EDITOR'S CHOICE

Bioinformatics and breast cancer: what can high-throughput genomic approaches actually tell us?

A H Sims

Correspondence to:
A H Sims, Applied Bioinformatics of Cancer Group, Breakthrough Research Unit, Edinburgh Cancer Research Centre, Western General Hospital, Crewe Road South, Edinburgh EH4 2XR, UK;
andrew.sims@ed.ac.uk

Accepted 18 January 2009
Published Online First 27 January 2009

ABSTRACT

High-throughput genomic technology has rapidly become a major tool for the study of breast cancer. Gene expression profiling has been applied to many areas of research from basic science to translational studies, with the potential to identify new targets for treatment, mechanisms of resistance and to improve on current tools for the analysis of prognosis. However, the sheer scale of the data generated along with the number of different protocols, platforms and analysis methods can make these studies difficult for clinicians to comprehend. Similarly, computational scientists and statisticians that may be called upon to analyse the data generated are often unaware of the processes involved in sample collection or the relevance and impact of genetics and pathological characteristics. There is a pressing need for better understanding of the challenges and limitations of microarray approaches, both in experimental design and data analysis. Holistic, whole-genome approaches are still relatively new and critics have been quick to highlight non-overlapping results from groups testing similar hypotheses. However, it is often subtle differences in the experimental design and technology that underpin the variation between these studies. Rather than indicating that the data are meaningless, this suggests that many findings are real, but highly context dependent. This review explores both the current state and potential of bioinformatics to bring meaning to high-throughput genomic approaches in the understanding of breast cancer.

The breast cancer field has been quick to embrace the potential of high-throughput genomic approaches.[1] The attraction of these methods is readily apparent given the opportunity to simultaneously measure variation in thousands of DNA sequences, mRNA transcripts, peptides or metabolites (fig 1) to give us a holistic view of the machinations of cellular processes. Breast cancer is an extremely complex disease, with many risk factors ranging from unavoidable genetic predisposition through to lifestyle choices such as diet and exercise.[2] In addition, the breast is a difficult tissue to study, as it is composed of several cell types and undergoes structural changes during the menstrual cycle, pregnancy and ageing.[2] From commonly used clinicopathologiocal characteristics (such as tumour size, lymph node involvement, subtype, grade and oestrogen receptor (ER) expression) it is clear that breast cancer is a highly heterogeneous disease. Molecular profiling has confirmed this and highlighted the underlying complexity[3–5] and it is therefore hardly surprising that different tumours respond to different treatments.

This review aims to highlight the importance of reasoned experimental design and sound statistical analysis, illustrating the many possible confounding factors and limitations that need be taken into account when considering the value of high-throughput studies. It will focus primarily on gene expression profiling, but many of the issues raised are also applicable to other "-omic" technologies, such as array CGH, miRNA and proteomics array based methods (fig 1).

EXPERIMENTAL DESIGN

The shift in approach from measuring the level of a single transcript or protein in a cohort of patients to simultaneously measuring the levels of thousands of genes or proteins brings with it a need to better understand the concepts of multiple testing and false discovery rates.[6] With conventional approaches, the level of a single gene or protein is measured to prove or disprove a hypothesis. One of the attractions of high-throughput methods is that they are data driven rather than hypothesis driven, so are not limited by prior knowledge. While these methods can be effectively used to prove or disprove a given hypothesis, there real value is in generating new ones. A consequence of having many more features (genes, transcripts, single nucleotide polymorphisms, peptides, etc) than the number of samples is that many of the apparently differentially expressed features may be due to chance, rather than real biological differences. Considering the heterogeneous nature and variability of samples it should not be unexpected that subgroups of data do not separate into well-defined clusters in low dimensional visualisations (fig 2). For a more detailed review of issues of dataset dimensionality and multiple testing the reader is directed to the review of Clarke *et al*.[6] Variation between gene signatures of the most changing genes can arise from differences in cohort selection (biological variables) and experimental bias (technical variables).

BIOLOGICAL VARIABLES

When the objective of a microarray experiment is to identify genes that are differentially expressed between groups of "experiment" and "control" samples it is essential that the phenotype under investigation (treatment, overexpressed gene, tumour or patient characteristic) represents the largest source of variation. If this is not the case, then the results will be compromised due to confounding factors (whether known or unknown). One way to minimise this potential problem is to make sure the two groups of samples being compared are as similar to each other as possible in every respect except the phenotype under test. An alternative approach is simply to use

