PLOS | ONE

RESEARCH ARTICLE

Improving Gene-Set Enrichment Analysis of RNA-Seq Data with Small Replicates

Sora Yoon[1], Seon-Young Kim[2,3], Dougu Nam[1,4]*

1 School of Life Sciences, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea, **2** Medical Genomics Research Center, Korea Research Institute of Bioscience & Biotechnology, Daejeon, Republic of Korea, **3** Department of Bioinformatics, University of Science and Technology, Daejeon, Republic of Korea, **4** Department of Mathematical Sciences, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea

* dougnam@unist.ac.kr

Abstract

Deregulated pathways identified from transcriptome data of two sample groups have played a key role in many genomic studies. Gene-set enrichment analysis (GSEA) has been commonly used for pathway or functional analysis of microarray data, and it is also being applied to RNA-seq data. However, most RNA-seq data so far have only small replicates. This enforces to apply the gene-permuting GSEA method (or preranked GSEA) which results in a great number of false positives due to the inter-gene correlation in each gene-set. We demonstrate that incorporating the absolute gene statistic in one-tailed GSEA considerably improves the false-positive control and the overall discriminatory ability of the gene-permuting GSEA methods for RNA-seq data. To test the performance, a simulation method to generate *correlated* read counts within a gene-set was newly developed, and a dozen of currently available RNA-seq enrichment analysis methods were compared, where the proposed methods outperformed others that do not account for the inter-gene correlation. Analysis of real RNA-seq data also supported the proposed methods in terms of false positive control, ranks of true positives and biological relevance. An efficient R package (AbsFilterGSEA) coded with C++ (Rcpp) is available from CRAN.

Citation: Yoon S, Kim S-Y, Nam D (2016) Improving Gene-Set Enrichment Analysis of RNA-Seq Data with Small Replicates. PLoS ONE 11(11): e0165919. doi:10.1371/journal.pone.0165919

Editor: Dongmei Li, University of Rochester, UNITED STATES

Received: July 27, 2016

Accepted: October 19, 2016

Published: November 9, 2016

Data Availability Statement: RNA-seq count table file for Pickrell data is available from ReCount database (http://bowtie-bio.sourceforge.net/recount/). RNA-seq count table for Li data is available from http://yeolab.github.io/papers/. RNA-seq count table for TCGA KIRC data is available within the R package SimSeq. The SimSeq R package is publicly available from CRAN archive and anyone can download and execute the package and use the data therein by typing the command "data(kidney)".

Funding: The genomics programs (NRF-2014M3C9A3068554 and 2014M3C9A3068555) of the National Research Foundation (NRF, http://

Introduction

The high-throughput cDNA sequencing technology (RNA-seq) has enabled an efficient and thorough analysis of the transcriptome in the cell [1, 2]. In particular, RNA-seq exhibited much lower background noise than the hybridization-based method (microarray), which resulted in an improved accuracy in quantitating the gene expression [3]. In spite of this advantage, the differential expression (DE) analysis of RNA-seq data between two sample groups has been a non-trivial task because of the varying sequencing depths and RNA compositions in each sample and the discrete nature of the read count data. To address the problem, various normalization methods have been developed to make the gene expression levels comparable between samples [4, 5], and a variety of methods have been developed to test the DE of each gene based on discrete probability models. [6–12].

www.nrf.re.kr/) of Korea, funded by the Ministry of Science, ICT, and Future Planning, the KRIBB Research Initiative. Basic Science Research Program through a National Research Foundation (NRF) grants funded by the Korean government (MOE) (2014R1A1A2056353). 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.

To interpret the DE analysis result, Gene Ontology (GO) terms or other gene-sets that share common functions have been used to assess the over-representation of a function in the DE genes, which may be called *GO analysis* [13, 14]. Another useful approach is the gene-set enrichment analysis (GSEA) [15]. Unlike GO analysis, GSEA does not use the cutoff threshold to identify the DE genes, but employs the (weighted) Kolmogorov-Smirnov (K-S) statistic to test whether genes contributing to the phenotype are 'enriched' in each gene-set. Thereby, GSEA is able to capture the subtle but coordinated changes in a gene-set and has been commonly used to find important pathways or functions in various diseases and cell conditions from microarray data [16–19].

In spite of the power of GSEA, the pathway analysis methods and tools for RNA-seq have only recently been devised based on methods developed for microarray [11, 20–22]. One of the issues in applying GSEA to RNA-seq data is how to normalize the read count data. voom method transforms the read counts into microarray-like data for which most linear-model based methods developed for microarray can be applied [11]. GSAAseqSP tool [21] deploys TMM or DESeq normalization methods [5, 7] which are able to address both the different depths and RNA compositions between samples. Another important issue is the small replicates: Despite the rapid decrease of the sequencing cost, it is still costly for most laboratories and only a few replicates have been produced for each sample condition [23]. Such small replicates prohibit from using the sample-permuting GSEA (GSEA-SP), but force to use the gene-permuting GSEA (GSEA-GP) which results in a great number of false positive gene-sets caused by the inter-gene correlation in the gene expression.

We demonstrate that the *absolute gene statistic* remarkably reduces the false positive rate and improves the overall discriminatory ability (ROC) of the GSEA-GP methods in analyzing RNA-seq read count data. This property has also been shown for microarray data [24]. RNA-seq read counts were modeled and simulated using discrete probability (negative binomial distribution) [6, 25], and a simulation method to generate 'correlated' read counts within a gene-set was newly developed to compare the performance of GSEA methods for RNA-seq data. Note that the inter-gene correlation has a critical effect on the performance of gene-set level analysis, but has not been considered so far for the counting data because of the lack of such a simulation method.

For a more accurate analysis, a one-tailed GSEA method that only considers the *positive* deviation in the K-S statistic was devised for the absolute enrichment analysis. Based on this result, we also propose filtering the GSEA-GP results with those obtained from the absolute GSEA-GP to effectively reduce false positives. The performances of the absolute GSEA and its filtering method were demonstrated for simulated and real RNA-seq data.

Materials and Methods

Absolute gene-permuting GSEA and filtering

In many RNA-seq experiments, the replicate size is not large enough to carry out GSEA-SP, in which case the GSEA-GP is used instead. However, the gene-permuting method generates a great number of false positives due to the inter-gene correlation in each gene-set [26–30]. Recently, it has been shown that incorporating the absolute gene statistic in GSEA-GP considerably reduces the false positive rate and improves the overall discriminatory ability in analyzing microarray data [24]. Therefore, we tested whether the absolute statistic exhibits a similar benefit in analyzing RNA-seq counting data. In addition to substituting the gene scores with their absolute values (i.e., flipping the sign of negative gene scores) [31], our absolute GSEA is modified as a one-tailed test by considering only the 'positive' deviation in the K-S statistic. We have two reasons for this modification. First, simply replacing the gene scores with their absolute values in GSEA can result in a few 'down-regulated' gene-sets which are meaningless in an *absolute* enrichment analysis. Removing these down-regulated gene-sets in itself may confer beneficial effect. Our one-tailed absolute GSEA does not yield any down-regulated gene-sets. Second, it provides more accurate background null distribution of gene-set scores: In the conventional GSEA algorithm, the maximum positive and negative deviation values are compared and only the larger absolute value between the two is selected for the gene-set score. This means the minor maximum deviation values are all excluded in constituting the gene-set null distributions. By taking only the positive deviation values (one-tailed K-S statistic), every gene-set participates in constituting the null distribution of absolute gene-set statistic even when the negative deviation is larger. In real RNA-seq data analysis, the one-tailed method tended to be more conservative than the two-tailed method.

