BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 18 2009, pages 2348–2354
doi:10.1093/bioinformatics/btp406

Gene expression

Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets

Galina V. Glazko[1,*] and Frank Emmert-Streib[2]

[1]Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642, USA and [2]Computational Biology and Machine Learning, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, 97 Lisburn Road, Belfast BT9 7BL, UK

Received on January 30, 2009; revised on April 2, 2009; accepted on June 29, 2009
Advance Access publication July 2, 2009
Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Recently, many univariate and several multivariate approaches have been suggested for testing differential expression of gene sets between different phenotypes. However, despite a wealth of literature studying their performance on simulated and real biological data, still there is a need to quantify their relative performance when they are testing different null hypotheses.

Results: In this article, we compare the performance of univariate and multivariate tests on both simulated and biological data. In the simulation study we demonstrate that high correlations equally affect the power of both, univariate as well as multivariate tests. In addition, for most of them the power is similarly affected by the dimensionality of the gene set and by the percentage of genes in the set, for which expression is changing between two phenotypes. The application of different test statistics to biological data reveals that three statistics (sum of squared t-tests, Hotelling's T^2, N-statistic), testing different null hypotheses, find some common but also some complementing differentially expressed gene sets under specific settings. This demonstrates that due to complementing null hypotheses each test projects on different aspects of the data and for the analysis of biological data it is beneficial to use all three tests simultaneously instead of focusing exclusively on just one.

Contact: Galina_Glazko@urmc.rochester.edu

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

1 INTRODUCTION

In recent years, there has been a considerable shift in attention from single components of molecular biological systems towards studies focusing on functionally related compartments. The reason for this change can be acknowledged, at least partly, to the fact that today's systems perspective (Kitano, 2001; Palsson, 2006) is generally considered as very beneficial to elucidate the collective functioning of biological processes and even of whole cells. In this context, it is no surprise that this reflects also in recent developments regarding the analysis of gene expression data (Emmert-Streib and Dehmer, 2008) and more specifically endeavors to detect differentially expressed pathways (Barry *et al*., 2008; Emmert-Streib, 2007; Mootha *et al*., 2003). The analysis of pathways (gene ontologies, or pre-selected gene sets) that are significantly differentially expressed between two phenotypes is intuitively appealing and there are two known reasons for this. First, by arranging genes into pathways the dimensionality of the data is reduced and as a consequence the number of statistical hypotheses to test. Second, the statement 'a gene is differentially expressed between two phenotypes' has less explanatory power compared to the statement 'a pathway is differentially expressed between two phenotypes'. However, the idea to look for differentially expressed pathways (gene sets in what follows) appeared with a different reasoning in mind. There is a general belief that in metabolic diseases changes in gene expression are moderate and cannot be detected for individual genes. For example, after correction for multiple tests there were no differentially expressed genes between type II diabetes positive and negative patients (Mootha *et al*., 2003). In contrast, the search for differentially expressed gene sets identified a set of genes involved in oxidative phosphorylation as coordinately decreased in human diabetic muscle (Mootha *et al*., 2003). In the latter work, Mootha and colleagues described the first algorithm ('Gene Set Enrichment Analysis', GSEA) focused on the expression changes in a set of genes as opposed to changes in the expression of individual genes. Since that time many approaches for the analysis of gene sets have been suggested (Kim and Volsky, 2005; Nettleton *et al*., 2008; Tomfohr *et al*., 2005) and their number is still growing (see Ackermann and Strimmer, 2009 for a review). The major difference between them was formulated by Goeman and Buhlmann (2007) in terms of the scope of the comparisons of these approaches. *Competitive* tests compare a gene set against the rest of all sets and *self-contained* tests answer the question whether two gene sets are differentially expressed between different phenotypes. In what follows we concentrate on the *self-contained* tests only [see Goeman and Buhlmann (2007) for further discussion]. Self-contained tests, in turn, are different in terms of whether they are multivariate and account for interdependencies among genes (e.g. Hotelling's T^2 test: Kong *et al*., 2006; Lu *et al*., 2005; Xiong, 2006; GlobalANCOVA: Hummel *et al*., 2008; $N-$statistic: Klebanov *et al*., 2007), or disregard existing complex correlation structure in a gene set and consider gene-level statistics only (e.g. weighted sum of t-tests: Tian *et al*., 2005; median-based or sign-tests: Jiang and Gentleman, 2007). Furthermore, for gene-level statistics a transformation of the test statistics is frequently used, to account for the presence of up- and down-regulated genes in

*To whom correspondence should be addressed.

 © The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from https://academic.oup.com/bioinformatics/article/25/18/2348/196468 by guest on 10 April 2024

a gene set (Ackermann and Strimmer, 2009). More importantly, for univariate and multivariate self-contained tests the underlying statistical hypotheses are different. For example, Hotelling's T^2 tests the equality of two multivariate mean vectors while N-statistic tests the equality of two multivariate distributions. A combination of univariate statistics (either transformed or not) studies whether the aggregate gene-level test score differentiates between two phenotypes (Jiang and Gentleman, 2007). We want to emphasize that due to these complementing null hypotheses each test projects on different aspects of the data.

