BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 3 2007, pages 306–313
doi:10.1093/bioinformatics/btl599

Gene expression

Extensions to gene set enrichment

Zhen Jiang* and Robert Gentleman

Computational Biology, 1100 Fairview Avenue. N. M2-B876, PO Box 19024, Seattle, WA 98109-1024, USA

Received on August 3, 2006; revised on October 24, 2006; accepted on November 21, 2006
Advance Access publication November 24, 2006
Associate Editor: Trey Ideker

ABSTRACT

Motivation: Gene Set Enrichment Analysis (GSEA) has been developed recently to capture changes in the expression of pre-defined sets of genes. We propose number of extensions to GSEA, including the use of different statistics to describe the association between genes and phenotypes of interest. We make use of dimension reduction procedures, such as principle component analysis, to identify gene sets with correlated expression. We also address issues that arise when gene sets overlap.

Results: Our proposals extend the range of applicability of GSEA and allow for adjustments based on other covariates. We have provided a well-defined procedure to address interpretation issues that can raise when gene sets have substantial overlap. We have shown how standard dimension reduction methods, such as PCA, can be used to help further interpret GSEA.

Contact: zjiang@fhcrc.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The classical approach to DNA microarray analysis has been to treat genes as independent agents, to apply some statistical test per gene and follow that up with some form of P-value correction method. Those genes whose adjusted P-values cross some predetermined threshold are deemed interesting and are followed up using other procedures. Such an approach can be criticized on a number of grounds. There is the arbitrariness of the cut-off, no matter how it is chosen, and in almost all experiments genes whose test statistics yield P-values that differ by a tiny amount are treated completely differently. By design this approach will find genes where the difference in mRNA abundance, between the conditions being studied, is large, but it will not detect a situation where the difference is small, but evidenced in a coordinated way in a set of related genes.

Gene Set Enrichment Analysis (GSEA) directly addresses these points. There is no need to use a cut-off. All genes assayed can be used in GSEA and only simple non-specific filtering, for variation across samples, is needed. GSEA aggregates the per gene statistics across genes within a gene set, thus making it possible to detect situations where all genes in a predefined set change in a small but coordinated way. Since it is likely that many relevant phenotypic differences are manifested by small but consistent changes in a set of genes GSEA is reasonable and seems likely to yield results. Furthermore, GSEA is likely to also detect the cases where the effect is due to large changes in a relatively few genes.

Examples of the efficacy of GSEA include Mootha *et al*. (2003) who used GSEA approach to identify PGC-1α-responsive genes involved in oxidative phosphorylation, and Majumder *et al*. (2004) who used the approach on prostate cancer to identify a seven member hypoxia-inducible factor 1 gene set.

In this paper, we consider GSEA from a slightly different perspective, develop the appropriate notation, and then provide a number of extensions of the methodology. These extensions include the use of linear models to adjust for other covariates, the use of a wide-variety of different statistics on each gene set, an explicit method to deconvolve the outputs when gene sets have substantial overlap, and the use of dimension reduction methods on gene sets that have been found to be interesting with respect to the likely coordinated activity of the genes involved.

2 MATERIALS AND METHODS

All methods are demonstrated on a large microarray dataset from a clinical trial in acute lymphoblastic leukemia (ALL) (Chiaretti *et al*., 2004). We will focus our attention on the patients with B-cell derived ALL, and in particular on comparing the group identified as having the BCR/ABL fusion gene (usually due to a t9;22 translocation) to those samples with no observed cytogenetics abnormalities, NEG. We make use of data from KEGG (Kanehisa and Goto, 2000) as our gene sets.

Our analysis procedures will aggregate information from different genes. Since expression values do not directly reflect the true mRNA abundance, we standardized the data, by gene, before applying GSEA.

2.1 Background

Subramanian *et al*. (2005) and Mootha *et al*. (2003) presented GSEA as a method to identify predefined gene sets that associate with the differences between phenotypes. They ranked all genes based on their association with the phenotype, then calculate an enrichment score for each gene set, and the maximal enrichment score (MES) is identified. The enrichment score combines the per gene associations with the phenotype and the distribution of the genes on the ranked list. A permutation technique was applied to generate the null distribution of the enrichment score. The P-value for the MES was then obtained with respect to the estimated null distribution.

Tian *et al*. (2005) and Kim and Volsky (2005) proposed a similar approach but instead of using the enrichment score, they used familiar two-sample statistics, such as the t-statistic. This approach can be viewed as an extension of GSEA that makes its application both simpler and richer. The test statistic for a gene set is an aggregate of the per gene test statistics of its members. They proposed using a permutation test to assess the significance of the statistics. As we note, there is also a parametric approximation that often works well.

*To whom correspondence should be addressed.

These two approaches follow a common idea of using combined information from individual genes, yet each approach has unique features. The main difference between the two methods is in the way they treat the genes that are not in the set. The approach of Subramanian *et al*. (2005) and Mootha *et al*. (2003) puts penalties on the non-member genes that are ranked between the genes in a gene set, especially when the member genes are clustered together, while the approach of Tian *et al*. (2005) ignores them. Our own approach is more similar to that of Tian *et al*. (2005).

