

Gene Pathways Associated With Prognosis and Chemotherapy Sensitivity in Molecular Subtypes of Breast Cancer

Takayuki Iwamoto, Giampaolo Bianchini, Daniel Booser, Yuan Qi, Charles Coutant, Christine Ya-Hui Shiang, Libero Santarpi, Junji Matsuoka, Gabriel N. Hortobagyi, William Fraser Symmans, Frankie A. Holmes, Joyce O'Shaughnessy, Beth Hellerstedt, John Pippen, Fabrice Andre, Richard Simon, Lajos Pusztai

Manuscript received June 3, 2010; revised November 19, 2010; accepted November 22, 2010.

Correspondence to: Lajos Pusztai, MD, DPhil, Department of Breast Medical Oncology, Unit 1354, The University of Texas M. D. Anderson Cancer Center, PO Box 301439, Houston, TX 77230-1439 (e-mail: lpusztai@mdanderson.org).

Background We hypothesized that distinct biological processes might be associated with prognosis and chemotherapy sensitivity in the different types of breast cancers.

Methods We performed gene set analyses with BRB-ArrayTools statistical software including 2331 functionally annotated gene sets (ie, lists of genes that correspond to a particular biological pathway or biochemical function) assembled from Ingenuity Pathway Analysis and Gene Ontology databases corresponding to almost all known biological processes. Gene set analysis was performed on gene expression data from three cohorts of 234, 170, and 175 patients with HER2-normal lymph node-negative breast cancer who received no systemic adjuvant therapy to identify gene sets associated prognosis and three additional cohorts of 198, 85, and 62 patients with HER2-normal stage I-III breast cancer who received preoperative chemotherapy to identify gene sets associated with pathological complete response to therapy. These analyses were performed separately for estrogen receptor (ER)-positive and ER-negative breast cancers. Interaction between gene sets and survival and treatment response by breast cancer subtype was assessed in individual datasets and also in pooled datasets. Statistical significance was estimated with permutation test. All statistical tests were two-sided.

Results For ER-positive cancers, from 370 to 434 gene sets were associated with prognosis ($P \leq .05$) and from 209 to 267 gene sets were associated with chemotherapy response in analysis by individual dataset. For ER-positive cancers, 131 gene sets were associated with prognosis and 69 were associated with pathological complete response ($P \leq .001$) in pooled analysis. Increased expression of cell cycle-related gene sets was associated with poor prognosis, and B-cell immunity-related gene sets were associated with good prognosis. For ER-negative cancers, from 175 to 288 gene sets were associated with prognosis and from 212 to 285 gene sets were associated with chemotherapy response. In pooled analyses of ER-negative cancers, 14 gene sets were associated with prognosis and 23 were associated with response. Gene sets involved in sphingolipid and glycolipid metabolism were associated with better prognosis and those involved in base excision repair, cell aging, and spindle microtubule regulation were associated with chemotherapy response.

Conclusion Different biological processes were associated with prognosis and chemotherapy response in ER-positive and ER-negative breast cancers.

J Natl Cancer Inst 2011;103:264-272

Breast cancer is a collection of molecularly and clinically distinct neoplastic diseases (1). The biggest molecular and clinical differences were observed between estrogen receptor (ER)-positive and ER-negative cancers that differ in the expression of thousands of genes and show distinct patterns of mutations and alterations in the DNA copy number (2-4). ER-negative cancers are more frequently of high histological grade, are more sensitive to chemotherapy, tend to metastasize to visceral organs, and recur earlier compared with ER-positive cancers (5-10). These two different types of breast cancers may arise from different epithelial precursor cells in the breast or may

represent different stages of differentiation arrest from a common stem cell (11,12).

In the past, biomarker discovery studies included all breast cancers in the analysis and the results were often contradictory between studies (13). In addition to numerous methodological issues, the conflicting results can be caused by the molecular heterogeneity of breast cancers. We hypothesize that different biological processes and molecular markers may be associated with prognosis and chemotherapy sensitivity in the different breast cancer subtypes. To examine this hypothesis, we performed gene set analyses to identify gene sets that are associated with prognosis

and chemotherapy sensitivity in ER-positive and HER2-normal (referred to hereafter as ER-positive) and ER-negative and HER2-normal (referred to hereafter as ER-negative) cancers separately. HER2-positive breast cancers were excluded from this study because survival and chemotherapy sensitivity of these patients are profoundly altered by the routine use of HER2-targeted treatments (14,15). We could not perform a separate analysis of HER2-positive cancers because of the small sample size of this cohort. We selected the ER- and HER2-based categorization of breast cancers because it is clinically relevant and widely used and because molecular class assignment has not been standardized. Individual breast cancers, especially if they are nonbasal tumors, can be assigned to different molecular classes depending on the method of classification used (16).

Gene set analysis is a statistical method to determine whether members of a particular gene set (ie, a list of genes that correspond to a particular biological pathway or biochemical function) preferentially occur toward the top or the bottom of a rank-ordered gene list where genes are ranked by the strength of their association with outcome (17,18). We tested 2331 functionally annotated gene sets that were assembled from Ingenuity Pathway Analysis (<http://www.ingenuity.com>) and Gene Ontology (<http://www.geneontology.org>) databases. These 2331 gene sets represent all known biological and metabolic pathways in human cells. The goals of this analysis were to determine which biological pathways are associated with prognosis and chemotherapy sensitivity in ER-positive and ER-negative cancers and whether these pathways were the same or different between these two different types of breast cancers. We used three different statistical comparison methods. 1) The primary analysis included identification of gene sets that were statistically significantly associated with outcome by ER status in individual patient datasets. We assessed the consistency of these findings across the different patient datasets. Consistent observations are more likely to represent true associations; however, an important limitation of this approach is that due to the small sample sizes of the individual patient datasets, these analyses have limited statistical power. 2) To increase the power of the analyses and to capture additional associations, we also performed pooled analyses of the prognostic datasets and of the neoadjuvant datasets. 3) Finally, we performed an interaction test to assess interaction between clinical outcome and gene sets by breast cancer subtype.

