4378–4391 *Nucleic Acids Research, 2013, Vol. 41, No. 8* *Published online 26 February 2013*
doi:10.1093/nar/gkt111

Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods

Leif Väremo, Jens Nielsen and Intawat Nookaew*

Department of Chemical and Biological Engineering, Chalmers University of Technology, Gothenburg 412 96, Sweden

Received October 11, 2012; Revised January 31, 2013; Accepted February 3, 2013

ABSTRACT

Gene set analysis (GSA) is used to elucidate genome-wide data, in particular transcriptome data. A multitude of methods have been proposed for this step of the analysis, and many of them have been compared and evaluated. Unfortunately, there is no consolidated opinion regarding what methods should be preferred, and the variety of available GSA software and implementations pose a difficulty for the end-user who wants to try out different methods. To address this, we have developed the R package Piano that collects a range of GSA methods into the same system, for the benefit of the end-user. Further on we refine the GSA workflow by using modifications of the gene-level statistics. This enables us to divide the resulting gene set *P*-values into three classes, describing different aspects of gene expression directionality at gene set level. We use our fully implemented workflow to investigate the impact of the individual components of GSA by using microarray and RNA-seq data. The results show that the evaluated methods are globally similar and the major separation correlates well with our defined directionality classes. As a consequence of this, we suggest to use a consensus scoring approach, based on multiple GSA runs. In combination with the directionality classes, this constitutes a more thorough basis for an enriched biological interpretation.

INTRODUCTION

The analysis of genome-wide expression data typically involves the task of compiling a list of the statistical significance of all genes over multiple conditions, enabling the identification of differentially expressed genes. A number of techniques and methods have been developed for this step of the gene expression analysis, but arriving at a list of significant genes, however, does not alone necessarily facilitate the biological interpretation of the data. To manually go through a list of individual genes in an attempt to map gene regulation to metabolic and biological functions and pathways is by far a simple task. To overcome this, methods based on statistical hypothesis tests have been developed that shift the analysis from individual genes to sets of genes. Gene set analysis (GSA), also generally referred to as gene set enrichment analysis (1–3) or gene set testing (4), has the advantage of incorporating existing biological knowledge into the expression analysis. As an example of a common approach, Gene Ontology (GO) terms (5) can be used to define gene sets, thus enabling the identification of, e.g. statistically significant biological processes, through the use of GSA. Gene sets are not restricted to GO terms, as they can be defined in an unlimited number of ways, correlating to anything from metabolic or signaling pathways to transcription factors and chromosomal positions: it is up to the researcher to use relevant gene sets for the question at hand.

The variety of GSA methods available can roughly be divided into three groups. First, a classic approach is to use a contingency table to test whether a gene set is overrepresented by a predefined list of significant genes, e.g. using Fisher's exact test, the χ^2 test or the hypergeometric test. These methods are popular among GO-based tools (6), including BiNGO (7), DAVID (8), GOstat (9) and GOEAST (10). A major drawback of this approach is the requirement of an *a priori* cut-off of gene significance. Second, a number of methods do not start from a gene list, but rather from the raw expression data using multivariate and global tests. Examples include Goeman's global test (11), Hotelling's T^2 (12), ANCOVA (13,14) and MANOVA (15). Third, a large group of the

*To whom correspondence should be addressed. Tel: +46 31 772 3854; Fax: +46 31 772 3801; Email: intawat@chalmers.se

GSA methods start from a list of gene-level statistics and, based on these statistics, calculate a gene set statistic for each gene set being analyzed. The popular gene set enrichment analysis (GSEA), first introduced by Mootha *et al.* (16) and subsequently improved and implemented by Subramanian *et al.* (17), is an example of such a method.

The third group of methods will be the focus of this article, for several reasons. Unlike the first group of methods, this approach does not require any *a priori* significance cut-off. Further on, by starting from user-defined gene-level statistics, these methods are not limited to a specific type of data. Microarray (continuous) as well as RNA-seq (discrete) data can be used for the analysis. Although we use gene expression throughout this article, it is also possible to base the gene-level statistics on genome-wide association studies [see, e.g. (18)], proteomics or metabolomics data, further increasing the flexibility of these GSA methods.

Several authors have reviewed and compared many of the available GSA methods (1–4,19–23); however, most of the articles only consider a handful of the methods at a time. Unfortunately for the end-user, although some recommendations exist, there is no unified opinion regarding what methods perform the best. Partly, this is perhaps because of the lack of a gold standard. Huang *et al.* (23) even suggest that, for the best results, the user should try multiple GSA tools, even those with similar statistical approaches, as different implementations may affect the results. Although a valid point, this procedure is far from practical for the general research community, as the vast amount of available tools, partly overlapping in terms of statistical methods, are implemented on a range of different platforms and programming languages, requiring the user to have high computer and programming skills, as well as plenty of time.

Ackermann and Strimmer (3) provide a modular framework that describes the key steps of GSA, where one important step is assessing the significance of each gene set. A common way to do this is to estimate the background distributions of the gene set statistics either by randomizing the genes or the sample labels. Here as well, the literature is unclear on the best procedure. For instance, Li *et al.* (24) prefer randomizing the genes, Goeman and Bühlmann (4) suggest randomizing the sample labels, whereas Nam and Kim (19) recommend using both approaches if possible.

After successfully assessing the significance of all gene sets, one is faced with the task of interpreting the GSA results. The meaning of the gene set *P*-values depends on the null hypothesis, which differs between methods and is connected to the choice of gene set statistic as well as the significance assessment method. This has previously been discussed by several authors (3,4,19,25), and Goeman and Bühlmann (4) introduced the distinction between competitive and self-contained tests. A competitive test identifies gene sets that are significantly affected by differential expression compared with the rest of the genes. On the other hand, a self-contained test considers only the genes in the gene set, comparing the association of differential expression to a given phenotype with that of randomly selected phenotypes. This is relatively clear; however, it is not always obvious for the user which definition of 'differential expression' that different methods use. Is it differential expression in general, without taking the direction of regulation into consideration, or is there a distinction between up- and downregulation? Further on, will a gene set affected by both up- and downregulation be cancelled out, or detected as significant? As an example, Hung *et al.* (1) suggest using the absolute values of the gene-level statistics, thus removing the information about the direction of change. In this case, a statistically significant gene set should be considered affected by differential expression in general, disregardful of whether the genes in the set are up- or downregulated. On the other hand, using the average of the gene-level statistics of a gene set as the gene set statistics would only detect gene sets that are affected by differential expression in a distinct direction.

In an attempt to address the practical issues discussed in the previous paragraphs, we refined the GSA workflow and introduced some new concepts. Our new workflow takes advantage of incorporating multiple GSA methods, simplifying the comparison of methods as well as laying the ground for a consensus scoring of gene sets. Furthermore, we introduce a new concept of handling the gene-level statistics to divide the resulting gene set *P*-values into three directionality classes: non-directional, mixed-directional and distinct-directional. In combination, this classification of the results can aid the interpretation of the biological meaning of a significant gene set. Our workflow is fully implemented as an R package for the community to use, collecting a range of GSA methods into the same platform and framework, thus reducing the required effort of the user to test different GSA methods.

The aim of this article is 3-fold: first, to introduce our new GSA workflow, including the handling of gene-level statistics, directionality classification of gene sets and consensus scoring. Second, to compare a range of selected GSA methods, as well as to investigate the impact of the individual components of GSA. This is carried out using microarray as well as RNA-seq data, from both yeast and human. Third, we wish to demonstrate the benefits brought into the analysis by consensus scoring and, in particular, the directionality classification of the results.

MATERIALS AND METHODS

In the following text, the separate steps of the GSA workflow will be described, starting with the methods for calculating gene set statistics, followed by the significance estimation. Next, the modification of gene-level statistics will be described together with the concept of the directionality classes, as well as the consensus scoring approach. Finally, the processing of the data sets used in this study is described. More details are available in the Supplementary Methods.

Gene set statistics

Based on literature review, we identified 11 methods fulfilling our criteria, and their definitions of gene set statistics are used in our workflow. These GSA methods take a list of gene-level statistics as input (e.g. *P*-values, *t*-values

or fold-changes), giving the user free flexibility of the process of going from raw expression data to assessing differential expression. Further on, these methods do not require any *a priori* cut-off of gene significance, thus making use of all the data. The 11 methods for calculating gene set statistics are listed next, together with a brief description (see Supplementary Methods for more details).

Fisher's combined probability test

Fisher's combined probability test (26) is a classical meta-analysis method that combines *P*-values from multiple individual statistical tests. The method is based on the sum of log transformed gene-level *P*-values. The significance of each Fisher gene set statistic can be estimated from a χ^2 distribution or by a permutation approach, as described in the next section.

Stouffer's method

Similar to Fisher's method, Stouffer's method (27) also combines *P*-values, but calculates the gene set statistic using the inverse normal cumulative distribution function. The significance of a Stouffer gene set statistic can be estimated using the normal cumulative distribution function, or by a permutation approach.

