Technical Report 003
Aug 22, 2001

Controlling the number of false discoveries: Application to high-dimensional genomic data

Edward L. Korn[*], James F. Troendle[⊥], Lisa M. McShane[*], Richard Simon[*]

[*]Biometric Research Branch, National Cancer Institute, Bethesda, Maryland 20892, USA

[⊥]Biometry and Mathematical Statistics Branch, National Institute of Child Health and Human Development, Bethesda, Maryland 20892 USA

Edward L. Korn is Head, Clinical Trials Section, Lisa M. McShane is Mathematical Statistician, and Richard Simon is Chief, Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD 20892 (Email: *korne@ctep.nci.nih.gov*). James F. Troendle is Mathematical Statistician, Biometry and Mathematical Statistics Branch, National Institute of Child Health and Human Development, Bethesda, Maryland 20892. This study utilized the high-performance computational capabilities of the Biowulf/LoBoS3 cluster at the National Institutes of Health, Bethesda, MD.

Abstract

Detailed genetic characterizations of specimens from healthy or diseased individuals may hold the key to predicting which healthy individuals will develop disease or which diseased individuals will respond to therapy. For example, cDNA microarrays allow simultaneous measurement of expression levels of thousands of genes on a single specimen, producing a "gene expression profile". Frequently an objective of such a study is to identify which genes among the thousands measured are differentially expressed in one group as compared to another. Statistically, this presents an enormous multiple comparisons problem. Here we propose two new statistical procedures for controlling the number of spurious findings.

We analyze a microarray data set consisting of measurements on approximately 9000 genes in paired tumor specimens, collected both before and after chemotherapy on 20 breast cancer patients. Our interest was to identify genes that were differentially expressed after chemotherapy as compared to before chemotherapy.

A straightforward approach to the identification of differentially expressed genes is to perform a univariate analysis of group mean differences for each gene, and then identify those genes that are most statistically significant. Using nominal significance levels (unadjusted for the multiple comparisons) will lead to the identification of many genes that truly are not differentially expressed, "false discoveries". However, control of the familywise error rate (e.g., using the Bonferonni inequality) seems too extreme. Since the identified genes will be further studied for biologic relevance, a reasonable strategy is to allow a small number of false discoveries, or a small proportion of the identified genes to be false discoveries. Although previous work has considered control for the *expected* proportion of false discoveries, we show these methods may be inadequate.

We propose two stepwise permutation-based procedures designed to control the *actual* number or proportion of false discoveries with specified confidence.

Applying these new methods to the breast tumor microarray example, we were able to identify 28 genes (more than twice the number identified by a Bonferroni procedure) that we can state with high confidence are differentially expressed comparing before to after chemotherapy. In addition, simulation studies evaluate the procedures and demonstrate that their use results in substantial gain in sensitivity to detect truly differentially expressed genes even when allowing as few as one or two false discoveries. The methods described are broadly applicable to the problem of identifying which variables of any large set of measured variables differ between pre-specified groups.

KEY WORDS: False discovery; Multiple comparisons; Stepwise procedure; Permutation method; Gene expression; Microarray

1. INTRODUCTION

Technological advances have made possible detailed genetic characterization of biological specimens. For example, cDNA microarray technology (Schena, Shalon, Davis, and Brown 1995) permits the simultaneous evaluation of expression levels of thousands of genes on a single specimen, generating a gene expression "profile" for that specimen. The cDNA microarray technology has now been applied to profile numerous human cancer specimens; it is hoped that gene expression profiles of tumors might aid in distinguishing aggressive from indolent tumors and might guide choice of therapies. Another exciting opportunity comes with the completion of the initial sequencing and analysis of the human genome. More than 1.4 million single nucleotide polymorphisms (SNPs) in the human DNA sequence have been now identified (International Human Genome Sequencing Consortium 2001). Differing patterns of SNPs might be related to risk of developing disease or predict response to, or toxicity from, drug therapies. A typical experimental approach in these settings would be to select specimens from two or more groups it was desired to compare, and then to measure a large number of characteristics on each specimen. Often one would want to identify characteristics that univariately are significantly different between the groups. The issue we address in this paper is how one can identify individual characteristics that are significantly different between groups of specimens while maintaining control over spurious findings amid the potentially enormous number of comparisons being made.

The particular example we consider in this paper consists of gene expression profiles obtained by cDNA microarray analysis of approximately 9000 genes for 40 paired breast tumor specimens. The specimens were collected on 20 breast cancer patients, before and after chemotherapy. Our interest is in identifying genes whose expression levels differed significantly after chemotherapy as compared to before. The example data are continuous, log-transformed

expression ratios that measure the relative abundance of each gene's mRNA in the test specimen compared to a reference sample using a two-color fluorescent probe hybridization system (Schena et al. 1995). We note, however, that the general approaches described in this paper are very broadly applicable to both continuous and discrete data, and to censored data.

If one were to simply conduct univariate tests of characteristics, for example gene expression levels, using conventional significance levels, there would be an enormous multiple comparisons problem. Some genes would likely be claimed significantly differentially expressed when, in truth, they were not differentially expressed. Such false claims of significance are often called "false discoveries". One can use a procedure to account for these multiple comparisons and control the probability of any false discovery. This overall probability of any error is usually referred to as the familywise error (FWE) rate. A Bonferroni adjustment to the p-values, or preferably less conservative stepwise procedures such as those described by Westfall and Young (1993, pp. 72-74) and Troendle (1996) can be used to control the familywise error rate. For example, Callow, Dudoit, Gong, Speed, and Rubin (2000) have applied the Westfall and Young method to microarray data. These procedures will guarantee that the probability of any false discovery is less than the designated significance level, e.g., .05. However, the criterion of not making any false discovery is too stringent for most microarray investigations, in which the identification of these genes will be followed by further study of them. On the other hand, making no adjustment for multiple comparisons could generate many false leads.

