6

Breast Cancer Prognostication and Risk Prediction in the Post-Genomic Era

Xi Zhao[1,2], Ole Christian Lingjærde[3,4] and Anne-Lise Børresen-Dale[1, 2]
[1]*Department of Genetics, Institute for Cancer Research, Oslo University Hospital, The Norwegian Radium Hospital, Montebello,*
[2]*Institute of Clinical Medicine, University of Oslo,*
[3]*Biomedical Research Group, Department of Informatics, Faculty of Mathematics and Natural Sciences, University of Oslo,*
[4]*Center for Cancer Biomedicine, University of Oslo, Norway*

1. Introduction

Today, breast cancer is appreciated as a group of molecularly distinct neoplastic disorders. Breast tumors are highly heterogeneous in pathology with respect to cell type and tissue origin. With the traditional diagnostic tools, patients with the same clinico-pathological parameters can have markedly different clinical courses. Individual tumors can frequently exhibit heterogeneous patterns of somatic mutations (Bamford et al. 2004, Stephens et al. 2009, Russnes et al. 2010) gene amplifications and deletions (Russnes et al. 2010), epigenetic profiles (Rønneberg et al. 2010), and gene expression portraits (Perou et al. 2000). Efforts to significantly impact cancer patient outcomes will require the development of robust strategies to subdivide such heterogeneous panels of cancers into biologically and clinically homogenous subgroups, for the purposes of personalizing treatment protocols and identifying optimal drug targets.

In this chapter, by reviewing published, as well as unpublished work, we outline the application of microarray expression profiling in breast cancer risk assessment; highlight the strategies of developing molecular classifiers and integrative strategies to improve risk stratification for breast cancer patients. We also discuss the limitations of the "first-generation" expression profiling as well as further methodologies.

2. Prognostication and risk prediction in breast cancer

Breast cancer is indeed a heterogeneous disease with large variation in clinical behavior. There exist a variety of prognostic factors associated with patient survival (such as susceptibility to metastasize) and some predictive markers (which can aid selection of relevant systemic therapy) for management of the breast cancer patient.

2.1 Clinico-pathological prognostic and predictive markers

Traditionally, the treatment decision for breast cancer patients is largely based on a number of histo-pathological features including tumor size, axillary lymph node status, histological

grade, TNM staging (Tumor size, regional lymph Nodes, distant Metastasis) and receptor status.

Tumor size

The size of a tumor is an established prognosis marker used in the clinic (Koscielny et al. 1984, Rosen et al. 1989, Carter et al. 1989, Page 1991). Tumors under 2 cm (T1) in diameter have a low risk of metastasis; tumors of 2-5 cm (T2) have a high risk of metastasis; tumors over 5 cm (T3) have a very high risk of metastasis. Tumor size carries independent prognosis value; both the axillary lymph node status and histological grade are related to tumor size (Rosen et al. 1989, Carter et al. 1989, Weigelt et al. 2005).

Axillary lymph node status

The axillary lymph node status (Carter et al. 1989, Rosen et al. 1989, Page 1991) is another established marker that has been used in clinic setting to characterize the risk of developing metastatic breast cancer. The presence of cancer cells in the lymph nodes increases the risk of metastatic breast cancer. The presence of over four lymph-node metastases is associated with very high metastatic risk.

Histological grade

Histological grade (Scarff et al. 1968) is a well-known histo-pathological parameter routinely used in the clinic to describe the similarity of breast cancer cells to normal breast tissue, and classify the cancer into well differentiated (low grade: Histological grade 1), moderately differentiated (intermediate grade: Histological grade 2), and poorly differentiated (high grade: Histological grade 3), reflecting progressively less normal appearing cells that have a worsening prognosis.

TNM staging

The TNM classification of malignant tumors (TNM) uses the size of the primary tumor (T), its nodal involvement (N), and the presence of distant metastases (M) to classify the progression of cancer into stage I to stage IV. Breast cancers classified as stage I are small and localized tumors, generally have good prognosis, while stage IV tumors are the most advanced and metastatic with poor prognosis. The staging system classifies breast tumors into groups with different prognosis profiles. Carcinoma *in situ* is indicated as stage 0 in the TNM classification. The stage of a cancer is one of the most important factors in determining prognosis and treatment options.

Receptor status

Protein expression of three receptors in breast cancer cells are routinely used in the clinic: estrogen receptor (ER), progesterone receptor (PgR) and Human Epidermal growth factor Receptor 2 (HER2; also known as HER2/neu, ErbB-2 or ERBB2). When treated with tamoxifen, breast cancer patients with tumors that are ER+ and/or PR+ have lower risks of mortality after their diagnosis compared to women with ER- and/or PgR-negative disease (Fisher et al. 1988). Determination the presence of the estrogen receptor is critical for the selection of the patients who could benefit form endocrine treatment (e.g. tamoxifen). Immunohistochemical (IHC) analysis is widely used to measure ER and PgR protein expression. HER2 is a protein involved in regulation of cellular growth giving higher

aggressiveness in breast cancers. HER2+ breast cancer had a worse prognosis (Slamon et al. 1989, Sotiriou and Pusztai 2009). Cells with none of these receptors are called triple negative. This type of breast cancer is clinically characterized as more aggressive and less responsive to standard treatment and associated with poorer overall patient survival (Dent et al. 2007, Chustecka 2007).

Some of the above traditional variables are combined into prognostic models (such as Adjuvant Online! and Nottingham Prognostic Index) for treatment decision-making about adjuvant systemic treatment of patients with early breast cancer.

Adjuvant Online! model

Adjuvant! Online (Ravdin et al. 2001) is a computer based model using patient age, comorbidity level, ER status, tumor grade, tumor size and number of positive lymph nodes to predict breast cancer specific mortality and recurrence risk, as well as the benefit of adjuvant therapy for women with early-stage breast cancer. Because Adjuvant! was directly derived from mortality data and because details of local therapy (surgery and initial radiation) can strongly influence local relapse rates more so than mortality, Adjuvant!'s estimates of mortality are more firmly based than those for relapse. Breast cancer outcome estimates made by Adjuvant! are for "patients who have unilateral, unicentric, invasive adenocarcinoma of the breast, who have undergone definitive primary breast surgery and axillary node staging, and who have no evidence of metastatic or known residual disease; no evidence of T4 features (extension to skin or chest wall); no evidence of inflammatory breast cancer. If they have had breast conserving therapy there should be plans for them to receive radiation therapy. They should not yet have received systemic therapy (neoadjuvant therapy), or radiation prior to their surgical staging." (Adjuvant! Breast Cancer Help Files http://www.adjuvantonline.com/breastnew.jsp Accessed on December 15, 2011).

Nottingham Prognostic Index

The Nottingham prognostic index (NPI) (Haybittle et al. 1982) is used to determine prognosis following surgery for breast cancer by integrating the size of the lesion; the number of involved lymph nodes; and the grade of the tumor. A prognostic index < 3.4 implies a good prognosis, in the range of [3.4, 5.4] a moderately good prognosis and > 5.4 a poor prognosis. It was established by the long-term follow-up in a dedicated breast unit of patients who did not receive adjuvant therapy, had a standard management.

2.2 Gene expression profiling approaches

Although the current diagnostic tools are valuable, breast cancer is still one of the most frequent cause of cancer death worldwide (Garcia et al. 2007). There is clearly a need for improved diagnostic tools that are highly sensitive and specific to stratify patients and predict risk of recurrence and therapeutic sensitivities on a continuous scale to aid individualized decision making for the treatment.

Microarrays where expression profiling of thousands of messenger RNA transcripts takes place in a single experiment have been evolving in the past decade to become an established approach in biological research. Genome-wide expression profiling has led to a better stratification of breast cancer and has been useful for outcome prediction.

2.2.1 Microarrays

Microarray technology allows genome-wide interrogation of mRNA expression by hybridization of labeled RNA (or cDNA) to complementary sequences that are arrayed on a chip. After washing off the excess, the array is processed by a laser scanner to produce an image of differential signal intensities. The intensity of each probe can then be linked to the RNA abundance of the corresponding gene.

The basic procedure in an experiment involves the isolation of RNA or messenger RNA (mRNA) from appropriate biological samples, reverse transcription of mRNA into complementary DNA (cDNA) and hybridization of the fluorescence-labeled cDNA to the microarray. After washing off the excess, the array is processed by a laser scanner to produce an image of differential signal intensities. Dual-channel microarrays are typically hybridized with cDNA prepared from test (e.g. tumor) and reference (e.g. normal). It provides a relative measurement level for the corresponding RNA molecule. The one-channel arrays provide intensity for each probe (or probe set) indicating a relative level of hybridization.

Microarray platforms can be classified with respect to their manufacturing (spotted cDNA or oligonucleotide) and hybridization quantification (single or dual-channel). In spotted microarrays, the probes are synthesized prior to deposition on the array surface and are then spotted onto the chip. A common approach utilizes an array of fine pins or needles controlled by a robotic arm that is dipped into wells containing DNA probes and then depositing each probe at specific locations on the array surface. In oligonucleotide microarrays, the probes are short DNA sequences designed to match parts of the sequence of known or predicted gene coding regions. Oligonucleotide arrays are produced by chemically synthesizing short oligonucleotide sequences directly onto the array surface. Sequences may be longer (60-mer probes such as the Agilent Design) or shorter (25-mer probes produced by Affymetrix). Longer probes are more specific to individual target genes, and shorter probes may be spotted in higher density across the array and are cheaper to manufacture. Other microarray platforms, such as Illumina bead-based platforms (San Diego, CA, USA), use microscopic beads, instead of the large solid support.

