RADIATION RESEARCH **165**, 745–748 (2006)
0033-7587/06 $15.00

COMMENTARY

Some Statistical Issues in Microarray Gene Expression Data

Matthew S. Mayo,[a,1] Byron J. Gajewski[b] and Jeffrey S. Morris[c]

[a] *Department of Preventive Medicine and Public Health, Center for Biostatistics and Advanced Informatics, Kansas Masonic Cancer Research Institute, and* [b] *Schools of Allied Health and Nursing, Center for Biostatistics and Advanced Informatics, University of Kansas Medical Center, Kansas City, Kansas; and* [c] *Department of Biostatistics and Applied Mathematics, The University of Texas M.D. Anderson Cancer Center, Houston, Texas*

Mayo, M. S., Gajewski, B. J. and Morris, J. S. Some Statistical Issues in Microarray Gene Expression Data. *Radiat. Res.* 165, 745–748 (2006).

In this paper we discuss some of the statistical issues that should be considered when conducting experiments involving microarray gene expression data. We discuss statistical issues related to preprocessing the data as well as the analysis of the data. Analysis of the data is discussed in three contexts: class comparison, class prediction and class discovery. We also review the methods used in two studies that are using microarray gene expression to assess the effect of exposure to radiofrequency (RF) fields on gene expression. Our intent is to provide a guide for radiation researchers when conducting studies involving microarray gene expression data.

INTRODUCTION

The use of DNA microarray gene expression data (*1–4*) in health research has exploded over the last few years. This technology is useful for making inferences about the genomic profile of an individual for use in risk assessment and/or detection of effects. As exemplified by the two papers appearing in this issue (*5, 6*), researchers are using microarray gene expression data to understand the influences of non-ionizing as well as ionizing radiation on the genotype. Therefore, since there may be a strong future for the use of microarray gene expression data in radiation research, the radiation research community may expect to see more research papers using this technology. It is vital that radiation researchers understand the fundamental statistical issues so that one can decide whether the research follows good, sound, fundamental statistical principles. The purpose of this commentary is to provide very basic guidance (Table 1) as to what should be the minimum standards for such work, statistically, for microarray gene expression data studies. Two of the biggest issues for assessing the validity of an analysis of microarray data studies are preprocessing the data and the analysis of the processed data (*7*).

[1] Address for correspondence: Department of Preventive Medicine and Public Health, Center for Biostatistics and Advanced Informatics, Kansas Masonic Cancer Research Institute, University of Kansas Medical Center, Kansas City, KS; e-mail: mmayo@kumc.edu.

PREPROCESSING DATA

The basic analytical data for microarrays involve an n-by-p table of data, where the n rows each correspond to an array and the p columns each correspond to a gene. Given this table, any of a number of different analyses can be done to identify genes differentially expressed across experimental conditions, find clusters of samples or genes with similar expression levels, or build predictive models based on the gene expression levels of sets of genes. To obtain this table, however, a number of preprocessing steps must first be taken to screen out poor-quality arrays, normalize the expression levels across arrays and filter out confounding effects, and quantify the gene expression values from the raw array data. If these steps are not performed properly, the table of expression levels may not be accurate, preventing the possibility of obtaining valid conclusions from the study. The specifics of these steps tend to differ across different platforms, e.g. between Affymetrix oligonucleotide arrays and glass cDNA arrays.

Some comprehensive preprocessing packages have been developed for analyzing Affymetrix data (*8, 9*). For glass cDNA arrays, various image processing algorithms are necessary to quantify the spots and adjust for background effects. For all platforms, various normalization procedures must be used to filter out systematic biases that can occur within the experiment (*10*). For example, with cDNA arrays, dye bias is an important factor that must be accounted for (*11*). Yang *et al.* (*11*) describe and compare several types of statistical normalizing techniques including global, intensity-dependent, within-print tip group, scale and composite normalization techniques. Kerr, Martin and Churchill (*12*) discuss ANOVA-based methods that can simulta-

TABLE 1
Basic Items to Consider when Reviewing the Merits of a Microarray Study

1. Preprocessing of the data
Screened for poor arrays?
Normalized expression levels across arrays?
e.g. Global
Intensity-dependent
Within-print tip group
Scale
Composite
Filtered confounding effects?
ANOVA-based methods and linear models
Quantify the gene expression table: produce n-by-p table.
2. Type of microarray study (analyze the n-by-p table)
Class comparison:
Do gene expression profiles differ between predefined groups?
Key issue: Did the researchers address the False Discovery Rate (FDR)?
FDR = average $F/(F + T)$
F: False positives
T: True positives
Class prediction:
Predict the prespecified group that expression profile belongs.
Key issue: Did the researchers perform cross validation to see how well their prediction model works?
Class discovery:
No prespecified group. Do the gene expressions cluster into groups?
3. Keeping up with the technology (http://genomebiology.com/2004/5/10/R80)

neously model some of these effects using linear models, and Wolfinger *et al.* (*13*) do the same using linear mixed models.

