BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 3. 189–197
Advance Access publication January 17, 2008
doi:10.1093/bib/bbn001

Gene-set approach for expression pattern analysis

Dougu Nam and Seon-Young Kim

Submitted: 7th November 2007; Received (in revised form): 28th December 2007

Abstract

Recently developed gene set analysis methods evaluate differential expression patterns of gene groups instead of those of individual genes. This approach especially targets gene groups whose constituents show subtle but coordinated expression changes, which might not be detected by the usual individual gene analysis. The approach has been quite successful in deriving new information from expression data, and a number of methods and tools have been developed intensively in recent years. We review those methods and currently available tools, classify them according to the statistical methods employed, and discuss their pros and cons. We also discuss several interesting extensions to the methods.

Keywords: *gene set analysis; DNA microarray; differential expression of genes*

INTRODUCTION

Since the advent of microarray technology, it has been of primary interest to identify differentially expressed genes (DEGs) and elucidate related biological processes. To this aim, a large number of statistical methods and tools have been developed in past decade. The most widely used approach, namely individual gene analysis (IGA), evaluates the significance of individual genes between two groups of samples compared. IGA typically yields a list of altered genes from a cutoff threshold. The list is then investigated with biologically defined gene sets derived from Gene Ontology or some pathway databases to assess the enrichment of specific biological themes in the list. Khatri and Draghici [1] and Rivals *et al.* [2] reviewed in detail the methods and tools for IGA.

The main problems of IGA originate from the use of the cutoff threshold value. First, the final result of IGA is significantly affected by the selected threshold, which is normally chosen arbitrarily [3]. Pan *et al.* [4], while analyzing three microarray datasets, showed that different choices of the threshold value severely alter the biological conclusions (enrichment of specific function categories in the gene list). Second, many genes with moderate but meaningful expression changes are discarded by the strict cutoff value, which leads to a reduction in statistical power [3, 5].

In recent years, gene set analysis (GSA) methods, free from the problems of the 'cutoff-based' methods, has received a great deal of attention (Figure 1). GSA directly scores pre-defined gene sets for differential expression and especially aims to identify gene sets with 'subtle but coordinated' expression changes that cannot be detected by IGA methods. The key principle is that even weak expression changes in individual genes gathered to a large gene set can show a significant pattern. From a biological perspective, GSA methods are promising because functionally related genes often display a coordinated expression to accomplish their roles in the cell. By changing the focus from individual genes to a set of genes or pathways, the GSA approach enables the understanding of cellular processes as an intricate network of functionally

Corresponding author. Seon-Young Kim, Functional Genomics Research Center, KRIBB, 111 Gwahangno, Yuseong-gu, Daejeon 305-806, Korea. Tel: 82-42-879-8116; Fax: 82-42-879-8110; E-mail: kimsy@kribb.re.kr

Dougu Nam received his PhD in applied mathematics from the Korea Advanced Institute of Science and Technology in 2002. He is a senior researcher at the Division of Industrial Mathematics, National Institute for Mathematical Sciences (NIMS). His main interest is the development of mathematical models and algorithms for bioinformatics and systems biology.

Seon-Young Kim received his PhD in microbiology from Seoul National University in 1998. He is a senior researcher at Functional Genomics Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB). His main interest is the development of bioinformatics tools for functional genomics research.