In single-channel microarrays, a single mRNA source is hybridized on a chip and comparison of RNA levels between samples is made *in silico* in a post-processing phase of the experiment. In dual-channel microarrays, two mRNA sources are used, each labeled with different fluors. The second mRNA source is usually either a common reference against which all samples in an experiment are compared to, or a sample coming from a tissue under an alternative condition (e.g. tumor versus normal).

2.2.2 Strategies to develop gene-expression prognostic signatures

In general, strategies to develop a gene-expression prognostic signature from microarray data include the so-called "*top-down*" and "*bottom-up*" approaches (Sotiriou and Pusztai 2009). In the first strategy, identification of genes associated with prognosis is carried out in a supervised fashion guided by known clinical outcomes without any *a priori* biologic assumption (van 't Veer et al. 2002, Wang et al. 2005), while in the bottom-up discovery approach, genes associated with a specific biologic phenotype or a deregulated molecular pathway are first identified and then subsequently correlated with the clinical outcome

(Chang et al. 2005, Chi et al. 2006). In addition, a *candidate-gene* approach, utilized in development of Oncotype DX® (Paik et al. 2004) is based on data from quantitative reverse-transcriptase-polymerase chain reaction (Q-RT-PCR); the technique selects genes of interest on the basis of existing biologic knowledge which are then combined into a multivariate predictive model.

In high-throughput molecular profiling, the number of genes is typically much larger than the number of samples ($p>>n$), which would run into the phenomenon commonly referred to as the curse of dimensionality (Bellman 1961). Feature selection and dimension reduction often become key steps in the microarray data analysis. However, feature selection in microarray data is a nontrivial task due to high dimensionality, correlation between variables (features) and sometimes high level of noise. Below, we outline the common strategies used in gene expression data for signature construction, including feature selection, unsupervised analysis and supervised learning (Figure 1). See Hastie et al. (2001) for a review of statistical learning methods for high-dimensional problems.

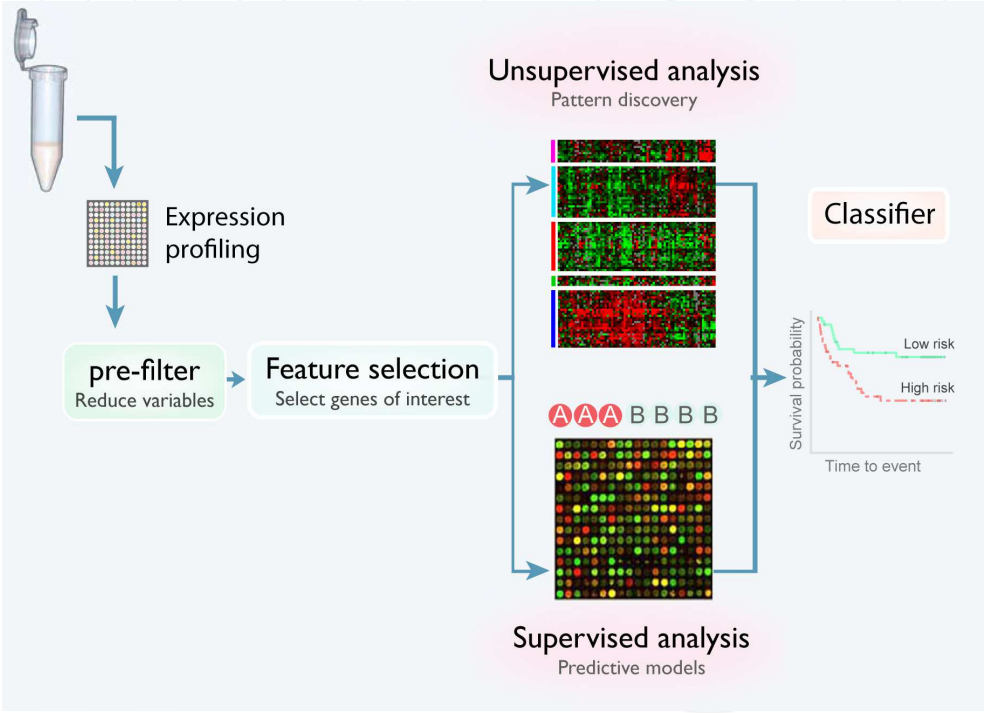


Fig. 1. Analytic pipeline for the development of a Gene-Expression Prognostic Signature.

Feature selection

Feature selection is different from quality filtering, where the reduction of the number of genes are carried out purely based on the quality of the measurement (such as signal-to-noise ratio, variance) without any information related to the outcome. A commonly used strategy for feature selection is to rank the genes (features) by their relevance to the outcome

of interest and then apply a cut-off to select the most interesting genes. Choice of model and statistical test depends on the characteristics of the outcome measurement, including whether it is quantitative (e.g. time to event) or categorical (e.g. tumor versus normal). Significance analysis of microarrays or SAM (Tusher et al. 2001) is one of the widest used methods in identifying differentially expressed genes in data arising from microarray experiments. It is a t-statistics variant that compares group means, adapted to a high-dimensional setting. Another popular approach is the empirical Bayes methods in the context of linear models (Smyth 2004). The empirical Bayes, SAM and other shrinkage methods (Smyth et al. 2005) are used to borrow information across genes to make the inferences stable when the number of arrays is small.

While the above filter methods assess the relevance of features, it ignores the effects of the selected feature subset on the accuracy of the model. The subset selection strategies assess subsets of features according to their relevance for a given model. These methods conduct a search for a good subset using the model itself as part of the evaluation function (e.g. stepwise regression by optimizing criteria such as AIC, BIC, Mallow's C_p, etc). Subset selection can also be achieved by cross-validation, where the samples are first split into $k > 1$ groups (or "folds") of roughly equal size. Suppose the goodness-of-fit is expressed by a loss function (which in the case of a likelihood based method would be minus the log of the likelihood). Then, for each model (variable subset) to be assessed, a cross-validation score is calculated by fitting the model to all samples except those in the j'th fold ($j = 1, 2, \ldots, k$), computing the loss function on the remaining fold, and adding together all the contributions. The model with the smallest cross-validation score is then selected.

In addition, methods embedded with internal variable selection (e.g. LASSO, PAM and decision tree) have also become popular tools for selecting a set of potential gene candidates from high-dimensional expression data. More complex model selection procedures such as double-loop k-fold cross-validation are sometimes used to determine several model parameters simultaneously.

Unsupervised pattern discovery

Unsupervised analyses are used to describe how the data are organized and find structures in the data. We only observe the features and do not use measurements of the outcome. Unsupervised methods such as hierarchical clustering, K-means clustering and self-organizing maps make it possible to identify groups of patients with similar gene expressions or groups of genes with similar expression pattern (co-expression gene cluster). The resulting patient groups are subsequently correlated with the clinical outcome as well as clinico-pathological parameters to assess the clinical relevance of the input gene markers.

Hierarchical clustering is an agglomerative approach in which single expression profiles are joined to form groups, which are further joined until the process has been carried to completion, forming a single hierarchical tree. There are several variations on hierarchical clustering that differ in how distances are measured between pairs of observations (distance metric) and between clusters (linkage criteria) as they are constructed. Hierarchical clustering is often criticized for giving ambiguous results because of sensitivity to data perturbation or clustering techniques used. The challenge is to select the algorithms appropriately so that the data is sensibly partitioned. Criteria such as the Gap statistic (Tibshirani et al. 2001), silhouette (Rousseeuw 1987), and bootstrap resampling (Suzuki and

Shimodaira 2004) are used to decide the optimal number, the quality and the reproducibility of the clusters, respectively. Often, external information about tumor characteristics (e.g. *TP53* mutation status, histological subtype, estrogen receptor status, etc) can be useful to evaluate to what extent the resulting tumor clusters fit with the existing prior knowledge.

Another commonly used unsupervised method is principal component analysis (PCA). It is a dimension reduction technique that produces linear combinations of the original variables to generate principal components (PCs) that are a set of uncorrelated variables. The first PC captures the highest variability presented in the data. PCA is useful for reduction of dimensionality by focusing on a few top principal components. Outliers can dominate the results of a principal components analysis.

The "*unsupervised*" approach aims at identifying subgroups of patients with similar gene expression pattern. The unsupervised learning process is not guided by any *a priori* biologic knowledge or clinical outcomes. In the "*supervised*" approach, markers associated with the outcome variable are identified. Validation (e.g. on an independent data set or using cross-validation procedure) is vital in the supervised learning process.

Supervised learning strategies

A supervised analysis aims to find a statistical relationship between input data (e.g. gene expressions) and output (e.g. response to a treatment or the survival of a patient). Supervised learning strategies can be further labeled according to whether outcome measurements are quantitative (regression) or qualitative (classification), as well as by whether models are designed to describe a current condition or predict future outcome based on a set of features (e.g. gene expression). A variety of models are available for regression and classification, respectively. Rather than elaborating each of these methods, we focus on the *regularization* approaches as the general remedy for the high-dimensionality in gene expression data.

