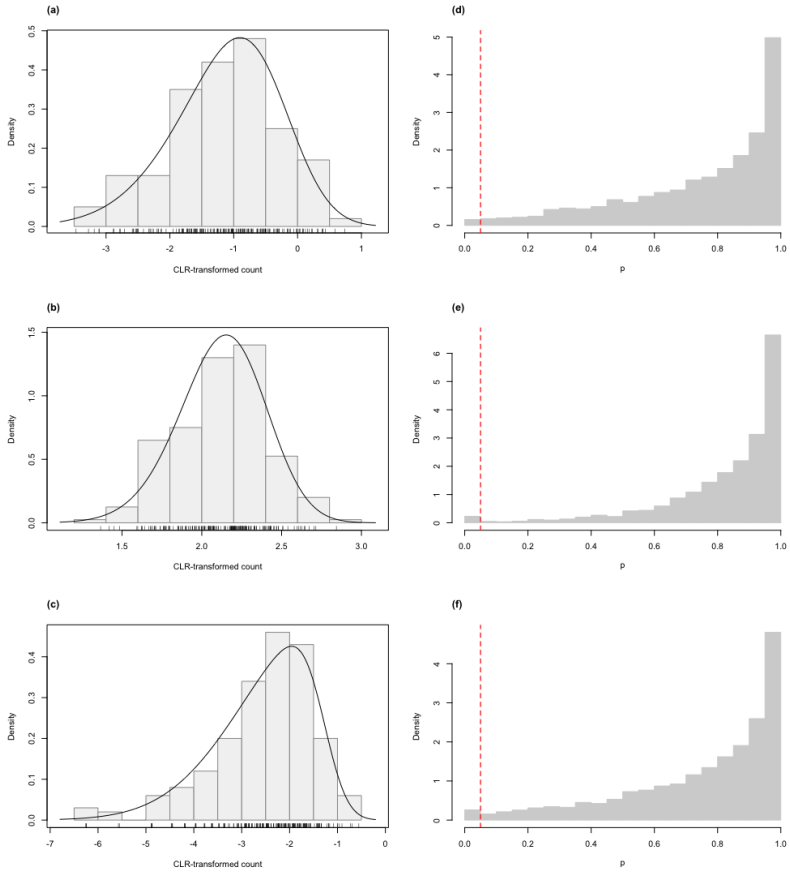


**Fig. 1**: Histograms of CLR-transformed counts for three selected genes with fitted skew-normal curve for (a) the Valentim dataset ($\hat{\mu} = -1.256$ (s.e. = 0.061), $\hat{\sigma} = 0.861$ (s.e. = 0.046) and $\hat{\gamma} = -0.450$ (s.e. = 0.172)); (b) the Kelmer dataset ($\hat{\mu} = 2.119$ (s.e. = 0.019), $\hat{\sigma} = 0.274$ (s.e. = 0.014) and $\hat{\gamma} = -0.232$ (s.e. = 0.196)); (c) the Zhou dataset ($\hat{\mu} = -2.564$ (s.e. = 0.073), $\hat{\sigma} = 1.040$ (s.e. = 0.057) and $\hat{\gamma} = -0.799$ (s.e. = 0.071)). Distribution of the $p$-values of the Kolmogorov-Smirnov goodness-of-fit tests of the skew-normal model for genes in the simulated (d) Valentim dataset; (e) Kelmer dataset; and (f) Zhou dataset. The skew-normal model gives good fit to about 99.2%, 98.9%, and 98.7% of the genes in (d), (e), and (f), respectively. The red dashed line corresponds to the threshold $p$-value of 0.05.

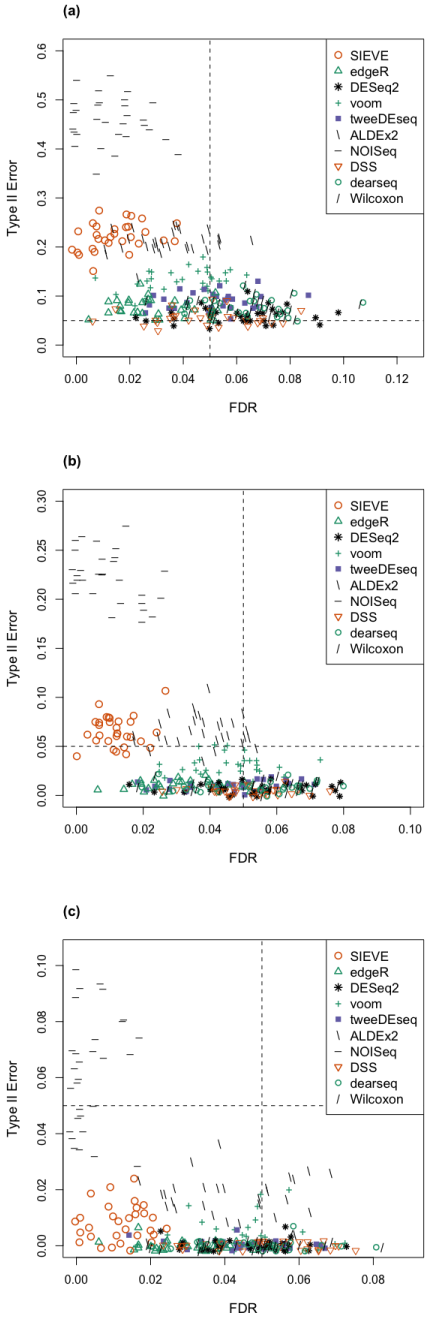


**Fig. 2**: Scatter plots of the probability of Type II error ($\beta$) against FDR for simulated data from the Valentim dataset (30 instances) for three sample size per group scenarios: (a) 30; (b) 50; (c) 100. Dashed lines represent the desired threshold FDR and $\beta$ of 0.05.

