BIOINFORMATICS

Vol. 00 no. 00 2006
Pages 1–7

Analyzing gene expression data in terms of gene sets: methodological issues

Jelle J. Goeman [a*], Peter Bühlmann [b]

[a] Leiden University Medical Center, Dept. of Medical Statistics and Bioinformatics, Postzone S5-P, P.O. Box 9600, 2300 RC Leiden, The Netherlands. [b] Seminar für Statistik, ETH Zurich, CH-8092 Zürich, Switzerland

ABSTRACT

Motivation: Many statistical tests have been proposed in recent years for analyzing gene expression data in terms of gene sets, usually from Gene Ontology. These methods are based on widely different methodological assumptions. Some approaches test differential expression of each gene set against differential expression of the rest of the genes, whereas others test each gene set on its own. Also, some methods are based on a model in which the genes are the sampling units, whereas others treat the subjects as the sampling units. This paper aims to clarify the assumptions behind different approaches and to indicate a preferential methodology of gene set testing.

Results: We identify some crucial assumptions which are needed by the majority of methods. P-values derived from methods that use a model which takes the genes as the sampling unit are easily misinterpreted, as they are based on a statistical model that does not resemble the biological experiment actually performed. Furthermore, because these models are based on a crucial and unrealistic independence assumption between genes, the p-values derived from such methods can be wildly anti-conservative, as a simulation experiment shows. We also argue that methods that competitively test each gene set against the rest of the genes create an unnecessary rift between single gene testing and gene set testing.

Contact: j.j.goeman@lumc.nl

1 INTRODUCTION

A successful microarray experiment typically results in a long list of differentially expressed genes. The gene list is usually not the end point of the analysis; it is the starting point of a complicated process of interpretation, in which the biologist will search for patterns in the differential expression. A list of differentially expressed genes is easier to interpret if the genes exhibit similarity in their functional annotation or chromosomal location.

In recent years many authors have proposed methods to formalize this interpretation process using statistical hypothesis tests. These methods group all genes that are annotated to the same annotation term together into sets and analyze the result of the microarray experiment in terms of these sets. This essentially shifts the level of analysis of the microarray experiment from single genes to sets of related genes. Such an analysis allows biologists to make use of previously accumulated biological knowledge in the analysis and makes a more biology-driven analysis of microarray data possible. The annotation terms are usually obtained from libraries such as Gene Ontology (Ashburner et al., 2000) or KEGG (Ogata et al., 1999). The sets of genes in this type of analysis are always given a priori and are constructed without reference to the data.

A great variety of methods has been proposed for testing differential expression of a gene set with a single test. The most popular method starts from the list of differentially expressed genes and tests whether the gene set is overrepresented in this list, using a test for independence in a 2×2 (contingency) table. This approach has been described with minor variations by many different authors (Hosack et al., 2003; Zeeberg et al., 2003; Al-Shahrour et al., 2004; Boyle et al., 2004; Beissbarth and Speed, 2004; Zhang et al., 2004; Lee et al., 2005; Pehkonen et al., 2005; Yi et al., 2006, among others). See Khatri and Drăghici (2005) for an overview.

Other authors have criticized this approach because it requires a strict cut-off for differential expression of individual genes. As an alternative they have proposed methods that use the whole vector of p-values. Breitling et al. (2004) and Al-Shahrour et al. (2005) use the same 2×2 tables, but tests simultaneously at many cut-off values. Mootha et al. (2003) test whether the ranks of the p-values of the genes in the gene set differ from a uniform distribution, using a weighted Kolmogorov-Smirnov test (see also Subramanian et al., 2005). Pavlidis et al. (2004) use a test based on the geometric mean of the p-values of the genes in the gene set. Barry et al. (2005) provide a general framework for post hoc testing based on p-values or other test statistics per gene.

A very different approach is used by a third group of authors, who do not start from the p-values per gene, but from the raw expression data. Goeman et al. (2004, 2005) test whether subjects with similar gene expression profiles have similar class labels, based on a logistic regression model. Conversely, Mansmann and Meister (2005) test whether subjects with similar class labels have similar expression profiles, based on an ANOVA model. Tomfohr et al. (2005) use a t-test after reducing the gene set to its first principal component.

Criticism of these methods in general has come from Khatri and Drăghici (2005) who pointed out limitations of the annotation databases used. More fundamental criticism has been given by Allison et al. (2006) who questioned the foundations and the validity of some approaches.

