PLOS ONE

RESEARCH ARTICLE

Differentially expressed heterogeneous overdispersion genes testing for count data

Yubai Yuan[1], Qi Xu[2], Agaz Wani[3], Jan Dahrendorff[3], Chengqi Wang[3], Arlina Shen[4], Janelle Donglasan[3], Sarah Burgan[3], Zachary Graham[3], Monica Uddin[3], Derek Wildman[3], Annie Qu[2]*

1 Department of Statistics, The Pennsylvania State University, State College, PA, United States of America, **2** Department of Statistics, University of California Irvine, Irvine, CA, United States of America, **3** Genomics Program, College of Public Health, University of South Florida, Tampa, FL, United States of America, **4** University of California Berkeley, Berkeley, CA, United States of America

* aqu2@uci.edu

OPEN ACCESS

Citation: Yuan Y, Xu Q, Wani A, Dahrendorff J, Wang C, Shen A, et al. (2024) Differentially expressed heterogeneous overdispersion genes testing for count data. PLoS ONE 19(7): e0300565. https://doi.org/10.1371/journal.pone.0300565

Editor: Andrea Tangherloni, Bocconi University: Universita Bocconi, ITALY

Received: March 5, 2023

Accepted: February 29, 2024

Published: July 17, 2024

Data Availability Statement: The datasets analysed during the current study are available in the NCBI's Gene Expression Omnibus (GEO) repository and are accessible through GEO Series accession number GSE219208 and link https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE219208. The experiment description and algorithm implementation are available via the following weblinks: https://github.com/xiaobai0518/DEHOGT.

Funding: MU (Monica Uddin) is awarded with grant: National Institutes of Health R01MD011728.

Abstract

The mRNA-seq data analysis is a powerful technology for inferring information from biological systems of interest. Specifically, the sequenced RNA fragments are aligned with genomic reference sequences, and we count the number of sequence fragments corresponding to each gene for each condition. A gene is identified as differentially expressed (DE) if the difference in its count numbers between conditions is statistically significant. Several statistical analysis methods have been developed to detect DE genes based on RNA-seq data. However, the existing methods could suffer decreasing power to identify DE genes arising from overdispersion and limited sample size, where overdispersion refers to the empirical phenomenon that the variance of read counts is larger than the mean of read counts. We propose a new differential expression analysis procedure: heterogeneous overdispersion genes testing (DEHOGT) based on heterogeneous overdispersion modeling and a post-hoc inference procedure. DEHOGT integrates sample information from all conditions and provides a more flexible and adaptive overdispersion modeling for the RNA-seq read count. DEHOGT adopts a gene-wise estimation scheme to enhance the detection power of differentially expressed genes when the number of replicates is limited as long as the number of conditions is large. DEHOGT is tested on the synthetic RNA-seq read count data and outperforms two popular existing methods, DESeq2 and EdgeR, in detecting DE genes. We apply the proposed method to a test dataset using RNAseq data from microglial cells. DEHOGT tends to detect more differently expressed genes potentially related to microglial cells under different stress hormones treatments.

Introduction

High-throughput sequencing of DNA fragments and mRNA-seq techniques are powerful tools based on next generation sequencing technologies [1] for monitoring RNA abundance to detect genetic variation. Specifically, for RNAseq, the sequenced RNA fragments are aligned with reference genome sequences, and the number of sequence fragments assigned to each

The website of National Institutes of Health is https://www.nih.gov/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

gene is counted for each sample. Then we can compare read counts between different biological conditions or between different genetic variants to infer genetic information based on biological systems of interest [2]. In the analysis of RNA-seq data, read counts do not have a prior upper bound, thus regression models based on a binomial distribution with a pre-specified number of trails do not apply [3]. Linear regression is therefore not feasible as count data is always a non-negative integer. More importantly, RNA-seq data presents high overdispersion, implying that the variance of the count can be much larger than its mean. Given that the sample sizes are typically small for RNA-seq analysis due to the cost and other factors, statistical modeling needs to address the large variation from the data and to improve the power of detecting differential gene expressions.

One fundamental clinical interest of applying RNA-seq analysis is to understand the mechanism of post-traumatic stress disorder (PTSD) formulation. PTSD is a common severe psychiatric disorder that develops following exposure to a life-threatening or traumatic experience [4]. PTSD is known to cause negative effect on an individual's life quality via the PTSD condition itself or the relevant comorbidities. Previous works [5, 6] show that only a small proportion of individuals experience traumatic events will develop PTSD. Meanwhile, the majority of people exposed to trauma are resilient even after repeated exposures to trauma [7]. In addition, various risk factors of PTSD have been identified such as low socio-economic status, social support and gender [8–10].

Significant individual heterogeneity of either response to trauma or the PTSD development originates from the individual epigenetic variability. Specifically, previous studies reveal the connection between PTSD and immune system functioning, and several genes such as FKBP5 involved with the immune system are also found to be differentially expressed among PTSD individuals [11–13]. In particular, previous work [14] has identified monocytes as a key cell type in differentiating male subjects with versus without lifetime PTSD. In addition, rodent studies have implicated peripheral monocytes in inducing anxiety-like behavior through trafficking of proinflammatory monocytes to the brain via activated microglia. Following this line of research, in this paper, we collect RNA-seq data from the well-designed lab experiments to investigate differential expression of genes in human microglia cells under different immune characteristic environments. This is an important step for understanding the role of microglia cells and immune-related genes in PTSD development.

The main challenge in analyzing microglial RNA-seq datasets lies in the high and heterogeneous overdispersion in the read counts. As an illustration, Fig 1 shows the histogram of the empirical RNA read counts from microglial data, where the read counts are highly spread out and the variance can be much larger than the mean. Several differential expression analysis methods have been developed to address the overdispersion issue in RNA-seq read counts. Among these methods, the DESeq2 [15] and EdgeR [16] are the most popular and are implemented and available using the R [17]. Specifically, the DESeq2 analyzes count data by using a shrinkage estimation for dispersions as well as fold changes to improve stability and interpretability of estimates. EdgeR is designed for the analysis of replicated count-based expression data, and is based on the method developed by Robinson and Smyth [18] using an overdispersed Poisson model to account for the read count variability. However, most existing methods adopt the shrinkage strategy when estimating the level of overdispersion by assuming that genes with similar expression strength have homogeneous dispersion levels. Although overdispersion shrinkage is a popular technique used to improve the estimation of variance for gene expression read count, it has been found that the overdispersion shrinkage also leads to the overestimation of the true biological variability in the data [19, 20]. Shrinking the estimates of gene-wise dispersion towards a common value might diminish the true differences in gene expression variability between different genes or conditions. In addition, the overdispersion

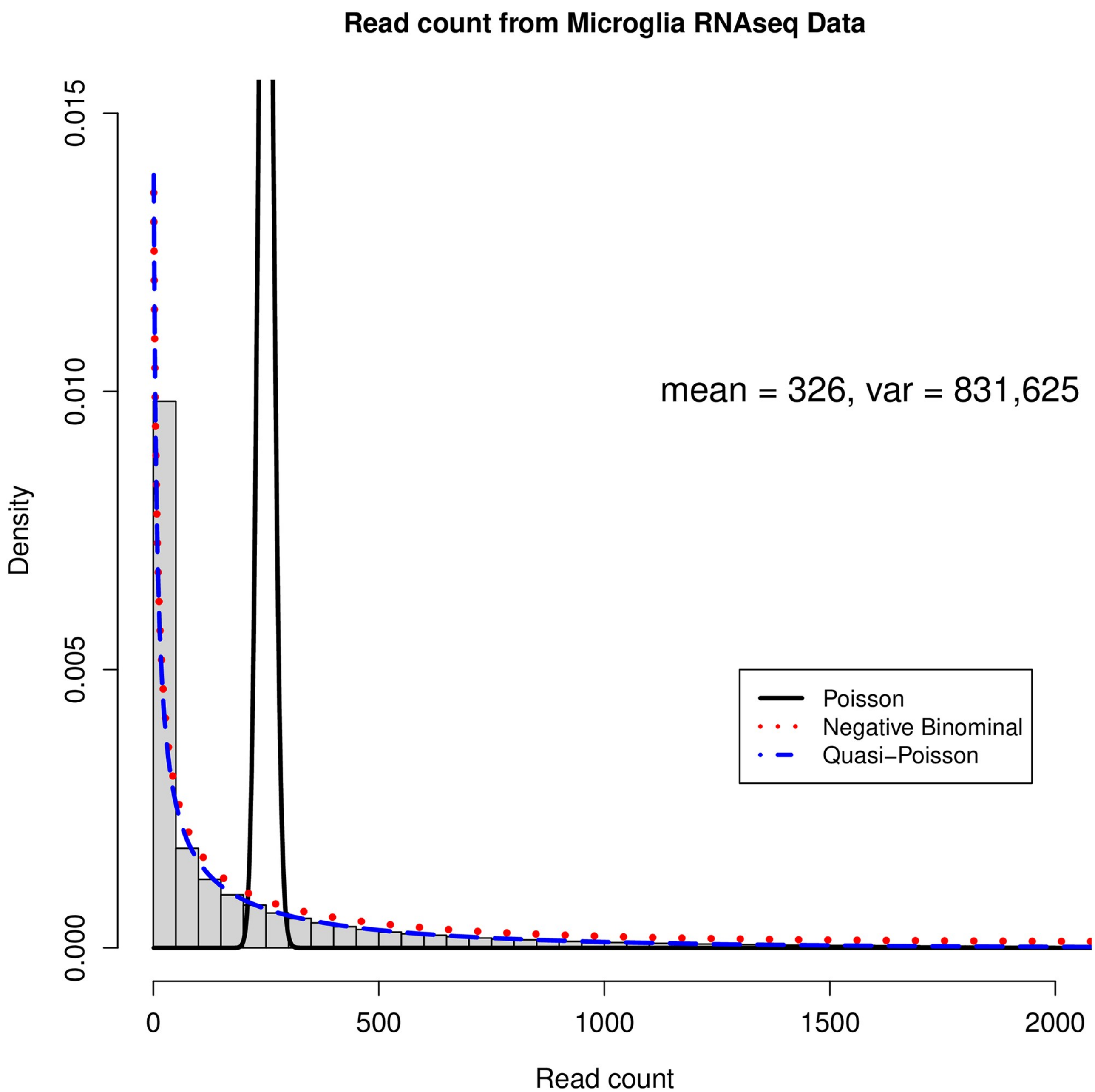

Fig 1. The overdispersion in the real RNA-seq count data.

https://doi.org/10.1371/journal.pone.0300565.g001

shrinkage can introduce bias in the estimation of variance for less expressed genes [21, 22], which leads to low sensitivity in detecting potential differentially expressed genes.

In this paper, we propose a new differential expression analysis framework based on generalized linear modeling. Compared with other popular RNA-seq analysis methods such as DESeq2 and EdgeR, the main advantages of the proposed method for differentially expressed heterogeneous overdispersion genes testing (DEHOGT) are as follows. First, our method jointly estimates the fold change and overdispersion parameters over samples from all treatment conditions, which increases the effective sample size and leads to more accurate inference. Second, and more importantly, our model adopts a within-sample independent structure among genes without assuming that genes with similar expression strength have homogeneous dispersion levels. Therefore, our method can better account for the heterogeneity in count dispersion and select more relevant genes. Third, our method allows for fully independent gene-wise inference and hence can achieve computational scalability to handle large gene datasets by implementing parallel computing. Finally, the proposed method enjoys the flexibility of adapting different overdispersion patterns by allowing different count generating distributions in the inference procedure.

Materials and methods

We develop a new differentially expressed gene testing procedure to account for the heterogeneity in gene-wise overdispersion levels. Traditionally, Poisson and multinomial distributions are used to model count data with large variance. However, the variance of RNA sequence counts tends to be much larger than that of the Poisson or multinominal distribution [23].

Overdispersion refers to the empirical phenomenon such that the RNA-seq data exhibits extra-Poisson variability, i.e., the variance of read counts is larger than the mean of read counts, compared with the traditional Poisson distribution model for count data where the mean is equivalent to the variance. Overlooking the overdispersion could result in biased and misleading inference about gene association to the response of interest. To overcome this limitation, we first introduce adaptive distribution modeling in this paper to analyze the overdispersed RNA-seq count data. We utilize a quasi-Poisson distribution and a negative binominal distribution as the read count, thus generating a distribution similar to the overdispersion pattern which is based on empirical data. Specifically, we denote Y as the random count response, and the quasi-Poisson distribution satisfies:

$$E(Y) = \mu, ; \mathrm{Var}(Y) = \theta\mu, \tag{1}$$

where $\mu > 0$ is the mean of Y, $\theta \geq 1$ denotes the overdispersion parameter, and larger θ indicates higher overdispersion level. Although μ is larger than 0, Y can be any nonnegative integer. Note that Poisson model assumes that the variance is equal to the mean, e.g., $\theta = 1$. In contrast, a quasi-Poisson distribution provides more flexibility to allow variance increases as a linear function of the mean. Accordingly, the quasi-Poisson regression generalizes the Poisson regression and is adopted to model an overdispersed count variable. The quasi-Poisson model is characterized by the first two moments, i.e., mean and variance. Besides the quasi-Poisson distribution, the negative binominal distribution can also be used to model overdispersed count data satisfying:

$$E(Y) = \mu, ; \mathrm{Var}(Y) = \mu + \mu^2/\theta, \tag{2}$$

where $\theta > 0$ is the overdispersion parameter, and smaller θ indicates higher overdispersion level. Similar to the quasi-Poisson distribution, the negative binominal distribution is characterized by the mean and variance while modeling the variance as a quadratic function of the mean. In the following, we denote the quasi-Poisson distribution and negative distribution as quasi- Poisson(μ, θ^{QP}) and negative- binomial(μ, θ^{NB}), respectively. In addition, we use NB and QP as the abbreviation of negative binomial and quasi-Poisson distribution. In Fig 1, we illustrate the distribution density functions by fitting the empirical read counts in our empirical data from microglia cells (see Methods) with 1) Poisson distribution, 2) negative binomial distribution, and 3) quasi-Poisson distribution, respectively. Compared with the Poisson distribution, both the negative binomial and quasi-Poisson distributions provide better approximation by capturing the overdispersion in read counts.

In addition, the read counts of a gene can be affected by other factors in an experiment other than its expression level in the RNA-seq. Therefore, instead of directly modeling the raw count data Y, we first perform count normalization, which makes the expression levels of genes more comparable and accurate between samples. We utilize the Trimmed Mean of M-values normalization (TMM) [24] adopted by EdgeR to compute the normalization factors that correct sample-specific biases. TMM is recommended for most RNA-Seq data where most genes are not differentially expressed across any pairs of the samples. Specifically, we first calculate the normalization factors as the median ratio of gene counts relative to the geometric mean per gene within a specific sample. The normalization factors account for two main non-expression factors; e.g., sequencing depth and RNA composition before between-sample comparison [24]. Consequently, we divide raw counts by sample-specific size factors to yield the effective read count for cross-sample comparisons.

The proposed DEHOGT workflow combines the above ingredients to identify differentially expressed genes. Compared with the two popular RNA-seq analysis methods DESeq2 and EdgeR, the main difference of the proposed method is at the model fitting step of the above algorithm, where the overdispersion parameters $\{\theta_i\}$ are estimated for each gene individually. The DESeq2 and EdgeR estimate the overdispersion parameters by pooling the samples from different genes under the assumption that genes with similar expression strength also share similar overdispersion levels. In contrast, the proposed method does not rely on the homogeneous dispersion assumption and can capture the heterogeneity in different genes' expression levels, especially when the overdispersion of gene is high. In addition, the proposed method allows one to choose different working distributions in Step 3 to model the RNA-seq count data to accommodate different associations between mean and variance presented in the empirical read count data. This provides us additional flexibility in modeling the overdispersion patterns to achieve more accurate read count fitting. Consequently, correctly specified read count overdispersion patterns can lead to higher statistical power of post-hoc testing to detect differentially expressed genes.