To get the most out of the many tests available one needs to know their relative power in different settings and account for different hypotheses they test. In this article, we compare the performance of univariate and multivariate tests on simulated and biological data with three questions in mind. First, not all genes in a gene set are expected to change their expressions between different phenotypes. The percent of genes changing their expression in a gene set, in the way that the entire gene set is called differentially expressed ('detection call'), is an important but currently unknown characteristic of a test performance. Second, genes in a gene set are functionally related and have complex correlation structure. Multivariate tests might have better power because they account for interdependences among genes considering the joint distribution of gene expression levels, in contrast to univariate tests, which test differences in the marginal distributions, but this hypothesis requires confirmation. The third question is an implication of the second: one might expect that because univariate and multivariate statistics test different null hypotheses that for real biological data they may result in completely different gene sets. There is a reason for concern here, because for example the application of Principal Component Analysis and gene-level tests resulted in a similar scenario (Jiang and Gentleman, 2007). In this article we answer the first two questions on simulated data, mimicking the stated conditions and study in detail the third one on two biological data sets.

In our analysis we compare five statistical tests, four tests representing popular choices in testing whether two gene sets are differentially expressed between different phenotypes (though testing different statistical hypotheses) and one test which has never been used in the context of gene expression sets before. We include two multivariate tests, Hotelling's T^2 test (Lu *et al*., 2005), and N-statistic (Klebanov *et al*., 2007), also known as non-parametric Cramer test (Baringhaus and Franz, 2004) and two univariate tests (different transformations of the t-statistic, see below). A fifth test, the multivariate Dempster's T_1 (Dempster, 1958) was recently considered by Srivastava and Du (2008) in the context of power study (see Srivastava and Du, 2008 for further details). For biological data we compare the performance of these tests with other popular approaches (Jiang and Gentleman, 2007; Liu *et al*., 2007).

2 METHODS

Consider a pair of biological conditions, such as 'case' versus 'control' or 'treated' versus 'untreated'. Suppose there are n_1 samples of measurements of p genes for the first, and n_2 samples of measurement of p genes for the second conditions. Let the two p-dimensional random vectors of measurements $X_1, \ldots, X_{n1}$ and $Y_1, \ldots, Y_{n2}$ be independent and identically distributed with the distribution functions F, G, mean vectors $\bar{X}$, $\bar{Y}$ and $p \times p$ covariance matrices S_x, S_y. We consider the problem of testing the hypothesis H: $F = G$ against a fixed alternative $F \neq G$.

2.1 Test statistics

2.1.1 Hotelling's T^2. Hotelling's T^2 does not test the hypothesis H: $F = G$. If F and G are multivariate normal distributions with common covariance matrix, the Hotelling's T^2 tests the simplified hypothesis $\bar{X} = \bar{Y}$. The corresponding test statistic is

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{X} - \bar{Y})S^{-1}(\bar{X} - \bar{Y})^T \quad (1)$$

where S is the pooled variance–covariance matrix of the measurements. The problem, however, is that if p is larger than $n_1 + n_2 - 2$ the covariance matrix becomes singular and cannot be inverted. In gene expression studies the number of samples is always much less than the number of observation and to calculate the inverse one needs to use additional steps. One obvious modification is to use a generalized matrix inverse instead of S^{-1}, and another one is the dimensionality reduction (e.g. Kong *et al*., 2006). Yet another possibility is to use the shrinkage estimator by Schafer and Strimmer (2005). We did not find significant differences between results obtained using the generalized matrix inverse (as implemented in Venables and Ripley, 1999) and the shrinkage estimator; in what follows the generalized inverse is used.

2.1.2 Dempster's T_1. For the same simplified hypothesis Dempster (1958) suggested an approximate T^2 test, avoiding the problem of singular matrix inverse

$$T_1 = \frac{\frac{n_1 n_2}{n_1 + n_2}(\bar{X} - \bar{Y})(\bar{X} - \bar{Y})^T}{trS} \quad (2)$$

asymptotically distributed as F-statistics under special conditions (see Srivastava and Du, 2008 for details).

2.1.3 N-statistic. For testing the hypothesis H: $F = G$ against a fixed alternative $F \neq G$ Klebanov *et al*. (2007) and Baringhaus and Franz (2004) proposed test statistic

$$N_{n_1,n_2} = \frac{n_1 n_2}{n_1 + n_2}\left[\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}L(X_i, Y_j) - \frac{1}{2n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}L(X_i, X_j) - \frac{1}{2n_2^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}L(Y_i, Y_j)\right] \quad (3)$$

Here we consider only $L(X, Y) = \|X - Y\|$, the Euclidean distance in R^p. In the original papers several other kernel functions L were suggested as well (see Baringhaus and Franz, 2004; Klebanov *et al*., 2007).

