Appl. Statist. (2003)
52, *Part* 3, *pp.* 365–376

Randomization tests for small samples: an application for genetic expression data

Gary L. Gadbury,

University of Missouri—Rolla, USA

Grier P. Page,

University of Alabama at Birmingham, USA

Moonseong Heo

Cornell University, White Plains, USA

and John D. Mountz and David B. Allison

University of Alabama at Birmingham, USA

[Received May 2002. Final revision January 2003]

Summary. An advantage of randomization tests for small samples is that an exact *P*-value can be computed under an additive model. A disadvantage with very small sample sizes is that the resulting discrete distribution for *P*-values can make it mathematically impossible for a *P*-value to attain a particular degree of significance. We investigate a distribution of *P*-values that arises when several thousand randomization tests are conducted simultaneously using small samples, a situation that arises with microarray gene expression data. We show that the distribution yields valuable information regarding groups of genes that are differentially expressed between two groups: a treatment group and a control group. This distribution helps to categorize genes with varying degrees of overlap of genetic expression values between the two groups, and it helps to quantify the degree of overlap by using the *P*-value from a randomization test. Moreover, a statistical test is available that compares the actual distribution of *P*-values with an expected distribution if there are no genes that are differentially expressed. We demonstrate the method and illustrate the results by using a microarray data set involving a cell line for rheumatoid arthritis. A small simulation study evaluates the effect that correlated gene expression levels could have on results from the analysis.

Keywords: Additivity; Microarray; Nonparametric test; Permutation; Randomization

1. Introduction

Microarrays are a recently developed technique that allows simultaneous measurement of gene expression on thousands of genes from an organism. Often, the interest is in determining whether genetic expression patterns are different for two or more groups of organisms that differ with respect to exposure to some environmental stimuli, genotype, age or some other characteristic. Various techniques have been proposed to quantify the difference in expression between groups for a single gene (see Kerr *et al.* (2000), Ideker *et al.* (2000), Tusher *et al.* (2001) and Allison *et al.* (2002)). Statistical tests for a difference that rely on parametric assumptions or the central

Address for correspondence: Gary L. Gadbury, Department of Mathematics and Statistics, University of Missouri—Rolla, Rolla, MO 65409, USA.
E-mail: gadburyg@umr.edu

 0035–9254/03/52365

limit theorem may be inappropriate for small samples. Small samples are common in microarray studies (e.g. Lee *et al.* (1999)).

This paper builds on an earlier work by Allison *et al.* (2002) who used a mixture of a uniform distribution and beta distributions to model a continuous distribution of *P*-values from statistical tests on differential gene expression between two experimental groups. This mixture model was used to answer various questions regarding microarray data analysis. First was a hypothesis that we call the 'global null hypothesis', stating that no genes are differentially expressed. This question was of initial interest to biologists with whom we were working so that the 'amount' of activity (differential gene expression) could somehow be quantified. If this global null hypothesis were rejected, then the mixture model could be used to quantify a false positive and false negative probability for any gene by using Bayesian posterior probability calculations.

Allison *et al*. (2002) relied on parametric inference techniques to produce a *P*-value, but the assumptions on which these tests rely may not always be met. A natural follow-up question regarded the use of 'nonparametric' randomization tests. However, these tests were initially considered inappropriate for small samples owing to low power related to the discreteness of the randomization distribution. Various methods to correct for multiple comparisons (Hochberg and Tamhane, 1987; Sidak, 1967; Zaykin *et al.*, 2002) often require 'significant' *P*-values that are impossible to achieve when using a discrete randomization distribution with a small number of outcomes.

Still, randomization tests have certain advantages and should not be considered to be unsuitable for small data sets such as occur in microarray studies. These tests eliminate some parametric assumptions and allow computation of exact *P*-values under the unit–treatment additivity assumption (Cox, 1992). This assumption states that units' outcomes under treatment are offset by a constant (often 0) from the same units' outcomes under control (see Neyman (1935) and Fisher (1935) for some interesting history concerning this assumption). Another advantage of randomization tests is that a random treatment assignment produces the probability mechanism that generates the notion of a *P*-value. A random sample from a larger, possibly infinite, population is not necessary.

In the context of genetic expression data from microarrays, we show that the discrete distribution of *P*-values from randomization tests on all genes can yield useful information to categorize genes into groups that are differentially expressed. This distribution is easy to interpret, and there is a simple test to compare this distribution with one that would be 'expected' if no genes were differentially expressed.