A reasonable compromise is to use a procedure that will allow some false discoveries, but not too many. A simple procedure is to lower the nominal significance level and appeal to Bonferroni, e.g., using an significance level of .001 would ensure in expectation at most 10 false discoveries with 10,000 variables. A slightly more complex procedure attempts to control for the

expected proportion of discoveries (identified genes) that are false discoveries (with the proportion set to 0 when no genes are identified): Order the univariate p-values from the k variables, $P_{(1)} < P_{(2)} < ... < P_{(k)}$. To keep the expected false discovery proportion less than γ (e.g., $\gamma = .10$), identify as differentially expressed those genes that are associated with the indices 1, 2, ..., i, where i is the largest index satisfying $P_{(i)}k < i\gamma$. This procedure is attributed to Eklund by Seeger (1968) and was studied by Benjamini and Hochberg (1995). Tusher, Tibshirani, and Chu (2001) present a procedure they call SAM (Significance Analysis of Microarrays) for estimating a false discovery rate from data, but they do not discuss the statistical properties of their procedure. Procedures targeting control of the expected number or proportion of false discoveries rather than the actual number or proportion can give a false sense of security. This is demonstrated in Tables 1 and 2 for simulated data with 10,000 variables. In Table 1, we consider using a univariate nominal significance level of .001. The expected number of false discoveries is less than or equal to 10, but the spread of the distribution of the actual number of false discoveries becomes quite large when the correlation between the variables increases. For example, there is a 10% chance of having 18 or more false discoveries with block correlation .5. The same problem arises when controlling the expected false discovery proportion. Table 2 displays the distribution of the actual false discovery proportion when using the simple procedure described above to control the expected false discovery proportion to be less than .10. Even with no correlation the results here are troubling: 10% of the time the false discovery proportion will be .29 or more.

In this paper we discuss methods for controlling the actual (rather than expected) number and proportion of false discoveries. We prove that our procedure for the former guarantees control, and our procedure for the latter achieves asymptotic control. That is, application of these methods will allow statements such as "with 95% confidence, the number of false discoveries does not

exceed 2" or "with approximate 95% confidence, the proportion of false discoveries does not exceed .10". In section 2 we describe our procedures for controlling the number and proportion of false discoveries and provide justifications for the algorithms. The methods are applied to the analysis of the pre-post chemotherapy breast cancer specimens in section 3. In section 4, we describe some limited simulation studies to assess our procedures and to compare them to procedures designed to control the FWE error rate. We end with a discussion in section 5.

2. CONTROLLING FALSE DISCOVERIES

We assume that for each variable we have performed an appropriate univariate statistical test of the null hypothesis that the distribution of that variable is the same across the groups, and we have obtained a p-value associated with that test. For example, if the data consist of continuous gene expression levels, t-tests or Wilcoxon tests might be performed for each gene. If the data consist of binary variables, each indicating the presence or absence of a particular marker, then chi-squared tests or Fisher's exact tests may be used. The tests may be for paired data (as in our breast cancer example) or unpaired data depending on the design. Let

$$P_{(1)} < P_{(2)} < ... < P_{(k)} \qquad (1)$$

be the ordered p-values from the univariate tests based on the k variables, and let $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$ denote the hypotheses in the corresponding order. For now, we assume that there are no ties in the p-values (as would be the case if the variables were continuous and parametric hypothesis tests were used). For any subset $T = \{t_1, t_2, ..., t_j\}$ of $K = \{1,2,...,k\}$, consider the multivariate permutation distribution of p-values $\{P'_{t_1}, P'_{t_2}, ..., P'_{t_j}\}$ that would be generated by permuting the group labels on the specimens and calculating the test statistics on the variables with indices in T using the permuted data. For paired data, these permutations are obtained by switching the

characteristic profiles within each pair. Therefore, for our paired breast tumor example, permutations are performed by switching the before and after gene expression profiles. For the unpaired or multi-group case, permutations are performed by shuffling the group membership labels. Note that in each case, the characteristic profiles measured on any given specimen remain intact so as to preserve the correlation among the measured characteristics.

We now present two procedures. Procedure A is designed to control the number of false discoveries, and Procedure B is designed to control the proportion of false discoveries. Both Procedures A and B described below are "step-down" permutation methods. That is, they proceed from smallest p-value to largest, which is equivalent to "stepping down" from the largest test statistic to the smallest.

2.1 Controlling the Number of False Discoveries

Procedure A: Suppose we wish to be 1-α confident that the number of false discoveries is $\leq u$ ($u >$ 0). Let $y^{\alpha}_{T,u}$ denote the α quantile of the distribution of the $(u+1)$st smallest of the $\{P'_{t_1}, P'_{t_2}, ..., P'_{t_j}\}$ under the multivariate permutation distribution. (Since the permutation distribution is discrete, the exact α quantile will typically be unobtainable, so we let $y^{\alpha}_{T,u}$ be the largest obtainable α^* quantile with $\alpha^* < \alpha$.) We reject null hypotheses sequentially, in order of smallest to largest p-value as follows:

Automatically reject $H_{(1)}, H_{(2)}, \ldots, H_{(u)}$. For $r >$ u, having rejected $H_{(r-1)}$, reject $H_{(r)}$ if $P_{(r)} < y^{\alpha}_{r,u}$ where

$$y^{\alpha}_{r,u} = min(\, y^{\alpha}_{\{(i_1),(i_2),\ldots,(i_u),(r),(r+1),\ldots,(k)\},u} \,|\, \{i_1 < i_2 < \ldots < i_u\} \subset \{1, 2, \ldots, \text{r-1}\}), \qquad (2)$$

and subscript (i_j) denotes the index associated with the i_j-th ordered p-value from among the full set of p-values. Once a hypothesis is not rejected, all further hypotheses are not rejected.

For $u = 0$, we define Procedure A to be the usual stepwise procedure for controlling the familywise error. That is, reject $H_{(1)}$ if $P_{(1)} < y^{\alpha}_{K,0}$, and having rejected $H_{(r-1)}$, reject $H_{(r)}$ if $P_{(r)} < y^{\alpha}_{\{(r),(r+1),\ldots,(k)\},0}$.

Proposition A: Procedure A controls the number of false discoveries to be less than or equal to u with confidence 1-α.

Proof of Proposition A: Let s be the rank of the p-value in the ordered list (1) corresponding to (u+1)st smallest p-value associated with a variable that satisfies the null hypothesis. Let K_0 be the set of indices corresponding to the variables that satisfy the null hypothesis. In particular, the variable associated with $H_{(s)}$ is in K_0 and there are u other variables in K_0 associated with u of the $H_{(1)}, H_{(2)}, \ldots, H_{(s-1)}$. Then

Pr(more than u true null hypotheses are rejected)

$$= \Pr(H_{(1)}, H_{(2)}, \ldots, H_{(s)} \text{ are rejected})$$

$$\leq \Pr(P_{(s)} < y^{\alpha}_{s,u})$$

$$\leq \Pr(P_{(s)} < y^{\alpha}_{K_0,u})$$

$$\leq \alpha,$$

where the penultimate inequality is true because $y^{\alpha}_{s,u} \leq y^{\alpha}_{K_0,u}$. The probability in the next-to-last line depends only on the marginal joint distribution of the p-values in K_0, in particular the (u+1)st smallest p-value in the set K_0, which is by definition $P_{(s)}$. Therefore, the permutation distribution of

the (u+1)st smallest of the $\{P'_t, t \in K_0\}$ can be used as a reference distribution, verifying the last inequality.

