Collection of Biostatistics Research Archive

COBRA Preprint Series

Year 2007 *Paper* 16

Improving GSEA for Analysis of Biologic Pathways for Differential Gene Expression across a Binary Phenotype

Irina Dinu, *Department of Public Health Sciences, School of Public Health, University of Alberta*
John D. Potter, *Division of Public Health Sciences, Fred Hutchinson Cancer Research Center*
Thomas Mueller, *Division of Nephrology & Transplantation Immunology, Faculty of Medicine and Dentistry, University of Alberta*
Qi Liu, *Department of Public Health Sciences, School of Public Health, University of Alberta*
Adeniyi J. Adewale, *Department of Public Health Sciences, School of Public Health, University of Alberta*
Gian S. Jhangri, *Department of Public Health Sciences, School of Public Health, University of Alberta*
Gunilla Einecke, *Division of Nephrology & Transplantation Immunology, Faculty of Medicine and Dentistry, University of Alberta*
Konrad S. Famulski, *Division of Nephrology & Transplantation Immunology, Faculty of Medicine and Dentistry, University of Alberta*
Philip Halloran, *Division of Nephrology & Transplantation Immunology, Faculty of Medicine and Dentistry, University of Alberta*
Yutaka Yasui, *Department of Public Health Sciences, School of Public Health, University of Alberta*

Improving GSEA for Analysis of Biologic Pathways for Differential Gene Expression across a Binary Phenotype

Irina Dinu, John D. Potter, Thomas Mueller, Qi Liu, Adeniyi J. Adewale, Gian S. Jhangri, Gunilla Einecke, Konrad S. Famulski, Philip Halloran, and Yutaka Yasui

Abstract

Gene-set analysis evaluates the expression of biological pathways, or a priori defined gene sets, rather than that of single genes, in association with a binary phenotype, and is of great biologic interest in many DNA microarray studies. Gene Set Enrichment Analysis (GSEA) has been applied widely as a tool for gene-set analyses. We describe here some critical problems with GSEA and propose an alternative method by extending the single-gene analysis method, Significance Analysis of Microarray (SAM), to gene-set analyses (SAM-GS). Specifically, we illustrate, in a simulation study, that GSEA gives statistical significance to gene sets that have no gene associated with the phenotype (null gene sets), and has very low power to detect gene sets in which half the genes are highly associated with the phenotype (truly-associated gene sets). SAM-GS, on the other hand, performs perfectly in the simulation study: none of the null gene sets is identified with statistical significance, while all of the truly-associated gene sets are. The two methods are also compared in the analyses of three real microarray datasets and relevant pathways, the diverging results of which clearly show the advantages of SAM-GS over GSEA, both statistically and biologically.

http://biostats.bepress.com/cobra/art16

1. Introduction

Some DNA microarray studies may target discovery of single genes that are associated with a phenotype. Useful statistical approaches have been proposed for such single-gene analyses, for example, Significance Analysis of Microarray (SAM) in (1). In many instances, however, the goal of studies is in the assessment of biologic pathways, or *a priori* defined gene sets, in association with a phenotype, i.e., gene-set analyses. Computationally, gene-set analyses require an additional consideration over single-gene analyses, namely, the incorporation of gene sets into an association measure. Mootha et al. (2) proposed Gene Set Enrichment Analysis (GSEA) for gene-set analysis, utilizing the Kolmogorov-Smirnov statistic to measure the degree of differential gene expression in a gene set across binary phenotypes. GSEA was revised in 2005 by the same research team, replacing the Kolmogorov-Smirnov statistic with its weighted version to avoid certain deficiencies in the original GSEA method (3).

We propose here an alternative approach, an extension of SAM, to gene-set analysis, called hereafter SAM-GS. This is motivated by our observation that GSEA, in both the original and revised versions, fails to satisfy certain required properties that a gene-set analysis method should satisfy: for example, a gene-set analysis should not indicate an association for a gene set in which no gene is associated with the phenotype. In this paper, we first illustrate the behavior of GSEA in relation to a few required properties of a gene-set analysis method and compare it with the behavior of SAM-GS, using a mouse-microarray kidney-transplant dataset. We then re-analyze, by SAM-GS, three DNA microarray datasets with which the application of GSEA was illustrated in (3), showing appreciable differences in the analysis results. The differences of the results are discussed from both biologic and statistical points of view, pointing out clear advantages of SAM-GS over GSEA.

2. Methods

<u>GSEA for gene-set analyses</u>

A gene-set analysis for an *a priori* defined set of genes S in a total of N genes (or probes) on a DNA microarray is a test of the null hypothesis that the expression pattern of S is not associated with a phenotype of interest, D. To simplify the discussion, we will consider only a phenotype with two categories, {0, 1}: e.g. presence or absence of a disease. As biologists are often interested in testing multiple gene-sets $\{S_1, \ldots, S_k\}$, we will also consider a gene-set analysis for multiple gene-sets, following our discussion of an individual gene-set.

The revised version of GSEA by (3), for *an individual gene-set***,** proceeds as follows.

