# Gene Expression Profiling in Human Breast Cancer – Toward Personalised Therapeutics?

Plamena N. Gabrovska, Robert A. Smith, Larisa M. Haupt and Lyn R. Griffiths[*]

*Genomics Research Centre and Griffith University Institute for Health and Medical Research, Parklands Drive, Southport, Queensland 4215, Australia*

**Abstract:** The most integrated approach toward understanding the multiple molecular events and mechanisms by which cancer may develop is the application of gene expression profiling using microarray technologies. As molecular alterations in breast cancer are complex and involve cross-talk between multiple cellular signalling pathways, microarray technology provides a means of capturing and comparing the expression patterns of the entire genome across multiple samples in a high throughput manner. Since the development of microarray technologies, together with the advances in RNA extraction methodologies, gene expression studies have revolutionised the means by which genes suitable as targets for drug development and individualised cancer treatment can be identified. As of the mid-1990s, expression microarrays have been extensively applied to the study of cancer and no cancer type has seen as much genomic attention as breast cancer. The most abundant area of breast cancer genomics has been the clarification and interpretation of gene expression patterns that unite both biological and clinical aspects of tumours. It is hoped that one day molecular profiling will transform diagnosis and therapeutic selection in human breast cancer toward more individualised regimes. Here, we review a number of prominent microarray profiling studies focussed on human breast cancer and examine their strengths, their limitations, clinical implications including prognostic relevance and gene signature significance along with potential improvements for the next generation of microarray studies.

**Keywords:** Breast cancer, microarray, clinical implications, prognostic, therapeutics.

## INTRODUCTION

A leading cause of death among women, breast cancer is one of the most common malignancies, with the sporadic form of the disease constituting more than 90% of all breast cancers. As the level of expression of every gene involved during the malignant transformation of a cell is controlled by the transfer of information encoded in its genetic blueprint to its environment *via* signal transduction and regulatory processes, we can examine this process at various stages [1].

Although estrogen receptor (ER) status is predictive of response to hormonal treatments, there are currently no clinically useful predictive markers of a patient's response to chemotherapy. As a result all patients who are eligible for chemotherapy receive the same treatment even though *de novo* drug resistance will result in the treatment failing in a large number of cases [2]. Most breast cancer patients treated with adjuvant chemotherapy do not get any benefit, yet are still exposed to the toxicity of the therapy. This consideration has led to the hypothesis that identifying predictors of prognosis may identify those patients who could benefit and those who could be spared from adjuvant chemotherapy. The expected benefit from such predictors would be to decrease acute and latent toxic effects and to reduce the cost associated with the treatment of the disease [3]. The underlying goal of improving systemic treatments of breast

cancer is to evolve from a shotgun approach of treating every patient with relatively non-specific cytotoxic chemotherapy or hormonal therapy to a more direct practice in which patients are treated with therapies aimed at specific molecular targets [4].

## MOLECULAR GENE EXPRESSION APPROACHES TO BREAST CANCER INVESTIGATION

The application of gene expression profiling using microarray technologies is an excellent, integrated, approach towards understanding the multiple molecular events and mechanisms by which cancer may develop [5]. As the process of oncogenesis involves the disruption of diverse cellular pathways including cell cycle, growth, survival and apoptosis, the high throughput of microarray analyses provides a powerful tool with which to examine multiple cellular processes simultaneously. Microarray hybridisation shows a snapshot of the complete cellular transcriptome on a single microarray chip, providing investigators with a global perspective of the complex interactions among thousands of genes [6]. RNA profiles obtained from microarray analysis are a static representation of the biological sample and yields the highest information and throughput of any classification assay. The identification of a gene signature specific to tumour cells may provide important etiological and diagnostic clues in the complex interactions of genes involved in cancer development.

The differentially regulated synthesis and degradation of RNA molecules forms hierarchal systems which determine organ, tissue and cell function through complex and interactive pathways. Therefore, molecular abnormalities

*Address correspondence to this author at the Genomics Research Centre, Griffith University Gold Coast Campus, Parklands Drive, Southport 4215, Queensland, Australia; Tel: + 61 7555 29158; Fax: + 61 7555 28908; E-mail: l.griffiths@griffith.edu.au

controlling the cellular gene transcription machinery may in turn contribute to the tumour phenotype [6]. For this reason, traditional techniques which focus on a single gene or a limited group of genes, limit the ability of researchers to identify complex gene interactions. Already, the identification of breast cancer molecular subtypes and the development of prognostic and predictive molecular signatures through gene expression profiling have resulted in a better appreciation of the biologic heterogeneity of breast cancer [7]. While studies have shown that the use of gene expression profiling holds tremendous potential as a new prognostic and predictive tool, to date, the clinical implementation of this technology has been limited. Here we review existing literature regarding the use of gene-expression based profiling in breast cancer and examine the

future clinical implications of this emerging technology.

## GENE EXPRESSION PROFILING IN HUMAN BREAST CANCER

A recent search of the PubMed database (Entrez, www.ncbi.nlm.nih.gov) identifies over 1500 publications with the key words "gene expression profiling using microarray in human breast cancer". To date, most of this research has focussed on the characterisation of transcriptional profiles of breast cancer utilising one of two platforms, cDNA microarrays or oligonucleotide arrays. Here, we review a combination of these (8 cDNA microarray studies and 13 oligonucleotide studies) in an effort to summarise gene expression studies utilising microarray

**Table 1.    Summary of cDNA Microarray Studies Investigating Human Breast Cancer, Including the Identified Key Markers and the Study Conclusions**

| Study | Tumour Sample Cohort | Number of Probes Used | Key Markers | Findings/Conclusion |
|---|---|---|---|---|
| Perou *et al.* (2000) | 42 | 8 102 | • Cluster included Ki-67 and PCNA <br> • STAT1 showed substantial variation in expression | • Classified tumours into 4 major groups <br> • Identification of a novel subgroup of basal-like breast tumours |
| Sorlie *et al.* (2001) | 78 | 8 102 | • Basal-like subtype characterized by high expression of keratins 5 and 17, laminin, and FABP7 <br> • ERBB21 subtype characterized by high expression of ERBB2 and GRB7 <br> • Luminal subtype A, demonstrated highest expression of the ER, GATA BP 3, X-BBP 1, TF3, HNF3a and LIV-1. | • Identification of novel luminal-type subclasses luminal A, luminal B and luminal C <br> • Basal-like, erbB2, luminal subtype BC worse clinical outcome: luminal A subtype good clinical outcome |
| Gruvberger *et al.* (2001) | 58 | 6 728 | • Top 5 genes higher expression in ER+ : ESR1, TFE3, GATA3, P28 and SFRS5 <br> • Top 5 genes higher expression in ER- : S100A8, LCN2, CDH3, PFKP and LAD1 <br> • Confirmed by SAGE: STC2 and CCND1 which had higher expression in ER+ and SLC7 A5 which had higher expression in ER- | • Developed class predictor of 100 genes that best distinguish ER+ and ER- breast tumours <br> • ER presence or absence determines distinct gene expression patterns |
| Hedenfalk *et al.* (2001) | 16 | 6 512 | • Tumours separated into 2 groups which manifested significantly different expression of the CYP1A1 | • Mutation in BRCA1 or BRCA2 gene results in gene expression profiles distinct from sporadic tumours |
| Sotiriou at al. (2003) | 99 | 4 336 | • GSTM3 emerged as an important survival marker both this study and van't Veer *et al.* (2002) <br> • PRAME, MYBL2 and BUB1 | • Gene profiles that relate to prognosis may help define new therapeutic targets <br> • Cell cycle regulation suggests continued use of anti-proliferatives is a rational approach. |
| Sorlie *et al.* (2003) | 115 | 13 691 | • Tested prognostic impact of the 231 markers published by van't Veer *et al.* on Norwegian cohort - performed less well (47%) in predicting recurrences within 5 years | • Classification of breast cancer based on gene expression profiling captures the molecular complexity of tumours |
| Ma *et al.* (2003) | 36 | 1 940 | • 39 genes with increased expression in IDC relative to DCIS <br> • CKS2, RRM2 and UBE2C showed increased expression in IDC relative to DCIS (validated by QRT-PCR) | • In contrast to tumour stage, different tumour grades are associated with distinct gene expression signatures |
| Zhao *et al.* (2004) | 60 | 13 671 | • Genes whose expression significantly differs between ILCs and IDCs: AOC3↑, ANXA1↑, F11R↓, AKR1C2↑, PTGS2↑, TNF↓, PKP3↓, PLA2G2A↑, ALOX15B↑, SAA2↑ | • Gene expression profiling has revealed distinct patterns of gene expression among ILCs and IDCs |

hybridisation to derive molecular profiles of human breast cancer. These studies are summarised in Tables **1** and **2** respectively.

## cDNA MICROARRAY STUDIES

cDNA microarrays are a competitive hybridization reaction where a sample of interest (e.g. tumour) is directly compared with a reference sample (e.g. normal tissue). In order to accomplish this, RNA isolated from the sample of interest and the reference sample are differentially labelled using distinguishable fluorescent dye-conjugated deoxyuridine triphosphates (d-UTPs) (usually cyanine-3 (Cy3) and cyanine 5 (Cy5)) by reverse transcription using random primers. The differentially labelled reference and tumour cDNAs are pooled and hybridised to a glass slide spotted with cDNA clones. A dual channel pair-wise fluorescence measurement is carried out for each spot and the relative abundance of the mRNA in the tumour reference mRNA sample is obtained as a ratio of Cy5/Cy3 intensities [5].