Thus, the technique that is reported here is an alternative 'sensitivity check' of the global null hypothesis that was initially considered by Allison *et al.* (2002). It can serve as an initial screen of a microarray experiment and may be best suited for experiments with possibly large numbers of genes that are differentially expressed. Moreover, it complements a number of other methods that use various metrics against some reference distribution to rank the genes that are 'most differentially expressed' and to quantify measures of false discovery (positive) rates FDR (see Benjamini and Hochberg (1995) or Storey (2002) for some background information on FDR). Mentioned earlier, Allison *et al.* (2002) used their fitted mixture model and Bayesian calculations to obtain the probability that a gene is a false positive when declared significant at a particular threshold (or a false negative if not declared). Tusher *et al.* (2001), in their statistical analysis of microarray methodology, used a permutation procedure to compute a reference distribution of a modified *t*-statistic and used this distribution to estimate FDR. Xu *et al.* (2002) used a method that is similar in concept to that of Tusher *et al.* (2001). van de Wiel (2003) has suggested that statistical analysis of microarray methodology may be less suited to situations when a large number of genes are differentially expressed and has proposed the use of a rank score. Lee and

Whitmore (2002) considered an analysis-of-variance model and evaluated sample size effects on power and type I error risk. The method reported herein does not deal directly with FDR or power but, instead, focuses on the global null hypothesis as an initial screen of a microarray experiment.

In the next section, we describe a case-study involving a microarray experiment utilizing a cell line from a subject with rheumatoid arthritis. The cell line was randomly divided into two groups of three units with each group receiving different 'treatments'. We then review some foundations of randomization-based inference as it pertains to genetic expression data from microarrays. We illustrate a randomization test to evaluate a global null hypothesis stating that there are no differences in gene expression between two groups. Next, we consider the effect that correlated gene expression levels might have on results and we conclude with a discussion.

2. Description of case-study

Microarrays for the measurement of gene expression levels became popular in the late 1990s (Nelson, 1996). These gene expression microarrays measure the level (or, more accurately, the relative level) of ribonucleic acid (RNA) abundance in a tissue sample from an organism. Although there are many types of microarrays, the two most common types are oligonucleotide arrays and spotted complementary deoxyribonucleic acid (DNA) arrays. In the case of oligonucleotide arrays, these are generally commercially prefabricated silicon 'chips' on which are preprinted oligonucleotide sequences for pieces of many genes. Oligonucleotide arrays tend to be expensive and, as they are a relatively new technology, little discretionary income has been allocated to their use. Obtaining tissue samples for certain microarray experiments can also be expensive. So studies using these arrays tend to rely on very small sample sizes (e.g. 3–5 cases per group). Complementary DNA arrays tend to be less expensive once they are established in a laboratory, but the initial set-up costs and the effort to achieve high quality and standardization can be large. Regardless of the type of microarray that is used, because of the large number of genes typically on an array (e.g. 1000–40000) and the small number of cases available, unique statistical challenges arise (Allison and Coffey, 2002).

Oligonucleotide arrays were used in a study of signalling in a human rheumatoid arthritis synovial fibroblast (RASF) cell line data set. RASFs are abnormal cells that are found in the synovium of individuals with rheumatoid arthritis (RA) and are associated with inflammation of the joints (Mountz *et al.*, 2001; Mountz and Zhang, 2001; Franz *et al.*, 2000). These cells can rapidly grow and divide, and they produce enzymes that can invade cartilage and bone around the joints (Tomita *et al.*, 2002; Wernicke *et al.*, 2002). As a result, RASFs are central to the pathology of RA. Tumour necrosis factor TNF-α, a cytokine (a class of secreted proteins that can stimulate a cell to grow or differentiate), is known to contribute to the development of RASFs (Mountz and Zhang, 2001; Handel *et al.*, 1995; Miyazawa *et al.*, 1998) and is a known cause for RASF cells to grow and divide and, thus, to increase the progression of RA (Handel *et al.*, 1995; Miyazawa *et al.*, 1998). Drugs that inhibit TNF-α are effective treatments for RA. TNF-α binds to the TNF-α receptor which is expressed on RASFs and causes phosphorylation and proteasome degradation of the inhibitor of nuclear factor κB (NF-κB), I-κB. This leads to the release and nuclear translocation of NF-κB that is sequestered in the cytoplasm of RASFs by I-κB. These events can be inhibited by transfecting the RASF with an adenovirus that contains a dominant negative mutant form of I-κB that cannot be phosphorylated. This mutant is AdI-κB-dominant negative (AdI-κB-DN).

The experiment that produced the case-study data sought to investigate changes in gene expression in response to the application of TNF-α in normal RASF cells transfected with a control vector compared with RASF cells where AdI-κB-DN has blocked the action of TNF-α. The hypothesis was that some of these differentially expressed genes are involved in the initial development of RASF cells or that they would be good targets for drugs to downmodulate the inflammatory activity or to increase the susceptibility of RASF to apoptotic (programmed cell death) signalling.

