BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 19 2014, pages 2747–2756
doi:10.1093/bioinformatics/btu374

Gene expression

Advance Access publication June 5, 2014

Gene set analysis: limitations in popular existing methods and proposed improvements

Pashupati Mishra[1,†,*], Petri Törönen[1,†], Yrjö Leino[2] and Liisa Holm[1]

[1]Institute of Biotechnology, University of Helsinki, Helsinki, Finland and [2]CSC - IT Center for Science, Ltd., Espoo, Finland

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Gene set analysis is the analysis of a set of genes that collectively contribute to a biological process. Most popular gene set analysis methods are based on empirical *P*-value that requires large number of permutations. Despite numerous gene set analysis methods developed in the past decade, the most popular methods still suffer from serious limitations.

Results: We present a gene set analysis method (mGSZ) based on Gene Set Z-scoring function (GSZ) and asymptotic *P*-values. Asymptotic *P*-value calculation requires fewer permutations, and thus speeds up the gene set analysis process. We compare the GSZ-scoring function with seven popular gene set scoring functions and show that GSZ stands out as the best scoring function. In addition, we show improved performance of the GSA method when the max-mean statistics is replaced by the GSZ scoring function. We demonstrate the importance of both gene and sample permutations by showing the consequences in the absence of one or the other. A comparison of asymptotic and empirical methods of *P*-value estimation demonstrates a clear advantage of asymptotic *P*-value over empirical *P*-value. We show that mGSZ outperforms the state-of-the-art methods based on two different evaluations. We compared mGSZ results with permutation and rotation tests and show that rotation does not improve our asymptotic *P*-values. We also propose well-known asymptotic distribution models for three of the compared methods.

Availability and implementation: mGSZ is available as R package from cran.r-project.org.

Contact: pashupati.mishra@helsinki.fi

Supplementary information: Available at http://ekhidna.biocenter.helsinki.fi/downloads/pashupati/mGSZ.html

Received on November 19, 2013; revised on May 22, 2014; accepted on May 30, 2014

1 INTRODUCTION

Inferring biological pathways from high-throughput gene expression datasets that are altered in a biological or a medical test is one of the major challenges in biosciences. The analysis of gene sets instead of individual genes results in significant reduction of noise and dimension and in greater biological interpretability. Furthermore, it enhances the statistical power of tests of association of phenotype with genetic variants by pooling signals of a set of genes linked to the same biological process. Annotation libraries like Gene Ontology (Ashburner *et al.*, 2000), KEGG pathways (Kanehisa and Goto, 2000) and MIPS functional Categories (Ruepp *et al.*, 2004) are some of the popular sources for gene sets. Several methods have been developed for gene set analysis.

According to Goeman and Bühlmann, 2007, gene set analysis methods can be categorized into two major categories, competitive and self-contained. Competitive methods focus on distinguishing the most significant gene sets among a dataset of gene sets (Dørum *et al.*, 2009; Efron and Tibshirani, 2006; Mootha *et al.*, 2003; Newton *et al.*, 2007; Subramanian *et al.*, 2005; Törönen *et al.*, 2009; Wu and Smyth, 2012), whereas self-contained methods focus on analysis of gene sets irrespective of other genes in the dataset (Dinu *et al.*, 2007; Wu *et al.*, 2010). In this article, we focus on competitive gene set analysis methods.

Recent research has addressed gene set analysis problems in datasets with fewer replicates and more than two sample groups (Wu and Smyth, 2012; Wu *et al.*, 2010). However, the fundamental problems in *P*-value calculation for gene set scores remain unaddressed. Most of the methods are still based on empirical *P*-value that requires a large number of permutations or rotations to be calculated accurately.

In this article, we first evaluate various gene set scoring functions for their optimality and then propose improvements to empirical *P*-value calculation. We start with a rigorous assessment and evaluation of popular gene set analysis methods: Gene Set Analysis (GSA) (Efron and Tibshirani, 2006), Gene Set Z-score (GSZ) (Törönen et al., 2009), Allez (Newton et al., 2007), methods based on Kolmogorov–Smirnov (KS) (Mootha *et al.*, 2003) and weighted Kolmogorov–Smirnov (wKS) (Subramanian *et al.*, 2005), and methods based on Wilcoxon Rank Sum (WRS) (Naeem *et al.*, 2012), sum of scores (SUM) (Tian *et al.*, 2005) and sum of squared scores (SS) (Dinu *et al.*, 2007) based on empirical *P*-value obtained from hundred thousand sample permutations. The selection of the methods was based on earlier comparisons (Ackermann and Strimmer, 2009; Naeem *et al.*, 2012). We then test whether empirical null distribution (gene set scores from permuted data) obtained from the compared methods can be modeled by suitable asymptotic distribution models. We point out major limitations in some of the methods and propose improvements on scoring functions, permutation models and *P*-value estimation. We show that the efficiency of GSA can be significantly improved by replacing the

*To whom correspondence should be addressed

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

max-mean statistics with the GSZ scoring function and by implementing asymptotic *P*-value. Similarly, the performance of Allez can be significantly improved by implementing sample permutation and asymptotic *P*-value estimation. The improved versions of GSA and Allez are called mGSA and mAllez, respectively. We show that GSZ outperforms all the compared methods including the improved ones in our evaluation.

We propose an asymptotic method of *P*-value estimation for GSZ, mGSA, mAllez and SUM. Asymptotic *P*-value estimation requires fewer permutations and significantly speeds up the gene set analysis process. Similar work that proposes asymptotic method for *P*-value calculation was done by Knijnenburg *et al*., 2009. Our method is different in that it requires fewer permutations. Irizarry *et al*., 2009 and Kim and Volsky, 2005 proposed fully parametric gene set analysis methods based on normal approximation of the gene set scores (gene independence assumption). However, mGSZ is semi-parametric in the sense that we approximated the empirical null distributions of gene set scores with well-known asymptotic distribution models for asymptotic *P*-value estimation. The implementation resulted in significant improvement in efficiency of the methods. For gene set analysis methods with no known asymptotic model, we implement an improved version of *P*-value calculation based on Phipson and Smyth, 2010.

P-value estimation in permutation-based gene set analysis methods uses either genes or samples or both as sampling unit. Despite its popularity, gene permutation can lead to flawed biological conclusion because of correlation among genes (Goeman and Bühlmann, 2007). For the competitive gene set analysis methods, however, we show that it is crucial to consider the results from gene permutation. Similar to Efron and Tibshirani, 2006, we show that the absence of gene permutation in competitive gene set analysis results in type I error. We argue in favor of using both gene and sample as sampling unit in competitive gene set analysis methods as proposed by Efron and Tibshirani (2006), Tian *et al*. (2005) and Törönen *et al*. (2009).