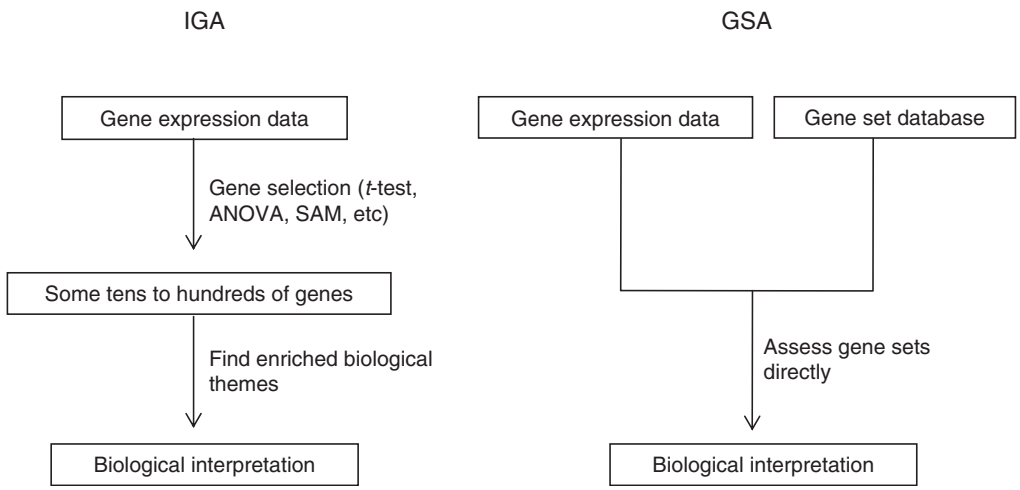

Figure 1: A schematic diagram comparing gene set analysis (GSA) with individual gene analysis (IGA). IGA is a two-step process which first selects some tens to hundreds of genes by an arbitrarily chosen cutoff and then, from the selected genes, infers the biological meaning of the gene expression data. In contrast, GSA is single-step process which in advance prepares gene sets from diverse sources as a testable hypothesis and then directly infers the biological meaning of gene expression data by applying either a sample or a gene randomization test.

related components [6]. Indeed, Mootha *et al.* [7], whose work has inspired the development of various GSA methods, developed gene set enrichment analysis (GSEA) to identify a significantly altered gene set between microarray samples of diabetic and normal muscles for which no single gene was found to be differentially expressed by IGA. They found that genes involved in oxidative phosphorylation, whose expression was only modestly decreased in human diabetic muscle, were correlated with body aerobic capacity and associated with variations in human metabolism [7]. Ben-Shaul *et al.* [5], while analyzing microarray data from the brains of MPTP-injected mice, showed that applying the GSA approach enabled the detection of many GO terms overlooked by IGA methods.

Another important problem with IGA is that all the statistical methods applied are based on the wrong assumption of independent gene (or gene-group) sampling, which increases false positive predictions. Some of the GSA methods also address this issue.

In this review, we compare and discuss the GSA methods and currently available tools that have been intensively developed in recent years to guide researchers to choose appropriate methods and tools for their own purposes. Additionally, several interesting extensions and applications of GSA are discussed. Researchers also may benefit from a recent review by Dopazo [6].

GSA METHODS AND THEIR NULL HYPOTHESES

Even before the work of Mootha *et al.* [7], a few works already applied the concept of cutoff-free group testing to analyze gene expression data. For example, Virtaneva *et al.* [8] prepared several functional categories (gene sets) using a SWISS-PROT database, calculated scores for the functional categories from expression data, and applied sample randomization to assess the significance of each category. Pavlidis *et al.* [9] developed a semi-supervised method to score each pre-defined gene class, and investigated three metrics: the degree of co-expression, the significance of expression profiles in the context of experimental designs and the learnability of gene class.

Thereafter, various GSA methods have been developed (Table 1) based on different null hypotheses and statistical methods [3, 10–13]. Tian *et al.* [10] classified two kinds of null hypotheses for testing the coordinated association of gene sets with a phenotype of interest. The first type hypothesizes the same level of association of a gene set with the given phenotype as the complement of the gene set (say, Q1). The second type only considers the genes within a gene set and hypothesizes that there is no gene in the gene set associated with the phenotype (say, Q2). The methods based on Q1 and Q2 were termed *competitive* and *self-contained*, respectively by Geoman and Buhlmann [11].