The data were produced from an experiment involving an RASF cell line. This cell line was composed of cells that had initially been obtained from a human and that had been changed so that they will grow outside humans (immortalization). Six samples were taken from the cell line. The AdI-κB-DN was added to three randomly selected samples with the other three receiving a control construct. After 15 h, TNF-α was added to all six samples. After a further 3 h, the RNA was extracted from the six samples. This 3-h time point resulted in intact RNA, in normal 2/1 ratio of the 28S and 18S ribosomal bands and no evidence of degradation (the data are not shown). The RNA was then labelled and run in the Affymetrix (Santa Clara, California) Hu95Av1 microarray. Each microarray chip was scanned on the same scanner by using Affymetrix gene scan analysis software, version 4. The result was a measure of genetic expression for 12625 genes on each of the six samples. Total RNA was prepared from 1×10^6 RASFs by the trizol extraction method (Gibco BRL, Rockville, Maryland). The RNA was quantitated by the optical density at 260 and 280 nm wavelength and adjusted to 100 mg ml^{-1}. Thus far, there is no consensus for the best normalization method (Hoffmann *et al.*, 2002). We performed mean normalization (Colantuoni *et al.*, 2002) for simplicity and to avoid losing information that could result from quantile–quantile normalization and similar methods.

3. Randomization tests for microarray data

Let Y_{ij}^k be a measure of genetic expression for the ith gene of the jth experimental unit to the kth treatment, $i = 1, \ldots, M$, $j = 1, \ldots, N$ and $k = t, c$. We initially assume that, for the ith gene, the bivariate potential outcomes (Rubin, 1974) $(Y_{ij}^t,\ Y_{ij}^c)$ represent a sample (not necessarily random) of size N, i.e. $j = 1, 2, \ldots, N$, from a superpopulation distribution (Sugden and Smith, 1984) $G_i(y^t, y^c)$ that is continuous. Once experimental units have been selected for a study, the $N \times 2$ matrix of bivariate potential outcomes for the ith gene are considered fixed, and $G_i(y^t, y^c)$ is no longer of immediate interest. The assumption that $G_i(y^t, y^c)$ is continuous provides that, for any ith gene, $(Y_{ij}^t,\ Y_{ij}^c)$ contain no ties for $j = 1, \ldots, N$.

Since a unit will receive, through random assignment, either the treatment t or the control c, only one of the two outcomes Y_{ij}^t or Y_{ij}^c is observable for the jth unit, depending on the treatment assignment outcome for that unit. The other outcome may be considered missing at random. Conceptually the 'true' individual effect of the treatment for the jth unit on the ith gene is defined as $D_{ij} = Y_{ij}^t - Y_{ij}^c$. The vector of true treatment effects (length N) on the ith gene is denoted by $D_i = Y_{i.}^t - Y_{i.}^c$. The quantity $\bar{D}_i = \bar{Y}_{i.}^t - \bar{Y}_{i.}^c$ is the true average effect of the treatment on the ith gene for the fixed set of experimental units, where

$$\bar{Y}_{i.}^t = \frac{1}{N} \sum_{j=1}^{N} Y_{ij}^t$$

and

$$\bar{Y}_{i.}^c = \frac{1}{N} \sum_{j=1}^{N} Y_{ij}^c.$$

Let $Z = (Z_1, \ldots, Z_N)'$ be a vector of length $N = 2n$ consisting of n 1s and n 0s, and let Ω be a collection of all possible $B = \binom{2n}{n}$ such vectors. We assume that $P(Z = z) = 1/B$ for all outcomes $z \in \Omega$. The estimated average treatment effect for the ith gene is

$$\bar{d}_i = \frac{1}{n} \sum_{j=1}^{2n} Y_{ij}^t Z_j - \frac{1}{n} \sum_{j=1}^{2n} Y_{ij}^c (1 - Z_j)$$

where $Z_j = 1$ indicates unit j received treatment t, and $Z_j = 0$ otherwise. The Z_j are the random components in $\bar{d}_i$.

Randomization tests are often applied under an additive model, $Y_{ij}^t = Y_{ij}^c + \tau_i$, where τ_i is a constant across all units, $j = 1, \ldots, N$. If the treatment effect varied with unit at a particular gene, it is said that unit–treatment interaction is present at that gene and this interaction is not identifiable in observable data without assumptions (Gadbury and Iyer, 2000). However, τ_i is allowed to vary across genes. The hypothesis test for the ith gene is

$$\begin{aligned} H_0^{(i)} &: \tau_i = 0, \\ H_a^{(i)} &: \tau_i \neq 0. \end{aligned} \tag{1}$$

Under the null hypothesis, the randomization distribution of $\bar{d}_i$ is discrete with B distinct outcomes (since $G_i(y^t, y^c)$ was assumed to be continuous) each with mass $1/B$. Let d_i^* be the observed value of $\bar{d}_i$ from the experiment. A two-tailed P-value for the above hypothesis test is defined to be

$$p_i = 2 \ \min\{P(\bar{d}_i \geqslant d_i^*), \ P(\bar{d}_i \leqslant d_i^*)\}. \tag{2}$$

Equal numbers in each treatment group assure this P-value to be exact. Otherwise, it may be conservative owing to possible skewness in the randomization distribution. More discussion of this issue is in Lehmann (1994), chapter 3, and in Cox (1977). A p_i is computed as twice the number of $\bar{d}_i$ in the randomization distribution that are equal to or more extreme than was observed—more extreme being those $\bar{d}_i$ in the tail of the distribution—divided by B.