Once significant gene sets are reported, it is useful to evaluate a gene set in more detail to see the behavior of the gene set. This can be done by visualizing the gene set score profile across the gene list as shown in the original GSEA article (Subramanian *et al*., 2005). In our opinion, it is even more relevant to compare gene set scores from original and permuted data. Original data correspond to differential gene expression test scores calculated from gene expression data with correct sample labels, and permuted data correspond to differential gene expression test scores calculated from gene expression data with permuted sample labels. Therefore, we represent an improvement on the visualization where we show the gene set scores from original data and a summary (e.g. percentiles) of the gene set scores from permuted data. This allows visualization of separation between gene set scores from original and permuted data.

We compared mGSZ with program packages of GSA (Efron and Tibshirani, 2006), Allez (Newton *et al*., 2007) and CAMERA (Wu and Smyth, 2012). We show that mGSZ clearly outperforms the other methods. ROAST (Wu *et al*., 2010) is a state-of-the-art self-contained gene set analysis method. Even though in principle self-contained methods cannot be compared with competitive methods, we compared mGSZ with ROAST to point out the differences in gene set analysis results from the two different approaches. Dørum *et al*., 2009 proposed rotation test for estimation of *P*-values for gene set analysis of datasets with very small sample size. We show that asymptotic *P*-values estimated from permutation test were equally good as that of rotation test with a dataset consisting of only three replicates in each group. Finally, we present a *R* statistical package 'mGSZ' that implements mGSZ and the other compared methods.

2 METHODS

In this section, we describe (i) Gene set scoring functions of the compared gene set analysis methods, (ii) Permutation model used in this work, (iii) Asymptotic method for *P*-value calculation, (iv) Series expansion for *P*-value calculation with EVD and GEVD models, (v) Evaluation tests for gene set scoring functions based on applicability of asymptotic *P*-value estimation, (vi) Evaluation tests for gene set scores based on empirical *P*-values, (vii) Comparison of mGSZ with state-of-the-art program packages, (viii) Evaluation of mGSZ with gene expression dataset with small sample size and (ix) Biological datasets used for the evaluation tests.

2.1 Gene set analysis methods

2.1.1 The Gene Set Z-score GSZ is a gene set scoring function that combines features from overrepresentation and shifted expression-based approaches. This is done by using a hypergeometric enrichment score that is weighted with the differential expression test scores (Törönen *et al*., 2009). GSZ takes as an input a gene list ordered based on differential gene expression test scores and classification of genes as member (MG) and non-member (NG) genes of gene sets (Supplementary Fig. S1). Given an ordered gene list, a threshold is placed in between every consecutive pair of genes. GSZ score is then calculated for every subset of the ordered gene list taken at each of the threshold positions (Supplementary Fig. S1). Let M denote the total number of genes in the ordered gene list and X_i, $i = 1, ..., M$ denote the differential gene expression test scores for i_{th} gene in the list. Let T_i denote the threshold position at i_{th} gene in the list and S_N be a subset of top N genes of the list taken at T_N. Then, GSZ uses the following function to calculate the difference between the sum of differential gene expression test scores for MG and NG genes of the analyzed gene set within the subset,

$$Diff_N = \sum_{\substack{i \in S_N \\ i \in MG}} X_i - \sum_{\substack{i \in S_N \\ i \in NG}} X_i. \quad (1)$$

The calculation is repeated for each threshold position, T_i and the largest absolute value is selected as the GSZ score of the analyzed gene set based on upregulated member genes. The ordered gene list is then inverted and the whole process is repeated analogously starting from the most downregulated genes for the GSZ score of the analyzed gene set based on the downregulated member genes. The largest absolute value of the GSZ scores based on upregulated and downregulated member genes is selected as the GSZ score for the analyzed gene set. The difference calculated in Equation 1 is unstable due to the biases caused by differences in gene set sizes, subset sizes and variances of the gene set scores in different subsets. GSZ method solves this problem by Z-score normalization (Fig. 5),

$$Z = \frac{Diff_N - E(Diff_N)}{\sqrt{D^2(Diff_N) + k}} \quad (2)$$

Where, $E(Diff_N)$ and $D^2(Diff_N)$ are the estimates for the expectation value and the variance, respectively, for $Diff_N$ (Equation 1) under the null hypothesis that the gene set members and non-members are distributed randomly across the gene list and k is a prior variance. The functions to calculate the estimates for the expectation value and the variance

(Equations 3 and 4) are based on two types of functions: (i) the functions that define the expectation value ($E(N_{MG})$) and the variance ($D^2(N_{MG})$) of the number of gene set members in the analyzed subset, and (ii) the functions that define the expectation value ($E(X)$) and the variance ($D^2(X)$) of the differential gene expression test scores in the selected subset of the gene list.

$$E(Diff_N) = 2E(X)E(N_{MG}) - NE(X) \quad (3)$$

$$D^2(Diff_N) = 4\left(\frac{D^2(X)}{N-1}(E(N_{MG})(N - E(N_{MG})) - D^2(N_{MG})) + E(X)^2 D^2(N_{MG})\right) \quad (4)$$

where,

X = differential gene expression test scores of the genes in the subset, S_N;

N_{MG} = number of gene set members in the subset, S_N;

N = total number of genes in the subset, S_N;

$E(X)$ = mean of differential gene expression test scores for the subset;

$D^2(X)$ = variance of differential gene expression test scores for the subset;

$E(N_{MG})$ = mean of the hypergeometric distribution of N_{MG};

$D^2(N_{MG})$ = variance of the hypergeometric distribution of N_{MG};

k = prior variance added to stabilize the scoring function behavior with small subsets.

The estimate of the expected value of $Diff_N$ (Equation 3) is a simple function of the expected value for hypergeometric distribution $E(N_{MG})$, the expected value for the differential gene expression test scores in the analyzed subset $E(X)$ and the size of the subset, N.

The estimate of the variance of $Diff_N$ (Equation 4) is a function of the expected value and the variance of hypergeometric distribution ($E(N_{MG})$, $D^2(N_{MG})$), the mean and the variance for the differential gene expression test scores in the subset of the gene list ($E(X)$,$D^2(X)$) and the size of the subset, N.

For the calculation of prior variance, k, we use the medians of the variance estimates over the ordered gene list obtained with the analyzed gene set and a gene set of size 10 (Supplementary Section S7).

GSZ is similar to the scoring function proposed by Efron and Tibshirani, 2006. Efron and Tibshirani, 2006 proposed a max-mean statistics, where the analyzed gene set is divided into two parts at $X = 0$ (positive regulation and negative regulation). This corresponds to GSZ when the threshold N is placed at $X_N = 0$ (Supplementary Section S8).

The scoring function proposed by Newton *et al*., 2007 calculates the mean value ($\overline{X}$) of the differential expression test scores for all the genes in the analyzed gene set followed by normalization. The calculation of the mean and variance estimates for $\overline{X}$ is practically the same equations as Equations 3 and 4 with the only difference caused by the absence of N. Thus, GSZ is identical to the scoring function proposed by Newton *et al*., 2007 if calculated using the whole gene list.