Table 1: Cutoff-free gene set analysis methods

Authors	Year	Name	Statistical test	Self-contained versus competitive	Gene versus ample randomization	Reference
Virtaneva *et al.*	2001		sample randomization	self-contained	sample	[8]
Pavlidis *et al.*	2002		gene randomization	competitive	gene	[9]
Mootha *et al.*	2003	GSEA	sample randomization	mixed	sample	[7]
Breslin *et al.*	2004	Catmap	gene randomization	competitive	gene	[3]
Goeman *et al.*	2004	globaltest	sample randomization	self-contained	sample	[17]
Smid *et al.*	2004	GO-Mapper	*z*-test	competitive	gene	[38]
Volinia *et al.*	2004	GOAL	gene randomization	competitive	gene	[39]
Barry *et al.*	2005	SAFE	sample randomization	competitive	sample	[19]
Beh-Shaul *et al.*	2005		Kolmogorov–Smirnov test	competitive	gene	[5]
Boorsma e al.	2005	T-profiler	*t*-test	competitive	gene	[15]
Kim *et al.*	2005	PAGE	*z*-test	competitive	gene	[14]
Lee *et al.*	2005	ErmineJ	sample randomization	competitive	gene	[16]
Subramanian *et al.*	2005	GSEA	sample randomization	mixed	gene	[25]
Tian *et al.*	2005	Q1, Q2	gene or sample randomization	competitive or self-contained	gene or sample	[10]
Tomfohr *et al.*	2005	PLAGE	sample randomization	self-contained	sample	[20]
Edelman *et al.*	2006	ASSESS	sample randomization	competitive	sample	[28]
Kong *et al.*	2006		Hotelling's T squared	self-contained	sample	[21]
Nam *et al.*	2006	ADGO	*z*-test	competitive	gene	[29]
Saxena *et al.*	2006	AE	sample randomization	competitive	sample	[31]
Scheer *et al.*	2006	JProGO	Fisher's exact test, Kolmogorov–Smirnov test, *t*-test, unpaired Wilcoxon's test	competitive	gene	[40]
Al-Shahrour *et al.*	2007	Fatiscan	Fisher's exact test, hypergeometric test	competitive	gene	[41]
Backes *et al.*	2007	GeneTrail	Fisher's exact test, hypergeometric test, sample randomization	competitive	gene or sample	[42]
Cavalieri *et al.*	2007	Eu.Gene Analyzer	Fisher's exact test, sample randomization	competitive	gene or sample	[43]
Dinu *et al.*	2007	SAM-GS	sample randomization	self-contained	sample	[22]
Efron *et al.*	2007	GSA	sample randomization	mixed	sample	[26]
Newton *et al.*	2007	Random set	*z*-test	competitive	gene	[44]

According to this classification, Catmap [3], PAGE [14], T-profiler [15], ErmineJ [16] and Q1 test of Tian *et al.* [10] are competitive methods, while the global test of Goeman *et al.* [17, 18], SAFE [19], Q2 test of Tian *et al.* [10], PLAGE [20], the multivariate approach of Kong *et al.* [21] and SAM-GS [22] are self-contained methods (Table 1).

The competitive methods test the relative enrichment of DEGs in a gene set compared with the background set, and target gene sets with coordinated expression changes. However, this feature of relativism can cause the peculiar behavior of a 'zero-sum game' [23, 24]. For example, in an extreme case where all the genes are down-regulated under an experimental condition, some gene sets can be considered up-regulated by the background distribution. Moreover, most of the competitive methods suffer from the invalid assumption of independent gene sampling [11]. On the other hand, the self-contained methods use only the information contained in the given gene set. They provide very powerful predictions due to the strong hypothesis of Q2. However, in this case only a single DEG can make the whole gene set significant so that the gene sets may not be 'enriched' with DEGs. In this sense, self-contained methods target a larger class of DEG sets than those targeted by competitive methods.

INTERPRETING GSEA PROCEDURE

Unlike the methods described above, the widely known GSEA [7] tests another kind of hypothesis that 'none of the gene sets considered is associated with the phenotype' (say, Q3). The object Q3 validates is the entire dataset, and it tests whether the dataset contains any gene set that is associated with

the phenotype, while Q1 and Q2 validate the significance of individual gene sets. The approach of GSEA utilizes a competitive statistic (Kolmogorov–Smirnov statistic) as a 'score function', if not as a test statistic, to represent the relative enrichment of DEGs in each gene set. Then, it tests the significance of the entire dataset by applying sample permutation to the scores. GSEA is considered a competitive method relative to individual gene sets, but is considered a self-contained method relative to the entire dataset (set of gene sets). Two ways to calculate the *P*-value from the sample permutation follow:

(1) For each observed score of a gene set, count the number of sample permutations for which a gene set with a better score than the observed is found.
(2) For N sample permutations, count the numbers of gene sets that have a better score than a given threshold for the real dataset and permuted datasets, say k_0 and $k_1, \ldots, k_N$, respectively, and then, the number of cases for which $k_i \geq k_0$, $i = 1,2, \ldots, N$.

The first method assigns *P*-values to each gene set, while the second method assigns *P*-values only to the entire dataset. The original GSEA method focuses on the highest-scoring single gene set so that the above two methods coincide, and hence the *P*-value can be given to both the highest-scoring gene set and the entire dataset. In subsequent works, however, only the first method was used to derive multiple significant gene sets [25, 26]. An interesting analog of GSEA in a self-contained analysis of a gene set is suggested by Geoman and Buhlmann [11]. They applied the second type of *P*-value calculation to assess the significance of the gene set, and not of its constituents. In spite of the complicated interpretation of the *P*-value in the GSEA procedure, their method seems to provide a legitimate approach for identifying subtle but coordinated expression changes without violating the dependence of gene expression.