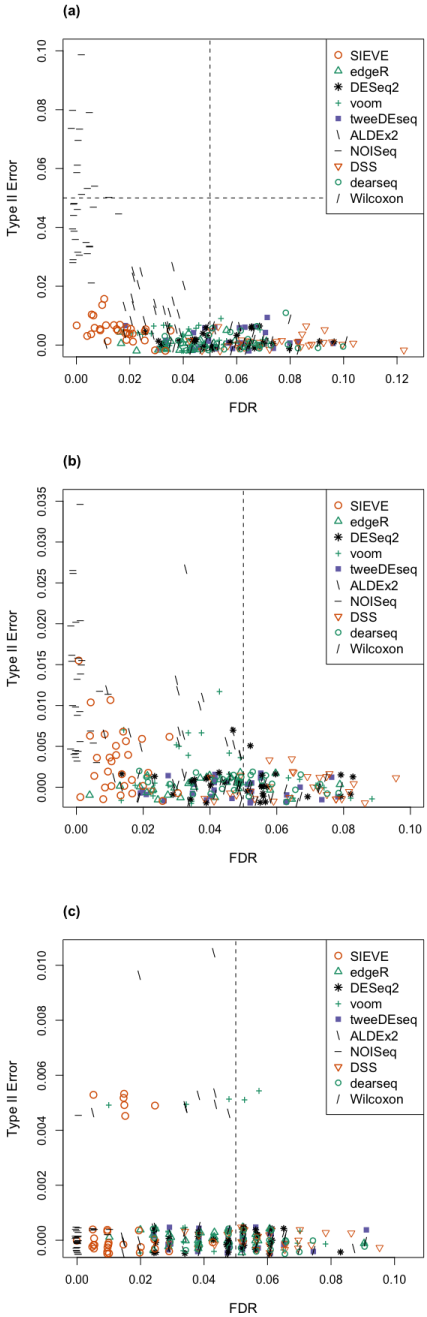


**Fig. 3**: Scatter plots of the probability of Type II error ($\beta$) against FDR for simulated data from the Kelmer dataset (30 instances) for three sample size per group scenarios: (a) 30; (b) 50; (c) 100. Dashed lines represent the desired threshold FDR and $\beta$ of 0.05.

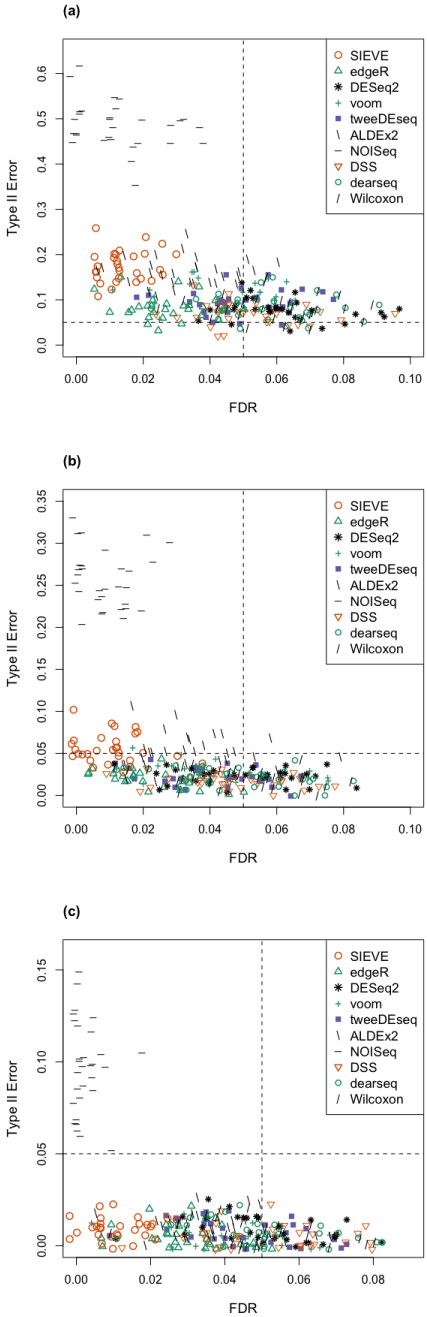


**Fig. 4**: Scatter plots of the probability of Type II error ($\beta$) against FDR for simulated data from the Zhou dataset (30 instances) for three sample size per group scenarios: (a) 30; (b) 50; (c) 100. Dashed lines represent the desired threshold FDR and $\beta$ of 0.05.

**Table 2**: The mean of false discovery rate (FDR$^*$), mean probability of Type II error ($\beta^\dagger$), and percentage of simulated instances with FDR $< 0.05$ and $\beta < 0.05$ ($\theta^\ddagger$) for each of the ten DE tests applied to data simulated from the Valentim dataset (30 instances) at three different sample sizes. Standard deviation in parentheses.

| Method | Sample size per group | | |
|---|---|---|---|
| | 30 | 50 | 100 |
| SIEVE | 0.014 (0.011) | 0.012 (0.006) | 0.012 (0.007)$^*$ |
| | 0.221 (0.028) | 0.066 (0.015) | 0.008 (0.007)$^\dagger$ |
| | 0% | 13.3% | 100%$^\ddagger$ |
| ALDEx2 | 0.034 (0.014) | 0.038 (0.011) | 0.041 (0.015) |
| | 0.210 (0.022) | 0.066 (0.016) | 0.017 (0.008) |
| | 0% | 10% | 66.7% |
| NOISeq | 0.013 (0.011) | 0.009 (0.008) | 0.004 (0.005) |
| | 0.458 (0.047) | 0.223 (0.026) | 0.062 (0.020) |
| | 0% | 0% | 30% |
| edgeR | 0.029 (0.013) | 0.031 (0.012) | 0.029 (0.010) |
| | 0.076 (0.018) | 0.010 (0.005) | 0.000 (0.001) |
| | 0% | 96.7% | 100% |
| DESeq2 | 0.062 (0.019) | 0.054 (0.016) | 0.047 (0.012) |
| | 0.060 (0.017) | 0.007 (0.005) | 0.000 (0.002) |
| | 6.7% | 40% | 50% |
| voom | 0.041 (0.014) | 0.045 (0.012) | 0.046 (0.014) |
| | 0.134 (0.021) | 0.031 (0.010) | 0.004 (0.006) |
| | 0% | 60% | 66.7% |
| DSS | 0.049 (0.018) | 0.051 (0.011) | 0.051 (0.015) |
| | 0.056 (0.015) | 0.004 (0.004) | 0.000 (0.000) |
| | 13.3% | 46.7% | 46.7% |
| dearseq | 0.063 (0.016) | 0.053 (0.015) | 0.047 (0.014) |
| | 0.077 (0.017) | 0.009 (0.005) | 0.000 (0.001) |
| | 0% | 33.3% | 50% |
| Wilcoxon | 0.047 (0.012) | 0.045 (0.012) | 0.045 (0.014) |
| | 0.123 (0.024) | 0.022 (0.009) | 0.001 (0.003) |
| | 0% | 76.7% | 73.3% |
| tweeDEseq | 0.052 (0.014) | 0.046 (0.014) | 0.044 (0.012) |
| | 0.089 (0.019) | 0.010 (0.005) | 0.000 (0.001) |
| | 0% | 56.7% | 73.3% |

At $n = 100$, only SIEVE has $\theta = 100\%$ for all three simulations (see Tables 2, S1, and S2). SIEVE, edgeR, and ALDEx2 have $\theta$ ranging from $66.7-100\%$, $70-100\%$ and $83.3-100\%$ for the three simulations, compared to $46.7-73.3\%$, $26.7-67.7\%$, and $43.3-66.7\%$ for DESeq2, voom, tweeDEseq, DSS, dearseq, and the Wilcoxon rank-sum test. NOISeq shows inconsistent performance, with $\theta$ ranging from 0% to 100%.

