Breast cancer classification: linking molecular mechanisms to disease prognosis
Atefeh Taherian-Fard*, Sriganesh Srihari* and Mark A. Ragan
Submitted: 20th February 2014; Received (in revised form): 7th May 2014

Abstract
Breast cancer was traditionally perceived as a single disease; however, recent advances in gene expression and genomic profiling have revealed that breast cancer is in fact a collection of diseases exhibiting distinct anatomical features, responses to treatment and survival outcomes. Consequently, a number of schemes have been proposed for subtyping of breast cancer to bring out the biological and clinically relevant characteristics of the subtypes. Although some of these schemes capture underlying molecular differences, others predict variations in response to treatment and survival patterns. However, despite this diversity in the approaches, it is clear that molecular mechanisms drive clinical outcomes, and therefore an effective scheme should integrate molecular as well as clinical parameters to enable deeper understanding of cancer mechanisms and allow better decision making in the clinic. Here, using a large cohort of \( \sim 550 \) breast tumours from The Cancer Genome Atlas, we systematically evaluate a number of expression-based schemes including at least eight molecular pathways implicated in breast cancer and three prognostic signatures, across a variety of classification scenarios covering molecular characteristics, biomarker status, tumour stages and survival patterns. We observe that a careful combination of these schemes yields better classification results compared with using them individually, thus confirming that molecular mechanisms and clinical outcomes are related and that an effective scheme should therefore integrate both these parameters to enable a deeper understanding of the cancer.

Keywords: breast cancer classification; molecular signature; biomarkers; clinical subtypes; gene expression

INTRODUCTION
With an estimated 1.38 million new cases and 458,000 deaths worldwide every year, breast cancer is the most common malignancy among women in both the developed and developing world [1, 2]. Historically, breast cancer was perceived as a single disease with varying histopathological features and responses to systemic treatment [3]. However, the advent of high-throughput platforms for gene expression profiling [4, 5] and whole-exome sequencing [6–9] have enabled studies that have challenged this view and brought to the fore the concept that breast cancer consists of a collection of different diseases that affect the same organ site but have different risk factors, clinical presentation, histopathological features, survival outcomes and responses to systemic therapies [3, 10, 11]. This heterogeneity poses a severe challenge for accurate diagnosis of patients for optimal dosage and extent of treatment and to estimate risk factors associated with
their disease. For example, a study analysing breast cancer cases in the United States reported between 1976 and 2008 estimated that ∼1.3 million women were over-diagnosed and over-treated as a result of regular mammogram screening during this 30-year period [12]; the long-term side effects of the treatments among survivors may be significant [2]. Such heterogeneity also complicates understanding of the underlying tumour biology—e.g. the genetic, epigenetic and host factors underpinning aberrant signalling pathways—for development of new therapeutic strategies. Consequently, a number of classification schemes have been devised to stratify breast tumours, attempting to comprehend the intricate biological mechanisms driving these tumours and to allow more effective decision making in clinical trials and treatments.

Although breast cancer has provided an excellent test bed for development and testing of classification schemes, their sheer number and lack of concordance have made it difficult to judge their applicability, and therefore standardize them. For example, while some schemes quantify molecular differences between tumours (molecular subtyping), others predict survival outcomes (prognosis) and therapeutic responses (prediction); it is not clear to what extent these molecular differences are related to clinical outcomes, or whether this relationship influences therapeutic strategies and clinical decisions.

Here, we review the different schemes developed for breast cancer classification including anatomical systems, gene expression and genomic signatures and multi-omic integrative models. Using gene expression profiles of ∼550 patients from The Cancer Genome Atlas (TCGA) [6], we evaluate 10 expression-based signatures for molecular subtyping and prognostic ability. We note that molecular characteristics strongly determine clinical outcomes, and therefore an effective predictive model should integrate both these parameters to enable deeper understanding of cancer mechanisms and development of better treatment strategies.

**BREAST CANCER CLASSIFICATION SCHEMES**

Based on the purpose and the type of information (dataset) used, breast cancer classification schemes can be roughly distributed into five classes, though these classes can be overlapping (Figure 1). Histological grading and TNM staging (tumour size, lymph node and metastatic spread based) are based on the physical or anatomical properties of the cancer and quantify its aggressiveness. The expression-based schemes predominantly use microarray or qRT-PCR expression profiles and can be divided into molecular subtyping, and prognostic and predictive gene signatures. The genomic profile-based schemes use mutational profiles of tumour genomes, and with the recent advent in high-throughput sequencing technologies have gained much interest. Network-based methods integrate molecular interactions with expression and mutational profiles (multi-omics integration) to capture network-aggregated properties of breast cancer.

