COMMENTARY

Mechanism-derived gene expression signatures and predictive biomarkers in clinical oncology

Edison T. Liu*

Genome Institute of Singapore, Genome 02-01, 60 Biopolis Street, Singapore

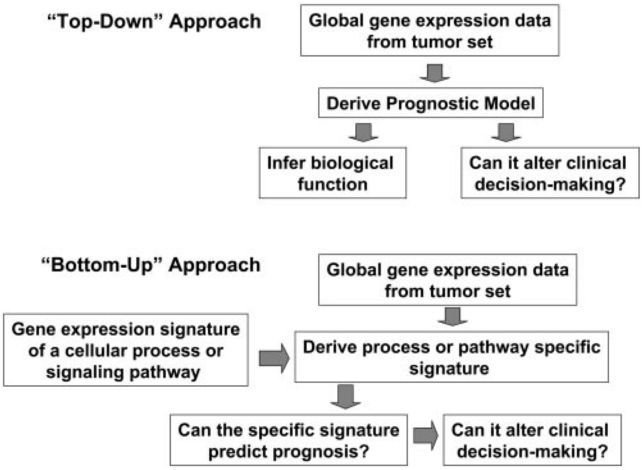
"Top-Down" Approach
Global gene expression data from tumor set
Derive Prognostic Model
Infer biological function
Can it alter clinical decision-making?
"Bottom-Up" Approach
Global gene expression data from tumor set
Gene expression signature of a cellular process or signaling pathway
Derive process or pathway specific signature
Can the specific signature predict prognosis?
Can it alter clinical decision-making?

Fig. 1. Schematic of the top-down vs. the bottom-up biomarker discovery approaches.

Medical scientists have always sought to uncover fundamental mechanistic explanations for human disease and to use this information to predict patient outcome and devise specific therapeutics. Although monogenetic diseases have been elucidated, the more common disorders often have complex or heterogeneous origins and involve the failure of multiple systems before disease is manifested. Breast cancer is an example. Many factors and genes have been implicated in the initiation of the disease (e.g., BRCA1, PTEN, P53, hormone exposure, irradiation, free radicals, etc.), but mortality is due to metastatic disease that requires invasion, evasion of immune surveillance, implantation in ectopic sites, continuous replication, cell migration, and angiogenesis (1). Capturing all of the genetic components that support these cellular processes has been a challenge for cancer cell biology.

The article by Chang *et al.* (2) in this issue of PNAS is interesting and novel on several fronts. First, it takes a mechanism-driven approach to prognostic biomarker discovery on a genome scale. Second, by focusing on biological mechanisms in the discovery process, Chang *et al.* have uncovered the catalog of genes involved in a potentially new cellular process that defines the clinical biology of breast cancer. Third, they have rendered these findings applicable for clinical decision making.

Mechanistically derived candidate biomarkers have been identified in the past, and the approach is not new. Rather than seeking biomolecules simply associated or correlated with a clinical outcome, the discovery process starts with a specific biological process and then asks whether perturbations of its components could predict cancer phenotype. The identification of the collection of mismatch repair enzymes involved in hereditary nonpolyposis colon cancer (HNPCC) is an example of a successful genomic discovery of evolutionarily related mismatch repair genes that were later confirmed as important in carcinogenesis (3, 4).

The application of microarray platforms to cancer biomarker discovery has used the standard design of assessing the correlations between individual gene expression and clinical outcomes such as survival. Biological plausibility of these candidate genes is considered only in retrospect and if a biochemical pathway is implicated (5, 6). By contrast, Chang and colleagues started with a specific physiologic mechanism, wound healing and its *in vitro* proxy, serum response of cultured fibroblasts, which they examined in exquisite detail in refs. 7–9. The operational definition of this expression response was genes that are seen in 50 fibroblast cultures whose expression changed after exposure to 10% serum. When the cell cycle-associated genes (10) were removed from this list, 512 core serum response (CSR) genes were identified and were considered representative of a "wound" signature. Using this expression cassette that confidently defines an *in vitro* core serum response, they posited that such a coordinated transcriptional cassette would have a role in the clinical behavior of cancer because migration and invasion are phenotypes seen both in wound response and invasive cancer. Their earlier work suggested that, even after eliminating growth-associated genes, the "activated" profile of the CSR gene cassette would predict for worse outcome in a variety of cancers.

