MICROARRAYS

Reading between the lines

A phenotype prediction tool helps 'fill in the blanks' for expression microarrays, extending their predictive power and uncovering once-hidden biases.

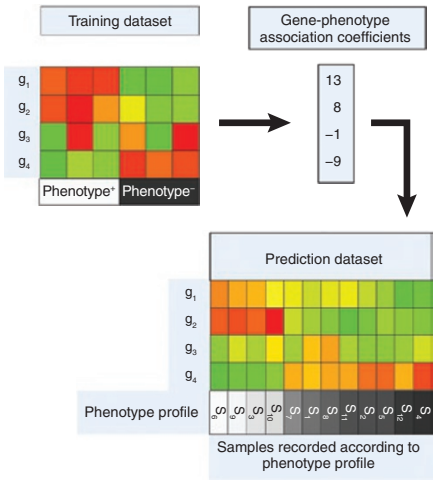
A well-designed expression microarray experiment can tell a great story about what happens to cells or tissues under different conditions, but at least one researcher is concerned that it might not be telling the right story.

Jasmine Zhou at the University of Southern California had been working with microarray data for some time before growing concerned with the extent to which many studies oversimplify how genetic variations manifest themselves. "In analyzing data, we realized that there's actually a lot of genomic data out there but very little phenotype information associated with [these] data," she says, adding that in many cases, studies simply focus on bipolar metrics such as cancerous versus noncancerous tissues.

To rectify this problem, members of her laboratory developed a system for identifying likely genetic correlates for particular phenotypes and applying those to predict phenotype in similar datasets in which such information is unavailable. The PhenoProfiler method begins with a training dataset, in which phenotype can be directly compared against gene expression data to identify 'signature genes' that associate significantly with phenotypic variation. The resulting correlations can then be used to derive predicted phenotypes from other microarray datasets.

In a recently published proof-of-concept study, Zhou's team—led by first authors Min Xu and Wenyuan Li—applied PhenoProfiler to 587 publicly available microarray datasets from the National Center for Biotechnology Information's Gene Expression Omnibus (GEO). In an initial series of training-prediction trials, they found that 81% of their predicted phenotype profiles were accurately reflected in the original samples.

These profiles also yielded unexpected predictive benefits. For example, Zhou's team analyzed datasets intended to distinguish between malignant grade III and highly invasive grade IV gliomas. PhenoProfiler proved effective at this task, but was also surprisingly capable of separating early stage grade II from grade III gliomas.



The PhenoProfiler method. A training dataset is used to identify significant associations (top) between 'signature genes' and phenotypic variation. These are then applied to other datasets to predict probable sample phenotypes (bottom). Reprinted with permission from the National Academy of Sciences, USA.

These capabilities also enabled detection of potentially confounding factors concealed in sample groups. For example, a rectal cancer study with 'training' and 'testing' sample sets had marked differences in extent of malignancy between the two groups rather than the intended random distribution of cancer prevalence and severity.

Phenotype prediction algorithms could help prevent such concealed biases. "It's not possible for scientists to look at all possible phenotypes of a sample," says Zhou. "[But] this will help scientists to identify phenotype context and get conclusions with less bias and get rid of confounding factors." As a first step, her group is making PhenoProfiler and their data available from their website.

However, Zhou also adds that the current state of phenotype data, often poorly structured, remains a major obstacle, and that better data annotation in future microarray studies is essential. "The more standardized the format, the more convenient it will be to use [these] data and the more powerful this method becomes," she says.

Michael Eisenstein

RESEARCH PAPERS
Xu, M. *et al.* Automated multidimensional phenotypic profiling using large public microarray repositories. *Proc. Natl. Acad. Sci. USA* **106**, 12323–12328 (2009).