Commentary

Microarrays and breast cancer clinical studies: forgetting what we have not yet learnt

Ahmed Ashour Ahmed and James D Brenton

Cancer Genomics Programme, Department of Oncology, University of Cambridge, Hutchison/MRC Research Centre, Cambridge, UK

Corresponding author: James D Brenton, jdb1003@cam.ac.uk

Published: 1 April 2005
Breast Cancer Research 2005, **7**:96-99 (DOI 10.1186/bcr1017)
This article is online at http://breast-cancer-research.com/content/7/3/96

See related commentary by Chang *et al*., page 100 [http://breast-cancer-research.com/content/7/3/100]

Abstract

This review takes a sceptical view of the impact of breast cancer studies that have used microarrays to identify predictors of clinical outcome. In addition to discussing general pitfalls of microarray experiments, we also critically review the key breast cancer studies to highlight methodological problems in cohort selection, statistical analysis, validation of results and reporting of raw data. We conclude that the optimum use of microarrays in clinical studies requires further optimisation and standardisation of methodology and reporting, together with improvements in clinical study design.

Introduction

By the time that a breast cancer is clinically apparent it has undergone multiple genetic and epigenetic primary carcinogenic events and further secondary molecular changes that ensure the adaptation of its cells to the changing microenvironment. The diversity of these genetic changes has made it difficult to classify breast cancer molecularly, and as a consequence there has been great enthusiasm for using genome-wide profiling methods to acquire a better understanding of the disease. This has led to an increasing number of studies using expression array profiling to improve the prediction of cancer prognosis [1–7]. Great things have been promised by exponents of these technologies [8]. How should we view the impact of current work?

Microarray technology

Irrespective of the questions being addressed in a profiling study, microarray techniques have inherent problems that lead to considerable data variability. Major sources of variability can arise from methods of RNA extraction [9,10], different types of probe preparation [9,11], probe labelling [12,13] and hybridisation [14,15]. It is also clear that varying the microarray platform, reference sample or segmentation method used for microarray image analysis leads to significant differences in data repeatability and gene discovery [16-18]. Although the MIAME (minimum information about a microarray experiment) report defines standards for information needed for reporting microarray experiments [19], it does not describe or quantify variabilities in the experiments. More studies addressing these experimental issues are urgently needed [20,21] along with efforts to define common standards for expression measurement controls. Guidelines are already emerging for best practice in using expression profiling for clinical trials [22].

The aim of supervised classification of microarray data is to detect genes that might prospectively predict defined outcomes. Existing studies in breast cancer have involved three steps: identifying a set of genes that are different between survival or drug response, refining this set for optimal classification within the sample set and finally validating the performance of the classifier genes on independent samples. Several studies have addressed these questions [1–7], but even before examining the technology a critical appraisal of the studies shows multiple methodological problems that make the interpretation of the results difficult.

Clinical study design

The problems can be summarised into four main categories: cohort selection, statistical analysis, validation of results and reporting of raw data. With the exception of the report by Chang and colleagues [5], studies were conducted as retrospective analyses of 'available' samples. Data collected retrospectively are inevitably incomplete, posing a complex problem in the interpretation of results [23,24]. Lack of detailed clinical information from paper records often means that important clinical predictors cannot be included in multivariate analysis to estimate the true predictive values of novel classifiers. This is exemplified by the studies from Ahr and colleagues [4] and van 't Veer and colleagues [2] that examined the association between a microarray classifier and prognosis without accounting for the effects of important clinical parameters such as performance status or treatment

modality. The use of 'available' samples may introduce significant heterogeneity into patient characteristics and unexpected temporal effects. van de Vijver and colleagues [3] used a 'validation set' (see below) containing patients treated with different modalities of surgery, chemotherapy and radiotherapy over 11 years. Each of these variables could introduce significant prognostic differences and make the estimation of the true independent effect of a molecular classifier difficult. A multi-variable analysis of data from van de Vijver and colleagues [3] clearly shows a highly significant decrease in hazard of recurrence in patients treated with chemotherapy in comparison with those who received no chemotherapy (hazard ratio of 0.37; $P < 0.001$). This confounding variable combined with the limited number of samples tested makes the microarray results difficult to interpret. Prospective studies that are much less sensitive to these sources of bias should be the priority for future research.