Gene scores. Four gene scores were considered for normalized read as follows:

1. Moderated *t*-statistic (mod-*t*): A modified two-sampe *t*-statistic

$$\tilde{t}_i = \frac{\mu_i^1 - \mu_i^2}{\tilde{s}_i \sqrt{\upsilon_i}}$$

where μ_i^n is the mean read count of *i*th gene, g_i in class *n*, and $\tilde{s_i}$ is a shrinkage estimation of the standard deviation of g_i. This statistic is useful for small replicate data and is implemented using the limma R package [32, 33]

2. Signal-to-Noise ratio (SNR): The SNR (S_i) is calculated as

$$S_i = \frac{\mu_i^1 - \mu_i^2}{\sigma_i^1 + \sigma_i^2}$$

where σ_i^n is the standard deviation of expression values of g_i in class *n*.

3. Zero-centered rank sum (Ranksum): This two-sample Wilcoxon statistic is introduced by Li and Tibshirani [34]. For g_i, the rank sum test statistic (T_i) is calculated as,

$$T_i = \sum_{j \in C_1} R_{ij} - \frac{n_1 \cdot (n+1)}{2}$$

where R_{ij} is the rank of expression level of j^{th} sample among all counts of g_i, C_1 is a set of sample indexes in the first phenotypic class, n_1 is the sample size of C_1 and n is the total sample size. Note that $E(T_i) = 0$.

4. Log fold-change (logFC): Log fold-change (logFC_i) for g_i is calculated as

$$\log FC_i = \log_2 \frac{\mu_i^1}{\mu_i^2}$$

Absolute GSEA. GSEA algorithm identifies functional gene-sets that show a coordinated gene expression change between given phenotypes from gene expression profiles. Given gene scores, GSEA implements a (weighted) K-S statistic to calculate the enrichment score (ES) of each pre-defined gene-set.

1. Enrichment score: Let S be a gene-set and r_i be the gene score of g_i. Then, the enrichment score ES(S) is defined as the maximum deviation of $p_{hit} - p_{miss}$ from zero, that is

$$ES(S) = \begin{cases} \max_i(p_{hit,i} - p_{miss,i}), & if\ |\max_i(p_{hit,i} - p_{miss,i})| \geq |\min_i(p_{hit,i} - p_{miss,i})| \\ \min_i(p_{hit,i} - p_{miss,i}), & if\ |\max_i(p_{hit,i} - p_{miss,i})| < |min_i(p_{hit,i} - p_{miss,i})| \end{cases}$$

   where

$$p_{hit,i} = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{|r_j|^q}{N_R}, \qquad p_{miss,i} = \sum_{\substack{g_j \in S^c \\ j \leq i}} \frac{1}{(N - N_H)}, \qquad N_R = \sum_{g_j \in S} |r_j|^q$$

   N is the total number of genes in the dataset, N_H is the number of genes included in S and q is a weighting exponent which is set as one in this study as recommended [15]. (For the classical K-S statistic, $q = 0$)

2. ES for one-tailed absolute GSEA: The absolute GSEA can be simply conducted by replacing the gene scores by their absolute values, but the orders of gene scores are quite different from the original GSEA algorithm in calculating the K-S statistic. For the one-tailed test, only the positive deviation ES(S) = $\max_i(p_{hit,i} - p_{miss,i})$ (right-tailed K-S statistic) is considered for the gene-set score even when the negative deviation is larger.

Then, the *gene permutations* are applied, and the corresponding ES's are calculated and normalized for evaluating the false discovery rate of each gene-set [15].

Filtering with absolute GSEA. To reduce the false positives in the GSEA-GP, we propose using the absolute GSEA-GP results for filtering false positives from the ordinary GSEA-GP results. In other words, only the gene-sets that are significant in both ordinary and the one-tailed absolute GSEA are considered significant. In this way, more reliable gene-sets with directionality can be obtained. In all the analyses presented in this paper, the same FDR cutoff is applied for both ordinary and absolute methods, but different cutoffs can also be considered for stricter or looser filtering.

Simulation of the read count data with the inter-gene correlation

Inter-gene correlation in each gene-set critically affects the performance of gene-permuting gene-set analysis methods (a.k.a. competitive analysis) [26, 35]. For microarray data, multivariate normal distributions have been used for modeling the inter-gene correlation [24, 29, 36], which cannot be directly applied for 'discrete' read count data. Here, we present a method to simulate read count data involving the inter-gene correlation within each gene-set. N = 10,000 genes are considered and the replicate sizes for the test and control groups are n_1 and n_2, respectively.

Step 1. Parameter estimation and read count generation: The read count X_{ij} of ith gene in jth sample has been modeled by an over-dispersed Poisson distribution, called negative binomial (NB) distribution [6, 7, 25] denoted by $X_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2)$ where μ_{ij} and $\sigma_{ij}^2 = \mu_{ij} + \varphi_i\mu_{ij}^2$ are the mean and variance, respectively, and $\varphi_i \geq 0$ is the dispersion coefficient for gene g_i. Here, $\mu_{ij} = s_j\mu_i$, where s_j is the ’size factor’ or ’scaling factor’ of sample j and μ_i is the expression level of g_i. For simplicity, we assume all the size factors $s_j = 1$ in this simulation study. To obtain realistic parameters, after filtering out genes with less than 10 average counts and normalization, 10,000 genes were randomly selected from the TCGA kidney RNA-seq dataset (denoted as TCGA KIRC) [37], and their mean and tag-wise dispersions were estimated using the edgeR package [25]. The read counts were generated using the R function ’rnbinom’ where the inverse of the estimated dispersion φ_i was input as the 'size' argument. This method generates read counts that are independent between genes.

Step 2. Generation of read count data with the inter-gene correlation: Given a gene-set S with K genes, the inter-gene correlation can be generated by incorporating a common variable within the gene-set. Let μ_i and φ_i, $i = 1,2,\ldots K$ be the mean and tag-wise dispersion of g_i in the gene-set and C_{ij} be the read count generated from these parameters (*Step 1*). Let $P_S = \{p_1, p_2, \cdots, p_{n_1+n_2}\}$ be probability values randomly sampled from the uniform distribution U(0,1). Then, for each g_i, the probability values in P_S are converted to a read count C^*_{ij}, $j = 1,2,\ldots$, n_1+n_2 using the inverse function of the individual gene's distribution $X_i \sim \mathrm{NB}(\mu_i,\varphi_i)$ such that $p_j \approx P(X_i \leq C^*_{ij})$. In short, C^*_{ij} are generated from the common uniform distribution via the gene-wise NB distribution. The 'correlated' read count for ith gene in jth sample is then obtained by the weighted sum of the original count C_{ij} and the 'commonly generated' count C^*_{ij} as follows:

$$\mathrm{M_{ij}} := [(1-\alpha)\cdot C_{ij} + \alpha\cdot C^*_{ij}]$$

where $\alpha \in [0,1]$ is the mixing coefcient that determines the strength of the inter-gene correlation and [] rounds the value to the nearest integer. One problem with this count is that its variance is reduced as much as $(2\alpha^2 - 2\alpha + 1)$ because

$$\mathrm{Var}(\mathrm{M_{ij}}) \approx (1-\alpha)^2\cdot \mathrm{V}(C_{ij}) + \alpha^2\cdot \mathrm{V}(C^*_{ij}) = (2\alpha^2 - 2\alpha + 1)\cdot \sigma^2_{ij}$$