Through hierarchical cluster analysis Perou and colleagues [8], analysed the expression profiles of primary breast tumours. Various 'gene clusters' were recognized as biologically distinct networks reflecting the phenotypic wiring of individual tumours. These molecular portraits revealed information on numerous biological levels, from broad tumourigenic properties to discrete biochemical pathways and intra-tumour tissue heterogeneity. These differences led to the discovery of an intrinsic gene subset able to distinguish between multiple new cancer subtypes on the basis of fundamental tumour properties associated with cell-type origin [9]. These subtypes, termed Luminal A/ER+, Luminal B/ER+, Normal Breast-like, ERBB2+, and Basal-like (termed the Perou–Sorlie subtypes), were later demonstrated to be stable and reproducible classes apparent in different patient populations, which correlated significantly with tumour recurrence and patient survival [9]. A set of 1753 genes was selected on the basis of a four-fold change in gene expression from the median abundance of transcript in the sample set. The largest distinct cluster of genes within the 1,753-gene cluster diagram was a group of genes whose levels of expression correlated with cellular proliferation rates. The expression profile of this cluster of genes varied widely among the tumour samples but was generally well correlated with the mitotic index. As one might expect, this cluster also included the genes encoding two widely used immunohistochemical markers of cell proliferation including Ki-67 and PCNA. In addition, a large cluster of genes regulated by the interferon pathway (including STAT1) showed substantial variation in expression among the tumours, as was previously observed in a smaller set of breast tumours [8].

In a subsequent study, Sorlie and colleagues [10] examined 78 breast tumours, 3 fibroadenomas and 4 normal breast samples to determine if a correlation between microarray-based tumour classification and clinical outcome could be established. The intrinsic subset of 476 genes resulted in a similar distribution based on tumour type as originally described [8], but the luminal subtype displayed two major sub-groups, the luminal type A, and another group, luminal types B and C. Luminal subtype A, demonstrated highest expression of the ER gene, GATA BP

3, X-BBP 1, TF3, HNF3a, and estrogen-regulated LIV-1. The basal-like subtype was characterized by increased expression levels of keratins 5 and 17, laminin, and FABP7. The ERBB21 subtype was characterized by high expression of both ERBB2 and GRB7. Interestingly, the luminal subtypes differed significantly in their outcome of disease-free and overall survival, suggesting that this classification had novel clinical implications.

In the study conducted by Gruvberger *et al.* [11], breast cancers were shown to segregate into clusters depending upon whether the tumours were ER+ or ER-, demonstrating the association of ER status with distinctive gene expression patterns. To determine the minimum number of genes whose expression pattern could be used to stratify tumours into ER+ and ER- groups, Gruvberger *et al.* [11] then analysed the gene expression profiles of breast cancers from 58 node-negative patients, of which 23 were ER+ and 24 were ER-. The expression patterns of the 3389 genes studied were demonstrated to be altered in a statistically significant manner. The genes with higher expression in ER+ compared to ER- tumours included ESR1, TFE3, GATA3, P28 and SFRS5. In contrast, genes with higher expression in the ER- compared to ER+ tumours included S100A8, LCN2, CDH3, PFKP and LAD1. These results were then confirmed by serial analysis of gene expression (SAGE) which also found STC2 and CCND1 to be more highly expressed in ER+ tumours and SLC7 A5 which had higher expression in the ER- tumours. The SAGE analysis also revealed that several genes that belonged to the ER set of signature genes were not regulated directly by estrogen signalling, suggesting that other regulatory pathways may be involved in the expression of those genes. The study developed a class predictor of 100 genes used to distinguish ER+ and ER- breast tumours and found that ER presence or absence determined distinct gene expression patterns.

In 2001, Hedenfalk *et al.* [12] compared gene expression profiles of sixteen tumours from each of three distinct tumour types: i) BRCA1 mutation carriers; ii) BRCA2-positive patients; and iii) sporadic tumours. The authors selected a sub-set of 51 genes, whose variance among samples best discriminated between the three groups. Secondly, the authors used a class prediction method to determine if gene expression profiles could correctly classify the BRCA1 and BRCA2 mutation-positive tumours. The BRCA tumours were separated into two groups using 60 genes. Interestingly, many of the genes with increased expression between the groups were ribosomal, possibly indicating different capacities for protein biosynthesis between these groups. Nine genes were differentially expressed between BRCA1 mutation-positive and BRCA1 mutation negative tumours, whereas 11 genes were differentially expressed in BRCA2-positive and BRCA2-negative tumours. The two groups also manifested significantly different expression of the CYP1A1. The authors then used three statistical tests to select 176 genes that distinguished between the BRCA1 mutation and BRCA2 mutation positive tumours. Of the genes selected, BRCA1 mutations appear to affect genes in the DNA repair and apoptotic pathway displaying a 'stress'-type state. An important observation of this experiment is that BRCA1 and BRCA2 mutations dictate cellular pathways distinct from those observed in sporadic tumours.

**Table 2.**   **Summary of Oligonucleotide Microarray Studies Used to Examine Human Breast Cancers Including the Identified Key Markers and the Study Conclusions**

| Study | Tumour Sample Cohort | Number of Probes Used | Key Markers | Findings/Conclusion |
|---|---|---|---|---|
| West *et al.* (2001) | 46 | 5642 | • Genes that contribute to discrimination of ER status: Trefoil factor 1 (pS2), ER, Cytochrome P450, subfamily IIB, Trefoil factor 3 and ILGF | • Gene expression profile could predict metastatic potential even in the absence of reportedly positive nodes |
| Van 't Veer *et al.* (2002) | 117 | 11 141 | • Genes which were significantly upregulated in ER-group (poor prognosis signature) include: cyclin E2, MCM6, MMP9, MMP1, RAB6B, PK428, ESMI and FLT1 - all considered potential drug targets. | • Signature that defines ER status can be used to decide on therapy and signature that reveals BRCA1 status may further improve diagnosis of hereditary breast cancer. |
| Van de Vijver *et al.* (2002) | 295 | 25 000 | • Data indicates that the ability to metastasize to distant sites is an early and inherent genetic property of breast cancer | • Classification of patients into high-risk and low-risk subgroups on basis of prognosis profile - useful means of guiding adjuvant therapy in patients with lymph-node–positive breast cancer |
| Huang *et al.* (2003) | 87 | 8 628 | • Genes implicated in interferon response found to overlap with genes found in study by West el al. (2001) including STAT1, MX1, IFIT1, ISG115, IFI27, and IFI44 | • Group analysis of lymph-node risk defines metagene patterns that can accurately predict high-risk versus low-risk cases, in both internal and external validation studies |
| Ma *et al.* (2004) | 60 | 14 457 | • 3 genes were identified as differentially expressed: HOXB13, IL17BR and EST AI240933 | • HOXB13:IL17BR expression ratio may be useful for identifying patients appropriate for alternative therapeutic regimens in early-stage breast cancer |
| Chang *et al.* (2005) | 295 | 11 378 | • Used the 4 major group signature derived by Perou *et al.* (2000) and the 70 gene prognosis signature by Ven't Veer *et al.* (2002) and Chang *et al.*'s (2004) CSR genes to predict "wound response signature" (activated *vs* quiescent) to predict clinical outcome. | • Both overall survival and distant metastasis-free survival diminished in patients whose tumours expressed the wound response signature. <br>• Wound response signature improves risk stratification independently of known clinico-pathogenic risk factors. |
| Dai *et al.* (2005) | 311 | - | • Many highly expressed genes in tumours of poor outcome are cell cycle-associated genes including STK6, STK12, CCNB2, CCNE2, BUB1, CDC6, CDC25A, CDC45L, MAD2L1, RBL2, E2F1, KNSL5, UBE2C, UBCH10, PKMYT1 and BIRC5 | • By combining ER expression level and age, identified group of patients with relatively poor outcome <br>• Suggest that cell proliferation is the driving mechanism associated with poor outcome |
| Farmer *et al.* (2005) | 49 | 12 331 | • Molecular apocrine represent 8-14% of tumours in this and 4 published studies <br>• Apocrine tumours express HMGCR, GHR, PRLR and EGFR | • Divided mammary tumour cells into 3 groups based on steroid receptor activity: luminal (ER+ AR+), basal (ER- AR-) and molecular apocrine (ER- AR+) |
| Miller *et al.* (2005) | 251 | 16 308 | • p53-inducible genes PERP, BAX and SFN <br>• HOXB13 and IL17BR (Ma *et al.* 2004) also predictive of disease-specific survival in patients treated with Tamoxifen monotherapy | • Show the primary importance of p53 functional status in predicting clinical breast cancer behaviour |
| Wang *et al.* (2005) | 286 | 12 331 | • Genes implicated in disease progression: CAPN2, ORP, DS phosphatases, Rho-GDP DI, TNF superfamily protein, CC 3, MTA protein, PP 1 and apoptosis regulator BCL-G <br>• Previously characterised prognostic genes cyclin E2 and CD44 were also the gene signature | • Identified 76-gene prognostic signature, which was a better predictor in pre-menopausal breast cancer as compared to postmenopausal cases <br>• Identification of patients at high risk of distant recurrence |
| Chin *et al.* (2006) | 145 | - | • Suggest FGFR1, TACC1, ADAM9, IKBKB, PNMT, GRB7, PROCC, FNTA, ACACA and NR1D1 as high-priority therapeutic targets | • Low-level CNAs appear to contribute to cancer progression by altering RNA and cellular metabolism |
| Naderi *et al.* (2006) | 135 | 22 575 | • Over expression of extracellular matrix genes OMD, SPARCL1, LUM and developmental genes NANOG, TIMELESS, SOX11, SHOX2 <br>• Genes overlaping with van't Veer *et al.* 2002 BM039, CTPS, PSMD2, BUB1, MAD2L1, PSMD7 <br>• Genes overlapping with Dai *et al.* 2005 BUB1, MAD2L1, PKMYT1, BM039, PSMD7 | • 29 genes that associated with survival <br>• Identified a prognostic classifier and gene set, which predicts the overall survival in three independent studies |
| Archarya *et al.* (2008) | 964 | 22 000 | • 6 prognostically significant clusters representing patterns of oncogenic pathway activation and tumour biology/ microenvironment <br>• Activation of TNF-a, RAS, CTNNB, and SRC associated with worst prognosis. | • Incorporation of gene expression signatures into clinical risk stratification can refine prognosis |