Reporter features

The reporter features method was first conceptually described by Patil and Nielsen (28) and generally extended by Oliveira *et al.* (29). At its core, the reporter features method is identical to Stouffer's method; however, the gene set statistics are corrected for the background distribution before significance estimation. The correction is performed by subtracting the mean and dividing by the standard deviation of randomly calculated gene set statistics. Once corrected, the significance of the reporter features gene set statistics is calculated in the same manner as for the Stouffer's method.

Parametric analysis of gene set enrichment

Kim and Volsky (30) developed parametric analysis of gene set enrichment (PAGE) as an alternative to the popular GSEA. The PAGE gene set statistic is based on the mean of the gene-level statistics of a gene set and corrected for the background, represented by all gene-level statistics. As for Stouffer's method and the reporter features method, the significance of a PAGE gene set statistic can be estimated from the cumulative normal distribution.

Tail strength

The tail strength method, proposed by Taylor and Tibshirani (31), is based on the mean of a set of gene-level *P*-values. Initially the genes are ranked according to their significance of differential expression, and the ranks are subsequently used as weights in the calculation of the gene set statistics. No theoretical null distribution can be used for the significance calculations, as in the previous methods. Instead, the null distribution is estimated using a permutation approach.

Wilcoxon rank-sum test

The Wilcoxon rank-sum test is a common choice for a non-parametric alternative of a significance test. The gene set statistic is calculated based on the sum of the ranks of the gene-level statistics, and the significance can be estimated from a normal distribution or using a permutation approach. The Wilcoxon rank-sum test is used for GSA in the popular limma R-package (32).

Gene set enrichment analysis

GSEA (16,17), as mentioned in the 'Introduction' section, is probably one of the most widespread methods for GSA. It uses a ranked list of gene-level statistics, so that significant genes are located at the two ends of the list (based on direction of differential expression). Next, a running sum is computed, starting with the first statistic and moving to the last. For each statistic that belongs to the gene set of interest, the sum is increased, otherwise it is decreased. The gene set statistic is the maximum deviation from zero of the running sum. The value that increases the running sum can be adjusted by a parameter. If the parameter is set to zero, the gene set statistic equals a Kolmogorov–Smirnov statistic; however, in this article we set the parameter to one, as recommended by the GSEA authors.

Mean, median and sum

Apart from the Wilcoxon rank-sum test, the limma R-package also offers the alternative to simply set the gene set statistic as the mean of the gene-level statistics, for the genes belonging to that gene set. Here, we also include the alternative to use the median or sum of the gene-level statistics. In all cases, a permutation approach can be used to estimate the gene set significance.

Maxmean statistic

Finally, Efron and Tibshirani (33) propose to use, what they call the maxmean statistic. This approach separates the positive and negative gene-level statistics belonging to a gene set, and for each of the two subsets, it calculates the absolute sum divided by the total number of genes in the set. The gene set statistic is then defined as the maximum of these two numbers. The significance, as for all the previous methods, can be calculated through a permutation approach. This will be explained in some more detail in the following section.

Assessing gene set significance

Each gene set statistic can be converted into a *P*-value that estimates the statistical significance of that gene set. By definition, the *P*-value of a gene set statistic is the probability to observe a new gene set statistic that is equal to or more extreme than the given gene set statistic. This probability can be estimated provided a null distribution, i.e. the probability distribution of the gene set statistic. For 5 of the 11 gene set statistics, theoretical null distributions, defined by continuous functions, can be used to estimate the *P*-values. These approaches are described in the Supplementary Methods. In all 11 cases, the null distributions can also be estimated by a permutation approach. This approach can be performed in two ways, either by randomizing the genes, referred to as gene sampling, or by randomizing the sample labels, referred to as sample permutation.

Gene sampling is carried out for each gene set by randomly taking a sample of genes (of the same number as in the gene set) and recalculating the gene set statistic. This is repeated a large number of times (e.g. 10 000 times) to give a discrete null distribution. The gene set *P*-value is simply the fraction of random gene set statistics that are equal to or more extreme (in general larger) than the original gene set statistic. It follows that the resolution of the gene set *P*-values (the number of possible values) is dependent on the number of permutations used.

Sample permutation is similar to gene sampling; however, in this case, the original sample labels are randomized, and all the gene-level statistics, and subsequently all the gene set statistics, are recalculated based on the new labeling. This procedure is also repeated a large number of times, and the *P*-values are calculated in the same way as described for the gene sampling.

The choice of permutation approach is tightly connected to the underlying null hypothesis. When using gene sampling, the association of a gene set with the phenotype is compared with the association of the rest of the genes to the phenotype. This is termed a competitive null hypothesis. On the other hand, by using sample permutation, the association of a gene set to the phenotype is compared with its association to random phenotypes. This is termed a self-contained null hypothesis. For a deeper understanding of the permutation approaches and the competitive and self-contained null hypothesis, please refer to the articles by Tian *et al*. (25) and Goeman and Bühlmann (4).

Classifying the results in terms of directionality of gene expression changes

The result of a GSA is a list of gene set *P*-values, indicating the significance of each gene set. In this article, we define three directionality classes: non-directional, mixed-directional and distinct-directional. Depending on the underlying statistical test, the resulting gene set *P*-values are assigned to the appropriate class. The non-directional class contains gene set *P*-values where the information about direction of differential expression is omitted, so that significant gene sets can be interpreted as affected by differential expression in general. For the mixed-directional class, a gene set can be significantly affected by differentially expressed genes in either or both directions. Two *P*-values for each gene set are reported, one for each direction. In the case of the mixed-directional class, a gene set can be both significantly affected by up- and downregulation, if it contains two subsets of genes that are coordinately regulated in opposite directions. Finally, the distinct-directional class aims to identify gene sets that are significantly affected by regulation in a distinct direction. Here as well, two *P*-values for each gene set are reported, one for each direction. If a gene set contains significantly differentially expressed genes in both directions, they will cancel out, and neither of the two distinct-directional *P*-values will be significant.

Modifications of gene-level statistics

The type of gene-level statistics and the chosen gene set statistic takes part in determining what kind of directionality class the analysis results in. Also, by modifying the gene-level statistics, *P*-values for different directionality classes can be calculated. If the gene-level statistics are *P*-values, they can be used unmodified. As they do not contain any directional information, the resulting gene set *P*-values will belong to the non-directional class. It is also possible to make two subsets of the gene-level statistics according to information on direction of regulation and run separate runs for each subset. In this case, the resulting gene set *P*-values will belong to the mixed-directional class. Finally, to use all data, but incorporating the directional information, we propose to perform a *P*-value transformation. In this step, the *P*-values of all the upregulated genes are scaled between 0 and 0.5, by dividing by 2. Next, the *P*-values of all the downregulated genes are scaled according to 1-p/2, so that they range between 0.5 and 1, and so that the significance order is swapped. The result will be that the most significantly upregulated genes will have values close to zero, whereas the most significantly downregulated genes will have values close to one. Running the GSA analysis with the transformed and scaled *P*-values will identify gene sets that are significantly affected by distinct upregulation. A second run, using *P*-values that are transformed and scaled in a reverse manner, will identify gene sets that are significantly affected by distinct downregulation.

A similar approach can be taken if the gene-level statistics are *t*-values (or similar). By using the absolute values the directional information can be discarded, and by subsetting, the mixed-directional class can be calculated. The *t*-values naturally contain information about direction; therefore, no transformation similar to the one for the gene-level *P*-values has to be performed. A more specific description of which modifications are used for each gene set statistic can be found in the Supplementary Methods.

Consensus scoring of gene sets

By using various combinations of gene-level statistics, gene set statistics, significance estimation methods and directionality classes, different unique GSA runs can be performed. Each run will produce a list of gene set *P*-values for some or all of the directionality classes. To achieve a consensus result, the different gene set *P*-value vectors belonging to the same class are aggregated to produce a consensus score for each gene set and class. The aggregation is based on ranking the gene sets according to their *P*-value and using rank aggregation approaches to yield a consensus score for each gene set. Two simple approaches are to use either the mean or the median of the ranks of a given gene set as the consensus score. We also include classical rank aggregation methods proposed by Borda (34) and Copeland (35). If multiple gene sets have identical *P*-values, they are given the minimum rank, i.e. the *P*-values (0.001, 0.002, 0.002, 0.002 and 0.003) would be given the ranks (1, 2, 2, 2 and 5), as input to the rank aggregation methods.

Robustness analysis

To investigate the robustness of the consensus scoring approach, two analyses were performed. First, the consistency of the consensus scores were evaluated when the input GSA runs originated from different numbers of gene permutations, i.e. using gene set *P*-value vectors with different resolution. Second, the robustness with regard to randomly selecting different subsets of GSA runs to aggregate was investigated. The detailed approach of the robustness analysis is described in the Supplementary Methods.