GSZ is identical to mGSZ with the addition of asymptotic *P*-values. For simplicity, we use the notation 'mGSZ' in the following text.

2.1.2 Modified methods Three of the methods in our evaluation were modified versions of GSZ (Törönen *et al*., 2009), GSA (Efron and Tibshirani, 2006) and Allez (Newton *et al*., 2007). The modifications were—GSZ: implementation of asymptotic *P*-value, GSA: replacement of max-mean statistics with GSZ statistics and implementation of asymptotic *P*-value and Allez: addition of sample permutation for asymptotic *P*-value calculation. We call the modified methods mGSZ, mGSA and mAllez, respectively.

2.1.3 Reference methods We also included methods based on WRS, SUM, sum of squared scores (SS), KS and wKS scoring functions. WRS statistics is the sum of ranks of member genes of the analyzed gene set in the whole gene list. SUM statistics is the simple sum of the differential gene expression test scores of the member genes of the analyzed gene set with background subtraction. SS statistics is simply the squared version of SUM.

2.1.4 Program packages We compared mGSZ, mGSA and mAllez with R package programs of GSA (Efron and Tibshirani, 2006), Allez (Newton *et al*., 2007), CAMERA (Wu and Smyth, 2012) and ROAST (Wu *et al*., 2010).

2.2 Permutation

Gene permutation involves permutation of the genes, whereas sample permutation involves permutation of the class labels. GSZ scoring function uses estimates of mean and standard deviation that takes into account the deviation in the gene expression data in both member and non-member genes. This feature of GSZ scoring function allows to compare the gene set scores calculated from the actual gene expression data with the gene set scores calculated from gene-wise permutations of the gene expression data implicitly without the need of separate runs of gene-wise permutations of the gene expression data. All the reference methods but SUM and SS implement gene permutation implicitly in their scoring functions. All the evaluations except detection of transcription factor (TF) activity in this article are based on sample permutation.

2.3 *P*-value calculation

The minimal obtainable *P*-value by empirical method and its resolution is directly proportional to the number of permutations. Permutation-based gene set analysis involves small *P*-values that require a large number of permutations to be calculated accurately. This is often computationally infeasible. An alternative way to calculate small *P*-values accurately without the need of a large number of permutations is to fit a suitable asymptotic distribution model to the empirical null distribution of gene set scores from permuted data with fewer permutations, estimate parameters of the fitted asymptotic model and use the estimated parameters to calculate *P*-values from the assumed cumulative distribution function. Asymptotic *P*-value calculation in this work is based on Gaussian distribution (*NORM*), Gamma distribution (*GAMMA*), Extreme value type I distribution (*EVD*) (Supplementary Section S1) and General extreme value distribution (*GEVD*) (Supplementary Section S1). Asymptotic distribution models were fitted on the empirical null distribution using distribution fitting methods in R packages, *MASS* (Venables and Ripley, 2002) and *ismev* (Heffernan and Stephenson, 2012). We used *NORM* for modeling gene set scores from SUM and mAllez because sums of normally distributed values should be normally distributed. We used *GAMMA* for modeling gene set scores data with skewed distribution, for example, SS. *EVD* and *GEVD* were used to model gene set score data composed of extreme values, for example, mGSZ and mGSA.

2.3.1 Series expansion *P*-values for *EVD* and *GEVD* are based on extraction which makes it difficult to calculate extremely small *P*-values. We solved this problem by deriving series expansions on the logarithm of a *P*-value for *EVD* and *GEVD*. Log *P*-values for *EVD* and *GEVD* are calculated as:

$$F(x) = -log(P\text{-}value_{EVD}) = -ln(1 - e^{-e^x}) \quad (5)$$

where,

$$x = -(z - \mu)/\beta \quad (6)$$

z is absolute mGSZ or mGSA score value for the analyzed gene set, and μ and β are location and scale parameters, respectively, for *EVD*.

and

$$G(x) = -log(P\text{-}value_{GEVD}) = -ln\left(1 - e^{-(1+ax)^{-1/a}}\right), a > 0 \quad (7)$$

where,

$$x = (z - \mu)/\beta \tag{8}$$

z is absolute mGSZ or mGSA score value for the analyzed gene set, and μ, β and a are the location, scale and shape parameters, respectively, for *GEVD*.

Expressions on the right hand side of Equations 5 and 7 are readily evaluated by ordinary numerical software for relatively small absolute values of x. However, already for values of the order of $|x| \approx 30$ the outer exponentials become so small that the results are distorted by rounding errors. To achieve more precise results, we derived the following expansions of $F(x)$ and $G(x)$ (Supplementary Section S2):

$$F^s(x) = -x + e^x/2 - e^{2x}/24 + e^{4x}/2880 - e^{6x}/181440 + \ldots \tag{9}$$

$$G^s(x) = \frac{\ln(1+ax)}{a} + \frac{(1+ax)^{-1/a}}{2} - \frac{(1+ax)^{-2/a}}{24} + \frac{(1+ax)^{-4/a}}{2880} - \frac{(1+ax)^{-6/a}}{181440} - \ldots \tag{10}$$

The implementation of series expansion for calculation of log *P*-value was based on a prespecified threshold:

$$\log(p_{\mathrm{EVD}}) = \begin{cases} F(x) & \text{if } x \geq -5 \\ F^s(x) & \text{if } x < -5 \end{cases} \tag{11}$$

$$\log(p_{\mathrm{GEVD}}) = \begin{cases} G(x) & \text{if } x < 5 \textit{ and } (1+ax) > 0 \\ G^s(x) & \text{if } x \geq 5 \textit{ and } (1+ax) > 0. \end{cases} \tag{12}$$

The prespecification of the threshold was based on the number of terms picked from the series expansion and the amount of error allowed. The number of terms used in our analysis is four, and for F(x) this means that the approximation error is at most of the order $1.78 * 10^{-15}$.

2.4 Datasets

2.4.1 TF deletion and overexpression data This dataset consists of 907 *Escherichia coli* microarrays taken from M3D Database (Faith *et al.*, 2008). The preprocessed (log_2 fold-change) dataset was obtained from Naeem *et al.*, 2012. The dataset has no biological replicates. The dataset consisted of knockout and overexpression experiments for 17 TFs targeting 949 genes.

2.4.2 P53 cancer data Protein p53 is a tumor suppressor protein that prevents development of cancer cells. The p53 dataset consists of 33 samples with mutated p53 gene and 17 samples with wild type p53 gene (Subramanian *et al.*, 2005).

2.4.3 Gender data Gender data consist of mRNA expression profiles from lymphoblastoid cell lines derived from 15 males and 17 females (Subramanian *et al.*, 2005).

2.4.4 Leukemia data Leukemia dataset consists of gene expression profiles of cells from 24 acute lymphoid leukemia patients and 24 acute myeloid leukemia patients (Armstrong *et al.*, 2002).