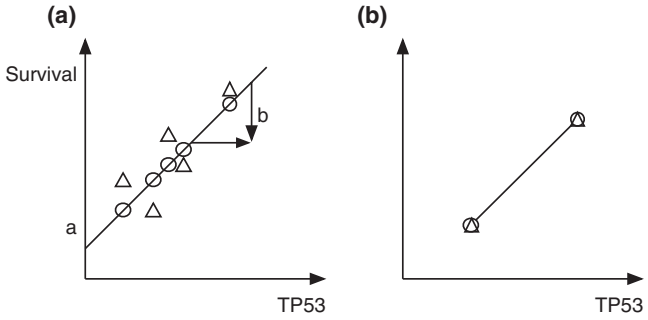
Defined criteria and endpoints

However, it is vital that both prospective and retrospective studies use clinically relevant criteria for categorising patients; these should be clearly defined and prospectively applied. Chang and colleagues [5] used median residual volume to measure tumour response to docetaxel in a prospective study of 24 patients with primary breast cancer, although pathological response is known to be the most important clinical outcome measure because it is strongly correlated with survival [25]. van de Vijver and colleagues [3] classified their breast cancers as positive or negative for oestrogen receptor on the basis of the expression array values and not a validated immunohistochemical test. This value was then used inconsistently as a categorical variable for examining association with the prognostic signature, and as a continuous variable in multivariate analysis to examine the independent effect of the signature on prognosis. Arbitrarily defined outcome measures that do not represent established clinical criteria are likely to increase subjective bias.

Statistical considerations

How can we decide whether a classifier might be a useful clinical test? The performance of any test is dependent upon the cut-off point used to discriminate between outcomes. van 't Veer and colleagues [2] and van de Vijver and colleagues [3] claim a correct classification rate of 83% for good prognosis. Similarly, Huang and colleagues [7] report a 90% accuracy for predicting outcome. However, these results were based on arbitrarily defined cut-off values. As these cut-off points were user defined they do not allow true estimation of the predictive power of the classifier and the use of differing values by van de Vijver and colleagues [3] is inappropriate and confusing. A more robust estimate is obtained by using sensitivity and specificity values obtained at multiple cut-off points to draw a receiver operating characteristics (ROC) curve. The area under the curve (AUC) is the best estimate of the performance of a classifier and this method was used by Chang and colleagues [5]: the reported area under the curve for their classifier was 0.96 (range 0 to 1).

Figure 1



A simple case of over-fitting. Consider that a researcher is studying the effect of TP53 expression level (x) on survival (y) of a group of breast cancer patients. **(a)** Simple regression: from knowing the expression level and survival (the variables) for each patient, the relationship between the two variables can be modelled with a simple univariable linear regression equation of the form $y = a + bx$, where a is the interception point with the y axis and b is the slope of the equation line. Applying this equation to a TP53 expression value will result in a new y value that corresponds to predicted survival. However, the equation seldom gives a perfect match between the real survival (triangles) and the predicted survival from the equation (circles) for any given x. In general, the closer the predicted values are to the real values, the better the equation (model) is in explaining the observations or the better the 'fit' of the model. The fit of the model is therefore used as a measure of its performance. **(b)** Over-fitting: an equation that is dependent on only two observations will always result in a line that passes between these two observations, giving an artificially perfect match between the predicted and the observed data. This represents meaningless good performance of a model or 'over-fitting'. This results from using too few observations (patients) per variable (gene) studied. To make a more complex 'multi-variable analysis' requires even more observations (patients) required to avoid over-fitting. In practice, a working ratio of 10 patients for every variable studied is recommended. However, in microarray studies few patients are evaluated for many thousands of genes.

Even with robust technology and rigorous analysis, the major challenge in the experimental design is the huge disproportion between the number of variables tested (gene expression values) and the number of samples. This inevitably leads to a high false-discovery rate and over-fitting of statistical models to the cohort under study (Fig. 1). It follows that appropriate validation of the classifier is an essential requirement in estimating the error of a classifier. Internal validation on the set from which a classifier was generated is usually performed. This is performed either by dividing the data into a test set (for obtaining a classifier) and a training set (for estimating the error) or by leaving one case out at a time, developing a model from the remaining cases (training set) and testing it on the omitted case (test set). In either method it is mandatory not to include all cases for developing a classifier before testing it on the training set because this results in overestimating the accuracy of a classifier. van 't Veer and colleagues [2] performed an internal validation on their data set with

(improperly) and without (properly) this distinct separation between training sets and test sets. The published sensitivity of their classifier of 73% was obtained when the internal validation was improperly done and only 59% when the validation was properly done (published as supplementary material) [26,27].