This paper addresses the questions raised by Allison et al. (2006), identifying and investigating some fundamental methodological differences that exist between gene set testing methods. We do not want to compare all available methods, or even to give a comprehensive account of all these methods. The aim of this paper is to discuss

*to whom correspondence should be addressed

some important methodological questions that arise when analyzing gene expression data in terms of gene sets. We focus on two methodological issues on which there is a clear disagreement. The first is the definition of the null hypothesis; the second is the calculation of the p-value.

Concerning the definition of the null hypothesis, we make a distinction between *competitive* and *self-contained* tests. A *competitive* test compares differential expression of the gene set to a standard defined by the complement of that gene set. A *self-contained* test, in contrast, compares the gene set to a fixed standard that does not depend on the measurements of genes outside the gene set. The competitive test is most popular: only Goeman et al. (2004, 2005), Mansmann and Meister (2005) and Tomfohr et al. (2005) present self-contained tests.

Concerning the calculation of the p-value, we make a distinction between *gene sampling* methods and *subject sampling* methods. The former bases the calculation of the p-value for the gene set on a distribution in which the gene is the sampling unit, while the latter takes the subject as the sampling unit. In both cases, the sampling units are assumed to be independent and identically distributed. Gene sampling methods are most popular, with only Goeman et al. (2004, 2005), Mansmann and Meister (2005), Tomfohr et al. (2005) and Mootha et al. (2003) using subject-sampling.

Because the focus of this paper is not on details of specific methods but on methodological issues, we do not compare published methods (see Díaz-Uriarte, 2005; Manoli et al., 2006), but we specifically construct methods that differ only with respect to the issue at hand. This is done on the basis of the 2×2 table overrepresentation methods, because these are most popular and easy to understand. The 2×2 table methods are competitive and gene-sampling methods. For purposes of comparison we will construct methods that are similar to the 2×2 table methods, except that they are self-contained, subject-sampling, or both. Section 2 describes the 2×2 table methods. Section 3 then studies competitive versus self-contained testing. Section 4 compares gene-sampling and subject-sampling.

2 2×2 TABLE METHODS

The general idea of 2×2 table methods is to search for an overrepresentation of the gene set among the differentially expressed genes, or, equivalently, an overrepresentation of differentially expressed genes among the genes in the gene set. There are minor differences in the methods proposed by various authors (Khatri and Drăghici, 2005), but we give a general description here.

First, a measure of differential expression is calculated for each gene. This is usually a p-value from a t-test or some other statistical test for differential expression of single genes. It can also be a simple measure such as fold change (Breitling et al., 2004). Next, a cut-off is found to separate differentially expressed from non-differentially expressed genes. This cut-off can be simple, such as the 100 genes with smallest p-values, or it can be more sophisticated, e.g. based on a multiple testing criterion such as Bonferroni or the False Discovery Rate (Benjamini and Hochberg, 1995).

Given the list of differentially expressed genes and the list of genes in the gene set, it is possible to fill a 2×2 table as indicated in Table 1. The table simply counts the number of genes on the microarray with every possible combination of the attributes "differentially expressed (yes/no)" and "in the gene set (yes/no)".

Table 1. A 2×2 table for assessing overrepresentation.

	diff. expr. gene	*non-diff. expr. gene*	*total*
in gene set	m_{GD}	m_{GD^c}	m_G
not in gene set	m_{G^cD}	$m_{G^cD^c}$	m_{G^c}
total	m_D	m_{D^c}	m

The p-value for overrepresentation of the gene set among the differentially expressed genes is subsequently calculated using a test for independence in the 2×2 table of Table 1. A number of different tests have been proposed for testing this independence, including the χ^2 test, the hypergeometric test (Fisher's exact test) and the binomial z-test for proportions. Each of these tests is equivalent to a procedure that finds the null distribution of a test statistic by randomly reassigning genes to the labels for being in the gene set and for being differentially expressed. The differences are in the choice of the test statistic and whether the random reassignment is done with or without replacement. These differences are not fundamental and tend to be unimportant in practice (Khatri and Drăghici, 2005). In this paper we use the hypergeometric test, which takes the size of the overlap between the gene set and the list of differentially expressed genes as the test statistic, and reassigns labels without replacement (i.e. it keeps the marginal totals in the table constant).

3 COMPETITIVE VS. SELF-CONTAINED TESTS

The main difference between competitive and self-contained tests lies in the formulation of the null hypothesis. Loosely, the null hypotheses can be formulated as follows. Let G be the gene set of interest and G^c its complement, then the competitive null hypothesis is

> H_0^{comp}: *The genes in G are at most as often differentially expressed as the genes in G^c*,

while the self-contained null hypothesis is

> H_0^{self}: *No genes in G are differentially expressed.*

Note that these hypotheses refer to the number of truly differentially expressed genes, not to the number of genes called differentially expressed, even though the empirical numbers of genes called differentially expressed will be used to test them.