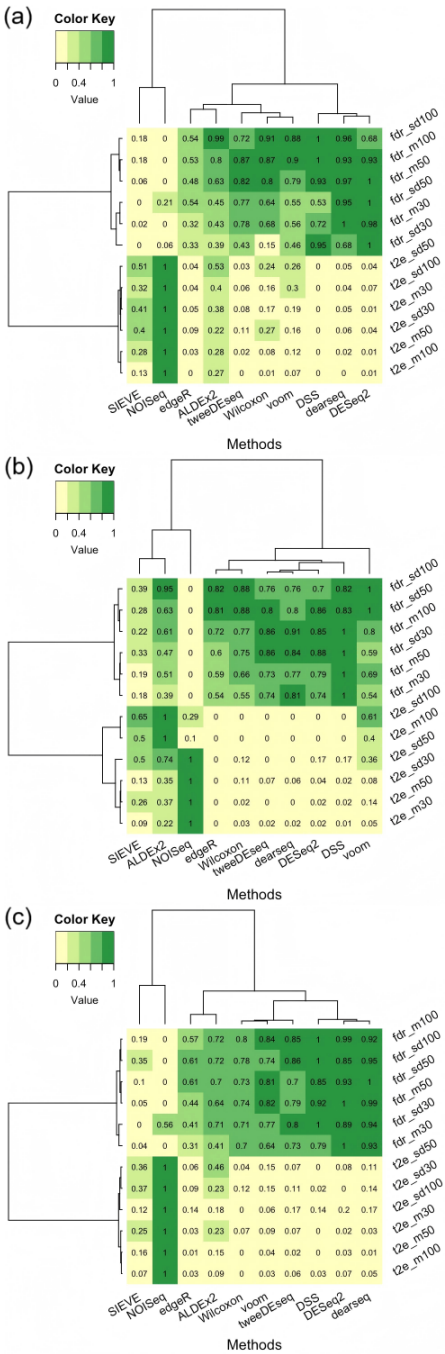


**Fig. 5**: Cluster heatmaps for showing similarity of DE methods (columns) with respect to summary statistics (rows) of FDR and probability of Type II error ($\beta$) at three sample size scenarios ($n = 30, 50, 100$) for the (a) Valentim dataset; (b) Kelmer dataset; and (c) Zhou dataset. The number in a cell represents the mean (over 30 instances) of the summary statistic of a specific feature. Abbreviations: fdr = FDR; m = mean; sd = standard deviation; t2e = probability of Type II error; numerical suffixes indicate corresponding sample size scenario.

The cluster heatmap (Fig. 5) shows that NOISeq has mean $\beta$ that is consistently the highest, in contrast with its mean FDR, which is consistently the lowest. SIEVE is similar to NOISeq with respect to mean FDR performance, but excels with substantially lower mean $\beta$. This suggests that SIEVE maintains the conservativeness of NOISeq without excessive compromise in $\beta$. Conversely, while other methods often have lower mean $\beta$ values, their mean FDR values are correspondingly higher.

SIEVE took longer time to run when sample size is smaller, but its speed improved for larger sample sizes (Table S3 in Supplementary Materials). At $n = 100$ for all three simulations, edgeR, DESeq2, voom, DSS, dearseq, and the Wilcoxon rank-sum took about 5s or less to complete, whereas NOISeq, ALDEx2 and SIEVE took about 11 to 46s. The slowest method was TweeDESeq, which was 14 to 17 times slower than the second (ALDEx2) and the third (SIEVE) slowest methods, respectively.

The result of the sensitivity analysis for different choice of pseudo-values shows that 0.5 generally provides better control of both FDR and $\beta$, compared to the smaller pseudo-values of 0.01 and 0.1, for data simulated from the Valentim, Kelmer, and Zhou datasets across three sample size scenarios (Table S4 in Supplementary Materials). Specificaly, at sample size of 30, all three pseudo-values give very similar $\beta$ values for all the simulated datasets. For the moderately large sample size of 50, using 0.5 as the pseudo-value tends to provide uniformly better control of both FDR and $\beta$.

## 3.2 Analysis of the Mayo RNA-Seq Dataset

The majority of the genes in both control (98.5%; 18385/18664) and AD groups (99.5%; 18571/18664) have CLR-transformed count data that are well-fitted ($p$-value $> 0.05$) by the skew-normal distribution (Fig. S2 in Supplementary Materials). We performed DE analysis of the Mayo RNA-Seq dataset by comparing SIEVE against edgeR, DESeq2, voom, tweeDEseq, ALDEx2, DSS, Wilcoxon, and NOISeq. We excluded dearseq because it is unable to provide output for the mean expression levels, rendering significance score method used inapplicable. Figure S4 (Supplementary Materials) shows the number of DE genes unique to a DE method or common to multiple DE methods (for complete list, see Table S10 in Supplementary Materials). DSS detected the most DE genes (4171), followed by DESeq2 (4155), edgeR (3942), voom (3636), tweeDEseq (3553), Wilcoxon (3330), ALDEx2 (3230), SIEVE (2773),

and NOISeq (2014). That NOISeq detected the least number of DE genes is consistent with findings from the simulation studies that show its poor control of probability of Type II error. Table S11 in Supplementary Materials shows the list of DE, DV (2276) and DS (1024) genes detected using SIEVE (see also Fig. S3 in Supplementary Materials). With respect to computational time, SIEVE completed the DE, DV, and DS tests concurrently in approximately 5.5 minutes. The run time for other methods (in increasing order) is as follows: 20s for DSS; about 30s for Wilcoxon; about 45s for edgeR, DESeq2, limma-voom and NOISeq; about 40 minutes for ALDEx2; and about 2.5 hours for tweeDEseq.

We defined the intersection of DE genes detected using each method as the high confidence gene set. Here, we excluded NOISeq because simulation studies suggested that it has a high false negative rate. Comparing SIEVE to three popular DE tests (i.e. edgeR, DESeq2 and voom), 98.7% (2736/2773) of DE genes identified by SIEVE are also detected by edgeR, DESeq2, or voom; 91.6% (2540/2773) are detected by edgeR, DESeq2, and voom; only 1.3% (37/2773) are uniquely detected by SIEVE. Furthermore, 93.1% (2583/2773), 94.3% (2615/2773) and 97.0% (2691/2773) of DE genes detected by SIEVE are also identified by edgeR, DESeq2 and voom, respectively. In comparison with ALDEx2, tweeDEseq, DSS and Wilcoxon, about 94.8% (2630/2733), 91.1% (2527/2773), 94.6% (2622/2773) and 91.2% (2528/2773) of DE genes identified by SIEVE are also detected by these four methods, respectively. Overall, 84.9% (2354/2773) of DE genes detected by SIEVE are also detected by all the other seven DE methods.