We summarize the proposed method (DEHOGT) for the RNAseq read count for detecting differentially expressed (DE) genes as follows. Assume that there exists a total of R different treatments and S samples where each treatment has multiple samples as replicated measurements. We index the gene and sample measurements as g and s such that $g = 1, 2, \cdots, N$ and $s = 1, 2, \cdots, S$. First, the read count data is modeled via one of the following generating distributions:

$$Y_{gs} \sim \text{quasi-Poisson}(K_s\mu_{gs}, \theta_g^{QP}),\ \ Y_{gs} \sim \text{negative-binomial}(K_s\mu_{gs}, \theta_g^{NB}),$$

where K_s denotes the normalization factor for the sth sample obtained by the TMM method. To determine the generating distribution, we check the overdispersion pattern between $E_s(Y_{gs})$ and $\mathrm{Var}_s(Y_{gs})$ from the empirical data. A better quadratic function fitting leads to the choice of a negative binomial distribution and a better linear relation fitting leads to the quasi-Poisson. Here we assume that the gene-wise dispersion parameter θ_g is constant across all samples to estimate the quasi-Poisson distribution θ^{QP}, or the negative binominal distribution θ^{NB}, by utilizing information from samples under different treatments.

To differentiate genes' read counts under different treatments, we model the genewise read count mean via the following generalized linear model:

$$\log \mu_{gs} = \beta^{(1)} x_g + \boldsymbol{\beta}_g^{(2)} \mathrm{T}_s^{\top},$$

where $\boldsymbol{\beta}_g^{(2)} = (\beta_{g1}^{(2)}, \beta_{g2}^{(2)}, \cdots, \beta_{gR}^{(2)})$ represents the fold change of the gth gene under R different treatments, and $\mathrm{T}_s \in \{0, 1\}^R$ is the dummy coding for the treatment membership of the sth sample such that $T_{sr} = 1$ when the sth sample belongs to treatment r, $r = 1, \cdots, R$. In addition, $x_g \in \mathbb{R}^p$ are the gene-wise covariates, so that our method can further adjust other non-expression factors to reduce the bias in inferring the genes' expression level, where R denotes a real number.

Given that $\{\boldsymbol{\beta}_g^{(2)}\}$ represents the gene-wise expression level under different treatments, we can infer whether the gth gene is differentially expressed under the two treatments r_1 and r_2 based on the linear hypothesis testing $H_0 : \boldsymbol{\beta}_{gr_1}^{(2)} - \boldsymbol{\beta}_{gr_2}^{(2)} = 0$. Then we can identify the DE genes under the treatment comparison pair (r_1, r_2) when the corresponding p-value is smaller than a specific cutoff. To control the type-I error of simultaneously testing on multiple genes, we adopt the Benjamini-Hochberg procedure [25] to adjust the gene-wise p-value, and control the false discovery rate. In addition to the adjusted p-value, the magnitude of the logfold change is also suggested as another criterion for choosing DE genes with a logfold change of $\log_2|\boldsymbol{\beta}_{gr_1}^{(2)} - \boldsymbol{\beta}_{gr_2}^{(2)}|$ larger than 1.5 [26, 27]. Therefore, we combine these two criteria, and select the DE genes with an adjusted p-value smaller than 0.05 and an absolute logfold change larger than 1.5.

Our proposed DEHOGT algorithm is summarized as follows:

Algorithm: DEHOGT

1. *(Input)*: For the ith gene $i = 1, \cdots, N$, input read counts $\{Y_{is}\}_{s=1}^S$ from S samples, the covariates x_i associated with the ith gene, and the treatment assignment for each sample $T_s \in \{0, 1\}^R$ from each gene, ($s = 1, \cdots, S$ where R is the number of treatments). Specifying the working distribution indicator I:

$$I = \begin{cases} 1, & \text{choose the quasi-Poisson,} \\ 2, & \text{choose the negative binominal.} \end{cases}$$

2. *(Read count normalization)*: Obtain normalization factor for the ith gene: $K_i = \mathrm{TMM}(\{Y_{is}\}_{s=1}^S)$ ($i = 1, \cdots, N$) where TMM denotes the Trimmed Mean of M-values normalization.
3. *(Fitting the generalized linear model)*: For the ith gene, estimate the fold change parameter $\boldsymbol{\beta}_i^{(2)}$ and the overdispersion parameters θ_i:

$$(\hat{\boldsymbol{\beta}}_i^{(2)}, \hat{\theta}_i) = \underset{\beta,\theta}{\operatorname{argmax}} \prod_{s=1}^{S} f_I(Y_{is}, K_s\mu_i(\beta), \theta_i), \tag{3}$$

$$\log \mu_i(\beta) = \beta^{(1)} x_i + \boldsymbol{\beta}_i^{(2)} T_s^\top, \tag{4}$$

where f_I denotes the probability density function of the chosen working distribution.
4. *(Post-hoc testing)*: For the ith gene and a specific interesting treatment pair (r_1, r_2), perform

$$H_0 : \hat{\boldsymbol{\beta}}_{gr_1}^{(2)} - \hat{\boldsymbol{\beta}}_{gr_2}^{(2)} = 0,$$

and obtain p-value p_i.
5. *(DE gene filtering)*: For the treatment pair (r_1, r_2), obtain gene-wise adjusted p value, using Benjamini-Hochberg [25] adjusting for

`false positive discovery:`

$$\{p_i^{adj}\}_{i=1}^N = \text{Benjamini-Hochberg}(\{p_i\}_{i=1}^N).$$

Select the ith gene if $p_i^{adj} < 0.05$ and $\log_2|\boldsymbol{\beta}_{ir_1}^{(2)} - \boldsymbol{\beta}_{ir_2}^{(2)}| > 1.5$.

6. *(Output)*: Set of differentially expressed genes and the corresponding fold change estimation $\hat{\boldsymbol{\beta}}^{(2)}$.

Read count normalization

The proposed method adopts the popular TMM normalization method because the TMM normalization method is more methodologically simpler and more computationally efficient compared with those normalization methods that incorporate gene length, which makes the proposed method more versatile and easier to apply across a wide range of datasets and experimental conditions. In addition, empirical studies [24] demonstrate that TMM normalization is more robust to low read counts and achieves better statistical test efficiency compared with traditional normalization methods such as RPKM, FPKM, and TPM. In the application scenario where gene length normalization is necessary, it is more straightforward for the user to add the pre-processing step of gene length normalization such as the RPKM, FPKM, and TPM methods, before the proposed RNA-seq analysis workflow. In addition, the proposed method is mainly designed for RNA-seq inter-sample analyses, i.e., identifying genes that are differentially expressed among different treatment conditions or disease states. Empirical studies [28] have found that the performance is similar whether the normalization methods accounting for gene length are utilized or not.

Working distribution selection

For choosing the appropriate distribution, we can investigate the empirical mean-variance relation of sample counts. Specifically, we denote the mean of read count Y as $E(Y) = \mu$, the quasi-Poisson distribution satisfies $Var(Y) = \theta\mu$, and the negative binomial distribution satisfies $Var(Y) = \mu + \mu^2/\theta$ where θ is the dispersion parameter. Consequently, we can determine the read count distribution based on whether the sample count mean and sample count variance follow a linear relation or quadratic relation.

In addition, we can determine the count distribution based on the biological questions to be answered by comparing sample counts under different conditions. Consider the generalized linear regression problem $Y \sim X\beta$ where X is the design matrix representing different experimental conditions, and β is the significance of genes. The regression coefficient can be estimated via iteratively weighted least squares $\hat{\beta} = (X^T WX)^{-1} X^T WY$, where W denotes sample weights. Notice that the quasi-Poisson and negative binomial distributions lead to different sample weights with quasi-Poisson distribution being $W = \text{diag}\left(\frac{\mu_1}{\theta}, \cdots, \frac{\mu_n}{\theta}\right)$, and the negative binomial distribution being $W = \text{diag}\left(\frac{\mu_1}{1+\theta\mu_1}, \cdots, \frac{\mu_n}{1+\theta\mu_n}\right)$ where $\mu_1, \cdots, \mu_n$ are the count means of n samples. Therefore, the quasi-Poisson distribution provides more weights on the samples with larger counts, while the negative binomial distribution gives similar weights. If the applications emphasize the analysis of highly expressed genes, we can choose quasi-Poisson distribution to improve the identification power [29], such as in the tasks of identifying housekeeping or marker genes of specific diseases or tissue, and gene functional pathway analysis. On the other hand, we can choose a negative binomial distribution if the gene expression levels are strongly overdispersed.

Overdispersion from technical variation

The proposed method aims to separate the biological variation due to biological heterogeneity from systematic variation due to treatment conditions when identifying DE genes. On the other hand, the technical variance is handled by appropriate quality control procedures in the RNA-seq data preprocessing step. Specifically, the raw RNA-seq read counts are preprocessed by the quality control, including identification and correction of low-read counts, sequence content correction, and normalization. Quality control is generally effective and standard in handling potential technical variation in RNA-seq data [30, 31], and the biological variance remains as the main source of overdispersion in the preprocessed read counts. Besides quality control, there exists other effective preprocessing method for handling technical variance and low-count samples, via mixture distribution modeling between random read counts and effective read counts [32].

Results

We compare the proposed DEHOGT method with two popular RNA-seq analysis methods DESeq2 [15] and EdgeR [16] in detecting differentially expressed genes on the simulated

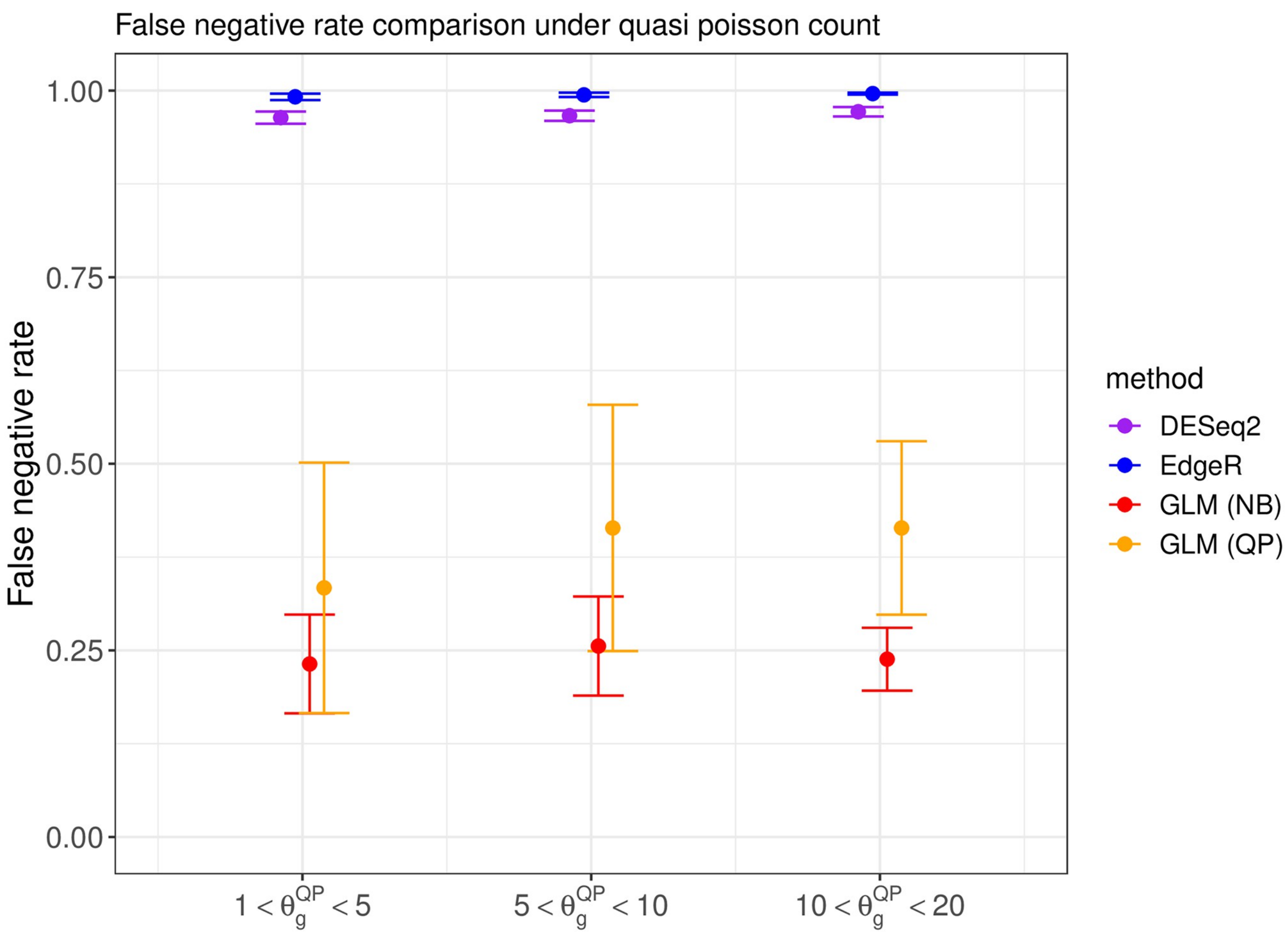

Fig 2. The false negative rate from different methods when the read counts follow the quasi-Poisson distribution with different overdispersion levels θ_g^{QP} in simulation setting 1. The bars represents the standard deviation of the false negative rate over repeated experiments.

https://doi.org/10.1371/journal.pone.0300565.g002

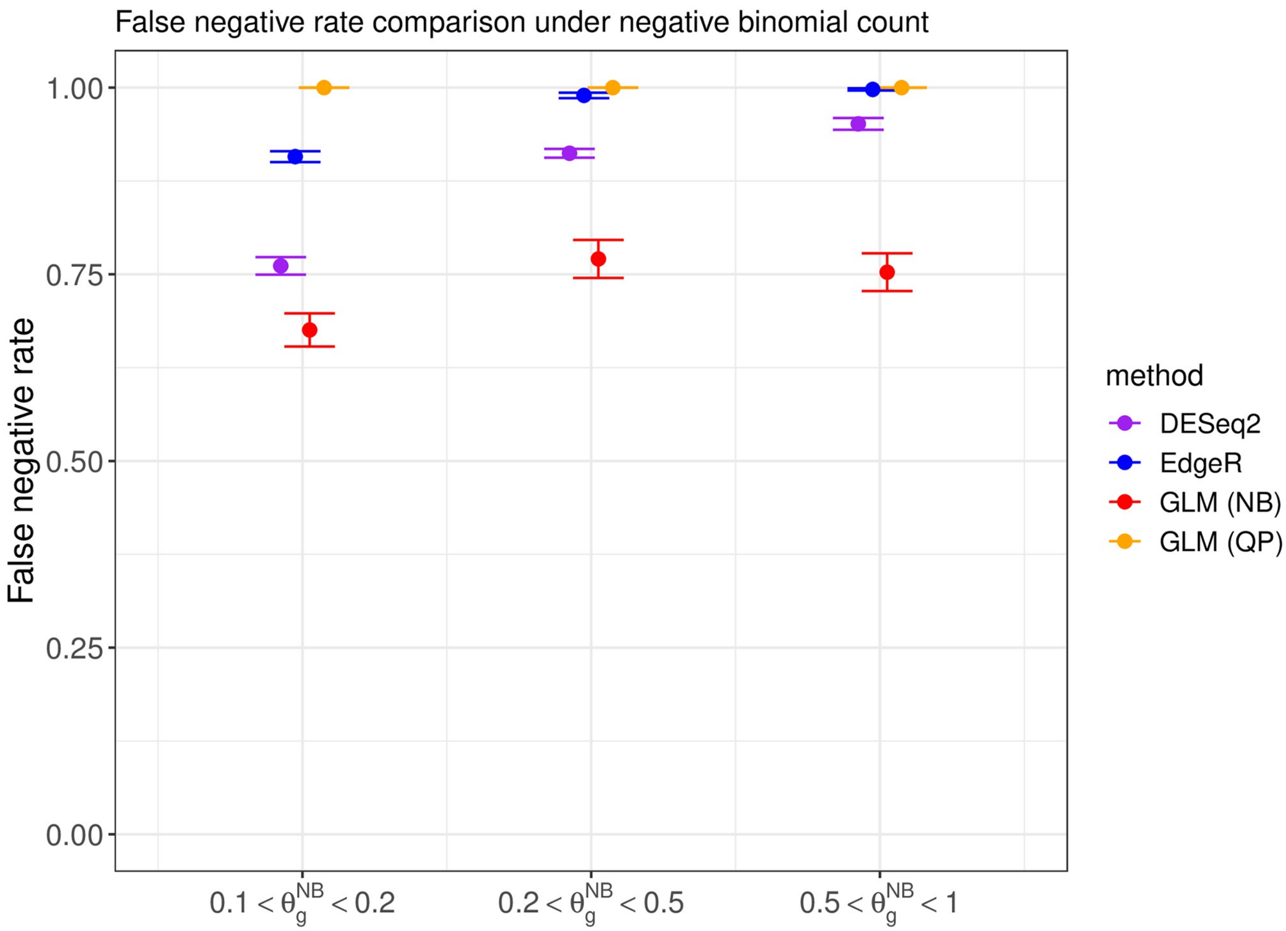

Fig 3. The false negative rate from different methods when the read counts follow the negative binomial distribution with different overdispersion levels θ_g^{NB} in simulation setting 1. The bars represents the standard deviation of the false negative rate over repeated experiments.

https://doi.org/10.1371/journal.pone.0300565.g003

read count data and microglia cell RNA-seq data. EdgeR and Deseq2 are the most popular RNA-Seq differential analysis methods, and serve as a benchmark in most RNA-Seq differential analysis method evaluation studies. In addition, we mainly investigate the performance of the proposed method when the number of samples is relatively small, as a small numbers of replicates occur commonly in many RNA-seq studies. Extensive numerical studies have demonstrated that DESeq2 outperforms other methods [33] in the scenario of a limited sample size. Furthermore, both proposed methods, DESeq2, and EdgeR, fall into the categories of parametric modeling with a negative binomial distribution, which makes for a fair comparison.