Materials and Methods

Patient Cohorts

Patient characteristics and datasets for the six cohorts studied are shown on Table 1. The prognostic analysis was performed on three different publicly available gene expression datasets from patients with stage I–II lymph node–negative breast cancers who received no systemic adjuvant therapy. We refer to these as the TRANSBIG (19), Wang (20), and Mainz (21) datasets. For gene set analyses, we compared gene expression data of patients with no distant metastasis to those that recurred. Only HER2-normal patients were included in our analysis, and HER2 overexpression was identified based on HER2 mRNA expression levels as described previously (22). In the final analysis, 234 patients were

CONTEXT AND CAVEATS

Prior knowledge

Estrogen receptor (ER)–positive and ER-negative cancers differ in the expression of thousands of genes and show distinct patterns of mutations and alterations in the DNA copy number.

Study design

Gene sets were obtained from two bioinformatics databases for this analysis, where a gene set is defined as a list of genes involved in a particular biological pathway or biochemical function. Three groups of patients with HER2-normal lymph node–negative breast cancer who received no systemic adjuvant therapy were studied to identify gene sets that were associated prognosis. Three additional groups of patients with HER2-normal stage I–III breast cancer who received preoperative chemotherapy were studied to identify gene sets associated with pathological complete response to therapy.

Contribution

More gene sets were associated with prognosis and with chemotherapy response in ER-positive breast cancers than ER-negative breast cancers in gene set analyses from both individual datasets and pooled datasets. In addition, the gene sets identified for these associations were different in the two types of breast cancer.

Implications

Different biological processes were associated with prognosis and chemotherapy response in ER-positive and ER-negative breast cancers.

Limitations

Sample sizes for individual ER-negative and ER-positive cancer data subsets were small. Patient in different neoadjuvant datasets were treated with different chemotherapy regimens.

From the Editors

from the Wang dataset (178 ER-positive and 56 ER-negative patients), 170 were from the TRANSBIG (124 ER-positive and 46 ER-negative patients), and 175 were from the Mainz (152 ER-positive and 23 ER-negative patients) datasets.

The chemotherapy sensitivity analysis was performed on three additional independent cohorts of patients with stage I–III breast cancer who received neoadjuvant chemotherapy (ie, the predictive datasets). These datasets are referred to as the MD Anderson Cancer Center/MicroArray Quality Control Consortium (MDACC/MAQC), MD Anderson Cancer Center/Institut Goussave Russy (MDACC/IGR), and US Oncology clinical trial 02103 (USO-02103). Results of the MDACC/MAQC dataset have been described previously (23) and include 198 HER2-normal patients (133 ER-positive and 65 ER-negative patients) treated with weekly paclitaxel (T) (80 mg/m²) for 12 treatments, followed by 5-fluorouracil (500 mg/m²), doxorubicin (50 mg/m²), and cyclophosphamide (500 mg/m²) for four treatments, given once every 21 days (FAC). The second cohort, the MDACC/IGR, included 85 HER2-normal patients (42 ER-positive and 43 ER-negative patients) treated with four courses of FAC chemotherapy. The third cohort, the USO-02103 included 62 HER2-normal patients (32 ER-positive and 30 ER-negative patients) who received four courses of 5-fluorouracil (500 mg/m²), epirubicin (100 mg/m²),

Table 1. Patient characteristics by study (or dataset) and breast cancer type*

Clinical and pathological information	Prognostic datasets						Neoadjuvant datasets					
	Wang		TRANSBIG		Mainz		MDACC/MAQC		MDACC/IGR		USO 02103	
	ER+/HER2-	ER-/HER2-	ER+/HER2-	ER-/HER2-	ER+/HER2-	ER-/HER2-	ER+/HER2-	ER-/HER2-	ER+/HER2-	ER-/HER2-	ER+/HER2-	ER-/HER2-
Patients, No.	178	56	124	46	152	23	133	65	41	41	32	29
Mean age, y (range)	—	—	46.5 (24-60)	45.2 (32-59)	—	—	51.9 (26-79)	51.4 (29-75)	49.7 (31-78)	47.3 (31-75)	47.5 (26-60)	50.0 (32-66)
Mean tumor size, mm (range)	—	—	20.7 (6-45)	24.7 (9-50)	20.2 (1-60)	26.2 (1-50)	20.1 (0-80)	13.6 (0-100)	—	—	63.8 (23-140)	69.0 (20-175)
Tumor size, † %												
T1	—	—	55.6	41.3	57.2	34.8	7.5	13.8	7.3	—	—	3.4
T2	—	—	43.4	58.7	41.4	65.2	60.2	49.2	56.1	51.2	34.4	31.0
T3	—	—	—	—	1.3	—	16.5	13.8	12.1	39.0	65.6	65.5
T4	—	—	—	—	—	—	15.0	21.5	24.4	9.8	—	—
Unknown	—	—	—	—	—	—	—	—	—	—	—	—
Lymph node status, ‡ %												
Negative	100	100	100	100	100	100	32.3	23.1	31.7	17.1	34.4	34.5
Positive	—	—	—	—	—	—	67.7	76.9	43.9	22.0	65.6	65.5
Unknown	—	—	—	—	—	—	—	—	24.4	61.0	—	—
Histological grade, %												
1	—	—	22.6	—	17.8	—	9.8	—	7.3	—	3.1	—
2	—	—	55.6	19.6	74.3	34.8	56.4	21.5	46.3	14.6	46.9	13.8
3	—	—	21.8	78.3	7.9	65.2	33.8	78.5	19.5	70.3	46.9	75.9
Unknown	—	—	—	2.1	—	—	—	—	26.8	14.6	—	10.3
Neoadjuvant chemotherapy	None	None	None	None	None	None	T/FAC	T/FAC	FAC or FEC	FAC or FEC	FEC/wTX	FEC/wTX
Adjuvant treatment	None	None	None	None	None	None	Variable	Variable	Variable	Variable	Variable	Variable
Event rate, § %	37.1	36.3	16.1	30.4	15.5	39.1	6.0	40.0	17.1	41.5	21.9	44.8
Data location (GEO accession)	GSE2034	GSE2034	GSE7390	GSE7390	GSE11121	GSE11121	GSE16716	GSE16716	GSE22093	GSE22093	GSE23988	GSE23988