2.5 Evaluation tests

2.5.1 Asymptotic approximation of empirical null distribution This test was designed to test the applicability of asymptotic *P*-value calculation in the compared methods. Gene set scores were calculated with the compared scoring functions with 500 sample permutations. Asymptotic distribution models (Section 1) were fitted to the empirical null distributions of the gene set scores and asymptotic *P*-values were calculated. Similarly, gene set scores were also calculated with the compared scoring functions with 100 000 sample permutations and empirical *P*-values were calculated. Empirical *P*-values calculated from 100 000 sample permutations were used as a reference of truth for evaluation of asymptotic *P*-values. Pearson correlation (*cor*) between log asymptotic *P*-values calculated from 500 sample permutations and log empirical *P*-values calculated from 100 000 sample permutations was calculated. Also, mean squared error (*mse*) of log asymptotic *P*-values calculated from 500 sample permutations was calculated against log empirical *P*-values calculated from 100 000 sample permutations. Although correlation indicates the magnitude and direction of a linear relationship between test *P*-values and the reference of truth, mean squared error indicates a difference between test *P*-values and the reference of truth. In addition to the whole list, the analysis was repeated on the signal rich region of *P*-values by calculating *cor* and *mse* also for the subset of *P*-values <0.10. This emphasizes the biologically interesting regions of the gene list. An optimality criterion was defined for the tested asymptotic models. Asymptotic models with *mse* <0.10 and *cor* >0.97 in both the whole list and the signal rich regions were considered to be optimal models for asymptotic *P*-value estimation.

2.5.2 Comparison of P-value estimation methods A comparison of asymptotic and empirical *P*-values was done for methods for which the empirical null distribution could be approximated with suitable asymptotic models. Asymptotic *P*-values were calculated with 500 sample permutations and compared with empirical *P*-values calculated with 500, 1000 and 2000 sample permutations. Correlation and mean squared error of the *P*-values were calculated against empirical *P*-values obtained from 100 000 sample permutations.

2.5.3 Evaluation of gene set scoring functions The objective of this evaluation was to rank the gene set scoring functions of the compared methods based on their overall performance on four different evaluation tests. The evaluation was based on empirical *P*-values calculated from 100 000 sample permutations.

Detection of TF activity. We evaluated the performance of the compared methods on a TF deletion and overexpression dataset. Targets of the TFs were treated as gene sets. The experimental perturbations of TFs provide the standard-of-truth for the evaluation. The experimentally perturbed TFs were considered positive gene sets. An ideal method should identify the perturbed TFs. The compared methods were run on TF dataset and TFs were assigned empirical P-values calculated from 100 000 gene permutations. Gene permutation was implemented because the dataset has no biological replicates. The TFs were ranked based on the empirical P-values. Area under curve (AUC) was calculated for each experiment. The compared methods were ranked based on the mean and standard *deviation* of the AUC.

Detection of relevant gene sets. Results obtained from TF deletion and overexpression data were verified using p53 and gender data. Forty gene sets strongly relevant to protein p53 and 10 gene sets strongly relevant to gender were selected and considered as relevant gene sets (Supplementary Tables S4 and S5). Gene sets reported by each of the methods were ranked based on empirical *P*-values calculated with 100 000 sample permutations. Subset of top gene sets (50 in p53 data and 20 in gender data) was taken. Relevant gene sets in the subset were labeled as 1 and irrelevant gene sets were labeled as 0. The cumulative sum was calculated for the subset.

False-positive signal test. The performance of the compared methods was further analyzed with real biological data but with null gene sets. The null gene sets were generated by randomly choosing the member genes while keeping the gene set sizes intact. The idea was to test if

some of the methods have a tendency to create false-positive results when gene sets have no biological signal.

Stability of gene set scores. GSZ scoring function is similar to KS-based scoring functions in that it calculates scores running across an ordered gene list. This allows us to compare mGSZ with KS and wKS based on the gene set score profile across an ordered gene list. Gene set score profiles of the top ranked gene set reported by wKS were calculated for original as well as 1000 sample permuted data with all the three methods. Instead of 1000 gene set score profiles, we plotted seven percentiles of gene set score profiles from permuted data.

2.6 Comparison of mGSZ with program packages

The main objective of this comparison was to test the performance of mGSZ as compared with some of the state-of-the-art gene set analysis methods. We also included mGSA and mAllez in the comparison to show the improved performance as compared with their original versions, GSA (Efron and Tibshirani, 2006) and Allez (Newton *et al*., 2007). The comparison was based on three tests: (i) Evaluation of top gene sets, (ii) False-positive signal test and (iii) *P*-value test where we compare the log *P*-values reported by each of the compared methods.

2.7 Evaluation of mGSZ on dataset with small sample size

Here, we tested the applicability of mGSZ to gene expression datasets with few replicates in each sample group. Gene set analysis methods based on empirical *P*-values are not suitable for datasets with few sample replicates (e.g. three), because the minimum obtainable *P*-value by empirical method is restricted by the small number of permutations that can be obtained from fewer sample replicates. Unlike permutation, number of rotations is not dependent on sample size (Dørum *et al*., 2009). For this evaluation, we developed two versions of mGSZ: one that implements permutation for generating empirical null distribution and the other that implements rotation for the same. In both cases, *P*-value was estimated asymptotically by fitting suitable asymptotic model to the empirical null distribution. We refer the two versions of mGSZ as mGSZpermutation and mGSZrotation, respectively. The only difference between the two versions is the number of permutations/rotations that can be generated for the empirical null distribution.

Both mGSZpermutation and mGSZrotation were implemented on a simulated gene expression datasets. For the simulated gene expression datasets, we generated data on 20 000 genes and six samples. The first three samples are the control group and the second three samples are the treatment group. Each consecutive non-overlapping block of 20 genes was considered to be a gene set with the total of 1000 gene sets. Each gene expression value was generated as i.i.d. $N(0,1)$. Then, the constant 1.0 was added to the first 10 genes in the every 10th gene set starting from 1st until 490th gene set. Thus, we generated 50 positive gene sets in the simulated dataset. The comparison of the results was based on evaluation of the top gene sets and *P*-value test, where we compare the log *P*-values reported by each of the methods.

3 RESULTS

3.1 Summary of the results

Table 1 summarizes the results obtained from the evaluation of the compared gene set scoring functions and *P*-value estimation methods mentioned in Section 2. Sections 3.2–3.4 describe the results in detail. Table 1 is based on approximate ranks. We use ties when the difference between methods is insignificant. Smaller ranks across the column denote good performance of a method, whereas bigger ranks denote the opposite. Stability test was relevant only to mGSZ, KS and wKS because the scoring functions of the other methods do not calculate running scores. mGSA leads mGSZ in the detection of TF activity with negligible difference. KS-based method showed surprisingly good empirical *P*-values leaving mGSZ as second best performer. However, KS lags behind in the other tests. Overall, mGSZ outperforms the other methods (Table 1).

