Commentary
Microarray foray

Robert J Coffey and David Threadgill

Vanderbilt University, Nashville, Tennessee, USA

Foray is defined as 'to pillage in search of spoils' and this is what Perou *et al* [1] have done in terms of analyzing gene expression in human breast cancer by microarray technology. The technology behind microarrays was developed over the past several years once it became apparent that new, more powerful analytical approaches were needed to utilize the flood of genomic data and resources being acquired through the various genome projects. Two dominant platforms have evolved; one revolves around *in situ* synthesis of oligonucleotide probes on support matrices and the other consists of physically stamping specific target DNAs onto solid supports. For economic, flexibility, and sensitivity reasons, the second platform has become favored generally by the academic research community. This latter technology, as well as the gene-clustering software that is crucial to data analysis, has been pioneered by the laboratories of Brown and Botstein at Stanford University [2–4]. Conceptually, the technology is easy to grasp. RNA from two separate samples is isolated and reverse transcribed with incorporation of a red and green fluor. Equivalent fluorescent intensity is hybridized to a chip that contains thousands of approximately 500–2000 base pair gene fragments, and relative gene expression is determined in a pairwise manner.

The Brown and Botstein group selected the subset of genes from a 5500 complementary DNA chip that were differentially expressed in human breast tumors compared with cultured normal human mammary epithelial cells by threefold or greater in three out of 26 arrays. This subset of 1200 genes was then analyzed by gene-clustering analysis. Several problems have arisen. A minor one – that a gene that may change dramatically in only one array – is ignored. In addition, the criteria of a threefold change in gene expression may exclude biologically important genes with expression that falls into the 'noise' range. A more substantive problem that needs to be addressed by the microarray com-

munity is the choice of an appropriate reference sample. To quantify differences, a gene must be expressed in the reference sample. Diversity and reproducibility are two critical issues. It is unlikely that the Stanford control (human mammary epithelial cells) exhibits sufficiently diversified gene expression to measure quantitatively all of the genes that were not expressed in the mammary epithelial cells. The ideal 'generic' control would express all genes and would be a renewable source, thus providing reproducibility. In a perfect world, all users of microarray technology would utilize the same 'generic' reference sample, and results could then be compared across experiments and across different laboratories. The microarray community might settle on RNA from a pool of cell lines of different cell types, or, even better, from pools of primary cell cultures that exhibit broadly diverse gene expression patterns.

Nevertheless, a wealth of data has been generated in the analysis of Perou *et al* [1]. Specific cell type signatures were identified. (The cynic will argue that a good pathologist can already provide this information at a fraction of the cost; however, consider that adequate diagnostic factors to identify the 30% of node-negative women with 1–2 cm breast tumors that will go on to develop metastases have not been identified with current standard methodologies.) Clusters, or groups, of coexpressed genes were identified that showed dramatically different expression levels across the tumor cohort tested. One example was a cluster of genes that are known to be regulated by the interferon/STAT1 signal transduction pathway. A potentially more important cluster of genes was the set of genes that was associated with cell proliferation. This set of genes was downregulated when cell growth was reduced *in vitro*, and was also shown to be highly expressed in all of the tumors tested that were deemed 'proliferative', as assessed by high Ki-67 staining index. One could potentially select one to two representative genes from each

'cluster' of genes, and profile paraffin-embedded tumor sections for these genes by whatever means that are available (eg *in situ* hybridization, immunohistochemistry).

What is the value of this technology in enhancing our understanding of cancer? With the advent of more dense arrays, we will determine the dimensions of the black box and perhaps identify some gray areas. The Stanford group is presently arraying 23 000 complementary DNAs on a single microarray (CM Perou, personal communication). Although this may only represent 20% of all human genes, it does represent a nonrandom set of selected genes and it is likely that most of the genes critical to the process of transformation are represented. (It will be worthwhile to examine the full set of expressed genes, but we predict its major value will be to identify tissue-specific and developmentally specific genes.) In our judgment, expression studies should be approached with an open mind, and in most instances investigators should simply let the observed patterns 'do the talking'. We do not favor boutique chips (eg chips composed of all known apoptotic or cell cycle genes), because it is often the genes that one does not suspect ahead of time that are the major players. After all, microarray technology was originally envisioned to free us from any preconceived 'prediction' of the biology and to let the biology provide us with insights to formulate better, unbiased hypotheses. Furthermore, major signal transduction pathways in mammalian cells do not appear to behave as discrete units transcriptionally. Also, there is much action taking place post-transcriptionally that is not addressed by this technology. Imaging mass spectrometry, and other proteomic approaches should provide complementary information.

Eventually, one would like to isolate nests of cells in a tissue from discrete stages in tumor progression and perform microarray analysis on each stage. Enhanced sensitivity and/or RNA amplification is needed, because 1.5–2 µg messenger RNA or 100 µg of total RNA is presently required. At the end of the day, one still must decide on which genes to study further. After this, the hard part starts. Ultimately, functional genomics equals good, old-fashioned cell biology. For more information on microarray technology and resources, visit our website [5].

Authors' affiliations: Robert J Coffey (Department of Medicine and Cell Biology, Vanderbilt University, Nashville, TN 37232, USA) and David Threadgill (Department of Cell Biology, Vanderbilt University, Nashville, TN 37232, USA)

Correspondence: Robert J Coffey, CC2201 Medical Center North, Vanderbilt University, Nashville, TN 37232, USA.
Tel: +1 615 343 6228; fax: +1 615 343 1591;
e-mail: coffeyrj@ctrvax.vanderbilt.edu

Reference

1. Perou CM, Jeffrey SS, Van de Rijn M, *et al*: **Distinctive gene expression patterns in human mammary epithelial cells and breast cancers.** *Proc Natl Acad Sci USA* 1999, **96**:9212–9217.
2. Schena M, Shalon D, Davis RW, Brown PO: **Quantative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270**:467–470.
3. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95**:14863–14868.
4. Eisen MB, Brown PO: **DNA arrays for analysis of gene expression.** *Methods Enzymol* 1999, **303**:179–205.
5. http://www.microarrays.com