BMC Bioinformatics

Proceedings

Testing for treatment effects on gene ontology

Taewon Lee*[1], Varsha G Desai[2], Cruz Velasco[3], Robert J Shmookler Reis[4] and Robert R Delongchamp[5]

Address: [1]Biometry Branch, Division of Personalized Nutrition and Medicine, National Center for Toxicological Research, 3900 NCTR Road, Jefferson, AR 72079, USA, [2]Center for Functional Genomics, Division of Systems Toxicology, National Center for Toxicological Research, 3900 NCTR Road, Jefferson, AR 72079, USA, [3]Louisiana State University Health Sciences Center, New Orleans, LA 70112, [4]Depts. of Geriatrics. Biochemistry & Molecular Biology, and Pharmacology/Toxicology, University of Arkansas for Medical Sciences, and VA Medical Center, Little Rock, AR 72205, USA and [5]Department of Epidemiology, University of Arkansas for Medical Sciences, College of Public Health, 4301 W Markham St, #820, Little Rock, AR 72205, USA

Email: Taewon Lee* - taewon.lee@gmail.com; Varsha G Desai - varsha.desai@fda.hhs.gov; Cruz Velasco - cvelas@lsuhsc.edu; Robert J Shmookler Reis - rjsr@uams.edu; Robert R Delongchamp - RDelongchamp@uams.edu

* Corresponding author

This article is available from: http://www.biomedcentral.com/1471-2105/9/S9/S20

Abstract

In studies that use DNA arrays to assess changes in gene expression, it is preferable to measure the significance of treatment effects on a group of genes from a pathway or functional category such as gene ontology terms (GO terms, http://www.geneontology.org) because this facilitates the interpretation of effects and may markedly increase significance. A modified meta-analysis method to combine p-values was developed to measure the significance of an overall treatment effect on such functionally-defined groups of genes, taking into account the correlation structure among genes. For hypothesis testing that allows gene expression to change in both directions, p-values are calculated under the null distribution generated by a Monte Carlo method.

As a test of this procedure, we attempted to distinguish altered pathways in microarray studies performed with Mitochips, oligonucleotide microarrays specific to mitochondrial DNA-encoded transcripts. We found that our analytic method improves the specificity of selection for altered pathways, due to incorporation of the inter-gene correlation structure in each pathway. It is thus a practical method to measure treatment effects on GO groups. In many actual applications, microarray experiments measure treatment effects under complicated design structures and with small sample sizes. For such applications to real data of limited statistical power, and also in computer simulations, we demonstrate that our method gives reasonable test results.

Introduction

The advent of DNA microarray technology has revolutionized genomic research and medicine because of its ability to simultaneously determine the expression levels of thousands of genes. However, the interpretation of large amounts of microarray gene expression data, and the abil-

ity to derive biologically meaningful conclusions from such data, have always been daunting tasks for statisticians. Because of the high volume and complex characteristics of microarray data, much of the initial work on their analysis has focused on development of data mining or data reduction methods to identify differentially expressed genes. Typically, the *p*-value of a test statistic is calculated for each gene, the genes are ranked according to these *p*-values, and a pre-specified significance criterion, such as the false discovery rate, is used to determine a cutoff which creates a category of differentially expressed genes [1-3].

Attempts to interpret individual genes in a list of significant genes are demanding and laborious. Therefore, recent efforts have focused on discovery of biological pathways rather than on individual gene function. Gene ontology terms (GO terms, http://www.geneontology.org) reflect gene groupings based on molecular function, biological process, or cellular structure/organelle. The interpretation of differentially expressed GO groups is generally simpler than the presentation of a list of statistically significant genes, and more resistant to erroneous conclusions that can arise from microarray artefacts.

Several statistical methods that combine the analysis of differential gene expression with biological databases have been proposed and implemented in computer packages for a more rapid interpretation of genome-wide expression data [4]. However, most such methods are based on a series of univariate statistical tests and do not properly account for the complex structure of gene interactions. The statistical significance of a GO group is commonly assessed by comparing the number of statistically significant genes in the group to the number expected by chance using Fisher's exact test, which is based on the hypergeometric distribution [5]. Fisher's exact test is used to compare these proportions to assess overrepresentation of significant genes in functional categories. This approach is not amenable to correction for correlations among *p*-values, since the test inherently assumes exchangeability among genes, an assumption which is not valid under arbitrary or actual correlation structures [6,7].

Hotelling's T^2 Statistic and permutation methods address the correlation structure among genes. Hotelling's T^2 statistic is not applicable when the sample size is smaller than the number of genes in a GO term [8]. Permutation methods, although quite valuable under appropriate conditions, are severely compromised by limited numbers of permutable sample pairs. In many cases, the design of microarray studies has a rather complicated structure intended to manage technical variation associated with differences among probes, dyes, and reagent batches by creating treatment blocks within these sources of variation [9]. Such cases are not suited to permutation methods.