Data sets and processing

We use two data sets to evaluate the workflow. The first data set is the human diabetes microarray data from Mootha *et al.* (16), comparing muscle biopsies from 17 normal glucose tolerant men to those from 18 men with type 2 diabetes. From the authors, we acquired normalized expression values for the 35 samples and used the R package limma (36) to fit linear models to each of the genes to describe the expression levels for the two groups. Further on, an empirical Bayes approach, as implemented in the *eBayes* function, was used to assess differential expression for each gene between the two groups. The resulting *t*-values, *P*-values and fold-changes were used in our gene set analysis. First, we used GO terms as gene sets, with at least 5 mapped gene-level statistics and at most 1000 mapped gene-level statistics. In this case, 6030 gene sets passed the size limits, and 17 016 of 22 283 gene-level statistics were mapped to at least one gene set. For the permutation-based significance assessment approaches, 1000 permutations were used. Second, as a case study, we used the same 149 gene sets that were originally used by Mootha *et al* (16). These gene sets include 113 metabolic pathways and 36 clusters of co-regulated genes. In this case, 9120 gene-level statistics were mapped to at least one gene set. For the significance assessment, 10 000 permutations were used for gene sampling and 1000 permutations for sample permutation.

The second data set is from a study comparing different RNA-seq–based transcriptome analyses with each other as well as with microarrays (37). For this, a case study was used comparing the transcriptome of *Saccharomyces cerevisiae* cultivated under batch conditions and in chemostats. From the authors, we acquired gene-level *P*-values and fold-changes for the comparison of the two groups. These statistics were acquired both from their microarray analysis and from their RNA-seq analysis. For the latter, the analysis with Stampy and Cuffdiff was used. For the microarray-based comparison, we also acquired *t*-values. GO terms were used as gene sets, with the same size limits as stated previously. In this case, 1436 gene sets passed the size limits, and 5439 of 5662 gene-level statistics were mapped to at least one gene set. For the gene sampling approach, 1000 permutations were used.

RESULTS

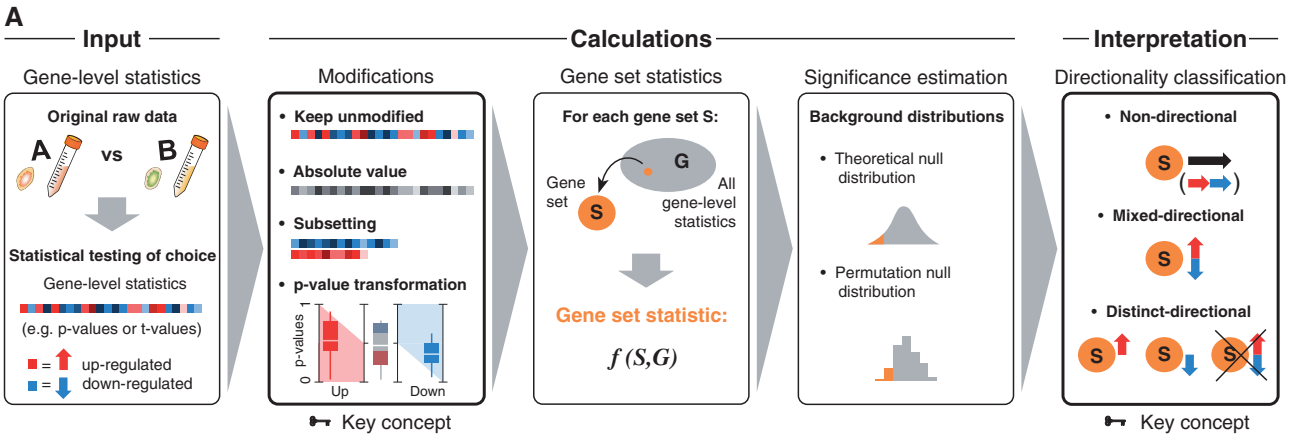
As outlined in the 'Introduction' section, GSA suffers from some practical issues. In summary, (i) there is no consolidated opinion regarding which gene set statistics are to be preferred; (ii) it is unclear whether to use sample permutation or gene sampling for significance assessment and if there is a great difference; (iii) from a biological point of view, it may not be obvious what a significant gene set means in terms of directionality; and (iv) the multitude of methods implemented are naturally spread among different platforms and programming languages, making it laborious for the average user to test different methods. To address these issues, we set out to refine the GSA workflow and use this to investigate the impact of the different components of GSA on the results.

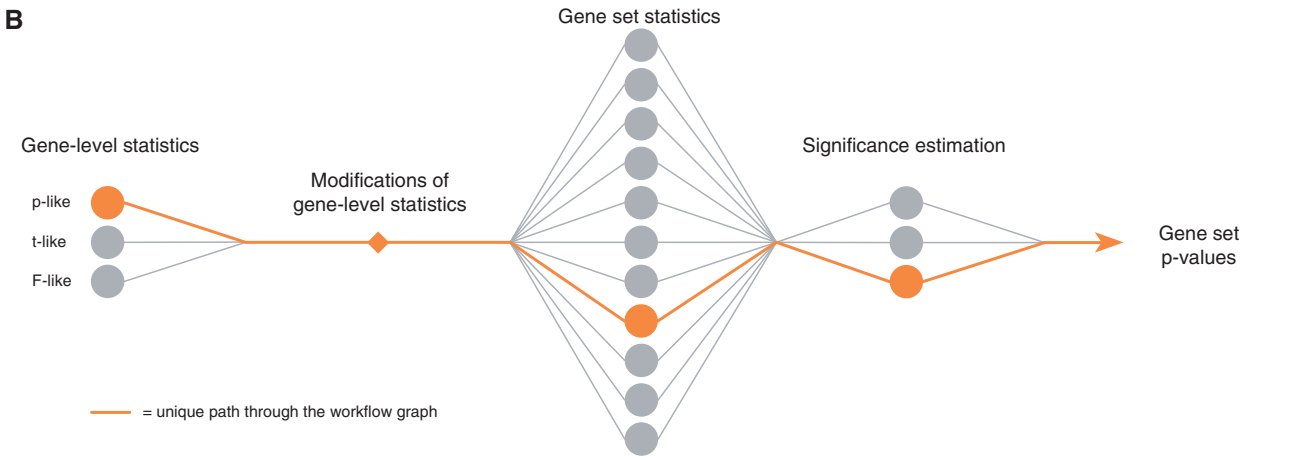
The refined GSA workflow

An overview of the refined GSA workflow is presented in Figure 1A. As described in the 'Materials and Methods' section, the input to the GSA is gene-level statistics originating from any statistical test of the user's choice. Next, the gene-level statistics are modified as possible (unmodified, absolute values, subsetting and *P*-value transformation) as described in the 'Materials and Methods' section, and one analysis is run for each modification. The purpose of the modifications is to enable parallel GSA runs resulting in gene set *P*-values of different directionality classes. For each possible modification, including the unmodified gene-level statistics, a gene set statistic is calculated according to one of the 11 possible methods. The significance of each gene set statistic is calculated either by using a theoretical null distribution (if possible for that statistic) or by a permutation approach (either gene sampling or sample permutation). The resulting gene set *P*-values will belong to one of the three directionality classes that we introduce in this article (non-directional, mixed-directional and distinct-directional). Gene sets that are found to be significant in the non-directional class should be interpreted as affected by differential expression in general, disregardful of if the direction of regulation have components of both up- and downregulation. Gene sets that are found significant in the mixed-directional class are given two *P*-values, one testing for upregulation and the other for downregulation. Here, a gene set can be significantly affected by upregulation disregardful of the extent of downregulation. As a consequence, in the mixed-directional class, a gene set can be found significant in both the upregulated and downregulated part. Finally, as a balance, the distinct-directional class detects gene sets that are significantly affected by differential expression in one distinct direction. A gene set that contains mainly significantly differentially expressed genes, but in both directions, will not turn up significant in this class.

The directionality classification is determined by the type of gene-level statistics and gene set statistics that are used as well as which modification of the gene-level statistics that is applied. The purpose of the classes is to aid the user in the interpretation step. By running the GSA with a combination of settings, it is possible for a gene set to be assigned *P*-values in each class, giving more information about direction and extent of regulation, than a single *P*-value would.

