

Meeting report

## Making sense of lung-cancer gene-expression profiles

Dennis A Wigle\*§, Ming Tsao†§ and Igor Jurisica\*¶

Addresses: Departments of \*Surgery, †Laboratory Medicine and Pathobiology, and ‡Computer Science, University of Toronto, Toronto, Ontario M5S 1A8, Canada. §Thoracic Oncology Site Group and ¶Division of Cancer Informatics, Princess Margaret Hospital, 610 University Avenue, Toronto, Ontario M5G 2M9, Canada.

Correspondence: Igor Jurisica. E-mail: ij@uhnres.utoronto.ca

Published: 30 January 2004

*Genome Biology* 2004, **5**:309

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2004/5/2/309>

© 2004 BioMed Central Ltd

---

A report on the Critical Assessment of Microarray Data Analysis (CAMDA'03) meeting and competition, Durham, USA, 12-14 November 2003.

---

The CAMDA meeting was started in 2000 by Simon Lin and Kimberly Johnson (Duke University Bioinformatics Shared Resource, Durham, USA). Patterned on the molecular modeling community's well-known Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment and related competitions in genomics, statistical genetics and computational toxicology, it allows participants to present analyses of common datasets to display novel methodology and results. All the speakers present an analysis of the same data, helping the audience gain an understanding of current capabilities and problems in the field. Conferences in previous years have concentrated on large datasets from experiments with model organisms. The added challenge of this year's competition was the provision of multiple clinical microarray datasets, with associated epidemiological information to heighten the stakes.

The organizers of this year's CAMDA meeting and competition [<http://www.camda.duke.edu>] had selected four lung cancer expression-profile databases as the data sources (Table 1), which had sparked clinical and pharmaceutical interest in addition to highlighting the latest approaches in microarray data mining. Over 150 statisticians, computer scientists, and biologists presented their analyses of the featured datasets. Jeffrey Morris (MD Anderson Cancer Center, University of Texas, Houston, USA) and his group were voted as winners by the audience for their approach to integrating data across different Affymetrix datasets (see below for details).

Relating critical assessment of microarray data analysis to clinical outcome for the deadliest of human cancers has been of particular interest to those profiling lung cancer. Despite more than 400 publicly available lung cancer expression profiles, the results so far have raised as many questions as answers. Many studies have clearly demonstrated correlations between expression profile and clinical outcome, but most of the gene sets that have been identified in this way have frustratingly little overlap with each other. The increasing volume of available data has not made the picture any clearer. The development of new methods for data mining, particularly when the data are analyzed in the context of associated clinical information, promises new inferences from existing data and may help to derive a true molecular system for further refining the stages of lung cancer.

Morris focused on the Affymetrix datasets from Michigan and Harvard (see Table 1) in an attempt to generate a

**Table 1**

**Data sources for CAMDA'03**

	Abbreviation used here	Reference
Affymetrix data	Harvard	Bhattacharjee <i>et al.</i> , <i>Proc Natl Acad Sci USA</i> 2001, <b>98</b> :13790-13795.
	Michigan	Beer <i>et al.</i> , <i>Nat Med</i> 2002, <b>8</b> :816-824.
cDNA spotted microarray data	Stanford	Garber <i>et al.</i> , <i>Proc Natl Acad Sci USA</i> 2001, <b>98</b> :13784-13789.
	Ontario	Wigle <i>et al.</i> , <i>Cancer Res</i> 2002, <b>62</b> :3005-3008.

meta-analysis of the combined data. He proposed a novel approach to the problem of analyzing two datasets in conjunction, even when they are on different versions of the Affymetrix human chip. Probe sets common to each array were identified and the data were combined for 4,101 different unigenes. Further filtering removed half the genes with the lowest mean expression levels across all samples, as well as genes with small standard deviations across the samples, to leave 1,036 genes for consideration. Multivariable Cox models were constructed for each of the 1,036 probe sets to look for genes providing prognostic information. A total of 26 genes were identified as predictors of patients' survival. Interestingly, none of these genes appeared in the list of the top 100 genes from the Michigan analysis, and only one was mentioned in the Harvard paper.