**Histological grading**

Very simply, histological grade is the description of the tumour based on the abnormality of tumour cells relative to normal cells when observed under a microscope. Typically, a numerical grade (1, 2, 3 or 4) is assigned to the tumour depending on the extent of this abnormality. Grade-1 tumour cells appear highly similar to normal cells, and these tumours tend to spread slowly. In contrast, cells of Grade 3 and 4 tumours appear highly dissimilar from normal cells, and these tumours grow rapidly and spread faster than lower-grade tumours.

The histological grading system most widely adopted in breast cancer is the Nottingham system [13, 14], and is included as part of prognostic indices such as the Nottingham Prognostic Index. This index combines tumour grade and lymph node stage of the TNM system (discussed next) for determining the treatment for breast cancer patients in the UK [15, 16].

**Tumour size, lymph node invasion and metastatic spread**

In the TNM system, T appended with a number (0–4) is used to describe the size and location of the tumour: T0—no evidence of tumour; T1—the invasive part of the tumour has size ≤20 mm and is carcinoma in situ, confined within the ducts or lobules of breast tissue; T2—the invasive part of the tumour is 20–50 mm; T3—the invasive part >50 mm; and T4—the tumour has grown into the chest wall and skin with signs of inflammation. Four stages are likewise recognized for lymph node invasion: N0—no cancer cells are found in the lymph nodes; N1—the cancer has spread to three nodes; N2—the cancer has spread to four to nine nodes;
and N3—the cancer has spread to ≥10 nodes. The spread is measured as distant metastasis with the following stages: M0—the cancer has not metastasized; and M1—there is evidence of metastasis to another body part. The cancer is staged by combining these T, N and M classifications. In breast cancer, there are five stages 0–4, of which stage 0 corresponds to non-invasive ductal carcinoma in situ and stages 1 through 4 are used for invasive breast cancer.

**Gene expression-based classification**

Molecular biology studies such as gene expression profiling have shown that response to treatment, and therefore clinical decision making, is not determined by anatomical factors (such as tumour size or lymph node status) *per se*, but rather by intrinsic molecular characteristics of the tumours [3, 10, 11, 17]. Consequently, a number of landmark studies [4, 5, 18–23] uncovered multi-gene expression markers that are independent of classical anatomical markers, from the compendia of genome-wide mRNA profiles of patients. These clinically motivated markers, also called gene signatures, correlate with the molecular characteristics of tumours (molecular subtypes) [4, 5], aggressiveness markers such as proliferation or grade [11, 22], survival outcomes (prognosis) [19–21] and response to therapy [17, 24, 25].

**Molecular or ‘intrinsic’ subtyping**

One of the first applications of microarray-based gene-expression analysis to the study of breast cancer was in the assessment of diversity at a molecular level. Starting with an initial set of 8102 genes from 65 tumour expression samples, 456 ‘intrinsic’ genes (those that varied more in expression between tumours than between repeated samples of the same tumour) were identified that hierarchically clustered the samples based on molecular characteristics [4]. Subsequent validation on an independent dataset of 78 breast cancers confirmed the robustness of this classification [5]. These seminal studies revealed that ER-positive and ER-negative tumours (ER: estrogen receptor) are molecularly different, and the intrinsic genes identified at least four distinct subtypes (luminal, HER2-enriched, basal-like and normal-like). Luminal tumours are mostly ER-positive, and are further classified into luminal-A, which are histologically low-grade, and luminal-B, which express lower levels of hormone receptors and are mostly high-grade. HER2-positive tumours show amplification and over-expression of the ERBB2 gene, and are mostly high grade. On the other hand, basal-like tumours are ER-negative, PR-negative (PR: progesterone receptor) and HER2-negative (hence ‘triple-negative’). These subgroups correspond reasonably well to clinical characterization on the basis of ER and HER2 status, as well as proliferation markers or histological grade [4, 5].

Although these intrinsic subtypes have been adopted to build breast cancer prognostic and therapeutic-response prediction models such as the PAM50 signature [18], the classification is limited by its close correspondence to ER, PR and HER2 status, and analyses have suggested that these do not have sufficient prognostic or predictive value [10].

