BioData Mining

METHODOLOGY Open Access

# Gene set analysis controlling for length bias in RNA-seq experiments

Xing Ren[1], Qiang Hu[2], Song Liu[2], Jianmin Wang[2] and Jeffrey C. Miecznikowski[1*]

*Correspondence:
jcm38@buffalo.edu
[1]Department of Biostatistics, SUNY University at Buffalo, 14214 Buffalo, USA
Full list of author information is available at the end of the article

## Abstract

**Background:** In gene set analysis, the researchers are interested in determining the gene sets that are significantly correlated with an outcome, e.g. disease status or treatment. With the rapid development of high throughput sequencing technologies, Ribonucleic acid sequencing (RNA-seq) has become an important alternative to traditional expression arrays in gene expression studies. Challenges exist in adopting the existent algorithms to RNA-seq data given the intrinsic difference of the technologies and data. In RNA-seq experiments, the measure of gene expression is correlated with gene length. This inherent correlation may cause bias in gene set analysis.

**Results:** We develop SeqGSA, a new method for gene set analysis with length bias adjustment for RNA-seq data. It extends from the R package GSA designed for microarrays. Our method compares the gene set maxmean statistic against permutations, while also taking into account of the statistics of the other gene sets. To adjust for the gene length bias, we implement a flexible weighted sampling scheme in the restandardization step of our algorithm. We show our method improves the power of identifying significant gene sets that are affected by the length bias. We also show that our method maintains the type I error comparing with another representative method for gene set enrichment test.

**Conclusions:** SeqGSA is a promising tool for testing significant gene pathways with RNA-seq data while adjusting for inherent gene length effect. It enhances the power to detect gene sets affected by the bias and maintains type I error under various situations.

**Keywords:** RNA-seq, Gene set analysis, Gene length bias, Maxmean statistic, Restandardization

## Background

Ribonucleic acid sequencing (RNA-seq) is a revolutionary tool for gene expression profiling. It has become an important alternative to traditional expression arrays in varieties of studies. How to adopt the existent algorithms for expression arrays to RNA-seq data is a challenge in data analysis. In microarrays the gene expression is a continuous number, while in RNA-seq it is a non-negative integer indicating the number of reads of a gene. More importantly some inherent biases in RNA-seq experiments need to be accounted for. Given the protocol of RNA-seq, it is reasonable to expect that a longer gene will have more counts than an equally expressed short gene. The length effect will cause bias in gene set analysis [1–3].

© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

We use the lymph node carcinoma of the prostate (LNCaP) cells RNA-seq data set [4] as an example to show the gene length bias. The data set contains the RNA sequencing profiles of 3 androgen-treated samples and 4 control samples. For every gene we calculate the average count of the 7 samples. Figure 1 shows the smooth scatterplot of gene length vs average count on a log-log scale. A smooth spline is fit to the scatterplot, suggesting a strong positive correlation between counts and gene length. The dependence of counts and gene length introduces a bias in the test of DE genes such that discoveries will favor long genes over short genes. We rank the genes by their length and divide them into 10 groups, with each group containing 10% of the total genes. We use the exact test in the R package edgeR to obtain the $p$-value (unadjusted) for every gene. Figure 2 shows the percentage of DE genes ($p$-value $< 0.05$) in each group increases with gene length. Given the definition in (1), a set consisting of primarily long genes will have a greater maxmean statistic than an equally expressed set of short genes. As a result, the test will have bias that favors sets of long genes.

There are a number of well-established gene set analysis methods for expression arrays. Most of these tests can be roughly classified into two groups, over-representation analysis (ORA) and functional class scoring (FCS) [5]. In ORA, the genes are labeled as differentially expressed (DE) or null based on thresholding their test statistics or $p$-values. Then we test the gene sets for over-representing the DE genes. The most common tests

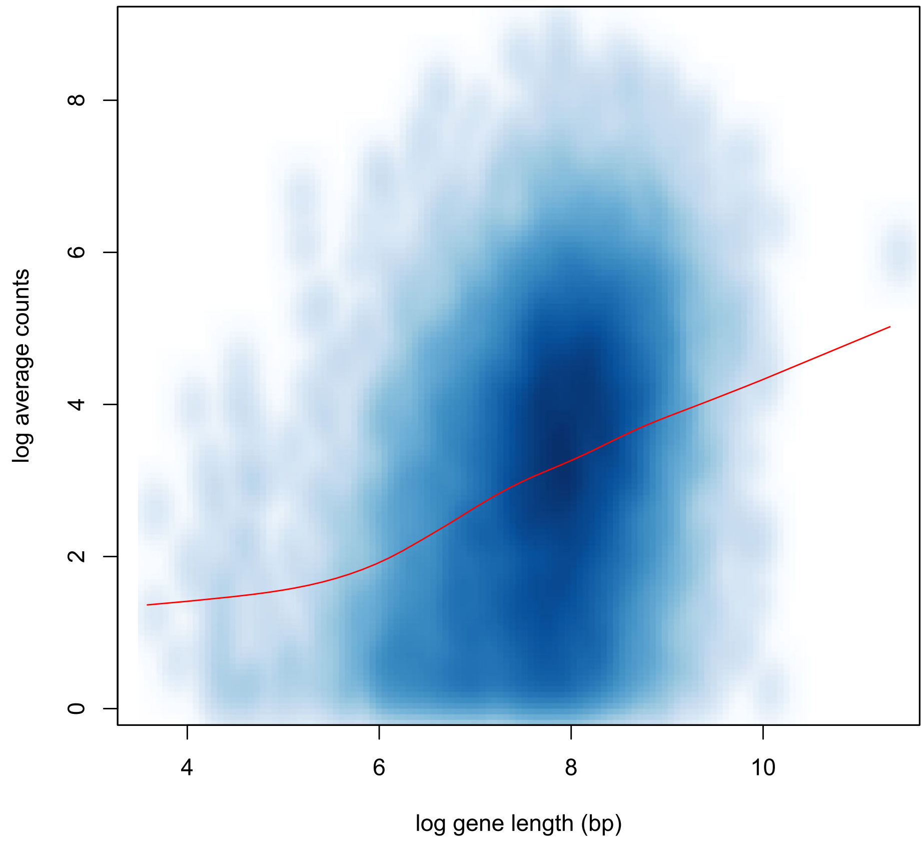

**Fig. 1** Correlation of count and gene length. The scatterplot of gene length vs average counts (log-log scale) for genes in the LNCaP data set [4]. *Darker color* indicates higher scatter density. The *red curve* is a smooth spline with 7 degrees of freedom, showing a positive correlation between count and gene length