TYPES OF MICROARRAY STUDIES

Once the data are preprocessed (normalized), the researcher moves to the analysis of the n-by-p table. To understand the analysis issues, one must understand the goals of the particular research project. There are three basic types of research goals when one is using microarray data. The first is *class comparison*, which is used when one is attempting to see which gene expression profiles differ between prespecified groups. Specifically, class comparison involves identifying a list of genes that are differentially expressed across the specified groups. For example, radiation researchers might be interested in seeing how radiofrequency (RF) fields associated with cellular phone use affect the gene expression profile. One might match subjects, animals or cells exposed to RF fields with similar specimens not exposed to RF fields. Then one would compare the gene expression profiles between the two groups.

The key statistical issue in class comparison involves controlling for the multiple comparisons. To appreciate this, consider an example in which there are 1,000 genes and 100 subjects in two groups. Using a two-sample t test with a significance level of 0.05, we would expect 50 of the gene expressions to be significant by chance alone, even if in fact the two groups had the same gene expression profiles. In some settings, classical multivariate statistical techniques such as Hotelling's T^2 can be used to assess differences in the multivariate setting in a way that controls for multiplicity. However, these methods require at least as many samples (n) as variables (p), which clearly does not hold in the setting of microarrays. Thus these approaches cannot be applied here. Another way to alleviate this problem is to adjust for multiple comparisons using a Bonferroni adjustment. But using a Bonferroni adjustment will result in a per comparison significance level of $0.05/1000 = 0.00005$, which might be too conservative because of the difficulty in achieving statistical significance. The conservatism comes from the fact that Bonferroni controls the experiment-wise error rate, the probability of even one false positive, at the alpha level. This criterion is usually not appropriate in the context of microarrays, since we are willing to have some false positive genes, and this stringent criterion leads to a large number of false negative genes whose group effects are not discovered. As a result, researchers have developed a new criterion, the False Discovery Rate (FDR), which is widely considered to be a more appropriate criterion in this context. Rather than controlling the probability of making even one false discovery, the FDR controls the expected proportion of false discoveries.

A landmark paper by Storey and Tibshirani (*14*) described the basics of FDR. Let F be the number of false positives and T be the number of true positives. The FDR for an experiment is the average $F/(F + T)$, or the ratio of the false positives divided by all positives. There are a number of methods available for controlling FDR, including significance analysis of microarrays (SAM) (*15*), Empirical Bayes methods (*15, 16*), beta-uniform mixture (BUM) (*18*), spacing LOESS histogram (SPLOSH) (*19, 20*), and other Bayesian methods (*21–23*). Most of these methods assume independence among the genes. Some recent research relaxes this assumption and can handle dependent genes (*24*). Some of these methods work by first computing a suitable test statistic, e.g. a t statistic, for each gene. Then they identify a cut point on the P values below which the gene is considered significant, while controlling the FDR at some level alpha. Another alternative to the FDR approach is to do permutation tests (*25, 26*). Without suitably addressing the multiple comparisons issue with these high-dimensional data, radiation researchers again could be directed down a dubious research path.

One debate in all statistical research is whether to use parametric statistical methods or non-parametric statistical methods for making class comparisons. For example, some parametric methods rely on assigning a normal distribution to the data and using, for example, a t test to compare genes across groups. Other approaches do not make distributional assumptions and instead use the rank sum test. Obviously one would like to use the technique that minimizes the FDR. Shedden *et al.* (*27*) compared seven methods for pro-

ducing Affymetrix expression scores. They found that the data processing method has a much bigger impact on FDR than the choice between using a parametric or a non-parametric technique. This may be because the *t* statistic is robust to distributional assumption.

The second type of research study, called *class prediction*, has a similar setup for comparison in that there is a prespecified set of groups. However, now the researcher is interested in using the microarray data to predict the group that they belong to. For example, does one's genotype predict the group of subjects exposed to radiation compared to the group of subjects that are unexposed?

