Editorial

Microarrays and Epidemiology: Not the Beginning of the End but the End of the Beginning...

Penelope M. Webb,[1] Melissa A. Merritt,[1,2] Glen M. Boyle,[1] and Adèle C. Green[1]

[1]Population Studies and Human Genetics Division, Queensland Institute of Medical Research and [2]School of Population Health, University of Queensland, Brisbane, Queensland, Australia

In recent years, molecular epidemiologists seem to have been seduced by the technological possibilities of conducting molecular analyses of human tissues for entire study populations. But has the promise of biotechnology-enhanced epidemiologic studies been fulfilled? Take for example the explosion in studies looking at genetic polymorphisms, particularly single nucleotide polymorphisms and cancer risk. It has been said (1) that, at most, 10% of some tens of thousands of research papers investigating single nucleotide polymorphisms in relation to cancer have been validated by other researchers and even fewer have resulted in clinical or public health outcomes. So is this the end of the road for such studies? Not according to Sellers commenting recently in *Cancer Epidemiology, Biomarkers & Prevention* (2). He contended that, although the glitter of the search for gene-environment interactions may be fading, we should not give up. Instead, we should be more realistic about our aims and be smarter about controlling the ''noise'' if we are to identify the occasional salient effect from innumerable spurious associations.

And what of other new laboratory technologies that are starting to make their way into epidemiologic and clinical studies? Another example is the ''microarray.'' A Medline search for the term ''microarray*'' returns ~20,500 hits, with more than 4,000 yearly since 2004. When coupled with the term ''epidemiol*,'' the number of papers immediately drops to ~220 and, of them, only a small minority report the use of microarrays in an epidemiologic context. A quick scan of mainstream epidemiology journals shows 21 hits in *Cancer Epidemiology, Biomarkers & Prevention*, 1 in the *American Journal of Epidemiology*, and none at all in either *International Journal of Epidemiology* or *Epidemiology*. Clearly, epidemiology as a discipline has not yet come to the microarray party... have we learned from the disappointments of the single nucleotide polymorphisms-and-cancer rush or are we holding back for lack of understanding of how to exploit this new technology? It is perhaps timely to consider more closely the possibilities offered by microarrays and how they might realistically be harnessed by cancer epidemiologists to gain better insights into cancer control.

So what exactly is a microarray? As the name suggests, it is an array of biological elements laid out in a known, uniform order on a solid support. Microarrays can be broadly classified into three general groups: DNA, protein, and tissue microarrays. The DNA array comprises tens of thousands of known DNA sequences with RNA extracted from the patient sample binding to the DNA on the array. On a protein array, proteins in the test sample bind to the antibodies on the array. Fluorescent dyes or other markers enable the binding to be visualized and quantified, allowing simultaneous assessment of the level of expression of thousands of genes/proteins in one sample. In contrast, a tissue microarray is a single paraffin block that contains tiny biopsies from tens or hundreds of patient samples. These can then be sectioned simultaneously for immunohistologic analysis or *in situ* hybridization, providing a rapid and convenient way to screen a large number of patient samples for a single tumor marker. The DNA expression microarray is the most developed of the high-throughput assays and the most widely used to date.

How can microarrays be exploited for epidemiology? Cancer epidemiologists look for plausible biological mechanisms to explain associations between potential risk factors and cancer. One immediate potential benefit is the possibility that molecular signatures from microarrays may allow distinction between subsets of histologically similar tumors. By reducing heterogeneity in our case groups, we may reduce the noise, making it easier to identify causal associations. Microarrays can also help identify the specific biological mechanisms that may be causally involved. Unlike most studies to date, which have focused on inherited polymorphisms that affect all cells in the body, microarrays can directly compare gene and/or protein expression in the specific target tissues that epidemiologists believe may be exposed to the risk factor of interest.

Tobacco smoke is a risk factor that has been studied in this way, on the grounds that if this exposure did not result in a detectable microarray signature, then it would be difficult to identify signatures associated with risk factors with more subtle effects (3). A comparison of gene expression in samples of airway epithelium from 34 current smokers and 23 never-smokers identified several genes whose expression differed significantly between the smokers and nonsmokers (4). Smoking has also been shown to influence gene expression signatures in circulating leukocytes, although these cells have only limited exposure to tobacco-related carcinogens (3). Furthermore, after only 2 years of smoking cessation, expression of the smoking-induced genes among 18 former smokers began to resemble that of never-smokers, although aberrant expression of selected genes persisted (4).

Some very small studies have attempted to look for gene expression patterns associated with other epidemiologic factors, with reports of differential gene expression between blood samples from liver transplant recipients with a body mass index <29 (n = 5) versus >29 (n = 7; ref. 5) and between primary cultures of human breast epithelial cells from nulliparous (n = 3) and parous women (n = 3; ref. 6). Such studies suffer from their lack not only of power but also of any control for potential confounding factors. Indeed, as pointed out by Potter (7, 8), one of the major limitations of microarray studies in the epidemiologic and clinical setting has been their inherent observational nature coupled to a lack of application of the standard approaches of epidemiology (i.e., assessment of chance, bias, and confounding).

Can these limitations be overcome? Chance and bias are major issues because, for pragmatic reasons, including cost and the challenges of fresh tissue collection, microarray studies often include small numbers of samples with limited supporting information. New developments allowing use of fixed tissue, such as the tissue arrays, will alleviate some of these

Cancer Epidemiol Biomarkers Prev 2007;16(4):637–8

doi:10.1158/1055-9965.EPI-07-0156

problems but, in doing so, may introduce others. The samples in a tissue array are usually <1.0 mm wide and thus may not accurately represent a heterogeneous tumor—use of multiple biopsies can reduce this problem but the challenge of interpreting possible discordant results remains.

