

# Commentary

## Defining the galaxy of gene expression in breast cancer

Edison T Liu<sup>1</sup> and Christos Sotiriou<sup>2</sup>

<sup>1</sup>Genome Institute of Singapore, Singapore

<sup>2</sup>Jules Bordet Institute, Brussels, Belgium

**Correspondence:** Edison T Liu, MD, Executive Director, Genome Institute of Singapore, 1 Science Park Road, #05-01 Science Park II, Singapore.  
Tel: +65 6827 5210; e-mail: [gislue@nus.edu.sg](mailto:gislue@nus.edu.sg)

Received: 6 March 2002

Revisions requested: 28 March 2002

Revisions received: 12 April 2002

Accepted: 12 April 2002

Published: 23 May 2002

*Breast Cancer Res* 2002, **4**:141-144

© 2002 BioMed Central Ltd  
(Print ISSN 1465-5411; Online ISSN 1465-542X)

### Abstract

Recent reports detailing the expression profiles of primary breast cancer have pointed to the utility of this approach in defining subclasses with distinct molecular configurations and clinical behaviour. Some of the subclasses can be predicted by current molecular tests: estrogen receptor status, p53 staining, and HER-2 overexpression. Others, however, are novel subgroups and may represent distinct cellular types. The results from two recent studies suggest common principles of classification by expression profiling. These principles are examined and the impact of these results on understanding the biology and the clinical behaviour of breast tumors is explored.

**Keywords:** arrays, breast cancer, expression profiling, gene expression

Expression profiling describes a group of technologies that simultaneously assess the transcriptional activity of a large number of genes. These technologies have ranged from direct PCR amplification-based approaches, to cDNA or to oligonucleotide two-dimensional arrays. Although originally thought to be simply a way to multiplex a collection of individual gene analyses ('large-scale Northern'), the reality is that expression arrays produce a composite picture rich with higher levels of biological information.

This complex picture is obtained through a statistical estimate of the coordinate expression of a large number of genetic elements relative to each other. Moreover, these expression 'cassettes' do not require precision in the analysis of individual gene expression, but rely on the 'consensus' expression (either up or down) of large gene sets. This primary principle of genome-wide expression profiling, that data from comprehensive sets of gene probes give greater biological information than the sum of the information from individual genes, is exemplified in several papers recently published on expression arrays and breast cancer.

Recent manuscripts by Sorlie *et al.* [1] and van't Veer *et al.* [2] appear to bear this principle out. Sorlie *et al.* investigated 78 breast cancers of different stages and histologic types, and then pursued a second analysis in individuals treated with unresectable T3/T4 breast cancer and doxorubicin monotherapy [1]. Van't Veer *et al.*, however, examined 117 node-negative breast cancers, with 18 breast cancers from women bearing germline *BRCA1* mutations [2]. Sorlie *et al.* interrogated their samples with a cDNA array containing 8102 genetic elements, and van't Veer *et al.* employed a 25,000-element oligo-based array. In examining the collective results of these two studies, several truths emerge.

The first noteworthy point is that the tumors can be clustered into groups in an unbiased manner by the composite expression profiles. The dominant factor partitioning the tumors is the estrogen receptor (ER) status. This has also been observed by other workers [3]. Furthermore, in both the studies of Sorlie *et al.* [1] and of van't Veer *et al.* [2], expression profile substructures could be discerned that appear to be related to a molecular point of origin.

ER = estrogen receptor; PCR = polymerase chain reaction.

For example, Sorlie *et al.* showed that erbB2 positivity appeared to be associated with a distinct expression cluster of ER-negative tumors [1]. Moreover, ER-positive tumors could be separated into expression groups that ultimately correlated with the frequency of *TP53* mutations (luminal subtype A with a 13% *TP53* mutation rate, and luminal subtype B and subtype C with 80% and 40%, respectively, bearing *TP53* mutations).

Van't Veer *et al.* did not investigate subgroup structure except to identify a set of 100 genes that could segregate positive tumors from sporadic tumors [2]. Possibly because of the small numbers of tumors studied, however, no further clusters could be discerned beyond this level of secondary structure.

An unresolved question arising from these studies is whether these expression cassettes segregating with ER status are due to the molecular/genetic origins of the breast cancer or whether they reflect the cell lineage of the tumor and are independent of the inducing genes/oncogenes. To address this, our group, in conjunction with Dr Jeff Greene at the National Cancer Institute (USA), examined the profiles of a number of mouse mammary cancers arising from animals bearing various onco-transgenes, using two different promoters targeting different compartments in the mammary gland. We found that the biochemical pathways served by the oncogenes determined the ultimate tumor clusters, and that the different promoters only altered the substructure and not the major cluster determinants (KV Desai *et al.*, manuscript submitted). This work suggests that both the genetic origins of a tumor and the cell lineage of origin can effectively determine the downstream expression profile. The genes of origin, however, may have more impact.