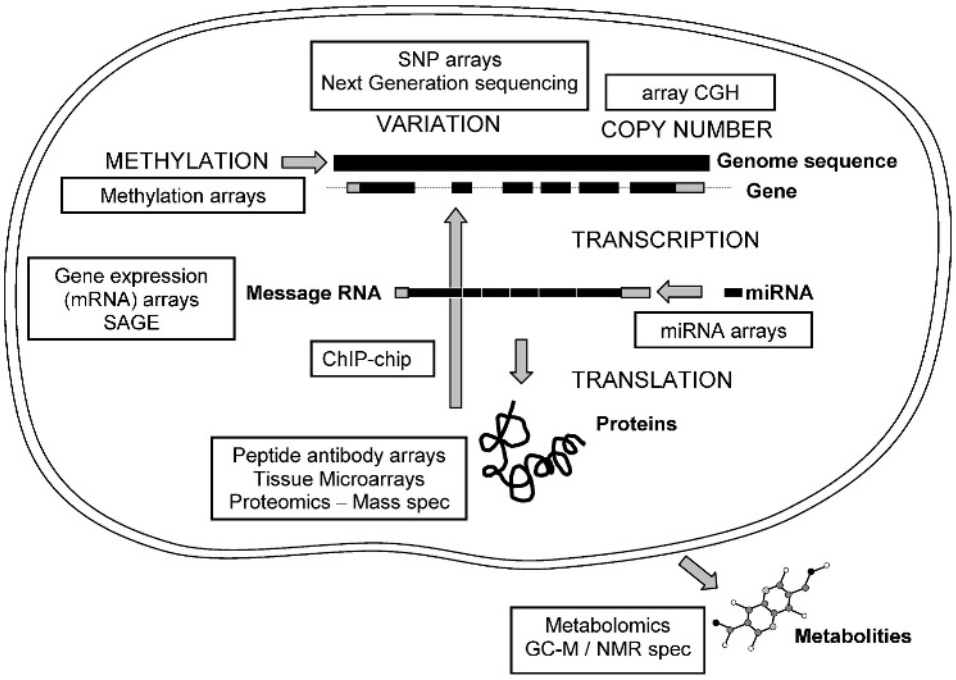

Figure 1 Summary of different types of high-throughput microarray and what they measure. There are now many different types of microarray that enable the measurement of many molecular variables in a holistic, systematic fashion. CGH, comparative genomic hybridisation; ChIP-chip, chromatin immunoprecipitation microarrays; GC-M, gas chromatography mass; NMR, nuclear magnetic resonance; SAGE, serial analysis of gene expression; SNP, single nucleotide polymorphism; spec, spectrometry.

huge numbers of samples, increasing the likelihood that the experimental variable is the only consistent difference between the two groups. Due to the high cost of these approaches and the scarcity of samples, the former is often the only feasible approach.

With cell line experiments it is relatively easy to minimise sources of variation by preparing samples in exactly the same way and following strict protocols to ensure that replicates are highly similar to each other. In this case, relatively few replicates are required to distinguish between the "experimental" and "control" samples (fig 2A). However with primary patient samples, molecular heterogeneity is a much bigger issue. Gene expression has been shown to be affected by so many variables that either highly specific entry criteria or large cohorts are required to distinguish between the "experimental" and "control" samples (fig 2B). An alternative approach, suited to looking for consistent changes in different individuals is to utilise "matched samples" of tumour[7] or normal tissue[8] taken from the same individual before and after an intervention (fig 2C). These paired studies have increased statistical power and the potential to predict which individuals will respond to the intervention. Another important consideration is the tissue composition of tumour material that is used as the starting point for extracting DNA, RNA, etc, for molecular analysis, with many studies now employing laser capture microdissection (fig 3). The need for such precision is highlighted by the reanalysis of the "normal-like" subtype described by Perou's and Sorlie's groups, in which histopathological examination of tumor samples categorised as "normal-like" revealed normal tissue contamination.[3–5 9 10]

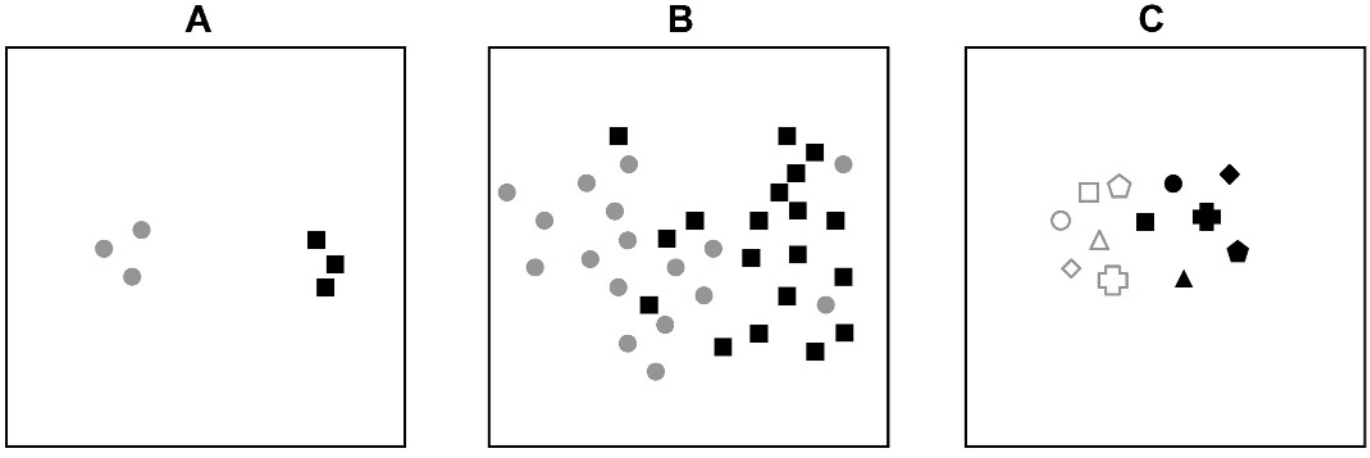

Figure 2 The level of variation across replicates or samples determines the numbers required to identify significantly differentially expressed genes that distinguish subgroups. The example plots represent hypothetical overall transcriptome similarity of samples by two-dimensional principle components analysis or multidimensional scaling. (A) The grey circle and black square replicates tightly cluster together, but are clearly distinct from each other. (B) The grey circles are less clearly separated from the black squares so greater numbers are required to identify consistent differences with the same level of confidence as in (A). (C) Matched samples, eg, before (filled symbols) and after (open symbols) measurements, can more easily identify common changes in expression relating to a particular treatment or procedure.