Table 3 gives the distribution of gene classes returned by SIEVE. The majority of genes are neither DE, DV, nor DS ($\sim$ 70%; 13030/18664). For the remaining 30% of genes, focusing solely on DE genes means that only about 49% (2773/5634) of them would be considered in downstream analyses. The ability of SIEVE to detect the remaining 51% (2861/5634) of DV and/or DS genes that cannot be detected using DE methods sets it apart from all current methods. Variation in the distribution of CLR-transformed gene counts between the control and the AD group with respect to mean, standard deviation and skewness can be seen in the examples given in Fig. S5 to S9 in the Supplementary Materials.

**Table 3**: Eight classes of genes identified by SIEVE for the Mayo RNA-Seq dataset. Abbreviations: 000 = non-DE, non-DV, and non-DS; 100 = DE, non-DV, and non-DS; 010 = non-DE, DV, and non-DS; 001 = non-DE, non-DV, and DS; 110 = DE, DV, and non-DS; 101 = DE, non-DV, and DS; 011 = non-DE, DV, and DS; 111 = DE, DV, and DS.

| Gene class | 000 | 100 | 010 | 001 | 110 | 101 | 011 | 111 |
|---|---|---|---|---|---|---|---|---|
| Number of genes | 13030 | 2284 | 1843 | 919 | 303 | 155 | 99 | 31 |

Table S5 (Supplementary Materials) summarizes the sets of pure DV, pure DS, and DVS genes detected using SIEVE that intersect with the DE gene sets identified by edgeR, DESeq2, voom, ALDEx2, tweeDEseq, DSS, and Wilcoxon. The key observation is that only 497/2861 = 17.4% of the genes in the union set of the pure DV, pure DS and DEV genes can be detected collaterally by the seven DE methods considered. This suggests that SIEVE detects a large number of genes with significant variability and skewness in gene expression distribution that standard DE methods fail to identify.

Figure 6 shows that DE, DV, and DS genes are represented among two alternative GUIDE trees for discriminating AD from the control state. The fibromodulin gene (*FMOD*; DE gene) is critical for enabling the identification of 70.7% (58/82) AD cases. A DS gene - tetratricopeptide repeat domain 7A (*TTC7A*), is crucial for enabling the prediction of the remaining AD cases (18.3%; 15/82) using the corticotropin releasing hormone gene (Fig. 6(a)). If *TTC7A* is removed, the resulting tree uses another DS gene - myotubularin related protein 7 gene (*MTMR7*) to predict the remaining 25.6% (21/82) AD cases (Fig. 6(b)). Overall, the first tree has estimated accuracy, sensitivity, and specificity of 87.5% (140/160), 89.0% (73/82) and 85.9% (67/78), respectively. For second tree, the estimated accuracy, sensitivity, and specificity is 89.4% (143/160), 96.3% (79/82) and 82.1% (64/78), respectively. See Supplementary Materials S5 for a brief summary of the biological relevance of genes in both trees.

The estimated generalization error of GUIDE was approximately the same ($\sim$ 17%) regardless of whether DE, DV and DS gene sets were used separately, or collectively. In contrast, the estimated generalization error of GUIDE using genes that are neither DE, DV nor DS genes was more than twice as large (39.4%; 63/160). This finding indicates that the gene sets detected using SIEVE are indeed informative.

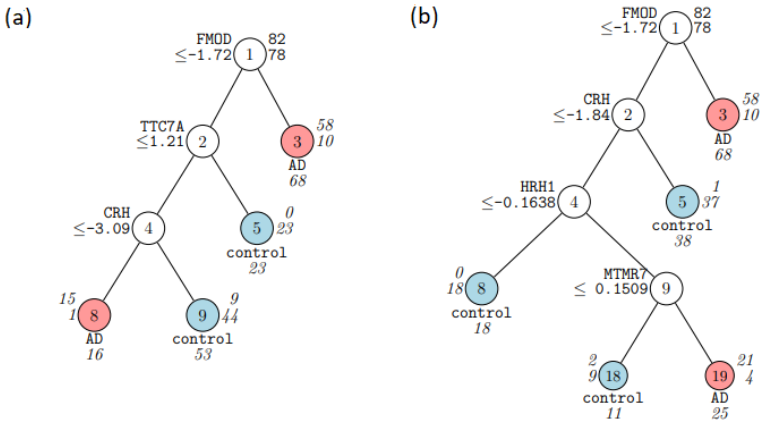


**Fig. 6**: GUIDE v.40.3 0.250-SE classification tree for predicting `label` using equal priors and unit misclassification costs. At each split, an observation goes to the left branch if and only if the condition is satisfied. Predicted classes and sample sizes (in *italics*) printed below terminal nodes; class sample sizes for `label` = `AD` and `control` beside nodes. (a) classification tree with *TTC7A*; (b) classification tree without *TTC7A*. `FMOD`: DE gene; `TTC7A`: DS gene; `CRH`: DE gene; `HRH1`: DV gene; `MTMR7`: DS gene.

The joint distribution of $\hat{\mu}_1$ and $\hat{\mu}_2$ appears approximately bivariate normal (Fig. 7(a)), with DE genes distributed on the periphery of the outermost probability ellipse. Similarly, the joint distribution of $\log_2 \hat{\sigma}_1$ and $\log_2 \hat{\sigma}_2$ is also approximately bivariate normal (Fig. 7(b)). The majority of genes have $\hat{\sigma}$ values between 0 and 1 in the control and the AD group, and are non-DV; DV genes generally have $\hat{\sigma}$ greater than 1, with the majority of them having smaller $\hat{\sigma}$ in the AD group compared to the control group (see also Li and Khang, 2023). Finally, the joint distribution of $\hat{\gamma}_1$ and $\hat{\gamma}_2$ appears to be bimodal (Fig. 7(c)), with DS genes distributed towards the upper left and bottom right quadrants. In the AD group, genes tend to have $\hat{\gamma}$ close to 0; thus, we may expect to see more genes with Gaussian-distributed CLR-transformed expression values. For genes in the control group, $\hat{\gamma}$ tends to be relatively more uniformly distributed; thus, distributions of CLR-transformed expression values are generally left or right-skewed.