In the first simulation setting, the discrepancy in expression level between the treatment and control group is weak for DE genes, while the average expression levels for both groups are high. In the second simulation setting, the expression discrepancy between the treatment and control group is strong for DE genes, while the average expression levels for both groups are low.

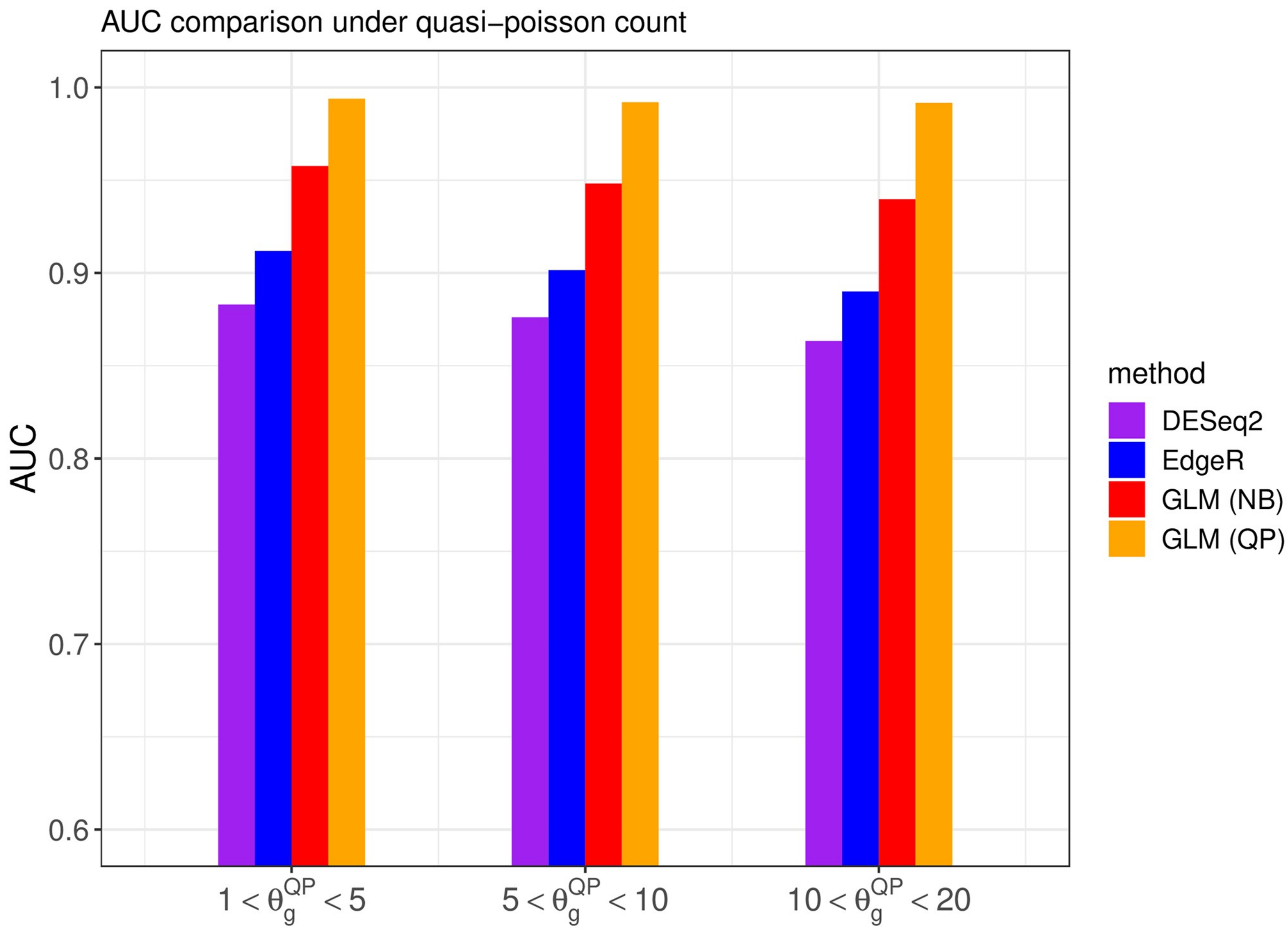

Fig 4. The area under the ROC curve from different methods when the read counts follow the quasi-Poisson distribution with different overdispersion levels θ_g^{QP} in simulation setting 1.

https://doi.org/10.1371/journal.pone.0300565.g004

Read count with low discrepancy of expression level

In the first setting, we simulate the read count data following the negative binomial and the quasi-Poisson distribution:

$$
\begin{aligned}
Y_{gs} &\sim QP(\text{ mean } = \mu_{gs}, \text{var} = \mu_{gs}\theta_g^{QP}),\ g = 1, \cdots, N,\ s = 1, \cdots, |S|, \\
Y_{gs} &\sim NB(\text{ mean } = \mu_{gs}, \text{var} = \mu_{gs}(1 + \mu_{gs}/\theta_g^{NB})),\ g = 1, \cdots, N,\ s = 1, \cdots, |S|,
\end{aligned}
$$

where $g \in \{1, \cdots, N\}$ denotes gene indexes and the total number of genes $N = 12, 500$. We use $G^{DE} \subset \{1, \cdots, N\}$ to denote the set of differentially expressed genes with $|G^{DE}| = 2500$. In addition, $s \in S$ and $|S| = 12$ denote the sample index with $S = S_1 \cup S_2$, $|S_1| = |S_2| = 6$, where S_1 and S_2 indicate the samples in the control group and the treatment group, respectively. Here the mean parameters μ_{gs} are similar to setting [34] in the RNA-seq data analysis. Specifically, the formulations are:

$$
\mu_{gs} = E\left[Y_{gs}\right] = \begin{cases} M_{gs} + \eta_g, & g \in G^{DE}, s \in S_2 \\ M_{gs}, & o.w \end{cases},
$$

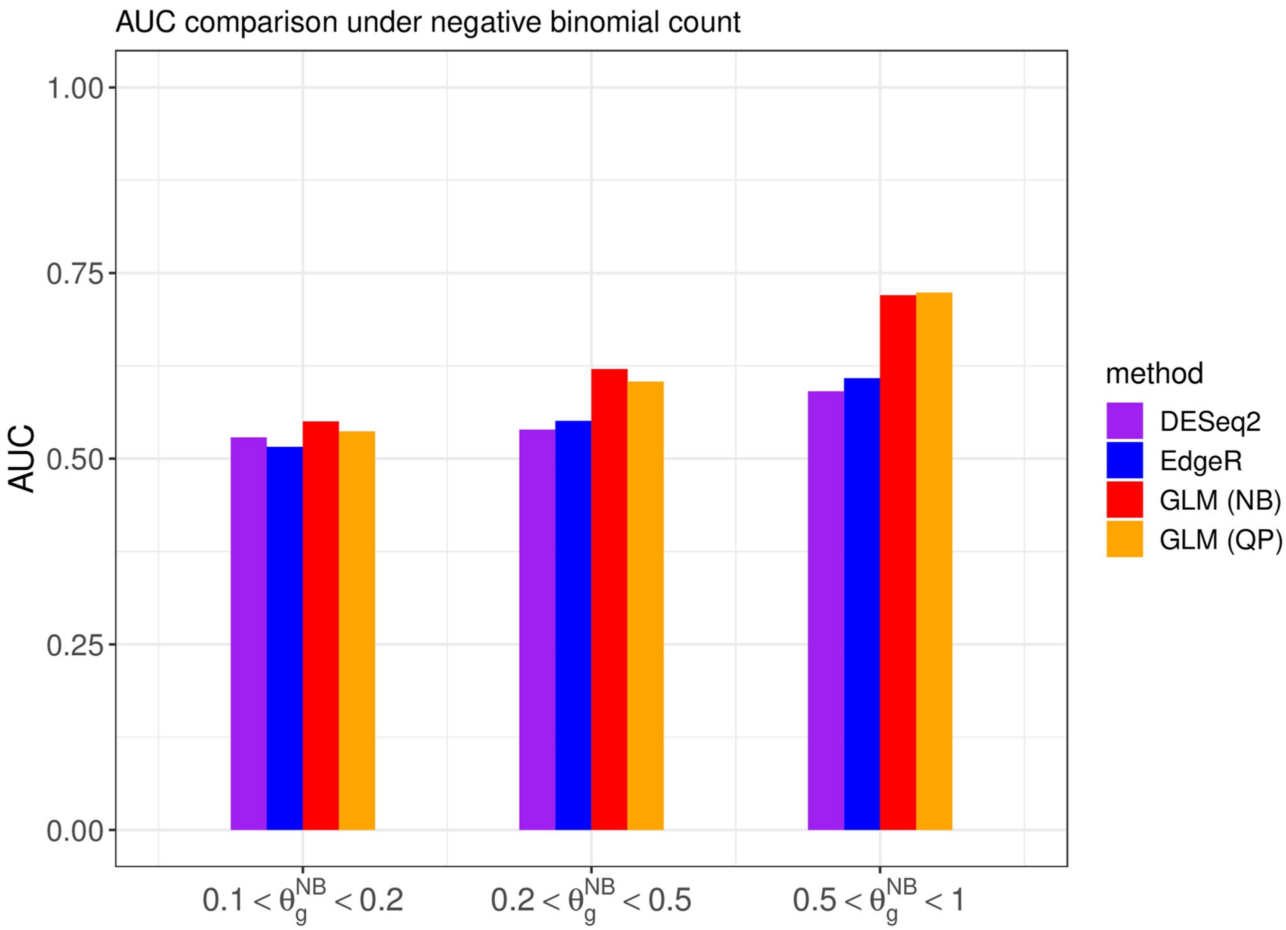

Fig 5. The area under the ROC curve from different methods when the read counts follow the negative binomial distribution with different overdispersion levels θ_g^{NB} in simulation setting 1.

https://doi.org/10.1371/journal.pone.0300565.g005

where we sample M_{gs} from Unif[0, U_s], and $U_s \sim$ Unif[600, 800] is the sample-wise sequencing depth. Furthermore, we sample η_g from exp(1/100) as the up-regulated signal of the differentially expressed genes. We consider three different overdispersion levels for the read counts from the quasi-Poisson distribution as

$$\theta_g^{QP} \sim \text{Unif}(1,5), \theta_g^{QP} \sim \text{Unif}(5,10), \theta_g^{QP} \sim \text{Unif}(10,20),$$

where a larger θ_g^{QP} indicates a greater overdispersion level. Similarly, we consider three overdispersion levels under the negative binomial read counts as

$$\theta_g^{NB} \sim \text{Unif}(0.1,0.2), \theta_g^{NB} \sim \text{Unif}(0.2,0.5), \theta_g^{NB} \sim \text{Unif}(0.5,1),$$

where a smaller θ_g^{NB} indicates a greater level of overdispersion.

We compare the performance of DESeq2 [15], EdgeR [16], and the proposed DEHOGT in identifying differentially expressed genes using an adjusted p value less than 0.05 and an absolute value of logfold change larger than 1.5. We first investigate the false negative rates from the comparison methods. The results under different data generations (quasi-Poisson or negative binomial) and different overdispersion levels are shown in Figs 2 and 3, which suggest that

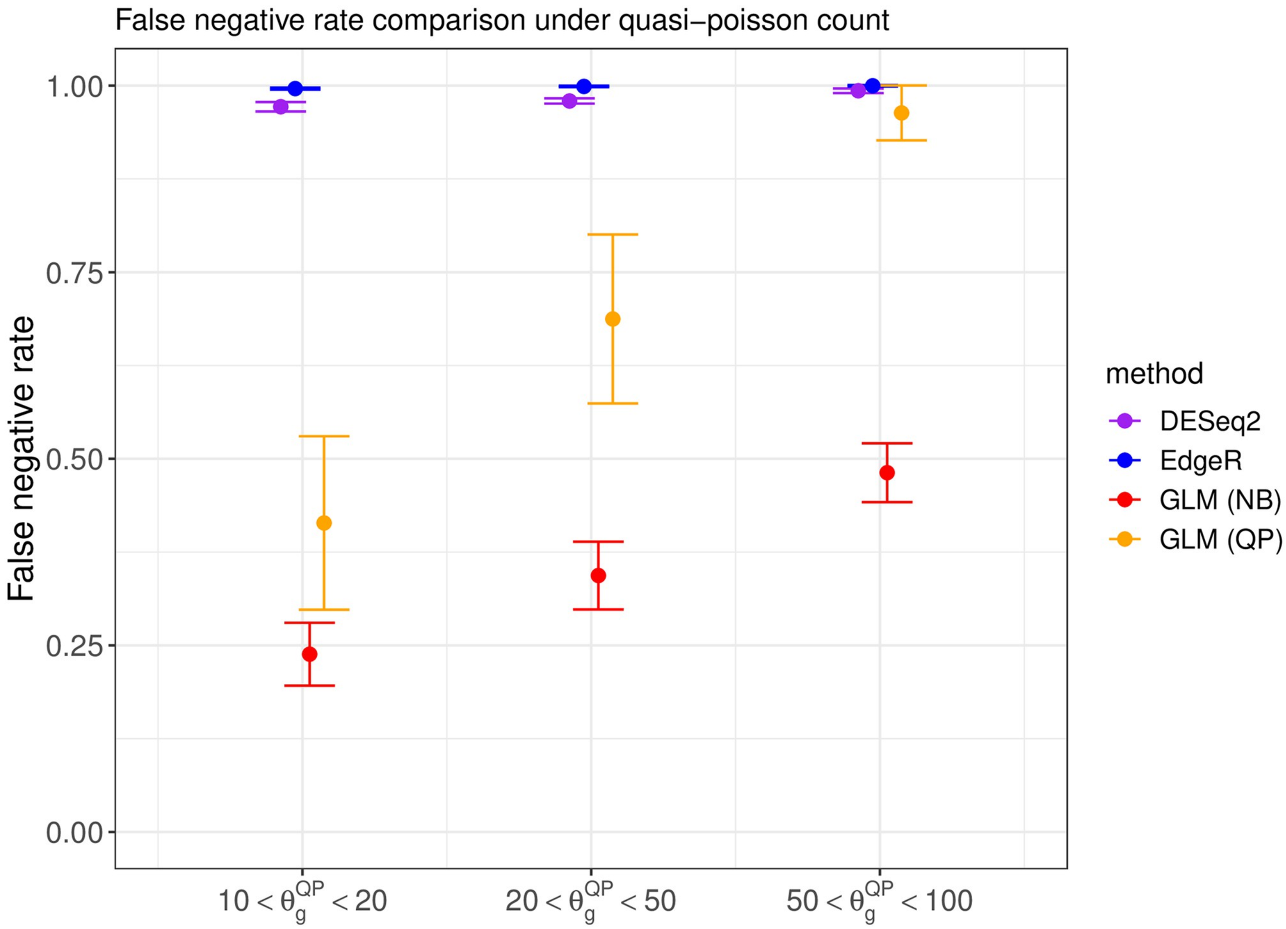

Fig 6. The false negative rate from different methods when the read count follow the quasi-Poisson distribution with different overdispersion levels θ^{QP}. The variance of FNR obtained from repeated experiments is illustrated using the bars.

https://doi.org/10.1371/journal.pone.0300565.g006

the proposed DEHOGT method reaches the lowest false negative rate over competing methods under different overdispersion levels, indicating that most of the genes selected by the proposed method are differentially expressed. Note that the DEHOGT (NB) under the true negative binominal setting always achieves the lowest false negative rate when the cutoff of the adjusted p-value is set as 0.05. This is because the p-values from DEHOGT under NB tend to be smaller than for DEHOGT under QP. The better performance of DEHOGT under QP for the ROC and AUC (area under the ROC curve) implies that we can select a p-value cutoff larger than 0.05, under which the false negative rate of DEHOGT under QP can be smaller than the false negative rate of DEHOGT under NB. Notice that we select DE genes when the corresponding p-value is smaller than a specified type I error rate where the gene-wise p-values are adjusted by the Benjamini-Hochberg procedure [25]. Therefore, the false positive rate from the proposed method is controlled at the same level as EdgeR and DESeq2 in comparing their identification power.