**First-generation prognostic signatures**

Over the past decade several groups have pursued the development of multi-gene prognostic signatures that classify patients with good prognosis, who hence can forgo chemotherapy, and those with
poor prognosis and metastasis risk. Here, we highlight three widely adopted signatures—MammaPrint, Wang-76 and OncotypeDX; for a comprehensive review readers are referred to [3, 11].

70-gene signature. MammaPrint (Agendia, Amsterdam, Netherlands) was the first successful prognostic gene signature, and is a microarray test approved by the US Food and Drug Administration for prognosis of patients with TNM stage 1 or 2, node-negative, invasive breast cancer of tumour size ≤50 mm. This signature was constructed from an empirical microarray analysis of 78 breast cancers from patients <55 years with node-negative tumours ≤50 mm [19]. A supervised analysis of 25,000 genes from the expression profiles of these patients identified a set of 70 genes that accurately predicted poor prognosis disease (development of distant metastasis within 5 years) on an independent cohort of 295 invasive breast cancers [26]. Subsequent studies confirmed the test’s prognostic potential in node-positive [27] and HER2-positive [28] tumours, and its correlation with chemotherapy sensitivity [29]. However, the discriminatory power of the signature for ER-negative cancers was noted to be very low [30].

76-gene signature. This signature was developed on the basis of supervised analysis of 115 breast cancers, of which 80 were ER-positive, but unlike in MammaPrint, ER-positive and ER-negative cancers were analysed separately [20]. This identified two separate sets of genes to predict poor prognosis, 60 genes for patients with ER-positive disease and 16 genes for ER-negative disease, which were then validated on an independent set of 171 patients. However, subsequent studies showed that this signature had the same limitations as MammaPrint, and the 16-gene signature did not have sufficient power to predict outcome for patients with ER-negative and HER2-positive cancers [31, 32].

OncotypeDX. In parallel with microarray-based signatures, OncotypeDX (Genomic Health, USA) was developed using qRT-PCR-based expression profiles [21], and is widely adopted for clinical practice in the United States. A mathematical function [recurrence score (RS)] in OncotypeDX uses a 21-gene expression profile to predict the risk of distant relapse at 10 years for patients with ER-positive, lymph node-negative cancers. The association between RS and distant relapse was examined retrospectively in 668 patients treated with tamoxifen, and RS predicted 10-year distant recurrence rates as 7, 14 and 30% for the low-risk, intermediate-risk and high-risk categories of patients, respectively [33]. In addition, the association of RS with benefit from adjuvant chemotherapy in ER-positive, node-negative, tamoxifen-treated patients was examined in 651 patients. Higher scores were associated with greater benefit from adjuvant chemotherapy, and more critically, lower scores were associated with a lack of even marginal benefit from chemotherapy [34].

Second-generation prognostic signatures
Analysis of breast tumours from large cohorts has revealed that although many genes, most of which are related to cell cycle and proliferation, predict the outcome of ER-positive cancers, fewer genes predict the outcome of ER-negative cancers, with the number strongly dependent on the dataset analysed [10, 35]. Genes involved in the immune response provide additional prognostic information for ER-negative and highly proliferative ER-positive cancers [22], and this has led to the development of immune response-based signatures [36]. Further, the analysis of genes expressed in the stromal compartment of breast cancers has led to the development of stroma-related prognostic signatures [22, 37].

Predictive signatures
Beyond prognostic classifiers, the challenge is to provide physicians with biomarkers that can predict response (or lack of response) to treatment [38]. OncotypeDX has been shown to be associated with benefit from adjuvant chemotherapy [10, 11, 17]. Many groups have focused on neoadjuvant therapy to estimate chemotherapy sensitivity. For example, a 30-gene signature developed in 82 breast cancer patients receiving neoadjuvant chemotherapy estimated the response of 51 independent patients with higher predictive value than clinical variables such as age, grade and ER status [24, 25]. Another approach to develop multi-gene classifiers of chemotherapy sensitivity is based on ‘metagenes’, that is, groups of co-expressed genes associated with a small number of biological processes. A retrospective microarray analysis of ER-negative breast cancers demonstrated that increased stromal metagene expression predicted resistance to chemotherapy [37].

Despite these promising initial results, signatures of chemotherapy sensitivity have so far seen limited use in clinical practice; only ER and HER2 are currently used as predictive markers (for selecting patients
likely to respond to endocrine therapy and trastuzumab, respectively).