When the number of explanatory variables genes (p) is large and even exceeds the number of individuals n used for training of the model, the fitted model typically performs well on the training data, but poorly on new observations. This is commonly referred to as overfitting and is a major concern in statistical analysis of high-dimensional data. In addition, a high degree of collinearity among the variables is likely to emerge, thereby leading to a situation in which the estimated regression coefficients may change substantially, even after slight perturbations of the training data.

In a linear model, dimension reduction techniques and penalized regression are the strategies to control and stabilize the variance of the estimates and further achieve better prediction rules. The primary goal in the penalized or regularized methods is to shrink the regression coefficients vector away from the ordinary least squares solution (in regression setting) and achieve improved the predictive performance through a bias-variance trade-off. Some widely used regression regularization methods such as ridge regression, partial least squares and principle components regression were compared in the studies by Frank and Friedman in 1993 (Frank and Friedman 1993) and Lingjærde and Christophersen in 2000 (Lingjærde and Christophersen 2000) . In Cox-ridge regression, the coefficients are estimated by maximization of the L2-penalized partial log-likelihood (using the Newton-Raphson procedure):

$$l_\lambda(\beta) = l(\beta) - \frac{1}{2}\lambda \sum_{i=1}^{p} \beta_i^2$$

where the first term is the partial log-likelihood and the second term is a penalty term in the form of a scaled L2 norm of the model coefficients (Verweij and Van Houwelingen 1994). Here, $\lambda > 0$ is a tunable penalty parameter that controls how much weight to put on the penalty function. The penalty parameter can be determined by the leave-one-out cross validation procedure proposed by Verweij and van Houwelingen (Verweij and Van Houwelingen 1993). Similar to ridge yet different, the lasso (Tibshirani 1996) is a penalized least squares method that imposes an L1 penalty on the regression coefficients. While ridge regression keeps all the predictors in the model, the lasso does both continuous shrinkage and automatic variable selection simultaneously. However, lasso is indifferent on the choice among a group of covariates that are strongly correlated. The elastic net penalty (Zou and Hastie 2005) was introduced as a compromise between ridge and lasso. The elastic-net simultaneously does automatic variable selection like the lasso and continuous shrinkage on the coefficients of correlated variables like ridge.

In a comparative study of survival prediction performance using microarray data (Bøvelstad *et al.* 2007), it has been found that Cox-ridge regression often outperforms other common regularization techniques for Cox regression, such as principal components regression, supervised principal components regression, partial least squares regression and the lasso.

Other supervised learning techniques, such as ensemble learning strategies (e.g. bagging, boosting and random forest) have also been applied to gene expression data analysis. The idea is to build a prediction model by combining the strengths of a number of weak learners. Refer to Hastie et al. (Hastie et al. 2001) for overview on a comprehensive collection of statistical learning methods.

2.2.3 Established gene signatures

Some of the established gene signatures with potential clinical usage are reviewed below. This review covers Intrinsic subtypes (Perou et al. 2000, Sørlie et al. 2001, Sørlie et al. 2003, Parker et al. 2009), MammaPrint® (van 't Veer et al. 2002), Wound-Response (Chang et al. 2004, Chang et al. 2005), 76-gene (Wang et al. 2005), Genomic Grade Index (Sotiriou et al. 2006), Oncotype DX® (Paik et al. 2004) and Hypoxia (Chi et al. 2006). For each of the gene signatures, we briefly describe the development procedures, the clinical characteristics for the targeted cohorts as originally intended (Table 1) and the critical requirements that are signature-specific for appropriate usage.

Expression-based molecular subtypes

The initial "intrinsic gene set" was found by searching genes that showed little variance within tumor samples (i.e., before and after neoadjuvant chemotherapy pairs), but high variance across different tumors (Perou et al. 2000). The signature that comprised the 496 intrinsic genes was further developed by unsupervised classifications that were based on clustering algorithms. The intrinsic signature was then used to classify breast tumors into five biological subgroups (luminal A, luminal B, HER2-enriched, basal-like, and normal-like) that show distinct clinical implications (Perou et al. 2000). There exist a couple of variants with different numbers of genes in the intrinsic gene set in subsequent publications (Sørlie et al. 2001, Sørlie et al. 2003, Hu et al. 2006, Perreard et al. 2006).

Signature	Predicted phenotype / Endpoint	Training Cohort	Validation Cohort
Intrinsic	Subtypes	Locally advanced BC	Consecutive BC
PAM50	Subtypes	Consecutive BC[1]	Consecutive BC
	5-year relapse	Node–	Node– & +
70gene	5 year distant metastasis	Node–	Node– & +
76gene	5 year distant metastasis	Node–	Node–
GGI	HG1- or HG3-like in HG2	ER+	Consecutive BC
WR	Active or Quiescent CSR	Representative BC[2]	Representative BC
	Population based prognosis		Consecutive BC
Hypoxia	Hypoxic or Non-hypoxic	Representative BC	Representative BC
RS	10 year distant metastasis	Tamoxifen-treated; ER+; Node–	Tamoxifen-treated; ER+; Node– & Node+

Table 1. Characteristics of the studied gene signatures and their clinically relevant breast cancer cohorts.

The molecular subtypes have profound impact in unveiling heterogeneities in breast cancers. The presence of distinct molecular entities suggests the existence of multiple "cells of origin"(Prat and Perou 2009). There has been a shift in how the subtypes are defined over time, such as including more proliferation-associated genes (Hu et al. 2006, Parker et al. 2009). This may partially explain the discordance between PAM50 and Intrinsic, with respect to LumA and LumB classification. In our study (Zhao et al. Unpublished), we compared the subtype classification between Intrinsic (Perou et al. 2000) and PAM50 (Parker et al. 2009) on a large breast cancer dataset (van Vliet et al. 2008) (n = 947). Overall, subtype assignments of the signatures were moderately correlated (Cohen's kappa, $\kappa = 0.54$) (Cohen 1960). Noticeably, nearly half of the LumA tumors by Intrinsic were assigned as LumB by PAM50, while the two signatures appeared to highly agree on classification of basal-like subtype tumors. Indeed, basal-like was the most concordant subtype with a Pearson correlation of 0.94 between Intrinsic and PAM50, followed by normal-like (0.85), LumA (0.68), LumB (0.55) and Her2-enriched (0.42). More specifically, basal-like was the most distinctly classified subtype across these two signatures, as only those samples for which the correlation to basal-centroid by Intrinsic was slightly larger than the second highest centroid correlation showed inconsistent calls by PAM50.

As previously pointed out (Sørlie et al. 2010), an important issue for the molecular subtyping of breast cancers is the need for a clear definition of the molecular subtypes of breast cancer and standardized analytical methods to identify them. Until a consistent taxonomy is established, it is expected for inconsistent results when comparing assignments by various approaches that do not comprise the same entities.

70-gene signature

The 70-gene prognosis profile or MammaPrint® (Agendia, Amsterdam, The Netherlands) (van 't Veer et al. 2002) has been trained on a cohort of *lymph-node-negative* patients:

[1] *Consecutive BC*: heterogeneous breast cancer cohort with consecutive clinical parameter distribution as reflected in the whole population of this disease.

[2] *Representative BC*: breast cancer dataset at hand carries representative features that are associated with a certain breast cancer subpopulation.

expression of a set of 70 prognostic markers that was identified in a ''supervised'' fashion based on their ability to predict freedom from tumor metastasis (favorable prognosis) over a five-year period in the same dataset. It was validated subsequently on NKI295, a larger cohort consisting both node negative and positive patients (van de Vijver et al. 2002) and another validation study (Mook et al. 2008) was done on cohorts of 241 patients with 1-3 positive lymph nodes. Despite the fact that part of the validation set in the original retrospective validation study (van de Vijver et al. 2002) was overlapped with the training set of the signature (van 't Veer et al. 2002), the 70-gene signature has been validated in the independent cohort by the TRANSBIG consortium (Buyse et al. 2006). Espinosa *et al.* (Espinosa et al. 2005) reproduced with quantitative reverse-transcriptase-polymerase chain reaction (Q-RT-PCR) the results obtained with a 70-gene expression profile.

The gene signature classifies patient into good or bad prognostic group by the average profile of previously determined 70 genes in tumors from patients with a good prognosis. A patient with a correlation coefficient of more than 0.4 was then assigned to the group with a good-prognosis signature and all other patients were assigned to the group with a poor-prognosis signature. The threshold was set to achieve a 10 percent rate of false negative results in the 78 tumors in the previous study (van 't Veer et al. 2002).

76-gene signature

The 76-gene signature (Veridex) (Wang et al. 2005, Foekens et al. 2006, Desmedt et al. 2007a) is designed to predict distant metastasis within five years for lymph-node-negative breast cancer patients. It was original developed based on 286 lymph-node-negative breast cancer patients (Wang et al. 2005) and validated on an independent multicentric population of 180 untreated N- breast cancer patients (Foekens et al. 2006) and another gene expression study of 198 node-negative breast cancer patients (Desmedt et al. 2007a) from the same Affymetrix U133a platform as in the original study (Wang et al. 2005).