We adopt some of the notation from Tian *et al*. (2005). Let i, j and k be the index of the genes, samples and gene sets, with $i = 1, \ldots, B$, $j = 1, \ldots, n$ and $k = 1, \ldots, K$, respectively. The association between the *i-th* gene and the phenotype is represented by z_i, and the association between the genes and the gene sets is presented in an incidence matrix A,

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1B} \\ \vdots & & & \vdots \\ a_{K1} & a_{K2} & \cdots & a_{KB} \end{pmatrix} \tag{2.1}$$

where

$$a_{ki} = \begin{cases} 0 & g_i \notin C_k \\ 1 & g_i \in C_k \end{cases}. \tag{2.2}$$

and C_k denotes the set of genes in the *k-th* gene set. There are situations where values other than 1, for a_{ki}, will be more appropriate. For example, one practical source of gene lists is other publications on the same disease or phenotype. Those papers often give a set of genes that are up-regulated and a second set that are down-regulated. Rather than treat these as two separate lists, all predictions can be accommodated by using a -1 in the corresponding elements of **A** for genes that are down regulated, a 1 for those that are up-regulated and a zero for genes that were not in the list. In other cases it may be more appropriate to use non-integer weights, perhaps based on some probability that a gene is differentially expressed, or the strength of evidence from the published paper. An example is given in Section 3.2.

The association between the genes and the phenotype is summarized in a vector **Z**,

$$\mathbf{Z} = (z_1, \ldots, z_B)^T, \tag{2.3}$$

where z_i is the observed test statistic for gene i. We denote gene sets as C_k and let n_k indicate the number of genes in C_k.

Tian *et al.* (2005) suggested using the average t-statistic of the members in a gene set as the statistic for that set. We generalize this definition. Let the vector of the gene set statistics be **X**, then **X** is the product of **A** and **Z** divided by the row sums of **A**, *rs(A)*,

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{Z}/rs(A) = f(\mathbf{A},\ \mathbf{Z}). \tag{2.4}$$

Using the approach of Tian *et al*. (2005), this definition has three components: the incidence matrix **A**, the per-gene statistic vector **Z** and a per-set summarization function *f*. Different choices for any one of them give variations of the method.

In Section 2.2 we propose several extensions including (1) for *f*: using the median or the sign test, (2) for **A**: adding signs or weights and (3) for **Z**: the use of a general linear model or a Bayesian approach. We also note that, because of the form of the aggregation, which is essentially the summation of estimated effects, it is important that those effects all be on essentially the same scale. This is one reason to use t-tests and when using other statistics care must by taken.

In Section 2.3, we introduce two new approaches that were not discussed by either Tian *et al*. (2005) or Subramanian *et al*. (2005). First we present a procedure to deal with overlap among gene sets and then we illustrate the application of dimension reduction methods.

2.1.1 Inference It is straightforward both to state and interpret a null hypothesis of no association between the observed phenotype and gene expression. This hypothesis can be tested in many different ways but for gene set enrichment it has been typical to permute the phenotype labels on the samples to generate a reference distribution. While some have proposed an approach that permutes the gene labels we do not advocate this since it is difficult to interpret the corresponding null hypothesis.

It is also possible to perform a parametric test of the hypothesis of no association. One advantage of using a t-statistic is provided by considering the following heuristic argument, first described to us by Dr T. P. Speed. Under the null hypothesis that there is no difference between the two groups being compared the t-statistics have a t distribution with degrees of freedom approximately $n - 2$ (the value depends on the form of the t-test used). If n is sufficiently large then these statistics have approximately a $N(0, 1)$ distribution, under the null hypothesis of no difference between the two groups. If the genes were independent then summing these over a gene set with n_k genes in it would yield a test statistic with a $N(0, n_k)$ distribution and dividing that statistic by the square root of n_k returns us to a $N(0, 1)$ distribution. Hence, the per set sums, divided by the square root of the gene set sizes can be compared to quantiles of the $N(0, 1)$. In practice this is both fast and reasonably reliable, but the assumption of independence of genes is not tenable. Here the null hypothesis is that there is no difference in the mean values of the two groups, and that is a different null hypothesis than that for a permutation approach.

2.2 Extensions

We now describe some extensions of the original concept of gene set enrichment. In some cases the extensions are quite simple, but even for these examples the results are compelling. In other cases the extensions are more substantial.

While most practitioners have used sums and averages to aggregate the test statistics per set, this is not the only approach that should be considered. We note that the average is not used universally in statistics as a means of measuring the center, largely because it is known to be susceptible to outliers. The median is another choice for a measure of the center. Other test statistics, such as the sign test can also be easily accommodated within the GSEA framework. The permutational method can be used to assess significance in these cases as well. We provide an example in Section 3.3.

Linear Modeling. The two sample t-statistic can also be obtained by fitting a linear model for each gene. We let

$$Y_{gi} = \mu_g + \beta_g X_i + \epsilon_{gi}, \tag{2.5}$$

where Y_{gi} is the vector of gene expression values for gene g and sample i, X_i is one or zero depending on the phenotype of sample i, and the ϵ_{gi} are assumed to be independent mean zero random variables with variance σ_g^2 (often assumed to have a Normal distribution). In this model μ_g represents the mean for the group with phenotype corresponding to $X_i = 0$, while β_g represents the difference in mean between that group and the group represented by $X_i = 1$. The t-statistic is equivalent to $\hat{\beta}_g/s_g$, where s_g is the natural estimate of σ_g. Adjusting for other variables, that are likely to affect expression values, can be handled using a more general regression equation, such as

$$Y_{gi} = \mu_g + \beta_{1g} X_{1i} + \beta_{2g} X_{2i} + \epsilon_{gi}, \tag{2.6}$$

where X_{2i} denotes the value of some additional covariate. The parameter β_{1g} then represents the mean difference in expression due to the phenotype after adjustment for X_2. We again make use of $\hat{\beta}_{1g}/s_{\beta_{1g}}$ as our standardized estimate of the phenotypic effect and these values can be used as **Z** in Equation (2.4).