The hypothesis H_0^{self} is almost invariably more restrictive than H_0^{comp}. The two null hypotheses are equivalent only in the case that none of the genes in G^c are truly differentially expressed, which is a highly unrealistic situation unless G^c is very small. In general, truth of H_0^{self} implies truth of H_0^{comp}.

In this section we discuss the merits of the two formulations of the null hypotheses using the example of the 2×2 table methods. To avoid the complicating issue of dependence of genes, which will be covered in detail in Section 4, we assume for simplicity that the p-values of all genes are independent.

The 2×2 table method tests the competitive null by comparing the proportions of genes called differentially expressed in G with the corresponding proportion in G^c, relying on the reasonable assumption that a larger proportion of truly differentially expressed genes in G will result in a higher probability that a randomly chosen gene in G will be called differentially expressed.

We can construct a self-contained counterpart of the 2×2 table method. This method tests H_0^{self} with a binomial test based on a

test statistic m_α, which is the number of genes in G with p-values smaller than α. Under H_0^{self} and assuming independence of genes, m_α should have a binomial $\mathcal{B}(m_G, \alpha)$ distribution, where m_G is the number of genes in G. Note that this test, like its null hypothesis, is self-contained in the sense that it does not use any information on genes in G^c. The binomial test for the self-contained null hypothesis in a multiple testing situation was first proposed by Tukey under the name of *higher criticism*. It has recently been developed into a more sophisticated method by Donoho and Jin (2004).

It is easy to compare the two procedures based on the two different null hypotheses. There are a few remarks to be made. Most of these have to do with the competitive nature of the competitive null, which pits each gene set against its complement in what Allison et al. (2006) called a "zero-sum game" (see also Damian and Gorfine, 2004).

The first remark is about power. A test based on the self-contained H_0^{self} will almost invariably have more power than a test based on the competitive H_0^{comp}. This follows immediately from the fact that the self-contained null is more restrictive than the competitive null, as noted above. As a consequence, a self-contained test will almost always reject the null hypothesis for more gene sets than a competitive null. This is especially the case in a data set in which there are many differentially expressed genes. In the competitive set-up the significance of the gene set G is "penalized" for the significance of the gene set G^c. Relative to the self-contained test, the competitive type of test can be said to voluntarily relinquish some power in order to make a stronger statement.

A second remark concerns the relationship between single gene testing and gene set testing. It can easily be seen that for a gene set containing only a single gene, Tukey's higher criticism will simply call the gene set significant whenever the single gene's p-value is below alpha. The self-contained test is therefore an immediate generalization of single gene testing to gene sets, in the sense that the two procedures are completely equivalent for singleton gene sets. This is a desirable property, which does not hold for the competitive test. On the contrary, the competitive test treats a singleton gene set very differently from a single gene, especially when there are many differentially expressed genes in G^c.

Thirdly, it is interesting to look at the set of all genes on the chip. This gene set can not be tested in a competitive way, simply because there is no complement to test the gene set against. In contrast, the set of all genes can be a very useful gene set to test with a self-contained test. It tests the global null hypothesis that there are no differentially expressed genes. Rejecting this null can be an interesting preliminary data quality check, as a failure to reject this null leaves little hope that anything can be found in the data. A self-contained test for the set of all genes can also have a useful prediction interpretation (Goeman et al., 2004).

The main objection that can be made against self-contained testing, on the other hand, is that it can sometimes be too powerful: in a situation in which there are many differentially expressed genes almost all gene sets may be called significant. Certainly, a direct application of Tukey's higher criticism to gene set testing in microarray data would lead to very large power. However, this overly large power is for a large part due to the strong independence assumption of the p-values that this procedure requires. This independence assumption is the subject of Section 4.

In the end, the issue of using a competitive or a self-contained test should depend on the biological interpretation of the null hypothesis. The self-contained null hypothesis that no gene in the gene set is differentially expressed always has a clear biological meaning. At the same time, it may not always be biologically interesting, e.g. when comparing cancer versus normal tissue: in such cases we may not expect the self-contained null hypothesis to be true for any gene set. The competitive null hypothesis on the other hand, although sometimes more relevant, is much more difficult to test because its definition is closely tied to a gene sampling model with independent genes. The gene sampling model is the subject of Section 4.

4 GENE VS. SUBJECT SAMPLING

The 2×2 table method and related methods are based on a model which uses the gene as the sampling unit. This approach is very different from the usual statistical setup, in which the subjects are taken as the sampling units (Klebanov and Yakovlev, 2006). It is instructive to compare the stochastic models. This comparison shows enormous differences not only in the assumptions underlying the respective models, but also in the interpretation of the resulting p-values.