* + = positive; - = negative; ER = estrogen receptor; FA(EG) = patients received chemotherapy regimen with 5-fluorouracil, doxorubicin (epirubicin), and cyclophosphamide; FEC/wTX = patients received chemotherapy regimen with four courses of 5-fluorouracil, doxorubicin, and cyclophosphamide, followed by four additional courses of weekly docetaxel and capecitabine; GEO = Gene Expression Omnibus; IGR = Institut Gustave Roussy; MAQC = MicroArray Quality Control project; MDACC = MD Anderson Cancer Center; T = tumor size according to TNM staging system; T/FAC = patients received chemotherapy regimen with paclitaxel and 5-fluorouracil, doxorubicin, and cyclophosphamide; USO 02103 = US Oncology PROTOCOL 02-103.

† Clinical tumor size for the MDACC/MAQC, MDACC/IGR, and USO 02103 studies and pathological tumor size for the Wang, TRANSBIG, and Mainz studies.

‡ Clinical lymph node status (MDACC/MAQC, MDACC/IGR, and USO 02103 studies) and pathological lymph node status (Wang, TRANSBIG, and Mainz studies).

§ Event rate = pathological complete response rate for the MDACC/MAQC, MDACC/IGR, and USO 02103 studies and distant event-free survival rate for the Wang, TRANSBIG, and Mainz studies.

and cyclophosphamide (500 mg/m²), given once every 21 days, followed by 12 weeks of docetaxel (35 mg/m²), given once weekly concomitant with capecitabine (850 mg/m² given twice daily for 14 days, repeated every 21 days) (FEC/wTX) (Table 1). Results from the last two datasets have not been previously published. For gene set analyses, we compared gene expression data of patients with a pathological complete response, defined as no invasive cancer after preoperative chemotherapy in the breast or lymph nodes, with those of patients with any residual invasive cancer.

All gene expression data were generated with Affymetrix U133A gene chips (Affymetrix, Inc, Santa Clara, CA), normalized with the MAS5 algorithm (<http://www.bioconductor.org>), with the mean expression centered to 600 (a standard value for pre-analytical processing of gene expression data) and log₂ transformed. Patients with an ESR1 mRNA expression (probe set 205225_at) level of greater than 10.18 were considered to be ER positive and those with a HER2 mRNA expression (probe set 216836_s_at) of greater than 12.54 were considered to be HER2 positive (22).

Bioinformatics and Statistical Analysis

Statistical analyses were performed with BRB-ArrayTools, version 3.9.0-Alfa (<http://linus.nci.nih.gov/BRB-ArrayTools.html>), and Ingenuity Pathway Analysis, version 7.6 (<http://www.ingenuity.com/>). Before gene set analysis, we removed probe sets that had average expression values of less than or equal to the lowest 15% of the expression distribution in each dataset to minimize noisy measurements. Gene sets with a minimum membership of 10 genes to a maximum of 100 genes were selected from the collection of gene sets corresponding to canonical cellular pathways in the Ingenuity Pathway Analysis database (<http://www.ingenuity.com>) and from the gene sets corresponding to all biological pathways in the Gene Ontology database (<http://www.geneontology.org>). The total number of gene sets that met these criteria and were included in this analysis was 2331 (2099 from Gene Ontology and 232 from Ingenuity Pathway Analysis databases). These 2331 gene sets represent essentially all pathways for known biological functions in eukaryotic cells. We used the Efron and Tibshirani (24) gene set analysis method to test whether gene sets were differentially expressed between the prognostic (recurrence vs no recurrence) and predictive (pathological complete response vs residual disease) outcome classes, with statistical significance being determined by a permutation test.

First, gene set analysis in each of the six datasets was performed separately for the ER-positive and ER-negative subset of tumors. In this analysis, each dataset served as an independent validation cohort for an observation made in the other datasets. Despite its simplicity, this analysis strategy had low power and high false-negative rates to detect statistically significant associations in the different ER subtypes because of the small sample sizes in each dataset and because of the unequal event rates across the subtypes (Table 1). To use the data more efficiently, we also calculated combined *P* values from the pooled prognostic and predictive datasets, respectively, by letting p_{ij} denote the *P* value for gene set *i* (where $i = 1-2331$) in the prognostic (or predictive) dataset *j* (where $j = 1, 2, \text{ or } 3$) for disease subtype *k* (where $k = 1$ for an ER-positive tumor and 2 for an ER-negative tumor), as calculated

by the Efron and Tibshirani (24) method. For each gene set *i*, we calculated a Fischer combined probability *P* value over the three datasets with 6 *df* (ie, two times the number of *P* values combined) by the following formula: $P_{ik} = -2[\ln(p_{i1k}) + \ln(p_{i2k}) + \ln(p_{i3k})]$, where ln is the natural logarithm (25).

An interaction test for clinical outcome and gene sets by ER subtype was also performed. The purpose of this test was to identify gene sets for which the differential expression among outcome classes (ie, good prognosis vs bad prognosis) is statistically significantly different between ER-negative breast cancers and ER-positive breast cancers. We computed a nominal statistical significance value, p_i , that measures differential expression between outcome classes for each gene *i* by using a univariate *t* test. These p_i values were computed separately for the two ER groups. We call them p_i values for group 1 (ER positive) and q_i values for group 2 (ER negative). Then for each gene set *k*, we calculated a summary statistic P_k of the p_i values for *i* in *k* for samples in group 1 by use of the maxmean statistic by Efron and Tibshirani (24) and a summary statistic Q_k of the q_i values for *i* in *k* for samples in group 2. We then tested the hypothesis that the two ER groups are equivalent with regard to between-class variations in gene expression for the genes in set *k*. We summarize the difference between groups using a test statistic $P_k - Q_k$. To obtain the null distribution of the test statistic $P_k - Q_k$, we randomly permuted the ER group labels, keeping the outcome class labels unchanged and recalculating $P[k]$, $Q[k]$, and the test statistic for the permuted data. We repeated 2000 times to tabulate the null permutation distribution of $P_k - Q_k$. We calculated a two-sided significance level for each gene set *k* using these null permutation distributions. The interaction *p* values for the three datasets were combined using Fischer combined probability method described above (24). All statistical tests were two-sided.