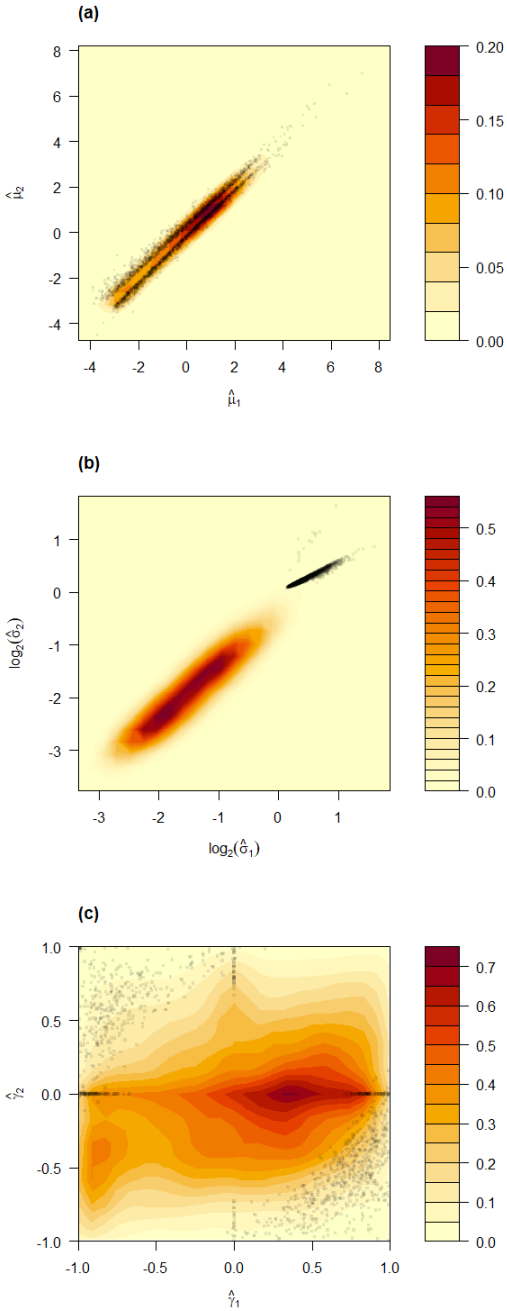


**Fig. 7**: Contour plots of (a) $\hat{\mu}_1$ vs. $\hat{\mu}_2$; (b) $\log_2 \hat{\sigma}_1$ vs. $\log_2 \hat{\sigma}_2$; and (c) $\hat{\gamma}_1$ vs. $\hat{\gamma}_2$, with density color keys on the right. The subscripts 1 and 2 denote the control group and AD group, respectively. Significant genes are indicated as black squares. Range of the estimated parameters: $\hat{\mu}_1 \in (-4.495, 8.457)$; $\log_2 \hat{\sigma}_1 \in (-3.344, 1.867)$; $\hat{\gamma}_1 \in (-0.994, 0.994)$; $\hat{\mu}_2 \in (-4.763, 8.190)$; $\log_2 \hat{\sigma}_2 \in (-3.769, 1.817)$; $\hat{\gamma}_2 \in (-0.993, 0.994)$.

## 3.3 Functional Enrichment Analysis

Table 4 presents concise groupings of 18 GO terms obtained using all five gene lists into seven biological aspects. For details, see Tables S6 to S10 in Supplementary Materials. The results indicate that an enrichment analysis that considers only conventional DE genes detects cell adhesion, extracellular structure organization, blood vessel development, and wound healing as enriched biological processes. Interestingly, consideration of non-DE but DV and/or DS genes reveals involvement of membrane protein localizations in AD pathogenesis, which is well-known, while simultaneously uncovering RNA catabolism. While the four enriched biological processes that are associated with pure DS gene list have FDR of about 0.61, they remain biologically interesting as studies of their involvement in AD have been reported. Importantly, enrichment analysis using the list of DE, DV, and DS genes from SIEVE not only recovers many hierarchically-related processes detected using just DE genes, or just DV and/or DS genes, but also uniquely detects the positive regulation of cell communication process. Thus, combining DE, DV, and DS genes may potentially produce a synergistic effect on enrichment analysis results. To summarize, our findings suggest that incorporating DV and DS analyses alongside DE analysis enables a more comprehensive understanding important biological processes involved in a disease.

A search of the AD literature reveals that many of these aspects in Table 4 have been explored based on subject matter considerations. Here, we briefly examine the literature that report the association of these biological aspects with AD pathophysiology.

(i) *Cellular protein localization*

The endoplasmic reticulum (ER) is a critical organelle where crucial cellular processes such as the biosynthesis, folding, modification and assembly of proteins occur. Under stress, protein biosynthesis demand may exceed protein-folding capacity at the ER lumen, thus resulting in the production of proteins that are partially folded, misfolded, or unfolded, which leads to a condition known as ER stress (Schwarz and Blower, 2016). Amyloid beta (A$\beta$) deposition in the ER is thought to produce ER stress, and excessive levels of such stress can trigger maladaptive unfolded protein response (UPR) activation that produces excessive autophagy and induction of cell death pathways (see

**Table 4**: Functional enrichment in GO biological processes for the DE, DV, and DS genes for the control vs. AD comparison. The size of a gene list refers to the number of gene symbols in the list that are unambiguously mapped to unique Entrez gene IDs. For hierarchical GO terms, the reference GO term is indicated by [0]; parent/child terms have a +/- sign, with the numerical suffix indicating the hierarchy levels above/below the reference GO term. The FDRs of GO terms returned using all gene lists except that of pure DS have order of magnitude less than $10^{-9}$. For the latter gene list, the FDRs of the GO terms are all about 0.61. A: DE, DV, DS genes (4907); B: high confidence DE genes (1981); C: non-DE, DV, DS genes (2497); D: pure DV genes (1578); E: pure DS genes (811).

| Biological aspects | GO ID | GO description | Gene list (size) | | | | |
|---|---|---|---|---|---|---|---|
| | | | A | B | C | D | E |
| Celluler protein localization | GO:0034613 | cellular protein localization [+4] | | | ✓ | | |
| | GO:0072599 | establishment of protein localization to endoplasmic reticulum [+2] | ✓ | | | | |
| | GO:0090150 | establishment of protein localization to membrane [+3] | | | ✓ | ✓ | |
| | GO:0045047 | protein targeting to ER [+1] | | | | ✓ | |
| | GO:0006614 | SRP-dependent cotranslational protein targeting to membrane [0] | | | ✓ | | |
| Cell communicatoin | GO:0022610 | biological adhesion [+2] | ✓ | | | | |
| | GO:0031589 | cell-substrate adhesion [0] | | ✓ | | | |
| | GO:0010647 | positive regulation of cell communication | ✓ | | | | |
| | GO:0070848 | response to growth factor | | ✓ | | | |
| Matrix biology | GO:0030198 | extracellular matrix organization | ✓ | | | | |
| Angiogenesis | GO:0001568 | blood vessel development | ✓ | ✓ | | | |
| | GO:0042060 | wound healing | | ✓ | | | |
| Regulation of gene expression | GO:0006412 | translation | | | ✓ | ✓ | |
| | GO:0006401 | RNA catabolic process | | | | ✓ | |
| Immune response | GO:0019884 | antigen processing and presentation of exogenous antigen | | | | | ✓ |
| | GO:0045619 | regulation of lymphocyte differentiation | | | | | ✓ |
| Oxidative stress | GO:0019363 | pyridine nucleotide biosynthetic process | | | | | ✓ |
| | GO:0019752 | carboxylic acid metabolic process | | | | | ✓ |

Ajoolabady et al, 2022). Costa et al (2018) showed that loss of the signal recognition particle (SRP) results in the mistargeting of genes bound for ER to the mitochondria, which damages mitochondria structure.