Many other highly sensitive methods are based on the invalid assumption of independent gene sampling for Q1 or on the strong Q2 hypothesis, and hence may not be directly compared with GSEA. Notwithstanding, improvements in the statistical power of GSEA have been achieved by combining various competitive score functions to the GSEA procedure. Originally, Mootha *et al.* [7] used an un-weighted Kolmogorov–Smirnov statistic, but they later improved it [25] by weighting each gene by the correlation with the phenotypes to prevent gene sets clustered in the middle of the ordered gene list from getting high scores [24]. Efron and Tibshirani [26] introduced five test statistics for a GSEA algorithm—mean, mean.abs, maxmean, GSEA and GSEA.abs—to test five simulated conditions and concluded that the maxmean statistic is the only method with consistently low *P*-values in all situations.

There have been some criticisms of GSEA. Damian and Gorfine [24] illustrated two examples in which the GSEA procedure seemed to behave peculiarly: the enrichment score can be influenced by the size of a gene set and by the presence or absence of lower-ranking sets. Surely, those two peculiar behaviors stem from the competitive nature of the GSEA procedure. The first example simply shows the typical properties of statistical scores that usually depend on the number of samples. Virtaneva *et al.* [8], Tian *et al.* [10] and Lee *et al.* [16] devised gene set scores that do not depend on the size of gene sets, and which one to use may depend on the preference of the user. However, it should be noted that subtle but coordinated expression changes are detectable only by taking into account the number of genes.

METHODS FOR *P*-VALUE CALCULATION

Many authors have discussed the differences between gene and sample randomization in inferring the statistical significance of gene set scores. Among them, Tian *et al.* [10] and Goeman and Buhlman [11] discussed the issue most comprehensively. Two different opinions exist on the gene versus sample randomization for calculating *P*-values. One group suggests that both gene and sample randomizations should be used because they test two different but complementary null hypotheses [10]. The other group insists that only sample randomization should be used to avoid inherent problems of the gene randomization method [11, 12, 22].

According to Goeman and Buhlman [11], gene randomization is problematic because the roles of samples and genes in classical statistical tests are reversed, which makes the interpretation of the *P*-value unclear. More importantly, the condition of independent gene sampling is not satisfied by the correlation structures among functionally related

genes. In particular, the latter problem has well been recognized by many researchers [3, 8, 10, 12, 13, 16, 19–21, 25, 27]. Indeed, Breslin *et al.* [3], Delongchamp *et al.* [27], Kim *et al.* [13] and Geoman and Buhlmann [11] using real or carefully designed simulated datasets showed that a gene set with highly correlated genes tends to show a much smaller *P*-value than that without correlation, which increases false positive predictions. For this reason, the current consensus favors sample randomization over gene randomization as the more appropriate statistical test.

However, sample randomization is not without a problem. First, it requires a certain level of sample replicates to attain deep levels of significance, but this condition often is not met in many datasets such as time-series data or those designed to investigate the effects of diverse drugs in multiple conditions. For this reason, Subramanian *et al.* [25] included the gene randomization option in their GSEA program and suggested using gene randomization to generate hypothesis when the number of samples is small. The second problem is that sample randomization methods often identify too many gene sets as significant when there are many DEGs in the dataset [10, 11, 26]. From sample randomization, one or only a few DEGs in a gene set can easily reject the strong hypothesis Q2, and the original purpose of GSA, i.e. discovering significant patterns represented simply by a set may not be fulfilled. This is clearly observed in the following simulation study.

A SIMULATION STUDY: COMPARISON OF Q1, Q2 AND Q3

This study compares the distribution of *P*-values obtained from the three hypotheses Q1, Q2 and Q3 on simulated data. We generated expression profiles of 2000 genes with two sample groups, each having 20 samples. The expression values were sampled from a standard normal distribution in both groups. For 600 randomly selected genes (30%), we added a random value between 0.5 and 1 to the second group to generate DEGs. The genes were divided into 100 gene sets, each of which contained 20 genes. To compare the difference of the three hypotheses, we commonly used the *average t-statistic* in a gene set as the score function [10]. Since the DEGs were chosen uniformly at random, no gene sets were expected to be 'enriched' with DEGs. Indeed, the competitive method (Q1) recognized no DEG sets so the *P*-values were distributed uniformly (Figure 2). However, the self-contained method (Q2) detected most of the gene sets (83%) as differentially expressed with a *P*-value cutoff of 0.05. The mixed approach, GSEA exhibited an intermediate performance.