Results

Gene Sets Associated With Prognosis or Pathological Complete Response From Individual Datasets From ER-Positive and ER-Negative Breast Cancers

We first determined which gene sets were associated with prognosis in the three prognostic datasets separately by using the gene set enrichment analysis method. This analysis was done separately for ER-positive and ER-negative cancers. At a permutation test *P* value threshold of less than or equal to .05, the numbers of gene sets that were statistically significantly associated with prognosis were higher among ER-positive cancers (range = 370–434 gene sets, depending on the cohort) than among ER-negative cancers (range = 175–288 gene sets) (Supplementary Table 1, available online). Next, we examined the overlap among the gene sets that were statistically significantly associated with prognosis in the individual datasets (Figure 1). Eighty-seven gene sets were associated with prognosis among ER-positive cancers in all three prognostic datasets. Only one gene set was associated with prognosis in all three datasets in ER-negative cancers (corresponding to Gene Ontology gene set GO:0016769, nitrogen transfer activity), and 58 additional gene sets were common to at least two datasets. There was no overlap between the 87 gene sets that were consistently associated with prognosis in ER-positive cancers and

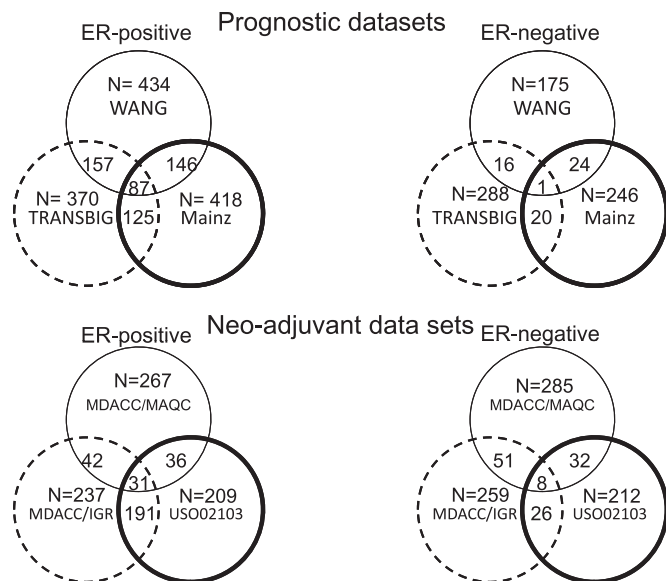


Figure 1. Venn diagrams of gene sets associated with prognosis and chemotherapy response in estrogen receptor (ER)-positive and ER-negative breast cancers. **Circles** represent individual datasets. The **outer portion of each circle** includes the total number of statistically significant gene sets, and the **overlapping portions** include the number of statistically significant ($P \leq .05$) prognostic or predictive (for response to chemotherapy) gene sets that are common to the respective datasets. A total of 2331 gene sets were assessed in the following six datasets with their Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>) data identification numbers in parentheses: Wang (GSE2034), TRANSBIG (GSE7390), Mainz (GSE11121), MD Anderson Cancer Center/MicroArray Quality Control Consortium (MDACC/MAQC; GSE22093), MD Anderson Cancer center/Institut Goustave Russy (MDACC/IGR; GSE22093), and US Oncology Protocol (USO; GSE23988).

the 59 gene sets that were associated with prognosis in two or more ER-negative cancers. In each dataset, the majority of gene sets that were associated with prognosis among ER-positive cancers were not statistically significantly associated with prognosis among ER-negative cancers and vice versa (Supplementary Table 2, available online).

We also examined what gene sets were associated with pathological complete response in each of the three the predictive datasets. The number of gene sets that were statistically significantly associated with pathological complete response (at a permutation test $P \leq .05$ level) was similar among ER-positive and ER-negative breast cancers in each of the three datasets (range = 209–267 gene sets in ER-positive cancers and 212–285 gene sets in ER-negative cancers) (Supplementary Table 3, available online). Among ER-positive cancers, 31 gene sets were predictive of chemotherapy response in all three datasets. Among ER-negative cancers, eight gene sets were predictive in all three cohorts, with an additional 93 gene sets that were common to at least two datasets (Figure 1). When the ER-positive predictive gene sets were compared with ER-negative predictive gene sets in individual datasets, very few were predictive in both ER-positive and ER-negative cancers (Supplementary Table 2, available online). This result indicated that most gene sets that were associated with chemotherapy response in ER-positive cancers were different from those associated with response in ER-negative cancers and vice versa. However, the limited overlap between these gene sets may

Table 2. Prognostic and predictive gene sets in pooled analyses*

Gene set and P value	No. of gene sets		
	ER+	ER-	Common Prognostic Neoadjuvant
Prognostic gene sets			
$\leq .05$	767	486	112
$\leq .01$	336	121	9
$\leq .005$	244	63	2
$\leq .001$	131	14	0
$\leq .0005$	99	8	0
Predictive gene sets			
$\leq .05$	524	573	85
$\leq .01$	202	163	2
$\leq .005$	140	105	0
$\leq .001$	69	23	0
$\leq .0005$	50	15	0
Prognostic and predictive gene sets with statistically significant P for interaction with ER status			
$\leq .05$			115
$\leq .01$			16
$\leq .005$			11
$\leq .001$			1
$\leq .0005$			1

* + = positive; - = negative; ER = estrogen receptor; P = Fischer combined probability P value. All statistical tests were two-sided.

also be caused by the limited and variable power of each analysis, which could lead to false-negative results in the individually small study cohorts.

