Bioconductor Project
Bioconductor Project Working Papers

Year 2006 *Paper* 12

Extensions to Gene Set Enrichment

Zhen Jiang* Robert Gentleman†

*Fred Hutchinson Cancer Research Center, zjiang@fhcrc.org

†rgentlem@fhcrc.org

Extensions to Gene Set Enrichment

Zhen Jiang and Robert Gentleman

Abstract

Motivation: Gene Set Enrichment Analysis (GSEA) has been developed recently to capture moderate but coordinated changes in the expression of sets of functionally related genes. We propose number of extensions to GSEA, which uses different statistics to describe the association between genes and phenotype of interest. We make use of dimension reduction procedures, such as principle component analysis to identify gene sets containing coordinated genes. We also address the problem of overlapping among gene sets in this paper.

Results: We applied our methods to the data come from a clinical trial in acute lymphoblastic leukemia (ALL) [1]. We identified interesting gene sets using different statistics. We find that gender may have effects on the gene expression in addition to the phenotype effects. Investigating overlap among interesting gene sets indicate that overlapping could alter the interpretation of the significant results.

Extensions to Gene Set Enrichment

Zhen Jiang, Robert Gentleman

August 21, 2006

1 Introduction

Gene set enrichment analysis (GSEA) is one of the more interesting tools to have been developed for the analysis of microarray data. In this paper we first consider the approach from a slightly different perspective, develop the appropriate notation and then provide a number of extensions of the methodology. These extensions cover a number of different important areas of application and show how one can make use of a wide variety of different statistics on each gene set, how to deconvolve the outputs when gene sets have substantial overlap, and how to inspect gene sets that have been found to be interesting with respect to the likely coordinated activity of the genes that have been identified.

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 p-values cross some predetermined threshold are deemed interesting and are followed up by a number of 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.

GSEA was designed to 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 such analyses 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 suggest some simple modifications to the approach, such as using different per gene set statistics. We also discuss the use of a more general linear model approach that can be used to adjust for other variables that are known to affect expression values, but which are often not of direct interest. We consider the interpretation of gene sets and show how there is potential for misinterpretation when subsets of genes are shared and finally we present some results regarding dimension reduction techniques applied on a per gene set basis.

2 Materials and Methods

All methods are demonstrated on a large microarray data set that come from a clinical trial in acute lymphoblastic leukemia (ALL). 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 analyzing.

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. First, they ranked all genes based on their association with the phenotype of interest differences. Then, they assumed that if the genes in a gene set have coordinated changes across phenotype, the distribution of the positions of these genes on the ranked list will show non-randomness. They developed an enrichment score to measure the non-randomness. This enrichment score combines the per gene association with the phenotype and the distribution of the genes on the ranked list. A permutation test was used to access the significance of the enrichment scores (more discussion in Section 2.1.1). As a result, gene sets will be deemed significant if most of its members have moderate association with the phenotype and are clustered within the ranked gene list.

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. A permutation test is also used to assess the significance of the statistics. As we note, there is a parametric approximation that often works well.

These two approaches follow a common idea of using combined information from individual genes, yet each of them 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], Tian et al. [2005] 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 ith 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 kth 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 $\mathbf{A}$ 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.

The association between the genes and the phenotype is summarized in a vector $\mathbf{Z}$,

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

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.

We now modify the gene set statistics a little from the definition in Tian et al. [2005]. Instead of using the average, we use the summation of the gene statistics as the per set statistic. Then, the vector of the gene set statistics, $\mathbf{X}$, is the inner product of the incidence matrix, $\mathbf{A}$, and the vector of the gene statistics, $\mathbf{Z}$.

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{Z} \tag{2.4}$$

The gene set statistic in Equation 2.4 has a general form. It is composed of two parts: the association/membership between genes and the gene sets, $\mathbf{A}$, and the association between genes and phenotype, $\mathbf{Z}$. Different realizations of them can give many variations of the method. Tian et al. [2005] proposed t-statistics, but any other test statistic could be used for $\mathbf{Z}$. In Section 2.2 we propose several modifications, including using different univariate test statistics, a more general linear models approach and a Bayesian approach.

We also note that because of the form of the aggregation, 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 in other cases care must by taken.

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. All permutation results reported here arise from permuting the group or phenotype labels and comparing the observed value of the test statistic to the empirical distribution of test statistics obtained from the permutations.

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 slightly 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.

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. Other per set summarizations, such as the median, or a sign test can 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.