The smallest $p_i, i = 1, \ldots, M$, will occur when the observations for the treatment group are separated from those of the control group. More formally, the two groups are separated at the ith gene if $Y_{ij}^t > Y_{ij'}^c$ or if $Y_{ij}^t < Y_{ij'}^c$ for every unit j assigned treatment and every unit j' assigned the control. Larger P-values occur when expression levels of the two groups have varying degrees of overlap. A P-value equal to 1 will occur when there is complete overlap, i.e. the expression levels are indistinguishable, as measured by $\bar{d}_i$, between the two groups.

Since p_i is a function of the data, it is a random variable (Sackrowitz and Samuel-Cahn, 1999). If the null hypothesis in expression (1) is true, then p_i takes values $2v/B$ where $v = 1, \ldots, B/2$ with equal probabilities $2/B$. We write $p_i \sim F(p)$, a discrete uniform distribution, under $H_0^{(i)}$.

Consider the global null hypothesis $H_0 : \tau_1 = \tau_2 = \ldots = \tau_M = 0$. The alternative is that there is at least one $\tau_i \neq 0$. For now we assume that, under H_0, the $p_i, i = 1, \ldots, M$, represent a random sample from $F(p)$. We later consider the effect of correlated expression levels in Section 4.

Let $p_1^*, \ldots, p_M^*$ be the observed values of the random sample, i.e. the actual P-value from the randomization test for each gene. Define the following $B/2$ statistics:

$$O_v = \sum_{i=1}^{M} I_{\{2v/B\}}(p_i^*), \qquad v = 1, \ldots, B/2,$$

where the term $I_{\{b\}}(a)$ is equal to 1 when $a = b$ and to 0 otherwise. O_v is a count of the number of p_i^* equal to $2v/B$. If hypothesis H_0 is true, the expected number for O_v will be $2M/B$ for all $v = 1, \ldots, B/2$. Thus, the null distribution of $(O_1, \ldots, O_{B/2-1})$ can be modelled using a

multinomial distribution with parameters M and equal probabilities $\pi_1 = \ldots = \pi_{B/2-1} = 2/B$. A test of H_0 can be constructed by using the test statistic

$$x = \sum_{v=1}^{B/2} \frac{(O_v - 2M/B)^2}{2M/B},$$

which is approximately a realization from a χ^2-distribution with $B/2 - 1$ degrees of freedom. Suppose that $X \sim \chi^2_{B/2-1}$. An approximate P-value of a test of the global null hypothesis H_0 is defined as $P(X \geqslant x)$.

4. Analysis of case-study

4.1. Results

For each gene, there are $B = 20$ outcomes in the randomization distribution. The distribution of two-tailed P-values will contain $(0.1, 0.2, \ldots, 1)$. When a hypothesis $H_0^{(i)}$ is true, each of these values occurs with probability 0.1. The numbers of P-values that equal each of these 10 values (out of $M = 12625$) are

$$(O_1, O_2, \ldots, O_{10}) = (2975, 1457, 1261, 1137, 1058, 1002, 934, 931, 957, 913). \tag{3}$$

Under H_0, we would expect each of these values to be 1262.5, or

$$(O_1, O_2, \ldots, O_9) \sim \text{multinomial}(12625, \pi_1 = \ldots = \pi_9 = 0.1).$$

The χ^2 test statistic from the case-study data is 2795 and is approximately a realization from a χ^2-distribution with 9 degrees of freedom. The P-value for the test of the global null hypothesis is nearly 0, indicating that hypothesis H_0 would be rejected. The observed counts in expression (3) show that the largest departure from a uniform distribution occurs at P-values equal to 0.1, the smallest possible value. We would expect, under H_0, only around 1263 P-values to fall in this category when, in fact, there were 2975. The proportion of genes with this property was 0.235 *versus* 0.1, the expected proportion under H_0. This technique has categorized genes into genes with expression levels of varying degrees of overlap, and it has suggested that there is strong evidence in the data that H_0 is not true.