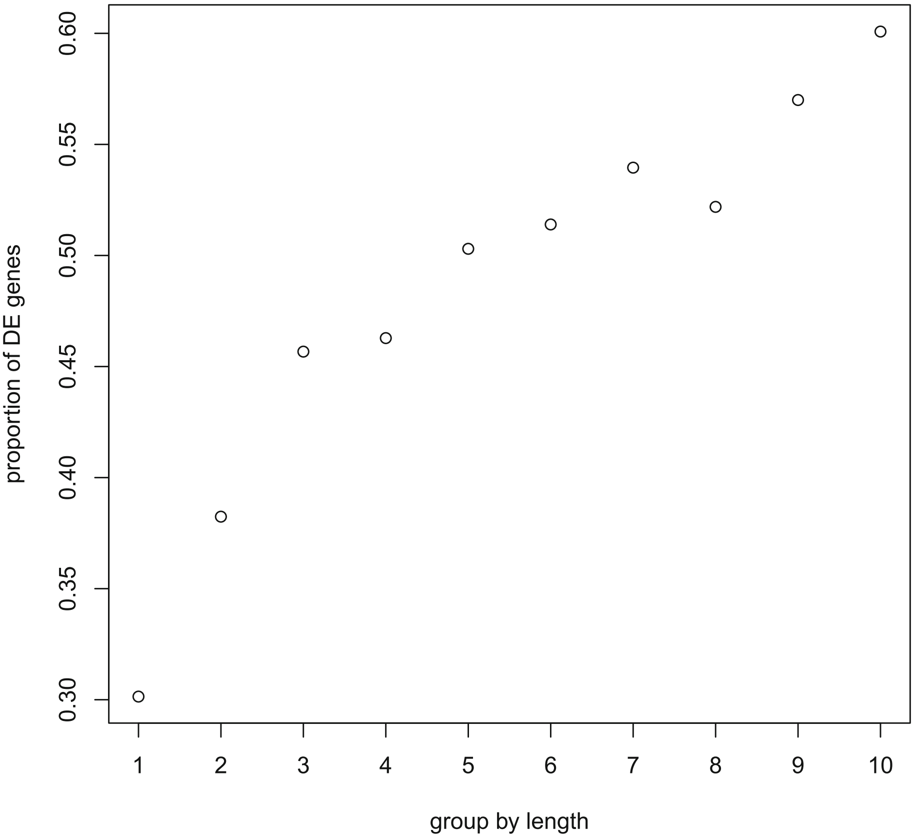

**Fig. 2** Gene length bias in RNA-seq data. The LNCaP data set [4] shows the probability of significant *p*-values ($p < 0.05$) increases as gene length. Group 1 is the genes of shortest length and Group 10 is the genes with longest length

are based on hypergeometric, Chi-square or binomial distribution [5]. Despite the wide usage, ORA approaches have a few limitations. First, some information is discarded in ORA as it ignores the continuous measurement of the test statistics and treats the data as binary outcomes (DE or null). Second, the test for over-representation assumes the genes are independent of each other, which is unlikely to hold given the complex interactions between genes, especially for genes in the same pathway.

The FSC approaches overcome the aforementioned limitations of ORA. First the test statistics are computed for individual genes, e.g. correlation [6], ANOVA [7], Q-statistic [8], signal-to-noise ratio [9], t-statistic [7, 10], z-statistic [11]. Then these gene-level statistics are aggregated into a pathway-level summary statistic e.g. Kolmogorov-Smirnov statistic [9, 12], sum, mean or median [13], the Wilcoxon rank sum [14] and the maxmean statistic [15]. By keeping the gene-level statistics, FSC can detect weak but coordinated changes in gene pathways. In assessing the significance of the pathway statistic, the hypothesis test can be classified into two major categories, self-contained or competitive. A self-contained test permutes sample labels and compares the pathway with its permutations, while a competitive test randomizes genes for each pathway, and compares the pathway with others. The self-contained test maintains the correlation structure between genes, but ignores the other pathways. On the other hand, the competitive test takes into account the other pathways but ignores the gene correlation.

Several gene set/pathway analysis methods have been developed for RNA-seq data, e.g. GOSeq [16], GSVA [17], SeqGSEA [18], CAMERA [19]. GOSeq employs a revised hypergeometric test for DE gene enrichment with the sampling probability adjusted to gene length. GSVA estimates variation of pathway activity over a sample population and it is able to detect subtle changes of gene expression in the pathway. SeqGSEA adopts DSGseq [20] and DESeq [21] for gene level test and uses the GSEA method [9] for functional enrichment analysis. CAMERA [19] estimates the inter-gene correlation and accordingly adjusts the gene set test statistic.

We propose a new functional scoring method SeqGSA for gene set analysis with RNA-seq data. Our method adopts the maxmean statistic in R package GSA [15] as the gene-set-level statistic. Comparing with the more commonly used method GSEA [9], it is argued that the maxmean statistic used in GSA is more powerful than the Kolmogorov-Smirnov statistic in GSEA [15]. The fundamental idea of GSA is that it combines the features of both self-contained test and competitive test through a restandardization procedure. As the original GSA, our method maintains the correlation of genes in permutation tests while also taking into account the competition of other random gene sets. We implement a flexible weighted restandardization scheme to adjust for the gene length bias. It works with any gene-level DE test e.g. [21–23]. We show in the presence of a length bias our method improves the power of identifying significant gene sets and controls type I errors. We also compare our method with GOSeq [16] and CAMERA [19], two representative methods for gene set enrichment test. We show that the maxmean statistic is more appropriate in detecting small changes of expression. Our method also maintains the type I error more accurately when genes are correlated.

## Methods

### Maxmean statistic and restandardization in GSA

In GSA the gene-level test statistics are first converted to $z$ statistics using quantile functions, and then the $z$ values are aggregated into a gene-set-level maxmean statistic. A restandardization procedure compares the maxmean statistic against permutations while also taking into account sets formed by random selections of genes. Given gene-level $z$ statistic $z_i, i = 1, \ldots, n$, let $\mathcal{S}$ be the indices of the gene set and $n_{\mathcal{S}}$ be the size of $\mathcal{S}$. The maxmean statistic $S$ is defined as,

$$\begin{aligned} S_+ &= \frac{1}{n_{\mathcal{S}}} \sum_{i \in \mathcal{S}} z_i I\{z_i > 0\}, \\ S_- &= -\frac{1}{n_{\mathcal{S}}} \sum_{i \in \mathcal{S}} z_i I\{z_i < 0\}, \\ \mathrm{S} &= \max(S_+, S_-). \end{aligned} \tag{1}$$

Assessing statistical significance of the maxmean statistic requires a null distribution. There are two operations considered to obtain such null distribution, randomization or permutation. For randomization, the null is constructed by randomly sampling a large number of gene sets (row sampling). The problem of randomization is that random sets do not maintain the correlation structure among genes as in real sets. For permutation, the null is estimated by column (sample labels) permutations so that genes still remain in the same set. It maintains the correlation structure. The problem of permutation is that it ignores the distribution of the maxmean statistics of the other gene sets. The restandardization procedure in GSA combines randomization and permutation. Given $z$ statistic $z_i$ ($i = 1, \ldots, n$) and gene set $\mathcal{S}$ with $n_{\mathcal{S}}$ genes, the algorithm works as follows,

1) Randomization: calculate the mean and standard deviation of the maxmean statistics of randomized gene sets, denoted by $\mu^{\dagger}$ and $\sigma^{\dagger}$.
2) Standardization: compute standardized maxmean $S^*$, $S^* = (S - \mu^{\dagger})/\sigma^{\dagger}$.
3) Permutation: permute sample labels $B$ times and compute the standardized maxmean for each permuted data as in 1) and 2), $S^{*b}, b = 1, \ldots, B$.

To assess how significant a gene set is associated with the outcome, the $p$-value is calculated by comparing $S^*$ against the permutations, $p = \sum_{b=1}^{B} I\{S^{*b} > S^*\}/B$. With restandardization, the null distribution for $S$ is essentially the permuted maxmean statistic rescaled by $\sigma^{\dagger}$ and relocated by $\mu^{\dagger}$. Therefore the maxmean statistic is compared against the permutations while also considering the distribution of the randomized sets.

### Restandardization weighted by gene length

In the randomization step of GSA, all genes are sampled with equal probability. In particular, two sets with the same number of genes will have the same $\mu^{\dagger}$ and $\sigma^{\dagger}$ regardless of the length of genes in the set. As the method was originally developed for microarrays, it does not consider the potential bias in RNA-seq, as shown in the introduction. To adjust for gene length bias we propose a weighted restandardization algorithm. The weighted algorithm uses the similarity of gene length as sampling weight in the randomization step of the restandardization. The empirical cumulative distribution function (CDF) of the gene length is:

$$\hat{F}(x) = \sum_{i=1}^{n} I\{l_i \le x\}/n\,, \tag{2}$$

where $l_i$ is the number of base pairs of gene $i$ and $n$ is the total number of genes. Note $l_i$ for gene length can be retrieved from public databases e.g. Ensembl (http://www.ensembl.org/) and refSeq (http://www.ncbi.nlm.nih.gov/refseq/).

