Computational prognostic indicators for breast cancer

Abstract: Breast cancer remains the leading cause of cancer-related mortality in women. Comprehensive genomics, proteomics, and metabolomics studies are emerging that offer an opportunity to model disease biology, prognosis, and response to specific therapies. Although many biomarkers have been identified through advances in data mining techniques, few have been applied broadly to make patient-specific decisions. Here, we review a selection of breast cancer prognostic indicators and their implications. Our goal is to provide clinicians with a general evaluation of emerging computational methodologies for outcome prediction.

Keywords: computational model, precision prognosis, tumor

Introduction

One in eight women develops breast cancer, the most common cause of malignancy in females. Although the majority of patients now survive for many years after initial diagnosis and therapy, a significant subpopulation remains at risk of metastatic relapse. These women have a median survival time of less than 1.5 years at the time of relapse, with only 10% expected to survive 10 years after diagnosis.

The ability to classify patients into clinically relevant subgroups to allow for precise therapy is urgently needed. Traditional tumor stratification is based largely on morphology, but is relevant in less than one-quarter of invasive breast carcinomas. However, novel computational models are emerging as powerful tools to address this deficiency. Recent work has integrated critical observations in the biology of breast cancer, including gene deletions, translocations, and locus amplification; biomarkers from high-throughput “-omics” technologies such as genomics, proteomics, and metabolomics; and long recognized outcome variables such as tumor size, histologic grade, axillary nodal status, and estrogen receptor (ER) status. We anticipate that these biomarkers will emerge as an effective molecular classification or guide for the determination of prognosis and the development of tailored therapy (reviewed by Gruver et al).

However, much work regarding validation of these approaches is still required. For example, single biomarkers have not proven highly informative in most women. Somatic mutations with the potential ability to act as a biomarker were noted to occur in single genes (e.g., TP53, PIK3CA, and GATA3) in <10% of 825 patients with primary breast cancers. Similarly, only three multi-gene signatures are incorporated into current clinical practice. Effective and clinically applicable prognostic indicators will also require the ability to address the emerging importance of the intra-tumoral heterogeneity of breast cancer.
This review will not address specific clinical prognostic variables as they have been reviewed extensively by others. Instead, we have chosen to evaluate emergent prognostic indicators for breast cancer, focusing on computational methodologies and strategies to identify a molecular indicator (a biomarker or set of biomarkers). We are interested particularly in statistical and computational methodologies based on cutting-edge “omics” technologies, moving from the conventional single oncogene/tumor suppressor gene strategy to one that exploits advances in systems biology. To achieve our goal, we performed a systematic search of all the English language literature regarding computational breast cancer prognosis, using a MEDLINE search for the period from January 2005 to July 2013. This review will begin with the utility of single molecular biomarkers and multi-gene signatures, followed by an evaluation of novel system-biology based analyses that infer information from high-throughput “omics” data. Importantly, we will discuss exciting possibilities in the development of new integrative methodologies.

Identification of prognostic molecular signatures

Examination of genetic variation, genomic association, or transcriptomic alterations in large study cohorts allows the selection of single or multiple prognostic markers that can stratify patients into groups showing distinct outcomes (Table S1). It may also allow the specific tailoring of care. Sample size and biological hypotheses are distinguishing features of prognostic marker studies. In terms of sample size, a practical clinically based prognostic index for patients with metastatic breast cancer was recently proposed by a retrospective analysis of 2,322 patients after primary treatment. In parallel, others have focused on using biological determinants. For example, the Cancer Genome Atlas Network has incorporated the four major breast cancer subtypes with five genetic and epigenetic factors (genomic DNA copy numbers, DNA methylation, exome and messenger RNA expression, microRNA sequencing, and reverse-phase protein expression). The results are promising for understanding the biological underpinning of subtype-specific prognostic indicators. However, there remains a critical gap between an integrative model of “omics”-data and clinical variables. Describing biological pathways underlining clinical biomarkers in a large sample size could provide the opportunity to significantly improve our understanding of the biology and heterogeneity of breast cancers. For instance, the enrichment patterns of gene sets associated with embryonic stem cell (ESC) identity in the expression profiles of various human tumor types is associated with poor outcome in breast cancer.

