clrDV: A differential variability test for RNA-Seq data based on the skew-normal distribution (#82667)

First submission

Guidance from your Editor

Please submit by **6 Apr 2023** for the benefit of the authors (and your token reward) .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Author notes
Have you read the author notes on the guidance page?

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

If this article is published your review will be made public. You can choose whether to sign your review. If uploading a PDF please remove any identifiable information (if you want to remain anonymous).

Files
Download and review all files from the materials page.

30 Figure file(s)
2 Latex file(s)
4 Table file(s)
1 Raw data file(s)
1 Other file(s)

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

clrDV: A differential variability test for RNA-Seq data based on the skew-normal distribution

Hongxiang Li [1], **Tsung Fei Khang** [Corresp. 1, 2]

[1] Institute of Mathematical Sciences, Universiti Malaya, Kuala Lumpur, Malaysia

[2] Universiti Malaya Centre for Data Analytics, Universiti Malaya, Kuala Lumpur, Malaysia

Corresponding Author: Tsung Fei Khang
Email address: tfkhang@um.edu.my

Background. Pathological conditions may result in certain genes having expression variance that differs markedly from control's. Finding such genes from gene expression data can provide invaluable candidates for therapeutic intervention. Under the dominant paradigm for modeling RNA-Seq gene counts using the negative binomial model, tests of differential variability are challenging to develop, owing to dependence of the variance on the mean.

Methods. Here, we describe clrDV, a statistical method for detecting genes that show differential variability between two populations. We present the skew-normal distribution for modeling gene-wise null distribution of centered log-ratio transformation of compositional RNA-seq data.

Results. Simulation results show that clrDV has false discovery rate and probability of Type II error that are on par with or superior to existing methodologies. In addition, its run time is faster than the closest competitor’s, and remains relatively constant for increasing sample size per group. Analysis of a large neurodegenerative disease RNA-Seq dataset using clrDV successfully recovers multiple gene candidates that have been reported to be associated with Alzheimer’s disease. Additionally, we find that most of the genes with differential variability have smaller relative gene expression variance in the Alzheimer’s disease population compared to the control population.

clrDV: A Differential Variability Test for RNA-Seq Data Based on the Skew-normal Distribution

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

[1]**Institute of Mathematical Sciences, Faculty of Science, Universiti Malaya, 50603 Kuala Lumpur, Malaysia**
[2]**Universiti Malaya Centre for Data Analytics, Universiti Malaya, 50603 Kuala Lumpur, Malaysia**

Corresponding author:
Tsung Fei Khang[2]

Email address: tfkhang@um.edu.my

ABSTRACT

Background. Pathological conditions may result in certain genes having expression variance that differs markedly from control's. Finding such genes from gene expression data can provide invaluable candidates for therapeutic intervention. Under the dominant paradigm for modeling RNA-Seq gene counts using the negative binomial model, tests of differential variability are challenging to develop, owing to dependence of the variance on the mean.
Methods. Here, we describe clrDV, a statistical method for detecting genes that show differential variability between two populations. We present the skew-normal distribution for modeling gene-wise null distribution of centered log-ratio transformation of compositional RNA-seq data.
Results. Simulation results show that clrDV has false discovery rate and probability of Type II error that are on par with or superior to existing methodologies. In addition, its run time is faster than the closest competitor's, and remains relatively constant for increasing sample size per group. Analysis of a large neurodegenerative disease RNA-Seq dataset using clrDV successfully recovers multiple gene candidates that have been reported to be associated with Alzheimer's disease. Additionally, we find that most of the genes with differential variability have smaller relative gene expression variance in the Alzheimer's disease population compared to the control population.

1 INTRODUCTION

1.1 Background

Finding patterns of gene expression variation that are associated with a biological condition of interest is the first step towards elucidating the molecular basis underlying a biological process. Currently, bulk tissue mRNA collected under specific biological conditions through RNA-sequencing (RNA-Seq) technologies remains an important approach for studying gene expression patterns. Typically, genes that show statistically and biologically meaningful difference in mean expression between conditions are of

interest. Indeed, pathological conditions frequently manifest as gene sets with altered mean mRNA expression levels. The identification of these genes is important for understanding how the functions of normal molecular pathways are perturbed (Van den Berge et al., 2019). Hence, detecting genes that are differentially expressed is a routine and main use of RNA-Seq data (Stark et al., 2019). To analyse differential gene expression, a multitude of statistical tests have been developed throughout the years. Methods such as edgeR (Robinson et al., 2010), DESeq2 (Love et al., 2014) and voom (Law et al., 2014) have become established, go-to methods for differential expression (DE) analysis.

To obtain a more complete picture of patterns of gene expression variation, we need to look beyond genes with significantly different mean expression (DE genes) between conditions (Gorlov et al., 2012). Genes that show differential variability (DV genes) are likely to be important as well because many biological phenomena are explained by changes in the variance, rather than the mean, of the distribution of gene expression level (de Jong et al., 2019). For example, genes that show differential variability between undifferentiated and differentiating states have been found to be related to body axis development, neuronal movement, and transcriptional regulation during the neural differentiation process (Ando et al., 2015). In cancer biology, DV genes are useful as biomarkers for predicting tumor progression and prognosis (Dinalankara and Corrada Bravo, 2015), and patient survival (Strbenac et al., 2016). Gorlov et al. (2012) found that genes with larger expression variance in tumors compared to normal cells show stronger association with clinically important features. In network biology, genes with high variability in expression correlate with their positions within the signaling network hierarchy (Komurov and Ram, 2010). Finally, increased gene expression variability is a common outcome of aging (Bahar et al., 2006; Stegeman and Weake, 2017). Standard DE analyses are likely to miss DV gene candidates, since they are not optimized for detecting differences in expression variability.

To date, only a few methods are available for finding DV genes using RNA-Seq data. In contrast, even in 2015, there were at least 20 methods for detecting DE genes (Khang and Lau, 2015). For testing differential variability of genes between two populations using RNA-Seq data, initial methods co-opted techniques from microarray data analysis. DiffVar (Phipson and Oshlack, 2014) is an empirical Bayes method that depends on the limma (Smyth, 2005) framework. Subsequently, negative binomial models became popular. MDSeq (Ran and Daye, 2017) uses the coefficient of dispersion (σ^2/μ) from the negative binomial 1 (NB1) generalized linear model as a measure for variability. The variance from the NB1 model is a function of the mean μ and the dispersion ϕ parameter ($\sigma^2 = \phi\mu$). The parameter μ is treated as a technical component, whereas ϕ is treated as a biological component and interpreted as a parameter for gene expression variability. de Jong et al. (2019) proposed a DV test that uses the generalized additive models for location, scale and shape (GAMLSS; (Rigby and Stasinopoulos, 2005)) framework for quantifying expression variability. GAMLSS is based on the negative binomial 2 (NB2) model, whereby the mean and the variance are related quadratically as $\sigma^2 = \mu + \phi\mu^2$. Recently, Roberts et al. (2022) developed DiffDist, a hierarchical Bayesian model based on the NB2 model. In their work, gene expression variability is measured using the dispersion parameter ϕ, which is treated as a log-normal prior. Subsequently, test of difference in dispersion between two conditions is based on the posterior distribution

simulated using Markov Chain Monte Carlo (MCMC).

In this paper, we wish to propose clrDV - a novel method for detecting DV genes between two conditions in RNA-Seq data that is based on a compositional data analysis framework. The method involves a log-ratio transformation of the raw gene counts, which results in a continuous variable. We show that the skew-normal distribution with centered parameters (Azzalini, 1985) is an appropriate model for the null distribution. Subsequently, we construct a Wald test statistic for testing differential variability. Through simulations, we show how well clr-DV performs compared to existing methods. Finally, we demonstrate the applied value of clrDV by using it to identify biologically meaningful genes in the analysis of a large RNA-seq dataset from a neurodegenerative disease study.

1.2 Motivation

The general idea of conducting a test of differential variability for RNA-Seq data involves testing the equality of variances (equivalently, standard deviations) between two populations. The variance parameter is embedded in some probability distribution that approximates the distribution of gene (more generally, transcript) counts, assuming the null hypothesis is correct. The standard approach models RNA- Seq data as a discrete random variable.

Before modeling can be done, the raw count data need to be normalized to account for variation in the sequencing depth of each sample. Commonly used methods include the trimmed mean of M values (TMM) (Robinson and Oshlack, 2010), the median-of-ratios method (Anders and Huber, 2010; Love et al., 2014), upper-quartile (Bullard et al., 2010), conditional quartile normalization (Hansen et al., 2012), etc. After this, a model that accounts for overdispersion commonly seen in RNA-Seq data (e.g. the NB distribution) is used, but alternative models are possible (Esnaola et al., 2013). Statistical tests of differential variability can then be based on estimators of suitable model parameters for representing expression variability.