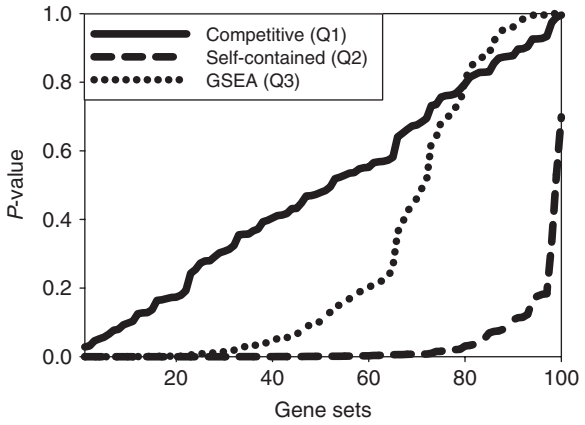

Figure 2: The *P*-value distributions of 100 gene sets for the three GSA approaches on simulated data. 10 000 permutations were performed for gene or sample randomizations on the average *t*-score.

From this result, we confirmed the criterion for choosing a statistical method for GSA. If the purpose is to find gene sets relatively enriched with DEGs, a competitive method based on Q1 should be used. However, if the purpose is to find gene sets clearly separated between the two sample groups, a self-contained method based on Q2 should be selected. Our group prefers the mixed approach (Q3) to avoid the clear drawbacks of the other methods, but recommends using all the methods simultaneously, if possible, with biological analyses.

CURRENT TOOLS AND GENE SET DATABASES

Our group reviewed currently available GSA tools (Table 2), which vary in statistical methods used (gene or sample randomization), the form of the provided tool (web server, standalone software, Excel add-ins, or command line scripts) and the applicable organisms. Because sample randomization generally takes much longer computation time than does gene randomization, most sample randomization-based methods are offered as standalone programs or command-line scripts, whereas

Table 2: Gene set analysis tools

Name	Organism[a]	Application Type	URL	Reference
ADGO	H, M, R, Y	Web server	http://array.kobic.re.kr/ADGO	[29]
ASSESS	H, M, R	Octave/Java standalone	http://people.genome.duke.edu/~jhg9/assess/	[28]
Babelomics	H, M, R, DM, S, C	Web server	http://www.babelomics.org	[45]
Catmap	H	Perl script	http://bioinfo.thep.lu.se/catmap.html	[3]
ErmineJ	H, M, R	Java standalone	http://www.bioinformatics.ubc.ca/ermineJ/	[16]
Eu.Gene Analyzer	H, M, R, Y	Windows/Unix standalone	http://www.ducciocavalieri.org/bio/Eugene.htm	[43]
FatiScan	H, M, R, Y, B, D, G, C, A, S, DM	Web server	http://fatiscan.bioinfo.cipf.es/	[41]
GAZER	H, M, R, Y	Web server	http://integromics.kobic.re.kr/GAzer/index.faces;	[13]
GeneTrail	H, M, R, Y, SA, CG, AT	Web server	http://genetrail.bioinf.uni-sb.de/	[42]
Global test	NA	R package	http://bioconductor.org/packages/2.0/bioc/html/globaltest.html	[17]
GOAL	H, M	Web server	http://microarrays.unife.it	[39]
GO-Mapper	H, M, R, Z, DM, Y	Windows standalone, Perl script	http://www.gatcplatform.nl/	[38]
GSA	H	R package	http://www-stat.stanford.edu/~tibs/GSA/	[26]
GSEA	H	Java standalone, R package	http://www.broad.mit.edu/gsea/	[25]
JProGO	Various prokaryotes	Web server	http://www.jprogo.de/	[40]
MEGO	H	Windows standalone	http://www.dxy.cn/mego/	[46]
PAGE	H, M, R, Y	Python script	From the author (kimsy@kribb.re.kr)	[14]
PLAGE	H, M	Web server	http://dulci.biostat.duke.edu/pathways/	[20]
SAFE	NA	R package	http://bioconductor.org/packages/2.0/bioc/html/safe.html	[19]
SAM-GS	NA	Windows Excel Add-In	http://www.ualberta.ca/~yyasui/homepage.html	[22]
T-profiler	Y, CA	Web server	http://www.t-profiler.org/	[15]

[a]H: *Homo sapiens*; M: *Mus musculus*; R: *Rattus norvegicus*; Y: *Sacchaomyces cerevisiae*; B: *Bos Taurus*; D: *Daniel rerio*; G: *Gallus gallus*; C: *Caenorhabditis elegans*; A: *Arabidopsis thaliana*; DM: *Drosophila melanogaster*; Z: *Zebra fish*; CA: *Candida albicans*; SA: *Staphylococcus aureus*; CG: *Corynebacterium glutamicum*; AT: *Arabidopsis thaliana*.