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 for gene g and sample i and X_i is one or zero depending on the phenotype of sample i and 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 can be obtained by testing the model parameter $\beta_g = 0$ and is equivalent to $\widehat{\beta}_g/s_g$, where s_g is natural estimate of σ_g^2. An natural extension is to adjust for other variables that are likely to affect expression values,

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

where X_{2i} denotes the value of the additional covariate in the model. The parameter β_{1g} then represents the mean difference in expression due to phenotype after adjustment for X_2. We again make use of $\widehat{\beta}_{1g}/s_{\beta_{1g}}$ as our standardized estimate of the phenotypic effect and these values are used as $\mathbf{Z}$ 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 could be very complex, including many variables and even high level interactions. Though we lose some degrees of freedom, including all appropriate variables in the linear model 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.

2.2.1 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 the Bayes' rule, the posterior probability of a gene with expression $\mathbf{x}$ to come from the DE group is

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

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 stronger associations with a phenotype of interest. The null hypothesis is that all the gene sets have the same association. 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 under the null hypothesis. Letting p_k denote the binomial probability parameter for gene set k, the null hypothesis can be written as:

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

where K is the number of gene sets. The alternative hypothesis is one-sided, since we are looking for gene sets with stronger association.

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

Under the null hypothesis, we estimate the parameter p as

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

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

where $\widehat{z}_g$ 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 of gene i is

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

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

Under the null hypothesis, the expected number of DE genes in the kth gene set is

$$\widehat{eDE}_k = |C_k| \widehat{p} \tag{2.13}$$

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

$$oDE_k = \sum_{g \in C_k} \widehat{z}_g \tag{2.14}$$

and, the probability of observing oDE_k or more DE genes is

$$\sum_{s=oDE_k}^{|C_k|} \binom{|C_k|}{s} \widehat{p}^s (1-\widehat{p})^{|C_k|-s} \tag{2.15}$$

If $|C_k|$ is large enough and $\widehat{p}$ is not too small, the binomial distribution can be approximated by a Normal distribution with parameters $\mu_k = |C_k| \cdot \widehat{p} = \widehat{eDE}$ and $\sigma_k = \sqrt{|C_k| \cdot \widehat{p}(1-\widehat{p})}$. The approximate p-value is

$$\Phi \left(\frac{oDE_k - \widehat{eDE}}{\sqrt{|C_k| \widehat{p}(1-\widehat{p})}} \right) \tag{2.16}$$

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.

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 49 and 79 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 Shared genes and aliasing 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 Consortium, 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. (Table 1 of supplimentary material). Among 2989 genes that with KEGG pathway annotation, about half of them involve 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.

Table 1: The frequency table for the number of KEGG pathways the genes on the HGU-95av2 gene chip belong to.

	Sets	Genes	Sets	Genes	Sets	Genes	Sets	Genes
1	1	1541	5	95	9	18	13	3
2	2	687	6	46	10	9	14	9
3	3	349	7	31	11	5	15	11
4	4	159	8	10	12	14	18	2

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. The effect is essentially the same. 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 use the following methods. For each pair of gene sets we can 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. To illustrate the different situations we present two examples. In the first example, Section 3.4.1, 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, Section 3.4.2, there seems to be no interpretation issue.

2.3.2 Dimension reduction per gene set

We consider the problem from the perspective of the samples. For each of the n samples and 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.

Principle component analysis (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 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 of the variation among data. For example, we chose 70%. 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. The gene sets generally have different sizes and this must be accounted for. We then checked the ratio between the number and the gene set size. The smaller this ratio is, the better the reduction is.

3 Applications

We will apply the methods discussed in the previous section to ALL data.

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 virtually any other 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 3446 genes/probes. Among them, there are 1138 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 our analyses we keep only the pathways with at least 10 genes. In the end, we have 1036 genes, 79 samples, and 76 pathways.

3.2 Simple Analyses

Computing the two group t-statistic of the gene expression for each gene. Permutation is used to access the significance of a pathway, using sum or mean are equivalent. We computed the mean of the t-statistic of the genes in each pathway. Using a permutation test with 5000 permutations, we obtained the p-values for each pathway. There were 14 significant pathways with a one-sided p-value less than 0.01. They are reported in Table 2 in supplimentary material. These pathways have higher gene expression levels in `BCR/ABL` versus `NEG` at the significant level of 0.01.

3.3 Median and sign test

We looked at using median and sign test of per gene statistics within a gene set as a per set statistic and compared with the results from using mean. Table 2 and Figure 1(a) show the comparison. The majority of findings for the mean and median are the same, except that 3 pathways are found by mean but not by median and 2 pathways are found by median but not by mean. The pathways reported differently by mean and median may suggest influential genes in these pathways.

For example, Figure 1(b) shows the t-test statistics for genes in the *mTOR signaling pathway*. This pathway is significant using mean but not using 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 2: Significant pathways reported by different statistics. The columns $\mathtt{p}^{Mn}$, $\mathtt{p}^{Md}$, and $\mathtt{p}^{ST}$ show the p-values using mean, median or 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.