In the 76-gene signature (Wang et al. 2005), a relapse score is calculated for ER+ and ER- samples using sum of the weighted log2-gene-expression of the 60 genes and 16 genes, respectively:

$$\sum_{i=1}^{60} w_i x_i \text{ (for ER positive sample)}$$

$$\sum_{j=1}^{16} u_j y_j \text{ (for ER negative sample)}$$

where i and j indicate markers for ER positive and ER negative group, respectively; w_i and u_j are the standardized Cox regression coefficients for ER positive and ER negative markers, respectively; x_i and y_j are the expression values of ER positive and ER negative markers, respectively.

Intuitively, the pre-derived constants in the relapse model (Wang et al. 2005) are likely platform dependent. Additionally, we observed that the 76-gene signature was unable to identify any Desmedt sample (Desmedt et al. 2007a) with good prognosis when applied on RMA- instead of MAS5-normalized data. The discrepancies suggested that the risk cutoffs

and possibly its original gene weights in the algorithm are sensitive to the data scale (Zhao et al. Unpublished).

Genomic Grade Index

The Genomic Grade Index (GGI) is a 97-gene measure of histologic tumor grade. The GGI was able to reclassify patients with histologic grade 2 tumors into two groups with distinct clinical outcomes similar to those of histologic grade 1 and 3, respectively (Sotiriou et al. 2006).

The Genomic Grade Index signature contains 128 Affymetix probes (representing 97 genes), of which 112 probes were with increased expression in histologic grade 3 tumors; and the remaining 16 probes with increased expression in histologic grade 1 tumors. The expressions of the 97 grade associated genes were further combined into the genomic grade index (GGI) by:

$$\sum_{j \in G_3} x_j - \sum_{j \in G_1} x_j$$

where x_j is the expression of either a grade 1 marker or grade 3 marker. The raw GGI scores were further scaled so that the mean of the GGI scores of histologic grade 1 tumors was − 1 and that of histologic grade 3 tumors was +1:

$$\text{GGI} = \text{scale}\left(\sum_{j \in G_3} x_j - \sum_{j \in G_1} x_j - \mathit{offset}\right)$$

High GGI is associated with decreased relapse-free survival in both untreated and tamoxifen-treated patients (Loi et al. 2007). In the original publication (Sotiriou et al. 2006) the GGI signature was proposed to classify histologic grade 2 samples (or samples neither HG1 nor HG3) into "HG1-like" & "HG3-like". Tumors with a negative GGI score were classified as "HG1-like"; 0 or a positive GGI score put a tumor into "HG3-like" category. In the subsequent study (Haibe-Kains et al. 2008a), the authors dichotomized the raw GGI into "low-risk" and "high-risk" group based on 33% quartile in two different populations, VDX and TRANSBIG, respectively: the third of the patients having the lowest GGI scores being defined as low-risk and the remaining patients as high-risk. The population based prognostic strategy for GGI signature particularly requires that the samples are a good representative of the population of breast cancer with consecutive clinical parameter distribution.

GGI has a standardization procedure using the information of histological grade so that the mean of the GGI scores of histologic grade 1 tumors was -1 and that of histologic grade 3 tumors was +1, which is likely to increase its robustness when transferred to another array-based expression dataset (Zhao et al. Unpublished).

Wound response

The wound response or core serum response (CSR) gene signature (Chang et al. 2004) was derived from the transcriptional response of normal fibroblasts to serum in cell culture. It has been shown to improve the risk stratification of early breast cancer over that provided by standard clinic pathological features, in that the development of distant metastases is

more likely among patients whose breast cancers have activated pathways for matrix remodeling, cell motility, and angiogenesis than among those whose cancers do not. The signature classifies tumors into two classes (*Activated* vs. *Quiescent*) through a centroid, which was built from the averaged fibroblast serum-induced expression pattern of the CSR genes (Chang et al. 2004, Chang et al. 2005).

Hypoxia signature

The epithelial hypoxia signature (Chi et al. 2006) consists of genes (253 image clones) that were consistently induced by hypoxia in cultured epithelial cells (HMECs and RPTECs). The 253 image clones were mapped to 168 Unigene clusters in the study (Chi et al. 2006). A "hypoxia score" was computed for a patient by averaging expression levels for the hypoxia response genes. Patients were assigned into high or low hypoxia response group by a cutoff hypoxia-score at zero (Nuyten et al. 2008). A positive score indicates *hypoxic* and non-positive score indicates *non-hypoxic*. Using published data sets, the authors found that the "high hypoxia response" group tends to be higher grade, and more likely to have p53 and oestrogen receptor deficiencies, and, most importantly, a significant association with a poorer prognosis in breast and ovarian cancer.

Oncotype DX®

Oncotype DX® (Genomic Health Inc., Redwood City, CA) (Paik et al. 2004) or the 21-gene-recurrence-score signature was developed from quantitative reverse transcription-polymerase chain reaction (Q-RT-PCR) assay to quantify the likelihood of distant recurrence at 10 years in adjuvant-tamoxifen-treated patients and further spare patients from adjuvant chemotherapy, in both node-negative (Paik et al. 2004) and node-positive disease(Albain et al. 2010). It includes 16 cancer-related genes that can be grouped into five different biological domains—proliferation, HER2 signaling, ER signaling, invasion and other—along with five reference genes. The linear combination of scores from these biological groups was computed and scaled into a Recurrence Score (RS), which is used to classifier a patient into categories of high risk (RS≥31), intermediate risk (18≤RS<31), and low risk of recurrence (RS<18).

Applying Oncotype DX® to a microarray-based dataset is not straightforward, which has often been underappreciated in existing studies (Fan et al. 2006, Loi et al. 2007). In Oncotype DX®, reference-normalized expression measurements ranged from 0 to 15, where one unit increase reflects approximately a 2-fold increase in RNA. The exact quantifications are hard to draw in the microarray-based measurements. We emphasize that only a pseudo RS based on the Oncotype DX® algorithm can be computed from microarray-based datasets.

Applicability of individual gene signatures

Translating the expression-based gene signatures to a new dataset is complicated by the heterogeneities derived from using several microarray, the differences of data processing procedures and the clinical uniqueness of each studied cohort.

We grouped the above gene signatures into two broad categories based on their associated approaches of summarizing expression values: *centroid-based* (Intrinsic, PAM50, 70-gene and WR) and *weighted average gene expression predictors* (76-gene, GGI, RS and Hypoxia).

For the 76-gene signature, the pre-derived constants in the relapse model are likely platform dependent. Ideally, applying this signature to a new dataset, one should follow the same protocol using the same platform with the same normalization procedure as in the original studies (Wang et al. 2005, Desmedt et al. 2007a). The Oncotype DX® (RS), another signature based on weighted sum method, also has potential issues related to the data scale in computing the recurrence score. Furthermore, the differences between the microarray and PCR technologies make the recurrence scores estimated from microarray experiments less optimal. GGI shares similarities with the 76-gene signature and Oncotype DX® in constructing risk estimation from gene expression pattern. However, GGI has a unique standardization procedure incorporating the information of histological grade, which likely increases its robustness when transferred to different microarray platforms. Generally, when the distribution of risk scores depends on platform and normalization procedure, as we found with some signatures, cutoffs for risk group assignment need to be recalibrated. The population-based strategy in which a fixed proportion of the population was assigned to each risk group is more general and applicable for a study with pure prognosis purpose on the new cohort. However, it particularly requires the samples to be representative of the population of breast cancer.

A previous study concluded that complex models are not better predictors of prognosis than simpler ones derived from gene expression studies (Haibe-Kains et al. 2008b). In general, we believe that successful models should be constructed in a robust way to tolerate cross-platform differences. This may explain why methods based on centroid correlations (such as subtype signatures and the 70-gene) or methods that transform the data into an invariant scale before computing the risk scores (such as GGI), have more consistent performances. We suspect that the weighted average fashion is more sensitive to the data scale and the issue of missing signature gene(s) in the data at hand.

2.3 Limitations of the "first-generation" expression profiling

Gene expression profiling has opened a door for personalized medicine. However, the "first-generation" gene signatures may offer no more than a snapshot of a tumor's gene expression profile that is most relevant for only a particular point in time. Meanwhile, tumor development is essentially Darwinian and tumor heterogeneity is dynamic as selective pressures change during the metastatic process. The complex structural network of the tumor system and the vital interactions of tumor cells with stromal and immune cells highlight the need for a cellular systems biology approach to cancer diagnostics, which combines multiplexed biomarker panels with informatics tools to produce a systemic readout relevant to patient prognosis. Comprehensive genomic analysis of tumor subpopulations of the host patient is likely the best way to effectively use gene signatures from both patient and tumor, so that treatment plans can be optimized.

