Two-Stage Testing in Microarray Analysis: What is Gained?

David B. Allison[1] and Christopher S. Coffey[2]

Department of Biostatistics[1,2], Section on Statistical Genetics[1]
& Clinical Nutrition Research Center[1]
University of Alabama at Birmingham
Birmingham, Alabama

Corresponding author: David B. Allison, Ph.D., Department of Biostatistics, Ryals Bldg, Suite 327, 1665 University Blvd, Birmingham, Alabama 35294. Phone: 205-975-9169. Fax: 205-975-2540. Email: Dallison@ms.soph.uab.edu.

This research was supported in part by NSF grant 0090286 and NIH grants R01ES09912, P30DK056336, R01DK56366, R01AG018922, P01AG11915, and U24DK058776.

Abstract

Microarray technology for gene expression studies offer powerful new technology for understanding changes in gene expression as a function of other observable or manipulable variables. However, microarrays also pose a number of new challenges. One of the most prominent of these is the difficulty in establishing a procedure for declaring whether a gene's expression level is associated with the independent variable that offers reasonable and specifiable false positive (type 1 error) and false negative (type 2 error) rates. Recently, Miller *et al* (1) offered a two-stage testing procedure to address these goals. However, information was not provided to indicate whether this procedure would or would not meet its objectives. Herein, we show mathematically that the two-stage procedure proposed does not provide benefits in terms of minimizing false negatives while controlling the false positive rate relative to standard single stage testing. Therefore, investigators are encouraged to consider alternative analytic approaches.

Introduction

The advent of microarray technology for gene expression measurements opens many exciting opportunities and challenges in aging research (2,3). One of the major challenges involves determining whether a sample numerical difference in gene expression among two or more groups, conditions, or tissues represents a 'statistically significant' difference (4). This is challenging in part because microarrays allow one to simultaneously test for differences in thousands of genes, thereby creating a problem of multiple inference if one stays in the frequentist (5) null hypothesis testing paradigm (6). For example, if differences in each gene expression are compared at the 0.05 significance level for a microarray containing 10,000 genes and the null hypothesis of no difference in gene expression were true for all genes, we would expect to observe approximately 500 'false-positives' or genes for which a statistically significant difference is observed when there is, in truth, no difference. This issue of multiple comparisons has long been a thorn in the side of researchers. There are many accepted strategies for adjusting the significance level to ensure that the probability of making any false positive is equal to or below the desired significance level. For example, a Bonferroni correction (8) divides the desired experiment-wise alpha level by the number of tests. Each individual test is then conducted at this Bonferroni adjusted alpha level. A criticism of this and similar approaches is the fact that by controlling the false positive rate, one becomes more likely to observe 'false negatives', or differences that fail to reach statistical significance when an actual difference exists. An alternative approach for meeting this challenge was suggested in a recent paper (1) which "advocates a two-stage design in which significance testing applied to exploratory data is used to guide a second round of hypothesis-testing experiments conducted in a separate set of experimental studies" (p. B52). Ideally, this method would control the experiment-wise alpha level or type 1 error rate (the probability of making *any* false positives in the study) while making fewer type 2 errors than the single stage design. However, as previously noted (7), evidence has yet to be offered that the two-stage design procedure either

controls the experiment-wise type 1 error rate or reduces the risk of type 2 errors. The purpose of this brief paper is to examine the type 1 error rate and power of the two-stage design and evaluate how these statistical properties compare to those of a single stage design.

The Two-Stage Approach

Miller *et al*'s approach (1) involves first testing for differences between two groups for all of k genes in one set of data (stage 1) using a stage 1 alpha level (which we will denote α_1) greater than the alpha level that would be required by a Bonferroni correction. No specific mention is given as to how to choose α_1, but Miller *et al* use $\alpha_1 = .001$ for an example with k = 10,000, suggesting perhaps that they mean for α_1 to be set somewhat below the more conventional .05. The k hypothesis tests are then performed, yielding some number of genes, m, with significant effects ($0 \leq m \leq k$). Then, at stage 2, a second independent set of data is gathered and only those m genes found to be significant at stage 1 are tested at level $\alpha_2 = .05/m$. Any gene significant at level α_2 "can be accepted as age sensitive" (1; p. B55). Presumably, although not explicitly stated, a 2-tailed hypothesis test is conducted at stage 1, while a 1-tailed test is conducted at stage 2. The stage 2 test should test only in the direction that the apparent effect was observed at stage 1 because it would make little sense to conclude that there is an effect on the basis of two random samples producing significant results in opposite directions.