(ii) *Matrix biology*

Constituents of the extracellular matrix (ECM) such as chondroitin sulfate proteoglycan and perineuronal net are potentially neuroprotective against tau lesions in AD pathogenesis (Sethi and Zaia, 2017). In particular, deglycosylation of perineuronal net has only very recently been shown to be associated with tauopathy-induced gliosis and neurodegeneration in mouse models (Logsdon et al, 2024).

(iii) *Cell communication*

Cell adhesion molecules (CAMs) have been shown to be linked to A$\beta$ metabolism as well as involved in neuroinflammatory responses and vascular changes (Wennström and Nielsen, 2012). Dysfunctional components of Wnt signalling have been shown to be increase neuronal susceptibility to A$\beta$ toxicity (Godoy et al, 2014). Enhancing Wnt signalling may potentially mitigate synaptic pathology in AD (Palomer et al, 2019).

(iv) *Angiogenesis*

The involvement of angiogenesis in AD pathophysiology was hypothesized as early as the early 2000s (Vagnucci and Li, 2003) and experimental support had accumulated since then (Jefferies et al, 2013). Lehrer and Rheinstein (2016) suggested the possibility that a fraction of AD cases may be caused by defects in brain wound healing. In a mouse model, overexpression of tau proteins induced changes in brain blood vessels that altered blood flow and led to cortical atrophy (Bennett et al, 2018). Major findings implicating vascular channel dysfunction (Taylor et al, 2022) and pericyte-associated dysregulation of blood flow in brains of AD patients were reported in recent years (Yang et al, 2022).

(v) *Regulation of gene expression*

Westmark and Malter (2007) demonstrated that translation of the mRNA of the amyloid precursor protein (APP), from which A$\beta$ peptide is derived from, is inhibited by the fragile X mental retardation protein (FMRP; Li et al, 2001). Additionally, the reduced expression of an FMRP-binding protein known as the cytoplasmic FMRP-interacting protein 2 (CYFIP2) leads to accumulation of phosphorylated tau proteins and A$\beta$ peptides (Ghosh et al, 2020). Meier et al (2016) showed that tau-ribosome interaction in the ER impairs global RNA translation as well as the synthesis of an important synaptic protein that leads to synaptic dysfunction.

(vi) *Immune response*

Neuroinflammation, whereby misfolded and aggregated proteins bind to surface receptors on microglia and astroglia to trigger innate immune response

involving release of inflammatory mediators, is recognized as one of the hallmarks of AD pathophysiology (Akiyama et al, 2000; Cameron and Landreth, 2010; Heneka et al, 2015). The effect of pathological differentiation of T cells during inflammation and the importance of antigen presentation through MHCII-positive microglia in AD pathophysiology were recently reviewed by Schetters et al (2018) and Moro-García et al (2018).

(vii) *Oxidative stress*

The mitochondria, which is critical for cellular energy metabolism and redox homeostasis, is a major target of oxidative damage. Indeed, oxidative stress brought about by reactive oxygen species eventually leads to a runaway spiral towards cellular degeneration in AD (Wang et al, 2014; Swerdlow et al, 2014). Pesini et al (2019) hypothesized that the pathogenesis of late onset AD in a subgroup of AD patients may be related to a decrease in the de novo synthesis of pyrimidine nucleotides, which leads to dysfunction of the oxidative phosphorylation system. Carboxylic acid metabolism is related to lipid metabolism, and lipid dyshomeostasis has been shown to be present during early stages of AD brains (Yin, 2023).

# 4 Discussion

In the Mayo RNA-Seq dataset, almost 50% (2773/5634) of genes that have significant difference in at least one aspect of the gene expression distribution can be detected using DE test. The affirms the value of conducting DE tests as a routine step in RNA-Seq data analysis. Importantly, we note that pure DV, pure DS, or DVS genes that constitute the remaining 50% (2861/5634) cannot be detected using DE tests. As complex diseases are multifaceted, geneomic studies that use DE genes for functional enrichment analysis would be disadvantaged by a narrow perspective. Here, querying the GO database using the gene lists generated by SIEVE substantially expands the landscape of biological processes that are potentially implicated in the pathogenesis of a disease, thus enabling their further exploration and investigation.

SIEVE is possibly limited in its applicability to RNA-Seq studies with small sample sizes, since it generally requires larger sample sizes (50 or more per group) for reliable estimation of the standard deviation and skewness parameters. As the fit of the skew-normal model to CLR-transformed data is

important, the result of testing of goodness-of-fit of the skew-normal model should be reported. Our present empirical analysis was necessarily restricted to the Mayo RNA-Seq dataset, owing to the need for depth of analysis. We encourage more future analyses using SIEVE to ascertain whether the skew-normal distribution is indeed a ubiquitous feature of CLR-transformed RNA-Seq data.

To summarize, SIEVE unlocks the potential of RNA-Seq data by allowing a richer class of genes to be characterized with respect to their mean, standard deviation, and/or skewness characteristics. With the availability of SIEVE, substantial published RNA-Seq datasets of complex diseases may be fruitfully reanalyzed to generate new hypotheses concerning involvement of hitherto unconsidered biological processes. Another possible application is the annotation of genes in gene-gene interaction networks using the gene classes generated by SIEVE, which can potentially facilitate interpretation through integration of domain knowledge.