In recent years, there has been an increasing call towards adopting a compositional data analysis (CoDA) framework for improving the analysis of RNA-Seq data. Indeed, in the closely related field of microbiome data analysis, CoDA forms the main theoretical framework of data analysis and differential abundance methods (Gloor et al., 2017). Nevertheless, the diffusion of CoDA approach into RNA-Seq data analysis is slow, possibly because established protocols for routine analyses such as differential expression analysis (e.g. DESEq2, edgeR) are all based on discrete count models such as the NB model. Quinn et al. (2018b) argued that next-generation sequencing abundance data should be viewed as inherently compositional because only a portion of genes may be sampled by sequencers, and cells are likely to be constrained in their capacity for mRNA production. Furthermore, Quinn et al. (2018a) showed the feasibility of applying ALDEx2 (Fernandes et al., 2014), a tool developed for differential abundance analysis in micriobiome studies under a CoDA framework, to differential expression analysis using RNA-Seq data. Encouragingly, they reported that ALDEx2 shows superior performance with respect to precision and recall when compared against edgeR and DESeq2. By removing the need to rely on assumptions that justify normalization protocols in standard count-based approaches, log-ratio based transformations of RNA-Seq data in compositional form is potentially more attractive and effective for differential expression

analyses (Quinn et al., 2019). More recently, McGee et al. (2019) developed absSimSeq - a novel simulation protocol for generating realistic RNA- Seq data using a compositional data framework.

The key step in processing compositional data involves log-ratio transformation, for which several variants are available. The simplest is the centered log-ratio (CLR) transformation, first proposed by Aitchison (1986). After CLR- transformation, the simplex space of the compositional data is transformed into the Euclidean space. It is then convenient to view CLR-transformed values as realizations of a continuous random variable. To be concrete, let X_{gi} be the read count for gene g and sample i, where $g = 1,2,\ldots,G$ and $i = 1,2,\ldots,n$. For a G-component composition $\{x_{1i},x_{2i},\ldots,x_{Gi}\}$, the CLR- transformation of X_{gi} is given by

$$\mathrm{CLR}(X_{gi}) = \log\left\{\frac{x_{gi}}{(\prod_{g'} x_{g'i})^{1/G}}\right\} = \log(x_{gi}) - \frac{1}{G}\sum_{g'=1}^{G}\log(x_{g'i}),$$

for $g' = 1,2,\ldots,G$. We call $\mathrm{CLR}(X_{gi})$ the relative gene expression, or CLR-transformed count, of gene g and sample i. A pseudo-value 0.5 is added if $x_{gi} = 0$ for any i. Thus, the main challenge for using CLR-transformed data to develop a test for differential variability is modeling them using a tractable probability distribution for which estimation of the variance parameter is practical.

2 MATERIALS AND METHODS

2.1 The skew-normal model for CLR-transformed data

We show that the null distribution of CLR-transformed count data approximately follows the skew-normal distribution (Azzalini, 1985; Azzalini and Capitanio, 2014) (see Supplementary Material S1). Denote the relative gene expression from gene g in sample i by Y_{gi}. Thus, Y_{gi} has a skew-normal distribution with centered parameters (CP), that is, $Y_{gi} \sim \mathrm{SN_C}(\mu_g,\sigma_g,\gamma_g)$, where μ_g is the mean, σ_g is the standard deviation, and γ_g is the skewness parameter, $g = 1,2,\ldots,G$ and $i = 1,2,\ldots,n$. 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 special case of $\gamma_g = 0$ results in a normal distribution with mean μ_g and variance σ_g^2. The probability density function of a skew-normal distribution with direct parameters (DP) is given by

$$f(y_{gi};\xi_g,\omega_g,\alpha_g) = \frac{2}{\omega_g}\phi\Big(\frac{y_{gi}-\xi_g}{\omega_g}\Big)\Phi\Big(\alpha_g\frac{y_{gi}-\xi_g}{\omega_g}\Big),$$

with location parameter $\xi_g \in \mathbb{R}$, scale parameter $\omega_g \in \mathbb{R}^+$, and skewness parameter $\alpha_g \in \mathbb{R}$; $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and the cumulative distribution function of the standard normal distribution, respectively. The skew-normal distribution with CP is derived from the DP form via the mapping (Azzalini and Capitanio, 2014)

$$\mu_g = \xi_g + b\omega_g\delta_g,\ \sigma_g = \omega_g\sqrt{1-b^2\delta_g^2},\ \gamma_g = \frac{4-\pi}{2}\frac{b^3\alpha_g^3}{\{1+(1-b^2)\alpha_g^2\}^{3/2}}; \qquad (1)$$

and the inverse mapping is provided by

$$\xi_g = \mu_g - b\omega_g\delta_g, \quad \omega_g = \frac{\sigma_g}{\sqrt{1-b^2\sigma_g^2}}, \quad \alpha_g = \frac{R}{\sqrt{b^2-(1-b^2)R^2}}, \tag{2}$$

where $b = \sqrt{2/\pi}$, $\delta_g = \alpha_g/\sqrt{1+\alpha_g^2}$, and $R = \sqrt[3]{2\gamma_g/(4-\pi)}$.

For a single sample, the log-likelihood function for $\boldsymbol{\theta}_g^{(D)} = (\xi_g, \omega_g, \alpha_g)^T$ is given by

$$\ell_1 = \log L(\boldsymbol{\theta}_g^{(D)}; y_{gi}) = c - \log\omega_g - \frac{(y_{gi}-\xi_g)^2}{2\omega_g^2} + \zeta_0\Big(\alpha_g\frac{y_{gi}-\xi_g}{\omega_g}\Big),$$

where c is a constant and $\zeta_0(\cdot) = \log\big\{2\Phi(\cdot)\big\}$. Taking $z_{gi} = (y_{gi}-\xi_g)/\omega_g$, we obtain the partial derivatives of ℓ_1:

$$\frac{\partial\ell_1}{\partial\xi_g} = \frac{z_{gi}}{\omega_g} - \frac{\alpha_g}{\omega_g}\zeta_1(\alpha_g z_{gi}), \ \frac{\partial\ell_1}{\partial\omega_g} = -\frac{1}{\omega_g} + \frac{z_{gi}^2}{\omega_g} - \frac{\alpha_g}{\omega_g}\zeta_1(\alpha_g z_{gi})z_{gi}, \ \frac{\partial\ell_1}{\partial\alpha_g} = \zeta_1(\alpha_g z_{gi})z_{gi};$$

thus the likelihood equations for a sample of size n are given by

$$\sum_{i=1}^n z_{gi} - \alpha_g\sum_{i=1}^n \zeta_1(\alpha_g z_{gi}) = 0, \ \sum_{i=1}^n {z_{gi}}^2 - \alpha_g\sum_{i=1}^n z_{gi}\zeta_1(\alpha_g z_{gi}) = n, \ \sum_{i=1}^n z_{gi}\zeta_1(\alpha_g z_{gi}) = 0, \tag{3}$$

where $\zeta_1(\cdot) = \phi(\cdot)/\Phi(\cdot)$. Numerical methods are necessary to solve these equations. Azzalini and Capitanio (2014) suggested that a sample size up to about 50 may be necessary for the skew- normal distribution. To initialize the search, method of moments (MM) estimates are chosen as starting points for the CP components in Equation (1). The MM estimators for the centered parameters are given by

$$\tilde{\mu}_g = \bar{Y}_g, \quad \tilde{\sigma}_g = S_g, \quad \tilde{\gamma}_g = \frac{M_{g,3}}{S_g^3}, \tag{4}$$

respectively, where $\bar{Y}_g$ is the sample mean, S_g is the sample standard deviation, and $M_{g,3}$ is the sample third central moment. By estimating the CP components in Equation (1) using Equation (4), and then converting them to DP components using Equation (2), we obtain the MM estimators of the DP components: $\bar{\xi}_g, \bar{\omega}_g$ and $\bar{\alpha}_g$. Subsequently, a search of the DP space where Equation (3) holds is done. Once $\hat{\boldsymbol{\theta}}_g^{(D)} = (\hat{\xi}_g, \hat{\omega}_g, \hat{\alpha}_g)$ is obtained, it is mapped to Equation (1) to get $\hat{\boldsymbol{\theta}}_g^{(C)} = (\hat{\mu}_g, \hat{\sigma}_g, \hat{\gamma}_g)$, the maximum likelihood estimators of the centered parameters.