GSEA Steps

1) Compute the Pearson correlation (or another metric) between each of the N genes with a phenotype D, where the correlation or another metric of the i^{th} gene is denoted by r_i.

2) Order the N genes by their correlation values from the maximum to the minimum (the ordered list is denoted by L).

3) Compute the Enrichment Score (ES): start with ES = 0; walk down the ranked list L, from the top rank (*i*=1) to the last rank (*i*=*N*), increasing ES by $|r_i|/\sum_{j\in S}|r_j|$ if the i^{th} gene belongs to the gene set *S*, and decreasing ES by $1/(N-|S|)$ otherwise, where $|S|$ is the number of genes in the set *S*.

4) Take the absolute value of the maximum deviation from zero of the ES values among the *N* genes as the test statistic for the gene set *S*.

5) Permute the labels of the phenotype *D* and repeat steps 1)- 4). Repeat until all (or a large number of) permutations are considered.

6) Statistical significance for the association of *S* and *D* is obtained by comparing the observed value of the test statistic from 3) and its permutation distribution from 5).

The initial version of GSEA proposed in Step (2) used $1/|S|$, instead of $|r_i|/\sum_{j\in S}|r_j|$, for increasing the ES for each gene in *S*. The use of $|r_i|/\sum_{j\in S}|r_j|$, or more generally $|r_i|^p/\sum_{j\in S}|r_j|^p$, was motivated by the need to reduce the ES values and the statistical significance of sets clustered near the middle of the ranked list (see Figure 1 and Table 1 in (3)). Although the modified version of GSEA was aimed at reducing the statistical significance of sets not exhibiting biologically relevant correlation with the phenotype, serious problems remain with GSEA as demonstrated below.

The proposed method, SAM-GS

The main aim of analyzing *an individual gene-set* is to distinguish between the two biologic conditions (phenotype) based on multivariate measurements of the expression of genes in the gene set. GSEA tests a null hypothesis that rankings of the genes in a gene set according to an association measure with the phenotype categories (e.g., correlation) are randomly distributed over the rankings of all genes, using Kolmogorov-Smirnov statistic. SAM-GS, on the other hand, tests a null hypothesis that the mean vectors of expressions of genes in a gene set does not differ by the phenotype of interest.

Our proposed SAM-GS method is based on individual t-like statistics from SAM, addressing the small variability problem encountered in microarray data, i.e., reducing the statistical significance associated with genes with very little variation in their expressions. SAM-GS for *an individual gene-set* can be summarized in a few steps.

SAM-GS Steps:
1) For each of the *N* genes, calculate the statistic *d* as in SAM for a single-gene analysis:

$$d_i = \frac{\overline{x}_1(i) - \overline{x}_2(i)}{s(i) + s_0},$$

where the 'gene-specific scatter' $s(i)$ is a pooled standard deviation over the two groups of the phenotype, and s_0 is a small positive constant that adjusts for the small variability encountered in microarray data (1).

2) Compute the *SAMGS* test statistic corresponding to set S:

$$SAMGS = \sum_{i=1}^{|S|} d_i^2$$

3) Permute the labels of the phenotype *D* and repeat 1) and 2). Repeat until all (or a large number of) permutations are considered.

4) Statistical significance for the association of *S* and *D* is obtained by comparing the observed value of the *SAMGS* statistic from 2) and its permutation distribution from 3).

Note that SAM-GS initially measures the gene-expression difference across the binary phenotype in each gene *i* of the gene set *S* using d_i, where the differences are standardized across the genes for their degrees of scatter with the denominators of d_i's, $\{s(i) + s_0\}$. It then summarizes these standardized differences in all the genes in the gene set *S* by *SAMGS*. The analysis of multiple gene sets can be accommodated in SAM-GS by estimating false discovery rates (FDRs) from p-values of individual sets using the q-value method of (6).

A Simulation Experiment

To illustrate the differences between SAM-GS and GSEA, we compared them on the following requisite properties for any method designed to perform a gene-set analysis:

(a) If the gene set *S* consists of genes that are consistently not correlated, or are variably weakly correlated and not correlated, with the phenotype *D* (e.g., all genes with small values of $|r|$), the method should not indicate that *S* is associated with *D*.
(b) If the gene set *S* consists of a mix of genes with high- and low-correlation with the phenotype, such that an appreciable subset of the genes in *S* are highly correlated with the phenotype *D* (e.g., half of the genes in *S* with large values of $|r|$), the method should indicate that *S* is associated with *D*.
(c) The size of the gene set *S* should not greatly alter the statistical significance in (a) and (b).

We performed two simulations tests to compare the performance of GSEA and SAM-GS.

Test 1: Sample n genes by a simple random sampling as a hypothetical gene set from a group of genes with low correlation with the phenotype, for example, genes with $|r| < 0.4$ or 0.1. Test the association of this *n*-gene set with the phenotype. Repeat 100 times to check property (a).

Test 2: Sample n genes by a stratified random sampling as a hypothetical gene set such that half of the genes in the set are highly correlated with the phenotype, for example, with $|r| \geq 0.6$ or 0.7, and the other half with $|r| < 0.6$ or 0.7. Test the association of this *n*-gene set with the phenotype. Repeat 100 times to check property (b).