A study by Sotiriou and colleagues in 2003 [13] correlated comprehensive gene expression patterns generated from cDNA microarrays with detailed clinico-pathological characteristics and clinical outcome in a group of 99 node negative and node positive breast cancer patients. Gene expression patterns were found to be strongly associated with ER status and moderately associated with grade but not associated with menopausal status, nodal status or tumour size. Hierarchical clustering separated the tumours into two main groups based on their ER status, which correlated well with basal and luminal characteristics. Statistical analysis identified 16 genes that were significantly associated with relapse-free survival at a stringent significance level of 0.001 to account for multiple comparisons. Of 231 genes previously reported by van't Veer *et al.* (2002) as being associated with survival, 93 probe elements overlapped with the set of 7,650 probe elements represented on the arrays used in this study. Hierarchical cluster analysis based on this set of probe elements segregated the population into two distinct subgroups with different relapse-free survival, with relapse-free survival (P < 0.05) represented by 11 unique genes. Genes involved in cell cycle, DNA replication and chromosomal stability were consistently elevated in the various poor prognostic groups. Another significant finding was the melanoma tumour antigen PRAME, along with MYBL2 and BUB1 were identified as being significantly differentially expressed between the two groups. In addition, glutathione S-transferase M3 emerged as an important survival marker in both studies.

In 2003, Sorlie and colleges [14] set out to refine their previously defined subtypes of breast tumours distinguished by distinct patterns of gene expression. They analysed a total of 115 malignant breast tumours by hierarchical clustering based on patterns of expression of 534 intrinsic genes resulting in one basal-like, one ERBB2-overexpressing, two luminal-like, and one normal breast tissue-like subgroups. They also reanalysed published data from two independent studies by van't Veer *et al.* (2002) and West *et al.* (2001) which are also discussed in the oligonucleotide microarray studies section of this review. They derived similar cluster analyses for the two published independent data sets representing different patient cohorts from different laboratories which uncovered some of the same breast cancer subtypes. The authors also tested the prognostic impact of the 231 markers published by van't Veer *et al.* (2002) on their cohort which performed less well (47%) in predicting recurrences within 5 years. Based on these findings they concluded that rather than individual tumour markers, patterns that distinguish subtypes appear to provide a more refined stratification.

The same year as Sorlie and colleagues [14] published their findings, Ma *et al.* (2003) [15] conducted a cDNA microarray study on 36 tissue specimens that demonstrated one or more pathological lesions: ADH (atypical ductal hyperplasia); DCIS (ductal carcinoma in situ); and IDC (invasive ductal carcinoma). They generated an *in situ* gene expression profile associated with the premalignant, pre-invasive, and invasive stages of human breast cancer. A paired t-test was performed on 11 patient-matched DCIS–IDC pairs to identify genes with increased expression in IDC relative to DCIS. A total of 39 genes were identified in the

100-gene grade III signature after adjusting the *P* value. Three of these genes were validated by QRT-PCR (CKS2, RRM2 and UBE2C). RRM2 may play a dual role in supporting rapid cell proliferation and promote invasive growth behaviour thus linking higher tumour grade (higher proliferation rate) and the DCIS-IDC transition (invasion). They revealed extensive similarities at the transcriptome level among the distinct stages of progression which suggest that gene expression alterations that show the potential for invasive growth are already present in the pre-invasive stage. In contrast to tumour stage, Ma *et al.* (2003) [15] suggest that different tumour grades are associated with distinct gene expression signatures. In their conclusions, they propose that a subset of genes associated with high tumour grade is quantitatively correlated with the transition from pre-invasive to invasive tumour growth.

Zhao *et al.* (2004) [16] conducted a comprehensive gene expression profiling study to determine if IDC and ILC (invasive lobular carcinoma) represent molecularly distinct entities and which genes (if any) are involved in the development of these distinct phenotypes. The study included 64 samples which ranged from ILCs, IDCs, and lymph node metastases to normal tissues. Eleven of 21 (52%) of the ILCs (referred to as "typical" ILCs) clustered together and displayed a different gene expression profile to the IDCs, whereas the other ILCs (the more "ductal-like" ILCs) were distributed between different IDC subtypes. Genes with significantly upregulated expression in ILCs when compared with IDCs included AOC3, ANXA1, AKR1C2, PTGS2, PLA2G2A, ALOX15B and SAA2. In contrast, F11R, TNF and PKP3 were shown to be significantly down-regulated in the ILCs. Combined, this data suggests that over half the ILCs differ from IDCs not only in their histological and clinical features but also in global transcription programs.

## OLIGONUCLEOTIDE MICROARRAY STUDIES

The preparation of oligonucleotide microarrays involves multiple steps [5]. Total RNA samples are reverse transcribed into cDNA using an oligo-dT primer with a T7 polymerase promoter sequence at the 5' end. Unlike cDNA microarrays, uniform labelling of the cDNA is not performed in oligonucleotide arrays. The cDNA is then converted to double-stranded DNA (dsDNA). This dsDNA is then used for *in vitro* transcription using T7 RNA polymerase to generate cRNA in the presence of biotinylated rNTPs for hybridisation to the chip. Signal is detected by the biotin streptavidin detection system where streptavidin is conjugated to a suitable fluorophore, the abundance of which can be determined using a laser scanner. Thus, oligonucleotide arrays have an additional step of target RNA amplification *via in vitro* transcription, leading to loss of the most linear relationship between the samples studied. In addition, a second loss of linearity occurs during the detection of the hybridized cDNA. Whether such processing leads to a bias in the overall data or a sub-set of genes is debatable. Amplification, however, allows one to perform array analyses on extremely small samples of low abundance template such as early tumour lesions and biopsies. The use of appropriate controls has been shown to reduce data bias [17].

In 2001, West and colleges [18] hybridised RNA from 46 tumour samples onto a custom made HuGeneFL array to develop patterns which classified breast tumours on the basis of ER status and lymph node involvement. They assessed the utility and validity in predicting the status of tumours. Through the use of cross validation determinations they determined that genes that contributed to discrimination of ER status included trefoil factor 1 (pS2), ER itself, cytochrome P450, subfamily IIB, trefoil factor 3 and insulin-like growth factor. The gene expression profile they developed was able to predict the metastatic potential of the tumours even in the absence of reportedly positive nodes. The authors conclude that the practical value of such an approach relies on the ability not only to assess the relative probabilities of clinical outcomes for future samples, but also to provide an honest assessment of the uncertainties associated with these predictive classifications on the basis of the selection of the specific gene subsets used for each validation analysis.

Although breast tumours may be classified similarly based upon traditional pathological criteria, response to treatment and clinical outcome vary widely. In 2002, van't Veer *et al.* [2] used microarray analyses to associate gene expression signatures with the history and clinical outcome of breast cancers. The investigation utilised 117 tumours from node-negative patients under the age of 55 years. Of the 25 000 genes analysed, approximately 5000 genes were regulated in a way that led to the classification of the tumours into 2 distinct groups containing 62 and 36 tumours, respectively. The authors subsequently used 78 tumours with known clinical outcome to determine a gene expression signature predictive of good and poor prognosis *via* artificial neural networks. In addition, they established a signature that identified the tumours of BRCA1 carriers. The poor prognosis signature consists of genes regulating cell cycle, invasion, metastasis and angiogenesis. Genes which were found to be significantly up-regulated in the ER- group (poor prognosis signature) included cyclin E2, MCM6, MMP9, MP1, RAB6B, PK428, ESMI and FLT1, which are all considered potential drug targets. As a consequence, the authors provided a strategy to identify patients who would potentially benefit from adjuvant therapy and established a molecular classification of tumours.