A modified meta-analysis method was developed by Delongchamp *et al.* [10] to combine *p*-values, and thus to measure the significance of an overall treatment effect on a group of genes, while taking into account the inter-genic correlation structure. The method is based on the fact that *p*-values follow a uniform distribution under the null hypothesis. Inverse-normal transformed *p*-values have a normal distribution and their sum over a set of genes also would follow a normal distribution, provided that the component *p*-values are independent. The test we have developed to measure the significance of overall treatment effect on genes within a GO category is based on a modification of this statistic, to reflect the actual correlation structure among genes sharing a GO term.

In this paper, we extend the method from a simple one-class *t*-test with the null hypothesis $H_0 : \mu = 0$ to a test for pair-wise contrast in a fixed-effects linear model. In the following sections, we describe in detail the extension of the methodology, with validation through computer simulations and application to two toxicogenomics studies designed to evaluate treatment effects on the levels of mRNA transcripts involved in mitochondrial function. We thus demonstrate that this methodology provides a practical approach to testing the significance of the treatment effects on gene classes defined by GO terms, and by extension on any other prior categorization of genes into functional subsets. Because many microarray experiments measure treatment effects under complicated design structures and with small sample sizes [9], we used a simulation study to determine whether the method gives reasonable results under these conditions.

Specific applications to toxicogenomics studies showed that the methodology has improved specificity in choosing significantly altered pathways or functional categories, and may thus assist in the understanding of molecular mechanisms of mitochondrial toxicity in the liver induced by anti-HIV drugs [11,12] and in assessing effects on mitochondrial function of weight-reducing dietary supplements, such as usnic acid [13].

Methods

Measurement of a treatment effect for each gene
Under a fixed-effects linear model, gene expression data for an arbitrary gene can be written as $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, where $\mathbf{y}$ and ε are $n \times 1$ random vectors, X is a known $n \times p$ design matrix of rank r, and β is a $p \times 1$ vector representing unknown parameters. The vector $\mathbf{y}$ denotes an observed measurement of expression, suitably transformed, for n biological samples, and ε is an error vector, distributed as $N_n (0, \sigma^2 \mathbf{I}_n)$, where σ^2 denotes the unknown within-treat-

ment variance among samples. The parameters β and σ^2 are assumed to be gene-specific. Statistical analyses are applied to one gene at a time, with a common design matrix, $\mathbf{X}$. The unbiased estimators of β and σ^2 are

$$\hat{\dagger} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\hat{\sigma}^2 = \frac{1}{n-r}(\mathbf{y} - \mathbf{X}\hat{\dagger})'(\mathbf{y} - \mathbf{X}\hat{\dagger}).$$

In many toxicogenomic studies, the significance of a treatment effect is tested under the null hypothesis $H_0 : \mathbf{c}\beta = 0$, where $\mathbf{c}\beta$ is a pair-wise contrast among treatments. Under the null hypothesis, $T = \dfrac{\mathbf{c}\hat{\dagger}}{\hat{\sigma}\sqrt{\mathbf{c}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}'}}$ has a *t*-distribution with n - r degrees of freedom, and the *p*-value to assess the significance of a treatment is calculated from this statistic.

Test for a gene group

A modified meta-analysis method of combining *p*-values was developed to measure the significance of an overall treatment effect on any group of genes by a one-class *t*-test [10]. The *p*-value calculated from the null hypothesis is a random variable with uniform distribution, which can be transformed to a suitable probability distribution. Inverse-normal transformed *p*-values, $z_k = \Phi^{-1}(1 - p_k) \sim N(0, 1)$, $k = 1, \cup, m$ have a normal distribution and their sum, $\sum_{k=1}^{m} z_k / \sqrt{m} \sim N(0,1)$, is also normally distributed when *p*-values are independent. Here, p_k represents a *p*-value for a gene in a GO group comprising m genes. The *p*-value for the sum of inverse-transformed *p*-values, $p = 1 - \Phi\left(\sum_{k=1}^{m} z_k / \sqrt{m}\right)$, gives the overall significance of the treatment effect on the GO group. We refer to this as the naïve estimate because it assumes independence among *p*-values.

