REVIEW ARTICLE

A Design and Statistical Perspective on Microarray Gene Expression Studies in Nutrition: The Need for Playful Creativity and Scientific Hard-Mindedness

Grier P. Page, PhD, Jode W. Edwards, PhD, Stephen Barnes, PhD, Richard Weindruch, PhD, and David B. Allison, PhD

From the Department of Biostatistics, Section on Statistical Genetics, the Department of Pharmacology and Toxicology and Purdue-UAB Botanical Center for Age-Related Disease, and the Clinical Nutrition Research Center, University of Alabama at Birmingham, Birmingham, Alabama, USA; and the Department of Medicine, University of Wisconsin–Madison, Madison, Wisconsin, USA

OBJECTIVES: Our purpose is to highlight some of the past and potential future uses of microarray in nutrition research, while also commenting on some aspects of the design conduct and analysis of microarray data that will leave to improved data quality.

METHODS: In this review article we outline some of the aspects of microarray experimentation that must be considered before and during these experiments. These topics include: identification of the experiment's objective (hypothesis), the experimental design, sample size, statistical analysis, data verification, data handling, and experimental interpretation.

RESULTS: In order to illustrate the principles we outline in this article we use the methods to layout the design of a microarray experiment to study one aspect of the observation that a diet high in soy is associated with lower rates of breast cancer.

CONCLUSIONS: Microarrays are a very powerful tool for studying virtually every nutrition-related disease and trait and can provide valuable insights that are not obtainable with other techniques. However, unless nutrition researchers conduct their studies with scientific hard-mindedness, the studies will be of lower power at least if not completely misleading. *Nutrition* 2003;19:997–1000. ©Elsevier Inc. 2003

KEY WORDS: Microarray, statistical analysis, design, power, soy, breast cancer

INTRODUCTION

Microarrays have been described as "the hottest thing in biology and medicine since the advent of the polymerase chain reaction a decade ago" (anonymous 2001). This technology emerged circa 1996[1,2] and had its first high-profile uses in 1998 and 1999.[3–5] What is "hot" is that microarrays allow simultaneous measurement of mRNA expression levels from thousands of genes.

DNA microarrays will make major contributions to nutritional research. Already, microarrays have been used to study retardation of aging by caloric restriction,[3,6] effects of high-fat diet,[7] effects of isoflavones,[8] and obesity and diabetes.[9] In the future, we envision the impact of microarrays on nutritional research increasing through applications like those mentioned and studies of genetically modified organisms, the impact of nutraceuticals on gene expression, and studies of the bioequivalence of foods and drugs.

Many researchers have embraced microarrays with the glee of "a kid in a candy shop." Many, however, have also suffered the effects of trying to overindulge on more data than they can digest. The effects of such overindulgence are preventable. Researchers need only to remain scientifically hard-minded by employing sound principles of experimental design, analysis, and interpretation. The solution to data overindulgence is not to condemn the pioneers of microarray research who entered this new realm without the benefit of existing sound methodologic approaches and analysis, nor is it to abandon microarrays as expensive "random number generators." Rather, we applaud the creativity and boldness of those who have tried to use a novel and complex technology. Microarrays are like any other hot new laboratory tools that have come before them; future use of microarrays lies in building on what has already been done to generate an ever more rigorous foundation for discovery and inference.

Our goal is to encourage nutritional scientists to forge ahead into microarray research while bearing in mind that microarrays are not different at a fundamental level from any other laboratory or experimental technique. Thus, they must consider clearly defining objectives, appropriately designing experiments, choosing adequate sample size, applying statistical rigor to data analysis, and paying close attention to data handling and management issues.[10]

Supported in part by NIH grants T32 AR007450, R01 DK56366, R01 DK54298, P50 AT00477, and U24 DK058776; NSF grants 0090286 and 0217651; and a grant from the University of Alabama at Birmingham.