TECHNICAL VARIABLES

Although the underlying principles of annealing and hybridisation of complementary sequences are the same for all gene expression approaches, there are some fundamental differences in the design and production of the microarray platforms. Early microarrays tended to be produced in individual laboratories from PCR products from cloned cDNA or synthetic oligonucleotides printed onto glass slides.[11 12] The technology has rapidly evolved and expanded to profile many other variables including genomic DNA mutations and copy number, methylation and microRNAs, protein antibody or tissue and cell lysates (fig 1). Availability of commercial microarrays has been facilitated by several companies including Affymetrix (Santa Clara, California, USA), Agilent Technologies (Santa Clara, California, USA) and Illumina (San Diego, California, USA) among others, improving comparison and consistency of results to some degree within studies using each particular platform.

The majority of peer-review journals have made it a prerequisite for publication that gene expression datasets are made publicly available, and this is facilitated by data repositories such as ArrayExpress[13] and NCBI Gene Expression Omnibus.[14] In addition to the raw data, authors have to supply details of the samples, platform and protocols used according to MIAME (Minimum Information About a Microarray Experiment) guidelines.[15] The requirement to make data available has improved the transparency of microarray studies and enabled meta-analysis. However, the many steps involved in the workflow of a typical microarray experiment (fig 3) often vary between studies and this can introduce bias. Although the lack of overlap in lists of significant genes from apparently similar studies has been well documented,[16] these discrepancies can usually be attributed to differences in the underlying technology such as probe sequence design or differences in the way the experiments were conducted. Nevertheless, Sorlie and coworkers[9] demonstrated that breast cancer subtypes are distinguishable at the *unsupervised* level (see fig 4) across three different microarray platforms. Where there is variation between the most differentially expressed genes identified by each array platform, there is normally a highly significant overlap at the pathway level. It is important to remember that all microarray results are highly dependent upon the information used to design them in the first place. A re-mapping exercise of microarray probesets with the latest genome annotation revealed a 30–50% discrepancy in the genes previously identified as differentially expressed, regardless of the analysis method employed.[17]

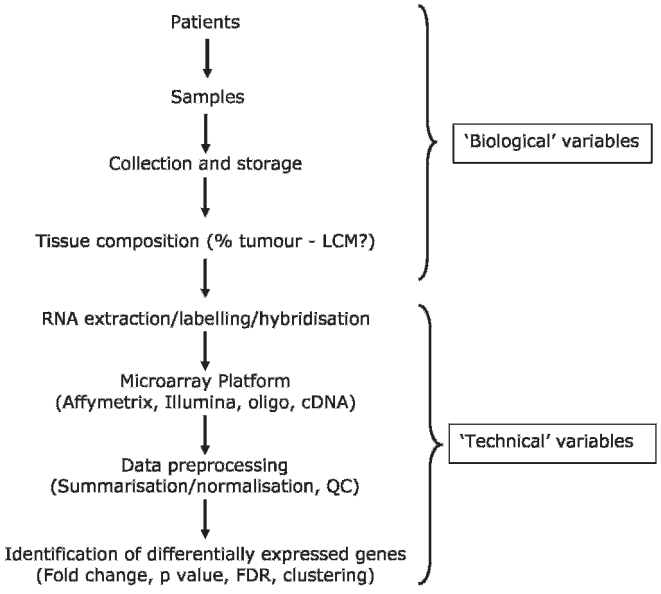

Figure 3 Overview of key steps in a microarray experiment. Biological and technical variables are introduced at many stages, and these will all have an impact on the final results. It is important that all these steps are clearly documented. FDR, false discovery rate; LCM, laser capture microdissection; QC, quality control.

SUBTYPING, CLASSIFICATION AND PROGNOSIS

Breast tumours can be segregated by many methods of histopathology and molecular pathology in order to predict prognosis or responsiveness to various therapies.[18] There have been three broad approaches to analysing gene expression microarrays in the breast cancer field (fig 4). The first of these is an *unsupervised* method of analysis, in which tumours are clustered into sub-groups by an ''intrinsic'' gene set that reflects differences in gene expression between tumours rather than within tumours,[3–5 9] without using selection criteria. The most striking molecular differences between luminal and basal-like subtypes have repeatedly been identified and validated with different technologies and platforms.[5 19–22] Identification of ''molecular apocrine'' tumours[21] and further subdivision of the ER-negative tumours into at least five different subtypes[23] has also been performed. The molecular subtypes identified are associated with significantly different clinical outcomes,[4 10] which are likely to best respond to different treatment approaches. A phase II trial of anti-androgen therapy in ER/progesterone receptor (PR)/Her2-negative, androgen-positive tumours derived from this type of study is now underway.