And what of confounding? The gene expression in any sample will be influenced by the combined exposures (e.g., age and smoking status) of the person from which it was collected. It cannot be assumed that tissues collected for analysis of a particular trait or factor will differ only according to that factor. As epidemiologists, we have models that allow us simultaneously to adjust our results for multiple potential confounders. Although these are certainly not perfect, they should bring us closer to the underlying truth in our usual dealings with hundreds or thousands of subjects and relatively few exposure variables. How then can we deal with the possibility of confounding in a study with tens of subjects and tens of thousands of genes? One simple approach is to step back to the cruder techniques used more widely before computing advances made logistic regression available at the press of a button. Restriction of the sample set to homogeneous groups where differences other than those of interest are minimized is one option, as is matching where individuals in the test group(s) are matched with comparable individuals in the comparison group. It is important to note, however, that, although matching improves comparability and efficiency of study design, it does not ensure control of confounding. In our own pilot study of this approach, we assessed gene expression in invasive ovarian cancer samples of the serous subtype according to subgroups of an exposure of interest, in this case, parity. It is well known that parous women have ~50% lower risk of ovarian cancer than nulliparous women, but the molecular effects underlying this protection are not well understood. Comparing ovarian cancer samples from nine women with zero or one pregnancy with nine women who had five or more pregnancies identified 238 differentially expressed genes (compared with 172 expected by chance). But this crude analysis did not account for differences in age or in the proportion of current smokers in the two groups, which, as shown above, could affect gene expression. When we restricted the analysis to nonsmokers over the age of 55, no significant differences in expression were detected, suggesting that the differences originally observed may have been artifacts due to factors other than parity and indicating that further investigation was essential.

The high ratio of genes to samples in a microarray study also means that reproducibility is crucial. Techniques exist to sift through the masses of data generated to identify the associations that are least likely to be due to chance, but it is impossible to rule this out completely. An added complication is the considerable variation in results obtained from gene expression studies using different platforms or technologies (9, 10). Many factors may contribute to this variation, including differences in sample preparation, hybridization, data normalization, and even the specific microarray platform used. It is anticipated that this variation will be reduced as standard platforms, probes, and operating procedures are adopted but, for the moment, comparison of results across different platforms is largely meaningless.

Given the possible problems with microarray technology and studies (both observational and experimental) to date, there have still been some remarkable developments and discoveries that may lead to real advances in cancer control. Recent successes include the use of expression levels of six genes, originally identified from separate microarray studies, to predict survival of patients with diffuse large B-cell lymphoma (11); identification of a panel of 254 genes whose expression is associated with metastatic dissemination of cutaneous melanomas (12); use of expression profiling to identify the origin of cancers of unknown primary site (13); and a validated 70-gene signature that adds independent prognostic information to clinicopathologic risk assessment for women with node-negative breast cancer (14). This last gene signature is currently being tested in a large randomized clinical trial.

In conclusion, microarray technology has the potential to add valuable information to clinical and epidemiologic studies of cancer but well-established epidemiologic principles should not be sacrificed in the process. Ideally, studies should be conducted prospectively with collection of large numbers of well-annotated samples to minimize bias and allow assessment of confounding. Attention must be paid to the array platform and methods to be used and validation is essential. If used judiciously, microarrays could provide a clearer picture of the cancers we are studying as well as enhancing our understanding of the molecular pathways underlying known exposure-cancer associations. This could lead to better options for cancer prevention, diagnosis, and treatment and our ultimate goal of cancer control.

References

1. Schmidt C. SNPs not living up to promise; experts suggest new approach to disease. J Natl Cancer Inst 2007;99:188–9.
2. Sellers TA. The beginning of the end for the epidemiologic focus on gene-environment interactions? Cancer Epidemiol Biomarkers Prev 2006;15:1059–60.
3. 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.
4. Spira A, Beane J, Shah V, et al. Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proc Natl Acad Sci U S A 2004;101:10143–8.
5. Skirton H. International Society of Nurses in Genetics (ISONG) Conference abstracts. Nurs Health Sci 2006;8:125.
6. Guo S, Russo IH, Russo J. Difference in gene expression profile in breast epithelial cells from women with different reproductive history. Int J Oncol 2003;23:933–41.
7. Potter JD. At the interfaces of epidemiology, genetics, and genomics. Nat Rev Genet 2001;2:142–7.
8. Potter JD. Epidemiology, cancer genetics, and microarrays: making correct inferences, using appropriate designs. Trends Genet 2003;19:690–5.
9. Bammler T, Beyer RP, Bhattacharya S, et al. Standardizing global gene expression analysis between laboratories and across platforms. Nat Methods 2005;2:351–6. Erratum in: Nat Methods 2005;2:477.
10. Jarvinen AK, Hautaniemi S, Edgren H, et al. Are data from different gene expression microarray platforms comparable? Genomics 2004;83:1164–8.
11. Lossos IS, Czerwinski DK, Alizadeh AA, et al. Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. N Engl J Med 2004;350:1828–37.
12. Winnepenninckx V, Lazar V, Michiels S. Gene expression profiling of primary cutaneous melanoma and clinical outcome. J Natl Cancer Inst 2006;98:472–82.
13. Tothill RW, Kowalczyk A, Rischin D, et al. An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. Cancer Res 2005;65:4031–40.
14. Buyse M, Loi S, van't Veer L, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 2006;98:1183–92.