If the variables or p-values are discrete, there can be ties in the p-values given in (1), but this does not present a problem. Regardless of the ordering of the tied variables in (1), if the hypothesis associated with the first variable in the order is rejected, then the hypotheses associated with the other tied variables will also be rejected because the minimization (2) will be over smaller sets for the other variables. In addition, which of the tied variables is considered first for rejection will not matter, as the permutation distribution will include all of them when considering the first rejection. Also, if the first of the tied variables fails to reject, the procedure ceases and no further hypotheses are rejected, so that the situation in which the first tied variable fails to reject, and the later tied variables do reject, need not be considered.

We have used univariate p-values in (1) and (2) upon which to base our permutation tests. Any univariate statistic could be used here to distinguish between the groups, e.g., the absolute difference between two group means or the absolute difference between two proportions for binary data. The only requirement is that the same statistic is used for the $\{P'_t, t \in K\}$ generated on the permuted data. Obviously, using a statistic that captures well the evidence of group differences should result in more power to make true discoveries.

The computational burden implicit in (2) can be large if $\binom{r}{u}$ is not small. A conservative procedure in these situations would be to set $y^{\alpha}_{r,u} = y^{\alpha}_{K,u}$. When $r << k$, we expect the conservatism of this less computationally intensive procedure to be minimal. We examine this in section 4 with some limited simulations.

As an alternative to fixing α in advance, one can obtain an "adjusted significance level" for each hypothesis. This adjusted significance level associated with hypothesis $H_{(r)}$ represents the smallest significance level at which $H_{(r)}$ would be rejected under the procedure. We first calculate the tentative adjusted significance level $p^*_{(r)}$ as follows. For $r \le u$, set $p^*_{(r)} = 0$. For $r > u$, $p^*_{(r)}$ is defined as the smallest α such that $P_{(r)} < y^{\alpha}_{r,u}$ where $y^{\alpha}_{r,u}$ is obtained from the generated permutation distribution. The adjusted significance level for $H_{(r)}$ is then defined as $max(p^*_{(1)}, p^*_{(2)},..., p^*_{(r)})$. This last maximization step is required because of the sequential nature of the rejection procedure.

2.2 Controlling the Proportion of False Discoveries

Intuitively, Procedure A could be applied in an iterative way to control the false discovery proportion (FDP) which we define as the number of null hypotheses rejected divided by the total number of hypotheses rejected. If no hypotheses are rejected, we define FDP = 0. Note that the expected value of the FDP is the false discovery rate (FDR) that has been previously studied by Benjamini and Hochberg (1995). Suppose that the goal is to have 1-α confidence that the false discovery proportion is less than some value γ. Then, if the *r*th ordered hypothesis is the last one to be rejected by the procedure, we want to be 1-α confident that no more than |[rγ]| false discoveries are among those *r* rejected hypotheses, where the notation |[*x*]| denotes the greatest integer less than or equal to *x*.

Procedure B: Suppose we wish to be 1-α confident that the proportion of false discoveries is no more than γ (γ > 0). Reject $H_{(1)}$ if $P_{(1)} < y^{\alpha}_{K,0}$. Having rejected $H_{(r-1)}$, reject $H_{(r)}$ if either $|[r\gamma]| > |[(r-1)\gamma]|$ or $P_{(r)} < y^{\alpha}_{r,|[r\gamma]|}$.

Proposition B: If the univariate tests are consistent, Procedure B asymptotically controls the proportion of false discoveries to be less than or equal to γ with confidence $1-\alpha$, where the asymptotic control is as sample size increases under fixed null and alternatives, and fixed number of variables.

Proof of Proposition B: Let S be the rank of the p-value in the ordered list (1) corresponding to the first time that if hypotheses 1, 2, . . . , S were rejected, then the false discovery proportion would be larger than γ at that point. S is a random variable that depends on the complete set of k tests. Note that the variable associated with $H_{(S)}$ is in K_0, and that there are $|[S\gamma]|$ other variables in K_0 associated with $|[S\gamma]|$ of the $H_{(1)}, H_{(2)}, \ldots, H_{(S-1)}$. Also note that $H_{(S)}$ cannot be an "automatic rejection" because if $|[S\gamma]| > |[(S-1)\gamma]|$, then the false discovery proportion would be larger than γ for hypothesis $(S-1)$. The proof of Procedure B follows from

$\Pr[\text{FDP} > \gamma]$

$\leq \Pr[H_{(1)}, H_{(2)}, \ldots, H_{(S)} \text{ are rejected}]$

$\leq \Pr[\, P_{(S)} < y^{\alpha}_{S,|[S\gamma]|}\,]$

$\leq \Pr[\, P_{(S)} < y^{\alpha}_{K_0,|[S\gamma]|}\,] \rightarrow \alpha$ under the stated asymptotics.

The first inequality is not an equality as it was in the proof of Proposition A because the procedure may continue rejecting beyond $H_{(S)}$ leading to the possibility that the false discovery proportion may subsequently fall below γ. If $|[S\gamma]|$ were a constant rather than a random quantity, then the last probability would be less than α without the need for asymptotics. However, it is not a constant, and it is possible to construct counterexamples with the last probability exceeding α. This does not

rule out the possibility that Procedure B does control at α the probability of FDP exceeding γ because there are several other inequalities in the proof. But, at present, we appeal to an asymptotic argument. If we consider asymptotics of increasing sample size with fixed number of variables and fixed null and alternative hypotheses, then S will approach a constant s with probability one since the p-values corresponding to the variables associated with nonnull hypotheses will appear first in the ordered list (1). Under these asymptotics, the last probability will approach a value less than or equal to α. We examine the small sample properties of this procedure in some limited simulation studies in section 4.

The issue of tied p-values is more subtle here than for Procedure A because of the automatic rejections, i.e. when $|[r\gamma]| > |[(r\text{-}1)\gamma]|$. If the p-value associated with an automatically rejected hypothesis is tied with the next p-value to be considered, then depending upon the arbitrary order of the variables with the tied p-values, the procedure might or might not reject the hypothesis immediately following the automatic rejection. We recommend in this situation that all possible ordering of the variables with tied p-values be considered, and the one that leads to the most rejections be used. If we have two tied p-values and consider both orderings, then we have two possible values of $y^{\alpha}_{S,|[S\gamma]|}$, say $y^{\alpha,1}_{S,|[S\gamma]|}$ and $y^{\alpha,2}_{S,|[S\gamma]|}$. Therefore, in the proof of Procedure B, we can replace the line $\Pr[P_{(S)} < y^{\alpha}_{S,|[S\gamma]|}]$ with $\Pr[P_{(S)} < y^{\alpha,1}_{S,|[S\gamma]|}$ or $P_{(S)} < y^{\alpha,2}_{S,|[S\gamma]|}]$. However, since $y^{\alpha,1}_{S,|[S\gamma]|}$ and $y^{\alpha,2}_{S,|[S\gamma]|}$ are each less than $y^{\alpha}_{K0,|[S\gamma]|}$, the next line in the proof follows.