This two-stage testing procedure was offered as one way of dealing with the problem of false positives (type 1 errors) that would result from multiple significance testing without correction and false negatives (type 2 errors) that would result from the use of a Bonferroni correction. However, concrete information indicating that this approach will achieve these goals has yet to be presented. First, although Miller *et al* do not state exactly to what overall alpha level this procedure holds the entire experiment, their use of .05 in determining α_2 suggests that perhaps they intend to achieve an overall experiment-wise alpha level of .05. No information

has been offered as to whether the proposed design actually controls the experiment-wise alpha level. Second, no information has been offered to indicate that this procedure is more powerful than simply testing all data together in a single stage design with a method that controls the experiment-wise alpha level. These questions are further evaluated below. It may be worth pointing out that the two-stage procedure under discussion is aimed at reducing purely stochastic threats to statistical inference and should be seen as distinct from constructive-type replications which have independent value in helping to eliminate non-stochastic threats to valid inference (9).

Type 1 Error (False Positive) Rate

Once one has reached stage 2, the probability of making one or more type 1 errors at this stage is equal to $1-(1-.05/m)^m$. This assumes one uses .05 as Miller *et al* do in their example (1). In fact, one is free to choose any alpha level by simply substituting the desired level for .05 in this formula. It should be noted that this formula is correct when all null hypotheses are true and all gene expression levels are independent. In reality, both of these conditions seem extremely unlikely to be met. To the extent that they are not met, the actual value of the stage 2 type I error rate will lie somewhere between zero and $1-(1-.05/m)^m$. Due to the known conservativeness of the Bonferroni correction, if the stage 2 testing were the only testing involved, this formula would hold the overall alpha rate of the entire experiment (which we will denote α_{ew}) to a value close to (though just slightly less than) .05. However, because we are testing in two stages, a type I error will be made only if genes for which no difference truly exists show significant differences (false positives) at both stages. To obtain the experiment-wise alpha level, first compute the alpha level at stage 2 given that m tests were significant at stage 1. Then, because the number of genes significant at stage 1 (M) is a random variable, the experiment-wise alpha level is obtained by computing the weighted sum over all possible outcomes, m, with the weights representing the probability of that outcome.

$$a_{ew} = P(M=0)(0) + \sum_{m=1}^{k}\left[P(M=m)\left(1-\left(1-.05/m\right)^{m}\right)\right]$$

$$a_{ew} = \sum_{m=1}^{k}\left[P(M=m)\left(1-\left(1-.05/m\right)^{m}\right)\right]$$

Finally, if the probability of rejecting each test equals α_1, the probability of observing m significant tests out of k independent tests can be described by the binomial distribution with parameters k and α_1. Taking this into account, the experiment-wise alpha level for the two-stage design can be written as:

$$a_{ew} = \sum_{m=1}^{k}\left[\left(\frac{k!}{m!(k-m)!}\right)\left(a_1^{m}(1-a_1)^{k-m}\right)\left(1-\left(1-.05/m\right)^{m}\right)\right]$$

Clearly, α_{ew} is affected by the choice of α_1. To demonstrate this dependence, consider the example used by Miller *et al* (1) where k=10,000, α_1 = .001 and α_2 = .05/m. Under this circumstance, α_{ew} = .049 which is very close to the level of .05 that might be desired. However, if α_1 were switched to .0001, a value still greater than the Bonferroni corrected value (.05/10000), then the overall type 1 error rate becomes only .0314. Furthermore, using the fact that the stage 2 alpha level will be less than or equal to .05, a simple bound on the experiment-wise alpha level for the two-stage design is:

$$a_{ew} \le (0.05)\,P(M \ge 1)$$

$$a_{ew} \le (0.05)\left[1-\left(1-a_1\right)^{k}\right].$$

This demonstrates that the two-stage design is often conservative, leading to an experimental-wise alpha level which is lower than that desired. Furthermore, as the choice of α_1 becomes smaller, the experiment-wise alpha level of the two-stage design becomes more conservative.

As the formula above shows and the example illustrates, the two-stage procedure fails to consistently hold the overall alpha level at .05. On the contrary, one can more easily achieve the goal of holding α_{ew} = .05 in a single stage design by simply setting $\alpha_1 = 1-(1-.05)^{1/k}$, and only conducting a stage 1 analysis. This provides a correction that is nearly equivalent to the

Bonferroni correction, but less conservative. Nevertheless, perhaps the two-stage procedure will reduce the type 2 error rate (i.e., increase power) relative to a one-stage procedure with $\alpha_1 = 1-(1-.05)^{1/k}$.