In Test 1, our simple random sampling from the no-or-weak correlation region creates gene sets that approximate to the null hypothesis such that their members are consistently not correlated with the phenotype (i.e., they have a mixture of genes with correlations between –0.1 and 0.1) or are variably weakly correlated and not correlated (correlations between –0.4 and 0.4, including many around zero). These gene sets should not be called significantly associated with the phenotype. In Test 2, as half of the genes in the gene set are highly correlated with the phenotype, these gene sets should be identified as significantly associated with the phenotype.

The two tests were performed based on the data from the mouse-microarray kidney transplant study. In this study, we compared two experimental groups of mouse kidney transplants: fully MHC mismatched allografts and MHC identical isografts. A more detailed description of the study is given in the Appendix. Briefly, in both groups, the kidneys have undergone the same surgical procedure of transplantation, but in addition the allograft develops the histologic lesions of rejection due to the immune response by the host, while the isograft does not develop these lesions due to an identical genetic background. We have studied a full timecourse between days 1 and 42 post transplant; the alloimmune response is fully developed at days 5-7, and the injury response in the isografts also peaks at days 5-6. To simplify the comparison between rejecting allografts and non-rejecting isografts, we have therefore selected the data from days 5 and 7 as the basis of this analysis. A total of 12 samples were analyzed: 3 samples each at day 5 and day 7 in allografts, 4 samples in day 5 isografts, and 2 samples in day 7 isografts. The microarray data were obtained by hybridizing RNA to Affymetrix MOE 430 2.0 microarrays. These arrays contain 45,099 probesets. We considered a gene set size *n* of 10, 30, 50, and 100.

Real data analyses

We compared the performance of the two methods, GSEA and SAM-GS, on the analyses of biologically defined gene sets using three microarray datasets considered in (3): male vs. female lymphoblastoid cells; p53 wild-type vs. mutant cancer cell lines; and ALL vs. AML leukemia cells. The comparison used GSEA results for the three examples, downloaded from GSEA web-page: http://www.broad.mit.edu/gsea. To run SAM-GS, we downloaded the datasets and gene-set subcatalogs C1 and C2 from the above address. The same datasets and subcatalogs were used for both GSEA and SAM-GS.

We did not analyze the lung adenocarcinoma data of three studies (Boston, Michigan, and Stanford studies) in (3) as such an analysis is methodologically problematic: the Michigan study included only patients with stage I or III lung adenocarcinoma, whereas the Boston and Stanford studies did not restrict the stages; the binary phenotype of interest, death, was defined using censored survival data, where the length of follow-up to ascertain death varied appreciably both by patient and across studies (the median follow-up was 49.9, 29.5, and 17.5 months in the Boston, Michigan, and Stanford studies, respectively), leading to inconsistent ascertainment of the binary phenotype (death) across patients and studies (patients with a longer follow-up had a higher chance of being ascertained to have died); and no adjustment was applied to control for possible differences across the studies in treatment, tumor characteristics, and demographics of the patients.

3. Results

Simulation experiments

We used two simulation tests to check whether GSEA and SAM-GS satisfy the properties of gene-set analysis methods described in the **Methods** Section. The distribution of the Pearson correlation coefficients of all genes with the phenotype in the mouse data is given in the Appendix. To run GSEA, we used the Desktop application downloaded from http://www.broad.mit.edu/gsea, and the options specified in (3), that is, the Pearson correlation of the gene expressions with the phenotype to rank the genes, and weighted ES with the default option $p = 1$.

Figure 1 shows the ES walk for one of the sets (GSEA p-value = 0), generated under Test 1. The results shown in Table 1 indicate that GSEA does not satisfy requisite properties (a) or (b). Moreover, GSEA does not satisfy property (c), as the performance of the method varies greatly with gene-set size.

[Figure 1 about here.]

[Table 1 about here.]

The results of these tests illustrate two situations where GSEA fails. One is where genes in a gene set cluster somewhere other than in the high-correlation region (e.g., all individual genes could have no or very low correlation with the phenotype but GSEA indicates that the total gene set is statistically significantly associated with the phenotype). In short, GSEA will indicate that gene sets with clear clustering are statistically significant, regardless of where the clustering occurs. The other situation is where a gene set has a mixture of highly correlated genes and weakly correlated genes. This mixture within a gene set seems biologically plausible: not all genes in a phenotype-associated pathway will show changes in relation to the phenotype. Indeed, this is the basis on which GSEA was originally applied to three lung cancer sets to show that lack of consistent genes did not mean lack of common gene sets. In such mixed situations, GSEA has very poor power.

To check whether SAM-GS satisfies the three requisite properties of a gene-set analysis, the same tests were applied as for GSEA, using the same randomly-sampled simulated gene sets. The results of Tests 1 and 2 displayed in Table 1 indicate that SAM-GS satisfies properties (a) and (b). Moreover, SAM-GS satisfies property (c), as its performance did not vary with the size of the gene set.