A number of clinical studies have correlated alterations in the expression of individual genes with disease outcome with often contradicting results. These include *cyclin D1, ER-α, UPA, PAI-1, HER-2* and *c-myc*, none of which were represented in van't Veer's set of 70 marker genes [2]. This discrepancy may be due to the fact that the van't Veer study determined expression at the level of transcription while the clinical studies measured protein levels. It is more probable however that these genes in isolation have only limited predictive power, highlighting the need for a multi-gene approach. Results from van't Veer *et al.* (2002) indicated that breast cancer prognosis can be determined from the gene expression profile of the primary tumours and that over-expressed genes in tumours with a poor prognosis profile are potential targets for the rational development of new cancer drugs. The classification of these targets may improve the efficacy of therapeutic development for all tumour types. The conclusions drawn from the van't Veer study suggest that small primary tumours negative for node metastases can display a poor prognosis signature but may already be programmed for metastasis. Thus, gene expression profiles may be utilized to determine prognostic classifiers in order to predict candidacy for specific therapeutic treatments.

In an effort to substantiate the classification based on this 70-gene signature, van de Vijver *et al.* [19] initiated a study using a cohort of 295 young breast cancer patients. All patients in this cohort were aged 53 or younger, but lymph-node positive patients were included. Interestingly, this was only a partially independent data series since 61 of the lymph-node negative patients from the original analysis were also included in this series. Using an optimized sensitivity threshold established in the original training series, 180 patients had a 'poor prognosis' signature and 115 had a 'good prognosis' signature with a mean overall 10-year survival rate of 54.6±4.4% and 94.5±2.6%, respectively. In multivariate analysis, the poor-prognosis signature was a more powerful predictor of distant metastases than other independent clinico-pathologic features. This data demonstrated the classification of patients into high-risk and low-risk subgroups on the basis of a prognostic profile may be a useful means by which to guide adjuvant therapy in patients with lymph-node positive breast cancers.

In 2003, Huang and colleges [20] aimed to predict nodal metastatic states and relapse for breast cancer patients. They analysed DNA microarray data from samples of primary breast tumours, using non-linear statistical analyses to assess multiple patterns of interaction of groups of genes that have predictive value for the individual patient, with respect to lymph node metastasis and cancer recurrence. They identified aggregate patterns of gene expression (metagenes) that associated with lymph node status and recurrence, and that were capable of predicting outcomes in individual patients with an approximate 90% accuracy. The metagenes defined distinct groups of genes, suggesting different biological processes underlie these two characteristics with breast cancer incidence. Genes implicated in interferon response found to overlap with genes found in study by West *el al.* (2001) [18] include STAT1, MX1, IFIT1, ISG115, IFI27, and IFI44. Multiple aggregate measures of gene expression profiles defined valuable predictive associations with lymph node metastasis and disease recurrence for individual patients. Group analysis of lymph-node risk defined metagene patterns that accurately predicted high-risk versus low-risk cases, in both internal and external validation studies.

In 2004, Ma *et al.* [21] generated gene expression profiles of hormone receptor-positive primary breast cancers in a set of 60 patients treated with adjuvant tamoxifen monotherapy. Tamoxifen significantly reduced tumour recurrence in several patients with early-stage ER+ breast cancers, yet markers predictive of treatment failure have not been identified. Three genes were identified as differentially expressed in both the LCM (laser capture microdissection) and whole tissue section analyses including HOXB13, IL17BR and EST AI240933. An expression signature predictive of disease-free survival was reduced to a two-gene ratio, HOXB13 versus IL17BR, which outperformed existing biomarkers. These findings indicate that the HOXB13:IL17BR expression ratio may be useful for

identifying patients who may benefit from alternative therapeutic regimens in early-stage breast cancers.

Work by Chang *et al.* (2005) [22] determined that the best validation of a gene signature's prognostic value is to test its ability to predict outcome in large independent data sets. In their study they examined a database of 295 breast cancer patients that had previously been used to identify and validate prognostic gene expression profiles defined by a set of 70 genes by van't Veer *et al.* (2002) [2] and Van de Vijver *et al.* (2002) [19]. They used the data set to test the reproducibility of the association between the wound response signature and breast cancer progression, and to investigate how the information from diverse gene expression signatures identified by various means might be integrated both biologically and for clinical use. Their results demonstrate that both overall survival and distant metastasis-free survival are markedly diminished in patients whose tumours expressed the wound-response signature. The wound-response signature improved risk stratification independently of known clinico-pathologic risk factors and other previously established prognostic signatures based on unsupervised hierarchical clustering (''molecular subtypes'') or supervised predictors of metastasis (''70-gene prognosis signature'').

In 2005, Dai and colleges [23] used 311 breast carcinoma samples (295 of which were from the Van de Vijver *et al.* study [19]) to show that within a subset of patients characterized by relatively high ER expression for their age, the occurrence of metastases was strongly predicted by a homogeneous gene expression pattern almost entirely consisting of cell cycle genes. Over-expression of this set of genes was clearly associated with an extremely poor outcome, with the 10-year metastasis-free probability only 24% for the poor prognosis group, compared with 85% for the good prognosis group. This gene expression pattern was not correlated with outcome in the other patient subpopulations. Examination of the molecular functions and biological processes of the 50 prognostic genes revealed that many of the highly expressed genes in tumours of poor outcome are cell cycle-associated genes (for example, STK6, STK12, CCNB2, CCNE2, BUB1, CDC6, CDC25A, CDC45L, MAD2L1, RBL2, E2F1, KNSL5, UBE2C, UBCH10, PKMYT1, and BIRC5). By combining the ER expression level and age, the authors identified a group of patients with relatively poor outcome. Within this group, a gene expression classifier identified a subgroup of patients with an almost 70% chance of metastasis. Importantly, this gene expression classifier suggests that cell proliferation is the driving mechanism associated with poor outcome. These results indicate that further refinements of diagnostic predictors may more often be generated by combining different informative clinical and molecular variables.

In 2005, Farmer *et al.* [24] examined a group of breast tumours with increased androgen signalling for correlation with a 'molecular apocrine' gene expression profile. They used tumour samples from 49 patients with large operable or locally advanced breast cancers. Principal component analysis and hierarchical clustering split the tumours into three groups: basal (ER- AR-), luminal (ER+ AR+) and a group classified as 'molecular apocrine' (ER- AR+). They determined that ERBB2 amplification was more common in the molecular apocrine group. To determine whether molecular apocrine tumours are present but under-reported in other breast cancer microarray studies, they identified genes able to discriminate between the three groups in their own data and correlated this expression pattern with four published data sets also discussed in this review (West *et al.,* 2001; van't Veer *et al.,* 2002; Huang *et al.,* 2003; Sorlie *et al.,* 2003). The luminal/apocrine/basal (LAB) genes were mapped in the other data sets and the top 90 genes for each pair-wise comparison in each data set retained. Genes that best split the three groups were identified, the average expression profile correlated and compared with the expression profile of individual tumours in the other four studies, suggesting that molecular apocrine tumours represent 8–14% of tumours in the studies. Their data also demonstrated that apocrine tumours express HMGCR, GHR, PRLR and EGFR. Overall, their results indicated the ability of microarray data to divide mammary tumour cells into three groups based on steroid receptor activity.

In an effort to prove whether the transcriptional fingerprint is a more definitive downstream indicator of p53 function, Miller *et al.* (2005) [25] analysed the transcript profiles of 251 p53-sequenced primary breast tumours. They identified a clinically embedded 32-gene expression signature that distinguishes p53-mutant and wild-type tumours of different histological grade that outperforms sequence-based assessments of p53 in predicting prognosis and therapeutic response. From their microarray hierarchal cluster analysis, the authors demonstrated that the p53-inducible genes PERP, BAX and SFN were all expressed at higher levels in the p53 mutants and the 26 mt-like tumours. They also determined that the previously implicated HOXB13 and IL17BR genes [21] were also predictive of disease-specific survival in patients treated with Tamoxifen monotherapy. These findings highlighted the primary importance of p53 functional status in predicting clinical breast cancer behaviour.

Both Wang *et al.* [26] and Naderi *et al.* [27] identified gene signatures predictive of distant metastases or survival. Wang *et al.* [26] identified a 76-gene signature consisting of 60 genes for patients positive for ER α and β and 16 genes for ER-negative patients. This signature showed 93% sensitivity and 48% specificity in a subsequent independent testing set of 171 lymph-node-negative patients. The gene profile was highly informative in identifying patients who developed distant metastases within 5 years, even when corrected for traditional prognostic factors in multivariate analysis. The 76-gene profile also represented a strong prognostic factor for the development of metastasis in the subgroups of 84 pre-menopausal patients and 79 patients with tumours of 10–20 mm, a group of patients for whom prediction of prognosis is especially difficult.

In Naderi *et al.*'s study [27] a prognostic signature of 70 genes that significantly correlated with survival was identified. Using two different prognostic classification schemes and measures, the 70-gene classifier was also found to be prognostic in two independent external data sets. Overall, the 70-gene set was prognostic in a 715 patient cohort. Finally, a common prognostic module of 29 genes that associated with survival in both the cohort and the two external data sets was identified. The evidence that some of

these 29 genes are part of a 'core' prognostic signature is strengthened by the fact that six (BM039, CTPS, PSMD2, BUB1, MAD2L1, PSMD7) were also identified as part of the 231 genes that correlated with the prognostic categories in the original van't Veer *et al.* study [2] and five (BUB1, MAD2L1, PKMYT1, BM039, PSMD7) overlap with a cell proliferation signature recently derived from the same data and associated with extremely poor outcome [23].