Under regular maximum likelihood estimation, certain data values can produce a divergent $\hat{\alpha}_g$. To overcome this problem, Azzalini and Arellano-Valle (2013) proposed a maximum penalized likelihood estimation ("Qpenalty") approach. A non-negative penalty term Q that penalizes the divergence of the skewness parameter α_g is formulated as $Q = c_1\log(1+c_2\alpha_g^2)$, where $c_1 \approx 0.87591$ and $c_2 \approx 0.85625$ (Azzalini and Arellano-Valle, 2013; Azzalini and Capitanio, 2014). Then, the maximum penalized likelihood for $\boldsymbol{\theta}_g^{(D)}$ is the penalized log-likelihood

$$\ell_p(\boldsymbol{\theta}_g^{(D)}) = \ell(\boldsymbol{\theta}_g^{(D)}; \boldsymbol{y}_g) - Q, \tag{5}$$

where $\boldsymbol{y}_g = (y_{g1}, y_{g2}, \ldots, y_{gn})$, $\ell(\boldsymbol{\theta}_g^{(D)}; \boldsymbol{y}_g)$ is the log- likelihood function with respect to the parameter vector $\boldsymbol{\theta}_g^{(D)}$:

$$\ell(\boldsymbol{\theta}_g^{(D)}; \boldsymbol{y}_g) = \text{constant} - n \log \omega_g - \sum_{i=1}^{n} \frac{(y_{gi} - \xi_g)^2}{2\omega_g^2} + \sum_{i=1}^{n} \zeta_0\big(\alpha_g \frac{y_{gi} - \xi_g}{\omega_g}\big).$$

The maximum penalized likelihood estimator (MPLE), $\tilde{\boldsymbol{\theta}}_g^{(D)}$, is a finite point that maximizes $\ell_p(\boldsymbol{\theta}_g^{(D)})$. The standard errors of $\tilde{\boldsymbol{\theta}}_g^{(D)}$ can be approximated from the corresponding penalized information matrix as $\text{Var}(\tilde{\boldsymbol{\theta}}_g^{(D)}) \approx -\ell_p''(\tilde{\boldsymbol{\theta}}_g^{(D)})^{-1}$.

The "MPpenalty" approach (Azzalini and Capitanio, 2014) defines the penalty function Q in Equation (5) as $-\log \pi_m(\alpha_g)$, where π_m is a prior distribution for the skewness parameter α_g. The matching prior (Cabras et al., 2012) for α_g, allowing for the presence of $\boldsymbol{\psi} = (\xi_g, \omega_g)$, is given by

$$\pi_m(\alpha_g) \propto \big(I_{\alpha_g\alpha_g}(\hat{\boldsymbol{\psi}}, \alpha_g) - I_{\alpha_g\boldsymbol{\psi}}(\hat{\boldsymbol{\psi}}, \alpha_g) I_{\boldsymbol{\psi}\boldsymbol{\psi}}(\hat{\boldsymbol{\psi}}, \alpha_g)^{-1} I_{\boldsymbol{\psi}\alpha_g}(\hat{\boldsymbol{\psi}}, \alpha_g)\big)^{1/2},$$

where the terms involved are specific blocks of the Fisher information matrix $\boldsymbol{I}$ of $\boldsymbol{\theta}_g^{(D)}$ (see Supplementary Material S1 for details). Since $\pi_m(0) = 0$, the matching prior penalty effectively penalizes $\alpha_g = 0$ with $Q = \infty$.

To perform parameter estimation and carry out related numerical tasks involving the skew-normal distribution, we used the `sn` (Azzalini, 2022) R package. Regular maximum likelihood estimation of parameters of the skew-normal model was first done using the function `selm()`. If NA values were returned, we used the maximum penalized likelihood estimation as implemented using the `Qpenalty` option. If NA values persisted, the `MPpenalty` option was used.

For RNA-Seq experiments comparing two populations, testing for differential variability is equivalent to testing the equality of the standard deviation of relative gene expressions in two populations, that is, $\sigma_{g,1} = \sigma_{g,2}$. For this purpose, we can use the Wald statistic

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

for $g = 1, 2, \ldots, G$, where $\hat{\sigma}_{g,j}, j = 1, 2$ are the maximum likelihood estimators of the standard deviation of the skew-normal distribution with centered parameters for population 1 and population 2, and $\text{Var}(\hat{\sigma}_{g,j}), j = 1, 2$ are the corresponding diagonal elements of the estimated Fisher information matrix of centered parameters $\boldsymbol{\theta}_g^{(C)} = (\mu_g, \sigma_g, \gamma_g)$. To control the false discovery rate (FDR) as a result of conducting multiple independent hypothesis tests across genes, we applied the Benjamini-Yekutieli procedure (Benjamini and Yekutieli, 2001). Note that in the context of samples, FDR is estimated as the sample proportion of false discoveries.

2.2 Data Description

In order to study the performance of clrDV and other existing methods with respect to FDR and probability of Type II error, it is necessary to simulate the null distribution with

realistic parameter values. For this purpose, we used two real RNA-Seq datasets. The first dataset (GEO accesion number: GSE123658) contains whole blood RNA-Seq data from from 39 Type 1 diabetes patients and 43 healthy donors (Leal Valentim et al., 2020), with 16,785 transcripts. The second dataset (GEO acccesion number: GSE150318) contains longitudinal gene expression data from 114 short-lived killfish *Nothobranchius furzeri* measured at 10 weeks and 20 weeks of age (Kelmer Sacramento et al., 2020), with 26,739 transcripts. Hereafter, we call these two datasets the "Valentim dataset" and the "Kelmer dataset".

For empirical assessment, we used the Mayo RNASeq dataset (Allen et al., 2016), which consists of 278 samples and 64,253 transcripts. In this study, RNA was isolated from the temporal cortex of brains of patients with four biological conditions: control ($n = 80$), Alzheimer's disease (AD; $n = 84$), progressive supranuclear palsy (PSP; $n = 84$) and pathologic aging ($n = 30$). We chose to compare the control group against the AD and the PSP group respectively, since the sample sizes in these groups are reasonably large and balanced.

2.3 Simulation study

Only transcripts that satisfy two conditions in each group were used for simulation: (i) average count-per- million (CPM) above 0.5; and (ii) less than 85% of samples have zero count. Then, 2000 of the filtered genes were randomly selected. For each gene, an NB2 model was fitted. We simulated 10% of the genes to be DV genes by multiplying their size parameter $(1/\phi)$ with a random value x, where $x \in (0.25, 0.5) \cup (2, 4)$. Counts were then simulated based on the fitted NB2 model, for six sample sizes $(50, 100, 125, 150, 200, 250)$ using the `polyester` (Frazee et al., 2015) R package. A total of 30 instances were thus simulated. Genes with BY-adjusted p-value < 0.05 were flagged as having differential variability.

The performance of clrDV against MDSeq, diffVar, and GAMLSS (Benjamini-Hochberg (BH) and Benjamini-Yekutieli (BY) variants) was evaluated by considering their FDR and probability of Type II error. Additionally, we also recorded the run time of each method. DiffDist was excluded from the evaluation since it needs to perform MCMC simulations to generate the posterior distribution. As such, it is computationally expensive to implement and difficult to justify as a choice for routine application. Indeed, running DiffDist on an RNA-Seq dataset with 43 samples per group and 23,416 transcripts, Roberts et al. (2022) reported that DiffDist took about three hours to complete, compared to 12 minutes for GAMLSS and 4 minutes for MDSeq .

2.4 Empirical assessment

We applied clrDV to the Mayo RNA-Seq dataset to assess its capacity for detecting DV genes that are contextually meaningful. Analysis using MDSeq and GAMLSS (BH and and BY variants) were also done. We dropped diffVar because this method performed poorly during the simulation stage. Volcano plots were used to inspect the biological effect size and statistical significance of all genes tested. Venn diagrams were used to identify sets of genes that are identically recovered by all three methods, by combinations of two methods, or uniquely recovered by a single method. Violin plots of selected DV genes were made to verify computational results.

2.5 Tools and computing environment

Computational tasks were done in a computer with a 1.80 GHz i5-8265U CPU and an 8GB RAM processor. R (R Core Team, 2022) (version 4.2.1) operating in Windows 10 was used. The complete list of R packages used is given in the Supplementary Material S2. ENSEMBL gene ID to gene symbol conversion was done using the application programming interface of the BioTools.fr website (Saurin, 2022).

3 RESULTS

3.1 Simulation study

We found that the skew-normal distribution with centered parameters fit the CLR-transformed count data well. Two examples are given in Figure 1. Additional examples can be readily inspected using the R codes provided. Figure 2 shows the scatter plots of probability of Type II error against FDR for analysis of the simulated Valentim dataset, for each of the six sample size per group scenarios. diffVar is clearly uniformly inferior to all other methods (mean probability of Type II error > 0.05 and FDR > 0.17, for all sample sizes).

For sample size of 50, all methods show relatively larger mean probability of Type II error (> 0.2); additionally, diffVar and GAMLSS-BH show high mean FDR (> 0.05). Against MDSeq, clrDV is uniformly superior with respect to mean FDR and mean probability of Type II error; against GAMLSS-BH, clrDV has uniformly superior mean FDR; against GAMLSS-BY, clrDV gives approximately similar mean FDR and mean probability of Type II error. When sample size is very large (250), clrDV, MDSeq and GAMLSS-BY give similar performance. With respect to computing speed, clrDV is substantially faster than GAMLSS (both BH and BY variants) as sample size increases (Supplementary Material Table S1). For the analysis of simulated data from the Kelmer dataset, we find clrDV to have comparable mean FDR and mean probability of Type II error (Figure 3) as MDSeq and GAMLSS-BY. However, clrDV computing time remains almost constant across the six sample sizes, whereas MDSeq and GAMLSS have computing times that increase with sample size (Supplementary Material Table S2). diffVar and GAMLSS-BH are inferior in controlling FDR across all six sample sizes.

3.2 Analysis of the Mayo RNA-Seq dataset

After filtering, sample sizes of the control, the AD and the PSP groups were 78, 82, and 84, respectively. For the AD and the control group comparison, a total of 18,664 transcripts were left; for the PSP and control comparison, 18,636 transcripts were left. For MDSeq and GAMLSS, we normalized the raw counts using TMM normalization.