The second point of note is that these expression cassettes can provide a refined estimate of prognosis, perhaps beyond those clinical indicators currently available to us. Sorlie *et al.* identified a luminal subgroup (subgroup A) of ER-positive tumors associated with the best outcome [1]. Interestingly, these tumors had the lowest rate of *TP53* mutations, suggesting that the ER status and *TP53* mutations significantly alter the expression profile of a tumor and may be the major determinants of breast cancer behavior [4].

The study of Sorlie *et al.* was not ideally suited to analyze the prognosis given the extensive and heterogeneous nature of the disease in these patients (T3/T4/N2). Van't Veer *et al.* addressed this problem by investigating a narrow subset of breast cancer patients: T1/T2, N0, and M0, all younger than 55 years of age, treated only with local regional therapies [2]. They found 231 genes significantly associated with disease outcome as defined by the presence of distant metastasis at the 5-year mark. Van't

Veer *et al.* could then subsequently collapse this list to a core set of 70 prognostic markers. These prognostic genes appear to have functions involved in the cell cycle, DNA repair, and chromosomal integrity.

Interestingly, van't Veer *et al.* tested the ability of this array-derived prognosis 'expression profile' to identify patients who would benefit from adjuvant chemotherapy [2]. The prognosis expression profile was then compared with other standard guidelines for treatment of node-negative breast cancer (National Institutes of Health and St Galen consensus guidelines). Although the array-based expression profile could correctly identify patients who would benefit from adjuvant chemotherapy, van't Veer *et al.* found that the greatest benefit was that this profile effectively reduced by about 60% the fraction of women unnecessarily receiving adjuvant chemotherapy.

Van't Veer *et al.* were concerned that their optimal 70-gene prognostic profile did not include the known prognostic markers such as HER-2, ER, cyclin D1 or plasminogen activating inhibitor-1. However, other than ER, the exclusion of the other known markers in the expression prognostic profile should actually not be a surprise. For example, HER-2 overexpression is not common in node-negative tumors (10–15%), and the impact on survival in this nonmetastatic form of breast cancer requires studying large numbers of subjects [5]. Moreover, we have assessed the impact of *myc* amplification and implied overexpression in over 700 node-positive breast cancer patients undergoing adjuvant chemotherapy, and we found no prognostic impact or effect on the efficacy of chemotherapy (ET Liu, unpublished results). The novel gene list of Van't Veer *et al.* may thus provide important predictive information.

There is no doubt that much has been learned about the genetic framework of breast cancer biology with the two advanced studies of Sorlie *et al.* [1] and of van't Veer *et al.* [2]. The beauty of this genetic/genomic approach is that one can amass large amounts of data but can immediately telescope the attack to individual genes, an approach that compares very favorably with interrogating the dots and bands from two-dimensional protein gels or with differential display experiments requiring significant downstream work.

However, it is fair to ask what impact this information may have either on clinical practice or on our conceptualization of the clinical disease. Cynics would question the current significance. Indeed, in the study by Sorlie *et al.*, the segregation of good and bad prognostic groups in the ER-positive tumors by the luminal classification could have been accomplished using ER, the p53 sequence, and the HER-2 status. Moreover, the gene profiling of the poor outcome tumors in the van't Veer *et al.* study has not provided any particularly new insights into why the tumors

behave badly. That some proliferation and repair genes are upregulated is not novel. Expression arrays are cumbersome, expensive, and not specific, in that they also look at the expression of nontumor-contaminating cells.

These concerns are understandable, although they were the same types of criticisms directed towards bone marrow transplantation and combination chemotherapy when both those experimental concepts were originally launched. The results of Sorlie *et al.* [1] and of van't Veer *et al.* [2], however, if applied and tested in clinical trials, may have some interesting twists in terms of significance. For example, there is now evidence that letrozole is more effective than tamoxifen in some forms of ER-positive breast cancer defined by concurrent HER-2 or epidermal growth factor receptor positivity [6]. Could this benefit be better defined using markers that also determine the luminal subtype B and subtype C? It also appears that ER-positive tumors that are wild type at the *TP53* locus, expressing high levels of GATA binding protein-3, hepatocyte nuclear factor 3 alpha, trefoil factor 3, and estrogen-regulated LIV-1, possibly have the best prognosis and may be spared any form of chemotherapy. Clinical trialists should quickly test these novel findings in confirmatory trials.