Pooled Analysis of Predictive and Prognostic Gene Sets

To increase the power of the analysis and to identify gene sets associated with outcome that may have been missed when datasets were analyzed separately (ie. to reduce false-negative results), we pooled all prognostic datasets and recalculated statistical significance for gene sets associated with recurrence in ER-positive and ER-negative cancers, respectively. In this pooled analysis among ER-positive cancers, 767 gene sets were statistically significantly associated with recurrence at Fischer combined probability P values of less than or equal to .05 and 131 gene sets were statistically significantly associated with recurrence at P values of less than or equal to .001 (Table 2). The 131 gene sets included 80 of the 87 prognostic gene sets that were also identified as prognostic in ER-positive cancers in all three datasets when these were examined separately. Among these 80 prognostic gene sets, four were consistently overexpressed in ER-positive cancers with good prognosis and 76 were overexpressed in ER-positive cancers with poor prognosis (Supplementary Table 4, available online). Similar pooled analysis for ER-negative cancers identified 486 gene sets that were statistically significantly associated with prognosis at a P threshold level of less than .05 and 14 gene sets that were statistically significantly associated with prognosis at a P threshold level of less than or equal to .001. These 14 gene sets included nine that were consistently overexpressed in ER-negative cancers with good prognosis and five showed statistically significant but inconstant directional association with prognosis (Supplementary Table 4,

available online). There was no overlap between the ER-positive and ER-negative prognostic gene sets at a combined probability P threshold level of less than or equal to .001 and only a 9% overlap of gene sets at a level of less than or equal to .05.

We also performed a similar pooled analysis on the neoadjuvant datasets (Table 2). For ER-positive cancers, 524 gene sets were statistically significantly associated with pathological complete response at a combined probability P value of less than or equal to .05 and 69 gene sets were statistically significantly associated with pathological complete response at a P value of less than or equal to .001. These 69 gene sets included 28 of the 31 gene sets that were predictive of pathological complete response in ER-positive cancers in all three datasets when analyzed individually. Of the 31 predictive gene sets that were identified by both the pooled and individual analyses in ER-positive cancers, only five were consistently overexpressed in cancers with pathological complete response in all three datasets. The remaining gene sets showed association with pathological complete response in some datasets but were associated with residual disease in other datasets (Supplementary Table 5, available online). This result indicated that there was much less consistency in these observations than in those identifying prognostic genes sets in ER-positive cancers. For ER-negative cancers in the pooled analysis of the three datasets, 573 gene sets were statistically significantly associated with pathological complete response at a P value of less than or equal to .05 and 23 gene sets were statistically significantly associated with pathological complete response at a P value of less than or equal to .001. These 23 gene sets included 20 of the 93 predictive gene sets that were detected in at least two of the three datasets when the analysis was performed separately for each dataset. Of these 20 gene sets, three were consistently overexpressed in ER-negative cancers, with a pathological complete response in all three datasets, 13 were overexpressed in ER-negative cancers that had residual cancer, and the remaining gene sets showed inconsistent associations across datasets (Supplementary Table 5, available online). There was no overlap between the ER-positive and ER-negative predictive gene sets at a P value of less than or equal to .001, and only 8% of gene sets were shared (ie, were predictive in both ER-negative and ER-positive cancers) at P value of less than or equal to .05.

Next, we tested interaction terms between gene sets and clinical outcome by ER status in the prognostic and predictive datasets. In the pooled prognostic dataset, 115 gene sets showed interaction with survival by ER status at a significance level of less than or equal to .05 by use of an interaction test to determine statistical significance and 16 gene sets showed interaction with survival by ER status at a statistical significance level of less than or equal to .01 (Table 2 and Supplementary Table 6, available online). In the pooled predictive dataset, 121 gene sets showed a statistically significant interaction with response by receptor status at a statistical significance level of less than or equal to .05 and 22 gene sets showed a statistically significant interaction with response by receptor status at a statistical significance level of less than or equal to .01 (Table 2 and Supplementary Table 7, available online). These results supported the finding that different gene sets are associated with good survival in ER-positive and ER-negative cancers. Similarly, different gene sets were associated with pathological complete response in ER-positive and ER-negative cancers.

Biological Functions Associated With Prognosis and Response to Chemotherapy Among ER-Positive and ER-Negative Breast Cancers

We used the Gene Ontology database functional annotation and the Ingenuity Pathways pathway mapping software to describe and plot the biological functions of the gene sets that were associated with clinical outcome in ER-positive or ER-negative cancers. On the basis of the Gene Ontology annotation, the 76 gene sets that were consistently associated with cancer recurrence (ie, poor prognosis) in ER-positive cancers were all involved in the regulatory and effector functions of cell division, including DNA replication, chromosome condensation and segregation, mitotic spindle formation, microtubule motor functions, and the regulation of G₁, S, and M phases of the cell cycle. The four gene sets that were associated with good prognosis in ER-positive cancers involved functions associated with B-cell and humoral immunity and complement activation. We used Ingenuity Pathway Analysis software tool to map all unique genes that consisted the 80 different prognostic gene sets in ER-positive cancers to cellular pathways. The most statistically significant cellular pathway, which also showed the highest membership hits, was the aryl hydrocarbon receptor pathway; other statistically significant cellular pathways that predicted prognosis in ER-positive cancers were related to cell cycle control (Figure 2). The aryl hydrocarbon receptor is a ligand-dependent transcription factor that can be activated by aromatic hydrocarbons (eg, dioxin) and regulates the transcription of genes encoding xenobiotic metabolizing enzymes (eg, cytochrome P450) and genes that regulate the cell cycle (eg, p21, p27, and cyclins A and E). It also inhibits the transcription factor activity of ER (26). Interestingly, the aryl hydrocarbon receptor itself is not highly expressed in ER-positive breast cancers (data not shown) and the association between this pathway and prognosis is primarily caused by the downstream cell cycle-related components of the pathway (Supplementary Figure 1, A, available online). Thus, the unifying theme of the prognostic pathways that were identified in ER-positive cancers through the Ingenuity Pathway Analysis was that almost all of the genes associated with poor prognosis were involved with DNA synthesis and cell proliferation.