3.2.1 Detection of genes with differential variability

Applying the procedure described in Section 2, we estimated the standard deviation of the CLR- transformed data, computed the Wald statistic and subsequently the BY-adjusted p-value for each tested gene. For the control vs. AD comparison, we detected a set of 4754 DV genes (see Supplementary Table S3 for complete list); for the control vs. PSP comparison, 4859 DV genes were detected (see Supplementary Table S4 for complete list). For the majority of DV genes, the estimated standard deviation in the control group is larger than the one in the treatment group (Figure 4). This observation

suggests that genes with decreased expression variability among patients with AD are far more common than those that show increased variability.

Figure 5 shows the number of significant DV genes identified by clrDV, MDSeq, GAMLSS-BH and GAMLSS- BY for the control vs. AD comparison (see Supplementary Table S5 for complete list). GAMLSS-BH detected the most DV genes (9926), followed by MDSeq (6924), and clrDV (4754). The high confidence gene set, defined as the intersection of DV genes from each method, contains genes with estimated $\log_2(\text{SD ratio})$ that is relatively large (> 0.5). About 99.8% $(4743/4754)$ of DV genes detected by clrDV are also identified by MDSeq or GAMLSS-BH; 92.0% $(4374/4754)$ are detected by both MDSeq and GAMLSS-BH; about 0.2% $(11/4754)$ are uniquely identified by clrDV. GAMLSS-BH identified very large numbers of DV genes in this dataset, but the majority of these are probably false positives, given its relatively poorer control of FDR as shown in the results of the simulation studies. Moreover, these DV genes have estimated $\log_2(\text{SD ratio})$ with relatively small magnitude, as indicated by the violin plots (Figure 5(c)).

Using GAMLSS-BY, only 6079 DV genes were detected, compared to 9926 using GAMLSS-BH. Thus, GAMLSS-BY primarily helps improve FDR by reducing the number of DV genes called. Between 61.7% $(4271/6924)$ and 89.8% $(4271/4754)$ of the DV genes detected by one method are detected by all three. About 97.0% $(4613/4754)$ of DV genes detected by clrDV are identified by one of other two methods, and 3.0% $(141/4754)$ of DV genes detected by clrDV are unique.

The result of the control vs. PSP comparison is similar (Figure 6; Supplementary Material Table S6). GAMLSS-BH also detected the most number of DV genes (9707), followed by MDSeq (6894), and clrDV (4859). Up to 99.4% $(4831/4859)$ of DV genes identified by clrDV are detected by MDSeq or GAMLSS-BH; about 89.1% $(4329/4859)$ are detected by both MDSeq and GAMLSS-BH; about 0.6% $(28/4859)$ are uniquely identified by clrDV. Using GAMLSS-BY, only 6024 DV genes were flagged. Approximately 95.9% $(4658/4859)$ of DV genes identified by clrDV are also identified by MDSeq or GAMLSS-BY; about 86.1% $(4186/4859)$ are detected by both MDSeq and GAMLSS-BY; about 4.1% $(201/4859)$ are uniquely detected by clrDV.

The violin plots (Figure 5 and Figure 6) suggest that the DV genes uniquely called by clrDV may be more likely to true positives, given that the magnitude of $\log_2(\text{SD ratio})$ is generally larger than 0.5. For those uniquely called by GAMLSS or MDSeq, the order of magnitude is generally below 0.5. With respect to run time, for the control vs. AD comparison, clrDV took about 7.5 minutes, compared to 6 minutes for MDSeq, and 13 minutes for GAMLSS; for the control vs. PSP comparison, clrDV took about 7 minutes, while MDSeq used 6 minutes, and GAMLSS used 15 minutes.

3.2.2 Biological significance of detected differential variability genes

In the control vs. AD comparison, four of the DV genes that have the largest estimated SD ratio above 1 are LTBP2, SLPI, C2orf40, and SLC47A1 (Figure 7). All four genes have been reported to be associated with Alzheimer's disease in the literature. The latent transforming growth factor (TGF)- beta binding proteins (LTBP) are important components of the extracellular matrix (Robertson et al., 2015). They interact with fibrillin microfibrils, and are known to be mediators of TGF-β functions (Rifkin et al., 2018), dysfunctions of which have been implicated in Alzheimer's disease (Das and

Golde, 2006). Then, the secretory leukocyte protease inhibitor protein (SLPI) is known to regulate the penetrance of frontotemporal lobar degeneration (FTLD) in patients who have mutations in the progranulin gene (Ghidoni et al., 2014). Loss of progranulin function has been found to enhance microglial neuroinflammation, which is implicated in Alzheimer's disease (Mendsaikhan et al., 2019). Podvin et al. (2016) found that C2orf40 is a neuroimmune factor in Alzheimer's disease. The SLC47A1 (solute carrier family 47 member 1) protein is expressed in both the kidney and the brain, and recent research has suggested a linkage between kidney diseases and Alzheimer's disease (Shi et al., 2018; Kelly and Rothwell, 2022).

We detected 74 genes from the SLC family in the high confidence DV gene set, including four members of the SLC39 family. Lang et al. (2012) demonstrated the modulating effect of dZip1, the ortholog of human SLC39 family transporter, on zinc ion uptake using a *Drosophila* model. Zinc is known to induce amyloid beta formation (Bush et al., 1994). Inhibition of dZip1 produces substantial reduction of amyloid beta peptide 42 (Aβ42) fibril deposits and less neurodegeneration in Aβ42-transgenic flies.

Two of the DV genes with estimated SD ratio substantially smaller than 1 are PELP1 and GP1BB (Figure 7). PELP1 mediates E2 inhibition of GSK3β, a neurodegenerative kinase signaling pathway in the brain (Thakkar et al., 2018). GSK3β is implicated in Alzheimer's disease as a key mediator of cell death (Llorens-Martin et al., 2014). The GP1BB gene produces glycoprotein 1b-beta (GPIbβ), a subunit of the GPIb-IX-V protein complex on the surface of platelet cells. Amyloid beta peptides are known to be actively released by platelets (Bush et al., 1990; Casoli et al., 2007). Visconte et al. (2020) recently reported that recruitment of GPIb-IX-V is required for fibrillar amyloid Aβ40 and Aβ42 to induce platelet aggregation. The study of the role of platelets and the pathogenesis of Alzheimer's disease is an active topic (Catricala et al., 2012).

We note that approximately half of genes in the high confidence gene set from the control vs. AD comparison (4271 genes) are also found in the high confidence gene sets from the control vs. PSP comparison (4186 genes). Altogether, 2149 DV genes are common to both comparisons. This observation is consistent with recent findings that transcriptomic changes are in AD and PSP relative to control are strongly correlated (Wang et al., 2022).

4 DISCUSSION

Our present work demonstrates that when analyzing gene expression data using the CoDA framework, the skew-normal distribution provides a natural way to model CLR-transformed data. The skew-normal distribution is a tractable model with mature computational support through the `sn` R package. A test of differential variability can therefore be based directly on the standard deviation parameter of the skew-normal distribution. Moreover, a test of differential expression that is based on the mean parameter can be derived as well. With these tests, it becomes possible to develop methods for detecting three classes of genes in two-population comparisons: (i) equal variance, different mean; (ii) equal mean, different variance; (iii) different mean, different variance. Although clrDV cannot differentiate genes of the second and the third type, inspection of violin plots should be useful for ascertaining whether the DV genes also appear to differ significantly in the mean of their relative expression level.

We observed that in the comparisons between control vs. AD and control vs. PSP, a majority of the DV genes identified by clrDV (between 86.1% and 92.0%) were already included in the high-confidence gene set, where the estimated $\log 2$(SD ratio) has a relatively large magnitude. Thus, it seems that clrDV alone should be able to recover most of the DV genes of interest.

The relative poorer performance of MDSeq and GAMLSS could be caused by the choice of normalization. It is known that incorrect normalization leads to inflated FDR in differential expression analyses (Evans et al., 2018), yet the assumptions that justify a normalization method are usually not testable. Since existing normalization methods have been developed for the purpose of finding differentially expressed genes, the assumptions that justify their use are probably suboptimal for differential variability tests. Consequently, the performance of existing count-based approaches for DV test is likely sensitive to the choice of normalization method. However, it is beyond the scope of the present work to optimize the choice of normalization step for these count-based methods.

On the aspect of practical application, we note that the R codes provided by de Jong et al. (2019) for GAMLSS are not sufficiently generic and require further user modifications to be suitable for routine use as a DV test. In addition, GAMLSS uses BH rather than BY as the default setting for multiple comparisons adjustment. For MDSeq, we found that it may occasionally encounter difficulties in estimating model parameters. In our analysis of the Mayo RNA-Seq dataset, we observed that 45 genes returned NA parameter estimates in the control vs. AD and the control vs. PSP comparisons. Given these findings, we believe clrDV is currently the most practical and effective method for researchers who wish to conduct differential variability test using RNA- Seq data.