3.2 Asymptotic approximation of empirical null distribution

We tested four distribution models—*EVD*, *GEVD*, *NORM* and *GAMMA*—with gene set scores from mGSZ of which *EVD* and *GEVD* turned out to be the optimal models based on the optimality criterion (Section 2.5.1) (Table 2). *GAMMA* and *NORM* did not pass the optimality criterion (Section 2.5.1) (Table 2). The same distribution models were tested with mGSA data and surprisingly, while *EVD*, *GEVD* and *GAMMA* turned out to be the optimal models, *NORM* did not pass the optimality criterion (Section 2.5.1) (Table 2). *NORM* was the best fitting model for SUM and mAllez data (Section 2.5.1) (Table 2). In case of mGSZ and mGSA, we selected the models that produced *P*-values with maximum *cor* and minimum *mse* with the reference of truth in both the whole data and the subset. In case of

Table 1. Summary of results from various evaluation tests

Evaluations	Datasets	mGSZ	mGSA	mAllez	WRS	SUM	SS	KS	wKS
P-value asymptotics	P53, gender	Yes	Yes	Yes	No	Yes	No	No	No
P-value estimation	P53	2	4	5	6	1	8	3	7
Detection of TF activity	TF	2	1	5	8	6	3	4	7
Detection of relevant gene sets	P53	1	3	6	7	4	2	8	5
	Gender	1	3	6	5	8	2	4	7
False-positive signal analysis	Leukemia	1	1	1	1	7	8	1	1
	P53	1	1	1	1	1	1	1	1
Stability of the scoring functions over gene list	P53, gender	1	*	*	*	*	*	3	2

Notes: First row indicates either presence (Yes) or absence (No) of asymptotic models for the methods, whereas in the rest of the rows, numbers indicate approximate rank of the methods based on the test results.
*Sign indicates that a test is not applicable to a particular method.

Table 2. Asymptotic distribution models fitted on the empirical null distribution generated by the compared methods with p53 data and their results

Methods	Models	*mse*	*mse(subset)*	*cor*	*Cor(subset)*	Comments
mGSZ	EVD	0.005	0.03	0.99	0.99	EVD and GEVD are the optimal models (EVD selected)
	GEVD	0.02	0.09	0.99	0.99	(Problem in parameter optimization for GAMMA
	GAMMA	0.07	**0.36**	0.98	0.98	distribution)
	NORM	**1.68**	**10.65**	**0.89**	**0.95**	
mGSA	EV	0.007	0.05	0.99	0.98	EVD, GEVD and Gamma are the optimal models
	GEVD	0.02	0.1	0.99	0.98	(EVD selected)
	GAMMA	0.009	0.06	0.99	0.98	
	NORM	**0.27**	**2.03**	0.98	**0.95**	
mAllez	NORM	0.01	0.03	0.99	0.98	Optimal (selected)
WRS	NORM	0.01	0.04	**0.97**	**0.95**	Not optimal
SS	GAMMA	**0.31**	**1.11**	**0.96**	**0.95**	Problem in parameter optimization
SUM	NORM	0.008	0.02	0.99	0.99	Optimal (selected)
KS	–	–		–		No suitable asymptotic model found
wKS	–	–		–		No suitable asymptotic model found

Notes: Scores not meeting the criteria for optimal model are highlighted.

mGSZ and mGSA, *EV* was selected as the model for asymptotic *P*-value calculation (Table 2),whereas *NORM* was the model of choice for asymptotic *P*-value calculation for mAllez and SUM (Table 2). Similar results were observed from the analysis with gender data (Supplementary Section S5).

3.3 *P*-value estimation methods: asymptotic and empirical

The results were highly in favor of asymptotic *P*-values for mGSZ, mGSA, mAllez and SUM with higher correlation and lower mean squared error, as compared with the empirical *P*-values (Table 3). As shown in Table 3, asymptotic *P*-values (*EV*) calculated against 500 permutations have very low *mse* and very high *cor* when compared with empirical *P*-values calculated against 100 000 permutations. This implies that 500 permutations are enough for asymptotic *P*-values.

3.4 Evaluation of gene set scoring functions

3.4.1 Detection of TF activity mGSZ and mGSA were the top ranking methods in identifying correct TF activity (Table 4). Analysis of five quantiles of AUC scores across the experiments in TF dataset showed that mGSZ and mGSA had the highest AUC score in the lowest quantile (0 quantile) (Fig. 1).

3.4.2 Detection of relevant gene sets *Test on p53 cancer data.* Top 50 gene sets reported by each of the compared methods were analyzed as described in section 2.5.3 (Fig. 2). mGSZ was the top ranking method, with SS being the closest competitor (Fig. 2).

Test on gender data. Top 20 gene sets reported by each of the compared methods were analyzed as described in section 2.5.3. The results were similar to that of p53 data (Fig. 3).

3.4.3 False-positive signal analysis We randomized curated gene sets from Subramanian *et al*., 2005 as described in Section 2.5.3 and analyzed their enrichment in the leukemia dataset

Table 3. Comparison of *P*-value calculation methods

Methods	*P*-value methods	*mse*	*mse (subset)*	*cor*	*cor (subset)*
mGSZ	***EV (500 perms)***	***0.005***	***0.03***	***0.99***	***0.99***
	EMP (500 perms)	0.06	0.31	0.95	0.88
	EMP (1000 perms)	0.04	0.22	0.97	0.93
	EMP (2000 perms)	**0.03**	**0.15**	**0.98**	**0.96**
mGSA	***EV (500 perms)***	***0.007***	***0.05***	***0.99***	***0.98***
	EMP (500 perms)	0.02	0.1	0.98	0.96
	EMP (1000 perms)	0.01	0.06	0.99	0.97
	EMP (2000 perms)	**0.007**	**0.04**	**0.99**	**0.98**
mAllez	***Gaussian (500 perms)***	***0.01***	***0.03***	***0.99***	***0.98***
	EMP (500 perms)	0.04	0.13	0.96	0.89
	EMP (1000 perms)	0.02	0.08	0.98	0.94
	EMP (2000 perms)	**0.01**	**0.05**	**0.99**	**0.96**
SUM	***Gaussian (500 perms)***	***0.008***	***0.02***	***0.99***	***0.99***
	EMP (500 perms)	0.04	0.14	0.96	0.89
	EMP (1000 perms)	0.03	0.1	0.97	0.93
	EMP (2000 perms)	**0.02**	**0.06**	**0.98**	**0.95**
WRS	**EMP (500 perms)**	**0.01**	**0.04**	**0.98**	**0.92**
	EMP (1000 perms)	0.006	0.02	0.99	0.97
	EMP (2000 perms)	**0.003**	**0.01**	**0.99**	**0.98**
SS	**EMP (500 perms)**	**0.11**	**0.38**	**0.95**	**0.91**
	EMP (1000 perms)	0.07	0.26	0.96	0.94
	EMP (2000 perms)	**0.05**	**0.17**	**0.98**	**0.95**
KS	**EMP (500 perms)**	**0.004**	**0.02**	**0.99**	**0.95**
	EMP (1000 perms)	0.002	0.01	0.99	0.98
	EMP (2000 perms)	***0.0005***	***0.003***	***0.99***	***0.99***
wKS	**EMP (500 perms)**	**0.02**	**0.18**	**0.97**	**0.91**
	EMP (1000 perms)	0.01	0.11	0.98	0.95
	EMP (2000 perms)	***0.01***	***0.08***	***0.99***	***0.96***

Note: Note that asymptotic *P*-value calculation methods outperform empirical *P*-value calculation methods. The comparison was based on p53 data. Bold and italic values highlight the difference between the *P*-values calculated with empirical method with maximum number of permutations (2000) in the comparison and *P*-values calculated with asymptotic method with minimum number of permutations (500) in the comparison.