Chin *et al.* [28] comprehensively profiled 101 tumours at the DNA and RNA level, which allowed the authors to better define the impact of specific genetic events on breast cancer phenotypes in relation to clinical outcome. Sixty-six genes deregulated by high-level amplifications were defined as potential therapeutic targets. This indicated that genomic profiles provide additional prognostic information as compared to what is available from transcriptomic profiles alone. The study listed nine potential therapeutic targets, which much like the prototype HER-2 oncogene, are activated by recurrent gene amplifications in breast cancer and may show an association with aggressive tumour types. Many of these are potential drug targets for small molecule inhibitors. Patients whose tumours had one or more DNA amplifications had a poor prognosis independent of the previously defined five major gene expression classes described by Perou *et al.* [8]. This study demonstrates that recurrent CNAs (copy number abnormalities) differ between tumour subtypes defined by expression pattern and that stratification of patients according to outcome can be improved by measuring both expression and copy number, especially high level amplification. Low-level CNAs appear to contribute to cancer progression by altering RNA and cellular metabolism.

The most current and, to date, largest clinical whole genome expression study was conducted in 2008 by Acharya *et al.* [29]. With a statistically favourable sample size of 964 clinically annotated breast tumour samples (573 in initial discovery set and 391 in the validation cohort), Acharya and colleagues looked at gene expression signatures and clinicopathological variables in breast cancers to determine a refined estimation of relapse free survival and sensitivity to chemotherapy. Since their work built on the current standard of care, i.e. clinicopathological risk stratification, their results provided an opportunity to better understand the biology underlying clinically relevant prognostic sub-phenotypes. Beyond prognosis this approach has the potential to dissect broad phenotypes while providing data to reveal novel therapeutic opportunities for patients at the highest risk of recurrence. Their results not only provided an opportunity to tailor targeted therapeutic approaches but also suggest opportunities for selection of therapies that may be the most effective in patients with specific sub-phenotypes. Potential limitations of the study by Acharya *et al.* [29] include the lack of data regarding hormonal therapy in some patients. The authors accounted for this by suggesting that the incorporation of the knowledge of hormonal receptor status (*via* the clinicopathological model) may function as a surrogate for response to hormonal therapy. In addition, the panel of signatures representing sensitivity to cytotoxic agents in their analysis was not exhaustive, with regiment-specific signatures for drug combinations proven to be effective in breast cancer proven to be more practical. A final limitation of the study was the small number of patients

within certain pathway clusters which hampered statistical comparisons. Taken together, these results provide preliminary evidence that incorporation of gene expression signatures into clinical risk stratification can be used as an effective tool to refine prognosis.

## SUMMARY

In summarising the 8 cDNA microarray studies and 13 oligonucleotide studies, several genes were found to overlap in a number of the studies. A total of 21 genes were found to be significantly differentially expressed in two or more of the studies (Table **3**). STAT1, a modulator of chemotherapy induced apoptosis [30] was not surprisingly differentially regulated in Perou *et al.* (2000), West *et al.* (2001) and Huang *et al.* (2003). Involved in cell proliferation and apoptosis [31], GRB7 was up-regulated in both the Sorlie *et al.* (2001) and Chin *et al.* (2006) studies. Interestingly, GATA3 (GATA BP 3), which has previously been found to inhibit breast cancer growth and pulmonary metastasis [32], was up-regulated in both Sorlie *et al.* (2001) and Gruvberger *et al.* (2001). GSTM1, involved in the metabolism of various carcinogens, emerged as a survival marker in both van't Veer *et al.* (2002) and Sotiriou *et al.* (2003). BUB1, involved in spindle checkpoint and preservation of correct ploidy through meiosis [33] was differentially expressed in van't Veer *et al.* (2002), Sotiriou *et al.* (2003), Dai *et al.* (2005) and Naderi *et al.* (2006).

Abnormal regulation of cell cycle control is one of the hallmarks of tumourigenesis [34], and UBE2C, involved in the cell cycle, was found to be over expressed in both Ma *et al.* (2003) and Dai *et al.* (2005). Since the TNF super-family of genes are involved in control of cell proliferation, differentiation and apoptosis [26], one would expect an over-expression of these genes; however in Zhao *et al.* (2004) TNF was down-regulated while in Archaya *et al.* (2008) it was associated with a worse prognosis. Another gene involved in cell cycle through the regulation of progression of cells [34] is cyclin E2 which was significantly up-regulated in van't Veer *et al.* (2002) and Wang *et al.* (2005). A regulator of translation initiation [35], IFIT1 was differentially expressed in both West *et al.* (2001) and Huang *et al.* (2003). A number of other genes were also found to overlap between these two studies and they all also appear to be involved in the interferon response including MX1, ISG115, IFI27 which is also a marker for epithelial proliferation [36] and IFI44, involved in T-cell function [37]. HOXB13 is pro-proliferative and pro survival role in ovarian cancer [38] while IL17BR has has involved in immunoregulatory activity [39] yet both of these genes were not only found to be significantly differentially expressed in both Ma *et al.* (2004) and Miller *et al.* (2005), but many other studies have also identified them as useful for identifying patients appropriate for alternative therapeutic regiments in early–stage breast cancer [40-42].

While its function is currently unknown [43], BM039 was differentially expressed in van't Veer *et al.* (2002), Dai *et al.* (2005) and Naderi *et al.* (2006). Meanwhile CTPS is involved in nucleic acid and phospholipid biosynthesis [44] and differentially expressed in both van't Veer *et al.* (2002) and Naderi *et al.* (2006). The protease subunits PSMD2, part of TNF signalling pathway [45] and PSMD7, part of the

**Table 3.    A Summary of cDNA and Oligonucleotide Microarray Studies Used to Investigate Human Breast Cancers**

| Gene | Function | Studies |
|---|---|---|
| STAT1 | Modulator of chemotherapy-induced apoptosis (Thomas *et al.* 2004) | Perou *et al.* 2000, West *et al.* 2001, Huang *et al.* 2003 |
| GRB7 | Cell proliferation and apoptosis (Han *et al.* 2001) | Sorlie *et al.* 2001, Chin *et al.* 2006 |
| GATA3, GATA BP 3 | Inhibits breast cancer growth and pulmonary breast cancer metastasis (Dynesborg *et al.* 2009) | Sorlie *et al.* 2001, Gruvberger *et al.* 2001 |
| GSTM1 | Metabolism of various carcinogens (Gudmundsdottir *et al.* 2001) | van't Veer at al 2002, Sotiriou *et al.* 2003 |
| BUB1 | Involved in spindle checkpoint and the preservation of correct ploidy through mitosis (Bernard *et al.* 1998) | van't Veer 2002, Sotiriou *et al.* 2003, Dai *et al.* 2005, Naderi *et al.* 2006 |
| UBE2C | Cell cycle (Ma *et al.* 2003) | Ma *et al.* 2003, Dai *et al.* 2005 |
| TNF, TNF super family, TNF a | Control of cell proliferation, differentiation and apoptosis (Want *et al.* 2005) | Zhao *et al.* 2004, Wang *et al.* 2005, Archaya *et al.* 2008 |
| Cyclin E2 | Regulates the progression of cells through the cell cycle (Gugas *et al.* 1999) | Van't Veer *et al.* 2002, Wang *et al.* 2005 |
| MX1 | Interferon target gene (Li *et al.* 2006) | West *et al.* 2001, Huang *et al.* 2003 |
| IFIT1 | Regulates translation initiation (Weichselbaum *et al.* 2008) | West *et al.* 2001, Huang *et al.* 2003 |
| IFI27, IFI44 | IFI27: novel marker of epithelial proliferation and cancer (Suomela *et al.* 2004)<br><br>IFI44: involved in interferon response, T-cell function gene (Sahar *et al.* 2005) | West *et al.* 2001, Huang *et al.* 2003 |
| ISG115 | Implicated in an interferon response (Huang *et al.* 2003) | West *et al.* 2001, Huang *et al.* 2003 |
| HOXB13 | Pro-proliferative and pro-survival role in ovarian cancer (Miao *et al.* 2007) | Ma *et al.* 2004, Miller *et al.* 2005 |
| IL17BR | Immunoregulatory activity (Goetz *et al.* 2006) | Ma *et al.* 2004, Miller *et al.* 2005 |
| BM039 | Currently unknown function (Lallier *et al.* 2007) | van't Veer *et al.* 2002, Dai *et al.* 2005, Naderi *et al.* 2006 |
| CTPS | Involved in nucleic acid and phospholipid biosynthesis (Kursula *et al.* 2006) | Van't Veer *et al.* 2002, Naderi *et al.* 2006 |
| PSMD2, PSMD7 | PSMD2: part of TNF signalling pathway (Dunbar *et al.* 1997)<br>PSMD7: ubiquitin proteasome pathway (Thompson *et al.* 2002) | Van't Veer *et al.* 2002, Dai *et al.* 2005, Naderi *et al.* 2006 |
| MAD2L1 | Involved in mitotic arrest (Percy *et al.* 2000) | Van't Veer *et al.* 2002, Dai *et al.* 2005, Naderi *et al.* 2006 |
| PKMYT1 | Involved in cell cycle progression and DNA damage response (Naderi *et al.* 2006) | Dai *et al.* 2005, Naderi *et al.* 2006 |

The table identifies the genes determined to be significantly differentially expressed in more than one of the studies examined.

proteasome pathway [46] were differentially expressed in van't Veer *et al.* (2002), Dai *et al.* (2005) and Naderi *et al.* (2006). Another gene which was significantly differentially expressed in all three of the same studies includes MAD2L1, involved in mitotic arrest [47]. The last gene found to be significantly differentially expressed was between Dai *et al.* (2005) and Naderi *et al.* (2006) which was PKMYT1 and is involved in cell cycle progression and DNA damage response [27].