If those genes with no overlap in expression levels between the two groups are of interest, we could now define a metric to quantify the distance of separation for those 2975 genes and then use this metric to sort genes by order of importance. The genes with the largest distance of separation may be ones that are most differentially expressed between the two treatment groups. Possible metrics would be a standardized mean difference, a ratio of means or a metric defined by Hodges and Lehmann (1963), among others. We chose the absolute value of a standardized mean difference using pooled variance across the two groups, although any meaningful metric can be used at this point. The absolute standardized mean difference for the 2975 genes with a P-value equal to 0.1 are plotted against the original number of the gene in the data matrix in Fig. 1.

The biological details of this experiment, and the biological interpretation of results, will be reported elsewhere (Zhang *et al.*, 2003), but for illustration we highlight a few interesting genes. If the genes in Fig. 1 were ranked from 1 (largest absolute standardized mean difference) to 2975 (smallest absolute standardized mean difference), certain genes ranked near the top were expected to be differentially expressed *a priori* of the experiment. There are genes associated with apoptosis (GGC-2 (rank 6), hIAP (rank 21) and DAXX (rank 35)) and several cytokines (GRO1 (rank 1), GRO2 (rank 19), Il-8 (rank 14), Il-15 (rank 3) and RANTES (rank 27)).

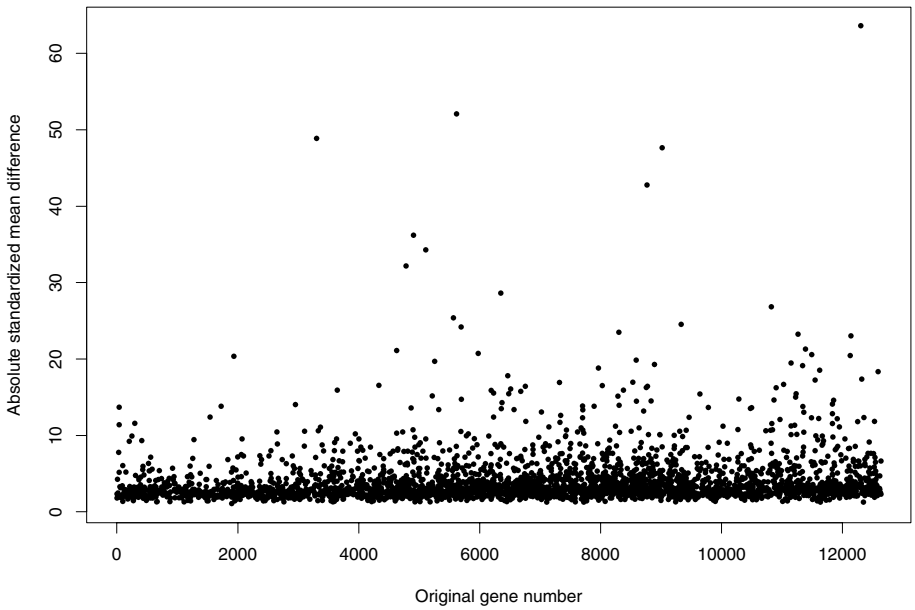

Fig. 1. Absolute standardized mean difference *versus* original gene number for the 2975 genes with a randomization *P*-value equal to 0.1

4.2. The effect of correlated expression levels

Simulations were used to assess the effect of correlated expression levels on the test of hypothesis H_0. Then, an empirical study of the $(12625)(12624)/2 = 79689000$ correlations in the case-study data was conducted and the results were compared with those of a simulated data set with a specified correlation structure.

4.2.1. Simulation study

A description of the simulation study follows.

(a) The data set from the case-study was used as a model for the mean and variance structure of simulated data. Six independent observations were generated from a 12625-dimensional multivariate normal distribution. The mean vector of length 12625 of this normal distribution had the same means as the mean of all six samples in the case-study data set, so there would be no mean difference in the generated data, except by chance. The 12625 variances in the normal distribution were obtained using the sample variance in the case-study data, at each gene, over all six samples.

(b) Correlation was incorporated similarly to the simulations described in Allison *et al.* (2002). A 500×500 equicorrelation matrix with 1 on the diagonal and ρ in the off-diagonal entries was used. Denote this matrix Σ_{500}. The 12625 square correlation matrix was block diagonal with Σ_{500} as the blocks and 0s everywhere else. So, genes within a block were correlated. The last block was of size 125 to match the size of the case-study data matrix.

(c) *P*-values were calculated using the randomization test on the simulated data set.

(d) The above steps were repeated 500 times. So, for each of the 10 possible *P*-values, there were 500 'counts' representing the number of *P*-values (a count out of 12625) falling into

each of the 10 categories. The result is a sampling distribution of 500 of these counts for each of 10 P-values.

(e) The mean and standard deviation of the 500 counts were recorded for each of the 10 P-values.

(f) The above steps were conducted at values of $\rho = 0, 0.3, 0.6$ and the results are shown in Table 1.

