4-21-2009

# GENE SET ENRICHMENT ANALYSIS MADE SIMPLE

Rafael A. Irizarry
*Johns Hopkins University, Bloomberg School of Public Health, Department of Biostatistics*, ririzarr@jhsph.edu

Chi Wang
*Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics*

Yun Zhou
*Johns Hopkins University School of Medicine, Department of Radiology*

Terence P. Speed
*University of California, Berkeley, Department of Statistics and Division of Genetics and Bioinformatics, Walter and Eliza Hall Institute of Medical Research, Melbourne*

# Gene Set Enrichment Analysis Made Simple

Rafael A. Irizarry*, Chi Wang, Yun Zhou, Terence P. Speed*

**Abstract**

Among the many applications of microarray technology, one of the most popular is the identification of genes that are differentially expressed in two conditions. A common statistical approach is to quantify the interest of each gene with a $p$-value, adjust these $p$-values for multiple comparisons, chose an appropriate cut-off, and create a list of *candidate genes*. This approach has been criticized for ignoring biological knowledge regarding how genes work together. Recently a series of methods, that do incorporate biological knowledge, have been proposed. However, many of these methods seem overly complicated. Furthermore, the most popular method, Gene Set Enrichment Analysis (GSEA), is based on a statistical test known for its lack of sensitivity. In this paper we compare the performance of a simple alternative to GSEA. We find that this simple solution clearly outperforms GSEA. We demonstrate this with eight different microarray datasets.

## 1 Introduction

The problem of identifying genes that are differentially expressed in two conditions has received much attentions from the statistical community and data analysts in general. Most of the work has focused on designing appropriate test statistics (Tusher, Tibshirani and Chu 2001, Smyth 2004) and developing procedures to account for multiple comparisons (Storey and Tibshirani 2003, Dudoit, Shaffer and Boldrick 2003). Most approaches follow a similar recipe: decide on a null hypothesis, test this hypothesis for each gene, produce a $p$-value, and attach a significance level that accounts for multiplicity. At the end, each gene receives a score which we use to decide if it is in our final

*To whom correspondence should be addressed

list of significant genes. Those on this final list are typically called *candidate genes* because further validation tests are commonly performed. In this paper, we refer to this as the *marginal* approach. A limitation of this approach is that genes that are known to be biologically associated are scored independently. Although many important discoveries have been made with this approach, the resulting gene lists do not always provide useful biological insights.

Recently, various approaches have been proposed to incorporate biological knowledge into the analysis. The vast majority of these have relied on the results from the marginal approach instead of starting from the original expression data. Because many of these marginal procedures have been useful and given the complicated nature of microarray data we view this as a correct first approach. In this paper we do not discuss nor propose methods that start from scratch.

There are currently two major types of procedure for incorporating biological knowledge into differential expression analysis. We will refer to these as the *over-representation* and the *aggregate score* approaches. In both, gene categories or *gene sets* are formed prior to the statistical analysis. The sets are formed by, for example, grouping genes that are part of the same cellular components, are essential for a biological process, or have the same molecular function. In many cases the gene sets target the condition that is being studied. However, it is more common to use category definitions from the Gene Ontology project (Lee, Braynen, Keshav and Pavlidis 2005). The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism (The Gene Ontology Consortium 2000).

Over-representation analysis can be summarized as follows: First, form a list of candidate genes using the marginal approach. Then, for each gene set, we create a two-by-two table comparing the number of candidate genes that are members of the category to those that are not members. The significance of over-representation can be assessed, for example, using the hypergeometric distribution or its binomial approximation. More elaborate approaches exists and a large number of over-representation methods have been published. Many of these have been implemented as web-tools. A comprehensive list can be found at
`http://www.geneontology.org/GO.tools.microarray.shtml`.

A limitation of the over-representation approach is that it ignores all the genes that did not make the list of candidate genes. Therefore, the results will be highly dependent on the cutoff used in constructing this list. In fact, examples can be found where very few, or even none, of the genes

in functional groups known to behave different in the two conditions survive the typical filters and therefore the groups are not detected as interesting. Mootha, Lindgren, Eriksson, Subramanian, Sihag, Lehar, Puigserver, Carlsson, Ridderstråle, Laurila, Houstis, Daly, Patterson, Mesirov, Golub, Tamayo, Spiegelman, Lander, Hirschhorn, Altshuler and Groop (2003) describes a particularly interesting example. The *aggregate score* approach, does not have this limitation. The basic idea is to assign scores to each gene set based on all the gene-specific scores for that gene set. There are various ways to calculate these aggregate scores (Pavlidis, Lewis and Noble 2002, Pavlidis, Qin, Arango, Mann and Sibille 2004, Mootha et al. 2003, Goeman, van de Geer, de Kort and van Houwelingen 2004, Goeman, Oosting, Cleton-Jansen, Anninga and van Houwelingen 2005, Kim and Volsky 2005, Subramanian, Tamayo, Mootha, Mukherjee, Ebert, Gillette, Pomeroy, Golub, Lander and Mesirov 2005, Tian, Greenberg, Kong, Altschuler, Kohane and Park 2005). In this paper we focus on the aggregate score method rather than the over-representation approach.