These studies provide early evidence that the transcriptional mechanisms of breast cancer may not only provide novel insights into the biology of cancer but could also accurately identify certain previously non-discernible clinical phenotypes. The result would be the defining of new molecularly informed classifications able to delineate novel disease entities associated with patient outcomes [9]. The summarised experimental approaches in this review demonstrate that patho-biological simulations performed *in vitro* can reveal transcriptional configurations predictive of tumour biology [9]. Together, these functional genomic strategies are changing the scientific process of breast cancer biomarker discovery, toward one that incorporates mechanistic parameters. Microarray technology has begun to re-define our understanding of the breast cancer onco-transcriptome and how it relates to tumour biology and behaviour. From this point forward, the correlations between pathological mechanisms and clinical endpoints can be explored.

## FUTURE CLINICAL IMPLICATIONS - PROGNOSTIC RELEVANCE OF GENE EXPRESSION

Several criticisms, questions and perspectives have been raised regarding the prognostic gene signatures reported to date. Primarily, some argue that most of the signatures add little information when compared to an optimal clinico-pathological score that would include *ER, HER-2* and *Ki67* data in addition to the conventional clinical parameters. It is also worthy to note that most of the genes included in the various published prognostic gene signatures are related to cell proliferation and the question then arises as to whether a

simpler biomarker for such parameter like *Ki67*, which has been measured routinely for decades, could have provided similar results. However, gene expression profiling studies suggest that measuring proliferation with a more objective, automated and quantitative assay may be more robust than less quantitative assays such as immunohistochemistry.

Another criticism of the technology relates to the fact that most of the predictors were generated using a mix of molecularly heterogeneous tumours. It has been reported by Perou *et al.* [8] that the breast cancer population is a mix of at least four different molecular classes. Since oncogenic events are different across molecular classes, some have suggested that optimal predictors should be set up in each molecular class. This was applied by Wang *et al.* [26] who developed the 76-gene signature to identify patients at a high risk of distant recurrence based on the prognostic genes separately identified in ER– and ER+ tumours. This study demonstrated that proliferation was the strongest parameter predicting clinical outcome in the ER+/HER-2– subgroup of patients only, whereas immune response and tumour invasion appear to be the main biological processes associated with prognosis in the ER–/HER-2– and HER-2+ subgroups, respectively. Combined, this data implies that the molecular background of the tumour should be taken into consideration when making predictions regarding prognosis.

Despite enthusiasm about the possible clinical implications of gene-expression profiling, a source of concern has been that comparisons of gene sets derived from the various studies show little overlap. This may be due to the different types of arrays used, the sample quality and the defined parameters used for data interpretation. In gene-expression profiling with DNA microarrays, short DNA fragments, each representing a single gene, are spotted in a grid arrangement onto a solid membrane often referred to as a "chip". RNA extracted from fresh frozen tissue is then hybridized to the spots with the amount of detectable hybridization corresponding to the level of gene expression. Subsequently, a computer algorithm is used to identify different gene-expression patterns. To date, this technology requires the availability of RNA from frozen tissue, although researchers are currently investigating whether sufficient RNA could be amplified from fixed tissue for use in DNA microarray analyses. Given the inherent instability of RNA, quality control and reproducibility have posed challenges to the use of microarray technology [7]. Differences in the handling and processing of samples can influence RNA integrity with significant alteration of gene-expression data. Microarray technology is susceptible to a number of potential errors not just at the time of sampling, pre-processing and processing but also at the time of calibration and analysis of the data.

## GENE SIGNATURE SIGNIFICANCE

Even though the sensitivity of gene signatures for prognostic purposes appears excellent, there is still an approximate 5–10% error rate, with patients who will present with metastatic relapse being classified into the predicted low risk of relapse group. Although third generation chemotherapy has been shown to reduce breast cancer deaths by ~30%, optimal chemotherapy may still provide benefit even in this positive prognostic population.

Nevertheless, most of the tumours classified as having a good prognosis are actually predicted to be resistant to chemotherapy.

To date, several gene signatures for prognostic purposes have been generated and some of them are currently being validated in prospective trials. Although they potentially increase the number of patients who could be spared adjuvant chemotherapy while still correctly identifying the high-risk patients, they present some limitations that need to be considered in order to generate more accurate 'second generation' gene signatures. As already discussed, while gene signatures clearly represent a major step forward in the molecular prediction of patient outcome and drug sensitivity, they are still showing constraints that are the basis for the development of second generation gene predictors. These constraints in terms of clinical relevance are summarised in Table **4**.

Some solutions to these issues have been presented by Desmendt *et al.* [3]. First, considering that most of the first generation gene signatures capture genes correlated to clinicopathological characteristics, the focus of several groups is on the development of molecular predictors that contain genes selected to provide additional information to these characteristics. Second, since it is now clear that breast cancer can be considered as four distinct molecular groups, the next generation molecular predictors should focus on homogenous classes, already undertaken by several groups including Desmendt *et al.* [3] and Loi *et al.* [48]. Third, predictors for metastatic relapse should be designed to predict both early and late relapses. Fourth, drug-specific predictors should be optimally defined either in randomised trials that compare two drugs, or based on the two series of similar patients with different treatment schedules. The latter approach was used to generate a predictor that is specific to ixabepilone B as opposed to paclitaxel in a study by Hess *et al.* [49]. Finally, DNA microarrays used to generate first generation predictors contained between 22 and 44 K probe sets. Although some genes presented multiple probe sets, the vast majority of these probes were located in 3' end of the transcriptome. This implies that gene profiling actually detects the transcriptional activity, but is unable to determine which of the transcripts are expressed for a single gene. Next generation arrays investigated by Pando *et al.* [50], include splice arrays, which detect both gene expression and splicing events, thereby providing a more functional picture of the genomic programme in every patient.

While gene-expression profiling has already provided important insight into the biologic heterogeneity of breast cancer, optimal incorporation of these genomic tools into clinical practice is yet to be determined and will depend on trials such as TAILORx and MINDACT [7]. Importantly, controversy remains about how much refinement in prognostic genomic tools provide over the standard assessment of histological grade and other clinicopathologic features. The importance of grade-related genes in the prognostic power of most gene expression models has been demonstrated by Desmendt *et al.* [3], who hypothesized that proliferation may serve as the most important common denominator in predictive gene-expression signatures. As pointed out by Fan *et al.* [51] the 70-gene signature, the 21-gene recurrence-score, the wound-response and the two-

**Table 4.    Summarising the Benefits and Limitations of First Generation Microarray Technology. A List of Potential Solutions for the Design and Implementation of Second Generation Analyses Directed Toward Personalised Therapeutics**

| Beneficial Components of First Generation Microarray Technology | Constraints in Terms of Clinical Relevance of First Generation Microarray Technology | Potential Solutions for Second Generation Molecular Predictors |
|---|---|---|
| Feasible and reproducible across platforms<br><br>Validated in retrospective studies<br><br>Medical usefulness under evaluation in phase III trials | **Predictors for prognosis and drug sensitivity:**<br>• Added value to optimal clinico-pathological characteristics not proven<br>• Metric performances not optimal<br>• Generated regardless of breast cancer molecular classes<br>• Generated using the 1st generation of arrays, of poorer coverage and quality<br><br>**Predictors for prognosis:**<br>• Poor performance to predict metastases over 5 years<br><br>**Predictors for drug sensitivity:**<br>• Not generated based on direct comparison between two drugs, drug specificity to be determined | Predictors generated to provide additional prediction to an optimal clinico-pathological score: select probes that increase performance of clinical characteristics<br><br>Predictors generated for each molecular class<br><br>Use of latest generation of arrays to generate predictors: exon arrays and splice arrays, SNP arrays and CNA arrays<br><br>Predictor for late events<br><br>Generation of drug-specific predictor based on the randomised trials (interaction test) |

gene-expression ratio models are all significantly correlated with histological grade, although to some degree all of these models provide prognostic information beyond that obtained by histological grade alone. Moreover, the genomic grade index (GGI) developed by Sotiriou *et al.* [52] seems promising in its ability to separate the clinically ambiguous histological grade II tumours as well as the clinically heterogeneous ER-positive tumours into prognostically distinct sub-types. Given the difficulty in clinical decision making for patients with small ER-positive grade II tumours, further refinement of prognosis within this subgroup of patients would certainly prove useful. Additionally, genomic assessment of histological grade, ER, PR and HER-2 status may lead to improved objectivity and accuracy.