	ID	PW Name	p^{Mn}	p^{Md}	p^{ST}	Size
1	04514	Cell adhesio...	0.0000	0.0000	0.0008	38
2	04940	Type I diabe...	0.0018	0.0020	0.0013	20
3	04060	Cytokine-cyt...	0.0030	0.0050	0.0001	54
4	04610	Complement a...	0.0000	0.0004		14
5	04512	ECM-receptor...	0.0000	0.0004		15
6	04530	Tight juncti...	0.0000	0.0020		40
7	04520	Adherens jun...	0.0000	0.0034		34
8	04670	Leukocyte tr...	0.0002	0.0010		49
9	04080	Neuroactive ...	0.0002	0.0012		20
10	04510	Focal adhesi...	0.0006	0.0028		73
11	01430	Cell Communi...	0.0014	0.0004		12
12	03010	Ribosome		0.0080	0.0000	23
13	04360	Axon guidanc...	0.0004			38
14	04810	Regulation o...	0.0066			79
15	04210	Apoptosis	0.0096			46
16	04640	Hematopoieti...		0.0008		38
17	00190	Oxidative ph...			0.0001	59
18	00620	Pyruvate met...			0.0003	16
19	00230	Purine metab...			0.0027	58
20	04110	Cell cycle			0.0046	66
21	00310	Lysine degra...			0.0065	14
22	00071	Fatty acid m...			0.0065	14
23	00010	Glycolysis /...			0.0085	22

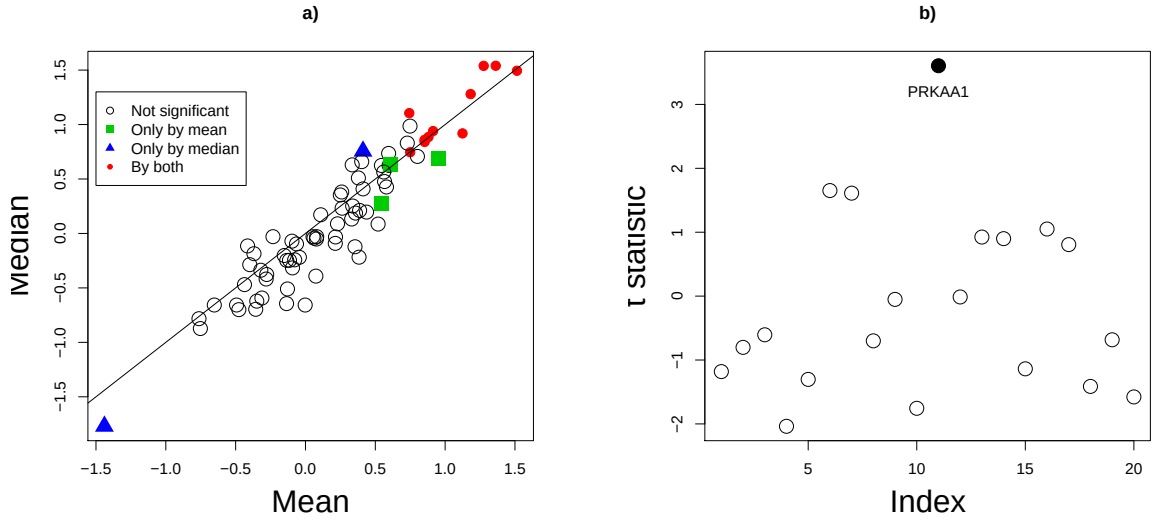

Figure 1: (a) Comparison of using mean or median as gene set statistic. (b) The t-statistic of genes in the *mTOR signaling pathway*. Most genes have t-statistic between -2 and 2, except one gene, PRKAA1 has t-statistic greater than 3.5 (shown in black).

The results from the sign test are much different from those of mean and median. We think this is because the mean and median are using the actual value of t-statistics whereas the sign test is using logical value. All the genes with higher t-statistics are treated the same whether they are higher by a small amount or a large amount.

If the mean is used, then our previous argument (Section 2.1.1) says a qq-plot can be used to graphically identify significant gene sets. We generate the qq-plot of our data in Figure 2(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.

3.3.1 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 3 reports all pathways significant at 0.01 level. The t-statistic adjusted for gender identified more significant pathways besides the ones that are reported by the un-adjusted t-statistic, suggesting that there may be important gender differences.

The qq-plots for the un-adjusted t-statistic and the t-statistic adjusted for gender of the pathways (Figure 2) identified the same pathways, such as, the *Cell adhesion molecules (CAMs)* pathway, the *Adherens junction* pathway, and the *Lysine degradation* pathway.