Correspondence to: Grier P. Page, PhD, Section on Statistical Genetics, Ryals Public Health Building, Suite 327, Department of Biostatistics, 1665 University Avenue, Birmingham, AL 35294-0022, USA. E-mail: gpage@ms.soph.uab.edu

0899-9007/03/$30.00
doi:10.1016/j.nut.2003.08.001

Experimental Objectives

It is not an uncommon experience for us when consulting with investigators who have conducted a microarray study of, for example, particular strains of lean versus obese mice, that they are not able to clearly state whether the objective of their study is to identify genes that were differentially expressed in response to obesity, whose differential expression led to obesity, or whose differential expression might be an early marker for obesity, and so on. Answering each of these questions might require a slightly different design. Also all too often, investigators have inadvertently allowed themselves to go down the slippery slope from hypothesis-generating studies to studies that might be termed "objective generating." By *objective generating* we mean studies in which the investigator has initiated activity without even a clear conception of what the exact objectives of the study are in the sole hopes that, by some mysterious process, the mass of data that will come out at the end of the microarray study will make all things clear. We do not believe that such outcomes are likely. However, the experimental objectives may include the generation of new objectives or hypotheses on interesting pathways or genes that may be highlighted as a result of the microarray study.

Although it may seem perfectly obvious to state this, the objectives of a study are likely to be met only if that study is designed in a way that is consistent with meeting those objectives, and a study can be designed to meet the objectives only if they can be clearly articulated before initiating the study. Thus, we urge investigators to articulate clearly what they hope to get out of microarray studies at the end so that those microarray studies can be designed from the beginning to meet those objectives.

Experimental Design

As we design experiments to compare specific treatments or conditions, variation in fluorescence or radioactivity measurements, which are biomarkers of the amount of mRNA, will reflect not only differences among treatment conditions of interest but also differences among the measured units (experimental units) that are not of experimental interest, i.e., "nuisance variables."[11–14] Microarray experiments can be affected by numerous nuisance variables including printer head, chip lot, reagent lot, day of extraction, scanners, and technicians.[13,15–17] Such variables can be controlled in two ways. A nuisance variable can be effectively removed by eliminating variation nuisance factors. For example, technician effects could be removed if the same technician extracts all mRNA samples. Even if a nuisance variable cannot be removed, its potential to bias results can still be reduced if the variation in the nuisance factor is spread evenly over each treatment group, i.e., treatments and nuisance variables are orthogonal. For example, if an experiment has three treatments and chips are processed on 2 d, the same number of chips from each treatment group should be processed on the first day, and the same number of chips from each treatment should be processed on the second day, i.e., each treatment group is equally represented on each day.[11,12,14,18]

In addition, many nutritionists may select single agents for study (e.g., vitamins and minerals), one food (e.g., soy protein), or one style of eating (e.g., low fat) as the focus of the experiment. In the latter cases, there may be multiple stimuli of gene expression and a failure to control for the variation in the non-nutritive components; thus, unless there is careful consideration of these factors, there can be extensive confounding of the results.

Sample Size

Sample size has a major effect on how confidently genes can be declared differentially expressed or not differentially expressed.[15,19] Choosing the best sample size for microarray research may not be well served by traditional power analysis, which is based on control of type I errors for detection of a specifically sized difference for a single hypothesis test (e.g., 80% power at $\alpha = 0.05$ to detect a 0.5 standard deviation reduction in a metabolite due to a treatment); although methods for calculating sample sizes have been developed along these lines,[20,21] this may not be the most appropriate manner to calculate sample size. First, every gene is differentially expressed a different amount, so it is difficult to specify a biologically meaningful amount of change to base power calculations on. Second, if one uses a Bonferroni correction, the sample size required per group for a chip that covers all expressed human genes would be more than 250 per group. Third, many genes likely will be expressed differentially and they will not be independent, which violates assumption underlying Bonferroni- and Sidak-type corrections for multiple testing. We believe it is more appropriate in microarray analysis to choose sample size based on control of the false discovery rate (FDR)[22] and the expected discovery rate (EDR). The FDR is an estimate of the expected proportion of genes declared significant that is in fact not differentially expressed, i.e., that are "false discoveries."[23,24] The EDR is the expected proportion of genes that is truly different between two conditions that are found to be significantly different. FDR and EDR values for planned studies may be computed for variable sample sizes by using data from previous studies.[25]