Table 1 shows two primary results. The first is that, as expected, the sampling distributions of counts are centred nearly on the expected value under the null hypothesis. However, the standard deviations of the distributions are larger for P-values near 0.1 or 0.2 or on the other extreme, 1.0, when there is dependence between genes. The second result is that the observed number of P-values from the analysis of the case-study data that are equal to 0.1 is 2975, and this number is far greater than all simulated sampling distributions at that P-value regardless of the value of ρ. An approximate 99% confidence interval for the number of P-values equal to 0.1, based on the case-study data, is (2852, 3098), indicating that the entire interval estimate is more than 5 standard deviations above the mean of the corresponding sampling distributions, including when $\rho = 0.6$. In the previous subsection, we rejected hypothesis H_0 and the simulation results provide some support for the claim that some genes are clearly differentially expressed between the two groups and it would take a rather unusual and extreme dependence between genes to explain this difference otherwise.

4.2.2. Empirical correlation study

Though we cannot estimate a (12625×12625)-dimensional correlation matrix from observed data on six units, we can compute a number of pairwise correlations to determine whether any apparent structure was present and to what extent this structure marked a departure from independence between genes. To do this, the group means were removed from each group, a pair (i, j) of genes were randomly selected such that $i < j$ and the sample correlation was computed. This sample correlation represented a point estimate of the (i, j) entry of the (12625×12625)-dimensional correlation matrix. Approximately 1% (797342) of pairs were sampled. This was

Table 1. Means and standard deviations (in parentheses) of simulated sampling distributions of the number of P-values (out of 12625) equal to $0.1, 0.2, \ldots, 1.0$ for values of $\rho = 0.0, 0.3, 0.6$ and the observed counts from the case-study

P-value	*Observed counts*	*Means for the following values of* ρ*:*		
		$\rho = 0.0$	$\rho = 0.3$	$\rho = 0.6$
0.1	2975	1262.3 (35.3)	1265.5 (136.7)	1246.8 (309.3)
0.2	1457	1264.2 (35.8)	1266.1 (74.5)	1252.7 (158.0)
0.3	1261	1263.7 (32.8)	1263.4 (52.3)	1252.8 (107.1)
0.4	1137	1261.7 (34.1)	1260.6 (36.7)	1261.6 (77.7)
0.5	1058	1261.2 (31.6)	1261.9 (40.5)	1264.9 (65.5)
0.6	1002	1261.3 (32.6)	1260.3 (44.0)	1267.9 (76.6)
0.7	934	1264.2 (33.4)	1262.1 (52.6)	1267.9 (95.1)
0.8	931	1263.8 (31.8)	1261.8 (58.7)	1271.1 (112.4)
0.9	957	1261.3 (33.8)	1258.9 (62.7)	1268.9 (130.4)
1.0	913	1261.1 (31.4)	1264.4 (64.8)	1270.4 (136.4)

Table 2. Summary percentiles for the distribution of 797342 pairwise sample correlations for four cases

Case	*Values of the following percentiles:*				
	5th	*25th*	*50th*	*75th*	*95th*
Actual data	−0.841	−0.445	0.003	0.448	0.842
Simulated, $\rho = 0.0$	−0.806	−0.404	0.0001	0.404	0.805
Simulated, $\rho = 0.3$	−0.800	−0.389	0.021	0.424	0.815
Simulated, $\rho = 0.6$	−0.798	−0.385	0.029	0.434	0.825

repeated for a simulated data set with $\rho = 0$, 0.3, 0.6 to compare the actual data set with simulated data with known correlation structures. So four cases were considered: the actual data and three simulated data sets. The 5th, 25th, 50th, 75th and 95th percentiles of the resulting four distributions are shown in Table 2.

The results in Table 2 suggest that there is no overall systematic pattern of correlation in the case-study data, but there may be multiple different patterns of correlation, potentially some strongly positive and others strongly negative. Our simulations used only positive values of ρ. The design of the simulation did not allow negative values since the equicorrelation block diagonal matrix was not positive definite for even $\rho = -0.3$.

5. Discussion

In this paper we presented a method for analysing microarray data from an experiment in which six units from a cell line were randomly divided into two treatment groups of size three units each. The interpretation of the *P*-value from a randomization test helped to answer the question 'how unusual are observed results compared with what might have happened if the treatment assignment outcome had been different, assuming that the treatment had no effect?'. A *P*-value for a single gene is not too informative given that the smallest *P*-value that could be realized in the case-study is 0.1, a value that is not considered 'significant' by most standards. However, the number of *P*-values equal to 0.1 out of 12625 (all genes) was compared with an 'expected' number under a global null hypothesis. It was then evident that far more genes fell into this category than would be expected under this null hypothesis. If the category representing the smallest *P*-value was of most interest, we could simply compare the number of *P*-values in this category with what would have been expected if no genes were differentially expressed. The χ^2-test of hypothesis H_0 would then simplify to a test of a proportion.