Single gene prognostic determinants

For specific tumor subtypes, the expression of single critical genes can serve as prognostic indicators (reviewed by Adam Maciejczyk). Germ-line mutations of BRCA1 are extensively used for early detection of familial breast cancer, and are predictive for 15%–20% of women with a family history of breast cancer, and 60%–80% of patients with combined breast and ovarian cancer. Additional prognostic factors include enhanced RAD21 and cohesin expression, which has been associated with resistance to chemotherapy in high-grade luminal, basal, and HER2 breast cancers. Transcription factor muscle segment homeobox 2 (Msx2) expression has been implicated in an increased likelihood of tumor cell death via apoptosis in invasive breast cancers. Enhanced expression of the anterior gradient-2 (AGR2) protein, which occurs in the presence of ER antagonist tamoxifen, confers poor prognosis in ER-positive breast cancers.

The most common computational method to assess these critical biomarkers is the hazard Cox regression model. Given a time after diagnosis for values of the predictor variables, the model produces a survival function for the probability that the binary event of interest (eg, death or survival at the endpoint) occurs. In this context, additional computational models to improve the prediction have been proposed, including Bayesian network analysis evaluating probabilistic relationships among candidate genes and support vector machine methodologies. A review of standard survival analyses and the use of the outstanding Bioconductor tool suite is available at http://cran.r-project.org/web/views/Survival.html.

Current status of multi-gene prognostic determinants

Multi-gene signatures are most often derived from transcriptomic microarray and sequence data. These signatures not only have potential for classification and prognosis, but they can also predict specific tumor sub-phenotypes such as resistance to radiation. Generally, these multi-gene models can classify patients into subgroups with either distinct outcomes or diverse treatment responses in an unsupervised manner. As an example, the DNA content of breast adenocarcinomas can be classified as either stable, conferring good prognosis, or unstable, conferring poor prognosis.
To identify multi-gene signatures or pathways, computational pattern learning algorithms have been successfully applied to transcript or mass spectrometry profiles. These algorithms can be roughly categorized into three groups: 1) unsupervised data mining, eg, hierarchical clustering,\(^7\) topographic projection,\(^14\) and other methodologies; 2) supervised classifiers, eg, decision trees,\(^35\) and Prediction Analysis of Microarray (PAM),\(^9\) and 3) semi-supervised learning models.\(^36\) In supervised algorithms, sample labels such as good- or poor-outcome are required to train the model before making decisions. In contrast, unsupervised algorithms are data-driven. In a semi-supervised model, the algorithm makes decisions based on both the raw data and the input labels, allowing partial labeling. Which method to choose depends on study purpose – for example, if a certain phenotype is known to be an important factor in identifying the multi-gene signature, one should select supervised methodologies to make use of that information.

Clinical application of prognostic determinants

In terms of single determinants, the expression levels of four genes (ER, PR, HER2, and Ki67), and combinations thereof, have been shown to have a strong prognostic impact (reviewed by Gökmen-Polar et al\(^37\)). Missense mutations, eg, within the tumor suppressor gene p53 or increased levels of urokinase-type plasminogen activator (uPA) and/or plasminogen activator inhibitor-1 (PAI-1), which indicates poor clinical outcome, have been included by the American Society of Clinical Oncology 2007 Update of Recommendations for the Use of Tumor Markers in Breast Cancer.\(^38\) Other single-gene determinants with somatic mutations are sensitive to targeted therapy; for instance, active ESR1 mutations in ER-positive metastatic breast cancer.\(^39\)