For a set $\mathcal{S}$, the process of constructing randomized set can be viewed as stepwise replacing genes in $\mathcal{S}$ with genes from all the genes. Instead of assigning equal probability to all genes, we let $q_{ij}$ be the probability of replacing gene $i$ in $\mathcal{S}$ by gene $j$, which is weighted by $1 - |\hat{F}(l_i) - \hat{F}(l_j)|$,

$$\begin{aligned} w_{ij} &= 1 - |\hat{F}(l_i) - \hat{F}(l_j)| \quad \forall \quad i \in \mathcal{S} \quad \text{and} \quad j = 1, \ldots, n\,, \\ q_{ij} &= \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}}\,. \end{aligned} \tag{3}$$

The probability $q_{ij}$ is large when gene $i$ and gene $j$ have similar length and it is small when their lengths are very different. For a set $\mathcal{S}$, let $q_{\mathcal{S}j}$ be the probability of selecting gene $j$ ($j = 1, 2, \ldots, n$) into a randomized set. Since $n_{\mathcal{S}} << n$ in practice, the following approximation can be made,

$$q_{\mathcal{S}j} \approx \frac{\sum_{i \in \mathcal{S}} q_{ij}}{\sum_{j=1}^{n} \sum_{i \in \mathcal{S}} q_{ij}}\,. \tag{4}$$

Random sets are constructed from a multinomial distribution with probability $q_{\mathcal{S}j}$ for gene $j$. To differentiate from $\mu^{\dagger}$ and $\sigma^{\dagger}$ in the previous section, we denote the mean and standard deviation of weighted randomization by $\mu_w^\dagger$ and $\sigma_w^\dagger$. We use $\mu_w^\dagger$ and $\sigma_w^\dagger$ to standardize the maxmean statistic of $\mathcal{S}$,

$$S_w^* = \left(S - \mu_w^\dagger\right)/\sigma_w^\dagger . \tag{5}$$

Permutation maxmean statistics $S_w^{*b}, b = 1, \ldots, B$, are computed in the same fashion from permuted data sets. The $p$-value for $\mathcal{S}$ is calculated by comparing $S_w^*$ and $S_w^{*b}$,

$$p = \sum_{b=1}^{B} I\left\{S_w^{*b} > S_w^*\right\}/B . \tag{6}$$

### Justification of weighted restandardization

In this section we present a model for RNA-seq testing that accommodates a gene length bias. Under the gene length bias, we show our weighted method has more power to detect gene sets consisting of short length genes.

We begin with the two group model proposed in [24] where we assume there are $n$ cases (genes) that are each either null or non-null with probability $p_0$ and $p_1$ and with $z$-values having density either $f_0(z)$ or $f_1(z)$. We assume $f_0(z)$ is the standard normal density and $f_1(z)$ is some no-null density. As mentioned in previous section, RNA-seq tests are not based on traditional $z$-values and hence we transform the $p$-values to $z$-values, via $z_i = \Phi^{-1}(p_i)$ where $\Phi(\cdot)$ is the standard normal CDF. We have,

$$z_i \sim f(z) = p_0 f_0(z) + p_1 f_1(z) , \tag{7}$$

where $f(z)$ is the mixture density composed of $f_0(z)$ and $f_1(z)$. The setting in (7) has proven very useful in microarray analysis, e.g. see [25–28]. To obtain more specific results we also assert the assumption of independence that $z_i$'s are independent.

We propose the following twist on this model for RNA-seq analysis. Assume genes are ordered according to the gene length, then in light of RNA-seq, we propose the following more general model,

$$z_i \sim p_0 f_0(z) + p_1 f_{1i}(z) . \tag{8}$$

In short, we allow each gene to follow a different alternative distribution. This flexibility allows us to characterize a gene length bias.

To accommodate the gene length into the model in (8), we assume $f_{1i}$ has the following form,

$$f_{1i}(z) = r f_{1i}^-(z) + (1-r) f_{1i}^+(z) , \tag{9}$$

where $f_{1i}^-$ and $f_{1i}^+$ are densities with means $\mu_{1i}^- < 0$ and $\mu_{1i}^+ > 0$, respectively. To impose the gene length bias we will further require

$$\mu_{1i}^+ \geq \mu_{1j}^+ \text{ and } \mu_{1i}^- \leq \mu_{1j}^- \quad \forall \quad i > j . \tag{10}$$

For our purposes it is not important to determine the parameters in (9) and (10), however, we will need the assumption in (10) to justify our method. The specification in (10) allows us to handle a gene length bias where, by nature of a gene's length, it is more likely to have extreme $z$ values than a DE gene of shorter length. For simplicity in the following calculations we also assume $f_{1i}$ is symmetric, that is,

$$f_{1i}(z) = f_{1i}(-z) \quad \forall \quad z . \tag{11}$$

The assumption in (11) is plausible, since a priori, it is reasonable to expect a DE gene to be equally likely to be overexpressed as it is to be underexpressed between two conditions. We proceed in the following manner, we derive the asymptotic distribution of the maxmean statistic in (1). We then examine the restandardization procedure and show that our method yields a more accurate restandardization mean for gene sets.

Under (7), by the Lindeberg-Feller theorem [29] if the Lindeberg condition holds, $S_+$ and $S_-$ in (1) asymptotically follow a bivariate normal distribution for adequately large $n_{\mathcal{S}}$,

$$\begin{pmatrix} S_+ \\ S_- \end{pmatrix} \sim N_2 \left\{ \begin{pmatrix} \mu_+ \\ \mu_- \end{pmatrix}, \begin{pmatrix} \sigma_+^2 & \mathrm{Cov}(S_+, S_-) \\ \mathrm{Cov}(S_+, S_-) & \sigma_-^2 \end{pmatrix} \right\}, \tag{12}$$

where $\mu_+ = E(S_+)$ and $\sigma_+^2 = V(S_+)$ and similarly for $S_-$. Note that under (8)

$$\begin{aligned} \mu_+ &= p_0 \int_0^\infty z\phi(z)dz + \frac{p_1}{n_{\mathcal{S}}} \sum_{i \in \mathcal{S}} \int_0^\infty z f_{1i}(z)dz \\ &= p_0 \sqrt{\frac{2}{\pi}} + \frac{p_1}{n_{\mathcal{S}}} \sum_{i \in \mathcal{S}} \int_0^\infty z f_{1i}(z)dz, \end{aligned} \tag{13}$$

and due to the symmetry of$f_{1i}$, $\mu_- = \mu_+$.

For completeness the other quantities in (12) are,

$$\begin{aligned} \sigma_+^2 = \sigma_-^2 &= \left(\frac{1}{n_{\mathcal{S}}}\right)^2 \sum_{i \in \mathcal{S}} V\left(z_i^+\right) \\ &= \left(\frac{1}{n_{\mathcal{S}}}\right)^2 \sum_{i \in \mathcal{S}} \left[ E\left(\left(z_i^+\right)^2\right) - \left(E\left(z_i^+\right)\right)^2 \right] \\ &= \left(\frac{1}{n_{\mathcal{S}}}\right)^2 \sum_{i \in \mathcal{S}} \left[ \left( \int_0^\infty \frac{\sqrt{z}}{2} \left(p_0\phi(\sqrt{z}) + p_1 f_{1i}(\sqrt{z})\right) dz \right) \right. \\ &\quad \left. - \left( p_0 \sqrt{\frac{2}{\pi}} + p_1 \int_0^\infty z f_{1i}(z)dz \right)^2 \right] \end{aligned} \tag{14}$$

$$\begin{aligned} \mathrm{Cov}(S_+, S_-) &= E(S_+ S_-) - E(S_+)E(S_-) \\ &= -\left(\frac{1}{n_{\mathcal{S}}}\right)^2 E\left[ \left( \sum_{i \in \mathcal{S}} z_i I\left(z_i > 0\right) \right) \left( \sum_{i \in \mathcal{S}} z_i I\left(z_i < 0\right) \right) \right] - (\mu_+)^2 \\ &= -\left(\frac{1}{n_{\mathcal{S}}}\right)^2 \left[ \sum_{i \neq j \in \mathcal{S}} E\left(z_i z_j I\left(z_i > 0\right) I\left(z_j < 0\right)\right) \right] - (\mu_+)^2 \\ &= -\left(\frac{1}{n_{\mathcal{S}}}\right)^2 \left[ \sum_{i \neq j \in \mathcal{S}} E\left(z_i I\left(z_i > 0\right)\right) E\left(z_j I\left(z_j < 0\right)\right) \right] - (\mu_+)^2 \\ &= -\left(\frac{1}{n_{\mathcal{S}}}\right)^2 \left[ \sum_{i \neq j \in \mathcal{S}} \left( p_0 \sqrt{2/\pi} + p_1 \int_0^\infty z f_{1i}(z)dz \right) \right. \\ &\quad \left. \times \left( p_0 \sqrt{2/\pi} + p_1 \int_{-\infty}^0 z f_{1j}(z)dz \right) \right] - (\mu_+)^2 \end{aligned} \tag{15}$$