In reality, genes in a GO group are likely to be functionally related and thus not independent. When the correlation structure among genes is known, we can make a simple adjustment of the naïve estimate. In this case, the test statistics T still has a standard normal distribution and we denote it as $T_k = \dfrac{\mathbf{c}\hat{\dagger}_k}{\sigma_k \sqrt{\mathbf{c}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}'}}$, for the k-th gene in a GO group. A common contrast vector, c, is used through all genes since we are measuring same contrast for each gene. The summary statistic for a GO group, $\mathbf{1}'\mathbf{z} = \sum_{k=1}^{m} z_k = \sum_{k=1}^{m} \Phi^{-1}(1 - p_k) = \sum_{k=1}^{m} T_k$ is also nor-

mally distributed and its variance is $$\text{var}(\mathbf{1}'\mathbf{z}) = \sum_{k=1}^{m} \text{var}(T_k) + 2\sum_{s>t} \text{cov}(T_s, T_t) = \mathbf{1}'\mathbf{R}\mathbf{1},$$ where $\mathbf{1}$ is m vector of 1s and $\mathbf{R}$ is the correlation matrix of $(\mathbf{y}_1, \mathbf{y}_2, \cup, \mathbf{y}_m)$. Note that

$$\begin{aligned} \text{cov}(T_s, T_t) &= \text{cov}\left(\frac{\mathbf{c}\dagger_s}{\sigma_s\sqrt{\mathbf{c}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}'}}, \frac{\mathbf{c}\dagger_t}{\sigma_t\sqrt{\mathbf{c}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}'}}\right) \\ &= \frac{1}{\sigma_s\sigma_t\mathbf{c}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}'}\text{cov}(\mathbf{c}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_s, \mathbf{c}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}_s) \\ &= \frac{1}{\sigma_s\sigma_t\mathbf{c}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}'}\mathbf{c}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{cov}(\mathbf{y}_s, \mathbf{y}_s)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}' \\ &= \frac{\sigma_{s,t}}{\sigma_s\sigma_t\mathbf{c}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}'}\mathbf{c}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}' \\ &= \frac{\sigma_{s,t}}{\sigma_s\sigma_t} = r_{s,t}, \end{aligned}$$

where $r_{s,t}$ is the *s*-th row and *t*-th column element of $\mathbf{R}$. Therefore, $p = 1 - \Phi\left(\dfrac{\mathbf{1}'\mathbf{z}}{\sqrt{\mathbf{1}'\mathbf{R}\mathbf{1}}}\right)$ is the appropriate *p*-value which corrects the naïve *p*-value, $p = 1 - \Phi\left(\dfrac{\mathbf{1}'\mathbf{z}}{\sqrt{m}}\right)$.

It follows that $p = 1 - \Phi\left(\left(\dfrac{\mathbf{1}'\mathbf{z}}{\sqrt{m}}\right)\sqrt{\dfrac{1}{1+(m-1)\bar{r}}}\right)$, where $\bar{r}$ is the average of off-diagonal elements of $\mathbf{R}$. The correction depends only on the average correlation, $\bar{r}$, and the correction tends to give a reduced significance when $\bar{r} > 0$. When $\mathbf{R}$ is unknown, we estimated the covariance, $\hat{\sigma}_{s,t} = \dfrac{1}{n-r}(\mathbf{y}_s - \mathbf{X}\hat{\dagger}_s)'(\mathbf{y}_t - \mathbf{X}\hat{\dagger}_t)$ to provide an average correlation coefficient $\bar{r}$ for the correction.

The correlation structure of *p*-values is different for a two-sided *t*-test, which allows gene expression changes in both directions, than for a one-sided situation. A two-tailed test, in which $p = 2(1 - \Phi(|z|))$, requires a different correction method, since the correlation among $|z_k|$, $k = 1, \cup, m$ differs from a one-sided test in which $p = 1 - \Phi(z)$. The null distribution of the summary statistics $\mathbf{1}'|\mathbf{z}|$ can be generated through Monte Carlo sampling from the null distribution of $\mathbf{z}$, MVN(0, cov($\mathbf{z}$)). When $\mathbf{z}_1, \cup, \mathbf{z}_n$ are random samples from MVN(0, cov($\mathbf{z}$)), the *p*-value for the observed value, $\Psi = \mathbf{1}'|\mathbf{z}|$, is computed as $p = \dfrac{1}{r}\sum_{l=1}^{r} \text{I}(\Psi > \mathbf{1}'|\mathbf{z}_l|)$, where I($A$) is an indicator function which gives 1 if A is true, or 0 otherwise. Here, cov($\mathbf{z}$)

has to be estimated from the data. The estimated correlation, $\hat{R}$ and its variation, $\bar{R}$, which has $\bar{r}$ for off-diagonal elements, are used to estimate cov(z).

Results

Simulation

The derivation of the method presented in the previous section is based on the known correlation matrix of the vector of dependent variable Y. When the correlation matrix is not known, we use an estimate of the correlation matrix. In reality, the correlation matrix is always unknown. The proposed method produces an approximately correct p-value for a group of genes. To demonstrate that the method gives not perfect but acceptably correct p-values, we present simulation results in this section. The validation is done by checking the cumulative distribution of p-values from the proposed methods under the null distribution. The *p*-values must have a uniform distribution, which should form a diagonal line in the following figures if *p*-values are correctly calculated.