The linear model is more flexible than the simple two-sample t-statistic. If the sample size is large enough, the linear model can be very complex, including many variables and interactions. Though we lose some degrees of freedom, including all appropriate variables in the linear model, it will provide more accurate estimates of the true effect due to the phenotype. It is important that the quantity being used as a test statistic have a distribution that is the same for all genes, unless there is some reason to prefer to work on a different scale. But typically, the observed values for gene expression data are not intrinsically meaningful and hence standardized estimates are preferred.

Posterior probability. We now provide a detailed discussion of an extension of the methodology to deal with a more complicated per-gene test statistic. We make use of the work of Newton *et al.* (2001) who developed a Bayesian approach to detect differentially expressed genes. This approach assumes a gene can come from one of the two groups, the equivalently expressed (EE) genes and the differentially expressed (DE) genes, with probabilities $1 - p$ and p, respectively. The gene expression from the two groups follows distributions $f_0(\cdot)$ and $f_1(\cdot)$, respectively. By Bayes' rule, the posterior probability of a gene with expression $\mathbf{e}$ to come from the DE group is

$$\frac{pf_1(\mathbf{e})}{pf_1(\mathbf{e}) + (1-p)f_0(\mathbf{e})}. \tag{2.7}$$

Using the posterior probability as per gene statistic, the gene set statistic $\mathbf{X}$ has a nice interpretation as the expected number of differentially expressed genes per set, and each component of $\mathbf{X}$ follows a Binomial distribution with parameters n_k and p_k, the later of which is unknown.

We are interested in finding gene sets having a strong association with a phenotype of interest. This association can be measured by the estimated number of DE genes in a gene set. But this number is related to the size of the gene set and we would naturally expect more DE genes in a larger set. We use the Binomial probability, p_k, which does not depend on the gene set size, to measure the association between a gene set and the phenotype. An interesting null hypothesis is that the probability of DE does not depend on gene set, which can be written as:

$$H_0: \quad p_1 = p_2 = \cdots = p_K = p, \tag{2.8}$$

where K is the number of gene sets. The alternative hypothesis is then there exists at least one gene set that is different from others:

$$H_a: \text{There exist at least one gene set}, k, \text{where } p_k \neq p. \tag{2.9}$$

Under the null hypothesis, we estimate the parameter p as

$$\hat{p} = \left(\sum_{g=1}^{N} \hat{z}_g\right)/N \quad \text{or,} \tag{2.10}$$

$$\hat{p} = \left(\sum_{k=1}^{m} \hat{p}_k\right)/K \quad \text{with} \quad \hat{p}_k = \left(\sum_{g \in C_k} \hat{z}_g\right)/|C_k|, \tag{2.11}$$

where $\hat{z}_g$ is the estimated posterior probability of gene g being differentially expressed. Equation 2.10 is the average of individual gene probabilities. It assumes that all the genes share the same probability of showing differential expression, which is a stricter null hypothesis than that of Equation 2.8. Equation 2.11 is the average of gene set probabilities, or it can be viewed as a weighted average of individual gene probabilities, where the weight for gene i is:

$$w_i = \left(\sum_{k: i \in C_k} \frac{1}{|C_k|}\right)/K. \tag{2.12}$$

Using this estimate, a gene has more weight if it belongs to a smaller gene set, or if it belongs to a larger number of gene sets.

Under the null hypothesis, the expected number of DE genes in the *k-th* gene set is:

$$\hat{n}_{e,k} = |C_k|\hat{p}, \tag{2.13}$$

the observed number of DE genes in the same gene set is:

$$n_{o,k} = \sum_{g \in C_k} \hat{z}_g, \tag{2.14}$$

and, the probability of observing $n_{o,k}$ or more DE genes is

$$\sum_{s=n_{o,k}}^{|C_k|} \binom{|C_k|}{s} \hat{p}^s (1-\hat{p})^{|C_k|-s}. \tag{2.15}$$

If $|C_k|$ is large enough, and $\hat{p}$ is not too small, the Binomial distribution can be approximated by a Normal distribution with parameters: $\mu_k = |C_k| \cdot \hat{p} = \hat{n}_{e,k}$ and $\sigma_k = \sqrt{|C_k| \cdot \hat{p}(1-\hat{p})}$. An approximate P-value can be obtained from:

$$\Phi\left(\frac{n_{o,k} - \hat{n}_{e,k}}{\sqrt{|C_k|\hat{p}(1-\hat{p})}}\right), \tag{2.16}$$

where Φ is the standard Normal distribution function.

One of the weaknesses of this approach is that the statistical algorithm detects differential expression without regard to direction. But if our goal is to detect coordinated changes in expression we should check to ensure that the estimated effects are in the same direction. For example, in a two sample comparison we would be interested in gene sets with many differentially expressed genes provided those samples from one phenotype, or condition, tended to have higher values than those from the other phenotype. So we propose that for each significant gene set, we check the change in the gene expression of each gene with posterior probability larger than the pre-selected cut-off. An example is shown in Section 3.2.

2.3 Interpreting the gene sets

The approach of computing a single test statistic per gene suggests a belief that all of the information that is contained in the gene set can be reduced first to a single number for each gene and then to a single number for all genes in the gene set. As this is not always the case, we discuss some extensions that can be used to help make more use of the available data.