**Pathway-based signatures**

The ability to capture breast cancer heterogeneity based on the activity of oncogenic pathways [39] has been demonstrated for molecular subtyping of breast cancer [40, 41]. The rationale behind these methods is that different oncogenic pathways are dysregulated in cancer subtypes that are fundamentally different in their molecular mechanisms, so by tracking the activities of these pathways it is possible to stratify tumours and predict clinical outcomes [42]. A cohort of 1143 tumours was classified into 17 subgroups using the expression profiles of genes in ER/PR, MYC, RAS, AKT, EGFR/TGFβ, STAT3/TNFα, P53-apoptosis and PI3K pathways, corresponding to the intrinsic subtypes including basal-like (subgroups 2, 5 and 8), luminal-A (subgroups 11 and 17), luminal-B (subgroups 3, 4, 6, 9 and 16) and HER2-enriched (subgroups 7 and 10) subtypes [40]. Different subgroups exhibited distinct patterns of pathway activity. For example, the basal subgroups exhibited low ER/PR and high Myc and Ras activity, whereas the luminal subgroups generally exhibited the reverse pattern. Further, pathway activity identified finer subtypes within the intrinsic subtypes. For example, among the basal subgroups, subtypes 2 and 5 exhibited low EGFR activity, whereas subtype 8 showed high EGFR expression. Validation on an independent dataset of 547 tumours indicated that the predicted subgroups corresponded well to the clinical properties of these tumours.

**Issues with gene expression-based signatures**

A comprehensive study comparing 47 published gene signatures with 1000 randomly generated gene sets strikingly found that most signatures were not more strongly associated with breast cancer outcome than were the random gene sets [23]. In fact, repeated trials showed that 11 (23%) of these signatures exhibited a weaker association than the top 5% of random sets. Similar results have been reported in other studies [43]. This high chance for discordance among the signatures has been attributed to the fact that expression data contain large numbers of highly correlated variables, and therefore different combinations of these variables can be selected to build similarly accurate prediction methods [11, 43]. A consequential limitation is that many of the signatures include genes which have no relevance to tumour biology, and thus seldom yield interesting insights into the mechanism of disease progression.

A meta-analysis [44] of 3000 tumours for recurrence found that most signatures separated the low-proliferative luminal-A tumours at low risk of recurrence, but these signatures were less informative for the ER-negative basal-like and HER2+ tumours; in fact, most signatures assigned a high risk of recurrence to almost all ER-negative tumours. This was because most of the signatures include a substantial proportion of cell-cycle progression genes. Although these genes are unquestionably important, including independent prognostic information (e.g. immune-related genes [44]) is important in the case of ER-negative tumours.

**Classification based on genomic profiling**

Advances in high-throughput whole-genome and -exome sequencing have revealed complex landscapes of cancer genomes, allowing comprehensive investigation into the molecular basis of cancers [6–9, 45, 46]. These studies have revealed a fundamental observation: there is considerable heterogeneity among the tumours of different patients, which could potentially have significant implications on how patients are stratified for clinical trials and treatments. However, much of this inter-patient genomic heterogeneity is because of the background mutations (‘passengers’), while only a limited number of genes (‘drivers’) are responsible for the molecular and clinical differences [45, 47]. By mapping these driver-event frequencies onto the molecular subtypes, we see that luminal-A tumours harbour a high frequency of PI3KCA mutations (45% cases), while luminal-B in TP53 and PI3KCA (29% each). On the other hand, TP53 mutations occur in a majority of basal-like tumours (80%), but are nearly absent in luminal tumours. HER2-enriched tumours show frequent copy-number amplification of HER2 (80%), and mutations in TP53 (72%) and PI3KCA (39%) [6].

Similarly, a study analysing 2000 breast tumours [7] found that copy-number drivers such as MYC, CCND1 and PPP2R2A/B account for significant differences between tumours, and identified up to 10 ‘integrative’ clusters that finely divide the intrinsic subtypes; each of these clusters correlates with distinct clinical outcomes. For example, an ER-positive
cluster corresponding to luminal tumours exhibited poor prognosis, representing a high-risk subgroup within luminal tumours, and this cluster showed amplification in CCND1.

At the time of writing, the COSMIC list (http://cancer.sanger.ac.uk/cancergenome/projects/census/) contains 19 genes implicated in breast cancer, and this list is likely to expand to 35–40 genes as the genes from the above sequencing studies are validated. However, in an analysis of 100 breast cancer genomes [48], tumours initiated from mutations in <10 genes (the minimum being 6) were observed, indicating that different combinations of these 19 driver genes could initiate tumours, and possibly different kinds of tumours, thus explaining in part the genetic basis for heterogeneity of breast cancer.