Gene-set analyses of the three datasets

We compared the performance of the two methods, GSEA and SAM-GS, on the analyses of biologically defined gene sets using three microarray datasets considered in (3): male vs. female lymphoblastoid cells; p53 wild-type vs. mutant cancer cell lines; and ALL vs. AML leukemia cells. The analysis results by GSEA and SAM-GS are summarized in

Table 2. In the sex-comparison analysis, both methods showed associations (FDR ≤ 0.01) with the three Y associated gene sets, the testis-expressed gene set (GSEA FDR = 0.02), and the gene set with genes that escape X inactivation. In addition, SAM-GS established an association with the chrXp22 gene set (SAM-GS FDR ≤ 0.01 vs. GSEA FDR = 1.00). In the p53-comparison analysis, SAM-GS and GSEA identified a subset of gene sets with an FDR ≤ 0.01 that includes the gene sets of hsp27, p53_UP (GSEA FDR = 0.013), p53 hypoxia, radiation sensitivity (GSEA FDR = 0.07), and p53 (BioCarta). However, SAM-GS identified additional 31 gene sets with an FDR ≤ 0.01, all of which had an FDR ≥ 0.49 for GSEA. These gene sets are shown in Table 3. In the leukemia-comparison dataset, the two methods gave even more discrepant results than the p53-comparison analysis. GSEA identified only five gene sets with an FDR ≤ 0.25 (none with an FDR ≤ 0.01), whereas all of the 182 gene sets were statistically significant (FDR ≤ 0.01) by SAM-GS. Note that the single-gene analysis showed that 80% of the single genes in this comparison had an FDR ≤ 0.25, which is in line with the gene-set analysis results of SAM-GS.

[Tables 2 and 3 about here.]

These discrepancies between the two methods are summarized along with the sensitivity and specificity of the GSEA p-value ≤ 0.05 and the area under the receiver operating characteristic curve of GSEA p-value in predicting the SAM-GS p-value ≤ 0.05 (Table 2).

4. Discussion

GSEA

Our Tests 1 and 2 showed that GSEA does not meet some requisite criteria for a gene-set analysis method. Although the gene sets in Tests 1 and 2 are randomly-sampled simulated sets, they are not unrealistic gene sets. For example, a Test 1 situation was encountered in the analysis of the sex dataset, where GSEA gave the "cell-cycle arrest genes" a p-value of 0.015 in association with sex (SAM-GS p-value = 0.84). No gene in this gene set has an absolute value of the Pearson correlation of 0.33 or greater, or the SAM p-value < 0.06: this clustering is thus identified incorrectly by GSEA as showing a significant association, failing Test 1. A Test 2 situation was encountered, for example, in the analysis of the leukemia dataset, where GSEA failed to identify the gene set "chr10q24", even though 13 of the 43 genes in the gene set had absolute values of the Pearson correlation of 0.5 or greater (4 genes greater than 0.7) and the chromosomal location of the gene set is biologically relevant given the role of *HOX11* in T-cell ALL. The use of GSEA is subject to appreciable false positive and negative findings, as illustrated by the two tests and the results shown in Table 2.

Another critical problem of GSEA is its *relative* ranking of genes in a gene set in relation to the other genes outside of the gene set. The use of a relative measure in GSEA, rather than an absolute measure, means that important information on the degree of association between each gene and the binary phenotype is discarded. For example, the leukemia dataset had 80% of its 10,056 individual genes with an FDR ≤ 0.25. Regardless of whether such clear differences in gene expression across the binary phenotype are determined by biology, or by more mundane (and biologically irrelevant) differences in sample collection or handling, a gene-set analysis of this dataset should find that many gene sets are associated with the phenotype. GSEA, however, found only five gene sets with an FDR ≤ 0.25 in the leukemia-comparison analysis, inconsistent with the single-gene analysis results. The cause of the inconsistency is the use of the relative ranking in GSEA. In contrast, SAM-GS found all gene sets in the leukemia dataset to have an FDR ≤ 0.01.

A related, perhaps less serious, issue with GSEA is that, when a single gene set is of biologic interest, the SAM-GS analysis requires measurement only of the expression of the genes in the gene set to construct the test statistic (except the calculation of s_0), whereas GSEA requires measurement of the expression of all genes to provide a relative ranking of all genes. The expression levels of the other genes should not affect the inference on a single gene set of interest, if the single set is, indeed, the only biologically relevant variable.

Another problematic aspect of GSEA is that its enrichment score considers genes with positive and negative associations with the phenotype in a contradictory manner. Specifically, positively associated genes increase the test statistic, whereas negatively associated genes decrease it, even if they are all members of a gene set and both are associated with the phenotype. Thus, a gene set with a mix of genes with positive and negative associations with the phenotype, although biologically plausible (for instance, due to feedback loops in pathways) is not appropriately evaluated for association with the phenotype by the enrichment score and, therefore, has an improperly low probability of being detected as a phenotype-associated gene set by GSEA.