The second two methods utilise *supervised* approaches based upon individual clinical follow-up data or characteristics of tumour biology such as ER status, grade or proliferation[17] (see fig 4). The lack of overlap (three genes) between the 70-gene signature of the Amsterdam group[24 25] (cDNA arrays) and the 76-gene signature of the Rotterdam group[26 27] (Affymetrix oligonucleotide arrays) has been claimed as evidence that genomic approaches based upon follow-up data are unreliable. But logically, the heterogeneity demonstrated by unsupervised approaches would preclude replicate findings from two modest-sized studies of different groups of samples. The disparity between the signatures can potentially be accounted for when examining the variations in the inclusion criteria (age, lymph node status, diameter of tumour, adjuvant treatment, etc), the platform (cDNA or oligonucleotide arrays, or quantitative reverse transcription PCR (qRT-PCR)) and different data analysis methods used in each study. Despite the clear differences in approach and a lack of consensus in the gene signatures generated, all three of the broad approaches outlined above (fig 4) have a similar capacity to predict prognosis. Evaluation of several signatures with a single test dataset demonstrated a high degree of overlap in the outcome predicted for individual patients.[20]

REPRODUCIBILITY, VALIDATION AND DATASET-SPECIFIC BIAS

The genes that make up a gene expression signature are by their nature dependent upon: patient and tumour characteristics, array platform, normalisation method, and statistical thresholds for gene selection or the classification algorithm employed (fig 3). Using a particular dataset to generate a predictive profile has its own inherent bias based upon its attributes. Ein-Dor *et al* demonstrated that many different but equally predictive lists of

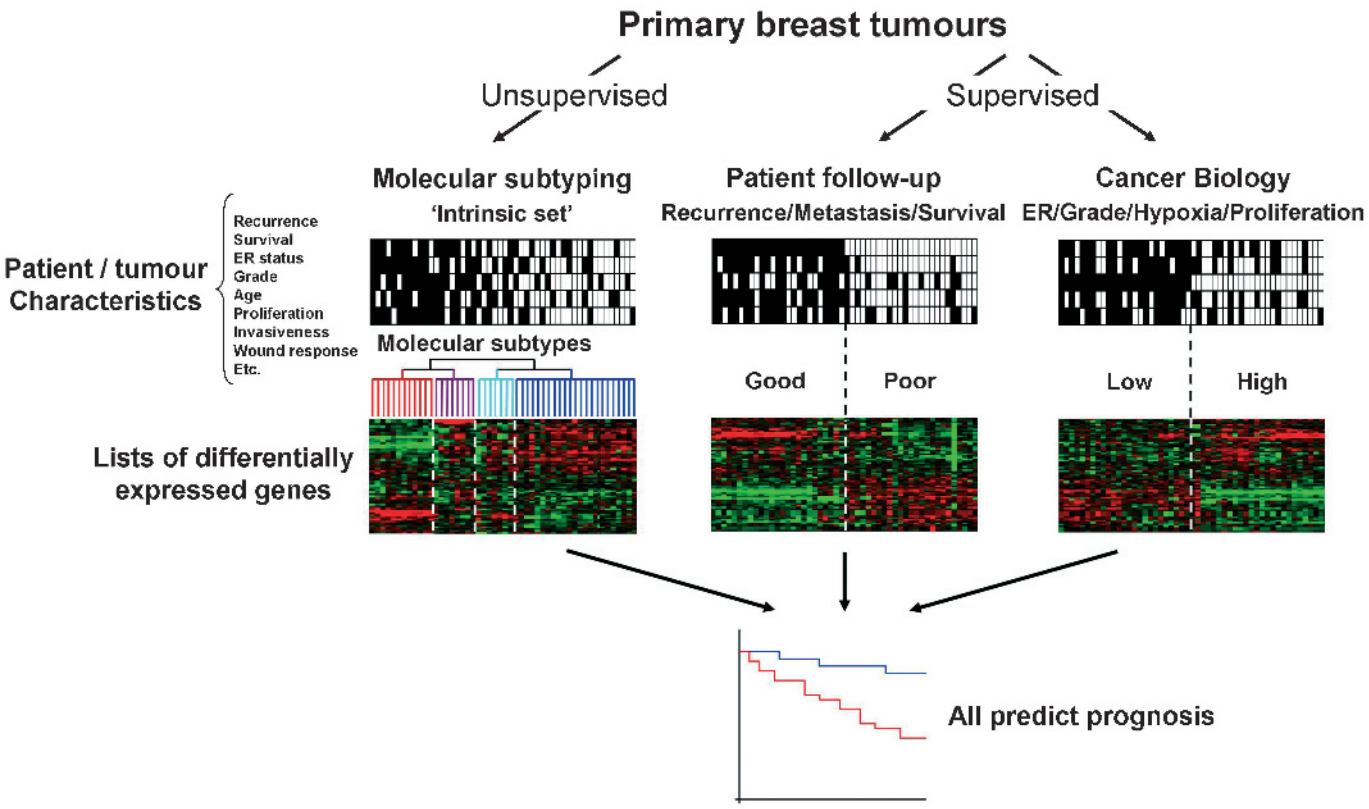

Figure 4 Concordance between different approaches to prognosis prediction. Regardless of the strategy used to identify lists of significantly differentially expressed each method can be used to predict prognosis. These profiles are inter-related and may be more accurate than existing single markers. ER, oestrogen receptor.