From the work in [30, 31] on order statistics, the pdf of $S$ is derived as,

$$\begin{aligned} f_S(s) = &\frac{1}{\sigma_+}\phi\left(\frac{-s+\mu_+}{\sigma_+}\right) \times \Phi\left(\frac{\eta\,(-s+\mu_+)}{\sigma_+\sqrt{1-\eta^2}} - \frac{-s+\mu_-}{\sigma^-\sqrt{1-\eta^2}}\right) \\ &+ \frac{1}{\sigma_-}\phi\left(\frac{-s+\mu_-}{\sigma_-}\right) \times \Phi\left(\frac{\eta(-s+\mu_-)}{\sigma_-\sqrt{1-\eta^2}} - \frac{-s+\mu_+}{\sigma_+\sqrt{1-\eta^2}}\right) \end{aligned} \tag{16}$$

where $\eta$ is the correlation of $S_+$ and $S_-$, $\phi$ is the pdf for the standard normal distribution. The mean of $S$ (computed via moment generating functions as in [32]) is

$$E(S) = \mu_+\Phi\left(\frac{\mu_+ - \mu_-}{\theta}\right) + \mu_-\Phi\left(\frac{\mu_- - \mu_+}{\theta}\right) + \theta\phi\left(\frac{\mu_- - \mu_+}{\theta}\right), \tag{17}$$

where $\theta = \left(\sigma_+^2 - 2\eta + \sigma_-^2\right)^{1/2}$ and $\sigma_+^2$, $\sigma_-^2$ and $\eta$ are given in (14) and (15). Hence, we can simplify (17) such that

$$\begin{aligned} E(S) =& \frac{1}{2}(\mu_+) + \frac{1}{2}(\mu_-) + \theta\phi(0), \\ =& \mu_+ + 0.40\theta \\ =& p_0\sqrt{\frac{2}{\pi}} + \frac{p_1}{n_{\mathcal{S}}}\sum_{i\in\mathcal{S}}\int_0^\infty zf_{1i}(z)dz + 0.40\theta\,, \end{aligned} \tag{18}$$

where each quantity in (18) has been previously derived.

We would like $S$ to be large for DE sets and small for null sets. Given (18) we see that, on average, $S$ will be large for a set of long genes in the presence of a strong gene length bias where $\int_0^\infty zf_{1i}(z)dz$ increases with the length of gene $i$. This event does not represent a true biological gene set enrichment and thus we would consider detecting this set a false positive. On the other hand, $S$ could also be large if $p_1$ is large for the genes in $\mathcal{S}$ relative to $p_1$ for the genes not in $\mathcal{S}$. This situation does reflect true biological enrichment and therefore we would like our procedure to detect significance in this setting.

So in light of the magnitude of $S$ and its randomized distribution for comparison here is the problem: consider a **biologically enriched** gene set $\mathcal{S}$ consisting of primarily short genes in the presence of a strong gene length bias. While $p_1$ is large for $\mathcal{S}$ it will not be detected because $1/n_{\mathcal{S}}\sum_{i\in\mathcal{S}}\int_0^\infty zf_{1i}(z)dz$ is relatively large for randomly assembled gene sets $\mathcal{S}'$. Likewise, for null sets with long genes $\int_0^\infty zf_{1i}(z)dz$ will be large relative to other random sets and we would like a scheme with a low probability of calling this null set significant. Our weighted randomization scheme is designed to increase (relative to the unweighted procedure) the probability of detecting the truly enriched gene set while decreasing the probability of detecting the unenriched gene sets.

Recall the algorithm compares the maxmean statistic against a distribution of permuted maxmean statistics that are centered and scaled according to the mean ($\mu^\dagger$) and variance ($\sigma^\dagger$) of maxmean statistics chosen at random. Hence in order for $\mathcal{S}$ to be enriched, $S$ must be significantly different than permuted maxmean values *and* random maxmean values. Hence, in light of the scheme above, we would like $\mu^\dagger$ to be small for truly differentially enriched sets and large for truly null sets. Heuristically the main difference between the two methods is in the computation of $\mu_w^\dagger$ and $\mu^\dagger$. Under the weighted resampling scheme, with $K$ resamplings, we have

$$\mu_w^\dagger = \frac{1}{K}\sum_{k=1}^{K} S_w'^k\,. \tag{19}$$

Hence, the difference between $\mu_w^{\dagger}$ and $\mu^{\dagger}$ can be studied by examining $E(S_w'^k)$ against $E(S'^k)$ from the unweighted approach.

From (18) we have

$$E(S_w) = \mu_{+w} + 0.40\theta_w \tag{20}$$

where for the unweighted resampling we have,

$$E(S) = \mu_{+} + 0.40\theta \tag{21}$$

We expect the difference between the second order effects $\theta_w$ and $\theta$ will be minor relative to the first order effects $(\mu_{+w}, \mu_{+})$ between the two methods.

From (13) with the weighted scenario we have

$$\mu_{+w} = p_0\sqrt{\frac{2}{\pi}} + \frac{p_1}{n_{\mathcal{S}_w'}} \sum_{i \in \mathcal{S}_w'} \int_0^{\infty} z f_{1i}(z) dz, \tag{22}$$

where the weighted scenario is based on choosing genes in $\mathcal{S}_w'$. The unweighted scenario is equivalent except the genes are from $\mathcal{S}'$. By design, the genes in $\mathcal{S}_w'$ are chosen to have a similar length to the genes in $\mathcal{S}$, while the genes in $\mathcal{S}'$ are chosen with equal probability. Hence if $i < j$ then in the presence of a gene length bias in (9) and (10), we have $\int_0^{\infty} z f_{1i}(z) dz < \int_0^{\infty} z f_{1j}(z) dz$. Therefore, if the weighted process is biased to choose genes of smaller length relative to the unweighted process, we are likely to obtain $\mu_{+w} < \mu_{+}$. Hence, on average $S_w'$ will be smaller relative to the unweighted scenario of $S'$ and thus the $p$-values in the weighted approach will be larger.

This feature produces the following effect: for enriched gene sets consisting primarily of short length genes, the weighted restandardization procedure will be more powerful than the unweighted procedure, while in the same way, the weighted method may also increase the type I error. On the surface this does not seem like an appealing solution. However, we will show in several simulations, our method increase the power while still maintaining the type I error under the controlled level.