## FROM GENE EXPRESSION TO CLINICAL EPIDEMIOLOGY

Gene expression profiling studies have claimed that molecular markers, identified by discovery-based proteomics and genomics research, provide a high degree of discrimination for cancer diagnosis and prognosis. According to Ransohoff *et al.* [53] some important claims about markers for diagnosis and prognosis have been unreliable and only weakly reproducible or not reproducible at all and the process of development seems slow and inefficient. If promising initial results are unreliable but are used as a foundation, the overall process of development may be inefficient. Although tools of phases and guidelines are useful, study design has the most critical role in addressing fundamental problems of reliability. To improve reliability, strong unbiased samples should be used rather than convenient samples however, this is not always possible. Methods to improve the reliability and efficiency of the process of marker development are being aggressively explored. Biganzoli *et al.* [54] suggests that it is relevant to promote the application of suitable study designs and statistical methods for the reliable assessment of data collected on tumour markers, either genomic or ''old'', along with a faster translation of basic research to medical decision-making. These goals can be most successfully met through the cooperation of clinicians, biomedical informaticians and biostatisticians. According to Hayden *et al.* [55] prognosis-focussed research can provide useful information to guide clinicians in their management of patients. To do this most effectively, researchers and readers of the medical literature should consider the objectives and approach of prognosis research studies to move their literature forward.

## CONCLUSIONS

Studies of gene expression profiling have altered our view of breast cancer and provided us with new tools for molecular diagnoses. Technical advances are rapid in this field, and the microarray platforms that were used to develop these signatures interrogated the "mRNA world" [56]. Pioneer studies have shown that gene signatures provide powerful information that should allow a better tailoring of treatment for breast cancer patients in the next five to ten years [3]. Building on the limitations exhibited by the first generation predictors, (Fig. **1**) the second generation of gene signatures are currently being developed. This new wave of predictors should present a higher accuracy and be more complementary to pre-existing and routine treatment decision tools.

The next generation of microarrays (e.g., tiling arrays, microRNA arrays, and direct sequencing of complementary DNA) will enable investigators to study the clinical and diagnostic potential of new RNA species, including microRNAs, RNA transcribed from non-coding DNA, pseudogenes, and antisense DNA strands. An exciting prospect of microarray-based technology is that multiple, distinct predictions including prognosis, ER and HER-2 status, and sensitivity to various treatment approaches are potentially generated from a single assay. This type of test would use information from different sets of genes from the same tissue for different predictions. This outcome is technically feasible and could substantially improve the cost-effectiveness of a multigene assay. To provide treatment recommendations that are truly molecularly tailored to individual patients, it is important to measure the risk of relapse and the probability of benefit from endocrine therapy and chemotherapy separately and to consider the preferences
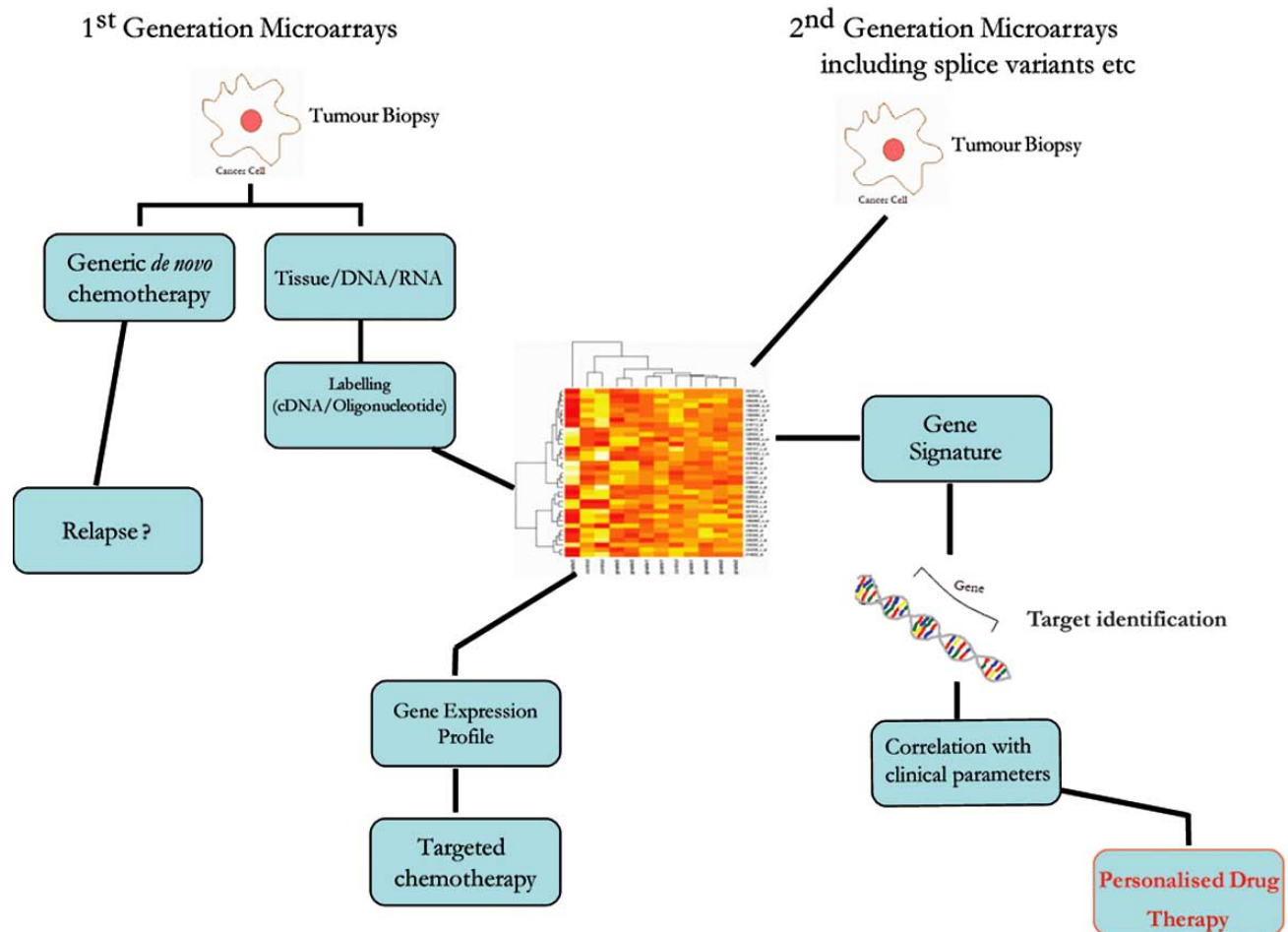
**Fig. (1).** Comparison of diagnostic and treatment options for breast cancer patients following identification of the tumour through biopsy and pathological diagnosis. Treatment regimes include **a**) traditional chemotherapy - not target specific and generally results in either *de novo* resistance to chemotherapy or relapse; **b**) 1st generation microarrays – tumour (stage/grade/current markers e.g. ER status) specific in combination with a gene expression profile which permits a more targeted chemotherapy approach; and **c**) 2nd generation microarrays – treatment regimes are based on a gene signature based on target identification including splice variants which, correlated with clinical parameters, result in a personalised drug and chemotherapy.

of patients within this context [56]. Another promising direction of research is to examine the hypothesis that different markers and biologic pathways may be involved in determining prognosis, response, and resistance to therapy in different molecular subgroups of breast cancers. As larger clinical data sets become available for gene expression analysis, it is conceivable that predictors of molecular class specific prognosis and treatment response will be developed.

As such, it is reasonable to postulate that at least initially, genomic results will be used in addition to and not instead of standard clinical and pathologic criteria [7]. Clinicians may rely on genomic tests as "tiebreakers" in selected cases. Whether these tools will be more broadly incorporated into clinical care is uncertain at present, but that represents the full potential of this technology. Further, it is unclear whether one or more tools, existing or yet to be developed, will emerge as optimal for widespread use. Technical and economic issues will need to be addressed and the manner in which genomic data is both presented and used in decision-making by physicians and patients must be optimised.

Gene expression arrays initially began as a method of multiplexing single gene discovery, similar to running several thousand quantitative RNA dot-blots. From this one-dimensional approach current expression profiling evolved to uncover pathway regulation of gene expression and to define molecular classes on the basis of integration of the total signals experienced by the cancerous cell. The ability to analyse and model complex systems is made possible by mathematical algorithms coupled with computational capacity which has been essential to this transition. The future of array-based expression genomics will lie in this analytical complexity and the advancement of such studies will have an immense impact on breast cancer research. The advantage will be the ability to be comprehensive yet precise at the same time and with time, the speed of discovery will be remarkable.

**ABBREVIATIONS**

BRCA-1/2  =  Breast cancer 1 & 2 gene

CAN       =  Copy number alterations

cDNA      =   Complementary deoxyribonucleic acid

CGH      =   Comparative genomic hybridisation

c-myc      =   Proto-oncogene originally isolated from an avian myelocytomatosis virus

DCIS      =   Ductal carcinoma *in situ*

DNA      =   Deoxyribonucleic acid

ER      =   Estrogen receptor

GGI      =   Genomic grade index

HER-2      =   Human EGF-like receptor 2

MINDACT   =   Microarray for Node-negative Disease may Avoid Chemotherapy

mRNA      =   messenger ribonucleic acid

PAI-1      =   plasminogen activator inhibitor 1

PR      =   Progesterone receptor

RNA      =   ribonucleic acid

SAGE      =   serial analysis of gene expression

TAILORx   =   Trial Assigning Individualized Options for Treatment

TP53      =   tumour protein 53

UPA      =   transcriptional regulation or urokinase

## REFERENCES

[1]    Zhang DH, Salto-Tellez M, Chiu LL, *et al*. Tissue microarray study for classification of breast tumors. Life Sci 2003; 73(25): 3189-99.

[2]    van 't Veer LJ, Dai H, van de Vijver MJ, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415(6871): 530-6.