We can obtain an adjusted significance level for each hypothesis using a method similar to that employed for Procedure A. This adjusted significance level associated with hypothesis $H_{(r)}$ represents the smallest significance level at which $H_{(r)}$ would be rejected under the procedure. We first calculate the tentative adjusted significance level $p^{*}_{(r)}$ as follows. If $|[r\gamma]| > |[(r\text{-}1)\gamma]|$, $p^{*}_{(r)} = 0$.

If $|[r\gamma]| = |[(r\text{-}1)\gamma]|$, then $p^*_{(r)}$ is defined as the smallest α such that $P_{(r)} < y^{\alpha}_{r,u}$ where $y^{\alpha}_{r,u}$ is obtained from the generated permutation distribution. The adjusted significance level for $H_{(r)}$ is then defined as $max(p^*_{(1)}, p^*_{(2)}, ..., p^*_{(r)})$.

3. APPLICATION TO MICROARRAY DATA FROM 20 PAIRED BREAST TUMORS

We demonstrate the methods on a subset of a previously published data set involving gene expression from cDNA microarrays using specimens from 65 breast tumors from 42 individuals (Perou et al. 2000). Our analysis is based on data from 20 individuals with specimens taken both before and after a 16-week course of doxorubicin chemotherapy. The microarray analyses generate two gene expression profiles for each specimen, one before, and one after chemotherapy. Each profile consists of log expression ratios measured on approximately 9000 genes. Based on a cluster analysis of these profiles, Perou et al. (2000, p. 747) note "Gene expression patterns in two tumor samples from the same individual were almost always more similar to each other than either was to any other sample." This similarity does not eliminate the equally interesting possibility of finding large and statistically significant differences in gene expression in the 20 paired pre vs. post chemotherapy specimens. Genes showing different expression before compared to after chemotherapy will henceforth be referred to as “differentially expressed”, and the goal of our analysis will be to identify these differentially expressed genes.

The primary data were obtained at <http://genome-www.stanford.edu/molecularportraits/>. Fluorescent intensities for the two labeled samples (test and reference samples) are recorded in two channels. Typically, a red and green image is produced to display the intensities in the two channels. For each channel and each spot (gene) on the array, both foreground (feature) and

background intensity measures are given, and a "signal" can be calculated as the foreground intensity minus the background intensity. An important first step in the analysis of microarray data is careful examination of the red and green images of individual spots on the array to determine the quality of the fluorescence measurements. Due to experimental artifacts such as dust specs, scratches, bubbles, or poor adherence of the DNA to the glass slide, some spots will need to be flagged and not used in the data analysis. This spot flagging had already been performed by the original investigators, and the flagging data was supplied. Data from spots flagged and not used by the original investigators were also not used here. Spots labeled as "EMPTY" were not used in any analyses.

A second data quality issue is the problem of low intensity spots. In particular, when foreground intensity is close to background intensity, and consequently the signal is low, it is known that the measurements tend to be unreliable. We have found it helpful to examine plots of signal in channel 1 versus signal in channel 2 for each array. Figure 1 shows such a plot for one of the arrays. A "fanning" of the point scatter in the lower left corner indicates the range in which signal is lost amid the noise. This plot and the many others examined (not shown) suggested that signal measurements less than 100 (2 on the log base 10 scale) should not be considered reliable. We chose to exclude from the analysis spots for which signal was less than 100 in both channels. If signal was less than 100 in only one channel, the spot was used with the signal in that channel set to 100.

Also noticeable in Figure 1 is a shift of the scatter of points to slightly below the 45 degree line. This shift typically results from differences in signal intensities in the two channels due to nuisance factors such as different physical properties of the two fluorescent dyes (for example, different efficiencies of dye incorporation or different rates of degradation), different photo-

multiplier tube settings for scanning, or differences in the starting amounts of the two fluorescently labeled samples; it indicates the need for data normalization. An expression ratio for each gene was formed as channel 2 divided by channel 1 signal for that gene. Ratios were then median normalized within each array by dividing the ratios by the median of the ratios from non-excluded spots for that array.

Genes for which data were missing from more than half of the 20 paired tumor specimens were eliminated from consideration. This left 8029 genes for analysis. All p-values were calculated on log transformed median-normalized expression ratios.

Table 3 shows the genes with the 28 smallest unadjusted paired t-test p-values for testing the null hypothesis that the mean pre and post chemotherapy expression of the gene is the same. Geometric means are also provided along with GenBank accession numbers. Information on these genes can be found by searching on the accession numbers at <http://www.ncbi.nlm.nih.gov/UniGene/>. It is interesting to note that the first 24 genes identified have more expression post-chemotherapy than pre-chemotherapy, and all of the specimens for 4 of the genes showed more expression after the chemotherapy (Figure 2). This list of 28 genes is the set identified by application of our Procedure B controlling the proportion of false discoveries to be no more than $\gamma = .10$ with approximate 95% confidence ($\alpha = .05$). Table 4 provides a comparison of the lists of genes identified under several methods: Bonferroni, step-down FWE control (equivalent to our Procedure A with $u = 0$ and $\alpha = .05$), Procedure A with $u = 1$ and $\alpha = .05$, Procedure A with $u = 2$ and $\alpha = .05$, and Procedure B with $\gamma = .10$ and $\alpha = .05$. Permutations in this paired data setting consist of switching the before- and after-chemotherapy expression profiles. Thus, there are 2^{20} possible permutations of the profile data. We used Monte Carlo sampling to randomly select 19,999 of these permutations. Adjusted p-values are given

under each method, with values in bold type denoting rejected hypotheses ($\alpha < .05$). Setting the confidence level at 95%, the number of rejections ranges from 11 to 28, with the Bonferroni method yielding the least and Procedure B, tied with Procedure A with $u = 2$, yielding the most. By exploiting the correlations among genes, the step-down FWE procedure was able to reject 6 more genes than the Bonferroni procedure. Allowing just one false discovery allowed identification of an additional 6 genes, and allowing 2 or 10% false discoveries allowed identification of an additional 5 genes, for a total of 28. The results presented in Table 4 all used the less computational conservative ($y^{\alpha}_{r,u} = y^{\alpha}_{K,u}$) versions of Procedures A and B. Calculation of all results in Table 4 took approximately 18 minutes on a 866 MHz pentium III processor with 1 GB of memory. The fully computational method was also run and took nearly 5 hours on the same computer. Comparing the conservative to the fully computational procedure, the resulting p-values were very similar and the same numbers of hypotheses were rejected.