Table 3: Gene set databases

Name	Organism[a]	Gene sets	Web address	Reference
ASSESS	H	Cytogenetic, pathway, motif	http://people.genome.duke.edu/~jhg9/assess/genesets.shtml	[28]
ErmineJ	H, M, R	GO	http://www.bioinformatics.ubc.ca/microannots/	[16]
GAzer	H, M, R, Y	GO, composite GO, InterPro, Pathways, TFBS	http://integromics.kobic.re.kr/GAzer/document.jsp	[13]
GSA	H	Tissue, cellular processes, cytobands, chromosome arms, 5Mb chromosomal tiles, cancer module	http://www-stat.stanford.edu/~tibs/GSA/	[26]
MSigDb	H	Cytobands, curated pathways, motif, computed	http://www.broad.mit.edu/gsea/msigdb/msigdb.index.html	[25]
PLAGE	H, M	KEGG and BioCarta pathways	http://dulci.biostat.duke.edu/pathways/misc.html	[20]

[a]H: *Homo sapiens*; M: *Mus musculus*; R: *Rattus norvegicus*; Y: *Saccharomyces cerevisiae*.

gene randomization-based methods are normally offered as web servers.

In GSA, as important as the algorithms are the gene sets. They are prepared using diverse sources of biological knowledge such as the gene ontology information, cytogenetic bands, pathways such as KEGG, GenMAPP, and Biocarta, cis-acting regulatory motifs and co-regulated genes in a microarray study. The limitations and challenges of the information on current annotation databases for IGA are thoroughly reviewed by Khatri and Draghici [1] focusing on incomplete knowledge, time-delayed curation, imprecise or incorrect electronic annotations, inability to predict new functions and semantic misclassification of annotations. The problems are all shared by GSA except for finding more relevant categories among overlapping gene sets for which GSA is able to assign different scores.

Table 3 shows currently downloadable gene set databases. Among them, GAzer [13] provides gene sets for four species, ErmineJ [16] for three species and the other three for human species only [25, 26, 28]. Molecular Signatures Database (MSigDB v2.1) provides 3337 gene sets in four categories: chromosomal locations, curated gene sets from canonical pathways and chemical and genetic perturbations, motif gene sets and computationally defined gene sets. GAzer provides gene sets for three GO categories, three composite GO categories [29], InterPro protein domain gene sets, pathway gene sets and transcription factor-binding site gene sets.

WHICH ONE TO USE?—A PRACTICAL GUIDELINE

Having discussed the pros and cons of existing GSA methods, our group now offers a practical guideline to experimental biologists for selecting an optimal tool, which of course depends on the type of experimental data. The first thing to consider is the type of organism. While most GSA tools support gene expression data for *Homo sapiens*, only a few web servers (Table 2) cover prokaryotes and many model organisms (e.g. *Caenorhabditis elegans*, *Arabidopsis thaliana* or *Drosophila melanogaster*). For rarely supported organisms, few choices exist other than available tools. For a human gene expression dataset with an enough number of samples (more than 10), GSEA is highly recommended because it is a statistically sound method based on sample randomization and provides a user-friendly, standalone program. Other similar tools such as GSA by Efron and Tibshirani [26] and SAFE provide potentially better statistical properties than GSEA, but they are currently offered as R packages that are difficult for most experimental biologists with few bioinformatics skills to use. For mouse, rat or yeast datasets for which the GSEA program is not available, web servers such as Babelomics, GAzer or GeneTrail are recommended. When the number of samples is small, gene randomization-based tools such as ErmineJ or GAzer are highly recommended.

EXTENSIONS TO GSA

GSA directly assesses the expression patterns of gene sets that are defined by shared biological themes while IGA assesses the significance of individual genes first and searches for the enriched biological themes later. In this sense, GSA is regarded as a theme-based approach. Although GSA is able to provide new information on subtle but coordinated expression patterns, it does not provide information as detailed as IGA. This theme-based approach, however, has great potential to derive much detailed information from expression data. Through the logical operations of gene sets (e.g. intersection) between different functional classifications, gene sets can be separated into more specific and smaller groups of genes, which facilitates a much more detailed analysis of expression patterns. This approach substantially complements the weak point of GSA. Indeed, Nam *et al*. [29] showed, by intersecting two GO categories, that a substantial portion of significant gene sets that have composite themes, can be newly derived. Jiang and Gentleman [30] additionally considered the subtracted gene sets between two gene sets as well as the intersection parts. Investigation in this direction is demanding and promising.