Several multi-gene–based commercial prognostic testing methodologies are now available. The MammaPrint\(^40\) 70-gene signature assay (Agenda, Inc., Irvine, CA, USA) stratifies patients’ outcomes and allows personalized therapeutic prediction.\(^7\)\(^40\)\(^41\) Two additional clinical decision-making assays, the Oncotype DX 21-gene assay\(^42\) (Genomic Health, Inc. (Redwood City, CA, USA) and a clinically updated version of the intrinsic subtype PAM50 assay\(^43\) (ARUP Laboratories, Salt Lake City, UT, USA), have been compared. Predicting prognostic intrinsic subtypes among 151 patients, investigators observed good agreement between the 21-gene and PAM50 assays for high and low prognostic risk assignment.\(^44\)

Overall, molecular prognostic determinants have been successfully developed as adjuvant tools for innovative diagnostic, prognostic, and therapeutic approaches. As shown by Albain et al,\(^45\) changes in treatment decision after review of multi-gene signatures occurred in 30% of individuals predominantly from chemotherapy plus endocrine therapy to endocrine therapy alone with an associated diminution in treatment-related toxicity.

However, there are several limitations in the utility of either single gene or multi-gene prognostic signatures in routine practice, including requirement of fresh or frozen tissues to measure uPA/PAI-1 and elaborate measurement compared with a gold standard or “housekeeping genes” for single gene determinants. Another limitation for both single and multi-gene signatures is that they are positive in only a specific subtype of breast cancer. For instance, the 70-gene assay can only identify potential chemotherapy benefits in high-risk patients.\(^45\) In another study, the discordance rate was approximately 30% between the clinicopathologic risk categories given by the 70-gene assay and the 21-gene assay.\(^45\) Both assays are particularly limited in assigning high-risk status to ER-negative patients.\(^46\)

How do we overcome these challenges?

Key issues with current multi-gene signature methodologies include specificity of prediction and standardization across diagnostic platforms. Li et al have reported that many randomly selected genes show predictive power for cancer prognosis in one dataset, while losing predictive power in other datasets.\(^7\) Therefore, an evaluation using independent datasets is necessary to assess a prognostic model or indicator. Recently, Venet et al\(^28\) compared 47 published breast cancer outcome signatures to mimic signatures made of random genes. Surprisingly, they found that 60% of reported gene signatures of identical size were equally predictive of “mimic” gene signatures. Of particular concern, 23% of the signatures were worse than random, and more than 90% of mimic signatures with 100 genes were significant outcome predictors.\(^28\) Thus, the deficiencies associated with the reproducibility of multi-gene–based prediction prevent effective usage at this time.

The complexity and discrepancy of single- or multi-gene signatures also precludes easy extraction of biologically and therapeutically relevant information. A clue may be provided by observations that suggest that although genetic alterations between patients differ, they frequently involve common pathways. It is therefore critical to identify relevant pathways involved in breast cancer progression and detect...
the corresponding indicators that are prognostic in different tumor subtypes or patients.

**Novel strategies for prognostic indicator development**

Incorporation of network or pathway information into prognostic biomarker discovery could significantly improve prediction performance. For this purpose, investigators have used a large number of supervised machine-learning approaches (ie, inferring knowledge from data with labeled samples) to take advantage of prior knowledge. For example, data mining on gene expression profiling reveals an association between gene expression, copy number, genetic markers, gene expression, clinical and text mining, and pathway. Gene expression, experiment-defined gene-set pairs, thus tolerating the diverse noise and differences observed from multiple technologies and laboratories. This semi-supervised approach successfully employed unlabeled gene expression data and achieved significant performance in gene expression-based outcome prediction for cancer patients. Identifies prognostic markers that can be used in combination to stratify breast cancer patients into groups of different risk ranks with high accuracy. Correlation between two biomarkers is a more useful prognostic factor than their individual expressions. Uses modern machine-learning method SVM to train six clinical prognostic variables (histological grade, tumor size, number of metastatic lymph nodes, estrogen receptors, lymphovascular invasion, local invasion of tumor, and number of tumors) into a prognostic model. A literature-proposed predictive gene list for breast cancer is benchmarked against a separate gene list to construct nonlinear topographic projection maps for prognosis. Cox proportional-hazards regression modeling, gene-set enrichment analysis. Based on a careful gene-set enrichment analysis, multiple gene-set signatures stratify samples into prognostic subgroups. Bayesian probability was employed in neural networks to model censored data.