4.1 Gene-sampling and subject sampling models

Classical statistical tests are based on a experimental design that samples subjects. Each subject gets the same fixed set of (gene expression) measurements. In the usual supervised setting the sample is assumed to consist of n independent realizations (for the n subjects) of

$$(X_1, Y_1), \ldots, (X_n, Y_n), \tag{1}$$

where X_i is the m-dimensional vector of the expression measurements of the i-th subject, and Y_i the corresponding response variable (usually a class label, e.g. treatment v.s. control). It is assumed that the n measurements of the different subjects are independent and identically distributed, but that different gene expression measurements within the same subject may be correlated. A replication of the experiment under a subject sampling model would involve a new sample of subjects, which are subjected to the same set of measurements, i.e. the experiment is repeated for new subjects, but with the same genes.

The model behind the 2×2 table methods is based on an urn model which turns the classical statistical setup around. The 2×2 table is filled with a sample of genes, each of which is drawn at random from a big urn of genes. Each gene is subjected to the same fixed set of two measurements. The first measurement (A) indicates whether the gene is part of the gene set or not; the second measurement (B) indicates whether the gene is in the list of differentially expressed genes, based on the p-value of that gene in the specific experiment performed. The sample is assumed to consist of m observations (for the m genes) of

$$(A_1, B_1), \ldots, (A_m, B_m). \tag{2}$$

The test that is subsequently performed assumes that the m measurements of the m different genes are all independent and identically distributed.

Essentially, the gene sampling urn model completely reverses the roles of samples and genes relative to the classical statistical setup. Instead of a sample of subjects which are given a fixed set of

measurements, we have a sample of measurements coming from a fixed set of subjects. A replication of the experiment under the urn model would therefore involve taking a new sample of genes and subjecting these genes to the same measurements, i.e. repeating the experiment for new genes and *the same subjects*.

Note that the sample size is very different in the two setups. The subject sampling approach has sample size equal to the number of subjects n, while the gene sampling approach uses a sample size equal to the number of genes m.

4.2 A subject sampling 2×2 table method

For comparison we construct a subject sampling analogue to the gene-sampling 2×2 table method. In general, this can be done by calculating the p-value by subject permutation instead of using the hypergeometric distribution, as proposed by Barry et al. (2005) (and by Mootha et al., 2003, for GSEA). This calculates a non-parametric permutation null distribution for the null hypothesis that X and Y are independent, under the assumption that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed. It is well known that for any test statistic calculated from $(X_1, Y_1), \ldots, (X_n, Y_n)$, its null distribution under these assumptions can be non-parametrically computed by the distribution of the same test-statistic based on $(X_1, Y_{\pi(1)}), \ldots, (X_n, Y_{\pi(n)})$, where the distribution is generated from all (or many randomly generated) permutations $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$.

The subject sampling analogue of the hypergeometric test is a subject permutation test based on the same test statistic that the hypergeometric distribution uses, namely the overlap m_{GD} between the set of significant genes and the gene set (see Table 1). Suppose that the data set has m_D differentially expressed genes, of which m_{GD} are in the gene set of interest. The algorithm is described in table 2.

Table 2. A subject sampling alternative to the 2×2 table method.

1. Permute the sample labels $Y_1, \ldots, Y_n$ N times.
2. For each permutation, recalculate the p-values for all genes based on the permuted data.
3. For each vector of permutation p-values, count how many genes in the gene set are among the m_D genes with the smallest p-values. Store these counts as $k_1, \ldots, k_N$.
4. Find the p-value of the gene set as the proportion of $k_1, \ldots, k_N$ which are greater than m_{GD}.

It should be noted that switching to subject permutation also changes the null hypothesis that is tested. The subject permutation null distribution is the complete null distribution that no gene in G and G^c is differentially expressed, which is a very specific case of the competitive null hypothesis, which is, in fact, also a self-contained null hypothesis. The alpha level of the test of table 2 is guaranteed for the complete null hypothesis, but is unclear for the competitive null hypothesis in general.

In a sense, the algorithm in table 2 is a hybrid form between gene sampling and subject sampling, as well as between competitive and self-contained testing. The test statistic is motivated by a gene sampling model, but the p-value is calculated using subject sampling. The test statistic is competitive in the sense that it involves the genes in G^c, but the actual null hypothesis tested is the complete null hypothesis, which is both competitive and self-contained. A completely self-contained and subject sampling alternative to the method of table 2 is given in Section 5.