2.3.1 Influence of time and ER status on gene signatures in breast cancer survival prediction

In Zhao et al. (Zhao et al. Unpublished), we assessed several prognostic gene signatures that have received the greatest interest and been validated in multiple studies. These include the *Intrinsic signature* (Perou et al. 1999, Perou et al. 2000, Sørlie et al. 2001, Sørlie et al. 2003), *PAM50* (Parker et al. 2009), 70-gene profile or MammaPrint® (Agendia, Amsterdam, The

Netherlands) (van 't Veer et al. 2002, van de Vijver et al. 2002, Mook et al. 2008, Buyse et al. 2006, Espinosa et al. 2005), 76-gene signature (Wang et al. 2005, Foekens et al. 2006, Desmedt et al. 2007a), Genomic Grade Index (GGI) (Sotiriou et al. 2006, Loi et al. 2007), wound response (WR) signature (Chang et al. 2004, Chang et al. 2005), hypoxia signature (Chi et al. 2006, Nuyten et al. 2008) and 21-gene-recurrence-score (RS) or Oncotype DX® (Genomic Health Inc., Redwood City, CA) (Paik et al. 2004).

The eight signatures were applied on an expression dataset (van Vliet et al. 2008) (n = 947) pooled from six published breast cancer datasets (Loi et al. 2007, Miller et al. 2005, Pawitan et al. 2005, Desmedt et al. 2007a, Minn et al. 2005, Chin et al. 2006) on Affymetrix Human Genome HG-U133A arrays. Survival predictions were fairly concordant across most gene signatures (Zhao et al. Unpublished). We found that these signatures generally performed better in ER positive than in ER negative breast cancers for prediction of distant metastasis free survival (Zhao et al. Unpublished). Cell proliferation seems to be the common driving force for the prognostication in ER positive breast cancers, while different biological mechanisms such as stress response and immune response (Rody et al. 2009, Teschendorff and Caldas 2008) may be crucial for risk stratification in ER negative tumors. The majority of the tested gene signatures are strong risk predictors especially during the first five years of follow-up for distant metastasis free survival and throughout the first 10 years for breast cancer specific survival. These indications are also in line with results from other studies (Desmedt et al. 2007b, Desmedt et al. 2008, Wirapati et al. 2008, Loi et al. 2007). It suggests that different molecular mechanisms are likely to be involved in the early and the late stage during the progression of the metastatic disease.

2.4 Combining multiple gene signatures likely to improve prognosis

Despite the fact that very few genes are shared among various gene signatures, most of gene signatures, evaluated in our own studies (Zhao et al. 2011) and by others (Fan et al. 2006, van Vliet et al. 2008, Reyal et al. 2008), have similar performances in survival risk assessment on the same breast cancer patients. This indicates that some common biological processes overlap across those gene signatures (Reyal et al. 2008, Yu et al. 2007, Desmedt et al. 2008), but more importantly they are likely to capture various biological aspects of breast cancer (Drier and Domany 2011). The combined information from multiple informative gene signatures is arguably more broadly applicable for survival prediction across heterogeneous tumor groups capturing a broad spectrum of biological aspects.

Methods such as decision-tree analysis have been explored to develop a combined predictor that showed improved performance than the individual gene signatures (Chang et al. 2005). In Zhao et al. (Zhao et al. 2011), an analytical framework (Fig. 1) was proposed to improve breast cancer risk stratification by integration of multiple informative gene signatures. We use the gene sets of eleven published gene signatures (Paik et al. 2004, Finak et al. 2008, Minn et al. 2005, van 't Veer et al. 2002, Wang et al. 2005, Sotiriou et al. 2006, van Vliet et al. 2008, Chi et al. 2006, Liu et al. 2007, Hu et al. 2006, Chang et al. 2004) to analyze breast cancer survival and relapse. To investigate the relationship between breast cancer survival and gene expression on a particular gene set, a Cox proportional hazards model is applied using partial likelihood regression with an L2 penalty to avoid overfitting and using cross-validation to determine the penalty weight. The fitted models are applied to an independent test set to obtain a predicted risk index (PI) for each individual and each gene

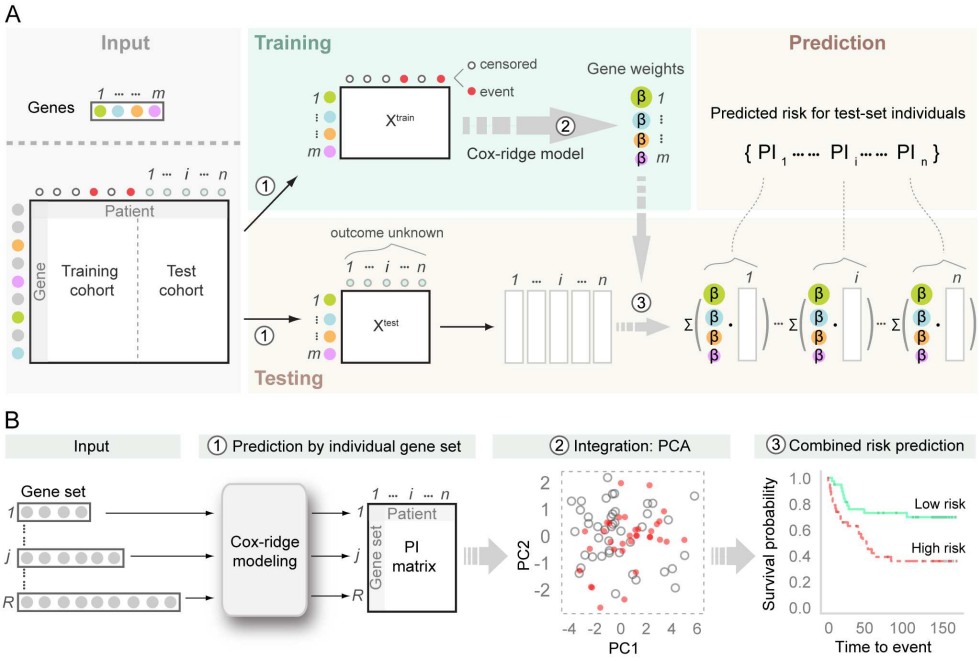


Fig. 1. Flowchart of the analysis showing integration of multiple informative gene signatures.

(A) Construction of the gene-set predictor/gene signature for risk prediction. Input: A set of genes of interest (gene *1*, …, *m*) which can be traced by the corresponding colors through out the diagram; gene expression data for training cohort and test cohort with genes placed in the rows and patients in the columns. *Step 1*. Gene identity mapping and extract expression matrix. *Step 2*. With available status of observing an event for the patients on the training set, a Cox model with L2 penalty is used to model the relationship of survival probability and gene expression pattern of the gene set. The coefficients or "gene weights" (β_1, …, β_m) associated with individual genes are estimated from the Cox-ridge model. Size of the bubble in the gene weights matrix reflects the importance of the corresponding gene for survival prediction. *Step 3*. A *Prognostic Index* (PI), the predicted risk score for a test patient *i* (*i* = *1*, …, *n*) is calculated by the sum of weighted gene expression from test patient *i* using the estimated gene weights from step2. **(B)** Integration of multiple gene signatures by dimension reduction. Input multiple gene sets of interest together with their gene expression data. *Module 1*: For *j*th gene set (*j* = 1, …, *R*), the procedure described in panel A is used to predict a risk score PI for individual test patient. The resulting PI matrix is positioned in *R* by *n* dimension representing the risk prediction of the *n* test patients by each of the *R* gene sets. *Module 2:* Integrate predictions from multiple gene signatures by dimension reduction using principal components analysis (PCA).

Module 3: Dichotomize the risk scores on PC1 by median (higher than median indicates high risk) resulting in two predicted risk groups for survival outcome. Image is taken from Ref. (Zhao et al. 2011).