Of these methods GSEA (Mootha et al. 2003, Subramanian et al. 2005) is by far the most popular. Surprisingly, GSEA is based on the Kolmogorov Smirnov (K-S) test which is well known for its lack of sensitivity and limited practical use. Subramanian et al. (2005) seem to have realized this and developed an ad-hoc modification of the K-S test. A further limitation of the K-S test and its modified versions, is that the null distribution of the score is hard to compute. Tian et al. (2005) proposed the use of the standard statistical approach for detecting shifts in center: a one sample $z$-test. Tian et al. (2005) propose the use of permutation tests for assessing the significance of the $z$-test. However, they do not explore the performance of the standard parametric approach. We find that using the one sample t-test along with a standard multiple comparison adjustment (Storey 2002) of the normal distribution $p$-value works well in practice. This procedure is extremely simple in comparison to GSEA and requires practically no computation time.

A possible advantage of GSEA, i.e. the K-S test, over the one sample $z$-test is that the latter is specifically designed to identify gene sets with mean shifts and the K-S test is designed to find general difference in the cumulative distribution. In principle, we want to be able to detect gene sets for which some members are up-regulated and others are down-regulated. The $z$-test is not sensitive to this change as there is no shift in mean. We therefore, propose the use of another standard statistical test useful for detecting changes in scale: the $\chi^2$ test.

In this paper we compare GSEA to the one sample $z$-test and $\chi^2$-test using all the datasets

described in Mootha et al. (2003) and Subramanian et al. (2005). In Section 2 we briefly describe the methods in question. In Section 3 we present the results from the comparison. Finally, in Section 4 we discuss these results describe some current work that we expect to improve upon our proposed method and give concluding remarks.

# 2 Methods

Most aggregate score approaches start with the results from a marginal analysis. For example, we may start with a $t$-statistic $t_i$ for each gene $i = 1, \ldots, N$. We then identify gene set $g$ with a subset $A_g \subset \{1, \ldots, N\}$. We want our score, say $E_g$ ($E$ for enrichment), to quantify how *different* the $t_i, i \in A_g$ are from the $t_i, i \notin A_g$. A second task is to assign a level of significance to each $E_g$. Most methods take the approach of defining a null hypothesis, calculating the null distribution, and assigning a level of significance. Because the score for dozens of gene sets are considered, the significance levels are adjusted for multiple comparisons. The competing methods differ in the way that *different* is quantified and the null hypothesis defined and calculated. Notice, that $t_i$ need not be a $t$-statistic. In fact the GSEA paper uses another statistics that summarized the signal to noise ratio for each gene. Because the resulting values are very similar to a $t$-statistic we refer to the $t_i$ as signal to noise value and $t$-statistic interchangeably.

Mootha et al. (2003) used a version of the Kolmogorov-Smirnov (K-S) statistic to test for differences in the distributions of the $t$-statistics related to members of a gene set compared to $t$-statistics from the rest of the genes. Because they were interested in comparing these scores across gene sets of different sizes, and then null distribution of the K-S statistic depends heavily on this size, Mootha et al defined a normalized K-S statistics as their score $E^{GSEA}$. To assess the significance of these scores a permutation test was performed. Specially, they permuted the sample labels and re-computed $E_g^{GSEA}$ 1000 times. In each permutation the maximum enrichment score was recorded. These 1000 values defined the null distribution and used to assign $p$-values. Mootha et al. (2003) for details.

Subramanian et al. (2005) seem to have noticed the lack of power of the K-S test, a well-known fact, and proposed an ad-hoc modification to improve this. Furthermore, in the original version of GSEA, an adjusted $p$-value was calculated only for the enrichment score of the top ranking set. In

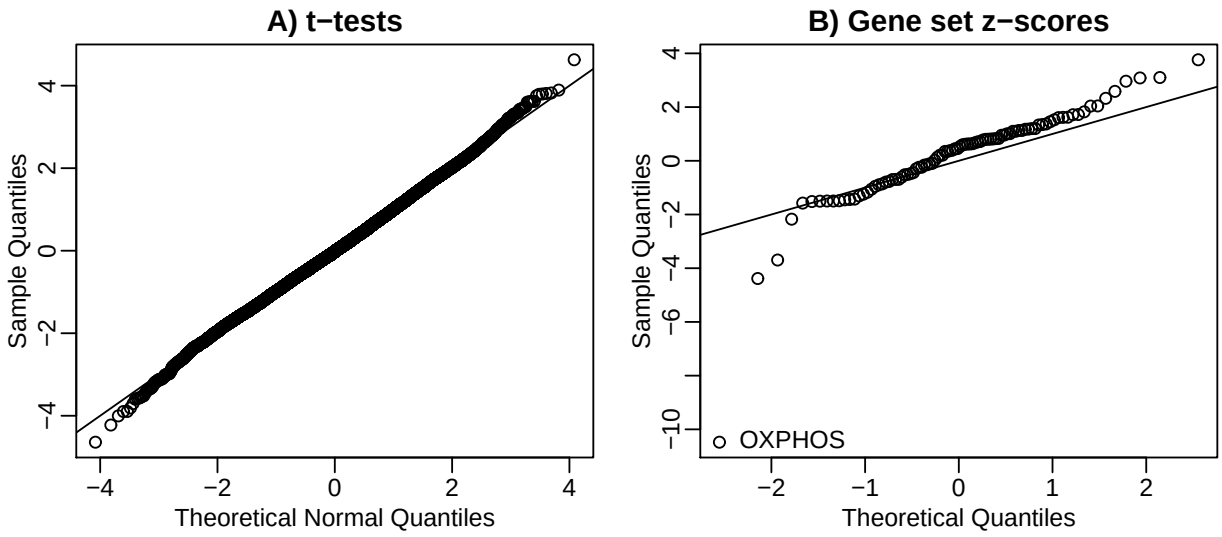


Figure 1: Quantile-quantile plots. A) For the diabetes data presented in Mootha et al. we plot the quantiles of the observed $t$-statistics versus the theoretical quantiles of the standard normal distribution. The identity line is shown. B) For the same data we show the enrichment score based on the $z$-test for the gene sets presented by Mootha et al. The score for the OXPHOS gene set is high-lighted.