We also investigate the overall DE gene discriminative power of different methods when the cutoff point of the adjusted p-value changes over the range from 0 to 1, as measured by the AUC (area under the ROC curve). Note that the AUC value is between 0 and 1, and a larger AUC value indicates that the algorithm can achieve an overall lower false positive rate and

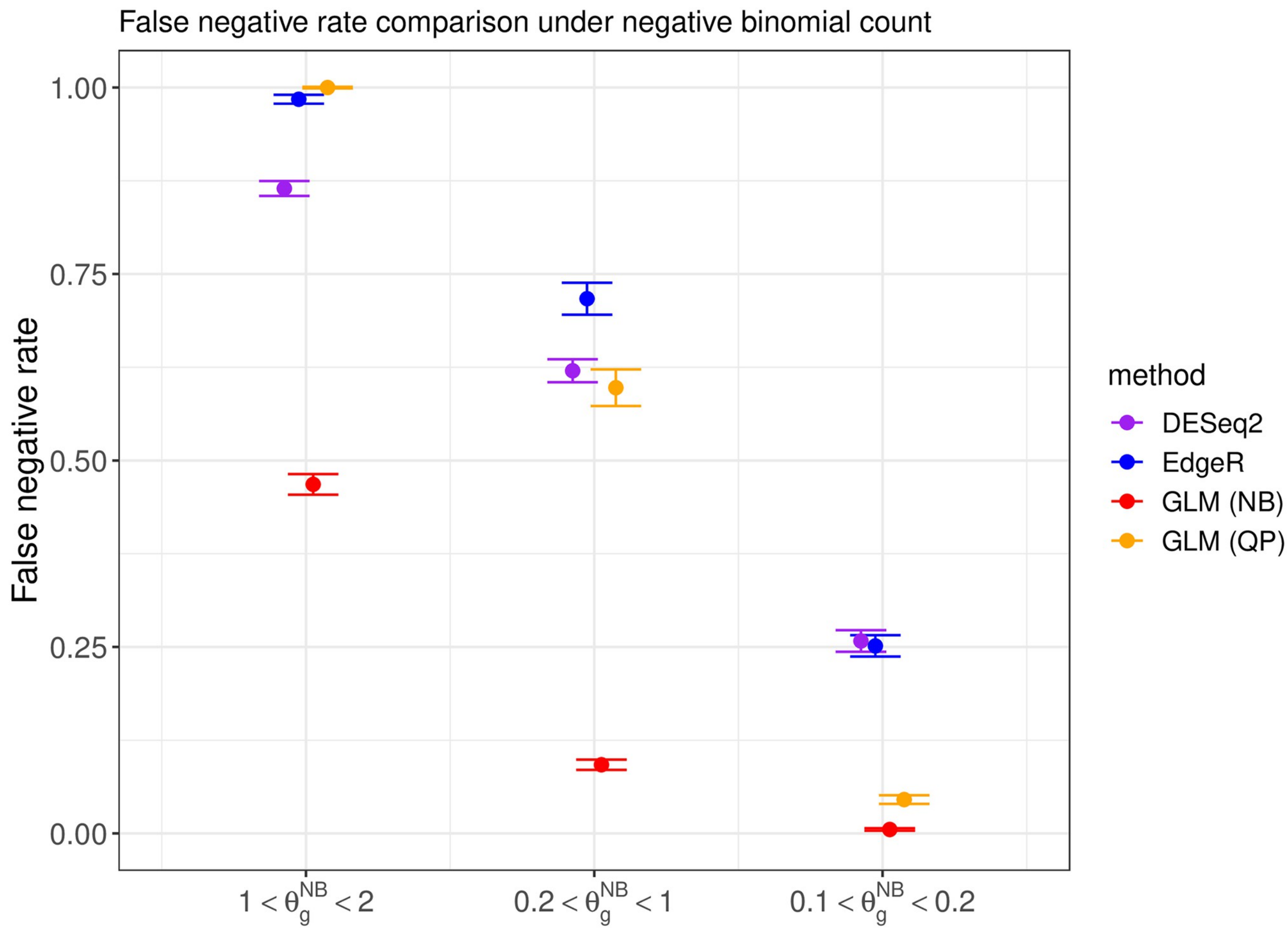

Fig 7. The false negative rate from different methods when the read count follows a negative binomial distribution with different overdispersion levels θ^{NB}.

https://doi.org/10.1371/journal.pone.0300565.g007

lower false negative rate simultaneously. The comparisons are shown in Figs 4 and 5, illustrating the AUC values for competing methods under different generating distributions and overdispersion levels.

The above results indicate that the proposed DEHOGT method outperforms both the DESeq2 and edgeR methods, and the proposed method can achieve the optimal AUC if the model is correctly specified. Specifically, DEHOGT under QP attains a higher AUC than DEHOGT under NB under varying θ_g^{QP} when the read counts are generated from the quasi-Poisson distribution. Similarly, DEHOGT (NB) attains higher AUC then DEHOGT (QP) under varying θ_g^{NB} if the read counts are generated from negative binomial distributions. Notice that the difference in average read counts between the treatment group and the control group is relatively small in this setting, which leads to a weak signal for detecting differentially expressed genes from different treatments. Therefore, this weak signal setting is fundamentally difficult for all differentially expressed gene detection methods to identify, which results in poor performance for both proposed method and the comparison methods.

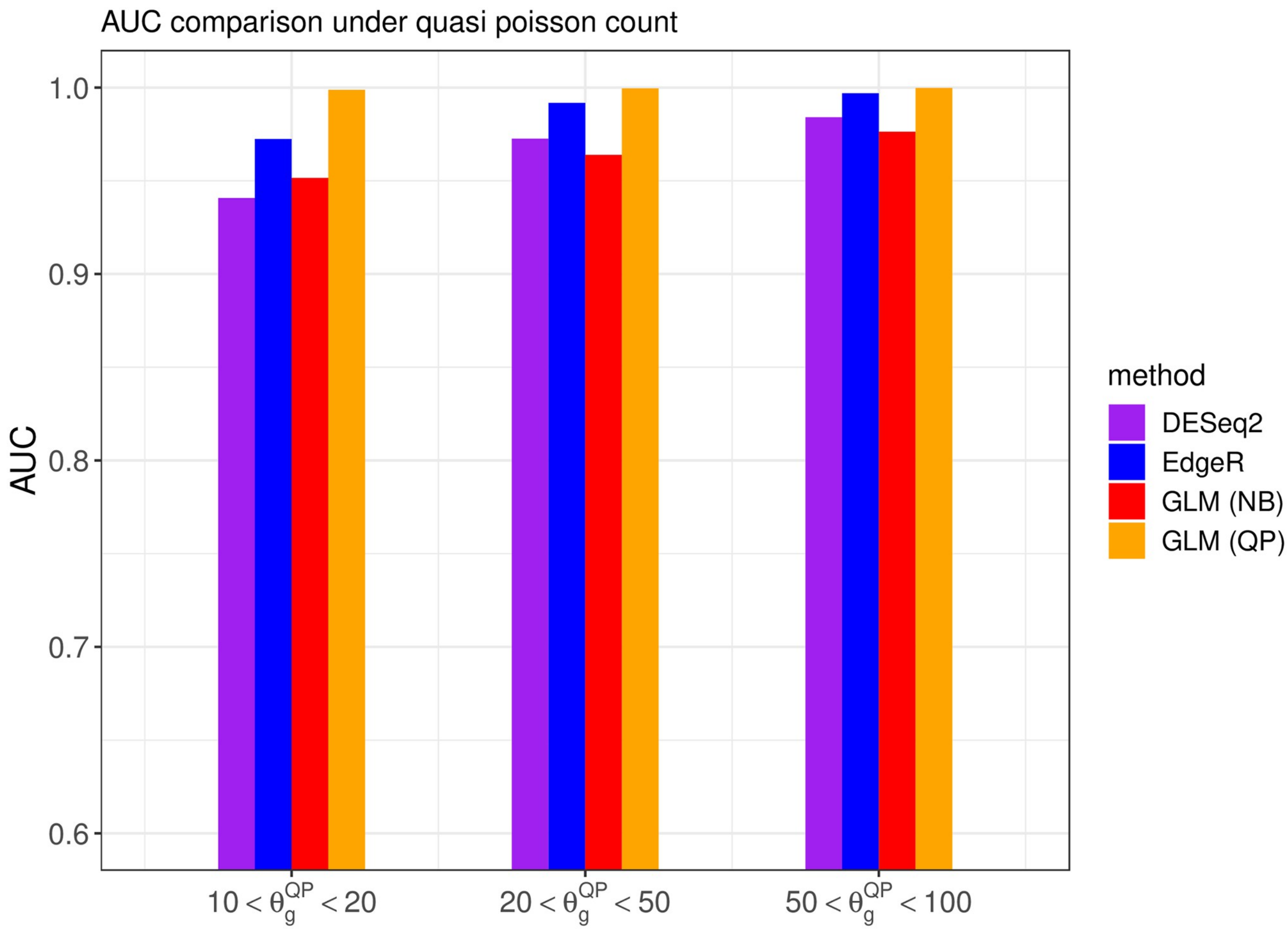

Fig 8. The AUC from different methods when the read count follows the quasi-Poisson distribution with different overdispersion levels θ^{QP}. The variance of FNR obtained from repeated experiments is illustrated using the bars.

https://doi.org/10.1371/journal.pone.0300565.g008

Read count with high discrepancy of expression level

In the second simulation setting, we simulate the read count data of the moderate overdispersion level in RNAseq read counts. Following the notations in simulation 1, we simulate the read count data from both the quasi-Poisson distribution and the negative binomial distribution as

$$\begin{aligned} Y_{gs} &\sim QP(\text{ mean } = \mu_{gs}, \text{var} = \mu_{gs}\theta_g^{\text{QP}}),\ g = 1, \cdots, N,\ s = 1, \cdots, |S|, \\ Y_{gs} &\sim NB(\text{ mean } = \mu_{gs}, \text{var} = \mu_{gs}(1 + \mu_{gs}/\theta_g^{\text{NB}})),\ g = 1, \cdots, N,\ s = 1, \cdots, |S|, \end{aligned}$$

where we choose $N = 10,000$ and $S = S_1 \cup S_2$, $|S_1| = |S_2| = 6$. The GE genes are randomly selected and $|G^{\text{DE}}| = 2,000$. We consider three different overdispersion levels for the read counts from the quasi-Poisson distribution as

$$\theta_g^{QP} \sim \text{Unif}(50, 100),\ \theta_g^{QP} \sim \text{Unif}(20, 50),\ \theta_g^{QP} \sim \text{Unif}(10, 20).$$

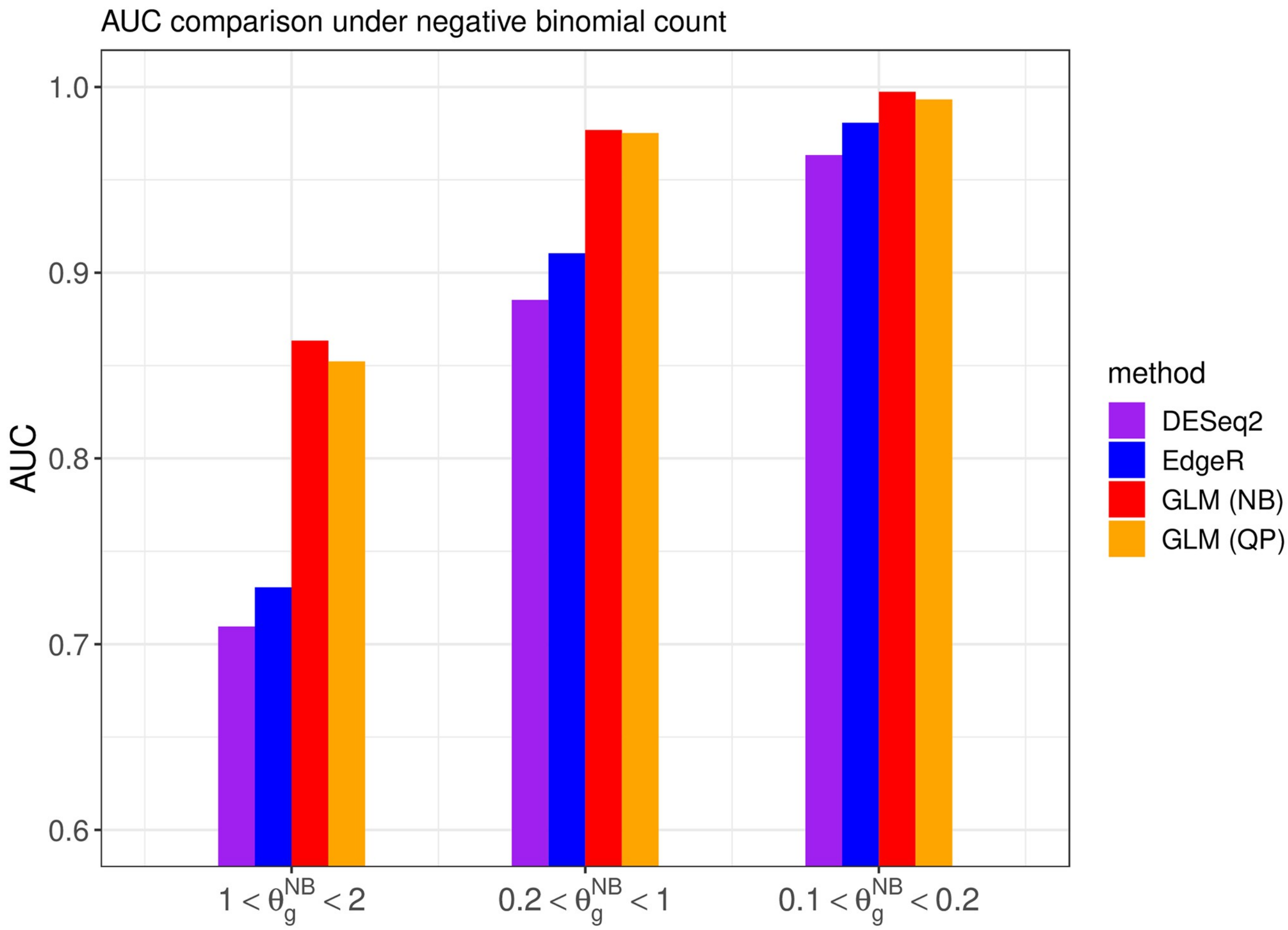

Fig 9. The AUC from different methods when the read counts follow the negative binomial distribution with different overdispersion levels θ^{NB}.