We begin with the observation that there is often substantial overlap between different gene sets. For example, if we use pathways, as defined by KEGG (Kanehisa and Goto, 2000), we find that the Leukocyte transendothelial migration pathway and the Regulation of actin cytoskeleton pathway contain 50 and 78 genes, respectively and there are 23 in both. Suppose that in an experiment there is an activation of the Leukocyte transendothelial migration pathway, but not of the Regulation of actin cytoskeleton pathway, we might still observe an extreme statistic for the Regulation of actin cytoskeleton pathway merely due to the genes that are shared between them. If undetected such an observation may mislead an investigator. We discuss approaches that can be used to better attribute the observed effect to the appropriate gene set.

There are several statistical methods that can be used to determine whether genes within a gene set show coordinated expression. We suggest using visualization methods and dimension reduction techniques, such as principal component analysis (PCA), (Mardia *et al.*, 1979; Johnson and Wichern, 1988).

2.3.1 Overlap among gene sets Whenever two gene sets contain at least one common gene there is the potential for problems in interpretation. The most extreme case occurs when two, or more, gene sets are identical. In such a case we say that the gene sets are aliased and the practical implication is that one cannot determine, from the available data, which gene set is responsible for the effect. While complete aliasing is unlikely there are circumstances where it can occur and partial overlap between gene sets is common and can cause similar problems in interpretation. In particular, due to the structure of GO (The Gene Ontology Consorium, 2000) if GO classifications are used to define gene sets there will always be nesting.

Since genes can be in many gene sets, the level of overlap can be quite substantial with many gene sets being involved. We studied the extent of overlap for KEGG pathways of genes on the Affymetrix HGU-95Av2 chip. (Supplementary Table S1). Among 3012 genes that with KEGG pathway annotation, about half of them are involved in multiple pathways. In this report, we restrict our attention to pairwise comparisons of gene sets, but note that there can be higher level interactions.

When trying to assess whether two gene sets are aliased we must consider the gene set restricted to the data being analyzed rather than the whole gene set. Thus, even though two gene sets are not themselves aliased, if a number of genes have been excluded from the analysis then the gene sets, restricted to the genes being analyzed, can be aliased. It is not possible to determine from the available data which of the two (or more) gene sets is responsible for the observed effect.

In cases where two gene sets have common genes, we propose the following approaches. First identify all gene sets which have a significant effect. Of these, identify those which have substantial overlap in gene membership. This could be operationalized as either more than l genes in common, or using some other criteria. Then for each such pair, decompose the genes involved into three disjoint parts: the genes unique to the first gene set, the genes unique to the second gene set and the genes found in both gene sets. These three parts can also be viewed as gene sets and hence can be analyzed via GSEA. Correct attribution of the observed effect will depend on which of these three new gene sets have significant effects. To illustrate the different situations we present two examples in Section 3.7. In the first example, we find that only one of the gene sets seems to be implicated in the differences between the phenotypes, the other is significant only due to those genes shared with the first gene set. In the second example, both sets are implicated.

2.3.2 Dimension reduction per gene set We consider the problem from the perspective of the samples. For each gene set C_k there are $|C_k| = n_k$ genes whose expression values we want to model. We can consider each sample to be represented by a point in n_k dimensional space. If the genes in gene set C_k show coordinated patterns of expression then the points in the space should display a pattern that reflects this observation. Gene sets, which can be reduced to two or three dimensions indicate situations where the constituent genes are likely to be co-regulated.

PCA (Mardia *et al*., 1979; Johnson and Wichern, 1988) is one of the popular tools for dimension reduction. We use it as an example to show how dimension reduction can help us finding interesting gene sets. Genes will be standardized [the median subtracted and divided by the median absolute deviation (MAD)] prior to the application of PCA.

We followed two approaches using principle components (PCs). First, we found the number of PCs needed to explain a certain percentage, e.g. 70% of the variation among data. Second, we applied the isotropic test [Chapter 8.4, Mardia *et al*. (1979)], on the expression data. The isotropic test identifies a value k such that the null hypothesis: the last $n - k$ PCs are equally important, is rejected for $k - 1$ but not for k. Then the number k is the suggested number of PCs to keep. Gene sets generally have different sizes and this must be accounted for in the analysis.

3 APPLICATIONS

We will apply the methods discussed in the previous section to ALL data (Chiaretti *et al*., 2004). In the use of GSEA on publicly available data, there is a risk of circularity since the data may have been used to help define the gene sets in the first place. Analysts should exercise caution when working with historical data. Since our analysis uses KEGG we are reasonably sure such circularity is not an issue here.

3.1 Data processing

Before applying the methods discussed in the previous section to the ALL data, it must be processed and filtered to some extent. We describe our choices but emphasize that users can substitute methods they prefer. We make use of these as we have found that they often provide a sound basis for analyses.

There are 37 samples for the BCR/ABL group and 42 samples for the NEG group. We first filtered the probes base on their expression variation. The probes with very little or no variation (IQR < 0.5) were filtered out, leaving 4149 probes. In some cases multiple probes map to a single gene, we retained the one with the largest t-test statistic between phenotype. Our reason for this approach is that we are looking for the best evidence we can find of gene set involvement. Since not all genes are accurately annotated, or arrayed, it seems reasonable to use the microarray probe for a gene with the best evidence for differences in phenotype. After this step, we were left with 3443 genes/probes. Among them, there are 1144 genes are annotated as members of one or more KEGG pathways.

Another practical issue that we need to deal with is the size of a gene set, or what might be termed the effective size of a gene set. This is a parameter and must be chosen by the user. In some cases it will be of interest to retain relatively small gene sets, but in most cases one will be interested in general descriptions and therefore larger gene sets are more helpful. We do emphasize that this size is not the size of the gene set that has been curated, but rather the size of the gene set when restricted to the genes that are going to be used in the analysis. For the analyses reported here we keep only the pathways with at least 10 genes. In the end, we have 1031 genes, 79 samples and 77 pathways.