The statistical problems associated with class prediction are classic in data mining. Data mining techniques can have a problem with creating models that over-fit the data, making it difficult to predict a future data set using the same decision rule. Researchers deal with this problem by splitting the data into two pieces, a training data set and a validation data set. The training data set is used to select the decision rules for prediction. Then the validation data set is used to test whether the model defined from the training data set is reasonable. One problem with such an approach is that the analyst must be very careful how training and validation data sets are selected. That is, a random sample should be selected—not a sample that works well for the answer desired. Another problem can occur if the analysts improperly select their model using the validation set. That is, they use the training set to get a model and test the model. The testing fails, so the analyst goes back to the training set and fits a new model then tests the new model. They repeat this process until they get a good fit on their training set. This is invalid because they are using their validation data to fit their model and also assess their prediction errors. The prediction errors obtained by such a procedure are not valid.

One way to fix this problem is to use cross-validation. For example, suppose that the researcher has 50 arrays; a fourfold cross-validation would be to select 40 arrays, fit a model, and then use the 10 other arrays to validate the model. The analyst would repeat this five times, getting a fourfold cross-validation of their model. The key to this methodology is that the user must perform the model selection at each of the steps to account for the uncertainty in the model selection process.

The third type of research study, called *class discovery* (*28*), involves no prespecified classes in the data. That is, the researchers are interested in finding some sort of a structure to the genome of a particular population. One way to do this is using clustering techniques. There are two ways to classify gene expression. One way is to cluster the samples. In this case one might be interested in seeing how radiation-exposed subjects and unexposed subjects cluster without specifying their classes. A second example of clustering would be to cluster the genes. In this case the researcher might be interested in identifying sets of genes with correlated expression levels. Clustering is useful for identifying relevant biological structure in the data and also for identifying structure caused by systematic biases in the data. For example, if the samples cluster strongly based on the day on which the arrays were run, that may indicate a strong day effect that must be accounted for in the analysis.

EXAMPLES IN THIS ISSUE OF *RADIATION RESEARCH*

In this issue of *Radiation Research*, there are two papers using microarray gene expression data. The study of effects of exposure to RF radiation on gene expression by Qutob *et al.* (*5*) is an example of class comparison. They used James-Stein shrinkage *F* tests to compare gene expressions in an exposed group to those in a control group. This technique is attractive since they used it to control the FDR. They normalized their expression values by using a technique in the library of a freeware statistical package called R. The second paper in this issue, by Whitehead *et al.* (*6*), is also a class comparison study of how exposures to different types of RF fields affect the genotype in mice. They normalized their array data using a scaling technique. The authors state that the GeneChip data were analyzed using a two-tailed *t* test and that the expected number of false positives was estimated from *t* tests on 20 permutations of the six sham RF-field-exposed samples. The use of the sham-sham experiments to estimate the number of false positives and control the overall type I error rate is an innovative and well-thought-out approach to estimate and control the FDR. One note of caution should be made; the lack of statistical significance should not be used to conclude a lack of effect. The lack of significance could be a function of the small sample sizes and the inability to detect differences, especially when using stringent methods to control the FDR. Thus a slightly more tempered title and conclusions might have been warranted for this manuscript. Neither manuscript details *a priori* power calculations for readers to determine if either study had a chance to detect clinically meaningful differences; this would have bolstered the conclusions that could have been drawn from each. The data from each of these studies could be used to aid the design of future studies to detect clinically meaningful differences in gene expression.

CONCLUSION

In this commentary we have outlined some of the basic statistical issues in microarray gene expression data. This technology is dynamic. The preprocessing and statistical analysis techniques are evolving daily. Many of the preprocessing and analysis methods for microarray data are detailed, and open-source software is made available as part of the Bioconductor project (*29*), which is continually updated as new methods are developed. This is an excellent source for methods for analyzing data from microarray experiments.

Keeping up with this technology can be daunting, but

the fundamentals are clear. By spending the necessary time properly designing your study (fundamental to *all* studies), preprocessing the data, and using a statistical analysis technique that corresponds to the particular goal of the study, radiation scientists can feel confident that their research findings are sound and will likely pass the scrutiny of peer review. For both papers in this issue (*5, 6*), the authors used appropriate methodologies to preprocess the data and analyze the data and used methods to control the FDR. They can conclude there is insufficient evidence to say that RF-field exposure alters gene expression.

ACKNOWLEDGMENT

Dr. Mayo's work was supported by the following grant: NCI R23 CA095835.