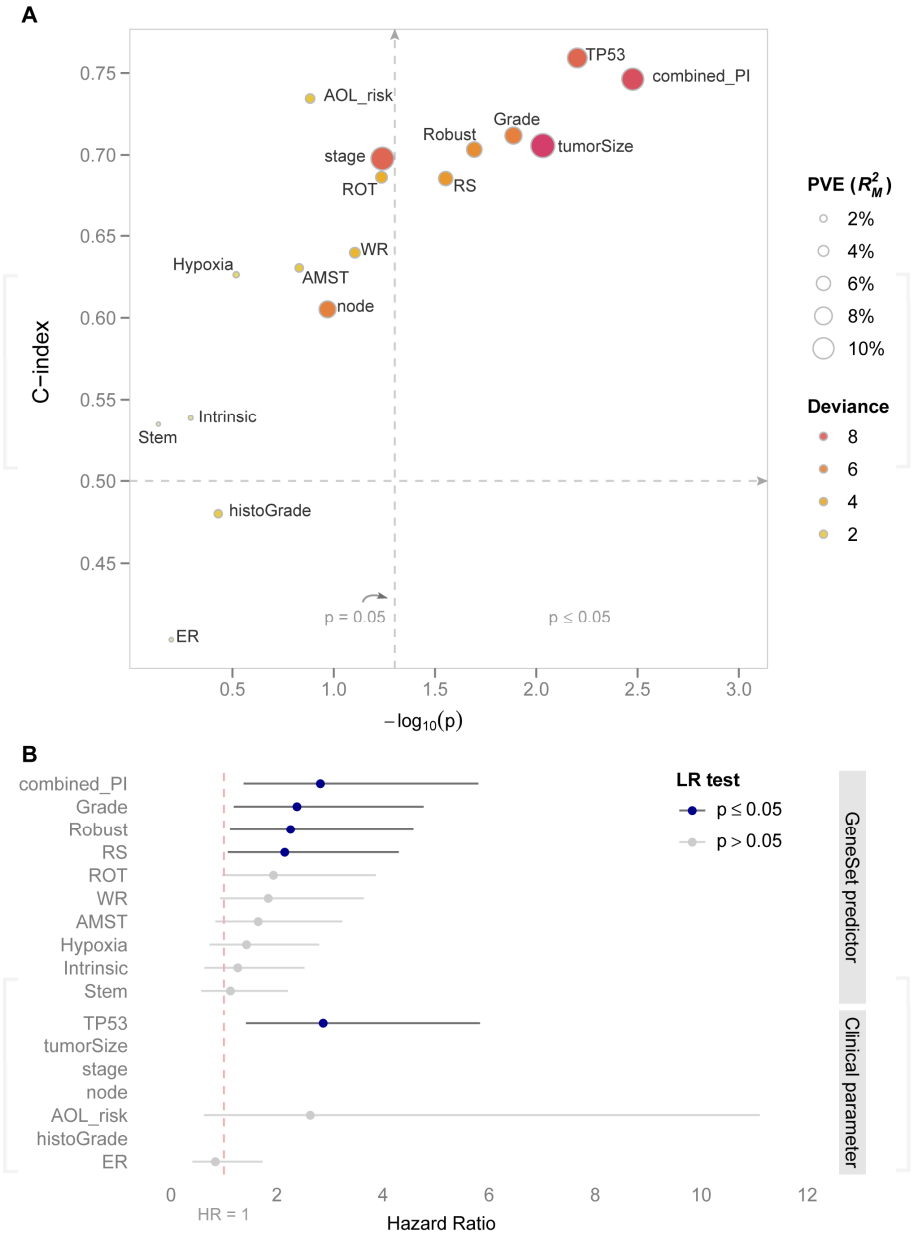


Fig. 2. Univariate comparison of predictors for systemic recurrence. Comparison of combined-PI risk predictor with clinical parameters and individual gene-set predictors using univariate Cox model. **(A)** Y-axis indicates C-index associated with individual predictor and X-axis indicates the p values (on minus log10 scale) from likelihood ratio test in univariate Cox model. C-index = 0.5 and the significant level: $\alpha = 0.05$ for the likelihood

ratio test are indicated by the dotted line. The size and the color of the bubble indicate the PVE and the deviance in univariate Cox model, respectively. The combined-PI risk predictor had the most significant marginal effect for predicting systemic recurrence (p = 0.003). It was associated with the second highest C-index score (C = 0.75) following *TP53* mutation status (C = 0.76). It had the second highest deviance (8.61) following tumor size (9.36), and the combined-PI predictor alone explained 10.6% of the variability as indicated by PVE, following tumor size (11.7%) and stage (11.1%) **(B)** X-axis indicates HR from the univariate Cox model and the 95% CIs are shown along with the point estimates. "LR test" stands for likelihood ratio test. Insignificant predictors (likelihood ratio test $p > 0.05$) are grayed out. To keep the results interpretable, only predictors with two levels are compared. The combined-PI risk predictor had the 2nd largest HR (2.82 with 95% CI 1.37—5.80) following *TP53* mutation status (2.87 with 95% CI 1.42—5.83). Image is taken from Ref. (Zhao et al. 2011).

set. Finally, principal components analysis of the gene signatures is used to derive a combined predictor from the first principal component. Based on a media cut, this combined model classifies test individuals into two risk groups with distinct survival characteristics (recurrence: p=0.003; breast cancer specific death: p=0.001). And it outperforms all the individual gene signatures, as well as Cox models based on traditional clinical parameters and the Adjuvant! Online for survival prediction (Fig. 2).

One weakness of this study is the fact that the training and test sets contain small sample sizes (training set: n = 123; test set: n = 80). The effect of the small sample size is reflected in the low degree of correlation between the PIs obtained by swapping the training and test sets. On the positive side, this study represents an elegant way to combine existing gene sets into a single predictor, without discretizing the survival times. It would be very interesting to see the performance of this classifier on a much larger cohort and explore different approaches for the integration step.

2.5 Clinical trials to conclude the clinical utility of gene expression signatures

To meet the requirements of a prognostic marker, the potential marker should be tested retrospectively in large patient cohorts with a long follow-up period. Subsequently, the findings should be validated by an independent group of experts, and, ideally, a prospective study should confirm the prognostic significance of the tested marker.

Ongoing clinical trials, MINDACT (Microarray In Node negative Disease may Avoid ChemoTherapy) (Cardoso and Van't Veer 2008) and TAILORx (Trial Assigning IndividuaLized Options for Treatment (Rx)) (Sparano and Paik 2008) have been launched to test the clinical usage of MammaPrint® (Agendia, Amsterdam, The Netherlands) (van 't Veer et al. 2002) and Oncotype DX® (Genomic Health Inc., Redwood City, CA) (Paik et al. 2004), respectively. MINDACT will directly compare the 70-gene signature (experimental arm) with Adjuvant! Online (clinico-pathological control arm) to determine whether to offer adjuvant chemotherapy in patients with node-negative breast cancer presenting with discordant risk estimation according to the two methods. It is estimated that 10-15% fewer women will be treated with chemotherapy in the experimental arm. In TAILORx, patients with low RS will be treated with hormonal therapy alone and patients with a high score will receive chemotherapy plus hormonal therapy. However, the 10-year results of both trials will not be available before the year 2020. These trials should provide level I evidence about the clinical relevance of applying gene-expression signatures to daily breast cancer patient management.

In addition, a phase II clinical trial design, the I-SPY 2 (investigation of serial studies to predict your therapeutic response with imaging and molecular analysis 2) (Barker et al. 2009), will test the idea of tailoring treatment by using molecular tests (estrogen receptor status, HER2 status, and the MammaPrint® (Mook et al. 2007, Cardoso and Van't Veer 2008) to identify patients who might benefit from investigational new drugs given along with standard neoadjuvant chemotherapy.

3. Conclusion

Breast cancer is markedly heterogeneous with respect to distinctive biological characteristics and clinical behavior. Many examples highlight that gene expression signatures have tremendous power to identify new cancer subtypes and to predict clinical outcomes. The genome-wide information of breast cancer provides overlapping clinico-pathological classifications, more importantly, adds prognostic accuracy and biological insights than relying on single biomarkers alone.

These signatures are more predictive in ER positive tumors, as seen from our study (Zhao et al. Unpublished) and others. Their low performances in ER negative group are in line with their limitation of assigning the high-risk category to almost all ER-negative patients (Sotiriou and Pusztai 2009, Wirapati et al. 2008). Moreover, their effects on survival prediction seem to decay with time (Desmedt et al. 2007b, Zhao et al. Unpublished), suggesting that different molecular mechanisms are likely involved during the development of early and late stages of the disease.

4. Future of personalized medicine in breast cancer

Genomic signatures play a significant role in individualized diagnosis, prognosis and therapeutic decision-making for cancer patients. In additional to mRNA expression profiling, other genetic information such as genomic complexity inferred from aCGH data (Russnes et al. 2010) also has possibilities to be translated into clinical applications for breast cancer. More recently, next generation DNA sequencing has been used to support the goals of personalized medicine. Charactering complete catalogues of the somatic alterations in cancer genomes holds great potential to discover informative biomarkers and develop targeted therapeutics (Chin et al. 2011).

Clinical and pathological factors such as axillary lymph node status, tumor size, histological grade, histological subtype, HER2 status, and hormone receptor status are still the most important factors for determining treatment. With increasing knowledge of specific genetic alterations and gene expression profiles of tumors, and the prognostic and predictive value of these genetic tumor characteristics, more individualized predictions of disease outcome and refined patient therapy are beginning to be realized.

Integration of clinical, pathological, genetic information derived from gene expression profiling, aCGH and massive parallel sequencing as well as metabolic profiles is a promising approach to achieve better breast cancer risk stratification and further to improve treatment decisions in breast cancer patients. Methods such as PARADIGM (Vaske et al. 2010) have been explored to infer patient-specific signaling pathway activities from integration of multi-dimensional cancer genomics data. Furthermore, the predicted pathway perturbations were able to stratify patients into clinically relevant subtypes (Vaske et al.

2010). With the advances in genomic technologies and the increased volume of high throughput data, it is imperative to develop approaches for integration of diverse biological information—DNA (and epigenetic changes), RNA, proteins and metabolites together with clinical, pathological information.

We look forward to the completion of the ongoing clinical trials to confirm the clinical utility of expression-based gene signatures in breast cancer. We anticipate that these results will facilitate the translation of other genetic information (such as genomic complexity inferred from aCGH data) (Russnes et al. 2010) into clinical applications for breast cancer. We particularly look forward to the impact of next generation DNA sequencing on diagnosis, prognosis and therapeutic decision-making.

5. References

Albain, K. S., W. E. Barlow, S. Shak, G. N. Hortobagyi, R. B. Livingston & I. Yeh (2010) Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. *The Lancet Oncology,* 11**,** 55-65.

Bamford, S., E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, P. Futreal & M. Stratton (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British journal of cancer,* 91**,** 355-358.

Barker, A., C. Sigman, G. Kelloff, N. Hylton, D. Berry & L. Esserman (2009) I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics,* 86**,** 97-100.