Table 4. Table showing average and standard deviation AUC of eight methods across all the experiments in TF deletion and overexpression dataset

No.	Scoring functions	AUC	STD
1	mGSA	0.9058	0.0725
2	mGSZ	0.9045	0.0698
3	SS	0.8843	0.1454
4	KS	0.8694	0.1099
5	mAllez	0.8683	0.1232
6	SUM	0.8593	0.1273
7	wKS	0.8562	0.1644
8	WRS	0.8138	0.1591

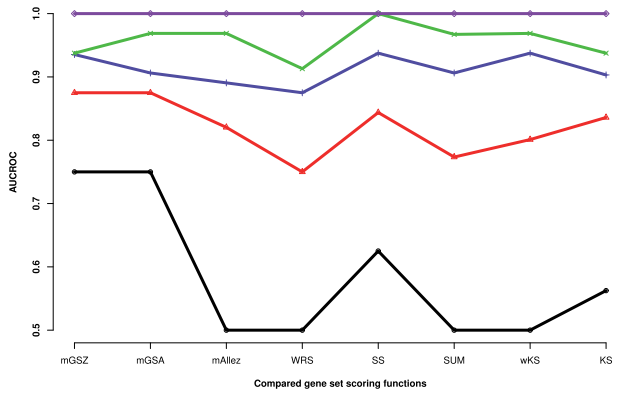

Fig. 1. Summary of *AUCROC* scores of the compared methods in TF dataset. The lines indicate five quantiles: black—0.00, red—0.25, blue–0.50, green–0.75 and purple–1.00

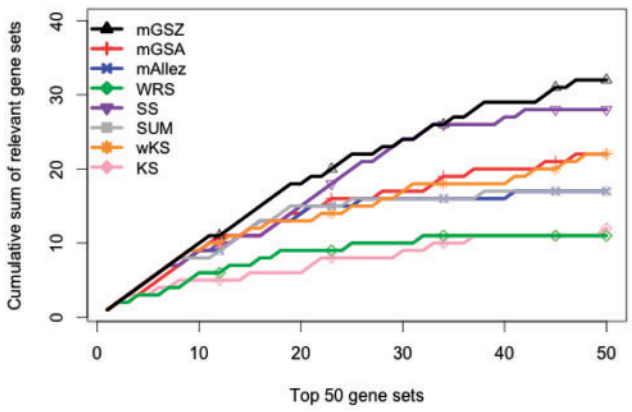

Fig. 2. Relevant gene sets identified by the compared methods. Figure represents cumulative count of biologically relevant gene sets (Supplementary Table S4) over the ranked list of top 50 gene sets reported by each of the compared methods. mGSZ (black) shows the best performance

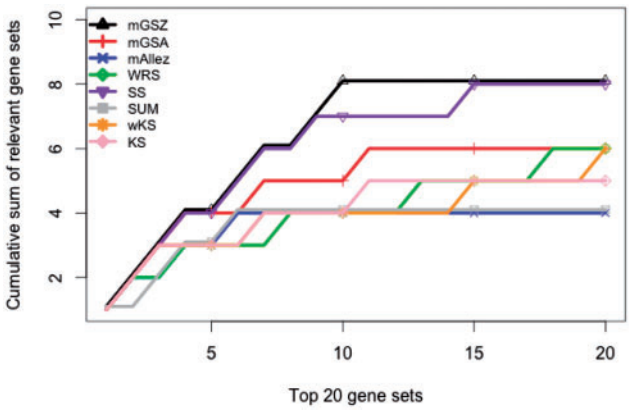

Fig. 3. Relevant gene sets identified by the compared methods. Figure represents cumulative count of biologically relevant gene sets (Supplementary Table S5) over the ranked list of top 20 gene sets reported by each of the compared methods. mGSZ (black) shows the best performance

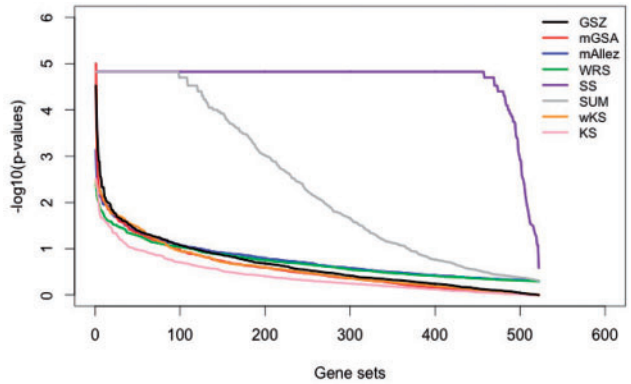

Fig. 4. Evaluation of methods with randomized gene sets. mGSZ and most other methods show quite similar behavior. SS and SUM, however, show strong noise signal

(Armstrong *et al.*, 2002) with the compared methods. Leukemia data are signal rich data and hence suitable to test methods that are sensitive to high signals. Surprisingly, SS identified >85% of the null gene sets as significant (Fig. 4). SUM is the second-worst performing method in this test (Fig. 4).

In addition to the leukemia data, we also performed the analysis on randomized p53 dataset with Gene Ontology genesets (Ashburner *et al.*, 2000). Even though mGSZ, mGSA, wKS and KS are more conservative, the difference to the other methods is significantly smaller (Supplementary Section S10).