To remove this factor, we use an inflated dispersion φ'_i derived from the equation

$$(2\alpha^2 - 2\alpha + 1)\cdot(\mu_i + \varphi'_i\mu_i^2) = \mu_i + \varphi_i\mu_i^2$$

$$\varphi_i\prime = \frac{1+\varphi_i\mu_i}{\mu_i\cdot(2\alpha^2 - 2\alpha + 1)} - \frac{1}{\mu_i}$$

instead of φ_i in generating C_{ij} and C^*_{ij}. The relationship between α and inter-gene correlation is shown in Fig 1.

Mining PubMed abstracts for scoring biological relevance of gene-sets in each tissue

To assess the biological relevance of gene-sets with the absolute GSEA filtering, a literature-based gene-set score was devised using the PubMed abstracts. Deregulated gene-sets may be featured by their member genes that are closely related to the tissue corresponding to the input data. For a significant gene-set S, its relevance with a specific tissue T is scored by the log geometric average of the abstract counts as follows:

$$L(S) = \frac{1}{K}\sum\nolimits_{i=1}^{K}\log(A_{T,i}) \tag{1}$$

where K is the gene-set size and $A_{T,i}$ is the number of PubMed abstracts where both the keywords related to the tissue T and the name of g_i co-occur. The literature mining was conducted using RISmed R package [38].

Processing RNA-seq data and gene-set size condition

For all the RNA-seq datasets analyzed in this study, raw read counts were normalized by the median method in DESeq [7], and then the lower five percentile of the normalized counts exclusive of the zero counts was used as an offset to stabilize the logFC score of genes with some low expression. Note that the offset has no effect on the other gene scores. The 'gene-set size' means the number of genes that are found in both the original gene-set and the RNA-seq dataset. When performing GSEA, the gene-set sizes between 10 and 300 inclusive were used.

AbsFilterGSEA R package

An efficient R package 'AbsFilterGSEA' that implements the GSEA-GP with or without absolute filtering and the absolute GSEA-GP was developed and is available from CRAN [39]. It accepts a raw read count matrix and normalizes it using DESeq median method [7]. It also accepts an already normalized dataset. The core GSEA was coded with C++ and the results were cross-checked with those from the original GSEA R-code [15]. It takes only several seconds to minutes depending on the number of gene-sets tested. The integration of C++ code to the R package was implemented using Rcpp package [40].

Results

Comparison of gene-permuting GSEA methods for simulated read count data

To compare the performance of GSEA methods for small replicates, twelve gene-permuting gene-set analysis methods were tested using simulated read count data incorporating the

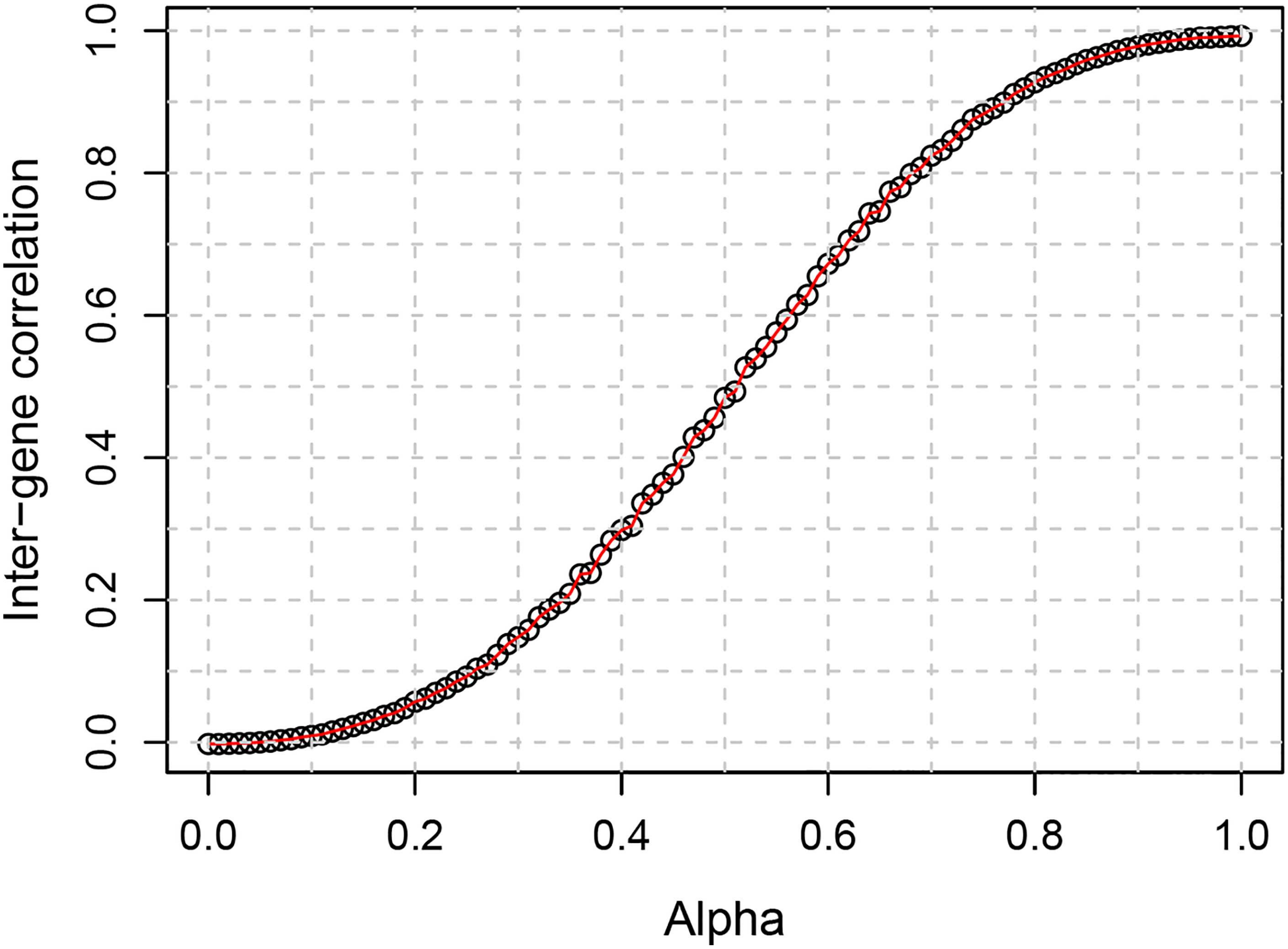

Fig 1. The relationship between the mixing coefficient (alpha) and the average inter-gene correlation.

doi:10.1371/journal.pone.0165919.g001

inter-gene correlation as described in Methods. The simulated read count data included 10,000 genes and 100 non-overlapping gene-sets each of which contained 100 genes.