A GSA according to our workflow can be depicted as a unique path through a graph, as in Figure 1B. However,





C

Gene-level statistics			Gene set statistics	Significance estimation			Directionality classes		
p-like	t-like	F-like		Theoretical null distr.	Gene sampling	Sample permutation	Non-directional	Mixed-directional	Distinct-directional
			Parametric methods						
✓			Fisher's combined probability test	✓	✓	✓	✓	✓	
✓			Stouffer's method	✓	✓	✓	✓	✓	✓
✓			Reporter features	✓	✓	✓	✓	✓	✓
	✓		PAGE	✓	✓	✓			✓
			Rank based methods						
✓			Tail strength		✓	✓	✓	✓	✓
✓	✓	✓	Wilcoxon rank-sum test	✓	✓	✓	✓	✓	✓
	✓		GSEA		✓	✓			✓
			Other methods						
✓	✓	✓	Mean		✓	✓	✓	✓	✓
✓	✓	✓	Median		✓	✓	✓	✓	✓
✓	✓	✓	Sum		✓	✓	✓	✓	✓
	✓		Maxmean		✓	✓	✓		

Figure 1. Overview of the refined GSA workflow. (**A**) The key steps in the workflow. Starting from user-defined gene-level statistics, and their possible modifications, gene set statistics are calculated according to the 11 available methods. Significance of the gene sets is estimated based on a theoretical null distribution or a permutation approach. The final gene set *P*-values will be separated into the defined directionality classes, facilitating the interpretation step. (**B**) A graph representation of our GSA workflow. Each path through the graph corresponds to a unique run of the GSA. The nodes represent the number of settings that can be made at each step. The resulting directionality classes depend on the combination of settings made. (**C**) Not all paths, or combinations of settings, are allowed. The table shows which settings are possible for each gene set statistic method. The highlighted settings correspond to the highlighted path in B and will result in two of the three possible *P*-value classes.

not all paths are possible, some restrictions exist depending on which gene set statistic is used. Figure 1C summarizes for each type of gene set statistic, which gene-level statistic it takes, which type of significance estimation that is possible to use, and which directionality classes it is possible to calculate *P*-values for.

The directionality classification clearly separates the results

We were interested in investigating the importance of the different steps of GSA and how they influence the results. We applied our workflow to the human diabetes microarray data from Mootha *et al.* (16), comparing 17 normal

glucose tolerant men to 18 men with type 2 diabetes. Using GO terms as gene sets, we ran our workflow with all possible combinations of settings (analogous to paths through the graph), using gene-level *P*-values and *t*-values as input. This resulted in 127 unique runs, and thus 127 unique gene set *P*-value vectors, each representing the results from a unique combination of gene-level statistics, gene set statistic types, significance estimation procedures and directionality classes. By performing principle component analysis (PCA) on the 127 vectors, each containing *P*-values for 6030 GO gene sets, the impact of the different components of GSA, on the variance in the resulting *P*-values, can be explored. Figure 2A shows a plot of the two first principle components, together explaining 79.7% of the variance. The main separation of the data seems to be consistent with our classification of *P*-values, shown by the five differently colored clusters. As expected, the runs testing for upregulated gene sets are separated from the downregulated ones, and the mixed-directional *P*-value class is closer to the non-directional, compared with the distinct-directional class. The latter is expected because the mixed-directional class may contain gene sets that are significantly affected by both up- and downregulation. These sets will also be significant if direction of regulation is disregarded, i.e. the non-directional class, whereas the gene sets will be non-significant in the distinct-directional class.

To further investigate the data, the directionality class factor was removed by dividing the 127 *P*-value vectors into five groups according to the clusters and rerunning the PCA on each of these groups. In Figure 2A, the smaller PCA plot of the non-directional class shows that the choice of gene set statistic partly explains the variance in the data within a class, illustrated by the scattered groups 1–6 (see Supplementary Figure S1 for the other classes). We also observe some separation depending on the choice of gene-level statistics: the mean and sum gene set statistics are separated into groups 3 and 5, respectively, depending on the gene-level statistic used. The factor that separates the data the least is the choice of significance estimation method (theoretical null distribution, gene sampling or sample permutation) as displayed by the black, green and purple boxes within each of the six scattered groups. Apparently, the choice of method for significance estimation plays a minor role from a global view.

To complement the principle component analysis, we also calculated the Spearman correlation for each pair of the 127 gene set *P*-value vectors. A heatmap of these results is shown in Figure 2B (see Supplementary Figure S2 for detailed row and column labeling). Based on their correlation, the runs were hierarchically clustered, and again, the five main clusters correlate to the directionality classes. The correlation heatmap also illustrates the relations between the different classes. The distinct-directional (up) class shows a strong inverse correlation to the distinct-directional (down) class and to the mixed-directional (down) class. The mixed-directional (up) class does, however, not show an inverse correlation to the mixed-directional (down) class, which is expected. Finally, the non-directional class correlates well to the mixed-directional class, but not to the distinct-directional class. Again, from a global view, focusing on the five larger squares making up the diagonal of the heatmap, we can conclude that the different GSA runs correlate well. In particular, we can lift out the correlation of the results when choosing between gene sampling and sample permutation. Supplementary Table S1 shows that the correlation between gene sampling and sample permutation for all GSA methods and directionality classes is very high. The average correlation is 0.98 and the minimum correlation, representing the worst case, is 0.96.

To not be constrained by one data set, we also applied our workflow on expression data from a case study comparing transcriptional differences in *S. cerevisiae* when cultivated in batch or chemostat, using both microarrays and RNA-seq (37). This allows us to investigate how the GSA results are affected by the use of two different platforms for assessing gene expression.

First, however, we noted that Fisher's and Stouffer's method, using theoretical null distributions for significance assessment, yield remarkably small *P*-values, in particular for large gene sets. This is true for the *S. cerevisiae* data sets, but not for the human diabetes data set (Supplementary Figure S3). The distributions of gene-level *P*-values of the human diabetes data, the *S. cerevisiae* microarray data and the *S. cerevisiae* RNA-seq data are compared in Supplementary Figure S4. Fisher's and Stouffer's method will consistently assign higher significance to larger gene sets if the gene-level *P*-value distributions are skewed towards small values. To illustrate this, Figure 3A shows boxplots of five simulated gene-level *P*-value distributions with increasing proportion of small *P*-values. The distribution of *P*-values is modeled by a mixture of the β distribution and the uniform distribution (38) (see Supplementary Methods for details). Figure 3B shows scatterplots of the gene set *P*-values, when using each of the five distributions as input, where it is obvious that Fisher's and Stouffer's methods show a clear dependency to gene set size. This property can be overcome by using the permutation-based null distributions in place of the theoretical ones. As the two *S. cerevisiae* data sets have *P*-value distributions that are similar to the rightmost of the simulated distributions (as opposed to the human diabetes data, which are similar to the leftmost distribution), the size-dependency of Fisher's and Stouffer's method will pose a problem. To avoid this, we simply removed the runs, in which Fisher's or Stouffer's method was used in combination with a theoretical null distribution for significance estimation before the continuing analysis.

Once these runs were removed, a PCA was run on the remaining 128 gene set *P*-value vectors, containing gene set *P*-values for 1436 GO terms. Figure 3C shows a plot of the first two principle components, together explaining 74.1% of the variance. Each point represents a run with a unique combination of platforms (RNA-seq or microarray), gene-level statistics, gene set statistic types, significance estimation procedures (theoretical null distribution or gene permutation) and directionality classes. Even though microarray and RNA-seq data are mixed, the results are well in line with those from the human