The question that arises as to which of these combinations give rise to tumours with distinct underlying disease mechanisms and prognosis. A key observation is that all of the driver genes can be classified into a limited number of known signalling pathways, and deregulation of different genes within the same pathway or sub-pathway leads to similar outcomes. This suggests that tumours could be stratified based on these pathways or networks of pathways [7, 45, 49, 50].

Network-based classification
Based on the observation that important aspects of inter-patient or inter-tumour heterogeneity can be summarized into common pathways, we can use molecular networks of protein–protein, functional and pathway interactions to integrate data from large-scale expression and mutation profiling studies [49, 50]. In a recent study [49], mutations were integrated through molecular networks using a ‘network-smoothing’ approach to identify distinct network modules associated with tumour subtypes and thereby stratify patients. Application of this network-based stratification method to ovarian cancer samples identified up to four subtypes that discriminated survival outcomes of patients better than earlier subtypes, based solely on expression data. In [50], a probabilistic approach was used to combine PPI networks, expression and mutation datasets, to track network modules that showed differential behaviour between tumour subtypes, and were related to disease prognosis. These network modules could associate tumour subtypes to disease mechanisms such as functional deficit in DNA damage through mutations in genes such as ATM, BRCA1 and BRCA2, or to those responsible for drug resistance such as fibroblast signalling genes [51]. It will be valuable to see what subtypes in breast cancer this method can identify.

EVALUATION OF EXPRESSION-BASED SIGNATURES
Because disease mechanism can be associated with clinical outcomes including survival times and response to therapy (above), an important goal of cancer classification should be to examine whether prognostic signatures are related to disease mechanism; that is, whether specific molecular markers play a functional role in the pathogenesis of cancer. For gene expression signatures the results have been modest, which is hardly surprising because different subsets of genes are highly correlated with each other and can equally discriminate samples for prognosis. Therefore, much of the functional interpretation of prognostic signatures should be treated with caution. On the other hand, deregulation in different oncogenic pathways has been shown to be responsible for the intrinsic differences between tumours. It remains to be determined if these differences amount to differences in clinical outcomes.

Here, we analyse this relationship between molecular mechanisms and clinical outcomes. We compile a list of oncogenic pathways implicated in breast cancer and widely used prognostic signatures, and evaluate them for their ability to classify breast tumours based on (i) molecular properties and (ii) survival outcomes. These experiments are by no means exhaustive, but reveal valuable insights into the link between molecular and clinical characteristics via these signatures and pathways.

Materials and methods
Our experiments were conducted in two parts, to evaluate these signatures (from now on we refer to all prognostic signatures and biological pathways as just signatures, unless specifically distinguished) for their ability to:

(i) classify breast tumours into known molecular or intrinsic subtypes (basal-like, HER2-enriched, luminal-A and luminal-B), ER-status (ER+ and ER−) and tumour stages (early and advanced); and
(ii) differentiate good and bad disease prognosis in patients.
We downloaded gene expression datasets of 547 breast cancer patients (as of April 2013) from TCGA (http://cancergenome.nih.gov/) [6]. Next, we compiled eight biological pathways implicated in breast cancer, four prognostic signatures and one blood-protein biomarker set (Table 1 and Supplementary Figure S7) from the literature [18–20, 52, 53], MSigDB (http://www.broadinstitute.org/gsea/msigdb/index.jsp) and KEGG databases [54]. For (i), we trained a support vector machine-classifier on each of these signatures and evaluated their classification performance using 10-fold cross-validation, benchmarked against the PAM50 signature used to assign the original labels in TCGA. A signature containing genes with high variability in expression across two or more subtypes is more likely to differentiate between the subtypes, and this in turn translates to higher classification accuracy. For (ii), we evaluated the prognostic ability of these signatures using Kaplan–Meier survival curves plotted using KM-Plotter (www.kmplot.com) and the UCSC Cancer Genomics Browser (https://genome-cancer.ucsc.edu/proj/site/hgHeatmap/). For a detailed description of our computational workflow, evaluation criteria and datasets, refer to Supplementary Materials (Supplementary Figure S1).

