Editorial

Microarrays and Epidemiology: Ensuring the Impact and Accessibility of Research Findings

Melissa A. Troester,[1,2] **Robert C. Millikan,**[1,2] **and Charles M. Perou**[2,3,4]

[1]Department of Epidemiology [2]Lineberger Comprehensive Cancer Center, [3]Department of Genetics, and [4]Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

Introduction

Gene expression microarrays have contributed to important discoveries in cancer biology in the past decade and, in some cases, have permanently altered our concepts of the disease. For example, microarrays led to the identification of the basal-like subtype of breast tumors (1-4), which was later discovered to be overrepresented in younger African-American women with breast cancer (5) and which shows distinct risk factor profiles (6). Microarrays allowed the discovery of molecular subtypes of breast cancer, but epidemiologic studies served as a necessary complement to the genomic studies, helping to address concerns about potential selection biases, generalizability, and population distributions of the genomic subtypes. Thus, interdisciplinary work at the interface of genomics and epidemiology benefits both (7). In a recent *Cancer, Epidemiology, Biomarkers & Prevention* (CEBP) editorial ''Microarrays and Epidemiology: The End of the Beginning not the Beginning of the End...'' Webb and coauthors initiated an important discussion about how microarrays can be used to further the field of cancer epidemiology (8). We seek to continue this discussion and to highlight specific ways in which investigators, peer reviewers, and editors might increase the impact of microarray-based epidemiologic research through improved data access.

Causal Contrasts and Genomic Data: Controlling Confounding and Avoiding Biases

One application of genomics is to examine correlations between gene expression and exposure to specific environmental risk factors. Here, confounding variables are a real concern. There are growing numbers of examples where microarray data are used as outcomes, actually thousands of outcomes, simultaneously. These studies have the potential to yield important insights about mechanisms of carcinogenesis. For example, Russo et al. (9) reported a genomic signature associated with parity status in histologically normal breast tissue and Andrew et al. (10) presented a lymphocyte gene expression signature associated with high versus low exposure to drinking water arsenic. Webb et al. (2007) identified several challenges and limitations associated with matching and restriction for confounding control in microarray analyses, but an approach that has received surprisingly little attention is application of statistical methods. There is no reason that logical model building and model selection strategies cannot be developed for high dimensional data, consistent with appropriate causal diagrams (7) and using backward selection methods used for single outcomes. Model selection could be done on a gene-by-gene basis using a rational algorithm that emphasizes parsimony and the impact of the covariate on the main effect estimate. Linear multivariable models are already being used in microarray analyses, and epidemiologists should ensure that the important methods and contributions of our field are not neglected in these analyses. Given the large percentage of genes (>10%) in the human genome whose expression seems to be affected by age (11), such methods would seem to be imperative for preventing artifactual signatures.

Many problems that become evident during analysis are actually problems stemming from study design. For example, in the search for gene expression signatures associated with risk of breast cancer, how are ''high risk'' and ''low risk'' conceptualized? Is it a conglomeration of known risk factors that are of interest (e.g., a Gail model score), or should a single factor (e.g., reproductive history) be examined? What are the implications of grouping heterogeneous exposure combinations? When does combining heterogeneous groups lead to exposure misclassification? If a rare exposure is the exposure of ultimate interest, how can subjects be selected to ensure adequate power with a smaller sample size? Some environmental and behavioral exposures are collected more accurately than others and the influence of error in exposure classification should be carefully considered during design and analysis of genomic studies. In addition, selection bias is an important concern for microarray studies (12), and although we may not be able to avoid some selection biases (e.g., the need for moderate quantities of carefully handled RNA may mean that larger tumors are overrepresented compared with smaller tumors where

Cancer Epidemiol Biomarkers Prev 2009;18(1):1–4

Received 9/16/08; revised 9/30/08; accepted 10/8/08.

Requests for reprints: Melissa A. Troester, University of North Carolina at Chapel Hill, Campus Box 7435, Chapel Hill, NC 27599. Phone: 919-966-7408; Fax: 919-966-2089. E-mail: troester@unc.edu

doi:10.1158/1055-9965.EPI-08-0867

Table 1. Guidelines for statistical analysis and reporting of microarray studies for epidemiology

Checklist		Comment
Do	Clearly state the causal contrasts being studied and how exposures were defined.	Consider and discuss potential exposure misclassification and its impact on findings.
Do	Identify potential confounders using appropriate causal diagrams and adjust for them using statistical methods.	Age may be an important confounder of exposure-gene expression associations.
Do	Discuss how the genomic methods used may have created selection biases.	
Do	Use class prediction on independent test sets and report sensitivity and specificity of predictors.	
Don't	Misuse standard terminology such as ''test set'' or unsupervised.	
Do	Report measures of internal reproducibility.	Report intraclass correlation coefficients for duplicates.
Do	Supplement technical validation (e.g., quantitative PCR) with external validation where possible.	Use publicly deposited data to confirm pathways affected by exposure.
Do	Deposit data in public databases at the time of review and before publication.	