3.2 Simple analyses

We first used the two-sample t-statistic as the per-gene statistic ($\mathbf{Z}$), the mean as the aggregation function (f) and permutation on the sample labels to obtain the significance of the gene sets. In the following sections, we illustrate the extensions described in Section 2.2. We used a permutation test with 5000 permutations and obtained the P-values for each pathway. There were 14 pathways with a P-value <0.01. They are reported in Table 1. These pathways have higher gene expression levels in BCR/ABL versus NEG at the significant level of 0.01.

3.3 Median/sign-test as f

In this section, we used different summaries of the evidence for each gene set. In one analysis we used the median of the t-statistics and in a second analysis we used the sign test on the observed per gene t-statistics for each gene set and compared with the results from using the mean. Table 1 contains the GSEA results for the ALL data using different methods to aggregate the single gene information into gene set information, the columns p^{Mn}, p^{Md} and p^{ST} show the P-values using the mean, the median or sign test to compute the gene set statistic. The rows are divided into six sections; the pathways found by all, two of the three, or only one of the three methods. If a pathway is reported significant by one test, the P-value for that test is listed in the table, otherwise the corresponding element is blank.

The majority of findings from using the mean or the median are the same, except for four pathways that are found by the mean but not by the median and 1 pathway found by the median but not by the mean. For example, Supplementary Figure S1(b) shows the t-statistics for genes in the mTOR signaling pathway. This pathway is reported significant using the mean but not using the median. The t-statistic for the gene PRKAA1 (shown as a black triangle) is much higher than all the others, suggesting that the median test is more reliable in this case.

Table 1. Significant pathways reported by different statistics

	ID	PW name	p^{Mn}	p^{Md}	p^{ST}	Size
1	04514	Cell adhesio...	0.0000	0.0004	0.0011	39
2	04940	Type I diabe...	0.0040	0.0052	0.0015	21
3	04610	Complement a...	0.0000	0.0008		14
4	04512	ECM-receptor...	0.0000	0.0008		15
5	04530	Tight juncti...	0.0000	0.0040		40
6	04080	Neuroactive...	0.0000	0.0044		21
7	04520	Adherens jun...	0.0000	0.0068		34
8	04510	Focal adhesi...	0.0004	0.0024		68
9	04670	Leukocyte tr...	0.0020	0.0032		50
10	01430	Cell Communi...	0.0028	0.0008		12
11	04060	Cytokine-cyt...	0.0060		0.0002	54
12	04360	Axon guidanc...	0.0008			38
13	05130	Pathogenic E...	0.0080			27
14	05131	Pathogenic E...	0.0080			27
15	04640	Hematopoieti...		0.0000		39
16	03010	Ribosome			0.0000	22
17	00620	Pyruvate met...			0.0005	16
18	00190	Oxidative ph...			0.0006	59
19	00230	Purine metab...			0.0075	57
20	04110	Cell cycle			0.0092	66

The columns p^{Mn}, p^{Md}, and p^{ST} show the *P*-values using the mean, the median or the sign-test, respectively, to compute the gene set statistic from per gene associations with phenotypic differences. The rows are divided into six sections. In each section are the pathways reported by all, two of the three, or only one of the three methods. If a pathway is reported significant by a method, the *P*-value is listed in the table. Otherwise the corresponding element is blank.

The results using the sign test as *f* are quite different from using the mean or the median. We think this is because the mean and median are using the actual values of *t*-statistics whereas the sign test is using logical values, where all the genes with higher *t*-statistics are treated the same whether they are higher by a small amount or a large amount.

When the mean is used, our argument, Section 2.1.1, suggests a qq-plot be used to graphically identify significant gene sets. We generate the qq-plot for our data in Supplementary Figure S2(a). The pathway statistics are quite close to the 45 degree line. We identify 3 pathways that are further away from the 45 degree line than others. They are Fatty acid metabolism (00071), Focal adhesion (04510) and Cell adhesion molecules (CAMs) (04514) in Table 1.

3.4 Linear modeling

For the ALL data, we fitted the model in Equation (2.6), with X_{2i} being the sex of the individual. We use the *t*-statistic $\hat{\beta}_{1g}/SE(\hat{\beta}_{1g})$ as gene statistic in GSEA. Table 2 reports all pathways significant at 0.01 level. The column $p^{\mathrm{lm.t}}$ lists the *P*-value obtained through linear modeling and the column p^{t} is the same as the p^{Mn} column in Table 1. The *t*-statistic adjusted for gender identified pathways, besides the ones that are reported by the unadjusted *t*-statistic, suggesting that there may be important gender differences.

The qq-plots for the unadjusted *t*-statistic and the *t*-statistic adjusted for gender of the pathways (Supplementary Figure S2) both identified the following pathways: CAMs pathway, the Adherens junction pathway, and the Lysine degradation pathway.

Table 2. Significant pathways and *P*-values reported using the adjusted *t*-statistic, $p^{\mathrm{lm.t}}$, and the un-adjusted *t*-statistic, p^{t}

	ID	PW name	$P^{\mathrm{lm.t}}$	P^{t}	Size
1	04510	Focal adhesi...	0.0000	0.0004	68
2	04512	ECM-receptor...	0.0000	0.0000	15
3	04514	Cell adhesio...	0.0000	0.0000	39
4	04670	Leukocyte tr...	0.0000	0.0020	50
5	04530	Tight juncti...	0.0000	0.0000	40
6	04360	Axon guidanc...	0.0000	0.0008	38
7	04610	Complement a...	0.0000	0.0000	14
8	04060	Cytokine-cyt...	0.0000	0.0060	54
9	04080	Neuroactive...	0.0000	0.0000	21
10	04520	Adherens jun...	0.0020	0.0000	34
11	01430	Cell Communi...	0.0040	0.0028	12
12	04940	Type I diabe...	0.0060	0.0040	21

The differences suggest that the gender has influence on the gene expression profile.