3.4.4 Stability of gene set scores Stability of the gene set scores from mGSZ, KS and wKS was compared using p53 cancer data as described in Section 2.5.3. The stability plot for mGSZ, based on permuted data shows that the middle part of the seven percentiles of the gene set score profiles of the permuted data stay quite stable across gene list positions (Fig. 5a). The minimum percentile shows smaller deviation from zero than the maximum percentile (Fig. 5a). This indicates that gene set scores coming from overrepresentation show stronger signal than those coming from underrepresentation. The maximum of the

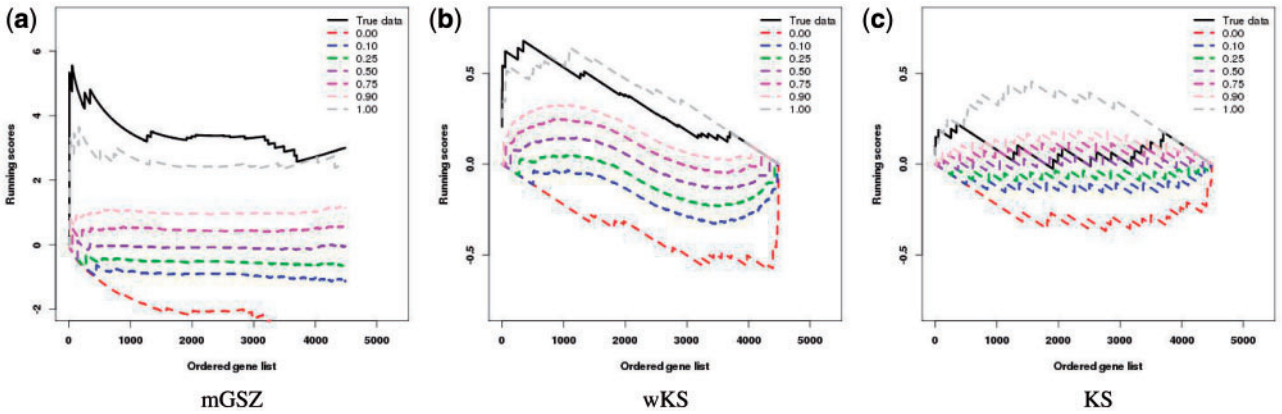

Fig. 5. Visualization of scoring function profiles for mGSZ, wKS and KS for gene set 'p53hypoxiaPathway' in p53 data. Figures also show seven percentiles (0, 10, 25, 50, 75, 90 and 100) of gene set score profiles obtained from the permuted data. Notice the clear separation of the gene set score profiles from original data with that of permuted data and stability of gene set score profiles from permuted data in case of GSZ-score

Table 5. Summary of results from various evaluation tests performed for comparison of mGSZ, mGSA and mAllez with the program packages (Section 2.1.4)

Evaluations	Datasets	mGSZ	mGSA	GSA	mAllez	Allez	CAMERA	ROAST
Detection of relevant gene sets	p53	1	2	3	6	7	4	5
	Gender	1	4	6	5	7	2	3
P-value resolution	p53	1	2	6	3	*	4	5
	Gender	1	3	4	2	*	5	6
False-positive signal analysis	Leukemia	2	2	2	2	*	1	6

Notes: Numbers indicate approximate rank of the methods based on the test results. *Sign indicates that a test is not applicable to a particular method.

gene set scores from original data comes quite early in the gene list, pointing that gene set members occur near the top of the ordered gene list (Fig. 5a). Notice that the separation between gene set score profile from original data and permuted data is quite clear across different threshold positions (Fig. 5a). The importance of our visualization is more highlighted when we analyze the behavior of wKS and KS with the same gene expression dataset and the same gene set. For wKS the gene set score profile from permuted data shows strange behavior with S-shaped profile (Fig. 5b) and for KS the middle area shows strong variation (Fig. 5c). The separation between gene set score profile from permuted data and original data is weak in both wKS and KS (Fig. 5c). This comparison is based on the strongest gene set reported by wKS.

Stability plot for mGSZ, wKS and KS based on gender data is discussed in Supplementary Section S6.

3.5 Comparison of mGSZ, mGSA and mAllez with program packages

We summarize the results of the comparison of mGSZ, mGSA and mAllez with program packages in Table 5. The notation used in Table 5 is same as in the Table 1. The results are described in detail in Supplementary Section S11. Two datasets (p53 and gender) were used for all evaluation tests except for false-positive signal analysis. For false-positive signal analysis, leukemia dataset was used for similar reason as mentioned in Section 2.5.3. CAMERA showed the best results in false-positive signal analysis. However, in the remaining two evaluation tests, mGSZ outperformed the other methods.

3.6 Evaluation of mGSZ on dataset with small sample size

No significant difference was observed between log *P*-values reported by mGSZrotation and mGSZpermutation with the simulated dataset (Supplementary Section S12, Supplementary Fig. 16a and b). We also evaluated the top gene sets reported by the two methods in both the simulated datasets, and the results were similar (Supplementary Section S12, Supplementary Figs 17a and b).

3.7 Evaluation of permutation methods

Table 6 demonstrates the effects of the use of gene or/and sample permutation methods in competitive gene set analysis methods. The table is the summary of the results from false-positive signal analysis of the compared gene set scoring functions (Section 3.4.3, Fig. 4) and comparison of mAllez with program package of Allez based on detection of relevant gene sets (Section 3.5, Table 5). The methods implementing only sample permutation

Table 6. Comparison of gene and sample permutation methods based on false-positive signal analysis and detection of relevant gene sets

Methods	Permutation		False-positive signal	Detection of relevant gene sets	Reference
	Sample	Gene			
mGSZ	Present	Present implicitly	Low	–	Figure 4
mGSA	Present	Present implicitly	Low	–	Figure 4
mAllez	Present	Present implicitly	Low	Improved as compared with Allez	Figure 4, Table 5
WRS	Present	Present implicitly	Low	–	Figure 4
SS	Present	Absent	High	–	Figure 4
SUM	Present	Absent	High	–	Figure 4
wKS	Present	Present implicitly	Low	–	Figure 4
KS	Present	Present implicitly	Low	–	Figure 4
Allez	Absent	Present	–	Ranked as worst	Table 5
GSA	Present	Present	–	Ranked lower than mGSZ and mGSA	Table 5

report high false-positive signal (Table 6). Note the ranks of mAllez and Allez based on detection of relevant gene sets in Table 5. The improved performance of mAllez as compared with Allez is clearly due to the addition of sample permutation (Table 6). Interestingly, GSA that explicitly implements both gene and sample permutations, ranks lower than mGSZ and mGSA (Table 6). This suggests that implicit implementation of gene permutation is at least as good as explicit implementation.

4 DISCUSSION

Gene set analysis methods aim to identify *a priori* defined gene sets associated with biological pathways that are altered in a biological or medical test. Despite the decade long research into the field, some of the most popular gene set analysis methods still suffer from serious limitations that need to be addressed and rectified. In this work, we have accessed and evaluated popular gene set analysis methods, pointed out limitations in the scoring functions, permutation methods and *P*-value calculation methods and proposed improvements.

mGSZ and mGSA are the best performers in detection of TF activity in TF data (Section 3.4.1). However, in case of p53 and gender data, mGSZ detected the highest number of relevant gene sets, closely followed by SS (Section 3.4.2). Overall, mGSZ stands out as the best performer in detection of relevant gene sets. Note that the evaluation tests, detection of TF activity in TF data and relevant gene sets in p53 and gender data were based on empirical *P*-values estimated with 100 000 permutations. Only the gene set scoring functions were varied in the compared methods. So, the variations in results were direct consequences of the gene set scoring functions.