[3]    Desmedt C, Ruiz-Garcia E, Andre F. Gene expression predictors in breast cancer current status, limitations and perspectives. Eur J Cancer 2008; 44(18): 2714-20.

[4]    Morris SR, Carey LA. Gene expression profiling in breast cancer. Curr Opin Oncol 2007; (6): 547-51.

[5]    Desai KV, Kavanaugh CJ, Calvo A, *et al*. Chipping away at breast cancer: insights from microarray studies of human and mouse mammary cancer. Endocr Relat Cancer 2002; (4): 207-20.

[6]    Nambiar PR, Boutin SR, Raja R, *et al*. Global gene expression profiling: a complement to conventional histopathologic analysis of neoplasia. Vet Pathol 2005; 42(6): 735-52.

[7]    Stadler ZK, Come SE. Review of gene-expression profiling and its clinical use in breast cancer. Crit Rev Oncol Hematol 2009; 69(1): 1-11.

[8]    Perou CM, Sorlie T, Eisen MB, *et al*. Molecular portraits of human breast tumours. Nature 2000; 406(6797): 747-52.

[9]    Miller LD, Liu ET. Expression genomics in breast cancer research: microarrays at the crossroads of biology and medicine. Breast Cancer Res 2007; 9(2): 206.

[10]   Sorlie T, Perou CM, Tibshirani R, *et al*. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 2001; 98(19): 10869-74.

[11]   Gruvberger S, Ringner M, Chen Y, *et al*. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. Cancer Res 2001; 61(16): 5979-84.

[12]   Hedenfalk I, Ringner M, Ben-Dor A, *et al*. Molecular classification of familial non-BRCA1/BRCA2 breast cancer. Proc Natl Acad Sci USA 2003; 100(5): 2532-7.

[13]   Sotiriou C, Neo SY, McShane LM, *et al*. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc Natl Acad Sci USA 2003; 100(18): 10393-8.

[14]   Sorlie T, Tibshirani R, Parker J, *et al*. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA 2003; 100(14): 8418-23.

[15]   Ma XJ, Salunga R, Tuggle JT, *et al*. Gene expression profiles of human breast cancer progression. Proc Natl Acad Sci USA 2003; 100(10): 5974-9.

[16]   Zhao H, Langerod A, Ji Y, *et al*. Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. Mol Biol Cell 2004 ; ( 6): 2523-36.

[17]   Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. Nature 2000; 405(6788): 827-36.

[18]   West M, Blanchette C, Dressman H, *et al*. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA 2001; 98(20): 11462-7.

[19]   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 2002; 347(25): 1999-2009.

[20]   Huang E, Cheng SH, Dressman H, *et al*. Gene expression predictors of breast cancer outcomes. Lancet 2003; 361(9369): 1590-6.

[21]   Ma XJ, Wang Z, Ryan PD, *et al*. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer Cell 2004; (6): 607-16.

[22]   Chang HY, Nuyten DS, Sneddon JB*, et al*. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc Natl Acad Sci USA 2005; 102(10): 3738-43.

[23]   Dai H, van't Veer L, Lamb J, *et al*. A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. Cancer Res 2005; 65(10): 4059-66.

[24]   Farmer P, Bonnefoi H, Becette V, *et al*. Identification of molecular apocrine breast tumours by microarray analysis. Oncogene 2005; 24(29): 4660-71.

[25]   Miller LD, Smeds J, George J, *et al*. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci USA 2005; 102(38): 13550-5.

[26]   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(9460): 671-9.

[27]   Naderi A, Teschendorff AE, Barbosa-Morais NL, *et al*. A gene-expression signature to predict survival in breast cancer across independent data sets. Oncogene 26(10): 1507-16.

[28]   Chin K, DeVries S, Fridlyand J, *et al*. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell 2006; (6): 529-41.

[29]   Acharya CR, Hsu DS, Anders CK, *et al*. Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer. JAMA 2008; 299(13): 1574-87.

[30]   Thomas M, Finnegan CE, Rogers KM, *et al*. STAT1: a modulator of chemotherapy-induced apoptosis. Cancer Res 2004; 64(22): 8357-64.

[31]   Han DC, Shen TL, Guan JL.The Grb7 family proteins: structure, interactions with other signaling molecules and potential cellular functions. Oncogene 2001; 20(44): 6315-21.

[32]   Dydensborg AB, Rose AA, Wilson BJ, *et al*. GATA3 inhibits breast cancer growth and pulmonary breast cancer metastasis. Oncogene 2009; (29): 2634-42.

[33]   Bernard P, Hardwick K, Javerzat JP. Fission yeast bub1 is a mitotic centromere protein essential for the spindle checkpoint and the preservation of correct ploidy through mitosis. J Cell Biol 1998; 143(7): 1775-87.

[34]   Gudas JM, Payton M, Thukral S, *et al*. Cyclin E2, a novel G1 cyclin that binds Cdk2 and is aberrantly expressed in human cancers. Mol Cell Biol 1999; (1): 612-22.

[35]   Weichselbaum RR, Ishwaran H, Yoon T, *et al*. An interferon-related gene signature for DNA damage resistance is a predictive marker for chemotherapy and radiation for breast cancer. Proc Natl Acad Sci USA 2008; 105(47): 18490-5.

[36]   Suomela S, Cao L, Bowcock A, *et al*. Interferon alpha-inducible protein 27 (IFI27) is upregulated in psoriatic skin and certain epithelial cancers. J Invest Dermatol 2004; 122(3): 717-21.

[37]   Sahar DE, Yang GP, Longaker MT, *et al*. Surgical application of cDNA microarray technique. Surgery 2005; 138(3): 399-403.

[38]   Miao J, Wang Z, Provencher H, *et al*. HOXB13 promotes ovarian cancer progression. Proc Natl Acad Sci USA 2007; 104(43): 17093-8.

[39]   Goetz MP, Suman VJ, Ingle JN, *et al*. A two-gene expression ratio of homeobox 13 and interleukin-17B receptor for prediction of

recurrence and survival in women receiving adjuvant tamoxifen. Clin Cancer Res 2006; 12(7 Pt 1): 2080-7.

[40]    Wang Z, Dahiya S, Provencher H, *et al.* The prognostic biomarkers HOXB13, IL17BR, and CHDH are regulated by estrogen in breast cancer. Clin Cancer Res2007; 13(21): 6327-34.

[41]    Jerevall PL, Brommesson S, Strand C, *et al.* Exploring the two-gene ratio in breast cancer--independent roles for HOXB13 and IL17BR in prediction of clinical outcome. Breast Cancer Res Treat 2008; (2): 225-34.

[42]    Jansen MP, Sieuwerts AM, Look MP, *et al.* HOXB13-to-IL17BR expression ratio is related with tumor aggressiveness and response to tamoxifen of recurrent breast cancer: a retrospective study. J Clin Oncol 2007; 25(6): 662-8.

[43]    Lallier TE, Spencer A. Use of microarrays to find novel regulators of periodontal ligament fibroblast differentiation. Cell Tissue Res 2007; 327(1): 93-109.

[44]    Kursula P, Flodin S, Ehn M, *et al.* Structure of the synthetase domain of human CTP synthetase, a target for anticancer therapy. Acta Crystallogr Sect F Struct Biol Cryst Commun 2006; 62(Pt 7): 613-7.

[45]    Dunbar JD, Song HY, Guo D, *et al.* Two-hybrid cloning of a gene encoding TNF receptor-associated protein 2, a protein that interacts with the intracellular domain of the type 1 TNF receptor: identity with subunit 2 of the 26S protease. J Immunol 1997; 158(9): 4252-9.

[46]    Thompson HG, Harris JW, Wold BJ, *et al.* Identification and confirmation of a module of coexpressed genes. Genome Res 2002; 12(10): 1517-22.

[47]    Percy MJ, Myrie KA, Neeley CK, *et al.* Expression and mutational analyses of the human MAD2L1 gene in breast cancer cells. Genes Chromosomes Cancer 2000; 29(4): 356-62.

[48]    Loi S, Sotiriou C, Buyse M, *et al.* Molecular forecasting of breast cancer: time to move forward with clinical testing. J Clin Oncol 2006; 24(4): 721-2; author reply 2-3.

[49]    Hess KR, Anderson K, Symmans WF, *et al.* Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. J Clin Oncol 2006; 24(26): 4236-44.

[50]    Pando MP, Kotraiah V, McGowan K, *et al.* Alternative isoform discrimination by the next generation of expression profiling microarrays. Expert Opin Ther Targets 2006; 10(4): 613-25.

[51]    Fan C, Oh DS, Wessels L, *et al.* Concordance among gene-expression-based predictors for breast cancer. N Engl J Med 2006; 10: 355(6): 560-9.

[52]    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; 15; 98(4): 262-72.

[53]    Ransohoff DF. How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. J Clin Epidemiol 2007; 60(12): 1205-19.

[54]    Biganzoli E, Boracchi P. Old and new markers for breast cancer prognosis: the need for integrated research on quantitative issues. Eur J Cancer 2004; 40(12): 1803-6.

[55]    Hayden JA, Cote P, Steenstra IA, *et al.* Identifying phases of investigation helps planning, appraising, and applying the results of explanatory prognosis studies. J Clin Epidemiol 2008; 61(6): 552-60.

[56]    Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. N Engl J Med 2009; 360(8): 790-800.