70 genes can be produced simply by changing the members of "training" and "test" sets.[28] It seems inevitable that gene signatures will perform less well with validation datasets than the ones used to generate the profile. A follow-up study to the 76-gene Rotterdam signature identified strong time dependence of the signature when validated with a cohort with longer median follow-up time (14 years) compared with the original study (8 years).[26] Many patient characteristics are known to affect gene expression (and other tumour features), including age[29] and race.[30] Anders *et al* demonstrated that breast cancer arising in younger women was more likely to involve PI3K, Myc and β-catenin, whereas the activation of Src and E2F deregulation was more associated with tumours in older women[31] Only genes that are clearly mechanistically different between particular groups of patients will be identified reproducibly between similarly defined cohorts.

Using a series of repeated validation datasets comparing breast cancer and normal breast cell lines (MCF7 and MCF10A), we recently examined the variability between datasets generated using different amounts of starting RNA, alternative protocols, different generations of Affymetrix GeneChip or scanning hardware. We demonstrated that systematic, multiplicative biases are introduced at the RNA, hybridisation and image-capture stages of a microarray experiment.[32]

DATA INTEGRATION AND META-ANALYSIS

Validation of new results with independent data is essential to establish that research findings are indeed "real". For example, meta-analyses of multiple experiments using different platforms has resulted in new predictive signatures that perform as well or better than the platform specific signature.[33 34] These approaches remove the inherent bias of a single microarray platform and are able to concentrate on genes that are consistently differentially expressed, regardless of the technology used. However, cross platform meta-analyses may be somewhat limited by the number of common genes represented. Cross platform normalisation[35] and distance weighted discrimination[36] methods have been put forward for comparing data from different types of microarrays.

One way to overcome the heterogeneity described above is to increase the size of studies by combining datasets; however this can make the problem of analysing the data even more daunting. The many breast cancer gene expression datasets now in the public domain represent a valuable resource for meta-analysis. However, dataset-specific bias precludes integration of published studies at the raw intensity level without some form of correction method (fig 5). In our study, simple batch mean-centring was sufficient to reconcile validation cell line and published breast tumour datasets, outperforming distance-weighted discrimination[36] and generating similar results to ComBat, an empirical Bayes method to adjust for batch effects.[37] Several meta-analysis studies have now been published, generally validating previous findings, emphasising "real" effects, consensus and improving clarity.[34 38–41]

We recently demonstrated that integrating up to six published breast cancer Affymetrix GeneChip datasets can increase the accuracy of prognosis prediction and that this can be improved further by removing systematic, multiplicative bias.[32] The most accurate prognosis predictions are generated when the test sets closely share the patient and tumour characteristics of the training sets. An alternative approach to building ever larger combined datasets representing the whole