We focused on permuting the mean differences between the two groups. This technique has a simple interpretation under unit–treatment additivity in that genes with little overlap of expression levels between the two groups produce the smaller *P*-values. Once these genes had been identified, we chose a standardized mean difference to rank them. It is important to note that if we had, instead, permuted *t*-statistics a different distribution of *P*-values would have resulted and a different set of genes may have been identified in the initial randomization test. Others have suggested that permuting *t*-statistics can sometimes lead to inappropriate results for genes with expression levels near background levels but with very small variances. As such, they have suggested adding a constant to the denominator of the *t*-statistic where the constant is computed using information borrowed from the variance of gene expression levels across all genes

(Tusher *et al.*, 2001; Efron *et al.*, 2001). Moreover, other nonparametric methods are available to produce a *P*-value for a gene such as the usual rank-based methods or the bootstrap (Efron and Tibshirani, 1993). How the different methods compare in terms of power and control of the type I error risk is a topic of on-going research (Brand *et al.*, 2003). How information from an initial screen (such as our test of the global null hypothesis) could be combined with the methods (such as those mentioned in Section 1) that quantify FDR for a set of genes is also an area of active interest. Finally, understanding the effect that correlated gene expression levels might have on the interpretation of quantities such as FDR or false negative probabilities remains to be explored.

The simulation study helped to reveal how correlated gene expression can affect the multinomial distribution of counts when there are no genes that are differentially expressed. The pattern shown in Table 1 suggests that the effect of certain correlated expression levels among genes on the distribution of *P*-values might be approached analytically. The correlation study also highlighted the difficulty in assessing the role that dependence plays when attempting to isolate specific genes that are differentially expressed between two groups, especially when the dependence structure is complex. We focused on a rather rigid dependence structure using equicorrelation matrices and normal distribution theory where the dependence is easily parameterized. An alternative would be to generate genetic expression data sequentially gene by gene by using a defined stochastic dependence structure. We have recognized, in earlier work (Allison *et al.*, 2002) and in this work, that the possibility of correlated expression levels is an important consideration, but it is often ignored in published methods to date.

The technique that is presented here has also been tested by using other data sets with equal numbers in each of two groups. Of course, the computing time increases as the group size increases. However, in one particular analysis with two groups of five units, there were 126 unique *P*-values in the exact randomization distribution, and the S-PLUS code that is required to compute exact *P*-values for 12625 genes executed in under 30 s on a laptop personal computer. As newer experiments using slightly larger samples are completed, the method here is sufficiently computationally efficient to use as an alternative to usual parametric procedures, and the code is relatively easy to implement in practice and is available from the authors.

Acknowledgements

This research was supported in part by National Institutes of Health grants R01AG011653, R01DK56366, P30DK56336, P20CA93753, U24DK058776, R01ES09912, P60AR048095, National Science Foundation grants 0090286 and 0217651, and a grant from the University of Alabama Health Services Foundation. We thank V. A. Samaranayake for helpful discussions on the initial aspects of this research, and we thank the Joint Editor, Associate Editor and referees for helpful suggestions that improved the manuscript.

References

Allison, D. B. and Coffey, C. S. (2002) Two stage testing in microarray analysis: what is gained? *J. Genont. Biol. Sci.*, **57**, B189–B192.

Allison, D. B., Gadbury, G. L., Heo, M., Fernandez, J. R., Lee, C. K., Prolla, T. A. and Weindruch, R. (2002) A mixture model approach for the analysis of microarray gene expression data. *Comput. Statist. Data Anal.*, **39**, 1–20.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

Brand, J. P. L., Gadbury, G. L., Beasley, T. M., Page, G. P., Long, J. D., Edwards, J. W. and Allison, D. B. (2003) A comparison of some non-parametric alternatives for inferential testing in microarray research. *Working Paper*.

Colantuoni, C., Henry, G., Zeger, S. and Pevsner, J. (2002) Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. *Biotechniques*, **32**, 1316–1320.

Cox, D. R. (1977) The role of significance tests. *Scand. J. Statist.*, **4**, 49–70.

Cox, D. R. (1992) Causality: some statistical aspects. *J. R. Statist. Soc.* A, **155**, 291–301.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Efron, B., Tibshirani, R. J., Storey, J. D. and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Statist. Ass.*, **96**, 1151–1160.

Fisher, R. A. (1935) Discussion on 'Statistical problems in agricultural experimentation' (by J. Neyman). *J. R. Statist. Soc.*, suppl., **2**, 154–157.

Franz, J. K., Pap, T., Hummel, K. M., Nawrath, M., Aicher, W. K., Shigeyama, Y., Muller-Ladner, U., Gay, R. and Gay, S. (2000) Expression of sentrin, a novel anti-apoptotic molecule, at sites of synovial invasion in rheumatoid arthritis. *Arth. Rheum.*, **43**, 599–607.