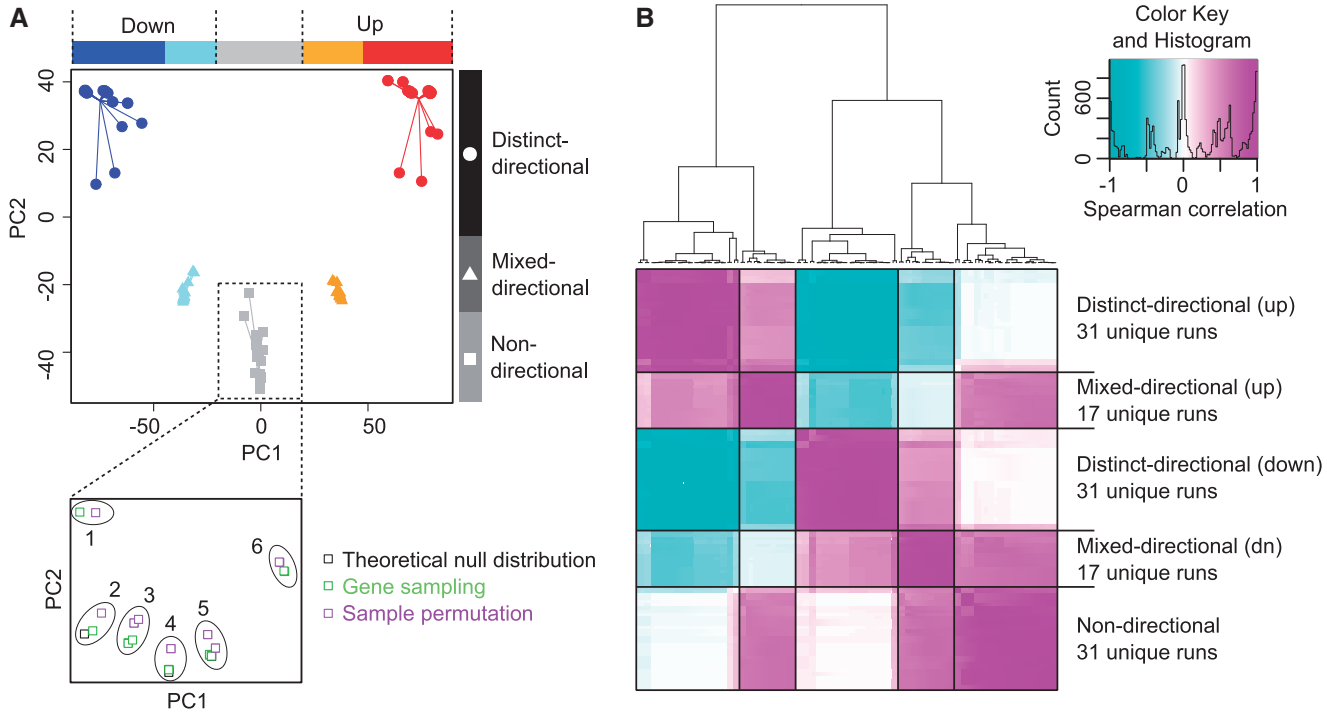


Figure 2. (**A**) A PCA plot of the 127 gene set *P*-value vectors from different runs of the GSA workflow using the human diabetes data. The two first principle components capture 79.7% of the variance. The major separation of the results is consistent with the defined *P*-value classes. The small plot shows the first two principle components of the 31 non-directional *P*-value vectors. The choice of gene-level statistic and gene set statistic seems to have a small importance, shown by the separation into six groups (1: maxmean; 2: Fisher's method; 3: tail strength, mean of *t*-values, sum of *t*-values; 4: reporter features, Stouffer's method; 5: Wilcoxon, mean of *P*-values, sum of *P*-values; 6: median). The choice of significance assessment method seems to play a minor role to the results. (**B**) A heatmap of the Spearman correlation of each pair of the 127 gene set *P*-value vectors. From a global view, the runs within each *P*-value class are highly correlated, shown by the five diagonal squares.

diabetes data aforementioned. The runs cluster according to the directionality classes, although there seems to be a minor separation of the two platforms within each cluster.

Consensus gene set scores by rank aggregation

Our observations from running the GSA workflow for all combinations of settings, are based on human and yeast expression data from both microarrays and RNA-seq. From this we conclude that the resulting gene set *P*-values do not vary much between runs with different settings, in fact they correlate well. From a biological point of view, in terms of interpreting the top significant gene sets, it is more relevant to use the information given by the combination of the directionality classes, than focusing on selecting one method. As a consequence of these two points, we propose to use all possible runs, or a larger subset of these, and combine the results for each directionality class. This means aggregating the information given by each of the five clusters rather than arbitrarily choosing one of the methods to base the biological interpretation on. The concept is outlined in Figure 4.

To aggregate the results, each gene set *P*-value vector is transformed into a rank vector, in total assigning each gene set with a set of ranks, one for each run. This turns the aggregation into a ranked voting problem, i.e. the process of selecting candidates based on ranked lists from multiple voters (in this case the different runs). This is not an uncommon problem, typically encountered during political elections, and thus there exists many methods to choose among. Naeem *et al.* (2) suggests using the average rank as a consensus score, in the case study provided later in the text, we instead use the median rank to tolerate outlier bias. In the Piano package, we also include the options of using the classical consensus methods proposed by Borda (34) and Copeland (35), respectively (see Supplementary Table S2 for a comparison). The reason to use the ranks rather than working with the *P*-values directly is to give equal weight to conservative methods (with few low *P*-values) and less conservative methods (with many low *P*-values).

The directionality classification enriches the biological interpretation

As a case study and an example of how the directionality classification and consensus scoring can enrich the biological interpretation, we used the human diabetes data set along with the 149 gene sets that were originally used for these data by the authors of the GSEA article (16). With this input we reran our workflow, yielding 127 runs. For each run sharing the same directionality class, the gene sets were ranked according to their *P*-values, and these ranked lists were aggregated using the median rank as a consensus score. Figure 5A shows boxplots of the gene set ranks given by each run for the distinct-directional (up) class (see Supplementary Figure S5 for the other classes), and the red lines represent the median ranks, i.e. the

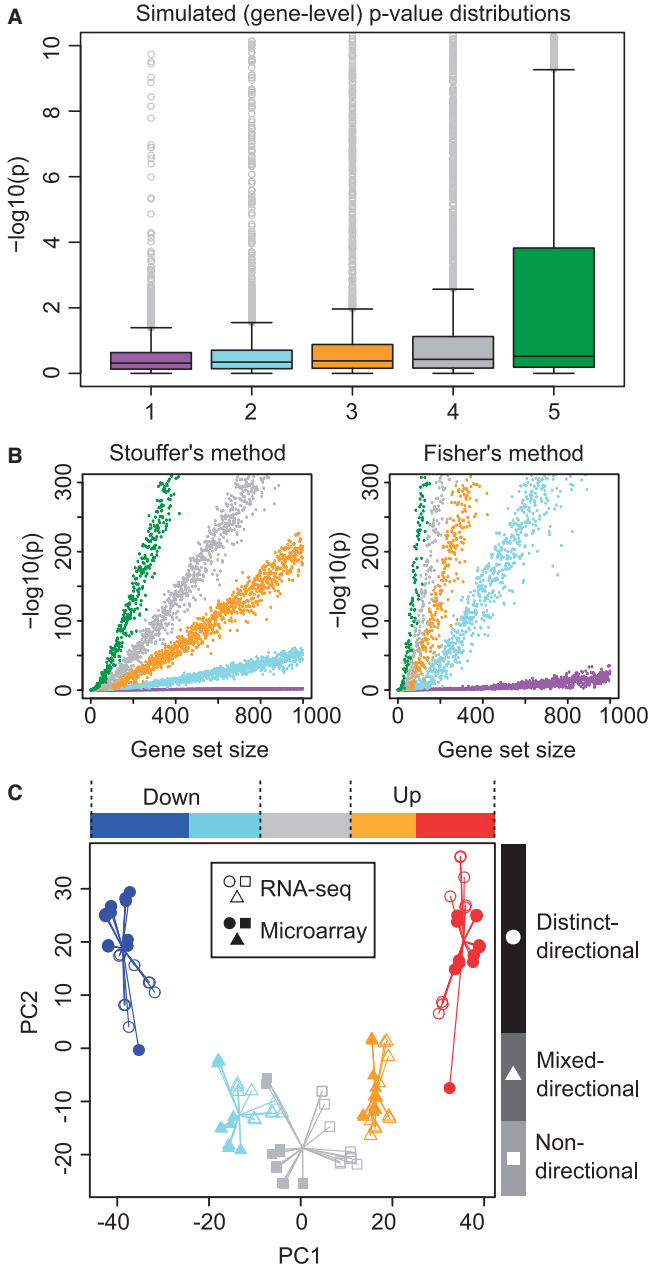


Figure 3. (**A**) Boxplots of five simulated *P*-value distributions (log-scaled) with increasing skewedness towards low *P*-values. Outliers extending beyond the plot borders are not shown. (**B**) Scatterplots of gene set *P*-values (log-scaled) against gene set size for Fisher's and Stouffer's methods, respectively, based on each of the five simulated gene-level *P*-value distributions. An increasing dependency of gene set size for the rightmost distributions can be observed. Each point represents the average of 100 random gene sets of the same size. (**C**) A PCA plot of the 128 gene set *P*-value vectors from different runs of the GSA workflow using the *S. cerevisiae* RNA-seq and microarray data. The first two principle components explain 74.1% of the variance. Although a mix of platforms, the major separation of the results is consistent with our directionality classification.