First, the false positive rates (FPRs; FDR < 0.1) of the GSEA-GP methods for the four gene statistics (mod-*t*, SNR, Ranksum and FC) and their absolute counterparts were measured using the simulated read count datasets with four different levels of inter-gene correlation, LOW (0~0.05), 0.1, 0.3 and 0.6 within each gene-set. Two, three and five replicates in each sample group were tested and no DE genes were included. This test was repeated twenty times and their average FPRs were depicted in Fig 2A and 2D for three and five replicates, respectively. The result for two-replicate case is available in Figure A in S1 File.

A recently developed competitive method, Camera combined with the voom normalization [29, 41], the bias-adjusted random-set method (RNA-Enrich) [22] as well as two preranked GSEA methods [15] were also compared. The preranked GSEA was implemented using the GSEA R-code [15] where the ranks of genes were determined according to either the *p*-values resulted from the differential expression analysis using edgeR [25] package or the simple absolute fold-changes of the normalized count data. Note that SeqGSEA [42] provides only sample–permuting GSEA which is not useful for small replicates, and GSAAseqSP [43] provides a gene-permuting GSEA method which is virtually the same as GSEA-GP described in this paper (We checked they yielded nearly the same results for the simulated count data). Although it is described that GSAAseqSP uses the absolute gene scores, they are only used for the step-sizes in K-S statistic, and it is far from the 'absolute' enrichment analysis.

The FPRs of GSEA-GP for the four ordinary gene statistics and the two preranked methods went up rapidly as the inter-gene correlation was increased. However, the increase rates of FPRs for the four absolute GSEA methods were considerably lower than those for the ordinary statistics. For example, when three replicates were used, even for a moderate inter-gene correlation 0.1, the FPRs for the original statistics were approximately 50% or higher while only a few false positive sets were detected for the absolute methods (1 ~ 3%). Camera yielded no false positives for each correlation level. Overall similar FPR trends were observed with five replicates. RNA-Enrich and the edgeR/preranked methods exhibited relatively better FPRs compared to the GSEA-GP and FC/Preranked methods.

Next, 20% of the gene-sets (20 gene-sets) in the data generated above were replaced with differentially expressed gene-sets to compare the power (true positive rate) and the overall discriminatory abilities (ROC). These gene-sets included 20~80% (uniformly at random in each gene-set) of DE genes whose mean counts in the test or control group were multiplied by 1.5~2.0 with which the read counts in the corresponding group were regenerated. Only weak inter-gene correlations between 0 and 0.05 were randomly assigned to the DE gene-sets and the four different inter-gene correlation levels were applied for the non-DE gene-sets. The corresponding powers and the area under the ROC curves (AUCs) were then obtained for the twelve methods compared (Fig 2B, 2C, 2E and 2F). The GSEA-GP methods and preranked GSEA with FCs had the highest level of power, but their AUCs rapidly declined as the inter-gene correlation level was increased because of their poor false positive controls. With the inter-gene correlation of 0.6, their performances were close to a random prediction (AUC$\approx$ 0.5). On the other hand, the absolute GSEA-GP methods and Camera were less affected by the inter-gene correlation level and exhibited stable and good AUCs. Among the absolute methods, the mod-*t* gene score resulted in best powers and AUCs. The ROC curves (average of 20 repetitions) of the twelve gene-permuting GSEA methods for the inter-gene correlation 0.3 are illustrated in Fig 3 and Figure B in S1 File.

For the two-replicate data, the FPRs were similar to those of triplicate case, but the powers and AUCs were rather lowered (Figure A in S1 File). While the mod-*t* still exhibited best powers and AUCs among the absolute methods, the power of SNR was considerably lowered,

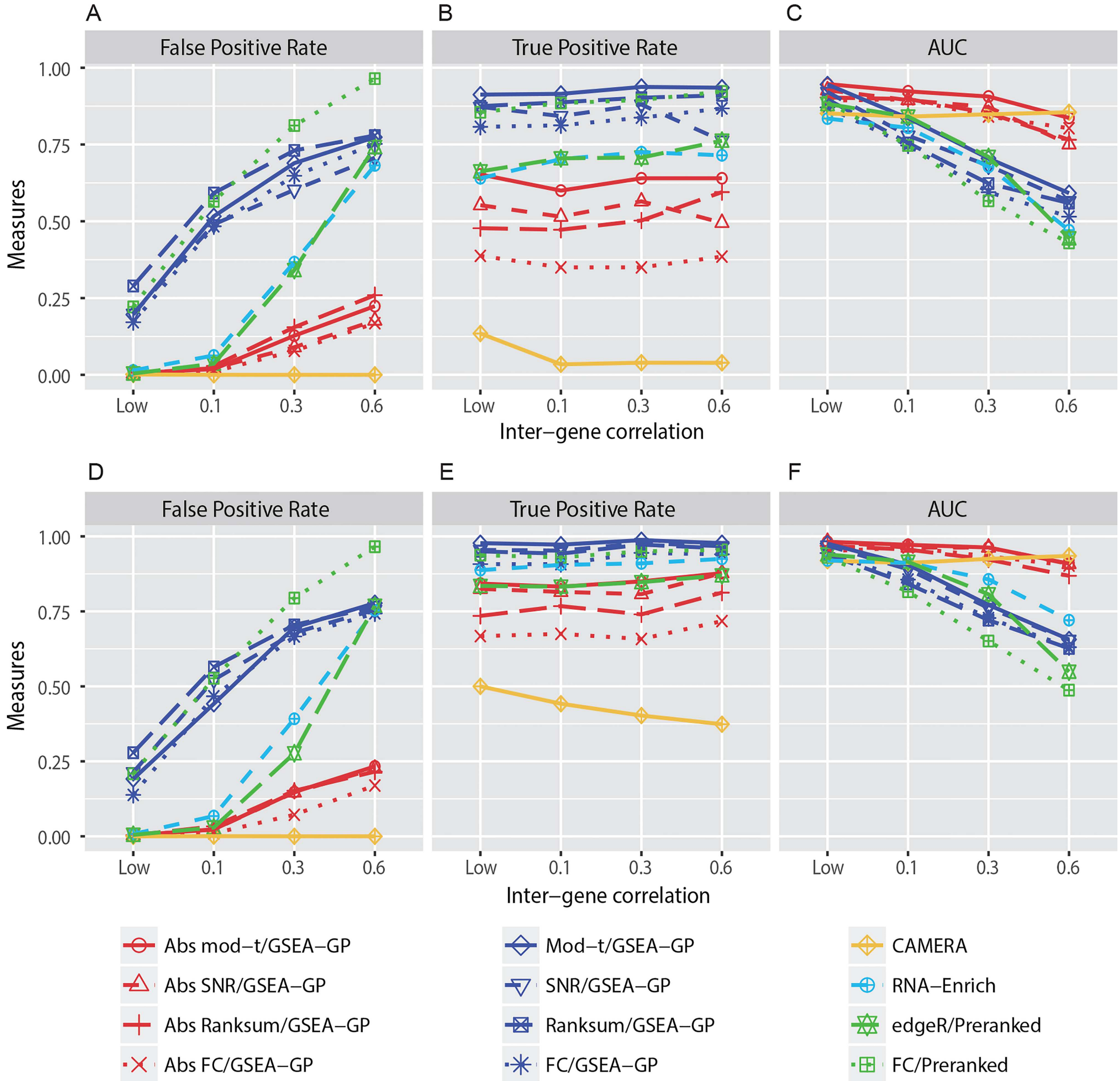

Fig 2. Performance comparison of gene-permuting GSEA methods for simulated read counts. GSEA-GP methods combined with eight gene statistics, (moderated *t*-statistic, SNR, Ranksum, logFC and their absolute versions), Camera combined with voom normalization, RNA-Enrich and two preranked GSEA methods for edgeR *p*-values and FCs were compared for false positive rate, true positive rate and area under the receiver operating curve using simulated read count data with three (A-C) and five replicates (D-F).