| Table 1 Methods to identify prognostic determinants based on gene expression in breast cancer |
|----------------------------------------|----------------------------------------|--------------------------------------------------|
| Data source                            | Method                                | Description                                      |
| Genome-wide gene expression            | RXA-GSP<sup>33</sup>                   | Summarizes the individualized relative expression between biological experiment-defined gene-set pairs, thus tolerating the diverse noise and differences observed from multiple technologies and laboratories. |
| LDS<sup>36</sup>                       |                                 | This semi-supervised approach successfully employed unlabeled gene expression data and achieved significant performance in gene expression-based outcome prediction for cancer patients. |
| MSS<sup>37</sup>                       |                                 | Identifies prognostic markers that can be used in combination to stratify breast cancer patients into groups of different risk ranks with high accuracy. |
| Correlation<sup>36</sup>               |                                 | Correlation between two biomarkers is a more useful prognostic factor than their individual expressions. |
| BCRSVM<sup>33</sup>                    |                                 | Uses modern machine-learning method SVM to train six clinical prognostic variables (histological grade, tumor size, number of metastatic lymph nodes, estrogen receptors, lymphovascular invasion, local invasion of tumor, and number of tumors) into a prognostic model. |
| PGL<sup>44</sup>                       |                                 | A literature-proposed predictive gene list for breast cancer is benchmarked against a separate gene list to construct nonlinear topographic projection maps for prognosis. |
| PAM<sup>9</sup>                        |                                 | PAM together with other conventional methods was used to define gene expression-based “intrinsic” subtypes that showed prognosis. |
| Cox proportional-hazards regression modeling, gene-set enrichment analysis<sup>17</sup> |                                 | Based on a careful gene-set enrichment analysis, multiple gene-set signatures stratify samples into prognostic subgroups. |
| Bayesian network analysis<sup>24</sup> |                                 | Bayesian probability was employed in neural networks to model censored data. |
| Gene expression, experiment-based gene signatures | Expression levels relative to a baseline condition, hierarchical clustering, “leave-one-out” cross-validation<sup>7</sup> | Top genes were selected to distinguish subtypes of breast cancers that show prognosis. |
| Gene expression, text mining           | eScience–Bayesian<sup>39</sup>       | Permits coherent integration of prior information and multiple data sources, such as gene expression and information derived from literature. |
| Gene expression, clinical and genetic markers | I-RELIEF<sup>17</sup>, an iterative method based on the feature selection algorithm called RELIEF | Integrated clinical variables with gene expression or biological pathway. |
| Gene expression, copy number           | iCluster<sup>38</sup>               | A likelihood-based, joint latent variable model for integrative clustering samples. |
| Gene expression, copy number, pathway  | PARADiGM<sup>36,49</sup>            | Integrates copy number, mRNA expression, and pathway interaction data into a personalized pathway-by-sample matrix that clusters patients into distinct prognostic subgroups. |

**Abbreviations:** BCRSVM, breast cancer recurrence prediction based on SVM; LDS, low density separation; MSS, multiple survival screening; PAM, prediction analysis of microarray; PARADiGM, pathway recognition algorithm using data integration on genomic models; PGL, predictive gene lists; RXA-GSP, relative expression analysis of gene set pair; SVM, support vector machine.
threshold value of 1 to dichotomize patients into prognostic groups. Pitroda et al piloted a tumor endothelium-derived inflammatory signature consisting of six genes that were associated with poor outcome in multiple cancers. Similarly, in multiple adult tumors, we found that histologically poorly differentiated tumors display an ESC-associated expression imbalance, ie, preferential overexpression of targets of ESC-associated transcription factors (NANOG, OCT, SOX2, and MYC) combined with underexpression of Polycomb-regulated genes. This finding is in agreement with the known significance of survival difference between patients with the expression pattern of these ESC-like signatures and other patients. More recently, the coactivation of the ESC marker MYC and oncogene HER2 was associated with an acquisition of a self-renewal phenotype that is associated with poor outcome.