Note also that permutation tests are not adequate in cases where the subjects were not sampled according to the simple sampling scheme given in (1), e.g. in time series or when covariates are present. This means that a subject-sampling equivalent of a specific 2×2 table method may not always exist.

4.3 Interpretation of the p-value

The interpretation of a p-value greatly depends on the sampling scheme on which the test is based. Because the gene sampling scheme is the mirror image of the subject sampling scheme, we will first review the interpretation of the classical subject sampling p-value and derive the interpretation of the gene sampling p-value by analogy.

The meaning of a p-value is related to hypothetical replications of the experiment performed. By definition, if the null hypothesis is true, no more than a fraction α of the replications of an experiment will yield a p-value smaller than α. This property of the p-value is the basis of all statistical inference based on it. However, as it is a statement about replications of the experiment, its meaning and interpretation are closely tied to the sampling scheme implied in the model.

In the classical subject sampling setup, replications of the experiment involve taking a new sample of subjects and measuring these subjects on the same variables. The interpretation of the p-value of a subject sampling method therefore relates to true biological replications of the experiment to new subjects. A significant p-value excludes random variation at the subject level as an explanation for the associations found, and therefore gives confidence that the same associations will be found for a new sample of subjects. On the other hand, the subject sampling p-value does not make any statement about replications to new genes: if a Gene Ontology term is represented on a chip by only a single gene, a very significant p-value for that singleton gene set does not say anything about other genes from the same Gene Ontology term.

In the gene sampling setup, the roles of genes and samples are reversed in the interpretation of the p-value. The interpretation of the p-value relates to replications of the urn experiment, which would take a new sample of genes and calculate their p-values for the same subjects. A significant p-value gives confidence that a similar association between the variables "membership of the gene set" and "being differentially expressed" will be found with these subjects on a new array with different genes. However, the gene sampling p-value does not say anything about biological replications of the experiment using different subjects.

This interpretation of the p-value of a gene sampling method can explain the radical claim made by Breitling et al. (2004). They proposed iGA, a variant of the 2×2 table method that bases the cut-off for differential expression of genes on fold change instead of p-value, and simultaneously looks at all possible choices of the cut-off. They claim that iGA even produces valid p-values when used on a single two-color array. In their abstract they write:

> "In the extreme, iGA can even produce statistically meaningful results without any experimental replication."

This statement is valid only in the context of the urn model. A significant p-value of iGA only indicates that the specific pair of subjects whose gene expression is measured on the two-color array tends to have consistently high fold changes for the genes in the gene set. However, this p-value does not say anything about the next pair of subjects.

Just as in Section 3, it is instructive to look at the relationship between gene set testing and single gene testing by considering a single gene as a singleton gene set. If the statement made about iGA were true, it would suddenly be possible to test for differential expression of single genes without any experimental replication by viewing the genes as singleton gene sets. This is against all common sense. If single gene tests are always based on a subject-sampling model, there is no real reason to base gene set tests on a widely different model.

However, the most important problem with the gene sampling urn model is that it does not mimic the actual biological experiment performed. A biological replication of the experiment always takes a new sample of subjects, not a new sample of genes. Biologists expect a p-value to measure the strength of the evidence based on the biological experiment actually performed and will interpret it in this context. Calculating a p-value based on a gene sampling urn model can too easily lead to wildly misleading interpretations, and should be discouraged in the strongest terms.

A related misleading aspect of the urn model is the apparent sample size, which is equal to the number of genes m in that model. This is not the same as the sample size of the biological experiment, which is equal to the number of subjects n. The urn model can therefore be seen as a model that artificially inflates the sample size, resulting in inflated power. This increase in power is not real, as it depends on a highly unrealistic assumption of independence between genes. This is the subject of Section 4.4.

4.4 The independence assumption

The gene-sampling model (2), on which all tests used in the 2×2 table methods are based, relies on the assumption that the observations (A_i, B_i) for each gene are independent and identically distributed. This is a highly unrealistic assumption for gene expression data.

It is well known that strong correlations between genes occur frequently in microarray data and that complete independence between any two gene expression measurements is rare, if only due to the presence of array effects. Correlations are especially expected between functionally related genes. As gene sets to be tested are usually chosen on the basis of functional annotation, it should be expected that many of the genes in a tested gene set are correlated.