The article by Chang *et al.* in this issue of PNAS goes several steps further to clinically operationalize this CSR gene cassette (8). Although hierarchical clustering can use the CSR genes to separate populations of tumors into prognostic groups, this approach cannot be applied to categorizing individual samples as would be necessary in the clinical setting. To accomplish this sorting, the authors developed a quantitative scale (called correlation score of this scalable wound signature) to assess how close a tumor's configuration resembled the activated profile. A threshold for this score can be selected for any sensitivity and specificity desired (2). The correlation score for the CSR gene cassette was found to be an independent variable in predicting survival when a variety of clinical parameters were measured. In clinical breast cancer treatment, the decision to treat with adjuvant chemotherapy depends on the probability of relapse as predicted by lymph node status, tumor size, and histologic grade. Several clinical guidelines taking account of these parameters, such as the National Institutes of Health or the St. Gallen consensus criteria, are used in clinical decision making (11). The wound response correlation score outperformed either consensus criteria in properly selecting patients who would not need adjuvant chemotherapy. Moreover, the combined application of two array-based prognostic scores further improved the ability to predict risk of metastatic disease.

The authors cite this decision making as a "bottom-up" approach that builds a prognostic predictor from defined expression modules assigned to a specific pathogenic mechanism. These guidelines are in

See companion article on page 3738.

*E-mail: liue@gis.a-star.edu.sg.

contrast to a "top-down" approach that identifies a predictive expression cassette by empiric association with disease outcome and makes no mechanistic assumptions (Fig. 1). If, indeed, this model-based bottom-up approach can be validated, then it is conceivable that tumor phenotype can be devolved to several mechanisms, each represented by specific transcriptional profiles. That cellular processes can be described (or even driven) by the coordinate expression of specific gene cassettes is clearly not a new biological concept; however, the ability to compute mathematically the likely presence of these cassettes has been a recent and welcomed contribution to the functional genomics field.

Community efforts such as the Alliance for Cell Signaling have provided large-scale expression data on specific cellular states for quantitative modeling, which were used to decipher the underlying dynamics of cellular regulatory networks operative in B lymphocytes (12, 13). Twelve major gene regulatory groups linked to definable gene ontologies could be abstracted into expression modules, which are the starting ground for constructing the integrated network of signaling pathways. This approach was used in an integrated analysis of 1,975 published microarrays encompassing 22 tumor types that uncovered 456 statistically significant expression modules (14). Certain modules were associated with specific tumor types: repression of a growth-inhibitory module in acute lymphoblastic leukemia, repression of steroid catabolism module in hepatocellular carcinoma, and activation of an osteoblastic module in breast cancer. Similarly, Gene Set Enrichment Analysis (GSEA) takes into account prior information about gene relationships in pathways in formulating association statistics. Mootha *et al.* (15) analyzed microarray data from diabetic muscle biopsies using GSEA and discovered that a set of genes involved in oxidative phosphorylation and activated by PGC-1α are coordinately decreased in human diabetic muscle. Thus, pathway inference through computational strategies is progressively feasible and valid.

An intriguing observation noted by Chang *et al.* (2) was that the 70-gene prognostic signature previously identified and validated in breast cancer (5) had minimal gene overlap with the CSR cassette and the two independently predicted for metastatic relapse in the same patient cohort. This finding suggests that different expression modules may independently contribute to the tumor phenotype, and we might expect more prognostic gene sets, especially from a bottom-up approach. For example, some of the independent cancer expression modules noted above (14) or the metastasis expression cassette (16) would be candidate modules for testing as potential prognostic or predictive marker sets. Other important cassettes to be explored might be cell proliferation or tumor grade-associated gene modules and pathway-specific modules, such as those induced by myc or p53 (L. Miller and E.T.L., unpublished data).

Different expression modules may independently contribute to the tumor phenotype.

Finally, the work by Chang *et al.* (2) provides a plausible thread that links the putative transcriptional response in wound healing with aggressive cancer behavior. Although this relationship has been observed on the macroscale level in clinical conditions such as cirrhosis of the liver and hepatocellular carcinoma, and in the similarities between wound keratinocytes and squamous cell carcinomas (17), the ability to capture the potential universe of specific genes operating in the two physiologic states (i.e., healing and cancer) moves us closer to uncovering fundamental mechanisms contributing to the cancer phenotype (1).