Bellman, R. 1961. Adaptive control processes. Princeton University Press, Princeton, NJ.

Bøvelstad, H. M., S. Nygard, H. L. Størvold, M. Aldrin, O. Borgan, A. Frigessi & O. C. Lingjærde (2007) Predicting survival from microarray data--a comparative study. *Bioinformatics,* 23**,** 2080-7.

Buyse, M., S. Loi, L. Van't Veer, G. Viale, M. Delorenzi, A. M. Glas & A. Saghatchian (2006) Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *JNCI Cancer Spectrum,* 98**,** 1183.

Cardoso, F. & L. Van't Veer (2008) Clinical application of the 70-gene profile: the MINDACT trial. *Journal of Clinical Oncology,* 26**,** 729.

Carter, C. L., C. Allen & D. E. Henson (1989) Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer,* 63**,** 181-187.

Chang, H. Y., D. S. A. Nuyten, J. B. Sneddon, T. Hastie, R. Tibshirani, T. Sørlie, H. Dai, Y. D. He, L. J. Van't Veer & H. Bartelink (2005) Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 3738.

Chang, H. Y., J. B. Sneddon, A. A. Alizadeh, R. Sood, R. B. West, K. Montgomery, J. T. Chi, M. van de Rijn, D. Botstein & P. O. Brown (2004) Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol,* 2**,** E7.

Chi, J. T., Z. Wang, D. S. Nuyten, E. H. Rodriguez, M. E. Schaner, A. Salim, Y. Wang, G. B. Kristensen, A. Helland, A. L. Borresen-Dale, A. Giaccia, M. T. Longaker, T. Hastie, G. P. Yang, M. J. van de Vijver & P. O. Brown (2006) Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. *PLoS Med,* 3**,** e47.

Chin, K., S. DeVries, J. Fridlyand, P. T. Spellman, R. Roydasgupta, W. L. Kuo, A. Lapuk, R. M. Neve, Z. Qian & T. Ryder (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell,* 10**,** 529-541.

Chin, L., J. N. Andersen & P. A. Futreal (2011) Cancer genomics: from discovery science to personalized medicine. *Nature medicine,* 17**,** 297-303.

Chustecka, Z. (2007) Survival Disadvantage Seen for Triple-Negative Breast Cancer.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and psychological measurement,* 20**,** 37-46.

Dent, R., M. Trudeau, K. I. Pritchard, W. M. Hanna, H. K. Kahn, C. A. Sawka, L. A. Lickley, E. Rawlinson, P. Sun & S. A. Narod (2007) Triple-negative breast cancer: clinical features and patterns of recurrence. *Clinical Cancer Research,* 13**,** 4429.

Desmedt, C., B. Haibe-Kains, P. Wirapati, M. Buyse, D. Larsimont, G. Bontempi, M. Delorenzi, M. Piccart & C. Sotiriou (2008) Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clinical Cancer Research,* 14**,** 5158.

Desmedt, C., F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang & M. d'Assignies (2007a) TRANSBIG Consortium. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res,* 13**,** 3207-14.

Desmedt, C., F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang & M. S. d'Assignies (2007b) Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clinical Cancer Research,* 13**,** 3207.

Drier, Y. & E. Domany (2011) Do Two Machine-Learning Based Prognostic Signatures for Breast Cancer Capture the Same Biological Processes? *PLoS One,* 6**,** e17795.

Espinosa, E., J. Vara, A. Redondo, J. Sanchez, D. Hardisson, P. Zamora, F. G. Pastrana, P. Cejas, B. Martinez & A. Suarez (2005) Breast cancer prognosis determined by gene expression profiling: a quantitative reverse transcriptase polymerase chain reaction study. *Journal of Clinical Oncology,* 23**,** 7278.

Fan, C., D. S. Oh, L. Wessels, B. Weigelt, D. S. Nuyten, A. B. Nobel, L. J. van't Veer & C. M. Perou (2006) Concordance among gene-expression-based predictors for breast cancer. *N Engl J Med,* 355**,** 560-9.

Finak, G., N. Bertos, F. Pepin, S. Sadekova, M. Souleimanova, H. Zhao, H. Chen, G. Omeroglu, S. Meterissian, A. Omeroglu, M. Hallett & M. Park (2008) Stromal gene expression predicts clinical outcome in breast cancer. *Nat Med,* 14**,** 518-27.

Fisher, B., C. Redmond, E. R. Fisher & R. Caplan (1988) Relative worth of estrogen or progesterone receptor and pathologic characteristics of differentiation as indicators of prognosis in node negative breast cancer patients: findings from National Surgical Adjuvant Breast and Bowel Project Protocol B-06. *Journal of Clinical Oncology,* 6**,** 1076.

Foekens, J. A., D. Atkins, Y. Zhang, F. C. G. J. Sweep, N. Harbeck, A. Paradiso, T. Cufer, A. M. Sieuwerts, D. Talantov & P. N. Span (2006) Multicenter validation of a gene expressionñbased prognostic signature in lymph nodeñnegative primary breast cancer. *Journal of Clinical Oncology,* 24**,** 1665.

Frank, I. & J. Friedman (1993) A statistical view of some chemometrics regression tools. *Technometrics,* 35**,** 109-135.

Garcia, M., A. Jemal, E. Ward, M. Center, Y. Hao, R. Siegel & M. Thun (2007) Global cancer facts & figures 2007. *Atlanta, GA: American Cancer Society,* 1.

Haibe-Kains, B., C. Desmedt, F. Piette, M. Buyse, F. Cardoso, L. Van't Veer, M. Piccart, G. Bontempi & C. Sotiriou (2008a) Comparison of prognostic gene expression signatures for breast cancer. *BMC Genomics,* 9**,** 394.

Haibe-Kains, B., C. Desmedt, C. Sotiriou & G. Bontempi (2008b) A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? *Bioinformatics,* 24**,** 2200.

Hastie, T., R. Tibshirani & J. Friedman (2001) The elements of statistical learning: data mining, inference, and prediction. *New York: Springer-Verlag,* 1**,** 371-406.

Haybittle, J., R. Blamey, C. Elston, J. Johnson, P. Doyle, F. Campbell, R. Nicholson & K. Griffiths (1982) A prognostic index in primary breast cancer. *British journal of cancer,* 45**,** 361.

Hu, Z., C. Fan, D. S. Oh, J. S. Marron, X. He, B. F. Qaqish, C. Livasy, L. A. Carey, E. Reynolds, L. Dressler, A. Nobel, J. Parker, M. G. Ewend, L. R. Sawyer, J. Wu, Y. Liu, R. Nanda, M. Tretiakova, A. Ruiz Orrico, D. Dreher, J. P. Palazzo, L. Perreard, E. Nelson, M. Mone, H. Hansen, M. Mullins, J. F. Quackenbush, M. J. Ellis, O. I. Olopade, P. S. Bernard & C. M. Perou (2006) The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics,* 7**,** 96.

Koscielny, S., M. Tubiana, M. Le, A. Valleron, H. Mouriesse, G. Contesso & D. Sarrazin (1984) Breast cancer: relationship between the size of the primary tumour and the probability of metastatic dissemination. *British journal of cancer,* 49**,** 709.

Lingjærde, O. C. & N. Christophersen (2000) Shrinkage structure of partial least squares. *Scandinavian Journal of Statistics,* 27**,** 459-473.

Liu, R., X. Wang, G. Y. Chen, P. Dalerba, A. Gurney, T. Hoey, G. Sherlock, J. Lewicki, K. Shedden & M. F. Clarke (2007) The prognostic role of a gene signature from tumorigenic breast-cancer cells. *N Engl J Med,* 356**,** 217-26.

Loi, S., B. Haibe-Kains, C. Desmedt, F. Lallemand, A. M. Tutt, C. Gillet, P. Ellis, A. Harris, J. Bergh & J. A. Foekens (2007) Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *Journal of Clinical Oncology,* 25**,** 1239.

Miller, L. D., J. Smeds, J. George, V. B. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar & E. T. Liu (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 13550.

Minn, A. J., G. P. Gupta, P. M. Siegel, P. D. Bos, W. Shu, D. D. Giri, A. Viale, A. B. Olshen, W. L. Gerald & J. Massague (2005) Genes that mediate breast cancer metastasis to lung. *Nature,* 436**,** 518-24.

Mook, S., M. K. Schmidt, G. Viale, G. Pruneri, I. Eekhout, A. Floore, A. M. Glas, J. Bogaerts, F. Cardoso, M. J. Piccart-Gebhart, E. T. Rutgers, L. J. Van't Veer & T. c. On behalf of the (2008) The 70-gene prognosis-signature predicts disease outcome in breast cancer patients with 1-3 positive lymph nodes in an independent validation study. *Breast Cancer Res Treat*.

Mook, S., L. J. Van't Veer, E. J. Rutgers, M. J. Piccart-Gebhart & F. Cardoso (2007) Individualization of therapy using Mammaprint: from development to the MINDACT Trial. *Cancer genomics & proteomics,* 4**,** 147.

Nuyten, D. S. A., T. Hastie, J. T. A. Chi, H. Y. Chang & M. J. van de Vijver (2008) Combining biological gene expression signatures in predicting outcome in breast cancer: An alternative to supervised classification. *European Journal of Cancer,* 44**,** 2319-2329.