Such correlations are problematic for tests used in 2×2 table methods. Correlations between gene expression measurements of genes tends to result in positive correlations between their (two-tailed) p-values, which in turn causes their indicators B_i of differential expression to be correlated. In turn, this results in overdispersion (see for example McCullagh and Nelder, 1989, Ch. 4.5) for the number of genes called differentially expressed. If p-values are positively correlated, the true null distribution of the hypergeometric test is not hypergeometric, but has much heavier tails. This can be understood by considering the probability that two genes are both called differentially expressed. This probability is much smaller when the genes are independent than when the same genes have positively correlated p-values. As a consequence, the use of the hypergeometric test is anti-conservative; it may greatly understate the true p-values if the genes in the gene set are not independent.

The anti-conservatism of the hypergeometric test may also be understood in a different way. The null hypothesis of the hypergeometric test assumes that the genes in the gene set are not unusually often differentially expressed, but also that the genes in the gene set are independent. Although designed to detect the first kind of deviation from the null hypothesis, the test also has power to detect the second. A significant result from the hypergeometric test may therefore indicate unusual differential expression of the genes in the gene set, but it may also simply indicate that the genes are dependent.

To quantify the anti-conservativeness of the 2×2 table method under dependence of genes, we conducted a small simulation experiment, simulating data under the null hypotheses with various degrees of dependence between genes. The simulation setup was as follows. We varied a correlation coefficient ρ from 0 to 1 in steps of 0.1. For each value of ρ we generated 5,000 independent data sets. Each data set had 10,000 genes for 20 subjects. The genes were divided into 100 gene sets of 100 genes each. Gene expression measurements were generated independently for each subject according to a multivariate normal distribution which had mean 0 and variance 1 for each gene, and for which the correlation between any two genes in the same gene set was taken equal to ρ, while the correlation between genes of different gene sets was taken as 0. The 20 subjects were divided into two groups of 10 each. The distribution of gene expression was independent of the group indicator, so that none of the genes was in reality differentially expressed.

On each data set, we performed a two-sided student t-test for each gene under the (valid) assumption of equal variance. This was followed up by a 2×2 table analysis for all gene sets based on the hypergeometric test and on a cut-off for significance of each t-test at 0.05. Together, this simulation setup gave a collection of 500,000 gene set p-values for each value of ρ, all generated under the null hypothesis so that there is no difference in differential expression between gene sets. We counted the number of rejections under various nominal α-levels of the hypergeometric test. The results are given in table 3.

From the table we note that the hypergeometric test keeps the α-level for uncorrelated genes, as expected. For $\rho = 0$ the test is even somewhat conservative due to the discrete nature of the test. Despite this conservatism, the test already becomes anti-conservative for very moderate correlations of 0.2 to 0.3, depending on the α-level. The anti-conservatism can grow to rejection rates up to 50,000 times the nominal level for some higher correlations and small α-levels. It is most pronounced in the tail of the distribution and for high correlations. The case $\rho = 1$ is an unrealistic but interesting extreme case in which all 100 genes in the gene set have the same expression, so that either all or none are called differentially expressed. This results in hypergeometric p-values of either (essentially) zero or exactly 1, the former occurring with probability 0.05 (the alpha-level of the original t-test), the latter with probability 0.95. Note that we focus especially on the extreme tail of the distribution in table 3, because that is the important part when correcting for multiple testing. Similar anti-conservatism was found by Breslin et al. (2004), who found that gene permutation gave much smaller p-values than subject permutation in several microarray data sets.

Table 3. Fraction rejected for the 2×2 table method (standard hypergeometric test) for various nominal levels of α, and for various degrees of correlation among the genes in each gene set. The table is based on 500,000 simulated gene sets. All simulations are under the null hypothesis.

correlation	*nominal α-level*						
ρ	0.1	0.05	0.01	0.001	0.0001	0.00001	0.000001
0	0.067	0.032	0.0061	0.00058	0.000036	0.000006	0.000000
0.1	0.068	0.033	0.0064	0.00061	0.000060	0.000006	0.000000
0.2	0.074	0.038	0.0088	0.0013	0.00023	0.000040	0.000012
0.3	0.094	0.058	0.022	0.0070	0.0028	0.0012	0.00058
0.4	0.12	0.088	0.047	0.023	0.013	0.0080	0.0050
0.5	0.15	0.12	0.078	0.049	0.033	0.024	0.018
0.6	0.17	0.14	0.10	0.075	0.057	0.046	0.037
0.7	0.17	0.15	0.12	0.097	0.080	0.067	0.058
0.8	0.16	0.15	0.13	0.11	0.094	0.084	0.075
0.9	0.14	0.13	0.12	0.10	0.095	0.088	0.083
1	0.050	0.050	0.050	0.050	0.050	0.050	0.050

5 ADAPTING EXISTING METHODS

In the previous sections we have studied the 2×2 table method which tests a competitive null hypothesis on the basis of a gene-sampling model. Using the 2×2 table methods as an example, we have identified some important problems in competitive methods as well as in gene sampling methods. On the basis of these we recommend to use methods which test a self-contained null hypothesis and base the calculation of the p-value on a subject-sampling model. There are two options for this.