A gene-set analysis utilizes existing biologic knowledge that maps single genes into gene sets or pathways. Because of the utilization of existing knowledge in the analysis, a well conducted gene-set analysis can be remarkably powerful. The p53 analysis illustrates this point. Although a very small proportion of individual genes had low p-values in the p53 dataset, SAM-GS indicated larger proportions of gene sets with low p-values. This is because a valid gene-set analysis would take into account a tendency among multiple genes in a gene set. Thus, even if the association of each gene with the phenotype is only moderate, a collection of such genes can be indicated as a phenotype-associated gene set; genes in a gene set need not have the same degree and direction of association with the phenotype for the gene set to be identified as statistically significant by SAM-GS.

In addition to the leukemia-comparison analysis discussed above, which showed an advantage of SAM-GS over GSEA empirically through the consistency of the gene-set analysis results with the single-gene analysis results, the other two DNA-microarray analyses (sex- and p53-comparison analyses) provided empirical biologic evidence supporting the advantage of SAM-GS over GSEA. Regarding the sex-comparison analysis, Subramanian et al. (3) specifically argue that they would not expect to find enrichment for bands on the X chromosome because most X-linked genes are randomly silenced in females and, therefore, are unlikely to show a male-female (gene-dose) difference. This argument has general merit; however, in the specific case of the chrXp22 gene set, it does not hold because, on the distal portion of the short arm of X, there is a cluster of genes that escape X-inactivation. Indeed, the top five genes of the chrXp22 gene set escape inactivation: two of the five are members of the X-inactivation-escape gene set whose FDR was ≤ 0.01 by both methods; and the other three have been reported to escape X-inactivation (7-9).

The differences in the results of the p53-comparison analysis illuminate biologically relevant performance differences between the two methods. It is appropriate to ask whether the 31 additional pathways identified by SAM-GS over GSEA are plausibly associated with the presence vs. absence of p53. Of the 31 gene sets, 11 actually involve p53 directly as a member. A further 6 gene sets directly involve the extrinsic and intrinsic apoptosis pathways (10), 3 involve the cell-cycle machinery, and 3 involve cytokines and/or JAK/STAT signaling (11). Each of these 12 gene sets, then, is in a direct, well-established relationship with aspects of p53 signaling. Of the remaining 8 gene sets, 6 have plausible, if less well established, links with p53. In the Ck1 pathway, cdk5 phosphorylates p53 so the presence vs. absence of p53 is likely to modify profoundly the effectiveness of this pathway (12). Ets1 (ets pathway) has been shown to be essential, in mouse embryonic stem cells, to maintain the ability to undergo UV-induced, p53-dependent apoptosis. Ets1, more broadly, may be necessary for p53-dependent gene transactivation (13). Akt and p53 are, respectively, essential to the transduction of anti-apoptotic and pro-apoptotic pathways. There is an integrated negative feedback loop whereby p53-dependent down regulation of Akt promotes cell death but cell survival signals will recruit Akt, leading to activation of Mdm2 and the inhibition of p53-dependent apoptosis (14). This may account, in part, for the association between the presence vs. absence of p53 and differences in the SA-TRKA receptor pathway. Proline oxidase is induced by p53 and mediates apoptosis via a calcineurin-dependent pathway (15). Coproporphyrinogen oxidase (CPO) is a key compound of the MAP 00860 porphyrin/chlorophyll metabolism gene set. It catalyzes a rate-limiting step in heme biosynthesis and may contribute to mitochondrial redox balance. It has recently been shown to be regulated by p53 (16). Finally, the Wnt and p53 pathways have also been shown to be linked via pro-apoptotic Dkk1, a wnt antagonist (17).

SAM-GS

Regarding the form of *SAMGS* test statistic, $\sum_{i=1}^{|S|} d_i^2$ is simply the L_2-norm of the t-like-statistic vector $d = (d_1, d_2, \cdots, d_{|s|})$, the length of the line segment joining the two phenotypes' mean gene-expression vectors of a gene set *S*. Our null hypothesis is that the mean vectors of expressions of genes in a gene set *S* do not differ by the phenotype of interest (i.e., this line-segment length is zero), a two-sample multivariate mean test in statistics. The classical multivariate statistics method for a two-sample mean test, Hotelling's T^2, addresses this question, but it cannot be applied when $|S| > n_1 + n_2 - 2$, where n_1 and n_2 are the sample sizes in the two groups defining the phenotype *D*. We would like to emphasize that this condition is often unmet in gene-set analyses of DNA microarray data. Dempster (4,5) introduced a test statistic for comparing highly multivariate samples of two groups, an alternative for Hotelling's T^2, when the number of multivariate measurements is large, relative to the sample sizes. Using Dempster's test in the context of microarray data, a potential candidate for a test statistic to be used in Step 2 of SAM-GS, would be the weighted Dempster's (WD) statistic:

$$WD = \sum_{i=1}^{|S|} d_i^2 / E[\sum_{i=1}^{|S|} d_i^2],$$

where $E[\sum_{i=1}^{|S|} d_i^2]$ in the denominator is the average of ($n_1 + n_2 - 2$) statistically-independent quantities that have the same mean and variance as the numerator $\sum_{i=1}^{|S|} d_i^2$ under the null hypothesis, created by an orthonormal transformation of multivariate gene expressions in the set *S*. This test statistic seems to have the advantage of taking into account the multivariate structure of the gene expression measurements in a gene set by dividing the numerator, the L_2-norm of the mean-vector difference, by its approximate expectation. However, since a permutation-based test is used, the denominator of WD statistic is unnecessary: as Dempster (5) stated, a permutation test based on the numerator only is equivalent to using the quotient. Given the computational simplicity and the use of permutation in SAM-GS, the L_2-norm used in *SAMGS* is preferred over WD.