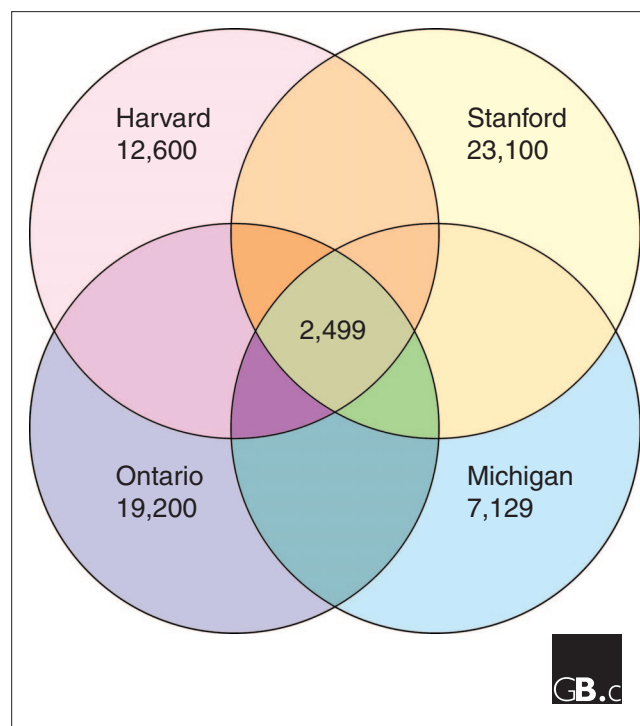
Two presentations from researchers at GlaxoSmithKline (GSK) also focused on a meta-analysis of the two Affymetrix datasets. They presented slightly different strategies, possibly representing views from research centers on different sides of the Atlantic. Xiwu Lin (GSK Biomedical Data Sciences Group, Collegetown, USA) integrated the data using both gene names and probe-set IDs. Initial principal component analysis revealed the two datasets to be completely non-overlapping, demonstrating the importance of normalization before further integrated analysis. A number of genes were shown to partition patients into groups with a high or low chance of survival, using the tree method for analyzing survival data in relation to significant genes identified by Cox modeling. These genes included *NME2* (encoding nucleoside diphosphate kinase B, a protein expressed in non-metastatic cells), *B2M* ( $\beta_2$ -microglobulin), *HSF1* (heat-shock transcription factor 1) and *PGK1* (phosphoglycerate kinase 1), the last of which has been recently identified in a number of genomic and proteomic studies as a potential indicator of cancer prognosis. Linda Robb (GSK Statistical Sciences Group, Stevenage, UK) discussed an approach that explores the data using principal component analysis before survival analysis and then uses Cox modeling for each gene to link variables associated with survival. Fisher's combined probability test (Fisher's meta-analysis) was used to combine *p*-values from the two analyses and to define a new modified *p*-value for the association of every gene with survival. This resulted in 33 genes with an adjusted *p*-value < 0.05. Many of these genes were not identified in the original publications.

In one of the few presentations that considered the data derived from spotted cDNA arrays (see Table 1), Geoff McLachlan (University of Queensland, Brisbane, Australia), showed evidence from the Stanford and Ontario studies that suggested that clustering of gene-expression data according to prognosis may be a more powerful predictor of disease outcome than current staging systems based on histopathology or extent of disease at presentation. Using the EMMIX-GENE algorithm developed by himself and colleagues [<http://www.maths.uq.edu.au/~gjm>], McLachlan was able

to identify genes correlated with clinical outcome. As with other analyses presented at the meeting, however, these genes did not appear to overlap well with ones identified in the original publications.

Overall, many presenters agreed that trying to combine data across different platforms to perform true meta-analyses was largely futile for anything other than the Affymetrix datasets. Even for these arrays, there are considerable challenges in trying to work across two different versions of the Affymetrix human gene chip. Spotted cDNA array data has the added challenge of variations in array construction, choice of reference RNA, data-normalization strategy and the extent of missing data, to name but a few of the possible variables. Despite this, Neil Hayes (Tufts University, Boston, USA) showed an interesting approach to data integration by creating ratios from Affymetrix data using gene-expression values from the reference samples used in the Stanford study. Whether such an approach will be robust enough for application to other cDNA studies remains to be determined.

Unfortunately, integrative analysis has shown that the gene overlap for the platforms used in the four datasets is soberingly low (the number of overlapping genes on the four



**Figure 1**  
Combining the four different microarray platforms used in the CAMDA competition gives 2,449 genes in common. The platforms used were Affymetrix HG-U95 (Harvard, with 12,600 probe sets), Stanford spotted arrays (with 23,100 clones), Toronto OCI 19k2 spotted arrays (Ontario, with 19,200 clones), and Affymetrix HuGeneFL (Michigan, with 7,129 probe sets). See Table 1 for references.

platforms used is shown in Figure 1). This is both good and bad news. The good news is that each study increases the overall search space for markers; the bad news is the limited chance for overlap in the set of differentially identified genes. Further complicating the picture in clinical studies is the potential for differences between tissue samples, which results in a further reduction in the chance of identifying overlapping markers.

The meeting also showed that computationally derived markers, even from multiple analyses on multiple datasets, clearly do not provide the level of confirmation necessary to translate into clinical utility. One can validate results in three ways: using different methodology, such as reverse-transcriptase PCR (RT-PCR), but still on the same RNA or tumor samples; using the same method but on different samples (such as more arrays on different tumors or patients); or using different method(s) on different samples. Only the third of these will provide validation that is considered adequate to translate into broad clinical impact.

Looking to the future, thoracic oncologists eagerly await data from the National Cancer Institute's Director's Challenge, which plans to collect and analyze Affymetrix data on over 600 lung cancer samples from multiple centers with associated clinical information in what will be the largest clinical microarray study to date in any tumor type. We hope this will provide clearer answers on which correlations between gene expression and clinical outcome are valid, so that identified markers can be incorporated into future clinical trials.