The first option is to use one of the proposed gene set testing methods that are based on classical statistical models, and which by construction already test a self-contained null hypothesis and calculate to p-value based on a subject sampling model that does not involve an independence assumption of genes. Such methods have been proposed by Goeman et al. (2004, 2005) based on the locally most powerful test of Goeman et al. (2006), by Mansmann and Meister (2005) based on an ANOVA model and by Tomfohr et al. (2005), based on principal components. These methods do not proceed in a post hoc fashion from the single gene p-values, but model the gene expression data directly.

The second option is to adapt an existing post hoc method to test a self-contained null hypothesis and to calculate the p-value using subject sampling. Each of these adaptations has already been demonstrated separately for the 2×2 table method. We can combine the two adaptations into a combined method, which is a subject-permutation version of Tukey's higher criticism. The algorithm is given in table 4. Fix some α beforehand and let m_{GD} be the number of genes in the gene set that have p-value below α.

Other methods may be similarly adapted to a self-contained null hypothesis and to subject sampling. An interesting method in the context of adaptation is GSEA (Mootha et al., 2003; Subramanian et al., 2005). This method uses a Kolmogorov-Smirnov test statistic to test whether the ranks of the p-values of the genes in the gene set can be a sample from a uniform distribution. To calculate the p-value they use subject permutation. This method is interesting because the Kolmogorov-Smirnov test statistic is motivated by a gene sampling model, whereas a subject sampling model is used for calculating the p-value. In this sense the method is similar to the hybrid method described in Section 4.2. It is interesting to see that GSEA is sometimes found to have low power, as can be seen from the GSEA user guide, which recommends 0.25 as the most suitable FDR threshold (www.broad.mit.edu/gsea). This low power may be due to the fact that the model and null hypothesis used to motivate the test statistic are different from the model and null hypothesis that are used when calculating the p-value. GSEA can easily be transformed to a self-contained test by calculating the Kolmogorov-Smirnov statistic on the basis of the p-values themselves, instead of on their ranks.

Table 4. Tukey's non-competitive subject sampling alternative to the 2×2 table method.

1. Permute the sample labels $Y_1, \ldots, Y_n$ N times.
2. For each permutation, recalculate the p-values for the genes in the gene set.
3. For each vector of permutation p-values, count how many genes in the gene set have p-value below α. Store these counts as $k_1, \ldots, k_N$.
4. Find the p-value of the gene set as the proportion of $k_1, \ldots, k_N$ which are greater than m_{GD}.

The method of Pavlidis et al. (2004) takes the arithmetic mean of the p-values as a test statistic per gene set and tests this by using gene label permutation. Their method is gene-sampling and uses a competitive null, but it may easily be transformed to a self-contained subject-sampling test by switching from gene permutation to subject permutation.

6 DISCUSSION

This paper has investigated methodological issues in methods that test for differential expression of gene sets. It has revealed some methodological aspects of popular methods that are inefficient or even incorrect from a statistical point of view. Although this paper looked specifically at supervised methods for gene set testing, similar problems occur in unsupervised settings, for example when using a hypergeometric test for testing overrepresentation of a GO term in a cluster of genes from a cluster analysis.

We have given strong arguments against models that take the genes as the independent sampling unit and therefore implicitly or explicitly assume that the genes are independent. We have argued

that because the statistical model underlying these p-values turns the actual experimental design on its head, the interpretation of the p-value changes radically from the traditional statistical one. This can easily lead to misunderstandings and false conclusions. Furthermore, we have shown that such tests do not give valid p-values when the genes on the microarray are correlated. The p-values may easily be falsely significant when the genes in the gene set are correlated, even when none of the genes is truly differentially expressed. We strongly recommend against the use of gene sampling models in gene set testing.

The issue of self-contained testing versus competitive testing is closely connected to the issue of gene versus subject sampling. A competitive null hypothesis is natural and easy to formulate in a gene sampling model, just as a self-contained null hypothesis is natural in a subject sampling model.

Methods for testing a self-contained null hypothesis are all based on a subject sampling model (Goeman et al., 2004, 2005; Mansmann and Meister, 2005; Tomfohr et al., 2005). The classical statistical combination of a subject sampling model and a self-contained null hypothesis gives the advantage of valid p-values, easy interpretability and a close relation to single gene testing, as single gene testing is also based on a self-contained null hypothesis and a subject sampling model.