Table 3. Significant pathways reported using posterior probability of DE as the per gene statistic

	ID	PW name	p^1	p^2	Size	B↑	B↓
1	04060	Cytokine-cyt...	0.0091	0.0028	54	25	8
2	04520	Adherens jun...		0.0057	34	15	7

The columns p^1 and p^2 are the *P*-values by using the average over all gene probabilities or using the average over all gene set probabilities as null hypothesis parameter. The columns B↓ and B↑ show the number of genes that have higher or lower in BCR/ABL, respectively, among the genes with posterior probability at least 0.01.

3.5 Posterior probability as gene statistic

For each gene, we estimated the probability of being differentially expressed using the EBarrays package (v1.3.0). Then we calculated the expected number of DE genes and the observed number of DE genes as in Equations (2.13) and (2.14). To get *P*-values for the pathways, we used two different methods. We estimated $\hat{p}$ by averaging over all genes [Equation (2.10)] and alternatively by averaging over all gene sets [Equation (2.11)]. Table 3 lists the results from the two methods. The columns p^1 and p^2 are the *P*-values by using the average over all genes or using the average over all gene set as the null hypothesis parameter. The columns B↓ and B↑ show the number of genes that have higher or lower expression in the BCR/ABL phenotype, respectively, among the genes with posterior probability at least 0.01 of that set. The cytokine–cytokine receptor interaction pathway is found by both approaches and the Adherens junction pathway is found only by the second approach.

We checked the direction of expression changes from BCR/ABL to NEG for the genes with posterior probability >0.01 in these pathways. In all cases more than two-thirds of the genes have higher expression in BCR/ABL phenotype. The Adherens junction pathway has about two-thirds of interesting genes showed higher expression in BCR/ABL phenotype. The cytokine–cytokine receptor interaction pathway and the Axon guidance pathway have about three quarters of interesting genes showing higher expression in BCR/ABL phenotype.

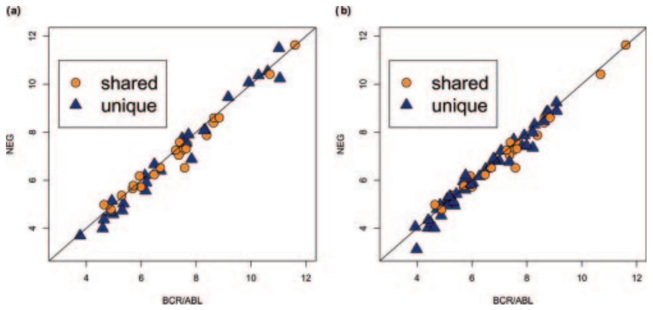

Fig. 1. Mean plots for **(a)** the Leukocyte transendothelial migration pathway. **(b)** the Focal adhesion pathway.

3.6 Incidence matrix

We make use of the analysis reported in Yeoh *et al*. (2002). Although their study was on pediatric patients, the type of cancer, ALL, was the same. We obtained a gene list from Yeoh *et al*. (2002) that was used to classify the BCR/ABL ALL subtype from other ALL subtypes by *t*-statistic (Supplementary Table 13 of Yeoh *et al*. (2002) at http://www.stjuderesearch.org/data/ALL1).

Yeoh *et al*. (2002) used the same gene chip as Chiaretti *et al*. (2004). Among the 40 genes they reported, 30 are higher in BCR/ABL and 10 are lower in BCR/ABL. After filtering genes for variance (Section 3.1), we were left with 10 genes from their list, 9 with higher values in BCR/ABL and 1 with a lower value. We put these genes in a gene set and used 1 for the up-regulated genes and -1 for the down-regulated genes. The resulting *P*-value was $<10^{-4}$, indicating a very strong concordance between the data from Chiaretti *et al*. (2004) and that of Yeoh *et al*. (2002).

3.7 Aliasing

Pathways have a substantial overlap (Supplementary Table S1). In this section, we describe an approach for dealing with aliasing and partially overlapping gene sets to interpret identified gene sets.

We consider the two pathways, the Leukocyte transendothelial migration pathway and the Focal adhesion pathway. Both pathways were found significant by *t*-statistic (Table 1). This pair of pathways has 23 genes in common.

In Figure 1, we present a graphical display of the two pathways. Each point in the graph represents a gene and its *x*- and *y*-values are the mean expression for that gene over all samples in the BCR/ABL and the NEG groups, respectively. The light colored dots are the genes in both pathways and the dark colored triangles are the genes unique in each pathway. Points that are above the 45-degree line have higher expression values in the NEG group while those that are below the 45-degree line have larger values in the BCR/ABL group.