The relation between FDR and EDR is complicated and a function of the sample size per group, the significance level (α), the magnitude of the effect, and the amount of biological heterogeneity in the samples. In general, for a constant sample size, choosing a smaller significance level will decrease the EDR (a negative result) and FDR (a positive result), and increasing the sample size will increase the EDR and decrease the FDR. In no case should fewer than five observations be made per experimental condition because this is the point where non-parametric statistical methods can be used, but the sample size may need to be far larger (40 or more) for a subtle effect in heterogeneous tissue samples from a multiracial, multi-gender, and multiage study of humans. An alternative method is to do a very small pilot project (three per group) and extrapolate the sample size needed to achieve a given FDR and EDR for a given experiment.[25]

Statistical Analysis

Objectives, experimental design, and the statistical analysis plan should be jointly constructed so that the data analysis answers the question asked or provides a powerful test of the hypothesis proposed with well-validated statistics. Fold change is very common in the microarray literature; it can be used to estimate the effect of treatment, but there are better quantitatives such as standardized mean difference, so fold change is not appropriate for inference. Valid inferential procedures will take into account the amount of variability in the gene and the sample size of the study; these procedures are commonly "frequentist" statistical methods, but Bayesian methods are becoming more prevalent.[26–29] The particular inferential statistical analysis used will depend on the question asked or hypothesis proposed. There are at least three classes of frequentist statistical analyses that have been applied to microarray data: 1) class descriptive methods or supervised analytical methods for identifying genes that differ in expression between two or more experimental conditions,[12,30] 2) class predictive methods for building models to predict which group new samples belong to,[31,32] and 3) class discovery or unsupervised analytical methods for dividing non-homogeneous samples into at least two groups.[4,33] Because microarray analysis is a very new technology with novel experimental and statistical issues, we strongly recommend collaboration with a biostatistician experienced in microarray study design and analysis throughout all phases of experimentation, in particular during experimental design and planning and getting guidance on the "wet" laboratory side of the experiment from investigators with previous experience in microarray experimentation.

Data Handling and Interpretation

Microarray experiments generate volumes of data that many biological researchers may not be accustomed to; a single Affymetrix chip will generate about 50 mB of data. After initial processing, each chip will provide 6000 to 60 000 numbers for analysis. After analysis, summary statistics, such as changes in expression and associated significance probabilities, will be available for all of the roughly 6000 to 60 000 genes on the original chips. Sorting through significance tests for 60 000 genes "manually" and trying to deduce biological principles is a Herculean task. Biology itself has evolved from a "one-gene, one-postdoc" mode of operation to whole genome analysis perhaps faster than anyone anticipated or any of us could have adequately prepared for. The volume of data generated by the new technology is overwhelming. Consideration should be paid to how data are stored, viewed, and interpreted.[34] When planning an experiment, one should be familiar with the pluses, minuses, and costs of existing online and bioinformatic resources that could aid interpretation, such as Genbank, Unigene, Locuslink, SOURCE,[35] OMIM (www.ncbi.nlm.nih.gov), KEGG (www.genome.ad.jp/kegg/),[36,37] Medminer (discover.nci.nih.gov/textmining/filters.html), Dragon (pevsnerlab.kennedykrieger.org/dragon.htm),[38] Onto-express (www.onto-express.org),[39] TIGR database (www.tigr.org), and Celera database (www.celera.com), to name a few.[40–42] Despite the availability of many online resources with basic data on individual genes, knowledge of their functions and interactions with other genes is still quite primitive, and many genes do not have available functional information. The "telling of the story" of a microarray experiment can be assisted only by these resources and requires considerable thoughtful interpretation from individuals with biological and statistical expertise. It is at this juncture that the "playful creativity" of nutrition researchers becomes critical to turn daunting volumes of data into valuable new biological and nutritional knowledge.