The simulation is conducted under a fairly common set of conditions for microarray studies, comprising three treatments with three samples (arrays) per treatment. Samples are generated from $N(\mu_i, \Sigma)$ for each treatment, $i = 1, 2\ 3,$. In this simulation, samples for two treatment have the same average values, $\mu_1 = \mu_2 = 1$, and samples for the other treatment have twice that average value $\mu_3 = 2$. *P*-values for the pair-wise contrast are calculated under the null hypothesis, $H_0 : \mu_1 = \mu_2$. A GO term is composed of $m = 20$ genes which have correlation structure generated randomly between 0.36 and 0.55. The standard deviations, $\sigma_i, i = 1, \cup, m$ for the genes, which are the diagonal elements of Σ, are generated randomly between 0.01 and 0.25. We iterated this procedure at least 10,000 times to observe the distribution of calculated *p*-values.

Figure 1 plots the cumulative distribution of *p*-values from a one-sided test when the number of samples is $n = 9$, i.e., 3 groups with 3 samples for each group. The naïve *p*-values, shown by the red line, clearly deviate from the diagonal line. Almost 30% of *p*-values are estimated to be less
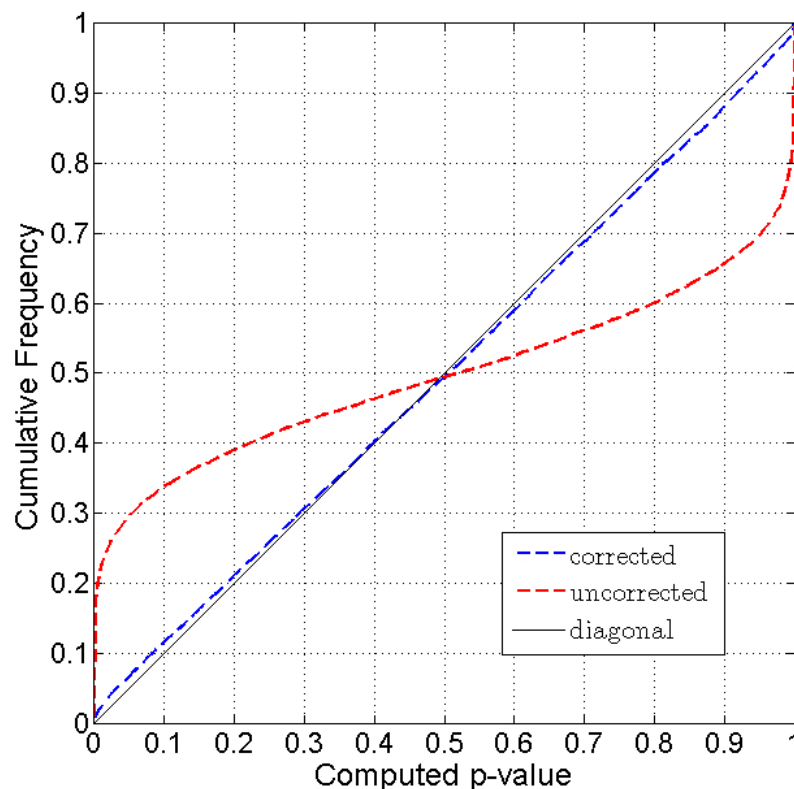

Figure 1
Cumulative distribution of p-values for one-sided test case with sample size n = 9. The naïve *p*-values (dashed red line) deviate from the diagonal line. Almost 30% of *p*-values are estimated to be less than 0.05. The corrected *p*-values (dashed blue line) fall very near the diagonal line.

than 0.05, indicating that the naïve *p*-values lead to a very high false-discovery rate. The corrected *p*-values (dashed blue line) fall very near the diagonal line. The corrected *p*-values thus have more specificity in choosing altered functional gene groups than the naïve *p*-values.

Figure 2 shows the simulation result for a two-sided case. *P*-values are calculated from the null distribution generated from Monte Carlo samples from MVN(0, cov($\mathbf{z}$)). Two estimates of cov($\mathbf{z}$) are used for the sampling. When $\bar{\mathbf{R}}$ is used, the empirical distribution of *p*-values is closer to a uniform distribution than when $\hat{\mathbf{R}}$ is used. The estimate of the average correlation, $\bar{r}$, is more robust than that of each element of $\hat{\mathbf{R}}$ when sample size is small (n = 3 for each group).