The technologies and, especially, the analytical approaches described in Chang *et al.* (2) and associated papers that are so common today are truly revolutionary for clinical medicine. The development of HER-2 as a single biomarker in breast cancer required >10 years from the first publication for several confirmatory studies of thousands of patients to be completed and for controversies to be resolved. The studies were usually conducted in tandem, one marker at a time. The identification of independent and validated prognostic and predictive gene sets accounting for 700–1,000 potential biomarkers all came out in a period of 3–4 years. This dramatic truncation of developmental time can be fundamentally attributed to the availability of the human genome sequence and the deposition of the complete data sets of array data and clinical information from these biomarkers studies in the public domain. The genome sequence allowed for the development of tools such as oligonucleotide expression arrays, where the probes are all computationally derived from the complete sequence. The availability of tumor array data for a number of breast cancer studies with the associated clinical outcomes has allowed for cross-validation of biomarker cassettes through database interrogations rather than generating an entirely new clinical study from scratch (18). Although confirmatory clinical studies designed to test a putative biomarker cassette remain the gold standard, such *in silico*-derived and database-enabled cross-validations give confidence to clinical trialists that there is a greater likelihood for success in the biomarkers presented to them for testing. All in all, the work discussed herein augurs well for the biomarker field.

I am supported by the Agency for Science, Technology, and Research (A*STAR) of Singapore.

1. Hanahan, D. & Weinberg, R. A. (2000) *Cell* **100,** 57–70.
2. Chang, H. Y., Nuyten, D. S. A., Sneddon, J. B., Hastie, T., Tibshirani, R., Sørlie, T., Dai, H., He, Y. D., van't Veer, L. J., Bartelink, H., *et al.* (2005) *Proc. Natl. Acad. Sci. USA* **102,** 3738–3743.
3. Fishel, R., Lescoe, M. K., Rao, M. R., Copeland, N. G., Jenkins, N. A., Garber, J., Kane, M. & Kolodner, R. (1993) *Cell* **75,** 1027–1038.
4. Nicolaides, N. C., Papadopoulos, N., Liu, B., Wei, Y. F., Carter, K. C., Ruben, S. M., Rosen, C. A., Haseltine, W. A., Fleischmann, R. D., Fraser, C. M., *et al.* (1994) *Nature* **371,** 75–80.
5. van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., *et al.* (2002) *N. Engl. J. Med.* **347,** 1999–2009.
6. Bullinger, L., Dohner, K., Bair, E., Frohling, S., Schlenk, R. F., Tibshirani, R., Dohner, H. & Pollack, J. R. (2004) *N. Engl. J. Med.* **350,** 1605–1616.
7. Chang, H. Y., Sneddon, J. B., Alizadeh, A. A., Sood, R., West, R. B., Montgomery, K., Chi, J. T., van de Rijn, M., Botstein, D. & Brown, P. O. (2004) *PLoS Biol.* **2,** E7.
8. Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C., Trent, J. M., Staudt, L. M., Hudson, J., Jr., Boguski, M. S., *et al.* (1999) *Science* **283,** 83–87.
9. Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., *et al.* (2002) *Nat. Med.* **8,** 68–74.
10. Whiteld, M. L., Sherlock, G., Saldanha, A. J., Murray, J. I., Ball, C. A., Alexander, K. E., Matese, J. C. Perou, C. M., Hurt, M. M., Brown, P. O. & Botstein, D. (2002) *Mol. Biol. Cell* **13,** 1977–2000.
11. Zujewski, J. & Liu, E. T. (1998) *J. Natl. Cancer Inst.* 90, 1587–1589.
12. Gilman, A. G., Simon, M. I., Bourne, H. R., Harris, B. A., Long, R., Ross, E. M., Stull, J. T., Taussig, R., Bourne, H. R., Arkin, A. P., *et al.* (2002) *Nature* **420,** 703–706.
13. de Bivort, B., Huang, S. & Bar-Yam, Y. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 17687–17692.
14. Segal, E., Friedman, N., Koller, D. & Regev, A. (2004) *Nat. Genet.* **6,** 1090–1098.
15. Mootha, V. K., Handschin, C., Arlow, D., Xie, X., St. Pierre, J., Sihag, S., Yang, W., Altshuler, D., Puigserver, P., Patterson, N., *et al.* (2004) *Proc. Natl. Acad. Sci. USA* **101,** 6570–6575.
16. Ramaswamy, S., Ross, K. N., Lander, E. S. & Golub, T. R. (2003) *Nat. Genet.* **33,** 49–54.
17. Pedersen, T. X., Leethanakul, C., Patel, V., Mitola, D., Lund, L. R., Dano, K, Johnsen, M., Gutkind, J. S. & Bugge, T. H. (2003) *Oncogene* **22,** 3964–3976.
18. Lin, C. Y., Strom, A., Vega, V. B., Kong, S. L., Yeo, A. L., Thomsen, J. S., Chan, W. C., Doray, B., Bangarusamy, D. K., Ramasamy, A., *et al.* (2004) *Genome Biol.* **5,** R66.