EXAMPLE

To better illustrate the principles we have outlined in this paper, we develop an example based on the observation that a diet high in soy is associated with lower rates of breast cancer. There are a great many questions around this topic we could ask. Should we look in breast tissues[43] or in the immune system[44]? Should we study humans, rats, or cell lines? Should we look at short- or long-term exposure and at what concentration? Should we use whole soy, processed soy, or a component? Any combination is possible and probably interesting; here we propose to make an initial study of changes in gene expression in human breast cancer cell lines after short-term exposure to three different doses of pure commercial grade genistein,[45,46] thus allowing us to look for any dose response relations in gene expression changes. Our goal is to identify potentially signaling pathways that are altered by genistein.

Because this is our initial study of genistein on breast cell lines, we use a well-studied cell line that we believe to be stable and has been used extensively in other microarray studies, e.g., one of the NCI-60 cell lines.[47,48] The use of a well-studied cell line allows us to compare the baseline expression with previous reported profiles to verify that the cell line is behaving correctly. Due to our greater experience with Affymetrix arrays,[3,6,49–51] we choose to spend a little more and use the Hu133A chip, which will give us measures of approximately 22 500 genes, including most well-characterized genes. We are unaware of any published studies of the effects of genistein on gene expression levels in breast cancer but are aware of studies on the effect of genistein on prostate cancer cell lines,[52] which found large changes in gene expression; thus we propose to use seven chips per group. The total experiment will have 28 chips (seven controls and seven at each of the three genistein concentrations), giving good power even if a few chips are lost during quality control[53] or the effect of genistein is less in a breast cell line than in a prostate cell line. A smaller sample size could be used ($n = 5$), but we would like a good list of targets and high confidence in our results. Based on our total sample size of 28, we purchase a single lot of chips to do the experiment and sufficient reagents to do the experiment plus 10%. For this experiment we use type 1 biological replication,[54] and multiple aliquots of the same cell line are grown and treated individually. Although not as generalizable as treating multiple cell lines, this is the easiest type of biological replication. All 28 aliquots are treated once per day at each concentration for 1 wk and harvested 24 h after the last genistein dose. We are unable to process all the samples on 1 d, so all are harvested on a single day, spun down, and snap frozen. The RNA is extracted from four samples from each condition on the first day and from three samples on the second day, with the RNA labeling process occurring similarly. We develop a plan with the microarray core to randomly assign the chips to fluidics system well and the order that the chips are scanned after leaving the hybridization oven. The data are collected on the same scanner using identical settings. The *.cel images are processed using the RMA[55] of Bioconductor to get good estimates of the expression levels for the genes. These images and the processed data are then entered into a SQL database.

This experiment has four defined groups in this study, and we would like to find those genes that are differentially expressed in response to genistein; thus, we will use class descriptive statistics. The first step of analysis will be a robust normalization procedure such as quantile-quantile.[56] Analysis of variance or a more robust non-parametric method such as a bootstrapping procedure will be used to detect whether there are differences in gene expression between the groups.[57] Alternatively, a linear test of trends may be applied to look for dose response in gene expression.[58] This is followed by a estimation of the posterior true positive, true negative, and false positive rates[24] and false discovery rates[22] to derive confidences that genes are differentially expressed.