Table 4. Results for the subsets in pathway pair of leukocyte transendothelial migration and focal adhesion

	Name	Size	Test statistic	*P*-value
1	Common	23	22.306	0.0030
2	Leukocyte tr. . .	27	20.593	0.0128
3	Focal adhesi. . .	45	33.422	0.0016

Table 5. Results for the subsets in pathway pair of the CAMs pathway and the type I diabetes mellitus pathway

	Name	Size	Test statistic	*P*-value
1	Common	14	19.602	0.0126
2	Cell adhesio. . .	25	30.520	0.0000
3	Type I diabe. . .	7	9.280	0.0048

We would like to make a few observations based on the content of these figures before proceeding with the discussion. First, those genes that are found in both pathways (colored orange) tend to have larger values in the BCR/ABL group and hence are mainly found below the 45-degree line. Those genes found only in the Focal adhesion pathway also tend to be below the 45-degree line, while those genes found only in the Leukocyte transendothelial migration pathway tend to be scattered above and below the line, with no apparent preference. Since GSEA detects the accumulated effect of genes within a gene set we suspect that the sub-group of genes unique to Leukocyte transendothelial migration will not be significant since the observed effects seem to cancel each other out.

We divided the genes in these two pathways into three parts, one for the genes unique to each pathway and one for the shared genes and then carried out GSEA on the three parts. The analysis was based on the permutation of sample labels, and the test results are summarized in Table 4. Genes unique to the Leukocyte transendothelial migration pathway exhibit a significant effect, as do those that are shared and most importantly the direction of the effect in both groups is the same. But for genes unique to the Focal adhesion pathway there seems to be no effect. This observation strongly suggests that the effect observed is due to the Focal adhesion pathway activation and not to the Leukocyte transendothelial migration pathway activation.

Next, we compare the CAMs pathway and the Type I diabetes mellitus pathway with 39 and 21 genes, respectively. There are 14 genes common to both pathways. We follow the same procedure described above and split the genes into three gene sets.

We generated the mean plots of the genes in these two pathways in the Supplementary Figure S2. Unlike our previous sample, we do not see any obvious patterns in these two plots. The permutation test results in Table 5 indicate that those genes that are common to the two pathways are not significant at the 0.01 level. The gene sets based on genes unique to each of the two pathway remain significant.

There is some rationale for believing this to be a more common situation. Genes which are shared among different pathways are likely to be regulated differently than those that are unique to a pathway. Genes that play a number of different roles will need to be expressed and translated when any of their associated functions are required, and hence are likely to be regulated by other mechanisms.

As illustrated in these two examples, modeling the genes shared between two pathways can improve our understanding and interpretation of the test results. In our first example we believe that the data are consistent with an activation or up-regulation of the Focal adhesion pathway in patients with BCR/ABL and that there is little evidence of the Leukocyte transendothelial migration pathway

Table 6. Pathways for which four or fewer PC's explain at least 70% of the variability

ID	PW name	Size	k	Ratio	PC1	PC2	PC3	k^Y	Ratio^Y	PC1^Y	PC2^Y	PC3^Y
03010	Ribosome	22	2	0.091	0.479	0.316	0.040	3	0.136	0.604	0.077	0.060
00251	Glutamate me...	11	4	0.364	0.367	0.200	0.123	4	0.364	0.367	0.200	0.123
00790	Folate biosy...	11	4	0.364	0.292	0.232	0.124	4	0.364	0.292	0.232	0.124
00071	Fatty acid m...	14	4	0.286	0.458	0.138	0.103	4	0.286	0.458	0.138	0.103
05040	Huntington's...	16	4	0.250	0.436	0.155	0.089	4	0.250	0.436	0.155	0.089
03320	PPAR signali...	11	4	0.364	0.464	0.143	0.092	4	0.364	0.464	0.143	0.092
00710	Carbon fixat...*	11	4	0.364	0.315	0.211	0.122	4	0.364	0.315	0.211	0.122
01031	Glycan struc...*	11	4	0.364	0.267	0.183	0.158	4	0.364	0.267	0.183	0.158
00350	Tyrosine met...*	14	4	0.286	0.326	0.197	0.115	4	0.286	0.326	0.197	0.115
00564	Glycerophosp...*	14	4	0.286	0.327	0.176	0.121	4	0.286	0.327	0.176	0.121
00051	Fructose and...*	15	4	0.267	0.309	0.172	0.137	4	0.267	0.309	0.172	0.137

Pathways are sorted by the number of principle components (k). Ratio is k/size, and PCi is the proportion of the variance explained by the i-th PC. The columns with superscript 'Y' are the PCA analysis results after removing th genes on the Y chromosome. Those labeled with a * are significant by permutation test.

involvement. In the second example, the genes in both pathways are likely to have higher expresion in patients with BCR/ABL.

3.8 PCA per gene set

We applied PCA to the gene expression values for each pathway seperately. The expression values were standardized by subtracting the median and dividing by the MAD.

Following the approach mentioned in Section 2.3.2, we obtained the number of PCs, k, needed to explain 70% of the variation and separately, the number of PCs that identified by the isotropic test. Table 6 reports the gene sets with $k \leq 4$. The column 'Ratio' gives the ratio of number k and the gene set size. The column 'PC1', 'PC2' and 'PC3' contain the proportion of variation for the first three PCs.

The number of PCs identified by the isotropic test tended to be quite large for all pathways. The reason could be that isotropic test is testing whether the last $n - k$ PCs are of the same importance, where n is the number of samples, and k is the number of PCs that were kept. In our data, it seems that although the last $n - k$ PCs are not important, they still cannot be considered equally important. For example, in Supplementary Figure S4 we plot the variation of the PCs for the Ribosome pathway. Except the first component, all the other components are not very important. But the isotropic test suggested keeping 13 PCs.

The Ribosome pathway appears to be very interesting. The first two components explain almost 80% of total variation of the gene set. The other components explain much lower percentage of the variation and the ratio is also very low. We know that some genes in the Ribosome pathway are sex related (e.g. SLC25A6 is on the Y chromosome). Figure 2a is the boxplot of the loadings for the first three PCs. The first PC is dominated by one gene. The biplot of the first two PCs (Fig. 2b) show that the first PC mostly explains the gender differences among the subjects. Also from this plot, we find four subjects with gender annotation mistakes. Two of them are recorded as females but the data indicates are male and two are recorded as males while the data indicates they are females. We made corresponding corrections.