## Results and discussion

### Simulation

To assess our method, we compare the type I error and power of the two versions of SeqGSA (weighted restandardization and the unweighted restandardization). The unweighted version is similar to the original GSA, except that the $t$ test is replaced with the exact negative binomial test of edgeR, as the latter is considered a more appropriate test for RNA-seq count data. In addition, we compare SeqGSA with two gene set test methods, GOSeq [16] and CAMERA [19]. GOSeq tests for higher proportion of DE genes in a set using a modified hypergeometric test, with adjustment to gene length bias. CAMERA compares the $t$ statistics of genes inside the set and genes outside the set with adjustment to the intra-correlation of the gene sets. It is embedded in the widely used limma package [33]. We conduct two simulations. In simulation 1, the count data is generated from a Poisson model in which the mean parameter is associated with gene length. The gene sets are constructed without consideration to intra-correlation. In simulation 2, we use the data of an RNA-seq experiment. The gene sets are constructed such that genes in a set are correlated. In both simulations, we let the size of gene set $n_{\mathcal{S}} = 50$. For GOSeq, the enrichment test requires a threshold of the exact test $p$-values for DE genes. For our simulations, the threshold is determined by controlling false discovery rate at 0.1 with the Benjamini-Hochberg procedure [34].

### *Simulation 1: sets of uncorrelated genes*

We randomly sample $n = 1000$ genes from the LNCaP data set [4] with gene length between 1000 and 3000 base pairs. Let $l_i(i = 1, 2, \ldots, 1000)$ be the length of gene $i$. The 1000 genes are grouped into 20 non-overlapping sets according to gene length, with 50 genes in every set. Thus set 1 contains the shortest 50 genes, set 2 contains the next shortest 50, and so on. This would generate length bias in gene set analysis. For each set, we let some number of genes be up-regulated and down-regulated. Let $\mu_i = l_i/10$. In control samples, count $x_{ij} \sim Poisson(\mu_i)$. In treatment samples, $x_{ij} \sim Poisson((1 + \beta)\mu_i)$ for up-regulated genes, $x_{ij} \sim Poisson((1 - \beta)\mu_i)$ for down-regulated genes and $x_{ij} \sim Poisson(\mu_i)$ for null genes, where $\beta$ ($\beta = 0.15, 0.3$) indicates the effect of DE at the gene level.

We first compare the type I error. Let $n_{de} = 2$ be the number of up- and down-regulated genes for all 20 sets. Thus all sets are considered as null. Table 1 summarizes the type I error ($p$-value $< 0.05$) in selected sets. Results shows the type I error is under control for all sets. In particular, under small DE effect ($\beta = 0.15$), comparing with the unweighted method, weighted SeqGSA reduces the type I error in sets of long genes (set 15 and set 20), but increases the error in sets of short genes (set 1). GOSeq has higher type I error for set 1 but it is very conservative for all other sets. Under large DE effect ($\beta = 0.3$), all methods are conservative in terms of type I error control and the difference is small. Similarly, CAMERA is very conservative under all settings.

Next we compare the power of identifying DE sets that are undermined by the length bias. We let $n_{de}$ be a greater number ($n_{de} = 6, 8, 10$) in set 1, thus set 1 becomes the DE set. An ideal test should have small $p$-values for set 1 and non-significant $p$-values for the others. Table 2 shows that the weighted algorithm increases the power of detecting set 1. In particular, SeqGSA (both weighted and unweighted algorithms) is more powerful under small DE effect ($\beta = 0.15$), while GOSeq is more powerful under more significant change of gene expression ($\beta = 0.3$). On the other hand, CAMERA was unable to detect such DE effect, indicating the test statistic of CAMERA is insensitive to small changes of expression under independence.

In addition we compare the power under no length bias, i.e. genes are grouped randomly thus no length bias exists among the gene sets. In such case, the weighted algorithm does not have advantage over the unweighted version. There is a slight compromise in the power, but the difference is negligible.

**Table 1** Type I error of simulation 1

| | $\beta = 0.15$ | | | | | $\beta = 0.3$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Set 1 | Set 5 | Set 10 | Set 15 | Set 20 | Set 1 | Set 5 | Set 10 | Set 15 | Set 20 |
| Weighted | 0.007 | 0.007 | 0.006 | 0.009 | 0.017 | 0 | 0 | 0 | 0.001 | 0.001 |
| Unweighted | 0.002 | 0.007 | 0.008 | 0.014 | 0.022 | 0 | 0 | 0 | 0 | 0.003 |
| GOSeq | 0.019 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CAMERA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Type I error of selected sets ($p$-value $< 0.05$) under different $\beta$. All four methods maintain the type I error under the controlled level (0.05)

**Table 2** Power of simulation 1

| | With length bias | | | | | | No length bias | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\beta = 0.15$ | | | $\beta = 0.3$ | | | $\beta = 0.15$ | | |
| $n_{de} =$ | 6 | 8 | 10 | 6 | 8 | 10 | 6 | 8 | 10 |
| Weighted | 0.36 | 0.48 | 0.77 | 0.81 | 0.95 | 0.98 | 0.61 | 0.93 | 0.96 |
| Unweighted | 0.29 | 0.45 | 0.75 | 0.77 | 0.90 | 0.96 | 0.61 | 0.94 | 0.97 |
| GOSeq | 0.23 | 0.38 | 0.56 | 0.97 | 1.0 | 1.0 | 0.96 | 1.0 | 1.0 |
| CAMERA | 0.01 | 0 | 0 | 0 | 0 | 0 | 0.02 | 0.01 | 0.02 |

Power of identifying set 1 as DE ($p$-value $< 0.05$) under different $\beta$ and length bias. The weighted SeqGSA increases the power of the unweighted procedure for detecting set 1. SeqGSA is more powerful with small DE effect ($\beta = 0.15$), while GOSeq is more powerful with large DE effect ($\beta = 0.3$). The two procedures of SeqGSA perform similarly under no length bias

### *Simulation 2: sets of correlated genes*

In simulation 2, we use a public data set with simulated gene sets that contain length bias. The data is a subset of the human prostate cancer data [35] downloaded from the European Bioinformatics Institute (EBI) server (http://www.ebi.ac.uk). There are 11 normal and 12 tumor samples in the subset data. The reads were aligned to the human genome reference sequence (version b37 from http://www.1000genomes.org) by TopHat v2.0.8 [36] and Bowtie v2.1.0 [37]. Gene length is retrieved with the getlength function in GOSeq.

Genes with average count less than 1 or with unretrievable length are filtered out. The 23396 remaining genes are divided into four groups by length with each group containing 25% genes. We use edgeR to calculate the $z$ values and label gene $i$ as DE for $|z_i| > \Phi^{-1}(0.975)$, where $\Phi^{-1}$ is the quantile function of standard normal distribution. The standard deviations of the $z$ values from group 1 to group 4 are 1.76, 2.02, 2.09, 2.08 respectively, showing the scale of $z$ values of group 1 is significantly smaller than others. As a result, the proportion of DE genes ($|z| > \Phi^{-1}(0.975)$) of the group 1 is lower than the others (26.7% vs 31.1%, 33.6%, 32.9%). Gene sets from group 1 will be affected by the length bias.

Different from simulation 1, the sets in simulation 2 are generated with intra-correlation. We first compare the type I error under correlation. A null gene set is constructed as follows. First a null gene is sampled from the null genes in group 1 as the hub gene [38]. The simulated gene set contains $n_S$ genes that are sampled from group 1 that have high correlation with the null hub gene (correlation coefficient $> 0.4$). The correlation violates the independence assumption of GOSeq, resulting in an inflated type I error (0.105). On the other hand, the gene correlation is taken into account in the permutation tests of SeqGSA, thus the type I errors are under control for both the weighted algorithm (0.014) and the unweighted algorithm (0.011).

Due to inflated type I error of GOSeq for correlated gene sets, we only compare SeqGSA to CAMERA. A DE hub gene is sampled from DE genes in group 1. The simulated gene set contains a high proportion of DE genes (40 to 70%) that are sampled from DE genes in group 1 and have high correlation with the DE hub gene (correlation coefficient $> 0.4$). The other genes in the set are sampled from the null genes in group 1. The weighted method improves the power for identifying the simulated DE sets (Table 3). Both the SeqGSA methods have higher power than CAMERA, suggesting the maxmean statistic is more sensitive to subtle change of expression on the gene-set level.