2.1.4 Univariate tests. Tian *et al*. (2005) suggested averaging the values of t-statistic for individual genes and testing the hypothesis of no association between gene sets and phenotypes with label permutations. This approach again tests the simplified hypothesis of no differences in mean expressions between two phenotypes. Here we consider two sign independent statistics: average absolute values of $t-$statistic and average squared t-statistic

$$\Sigma_{t,1} = \frac{1}{p}\sum_{i=1}^{p}|t_i|, \quad (4)$$

$$\Sigma_{t,2} = \frac{1}{p}\sqrt{\sum_{i=1}^{p}t_i^2}, \quad (5)$$

because the possibility that the same gene set has to include both, highly up- and down-regulated genes cannot be excluded.

2.2 Simulation setup

We simulated two samples of equal size, $N/2$ from the p-dimensional normal distribution $N(0, \Sigma)$ and $N(\mu, \Sigma)$ representing two biological conditions with different outcomes under the following global and local settings.

Downloaded from https://academic.oup.com/bioinformatics/article/25/18/2348/196468 by guest on 10 April 2024

(i) *Global settings*

(A) The number of genes (the dimensionality of the vector), p was set to mimic the typical number of genes in a biological pathway, $p = (20, 60, 100)$. The sample size was $N = 40$ for all simulations.

(B) In order to test the 'detection call' for different statistics under different local settings (see below) the parameter γ, indicating the proportion of genes in a pathway under alternative hypothesis, was introduced. That is in a given pathway only γp genes were changing their expression between phenotypes. For every p, γ was set to be 0.25, 0.5 or 0.75.

(C) In all experiments, the correlation coefficients between pairs of genes ($r_{ij} = s_{ij}/s_{ii}s_{ij}, i \neq j$) were set to 0.1, 0.5 or 0.9, respectively, reflecting the assumptions of approximately uncorrelated data, medium correlations and highly correlated gene expression data.

(ii) *Local settings*

(A) The mean vector for the first biological condition was fixed as $\mathbf{0}$ and all components μ_i of the mean vector, μ for the second biological condition were set to change from 0 to 2 with the step of 0.25, that is μ was varied from $\mu = (0, ..., 0)$ to $\mu = (2, ..., 2)$. The diagonal elements of the covariance matrix, i.e. variances of individual genes, were set to 1.

(B) For both biological conditions the mean vectors were set to $\mathbf{0}$ and for the second biological condition s_{ii}^2 were set to change from 1 to 5 with the step of 1.

For every fixed local setting (A or B) the samples were simulated under all varieties of global settings, giving in sum $3 \times 3 \times 3 \times 2 = 54$ different simulated data sets.

In order to assess how good the five tests under investigation control the Type I error rate we estimate it numerically from 1000 replications of the data set. We estimate the Type I error by the observed proportion of 1000 replications of the data set, simulated under the null hypothesis, where the alternative hypothesis was falsely accepted. We also estimate the empirical power by the observed proportion of 1000 replications of the data set, simulated under the alternative hypothesis, where the null hypothesis was correctly rejected. For all of these simulations we assumed a significance level $\alpha = 0.05$.

2.3 Biological data sets

To study the performance of different statistics on real biological data we consider two biological examples, frequently used in the literature with regards to differentially expressed gene sets.

2.3.1 Gene expression patterns in NCI-60 cell lines (p-53 data set). The first example comprises 50 samples of NCI-60 cell lines differentiated based on the status of the p53 gene: 17 cell lines carrying normal p53 gene and 33 cell lines classified as currying mutated p53 (Subramanian *et al*., 2005). This data set was also analyzed by Liu *et al*. (2007), to compare the performance of three methods, Global Test, ANCOVA Global Test and SAM-GS (see Liu *et al*., 2007 for detail). Expression data and pathways (C2 functional gene sets, as defined in Subramanian *et al*., 2005) were downloaded from GSEA web site ('Molecular Signature Database'). We excluded pathways with <15 genes, and studied differential expression of 369 sets. Data were normalized using Variance Stabilization (Huber *et al*., 2002) as recommended (Liu *et al*., 2007).

2.3.2 Acute lymphoblastic leukemia samples (ALL data set). The second example is a large data set from a clinical trial of ALL. Similar to Jiang and Gentleman (2007) we considered only two groups of patients with ALL: those having BCR/ABL fusion gene (37 cases) and those tested negative for this fusion (42 cases). Data were preprocessed as described (Jiang and Gentleman, 2007). As gene sets we considered KEGG (Kanehisa and Goto, 2000) pathways with more than 10 members.

For all statistics P-values were obtained from permutations (1000). For the sake of comparison to the results of Liu *et al*. (2007) and Jiang and Gentleman (2007), we fixed the same threshold for all P-values (0.001), followed as much as possible to their data preprocessing steps and did not consider correction for multiple testing. The later helps to avoid selection of a specific multiple testing procedure, influencing the results significantly (Dudoit and van der Laan, 2008). In these settings we present all pathways, found differentially expressed by at least one out of four tests (due to the similarity of Dempster's and $\Sigma_{t,2}$ in the simulation studies we dropped the former test for biological data. This similarity is actually expected, because Dempster's test is not truly multivariate in a sense it does not account for a comlex correlation structure).