A second interesting extension is the absolute enrichment analysis that takes into account both up and down-regulated genes in calculating enrichment scores [19, 22, 26, 31]. For example, many enrichment scores are functions of the difference of means between two sample groups, and hence genes altered in the opposite directions in a gene set will cancel each other to make the score insignificant. By using absolute values on the difference of means, gene sets with bi-directional changes can be identified. This idea is helpful in identifying homeostatic systems that have bi-directional expression changes to maintain the constancy of the system [31]. The maxmean score suggested by Efron and Tibshirani [26] and some self-contained methods such as the multivariate method [21] and SAM-GS [22] also can be used to identify gene sets with bi-directional changes.

A third extension is the sample-level application of GSA. Recently, Bild *et al*. [32] and Potti *et al*. [33] developed genomic signatures to identify the activation status of oncogenic pathways and predict the sensitivity to individual chemotherapeutic drugs. Edelman *et al*. [28] developed a formal statistical method, named ASSESS (analysis of sample-set enrichment scores), to measure the enrichment of each gene set in an individual sample, and suggested extending their method into personalized identification and treatment. PLAGE [20] and domain-enhanced analysis by Liu *et al*. [34] also are able to identify the enrichment of each gene set at the

individual sample level because they first calculate gene set scores for each sample. Moreover, after an adequate normalization of expression data, many gene-sampling-based methods also can be used to identify the enrichment of each gene set in individual samples [10, 14, 16].

Finally, another useful application of GSA is that the congruency between different datasets on the same biological question increases much more when compared at a gene set level than at an individual gene level. This point was briefly mentioned by Kim and Volsky [14] and Subramanian *et al*. [25] and was later systematically investigated by Cheadle *et al*. [35] and Manoli *et al*. [36]. Additionally, GSA methods are less sensitive to the pre-processing of microarray data than are IGA methods [14, 37].

Key Points

- GSA methods are advantageous over IGA methods in many aspects; results of GSA are not affected by arbitrarily chosen cutoffs and GSA methods can detect many subtle changes in gene sets and pathways overlooked by IGA methods.
- GSA methods can be classified as either competitive or self-contained methods according to the hypotheses tested. The appropriate method is chosen depending on either the number of the samples or the property of the DEG sets the user wants to find.
- The *P*-values in a GSA can be computed by gene or sample randomization. Sample randomization provides statistically sound *P*-values.
- Using overlapping gene sets in GSA substantially improves the analysis.

Acknowledgements

The authors thank Dr Chu IS, Dr Kim SB and Dr Kim SK for insightful discussions. Supports were received from the Korea Research Institute of Bioscience and Biotechnology research initiative program [KGS2603611] and from the Korea Research Council of Fundamental Science & Technology (KRCF), Grant No. C-RESEARCH-07-08-NIMS.

References

1. Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 2005;**21**:3587–95.
2. Rivals I, Personnaz L, Taing L, *et al*. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* 2007;**23**:401–7.
3. Breslin T, Eden P, Krogh M. Comparing functional annotation analyses with Catmap. *BMC Bioinformatics* 2004;**5**:193.
4. Pan KH, Lih CJ, Cohen SN. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proc Natl Acad Sci USA* 2005;**102**:8961–5.
5. Ben-Shaul Y, Bergman H, Soreq H. Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics* 2005; **21**:1129–37.
6. Dopazo J. Functional interpretation of microarray experiments. *OMICS* 2006;**10**:398–410.
7. Mootha VK, Lindgren CM, Eriksson KF, *et al*. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003;**34**:267–73.
8. Virtaneva K, Wright FA, Tanner SM, *et al*. Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc Natl Acad Sci USA* 2001; **98**:1124–9.
9. Pavlidis P, Lewis DP, Noble WS. Exploring gene expression data with class scores. *Pac Symp Biocomput* 2002; 474–85.
10. Tian L, Greenberg SA, Kong SW, *et al*. Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A* 2005;**102**:13544–9.
11. Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 2007;**23**:980–7.
12. Ye C, Eskin E. Discovering tightly regulated and differentially expressed gene sets in whole genome expression data. *Bioinformatics* 2007;**23**:e84–90.
13. Kim SB, Yang S, Kim SK, *et al*. GAzer: gene set analyzer. *Bioinformatics* 2007;**23**:1697–9.
14. Kim SY, Volsky DJ. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* 2005;**6**:144.
15. Boorsma A, Foat BC, Vis D, *et al*. T-profiler: scoring the activity of predefined groups of genes using gene expression data. *Nucleic Acids Res* 2005;**33**:W592–5.
16. Lee HK, Braynen W, Keshav K, *et al*. ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics* 2005;**6**:269.
17. Goeman JJ, van de Geer SA, de Kort F, *et al*. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004;**20**:93–9.
18. Goeman JJ, Oosting J, Cleton-Jansen AM, *et al*. Testing association of a pathway with survival using gene expression data. *Bioinformatics* 2005;**21**:1950–7.
19. Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* 2005;**21**: 1943–9.
20. Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics* 2005;**6**:225.
21. Kong SW, Pu WT, Park PJ. A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics* 2006;**22**:2373–80.
22. Dinu I, Potter JD, Mueller T, *et al*. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics* 2007;**8**:242.
23. Allison DB, Cui X, Page GP, *et al*. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* 2006;**7**:55–65.