The L_1-norm of $d = (d_1, d_2, \cdots, d_{|s|})$, $\sum_{i=1}^{|S|} | d_i |$, can be considered, similar to Chung and Fraser (18) who proposed the L_1-norm as an alternative to Dempster's use of the L_2-norm. While one might expect the two norms to give similar performances overall, since the L_1-norm would be less sensitive to extreme values than the L_2-norm, the L_1-norm may be less powerful in detecting a gene-set with a small number of genes being strongly associated with the phenotype. Test 2 simulation above confirmed this point: as the proportion of genes in a gene set that are correlated with the phenotype ($| r |\geq 0.6$) becomes smaller than approximately 30%, the two norms performs differently and the L_1-norm is less powerful in detecting the gene set being associated with the phenotype (data not shown).

To account for multiple comparisons (statistical testing of many hypotheses) when multiple gene sets are to be tested, SAM-GS takes the same approach as SAM, estimating a q-value, an upper limit for the FDR, for each gene set. The q-value of a gene set can be determined solely from the p-values of all gene sets tested (6). The collection of p-values of all gene sets contains information, not only on the statistical significance of each gene for its association with the phenotype, but also on the proportion of gene sets that are not associated with the phenotype, the "null gene-set proportion." Note that the null gene-set proportion is determined by biology: the phenotype is either biologically associated or not associated with each gene set. However, the p-value is a function of sample sizes and noise levels in gene-expression measurements as well as the degree of underlying biological associations. Thus, even if a strong biologic association between a gene set and the phenotype exists, because of small sample sizes and/or high measurement noise levels (features of many DNA microarray observations and experiments), the p-value of the gene set can be large. This is another aspect of the p53 analysis discussed above, where many gene sets have low FDR estimates in spite of the fact that the p-values are not correspondingly low: this is due to an estimated small null gene-set proportion which lowers FDR estimates.

In conclusion, GSEA has some serious problems as a method for gene-set analysis, potentially leading to unnecessarily high false-positive and false-negative discovery rates. SAM-GS, based on the SAM t-like statistic, on the other hand, is statistically sound and has advantages, as illustrated in this paper, from both statistical and empirical biologic perspectives.

An Excel Add-In for performing SAM-GS is available for public use at http://www.ualberta.ca/~yyasui/homepage.html.

Acknowledgements

JDP is supported by grant CA074794 – the Seattle Colorectal Family Registry – from the National Cancer Institute. The Halloran Lab. is supported by Genome Canada, Genome Alberta, Roche Molecular Systems, Hoffmann La-Roche Canada, University of Alberta Hospitals Foundation, Alberta Innovation and Science, Roche Organ Transplant Research Foundation, the Canadian Institutes of Health Research, Kidney Foundation of Canada, Roche Germany, Astellas Canada, and the Muttart Foundation. PH and YY are Canada Research Chairs in Transplant Immunology and Biostatistics/Epidemiologic Methods, respectively, supported by the Canada Research Chair Program. ID and YY are supported by the Alberta Heritage Foundation for Medical Research and YY is supported by Canadian Institute of Health Research.

Reference

1. Tusher, V. G., Tibshirani, R. & Chu, G. (2001) *Proc Natl Acad Sci U S A* **98,** 5116-21.
2. Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D. & Groop, L. C. (2003) *Nat Genet* **34,** 267-73.
3. 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. & Mesirov, J. P. (2005) *Proc Natl Acad Sci U S A* **102,** 15545-50.
4. Dempster, A. P. (1958) *The Annals of Mathematical Statistics* **29,** 995-1010.
5. Dempster, A. P. (1960) *Biometrics* **16,** 41-50.
6. Storey, J. D. (2002) *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64,** 479-98.
7. Yen, P.H., Ellison, J., Salido, E.C., Mohandas, T., & Shapiro, L. (1992) *Hum Mol Genet* **1**, 47-52.
8. Goodfellow, P., Pym, B., Mohandas, T., & Shapiro, L.J. (1984) *Am J Hum Genet*. **36**, 777-82.
9. Craig, I.W., Mill, J., Craig, G.M., Loat, C., & Schalkwyk, L.C. (2004) *European Journal of Human Genetics* **12**, 639–46.
10. Cory, S. & Adams, J.M. (2002) *Nat Rev Cancer* **2,** 647-56.
11. Verma, A., Kambhampati, S., Parmar, S. & Platanias, L.C. (2003) *Cancer Metastasis Rev* **22**, 423-34.
12. Zhang, J., Krishnamurthy, P.K. & Johnson, G.V. (2002) *J Neurochem* **81,** 307-13.
13. Xu, D., Wilson, T.J., Chan D., Luca E.D, Zhou, J. Hertzog P.J. & Kola, I. (2002) *Embo J* **21**, 4081-93.
14. Gottlieb, T.M., Leal, J.F.M., Seger, R., Taya, Y. & Oren, M. (2002) *Oncogene* **21**, 1299-303.
15. Rivera, A. & Maxwell, S.A. (2005) *J Biol Chem.* **12,** 29346-54.
16. Mann, K. & Hainaut, P. (2005) *Oncogene* **24**, 3964-75.
17. Shou, J., Ali-Osman, F., Multani, A.S., Pathak, S., Fedi, P. & Srivenugopal, K.S. (2002) *Oncogene* **21**, 878-89.
18. Chung, J. H. & Fraser, D. A. S. (1958) *Journal of the American Statistical Association* **53,** 729-35.