Subramanian et al. (2005), after normalizing the test statistic for each gene set, the FDR $q$-value for each gene set was calculated and used to select candidate gene sets. The end results is a rather complicated method that takes minutes to run on a typical laptop computer.

Determining if two sets of numbers have different distribution is certainly not a new problem. Many solutions exist. The K-S test is one that has not been used in many (or any) other applications, so why use it here? Let us start with the most basic statistical approach: test for a shift in center/mean as proposed by Tian et al 2005. If, under the null hypothesis, the $t_i$ are normally distributed with mean 0 and standard deviation 1, inference can be done with a one sample $z$-test. For a robust version we could use a Wilcoxon test. When enough replicates are available in each condition we expect the $t$-statistics to follow a standard normal distribution under the null-hypothesis of no difference between the conditions. The data presented by Mootha et al seem to satisfy this assumption. Figure 1A shows a quantile-quantile plot comparing the t-tests used in Mootha et al. to a standard normal distribution. Figure 2 shows this quantile-quantile plot for all datasets

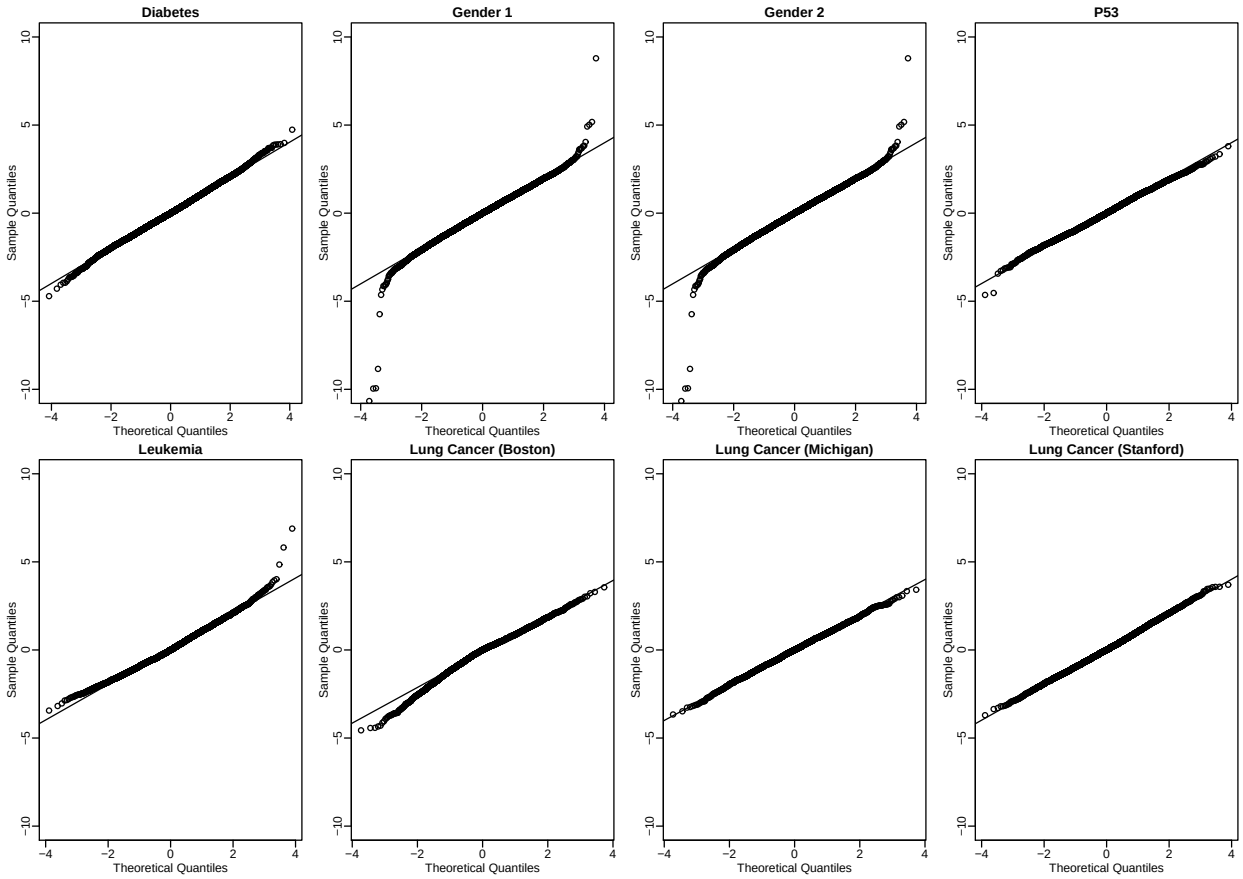


Figure 2: As Figure 1 but for all the datasets presented in Mootha et al. 2003 and Subramanian et al. 2005. The identity line is shown.

in Subramanian et al. (2005). Barring a few outliers, which are likely associated to differentially expressed genes, the assumption appears appropriate in all datasets. If we assume that these tests are independent (under the null) then for any given gene set the $z$-score:

$$E_g^z = \sqrt{N_g}\,\bar{t}, \text{ with } \bar{t} = \frac{1}{N_g}\sum_{i \in A_g} t_i, \tag{1}$$

with $N_g$ the number of genes in $A_g$, also follows a standard normal distribution. This implies that we can easily obtain a $p$-value.