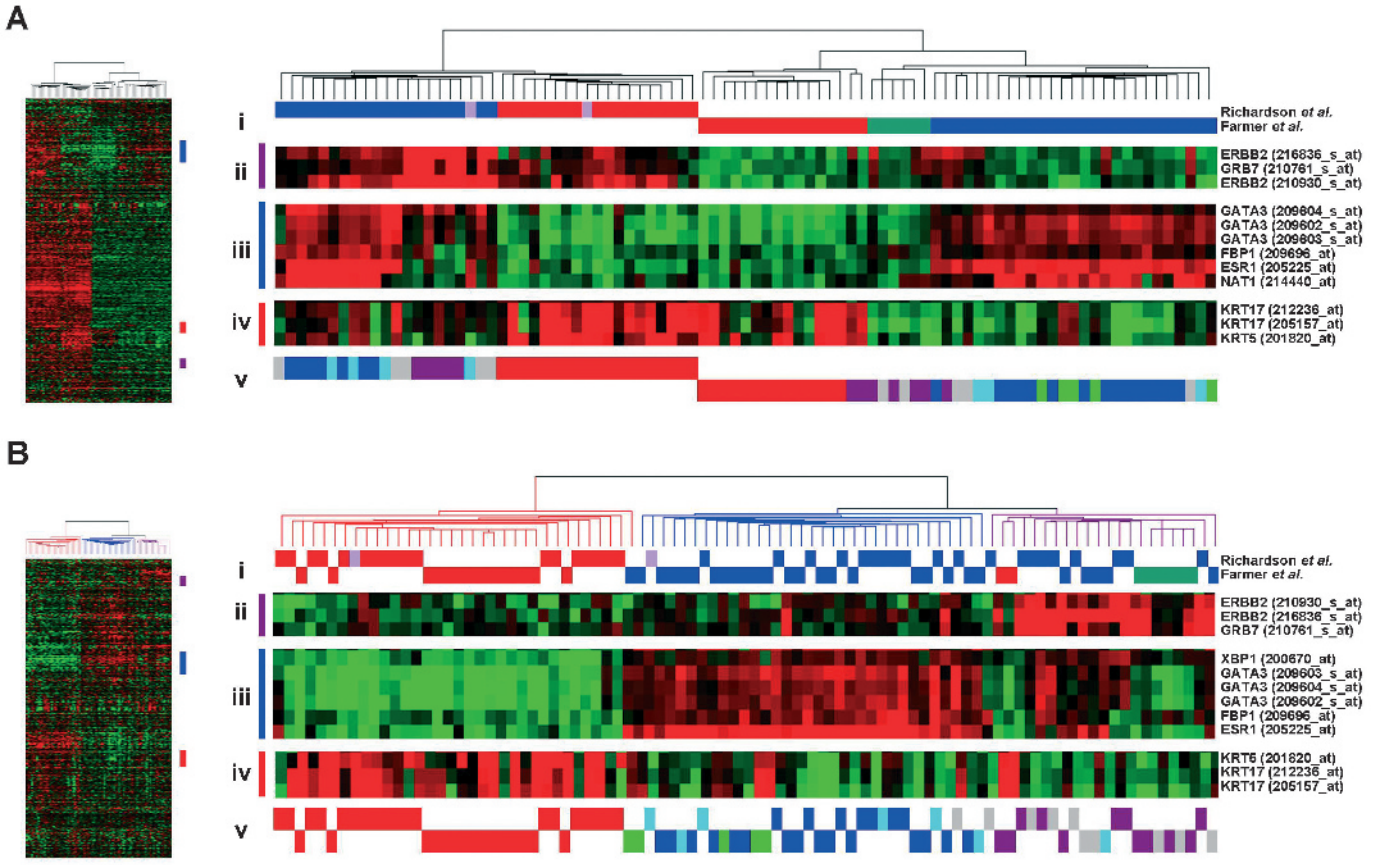

Figure 5 Dataset-specific bias must be removed for integration of gene expression data.[32] Combining breast tumour gene expression profiles generated by two published studies. (A) Before mean batch-centering. (B) After mean batch-centering. Hierarchical clustering of tumours based upon 640 probesets representing Sorlie *et al*[5] "intrinsic" genes. Thumbnails show all 640 probesets. (i) Tumours classified by Richardson *et al*[22]: red, basal-like; blue, non-basal like, pink, BRCA1; tumours classified by Farmer *et al*[21]: red, basal; blue, luminal; green, apocrine. Clusters of genes associated with the "Sorlie subtypes" are highlighted as follows: (ii) ERBB2 gene cluster, (iii) luminal A gene cluster, (iv) basal gene cluster. (v) Centroid prediction was used to assign the tumours to the five Norway/Stanford subtypes: basal (red), luminal A (dark blue), luminal B (light blue), ERBB2 (purple), normal-like (green), unassigned (grey).

breast cancer population would be to concentrate on generating gene expression classifiers for separate clearly defined groups of patients based on commonly used clinicopathological parameters. However, strict entry criteria would severely restrict the number of suitable patients/tumours eligible for inclusion and this approach could take no account of possible unknown confounding factors. In clinical practice, single sample predictors[10] are required, applicable to large groups of patients and our work strongly suggests that these will be best generated from the largest possible cohorts (or integrated datasets). It is essential that researchers are aware that differences in dataset composition can also have dramatic effects on meta-analysis and it may not always be appropriate to combine datasets if they have been subject to different entry criteria or treatments.[32]

WILL HIGH-THROUGHPUT APPROACHES MAKE IT TO THE CLINIC?

Clinicians have to chose the most appropriate treatment for individuals; however many of the disease parameters currently used are qualitative rather than quantitative. Prognostic models such as The International Consensus Guidelines of St Gallen[42] and the Nottingham Prognostic Index[43] are used to guide treatment decisions. While these models may be able to predict proportions of the population in which an outcome may occur with reasonable accuracy, they cannot identify in which women the outcome will occur; the inevitable consequence of this is either overtreatment or inadequate treatment. Following the National Comprehensive Cancer Network guidelines can result in unnecessary chemotherapy for up to 80% of some of the better prognosis subgroups. For molecular signatures to have any true value in treatment selection they must be reliably validated to outperform or add value to existing clinical guidelines.[1] Traditional classifications of tumours may provide clear-cut treatment options in high-risk and low-risk cases, but often tumours fall into an "intermediate" group; it is in these borderline cases where improvements are most urgently required. In these cases the "safe" option is to overtreat, benefiting a relatively small minority of cases and exposing the rest to side effects unnecessarily. Conversely, a more conservative approach may avoid unwarranted treatment and additionally reduce costs, but some women that would benefit may go untreated. Studies that examine links between gene expression and known prognostic factors such as grade[44] and ER status[45] may be beneficial for this intermediate group.

Two clinical tests based upon gene expression profiling studies are already commercially available and being evaluated in large multicentre, multinational trials. The TAILORx study, sponsored by the National Cancer Institute, will test OncotypeDX,[46] a 21-gene qRT-PCR recurrence score algorithm