doi:10.1371/journal.pone.0165919.g002

which necessitates the moderated gene statistic in GSEA of small replicate data. Lastly, different inter-gene correlations were randomly assigned for gene-sets in a dataset, and two, three and five replicate cases were tested (Figure A in S1 File). The absolute *mod-t* absolute method still exhibited best AUCs in most cases and exhibited overall similar trends as the identical inter-gene correlation cases.

Overall, these results indicate that the absolute GSEA-GP provides an excellent false positive control and improves the overall discriminatory ability of GSEA-GP. Although the

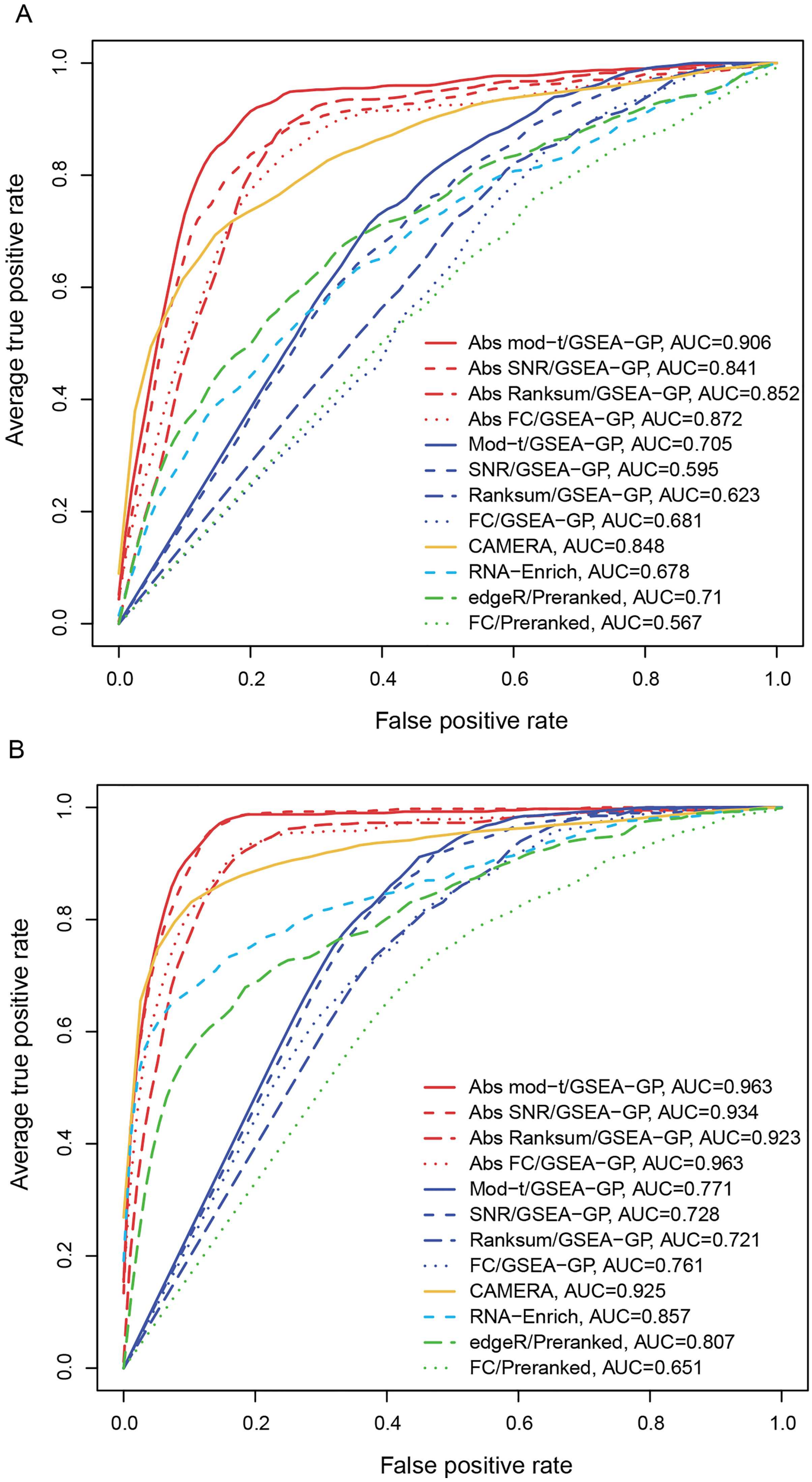
A
Average true positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
Abs mod−t/GSEA−GP, AUC=0.906
Abs SNR/GSEA−GP, AUC=0.841
Abs Ranksum/GSEA−GP, AUC=0.852
Abs FC/GSEA−GP, AUC=0.872
Mod−t/GSEA−GP, AUC=0.705
SNR/GSEA−GP, AUC=0.595
Ranksum/GSEA−GP, AUC=0.623
FC/GSEA−GP, AUC=0.681
CAMERA, AUC=0.848
RNA−Enrich, AUC=0.678
edgeR/Preranked, AUC=0.71
FC/Preranked, AUC=0.567
B
Average true positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
Abs mod−t/GSEA−GP, AUC=0.963
Abs SNR/GSEA−GP, AUC=0.934
Abs Ranksum/GSEA−GP, AUC=0.923
Abs FC/GSEA−GP, AUC=0.963
Mod−t/GSEA−GP, AUC=0.771
SNR/GSEA−GP, AUC=0.728
Ranksum/GSEA−GP, AUC=0.721
FC/GSEA−GP, AUC=0.761
CAMERA, AUC=0.925
RNA−Enrich, AUC=0.857
edgeR/Preranked, AUC=0.807
FC/Preranked, AUC=0.651

Fig 3. Average receiver operating characteristic (ROC) curves. The average ROC curves (20 repetitions) of the twelve gene-permuting GSEA methods applied to simulation data with the inter-gene correlation of 0.3 for (A) three and (B) five replicate cases.

doi:10.1371/journal.pone.0165919.g003

ordinary GSEA-GP methods exhibited best powers, the true positives are overwhelmed by the prohibitively high rate of false positives resulting in very poor ranks of true positives (AUCs). In general, the false positive control and power may be regarded as a tradeoff between different methods, but the overall gain is represented by ROC analysis which demonstrated a clear improvement with the absolute GSEA methods. Compared with Camera, the absolute methods yielded a little more false positives, but exhibited better power and overall discriminatory ability (correlation$\leq$0.3). For small replicate datasets, not all of the true positives may be prioritized perfectly by any method, but it would be important to discern some of the truly altered gene-sets reliably. The proposed absolute GSEA approach provides a simple and reasonable solution for this purpose. However, if minimizing true negatives is of main interest, we recommend investigating all the 'significant' gene-sets (typically hundreds) resulted from the ordinary GSEA-GP or generating more samples for GSEA-SP.