3 RESULTS

3.1 Simulation studies

3.1.1 Estimated Type I error rate. Table 1 presents the results of our simulations to estimate the attained significance levels of the five tests. As can be seen all tests provide rather good estimates of $\alpha = 0.05$ when $\mu = 0$ under different parameter settings. It should be noted that Hotelling's T^2 always provides slightly conservative estimates of Type I errors, while all other tests are more liberal.

3.1.2 The empirical power of tests when the mean expression vector changes. Figures 1–3 show the power curves of multivariate and univariate tests under the local setting A. Generally, among all factors (namely dimensionality, detection call gamma and pairwise correlations), the correlations impact the power of the tests in the most effective way.

When the pairwise correlation is set to 0.1 (Fig. 1) the power of two univariate tests namely $\Sigma_{t,1}, \Sigma_{t,2}$and two multivariate tests, namely N-statistic, T_1, is virtually the same (the power of $\Sigma_{t,1}$ is lower when the detection call is equal to 0.25; Fig. 1, left-most column). The pathway's dimensionality slightly influences the slope of the power curve. When the detection call is set to 0.5 all four statistics reach $\sim$100% power when the mean expressions are 0.5 different given the pathway size is 100, 90% power given the pathway size is 60 and 80% power given the pathway size is 20 (Fig. 1, middle column). The detection call influences the power in

Table 1. Attained significance levels of the Hotelling's T^2, N-statistic $\Sigma_{t,1}$, $\Sigma_{t,2}$ and T_1

Parameters	T^2	N-statistic	$\Sigma_{t,1}$	$\Sigma_{t,2}$	T_1
$r_{ij} = 0.1$					
$p = 20$	0.048	0.058	0.051	0.046	0.049
$p = 60$	0.046	0.050	0.051	0.049	0.048
$p = 100$	0.043	0.055	0.058	0.063	0.061
$r_{ij} = 0.5$					
$p = 20$	0.048	0.053	0.051	0.051	0.049
$p = 60$	0.048	0.044	0.053	0.048	0.049
$p = 100$	0.043	0.056	0.052	0.053	0.054
$r_{ij} = 0.9$					
$p = 20$	0.048	0.057	0.049	0.048	0.047
$p = 60$	0.048	0.046	0.051	0.050	0.050
$p = 100$	0.044	0.056	0.056	0.055	0.056

Downloaded from https://academic.oup.com/bioinformatics/article/25/18/2348/196468 by guest on 10 April 2024

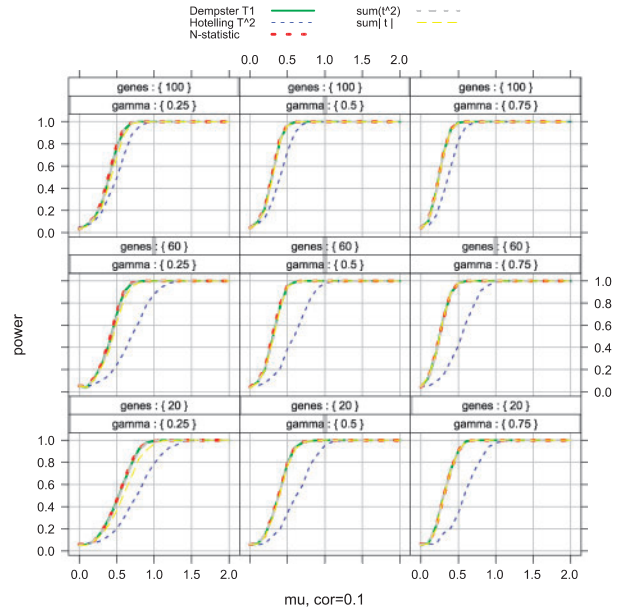

Fig. 1. The power curves of five tests. The mean expression vector is changing, the variance is fixed to 1, for correlation of 0.1 among genes. Parameter 'genes' is a number of genes in a gene set. Parameter 'gamma' corresponds to detection call.

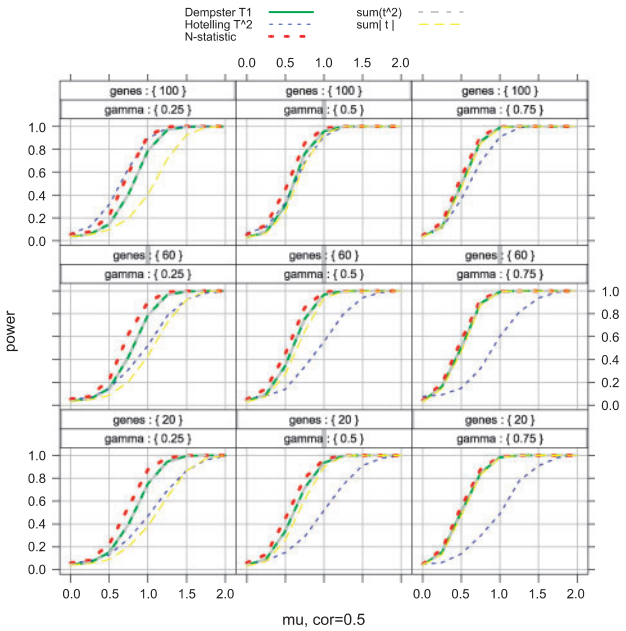

Fig. 2. The power curves of five tests. The mean expression vector is changing, the variance is fixed to 1, for correlation of 0.5 among genes. Parameter 'genes' is a number of genes in a gene set. Parameter 'gamma' corresponds to detection call.

a similar manner. When the pathway size is set to 60, the power of all four statistics reaches ∼100% power when the mean expressions are 0.5 different given the detection call is 0.75, 90% power given the detection call is 0.5 and 70% power given the detection call is 0.25 (Fig. 1, middle row).