Power & False Negatives

Using the between groups t-test as Miller et al (1) discuss and assuming conditions for its use are valid (i.e., normality, homoscedasticity, independence of observations), an effect size for a specific gene can be expressed as $d = \frac{m_1 - m_2}{s}$, where μ_1 is the population (not sample) mean level of gene expression for one group of subjects (group 1), μ_2 is the population mean for the second group (group 2), and σ is the common within-group standard deviation. Then, assuming equal numbers of subjects per group, and denoting the total number of subjects by 2n, the non-centrality parameter for the t-distribution with 2n-2 degrees of freedom (df) for testing the between group difference is $w_1 = \sqrt{\frac{d^2 n}{2}}$. Let $t_{\nu,\alpha}$ represent the value that cuts off the upper 100*α percentile of the central t-distribution with ν degrees of freedom and let $F(x,\nu,\omega)$ denote the cumulative distribution function at the point x of a noncentral t-distribution with ν degrees of freedom and noncentrality parameter, ω. The power for single-stage testing can be written as $P_1 = 1 - F\left(t_{a_1/2}, n-2, w_1\right)$.

If one splits one's sample into two non-overlapping sub-samples to be used in the two stages, then the power calculation is somewhat more complex. Miller *et al* (1) did not state how the subjects should be divided between the two stages, but for our subsequent calculations, we will assume subjects are divided equally between the two stages. Because a significant result will occur only if the gene showed statistically significant differences at *both* stages, the power for two-stage testing will equal the product of the probability (power) of obtaining a significant result at stage 1 and the probability (power) of getting a significant result at stage 2 given a

significant result at stage 1. The probability of getting a significant result at stage 1 can be derived just as above for single-stage testing, with the exception that the sample size will now be half of that used in the single stage design. As a consequence, the non-central t-distribution will now have $w_2 = \sqrt{\frac{d^2 n}{4}}$ and df=n-2. Once a significant result is obtained at stage 1, the conditional probability of getting a significant result at stage 2 depends on how many other genes (C) were declared significant at stage 1. Because C is a random variable, we must then sum these conditional powers over all possible outcomes for C, weighting by the probability of that outcome. Then, the power for two-stage testing can be written as

$$P_2 = \left[1 - F\left(t_{a_1/2}, n-2, w_2\right)\right] \sum_{c=0}^{k-1} \left\{ P\left(C = c\right) \left[1 - F\left(t_{.05/(c+1)}, n-2, w_2\right)\right]\right\}.$$

Note that the third term within the square brackets on the right side of the equation is for the 1-tailed test at stage 2. For any given value of c, $P\left(C = c\right)$ depends on the power of the tests for the other genes which may vary from one data set to the next and will be unknown. Therefore, in order for us to calculate $P\left(C = c\right)$ we need to assume some model. For simplicity, we assume that the null hypothesis is true for all genes except the one for which we are calculating power. Were we to assume the null hypothesis is not true for other genes as well, we would increase the probability of declaring a larger number of genes significant at stage 1. As more genes are declared significant at stage 1, the stage 2 alpha level used for each individual test will become smaller, hence reducing the stage 2 power for that test. As a consequence, the overall power will be smaller than it would be under the assumption that the null hypothesis is true for all genes except the one of interest. That is, by assuming that the null hypothesis is true for all genes except one for which we are calculating power we are, for any particular effect size and sample size, deriving the maximum possible power for the Miller *et al* two-stage procedure

(1). Using the same reasoning as for M above, C will follow a binomial distribution with parameters k-1 and α_1 and we can write the power for the two-stage design as:

$$P_2 = \left[1 - F\left(t_{a_1/2}, n-2, w\right)\right] \sum_{c=0}^{k-1} \left\{ \left(\frac{(k-1)!}{c!(k-1-c)!} \right) \left(a_1^c (1-a_1)^{k-1-c} \right) \left[1 - F\left(t_{.05/(j+1)}, n-2, w\right)\right] \right\}$$