Figure 3 shows the distribution of *p*-values for a case with larger samples. The simulation for one-class *t*-test with

sample size n = 25 was conducted as above, to compare several methods for two sided tests. The empirical distributions of *p*-values were generated from 500,000 iterations. We looked at small *p*-values between 0.1 and 0.001 on a log scale. Figure 3 shows that the Hotelling T^2 test gives the smallest difference from the uniform distribution. The Hotelling T^2 test is applicable when the number of sample is larger than the number of genes in a group. When we have a reasonably correct estimate of $\mathbf{R}$, the $\hat{\mathbf{R}}$ method is a little better than the $\bar{\mathbf{R}}$ method which uses an approximation of $\mathbf{R}$. Both the $\hat{\mathbf{R}}$ method and the $\bar{\mathbf{R}}$ method give quite accurate p-values with reference to the p-values from the true correlation matrix, $\mathbf{R}$.

Examples
We present two real-world examples based on data from Mitochip, a mitochondria-specific mouse oligonucleotide microarray which was developed by Dr. Varsha Desai at
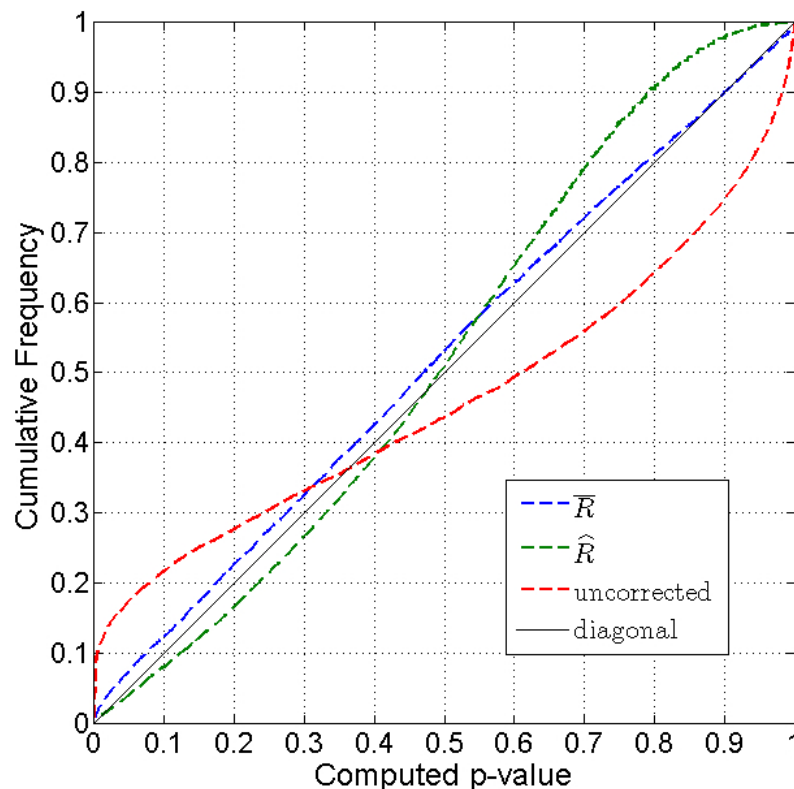

Figure 2
Cumulative distribution of p-values for two-sided test case with sample size n = 9. *P*-values calculated from random samples based on $\bar{\mathbf{R}}$ (dashed blue line) and $\hat{\mathbf{R}}$ (dashed green line) give reliable corrections, while the naïve *p*-value (dashed red line) overstates the significance of the test.
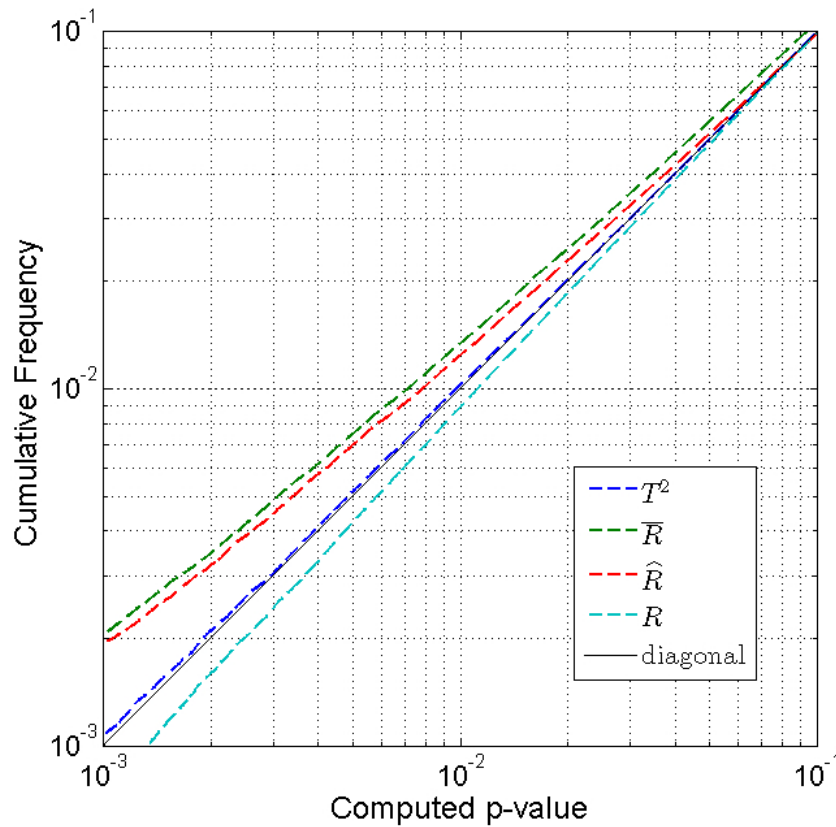
Figure 3