5 CONCLUSIONS

Variability of gene expression at aberrant levels is one of the hallmarks of disrupted or dysregulated biological processes. Hence, detection of genes with differential variability should accompany routine differential expression analysis to expand the pool of potential therapeutic intervention targets. clrDV offers a novel approach for identifying DV genes in RNA-seq data. By modeling the null distribution of centered log- ratio transformed RNA-Seq data using a skew-normal distribution, clrDV can detect genes with expression variance that differs significantly between two populations. Simulation results demonstrate that clrDV has a comparable or superior false discovery rate and probability of Type II error compared to existing methods, while also having a faster run time for larger sample sizes per group. Applying clrDV to the Mayo RNA-seq dataset, we identified several genes associated with Alzheimer's disease, many of which had smaller relative gene expression variance in the Alzheimer's disease population compared to the control population. Crucially, the compositional data analysis framework used in this work can be extended to create statistical tests for differential expression and differential skewness using RNA-seq data. Results from this extension will be reported elsewhere.

CODE AVAILABILITY

We have created an R package called `clrDV` to perform the differential variability test described here. The R package and codes for reproducing the analyses in this study are

available at https://github.com/Divo-Lee/clrDV.

DATA AVAILABILITY

The present work did not generate no new datasets. We used datasets published by other researchers described in Section 3.2. The RNA-Seq datasets with GEO accession numbers GSE123658 and GSE150318 are freely available in the NCBI Gene Omnibus Expression database. We obtained permission from AD Knowledge Portal (accessible at https://adknowledgeportal.org) to access and use the Mayo RNASeq dataset for research purpose.

ACKNOWLEDGMENTS

We wish to thank Dr. C.Y. Ung for helpful discussions. We are grateful to Professor A. Azzalini for developing the theoretical foundations as well as computational tools for the skew normal distribution over many years, without which the present work would not be possible. 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.

Author contributions

Conceptualization: Tsung Fei Khang; Methodology: all authors; Formal analysis and investigation: all authors; Writing - original draft preparation: all authors; Writing - review and editing: all authors; Supervision: Tsung Fei Khang.

Financial disclosure

None reported.

Conflict of interest

The authors declare no potential conflict of interests.

REFERENCES

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

Allen, M., Carrasquillo, M. M., Funk, C., Heavner, B. D., Zou, F., Younkin, C. S., Burgess, J. D., Chai, H. S., Crook, J., Eddy, J. A., et al. (2016). Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Scientific Data*, 3(89).

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11:R106.

Ando, T., Kato, R., and Honda, H. (2015). Differential variability and correlation of gene expression identifies key genes involved in neuronal differentiation. *BMC Systems Biology*, 9:Article no. 82.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12(2):171–178.

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. Available at: `http://azzalini.stat.unipd.it/SN/`.

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

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

Bahar, R., Hartmann, C. H., Rodriguez, K. A., Denny, A. D., Busuttil, R. A., Dollé, M. E. T., Calder, R. B., Chisholm, G. B., Pollock, B. H., and Klein, C. A. (2006). Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature*, 441:1011–1014.

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

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(96).

Bush, A. I., Martins, R. N., Rumble, B., Moir, R., Fuller, S., Milward, E., Currie, J., Ames, D., Weidemann, A., Fischer, P., Multhaup, G., Beyreuther, K., and Masters, C. L. (1990). The amyloid precursor protein of Alzheimer's disease is released by human platelets. *Journal of Biological Chemistry*, 265(26):15977–15983.

Bush, A. I., Pettingell, W. H., Multhaup, G., d'Paradis, M., Vonsattel, J. P., Gusella, J. F., Beyreuther, K., Masters, C. L., and Tanzi, R. E. (1994). Rapid induction of Alzheimer A beta amyloid formation by zinc. *Science*, 265(5177):1464–1467.

Cabras, S., Racugno, W., Castellanos, M. E., and Ventura, L. (2012). A matching prior for the shape parameter of the skew-normal distribution. *Scandinavian Journal of Statistics*, 39(2):236–247.

Casoli, T., Di Stefano, G., Giorgetti, B., Grossi, Y., Balietti, M., Fattoretti, P., and Bertoni-Freddari, C. (2007). Release of beta-amyloid from high-density platelets: implications for Alzheimer's disease pathology. *Annals of the New York Academy of Sciences*, 1096:170–178.

Catricala, S., Torti, M., and Ricevuti, G. (2012). Alzheimer disease and platelets: how's that relevant. *Immunity and Ageing*, 9(1):20.

Das, P. and Golde, T. (2006). Dysfunction of TGF-β signaling in Alzheimer's disease. *Journal of Clinical Investigations*, 116(11):2855–2857.

de Jong, T. V., Moshkin, Y. M., and Guryev, V. (2019). Gene expression variability: the other dimension in transcriptome analysis. *Physiological Genomics*, 51(5):145–158.

Dinalankara, W. and Corrada Bravo, H. (2015). Gene expression signatures based on variability can robustly predict tumor progression and prognosis. *Cancer Informatics*, 2015:71–81.

Esnaola, M., Puig, P., Gonzalez, D., Castelo, R., and Gonzalez, J. R. (2013). A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments. *BMC Bioinformatics*, 14(1):1–22.

Evans, C., Hardin, J., and Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, 19(5):776–792.

Fernandes, A. D., Reid, J. N. S., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., and Gloor, G. B. (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(1):1–13.

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

Ghidoni, R., Flocco, R., Paterlini, A., Glionna, M., Caruana, L., Tonoli, E., Binetti, G., and Benussi, L. (2014). Secretory leukocyte protease inhibitor protein regulates the penetrance of frontotemporal lobar degeneration in progranulin mutation carriers. *Journal of Alzheimer's Disease*, 38:533–539.

Gloor, G., Macklaim, J., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology*, 8:2224.

Gorlov, I. P., Byun, J., Zhao, H., Logothetis, C. J., and Gorlova, O. Y. (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*, 10:1241013.

Hansen, K. D., Irizarry, R. A., and Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, 13(2):204–216.

Kelly, D. M. and Rothwell, P. M. (2022). Disentangling the relationship between chronic kidney disease and cognitive disorders. *Frontiers in Neurology*, 13:Article 830064.

Kelmer Sacramento, E., Kirkpatrick, J., Mazzetto, M., Baumgart, M., Bartolome, A., Di Sanzo, S., Caterino, C., Sanguanini, M., Papaevgeniou, N., Lefaki, M., et al. (2020). Reduced proteasome activity in the aging brain results in ribosome stoichiometry loss and aggregation. *Molecular Systems Biology*, 16(6):e9596.

Khang, T. F. and Lau, C. Y. (2015). Getting the most out of RNA-seq data analysis. *PeerJ*, 3:e1360.

Komurov, K. and Ram, P. T. (2010). Patterns of human gene expression variance show strong associations with signaling network hierarchy. *BMC Systems Biology*, 4:154.

Lang, M., Wang, L., Fan, Q., Xiao, G., Wang, X., Zhong, Y., and Zhou, B. (2012). Genetic inhibition of solute-linked carrier 39 family transporter 1 ameliorates Aβ pathol-

ogy in a *Drosophila* model of Alzheimer's disease. *PLoS Genetics*, 8(4):e1002683.
Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29.
Leal Valentim, F., Mariotti-Ferrandiz, E., Klatzmann, D., Six, A., and Konza, O. (2020). Transimmunom whole blood RNA-seq data from type 1 diabetic patients and healthy volunteers. Unpublished GEO dataset. GEO accesion number: GSE123658.
Llorens-Martin, M., Jurado, J., Hernandez, F., and Avila, J. (2014). GSK-3β, a pivotal kinase in Alzheimer disease. *Frontiers in Molecular Neuroscience*, 7:46.
Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550.
McGee, W. A., Pimentel, H., Pachter, L., and Wu, J. Y. (2019). Compositional data analysis is necessary for simulating and analyzing RNA-Seq data. *bioRxiv*, 564955:doi: https://doi.org/10.1101/564955.
Mendsaikhan, A., Tooyama, I., and Walker, D. G. (2019). Microglial progranulin: Involvement in Alzheimer's disease and neurodegenerative diseases. *Cells*, 8(3):230.
Phipson, B. and Oshlack, A. (2014). DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biology*, 15(9):465.
Podvin, S., Miller, M. C., Rossi, R., Chukwueke, J., Donahue, J. E., Johanson, C. E., Baird, A., and Stopa, E. G. (2016). The orphan C2orf40 gene is a neuroimmune factor in Alzheimer's disease. *JSM Alzheimer's Disease and Related Dementia*, 3(1):1020.
Quinn, T. P., Crowley, T. M., and Richardson, M. F. (2018a). Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. *BMC Bioinformatics*, 19(274).
Quinn, T. P., Erb, I., Gloor, G., Notredame, C., Richardson, M. F., and Crowley, T. M. (2019). A field guide for the compositional analysis of any-omics data. *GigaScience*, 8(9):giz107.
Quinn, T. P., Erb, I., Richardson, M. F., and Crowley, T. M. (2018b). Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, 34(16):2870–2878.
R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Ran, D. and Daye, Z. J. (2017). Gene expression variability and the analysis of large-scale RNA-seq studies with the MDSeq. *Nucleic Acids Research*, 45(13):e127–e127.
Rifkin, D. B., Rifkin, W. J., and Zilberberg, L. (2018). LTBPs in biology and medicine; LTBP diseases. *Matrix Biology*, 71-72:90–99.
Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.
Roberts, A. G. K., Catchpoole, D. R., and Kennedy, P. J. (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.
Robertson, I. B., Horiguchi, M., Zilberberg, L., Dabovic, B., Hadjiolova, K., and Rifkin, D. B. (2015). Latent TGF-β-binding proteins. *Matrix Biology*, 47:44–53.
Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor