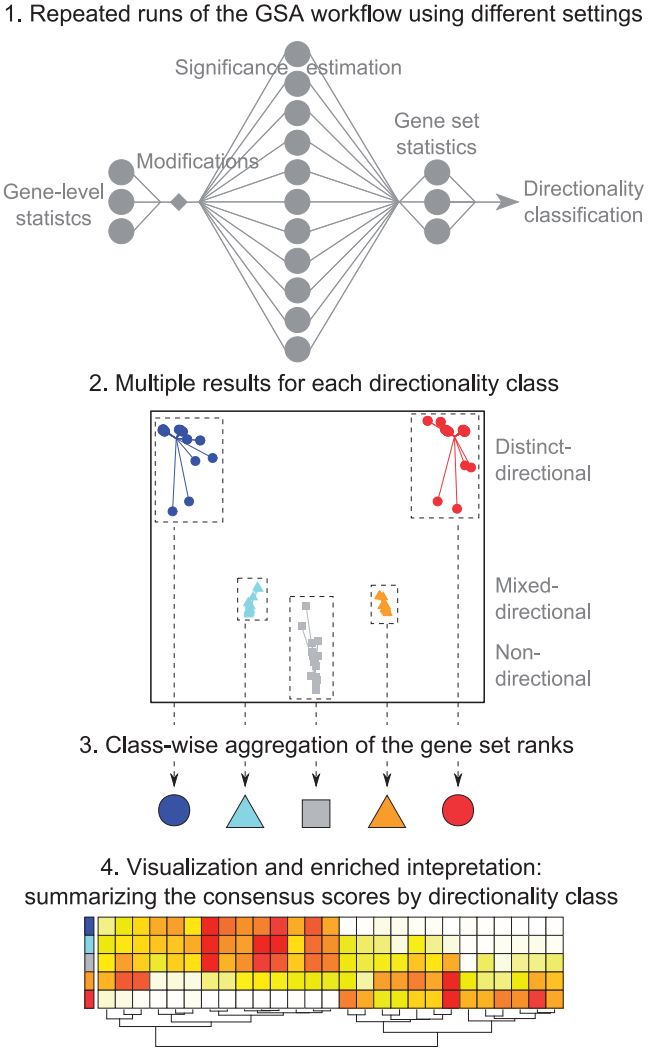


Figure 4. Outline of the concept of result aggregation of the directionality classes. (1) Initially, repeated runs of the GSA workflow are performed, resulting in multiple gene set *P*-value vectors, each representing a unique combination of settings (gene-level statistic, significance estimation method, gene set statistic and directionality class). (2) Based on their directionality class, the *P*-value vectors are grouped and converted to ranks and finally (3) aggregated classwise. (4) This constitutes a more thorough basis for the biological interpretation of the gene set results.

consensus score for each gene set. Figure 5B shows a heatmap of the consensus scores of the top-ranked gene sets for each of the directionality classes. Although the use of consensus scores will select the most significant gene sets based on all GSA runs, the scores themselves do not convey the absolute statistical significance of those gene sets. Hence, for one data set or study, a gene set with consensus score 10 may be highly significant, whereas a gene set with the same score in another data set may be less significant. To complement the consensus scores, we report the median *P*-value for each gene set, along with its consensus score, as a guide for which gene sets that should be further considered in the biological interpretation (Supplementary Figure S6). Figure 5C shows the median gene set *P*-values plotted against the consensus ranks. The gene sets in quadrant *b* are all top-ranked and have a median *P*-value <0.05. The few gene sets in quadrant *a* are selected among the top-ranked, but they have *P*-values slightly >0.05. In principle, for a desired median *P*-value as cut-off for significance, one wishes to adjust the horizontal line (consensus rank) so that quadrant *a* contains as few gene sets as possible.

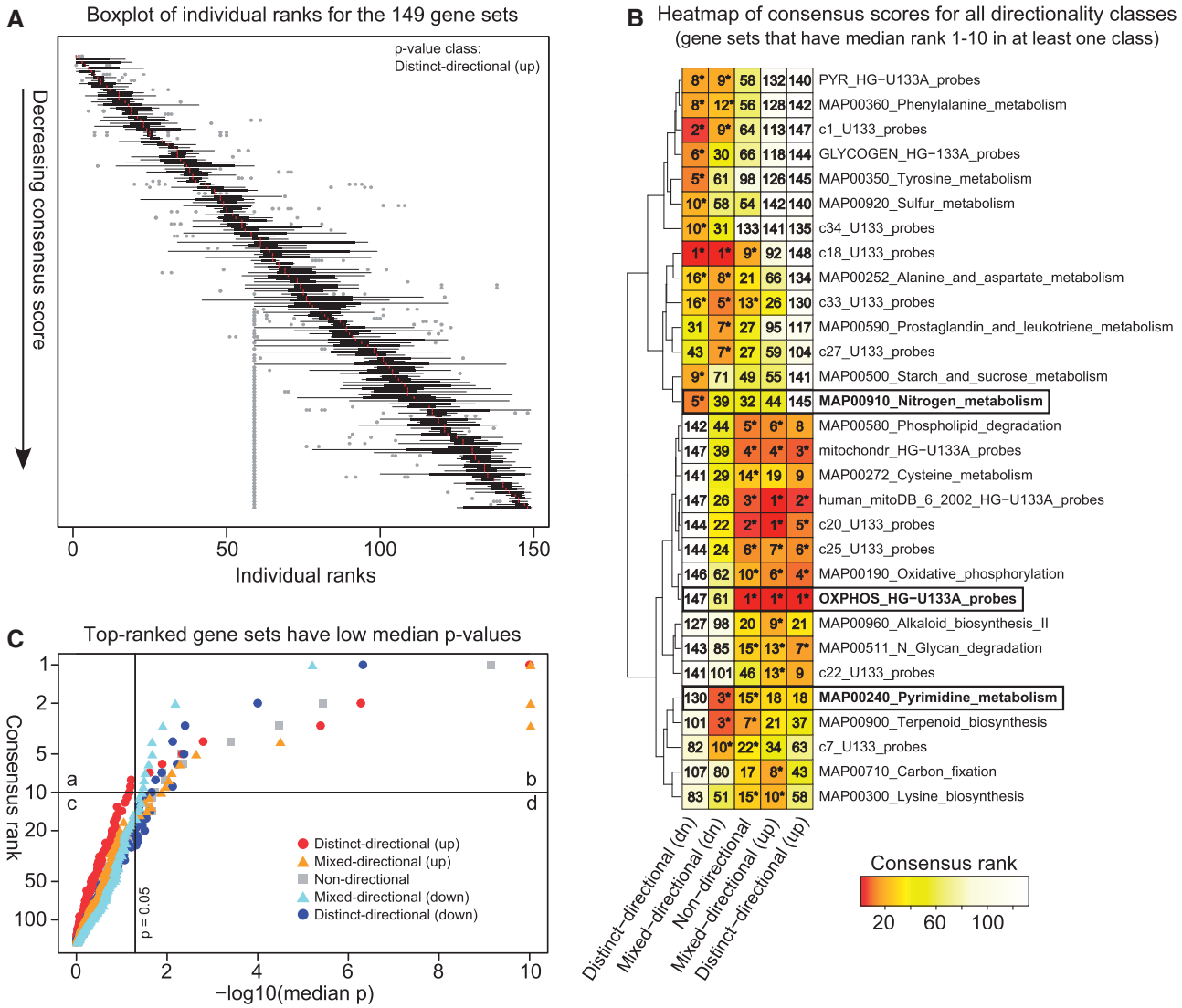


Figure 5. (**A**) Boxplots of individual ranks for the 149 Mootha gene sets representing metabolic pathways and co-regulated genes (c1–c36). Each box shows the ranks given in a specific gene set by the different runs of the GSA workflow, for the distinct-directional class (see Supplementary Figure S5 for the other classes). The gene sets are sorted, from top to bottom, based on their median rank, i.e. the consensus score. All outliers are shown as gray points. The vertically aligned outliers of the bottom ranked gene sets originate from the GSEA runs, which score gene sets with positive and negative gene set statistics separately. In this case, the negative gene sets are not scored and thus ranked equally and last. (**B**) A heatmap of the consensus scores of selected gene sets. All gene sets that received a median rank <10, in at least one class, are included. The three gene sets selected by a black rectangle are further discussed in the text. Gene sets with median *P*-values <0.05 are marked with an asterisk, and all corresponding median gene set *P*-values are shown in Supplementary Figure S6. (**C**) For all gene sets and each directionality class, the median *P*-value over all runs of the GSA workflow is plotted against the consensus rank (*y*-axis is logarithmic). For a chosen consensus score-based cut-off (horizontal line), one can observe the corresponding median *P*-value cut-off (vertical line) that all selected gene sets pass, i.e. for which they all end up in quadrant *b*. The *P*-values that are <1e-10 are set to this value in the plot.

The use of consensus scores and directionality classes enables us to get a more varied picture of how the gene sets are affected by differential expression. To explain this, we will use the three gene sets marked in Figure 5B as examples. The OXPHOS_HG-U133A_probes gene set has a consensus score (median rank) of 1 in the non-directional, mixed-directional (up) and distinct-directional (up) class. The corresponding median *P*-values are also highly significant for these classes (Supplementary Figure S6). This tells us that the gene set is in general highly regulated, as the non-directional class identifies gene sets that contain a high amount of significant genes, although not taking direction into account. Further on, if we are just interested of the part of the gene set that contains upregulated genes (upregulated in the normal glucose tolerant group compared with the type 2 diabetes group), we still find that this part is significantly differentially expressed (as reported by the mixed-directional up class). This tells us that there is an important component of upregulation, even though there could also be an important component of downregulation simultaneously. For the OXPHOS_HG-U133A_probes gene set, this is not the case, as the mixed-directional (down) class reports a median rank of 61. Finally, the

distinct-directional (up) class tells us that this gene set is coordinately upregulated (in the normal glucose tolerant group), all in correlation with the main results of Mootha *et al*. (16). Note that a gene set can receive good scores in the distinct-directional class, of one direction, and still get a good score in the mixed-directional class, of the other direction. This only means that, as a whole, the gene set is regulated in a distinct direction, but it can contain a small subset of genes that compose a component of regulation in the opposite direction. This is the case for the MAP00240_Pyrimidine_metabolism gene set.