Comparison of two-sided tests with sample size n = 25. Hotelling T^2 test (dashed blue line) gives the smallest difference from the uniform distribution. When we have enough number of samples to have a reasonably correct estimate of **R**, the $\hat{R}$ method (dashed red line) is a little better than the $\bar{R}$ method (dashed green line). Both the $\hat{R}$ method and the $\bar{R}$ method give quite accurate p-values compared to the p-values from the true correlation matrix, **R** (dashed cyan line).

---

the National Center for Toxicological Research [11]. Mitochip measures the levels of mRNA for 542 mitochondrial and nuclear genes associated with mitochondrial structure and function. Each Mitochip includes 9 housekeeping genes and 9 *Arabidopsis* plant genes to serve as positive and negative control genes, respectively. We considered 317 relevant GO groups related to mitochondrial functions,

based on a database from Mouse Genome Informatics (MGI, http://www.informatics.jax.org).

Table 1 shows the design of an experiment to test the effects of zidovudine (AZT) and lamivudine (3TC) on mouse-liver gene expression. AZT is an anti-HIV drug used to reduce mother-to-child transmission of the virus. AZT

Table 1: Experimental design for the AZT and 3TC effects on mouse-liver gene expression.

Genotype		(+/-)			(+/+)	
Treatment	Vehicle	AZT 240 mg/kg/d	AZT+3TC 160+100 mg/kg/d	Vehicle	AZT+3TC 160+100 mg/kg/d
Batch 1	A1	B1	C1	D1	E1
Batch 2	A2	B2	C2	D2	E2
Batch 3	A3	B3	C3	D3	E3

Fifteen samples are collected and assayed for gene expression using the experimental design

is reported to produce severe adverse effects, and shows more toxicity when AZT is applied in combination with 3TC. Adverse effects are believed to be due to drug-induced mitochondrial disfunction [14].

Oxidative phosphorylation is a key mitochondrial function that requires the electron transport assembly of four protein complexes (I, II, III, IV) to catalyze sequential oxidation/reduction reactions, and complex V to generate ATP. Several clinical and animal studies have investigated the effect of nucleoside reverse transcriptase inhibitors (NRTIs), analogs such as AZT, on mitochondrial respiratory chain complexes. These studies suggest that there is a deficit in one of the components of complexes I and IV in skeletal muscle of children perinatally exposed to antiretroviral nucleoside analogues [15].

Table 2 shows *p*-values indicating the significance of treatment effects on the GO groups related to oxidative phosphorylation and apoptosis. The two-sided correction method detects significant effects on genes encoding components of complexes III and IV, whereas the naïve method finds that genes in all 5 complexes are significantly affected. This demonstrates that the two-sided correction method is more specific in finding significantly affected gene groups, although of course the "true" answer is not known *a priori*. Although Fisher's exact test also detects significant alteration in genes of complex IV, this test appears to be detecting a gene group that is different from the other groups, rather than registering treatment effects directly. The one-sided test is not applicable since it seems that gene expression changes in both directions after AZT and 3TC treatment.

Usnic acid is a lichen metabolite used as a weight-loss dietary supplement due to its uncoupling action on mitochondria. However, its use has been associated with severe liver disorders in many individuals. Animal studies conducted thus far have evaluated effect of usnic acid on mitochondria, primarily by measuring the rate of oxygen consumption and/or ATP generation. Generation of ATP requires tight coupling of electron transport with oxidative phosphorylation, maintained through a proton gradi-

ent across the inner mitochondrial membrane. An important finding of the study is a lack of usnic acid effect on complex V, despite a significant up-regulation of all four complexes of the electron transport chain. Usnic acid is a known uncoupler that is highly lipophilic in both neutral and anionic forms due to its numerous carbonyl groups that absorb the negative charge of the anion by resonance stabilization. This lipophilicity of usnic acid and the usniate anion allows easy passage of both entities through the mitochondrial membranes by passive diffusion into the matrix where it is ionized, releasing a proton into the matrix. The resulting usniate anion can then diffuse back into the inter-membrane space where it binds to the proton on the acidic side of the inner membrane to re-form usnic acid which can then diffuse back into the matrix. The resulting cycle causes proton leakage that eventually can dissipate the proton gradient across the inner membrane, disrupting the tight coupling between electron transport and ATP synthesis. This model would explain the absence of gene-expression changes associated with complex V in usnic acid-treated mice, despite the increased electron transport by complexes I – IV. It may also explain the decline in ATP level in spite of increased oxygen consumption.