Figure 1. An illustration of a statistically-significant GSEA result with 100 genes selected at random from weakly correlated genes with $| r | < .4$.

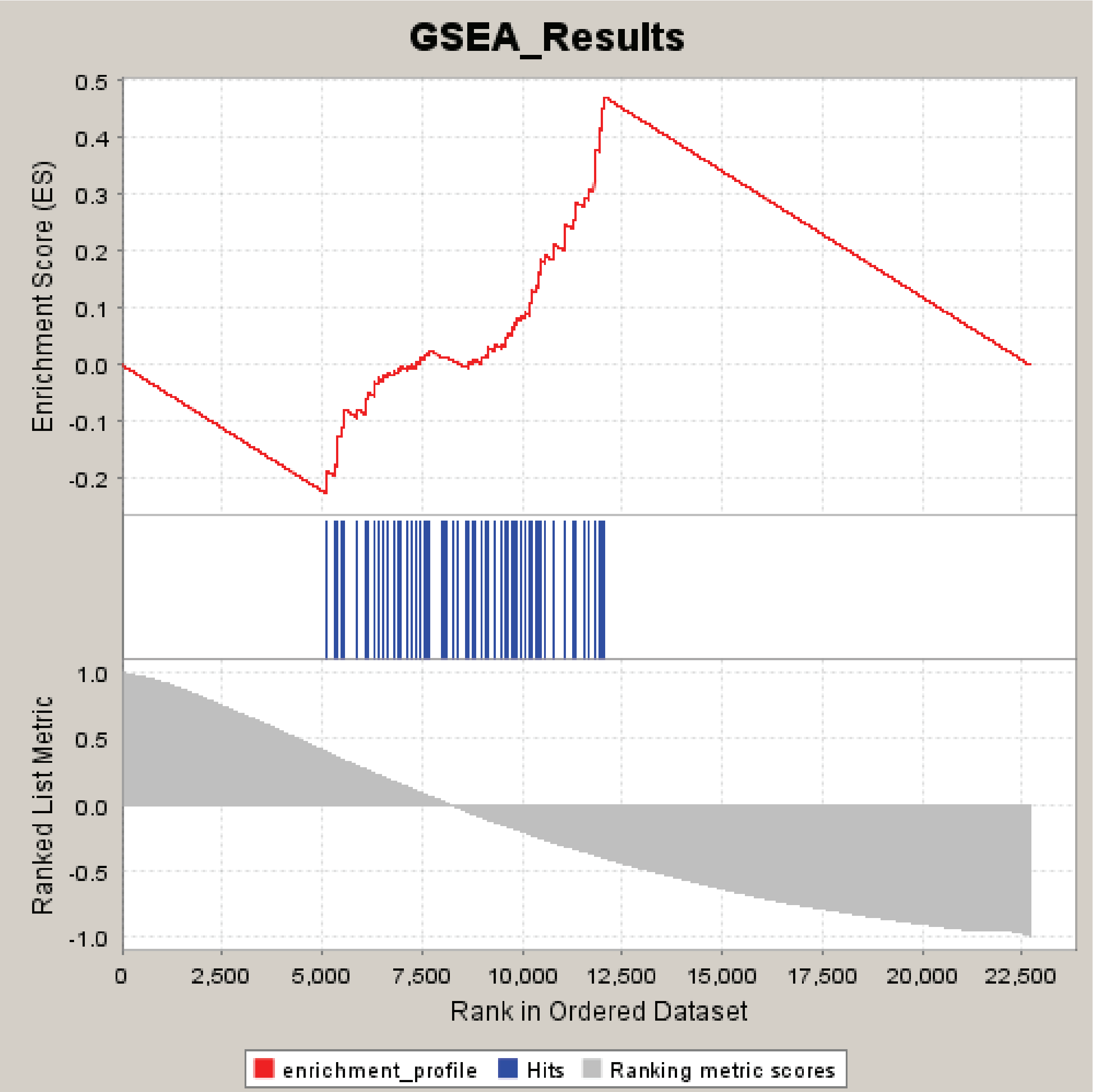

Table 1. Performance of GSEA and SAM-GS in testing the statistical significance (p-value ≤ 0.05) of hypothetical gene sets in relation to the phenotype in a mouse-microarray study.