Despite the good performance in detection of relevant gene sets with p53, gender and TF data, SS as well as SUM had significantly high type 1 error with randomized leukemia dataset (Section 3.4.3). SS and SUM are the only methods in our comparison that lack gene permutation. Because of that the methods do not compare the member and non-member genes of the analyzed gene set and thus lose the essence of competitive GSA. Note that self-contained gene set analysis methods do not require gene permutations as they calculate gene set scores with no reference to genes other than member genes of the analyzed gene sets. The remaining six methods have gene permutation included implicitly in their scoring functions and take into consideration the results from gene permutations implicitly. This clearly points to the importance of considering results from gene permutation in competitive gene set analysis methods (Section 3.7, Table 6) as suggested by Efron and Tibshirani, 2006, Tian *et al.*, 2005 and Törönen *et al.*, 2009.

EVD and *GEVD* turned out to be the best fitting models for mGSZ data which is quite natural considering that the mGSZ score is the highest absolute value in a profile. For mGSA, *EVD*, *GEVD* and *GAMMA* turned out to be the best fitting models. In case of mAllez and SUM, *NORM* was the best fitting model. Asymptotic *P*-values calculated for mGSZ, mGSA, mAllez and SUM scores with the best fitting models with 500 sample permutations were more accurate than empirical *P*-values calculated with 2000 sample permutations, clearly demonstrating the superiority of asymptotic *P*-value over empirical *P*-value (Section 3.3). Surprisingly, despite the poor performance on the other tests, empirical *P*-value estimation method for KS with 500 sample permutations shows exceptionally high accuracy (Section 3.3).

Comparison of mGSZ with the state-of-the-art gene set analysis methods showed the best overall performance of mGSZ. mGSZ is the best method among the compared methods based on the results from the detection of relevant gene sets from p53 and gender datasets and comparison of resolution of the *P*-values assigned to the top gene sets from p53 and gender datasets (Section 3.5). The comparison of mGSZ with ROAST was included to illustrate the differences between self-contained and competitive methods with a suitable dataset. In a signal rich dataset like leukemia, self-contained methods like ROAST fail to provide biologically conclusive results. Randomization of gene sets in leukemia dataset randomizes the member genes of the gene sets. However, as 79.9% of the individual genes are differentially expressed in the leukemia data (Dinu *et al.*, 2007), ROAST identifies a significantly large number of null gene sets as significant. This effect is not seen in competitive methods as competitive methods compare the member genes with non-

member genes of the analyzed gene sets, and thus the results remain stable in datasets with varying level of biological signals. Interestingly, both mGSA and mAllez rank higher than GSA and Allez in detection of the relevant gene sets in p53 and gender dataset. Higher rank of mGSA as compared with GSA suggests that replacement of max-mean statistics in GSA with GSZ scoring function and implicit implementation of gene permutation improved the performance of the method. Note that GSA implements gene permutation explicitly. Whereas, higher rank of mAllez as compared with Allez suggests that implementation of sample permutation in addition to gene permutation in Allez improved the performance of the method, once again highlighting the importance of implementing both gene and sample permutations in competitive gene set analysis (Section 3.7, Table 6).

Negligible difference was observed between the asymptotic *P*-values reported by mGSZ with rotation test and mGSZ with permutation test (Section 3.6) with simulated gene expression dataset with three replicates in each sample group. In addition, both the methods identified same number of positive gene sets in the simulated dataset. The results suggest that mGSZ with permutation test is applicable to gene expression datasets with sample size as small as three.

Improved visualization of scoring function profiles of a gene set (the most significant gene set reported by wKS) in the case of mGSZ, KS and wKS revealed a major pitfall in KS based methods, not seen in the original visualization used by Subramanian *et al*., 2005 (Section 3.4.4).

Based on the results of our evaluation, we therefore propose: (i) use of GSZ scoring function instead of max-mean statistics in GSA (Efron and Tibshirani, 2006), (ii) implementation of both gene and sample permutation in competitive gene set analysis methods and (iii) use of asymptotic *P*-values for the methods based on GSZ, mGSA, mAllez and SUM. These improvements are implemented in our R package, mGSZ. mGSZ takes about 10 min for a gene set data with 522 gene sets and p53 data with 500 permutations in Mac book Pro 10.7.5.

ACKNOWLEDGEMENT

We thank Robert Küffner for providing preprocessed TF deletion and overexpression data.

Funding: This work was supported by Biocentrum Helsinki, Finland.

Conflicts of Interest: none declared.

REFERENCES

Ackermann,M. and Strimmer,K. (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10**, 47.

Armstrong,S.A. *et al.* (2002) Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet*., **30**, 41–47.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet*., **25**, 25–29.

Dinu,I. *et al.* (2007) Improving gene set analysis of microarray data by sam-gs. *BMC Bioinformatics*, **8**, 242.

Dørum,G. *et al.* (2009) Rotation testing in gene set enrichment analysis for small direct comparison experiments. *Stat. Appl. Genet. Mol. Biol*., **8**, Article34.

Efron,B. and Tibshirani,R. (2006) On testing the significance of sets of genes. *Ann. Appl. Stat*., **1**, 107–129.

Faith,J.J. *et al.* (2008) Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. *Nucleic Acids Res*., **36**, D866–D870.

Goeman,J.J. and Bühlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Heffernan,J.E. and Stephenson,A.G. (2012) *ismev: An Introduction to Statistical Modeling of Extreme Values*. Vienna, Austria: R foundation for Statistical Computing. R package version 1.39.

Irizarry,R.A. *et al.* (2009) Gene set enrichment analysis made simple. *Stat. Methods Med. Res*., **18**, 565–575.

Kanehisa,M. and Goto,S. (2000) Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **28**, 27–30.

Kim,S.Y. and Volsky,D.J. (2005) Page: parametric analysis of gene set enrichment. *BMC Bioinformatics*, **6**, 144.

Knijnenburg,T.A. *et al.* (2009) Fewer permutations, more accurate p-values. *Bioinformatics*, **25**, i161–i168.

Mootha,V.K. *et al.* (2003) Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet*., **34**, 267–273.

Naeem,H. *et al.* (2012) Rigorous assessment of gene set enrichment tests. *Bioinformatics*, **28**, 1–7.

Newton,M.A. *et al.* (2007) Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann. Appl. Stat*., **1**, 85–106.

Phipson,B. and Smyth,G.K. (2010) Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol*., **9**, 39.

Ruepp,A. *et al.* (2004) The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*., **32**, 5539–45.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Tian,L. *et al.* (2005) Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci. USA*, **102**, 13544–13549.

Törönen,P. *et al.* (2009) Robust extraction of functional signals from gene set analysis using a generalized threshold free scoring function. *BMC Bioinformatics*, **10**, 307.

Venables,W.N. and Ripley,B.D. (2002) *Modern Applied Statistics with S-plus*. 4th edn. Springer, New York. ISBN 0-387-95457-0.

Wu,D. and Smyth,G.K. (2012) Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res*., **40**, e133.

Wu,D. *et al*. (2010) Roast: rotation gene set tests for complex microarray experiments. *Bioinformatics*, **26**, 2176–2182.