With appropriate $p$-values calculated we have numerous multiple comparison adjustment methods to choose from and do not need to perform permutation tests. Tian et al argue that the normality assumption is not appropriate because we expect the $t_i$ to be correlated even under the null hypothesis. However, they do not appear to have tested this empirically. We find that assuming the $E_g^z$ are

normally distributed under the null hypothesis is in fact a useful approximation for all the examples we examined. For example, Figure 1B shows the $z$-score for the dataset presented in Mootha el al. for the same gene sets they considered. Notice that the obvious outlier in Figure 1B, is the OXPHOS gene set discovered to be important by Mootha et al. Thus, the discovery that merited their publication would have been made with a statistical method that could be explained in one paragraph instead of several pages.

A possible limitation of the one sample $z$-test is that it will not detect changes in scale. A gene set where half the gene sets are up regulated and the other half are down regulated may have no mean shift but is certainly interesting from a biological standpoint. The standard test for scale change, i.e. the $\chi^2$-test, is useful for this. We define a standardized $\chi^2$-test that permits us to compare gene sets of different sizes and different mean shifts:

$$E_g^{\chi^2} = \frac{\sum_{i \in A_g} (t_i - \bar{t})^2 - (N_g - 1)}{2(N_g - 1)}. \tag{2}$$

For gene sets that are large enough, say $> 20$, $E_g^{\chi^2}$ follows a standard normal distribution as well. Thus computing $p$-values and adjusting these is just as straight forward as for the $z$-test.

# 3 Results

We computed the $z$-score and normalized $\chi^2$ for all gene sets and all datasets presented in Mootha et al. (2003) and Subramanian et al. (2005). We used the latest version of GSEA. We adjusted for multiple comparisons using Storey's $q$-value (Storey 2002). We compared these to the $q$-values computed using GSEA. Table 1 shows all the gene sets achieving a GSEA $q$-values of less than $0.25$, as done by Subramanian et. al. With the exception of only three cases out of 4139, all gene sets found by GSEA to have $q$-values $< 0.025$ were either in the top 10 gene sets or had a $q$-value less than 0.05 for either the $z$-test or the $\chi^2$ test. The three cases are highlighted with bold letters in Table 1. Notice that all three were found in the Michigan Lung Cancer dataset.

Figure 3 shows two gene sets: the GO ROS group in the Michigan Lung Cancer dataset and the GLUT DOWN gene set in one of the Gender datasets. GO ROS would be considered interesting in the Michigan Lung Cancer study by GSEA but not by the simpler methods. GLUT DOWN would be considered interesting in the Gender data set by the $z$-test but not by GSEA. The only