In Table 3, only the two-sided correction method enables us to explain the function of usnic acid as described above. The one-sided correction method gives *p*-values similar to those in the two-sided correction method, but this is likely to be due to most of the gene expression changes entailing up-regulation. When the direction of gene expression change due to a treatment is known, then the one-sided correction method is the appropriate choice; it also needs less computation time than the two-sided correction method which employs Monte Carlo sample generation.

Discussion

In many studies that use microarray data, the number of samples is small as in the first example shown above. While the number of samples in the simulations is only 3 for each group, the distribution of corrected *p*-values

Table 2: Effects of AZT and 3TC on oxidative phosphorylation and apoptosis.

gene group	# genes	up	# P<0.05	Fisher's exact	Not corrected	One-sided correction	Two-sided correction
Complex1	29	18	10	0.460	7.41E-6	0.600	0.049
Complex2	3	2	1	0.688	0.017	0.828	0.043
Complex3	7	5	3	0.401	5.3E-5	0.559	0.002
Complex4	13	8	8	0.026	1.25E-07	0.641	0.001
Complex5	14	6	6	0.273	0.0003	0.982	0.022
apoptosis	18	10	7	0.347	2.92E-5	0.592	0.005

P-values calculated from four methods are presented for a comparison. The number of genes, up-regulated genes, and significantly expressed genes in each gene group are presented

Table 3: Effects of usnic acid on phosphorylation and apoptosis.

gene group	# genes	up	# P<0.05	Fisher's exact	Not corrected	One-sided correction	Two-sided correction
Complex1	37	31	12	0.517	1.32E-12	0.027	0.022
Complex2	3	2	1	0.680	0.035	0.029	0.014
Complex3	7	5	2	0.704	0.019	0.055	0.042
Complex4	18	17	11	0.008	7.08E-10	0.006	0.003
Complex5	17	13	4	0.838	0.001	0.044	0.051
apoptosis	19	14	11	0.014	8.05E-10	0.008	0.0004

approximates a uniform distribution. The estimation for the correlation, $\hat{\mathbf{R}}$, might not be very close to the true correlation, $\mathbf{R}$. However, the correction methods that depend only on the average correlation, $\bar{r}$, are more robust because the estimation of $\bar{r}$ is more robust.

For the one-sided test, the correction for the correlation depends only on $\bar{r}$, the average of off-diagonal elements of the correlation matrix. The corrections using $\bar{\mathbf{R}}$ or $\hat{\mathbf{R}}$ are equivalent for the one-sided test method. Important points in choosing a correlation estimate for the two-sided test are the following; 1) The correlation estimate should be robust for small sample sizes, and 2) The correlation estimate should preserve $\bar{r}$. $\bar{\mathbf{R}}$ satisfies these two conditions.

When the direction of gene expression change is pre-specified, the one-sided test is a good choice since it is easy and fast to calculate p-values. However, the two-sided test is the one we have to use in most cases, because it is usually not possible to pre-specify how individual genes will respond to treatment in the exploratory context. When we have a small number of samples to estimate the correlation, the $\bar{\mathbf{R}}$ method gives a robust result. Since $\bar{\mathbf{R}}$ misrepresents the true correlation, and gives biased p-values, the $\hat{\mathbf{R}}$ method works better for larger sample sizes. This is seen in Figure 3, where the $\hat{\mathbf{R}}$ method is better than the $\bar{\mathbf{R}}$ method. We hesitate to present a specific threshold sample size as sufficient for a converged correlation estimate, since it varies with respect to several conditions, such as the number of genes in a group, the variation of the elements of the correlation matrix, etc. The simulation result in Figure 3 shows that both methods give quite accurate p-values compared to the p-values from the true correlation matrix, $\mathbf{R}$. The Hotelling T^2 test is the best choice whenever it is applicable.

In Table 1, the distribution of animals from different treatment groups (A-E) in three batches (1–3) gives no per-mutable pairs. In this case, randomization methods are not applicable. Even though randomization methods inherently take into account the correlation structure among genes, they may not be practical when the design of the experiment is complicated and the number of samples per group is small, reducing the numbers of permutable pairs.