Once the data are analyzed we use Source and Dragon databases to annotate the genes and Onto-express on an interesting subset of genes to get the significance of the different GO functional categories. This is followed by extensive literature review and the potential use of other software packages. Some of the most interesting genes will be verified by reverse transcriptase quantitative polymerase chain reaction or other method to establish the validity of some of the results.[59] This hopefully leads to the discovery of several interesting pathways for us to study in the future. The results are then written for publication, and the data are potentially deposited in one of the public databases.[60]

CONCLUDING COMMENTS

Microarrays will allow nutritional scientists to conduct many exciting, interesting, and insightful projects. However, great effort will be needed in careful planning, experimental design, analysis, and interpretation of microarray experiments to turn mountains of data generated by this hot new technology into credible and useful biological knowledge.

REFERENCES

1. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. Science 1996;274:610
2. Lockhart DJ, Ding H, Byrne M, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996;14:1675
3. Lee C-K, Kloop RG, Weindruch R, Prolla TA. Gene expression profile of aging and its restriction by caloric restriction. Science 1999;285:1390
4. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863
5. Perou CM, Jeffrey SS, van de RM, et al. Distinctive gene expression patterns in

human mammary epithelial cells and breast cancers. Proc Natl Acad Sci USA 1999;96:9212
6. Lee CK, Allison DB, Brand J, Weindruch R, Prolla TA. Transcriptional profiles associated with aging and middle age-onset caloric restriction in mouse hearts. Proc Natl Acad Sci USA 2002;99:14988
7. Sreekumar R, Unnikrishnan J, Fu A, et al. Impact of high-fat diet and antioxidant supplement on mitochondrial functions and gene transcripts in rat muscle. Am J Physiol Endocrinol Metab 2002;282:E1055
8. Chen CC, Shieh B, Jin YT, et al. Microarray profiling of gene expression patterns in bladder tumor cells treated with genistein. J Biomed Sci 2001;8:214
9. Nadler ST, Attie AD. Please pass the chips: genomic insights into obesity and diabetes. J Nutr 2001;131:2078
10. Lee ML, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci USA 2000;97:9834
11. Churchill GA. Fundamentals of experimental design for cDNA microarrays. Nat Genet 2002;32:490
12. Kerr MK, Churchill GA. Statistical design and the analysis of gene expression microarray data. Genet Res 2001;77:123
13. Coombes KR, Highsmith WE, Krogmann TA, Baggerly KA, Stivers DN, Abruzzo LV. Identifying and quantifying sources of variation in microarray data using high-density cDNA membrane arrays. J Comput Biol 2002;9:655
14. Yang YH, Speed T. Design issues for cDNA microarray experiments. Nat Rev Genet 2002;3:579
15. Lee MT, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci USA 2000;97:9834
16. Kerr MK, Churchill GA. Statistical design and the analysis of gene expression microarray data. Genet Res 2001;77:123
17. Mirnics K. Microarrays in brain research: the good, the bad and the ugly. Nat Rev Neurosci 2001;2:444
18. Kerr MK, Churchill GA. Experimental design for gene expression microarrays. Biostatistics 2001;2:183
19. Allison DB, Coffey CS. Two-stage testing in microarray analysis: what is gained? J Gerontol A Biol Sci Med Sci 2002;57:B189
20. Pan W, Lin J, Le CT. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. Genome Biol 2002;3:22
21. Hwang D, Schmitt WA, Stephanopoulos G, Stephanopoulos G. Determination of minimum sample size and discriminatory expression patterns in microarray data. Bioinformatics 2002;18:1184
22. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 1995;57:289
23. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001;98:5116
24. Allison D, Gadbury G, Heo M, et al. A mixture model approach for the analysis of microarray gene expression data. Comput Stat Data Anal 2002;39:1