The MAP00240_Pyrimidine_metabolism gene set has a score of 18 for the distinct-directional (up) class, as compared with 130 for the distinct-directional (down) class. Naturally, this tells us that the gene set is coordinately upregulated. However, in the mixed-directional (down) class the gene set is given a score of 3. Combined, this information tells us that the small subset of downregulated genes is highly significant (median *P*-value of 0.0069), but as a whole, the upregulation overrides this downregulated component.

As a third example of the benefits with the directionality classes, during the interpretation step, we will use the MAP00910_Nitrogen_metabolism gene set. This gene set receives a good score in the distinct-directional (down) class (median rank of 5 and median *P*-value of 0.0075), but not in the mixed-directional (down) class or any of the other classes. This means that the gene set as a whole is biased towards downregulation. In fact, 22 genes are downregulated, whereas only 8 are upregulated. However, as we can see from the mixed-directional (down) class, the downregulated genes on their own are not significant.

Robustness analysis

The previous section demonstrates the advantages with the directionality classification during the biological interpretation. An important question that remains to be answered is if the consensus scoring approach is robust, i.e. if it generates comparable results with regard to what input it is given. We, therefore, investigated two issues regarding how consistent the consensus scoring is: first by using runs of the GSA workflow with different numbers of permutations during the *P*-value calculations and second by using randomly selected runs.

Starting with the first issue, the concern is that when using a low number of permutations for the gene set *P*-value calculations, the resolution of the *P*-values will be low, resulting in that several gene sets may share the exact same *P*-value. Consequently, these gene sets will receive the same rank (see Supplementary Figure S7A for different numbers of permutations) when used as input to the consensus scoring algorithm. The question is if such a case would produce similar consensus scores when compared with the case of using GSA runs with a high number of permutations as input. To investigate this, we reran the GSA workflow, based on the human diabetes data with the original Mootha gene sets, for a selection of methods (Supplementary Methods) using different numbers of gene permutations (500, 750, 1000, 1500, 2000, 5000 and 10 000). The results from the runs in each of the seven permutation groups were aggregated so that each group resulted in a consensus score vector (actually one for each directionality class) for the gene sets. These consensus score vectors can be compared by calculating the Spearman correlation between all possible pairs, and it turns out that they correlate well. The minimum pairwise correlation (over all directionality classes and the four rank aggregation methods) is 0.997, showing that the correlation is high even in the worst case (Table 1). Next, we aggregated the runs again, but this time by randomly choosing methods from the different permutation groups, so that results that were aggregated into consensus scores originated from a mix of permutation numbers. This was repeated 1000 times, each time using a different random set of runs from the seven groups. Again, the resulting 1000 consensus score vectors for each directionality class correlate well (minimum correlation is 0.998), confirming that the consensus scoring approach is robust and not heavily influenced by mixing methods using different numbers of permutations for significance estimation (Table 1).

Regarding the second issue, the concern is that the consensus scores will differ depending on which GSA runs are chosen to be aggregated. To test this, we used the results from the case study on the human diabetes data. From these results, we randomly selected 95% of the GSA runs as input to the consensus scoring algorithm and repeated this 1000 times, thus generating 1000 consensus score vectors for each directionality class. As it turns out, the correlation of these results is high, and the minimum correlation is 0.994. This approach was also repeated when randomly using 85 and 70% of the GSA runs, showing similar results (see Table 1 for details). The robustness analysis, using 95, 85 and 70% of the runs, was also performed for the *S. cerevisiae* data set and the human diabetes data set with GO terms (Supplementary Table S3).

All the aforementioned results on robustness are on a global scale, i.e. comparing all gene sets. In practice, it may be interesting to also focus on the top-ranked gene sets and compare these between runs of the consensus scoring algorithm during the robustness analysis. Reassuringly, it turns out that the vast majority of the top-ranked gene sets, i.e. the ones that would be selected for further biological interpretation, are highly ranked in all consensus scoring runs, regardless of the input (i.e. which GSA runs and how many permutations). See Supplementary Figure S7 for full details.

The Piano R package for enriched gene set analysis

The workflow described in this article, as well as the consensus scoring approach, is implemented as the R package Piano (Platform for Integrated Analysis of Omics data). Additionally, Piano includes functions for importing gene set collections of various formats, including the Cytoscape sif-format, the gmt-files available from the Molecular Signatures Database (17) and genome-scale metabolic models in the SBML-format, e.g. available through the web-based BioMet ToolBox (39) and the Human

Table 1. Robustness of the consensus scoring approach

Robustness analysis	Mean rank	Median rank	Borda	Copeland
Different number of permutations	0.998	0.997	0.998	0.998
Mixed number of permutations	0.999	0.998	0.998	0.998
95% of the GSA runs	0.997	0.994	0.997	0.997
85% of the GSA runs	0.991	0.978	0.991	0.991
70% of the GSA runs	0.978	0.968	0.979	0.978

Minimum Spearman correlation (over all directionality classes) between repeated runs of the consensus scoring algorithm, using varying input, for each of the four rank aggregation methods. The first row shows the correlation when using the same number of permutations for the input GSA runs, but varying this fixed number between seven different consensus score runs (for each directionality class). The second row shows the results from randomly selecting the number of permutations for each input GSA run, creating an input with mixed numbers of permutations. This is repeated 1000 times for each directionality class. The last three rows are based on randomly selecting a percentage of the GSA runs as input, and this is repeated 1000 times for each directionality class. The results are for the human diabetes data and the 149 Mootha gene sets, similar results for the *S. cerevisiae* data set and for the human diabetes data set using GO terms are presented in Supplementary Table S3.

Metabolic Atlas (40) for microorganisms and human cell types, respectively. The package contains implementations of all the gene set statistic methods described in this article, serving as a good platform for testing different gene set analyses using the same set-up. The package also includes several functions for result visualization, including a network-based plot showing overlapping gene sets and their significance. Finally, Piano also contains functions for the full analysis of microarray data, if the user wants a fully integrated GSA starting from raw expression data. See Supplementary Figure S8 for an overview of the functions in the Piano R package. Piano is available, together with a user manual, for download at www.sysbio.se/piano.

DISCUSSION

In this article, we have addressed the practical issues with gene set analysis based on the questions raised in the 'Introduction' section. We have defined a new workflow for GSA where we include a step of using different modifications of the gene-level statistics. We have also proposed to separate the GSA results into three classes, based on the choices made at each step of the analysis, describing different aspects of gene expression directionality. As our workflow is fully implemented in the Piano R package, it serves well as a platform for testing and evaluating different GSA runs in a simple way. However, its major use should be to run repeated analyses with different settings and use the consensus scoring approach, and in particular the directionality classes to interpret the results.

In this study, we initially use the Piano package to evaluate the impact of the separate components of the GSA workflow. The primary observation is that the different methods produce comparable results. However, it should be mentioned that the choice of gene-level statistics as well as gene set statistics do influence the results to some extent, although from a global point of view, they still correlate well. Regarding the three significance estimation methods (theoretical null distribution, gene sampling and sample permutation), we did not observe any great differences between the results. In particular, the gene set *P*-values show a high correlation when comparing gene sampling with sample permutation. The lack of great difference between these two significance estimation methods means that the choice between them should be based on which is more practical and which is statistically theoretically correct. However, these two issues, in particular the latter, need to be further investigated. From a theoretical point of view, it is of course important to consider the change of null hypothesis depending on the choice of significance estimation method, but in general and from a practical point of view, we prefer using gene sampling. The reason for this is that permutated null distributions do not rely on the assumptions that the theoretical null distributions make on the gene-level statistics. These assumptions may pose a problem for some gene-level data, as we showed for the case of Fisher's and Stouffer's method. Further on, sample permutation takes a considerable higher amount of computational time compared with gene sampling, as the step of recalculating the gene-level statistics has to be included. Furthermore, to sufficiently permute the sample labels to generate an appropriate background of gene-level statistics, a proper number of samples have to be available. For small-scale experiments with few replicates, this may often not be the case.

The similarity of the different methods, or actually, the different runs of the GSA workflow, leads to a difficulty in selecting the best method. Instead, perhaps other factors will become more guiding, such as availability of software in a familiar system. As a consequence of this, we propose to use a combination of methods and use the consensus scoring approach to find important gene sets. This would normally be a somewhat laborious task but is easily performed with the implementations in Piano. Regarding runtime, of all the methods implemented, the GSEA and Wilcoxon rank-sum test take the longest time, up to several hours if the gene set collection is large. However, it is possible to run the consensus approach with a majority of the methods within a reasonable time. As an example, running the *S. cerevisiae* microarray data with 1436 GO term gene sets, using all methods except GSEA and Wilcoxon, takes ~15 min on an ordinary desktop computer. Future efforts should include improving the implementations of the slower methods with regard to decreasing the runtime.