24. Damian D, Gorfine M. Statistical concerns about the GSEA procedure. *Nat Genet* 2004;**36**:663; author reply 663.
25. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;**102**:15545–50.
26. Efron B, Tibshirani R. On testing the significance of sets of genes. *Ann Appl Stat* 2007;**1**:107–29.
27. Delongchamp R, Lee T, Velasco C. A method for computing the overall statistical significance of a treatment effect among a group of genes. *BMC Bioinformatics* 2006; **7**(Suppl 2):S11.
28. Edelman E, Porrello A, Guinney J, *et al*. Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics* 2006;**22**:e108–116.
29. Nam D, Kim SB, Kim SK, *et al*. ADGO: analysis of differentially expressed gene sets using composite GO annotation. *Bioinformatics* 2006;**22**:2249–53.
30. Jiang Z, Gentleman R. Extensions to gene set enrichment. *Bioinformatics* 2007;**23**:306–13.
31. Saxena V, Orgill D, Kohane I. Absolute enrichment: gene set enrichment analysis for homeostatic systems. *Nucleic Acids Res* 2006;**34**:e151.
32. Bild AH, Yao G, Chang JT, *et al*. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 2006;**439**:353–7.
33. Potti A, Dressman HK, Bild A, *et al*. Genomic signatures to guide the use of chemotherapeutics. *Nat Med* 2006;**12**: 1294–300.
34. Liu J, Hughes-Oliver JM, Menius JA Jr. Domain-enhanced analysis of microarray data using GO annotations. *Bioinformatics* 2007;**23**:1225–34.
35. Cheadle C, Becker KG, Cho-Chung YS, *et al*. A rapid method for microarray cross platform comparisons using gene expression signatures. *Mol Cell Probes* 2007; **21**:35–46.
36. Manoli T, Gretz N, Grone HJ, *et al*. Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics* 2006;**22**:2500–6.
37. Raghavan N, De Bondt AM, Talloen W, *et al*. The High-level similarity of some disparate gene expression measures. *Bioinformatics* 2007;**23**:3032–8.
38. Smid M, Dorssers LC. GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate gene ontology terms. *Bioinformatics* 2004;**20**: 2618–25.
39. Volinia S, Evangelisti R, Francioso F, *et al*. GOAL: automated gene ontology analysis of expression profiles. *Nucleic Acids Res* 2004;**32**:W492–9.
40. Scheer M, Klawonn F, Munch R, *et al*. JProGO: a novel tool for the functional interpretation of prokaryotic microarray data using Gene Ontology information. *Nucleic Acids Res* 2006;**34**:W510–5.
41. Al-Shahrour F, Arbiza L, Dopazo H, *et al*. From genes to functional classes in the study of biological systems. *BMC Bioinformatics* 2007;**8**:114.
42. Backes C, Keller A, Kuentzer J, *et al*. GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Res* 2007;**35**: W186–92.
43. Cavalieri D, Castagnini C, Toti S, *et al*. Eu.Gene analyzer a tool for integrating gene expression data with pathway databases. *Bioinformatics* 2007;**23**:2631–2.
44. Newton MA, Quintana FA, den Boon JA, *et al*. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann Appl Stat* 2007; **1**:85–106.
45. Al-Shahrour F, Minguez P, Tarraga J, *et al*. BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res* 2006;**34**: W472–6.
46. Tu K, Yu H, Zhu M. MEGO: gene functional module expression based on gene ontology. *Biotechniques* 2005;**38**: 277–83.