NOTE: This table is formatted as a supplement to a table presented by Dupuy and Simons (13).

the entirety of the tissue is used for diagnostic purposes), we should remain aware of them and discuss their potential impacts openly. Investigators should be clear at the outset of the design and analysis stages about the contrasts they are setting up to avoid interpretation errors and to establish the ideal study design for investigating their hypotheses.

Developing Formal Guidelines for Review of Observational Genomic Studies?

It may be useful to establish some formal guidelines or standards for microarray-based epidemiology. An excellent review, including guidelines on statistical analysis and reporting of microarray research, has been published previously (13). The ''checklist'' presented therein is recommended reading for microarray users and also for peer reviewers and editors. Table 1 supplements the checklist of Dupuy and Simon with ''Dos'' and ''Don'ts'' that seem most critical for microarray studies in epidemiology. Guidelines such as ''Don't use any information from the test set for developing the classifier'' (13) seem straightforward, but thinking and reporting clearly about statistical methods for microarrays is challenging. In preparing this editorial, we found several examples of recent, high-impact articles where ''test'' sets were used to refine or redefine predictors. It is also not uncommon for important terminology to be misused. For example, consider the use of ''unsupervised'' and ''supervised'' clustering. The former uses all genes or filtered genes but does not select or rank genes based on a relation with exposure or outcome. Supervised clustering explicitly preselects genes based on their association with a class variable. Proper use of these terms is essential to allow readers to appropriately interpret cluster figures. Simple mislabeling of a figure as unsupervised can lead to vastly different conclusions about the importance of a variable in determining gene expression.

In studies where predictive value or class discrimination based on the gene set is the goal, cross-validation methods could be used and estimates for the classification algorithms should be reported. We support the proposal of Dupuy and Simon that the sensitivity and specificity of the predictor (for a binary outcome) should be reported where possible. For example, sensitivity, specificity, and 95% confidence intervals for each were reported for training and test sets in a recent CEBP article to identify signatures of tobacco smoke exposure in leukocytes (14). In addition, the fully specified classifier should be reported with all gene names and fully detailed algorithms, so that it can be used and evaluated by others (for example, see ref. 15). Reporting of gene lists or clusters, or even ontological analyses, derived from a single data set will not help to advance the field if these reports are not accompanied by predictors and data that other investigators can use to conduct further analyses and compare results across platforms.

In epidemiology, the internal reproducibility of biomarker measurements is often reported as part of a standard Materials and Methods section. Although the reporting of assay (or internal) reproducibility has not been widely practiced in genomics research, it is often the case that replicate microarrays are done on the same sample (either two replicates on the same RNA preparation or two separate preparations of RNA from the same sample). Wherever such replicates are available, an intraclass correlation coefficient can be computed for the duplicates and should be reported (16).

The Necessity of Public Data Access for Establishing Reproducibility and Increasing Study Impact

The ultimate standard for genomic research is external reproducibility. To ensure that microarray research is not at the ''beginning of the end'' (8), reproducibility and synthesis across studies is essential. As the number of epidemiologic studies aimed at identifying signatures associated with a particular class increases, how will the reproducibility of the findings be assessed? The most important step that authors can take to increase the impact and accessibility of their research is to adhere to Minimum Information About a Microarray Experiment (MIAME) principles. This means that the primary (i.e., raw) microarray data are to be made publicly available at the time of publication. Editorial policies for the AACR journals in general, and CEBP in particular, require authors to deposit their data in public databases in compliance with MIAME;[5] yet, none of the seven gene expression microarray articles published in CEBP between January 2007 and July 2008 provided either a Gene Expression Omnibus[6] or ArrayExpress[7] accession. One of the seven articles did note that the data would be available publicly, and we were able to find this data set by searching Gene Expression Omnibus. CEBP reviewers and editors could play an important role in encouraging or enforcing compliance with MIAME standards. In an era where findings from genomic association studies have been difficult to replicate, publicly available microarray data sets could afford unprecedented opportunities for meta-analysis and systematic assessments of reproducibility. Public gene expression data, in combination with public genome-wide association study data such as the Cancer Genetic Markers of Susceptibility data,[8] will allow integration of phenotypic information with insights about inherited susceptibility. All associated clinical, biological, and/or epidemiologic variables explored in any given article need to be made available and easily linked to each sample. In fact, the Gene Expression Omnibus data tables allow for inclusion of these variables in addition to the gene expression data, and thus, the best practice is to include all of the variables associated with each sample when uploaded to the public repositories. This does not mean that authors need to publish their entire data sets with the first instance of publishing their gene expression data, but all variables used in the published analyses should be available for secondary analyses at the time of review and publication.