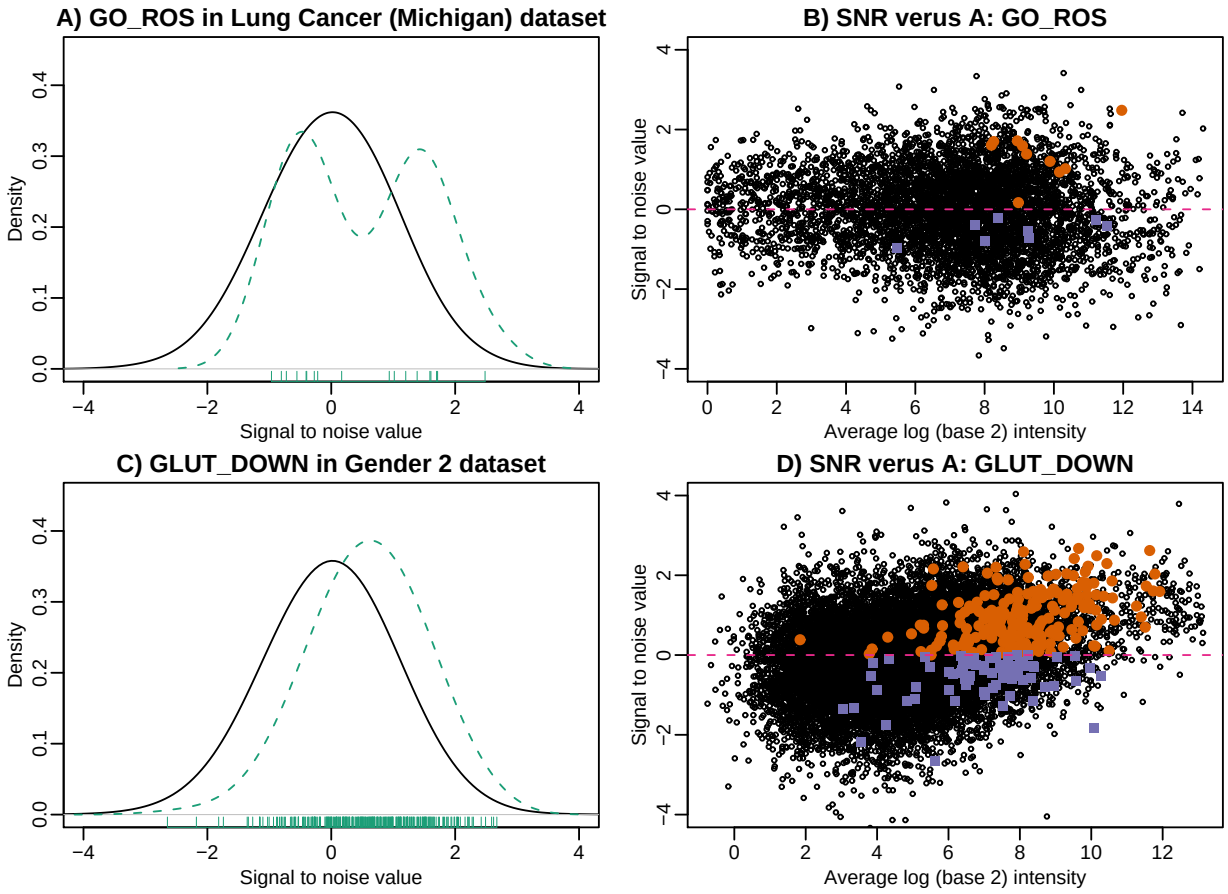


Figure 3: Gene sets showing disagreement between GSEA and the $z$-test. A) Empirical density estimate of the the signal to noise values for the GO ROS group (dashed lines) and the rest of the genes (solid line). The ticks on the x-axis show the actual observations. This particular group had a small GSEA $q$-value but a $z$-test and $\chi^2 > 0.25$. B) For each gene, signal to noise values plotted against the average intensity for the same dataset as in A). The values for the GO ROS gene set are highlighted. Circles denote the up-regulated genes in the gene set and squares denote the down-regulated genes. C) As A) but for the GLUT DOWN gene set in the Gender data set. The $z$-test approach results in a very small $q$-value ($< 0.001$) for this gene set but a GSEA $q$-value larger than 0.25. D) As B) but for the data described in C).

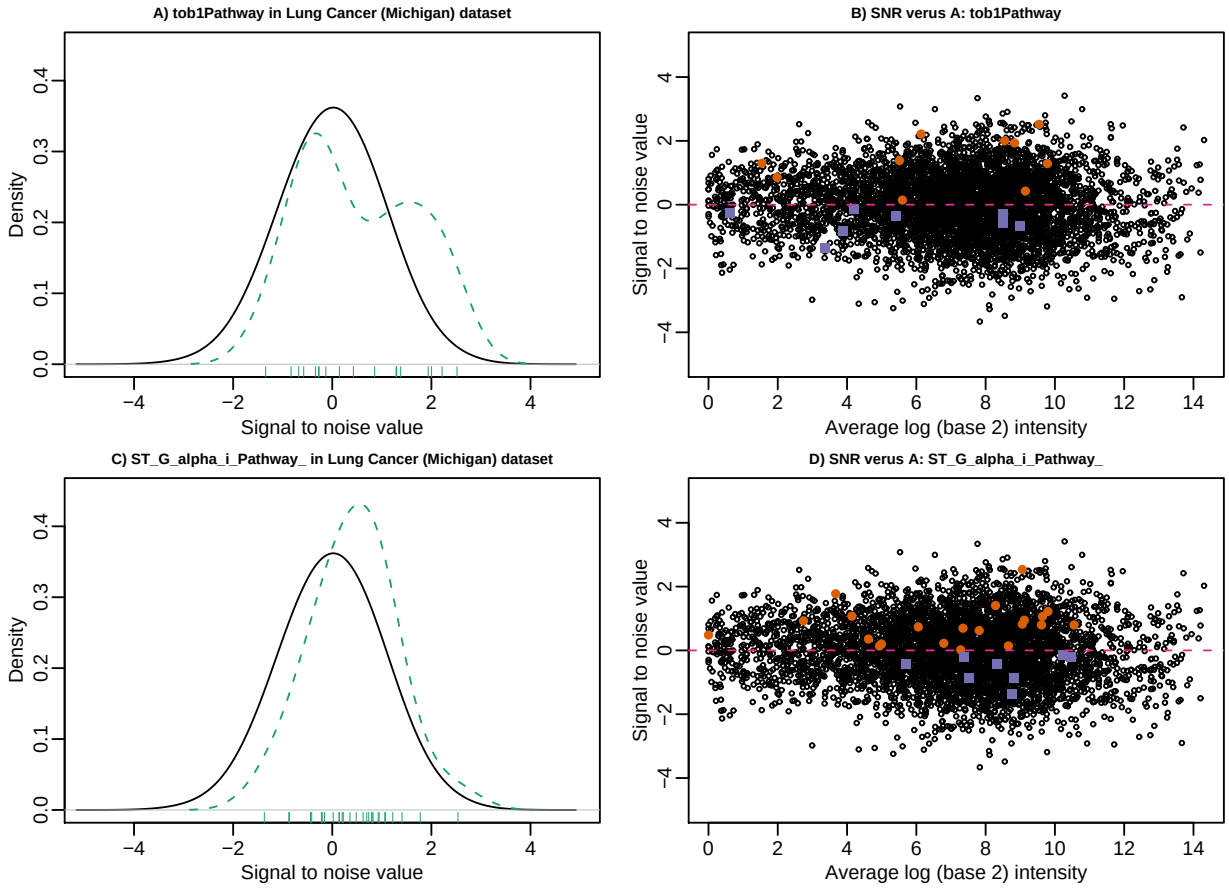


Figure 4: As Figure 3 but for the two other gene sets found by GSEA and not by the $z$-test or $\chi^2$-test.

interesting feature of the GO ROS group is a gap (no observations of $t_i$) between 0.5 and 1. We do not consider this to be interesting enough to merit detection. On the other hand the GLUT DOWN has a clear shift in mean. Figure 4 shows the other two gene sets found by GSEA and not by the other methods. They do not appear interesting in any way.