3.3.2 Posterior probability as gene statistic

For each gene, we estimated the probability of being differentially expressed using the `EBarrays` package. Then we calculated the expected number of DE genes and the observed number of DE genes as in equation (2.13) and (2.14). To get p-values for the pathways, we used two different methods: estimate $\widehat{p}$ by averaging over all genes (Equation 2.10) and estimate $\widehat{p}$ by averaging over all gene sets (Equation 2.11). Table 4 lists the results from the two methods. The *Cytokine-cytokine receptor interaction* pathway is found by both approaches and the *Adherens junction* pathway and the *Axon guidance* pathway are found only by the

Table 3: Significant pathways and p-values reported using the adjusted t-statistic,$\mathtt{p}^{adj.t}$, and the un-adjusted t-statistic, $\mathtt{p}^{t}$. The differences suggest that the gender has influence on the gene expression profile.

	ID	PW Name	$\mathrm{p}^{adj.t}$	p^{t}	Size
1	04510	Focal adhesi...	0.0000	0.0006	73
2	04512	ECM-receptor...	0.0000	0.0000	15
3	04514	Cell adhesio...	0.0000	0.0000	38
4	04630	Jak-STAT sig...	0.0000		57
5	04520	Adherens jun...	0.0000	0.0000	34
6	04530	Tight juncti...	0.0000	0.0000	40
7	04650	Natural kill...	0.0000		61
8	03010	Ribosome	0.0000		23
9	04060	Cytokine-cyt...	0.0000	0.0030	54
10	04660	T cell recep...	0.0000		51
11	04810	Regulation o...	0.0000	0.0066	79
12	04670	Leukocyte tr...	0.0000	0.0002	49
13	04940	Type I diabe...	0.0000	0.0018	20
14	04350	TGF-beta sig...	0.0000		35
15	04010	MAPK signali...	0.0000		109
16	04610	Complement a...	0.0000	0.0000	14
17	04612	Antigen proc...	0.0000		41
18	04360	Axon guidanc...	0.0000	0.0004	38
19	04210	Apoptosis	0.0010	0.0096	46
20	04080	Neuroactive ...	0.0010	0.0002	20
21	04620	Toll-like re...	0.0020		35
22	05120	Epithelial c...	0.0020		23
23	00310	Lysine degra...	0.0020		14
24	00230	Purine metab...	0.0020		58
25	04330	Notch signal...	0.0030		14
26	00071	Fatty acid m...	0.0030		14
27	00190	Oxidative ph...	0.0030		59
28	00240	Pyrimidine m...	0.0050		39
29	04310	Wnt signalin...	0.0060		52
30	00620	Pyruvate met...	0.0060		16
31	04020	Calcium sign...	0.0070		46
32	00010	Glycolysis /...	0.0090		22

Table 4: Significant pathways reported using posterior probability as gene statistic. The columns $\mathtt{p}^1$ and $\mathtt{p}^2$ 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 $\mathtt{B}\downarrow$ and $\mathtt{B}\uparrow$ show the number of genes that are higher or lower in `BCR/ABL`, respectively, among the genes with posterior probability at least 0.01

	ID	PW Name	p^1	p^2	Size	B↑	B↓
1	04060	Cytokine-cyt...	0.0080	0.0018	54	25	8
2	04520	Adherens jun...		0.0041	34	15	7
3	04360	Axon guidanc...		0.0092	38	15	5

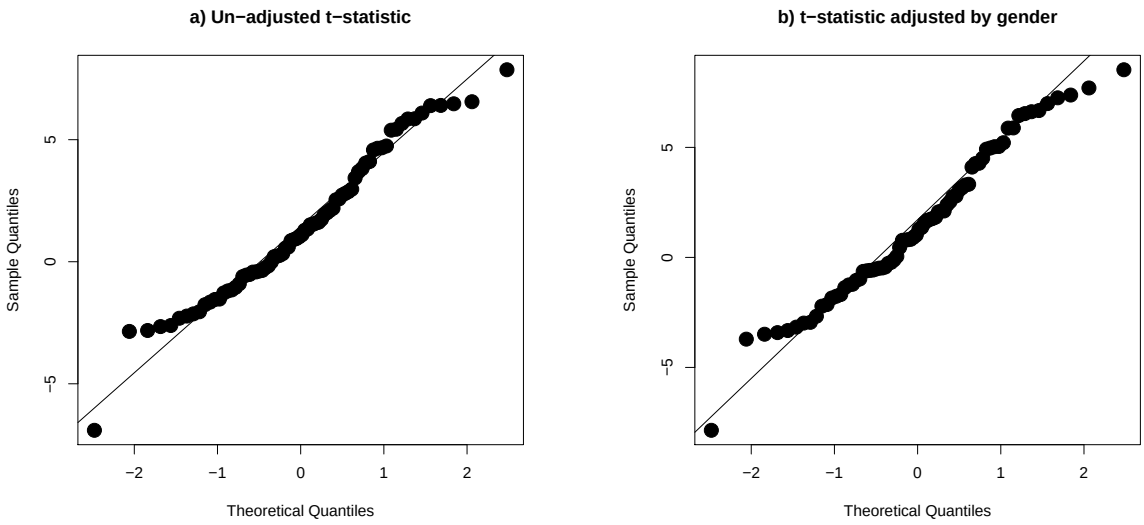

Figure 2: qq-plots of the pathway statistics for the `ALL` data. (a) Two-sample t-statistic. (b) The t-statistic from linear model, with an adjustment for sex.

second approach.