The five gene sets that were consistently predictive of chemotherapy response in ER-positive breast cancers were involved in microtubule motor activity and cell cycle regulation on the basis of their Gene Ontology annotation. Ingenuity Pathway Analysis also identified chemokine signaling pathways in addition to cell cycle-related cellular pathways as being predictive of chemotherapy response in ER-positive cancers (Figure 3). The expression of genes involved in signaling from chemokine receptor-3 (*CCR3*, primarily expressed in T cells and eosinophils), chemokine receptor-5 (*CCR5*, primarily expressed in T cells and macrophages), and interleukin-8 (a major mediator of inflammatory response and a chemotactic factor that attracts neutrophils, basophils, and T cells) were higher in highly chemotherapy-sensitive ER-positive cancers.

In ER-negative breast cancers, all nine statistically significant prognostic gene sets that were consistent across datasets were associated with better prognosis and were involved in glyco- and sphingolipid metabolism, according to functional annotation in

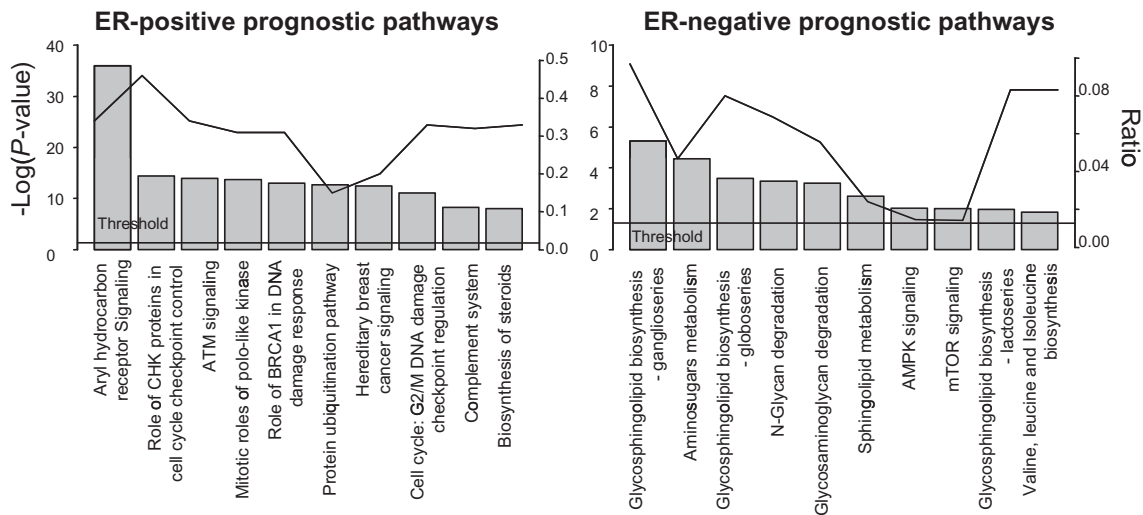


Figure 2. Cellular pathways associated with prognosis in different types of breast cancer. **A)** Estrogen receptor (ER)-positive breast cancers. **B)** ER-negative breast cancers. All unique genes from all statistically significant prognostic gene sets were examined by an Ingenuity Pathway Analysis, and data for the 10 most statistically significant pathways are presented. The left y-axis corresponds to data for the bars;

these data are logarithm of *P* values that were calculated by Fisher exact test, with a threshold for statistical significance set at .05. The right y-axis corresponds to data in the line graphs; these data are the ratio of the number of molecules in a given pathway that meet the twofold change cutoff criteria in either direction divided by total number of molecules that make up that pathway.

the Gene Ontology database. However, pathway-level associations for these genes by Ingenuity Pathway Analysis were modest, with strong associations observed only for a few genes that were members of much larger metabolic networks (Supplementary Figure 1, B and C, available online). No gene set was consistently and statistically significantly associated with poor prognosis among ER-negative cancers in these datasets. A lack of association between any gene set involved in cell proliferation and poor prognosis was notable in ER-negative cancers, and this result was quite different from that in ER-positive cancers. High sensitivity to chemotherapy in ER-negative cancers was consistently associated

with higher expression of gene sets that included microtubule spindle formation, DNA repair, and cellular aging Gene Ontology categories. Poor response to chemotherapy was associated with gene sets that included various cellular functions, such as metabolism of xenobiotics, lipid metabolism, and G-protein-coupled receptor signaling. Ingenuity Pathway Analysis applied to the unique genes included in these predictive gene sets also identified G-protein signaling, fatty acid, and xenobiotic metabolism as statistically significant pathways associated with chemotherapy sensitivity (Figure 3 and Supplementary Figure 1, D and E, available online).