562 package for differential expression analysis of digital gene expression data. *Bioinfor-*
563 *matics*, 26(1):139–140.
564 Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential
565 expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25.
566 Saurin, A. (2022). Bioinformatics tools for genomics and transcriptomics
567 analyses: ENSEMBL ID to Gene Symbol Converter. Available at:
568 https://www.biotools.fr/human/ensembl_symbol_converter. Accessed: 31 August
569 2022.
570 Shi, Y., Liu, Z., Shen, Y., and Zhu, H. (2018). A novel perspective linkage between kid-
571 ney function and Alzheimer's disease. *Frontiers in Cellular Neuroscience*, 12:Article
572 384.
573 Smyth, G. K. (2005). limma: Linear models for microarray data. In Gentleman,
574 R., Carey, V. J., Huber, W., Irizarry, R. A., and Dudoit, S., editors, *Bioinformatics*
575 *and Computational Biology Solutions Using R and Bioconductor*, pages 397–420.
576 Springer, NY.
577 Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years.
578 *Nature Reviews Genetics*, 20:631–656.
579 Stegeman, R. and Weake, V. M. (2017). Transcriptional signatures of aging. *Journal of*
580 *Molecular Biology*, 429(16):2427–2437.
581 Strbenac, D., Mann, G. J., Yang, J. Y. H., and Ormerod, J. T. (2016). Differential distribu-
582 tion improves gene selection stability and has competitive classification performance
583 for patient survival. *Nucleic Acids Research*, 44(13):e119.
584 Thakkar, R., Sareddy, G. R., Zhang, Q., Wang, R., Vadlamudi, R. K., and Brann, D.
585 (2018). PELP1: a key mediator of oestrogen signalling and actions in the brain.
586 *Journal of Neuroendocrinology*, 30(2):e12484.
587 Van den Berge, K., Hembach, K. M., Soneson, C., Tiberi, S., Clement, L., Love, M. I.,
588 Patro, R., and Robinson, M. D. (2019). RNA sequencing data: Hitchhiker's guide to
589 expression analysis. *Annual Review of Biomedical Data Science*, 2:139–173.
590 Visconte, C., Canino, J., Vismara, M., Guidetti, G. F., Raimondi, S., Pula, G., Torti,
591 M., and Canobbio, I. (2020). Fibrillar amyloid peptides promote platelet aggregation
592 through the coordinated action of ITAM- and ROS-dependent pathways. *Journal of*
593 *Thrombosis and Haemostasis*, 18(11):3029–3042.
594 Wang, X., Allen, M., İş, O., Reddy, J. S., Tutor-New, F. Q., Casey, M. C., Carrasquillo,
595 M. M., Oatman, S. R., Min, Y., Asmann, Y. W., Funk, C., Nguyen, T., Ho, C. C. G.,
596 Malphrus, K. M., Seyfried, N. T., Levey, A. I., Younkin, S. G., Murray, M. E., Dickson,
597 D. W., Price, N. D., Golde, T. E., and Ertekin-Taner, N. (2022). Alzheimer's disease
598 and progressive supranuclear palsy share similar transcriptomic changes in distinct
599 brain regions. *Journal of Clinical Investigation*, 132(2):e149904.

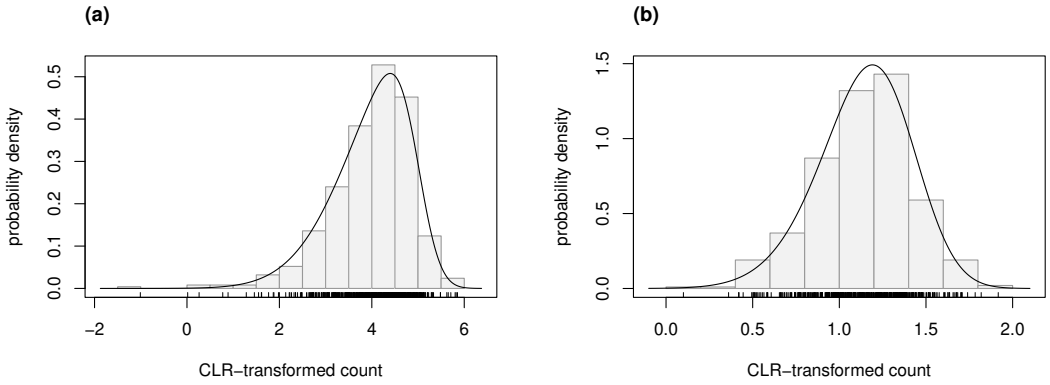

Figure 1. Histograms of CLR-transformed counts for two genes with fitted skew-normal curve for (a) the Valentim dataset ($\hat{\mu} = 3.968$ (s.e. $= 0.038$), $\hat{\sigma} = 0.858$, (s.e. $= 0.030$) and $\hat{\gamma} = -0.732$ (s.e. $= 0.055$)); (b) the Kelmer dataset ($\hat{\mu} = 1.140$ (s.e. $= 0.012$), $\hat{\sigma} = 0.275$ (s.e. $= 0.009$) and $\hat{\gamma} = -0.336$ (s.e. $= 0.107$)).

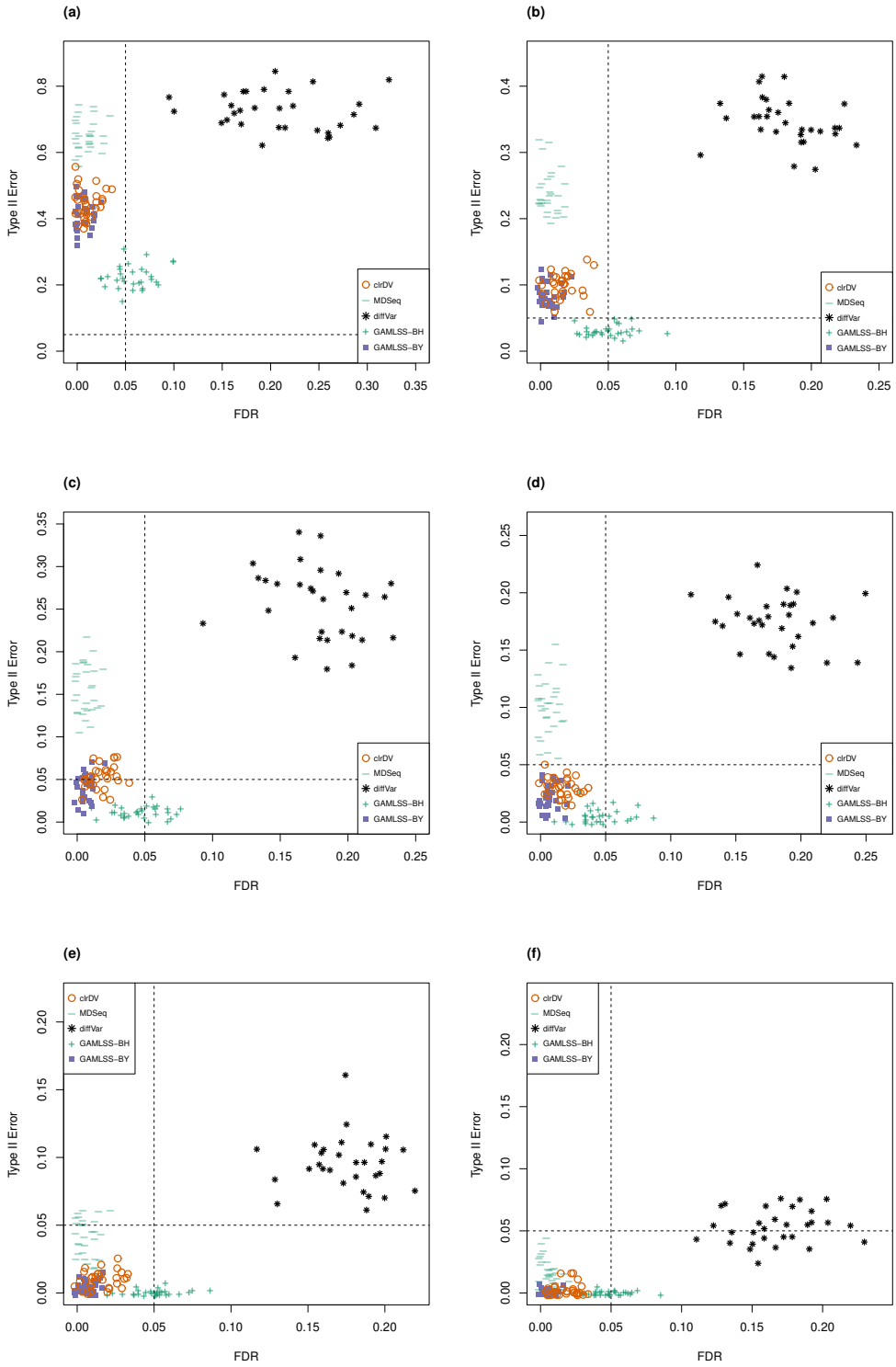

Figure 2. Scatter plots of probability of Type II error vs. FDR for simulation study of the Valentim dataset (30 instances) for samples size per group of (a) 50, (b) 100, (c) 125, (d) 150, (e) 200, and (f) 250. Dashed lines represent probability of Type II error and FDR of 0.05.