Randomization methods that permute class labels can adjust p-values for the correlation structure among genes. Randomization methods choose a summary statistic (e.g. enrichment score (ES) in [16], average z-score in [17]), which reflects the degree to which a set of genes is enriched. When the significance of the summary statistic is measured by permuting class labels, the method preserves gene-gene correlations and when applicable, would give similar result to the presented method. Randomization methods that permute gene labels, such as Fisher's exact test, do not preserve the correlation structure and misrepresent the group's significance.

Conclusion
We have presented a method to test the significance of expression changes within a group of genes, while considering the correlation structure among genes in each group. This method will enable the rapid detection of microarray evidence indicating altered cell functions or pathways, and will facilitate the interpretation of microarray outcomes. Application of the method to real data shows that it is an improved, practical method to evaluate the effects of treatments on functional classes of genes such as those based on Gene Ontology descriptors.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
TL conducted the simulations, analyzed the data and wrote the manuscript. VGD conducted the microarray experiments using Mitochip and made the biological interpretations. CV reviewed the literature. RJSR gave valuable suggestions on the preparation of the manuscript. RRD directed the methodology development, data analy-

sis, and manuscript preparation. All authors read and approved the final manuscript.

References

1. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and power-ful approach to multiple testing.** *Journal of the Royal Statistical Society Series B* 1995, **57:**289-300.
2. Storey JD, Taylor JE, Siegmund D: **Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach.** *Journal of the Royal Statistical Society, Series B* 2004, **66:**187-205.
3. Allison DB, Gadbury GL, Heo M, Fernandez JR, Lee C-K, Prolla TA, Weindruch R: **A mixture model approach for the analysis of microarray gene expression data.** *Computational Statistics and Data Analysis* 2002, **39:**1-20.
4. Khatri P, Draghici S: **Ontological analysis of gene expression data: current tools, limitations, and open problems.** *Bioinformatics* 2005, **21:**3587-3595.
5. Draghci S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA: **Global functional profiling of gene expression.** *Genomics* 2003, **81:**98-104.
6. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: **Discovering statistically significant pathways in expression profiling studies.** *PNAS* 2005, **102:**13544-13549.
7. Pavlidis P, Qin J, Arango V, Mann JJ, Sibille E: **Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex.** *Neurochemical Research* 2004, **29:**1213-1222.
8. Kong SW, Pu WT, Park PJ: **A multivariate approach for integrating genome-wide expression data and biological knowledge.** *Bioinformatics* 2006, **22:**2373-80.
9. Delongchamp RR, Velasco C, Desai VG, Lee T, Fuscoe JC: **Designing toxicogenomics studies that use DNA array technology.** *Bioinformatics and Biology Insights* 2007 in press.
10. Delongchamp RR, Lee T, Velasco C: **A method for computing the overall statistical significance of a treatment effect among a group of genes.** *BMC Bioinformatics* **7(Suppl 2):**S11. 2006 Sep 6
11. Desai VG, Lee T, Delongchamp RR, Moland CL, Branham WS, Fuscoe JC, Leakey JEA: **Development of mitochondria-specific mouse oligonucleotide microarray and validation of data by real-time PCR.** *Mitochondrion* 2007, **7(5):**322-9.
12. Desai VG, Lee T, Delongchamp RR, Leakey JEA, Lewis SM, Lee F, Moland CL, Branham WS, Fuscoe JC: **NRTI-induced expression profile of mitochondrial genes in the mouse liver.** *Mitochondrion* 2008, **8(2):**181-195.
13. Joseph A, Lee T, Moland CL, Branham WS, Fuscoe JC, Leakey JEA, Allaben W, Lewis SM, Ali AA, Desai VG: **Effect of usnic acid on mitochondrial functions as measured by mitochondria-specific oligonucleotide microarray in liver of B6C3F$_1$ mice.** *Biochemical Pharmacology* 2008 in press.
14. Cherry CL, Wesselingh SL: **Nucleoside analogues and HIV: the combined cost to mitochondria.** *J Antimicrob Chemother* 2003, **51:**1091-1093.
15. Barret B, Tardieu M, Rustin P, Lacroix C, Chabrol B, *et al.*: **For the French perinatal cohort study group. Persistent mitochondrial dysfunction in HIV-1-exposed but uninfected infants: clinical screening in a large prospective cohort.** *AIDS* 2003, **17:**1769-1785.
16. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **A knowledge-based approach for interpreting genome-wide expression profiles.** *PNAS* 2005, **102:**15545-15550.
17. Efron B, Tibshirani R: **On testing the significance of sets of genes.** *Tech report* 2006 [http://www-stat.stanford.edu/~tibs/GSA/].