The reason for using gene set ranks in the result aggregation, rather than the gene set *P*-values directly, is to give equal weight to conservative and less conservative methods. A limitation of this is that it may become difficult to compare the results of different experiments, based solely on the consensus scores. It should be kept in mind that a rank of one, only means that this gene set is the most significant; however, it does not *per se* say anything about the level of significance. Hence, it is important to evaluate the consensus scores in combination with the gene set *P*-values, as described in Figure 5 and Supplementary Figure S6, to not lose the information about statistical significance. Apart from aggregating the results of different methods, the consensus scoring approach can enable the generation of gene set results for all of the directionality classes (which not all methods can do on their own). An important issue is the robustness of the aggregation and we investigated this with regard to the issue of using GSA results with different *P*-value resolutions (derived from the use of different numbers of permutations during significance estimation) and with regard to randomly choosing a subset of the GSA results. Reassuringly, in both cases, we show that the consistency of the consensus scoring is high.

The directionality classification aims to give comprehensible information on the effect of gene expression directionality in the context of gene sets, as well as categorize what kind of directionality information is captured by the *P*-values given by different GSA methods. If the goal is to analyze gene sets without the interest in expression directionality, the non-directional class is the choice. If the goal is to incorporate directionality information in the analysis, the mixed-directional and distinct-directional classes should be used. The mixed-directional class gives information on the significance of the separate subsets of up- and downregulated genes in a gene set, disregarding the relative amount of genes in each of the two subsets. This can for instance be important for the biological interpretation of large gene sets, experiencing complex regulation with components of regulation in both directions. The distinct-directional class, on the other hand, incorporates the directionality information on the gene set as whole to assess whether there is a consolidated significance in one distinct direction. Of course, it is encouraged to use the combined information given by all three directionality classes in the interpretation step.

The important point with using the directionality classification and consensus scoring is demonstrated by the three example gene sets (OXPHOS_HG-U133A_probes, MAP00240_Pyrimidine_metabolism and MAP00910_Nitrogen_metabolism), as mentioned in the end of the 'Results' section. The information given by the combination of *P*-values of different directionality classes is superior of that given by only a single *P*-value. For instance, if one would use the mean gene statistic starting from gene-level *t*-values (resulting in the distinct-directional class), the interpretation would be that nitrogen metabolism is downregulated. However, if the gene-level *P*-values were used instead (resulting in the mixed-directional class), the interpretation would completely change to that nitrogen metabolism is not affected at all. These are of course two different results. The benefit of the directionality classes comes from using them in combination, to get the whole picture. This can be achieved by combining the results from different runs of the GSA workflow.

In conclusion, our refined and implemented GSA workflow can be used to analyze gene expression data, as well as proteomics, metabolomics and genome-wide association data. Different methods can easily be run in parallel, enabling comparison and assessment of the variation of the results and the possibility of calculating gene set consensus scores. In combination with the directionality classes, this constitutes a more thorough basis for the biological interpretation of the gene set results.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–3, Supplementary Figures 1–8 and Supplementary Methods.

ACKNOWLEDGEMENTS

The authors acknowledge Tobias Österlund for the valuable comments and the assistance regarding the *P*-value transformation. We are grateful to the reviewers for their valuable feedback. The computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at C3SE. We also acknowledge Gothenburg Bioinformatics Network (GOTBIN).

FUNDING

Knut and Alice Wallenberg foundation; Chalmers foundation; Bioinformatics Infrastructure for Life Sciences (BILS). Funding for open access charge: Chalmers Library.

Conflict of interest statement. None declared.

REFERENCES

1. Hung,J.H., Yang,T.H., Hu,Z., Weng,Z. and Delisi,C. (2012) Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief. Bioinform.*, **13**, 281–291.
2. Naeem,H., Zimmer,R., Tavakkolkhah,P. and Kuffner,R. (2012) Rigorous assessment of gene set enrichment tests. *Bioinformatics*, **28**, 1480–1486.
3. Ackermann,M. and Strimmer,K. (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10**, 47.
4. Goeman,J.J. and Buhlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.
5. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25.
6. Rivals,I., Personnaz,L., Taing,L. and Potier,M.C. (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.
7. Maere,S., Heymans,K. and Kuiper,M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.

8. Dennis,G. Jr, Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, R60.
9. Beißbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.
10. Zheng,Q. and Wang,X.J. (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.*, **36**, W358–W363.
11. Goeman,J.J., Van De Geer,S.A., De Kort,F. and Van Houwelingen,H.C. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.
12. Kong,S.W., Pu,W.T. and Park,P.J. (2006) A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, **22**, 2373–2380.
13. Mansmann,U. and Meister,R. (2005) Goeman's global test versus an ANCOVA approach. *Methods Inf. Med.*, **44**, 449–453.
14. Hummel,M., Meister,R. and Mansmann,U. (2008) GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics*, **24**, 78–85.
15. Tsai,C.A. and Chen,J.J. (2009) Multivariate analysis of variance test for gene set analysis. *Bioinformatics*, **25**, 897–903.
16. Mootha,V.K., Lindgren,C.M., Eriksson,K.F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstrale,M. and Laurila,E. (2003) PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.
17. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *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.
18. Perry,J.R., McCarthy,M.I., Hattersley,A.T., Zeggini,E. the Wellcome Trust Case Control Consortium, Weedon,M.N. and Frayling,T.M. (2009) Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. *Diabetes*, **58**, 1463–1467.
19. Nam,D. and Kim,S.Y. (2008) Gene-set approach for expression pattern analysis. *Brief. Bioinform.*, **9**, 189–197.
20. Song,S. and Black,M.A. (2008) Microarray-based gene set analysis: a comparison of current methods. *BMC Bioinformatics*, **9**, 502.
21. Fridley,B.L., Jenkins,G.D. and Biernacka,J.M. (2010) Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS One*, **5**, e12693.
22. Cao,W., Li,Y., Liu,D., Chen,C. and Xu,Y. (2011) Statistical and biological evaluation of different gene set analysis methods. *Procedia Environ. Sci.*, **8**, 693–699.
23. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
24. Li,J., Wang,L., Xu,L., Zhang,R., Huang,M., Wang,K., Xu,J., Lv,H., Shang,Z. and Zhang,M. (2012) DBGSA: a novel method of distance-based gene set analysis. *J. Hum. Genet.*, **57**, 642–653.
25. Tian,L., Greenberg,S.A., Kong,S.W., Altschuler,J., Kohane,I.S. and Park,P.J. (2005) Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci. USA*, **102**, 13544–13549.
26. Fisher,R.A. (1932) *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.
27. Stouffer,S.A., Suchman,E.A., Devinney,L.C., Star,S.A. and Williams,R.M. Jr (1949) *The American Soldier: Adjustment During Army Life*. Princeton University Press, Oxford, England.
28. Patil,K.R. and Nielsen,J. (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc. Natl Acad. Sci. USA*, **102**, 2685–2689.
29. Oliveira,A.P., Patil,K.R. and Nielsen,J. (2008) Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. *BMC Syst. Biol.*, **2**, 17.
30. Kim,S.Y. and Volsky,D.J. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, **6**, 144.
31. Taylor,J. and Tibshirani,R. (2006) A tail strength measure for assessing the overall univariate significance in a dataset. *Biostatistics*, **7**, 167–181.
32. Smyth,G.K. (2005) Limma: linear models for microarray data. In: Gentleman,R., Carey,V., Dudoit,S., Irizarry,R. and Huber,W. (eds), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, pp. 397–420.
33. Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107–129.
34. de Borda,J.C. (1781) Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*.
35. Copeland,A.H. (1951) A reasonable social welfare function. *Seminar on Mathematics in Social Sciences*. University of Michigan, Michigan.
36. Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**
37. Nookaew,I., Papini,M., Pornputtpong,N., Scalcinati,G., Fagerberg,L., Uhlén,M. and Nielsen,J. (2012) A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **40**, 10084–10097.
38. Allison,D.B., Gadbury,G.L., Heo,M., Fernández,J.R., Lee,C.K., Prolla,T.A. and Weindruch,R. (2002) A mixture model approach for the analysis of microarray gene expression data. *Comput. Stat. Data Anal.*, **39**, 1–20.
39. Cvijovic,M., Olivares-Hernández,R., Agren,R., Dahr,N., Vongsangnak,W., Nookaew,I., Patil,K.R. and Nielsen,J. (2010) BioMet Toolbox: genome-wide analysis of metabolism. *Nucleic Acids Res.*, **38**, W144–W149.
40. Agren,R., Bordel,S., Mardinoglu,A., Pornputtapong,N., Nookaew,I. and Nielsen,J. (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput. Biol.*, **8**, e1002518.