We checked the direction of expression changes from `BCR/ABL` to `NEG` for the genes with posterior probability higher than 0.01 in these pathways. In all cases more than two thirds of the genes are larger in one 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.

3.3.3 Modifying the 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 (Table 13 in the supplemental material of Yeoh et al. [2002] at http://www.stjuderesearch.org/data/ALL1).

Yeoh et al. [2002] used the same gene chip as ours. 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 is less than 10^{-4}, indicating a very strong concordance between our data and that of Yeoh et al. [2002].

3.4 Aliasing

Pathways have a fairly large amount of overlap and there are many different pathways that share a large number of genes with other pathways (Table 1). In this section, we emphasize the idea of dealing with aliasing and partially overlapping gene sets to understand the relationship among gene sets.

3.4.1 Example of largely overlapping gene sets

We consider the two pathways, the *Leukocyte transendothelial migration* pathway and the *Regulation of actin cytoskeleton* pathway. Both pathways were found significant by t-test (Table 2). This pair of pathways has

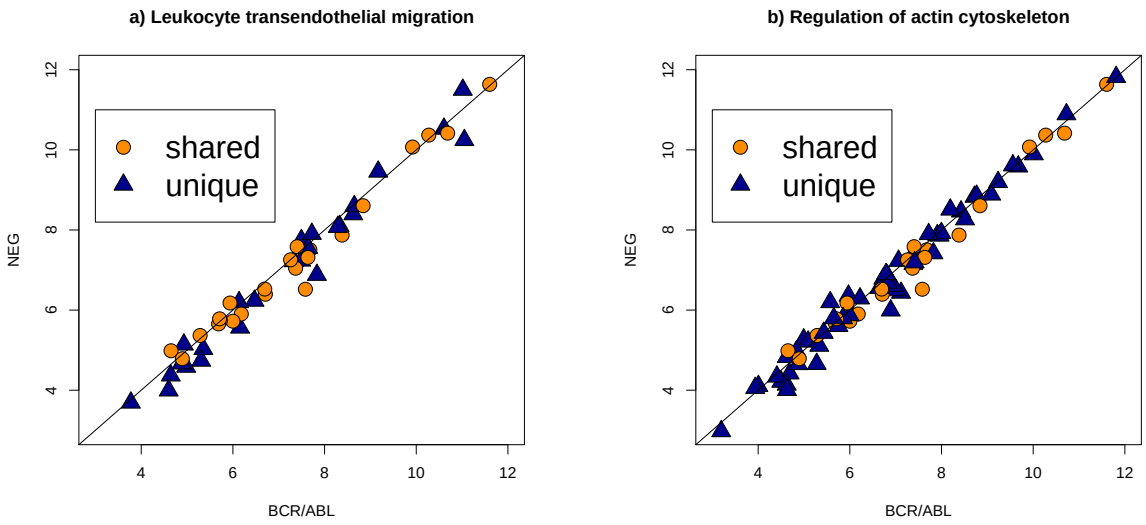

Figure 3: Mean plots for (a) the Leukocyte transendothelial migration pathway. (b) the Regulation of actin cytoskeleton pathway.

Table 5: Results for the subsets in pathway pair of Leukocyte transendothelial migration and Regulation of actin cytoskeleton

	Name	Size	Test Statistic	p-value
1	Common	23	19.0716	0.0042
2	Leukocyte tr...	26	25.6994	8e-04
3	Regulation o...	56	29.1866	0.0174

23 genes in common.

In Figure 3(a) and Figure 3(b), we present a graphical display of the two pathways. Each point represents a gene and the position on the x-axis is the mean expression value for that gene over all samples in the `BCR/ABL` group, while the value on the y-axis represents the mean expression value for that gene in the `NEG` group. Points that lie above the 45-degree line have higher expression values in the `NEG` group while those that lie below the 45-degree line have larger values in the `BCR/ABL` group. Genes that are found in both pathways are colored orange while genes unique to one of the two pathways are dark blue.

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 *Leukocyte transendothelial migration* pathway also tend to be below the 45-degree line, while those genes found only in the *Regulation of actin cytoskeleton* 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 *Regulation of actin cytoskeleton* will not be significant since the observed effects seem to cancel each other out.