Neither of the two types of internal validation is a substitute for independent validation on different data sets. Only three reports attempted such validation in breast cancer studies [2,3,5]. van 't Veer and colleagues [2] and Chang and colleagues [5] performed only a limited validation on 15 and 6 patients, respectively. Although van de Vijver and colleagues [3] reported a validation of the classifier of van 't Veer and colleagues [2] on 151 patients with lymph-node-negative disease, 61 patients were in fact taken from the original study. It is therefore unclear how applicable these classifiers are to the wider population at risk.

Reproducible analysis

These criticisms underscore the importance of comprehensive reporting of the raw data so that results can be compared and possibly validated with different studies. Sorlie and colleagues [1] published both microarray image files as well as individual feature intensity values, allowing full reinterpretation of their data. This example has not been followed by subsequent researchers. For example, van 't Veer and colleagues [2] merely reported average outcome correlations for 232 genes of their classifier and not the original raw data. Sotiriou and colleagues [6] identified 56 overlapping genes between their set of 485 differentially expressed genes and those reported by van 't Veer and colleagues [2]. Because the raw data for all the genes in the latter study are not available, it is difficult to exclude a random effect as the cause of this overlap. In addition, most descriptions of analysis methods in published papers are inadequate (for example see [28]). Analysis tools such as the open-source statistical language R and its microarray-specific Bioconductor packages are essentially high-level programming environments that oblige the user to enter declarations and expressions to analyse data [29,30]. This type of interaction makes it relatively easy to output detailed transcripts that contain both commands and data, and therefore allow reproducible analyses [31]. Analysis methods based on using software with graphical user interfaces are harder to record, but as a minimum, significant intermediate calculations and data objects should be submitted as supplementary information so that cross-checking by the reader is possible. Finally, to make the best use of microarray data sets, individual patient data should be anonymously reported and electronically accessible. The use of controlled vocabulary and standardised indices is critical for the reuse of clinical information.

Conclusion

Microarray profiling has, unquestionably, been established as a powerful tool in unravelling mechanistic insights into tumour biology. We argue here that the optimum use of such a technique in clinical studies requires the further optimisation and standardisation of reporting procedures coupled with carefully planned prospective studies. It is important to underscore the difference between validating a classifier and justifying its use in clinical practice. The latter requires evidence of significant improvement of clinical outcome for patients when a classifier is used to guide management. This ultimately requires testing a classifier in a randomised prospective trial to prove that a 'classifier-informed' management yields a better clinical outcome than a 'classifier-blind' arm. However, we argue that the data produced so far may be too preliminary to launch large-scale expensive phase III studies. Many of the methodological problems in identifying prognostic factors are not new and have been successively ignored by the clinical community over the past 20 years. The great danger of using new technology with newer problems is that these older lessons are quickly forgotten.

Competing interests

The author(s) declare that they have no competing interests.

Acknowledgements

AAA is a Medical Research Council Fellow and a Sackler Fellow. JDB is a Cancer Research UK Senior Clinical Research Fellow.

References

1. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, *et 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 AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, *et al.*: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415:**530-536.
3. van de Vijver MJ, He YD, van 't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, *et al.*: **A gene-expression signature as a predictor of survival in breast cancer.** *N Engl J Med* 2002, **347:**1999-2009.
4. Ahr A, Karn T, Solbach C, Seiter T, Strebhardt K, Holtrich U, Kaufmann M: **Identification of high risk breast-cancer patients by gene expression profiling.** *Lancet* 2002, **359:**131-132.
5. Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Elledge R, Mohsin S, Osborne CK, Chamness GC, Allred DC, *et al.*: **Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer.** *Lancet* 2003, **362:**362-369.
6. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET: **Breast cancer classification and prognosis based on gene expression profiles from a population-based study.** *Proc Natl Acad Sci USA* 2003, **100:** 10393-10398.
7. Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, *et al.*: **Gene expression predictors of breast cancer outcomes.** *Lancet* 2003, **361:** 1590-1596.
8. Liu ET, Karuturi KR: **Microarrays and clinical investigations.** *N Engl J Med* 2004, **350:**1595-1597.
9. Wang E, Miller LD, Ohnmacht GA, Liu ET, Marincola FM: **High-fidelity mRNA amplification for gene profiling.** *Nat Biotechnol* 2000, **18:**457-459.
10. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JC, *et al.*: **Distinctive gene expression patterns in human mammary epithelial cells and breast cancers.** *Proc Natl Acad Sci USA* 1999, **96:** 9212-9217.