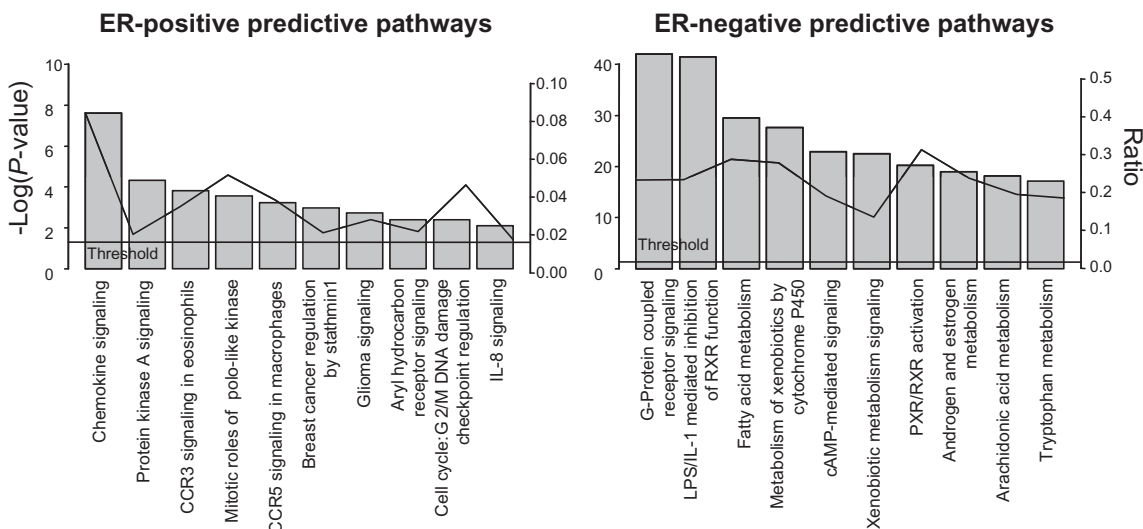


Figure 3. Cellular pathways associated with chemotherapy response in different types of breast cancer. **A)** Estrogen receptor (ER)-positive breast cancers. **B)** ER-negative breast cancers. All unique genes from all statistically significant predictive gene sets were examined by an Ingenuity Pathway Analysis, and data for the 10 most statistically significant pathways are presented. The left y-axis corresponds to data for

the bars; these data are the logarithm of *P* values that were calculated by Fisher exact test, with a threshold for statistical significance set at .05. The right y-axis corresponds to data in the line graphs; these data are the ratio of the number of molecules in a given pathway that meet the twofold change cutoff criteria in either direction divided by total number of molecules that make up that pathway.

Discussion

Our findings support the hypothesis that distinct gene pathways and biological processes are associated with prognosis and chemotherapy sensitivity in different subtypes of breast cancers. Previous attempts to discover prognostic or predictive markers for breast cancer usually analyzed all patients together. More recently, it is increasingly recognized that individual biomarkers (eg, Ki 67, expression of the microtubule-associated protein tau) may have different predictive or prognostic values in ER-positive and ER-negative cancers and different markers may be predictive of the same outcome in different types of breast cancers (19,27–29). In this large-scale systematic analysis of all known biological processes, we found that gene pathways that are associated with prognosis in ER-positive and ER-negative cancers were substantially different. Interestingly, many more pathways were associated with prognosis in ER-positive cancers than in ER-negative cancers, and these associations were consistently observed across multiple independent datasets. We identified fewer prognostic gene sets in ER-negative cancers than in ER-positive cancers and those gene sets were also less consistent across datasets. Thus, it appears to be easier to discover mRNA-based biomarkers of prognosis for ER-positive cancers than for ER-negative tumors. Indeed, many prognostic markers have been developed for this group of patients, including several that are used routinely in clinical practice (eg, Oncotype DX Recurrence Score; Genomic Health, Inc, Redwood City, CA; and MammaPrint; Agendia, Inc, Amsterdam, the Netherlands) (30). No similar, robust, and validated prognostic marker exists for ER-negative cancers. Because fewer gene sets are associated with prognosis in this group and because the associations are often inconsistent between datasets, a substantially larger sample size than is currently available may be needed to identify robust mRNA profile-based biomarkers that are associated with prognosis for ER-negative cancers.

Importantly, there was limited overlap among the prognostic gene sets for ER-positive cancers and for ER-negative cancers, which indicates that different biological processes appear to be associated with prognosis in these different types of cancers. At the functional level in ER-positive cancers, gene sets that were involved in mitosis, cell cycle, DNA replication, and chromosome duplication pathways were associated with poor prognosis; and gene sets that were involved with immune and inflammatory responses were associated with good prognosis. Proliferation-related gene sets were also associated with increased chemotherapy sensitivity in ER-positive breast cancers but not in ER-negative breast cancers. This result is consistent with the clinical phenomenon that ER-positive breast cancers that have the worst prognosis without systemic adjuvant therapy include many cancers that are the most sensitive to chemotherapy.

In ER-negative cancers, gene sets that were involved in glycolipid, sphingolipid, and fucose metabolism pathways were associated with better prognosis. It is important to note that the individual genes in these gene sets were only modestly associated with prognosis and so these pathway-level associations would not have been identified through gene-by-gene comparisons. It is biologically plausible that these metabolic pathways are involved with the prognosis of ER-negative cancers. Fucose-containing glycans

play important roles in selectin-mediated cell adhesion, and the fucosylation has been shown to modulate signaling by various receptors, including those in the epidermal growth factor receptor and Notch receptor families (31). Sphingolipids also play important roles in cell signaling and regulation of apoptosis (32). The cellular pathways in ER-negative cancers that were most consistently associated with resistance to chemotherapy contained genes that are involved in fatty acid and xenobiotic metabolism, including a repertoire of cytochrome P450 enzymes and components of G-protein-coupled signaling. These observations indicate that it might be possible to increase the sensitivity of ER-negative cancers to chemotherapy by inhibiting CYP450 enzymes or G-protein-mediated signaling.

This study had several limitations. The individual sample sizes for ER-negative and ER-positive cancer subsets were small in each of the prognostic and predictive datasets. Patient included in the three different neoadjuvant datasets were treated with different chemotherapy regimens in each of the datasets. The limited consistency between predictive gene sets across datasets for any given ER cohort may, therefore, indicate that some gene sets represented treatment type-specific associations or may reflect false discoveries because of small sample sizes and limited power. Unfortunately, any analysis of currently publicly available breast cancer microarray data is affected by the uneven sample sizes of the molecular subsets included in these datasets. The small subset sizes and the variable event rates across the subsets increase the risk for false-negative findings and spurious conclusion of differential marker effect by subgroup. Different sample preparation and pre-analytical variations from dataset to dataset can also reduce reproducibility and generalizability of observations across datasets.

Despite these limitations, overall our observations are consistent with the hypothesis that prognosis and chemotherapy response are associated with different biological processes in ER-positive and ER-negative breast cancers. Future biomarker studies will need to examine the prognostic and predictive value of proposed markers separately and prospectively in different breast cancer subtypes.