### Evaluation for molecular classification of breast tumours

Figure 2 shows that most signatures were able to classify breast tumours with \( \sim 70\% \) accuracy, with Notch, DDR, Cell cycle, PI3K and Ras showing the best performance among the pathways, and MammaPrint and OncotypeDX the best among the prognostic signatures. This is interesting given the fact that most signatures share few genes (Supplementary Figure S7). Most signatures were also able to classify tumours based on ER status with \( \geq 80\% \) accuracy (Supplementary Figure S3). This was despite ER not being a part of all the signatures; this is likely due to the presence of several ER targets (e.g. GATA3) or regulators of ER (e.g. FOXA1) in these signatures. However, most signatures showed a relatively lower performance for tumour-stage classification (Supplementary Figure S4). A cross-labelling experiment (Supplementary Tables S1 and S2) predicting the stage of tumours from different molecular subtypes showed that while 100% of all subtypes were designated as early stage by the prognostic signatures, about 10–17% of tumours (the highest being HER2+ and luminal-B, 17%) were designated as advanced stage by PAM50 and the biological pathways, indicating an extent of disagreement between prognostic signatures and pathways. We hypothesize that dysregulation of oncogenic pathways occurs before manifesting as clinical outcomes, and therefore might reflect the disease in its advanced stages, whereas genes involved in symptomatic changes mostly comprise the prognostic signatures, and are still reflecting early stages.

### Evaluation for prognosis estimation

Figure 3 shows the survival plots (survival percentage versus days to death) for all signatures. As expected, the widely adopted OncotypeDX showed the best differentiating ability between good and bad prognosis diseases (log-rank test \( P < 0.0021 \)), followed by MammaPrint \( (P < 0.0043) \), Wnt \( (P < 0.019) \), Notch \( (P < 0.023) \) and PI3K \( (P < 0.023) \) showed reasonably good prognostic ability among the oncogenic pathways. Interestingly, the blood signature \( (P < 0.0025) \) performed better than the pathways, indicating this could be a first-level, easy-to-use test to determine prognosis of the disease in patients.

### Combining molecular and prognostic signatures

An interesting observation from our experiments is that biological pathways, which are typically not employed to estimate prognosis, show reasonably good
performance in differentiating good and bad survival outcomes, and likewise the prognostic signatures show good performance in molecular subtyping of tumours. There is thus considerable association between molecular and clinical features. In fact, the four molecular subtypes are known to associate with distinct survival outcomes (Supplementary Figure S6), with basal-like and HER2+ tumours characterized by high aggressiveness and low survival, compared with luminal-B and luminal-A tumours.

On this basis, we combined the top-performing pathways and prognostic signatures via a feature-selection approach and assessed the performance of the combined signature vis-à-vis the individual signatures (Supplementary Figure S8). Figure 4 shows that the combined signature, although containing only 12 genes (Table 2), showed a better classification performance than the individual signatures did, and was second only to PAM50. Further, at least four of these 12 genes showed significant differentiation ability in survival curves (Supplementary Figure S9). PI3K/RAS and cell cycle genes are over-represented in this combined signature. PI3K/RAS are frequently over-expressed in aggressive breast tumours, especially basal-like, whereas cell cycle genes are markers of high proliferative rate in tumours. Over-expression of EGFR and of CKS1B is associated with poor survival, and these genes have been explored as molecular targets in breast tumour therapy [55, 56].

However, there are a few caveats. A blind combination of all signatures brings in noise, which considerably reduces the performance of the combined signature. Therefore, the use of a feature selection method, which essentially selects non-redundant genes and hence reduces the noise, is critical to this combination procedure. However, the choice of the feature selection method also matters here. We tested several methods and found Forward Selection (FS) [57] to produce a combined signature that performs better than any of the component signatures individually (Figure 4). This combined signature is not necessarily the best-performing signature in general (Discussion).

Classifying familial breast tumours
The molecular subtypes have been defined predominantly based on expression datasets from sporadic breast tumours, which constitute 93–95% of all breast tumours. These tumours originate during the lifetime of patients and are typically diagnosed in women >40 years. On the other hand, familial breast tumours (the remaining 5–7%) show high risk of predisposition right from birth and typically diagnosed in women <40 years owing to inherited defects in breast cancer risk genes including BRCA1, BRCA2 or ATM. Being rarer and mostly restricted to families, much less is known about the molecular and clinical characteristics of these tumours. Previous studies [58, 59] have noted that BRCA1 and BRCA2 tumours show distinct
Figure 3: Survival plots. Plots of survival percentage versus days to death based on the expression levels of genes in the signatures at $P < 0.05$ from the UCSC Cancer Browser (https://genome-cancer.ucsc.edu). The red curve indicates survival when the genes show higher expression than mean, green when the genes show lower expression, and black when the expression is close to the mean. The permutation $P$-values of the log-rank test between the risk groups (red and green) were: Oncotype DX: 0.0021, MammaPrint: 0.0043, PAM50: 0.0087, Blood: 0.0025, Cell cycle: 0.065, DDR: 0.039, Notch: 0.023, PI3K: 0.023 and Wnt: 0.019.