11. Hu L, Wang J, Baggerly K, Wang H, Fuller GN, Hamilton SR, Coombes KR, Zhang W: **Obtaining reliable information from minute amounts of RNA using cDNA microarrays.** *BMC Genomics* 2002, **3:**16.
12. 't Hoen PA, de Kort F, van Ommen GJ, den Dunnen JT: **Fluorescent labelling of cRNA for microarray applications.** *Nucleic Acids Res* 2003, **31:**e20.
13. Naderi A, Ahmed AA, Barbosa-Morais NL, Aparicio S, Brenton JD, Caldas C: **Expression microarray reproducibility is improved by optimising purification steps in RNA amplification and labelling.** *BMC Genomics* 2004, **5:**9.
14. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, *et al.*: **Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.** *Nat Biotechnol* 2001, **19:**342-347.
15. Taylor S, Smith S, Windle B, Guiseppi-Elie A: **Impact of surface chemistry and blocking strategies on DNA microarrays.** *Nucleic Acids Res* 2003, **31:**e87.
16. Lee JK, Bussey KJ, Gwadry FG, Reinhold W, Riddick G, Pelletier SL, Nishizuka S, Szakacs G, Annereau JP, Shankavaram U, *et al.*: **Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells.** *Genome Biol* 2003, **4:**R82.
17. Zorn KK, Jazaeri AA, Awtrey CS, Gardner GJ, Mok SC, Boyd J, Birrer MJ: **Choice of normal ovarian control influences determination of differentially expressed genes in ovarian cancer expression profiling studies.** *Clin Cancer Res* 2003, **9:**4811-4818.
18. Ahmed AA, Vias M, Iyer NG, Caldas C, Brenton JD: **Microarray segmentation methods significantly influence data precision.** *Nucleic Acids Res* 2004, **32:**e50.
19. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, *et al.*: **Minimum information about a microarray experiment (MIAME) – toward standards for microarray data.** *Nat Genet* 2001, **29:** 365-371.
20. Park PJ, Cao YA, Lee SY, Kim JW, Chang MS, Hart R, Choi S: **Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference.** *J Biotechnol* 2004, **112:**225-245.
21. Tan PK, Downey TJ, Spitznagel EL Jr, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC: **Evaluation of gene expression measurements from commercial microarray platforms.** *Nucleic Acids Res* 2003, **31:**5676-5684.
22. Tumor Analysis Best Practices Working Group: **Expression profiling – best practices for data generation and interpretation in clinical trials.** *Nat Rev Genet* 2004, **5:**229-237.
23. Altman DG: **Systematic reviews of evaluations of prognostic variables.** *BMJ* 2001, **323:**224-228.
24. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Jones DR, Heney D, Burchill SA: **Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future.** *Br J Cancer* 2003, **88:**1191-1198.
25. Chollet P, Amat S, Cure H, de Latour M, Le Bouedec G, Mouret-Reynier MA, Ferriere JP, Achard JL, Dauplat J, Penault-Llorca F: **Prognostic significance of a complete pathological response after induction chemotherapy in operable breast cancer.** *Br J Cancer* 2002, **86:**1041-1046.
26. Simon R: **Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data.** *Br J Cancer* 2003, **89:**1599-1604.
27. Simon R: **When is a genomic classifier ready for prime time?** *Nat Clin Pract Oncol* 2004, **1:**4-5.
28. Tibshirani RJ, Effron B: **Pre-validation and inference in microarrays.** *Statist Applic Genet Mol Biol* 2002, **1**(1):article 1.
29. Ihaka R, Gentleman R: **R: a language for data analysis and graphics.** *J Comput Graph Stat* 1996, **5:**299-314.
30. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, *et al.*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5:**R80.
31. Ruschhaupt M, Huber W, Poustka A, Mansmann U: **A compendium to ensure computational reproducibility in high-dimensional classification tasks.** *Statist Applic Genet Mol Biol* 2004, **3**:article 37.