However, the lack of standards for measurement across various studies, assays, and diagnostic platforms makes it difficult to translate individual findings into clear clinical applications. A popular solution is to compare the median scores of a calculated indicator in a measured cohort, whereas adding new patients changes the estimation of median score and thus the final decision.

Relative expression analysis of gene-set pairs (RXA-GSP), is a novel methodology that we have pioneered to translate prognostic gene targets for individualized treatment planning. RXA-GSP is built on three principles: 1) each individual has both favorable and unfavorable prognostic factors for breast cancer. It is the imbalance that determines the individual outcome (Figure 1). For example, NANOG expression is an indicator of poor prognosis stimulating the growth and metastasis of breast cancer cells, whereas KLF4 is a favorable prognostic indicator inhibiting these processes. 2) The correlation between two factors is, at least in some cases, more significant than the overall expression of either. For example, SIRT1 and DBC1 are both overexpressed in breast tumor tissue, but the correlation between their levels of expression is diminished. 3) We employ a hypothesis-based and experimentally derived approach to the identification of candidates.

Using RXA-GSP, a resultant indicator is computed as the optimal combination of candidate genes that was trained from several large cohorts, and validated in independent cohorts. Considering the heterogeneity of breast cancer in each individual patient along with the relative expression of gene-set pairs and the correlations between them, RXA-GSP allows the computationally driven derivation of individualized prognostic indicators, bridging cancer biology with the clinic through gene expression analysis.

**Next generation integrative prognostic methods**

To date, the majority of studies focused on prognosis have integrated clinical variables with gene expression or biological pathway data. However, attractive opportunities are now available to integrate comprehensive genotype, metabolomics, and phenotype data. Success should facilitate robust prognostic and therapeutic prediction for each patient. Indeed, such success in this endeavor will have significant impact across biomedicine.

However, this potential is limited by the challenges of the sheer size and complexity of high-throughput data resources, often resulting in significant imprecision in data usage. Classical statistical models are frequently insufficient for integrating “omics”-data with other resources, so specific problems are dealt with on a case-by-case basis due to the lack of a coherent overarching method of analysis. Indeed, we and others, have proposed the development of integrated breast cancer prognostic indicators to provide insights into significant variables.

Eklund et al have addressed this problem through an eScience–Bayes approach, complementing a Bayesian probability models’ ability to incorporate gene expression with text mining, modeling highly complex problems with high performance computing. Applying this model to several sets of independent gene expression data results in consistently accurate prediction of breast cancer metastases across cohorts. Though such methods show promise for tailoring specific models to complex problems with high accuracy, their application is cumbersome, as they require manual selection of genes for which we have prior knowledge. This deficiency is related to a lack of a standard for displaying

**Figure 1** Illustration of RXA-GSP method.

**Notes:** This prognostic indicator is the ratio between scores (eg, expression values) of poor prognostic markers versus that of good prognostic markers. It has the ability to integrate different scales of data, bridging cancer biology with the clinic by employing both hypothesis-based and experimentally derived gene-set selection.

**Abbreviation:** RXA-GSP, relative expression analysis with gene-set pairs.
scientific text in formats that are computationally readable. Another obvious weakness is that a microarray represents a single snapshot of the patient.

New avenues of patient-specific tumor phenotyping are also facilitating prognosis prediction. These methodologies include the evaluation of risk associated with inherited genetic variation or genome-wide association studies (GWAS). Similarly, epigenetic changes in breast cancer cells (such as modifications in histone acetylation and DNA methylation at specific gene regulatory elements) have been implicated in breast oncogenesis. These observations, coupled with the emerging use of epigenetic-modifying agents suggest an era of genomics-based therapy selection. For example, stromal/microenvironmental elements affect the prognosis, modifying tumor invasion and/or drug targeting.

Thus, the effectiveness of interpretations of multi-gene assays likely depends on adjustment for other factors. A breakthrough work, Pathway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM), integrates copy number, mRNA expression, and pathway interaction data to a personalized pathway-by-sample matrix that clusters patients into distinct prognostic subgroups. This integration algorithm has been recently applied to depict comprehensive molecular portraits of human breast tumors.