https://doi.org/10.1371/journal.pone.0300565.g009

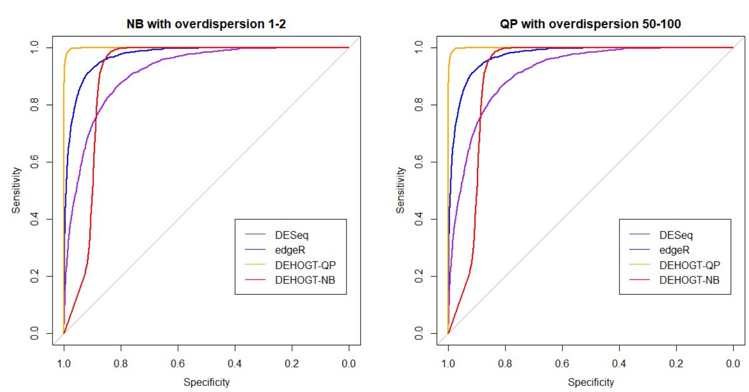

Fig 10. The ROC curve from different methods when the read counts follow the negative binomial distribution with $\theta_g^{NB} \in (1, 2)$ and quasi-Poisson distribution with $\theta_g^{QP} \in (50, 100)$.

https://doi.org/10.1371/journal.pone.0300565.g010

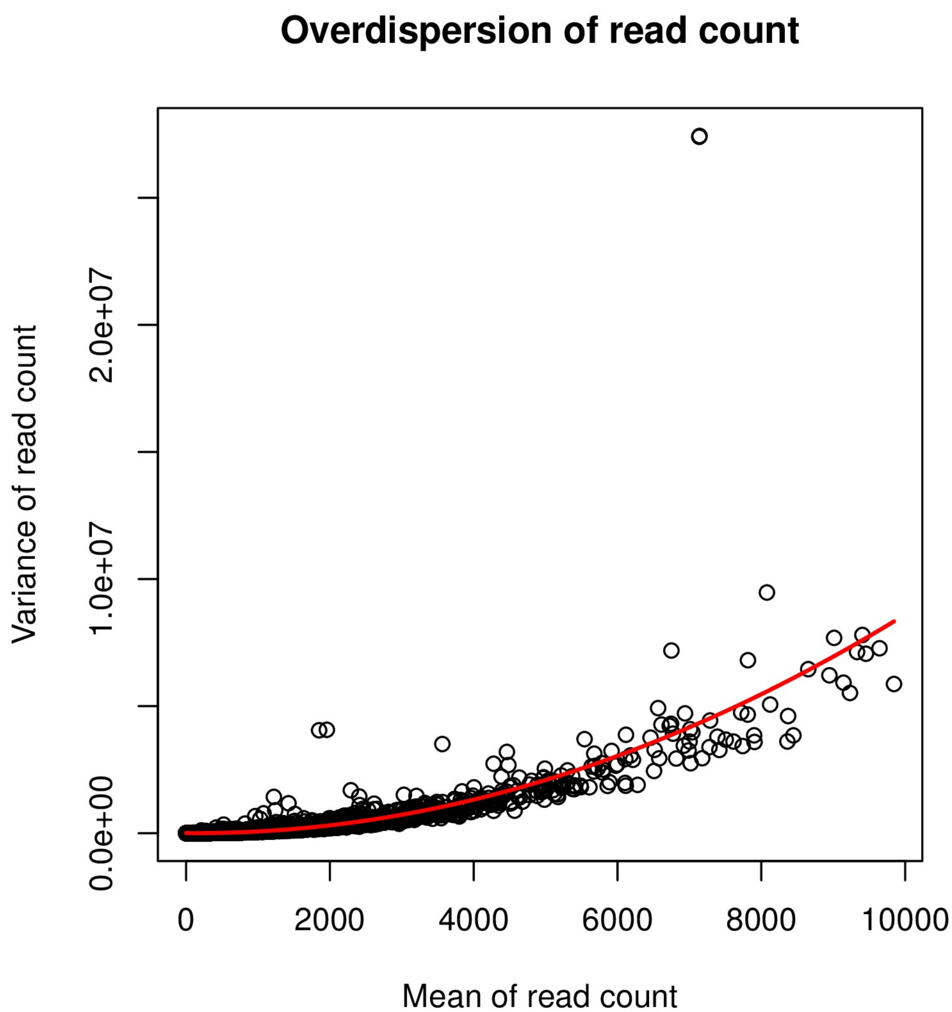

Fig 11. The dispersion of genewise RNA read counts. Each dot corresponds to a sample count from a specific gene.

https://doi.org/10.1371/journal.pone.0300565.g011

Similarly, we also consider three overdispersion levels under negative binomial read counts as

$$\theta_g^{NB} \sim \text{Unif}(0.1, 0.2),\ \theta_g^{NB} \sim \text{Unif}(0.2, 1),\ \theta_g^{NB} \sim \text{Unif}(1, 2).$$

We differentiate DE genes and non-DE genes with different sample means such that

$$\mu_{gs} \sim \quad \text{Unif}(1, 500),\ s \in S,\ g \notin G^{DE},$$

$$\mu_{gs} \sim \begin{cases} \text{Unif}(1, 500),\ s \in S_1, g \in G^{DE} \\ \text{Unif}(1, 500) + \lceil 3.5 \times \bar{\mu}_{gS_1} \rceil,\ s \in S_2, g \in G^{DE} \end{cases}$$

where $\lceil \cdot \rceil$ is the ceiling function, and $\bar{\mu}_{gS_1} = \frac{1}{|S_1|} \sum_{s \in S_1} \mu_{gs}$.

Notice that the expression discrepancy between the treatment and control group is strong for DE genes, while the average expression levels for both groups are low. To select the DE genes, we follow the selection criterion in the previous simulation such that the absolute value of log2fold change is larger than 1.5 and the adjusted p value is smaller than 0.05. We first

Table 1. The number of selected DE genes from microglia RNA-seq data under different treatment comparisons.

	Treatment Pairs						
Methods	**dexh3 vs control**	**dexh6 vs control**	**corth3 vs control**	**corth6 vs control**	**dexvh3 vs dexh3**	**dexvl3 vs dexl3**	**cortvh3 vs corth3**
DESeq2	221	45	166	1	255	118	112
EdgeR	419	115	308	3	469	180	216
DEHOGT	981	237	355	582	383	148	256

https://doi.org/10.1371/journal.pone.0300565.t001

investigate the false negative rates from different methods, and the results are illustrated in Figs 6 and 7.

The numerical results illustrates that the proposed method DEHOGT has a lower false negative rates than DESeq2 and EdgeR under different read count generation distributions and different overdispersion levels. Specifically, when the read count distribution is correctly specified, our method consistently achieves lower false negative rate than the EdgeR and DESeq2. More importantly, the improvement from the DEHOGT increases as the degree of overdispersion in the read count increases for both quasi-Poisson and negative binominal distributions.

We also investigated the overall discriminative power of the DE gene using different methods when the adjusted p-value cutoff varies between 0 and 1 instead of using 0.05. The overall classification performance is measured by the AUC. The Figs 8 and 9 illustrate the AUC from competing methods under different settings of read counts.

The above results show that the proposed DEHOGT method achieves a higher AUC in detecting the DE genes than the DESeq2 and EdgeR, indicating that our method offers a better balance between decreasing false positive rate and false negative rate. In addition, the

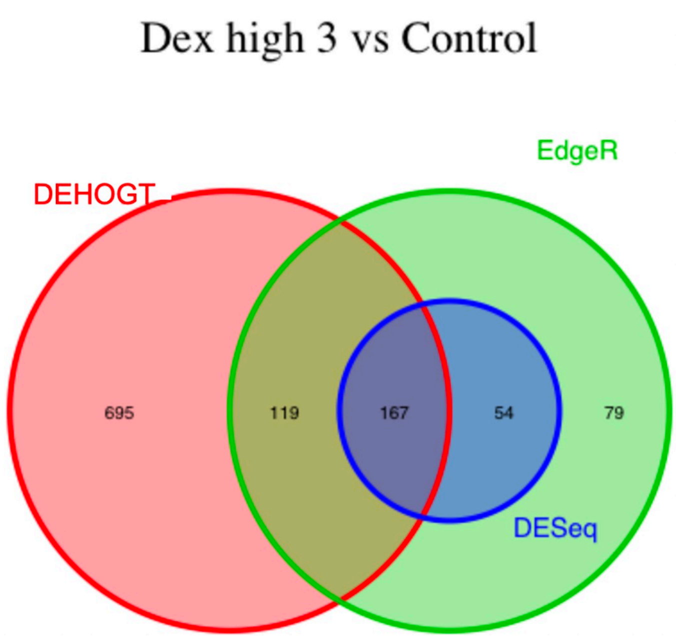

Fig 12. The selected DE genes from DEHOGT, DESeq2, EdgeR under treatment comparison dexh3 versus control.

https://doi.org/10.1371/journal.pone.0300565.g012

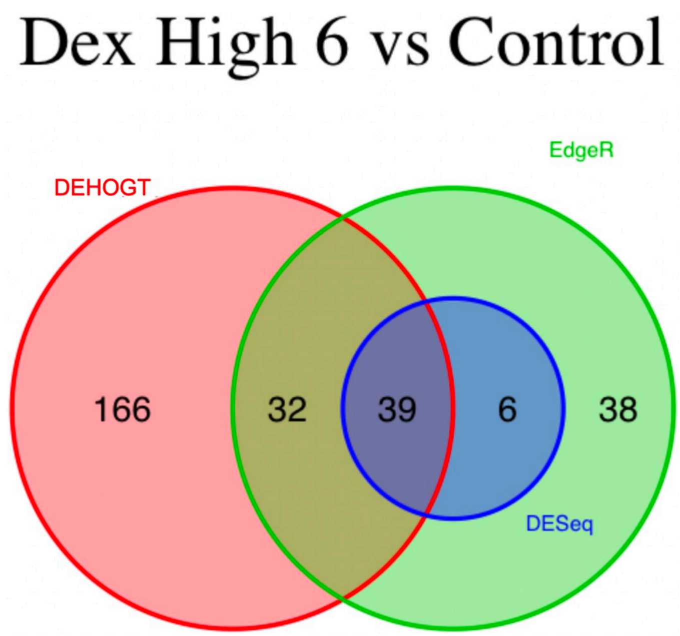

Fig 13. The selected DE genes from DEHOGT, DESeq2, EdgeR under treatment comparison dexh6 versus control.

https://doi.org/10.1371/journal.pone.0300565.g013

improvement from our method is more significant as the overdispersion level increases, which is consistent with the aforementioned false negative rate comparison. A higher AUC from the DEHOGT method also implies that it can be more robust against the selection of different cut-off of p-value for DE genes.

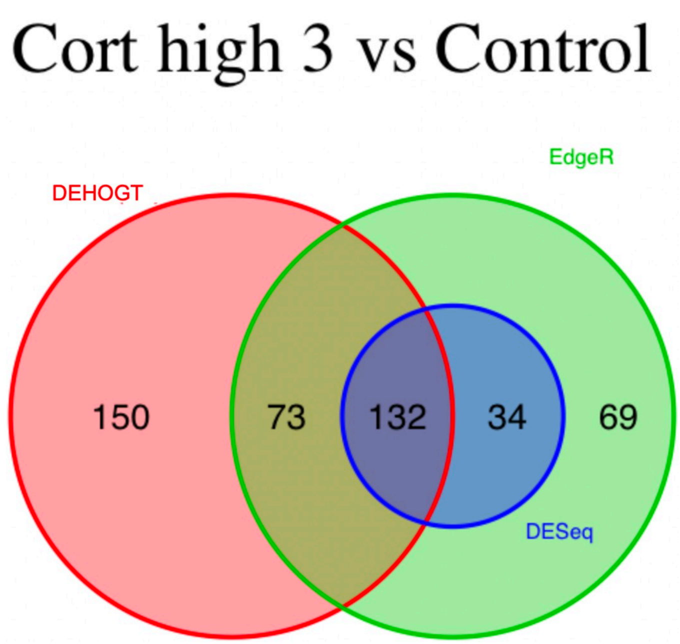

Fig 14. The selected DE genes from DEHOGT, DESeq2, EdgeR under treatment comparison cortvh3 vs corth3.

https://doi.org/10.1371/journal.pone.0300565.g014

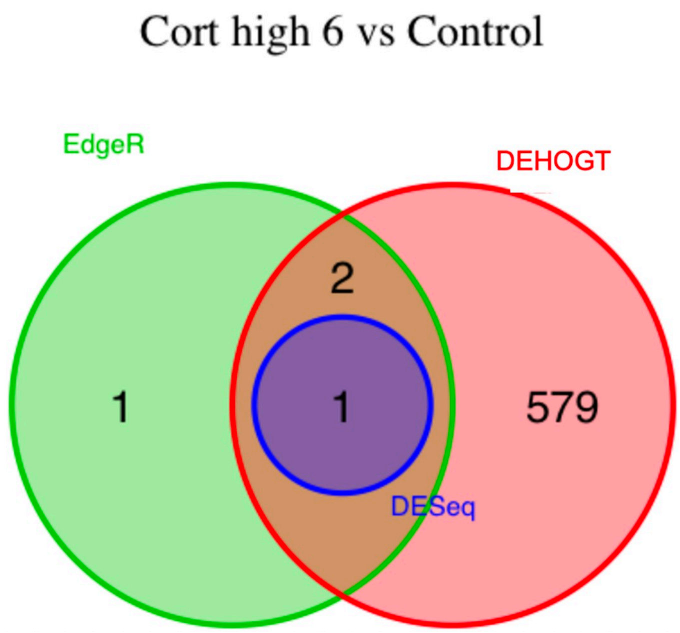

Fig 15. The selected DE genes from DEHOGT, DESeq2, EdgeR under treatment comparison corth6 versus control.

https://doi.org/10.1371/journal.pone.0300565.g015

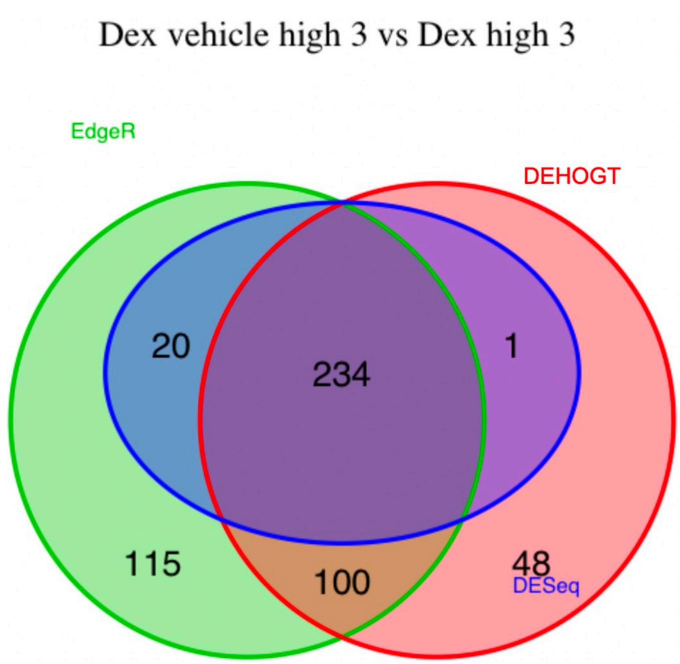

Fig 16. The selected DE genes from DEHOGT, DESeq2, EdgeR under treatment comparison dexvh3 versus dexh3.

