Efforts Aimed at Reducing Noise, Data Overload in Microarrays

Microarrays have helped researchers identify previously unrecognized subtypes of cancers, and more recently they have been put to the test to determine their ability to identify cancers with better or worse prognosis (see News, Vol. 97, No. 5, p. 331, “Trial and Error: Prognostic Gene Signature Study Design Altered”). Now, researchers are working to find the best way to take the tool to a new level of complexity by asking it to help them identify genes involved in the basic biology of tumors.

Experts in the field expect that the approach will work—but caution that it won’t be entirely straightforward. “For me, prediction is something we can often do without understanding the underlying biology, and that is much more difficult,” said Jill Mesirov, Ph.D., director of computational biology and bioinformatics at the Broad Institute at the Massachusetts Institute of Technology and Harvard in Cambridge, Mass.

The problem boils down to issues of noise in the data and the ability to demonstrate biological relevance. For Mesirov, the use of gene sets, which are sometimes referred to as metagenes, can help address both problems.

If, instead of analyzing the data in terms of individual genes, an investigator looks for gene sets that are enriched in a given tumor type, the data are likely to be more reproducible because the signal-to-noise ratio improves when 400 gene sets are analyzed versus 10,000 genes. Thus, genes that wouldn’t show up very well individually may do so if they
are coordinately expressed and biologically important.

When she speaks to biologists, Mesirov points out that the biggest problem in many array experiments is that scientists end up with either too many differentially expressed genes—or none. If they have too many, they can cherry-pick the genes on the list that look most interesting to them based on prior knowledge, but those aren’t necessarily the most important, and therefore the approach can be misleading.

The quintessential example of gene set analysis comes from a diabetes study led by the Broad Institute’s Vamsi Mootha, in which Mesirov’s group participated several years ago. They performed microarray analysis on muscle biopsy samples from patients with diabetes and from control subjects who had normal glucose tolerance. At the individual gene level, there were no statistically significant differences in the expression data. When they used gene set enrichment analysis, they found a statistically significant decrease in the genes in the oxidative phosphorylation pathway. Individually, the expression level of each gene decreased between the control and diabetic samples by only 15%–20%, but because there were approximately 100 genes in the set, the difference became statistically significant.

The other advantage of gene sets, said Mesirov, is that they often come with substantial biological information, which provides a head start in a functional analysis. Of course, the output data are only as good as the data used to derive the gene set, cautioned Mesirov, which means that evaluating the strength of those data before intertwining them with the current experiment pays off. (Her team bundles several already annotated gene sets in the software they have developed, and she regularly asks researchers to send her new sets so she can expand that collection.)

“Everybody who has a scanner and can extract RNA is producing microarray data,” said Dennis Slamon, M.D., Ph.D., professor of Hematology and Oncology at the University of California Medical School in Los Angeles. “That is part of the problem with the field—no one is separating the wheat from the chaff very well.”

To get around the problem in his own laboratory, he now relies on constraints-based analyses, in which he first separates tumor samples into known breast cancer subtypes, including Her2 status, estrogen receptor status, or BRCA1 or -2 status, and triple-negative disease, which lacks all three markers. Working from that starting place, Slamon can discern pathway or gene expression differences that arise in one tumor type versus another, which may mean that the gene or pathway is involved in tumorigenesis rather than in the final tumor phenotype. In other words, a gene set upregulated in all of the tumor types may be a signature for a late-stage disease phenotype, such as aggressiveness or invasiveness, but it is unlikely to be causal in the early stages of the disease, as that gene expression pattern occurs in tumors that have different underlying genetic problems.

Using this strategy, his team found that the vascular endothelial growth factor (VEGF) is dramatically upregulated in Her2-positive cancers. VEGF is also upregulated in some of the tumors from other breast cancer classes, but the consistency of the upregulation in Her2 tumors led his group to think it wasn’t just a bystander, but part of the underlying problem in this pathology.

“It’s interesting that you can make the intellectual link between Her2 and VEGF, but you still need to go back and do the biology,” said Slamon. To do this, his team looked to see if the Her2–VEGF correlation held up in a variety of samples. They also found that treating cells with trastuzumab (Herceptin), an antibody against Her2/neu protein, caused a drop in VEGF expression and that patients with higher VEGF expression tended to have more aggressive disease.

From these and other preclinical data, which suggested a causative role for VEGF in the Her2 breast cancer phenotype, the team tested a combination of trastuzumab and a recombinant monoclonal antibody against VEGF in a phase I trial with nine patients with Her2-positive cancer. Two patients had a complete response, three had partial responses, and there were no unexpected toxicities, according to data Slamon presented earlier this year at the annual meeting of the American Association for Cancer Research. The team has now launched a 50-patient phase II trial.

Experts agree that, to obtain that kind of success, researchers must use a reasonable number of samples. Just what that number is, though, is unclear, especially at the outset of an experiment because the “right” number will be determined in part by the expression level of the genes under study.

David Bowtell, Ph.D., director of research and professor at the Peter MacCallum Cancer Institute in Melbourne, Australia, and his group recently published a study that used microarrays to categorize tumors of unknown primary origin. During that study, they looked at the number of samples required to derive a reproducible signature that could define the tissue of origin of a tumor. Their data show that although 10 samples were enough to adequately represent a relatively homogeneous tumor type such as colon cancer, they needed substantially more samples from histologically variable cancers, such as ovarian and lung, to obtain a reproducible signature.

To gain enough ovarian tumor samples and to have clean, complete clinical data that go along with them, Bowtell is leading the Australian Ovarian Cancer Study, which aims to collect...
1,000 fresh-frozen tumor samples by 2006. They already have about 750 and are in the midst of testing 500 on arrays. The team plans a conventional top-down array approach, looking for molecular indicators of response, for example, but they also plan to look for enrichment of gene sets.

Looking at the field now, Mesirov, Slamon, and Bowtell agreed that the shifts from single-gene analysis to gene sets and from correlates of response to searches for the biological underpinnings of cancer reflect the maturation of the tool and the field. “When we first started this approach, we used unsupervised hierarchical clustering to analyze the data and hoped that it would fall out in useful fashion,” said Bowtell. “Then we used supervised clustering to relate genes to the thing we wanted to find—for example, outcome versus gene profile. The unsupervised way seemed pure, but because of the problem of sample number and gene number, the noise could obscure the signal. Given that a supervised approach predefines associations, it is critical that these are independently validated.” Now, he says, the microarray approach is being further refined with the use of gene sets, for example, which if reproducible could lead to important biological insights.

The key for Slamon, though, is relatively straightforward: With enough good samples, including strong clinical annotation, strong biological signals will shine through. “If genes are really critical and common, you should be able to find them in a few samples consistently,” he said.

—Rabiya S. Tuma