# 5 Acknowledgements

The present work forms part of the PhD research of HL while at Universiti Malaya. Part of the manuscript was prepared while TFK was on sabbatical leave at the Institute for Chemical Research, Kyoto University. The results published here are in whole or in part based on data obtained from the AD Knowledge Portal. The Mayo RNAseq study data was led by Dr. Nilüfer Ertekin-Taner, Mayo Clinic, Jacksonville, FL as part of the multi-PI U01 AG046139 (MPIs Golde, Ertekin-Taner, Younkin, Price). Samples were provided from the following sources: The Mayo Clinic Brain Bank. Data collection was supported through funding by NIA grants P50 AG016574, R01 AG032990, U01 AG046139, R01 AG018023, U01 AG006576, U01 AG006786, R01 AG025711, R01 AG017216, R01 AG003949, NINDS grant R01 NS080820, CurePSP Foundation, and support from Mayo Foundation. Study data includes samples collected through the Sun Health Research Institute Brain and Body Donation Program of Sun City, Arizona. The Brain and Body Donation Program is supported by the National Institute of Neurological Disorders and Stroke (U24 NS072026 National Brain and Tissue Resource for Parkinsons Disease and Related Disorders), the National Institute on Aging (P30 AG19610 Arizona Alzheimers Disease Core Center), the Arizona Department of Health Services (contract 211002, Arizona Alzheimers Research Center), the Arizona Biomedical Research Commission (contracts 4001, 0011, 05-901 and

1001 to the Arizona Parkinson's Disease Consortium) and the Michael J. Fox Foundation for Parkinsons Research.

# 6 Funding

The present work is partially supported by the Yunnan Key Laboratory of Modern Analytical Mathematics and Applications (No. 202302AN360007) and the Yunnan Cross-integration Innovation Team of Modern Applied Mathematics and Life Sciences (No. 202405AS350003).

# References

Aitchison J (1986) The Statistical Analysis of Compositional Data. Chapman & Hall, London

Ajoolabady A, Lindholm D, Ren J, et al (2022) ER stress and UPR in Alzheimer's disease: Mechanisms, pathogenesis, treatments. Cell Death & Disease 13(8):706

Akiyama H, Barger S, Barnum S, et al (2000) Inflammation and Alzheimer's disease. Neurobiology of Aging 21(3):383–421

Allen M, Carrasquillo MM, Funk C, et al (2016) Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. Scientific Data 3:160,089

Azzalini A (2022) The R package `sn`: The skew-normal and related distributions such as the skew-$t$ and the SUN (version 2.1.0). Università degli Studi di Padova, Italia, URL https://cran.r-project.org/package=sn, home page: http://azzalini.stat.unipd.it/SN/

Azzalini A, Arellano-Valle RB (2013) Maximum penalized likelihood estimation for skew-normal and skew-t distributions. Journal of Statistical Planning and Inference 143(2):419–433

Azzalini A, Capitanio A (2014) The Skew-Normal and Related Families. Cambridge University Press

Bahar R, Hartmann CH, Rodriguez KA, et al (2006) Increased cell-to-cell variation in gene expression in ageing mouse heart. Nature 441:1011–1014

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 29(4):1165–1188

Bennett RE, Robbins AB, Hu M, et al (2018) Tau induces blood vessel abnormalities and angiogenesis-related gene expression in P301L transgenic mice and human Alzheimer's disease. Proceedings of the National Academy of Sciences 115(6):E1289–E1298

Cameron B, Landreth GE (2010) Inflammation, microglia, and Alzheimer's disease. Neurobiology of Disease 37(3):503–509

Casellas J, Varona L (2012) Modeling skewness in human transcriptomes. PLoS One 7(6):e38,919

Cheung VG, Conlin LK, Weber TM, et al (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. Nature Genetics 33:422–425

Church BV, Williams HT, Mar JC (2019) Investigating skewness to understand gene expression heterogeneity in large patient cohorts. BMC Bioinformatics 20(Supplement 24):668

Corrada Bravo H, Pihur V, McCall M, et al (2012) Gene expression anti-profiles as a basis for accurate universal cancer signatures. BMC Bioinformatics 13:272

Costa EA, Subramanian K, Nunnari J, et al (2018) Defining the physiological role of SRP in protein-targeting efficiency and specificity. Science 359(6376):689–692

Esnaola M, Puig P, Gonzalez D, et al (2013) A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments. BMC Bioinformatics 14:254

Fernandes AD, Macklaim JM, Linn TG, et al (2013) ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. PLoS One 8(7):e67,019

Fernandes AD, Reid JN, Macklaim JM, et al (2014) Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome 2:15

Frazee AC, Jaffe AE, Langmead B, et al (2015) Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics 31(17):2778–2784

Gauthier M, Agniel D, Thiébaut R, et al (2020) dearseq: a variance component score test for RNA-seq differential analysis that effectively controls the false discovery rate. NAR Genomics and Bioinformatics 2(4):lqaa093

Geiler-Samerotte KA, Bauer CR, Li S, et al (2013) The details in the distributions: why and how to study phenotypic variability. Current Opinions in Biotechnology 24:752–759

Ghosh A, Mizuno K, Tiwari SS, et al (2020) Alzheimer's disease-related dysregulation of mRNA translation causes key pathological features with ageing. Translational Psychiatry 10:192

Godoy JA, Rios JA, Zolezzi JM, et al (2014) Signaling pathway cross talk in Alzheimer's disease. Cell Communication and Signaling 12:23

Gorlov IP, Byun J, Zhao H, et al (2012) Beyond comparing means: the usefulness of analyzing interindividual variation in gene expression for identifying genes associated with cancer development. Journal of Bioinformatics and Computational Biology 10(02):1241,013

Hardin J, Wilson J (2009) A note on oligonucleotide expression values not being normally distributed. Biostatistics 10(3):446–450

Hasegawa Y, Taylor D, Ovchinnikov DA, et al (2015) Variability of gene expression identifies transcriptional regulators of early human embryonic development. PLoS Genetics 11(8):e1005,428

Heneka MT, Carson MJ, El Khoury J, et al (2015) Neuroinflammation in Alzheimer's disease. The Lancet Neurology 14(4):388–405

Ho JW, Stefani M, Dos Remedios CG, et al (2008) Differential variability analysis of gene expression and its application to human diseases. Bioinformatics 24(13):i390–i398

Jefferies WA, Price KA, Biron KE, et al (2013) Adjusting the compass: new insights into the role of angiogenesis in Alzheimer's disease. Alzheimer's Research & Therapy 5:64

de Jong TV, Moshkin YM, Guryev V (2019) Gene expression variability: the other dimension in transcriptome analysis. Physiological Genomics 51(5):145–158

Kelmer Sacramento E, Kirkpatrick JM, Mazzetto M, et al (2020) Reduced proteasome activity in the aging brain results in ribosome stoichiometry loss and aggregation. Molecular Systems Biology 16(6):e9596

Law C, Chen Y, Shi W, et al (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology 15:R29

Leal Valentim F, Mariotti-Ferrandiz E, Klatzmann D, et al (2020) Transimmunom whole blood RNA-seq data from type 1 diabetic patients and healthy volunteers. Unpublished GEO dataset. GEO accesion number: GSE123658.