In summary, we believe biological hypothesis-driven computational models that integrate current significant variables in multiple facets, including transcriptomic alterations, chromatin modification, and stroma response, will provide significant insight not only into the biology of breast cancer, but also into determining the prognosis of individual patients.

Conclusion
Challenges remain for breast cancer prognosis due to its complexity and the lack of standardization among models and “omics”-datasets. Specifically, what impedes the convenience and simplicity necessary for clinical application is the lack of precision molecular indicators that help individualized therapeutic decisions. Ideally, a prognostic indicator performs in a specific, sensitive, inexpensive, and easy manner. We expect an integrative computational model of “omics”-data and clinical variables to significantly improve our understanding of the biology and heterogeneity of breast cancers. This is likely to be achieved through either enlarging the sample size or characterizing a specific subtype with given phenotypes. The latter has been recently recognized, though controversial, hypothesizing a collection of genuinely variable malignancies that happen to originate from breast epithelium.

In conclusion, a prognostic indicator should clearly delineate the risk group for an individual, in terms of tumor growth, invasion, and metastatic potential, and/or the likelihood of response to a given therapeutic modality.

Acknowledgment
This research was supported by NIH grant R21 CA167305-01A1. We appreciate Zakir Gowani for his careful proofreading.

Disclosure
The authors report no conflicts of interest in this work.