https://doi.org/10.1371/journal.pone.0300565.g016

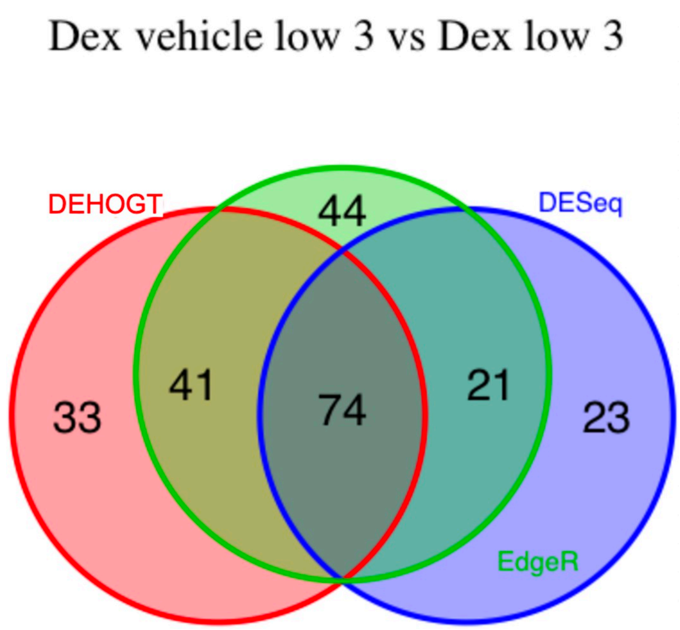

Fig 17. The selected DE genes from DEHOGT, DESeq2, EdgeR under treatment comparison dexvl3 vs dexl3.

https://doi.org/10.1371/journal.pone.0300565.g017

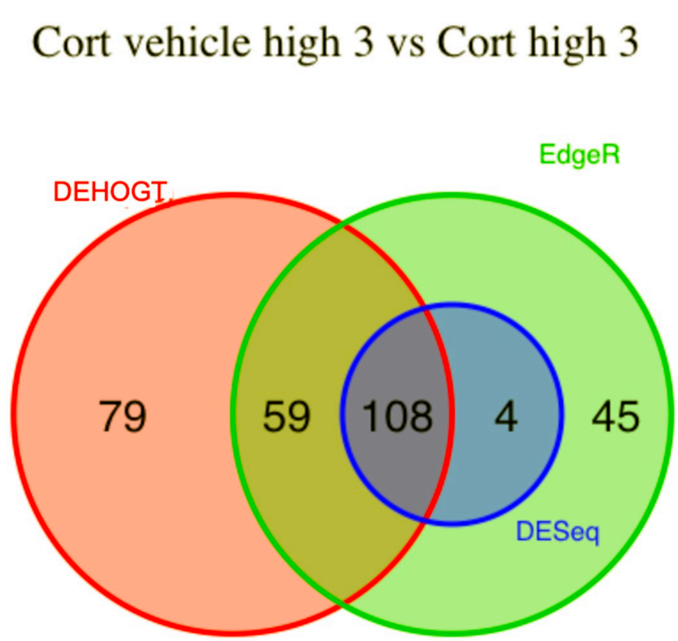

Fig 18. The selected DE genes from DEHOGT, DESeq2, EdgeR under treatment comparison corth3 versus control.

https://doi.org/10.1371/journal.pone.0300565.g018

Table 2. Top 15 unique DE genes unique selected by DEHOGT, EdgeR, and DESeq2.

	Methods		
Treatments	DEHOGT	EdgeR	DESeq2
dexh3 vs control	BACE1, H19, LOC102724852, VSIR, SLC4A4	PSG11, WNT7B,KLF15, SNORA74A, IGFN1	
	ITGB3, LTBP2, ELF1, EPS8, DUBR	TNFRSF11B, EPB41L3, LCT-AS1, MYO5B, MMP1	
dexh6 vs control	SLC16A12, GCNT1, SLC7A2, FGD4, AOX1	AMIG02, SRPX2, PADI1, STAMBPL1, SHC3	
	DUBR, DGKI, NFKBIA, EMILIN3, KCNIP3	KCNJ8, CYP24A1, KRT18P29, FAP, SOX3	
	MARCHF1, TXNIP, IGFBP7, EPSTI1, PLXNA2	LTBR, SLC28A3, TRHDE, LINC00402, MY05B	
corth3 vs control	PRKACB, ABCB1, KLF9, SLC16A12, LTBP1	RGMB, CTSC, FGF1, PSG4, ST6GAL2	
	UGT2B7, MARCHF1, LAMA5, H19, LOC102724852	FOSL1, SERPINE2, TRPC4, CREB5, NPPB	
	KCNIP3, COL5A1, NFKBIA, MEGF6, FGD4	ADAMTS1, ADAMTSL1, NLRP10, EVA1A, AGTR1	
corth6 vs control	ROR1, NID2, H3–2, ACTBL2, CHST2	NHS	
	EMILIN3, INAVA, RPEL1, PTGER2, MROH6		
	BBS12, CAMSAP3, TYSND1, TMEM116, MRPL38		
dexh3 vs dexvh3	GCNT1, ABCB1, SLC26A2, ITGB3, LTBP2	MAP3K7CL, QSOX1, SMURF2, SLC8A1, CLDN11	
	IGFBP7, COL5A1, DUBR, FKBP5, ADAMTS7	KHDRBS3, CREB5, GPRC5A, NR2F2, KRT17	
	EFEMP1, PLXNA2, CPM, LPIN3, BIRC3	ADAMTSS,NR2F2-AS1,HIF1A-AS3,SRPX2, KCNMA1	
dexl3 vs dexvl3	AGTR1, SCN9A, ABCB1, ELOVL6, SPSB1	PLAUR, KRT17, AJAP1, MARCH, COL13A1	TGFBI, CPA4, CHST2, DGKI, KRT18P11
	DEPTOR, SPOCK1, NCAM2, CLDN1, KRT18P29	ITGB3, EPSTI1, COL4A4, DNER, TGFB3	SNORA74A, LOC102724434, HNRNPA1P33, KRTAP4–8, RHEX
	ITGA7, USP44, SRP14-DT, DPYD, STARD8	HAS3, ANGPTL4, HCN3, ALPP, DNAH8	SENCR, F8, PALMD, CCL26, KRTAP2–4
corth3 vs cortvh3	SLC26A2, MARCHF1, DOCK4, EIF1B-AS1, CREB5	EFHD1, SLC1A3, SAA1, WFDC21P, ITGA1	
	EVA1A, PDZK1, COL13A1, FRMD6-AS1, FBXW4	CEBPD, GJD2, EGR2, KDR, GSX2	
	CYP24A1, AJAP1, NUPR1, ATP6V1G2, MGST2	CCDC30, ASAH1-AS1, GGTLC3, LINC00886, LOC100506207	

https://doi.org/10.1371/journal.pone.0300565.t002

We also illustrate the ROC curves in Fig 10 for two representative cases where read counts follow the negative binomial distribution with $\theta^{NB} \in (1, 2)$, and the quasi-Poisson distribution with $\theta^{NB} \in (50, 100)$, respectively.

Application on microglia RNA-seq read count data

In this subsection, we apply the proposed DEHOGT method, DESeq2, and EdgeR in the study of post-traumatic stress disorder described in the Introduction section. Specifically, we aim to identify differentially expressed genes from microglia cells that are relevant to the PTSD progress. The RNA-seq data were collected by Uddin research team and Wildman lab at the University of the South Florida. The research performed in-vitro experiments on microglial cells

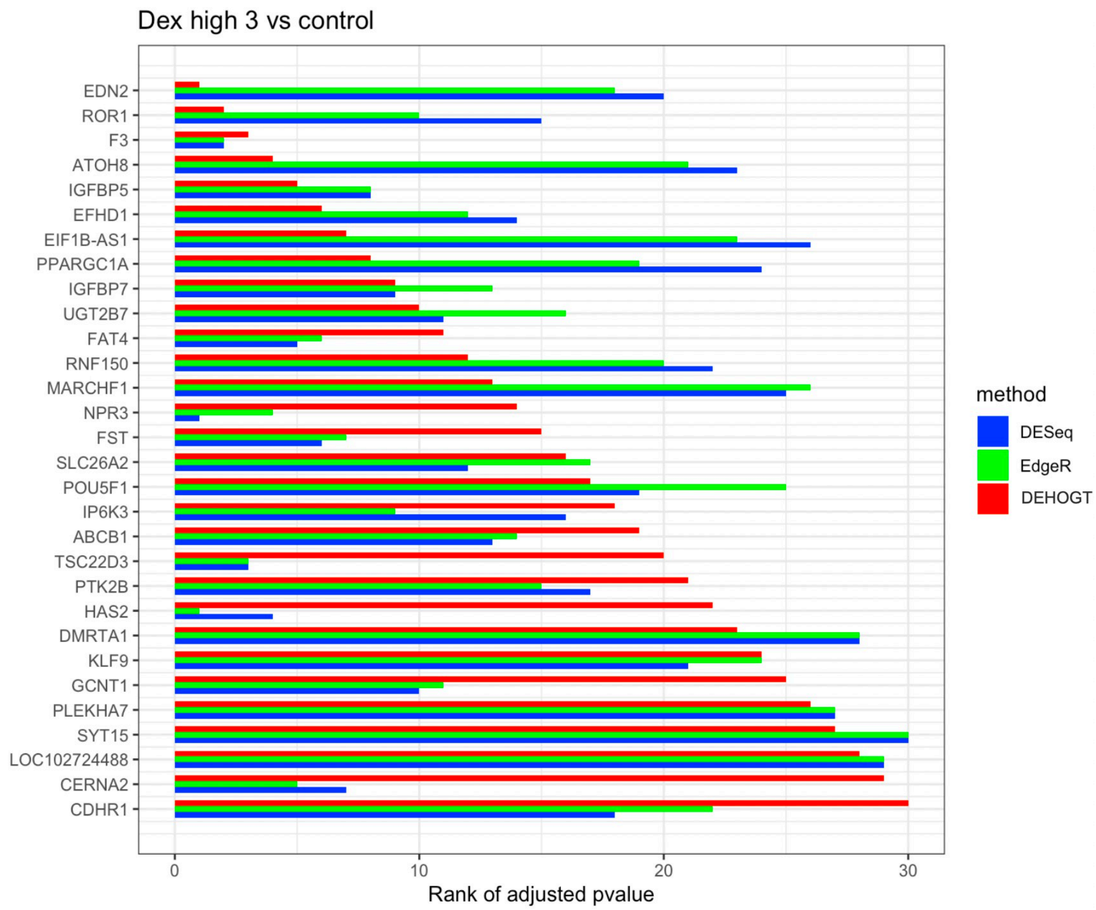

Fig 19. The rank of p-value of selected genes under treatment comparison dexh3 versus control, a shorter bar indicates a smaller p-value (more significantly differently expressed).

https://doi.org/10.1371/journal.pone.0300565.g019

which utilized stress hormones to imitate immune environments similar to PTSD. The function of stress hormones is to adjust the human interior environment, provide energy, and increase heart rate when experience stress [35]. The experiments exposed microglial cells to dexamethasone (dex) and hydrocortisone (cort) serving as stress hormones. The alcohol is also utilized as an additional control treatment to validate if changes in gene expressions are due to the exposure to stress hormones or just a random treatment (alcohol). Specifically, the experiments grew microglial cells under one of the four treatments: hydrocortisone, dexamethasone, alcohol (vehicle), or control. After exposure of three days, RNA-seq data was extracted from the cells on the third day and on the final day of the washout period (day 6), respectively. The goal of study is to identify the genes that are differentially expressed in microglia cells when exposed to different hormones and to determine if the dose of the hormone affects gene expression levels.

More specifically, there are a total of 20,052 expressed genes after quality control preprocessing. There is a total of 9 different treatments with the combination of media (dex, cort, vehicle, and control) and dosage (low and high): dex high, dex low, cort high, cort low, dex vehicle high, dex vehicle low, cort vehicle high, cort vehicle low, and control. On day 3 (time point 3), three repeated samples are collected under each treatment. On day 6 (time point 6),

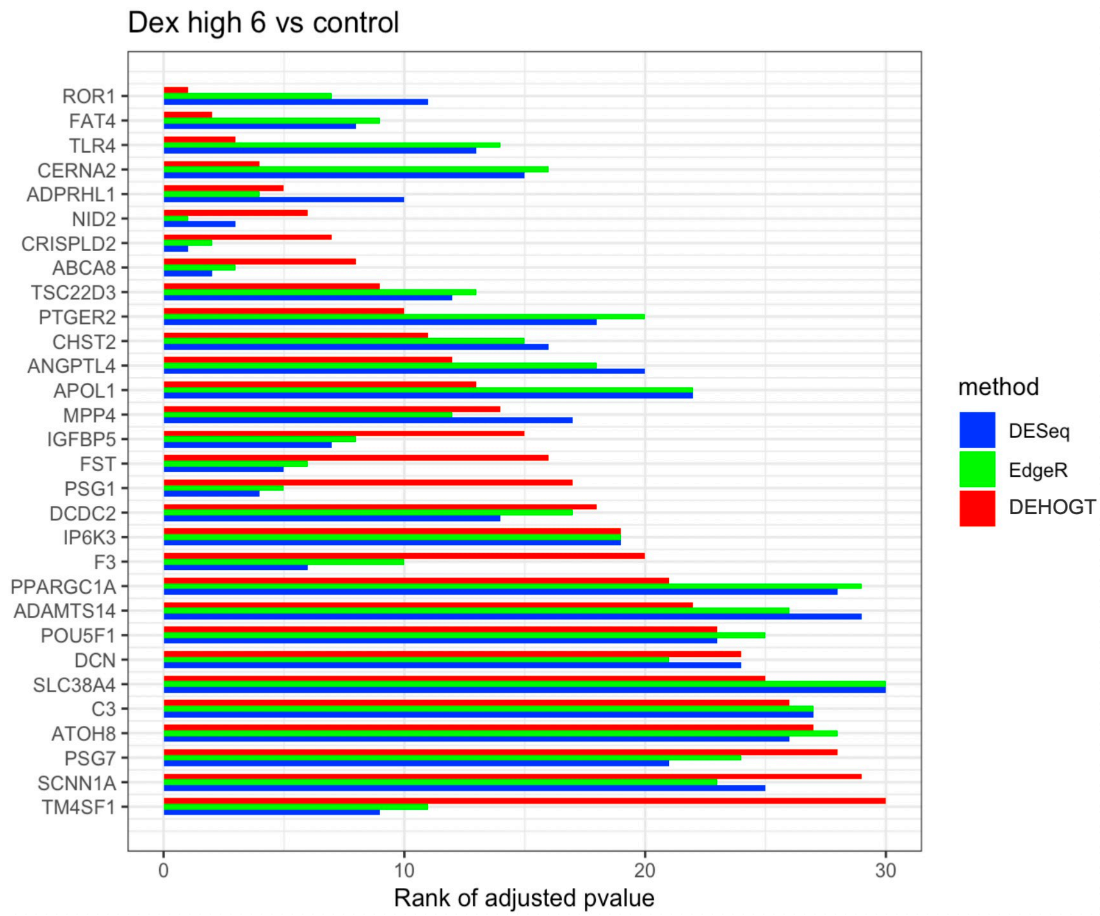

Fig 20. The rank of p-value of selected genes under treatment comparison dexh6 versus control, and shorter bar indicates a smaller p-value.

https://doi.org/10.1371/journal.pone.0300565.g020

three repeated samples are collected under treatments dex high, dex low, cort high, and cort low, and one sample under dex vehicle high, dex vehicle low, cort vehicle high, and cort vehicle low.

We first investigated the level of empirical dispersion in the microglia RNA-seq read counts. Specifically, we examine the relation between sample count mean and sample count variance across all genes. Fig 11 illustrates a quadratic growth of count variance over count mean. In addition, we fit a quadratic regression on count variance over count mean, where an adjusted R^2 coefficient reaches 0.66. Therefore, we choose to use a negative binomial distribution as the read counts generating process in the proposed DEHOGT method.

We utilize DEHOGT, DESeq2, and EdgeR to select DE genes under the following 7 treatment comparison pairs: dex high at time point 3 and control (dexh3 vs control), dex high at time point 6 and control (dexh6 vs control), cort high at time point 3 and control (corth3 vs control), cort high at time point 6 and control (corth6 vs control), dex vehicle high and dex high at time point 3 (dexvh3 vs dexh3), dex vehicle low and dex low at time point 3 (dexvl3 vs dexl3), cort vehicle high and cort high at time point 3 (cortvh3 vs corth3). In selecting DE genes between the two treatments, we follow the criterion in Section 2 in that the adjusted p value is smaller than 0.05, and the log2fold change is larger than 1.5.