(derived from gene expression array studies) that can be performed on formalin-fixed, paraffin-embedded tissue. The study will enrol more than 10 000 women with hormone-positive (ER positive and/or PR positive), ERBB2-ngative and node-negative breast cancer to determine which women should receive adjuvant chemotherapy in addition to hormone therapy. In a study of archival material from 4964 lymph-node-negative breast tumours that were not treated with chemotherapy, the Recurrence Score was strongly associated with risk of breast cancer death among ER-positive, tamoxifen-treated and untreated patients.[47] In the B-20 study, recurrence score not only quantified the likelihood of breast cancer recurrence in women with node-negative, ER-positive breast cancer, but also predicted the magnitude of chemotherapy benefit.[48] However, in a study of 149 patients who were not treated with adjuvant therapy, the 21 gene-based recurrence score was not predictive of distant disease recurrence, highlighting the importance of cohort selection.[49] OncotypeDX has been added to the list of approved American Society of Clinical Oncology markers[50] and it is anticipated that 60 000 OncotypeDX tests will be performed in 2008. With tests costing thousands of dollars this could have implications for health service providers, although this would be set against reducing the cost of unnecessary treatment.[1] The US Food and Drug Administration has approved the Mammaprint clinical test that was developed by Agendia (Huntington Beach, California, USA) from the 70-gene signature.[25] While the assay has been validated by this group,[24 51] concerns regarding the design and statistical analysis used to derive the original 70-gene signature have been raised[28 52] These issues have largely been incorporated into the prospective MINDACT (Microarray in Node-Negative Disease May Avoid Chemotherapy) clinical trial of 6000 patients.[53 54] The TRANSBIG consortium also used the same 70-gene validation samples to evaluate two other gene expression signatures with potential prognostic value that were developed, using the Affymetrix microarray platform: the 76-gene Veridex/Rotterdam signature[27 55] and the Genomic Grading Index.[44] This retrospective validation was recently published,[26] concluding that the three signatures performed in a similar way, all being superior to the classical clinicopathological methods.

One consequence of moving towards ''individualised treatment'' is that it can be difficult to identify appropriate numbers of patients with similar characteristics that have been exposed to the same treatment regimen to adequately statistically power a study. While high-throughput expression profiling methods are not yet fully evaluated, they clearly have great potential that needs to be carefully validated before they become standard prognostic tools. In the meantime, they are generating a large amount of valuable data that are gradually improving our understanding of the molecular changes that are associated with breast cancer development, progression and treatment.

CONCLUSIONS

Issues of cohort selection and choice of appropriate analysis methods are central to breast cancer studies using high-throughput genomic approaches. Ultimately, bioinformatics seeks to bring meaning to biological data so that it can be comprehended in the context of current knowledge, allowing new hypotheses to be generated and tested.

Acknowledgements: AHS is very grateful for funding from Breakthrough Breast Cancer.

Competing interests: None.

Take-home messages

- High-throughput genomic approaches have the potential to significantly improve our understanding of breast cancer.
- Breast cancer is a highly heterogeneous disease and many biological variables can affect the data generated, so careful experimental design is required for meaningful results.
- There are many different protocols and analysis methods; understanding which is most appropriate is a considerable challenge.
- Results need to be demonstrated to be statistically robust, combining datasets and performing meta-analyses can help us to identify consensus findings.

REFERENCES

1. **Sims AH,** Ong KR, Clarke RB, *et al*. High-throughput genomic technology in research and clinical management of breast cancer. Exploiting the potential of gene expression profiling: is it ready for the clinic? *Breast Cancer Res* 2006;**8**:214.
2. **Howell A,** Sims AH, Ong KR, *et al*. Mechanisms of Disease: prediction and prevention of breast cancer—cellular and molecular interactions. *Nat Clin Pract Oncol* 2005;**2**:635–46.
3. **Perou CM,** Sorlie T, Eisen MB, *et al*. Molecular portraits of human breast tumours. *Nature* 2000;**406**:747–52.
4. **Sorlie T,** Perou CM, Tibshirani R, *et al*. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* 2001;**98**:10869–74.
5. **Sorlie T,** Tibshirani R, Parker J, *et al*. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* 2003;**100**:8418–23.
6. **Clarke R,** Ressom HW, Wang A, *et al*. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer* 2008;**8**:37–49.
7. **Miller WR,** Larionov A, Anderson TJ, *et al*. Predicting response and resistance to endocrine therapy: profiling patients on aromatase inhibitors. *Cancer* 2008;**112**:689–94.
8. **Kendall A,** Anderson H, Dunbier AK, *et al*. Impact of estrogen deprivation on gene expression profiles of normal postmenopausal breast tissue in vivo. *Cancer Epidemiol Biomarkers Prev* 2008;**17**:855–63.
9. **Sorlie T,** Wang Y, Xiao C, *et al*. Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: Gene expression analyses across three different platforms. *BMC Genomics* 2006;**7**:127.
10. **Hu Z,** Fan C, Oh DS, *et al*. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* 2006;**7**:96.
11. **Lockhart DJ,** Dong H, Byrne MC, *et al*. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996;**14**:1675–80.
12. **DeRisi JL,** Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997;**278**:680–6.
13. **Brazma A,** Kapushesky M, Parkinson H, *et al*. Data storage and analysis in ArrayExpress. *Methods Enzymol* 2006;**411**:370–86.
14. **Barrett T,** Suzek TO, Troup DB, *et al*. NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res* 2005;**33**:D562–6.
15. **MGED Society**. http://www.mged.org (accessed 7 July 2009).
16. **Tan PK,** Downey TJ, Spitznagel EL Jr, *et al*. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* 2003;**31**:5676–84.
17. **Dai M,** Wang P, Boyd AD, *et al*. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 2005;**33**:e175.
18. **Sims AH,** Howell A, Howell SJ, *et al*. Origins of breast cancer subtypes and therapeutic implications. *Nat Clin Pract Oncol* 2007;**4**:516–25.
19. **Sorlie T,** Perou CM, Fan C, *et al*. Gene expression profiles do not consistently predict the clinical treatment response in locally advanced breast cancer. *Mol Cancer Ther* 2006;**5**:2914–8.
20. **Fan C,** Oh DS, Wessels L, *et al*. Concordance among gene-expression-based predictors for breast cancer. *N Engl J Med* 2006;**355**:560–9.
21. **Farmer P,** Bonnefoi H, Becette V, *et al*. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene* 2005;**24**:4660–71.
22. **Richardson AL,** Wang ZC, De Nicolo A, *et al*. X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell* 2006;**9**:121–32.
23. **Teschendorff AE,** Miremadi A, Pinder SE, *et al*. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biol* 2007;**8**:R157.
24. **van de Vijver MJ,** He YD, van't Veer LJ, *et al*. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 1999–2009, 2002:347.
25. **van't Veer LJ,** Dai H, van de Vijver MJ, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;**415**:530–6.