We divided these two pathways into three gene sets, one set for the genes unique to each pathway and one set for the shared genes and then carried out GSEA on the three gene sets. The analysis was based on the permutation of sample labels, and the test results are summarized in Table 5. Genes unique to the *Leukocyte transendothelial migration* pathway exhibit a significant effect, as do those that are shared

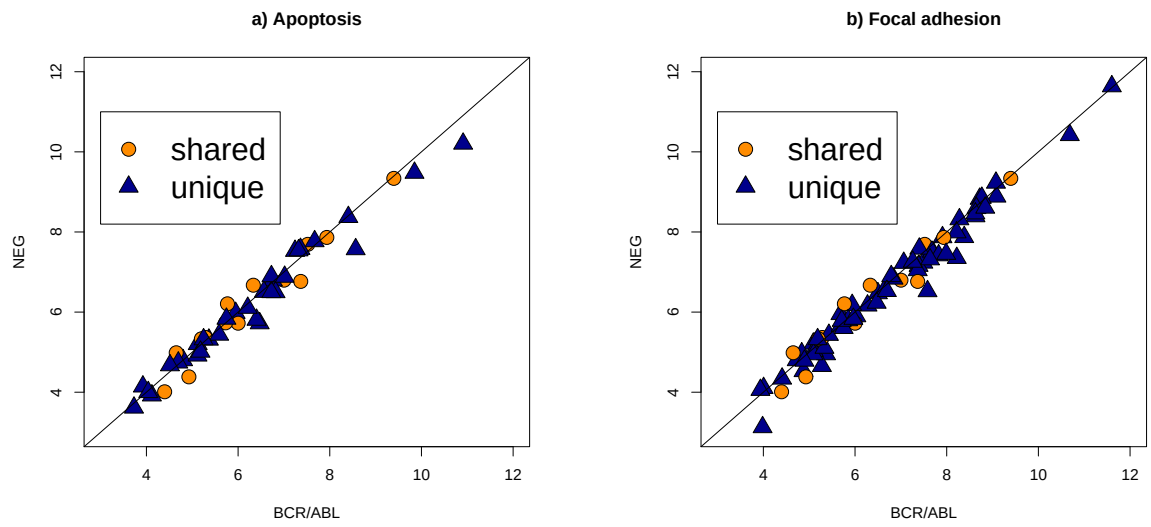

Figure 4: Mean plots for (a) the Apoptosis pathway. (b) the Focal adhesion pathway.

Table 6: Results for the subsets in pathway pair of the Apoptosis pathway and the Focal adhesion pathway

	Name	Size	Test Statistic	p-value
1	Common	14	3.3012	0.2318
2	Apoptosis	32	21.7289	0.0066
3	Focal adhesi...	59	51.4231	4e-04

and most importantly the direction of the effect in both groups is the same. But for genes unique to the *Regulation of actin cytoskeleton* pathway there seems to be no effect. This observation strongly suggests that the effect observed is due to the *Leukocyte transendothelial migration* pathway activation and not to the *Regulation of actin cytoskeleton* pathway activation.

3.4.2 A second example

In this section, we compare the *Apoptosis* pathway and the *Focal adhesion* pathway with 46 and 73 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 Figure 4. Unlike our previous sample, we do not see any obvious patterns in these two plots. The permutation test results in Table 6 indicate that those genes that are common to the two pathways are not significant. 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 *Leukocyte transendothelial migration* pathway in patients

with BCR/ABL and that there is little evidence of the *Regulation of actin cytoskeleton* pathway involvement. In the second example, the two pathways are likely to be independently activated in patients with BCR/ABL.

3.5 Principle component analysis (PCA) per gene set

We applied the PCA to the gene expression of each pathway. The expression values were standardized by subtracting the median and dividing by the median absolute deviation (MAD).

Following the approach mentioned in section 2.3.2, we obtained the number of PCs needed to explain 70% of the variation and separately, the number of PCs that identified by the isotropic test. Table 7 reports the gene sets which need at most 4 components to keep at least 70% of variation. We report the number k, the ratio of k to pathway size, and the proportion of variation for the first three PCs of these pathways.

Then, we applied the isotropic test to the pathways. The value of k identified by the isotropic test were 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 Figure 5 except the first component, all the other components are not very important. But the isotropic test suggested keeping 13 PCs.

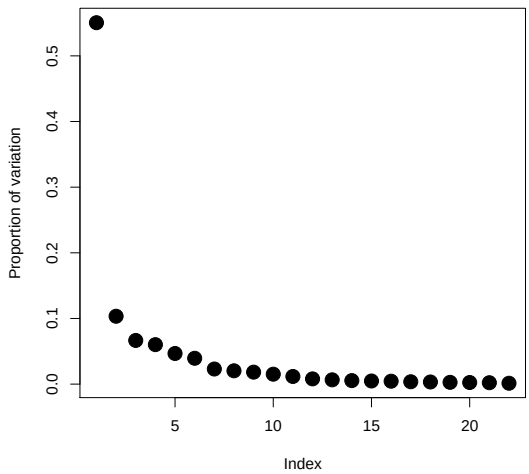

Figure 5: The eigenvalues of the genes in Ribosome pathway, the genes on the Y chromosome are excluded.