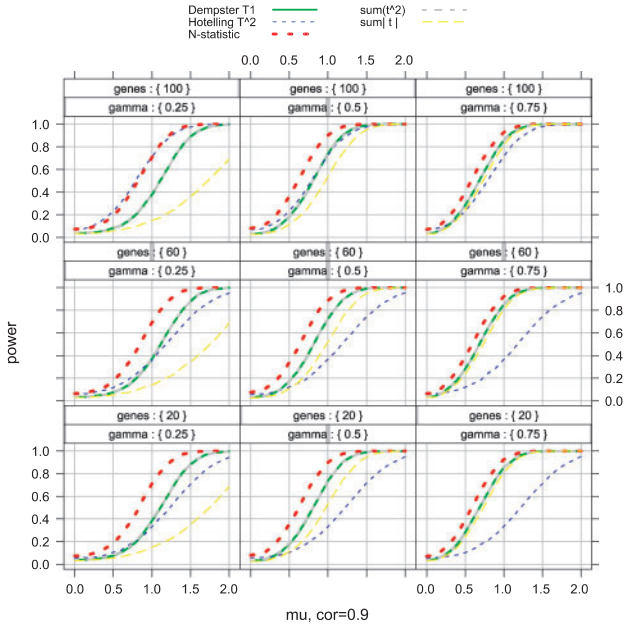

Fig. 3. The power curves of five tests. The mean expression vector is changing, the variance is fixed to 1, for correlation of 0.9 among genes. Parameter 'genes' is a number of genes in a gene set. Parameter 'gamma' corresponds to detection call.

When pairwise correlations are small (0.1) in all settings Hotelling's T^2 has lower power in comparison to other tests. The dimensionality of the pathway influences the power of T^2 in the similar way as the power of all other statistics. In contrast, the detection call influences the power of T^2 to a smaller extent: to reject the null hypothesis T^2 needs only several components of the mean vector to be significantly different (see below).

When the pairwise correlations are set to 0.5 or 0.9 (Figs 2 and 3) the power of all statistics is generally lower, but the power of $\Sigma_{t,1}$ is influenced the most when the detection call is set to 0.25. For the pathway size of 100 and a detection call of 0.75, the power curves for other four statistics are nearly identical (Figs 2 and 3, upper rows). However, under the presence of correlations even the best-performing N-statistic reaches the power of 100% only when the mean expressions are 1.5 different.

An interesting observation is that there is a narrow area of parameter values where Hotelling's T^2 is the best statistics (Figs 2 and 3, top left). For higher correlations (>0.1) low gamma (<0.5) and large gene sets (>60) there are intervals of mean differences for which Hotelling's T^2 slightly outperforms N-statistic. The difference in power is small but we will see in the results section for the biological data that this effect is relevant. Aside from these special parameter settings Hotelling's T^2 gives almost always the lowest power compared to all other tests.

3.1.3 The empirical power of tests when the mean expression vector is fixed and the variances are changing. Figure 4 shows the power curves of multivariate and univariate tests under the local setting B. Only N-statistic has the power to test the full hypothesis $F = G$ against a fixed alternative $F \neq G$ (Fig. 4). All other statistics have no power at all. This is expected, because the other tests were designed for testing different hypothesis. For a pathway, changes in the

Downloaded from https://academic.oup.com/bioinformatics/article/25/18/2348/196468 by guest on 10 April 2024

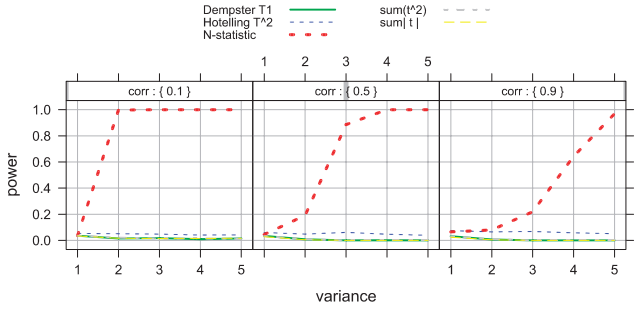

Fig. 4. The power curves of five tests. The mean expression vector is fixed, the variance is changing from 1 to 5 for correlation of 0.1, 0.5 and 0.9 among genes.