25. Gadbury G, Page G, Edwards J, et al. Power analysis and sample size estimation in the age of high dimensional biology: a parametric bootstrap approach and examples from microarray research. 2004 Stat Methods Med Res (in press)
26. Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. Genet Epidemiol 2002;23:70
27. Broet P, Richardson S, Radvanyi F. Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. J Comput Biol 2002; 9:671
28. Medvedovic M, Sivaganesan S. Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics 2002;18:1194
29. Townsend JP, Hartl DL. Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments. Genome Biol 2002;3:71
30. Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. J Comput Biol 2000;7:819
31. Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999;96:2907
32. Dudoit S, Fridlyland J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002;97:77
33. Brazma A, Vilo J. Gene expression data analysis. FEBS Lett 2000;480:17
34. Cheung KH, White K, Hager J, et al. A microarray database for large-scale gene expression analysis. Proc AMIA Symp 2002;7:140
35. Diehn M, Sherlock G, Binkley G, et al. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Res 2003;31:219
36. Kanehisa M. The KEGG database. Novartis Found Symp 2002;247:91
37. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res 2002;30:42
38. Bouton CM, Pevsner J. DRAGON. Database referencing of array genes online. Bioinformatics 2000;16:1038
39. Khatri P, Draghici S, Ostermeier GC, Krawetz SA. Profiling gene expression using onto-express. Genomics 2002;79:266
40. Palsson BO. What lies beyond bioinformatics? Nat Biotechnol 1997;15:3
41. Brazhnik P, de la FA, Mendes P. Gene networks: how to put the function in genomics. Trends Biotechnol 2002;20:467
42. de la FA, Brazhnik P, Mendes P. Linking the genes: inferring quantitative gene networks from microarray data. Trends Genet 2002;18:395
43. Lamartiniere CA. Protection against breast cancer with genistein: a component of soy. Am J Clin Nutr 2000;71:1705S-7S
44. Watanabe S, Uesugi S, Kikuchi Y. Isoflavones for prevention of cancer, cardiovascular diseases, gynecological problems and possible immune potentiation. Biomed Pharmacother 2002;56:302
45. Sarkar FH, Li Y. Mechanisms of cancer chemoprevention by soy isoflavone genistein. Cancer Metastasis Rev 2002;21:265
46. Li Y, Sarkar FH. Gene expression profiles of genistein-treated PC3 prostate cancer cells. J Nutr 2002;132:3623
47. Ross DT, Scherf U, Eisen MB, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 2000;24:227
48. Scherf U, Ross DT, Waltham M, et al. A gene expression database for the molecular pharmacology of cancer. Nat Genet 2000;24:236
49. Lee CK, Weindruch R, Prolla TA. Gene-expression profile of the ageing brain in mice. Nat Genet 2000;25:294
50. Kayo T, Allison DB, Weindruch R, Prolla TA. Influences of aging and caloric restriction on the transcriptional profile of skeletal muscle from rhesus monkeys. Proc Natl Acad Sci USA 2001;98:5093
51. Weindruch R, Kayo T, Lee CK, Prolla TA. Gene expression profiling of aging using DNA microarrays. Mech Ageing Dev 2002;123:177
52. Li Y, Sarkar FH. Gene expression profiles of genistein-treated PC3 prostate cancer cells. J Nutr 2002;132:3623
53. Li C, Hung WW. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol 2001;2:32
54. Yang YH, Speed T. Design and analysis of comparative microarray experiments. In: Speed T, ed. *Statistical analysis of gene expression microarray data.* Boca Raton: Chapman & Hall/CRC, 2003:35
55. Irizarry R, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4:249
56. Bolstad B, Irizarry R, Astrand M, Speed T. A comparison of normalization methods for high density oligonucleotide array based on variance and bias. Bioinformatics 2003;19:185
57. Wolfinger RD, Gibson G, Wolfinger ED, et al. Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 2001;8:625
58. Lee Y, Nelder J. Hierarchical generalized linear models. J R Stat Soc B 1996; 58:619
59. Rockett JC. To confirm or not to confirm (microarray data)—that is the question. Drug Discov Today 2003;8:343
60. Ball CA, Sherlock G, Parkinson H, et al. An open letter to the scientific journals. Bioinformatics 2002;18:1409