26. **Desmedt C,** Piette F, Loi S, *et al*. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res* 2007;**13**:3207–14.
27. **Wang Y,** Klijn JG, Zhang Y, *et al*. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005;**365**:671–9.
28. **Ein-Dor L,** Kela I, Getz G, *et al*. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 2005;**21**:171–8.
29. **Anders CK,** Hsu DS, Broadwater G, *et al*. Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression. *J Clin Oncol* 2008;**26**:3324–30.
30. **Carey LA,** Perou CM, Livasy CA, *et al*. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA* 2006;**295**:2492–502.
31. **Anders CK,** Acharya CR, Hsu DS, *et al*. Age-specific differences in oncogenic pathway deregulation seen in human breast tumors. *PLoS ONE* 2008;**3**:e1373.
32. **Sims AH,** Smethurst GJ, Hey Y, *et al*. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets – improving meta-analysis and prediction of prognosis. *BMC Med Genomics* 2008;**1**:42.
33. **Warnat P,** Eils R, Brors B. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics* 2005;**6**:265.
34. **Shen R,** Ghosh D, Chinnaiyan AM. Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics* 2004;**5**:94.
35. **Shabalin AA,** Tjelmeland H, Fan C, *et al*. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics* 2008;**24**:1154–60.
36. **Benito M,** Parker J, Du Q, *et al*. Adjustment of systematic microarray data biases. *Bioinformatics* 2004;**20**:105–14.
37. **Johnson WE,** Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;**8**:118–27.
38. **Wirapati P,** Sotiriou C, Kunkel S, *et al*. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res* 2008;**10**:R65.
39. **Acharya CR,** Hsu DS, Anders CK, *et al*. Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer. *JAMA* 2008;**299**:1574–87.
40. **Ben-Porath I,** Thomson MW, Carey VJ, *et al*. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat Genet* 2008;**40**:499–507.
41. **Teschendorff AE,** Naderi A, Barbosa-Morais NL, *et al*. A consensus prognostic gene expression classifier for ER positive breast cancer. *Genome Biol* 2006;**7**:R101.
42. **Goldhirsch A,** Glick JH, Gelber RD, *et al*. Meeting highlights: international expert consensus on the primary therapy of early breast cancer 2005. *Ann Oncol* 2005;**16**:1569–83.
43. **Haybittle JL,** Blamey RW, Elston CW, *et al*. A prognostic index in primary breast cancer. *Br J Cancer* 1982;**45**:361–6.
44. **Sotiriou C,** Wirapati P, Loi S, *et al*. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 2006;**98**:262–72.
45. **Oh DS,** Troester MA, Usary J, *et al*. Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. *J Clin Oncol* 2006;**24**:1656–64.
46. **Paik S,** Shak S, Tang G, *et al*. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004;**351**:2817–26.
47. **Habel LA,** Shak S, Jacobs MK, *et al*. A population-based study of tumor gene expression and risk of breast cancer death among lymph node-negative patients. *Breast Cancer Res* 2006;**8**:R25.
48. **Paik S,** Tang G, Shak S, *et al*. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J Clin Oncol* 2006;**24**:3726–34.
49. **Esteva FJ,** Sahin AA, Cristofanilli M, *et al*. Prognostic role of a multigene reverse transcriptase-PCR assay in patients with node-negative breast cancer not receiving adjuvant systemic therapy. *Clin Cancer Res* 2005;**11**:3315–9.
50. **Harris L,** Fritsche H, Mennel R, *et al*. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *J Clin Oncol* 2007;**25**:5287–312.
51. **Buyse M,** Loi S, van't Veer L, *et al*. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* 2006;**98**:1183–92.
52. **Lonning PE,** Sorlie T, Borresen-Dale AL. Genomics in breast cancer-therapeutic implications. *Nat Clin Pract Oncol* 2005;**2**:26–33.
53. **Bogaerts J,** Cardoso F, Buyse M, *et al*. Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. *Nat Clin Pract Oncol* 2006;**3**:540–51.
54. **Cardoso F,** Van't Veer L, Rutgers E, *et al*. Clinical application of the 70-gene profile: the MINDACT trial. *J Clin Oncol* 2008;**26**:729–35.
55. **Foekens JA,** Atkins D, Zhang Y, *et al*. Multicenter validation of a gene expression-based prognostic signature in lymph node-negative primary breast cancer. *J Clin Oncol* 2006;**24**:1665–71.