**Table 3** Power of simulation 2

| Proportion of DE genes | 0.4 | 0.5 | 0.6 | 0.7 |
|---|---|---|---|---|
| Weighted | 0.22 | 0.53 | 0.71 | 0.87 |
| Unweighted | 0.16 | 0.41 | 0.67 | 0.73 |
| CAMERA | 0.09 | 0.19 | 0.26 | 0.39 |

Power of identifying DE sets (with correlation) in group 1 ($p$-value $< 0.05$). The DE sets are constructed with a high proportion of DE genes in group 1

#### *Simulation 3: power on sets of long genes*

As the weighted method samples more often from genes with similar length, we expect it should also result in a decrease of the power on sets with long genes. In simulation 3 we evaluate how significant the effect is. The same strategy of simulation 2 on was run on group 3 and 4 for the long genes. Results show although there is a slight decrease in the power for group 3 and 4 overall, the decrease is very minimal compared with the increase on the short gene sets in simulation 2 (Table 4). Therefore the weighted method should increase the overall power of all gene sets.

### Application to RNA-seq data set

In this section we first assess the effectiveness of gene length adjustment of SeqGSA with true biological gene sets. Again we use the LNCaP data set [4]. The genes are mapped to the GO categories [39] via R package biomaRt [40] on ENSEMBL (www.ensembl.org) homo sapiens database. The mapped GO categories are filtered by their size. The 1585 categories with 20 to 200 mapped genes are used for the analysis. Small categories are filtered out to prevent outliers in gene length for small gene sets. Large categories are filtered out to avoid slow computation. These categories are grouped into 25 bins of approximately equal size by their median gene length. Figure 3 compares the correlation between the proportion of DE categories and the median length of the bin. A straight line is fit with the slope and its $p$-value shown on the figure. For the unweighted method the correlation is on the borderline of significance ($p = 0.068$) while for weighted method the correlation is not significant ($p = 0.117$).

Next we apply the methods to another public RNA-seq data set. The data set is from the Center for the Study of Human Polymorphisms (CEPH) HapMap samples [41]. It contains 17 female and 24 male human B-cell samples. Genes with average count less than 1 or with unretrievable gene length by the getlength function in GOSeq are filtered out. We map the genes to the 229 KEGG pathways via the getgo function in GOSeq. Only 2549 genes can be mapped to the KEGG pathways. We further filter out pathways with 5 or fewer mapped genes. We run SeqGSA and GOSeq to test DE between male and female group on the remaining 208 pathways. Controlling FDR at 0.1, the weighted SeqGSA identifies the hsa04810 pathway, regulation of actin cytoskeleton. The pathway contains 214 genes, 94 of which are mapped to the data set. It is associated with multiple genetic diseases

**Table 4** Power of simulation 3

| Proportion of DE genes | 0.4 | 0.5 | 0.6 | 0.7 |
|---|---|---|---|---|
| Weighted | 0.45 | 0.75 | 0.90 | 0.95 |
| Unweighted | 0.49 | 0.76 | 0.88 | 0.95 |

Power of identifying DE sets (with correlation) in group 3 and 4 ($p$-value $< 0.05$)

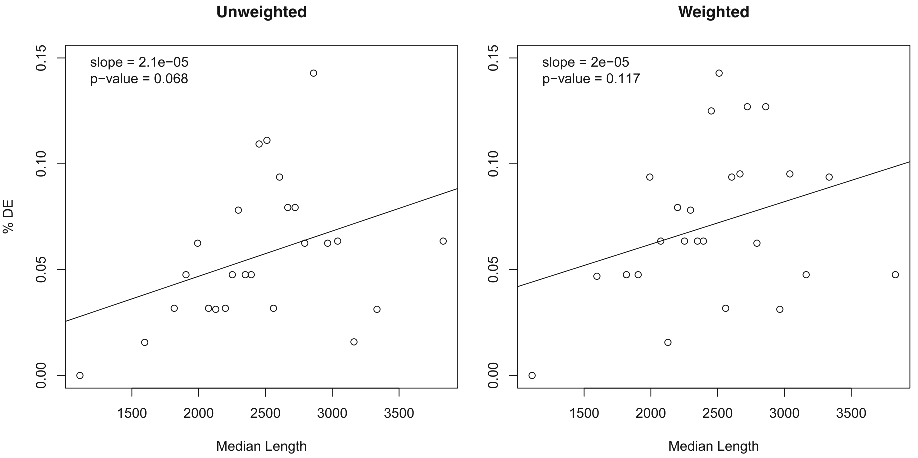

**Fig. 3** Proportion of DE of GO categories for unweighted and weighted SeqGSA. Genes in the LNCaP data set in [4] are mapped to the GO categories. The 1585 mapped GO categories are grouped by median gene length from low to high. For the unweighted method (*left*), the correlation of the proportion of DE categories and gene length is on the borderline of significance ($p = 0.068$). For the weighted method (*right*), the correlation is not significant ($p = 0.117$)

including sex-linked disorders as the non-syndromic X-linked mental retardation. Studies have shown the pathway is related to sex steroids regulation of cell morphology and tissue organization that may play an important role in gender-specific differences of brain dysfunction [42, 43]. Figure 4 is the heatmap of the hsa04810 pathway for the 41 samples. The genes are ordered by their $z$-values. The significance of the pathway is driven by gene TMSB4Y (Ensemble gene ID ENSG00000154620) that lies on the forward strand of chromosome Y. Its homolog on chromosome X escapes X inactivation and encodes an actin sequestering protein (provided by RefSeq, Jul 2008). The TMSB4Y gene is shown in relation to multiple biological activities including actin polymerization and depolymerization in non-muscle cells [44], activation of natural killer cell cytotoxicity [45] and minor histocompatibility antigen encoding [46], etc. The unweighted procedure also identifies the same pathway. Figure 5 compares the standardized maxmean statistic for hsa04810 pathway and its permutation statistics of the two algorithms. The two versions of permutation statistics overlap, while the standardized maxmean $S_w^*$ (red tick) by the weighted method falls further away from the permutation comparing with the unweighted $S^*$ (black tick), suggesting that our weighted method is more powerful in finding the pathway. On the other hand, GOSeq and CAMERA do not find any pathway at FDR $= 0.1$. This example demonstrates our method is easy to implement and gives favorable results in detection of small but coordinated change in gene sets.

## Discussion

We propose SeqGSA, an extension of the GSA method for RNA-seq data. To calculate maxmean statistics from the RNA-seq count data, we replace the $t$ test in the original GSA with a count based test, e.g. the negative binomial test in edgeR. The users can also substitute it with custom defined methods for the gene-level test. More importantly we propose a weighted restandardization approach to accommodate the gene length bias in RNA-seq data. Different from traditional expression arrays, gene length effect generally exists in RNA-seq data sets. To compute the null distribution of the maxmean statistic,