Comparison of GSEA methods for RNA-seq data

The performances of GSEA methods were compared for published RNA-seq datasets in several aspects. First, two RNA-seq datasets denoted by Pickrell and Li data, respectively, were analyzed for comparing power and accuracy as follows:

The Pickrell data were generated from the lymphoblastoid cell lines of 69 unrelated Nigerian individuals (29 male and 40 female) [44]. To analyze the chromosomal differences in expression between male and female, MSigDB C1 (cytogenetic band gene-sets) [45–47] was used for analysis. The GSEA-SP with SNR gene score was applied for the total dataset which resulted in two significant gene-sets 'chryq11' (FDR = 0.00143) and 'chrxp22' (FDR = 0.0514) both of which were sex-specific. These two gene-sets were significantly up-regulated in male and female groups, respectively. Since the GSEA-SP controls the false positives well, these two gene-sets were regarded as true positives. Then, five samples were randomly selected from each group to constitute a small replicate dataset and GSEA-GP methods with or without absolute filtering, Camera, edgeR/Preranked methods were compared for this small replicate dataset. This process was repeated ten times. Using mod-*t* and logFC as the gene scores, on average, the GSEA-GP yielded 48.2 and 20.4 significant (FDR<0.25) gene-sets including 1.6 and 1.1 true positives, respectively. On the other hand, GSEA-GP with the absolute filtering resulted in only 2.6 and 3.5 significant gene-sets which included 1.11 and 1 true positives for the mod-*t* and logFC gene scores, respectively. For these five-replicate datasets, Camera did not detect any significant gene-set, and the edgeR/Preranked detected as many as 137.4 which included 1.8 true positives. This result implies that the absolute filtering method effectively reduces the false positives resulted from GSEA-GP while maintaining a good statistical power.

A similar trend was observed with the Li dataset. The Li data [48] were generated from LNCaP cell lines with three samples treated with dihydrotestosterone (DHT) and four control samples. The MSigDB C2 (curated gene-set) was used for analysis and the six gene-sets containing the term 'androgen' were regarded as potential true positives since DHT is a kind of androgen, though there can be other truly altered gene-sets. When the GSEA-SP with mod-*t* and logFC gene score was applied for this small replicate dataset, as expected, only one and no 'androgen' gene-set was significant (FDR<0.1), respectively. On the other hand, GSEA-GP with mod-*t* and logFC gene scores yielded as many as 187 and 569 significant gene-sets, respectively, which included four 'androgen' gene-sets with FDR$\leq$0.0067. When the absolute

Table 1. Significant gene-sets detected by the absolute GSEA-GP filtering (FDR<0.1) with the mod-*t* score (DHT-treated and control LNCaP cell line).

Gene-set name	FDR	Literature Score
NELSON_RESPONSE_TO_ANDROGEN_DN	**0**	**2.15**
NELSON_RESPONSE_TO_ANDROGEN_UP	**0**	**1.87**
WANG_RESPONSE_TO_ANDROGEN_UP	**1.63×10^{-4}**	**1.37**
PIONTEK_PKD1_TARGETS_UP	2.79×10^{-4}	1.78
REACTOME_AMINO_ACID_SYNTHESIS_AND_INTERCONVERSION_TRANSAMINATION	2.27×10^{-2}	1.94
WANG_RESPONSE_TO_FORSKOLIN_UP	3.02×10^{-2}	1.42
VALK_AML_CLUSTER_11	5.12×10^{-2}	1.17
HUPER_BREAST_BASAL_VS_LUMINAL_UP	4.40×10^{-2}	1.53

doi:10.1371/journal.pone.0165919.t001

filtering was applied, the numbers of significant gene-sets were dramatically reduced to *eight* (Table 1) and 242, which included three and four ’androgen’ gene-sets, respectively. Of note, the top three gene-sets were ‘androgen’ terms for the mod-*t* score. The absolute GSEA filtering with SNR score provided a similar result. Camera detected only two ’androgen’ gene-sets within 101 significant gene-sets with FDR = 0.00836 and 0.0195, respectively. RNA-Enrich and edgeR/Preranked were so sensitive for this dataset that 1108 and 782 sets were significant (FDR<0.1). RNA-Enrich and edgeR/Preranked detected *four* and three androgen terms within top 52 and 91 gene-sets.

Overall, the results for real data analysis were concordant with the simulation results. GSEA-GP yielded a large number of significant gene-sets most of which seemed to be false positives. The absolute filtering method considerably reduced false positives at the cost of small loss of power. Camera exhibited a strict false positive control, but its power was relatively weak. In particular, the absolute filtering with mod-*t* score exhibited a high precision and a good power in both datasets. The absolute filtering with the two-tailed absolute GSEA-GP yielded a little more liberal results which are available from Supporting Information.

Effects of the absolute filtering on false positive control and biological relevance

Here, the effects of the absolute filtering were analyzed for real data in two other aspects. The first one is the false positive rate as investigated with the variance inflation factor (VIF). The FPR of a competitive gene-set analysis method is known to be determined by VIF which is defined as:

$$\mathrm{Var}(\text{gene set statistic}) = \mathrm{Var}_{\text{i.i.d.}}(\text{gene set statistic}) \times \mathrm{VIF}$$

where $\mathrm{Var}_{\text{i.i.d.}}$ is the variance of a gene-set statistic under the assumption that genes in each gene-set have independent expression values. For a linear gene-set statistic, the VIF is explicitly represented as a function of the gene-set size (K) and the average inter-gene correlation ($\bar{\rho}$) [29, 49] as follows:

$$VIF = 1 + (K - 1)\bar{\rho} \tag{2}$$

To compare the FPRs of the GSEA-GP and the absolute GSEA-GP methods, VIF distributions (2) of the significant gene-sets were compared for two TCGA RNA-seq datasets (KIRC and BRCA tumor vs. normal) [50]. These datasets comprise a large number of cancer and normal samples (144 for KIRC and 216 for BRCA in total, respectively) with which the average inter-gene correlation can be reliably estimated. In each dataset, five cancer samples and five normal samples were randomly drawn to constitute a small replicate dataset, to which GSEA-GP was applied for the MSigDB C2 curated gene-sets using the gene scores logFC and absolute logFC, respectively. Then, the VIFs were compared between two classes of significant gene-sets as follows: One is the gene-sets that are significant only in the ordinary GSEA-GP (class A) and the other is those that are significant in both the ordinary and absolute GSEA-GP methods (class B). Note that the total samples in each dataset were used to calculate $\bar{\rho}$. This process was repeated ten times and the corresponding VIF distributions were compared (Figure C in S1 File). In all cases, VIFs of class B were significantly smaller than those for class A which implies smaller FPRs in the absolute GSEA-GP method. All the ten randomly drawn sub-datasets exhibited significantly smaller VIFs in class B in both the TCGA datasets. (Wilcoxon ranksum p-value$<$0.05; smallest p-value 9.77E-27 for KIRC and 7.56E-33 for BRCA dataset). This indicates the absolute filtering method substantially reduces the false positives in real data analysis.