Lehrer S, Rheinstein PH (2016) A derangement of the brain wound healing process may cause some cases of Alzheimer's disease. Discovery Medicine 22(119):43

Li H, Khang TF (2023) clrDV: A differential variability test for RNA-Seq data based on the skew-normal distribution. PeerJ 11:e16,126

Li J, Tibshirani R (2013) Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Statistical Methods in Medical Research 22(5):519–536

Li Y, Ge X, Peng F, et al (2022) Exaggerated false positives by popular differential expression methods when analyzing human population samples. Genome Biology 23:79

Li Z, Zhang Y, Ku L, et al (2001) The fragile X mental retardation protein inhibits translation via interacting with mRNA. Nucleic Acids Research

29(11):2276–2283

Liao Y, Wang J, Jaehnig EJ, et al (2019) WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Research 47(W1):W199–W205

Logsdon AF, Foresi B, Hu SJ, et al (2024) Perineuronal net deglycosylation associates with tauopathy-induced gliosis and neurodegeneration. Journal of Neurochemistry https://doi.org/10.1111/jnc.16067

Loh WY (2009) Improving the precision of classification trees. Annals of Applied Statistics 3(2):1710–1737

Loh WY (2022) GUIDE (version 40.3). Available at: https://pages.cs.wisc.edu/ loh/guide.html

Love M, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15(12):550

Mar JC (2019) The rise of the distributions: why non-normality is important for understanding the transcriptome and beyond. Biophysical Reviews 11:89–94

Mar JC, Matigian NA, Mackay-Sim A, et al (2011) Variance of gene expression identifies altered network constraints in neurological disease. PLoS Genetics 7(8):e1002,207

Marko NF, Weil RJ (2012) Non-gaussian distributions affect identification of expression patterns, functional annotation, and prospective classification in human cancer genomes. PLoS One 7(10):e46,935

McCarthy DJ, Chen Y, Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research 40(10):4288–4297

Meier S, Bell M, Lyons DN, et al (2016) Pathological tau promotes neuronal damage by impairing ribosomal function and decreasing protein synthesis. Journal of Neuroscience 36(3):1001–1007

Moro-García MA, Mayo JC, Sainz RM, et al (2018) Influence of inflammation in the process of T lymphocyte differentiation: proliferative, metabolic, and oxidative changes. Frontiers in Immunology 9:328,039

Palomer E, Buechler J, Salinas PC (2019) Wnt signaling deregulation in the aging and Alzheimer's brain. Frontiers in Cellular Neuroscience 13:227

Pesini A, Iglesias E, Bayona-Bafaluy MP, et al (2019) Brain pyrimidine nucleotide synthesis and Alzheimer disease. Aging 11(19):8433

Phipson B, Oshlack A (2014) DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. Genome Biology 15:465

Pritchard CC, Hsu L, Delrow J, et al (2001) Project normal: defining normal variance in mouse gene expression. Proceedings of the National Academy of Sciences 98(23):13,266–13,271

Quinn TP, Crowley TM, Richardson MF (2018) Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. BMC Bioinformatics 19:274

R Core Team (2022) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/

Ran D, Daye ZJ (2017) Gene expression variability and the analysis of large-scale RNA-seq studies with the MDSeq. Nucleic Acids Research 45(13):e127

Rasch D, Teuscher F, Guiard V (2007) How robust are tests for two independent samples? Journal of Statistical Planning and Inference 137(8):2706–2720

Rigby RA, Stasinopoulos DM (2005) Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society C: Applied Statistics 54(3):507–554

Roberts AGK, Catchpoole DR, Kennedy PJ (2022) Identification of differentially distributed gene expression and distinct sets of cancer-related genes identified by changes in mean and variability. NAR Genomics and

Bioinformatics 4(1):lqab124

Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11:R25

Robinson MD, Smyth GK (2007) Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23(21):2881–2887

Saurin A (2022) Bioinformatics tools for genomics and transcriptomics analyses: ENSEMBL ID to Gene Symbol Converter. Available at: https://www.biotools.fr/human/ensembl_symbol_converter. Accessed: 31 August 2022

Schetters ST, Gomez-Nicola D, Garcia-Vallejo JJ, et al (2018) Neuroinflammation: microglia and T cells get ready to tango. Frontiers in Immunology 8:1905

Schwarz DS, Blower MD (2016) The endoplasmic reticulum: structure, function and response to cellular signaling. Cellular and Molecular Life Sciences 73:79–94

Sethi MK, Zaia J (2017) Extracellular matrix proteomics in schizophrenia and Alzheimer's disease. Analytical and Bioanalytical Chemistry 409:379–394

Smyth G (2005) Limma: Linear models for microarray data. In: Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, p 397–420

Srivastava S, Chen L (2010) A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. Nucleic Acids Research 38(17):e170

Swerdlow RH, Burns JM, Khan SM (2014) The Alzheimer's disease mitochondrial cascade hypothesis: progress and perspectives. Biochimica et Biophysica Acta 1842(8):1219–1231

Tarazona S, García-Alcalde F, Dopazo J, et al (2011) Differential expression in RNA-seq: a matter of depth. Genome Research 21(12):2213–2223

Tarazona S, Furió-Tarí P, Turrà D, et al (2015) Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic

Acids Research 43(21):e140

Taylor JL, Pritchard HAT, Walsh KR, et al (2022) Functionally linked potassium channel activity in cerebral endothelial and smooth muscle cells is compromised in Alzheimer's disease. Proceedings of the National Academy of Sciences 119(26):e2204581,119

Thomas R, de la Torre L, Chang X, et al (2010) Validation and characterization of DNA microarray gene expression data distribution and associated moments. BMC Bioinformatics 11:576

Vagnucci AH, Li WW (2003) Alzheimer's disease and angiogenesis. The Lancet 361(9357):605–608

Wang X, Wang W, Li L, et al (2014) Oxidative stress and mitochondrial dysfunction in Alzheimer's disease. Biochimica et Biophysica Acta 1842(8):1240–1247

Wennström M, Nielsen HM (2012) Cell adhesion molecules in Alzheimer's disease. Degenerative Neurological and Neuromuscular Disease 2:65–77

Westmark CJ, Malter JS (2007) FMRP mediates mGluR5-dependent translation of amyloid precursor protein. PLoS Biology 5(3):e52

Xiao Y, Hsiao TH, Suresh U, et al (2014) A novel significance score for gene selection and ranking. Bioinformatics 30(6):801–807

Yang AC, Vest RT, Kern FN, et al (2022) A human brain vascular atlas reveals diverse mediators of Alzheimer's risk. Nature 603(7903):885–892

Yin F (2023) Lipid metabolism and Alzheimer's disease: clinical evidence, mechanistic link and therapeutic promise. The FEBS Journal 290(6):1420–1453

Zhang B, Kirov S, Snoddy J (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Research 33(suppl_2):W741–W748

Zhou Y, Gallins PJ, Etheridge AS, et al (2022) A resource for integrated genomic analysis of the human liver. Scientific Reports 12:15,151