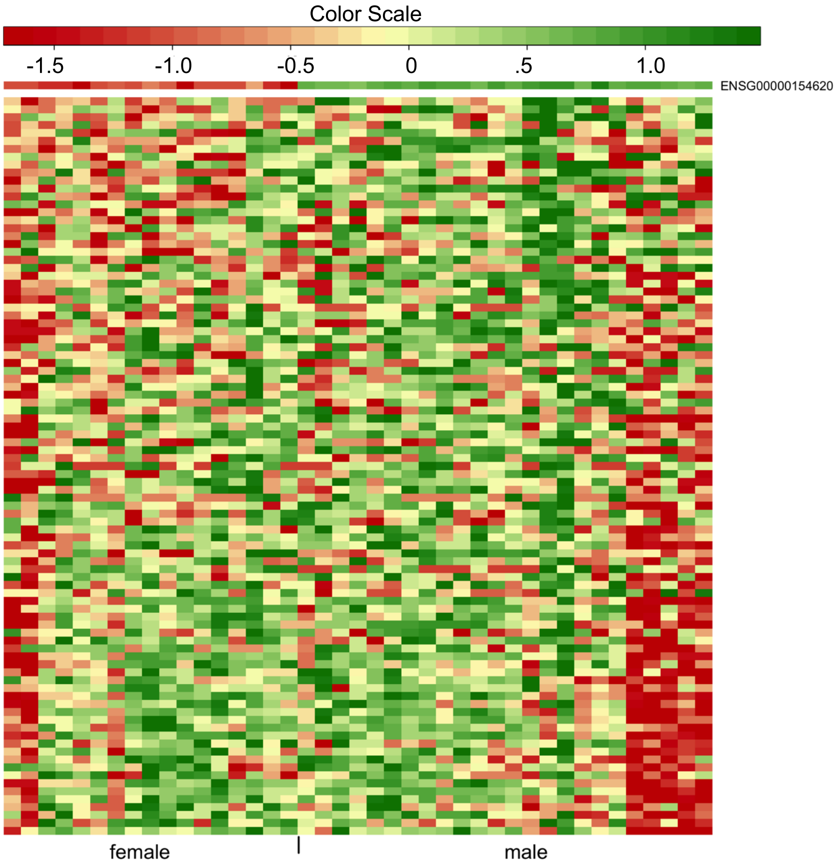

**Fig. 4** Heatmap of hsa04810 KEGG pathway. The heatmap of the 94 mapped genes in the hsa04810 KEGG pathway for the CEPH HapMap data [41]. The count data is log-transformed and row-scaled

GSA employs a randomization strategy that randomly samples genes with equal weights. However, having equal sampling weights ignores the length effect in the data. To adjust for the bias, we use the similarity of gene length as the sampling weight instead, so that the randomized sets will be more likely to consist of genes with similar length. This reduces the bias and yields more accurate null distribution of the maxmean statistics. A DE set comprised mainly of short genes will be more likely called significant in our weighted approach. On the other hand, a null set comprised of long genes will be less significant so that type I error is reduced for these sets. We show by simulations that our method reduces the length bias in RNA-seq data. In addition, we show that the weighted and non-weighted approaches have very similar results when there is no length bias present in the data.

There have been several other methods proposed to remove the length bias in enrichment testing with RNA-seq data [2, 3, 16]. In those works, the enrichment test employs the modified hypergeometric test, Wilcoxon rank sum test or logistic regression. These methods fall in the category of over-representation analysis (ORA) that compares the set against randomized sets. A fundamental difference of our method is that it works under the permutation scheme while also taking into account gene set randomization. An

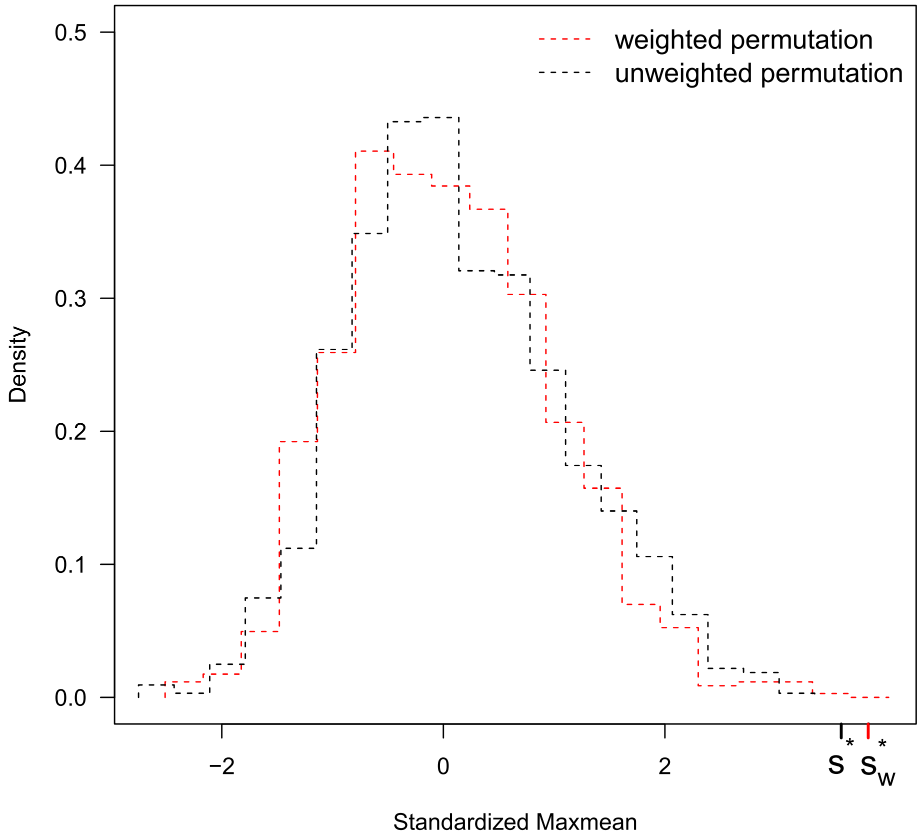

**Fig. 5** Standardized maxmean and permutations of hsa04810 pathway. The standardized maxmean statistic and permutations of the hsa04810 pathway for the CEPH HapMap data [41]. The standardardized maxmean $S^*_W$ (*red*) by the weighted method falls on the right of the unweighted $S^*$ (*black*), while the permutation statistics of the two methods overlap

advantage of the permutation test is that it considers the correlation in a gene set and the correlation structure is maintained during sample label permutations. We compare our method against GOSeq in two simulations. GOSeq also implements a procedure for length bias adjustment. In GOSeq, the adjustment occurs on the gene level. A spline is fit to the proportion of DE genes and gene length. This spline is used to correct the sampling probability of the hypergeometric test. In our method, we adjust the bias on the gene set level. We recalculate the randomization mean and standard deviation based on weighted sampling so that it compares against gene sets of similar length. In the simulations, we control the level of DE by the parameter $\beta$. The value of $\beta$ is chosen to represent two settings, $\beta = 0.15$ for weak DE and $\beta = 0.30$ for moderate DE, each representing the strength of the two methods. In simulation 1, the gene sets are sampled independently. We show that our method is more powerful in finding sets of genes with weak changes. In particular, when expression of genes is weakly changed, the test statistics for many DE genes fall below the threshold of multiple test adjustment in GOSeq. As a result, the power of detecting set of weakly DE genes is undermined. On the other hand, our method aggregates the weak signals of individual genes and increases the power of detecting such DE sets. When there is no correlation among the simulated gene sets, the type I errors of all methods are under the controlled level. In simulation 2, we use the data of an RNA-seq experiment and compare the three methods on simulated gene sets. The gene sets are constructed such that the genes are strongly correlated. This simulates the fact that many of the pre-defined gene sets are identified by gene-gene correlation. The hypergeometric test of GOSeq is based on the assumption that genes are independent. As a result of violated assumption, the type I error of GOSeq is inflated. On the other hand, the permutation test in SeqGSA takes into account the correlation structure within a gene set and thus maintains the type I error more accurately. We also compare SeqGSA with CAMERA, a competitive gene set test also based on sample permutations with adjustment to gene-set intra-correlation. We show that the maxmean statistic in SeqGSA is more sensitive to subtle but synchronized changes in the gene sets, which has been shown as one of the advantages of the original GSA method [15]. In both simulations, CAMERA has limited power to detect small to moderate changes on the gene-set level.