The second aspect is the tissue-specific relevance score (1). As the above case, five samples were randomly selected from each group of the KIRC and BRCA datasets ten times, and the literature relevance scores between the class A and B sets were compared (Figure C in S1 File). As a result, for all the ten sub-datasets, the relevance scores in class B were significantly larger for both the KIRC and BRCA datasets (smallest p-value: 1.99E-17 and 2.02E-19, respectively).

In addition, the ratios of cancer-related gene-sets (defined as those sets containing one of following keywords such as 'cancer', 'tumor' and 'carcinoma' in their names) were significantly higher in class B compared to those of class A. On average, the ratio of cancer-related terms in class A and class B were 10.0% and 16.3% in KIRC, and 12.7% and 20.3% in BRCA datasets, respectively (Figure C in S1 File). These results indicate that the absolute filtering method tends to result in more reliable and biologically relevant gene-sets.

Discussion

Since the advent of RNA-seq technology until recently, various methods to identify DE genes from the RNA-seq read count data have been developed [6, 25, 41, 51]. One notable feature shared by DE analysis methods is that they yield quite a number of DE genes. RNA-seq is known to provide a much improved resolution in quantitating gene expression compared to that of microarray [2], which may have increased the sensitivity of DE analysis for RNA-seq data.

With the increased resolution and sensitivity, the pathway analysis or GSEA are expected to play a crucial role in genomic studies with their ability to detect the 'subtle but coordinated' changes in a gene-set [15]. However, in many cases, only GO analysis has been applied for interpreting RNA-seq data [52]. The low application rate of pathway analysis or GSEA for RNA-seq may be ascribed to the lack of tools that are specifically designed for RNA-seq data. The popularly used GSEA software [15] developed for microarray analysis can be used for RNA-seq data by normalizing the read count data 'appropriately' or simply applying the gene-permuting method (preranked GSEA) after ranking the gene differential scores using another software (e.g. edgeR or DESeq).

Since the majority of RNA-seq experiments have generated only small replicates, the preranked GSEA methods were often used for function and pathway analysis. However, gene-permuting methods usually result in a great number of false positives due to the inter-gene correlation whatever the replicate sizes are. To date, Camera [29] has been the only method to control the false positive gene-sets caused by the inter-gene correlation in analyzing small replicate read count data, but its statistical power was quite weak. In this study, we showed one-tailed absolute GSEA manifests an excellent false positive control and a good statistical power for analyzing small replicate RNA-seq data. For absolute GSEA, it is natural to consider the one-tailed test, while the conventional GSEA applies two tailed test. Simply flipping the negative gene score is still a two tailed test, so we devised one-tailed version for the absolute enrichment by taking only the positive deviation in K-S statistic. When we compared one-tailed and two-tailed absolute GSEA in simulation tests, both of the methods exhibited nearly the same performances and were effective in reducing false positives. This is not an unexpected result because the absolute GSEA already have larger positive deviations than their negative counterparts in most gene-sets. However, when we compared the two methods in the two real RNA-seq datasets (Pickrell and Li data), the one-tailed method seemed to be more conservative.

To compare the performance of GSEA methods, read count data incorporating the inter-gene correlation were newly designed and simulated. It is crucial to consider the inter-gene correlation in evaluating gene-set analysis methods. The analysis results for the simulated and RNA-seq data commonly demonstrated the effectiveness of the suggested method. As such, the method and tool presented in this paper may facilitate the pathway analysis of RNA-seq data with small replicates.

Supporting Information

S1 File. Contains Supporting Figures A, B and C and Comparison of one-tailed and two-tailed absolute GSEA results.
(PDF)

Author Contributions

Formal analysis: DN SY.

Funding acquisition: DN SYK.

Investigation: SY DN.

Methodology: DN SY SYK.

Project administration: DN SYK.

Software: SY.

Supervision: DN SYK.

Validation: SY DN.

Visualization: SY.

Writing – original draft: DN SY.

Writing – review & editing: SYK.

References

1. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008; 320(5881):1344–9. doi: 10.1126/science.1158441 PMID: WOS:000256441100046.
2. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009; 10(1):57–63. doi: 10.1038/nrg2484 PMID: WOS:000261866500012.
3. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008; 18(9):1509–17. Epub 2008/06/14. doi: 10.1101/gr.079558.108 PMID: 18550803; PubMed Central PMCID: PMCPmc2527709.

4. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013; 14(6):671–83. Epub 2012/09/19. doi: 10.1093/bib/bbs046 PMID: 22988256.

5. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010; 11(3):R25. Epub 2010/03/04. doi: 10.1186/gb-2010-11-3-r25 PMID: 20196867; PubMed Central PMCID: PMCPmc2864565.

6. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007; 23(21):2881–7. Epub 2007/09/21. doi: 10.1093/bioinformatics/btm453 PMID: 17881408.

7. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010; 11(10). ARTN R106 doi: 10.1186/gb-2010-11-10-r106 PMID: WOS:000287378900008.

8. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010; 11:94. Epub 2010/02/20. doi: 10.1186/1471-2105-11-94 PMID: 20167110; PubMed Central PMCID: PMCPmc2838869.

9. Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 2013; 22(5):519–36. Epub 2011/12/01. doi: 10.1177/0962280211428386 PMID: 22127579; PubMed Central PMCID: PMCPmc4605138.

10. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 2013; 14(9):R95. Epub 2013/09/12. doi: 10.1186/gb-2013-14-9-r95 PMID: 24020486; PubMed Central PMCID: PMCPmc4054597.

11. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014; 15(2):R29. Epub 2014/02/04. doi: 10.1186/gb-2014-15-2-r29 PMID: 24485249; PubMed Central PMCID: PMCPMC4053721.

12. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15(12):550. Epub 2014/12/18. doi: 10.1186/s13059-014-0550-8 PMID: 25516281; PubMed Central PMCID: PMCPMC4302049.

13. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. 2003; 4(4):R28. Epub 2003/04/19. PMID: 12702209; PubMed Central PMCID: PMCPMC154579. doi: 10.1186/gb-2003-4-4-r28

14. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009; 37(1):1–13. Epub 2008/11/27. doi: 10.1093/nar/gkn923 PMID: 19033363; PubMed Central PMCID: PMCPMC2615629.

15. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(43):15545–50. doi: 10.1073/pnas.0506580102 PMID: WOS:000232929400051.

16. Carver BS, Chapinski C, Wongvipat J, Hieronymus H, Chen Y, Chandarlapaty S, et al. Reciprocal Feedback Regulation of PI3K and Androgen Receptor Signaling in PTEN-Deficient Prostate Cancer. Cancer Cell. 2011; 19(5):575–86. doi: 10.1016/j.ccr.2011.04.008 PMID: WOS:000290932400004.

17. Schwarz JK, Payton JE, Rashmi R, Xiang T, Jia YH, Huettner P, et al. Pathway-Specific Analysis of Gene Expression Data Identifies the PI3K/Akt Pathway as a Novel Therapeutic Target in Cervical Cancer. Clinical Cancer Research. 2012; 18(5):1464–71. doi: 10.1158/1078-0432.Ccr-11-2485 PMID: WOS:000301040700030.