In the effort to eliminate the gender effects on the gene expression, we removed all the genes on the Y chromosome and repeated the PCA analysis on the remaining genes. The results are summarized in Table 6. Most results are the same except for the Ribosome pathway. The plots of the new PCs for the Ribosome pathway are shown in the Supplementary Figure S3. The first PC is no longer dominated by one gene. It is quite closely related to the second PC obtained before removing the Y genes. The difference is a small gender effect. Removing the genes on the Y chromosome is not enough to eliminate the gender difference in data, but it now is not the most important source of variation in the data.

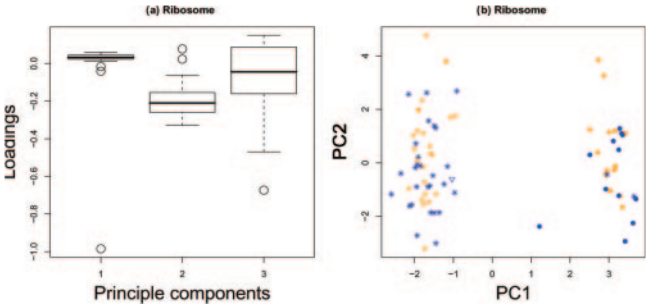

Fig. 2. The PCA results for the Ribosome pathway. **(a)** Boxplots of the loadings for the first three components of Ribosome pathway. The first component is dominatd by one gene, ribosomal protein S4, Y-linked 1, which is a Y chromosome gene. **(b)** Biplot of the first two PCs of the Ribosome pathway. The points in orange and blue represent samples with BCR/ABL and NEG phenotypes, respectively. The star and bullet symbols represents male and female, respectively. There is one sample with missing sex annotation, represented by a triangle. (We predict it to be a sample from a male subject.)

Another way to reduce the variation from gender differences is to adopt the ideas in Section 3.2 and fit a linear model of gene expression on gender. The residuals from this model should be free of gender differences and the PCA techniques can be applied to the residuals.

A permutation test could also be applied. In this case, we permute the labels of the genes in the data. We realize that this is contrary to our advice in Section 2.1.1, but for this analysis permutation of the sample labels has no effect, and the only way in which to generate a permutational distribution is to permute the labels on the genes. For each permutation we performed PCA on the new gene sets to estimate the null distribution of k, the number of PCs needed to explain

70% of the variation in the gene set. We used 1000 permutations and obtained 54 pathways with P-values <0.01, including 5 out of the 11 pathways in Table 6 (marked with star).

4 DISCUSSION

GSEA, as presented in Subramanian *et al*. (2005) and Tian *et al*. (2005), provides a valuable and useful tool for the analysis of genomic data. In this report we have discussed a number of extensions of the original proposal.

The extensions include the use of linear modeling and posterior probabilities. The t-statistic and posterior probability both measure the strength of the association between gene expression and phenotype. The advantages of the t-statistic are that it contains the information of the direction in which the gene expression has changed, and since the approximate distribution is known, QQ-plots can be used to visually inspect the distribution, which seems reasonably reliable. The posterior probability approach has a nice interpretation but it ignores the direction in which the gene expression altered. The choice between t-statistic and Baysian posterior probability depends on researcher's belief about the underlying biology. If the researcher does not care about the direction of change, the Bayesian approach can be a very good choice. But if the direction is important, the t-statistic is a better choice.

The extensions of gene set aggregation functions include the use of the median and the sign-test. Compared to the mean, the median is a more robust measure of the center. The sign-test is rather different than the other two and favors direction over magnitude.

In addition we have shown how to address issues of aliasing, where two or more gene sets overlap. In our experience this is not merely an academic exercise, almost all experiments we have analyzed suffer from some form of aliasing. We have specifically addressed pair-wise overlap, mainly because it is directly interpretable, higher order interactions and overlap are both harder to model, and to interpret. Investigators should always be aware of this problem. After having the list of interesting gene sets, investigators should always check for overlap. If there are sets with significant amounts of overlap, the investigators should follow the procedure provided here to better interpret their data.

Finally, we have considered a simple method of examining the amount of collinearity among the gene sets using PCA. Again, in our examples, the application of these methods was very fruitful. It helped to identify some potential underlying problems and to identify gene sets where there appears to be coordinated behavior of the constituent genes.

We also remark that while GSEA approach has largely been applied to microarrays, there is nothing special about microarray data and could just as easily be applied to any other high-throughput data streams where the variables can be grouped in relevant ways a priori.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article were provided by FHCRC Development - Computational Biology (R. Gentleman).

Conflict of Interest: none declared.

REFERENCES

Chiaretti,S. *et al*. (2004) Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, **103**, 2771–2778.

Johnson,R.A. and Wichern,D.W. (1988) *Applied Multivariate*. Prentice Hall, New Jersey.

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

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

Majumder,P.K. *et al*. (2004) mTOR inhibition reverses akt-dependent prostate intraepithelial neoplasia through regulation of apoptotic and HIF-1-dependent pathways. *Nat. Med*., **10**, 594–601.

Mardia,K., Kent,J. and Bibby,J. (1979) *Multivariate Analysis*. Academic Press.

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

Newton,M. *et al*. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol*., **8**, 37–52.

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

The Gene Ontology Consortium. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

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

Yeoh,E.-J. *et al*. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.