There are some limitations to our length-weighted method. First there are other sources of bias existent in the RNA-seq data, such as “GC-content”. It has been shown that genes with a large number of guanine (G) and cytosine (C) bases are preferentially read by sequencing machines and the effect may not be monotone [47]. A solution to such bias issues is to correct the bias at the gene level by modeling the number of reads or test statistics by gene length or GC content. However this method has its own problem since the true expression level is unknown. Without a comparison experiment, it is difficult to tell whether the difference in reads and test statistics results from true biological expression or biases. Second, the resampling step in SeqGSA weighted by the empirical CDF of gene length is a simple solution, but it does not completely remove the length bias. To improve the performance of the weighted restandardization, we may consider using tunable parameters to adjust the weight. Determining tuning parameters requires further exploration and assumptions. Third, our method improves the power for sets comprised of short genes, but it may slightly compromise the power to detect enriched sets comprised of long genes. This compromise is common for all methods of length bias adjustment. We feel this comprise is scientifically justifiable since the gain in power for small length gene pathways is much larger than the loss of power for large length gene pathways. The effect of this bias on the family-wise error rate or false discovery rate in multiple gene set testing needs further investigation and is part of our future work on this topic. Fourth, in this paper we focus on the exact binomial test in edgeR for gene-level test. However, there are many other tests that can be considered, e.g. see [21, 23, 48–51]. Alternatively, log fold change can be considered as it is not affected by length. This motivates a question for future work on an optimal test statistic for gene set analysis. Last, the computation performance of the weighted algorithm is slightly slower than the unweighted algorithm as it estimate mean and standard deviation using different weights, but the speeds are comparable as the most computationally intense step, computing the permutation $z$ values, is the same for both methods.

## Conclusions

We develop a gene set analysis method for RNA-seq data affected by gene length bias. This novel approach is designed to enhance the power to detect DE gene sets comprised of mainly small length genes. Importantly, we justify our method and demonstrate that it controls the type I error comparing to a representative ORA method for RNA-seq. Also, we show that without the presence of a gene length bias, our approach still performs nearly the same as the original unweighted algorithm in GSA. We expect our gene set analysis method will be of great utility to researchers performing gene set analysis with RNA-seq datasets.

**Acknowledgements**
None.

**Funding**
This work was supported by Roswell Park Cancer Institute and National Cancer Institute (NCI) grant #P30 CA016056.

**Availability of data and materials**
SeqGSA can be accessed at https://github.com/xingrenub/SeqGSA.

**Authors' contributions**
XR, SL, JW, JM designed the experiments and simulations. QH performed the sequence alignment and bioinformatics. XR performed the statistical analysis and wrote the manuscript. JM conceived the study and helped to draft the manuscript. All authors read and approved the final manuscript.

**Competing interests**
The authors declare that they have no competing interests.

**Consent for publication**
Not applicable.

**Ethics approval and consent to participate**
Not applicable.

**Author details**
[1]Department of Biostatistics, SUNY University at Buffalo, 14214 Buffalo, USA. [2]Department of Biostatistics and Bioinformatics, Roswell Park Cancer Institute, 14263 Buffalo, USA.

Received: 23 March 2016 Accepted: 11 January 2017
Published online: 06 February 2017

**References**

1. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biol Direct. 2009;4(1):14.
2. Gao L, Fang Z, Zhang K, Zhi D, Cui X. Length bias correction for RNA-seq data in gene set analyses. Bioinformatics. 2011;27(5):662–9.
3. Mi G, Di Y, Emerson S, Cumbie JS, Chang JH. Length Bias Correction in Gene Ontology Enrichment Analysis Using Logistic Regression. PloS one. 2012;7(10):e46128.
4. 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. 2008;105(51):20179–84.
5. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.
6. Pavlidis P, Qin J, Arango V, Mann JJ, Sibille E. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. Neurochem Res. 2004;29(6):1213–22.
7. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. Bioinformatics. 2005;21(13):2988–93.
8. Goeman JJ, Van De Geer SA, De Kort F, Van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics. 2004;20(1):93–9.
9. 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. Proc Natl Acad Sci USA. 2005;102(43):15545–50.
10. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci USA. 2005;102(38):13544–9.
11. Kim SY, Volsky DJ. PAGE: parametric analysis of gene set enrichment. BMC Bioinforma. 2005;6(1):144.
12. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1$\alpha$-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34(3):267–73.
13. Jiang Z, Gentleman R. Extensions to gene set enrichment. Bioinformatics. 2007;23(3):306–13.
14. Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics. 2005;21(9):1943–9.
15. Efron B, Tibshirani R. On testing the significance of sets of genes. Ann Appl Stat. 2007;1(1):107–29.
16. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010;11(2):1–12.
17. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinforma. 2013;14(1):7.
18. 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.

19. Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. 2012;40(17):e133.
20. Wang W, Qin Z, Feng Z, Wang X, Zhang X. Identifying differentially spliced genes from two groups of RNA-seq samples. Gene. 2013;518(1):164–70.
21. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.
22. 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.
23. Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics. 2012;13(3):523–38.
24. Efron B. Large-scale inference: empirical Bayes methods for estimation, testing, and prediction. vol. 1. Cambridge: Cambridge University Press University Printing House; 2010.
25. Pounds S, Morris SW. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of *p*-values. Bioinformatics. 2003;19(10):1236–42.
26. Efron B, et al. Microarrays, empirical Bayes and the two-groups model. Stat Sci. 2008;23(1):1–22.
27. Gold DL, Miecznikowski JC, Liu S. Error control variability in pathway-based microarray analysis. Bioinformatics. 2009;25(17):2216–21.
28. Miecznikowski JC, Gaile DP. A novel characterization of the generalized family wise error rate using empirical null distributions. Stat Appl Genet Mol Biol. 2014;13(3):299–322.
29. Billingsley P. Probability and measure. Hoboken: John Wiley & Sons; 2008.
30. Basu A, Ghosh J. Identifiability of the multinormal and other distributions under competing risks model. J Multivar Anal. 1978;8(3):413–29.
31. Ker AP. On the maximum of bivariate normal random variables. Extremes. 2001;4(2):185–90.
32. Cain M. The moment-generating function of the minimum of bivariate normal random variables. Am Stat. 1994;48(2):124–5.
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):gkv007.
34. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.
35. Ren S, Peng Z, Mao JH, Yu Y, Yin C, Gao X, et al. RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. Cell Res. 2012;22(5):806–21.
36. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9): 1105–11.
37. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. 9. 2012;4:357–9.
38. Langfelder P, Mischel PS, Horvath S. When is hub gene selection better than standard meta-analysis? PloS ONE. 2013;8(4):e61505.
39. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9.
40. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21(16):3439–40.
41. Cheung VG, Nayak RR, Wang IX, Elwyn S, Cousins SM, Morley M, et al. Polymorphic cis- and trans-regulation of human gene expression. PLoS Biol. 2010;8(9):e1000480.
42. Sanchez A, Flamini M, Polak K, Palla G, Spina S, Mannella P, et al. Actin cytoskeleton remodelling by sex steroids in neurones. J Neuroendocrinol. 2012;24(1):195–201.
43. Giretti MS, Simoncini T. Rapid regulatory actions of sex steroids on cell movement through the actin cytoskeleton. Steroids. 2008;73(9):895–900.
44. Li X, Zimmerman A, Copeland NG, Gilbert DJ, Jenkins NA, Yin HL. The mouse thymosin $\beta$4 gene: structure, promoter identification, and chromosome localization. Genomics. 1996;32(3):388–94.
45. Lee Hr, Yoon SY, Kang HB, Park S, Kim KE, Cho YH, et al. Thymosin beta 4 enhances NK cell cytotoxicity mediated by ICAM-1. Immunol Lett. 2009;123(1):72–6.
46. Torikai H, Akatsuka Y, Miyazaki M, Warren EH, Oba T, Tsujimura K, et al. A novel HLA-A* 3303-restricted minor histocompatibility antigen encoded by an unconventional open reading frame of human TMSB4Y gene. J Immunol. 2004;173(11):7046–54.
47. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012;13(2):204–16.
48. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2010;26(1):136–8.
49. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinforma. 2010;11(1):422.
50. Zhou YH, Xia K, Wright FA. A powerful and flexible approach to the analysis of RNA sequence count data. Bioinformatics. 2011;27(19):2672–8.
51. Di Y, Schafer DW, Cumbie JS, Chang JH. The NBP negative binomial model for assessing differential gene expression from RNA-seq. Stat Appl Genet Mol Biol. 2011;10(1):.