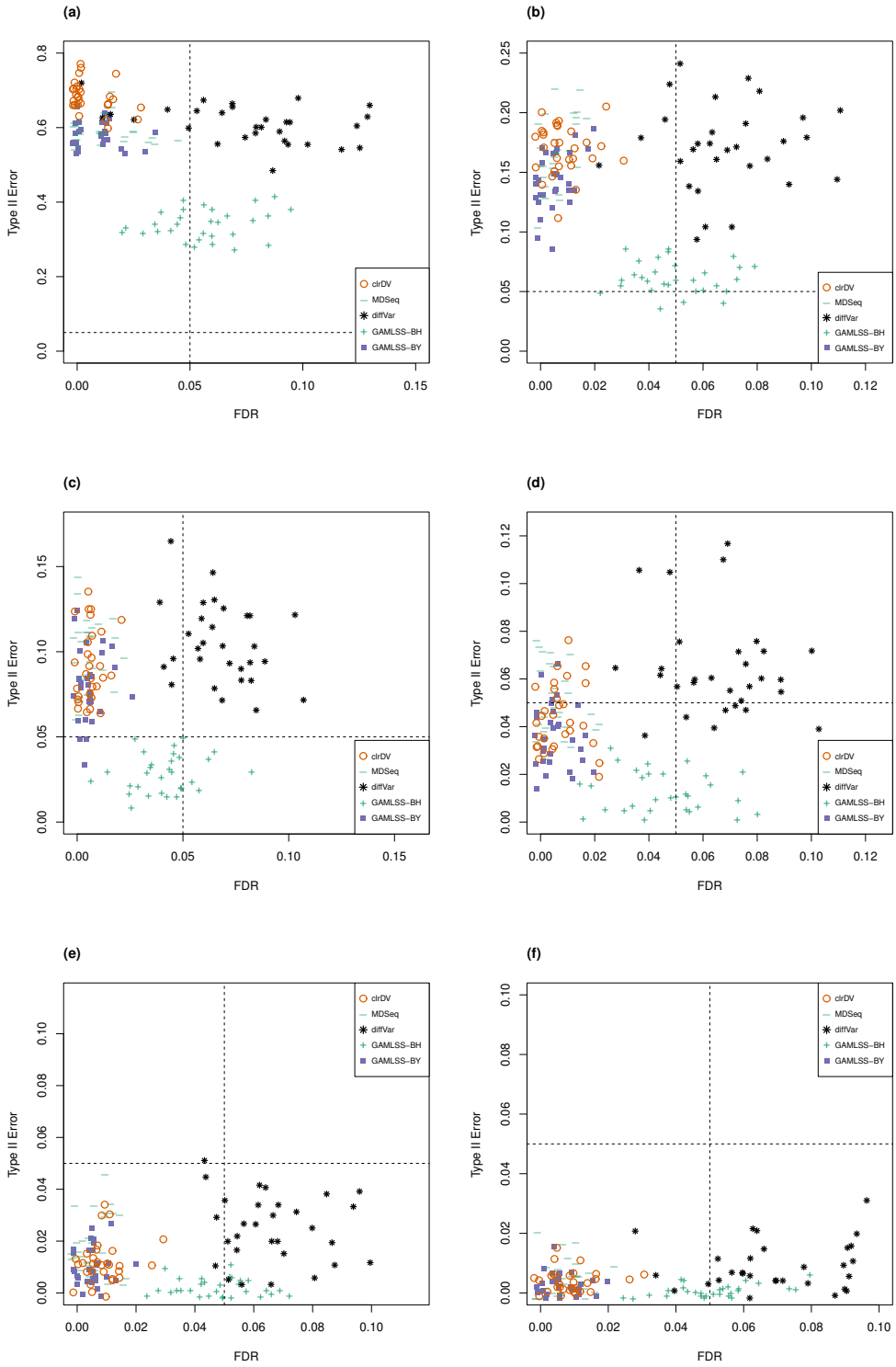

Figure 3. Scatter plots of probability of Type II error vs. FDR for simulation study of the Kelmer dataset (30 instances) for samples size per group of (a) 50, (b) 100, (c) 125, (d) 150, (e) 200, and (f) 250. Dashed lines represent probability of Type II error and FDR of 0.05.

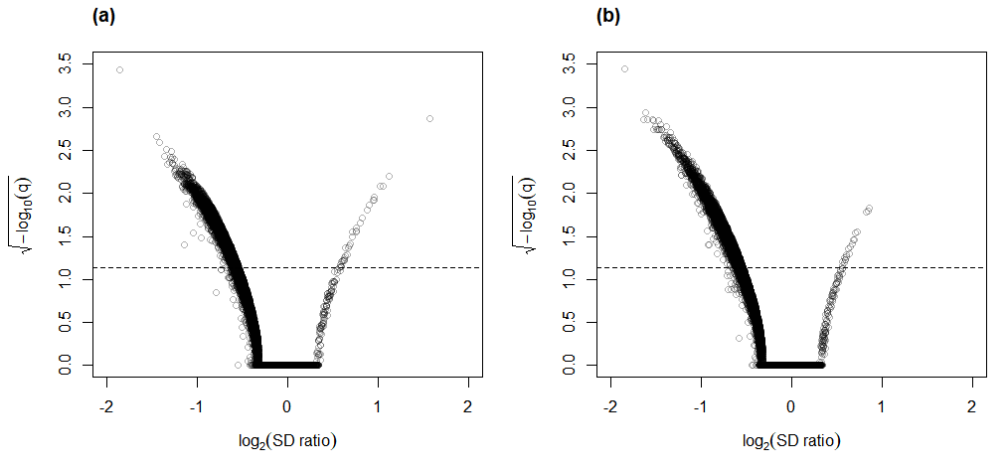

Figure 4. Volcano plots for (a) control vs. AD and (b) control vs. PSP comparisons for the Mayo RNA-Seq dataset. Dashed line represents the threshold of BY- adjusted p-value (q) at 0.05 for flagging DV genes. The number of DV genes with $\log_2(\text{SD ratio}) > 0$ and $\log_2(\text{SD ratio}) < 0$ respectively: (a) 32 and 4722; (b) 19 and 4840.