Quantitative Results

Having derived expressions for power using both single-stage and two-stage testing, we can begin to compare their relative power. Let us return to Miller *et al* 's example of 10,000 genes. To maintain α_{ew} = .05 with single-stage testing of 10,000 genes requires a per test alpha of $\alpha_1 = 5.13 \times 10^{-6}$. Figures 1 and 2 compare the power for single-stage and two-stage testing with two different levels of α_1 (.001, .0001) across a range of effect sizes. Figure 1 corresponds to the Miller *et al* example of 10 subjects per group, while Figure 2 demonstrates the same results for a study with 20 subjects per group. The figures clearly demonstrate that Miller *et al* 's two-stage procedure does not achieve the goal of providing a method that reduces false negatives (i.e., increases power). In fact, it can even exacerbate the very problem it is intended to alleviate. For example, under the scenario considered in Miller et al's paper (α_1 = .001, δ = 3.0, n = 10), the power for a single-stage test is .614 (calculations were conducted using SAS/IML) but the power is only .409 for the two-stage approach.

It may not be all that surprising that the two-stage approach results in reduced power because one can see that it creates a kind of paradox in which identical information is treated differently depending on the sequence in which it occurs. For example, assume $\alpha_1 < \alpha_2$ and that the p-value obtained for testing a gene in the first stage is less than α_1 while the p-value obtained at the second stage is between α_1 and α_2. Using the two-stage approach, differences for this gene would be declared significant. In contrast, a gene for which the p-value at stage 1 is between α_1 and α_2 and the p-value at stage 2 is less than α_1 would not be declared significant. Yet the two situations offer equivalent evidence against the null hypothesis. The fact

that evidence against the null hypothesis in this second situation is ignored shows that information is being discarded and it is therefore not surprising that power is lost.

Conclusion

In conclusion, the two-stage procedure offered by Miller *et al* (1) often fails to hold the overall experiment-wise type 1 error rate to some desired alpha level. Many choices of α_1, the per-test alpha level at stage 1, will result in an overall experiment-wise type 1 error rate that is overly conservative. Less conservative, single stage methods of holding the overall type 1 error rate to any desired level exist (10). Moreover, this two-stage method can also exacerbate the false negative (type 2 error) rate, that is, decrease power compared to a single stage method.

While the two-stage design does not fare well when compared statistically to the single stage design, there may be non-statistical concerns which increase the attractiveness of the two-stage design. For example, it is possible that such a two-stage approach could improve power *per dollar spent on a study* if, at stage 2, one needed only to assay a subset of all genes on the array used in stage 1 and the cost for assaying a subset was less than the cost for assaying the entire set. In such situations, both the costs and required resources per subject in stage 2 (and hence the entire study) might be substantially reduced. Furthermore, the two stage design proposed by Miller *et al* represents only one possible type of two (or more generally *multi*) stage design. It is possible that other two-stage designs could be proposed which have better statistical properties and compare more favorably to the single stage design.

In attempting to interpret these results, one question that may be of primary interest to researchers regards the size of the mean differences represented by the effect sizes shown in Figures 1 and 2. To address this issue, we can offer the following information. Writing from the social science perspective, Cohen (11) defined 'small,' 'medium,' and 'large' effect sizes as values of 0.20, 0.50, and 0.80, respectively. By this standard, an effect size of 3.0, as in the example considered by Miller *et al* (1) is extremely large. However, in basic laboratory research, effect sizes are often much larger. For example, consider a study of a knock-out mouse model

of hereditary haemochromatosis (12). When knock-out mice were compared to wild-type mice the "iron concentration in livers were: 170 +/- 15 mu g/g (mean +/- SD) in controls and 1010 +/- 50 mu g/g in beta 2m (-/-) mice." This represents an effect size of 22.6. Unfortunately, because fold-change is not a statistic that takes within-group variability into account, there is no way to directly translate an effect size expressed as a standardized mean difference into a specific fold-change value.

Finally, the two-stage procedure discussed operates from a strictly frequentist point of view under the seemingly implausible assumption of the null hypothesis being true for all genes studied. Alternatives to a strict frequentist approach exist (e.g., 4; 13; 14) and are seen by many (e.g., 15) to be preferable when conducting many tests and a global null hypothesis seems untenable.