References

1. Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. *N Engl J Med*. 2009;360(8):790–800.
2. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747–752.
3. Andre F, Job B, Dessen P, et al. Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array. *Clin Cancer Res*. 2009;15(2):441–451.
4. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*. 2001;98(19):10869–10874.
5. Hess KR, Pusztai L, Buzdar AU, Hortobagyi GN. Estrogen receptors and distinct patterns of breast cancer relapse. *Breast Cancer Res Treat*. 2003; 78(1):105–118.
6. Liedtke C, Mazouni C, Hess KR, et al. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *J Clin Oncol*. 2008;26(8):1275–1281.
7. Rouzier R, Perou CM, Symmans WF, et al. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin Cancer Res*. 2005;11(16):5678–5685.
8. Rakha EA, El-Sayed ME, Green AR, et al. Biologic and clinical characteristics of breast cancer with single hormone receptor positive phenotype. *J Clin Oncol*. 2007;25(30):4772–4778.

9. Teschendorff AE, Naderi A, Barbosa-Morais NL, et al. A consensus prognostic gene expression classifier for ER positive breast cancer. *Genome Biol.* 2006;7(10):R101.
10. Paik S, Tang G, Shak S, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J Clin Oncol.* 2006;24(23):3726–3734.
11. Polyak K. Breast cancer: origins and evolution. *J Clin Invest.* 2007;117(11):3155–3163.
12. Sims AH, Howell A, Howell SJ, Clarke RB. Origins of breast cancer subtypes and therapeutic implications. *Nat Clin Pract Oncol.* 2007;4(9):516–525.
13. Harris L, Fritsche H, Mennel R, et al. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *J Clin Oncol.* 2007;25(32):5287–5312.
14. Slamon DJ, Leyland-Jones B, Shak S, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that over-expresses HER2. *N Engl J Med.* 2001;344(11):783–792.
15. Romond EH, Perez EA, Bryant J, et al. Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *N Engl J Med.* 2005;353(16):1673–1684.
16. Weigelt B, Mackay A, A'Hern R, et al. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol.* 2010;11(4):339–349.
17. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005;102(43):15545–15550.
18. Jiang Z, Gentleman R. Extensions to gene set enrichment. *Bioinformatics.* 2007;23(3):306–313.
19. Buyse M, Loi S, van't Veer L, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst.* 2006;98(17):1183–1192.
20. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet.* 2005;365(9460):671–679.
21. Schmidt M, Bohm D, von Torne C, et al. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res.* 2008;68(13):5405–5413.
22. Gong Y, Yan K, Lin F, et al. Determination of oestrogen-receptor status and ERBB2 status of breast carcinoma: a gene-expression profiling study. *Lancet Oncol.* 2007;8(3):203–211.
23. Shi L, Campbell G, Jones WD, et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol.* 2010;28(8):827–838.
24. Efron B, Tibshirani R. On testing the significance of sets of genes. *Ann Appl Stat.* 2007;1(1):107–129.
25. Fisher RA. *Statistical Methods for Research Workers.* Edinburgh, UK: Oliver and Boyd; 1925.
26. Marlowe JL, Puga A. Aryl hydrocarbon receptor, cell cycle regulation, toxicity, and tumorigenesis. *J Cell Biochem.* 2005;96(6):1174–1184.
27. Yu JX, Sieuwerts AM, Zhang Y, et al. Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer.* 2007;7:182.
28. Tordai A, Wang J, Andre F, et al. Evaluation of biological pathways involved in chemotherapy response in breast cancer. *Breast Cancer Res.* 2008;10(2):R37.
29. Pusztai L, Jeong JH, Gong Y, et al. Evaluation of microtubule-associated protein-tau expression as a prognostic and predictive marker in the NSABP-B28 randomized clinical trial. *J Clin Oncol.* 2009;27(26):4287–4429.
30. Marchionni L, Wilson RF, Wolff AC, et al. Systematic review: gene expression profiling assays in early-stage breast cancer. *Ann Intern Med.* 2008;148(5):358–369.
31. Becker DJ, Lowe JB. Fucose: biosynthesis and biological function in mammals. *Glycobiology.* 2003;13(7):41R–53R.
32. Spiegel S, Milstien S. Sphingosine 1-phosphate, a key cell signaling molecule. *J Biol Chem.* 2002;277(29):25851–25854.

Funding

The Breast Cancer Research Foundation (L.P. and W.F.S.), The MD Anderson Cancer Center Faculty Incentive Funds (W.F.S.), and the Commonwealth Cancer Foundation (L.P. and W.F.S.). C.C. received support from l'Association pour la Recherche sur le Cancer, Lavoisier program le Ministère Français des Affaires Étrangères et Européennes, le Collège National des Gynécologues et Obstétriciens Français, and la Fondation de France and the la Fédération Nationale des Centres de Lutte Contre le Cancer.

Notes

The authors had full responsibility for the design of the study; the collection, the analysis, and interpretation of the data; the decision to submit the article for publication, and the writing of the article.

Dr. F. A. Holmes is a member of the speaker's bureau for Novartis (makers of Zometa) and Genentech (makers of Avastin and Herceptin).

Affiliations of authors: Department of Breast Medical Oncology (TI, GB, DB, CC, CY-HS, LS, GNH, LP), Department of Bioinformatics and Computational Biology (YQ), and Department of Pathology (WFS), The University of Texas M. D. Anderson Cancer Center, Houston, TX; Department of Gastroenterological Surgery and Surgical Oncology, Okayama University, Okayama, Japan (TI, JM); Translational Research Unit, Department of Oncology, Hospital of Prato and Istituto Toscano Tumori, Prato, Italy (LS); Texas Oncology, US Oncology, Houston, TX (FAH); Baylor Sammons Cancer Center, Texas Oncology, US Oncology, Dallas, TX (JO, JP); Texas Oncology, US Oncology, Austin, TX (BH); Department of Medical Oncology, Institute Gustave Roussy, Villejuif, France (FA); Biometrics Branch, National Cancer Institute, Rockville, MD (RS).