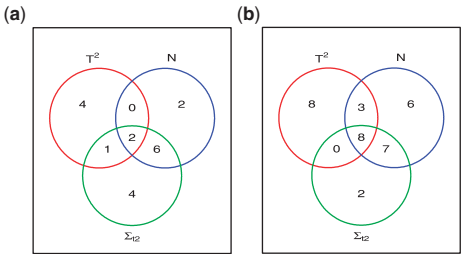

Fig. 5. Agreement among tests statistics. Venn diagrams for (**a**) the p53 data set and (**b**) the ALL data set.

Table 2. Gene sets in the p53 data sets with $P \leq 0.001$, discovered in at least one test (Hotelling's T^2, N-statistic, $\Sigma_{t,1}$, $\Sigma_{t,2}$)

Gene set ID	T^2	N-statistics	$\Sigma_{t,1}$	$\Sigma_{t,2}$	γ^a
Cell_Cycle	≤0.001	0.017	0.073	0.05	12.3
Cell_cycle_checkpointII	≤0.001	0.012	0.096	0.101	12.5
Cell_cycle_regulator	0.111	≤0.001	0.009	0.008	21.4
CR_CELL_CYCLE	≤0.001	0.022	0.078	0.052	11.3
DNA_DAMAGE_SIGNALLING	≤0.001	0.003	0.016	0.006	18.0
g2Pathway	0.076	≤0.001	0.013	0.007	22.7
badPathway	0.197	≤0.001	≤0.001	≤0.001	26.8
bcl2family_and_reg_network	0.807	0.003	≤0.001	≤0.001	21.3
ceramidePathway	0.05	≤0.001	≤0.001	≤0.001	27.7
chemicalPathway	0.129	0.001	0.002	≤0.001	20.5
CR_DEATH	≤0.001	0.012	0.007	≤0.001	15.3
hivnefPathway	0.011	0.009	≤0.001	≤0.001	18.2
mitochondriaPathway	0.005	≤0.001	≤0.001	≤0.001	30.3
p53hypoxiaPathway	0.247	0.018	≤0.001	≤0.001	27.5
p53Pathway	0.009	≤0.001	≤0.001	≤0.001	32.5
radiation_sensitivity	0.382	0.02	≤0.001	≤0.001	27.1
SA_FAS_SIGNALLING	0.009	≤0.001	≤0.001	≤0.001	34.8
SA_G1_AND_S_PHASES	≤0.001	≤0.001	≤0.001	≤0.001	37.5
SA_PROGRAMMED_CELL_DEATH	≤0.001	≤0.001	≤0.001	≤0.001	37.5

[a]The percent of genes, changing their expression between two phenotypes in a gene set.

variance of the pathway given the same mean vector might indicate the presence of differential regulation under different phenotypes, and should be as interesting as changes in the average expression itself. It appears that this case can be detected by the N-statistic while all other test statistics are insensitive.

3.2 The analysis of biological data

3.2.1 The power of tests to detect differentially expressed gene sets for p-53 data set. In sum, different statistics detected 19 gene sets at the given significance level ($P < 0.001$), differentially expressed between cancer cell lines with and without p53 mutation (Table 2). Among them, 13 were also reported in Liu *et al*. (2007) study and six were found in this study for the first time. We note that the sum of squared t-statistics has the highest power [13 significantly differentially expressed pathways, Table 2, Fig. 5a] and the result of $\Sigma_{t,1}$ is a subset of the result of $\Sigma_{t,2}$.

While the results of N-statistic and two univariate tests almost coincide at the more liberal significance level (e.g. $P < 0.01$), for several pathways P-values from Hotelling's T^2 are rather high. On the other hand, four gene sets (Cell_Cycle, cell_ cycle_checkpointII, DNA_damage_signalling and CR_CELL_CYCLE) were reported exclusively by Hotelling's T^2 and cannot be detected by other statistics even at the 0.01 liberal threshold (except DNA_damage_signalling, detected at 0.01 threshold by $\Sigma_{t,2}$). These sets represent the major of the functional targets of p53 activity, namely the regulation of cell cycle progression and DNA damage signaling, and include p53 itself and its multiple targets. Therefore, one may consider the failure of other statistics to identify correctly these sets as a false negative error. In what follows we provide plausible, yet empirical explanation of the reasons behind sporadically high P-values of Hotelling's T^2 and unexpected false negative errors of other statistics.

First, we note that generally when P-values from Hotelling's T^2 are high, the distributions of pairwise correlation coefficients in two groups are rather different. For example, for bcl2family_and_reg_network the P-value of Hotelling's T^2 is the largest: 0.807 (Table 2). For this set the distributions of pairwise correlations in two groups are drastically different and average pairwise correlations are 0.027 and −0.004, respectively (Supplementary Fig. 1). For two other gene sets with the largest P-values from Hotelling's T^2 (radiation_sensitivity and p53hypoxiaPathway, P-values are 0.382 and 0.247, respectively) the situation is similar (Supplementary Fig. 1). It might be that in these cases the assumption of equal covariance for two samples is violated and the power of the test is dropped.