Page, D. L. (1991) Prognosis and breast cancer: recognition of lethal and favorable prognostic types. *The American Journal of Surgical Pathology,* 15**,** 334.

Paik, S., S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant & N. Wolmark (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med,* 351**,** 2817-26.

Parker, J. S., M. Mullins, M. C. U. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He & Z. Hu (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology,* 27**,** 1160.

Pawitan, Y., J. Bjöhle, L. Amler, A. L. Borg, S. Egyhazi, P. Hall, X. Han, L. Holmberg, F. Huang & S. Klaar (2005) Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research,* 7**,** R953-R964.

Perou, C. M., S. S. Jeffrey, M. van de Rijn, C. A. Rees, M. B. Eisen, D. T. Ross, A. Pergamenschikov, C. F. Williams, S. X. Zhu, J. C. F. Lee, D. Lashkari, D. Shalon, P. O. Brown & D. Botstein (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences,* 96**,** 9212-9217.

Perou, C. M., T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lonning, A. L. Børresen-Dale, P. O. Brown & D. Botstein (2000) Molecular portraits of human breast tumours. *Nature,* 406**,** 747-52.

Perreard, L., C. Fan, J. F. Quackenbush, M. Mullins, N. P. Gauthier, E. Nelson, M. Mone, H. Hansen, S. S. Buys & K. Rasmussen (2006) Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. *Breast Cancer Res,* 8**,** R23.

Prat, A. & C. M. Perou (2009) Mammary development meets cancer genomics. *Nature medicine,* 15**,** 842-844.

Ravdin, P., L. Siminoff, G. Davis, M. Mercer, J. Hewlett, N. Gerson & H. Parker (2001) Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *Journal of Clinical Oncology,* 19**,** 980.

Reyal, F., M. Van Vliet, N. Armstrong, H. Horlings, K. De Visser, M. Kok, A. Teschendorff, S. Mook, L. Van't Veer & C. Caldas (2008) A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. *Breast Cancer Res,* 10**,** R93.

Rody, A., U. Holtrich, L. Pusztai, C. Liedtke, R. Gaetje, E. Ruckhaeberle, C. Solbach, L. Hanker, A. Ahr, D. Metzler, K. Engels, T. Karn & M. Kaufmann (2009) T-cell metagene predicts a favorable prognosis in estrogen receptor-negative and HER2-positive breast cancers. *Breast Cancer Res,* 11**,** R15.

Rønneberg, J. A., T. Fleischer, H. K. Solvang, S. H. Nordgard, H. Edvardsen, I. Potapenko, D. Nebdal, C. Daviaud, I. Gut, I. Bukholm, N. B., B.-D. A.L., T. J. & K. V. (2010) Methylation profiling with a panel of cancer related genes: Association with estrogen receptor, TP53 mutation status and expression subtypes in sporadic breast cancer. *Molecular Oncology*.

Rosen, P. P., S. Groshen, P. E. Saigo, D. W. Kinne & S. Hellman (1989) Pathological prognostic factors in stage I (T1N0M0) and stage II (T1N1M0) breast carcinoma: a study of 644 patients with median follow-up of 18 years. *Journal of Clinical Oncology,* 7**,** 1239.

Rousseeuw, P. J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics,* 20**,** 53-65.

Russnes, H., H. Vollan, O. Lingjærde, A. Krasnitz, P. Lundin, B. Naume, T. Sørlie, E. Borgen, I. Rye & A. Langerød (2010) Genomic Architecture Characterizes Tumor Progression Paths and Fate in Breast Cancer Patients. *Science Translational Medicine,* 2**,** 38ra47.

Scarff, R. W., H. Torloni & W. H. Organization. 1968. *Histological typing of breast tumours*. World Health Organization.

Slamon, D. J., W. Godolphin, L. A. Jones, J. A. Holt, S. G. Wong, D. E. Keith, W. J. Levin, S. G. Stuart, J. Udove & A. Ullrich (1989) Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science,* 244**,** 707.

Smyth, G. K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology,* 3**,** 3.

Smyth, G. K., J. Michaud & H. S. Scott (2005) Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics,* 21**,** 2067.

Sørlie, T., E. Borgan, S. Myhre, H. K. Vollan, H. Russnes, X. Zhao, G. Nilsen, O. C. Lingjærde, A. L. Børresen-Dale & E. Rødland (2010) The importance of gene-centring microarray data. *The Lancet Oncology,* 11**,** 719-720.

Sørlie, T., C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. Eystein Lonning & A. L. Børresen-Dale (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A,* 98**,** 10869-74.

Sørlie, T., R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lonning, P. O. Brown, A. L. Borresen-Dale & D. Botstein (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A,* 100**,** 8418-23.

Sotiriou, C. & L. Pusztai (2009) Gene-expression signatures in breast cancer. *New England Journal of Medicine,* 360**,** 790-800.

Sotiriou, C., P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart & M. Delorenzi (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst,* 98**,** 262-72.

Sparano, J. A. & S. Paik (2008) Development of the 21-gene assay and its application in clinical practice and clinical trials. *Journal of Clinical Oncology,* 26**,** 721.

Stephens, P. J., D. J. McBride, M. L. Lin, I. Varela, E. D. Pleasance, J. T. Simpson, L. A. Stebbings, C. Leroy, S. Edkins & L. J. Mudie (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature,* 462**,** 1005-1010.

Suzuki, R. & H. Shimodaira. 2004. An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: How accurate are these clusters. 34.

Teschendorff, A. E. & C. Caldas (2008) A robust classifier of high predictive value to identify good prognosis patients in ER-negative breast cancer. *Breast Cancer Res,* 10**,** R73.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)***,** 267-288.

Tibshirani, R., G. Walther & T. Hastie (2001) Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 63**,** 411-423.

Tusher, V. G., R. Tibshirani & G. Chu (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America,* 98**,** 5116.

van 't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards & S. H. Friend (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature,* 415**,** 530-6.

van de Vijver, M. J., Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend & R. Bernards (2002) A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med,* 347**,** 1999-2009.

van Vliet, M. H., F. Reyal, H. M. Horlings, M. J. van de Vijver, M. J. Reinders & L. F. Wessels (2008) Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability. *BMC Genomics,* 9**,** 375.

Vaske, C. J., S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler & J. M. Stuart (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics,* 26**,** i237.

Verweij, P. J. & H. C. Van Houwelingen (1993) Cross-validation in survival analysis. *Stat Med,* 12**,** 2305-14.

Verweij, P. J. & H. C. Van Houwelingen (1994) Penalized likelihood in Cox regression. *Stat Med,* 13**,** 2427-36.

Wang, Y., J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins & J. A. Foekens (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet,* 365**,** 671-9.

Weigelt, B., J. L. Peterse & L. J. van't Veer (2005) Breast cancer metastasis: markers and models. *Nature Reviews Cancer,* 5**,** 591-602.

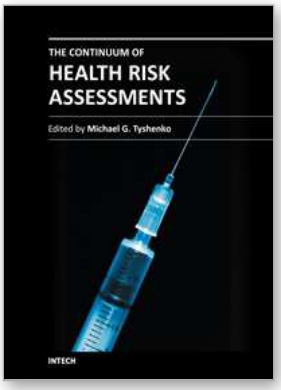
Wirapati, P., C. Sotiriou, S. Kunkel, P. Farmer, S. Pradervand, B. Haibe-Kains, C. Desmedt, M. Ignatiadis, T. Sengstag & F. Schutz (2008) Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res,* 10**,** R65.

Yu, J. X., A. M. Sieuwerts, Y. Zhang, J. W. Martens, M. Smid, J. G. Klijn, Y. Wang & J. A. Foekens (2007) Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer,* 7**,** 182.

Zhao, X., E. A. Rødland, T. Sørlie, B. Naume, A. Langerød, A. Frigessi, V. N. Kristensen, A. L. Børresen-Dale & O. C. Lingjærde (2011) Combining Gene Signatures Improves Prediction of Breast Cancer Survival. *PLoS One,* 6**,** e17845.

Zhao, X., E. A. Rødland, T. Sørlie, H. K. M. Vollan, V. N. Kristensen, O. C. Lingjærde & A. L. Børresen-Dale (Unpublished) Systematic assessment of prognostic gene signatures for breast cancer shows distinct influence of time and ER status.

Zou, H. & T. Hastie (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 67**,** 301-320.



The Continuum of Health Risk Assessments

Edited by Dr. Michael G. Tyshenko

ISBN 980-953-307-582-7
Hard cover, 194 pages
Publisher InTech
Published online 16, May, 2012
Published in print edition May, 2012

This book presents a collection of health risk assessments for known and emerging hazards that span a continuum. Case studies for existing health risks include psychoactive drug usage in delivery truck drivers and using look-back risk assessment for accidental syringe re-use in healthcare settings. Case studies for emerging risks include precautionary actions to safeguard blood supplies; nanoparticle deposition in the lung; and the epistemic issues surrounding genetically modified organism risk assessments. The final section of the book deals with advancing health risk assessment analyses through a post-genomics lens and provides case studies on personalized genomics, new data analyses and improving in silico models for risk assessment. These case studies provide much insight into the ongoing evolution of health risk assessments.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821