References


Supplementary material

Table S1 Selected instances from literature search

<table>
<thead>
<tr>
<th>1st author</th>
<th>Year</th>
<th>Journal</th>
<th>Pubmed ID</th>
<th>Cancer biology underpinning</th>
<th>Related clinical indicator</th>
<th>Major contribution/conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>van ’t Veer⁴</td>
<td>2002</td>
<td>Nature</td>
<td>11823860</td>
<td>Gene expression patterns</td>
<td>Lymph node status</td>
<td>Provides a powerful tool to tailor adjuvant systemic treatment that could greatly reduce the cost of BC treatment.</td>
</tr>
<tr>
<td>Rennstam²</td>
<td>2003</td>
<td>Cancer Research</td>
<td>14695203</td>
<td>Chromosomal copy number aberrations</td>
<td></td>
<td>Patterns of copy number gains and losses define BCs with distinct clinicopathological features and patient prognosis.</td>
</tr>
<tr>
<td>Paik⁴</td>
<td>2004</td>
<td>New England Journal of Medicine</td>
<td>15591335</td>
<td>Gene expression patterns</td>
<td>Node-negative, ER-positive, tamoxifen treatment</td>
<td>A novel recurrence score based on 21 genes to quantify the likelihood of distant recurrence in patients as well as overall survival time.</td>
</tr>
<tr>
<td>Kronenwett⁴</td>
<td>2006</td>
<td>Cancer Epidemiology, Biomarkers and Prevention</td>
<td>16985023</td>
<td>Genomic stability</td>
<td></td>
<td>Objective classification of BCs into stable and unstable subtypes that are a prognostic indicator independent of established clinical factors.</td>
</tr>
<tr>
<td>Suh⁶</td>
<td>2007</td>
<td>Clinical Cancer Research</td>
<td>17200346</td>
<td>CLUC4 (chloride intracellular channel 4)</td>
<td></td>
<td>Reactivation and restoration of CLUC4 in tumor cells or the converse in tumor stromal cells could provide a novel approach to inhibit tumor growth.</td>
</tr>
<tr>
<td>Conlin³</td>
<td>2007</td>
<td>Molecular Diagnosis and Therapy</td>
<td>18078353</td>
<td>Oncotype DX recurrence score assay</td>
<td>Lymph node negative, ER-expressing BC</td>
<td>The Oncotype DX assay and others aim to help improve risk classification and recurrence prediction and optimize selection of patients for adjuvant chemotherapy.</td>
</tr>
<tr>
<td>Rodriguez⁸</td>
<td>2008</td>
<td>Carcinogenesis</td>
<td>18499701</td>
<td>HOXB13 (homeobox B13) and IL17BR (interleukin 17 receptor B)</td>
<td>Estrogen signaling</td>
<td>Hypermethylation of HOXB13 is a later event of tumor progression and a prognostic indicator of advanced BC.</td>
</tr>
<tr>
<td>Wei⁸</td>
<td>2008</td>
<td>Molecular Carcinogenesis</td>
<td>18176935</td>
<td>H3K27me3</td>
<td></td>
<td>Loss of H3K27me3 is a predictor of poor outcome in BCs.</td>
</tr>
<tr>
<td>Kim⁹</td>
<td>2008</td>
<td>Annals of Oncology</td>
<td>17956886</td>
<td>CDKs</td>
<td>Patients recruited for study underwent mastectomy or breast-conserving surgery</td>
<td>Tumors with high CDK1SA and high CDK2SA showed significantly poorer 5-year relapse-free survival than those with low CDK1SA and low CDK2SA, respectively.</td>
</tr>
<tr>
<td>Han¹¹</td>
<td>2008</td>
<td>Nature</td>
<td>18337816</td>
<td>SATB1 (SATB homeobox 1)</td>
<td></td>
<td>SATB1 is a genome organizer that tethers multiple genomic loci and recruits chromatin-remodeling enzymes to regulate chromatin structure and gene expression.</td>
</tr>
<tr>
<td>Ben-Porath¹²</td>
<td>2008</td>
<td>Nature Genetics</td>
<td>18443585</td>
<td>Stem cell genetic expression signatures</td>
<td></td>
<td>Detailed characterization of the stem-cell regulatory networks active in cancer is likely to yield powerful diagnostic and prognostic markers.</td>
</tr>
<tr>
<td>Parker¹¹</td>
<td>2009</td>
<td>Journal of Clinical Oncology</td>
<td>19204204</td>
<td>A 50-gene set (PAM50)</td>
<td>“Intrinsic” subtypes, pathologic staging, histologic grade</td>
<td>The intrinsic subtype and risk predictors based on the PAM50 gene set adds significant prognostic and predictive value.</td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>1st author</th>
<th>Year</th>
<th>Journal</th>
<th>Pubmed ID</th>
<th>Cancer biology underpinning</th>
<th>Related clinical indicator</th>
<th>Major contribution/conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sung</td>
<td>2010</td>
<td>Cancer Science</td>
<td>20412117</td>
<td>SIRT1 (sirtuin 1) and CCAR2 (also known as DBC1, deleted in breast cancer 1, KIAA1967)</td>
<td>Luminal subtype, ER and PR expressions</td>
<td>Correlation between SIRT1 and DBC1 is a more useful prognostic factor than their individual expressions. Correlation between the two is decreased in tumor cells.</td>
</tr>
<tr>
<td>Gevensleben</td>
<td>2010</td>
<td>International Journal of Molecular Medicine</td>
<td>21042777</td>
<td>70-gene expression profile MammaPrint®</td>
<td>Size, age, histological grade, hormone receptor status, peritumoral vascular invasion and HER2 status</td>
<td>Gene signature MammaPrint® is shown to provide additional independent prognostic information.</td>
</tr>
<tr>
<td>Lanigan</td>
<td>2010</td>
<td>Breast Cancer Research</td>
<td>20682066</td>
<td>Msx2 (msx homeobox 2)</td>
<td></td>
<td>Msx2 expression results in improved outcome for BCs, possibly by increasing the likelihood of tumor cell death by apoptosis.</td>
</tr>
<tr>
<td>Creighton</td>
<td>2010</td>
<td>Breast Cancer Research</td>
<td>20569503</td>
<td>PI3K pathway</td>
<td>Luminal ER⁺ breast tumors</td>
<td>Luminal B tumors have hyperactive GFR/PI3K signaling associated with lower ER levels, which has been correlated with resistance to endocrine therapy. Targeting PI3K in these tumors may reverse loss of ER expression and signaling and restore hormonal sensitivity.</td>
</tr>
<tr>
<td>Xu</td>
<td>2011</td>
<td>Breast Cancer Research</td>
<td>21255398</td>
<td>RAD21 (RAD21 homolog [S. pombe])</td>
<td>N/A</td>
<td>RAD21 expression confers poor prognosis and resistance to chemotherapy in high-grade luminal, basal, and HER2 BCs.</td>
</tr>
<tr>
<td>Littlepage</td>
<td>2012</td>
<td>Cancer Discovery</td>
<td>22728437</td>
<td>ZNF217 (zinc finger protein 217)</td>
<td>Amplification of the human chromosomal region 20q13</td>
<td>ZNF217 (amplified in numerous cancers) is a poor prognostic indicator and therapeutic target in patients with BC and may be a strong biomarker of triciribine treatment efficacy in patients.</td>
</tr>
<tr>
<td>Kim</td>
<td>2012</td>
<td>Journal of Breast Cancer</td>
<td>22807942</td>
<td>Gene expression patterns</td>
<td>Histological grade, size, number of metastatic lymph nodes, ER, lymphovascular invasion, local invasion of tumor, and number of tumors</td>
<td>As the selected prognostic factors can be easily obtained in clinical practice, the proposed model might prove useful in the prediction of BR recurrence.</td>
</tr>
<tr>
<td>Faryna</td>
<td>2012</td>
<td>FASEB Journal</td>
<td>22930747</td>
<td>Aberrant DNA methylation</td>
<td>Low-grade ER- and/or PR-positive BC</td>
<td>Early methylation changes are frequent in the low-grade pathway of BC and may be useful in the development of prognostic markers.</td>
</tr>
<tr>
<td>Fasching</td>
<td>2012</td>
<td>Human Molecular Genetics</td>
<td>22532573</td>
<td>TOX3 (TOX high mobility group box family member 3)</td>
<td></td>
<td>With the exception of rs3803662 (TOX3), there was no evidence that any of the SNPs associated with BC susceptibility were associated with BC survival.</td>
</tr>
<tr>
<td>Huang</td>
<td>2013</td>
<td>Cell and Bioscience</td>
<td>23497677</td>
<td>MACC1 (metastasis associated in colon cancer 1)</td>
<td>Clinicopathologic features</td>
<td>The over expression of MACC1 in BR is significantly correlated with adverse clinicopathological features, including metastasis and patient survival.</td>
</tr>
</tbody>
</table>
Table S1 (Continued)