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. The ratio between the number of PCs needed for 70% of variation to the pathway size is also very low. We know that some genes in the *Ribosome* pathway are sex related (for example, PRKY is on the Y chromosome). The results of the PCA analysis of this pathway (Figure 6) are mostly the gender differences (first component) and the between subject variation (second component). We believe there are four subjects with gender annotation mistakes. Two of them are recorded as females with data indicated as males and two of them are recorded as males with data indicated as females. We made corresponding corrections.

We removed all the genes on the Y chromosome and repeated the PCA analysis on the remaining genes. The results (Table 7) did not change except for the *Ribosome* pathway. The plots of the new PCs for the *Ribosome* pathway are shown in Figure 7. The first PC is no longer dominated by one gene. The plot of the loadings of the first PC, without genes on the Y chromosome, against the loadings of the second PC with Y genes (Figure 7(b)) showed that the new first PC is the old second PC plus a small gene effects. All

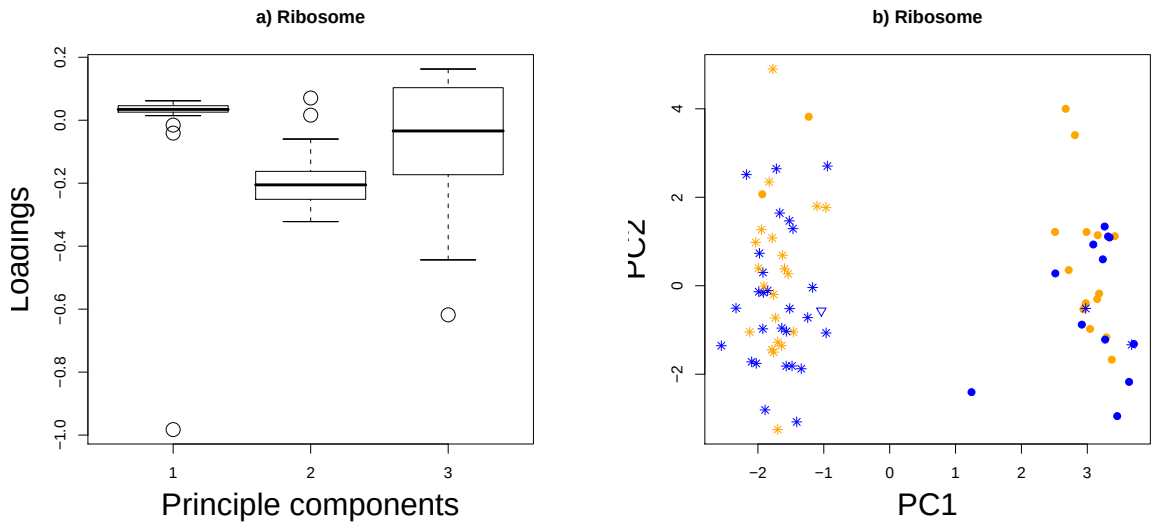

Figure 6: The PCA results for the Ribosome pathway. (a) Boxplots of the loadings for the first three components of Ribosome pathway. The first component is dominated by one gene, ribosomal protein S4, Y-linked 1, which is a Y chromosome gene. (b) Biplot of the first two PCs of the Ribosome pathway. The 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.)

the female samples are below the -45 degree line and all the male samples are above the -45 degree line. 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.

Another way to reduce the variation from gender differences is to adopt the ideas in Section 3.3.1 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.

	ID	PW Name	Size	k	Ratio	PC1	PC2	PC3	k^Y	$Ratio^Y$	$PC1^Y$	$PC2^Y$	$PC3^Y$
1	03010	Ribosome *	23	2	0.087	0.476	0.309	0.043	3	0.130	0.587	0.081	0.063
2	04320	Dorso-ventra...	12	3	0.250	0.316	0.262	0.128	3	0.250	0.316	0.262	0.128
3	00650	Butanoate me...	14	3	0.214	0.376	0.252	0.090	3	0.214	0.376	0.252	0.090
4	00071	Fatty acid m... *	14	3	0.214	0.458	0.139	0.103	3	0.214	0.458	0.139	0.103
5	00251	Glutamate me...	11	4	0.364	0.367	0.201	0.122	4	0.364	0.367	0.201	0.122
6	00051	Fructose and...	15	4	0.267	0.311	0.170	0.137	4	0.267	0.311	0.170	0.137
7	00310	Lysine degra...	14	4	0.286	0.429	0.153	0.091	4	0.286	0.429	0.153	0.091
8	05040	Huntington's... *	16	4	0.250	0.435	0.155	0.089	4	0.250	0.435	0.155	0.089
9	01031	Glycan struc... *	11	4	0.364	0.265	0.185	0.157	4	0.364	0.265	0.185	0.157
10	00790	Folate biosy... *	12	4	0.333	0.276	0.233	0.119	4	0.333	0.276	0.233	0.119
11	00564	Glycerophosp... *	14	4	0.286	0.329	0.174	0.120	4	0.286	0.329	0.174	0.120
12	00340	Histidine me... *	12	4	0.333	0.255	0.240	0.131	4	0.333	0.255	0.240	0.131
13	00710	Carbon fixat... *	11	4	0.364	0.317	0.209	0.121	4	0.364	0.317	0.209	0.121
14	00350	Tyrosine met... *	15	4	0.267	0.323	0.203	0.112	4	0.267	0.323	0.203	0.112