Gadbury, G. L. and Iyer, H. K. (2000) Unit-treatment interaction and its practical consequences. *Biometrics*, **56**, 882–885.

Handel, M. L., McMorrow, L. B. and Gravallese, E. M. (1995) Nuclear factor-kappa B in rheumatoid synovium: localization of p50 and p65. *Arth. Rheum.*, **38**, 1762–1770.

Hochberg, Y. and Tamhane, A. C. (1987) *Multiple Comparison Procedures*. New York: Wiley.

Hodges, J. L. and Lehmann, E. L. (1963) Estimates of location based on rank tests. *Ann. Math. Statist.*, **34**, 598–611.

Hoffmann, R., Seidl, T. and Dugas, M. (2002) Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol.*, **3**, research0033.1–research0033.11.

Ideker, T., Rhorsson, V., Siegel, A. F. and Hood, L. E. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.*, **7**, 805–817.

Kerr, M. K., Martin, M. and Churchill, G. A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Lee, C., Klopp, R. G., Weindruch, R. and Prolla, T. A. (1999) Gene expression profile of aging and its retardation by caloric restriction. *Science*, **285**, 1390–1393.

Lee, M. T. and Whitmore, G. A. (2002) Power and sample size for DNA microarray studies. *Statist. Med.*, **21**, 3543–3570.

Lehmann, E. L. (1994) *Testing Statistical Hypotheses*. New York: Chapman and Hall.

Miyazawa, K., Mori, A., Yamamoto, K. and Okudairi, H. (1998) Transcriptional roles of CCAAT/enhancer binding protein-beta, nuclear factor-kappa B and C-promoter binding factor 1 in interleukin (IL)-1 beta-induces IL-6 synthesis by human rheumatoid fibroblast-like synoviocytes. *J. Biol. Chem.*, **273**, 7620–7627.

Mountz, J. D., Hsu, H. C., Matsuki, Y. and Zhang, H. G. (2001) Apoptosis and rheumatoid arthritis: past, present and future direction. *Curr. Rheum. Rep.*, **3**, 70–78.

Mountz, J. D. and Zhang, H. G. (2001) Regulation of apoptosis of synovial fibroblasts. *Curr. Directns Autoimmun.*, **3**, 216–239.

Nelson, N. (1996) Microarrays pave the way to 21st century medicine. *J. Natn. Cancer Inst.*, **88**, 1803–1805.

Neyman, J. (1935) Statistical problems in agricultural experimentation (with discussion). *J. R. Statist. Soc.*, suppl., **2**, 107–180.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.

Sackrowitz, H. and Samuel-Cahn, E. (1999) P values as random variables—expected p values. *Am. Statistn*, **53**, 326–331.

Sidak, Z. (1967) Rectangular confidence regions for the means of multivariate normal distributions. *J. Am. Statist. Ass.*, **62**, 626–633.

Storey, J. D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc.* B, **64**, 479–498.

Sugden, R. A. and Smith, T. M. F. (1984) Ignorable and informative designs in survey sampling inference. *Biometrika*, **71**, 495–506.

Tomita, T., Nakase, T., Kaneko, M., Shi, K., Takahi, K., Ochi, T. and Yoshikawa, H. (2002) Expression of extracellular matrix metalloproteinase inducer and enhancement of the production of matrix metalloproteinases in rheumatoid arthritis. *Arth. Rheum.*, **46**, 373–378.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natn Acad. Sci.*, **98**, 5116–5121.

van de Wiel, M. A. (2003) Significance analysis of microarrays using rank scores. To be published. (Available from `http://www.eurandom.nl/Past%20years/reports/2002/005mwreport.pdf`.)

Wernicke, D., Schulze-Westhoff, C., Brauer, R., Petrow, P., Zacher, J., Gay, S. and Gromnica-Ihle, E. (2002) Simulation of collagenase 3 expression in synovial fibroblasts of patients with rheumatoid arthritis by contact with a three-dimensional collagen matrix or with normal cartilage when coimplanted in NOD/SCID mice. *Arth. Rheum.*, **46**, 64–74.

Xu, X. L., Olson, J. M. and Zhao, L. P. (2002) A regression-based method to identify differentially expressed genes in microarray time course studies and its application in an inducible Huntington's disease transgenic model. *Hum. Molec. Genet.*, **11**, 1977–1985.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H. and Weir, B. S. (2002) Truncated product method for combining p-values. *Genet. Epidem.*, **22**, 170–185.

Zhang, H., Hyde, K., Page, G. P., Brand, J. P. L., Allison, D. B. and Mountz, J. D. (2003) NF-κB regulated genes in RASF. To be published.