Test	**Pearson correlation of genes in the gene set with the phenotype**	**Methods**	**Set Size**			
			10	**30**	**50**	**100**
1	$\vert r\vert<.1$	GSEA	61%	100%	100%	100%
		SAM-GS	0%	0%	0%	0%
	$\vert r\vert<.4$	GSEA	0%	42%	76%	99%
		SAM-GS	0%	0%	0%	0%
2	Half of genes with $\vert r\vert\geq .6$, other half with $\vert r\vert<.6$.	GSEA	1%	6%	15%	21%
		SAM-GS	100%	100%	100%	100%
	Half of genes with $\vert r\vert\geq .7$, other half with $\vert r\vert<.7$.	GSEA	9%	22%	39%	74%
		SAM-GS	100%	100%	100%	100%

Table 2. Results of the analyses of three real datasets by GSEA and SAM-GS

Dataset	% of probes with FDR* ≤ 0.25	# of gene sets with FDR ≤ 0.01		# of gene sets with FDR ≤ 0.25		Sensitivity/ Specificity (AUC[†]) of GSEA[‡]
		GSEA	SAM-GS	GSEA	SAM-GS	
Sex	0.1%	4	5	6	5	0.78 / 0.98 (0.94)
p53	0.3%	3	36	6	308	0.21 / 0.94 (0.68)
Leukemia	79.9%	0	182	5	182	0.06 / NA[§] (NA[§])

[*] FDR = False discovery rate estimate

[†] AUC = Area under the ROC curve

[‡] Taking SAM-GS p≤0.05 as the target to be predicted

[§] All gene sets in the leukemia dataset had p≤0.05

Table 3. The 31 gene sets for which SAM-GS and GSEA strongly disagreed (SAM-GS FDR ≤ 0.01, but GSEA FDR ≥ 0.49) in the p53-comparison analysis

Gene Set	GSEA		SAM-GS		p53 LINK
	FDR	p-value	FDR	p-value	
ATM Pathway	0.87	0.21	≤ 0.01	< 0.001	Pathway member
BAD Pathway	0.57	0.04	≤ 0.01	< 0.001	Apoptosis
Calcineurin Pathway	0.84	0.13	≤ 0.01	< 0.001	p53-induced proline oxidase mediates apoptosis via a calcineurin-dependent pathway (15)
Cell cycle regulator	0.90	0.29	≤ 0.01	< 0.001	Cell cycle
Mitochondria pathway	0.88	0.32	≤ 0.01	< 0.001	Apoptosis
p53 signaling pathway	0.51	0.01	≤ 0.01	< 0.001	Pathway member
Raccycd Pathway	0.83	0.56	≤ 0.01	< 0.001	Cell cycle
SA_TRKA_RECEPTOR	0.83	0.34	≤ 0.01	< 0.001	Integrated negative feedback loop between Akt and p53 (14)
bcl2family and reg. network	0.83	0.42	≤ 0.01	0.001	Apoptosis
Cell cycle arrest	0.98	0.49	≤ 0.01	0.001	Cell cycle
Ceramide Pathway	0.88	0.30	≤ 0.01	0.001	Apoptosis
DNA DAMAGE SIGNALLING	0.85	0.23	≤ 0.01	0.002	Pathway member
SIG_IL4RECEPTOR_IN_B_LYMPHOCYTES	0.93	0.27	≤ 0.01	0.002	Cytokines; JAK/STAT signaling
Cell cycle Pathway	0.89	0.72	≤ 0.01	0.003	Pathway member
G2 Pathway	0.81	0.50	≤ 0.01	0.003	Pathway member
Chemical Pathway	0.53	0.04	≤ 0.01	0.005	Pathway member
Drug resistance and metabolism	0.86	0.08	≤ 0.01	0.005	Pathway member
G1 Pathway	0.81	0.37	≤ 0.01	0.005	Pathway member
Breast cancer estrogen signaling	1.00	0.85	≤ 0.01	0.006	Pathway member
Ca_nf_at_signaling	0.78	0.08	≤ 0.01	0.007	Apoptosis (and cytokines)
Cytokine Pathway	0.53	0.05	≤ 0.01	0.007	Cytokines
ST_Interleukin_4_Pathway	0.84	0.07	≤ 0.01	0.007	Cytokines; JAK/STAT signaling
CR_DEATH	0.86	0.31	≤ 0.01	0.008	Pathway member
MAP00860: Porphyrin & chlorophyll metabolism	0.92	0.29	≤ 0.01	0.010	CPO regulated by p53 (16)
Ck1 Pathway	0.49	0.02	≤ 0.01	0.011	Cdk5 phosphorylates p53 (12)
Hivnef Pathway	0.95	0.48	≤ 0.01	0.011	Apoptosis
Ets Pathway	0.79	0.45	≤ 0.01	0.012	Ets1 required for p53 transcriptional activation in

					UV-induced apoptosis (13)
ST_Wnt_Ca2_cyclic_GMP_Pathway	0.80	0.13	≤ 0.01	0.012	At least one known link between wnt and p53 (17)
Chrebp Pathway	0.84	0.42	≤ 0.01	0.013	unknown
GPCRs_Class_A_Rhodopsin-like	0.60	0.04	≤ 0.01	0.013	unknown
ST_Fas_Signaling_Pathway	0.80	0.52	≤ 0.01	0.013	Pathway member