Many genes from the Wang-76 signature were missing in the dataset. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.
molecular phenotypes, with BRCA1 tumours being predominantly ER-negative and developing into basal-like tumours, and BRCA2 tumours being predominantly ER-positive and developing into luminal-like tumours.

Here, we applied our classifier trained on the molecular subtypes from sporadic breast cancer data (TCGA) to an independent dataset [58] of 19 BRCA1 and 30 BRCA2 familial tumours, using PAM50 and our 12-gene signature (Supplementary Table S4). PAM50 classified about a quarter (26%) of the BRCA1 tumours as basal-like and the remaining (74%) as luminal, but most (96%) BRCA2 tumours as luminal. On the other hand, our 12-gene signature classified all familial tumours as luminal. This is because the signature lacked ER-related genes (e.g. ESR1 and FOXA1) that are present in PAM50 and which are required for distinguishing basal-like (i.e. ER-negative) from luminal (i.e. ER-positive) tumours.

**DISCUSSION**

Over the past 15 years, a significant number of gene expression signatures in cancer have been published. The query ‘cancer gene expression signature’ in PubMed returns >500 results for 2013 alone. The sheer number of such studies makes it challenging to judge the applicability of these signatures. Although signatures such as OncotypeDX and MammaPrint are being used in the clinic, it is still not clear to what extent these signatures add prognostic or predictive value to physical or anatomical characteristics such as age, grade, nodal involvement and tumour size [60]. Further, as Venet et al. [23] and others have shown, a considerable number of these signatures in fact do not correlate to the underlying tumour biology. Even in our analysis, we found that pathways with no direct evidence for involvement in breast cancer, and sets of randomly selected genes, or genes showing no or minimum variation in expression, performed reasonably well (accuracy ~0.70) in molecular subtyping (Supplementary Figure S10). Therefore, identifying a single most-effective signature is challenging, and the functional and retrospective validation of signatures is critical before applying them to patients.

Diagnosis and prediction, although primarily handled by medical practitioners and pathologists, are of multidisciplinary concerns. An effective prediction strategy should integrate anatomical, histopathological, prognostic and molecular parameters, and this should be a continuous process enabled through constant feedback obtained from patient monitoring in the clinic to molecular profiling and molecular subtyping.
bioinformatic analysis (Figure 5). As patients are stratified and treated, any deviation from the expected course of response to the treatment should be captured and effectively traced back to molecular mechanisms, i.e. genes and pathways, driving the unexpected response. Only such an integrative feedback-based approach can ensure deeper understanding of cancer mechanisms and aid the development of effective drugs and treatment for patients.

CONCLUSION

With its highly heterogeneous characteristics, breast cancer provides a challenging test bed for differentiating the different diseases (subtypes) that constitute the cancer. Although multiple classification signatures have been proposed, their sheer number and lack of concordance pose a challenge to judge their applicability. Although it is increasingly clear that intrinsic molecular properties of the disease are associated with clinical outcomes, the functional