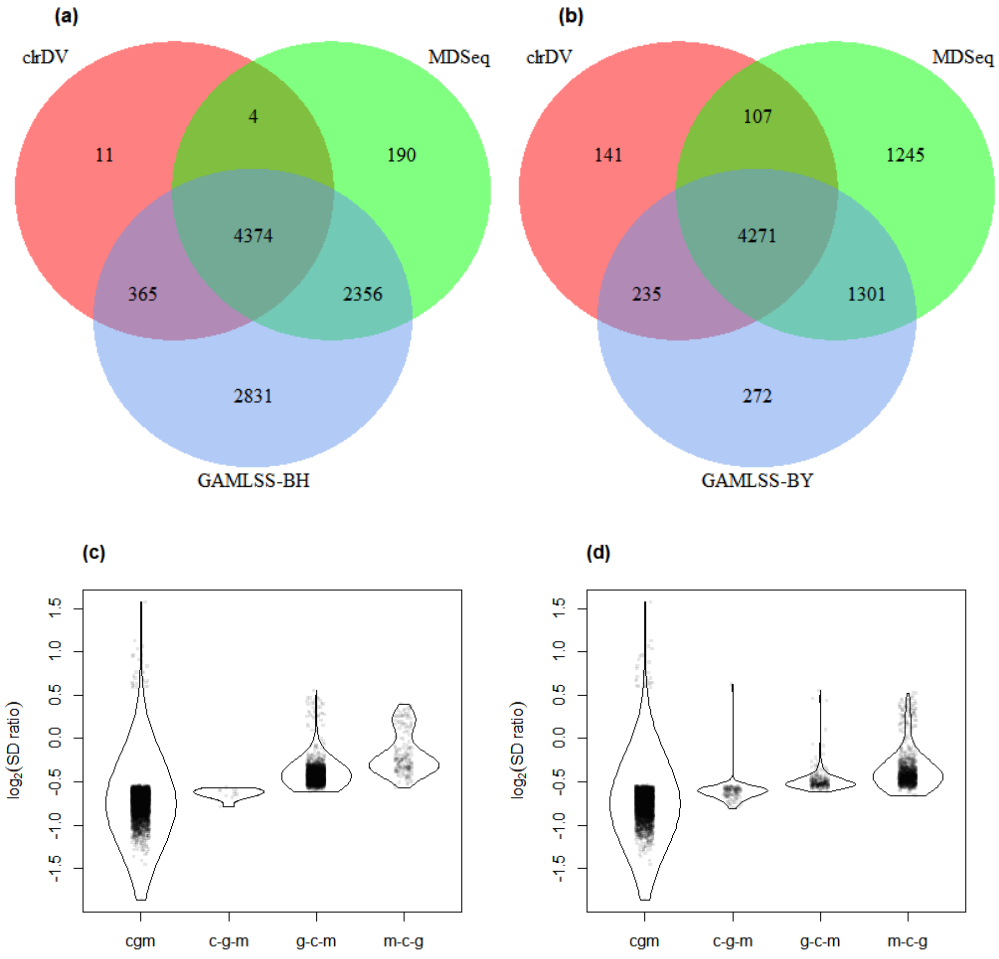

Figure 5. Venn diagrams of DV genes detected by clrDV, MDSeq and (a) GAMLSS-BH; (b) GAMLSS-BY for the control vs. AD comparison. Violin plots of the distribution of estimated $\log_2(\text{SD ratio})$ of the DV genes detected using clrDV, MDSeq and (c) GAMLSS-BH; (d) GAMLSS-BY. Abbreviations: cgm = DV genes detected by clrDV, GAMLSS and MDSeq; c-g-m = DV genes detected by clrDV only; g-c-m = DV genes detected by GAMLSS-BH only; m-c-g = DV genes detected by MDSeq only.

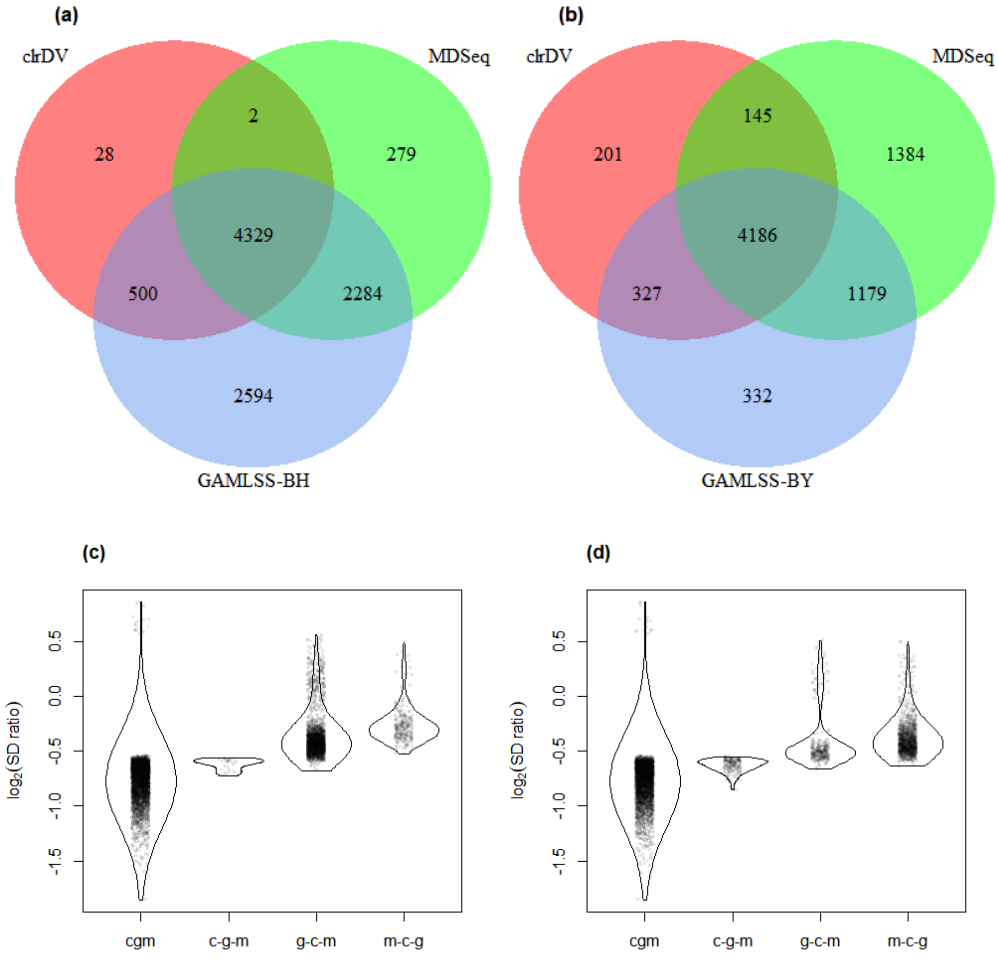

Figure 6. Venn diagrams of DV genes detected by clrDV, MDSeq and (a) GAMLSS-BH; (b) GAMLSS-BY for the control vs. PSP comparison. Violin plots of the distribution of estimated $\log_2$(SD ratio) of the DV genes detected using clrDV, MDSeq and (c) GAMLSS-BH; (d) GAMLSS-BY. Abbreviations: cgm = DV genes detected by clrDV, GAMLSS and MDSeq; c-g-m = DV genes detected by clrDV only; g-c-m = DV genes detected by GAMLSS-BH only; m-c-g = DV genes detected by MDSeq only.

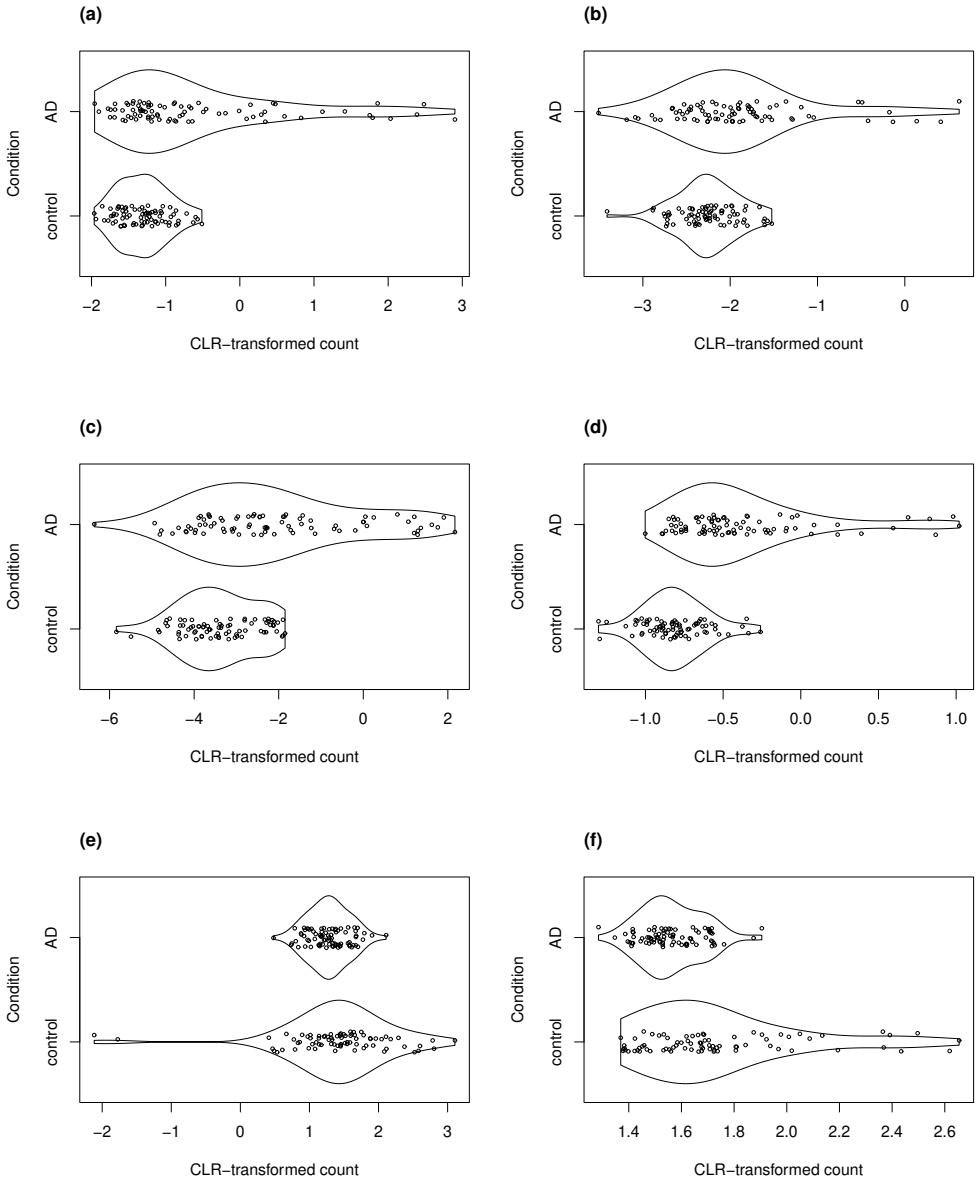

Figure 7. Violin plots of selected DV genes detected in the control vs. AD comparison. (a) SLC47A1, (b) C2orf40, (c) SLPI, (d) LTBP2 have the largest SD ratio (> 2); (e) GP1BB and (f) PELP1 have SD ratio about 0.4.