18. Li HL, Chiappinelli KB, Guzzetta AA, Easwaran H, Yen RWC, Vatapalli R, et al. Immune regulation by low doses of the DNA methyltransferase inhibitor 5-azacitidine in common human epithelial cancers. Oncotarget. 2014; 5(3):587–98. PMID: WOS:000332701300003. doi: 10.18632/oncotarget.1782

19. Nam D, Kim J, Kim SY, Kim S. GSA-SNP: a general approach for gene set analysis of polymorphisms. Nucleic acids research. 2010; 38(Web Server issue):W749–54. Epub 2010/05/27. doi: 10.1093/nar/gkq428 PMID: 20501604; PubMed Central PMCID: PMCPmc2896081.

20. Wang X, Cairns MJ. Gene set enrichment analysis of RNA-Seq data: integrating differential expression and splicing. Bmc Bioinformatics. 2013; 14. Artn S16 doi: 10.1186/1471-2105-14-S5-S16 PMID: WOS:000318816300016.

21. Xiong Q, Mukherjee S, Furey TS. GSAASeqSP: A Toolset for Gene Set Association Analysis of RNA-Seq Data. Scientific Reports. 2014; 4. ARTN 6347 doi: 10.1038/srep06347 PMID: WOS:000341935800002.

22. Lee C, Patil S, Sartor MA. RNA-Enrich: a cut-off free functional enrichment testing method for RNA-seq with improved detection power. Bioinformatics. 2015. Epub 2015/11/27. doi: 10.1093/bioinformatics/btv694 PMID: 26607492.

23. Feng JX, Meyer CA, Wang Q, Liu JS, Liu XS, Zhang Y. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. Bioinformatics. 2012; 28(21):2782–8. doi: 10.1093/bioinformatics/bts515 PMID: WOS:000310155300010.

24. Nam D. Effect of the absolute statistic on gene-sampling gene-set analysis methods. Stat Methods Med Res. 2015. doi: 10.1177/0962280215574014 PMID: 25733546.

25. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26(1):139–40. doi: 10.1093/bioinformatics/btp616 PMID: WOS:000273116100025.

26. Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007; 23(8):980–7. doi: 10.1093/bioinformatics/btm051 PMID: WOS:000246293000009.

27. Nam D, Kim SY. Gene-set approach for expression pattern analysis. Brief Bioinform. 2008; 9(3):189–97. Epub 2008/01/19. doi: 10.1093/bib/bbn001 PMID: 18202032.

28. Newton MA, Quintana FA, Den Boon JA, Sengupta S, Ahlquist P. Random-Set Methods Identify Distinct Aspects of the Enrichment Signal in Gene-Set Analysis. Annals of Applied Statistics. 2007; 1(1):85–106. doi: 10.1214/07-Aoas104 PMID: WOS:000261050400005.

29. Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Research. 2012; 40(17). ARTN e133 doi: 10.1093/nar/gks461 PMID: WOS:000309464300007.

30. Nam D. De-correlating expression in gene-set analysis. Bioinformatics. 2010; 26(18):i511–i6. doi: 10.1093/bioinformatics/btq380 PMID: WOS:000281714100016.

31. Saxena V, Orgill D, Kohane I. Absolute enrichment: gene set enrichment analysis for homeostatic systems. Nucleic Acids Research. 2006; 34(22). ARTN e151 doi: 10.1093/nar/gkl766 PMID: WOS:000243191500004.

32. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004; 3(1):1–25.

33. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43(7):e47. Epub 2015/01/22. doi: 10.1093/nar/gkv007 PMID: 25605792; PubMed Central PMCID: PMCPmc4402510.

34. Li J, Tibshirani R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Statistical Methods in Medical Research. 2013; 22(5):519–36. doi: 10.1177/0962280211428386 PMID: WOS:000325863700005.

35. Nam D, Kim SY. Gene-set approach for expression pattern analysis (vol 9, pg 189, 2008). Briefings in Bioinformatics. 2008; 9(5):450-. doi: 10.1093/bib/bbn030 PMID: WOS:000258471200009.

36. Song S, Black MA. Microarray-based gene set analysis: a comparison of current methods. Bmc Bioinformatics. 2008; 9. Artn 502 doi: 10.1186/1471-2105-9-502 PMID: WOS:000262159300001.

37. Cancer Genome Atlas Research N. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013; 499(7456):43–9. doi: 10.1038/nature12222 PMID: 23792563; PubMed Central PMCID: PMC3771322.

38. Kovalchik S. RISmed: Download Content from NCBI Databases. 2015.

39. R.Core.Team. R: A Language and Environment for Statistical Computing. 2015.

40. Eddelbuettel D, Francois R. Rcpp: Seamless R and C plus plus Integration. J Stat Softw. 2011; 40(8):1–18. PMID: WOS:000289457800001.

41. Law CW, Chen YS, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology. 2014; 15(2). ARTN R29 doi: 10.1186/gb-2014-15-2-r29 PMID: WOS:000336256600012.

42. Wang X, Cairns MJ. SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing. Bioinformatics. 2014; 30(12):1777–9. Epub 2014/02/19. doi: 10.1093/bioinformatics/btu090 PMID: 24535097.

43. Xiong Q, Mukherjee S, Furey TS. GSAASeqSP: a toolset for gene set association analysis of RNA-Seq data. Sci Rep. 2014; 4:6347. Epub 2014/09/13. doi: 10.1038/srep06347 PMID: 25213199; PubMed Central PMCID: PMCPmc4161965.

44. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464(7289):768–72. doi: 10.1038/nature08872 PMID: WOS:000276205000047.

45. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011; 27(12):1739–40. doi: 10.1093/bioinformatics/btr260 PMID: WOS:000291261300036.

46. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res. 2015; 43(D1):D1079–D85. doi: 10.1093/nar/gku1071 PMID: WOS:000350210400158.

47. Jiang H, Bivens NJ, Ries JE, Whitworth KM, Green JA, Forrester LJ, et al. Constructing cDNA libraries with fewer clones that contain long poly(dA) tails. Biotechniques. 2001; 31(1):38-+. PMID: WOS:000169804500005.

48. Li H, Lovci MT, Kwon YS, Rosenfeld MG, Fu XD, Yeo GW. Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. Proc Natl Acad Sci U S A. 2008; 105(51):20179–84. doi: 10.1073/pnas.0807121105 PMID: 19088194; PubMed Central PMCID: PMCPMC2603435.

49. Barry WT, Nobel AB, Wright FA. A Statistical Framework for Testing Functional Categories in Microarray Data. Annals of Applied Statistics. 2008; 2(1):286–315. doi: 10.1214/07-Aoas146 PMID: WOS:000261057700019.

50. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490(7418):61–70. doi: 10.1038/nature11412 PMID: WOS:000309446800032.

51. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks (vol 7, pg 562, 2012). Nat Protoc. 2014; 9(10):2513-. doi: 10.1038/nprot1014-2513a PMID: WOS:000343227800017.

52. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biology. 2010; 11(2). ARTN R14 doi: 10.1186/gb-2010-11-2-r14 PMID: WOS:000276434300013.