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Description</th>
<th>Source signature</th>
<th>Biological function of the encoded protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDC45</td>
<td>Cell division cycle 45</td>
<td>Cell cycle</td>
<td>An essential protein for initiation of DNA replication, which plays an important role in loading the DNA polymerase alpha onto chromatin.</td>
</tr>
<tr>
<td>EGFR</td>
<td>Epidermal growth factor receptor</td>
<td>PI3K</td>
<td>A cell surface protein that binds to epidermal growth factor. This binding induces receptor dimerization and tyrosine autophosphorylation and leads to many cellular responses, including changes in gene expression, cytoskeletal rearrangement, anti-apoptosis and increased cell proliferation.</td>
</tr>
<tr>
<td>GRB7</td>
<td>Growth factor receptor-bound protein 7</td>
<td>OncotypeDX</td>
<td>An adaptor protein that is known to interact with a number of receptor tyrosine kinases and signalling molecules, which plays a role in the integrin signalling pathway and cell migration by binding with focal adhesion kinase (FAK).</td>
</tr>
<tr>
<td>CHRM2</td>
<td>Cholinergic receptor, muscarinic 2</td>
<td>PI3K</td>
<td>Belongs to G protein-coupled receptors family. It binds to acetylcholine and is associated to several cellular responses including adenylate cyclase inhibition, phosphoinositide degeneration and potassium channel mediation.</td>
</tr>
<tr>
<td>SCUBE2</td>
<td>Signal Peptide, CUB and EGF-like domain containing protein 2</td>
<td>OncotypeDX</td>
<td>Has GO annotation related to calcium ion binding activity and has been associated to lung and best cancer.</td>
</tr>
<tr>
<td>FGFR4</td>
<td>Fibroblast growth factor receptor 4</td>
<td>PI3K</td>
<td>Belongs to fibroblast growth factor receptor family; mitogenic signalling molecules that have roles in angiogenesis, wound healing, cell migration, neural outgrowth and embryonic development.</td>
</tr>
<tr>
<td>SHC4</td>
<td>Src Homology 2 Domain-Containing-Transforming Protein</td>
<td>RAS</td>
<td>Has GO annotation related to receptor tyrosine kinase binding and protein domain-specific binding.</td>
</tr>
<tr>
<td>IGFIR</td>
<td>Insulin-like growth factor 1 receptor</td>
<td>RAS/PI3K</td>
<td>Binds insulin-like growth factor with a high affinity. With a tyrosine kinase activity, it is highly over-expressed in most malignant tissues where it functions as an anti-apoptotic agent by enhancing cell survival.</td>
</tr>
<tr>
<td>CKS1B</td>
<td>CDC28 protein kinase regulatory subunit 1B</td>
<td>Cell cycle</td>
<td>Binds to the catalytic subunit of the cyclin-dependent kinases and is essential for their biological function.</td>
</tr>
<tr>
<td>CDKN3</td>
<td>Cyclin-dependent kinase inhibitor 3</td>
<td>Cell cycle</td>
<td>Dephosphorylates and eventually prevents the activation of CDK2 kinase. May play a role in cell cycle regulation.</td>
</tr>
<tr>
<td>LAMC2</td>
<td>Laminin, gamma 2</td>
<td>PI3K</td>
<td>Belongs to an extracellular matrix glycoprotein family. Have been associated to a wide variety of biological processes including cell adhesion, differentiation, migration, signalling, neurite outgrowth and metastasis.</td>
</tr>
<tr>
<td>PRKX</td>
<td>Protein kinase, X-linked</td>
<td>RAS</td>
<td>A serine/threonine protein kinase regulated by and mediating cAMP signalling in cells and has multiple functions in cellular differentiation and epithelial morphogenesis, also involves in angiogenesis through stimulation of endothelial cell proliferation, migration and vascular-like structure formation.</td>
</tr>
</tbody>
</table>

Note: The gene descriptions and biological functions were obtained from GeneCards™ (http://www.genecards.org/).
interpretation of most of these signatures is not clear. Here, we report a simple experiment to evaluate these signatures for their capability to classify breast cancer based on molecular subtypes and disease prognosis, and attempt to understand the association between molecular properties and clinical outcomes. We conclude that an effective prediction model should integrate anatomical, histopathological, molecular and prognostic parameters to enable better drug discovery as well as patient treatment strategies.

SUPPLEMENTARY DATA
Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points
- The heterogeneity of breast cancer poses a challenging test bed for the development of classification techniques.
- Many of the gene expression-based signatures proposed for breast cancer classification show considerable disagreement and lack functional interpretation.
- Biological pathways including DNA-damage response, cell cycle and oncogenic pathways together with a few prognostic signatures such as OncotypeDX and MammaPrint show good performance in classifying breast cancer and estimating clinical outcomes.
- An effective prediction model should combine anatomical, histopathological, molecular and prognostic parameters through a continuous feedback-based approach to enable effective drug discovery and treatment strategies.

ACKNOWLEDGEMENTS
We thank Dr Peter T. Simpson, Dr Stefan Maetschke and Professor Kum Kum Khanna for valuable discussions, and Dr Lachlan Coin for the feature selection software.

FUNDING
Australian National Health and Medical Research Council (NHMRC) (grant no. 1028742 to Dr Peter T. Simpson and M.A.R.).

References
Breast cancer classification