Table 7: Pathways for which four or fewer PC's explain at least 70% of the variability. 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 ith PC. The columns with superscript "Y" are the PCA analysis results after removing the genes on the Y chromosome. Those labeled with a * are significant by permutation test.

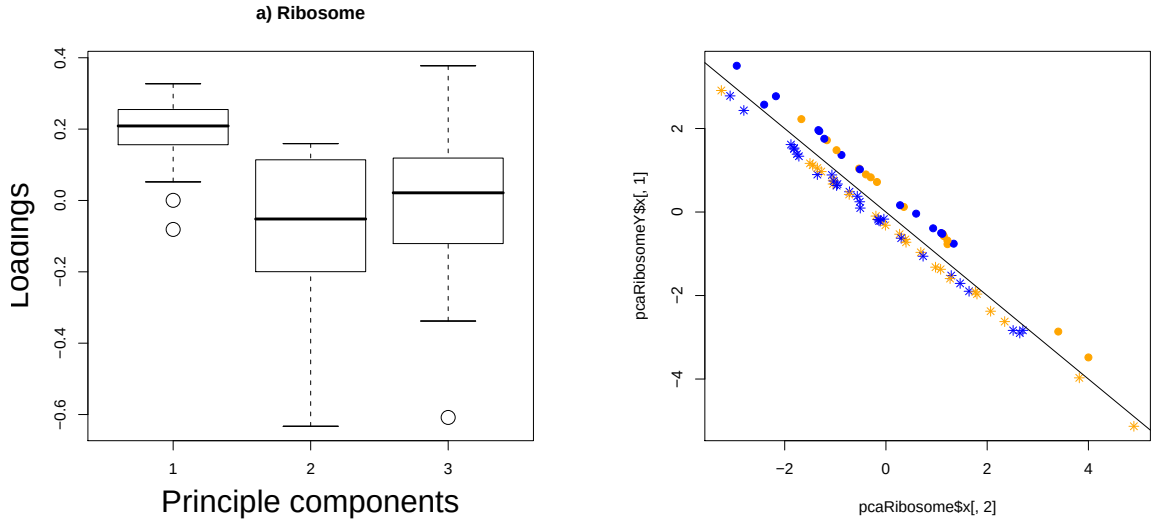

Figure 7: The PCA results for the Ribosome pathway after removing the Y chromosome genes. The gender annotation has been corrected. (a) Boxplots of the loadings for the first three principle components. (b) Biplot of the first two PCs.

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 53 pathways with p-values less than 0.01, including 9 out of the 14 pathways in Table 7 (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, we have shown that any per category summary statistic can be used and that the usual use of a t-test can easily be extended to any linear modeling situation, where standardized estimates of the effects can be employed in computing the per category measure of change.

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 overlaps are both harder to model, and to interpret.

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 does appear 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*.

References

Richard A. Johnson and Dean W. Wichern. *Applied Multivariate*. Prentice Hall, 1988.

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

Seon-Young Kim and David J Volsky. Page: Parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6;144, June 2005.

P. K. Majumder, P. G. Febbo, R. Bikoff, et al. mTOR inhibition reverses akt-dependent prostate intraepithelial neoplasia through regulation of apoptotic and HIF-1-dependent pathways. *Nature Medicine*, 10: 594–601, 2004.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

V. K. Mootha, C. M. Lindgren, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34:267–273, 2003.

M.A. Newton, C.M. Kendziorski, C.S. Richmond, F.R. Blatterner, and K.W. Tsui. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8:37–52, 2001.

A. Subramanian, P. Tamayo, V. K. Mootha, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102 (43):15545–15550, 2005.

The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25: 25–29, 2000.

L. Tian, S. A. Greenberg, S. W. Kong, et al. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences*, 102(38):13544–13549, 2005.

E-J. Yeoh, M. E. Ross, S. A. Shurtleff, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133–143, March 2002.