4. SIMULATION STUDIES

We conducted several simulations to assess the performance of Procedures A and B and to compare them to the procedure to control the FWE rate. Specifically, we evaluated the same procedures that we applied to the breast tumor data in the previous section. Similar to the paired breast tumor example, we considered cases in which 8000 hypotheses were to be tested, and we assumed 20 paired specimens. The paired difference data were generated as multivariate Gaussian with a variety of block correlation structures. The blocks of correlated variables can be viewed, for example, as representing genes that are in the same pathway or that are co-regulated. Variables are independent between blocks. Both positive and negative correlations were considered. We varied the number of differentially expressed genes (nonnull variables), and the patterns of correlations

among those differentially expressed genes. Each simulation was repeated 10000 times. For purposes of computational feasibility, 99 resamplings were performed at each stage, and the conservative method of critical value calculation ($y^{\alpha}_{r,u} = y^{\alpha}_{K,u}$) was used except where noted otherwise. The effect of using 99 resamplings was investigated in some limited simulation studies, and the results were found not to differ substantially from using 199 resamplings.

In Table 5, performance of the methods is assessed under a block exchangeable correlation structure with 30 nonnull variables out of 8000. All procedures satisfied the targeted 95% confidence, with very little conservatism. The increased sensitivity afforded by allowing even one or two false discoveries is striking, for example an absolute increase of over 20% as compared to the procedure controlling the FWE when the within-block correlation was 0 or .5. Several variants of these cases were also examined and are now described (results not shown). For the results in Table 5, the 30 nonnull variables were all placed within the same block, that is, they were correlated. For the case with correlation .5, when we distributed the nonnull variables one per block (uncorrelated), the results were very similar. The sensitivities and levels were also very similar when a block size of 20 rather than 100 was used. An interesting result was obtained when the number of nonnull variables was reduced to 5. In this case, the performance of Procedure A to control for no more than 1 or 2 errors was essentially the same as before, but the sensitivity advantage of Procedure B was lost, as it became nearly equivalent to the FWE procedure; even one false rejection would result in 20% false discoveries, which exceeds the allowable 10%. A simulation under the complete null, i.e., no nonnull variables, with block correlation .5 verified that stated confidence levels were maintained by each procedure. Also, the mean numbers of false discoveries were .05, 1.07, and 2.09 for the procedures controlling for 0, 1, and 2 false discoveries, respectively.

A case of negatively correlated variables in combination with positively correlated variables was also considered. As in Table 5, there were 30 nonnull variables out of 8000. Results are presented in Table 6, and they are very similar to those given in Table 5. Confidence levels are maintained, and substantial sensitivity is gained by using the procedures that allow a few false discoveries.

We also assessed the difference between the conservative procedure and the fully computational procedure for the case of block correlation structure with correlation .5 and 30 nonnull variables. We found that the two computational methods had essentially the same performance in terms of sensitivity to detect nonnull variables.

5. DISCUSSION

This paper has considered both the control of the absolute number of false discoveries as well as the control of the false discovery proportion. Investigators may prefer one or the other of these types of control. For example, if an investigator considers 10 false discoveries out of 100 discoveries acceptable but not 10 false discoveries out of 12 discoveries, then he or she would be more interested in controlling the false discovery proportion. On the other hand, if following up 10 false discoveries is considered acceptable to find even one true discovery, then control of the number of false discoveries would be more appropriate and more sensitive. Also, when it is expected that there are few truly differentially expressed genes (nonnull hypotheses), the discreteness of the false discovery proportion should be kept in mind when setting the bound γ. For example, setting $\gamma = .10$ with 5 nonnull variables is essentially equivalent to allowing no false discoveries.

The sensitivity gains from allowing even one or two false discoveries or a small proportion of false discoveries can be very large. This was clearly evident in Table 5 in the cases of low to moderate correlation where the sensitivity to detect nonnull variables increased from approximately 65% to about 85% by allowing just one false discovery, and to over 90% if two false discoveries were allowed. In the breast cancer example, the number of identified genes allowing either 2 or 10% false discoveries was more than 1.5 times the number identified when controlling the FWE and more than 2.5 times the number identified by the Bonferroni procedure.

The main potential downside of these procedures is the computational burden, but we don't view this as a serious disadvantage. In section 2.1 we suggest the computational simplification of setting $y^{\alpha}_{r,u} = y^{\alpha}_{K,u}$ when $r << k$. In our breast tumor example, the slight conservatism that this might have introduced did not alter the number of differentially expressed genes we were able to

identify compared to using the fully computational procedure. Applying all of the conservative procedures on our example data required only 18 minutes in computer time. Also, our simulations which all used the conservative simplification did not suggest any substantial conservatism in the confidence levels or significant reduction in sensitivity for the cases we considered. The degree of conservatism would be most severe in a situation in which the number of nonnull variables is large and a moderately large number or proportion of false discoveries is allowed. Such situations are not likely to be the norm, but further work might be needed to develop appropriate computational simplifications or computationally more efficient algorithms for those cases. For example, for Procedure A, a potentially useful hybrid approach might be to use $y^{\alpha}_{r,u}$ up to a certain value of r, say $r^{*} = 20$ or 30, and then switch to the fixed critical value $y^{\alpha}_{r^{*},u}$ thereafter. In any case, undoubtedly a substantial amount of time has been invested to measure the large number of variables on the many specimens, so even if a statistical procedure requires a few hours to complete, the time seems well spent to improve the sensitivity for detecting important variables.

Procedures A and B are immediately generalizable to a wide range of situations. In our example and simulations, we based the permutation procedures on parametric paired t-test p-values. However, the procedures did not depend on any of the parametric assumptions. Any orderable univariate statistics can form the basis for the permutation procedure. With appropriate choice of univariate statistics, these methods apply to discrete data or even to censored data provided that the censoring distributions are the same in the groups. Unpaired or multi-group sampling designs are handled by appropriately modifying the permutation method.

REFERENCES

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

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., Rubin, E. M. (2000), "Microarray expression profiling identifies genes with altered expression in HDL-deficient mice," *Genome Research*, 10, 2022-2029.

International Human Genome Sequencing Consortium (2001), "Initial sequencing and analysis of the human genome," *Nature*, 409, 860-921.

Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A., Brown, P. O., Botstein, D. (2000), "Molecular portraits of human breast tumors," *Nature*, 406, 747-752.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995), "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, 270, 467-470.

Seeger, P. (1968), "A note on a method for the analysis of significance en masse," *Technometrics*, 10, 586-593.

Troendle, J. F. (1996), "A permutational step-up method of testing multiple outcomes," *Biometrics*, 52, 846-859.