Methods for testing a competitive null hypothesis are usually based on a gene sampling model and suffer from the same validity problems as described for 2×2 table methods above. A few methods such a GSEA (Mootha et al., 2003) and the method of table 2 (see also Barry et al., 2005) are hybrid in the sense that they motivate their test statistic on the basis of a gene sampling model, but calculate their p-value in a subject sampling manner. The discrepancy between the two models makes the statistical properties of the test unclear and its interpretation difficult. These problems are unavoidable, as the definition of the competitive null hypothesis is intimately tied to the gene sampling model, whereas valid p-values are easily available for subject sampling only.

REFERENCES

Al-Shahrour, F., R. Díaz-Uriarte, and J. Dopazo (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics 20*(4), 578–580.

Al-Shahrour, F., R. Díaz-Uriarte, and J. Dopazo (2005). Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics 21*(13), 2988–2993.

Allison, D. B., X. Q. Cui, G. P. Page, and M. Sabripour (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics 7*(1), 55–65.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics 25*(1), 25–29.

Barry, W. T., A. B. Nobel, and F. A. Wright (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics 21*(9), 1943–1949.

Beissbarth, T. and T. P. Speed (2004). GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics 20*(9), 1464–1465.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological 57*(1), 289–300.

Boyle, E. I., S. A. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock (2004). GO-TermFinder: open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics 20*(18), 3710–3715.

Breitling, R., A. Amtmann, and P. Herzyk (2004). Iterative group analysis (iGA): A simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics 5*, 34.

Breslin, T., P. Eden, and M. Krogh (2004). Comparing functional annotation analyses with catmap. *BMC Bioinformatics 5*, 193.

Damian, D. and M. Gorfine (2004). Statistical concerns about the GSEA procedure. *Nature Genetics 36*(7), 663–663.

Díaz-Uriarte, R. (2005). Supervised methods with genomic data: a review and cautionary review. In F. Azuaje and J. Dopazo (Eds.), *Data Analysis and Visualization in Genomics and Proteomics*, pp. 193–214. Chichester: Wiley.

Donoho, D. and J. S. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics 32*(3), 962–994.

Goeman, J. J., J. Oosting, A. M. Cleton-Jansen, J. K. Anninga, and J. C. van Houwelingen (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics 21*(9), 1950–1957.

Goeman, J. J., S. A. van de Geer, F. de Kort, and J. C. van Houwelingen (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics 20*(1), 93–99.

Goeman, J. J., S. A. van de Geer, and J. C. van Houwelingen (2006). Testing against a high-dimensional alternative. *Journal of the Royal Statistical Society Series B-Statistical Methodology 68*(3), 477–493.

Hosack, D. A., G. Dennis, B. T. Sherman, H. C. Lane, and R. A. Lempicki (2003). Identifying biological themes within lists of genes with EASE. *Genome Biology 4*(10), R70.

Khatri, P. and S. Drăghici (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics 21*(18), 3587–3595.

Klebanov, L. and A. Yakovlev (2006). Treating expression levels of different genes as a sample in microarray data analysis: is it worth a risk? *Statistical Applications in Genetics and Molecular Biology 5*(1), article 9.

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

Manoli, T., N. Gretz, H. J. Grone, M. Kenzelmann, R. Eils, and B. Brors (2006). Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics 22*(20), 2500–2506.

Mansmann, U. and R. Meister (2005). Testing differential gene expression in functional groups: Goeman's global test versus an ANCOVA approach. *Methods of Information in Medicine 44*(3), 449–453.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models* (2nd ed.). Boca Raton: Chapman & Hall.

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

Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research 27*(1), 29–34.

Pavlidis, P., J. Qin, V. Arango, J. J. Mann, and E. Sibille (2004). Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical Research 29*(6), 1213–1222.

Pehkonen, P., G. Wong, and P. Toronen (2005). Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC Bioinformatics 6*, 162.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America 102*(43), 15545–15550.

Tomfohr, J., J. Lu, and T. B. Kepler (2005). Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics 6*, 225.

Yi, M., J. D. Horton, J. C. Cohen, H. H. Hobbs, and R. M. Stephens (2006). Wholepathwayscope: a comprehensive pathway-based analysis tool for high-throughput data. *BMC Bioinformatics 7*, 30.

Zeeberg, B. R., W. M. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, and J. N. Weinstein (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology 4*(4), R28.

Zhang, B., D. Schmoyer, S. Kirov, and J. Snoddy (2004). GO Tree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics 5*, 16.