Subramanian et al. (2005) pointed out that there is very little agreement in the results obtained from the three lung cancer datasets they studied. They demonstrate the advantages of GSEA over the marginal approach by showing better agreement between aggregate scores as compared to marginal ones. We created lists of the top gene sets for these three studies using four different approaches: the top 30 gene sets (lowest q-values) in each group as found by GSEA and the $z$-test, all the gene sets with FDR $< 0.25$ for GSEA, and all the gene sets with FDR $< 0.05$ for the $z$-test.

Table 1: For each of the eight datasets studies by Mootha et al. 2003 and Subramanian et al. 2005 we found the gene sets for which GSEA reports a $q$-value of 0.25 or less. Note that the Stanford dataset had no gene sets passing this requirement. For the rest we show the $q$-values obtained for these same gene sets when using the $z$-test and the $\chi^2$-test. The ranks of the gene sets obtained with each of these three methods, within the dataset, are also shown. There are only three examples for which the q-value was larger than 0.05 and the rank was larger than 10 in both the $z$-test and the $\chi^2$- test. These are shown in bold.

| | | | GSEA | | $z$-test | | $\chi^2$ test | |
|---|---|---|---|---|---|---|---|---|
| Study | Gene set | Size | q-value | Rank | q-value | Rank | q-value | Rank |
| Diabetes | MAP00360 Phenylalanine metabolism | 23 | 0.06 | 2 | 0.07 | 9 | 0.6 | 46 |
| Diabetes | MAP00910 Nitrogen metabolism | 30 | 0.3 | 3 | <0.01 | 6 | 0.6 | 43 |
| Diabetes | OXPHOS HG-U133A probes | 114 | 0.04 | 1 | <0.001 | 1 | 0.6 | 66 |
| Gender 1 | chrY | 40 | <0.001 | 1 | <0.001 | 1.5 | <0.001 | 2.5 |
| Gender 1 | chrYp11 | 18 | <0.001 | 3 | <0.001 | 3 | <0.001 | 2.5 |
| Gender 1 | chrYq11 | 16 | <0.001 | 2 | <0.001 | 1.5 | <0.001 | 2.5 |
| Gender 2 | XINACT MERGED | 20 | <0.001 | 1 | <0.001 | 6 | <0.001 | 2 |
| Gender 2 | GNF FEMALE GENES | 85 | 0.05 | 3 | <0.001 | 7 | <0.001 | 2 |
| Gender 2 | TESTIS GENES | 73 | 0.02 | 2 | <0.001 | 2.5 | <0.001 | 2 |
| P53 | rasPathway | 22 | 0.2 | 6 | <0.01 | 5 | 0.9 | 123 |
| P53 | p53hypoxiaPathway | 20 | <0.001 | 2 | 0.03 | 22 | <0.001 | 1 |
| P53 | hsp27Pathway | 15 | <0.001 | 2 | 0.01 | 14 | 0.4 | 40 |
| P53 | p53Pathway | 16 | <0.001 | 2 | <0.01 | 4 | <0.001 | 2 |
| P53 | P53 UP | 40 | 0.01 | 4 | <0.001 | 2 | <0.001 | 6 |
| P53 | radiation sensitivity | 26 | 0.08 | 5 | 0.02 | 16 | <0.001 | 3 |
| Leukemia | chr6q21 | 31 | 0.01 | 1 | <0.001 | 2 | 0.8 | 23 |
| Leukemia | chr5q31 | 59 | 0.05 | 2 | 0.03 | 7 | 0.1 | 86 |
| Leukemia | chr13q14 | 31 | 0.06 | 3 | 0.2 | 16 | 0.4 | 7 |
| Leukemia | chr14q32 | 64 | 0.08 | 5 | <0.01 | 3 | <0.01 | 2 |
| Leukemia | chr17q23 | 39 | 0.07 | 4 | <0.01 | 4 | 0.7 | 18 |
| Boston | p53hypoxiaPathway | 19 | 0.05 | 1 | <0.001 | 13 | <0.01 | 18 |
| Boston | Aminoacyl tRNA biosynthesis | 15 | 0.1 | 5 | <0.001 | 12 | 0.2 | 63 |
| Boston | INSULIN 2F UP | 113 | 0.1 | 2 | <0.001 | 2.5 | <0.01 | 22 |
| Boston | tRNA Synthetases | 16 | 0.2 | 7 | <0.001 | 9 | 0.3 | 91 |
| Boston | LEU DOWN | 124 | 0.1 | 4 | <0.001 | 2.5 | <0.01 | 27 |
| Boston | HTERT UP | 104 | 0.1 | 3 | <0.001 | 5 | 0.05 | 38 |
| Boston | GLUT DOWN | 199 | 0.2 | 6 | <0.001 | 2.5 | <0.001 | 8 |
| Boston | cell cycle checkpoint | 19 | 0.2 | 8 | <0.001 | 16 | 0.3 | 98 |
| Michigan | amiPathway | 22 | 0.01 | 3.5 | <0.001 | 6.5 | 1 | 208.5 |
| Michigan | cskPathway | 22 | 0.01 | 3.5 | <0.001 | 6.5 | 1 | 208.5 |
| Michigan | badPathway | 19 | <0.01 | 2 | 0.03 | 29 | 0.9 | 151 |
| Michigan | Il12Pathway | 22 | 0.05 | 6 | 0.01 | 23 | 0.9 | 79 |
| Michigan | no2il12Pathway | 16 | 0.08 | 7 | 0.02 | 25 | 1 | 246 |
| Michigan | **GO ROS** | 18 | 0.09 | 8 | 0.06 | 54 | 0.9 | 156 |
| Michigan | **tob1Pathway** | 18 | 0.2 | 17 | 0.06 | 53 | 0.9 | 69 |
| Michigan | HEMO TF LIST JP | 66 | 0.2 | 13 | <0.01 | 18 | 1 | 245 |
| Michigan | ctla4Pathway | 16 | 0.2 | 20 | <0.01 | 10 | 0.9 | 26 |
| Michigan | **ST G alpha i Pathway** | 29 | 0.2 | 16 | 0.05 | 50 | 0.9 | 68 |
| Michigan | MAP00010 Glycolysis Gluconeogenesis | 45 | <0.01 | 1 | <0.001 | 8 | 0.9 | 30 |
| Michigan | vegfPathway | 21 | 0.03 | 5 | <0.01 | 17 | 1 | 173 |
| Michigan | INSULIN 2F UP | 113 | 0.2 | 9 | <0.001 | 2 | 0.9 | 65 |
| Michigan | insulin signalling | 77 | 0.2 | 10 | 0.04 | 39 | 0.9 | 8 |
| Michigan | HTERT UP | 104 | 0.2 | 12 | <0.001 | 5 | 0.3 | 4 |
| Michigan | MAP00251 Glutamate metabolism | 18 | 0.2 | 14 | 0.01 | 21 | 0.9 | 19 |
| Michigan | ceramidePathway | 18 | 0.2 | 15 | <0.01 | 19 | 0.9 | 111 |
| Michigan | p53 signalling | 65 | 0.2 | 11 | <0.01 | 11 | 0.9 | 60 |
| Michigan | tRNA Synthetases | 16 | 0.2 | 18 | <0.01 | 14 | 0.9 | 55 |
| Michigan | MAP00970 Aminoacyl tRNA biosynthesis | 15 | 0.2 | 19 | <0.01 | 16 | 0.9 | 73 |