The presence of false negative errors for the other statistics can be explained if we introduce the measure of the percentage of individually changing genes in a gene set. Let us measure this quantity using absolute values of the t-statistic and consider that the gene is changing if its t-statistic is more than 1.96 (the factual value of the threshold here does not matter; it should simply reflect the presence of change). For the gene sets, detected by Hotelling's T^2 only (Cell_Cycle, cell_cycle_checkpointII, DNA_ damage_signalling and CR_CELL_CYCLE) this measure is much lower (12.3, 12.5, 18.0 and 11.3%, respectively), as compared to other gene sets (Table 2 and Supplementary Fig. S1). Thus the other statistics cannot detect the overall expression changes in a set when only few genes are actually changing, in contrast with the high sensitivity of Hotelling's T^2.

Downloaded from https://academic.oup.com/bioinformatics/article/25/18/2348/196468 by guest on 10 April 2024

There are two pathways, reported by univariate tests only, even if we consider the more liberal significance level (e.g. $P < 0.01$): radiation_sensitivity and p53hypoxiaPathway. These pathways were also reported by Liu *et al.* (2007), include p53 and its targets, and should be considered as true positives.

3.2.2 The power of tests to detect differentially expressed sets for ALL data set. Overall for the ALL data set the four tests gave more homogeneous results compared to the p53 data set, probably because of the larger ALL sample size (79 slides) of the ALL data set. Thirty-five KEGG pathways were found differentially expressed by at least one of the four tests at the given significance level ($P \leq 0.001$) (Supplementary Table 1). Multivariate and univariate tests detected simultaneously eight pathways (Fig. 5b). In four cases, P-values of $\Sigma_{t,1}$ and in four cases Hotelling's T^2 were larger than 0.05 (Supplementary Table 1). As before, the failure of $\Sigma_{t,1}$ to detect differential expression among sets can be explained by the small percent of individual genes in a set actually changing their expression (Supplementary Table 1 and Supplementary Fig. S2). The failure of Hotelling's T^2 test might be related to the violated assumption of equal covariance for two samples (Supplementary Fig. S2). Among 35 sets, 13 were also reported by Jiang and Gentleman (Jiang and Gentleman, 2007) as differentially expressed.

4 DISCUSSION

The analysis of differentially expressed gene sets is an effective way to overview the underlying biological trends in gene expression data sets. There is a rich body of statistical tests available for finding gene sets, differentially expressed between two phenotypes. Several comparative studies address the relative performance of these tests (Jiang and Gentleman, 2007; Liu *et al*., 2007; Song and Black, 2008), one suggests a special treatment of gene sets analyses for prokaryotes (Tintle *et al*., 2008) and one builds a complete 'taxonomy' of the testing procedures (Ackermann and Strimmer, 2009). Here we emphasize that the most fundamental difference among these approaches is formulated in terms of the null hypothesis they test.

We have studied the relative power of popular univariate and multivariate tests for several simulated and two biological data sets. We considered two gene-level statistics, the average of the absolute and the squared t-tests for individual genes in a gene set, and three multivariate statistics, Hotelling's T^2, N-statistic and Dempster T_1. Although for three of these statistics the tested null hypothesis is different, their relative performance on simulated data is similar. All tests perform reasonably well in estimating the Type I error rate. Among the three parameters varied in simulations (the magnitude of pairwise correlations among gene expressions, the number of genes changing their expression in a set and the size of a gene set) the magnitude of pairwise correlations has the largest influence on the power of all tests. Despite the general belief that multivariate tests, accounting for a complex interdependence structure between genes might have a better power compared to univariate tests this study demonstrates that the use of multivariate statistics does not lead to a substantial gain in power when correlations are present and high. When correlations are low both, the number of genes changing their expression in a set and the size of a set have only slight influence on the power (except Hotelling's T^2), in contrast to the case of high correlations. For the poor performance of Hotelling's T2 there are two possible explanations in our opinion. First, the numerical estimation of the generalized inverse may be unstable for conditions relevant for the analysis of microarray data. Second, this may depend directly or indirectly on an altered correlation structure in both conditions because they form the underlying basis for the generalized inverse of the combined covariance matrix. For both reasons additional studies are necessary. On the other hand, the power of Hotelling's T^2 rises quickly when the number of genes changing in a set and the set's size increase when correlations are low. In sum, the performance of all tests coincides when the correlations are low, the gene set size is large and the percent of genes changing their expression is high. However, the beneficial combinations of all these factors may rarely happen in true biological data and the performance of these tests might be different for real data set.

The analysis of biological data again demonstrates that there are some aspects in gene expression data that cannot be efficiently modeled. From the Venn diagrams shown in Figure 5a and b one can see that Hotelling's T^2 is a more important test for biological data than one might expect from the simulation results in Section 3. This is not only a surprise but strongly indicates that the simulation technique lacks important characteristics from real biological data. Also, for the p53 data set, Σ_{t2} has slightly higher power than N-statistic, while N-statistic always slightly outperforms all other tests on simulated data. On the other hand, for simulated as well as biological data, the results of Σ_{t1} are subsets of the results of Σ_{t2}. Similarly, the good performance of Hotelling's T^2 when the number of genes changing their expression in a gene set is small is captured in both, simulated and biological data. That is, the simplified model assuming a multivariate normal distribution for gene expressions adequately reflects some, but not all properties of the biological data.