<table>
<thead>
<tr>
<th>1st author</th>
<th>Year</th>
<th>Journal</th>
<th>Pubmed ID</th>
<th>Cancer biology underpinning</th>
<th>Related clinical indicator</th>
<th>Major contribution/conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang⁵⁵</td>
<td>2013</td>
<td>PLoS One</td>
<td>23441166</td>
<td>Mediation of transcription factor GRHL2 (grainyhead-like 2 [Drosophila]) on its targets is prognostic in BC</td>
<td>Histological grade</td>
<td>Proposed the RXA-GSP (relative expression analysis with gene-set pairs) method, shows promise as both a valid prediction model as well as high potential for clinical utility.</td>
</tr>
<tr>
<td>Nagata⁵⁶</td>
<td>2014</td>
<td>Breast Cancer</td>
<td>22528804</td>
<td>Induced pluripotent stem cell inducing factors</td>
<td></td>
<td>Strong expression of NANOG is an indicator of poor prognosis for BC patients, whereas KLF4 is a favorable prognostic indicator.</td>
</tr>
</tbody>
</table>

Note: Each key word (genomic, transcriptional, epigenetic, sequence, novel) respectively together with “breast cancer” and “prognostic indicator” was searched in PubMed, from Jan 2005–July 2013.

Abbreviations: BC, breast cancer; eR, estrogen receptor; PAM, prediction analysis of microarray; RXA-GSP, relative expression analysis of gene set pair; SNPs, single-nucleotide polymorphisms.

References