REFERENCES

1. Miller RA, Galecki A, Shmookler-Reis RJ. Interpretation, design, and analysis of gene array expression experiments. *J Gerontol A-Biol* 2001; **56**: B52-B57.
2. Weindruch R, Kayo T, Lee CK, Prolla TA. Microarray profiling of gene expression in aging and its alteration by caloric restriction in mice. *Journal of Nutrition* 2001; **131**: 918S-923S.
3. Pilarsky CP, Schmitt AO, Dahl E, Rosenthal A. Microarrays - chances and challenges. *Curr Opin Mol Ther* 1999; **1**: 727-736.
4. Allison DB, Gadbury G, Heo M, Fernandez J, Lee C-K, Prolla TA, Weindruch, R. Statistical methods for the analysis of microarray gene expression data: a mixture model approach. *Computational Statistics & Data Analysis*. (In press.).
5. Neyman, J. Frequentist probability and frequentist statistics. *Foundations Of Probability And Statistics.* I. Synthese 1977; **36**: 97-131.
6. Thomas DC, Siemiatycki J, Dewar R, Robins J, Goldberg M, Armstrong BG. The Problem Of Multiple Inference In Studies Designed To Generate Hypotheses. *American Journal Of Epidemiology* 1985; **6**: 1080-1095.
7. Prolla TA, Allison DB, Weindruch R. Response to: "Interpretation, Design and Analysis of Gene Array Expression Experiments". *J Gerontol A-Biol* 2001; **56**: B327-B329.
8. Bland JM, Altman DG. Multiple Significance Tests - The Bonferroni Method. *Brit Med J* 1995; **310:** 170-170.
9. Lykken, DT. Statistical significance in psychological research. *Psychological Bulletin* 1968; **70**: 151-159.

10. Hochberg Y, Tamhane, AC. Multiple comparison procedures. In: Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons; 1987.
11. Cohen J. Statistical Power Analysis for the Behavioral Sciences (2nd edition). Hillsdale, NJ: Erlbaum; 1988.
12. Moos T, Trinder D, Morgan EH. Cellular Distribution of Ferric Iron, Ferritin, Transferrin, and Divalent Metal Transporter 1 (DMT1) in Substantia Nigra and Basal Ganglia of Normal and Beta 2-Microglobulin Deficient Mouse Brain. *Cell Mol Biol* 2000; **46**: 549-561.
13. Lee MLT, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *P Natl Acad Sci USA* 2000; **97**: 9834-9839.
14. Manduchi E, Grant GR, McKenzie SE, Overton GC, Surrey S, Stoeckert CJ. Generation of patterns from gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics* 2000; **16**: 685-698.
15. Steenland K, Bray I, Greenland S, Boffetta P. Empirical Bayes adjustments for multiple results in hypothesis-generating or surveillance studies. *Cancer Epidem Bio* 2000; **9**: 895-903.

FIGURE 1. COMPARING POWER OF THE TWO DESIGNS (n=10)

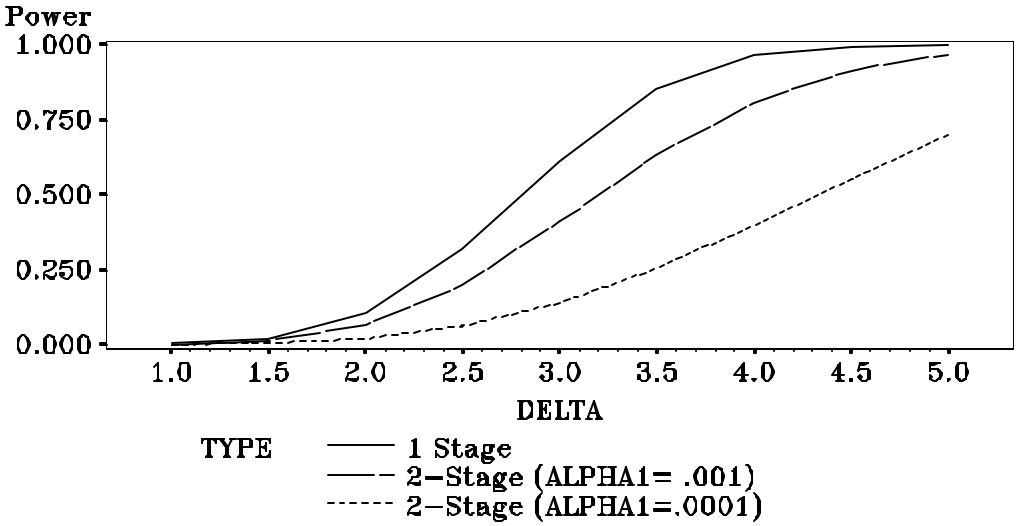

FIGURE 2. COMPARING POWER OF THE TWO DESIGNS (n=20)