Besides the advantages in gene discovery, these array experiments are clearly changing how we view and approach biological and clinical data. We predict that these changes will be profound. This massively parallel and highly comprehensive array technology has forced laboratory scientists to adopt a statistical and population approach to the data, and has forced clinical scientists to seek novel and more biological approaches to their clinical trials.

Reductionism has been the fundamental framework of our scientific enterprise in both basic and clinical sectors: study one gene/marker, hold all else constant, and assess an outcome/read out. The subtext in both the Sorlie *et al.* [1] and van't Veer *et al.* [2] studies is that not one marker or gene could correctly classify the tumors in terms of clinical outcome. Instead, the composite analysis of multiple genetic elements (from 70 to over 1000) may be necessary to provide the needed resolution for tailored and personalized medicine. But this conceptual shift will require us to overcome some old thinking and to address new realities.

The statistical analysis of the data output from such comprehensive genomic approaches will be daunting. Experimentally, it is not feasible to test the significance of each gene in an array experiment using standard logistic regression. Thus, a move from assessing each gene as an independent marker to developing robust approaches that take into account the biological relationships between markers might be necessary.

We should explore the development of a 'consensus' read-out of large marker profiles where the composite information of multiple markers can be defined as a single parameter. This has previously been attempted through the use of 'prognostic indices' that mathematically combine the predictive capabilities of many parameters into one measure. The use of such indices that depict data clusters with potential biological meaning may, however, be necessary to rationally interpret array-based experiments across clinical trials. For example, how does one plan and interpret a trial to validate van't Veer *et al.*'s observations where the readout is not just one marker but between 70 and 100 markers? This all suggests that new statistical approaches and new study designs are necessary to integrate genomic data in clinical trials.

Whole genomic expression profiling will need to be tested in other settings in breast cancer (e.g. metastatic disease, node-positive breast cancer, etc.) and in larger validation studies, especially if survival is to be assessed. But the integration of such quickly changing, highly complex technology with the more cumbersome phase III clinical trials presents some serious experimental challenges. For example, technology migration might cause a situation where, every year that a clinical trial is accruing patients and tissue samples, the molecular platform adopts a new, and significantly better, format. Future clinical trialists will face criticism either for analyzing their samples with pitifully outdated technology or for using normalization schemes that are artificial at best. We suspect that a novel study design taking into account progressive advances in technology will be needed.

The problem that tumor marker studies currently have in standardizing probes and scoring will worsen. Array preparation has not been standardized. In addition, most accessible expression array platforms still require reference RNA, and no common reference source has been agreed upon. Even the format of the data fields is not consistent from group to group. There will need to be some common standards in the field so that trials data can be interchangeable.

In truth, all these concerns are surmountable. But what we are witnessing in molecular biology is what astronomers have lived with for some time: that studying both the micro and the macro is needed to gain the conceptual big picture. One can learn much about the cosmos by studying the rock samples from the moon and the rings of Saturn, but some of the greatest wonders, such as the origin of the stars, have been uncovered by observing the galaxies in the universe. Happily, we are now at the beginnings of such galactic studies in biology.

## References

1. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proc Natl Acad Sci USA* 2001, **98**:10869-10874.
2. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415**:530-535.
3. Gruvberger S, Ringner M, Chen Y, Panavally S, Saal LH, Borg A, Ferno M, Peterson C, Meltzer PS: **Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns.** *Cancer Res* **61**:5979-5984.
4. Bergh J, Norberg T, Sjogren S, Lindgren A, Holmberg L: **Complete sequencing of the p53 gene provides prognostic information in breast cancer patients, particularly in relation to adjuvant systemic therapy and radiotherapy.** *Nat Med* 1995, **1**:1029-1034.
5. Paik S, Hazan R, Fisher ER, Sass RE, Fisher B, Redmond C, Schlessinger J, Lippman ME, King CR: **Pathological findings from the National Surgical Adjuvant Breast and Bowel Project: prognostic significance of erbB2 protein in primary breast cancer.** *J Clin Oncol* 1990, **8**:103-112.
6. Ellis MJ, Coop A, Singh B, Mauriac L, Llombert-Cussac A, Janicke F, Miller WR, Evans DB, Dugan M, Brady C, Quebe-Febeling E, Borgs MJ: **Letrozole is more effective neoadjuvant endocrine therapy than tamoxifen for erbB-1 and erbB-2 positive estrogen receptor positive primary breast cancer: evidence from a phase III randomized trial.** *J Clin Oncol* 2001, **19**:3808-3816.