Tusher, V., Tibshirani, R., and Chu, G. (2001), "Significance analysis of microarrays applied to transcriptional responses to ionizing radiation," *Proceedings of the National Academy of Sciences*, 98, 5116-5121.

Westfall, P. H. and Young, S. S. (1993), *Resampling-Based Multiple Testing,* New York: Wiley, pp. 72-74.

Table 1. Simulated distribution of the number of false discoveries for 10,000 hypotheses tested using univariate nominal significance levels of .001.

Correlation, ρ	Mean	Percentile				
		10th	25th	50th	75th	90th
0	10.0	6	8	10	12	14
.5	9.9	3	5	8	13	18
.8	9.9	0	1	4	12	27

NOTE: Test statistics are simulated under a global null as standard normal with block diagonal correlation matrix, block size 100 and pairwise correlation ρ within a block. The number of simulated data sets is 10,000.

Table 2. Simulated distribution of the false discovery proportion for 10,000 hypotheses tested using a simple step-up procedure to control the expected false discovery rate to be no more than 10%.

		Percentile				
Correlation, ρ	Mean	10th	25th	50th	75th	90th
0	.098	0	0	0	.18	.29
.5	.090	0	0	0	.13	.30
.8	.055	0	0	0	0	.17

NOTE: Simulated test statistics are normally distributed with block diagonal correlation matrix, block size 100, and pairwise correlation ρ within a block. The first 10 test statistics have standardized mean of 4, and the remaining have mean zero. The number of simulated data sets is 10,000. The step-up procedure applied is that described and studied by Benjamini and Hochberg (1995).

Table 3. Genes identified when controlling false discovery proportion to be no more than approximately 10% with approximately 95% confidence.

Gene rank	Accession number	Number of patients	Gene expression ratio (geometric mean)			Unadjusted p-value[a]
			Pre-chemo	Post-chemo	Ratio of Post/Pre	
1	AA478553	19	0.74	2.21	2.98	.0000002
2	N23941	20	1.27	2.18	1.72	.0000005
3	W96134	20	1.66	3.08	1.85	.0000006
4	N95402	20	1.19	2.07	1.74	.0000011
5	AA040944	20	0.42	1.79	4.21	.0000011
6	AA442853	20	1.53	2.61	1.71	.0000012
7	AA134757	20	2.32	4.69	2.02	.0000021
8	AA418077	20	1.75	3.38	1.94	.0000027
9	R12840	20	0.50	1.77	3.55	.0000052
10	AI831083	20	1.24	2.48	2.00	.0000057
11	AA044993	20	0.65	1.37	2.09	.0000057
12	AA031596	20	1.19	2.06	1.73	.0000067
13	AA454868	18	2.95	5.05	1.71	.0000067
14	AA598794	20	0.72	1.52	2.10	.0000077
15	T74141	19	8.28	16.44	1.98	.0000097
16	H21041	20	0.95	1.83	1.93	.0000110
17	AA133129	20	0.68	1.37	2.01	.0000119
18	AA485377	20	0.52	1.33	2.56	.0000221
19	AA167222	20	4.21	9.62	2.29	.0000293
20	AA287695	19	1.27	2.18	1.72	.0000325
21	AA489234	20	1.31	1.95	1.49	.0000582
22	AA293362	20	1.71	2.94	1.72	.0000600
23	N94487	20	1.19	2.39	2.01	.0000658
24	H86754	19	2.00	3.79	1.89	.0000974
25	AA430629	19	1.82	1.25	0.69	.0001050
26	AA485743	20	1.20	0.92	0.77	.0001155
27	H82948	19	0.84	0.72	0.86	.0001209
28	H05099	19	0.97	0.77	0.80	.0001309

[a]Based on a paired Student's t-test with degrees of freedom equal to one less than the number of patients with paired data available for that gene.

Table 4. Adjusted p-values for genes identified in Table 3.

Gene rank	Accession number	Adjusted P-value				
		Bonferroni	Step-down FWE control (Procedure A with $u = 0$)	Procedure A: Number of false discoveries not to exceed 1 ($u = 1$)	Procedure A: Number of false discoveries not to exceed 2 ($u = 2$)	Procedure B: False discovery proportion not to exceed .10 ($\gamma = .10$)
1	AA478553	**.0016**	**.0010**	**.0000**	**.0000**	**.0010**
2	N23941	**.0040**	**.0019**	**.0002**	**.0000**	**.0019**
3	W96134	**.0048**	**.0022**	**.0002**	**.0001**	**.0022**
4	N95402	**.0088**	**.0035**	**.0002**	**.0001**	**.0035**
5	AA040944	**.0088**	**.0035**	**.0002**	**.0001**	**.0035**
6	AA442853	**.0096**	**.0038**	**.0002**	**.0001**	**.0038**
7	AA134757	**.0169**	**.0068**	**.0002**	**.0001**	**.0068**
8	AA418077	**.0217**	**.0088**	**.0003**	**.0001**	**.0088**
9	R12840	**.0418**	**.0176**	**.0005**	**.0002**	**.0176**
10	AI831083	**.0458**	**.0189**	**.0007**	**.0002**	**.0176**
11	AA044993	**.0458**	**.0189**	**.0007**	**.0002**	**.0176**
12	AA031596	.0538	**.0223**	**.0008**	**.0002**	**.0176**
13	AA454868	.0538	**.0223**	**.0008**	**.0002**	**.0176**
14	AA598794	.0618	**.0254**	**.0010**	**.0002**	**.0176**
15	T74141	.0779	**.0318**	**.0016**	**.0003**	**.0176**
16	H21041	.0883	**.0375**	**.0021**	**.0003**	**.0176**
17	AA133129	.0955	**.0416**	**.0023**	**.0003**	**.0176**
18	AA485377	.1774	.0753	**.0070**	**.0014**	**.0176**
19	AA167222	.2352	.1019	**.0118**	**.0022**	**.0176**
20	AA287695	.2609	.1137	**.0143**	**.0026**	**.0176**
21	AA489234	.4673	.2039	**.0386**	**.0094**	**.0176**
22	AA293362	.4817	.2090	**.0409**	**.0099**	**.0176**
23	N94487	.5283	.2267	**.0472**	**.0121**	**.0176**
24	H86754	.7820	.3158	.0879	**.0276**	**.0276**
25	AA430629	.8430	.3352	.0985	**.0318**	**.0318**
26	AA485743	.9273	.3591	.1127	**.0382**	**.0382**
27	H82948	.9707	.3724	.1210	**.0423**	**.0423**
28	H05099	1.00	.3959	.1352	**.0492**	**.0492**

NOTE: All permutation procedures used 19,999 resampled permutations. The computationally conservative versions of Procedures A and B, i.e., $y^{\alpha}_{r,u} = y^{\alpha}_{K,u}$, were used. Univariate p-values upon which all procedures were based were computed using a paired Student's t-test with degrees of freedom equal to one less than the number of patients with paired data available for that gene. P-values in bold type correspond to tests that would be rejected under the given procedure at level .05.