Accessible data and detailed methods mean more citations and greater impact for articles, but it also means that data sets can be integrated. Webb et al. raised a concern about making comparisons across platforms, arguing that integration across platforms was ''largely meaningless.'' Certainly, there are important challenges in merging cancer data sets (17); however, there are successful examples of integrating data sets to establish reproducible biology (15). Gene lists from similar studies may have little overlap, but individual genes should be thought of as markers for important pathways and underlying biology (2). The test of reproducibility should not be recapitulation of a list or some percentage of a list, but reproducible predictive ability. Does the gene list associate with the same phenotype or class in other data sets? This shift to thinking about the relevant pathways and biology instead of individual genes has been accompanied by development of important analytic tools that use gene sets as the fundamental unit of analysis (18-20). The notion of a gene set as the functional unit instead of the individual gene also means that measuring RNA expression by another method (such as quantitative PCR) may offer value in assessing technical reproducibility, but does little to establish biological reproducibility. Biological reproducibility is our greatest challenge, and this challenge can be addressed if fully annotated data sets are made available to the scientific community.

As interdisciplinary research at the interface of gene expression microarrays and epidemiology continues to develop, reviewers and editors are increasingly being asked to possess expertise in both epidemiology and genomic methods. Would having guidelines on statistical and epidemiologic analyses help to support the advancement of the field and ensure high-impact research? We suggest that guidelines and automatic public release of the primary data will greatly assist the sound usage of microarray data in epidemiologic studies and ''avoid the unrealistic hype and excessive skepticism'' (13) that microarray-based clinical investigations have experienced.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

References

1. Sørlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001;98:10869–74.
2. Hu Z, Fan C, Oh DS, et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics 2006;7:96.
3. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature 2000;406:747–52.
4. Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A 2003;100:8418–23.
5. Carey LA, Perou CM, Livasy CA, et al. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. JAMA 2006;295:2492–502.
6. Millikan RC, Newman B, Tse CK, et al. Epidemiology of basal-like breast cancer. Breast Cancer Res Treat 2008;109:123–39.
7. Millikan R. The changing face of epidemiology in the genomics era. Epidemiology 2002;13:472–80.
8. Webb PM, Merritt MA, Boyle GM, Green AC. Microarrays and epidemiology: not the beginning of the end but the end of the beginning. Cancer Epidemiol Biomarkers Prev 2007;16:637–8.
9. Russo J, Balogh GA, Russo IH. Full-term pregnancy induces a specific genomic signature in the human breast. Cancer Epidemiol Biomarkers Prev 2008;17:51–66.
10. Andrew AS, Jewell DA, Mason RA, et al. Drinking-water arsenic exposure modulates gene expression in human lymphocytes from a u s. Population. Environ Health Perspect 2008;116:524–31.
11. Tan Q, Zhao J, Li S, et al. Differential and correlation analyses of

[5] http://cebp.aacrjournals.org/misc/ifora.shtml
[6] http://www.ncbi.nlm.nih.gov/geo/
[7] http://www.ebi.ac.uk/aerep/
[8] https://caintegrator.nci.nih.gov/cgems/

microarray gene expression data in the CEPH Utah families. Genomics 2008;92:94 – 100.

12. Potter JD. At the interfaces of epidemiology, genetics and genomics. Nat Rev Genet 2001;2:142 – 7.
13. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 2007;99:147 – 57.
14. Lampe JW, Stepaniants SB, Mao M, et al. Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. Cancer Epidemiol Biomarkers Prev 2004;13:445 – 53.
15. Fan C, Oh DS, Wessels L, et al. Concordance among gene-expression-based predictors for breast cancer. N Engl J Med 2006;355:560 – 9.
16. Hu Z, Troester M, Perou CM. High reproducibility using sodium hydroxide-stripped long oligonucleotide DNA microarrays. Biotechniques 2005;38:121 – 4.
17. Lusa L, McShane LM, Reid JF, et al. Challenges in projecting clustering results across gene expression-profiling datasets. J Natl Cancer Inst 2007;99:1715 – 23.
18. Segal E, Shapira M, Regev A, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 2003;34:166 – 76.
19. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545 – 50.
20. Efron B, Tibshirani R. On testing the significance of sets of genes. Ann Appl Stat 2007;1:107 – 29.