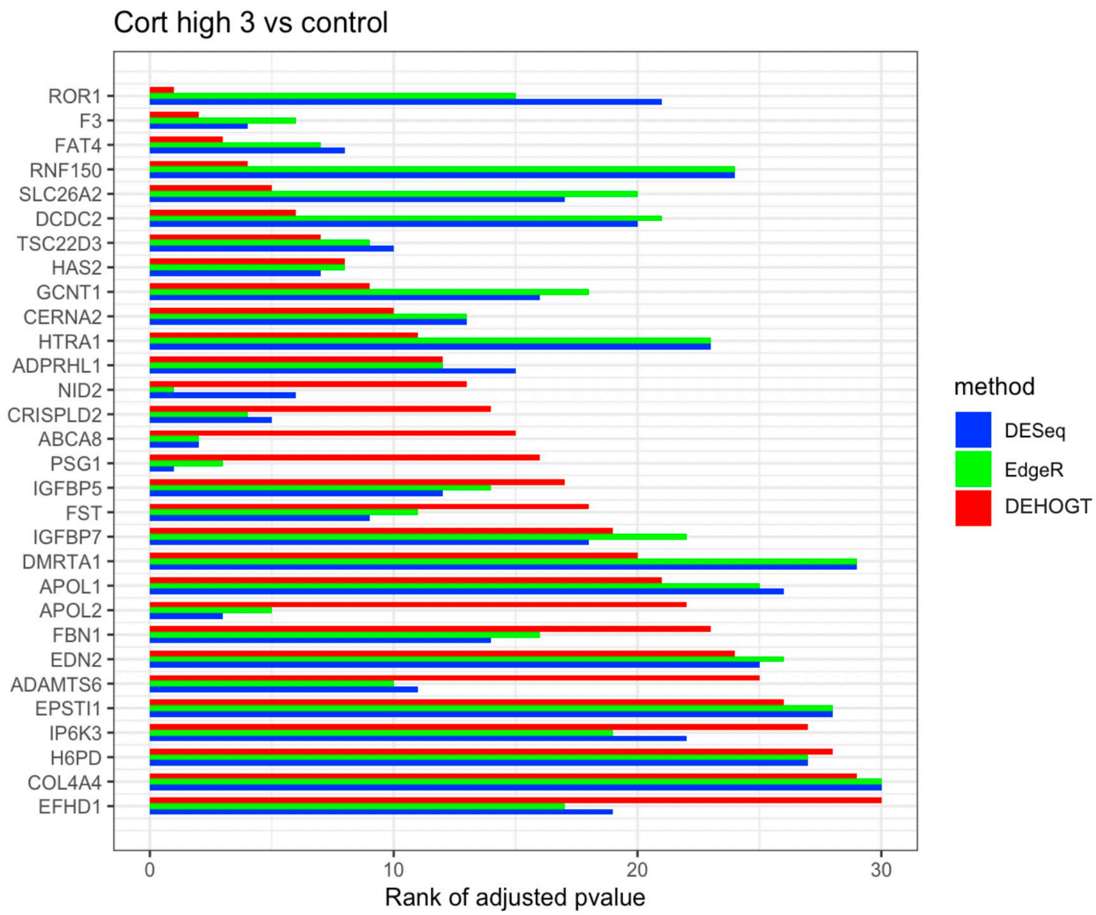

Fig 21. The rank of p-value of selected genes under treatment comparison corth3 versus control, and shorter bar indicates a smaller p-value.

https://doi.org/10.1371/journal.pone.0300565.g021

We first illustrate the number of DE genes selected by competing methods. Table 1 shows that the proposed method tends to select more genes than the other two methods, especially compared to the DESeq2. In the exploratory stage, it is critical to include as many relevant genes as possible for the downstream analysis. The DEHOGT method is more effective in reducing the false negative rate in detecting PTSD-related genes by identifying a larger candidate pool of DE genes.

We conduct detailed analysis for the DE genes based on three methods for each treatment pair. In general, we investigate the overlapping in DE genes from three methods, where the findings are illustrated via the Venn diagram in Figs 12–18. Notice that the proposed DEHOGT method selects more DE genes than DESeq2 and EdgeR for all pairwise comparisons between treatments except dexvh 3 vs dexh 3 and dexvl 3 vs dexl 3, demonstrating that the proposed method can identify more DE genes to reduce the potential risk of missing underlying relevant genes. In the following, we provide an interpretation for the treatment pair dexh6 versus control. The interpretation of other treatment pairs can be conducted similarly. The Venn diagram in Fig 13 shows that all the DE genes selected by the DESeq2 are also selected by EdgeR, and 86.7% of the DE genes selected by DESeq2 are also selected by DEHOGT. In addition, 61.7% of the DE genes selected by EdgeR are detected by DEHOGT.

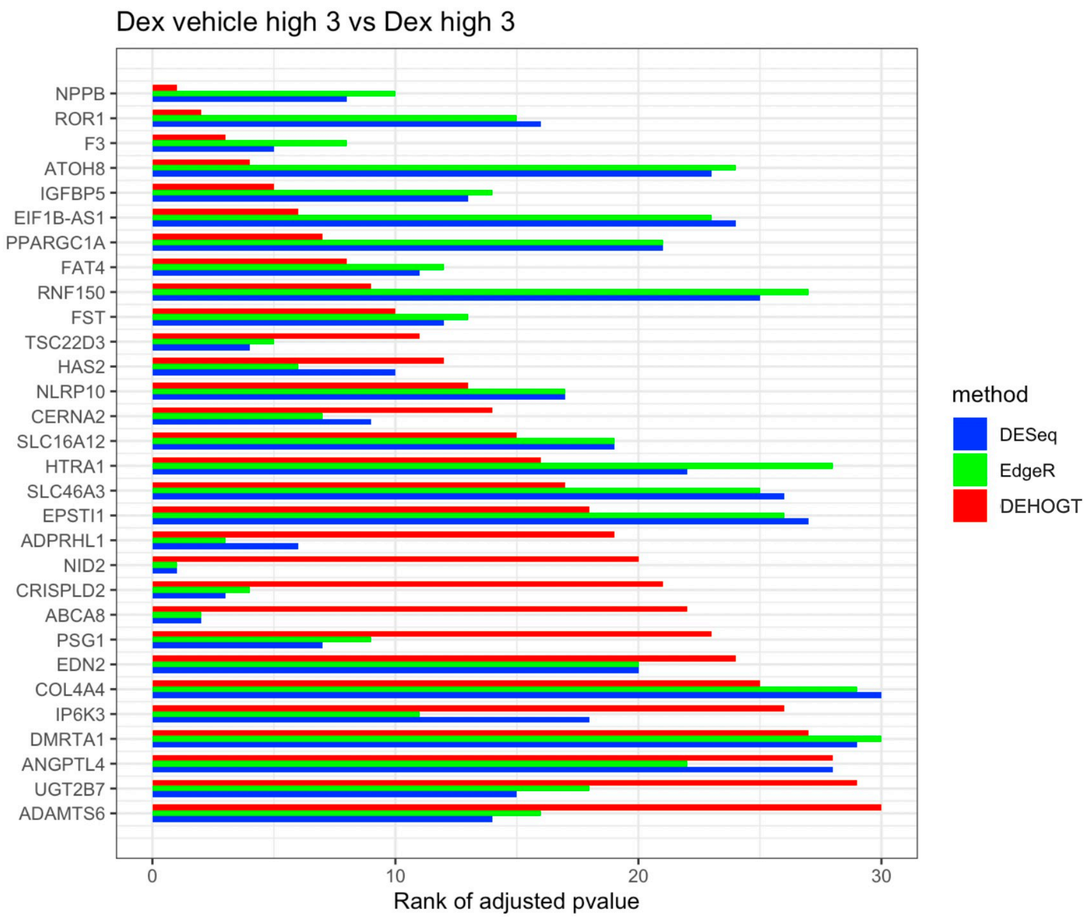

Fig 22. The rank of p-value of selected genes under treatment comparison dexvh versus dexh, and shorter bar indicates a smaller p-value.

https://doi.org/10.1371/journal.pone.0300565.g022

As illustrative examples for interpreting the results of analyzing Microglia RNA-seq data, we select three genes *CRISPLD2*, *TSC22D3*, and *PSG1* which are differentially expressed under the three treatment comparisons: dexvh 3 versus dexh 3, dexvl 3 versus dexl 3, and cortvh3 versus corth3. Specifically, the glucocorticoid-responsive gene *CRISPLD2* is found to be differentially expressed in read counts from an RNA-seq experiment with muscle cells exposed to dexamethasone [36]. The another glucocorticoid-responsive gene *TSC22D3 (GILZ)* is found to be differentially expressed under gonorrhea or chlamydia exposure based on many animal and human gene studies that examine different cell types [37, 38]. These evidences support the fact that *TSC22D3* serves as a mediator for the anti-inflammatory activity of gonorrhea or chlamydia summarized in [39]. The gene *PSG1* is found to activate the underlying beta 1 (TGF-β1) known as transforming growth factor, which is an essential cytokine process in suppression and immunoregulation of inflammatory T cells [40, 41].

We also list the significant DE genes uniquely selected by the three methods in Table 2, which demonstrates that most of the DE genes identified by DESeq2 are also selected by DEHOGT and EdgeR. Specifically, the gene *FKBP5* is identified by the proposed method but not identified by the other methods under the comparison dexh3 versus dexvh3. The gene *FKBP5* is a co-chaperone adjust the activity of glucocorticoid receptor. *FKBP5* is known as an

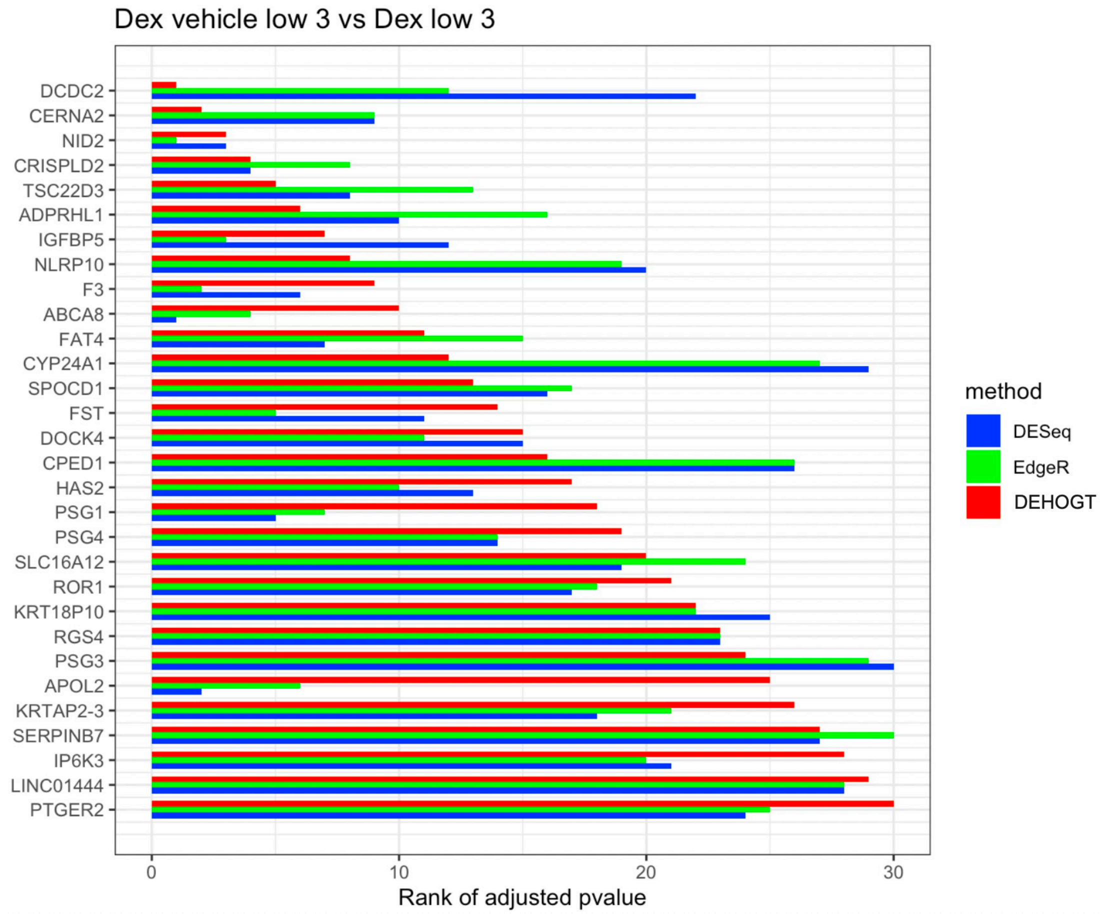

Fig 23. The rank of p-value of selected genes under treatment comparison dexvl versus dexl, and shorter bar indicates a smaller p-value.

https://doi.org/10.1371/journal.pone.0300565.g023

important modulator of responding stress. In many studies using different cell types, the dysregulation phenomenon of *FKPB5* is found in many stress-related psychopathologies via investigating single nucleotide polymorphisms [42, 43], gene expression [44], and DNA methylation profiles [45].

In addition, we examine the most significant DE genes among the overlaps of the three methods in Figs 19–24. For treatment pair dexh6 versus control, Fig 20 lists the 30 most significant DE genes which are overlapping for all three methods, and the bar charts with different colors represent the rank of p-values from the three methods. A shorter bar indicates a smaller p-value and therefore a more significantly differentially expressed genes under dex high and control comparison. The DEHOGT method selects genes *ROR1, FAT3, TLR4, CERNA2, ADPRHL1, NID2, CRISPLD2*, and *ABCA8* as the top 8 significant DE genes, and these genes are also among the top significant DE genes selected by EdgeR and DESeq2. In general, our method provides a list of the top significant DE genes which is consistent with the DESeq2 and EdgeR in comparing dexh6 versus control. This similarity of results of the three methods confirms the association between PTSD and the top DE genes which are identified by the DEHOGT. In particular, the previously mentioned genes *TSC22D3* and *PSG1* are identified by all three methods for the vehicle treatment comparisons: dexvh 3 versus dexh 3, dexvl 3 versus

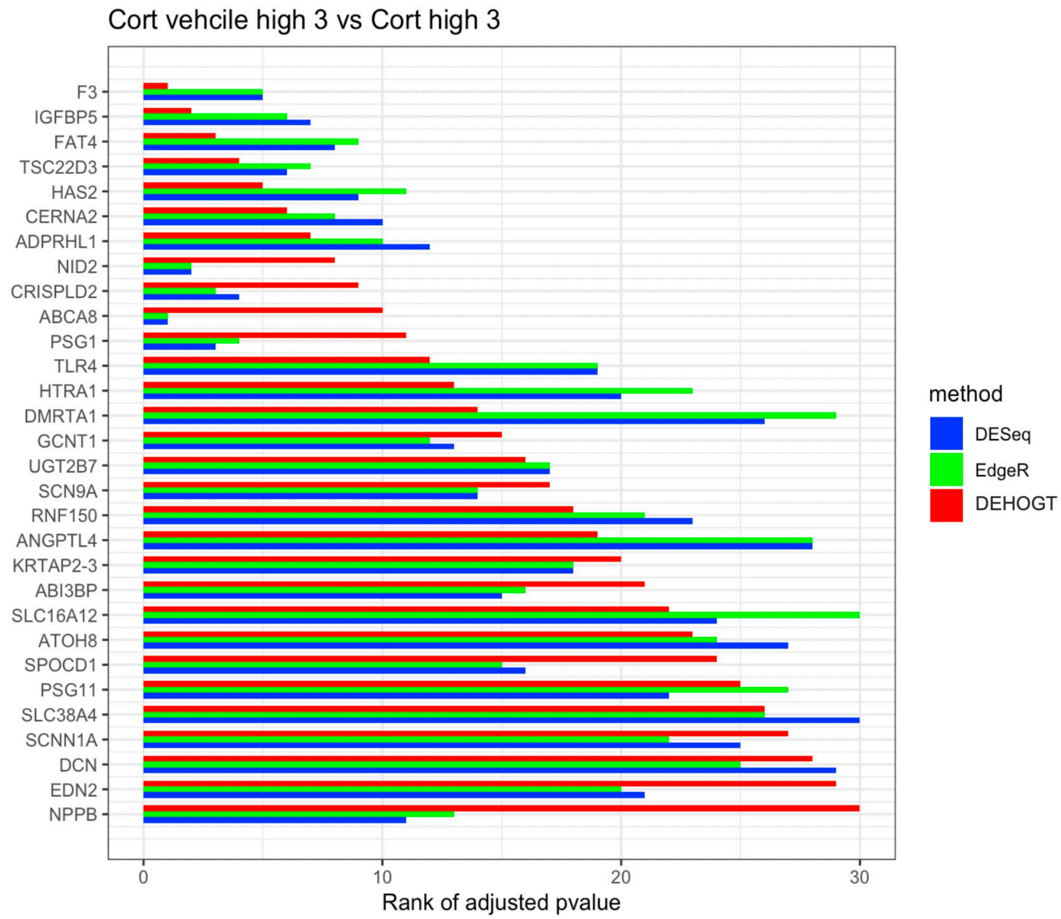

Fig 24. The rank of p-value of selected genes under treatment comparison cortvh versus corth, and shorter bar indicates a smaller p-value.

https://doi.org/10.1371/journal.pone.0300565.g024

dexl 3, and cortvh3 versus corth3. These results provide evidence of further need to explore their roles in the formulation of PTSD.