The intersection of significant gene sets, found by different tests in p53 and ALL data is substantial. However, Hotelling's T^2 is able to find gene sets which are not discovered by other tests, where only about 11–12% of all genes are changing their expression. This is because Hotelling's T^2 involves all variables symmetrically and can equally detect changes for a single variable, for all, or for the subset. One can expect that the sets reported exclusively by Hotelling's T^2 constitute false negative errors for other tests in the case of p53 data, because these sets directly include p53 together with its functional targets. We also note that for the p53 and ALL data sets, on the average only 20–25% of genes in the gene sets are actually changing their expression (Table 2, Supplementary Table 1). This observation adds more evidence to the motivation of (Mootha *et al*., 2003) for studying gene sets instead of individual genes.

Here we studied two univariate and three multivariate self-contained tests for detecting differential expression of gene sets. All these tests can be distinguished by their underlying null hypotheses, leaving only three conceptually different statistical tests with respect to the null hypotheses. It should be noted that these three null hypotheses cover the vast majority of the current universe of self-contained tests. The three best-performing tests for these hypotheses (sum of squared t-tests, Hotelling's T^2 and N-statistic), find common but also complementing gene sets differentially expressed. Due to complementing null hypotheses, each test projects on different aspects of the data and for this reason, their simultaneous use for the analysis of biological data leads to an increased power as compared to individual tests.

Downloaded from https://academic.oup.com/bioinformatics/article/25/18/2348/196468 by guest on 10 April 2024

ACKNOWLEDGEMENTS

The authors like to thank Earl Glynn for his comments on the manuscript.

Funding: Grants from NIH (R21HG004648, GM079259) and an Alfred P. Sloan Research Fellowship (to G.G.).

Conflict of Interest: none declared.

REFERENCES

Ackermann,M. and Strimmer,K. (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10**, 47.

Baringhaus,L. and Franz,C. (2004) On a new multivariate two-sample test. *J. Multivariate Anal.*, **88**, 190–206.

Barry,W.T. *et al*. (2008) A statistical framework for testing functional categories in microarray data. *Ann. Appl. Stat.*, **2**, 286–315.

Dempster,A.P. (1958) A high dimentional two sample significance test. *Ann. Math. Statist.*, **29**, 995–1010.

Dudoit,S. and van der Laan,M.J. (2008) *Multiple Testing Procedures with Applications to Genomics.* Springer, Berlin.

Emmert-Streib,F. (2007) The chronic fatigue syndrome: a comparative pathway analysis., *J. Comput. Biol.*, **14**, 961–972.

Emmert-Streib,F. and Dehmer,M.E. (2008) *Analysis of Microarray Data: A Net-work-Based Approach.* Wiley-VCH, New York.

Goeman,J.J. and Buhlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Huber,W. *et al*. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18** (Suppl. 1), S96–S104.

Hummel,M. *et al*. (2008) GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics*, **24**, 78–85.

Jiang,Z. and Gentleman,R. (2007) Extensions to gene set enrichment. *Bioinformatics*, **23**, 306–313.

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

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

Kitano,H. (2001) *Foundations of Systems Biology.* The MIT Press, Cambridge.

Klebanov,L. *et al*. (2007) A multivariate extension of the gene set enrichment analysis. *J. Bioinform. Comput. Biol.*, **5**, 1139–1153.

Kong,S.W. *et al*. (2006) A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, **22**, 2373–2380.

Liu,Q. *et al*. (2007) Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, **8**, 431.

Lu,Y. *et al*. (2005) Hotelling's T2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics*, **21**, 3105–3113.

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

Nettleton,D. *et al*. (2008) Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics*, **24**, 192–201.

Palsson,B.O. (2006.) *Systems Biology: Properties of Reconstructed Networks.* Cambridge University Press, Cambridge.

Schafer,J. and Strimmer,K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article 32.

Song,S. and Black,M.A. (2008) Microarray-based gene set analysis: a comparison of current methods. *BMC Bioinformatics*, **9**, 502.

Srivastava,M.S. and Du,M. (2008) A test for the mean vector with fewer observations than the dimension. *J Multivariate Anal.*, **99**, 386–402.

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

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

Tintle,N.L. *et al*. (2008) Gene set analyses for interpreting microarray experiments on prokaryotic organisms. *BMC Bioinformatics*, **9**, 469.

Tomfohr,J. *et al*. (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, **6**, 225.

Venables,W.N. and Ripley,B.D. (1999) *Modern Applied Statistics with S-PLUS.* Springer, Berlin.

Xiong,H. (2006) Non-linear tests for identifying differentially expressed genes or genetic networks. *Bioinformatics*, **22**, 919–923.

Downloaded from https://academic.oup.com/bioinformatics/article/25/18/2348/196468 by guest on 10 April 2024