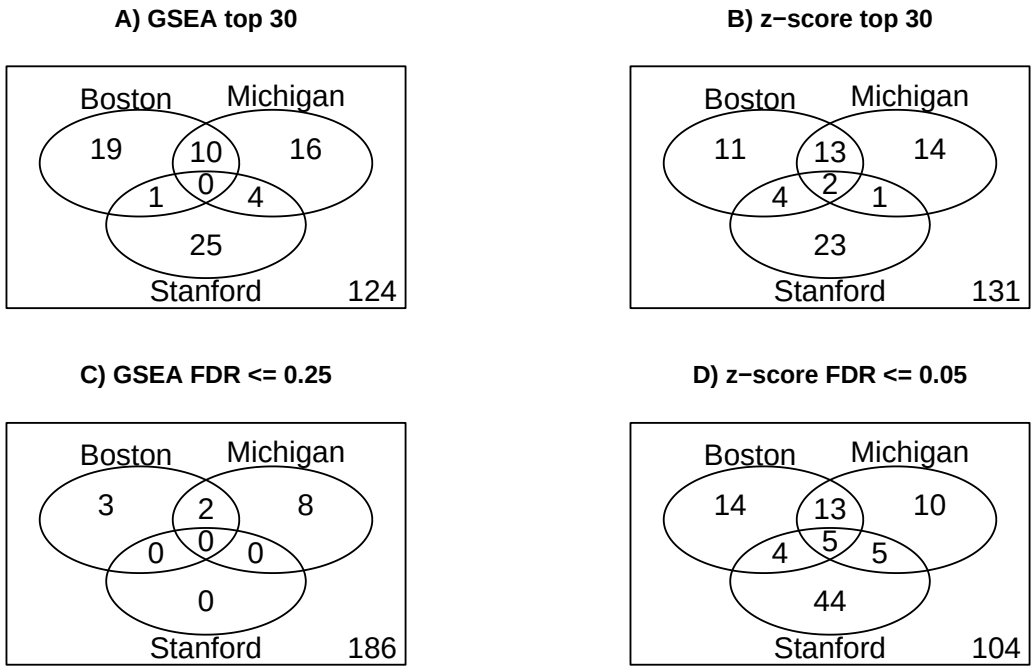


Figure 5: Gene set agreement, shown with Venn diagram, in lung cancer dataset. The numbers in the lower right corners are the number of gene sets that were not in any list. A) Agreement among top 30 gene sets ranked by their GSEA $q$-value. B) As A) but for the $z$-test. C) Agreement among gene sets achieving a GSEA $q$-value smaller than 0.25. D) As C) but for gene sets achieving a $q$-value smaller than 0.05 with the $z$-test.

Figure 5 shows Venn diagrams for the results. It is clear that much better agreement is found with the $z$-tests than with GSEA.

# 4 Discussion

We have compared GSEA to two very simple procedures based on standard statistical approaches: the one sided $z$-test and the $\chi^2$ test. We found that the simpler methods outperformed GSEA in assessments based on the eight datasets used in the GSEA papers and a simulation study. The great majority of gene sets found by GSEA to be interesting are also found by the $z$-test. Notice that if we expect gene sets to be interesting due to mean shifts then it is no surprise that the $z$-test outperforms GSEA since statistical theory predicts this test to be much more powerful than the K-S test. In fact, this is one reason we use the 0.05 cut-off, instead of 0.25, for the $z$-test $q$-value.