Discussion

In this paper, we propose a new differential expression analysis procedure of RNA-seq data based on generalized linear modeling. In our simulation study, we demonstrate that the proposed method achieves better performance in detecting DE genes compared with the EdgeR and DESeq2 methods, especially when the per-treatment sample size is relatively small. The numerical experiments suggests that DEHOGT is less conservative in selecting DE genes due to adopting the individual fitting procedure. This property enables our method to have improved performance in controlling the false negative rate which is more critical for downstream analysis.

We further apply our method and compare it with EdgeR and DESeq2 on a real application in a microglia RNA-seq dataset collected by our team. Specifically, our method identifies more potential genes which may be potentially more relevant to PTSD than either EdgeR and DESeq2. In addition, the cross-validation among EdgeR, DESeq2 and the proposed method provides a rich and robust candidate pool for genes relevant to PTSD. These results were obtained in the microglia dataset despite having issues of overdispersion and small sample size.

The popular existing methods DESeq2 and EdgeR identify differentially expressed genes by adopting an aggregate estimation strategy for read count overdispersion levels, which relies on the key assumption that genes with similar expression levels have similar overdispersion levels. The numerical results in this paper indicate that this assumption might be questionable under the scenario when heterogeneity of gene expression level is high. The violation of this assumption can undermine the detection power of methods based on aggregate estimators of overdispersion, especially when the overdispersion level is high. In contrast, estimating overdispersion levels for each gene separately can be more robust under high heterogeneity in gene expressions. On the other hand, the proposed independent estimation scheme integrates samples from different treatments instead from different genes, which might lose a certain amount of statistical testing power especially when the sample size is small. One direction worth of further exploration is to incorporate neighborhood similarity structures among genes such that the overdispersion estimation of a specific gene can borrow the information of samples from correlated genes, therefore we can increase the effective sample size for estimating overdispersion levels. A potential strategy could utilize gene-wise covariate variables or develop an adaptive fused-type penalty on gene overdispersion levels.

Conclusion

DEHOGT is a general workflow for identifying differentially expressed genes based on overdispersed RNA-seq read count data. DEHOGT adopts a joint estimation of logfold changes that incorporates samples from all treatments simultaneously to utilize cross-treatment information. In addition, the proposed method takes advantage of within-treatment independence structures among genes to increase the effective sample size, which leads to stronger power in detecting DE genes. Furthermore, our method enjoys flexibility in utilizing different read count generating distributions instead of fixing only one negative binominal distribution as in the popular methods such as EdgeR and DESeq2. This allows us to choose a generating distribution adopted to the empirical dispersion level. Therefore, DEHOGT has the potential to be applied for other genetic datasets with similar challenges of heterogeneous overdispersion levels.

Supporting information

S1 Appendix. Supplementary materials. Numerical comparison with Limma-voom method, application on murine alveolar macrophages dataset, microglia cell experiment design, RNA-seq data preprocessing, and gene ontology analysis on Microglia cell dataset.
(PDF)

Acknowledgments

The authors would like to acknowledge the USF Genomics Core for their support of the microglia cell experiment. The authors would like to acknowledge Editor and reviewers for their suggestions and helpful feedback.

Author Contributions

Conceptualization: Chengqi Wang, Monica Uddin, Derek Wildman, Annie Qu.

Data curation: Agaz Wani, Jan Dahrendorff, Chengqi Wang, Arlina Shen, Janelle Donglasan, Sarah Burgan, Zachary Graham.

Formal analysis: Yubai Yuan, Qi Xu, Arlina Shen, Janelle Donglasan.

Funding acquisition: Monica Uddin.

Investigation: Agaz Wani.

Methodology: Yubai Yuan, Qi Xu.

Resources: Agaz Wani, Sarah Burgan, Zachary Graham.

Software: Qi Xu.

Validation: Jan Dahrendorff.

Visualization: Yubai Yuan, Qi Xu.

Writing – original draft: Yubai Yuan, Jan Dahrendorff, Derek Wildman, Annie Qu.

Writing – review & editing: Yubai Yuan, Monica Uddin, Annie Qu.

References

1. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends in Genetics. 2008; 24(3):133–141. https://doi.org/10.1016/j.tig.2007.12.007 PMID: 18262675
2. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods. 2008; 5(7):621–628. https://doi.org/10.1038/nmeth.1226 PMID: 18516045
3. Zhang H, Pounds SB, Tang L. Statistical methods for overdispersion in mRNA-Seq count data. The Open Bioinformatics Journal. 2013; 7(1). https://doi.org/10.2174/1875036201307010034
4. Yehuda R. Post-traumatic stress disorder. New England Journal of Medicine. 2002; 346(2):108–114. https://doi.org/10.1056/NEJMra012941 PMID: 11784878
5. Kessler RC, Aguilar-Gaxiola S, Alonso J, Benjet C, Bromet EJ, Cardoso G, et al. Trauma and PTSD in the WHO world mental health surveys. European Journal of Psychotraumatology. 2017; 8 (sup5):1353383. https://doi.org/10.1080/20008198.2017.1353383 PMID: 29075426
6. Mills KL, McFarlane AC, Slade T, Creamer M, Silove D, Teesson M, et al. Assessing the prevalence of trauma exposure in epidemiological surveys. Australian & New Zealand Journal of Psychiatry. 2011; 45 (5):407–415. https://doi.org/10.3109/00048674.2010.543654 PMID: 21189046
7. Zohar J, Fostick L, Cohen A, Bleich A, Dolfin D, Weissman Z, et al. Risk factors for the development of posttraumatic stress disorder following combat trauma: A semiprospective study. The Journal of Clinical Psychiatry. 2009; 70(12):18399. https://doi.org/10.4088/JCP.08m04378blu PMID: 19852906
8. Yehuda R, LeDoux J. Response variation following trauma: a translational neuroscience approach to understanding PTSD. Neuron. 2007; 56(1):19–32. https://doi.org/10.1016/j.neuron.2007.09.006 PMID: 17920012
9. Brewin CR, Andrews B, Valentine JD. Meta-analysis of risk factors for posttraumatic stress disorder in trauma-exposed adults. Journal of Consulting and Clinical Psychology. 2000; 68(5):748. https://doi.org/10.1037/0022-006X.68.5.748 PMID: 11068961
10. Lowe SR, Galea S, Uddin M, Koenen KC. Trajectories of post traumatic stress among urban residents. American Journal of Community Psychology. 2014; 53(1):159–172. https://doi.org/10.1007/s10464-014-9634-6 PMID: 24469249
11. Sarapas C, Cai G, Bierer LM, Golier JA, Galea S, Ising M, et al. Genetic markers for PTSD risk and resilience among survivors of the World Trade Center attacks. Disease Markers. 2011; 30(2-3):101–110. https://doi.org/10.3233/DMA-2011-0764 PMID: 21508514
12. Yehuda R, Cai G, Golier JA, Sarapas C, Galea S, Ising M, et al. Gene expression patterns associated with post traumatic stress disorder following exposure to the World Trade Center attacks. Biological Psychiatry. 2009; 66(7):708–711. https://doi.org/10.1016/j.biopsych.2009.02.034 PMID: 19393990
13. Mehta D, Gonik M, Klengel T, Rex-Haffner M, Menke A, Rubel J, et al. Using polymorphisms in FKBP5 to define biologically distinct subtypes of posttraumatic stress disorder: evidence from endocrine and gene expression studies. Archives of General Psychiatry. 2011; 68(9):901–910. https://doi.org/10.1001/archgenpsychiatry.2011.50 PMID: 21536970

14. Kim GS, Smith AK, Xue F, Michopoulos V, Lori A, Armstrong DL, et al. Methylomic profiles reveal sex-specific differences in leukocyte composition associated with post-traumatic stress disorder. Brain, Behavior, and Immunity. 2019; 81:280–291. https://doi.org/10.1016/j.bbi.2019.06.025 PMID: 31228611

15. Anders S, Huber W. Differential expression analysis for sequence count data. Nature Precedings. 2010; p. 1–1. https://doi.org/10.1186/gb-2010-11-10-r106 PMID: 20979621

16. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26(1):139–140. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308

17. Chambers JM. Software for data analysis: programming with R. Springer. 2008; 2(1).

18. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007; 23(21):2881–2887. https://doi.org/10.1093/bioinformatics/btm453 PMID: 17881408

19. Wu H, Wang C, Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. Biostatistics. 2013; 14(2):232–243. https://doi.org/10.1093/biostatistics/kxs033 PMID: 23001152

20. Wu H, Zhang Y, Long JD. Longitudinal beta-binomial modeling using GEE for overdispersed binomial data. Statistics in Medicine. 2017; 36(6):1029–1040. https://doi.org/10.1002/sim.7191 PMID: 27917499

21. Mou T, Deng W, Gu F, Pawitan Y, Vu TN. Reproducibility of methods to detect differentially expressed genes from single-cell RNA sequencing. Frontiers in Genetics. 2020; 10:1331. https://doi.org/10.3389/fgene.2019.01331 PMID: 32010190

22. Landau WM, Liu P. Dispersion estimation and its effect on test performance in RNA-seq data analysis: a simulation-based comparison of methods. PloS One. 2013; 8(12):e81415. https://doi.org/10.1371/journal.pone.0081415 PMID: 24349066

23. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-Seq. Nature Biotechnology. 2013; 31(1):46–53. https://doi.org/10.1038/nbt.2450 PMID: 23222703

24. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-Seq data. Genome biology. 2010; 11(3):1–9. https://doi.org/10.1186/gb-2010-11-3-r25 PMID: 20196867

25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995; 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

26. Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, Collins PJ, Chu TM, Bao W, et al. Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nature Biotechnology. 2006; 24(9):1140–1150. https://doi.org/10.1038/nbt1242 PMID: 16964228

27. Peart MJ, Smyth GK, Van Laar RK, Bowtell DD, Richon VM, Marks PA, et al. Identification and functional significance of genes regulated by structurally different histone deacetylase inhibitors. Proceedings of the National Academy of Sciences. 2005; 102(10):3697–3702. https://doi.org/10.1073/pnas.0500369102 PMID: 15738394

28. Smid M, Coebergh van den Braak RR, van de Werken HJ, van Riet J, van Galen A, de Weerd V, et al. Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq data performs similarly in intersample analyses while improving intrasample comparisons. BMC Bioinformatics. 2018; 19:1–13. https://doi.org/10.1186/s12859-018-2246-7 PMID: 29929481

29. Ver Hoef JM, Boveng PL. Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology. 2007; 88(11):2766–2772. https://doi.org/10.1890/07-0043.1 PMID: 18051645

30. Cotton RG, Horaitis O. Quality control in the discovery, reporting, and recording of genomic variation. Human Mutation. 2000; 15(1):16–21. https://doi.org/10.1002/(SICI)1098-1004(200001)15:1%3C16::AID-HUMU6%3E3.0.CO;2-S PMID: 10612817

31. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012; 28(16):2184–2185. https://doi.org/10.1093/bioinformatics/bts356 PMID: 22743226

32. Deyneko IV, Mustafaev ON, Tyurin A–, Zhukova KV, Varzari A, Goldenkova-Pavlova IV. Modeling and cleaning RNA-seq data significantly improve detection of differentially expressed genes. BMC Bioinformatics. 2022; 23(1):488. https://doi.org/10.1186/s12859-022-05023-z PMID: 36384457

33. Li D, Zand MS, Dye TD, Goniewicz ML, Rahman I, Xie Z. An evaluation of RNA-seq differential analysis methods. PLoS One. 2022; 17(9):e0264246. https://doi.org/10.1371/journal.pone.0264246 PMID: 36112652

34. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-Seq data. BMC Bioinformatics. 2013; 14(1):1–18. https://doi.org/10.1186/1471-2105-14-91 PMID: 23497356

35. Ranabir S, Reetu K. Stress and hormones. Indian Journal of Endocrinology and Metabolism. 2011; 15 (1):18. https://doi.org/10.4103/2230-8210.77573 PMID: 21584161

36. Himes BE, Jiang X, Wagner P, Hu R, Wang Q, Klanderman B, et al. RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells. PloS One. 2014; 9(6):e99625. https://doi.org/10.1371/journal.pone.0099625 PMID: 24926665

37. Cari L, Ricci E, Gentili M, Petrillo MG, Ayroldi E, Ronchetti S, et al. A focused Real Time PCR strategy to determine GILZ expression in mouse tissues. Results in Immunology. 2015; 5:37–42. https://doi.org/10.1016/j.rinim.2015.10.003 PMID: 26697291

38. Franco LM, Gadkari M, Howe KN, Sun J, Kardava L, Kumar P, et al. Immune regulation by glucocorticoids can be linked to cell type–dependent transcriptional responses. Journal of Experimental Medicine. 2019; 216(2):384–406. https://doi.org/10.1084/jem.20180595 PMID: 30674564

39. Ronchetti S, Migliorati G, Riccardi C. GILZ as a mediator of the anti-inflammatory effects of glucocorticoids. Frontiers in Endocrinology. 2015; 6:170. https://doi.org/10.3389/fendo.2015.00170 PMID: 26617572

40. SNYDER SK, WESSELLS JL, WATERHOUSE RM, DVEKSLER GS, WESSNER DH, WAHL LM, et al. Pregnancy-specific glycoproteins function as immunomodulators by inducing secretion of IL-10, IL-6 and TGF-β1 by human monocytes. American Journal of Reproductive Immunology. 2001; 45(4):205–216. https://doi.org/10.1111/j.8755-8920.2001.450403.x PMID: 11327547

41. Blois SM, Sulkowski G, Tirado-González I, Warren J, Freitag N, Klapp BF, et al. Pregnancy-specific glycoprotein 1 (PSG1) activates TGF-β and prevents dextran sodium sulfate (DSS)-induced colitis in mice. Mucosal Immunology. 2014; 7(2):348–358. https://doi.org/10.1038/mi.2013.53 PMID: 23945545

42. Binder EB. The role of FKBP5, a co-chaperone of the glucocorticoid receptor in the pathogenesis and therapy of affective and anxiety disorders. Psychoneuroendocrinology. 2009; 34:S186–S195. https://doi.org/10.1016/j.psyneuen.2009.05.021 PMID: 19560279

43. Appel K, Schwahn C, Mahler J, Schulz A, Spitzer C, Fenske K, et al. Moderation of adult depression by a polymorphism in the FKBP5 gene and childhood physical abuse in the general population. Neuropsychopharmacology. 2011; 36(10):1982–1991. https://doi.org/10.1038/npp.2011.81 PMID: 21654733

44. Ising M, Maccarrone G, Brückl T, Scheuer S, Hennings J, Holsboer F, et al. FKBP5 gene expression predicts antidepressant treatment outcome in depression. International Journal of Molecular Sciences. 2019; 20(3):485. https://doi.org/10.3390/ijms20030485 PMID: 30678080

45. Klengel T, Mehta D, Anacker C, Rex-Haffner M, Pruessner JC, Pariante CM, et al. Allele-specific FKBP5 DNA demethylation mediates gene–childhood trauma interactions. Nature Neuroscience. 2013; 16(1):33–41. https://doi.org/10.1038/nn.3275 PMID: 23201972