Received: February 24, 2006; accepted: March 7, 2006

REFERENCES

1. R. Simon, M. D. Radmacher and K. Dobbin, Design of studies using DNA microarrays. *Genet. Epidemiol.* **23**, 21–36 (2002).
2. A. Tefferi, E. D. Wieben, G. W. Dewald, D. A. H. Whiteman, M. E. Bernard and T. C. Spelsberg, Primer on medical genomics part II: Background principles and methods in molecular genetics. *Mayo Clinic Proc.* **77**, 785–808 (2002).
3. C. P. Lorentz, E. D. Wieben, A. Tefferi, D. A. Whiteman and G. W. Dewald, Primer on medical genomics part I: History of genetics and sequencing of the human genome. *Mayo Clin. Proc.* **77**, 773–782 (2002).
4. P. L. Elkin, Primer on medical genomics Part V: Bioinformatics. *Mayo Clinic Proc.* **78**, 57–64 (2003).
5. S. S. Qutob, V. Chauhan, P. V. Bellier, C. L. Yauk, G. R. Douglas, L. Berndt, A. Williams, G. B. Gajda, E. Lemay and J. P. McNamee, Microarray gene expression profiling of a human glioblastoma cell line exposed *in vitro* to a 1.9 GHz pulse-modulated radiofrequency field. *Radiat. Res.* **165**, 636–644.
6. T. D. Whitehead, E. G. Moros, B. H. Brownstein and J. L. Roti Roti, Gene expression does not change significantly in C3H 10T^{1}/2 cells after exposure to 847.74 CDMA or 835.62 FDMA radiofrequency radiation. *Radiat. Res.* **165**, 626–635.
7. S. Huang, H. R. Qian, C. Geringer, C. Love, L. Gelbert and K. Bemis, Assessing the variability in GeneChip data. *Am. J. Pharmacogenomics* **3**, 279–290 (2003).
8. C. Li and W. Hung Wong, Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biol.* **2,** RESEARCH0032 (2001).
9. R. A. Irizarry, B. Hobbs, F. Collin, D. Y. Beazer-Barclay, K. J. Antonellis, U. Scherf and T. P. Speed, Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
10. T. Park, S. G. Yi, S. H. Kang, S. Y. Lee, Y. S. Lee and S. Simon, Evaluation of normalization methods for microarray data. *BMC Bioinformatics* **4**, 33 (2003).
11. Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai and T. P. Speed, Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, 1–10 (2002).
12. M. K. Kerr, M. Martin and G. A. Churchill, Analysis of variance for gene expression microarray data. *J. Computat. Biol.* **7**, 819–837 (2000).
13. R. D. Wolfinger, G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari and R. S. Paules, Assessing gene significance from cDNA microarray expression data via mixed models. *J. Computat. Biol.* **8**, 625–637 (2001).
14. J. D. Storey and R. Tibshirani, Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
15. G. V. Tusher, R. Tibshirani and G. Chu, Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).
16. B. Efron, R. Tibshirani, J. D. Storey and V. Tusher, Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **96**, 1151–1160 (2001).
17. M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner and K. W. Tsui, On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.* **8**, 37–52 (2001).
18. S. Pounds and S. W. Morris, Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* **19**, 1236–1242 (2003).
19. S. Pounds and C. Cheng, Improving false discovery rate estimation. *Bioinformatics* **20**, 1737–1745 (2004).
20. S. Pounds and C. Cheng, Sample size determination for the false discovery rate. *Bioinformatics* **21**, 4263–4271 (2005).
21. G. Parmigiani, E. S. Garrett, R. Anbazhagan and E. Gabrielson, A statistical framework for expression-based molecular classification in cancer. *J. Royal Stat. Soc.* **64**, 717–736 (2002).
22. J. D. Storey, The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Stat.* **31**, 2013–2035 (2003).
23. P. Muller, G. Parmigiani, C. P. Robert and J. Rousseau, Optimal sample size for multiple testing: The case of gene expression microarrays. *J. Am. Stat. Assoc.* **99**, 990–1001 (2004).
24. Y. Benjamini and D. Yekutieli, The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
25. P. H. Westfall and S. S. Young, *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment.* Wiley, New York, 1993.
26. S. Dudoit, Y. H. Yang, M. J. Callow and T. S. Speed, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiment. *Stat. Sin.* **12**, 111–139 (2002).
27. K. Shedden, W. Chen, R. Kuick, D. Ghosh, J. Macdonald, K. R. Cho, T. J. Giordano, S. B. Gruber, E. R. Fearon and S. Hanash, Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data. *BMC Bioinformatics* **6**, 26 (2005).
28. S. Varma and R. Simon, Iterative class discovery and feature selection using Minimal spanning trees. *BMC Bioinformatics* **5**, 126 (2004).
29. R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge and J. Zhang, Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80, 1–16 (2004).