Table 5. Simulation results applying Procedures A and B at level .05 to sets of 20 pairs of profiles generated under block-exchangeable correlation structure.

	Step-down FWE (Proc. A, $u = 0$)	Procedure A: FD[a] ≤ 1 ($u = 1$)	Procedure A: FD ≤ 2 ($u = 2$)	Procedure B: FDP[b] $\leq .10$ ($\gamma = .10$)
$\rho = 0$				
Sensitivity (in %) to detect nonnull	65.90	87.07	92.72	93.42
% of simulations with FD > 0	**5.04**	29.50	55.37	63.56
% of simulations with FD > 1	0.19	**4.74**	19.25	30.62
% of simulations with FD > 2	0.00	0.53	**4.68**	12.29
% of simulations with FDP > .10	0.02	0.20	1.45	**4.51**
$\rho = .5$				
Sensitivity (in %) to detect nonnull	66.64	85.57	91.10	87.08
% of simulations with FD > 0	**5.04**	22.49	39.35	52.50
% of simulations with FD > 1	0.34	**4.83**	13.53	26.90
% of simulations with FD > 2	0.07	1.29	**5.08**	12.54
% of simulations with FDP > .10	0.63	1.26	3.32	**4.58**
$\rho = .9$				
Sensitivity (in %) to detect nonnull	82.12	88.45	90.93	87.73
% of simulations with FD > 0	**4.79**	8.90	11.66	70.53
% of simulations with FD > 1	2.52	**4.76**	6.72	65.67
% of simulations with FD > 2	1.82	3.44	**4.71**	59.17
% of simulations with FDP > .10	1.92	3.53	4.53	**4.87**

NOTE: Paired difference data were generated as 8000 Gaussian variables, each with variance 1, in 80 blocks of 100 correlated variables each. Pairwise correlation between variables was ρ within a block and 0 between blocks. Thirty of the variables out of the 8000 tested had standardized mean equal to 1.5, and the remaining had mean zero. Percentages in bold type have nominal value 5.00%.

[a]FD = number of false discoveries

[b]FDP = false discovery proportion

Table 6. Simulation results applying Procedures A and B at level .05 to sets of 20 paired profiles generated under correlation structure with mixed positive and negative correlations.

	Step-down FWE (Proc. A, $u = 0$)	Procedure A: FD[a] ≤ 1 ($u = 1$)	Procedure A: FD ≤ 2 ($u = 2$)	Procedure B: FDP[b] $\leq .10$ ($\gamma = .10$)
Sensitivity (in %) to detect nonnull	67.59	85.17	90.60	90.10
% of simulations with FD > 0	**4.85**	20.57	35.28	44.31
% of simulations with FD > 1	0.55	**4.80**	12.25	18.30
% of simulations with FD > 2	0.16	1.60	**5.15**	8.41
% of simulations with FDP > .10	0.25	1.14	3.17	**4.75**

NOTE: Paired difference data were generated as 8000 Gaussian variables in 80 blocks of 100 correlated variables each. Within each block, the set of variables is divided into thirds (33|33|34). Any pair of variables within the same third have positive correlation equal to 2/3, and between thirds, variables have negative correlation –1/3. Between blocks, variables are independent. Of the 100 variables in the first block, 10 variables in the first third, 10 in the second third, and 10 in the last third each had standardized mean equal to 1.5. All remaining variables had mean zero. This resulted in 30 correlated, nonnull variables, each with standardized mean equal to 1.5. Percentages in bold type have nominal value 5.00%.

[a]FD = number of false discoveries

[b]FDP = false discovery proportion

Figure Legends

Figure 1. Logarithm (base 10) of signal in channel 2 plotted versus logarithm of signal in channel 1 for the genes having measurements in both channels in microarray experiment #25 in the Perou data set. Signal is computed as foreground intensity minus background intensity. Dashed line represents the 45 degree line.

Figure 2. For each of the genes given in Table 3, plotted points are the ratios of the post-chemotherapy to pre-chemotherapy gene expression for each of 18-20 patients, and arrows are the geometric means of the post/pre ratios. A) The 17 most significant genes, comprising the set identified as differentially expressed by the level .05 step-down FWE procedure, or equivalently Procedure A allowing no false discoveries with 95% confidence. B) The 18^{th} through 28^{th} most significant genes, which together with the first 17 comprise the set identified as differentially expressed by Procedure B allowing no more than 10% false discoveries with approximate 95% confidence.