An argument for GSEA could be that some gene sets are interesting for reasons other than mean shifts, such as scale changes. For many of these cases the $\chi^2$ test was able to identify them as interesting. The only three gene sets not found by either the $z$-test or $\chi^2$ test are shown in Table 1, Figure 3 and Figure 4. For all three it is hard to argue that they are interesting in anyway. We notice that all three gene sets are small in size as compared to other gene sets and have unexpected gaps in the observations of the signal to noise values. It is possible that the *ad-hoc* modification of the K-S test is biased in favor of small gene sets.

Another advantage of the method presented here is that it can be easily extended to application other than the comparison of two conditions. There is no need for the statistics used to compute the enrichment scores described here, equation (1) and (2), to be $t$-statistics. Any statistics that we expect to follow a standard normal distribution can be used. For example, another common applications of microarrays examines cancer survival data. In these cases the summary statistics is commonly a parameter estimate from a Survival model. The standard normal approximation is a common approximation of the standardized versions of these estimates. Tian et al. (2005) argue against the use of the normal approximation for the averaged $t$-tests and propose the use of permutation-based tests. A disadvantage of their proposed permutation tests is that they are not easily extended to cases other than comparison of two conditions. Tian et al. (2005) correctly point out that if the $t$-statistics are correlated under the null hypothesis, the assumption that the $z$-score is normal with standard deviation 1 is incorrect. We did not find this to be a problem in practice. Furthermore, we find that the the average correlation in gene sets is of the order of 0.1 (data not shown), which only corresponds to a 5% inflation of the score. A correction factor can easily be inserted at the appropriate place.

An entirely parametric approach, as the one described here, has been previously proposed by Kim and Volsky (2005). Their approach, referred to as PAGE, ignores the marginal $t$-tests, and computes a $t$-test based on the effect sizes (log fold changes) within each gene set. A limitation of this approach is that it does not take into account the gene-specific variances. This is problematic because different genes are known to result in measurements with different variances (Kendziorski, Irizarry, Chen, Haag and Gould 2005). Furthermore, PAGE is restricted to applications of comparing two conditions. However, we expect PAGE to outperform GSEA as well.

We have made an argument against the use of GSEA. Methods that are much simpler, require

hardly any computation time, and can be easily implemented in any data analysis package, have been demonstrated to outperform GSEA. However, we do not think the methods we have described here are a final solution. We describe them here because they are an obvious first step that has been ignored. Efron and Tibshirani (2007) have proposed an approach that includes a statistic that specifically targets gene sets with only a fraction of the genes differentially expressed and a novel permutation approach. Falcon and Gentleman (2006) has developed methodology that takes into account the fact that overlap exists between the different gene sets. These approaches certainly seem promising.

# References


Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003), "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, 18.

Efron, B., and Tibshirani, R. (2007), "On testing the significance of sets of genes," *Ann. Appl. Stat.*, 1(1), 107–129.

Falcon, S., and Gentleman, R. (2006), "Using GOstats to Test Gene Lists for GO Term Association," *Bioinformatics*, 567(1).

Goeman, J. J., Oosting, J., Cleton-Jansen, A. M., Anninga, J. K., and van Houwelingen, H. C. (2005), "Testing association of a pathway with survival using gene expression data," *Bioinformatics*, 21(9), 1950–7.

Goeman, J., van de Geer, S., de Kort, F., and van Houwelingen, H. (2004), "A global test for groups of genes: testing association with a clinical outcome," *Bioinformatics*, 20, 93–99.

Kendziorski, C., Irizarry, R. A., Chen, K.-S., Haag, J. D., and Gould, M. N. (2005), "On the utility of pooling biological samples in microarray experiments," *PNAS*, 102(12), 4252–4257.

Kim, S. Y., and Volsky, D. J. (2005), "PAGE: Parametric Analysis of Gene Set Enrichment," *BMC Bioinformatics*, 6(144).

Lee, H. K., Braynen, W., Keshav, K., and Pavlidis, P. (2005), "ErmineJ: Tool for functional analysis of gene expression data sets," *BMC Bioinformatics*, 6, 269.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003), "PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nat Genet*, 34(3), 267–273.

Pavlidis, P., Lewis, D. P., and Noble, W. S. (2002), "Exploring gene expression data with class scores," *Pac. Symp. Biocomput.*, pp. 474–485.

Pavlidis, P., Qin, J., Arango, V., Mann, J. J., and Sibille, E. (2004), "Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex," *Neurochem Res*, 29, 1213–1222.

Smyth, G. K. (2004), "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments," *Statistical Applications in Genetics and Molecular Biolog 3*, 1.

Storey, J. D. (2002), "A direct approach to false discovery rates," *J Royal Statistical Soc B*, 64.

Storey, J. D., and Tibshirani, R. (2003), "Statistical significance for genomewide studies," *PNAS*, 100(16).

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Pomeroy, A. P. S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005), "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proc Natl Acad Sci*, 102(43), 15545–50.

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

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005), "Discovering statistically significant pathways in expression profiling studies," *Proc Natl Acad Sci*, 102(38), 13544–9.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001), "Significance analysis of microarrays applied to the ionizing radiation response," *PNAS*, 98.