Figure 1

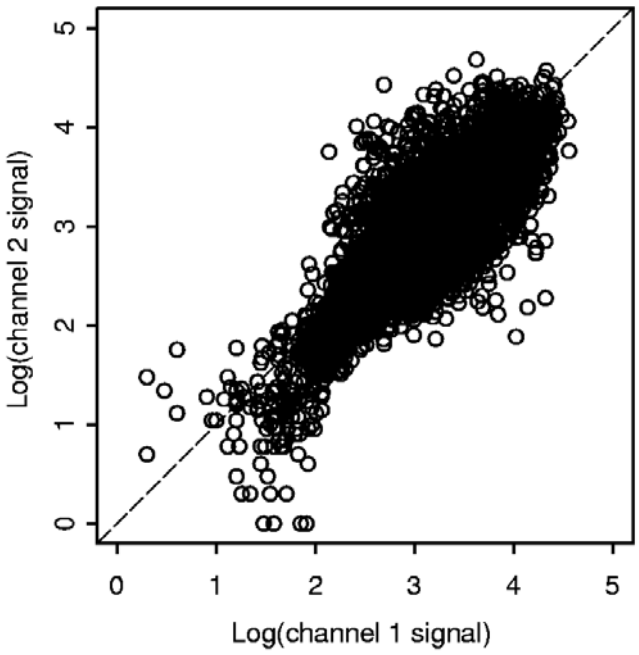


Figure 2A

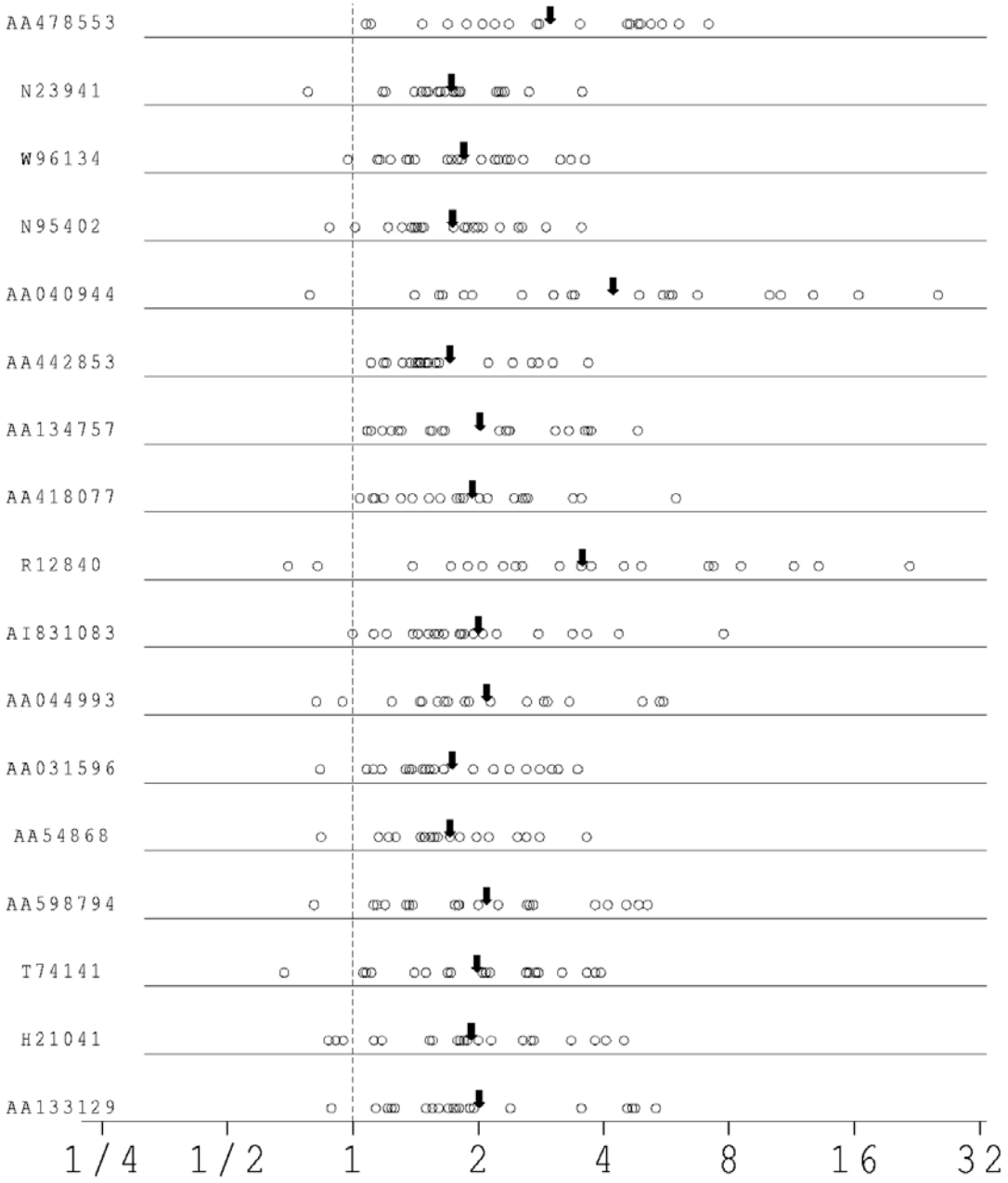
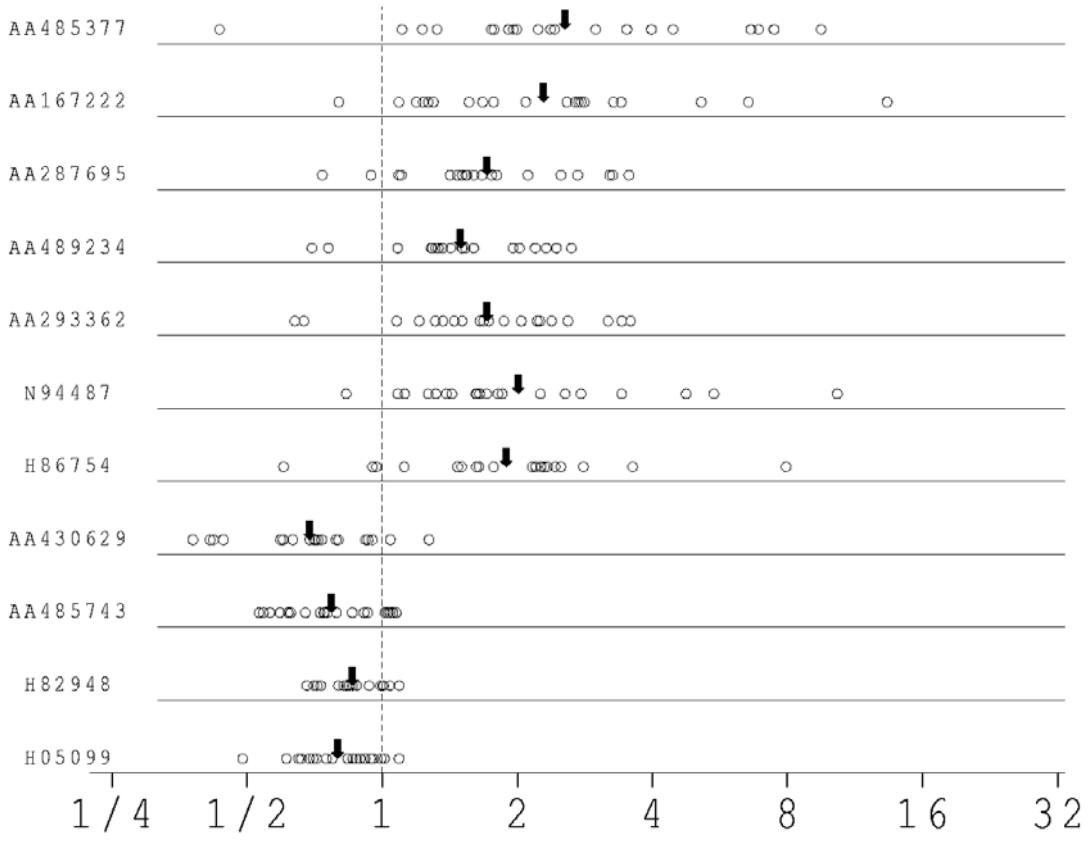

AA478553
N23941
W96134
N95402
AA040944
AA442853
AA134757
AA418077
R12840
AI831083
AA044993
AA031596
AA54868
AA598794
T74141
H21041
AA133129
1/4
1/2
1
2
4
8
16
32

Figure 2B


AA485377
AA167222
AA287695
AA489234
AA293362
N94487
H86754
AA430629
AA485743
H82948
H05099
1/4
1/2
1
2
4
8
16
32