Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Liu R, Wang X, Chen GY, et al. The prognostic role of a gene signature from tumorigenic breast-cancer cells. N Engl J Med 2007;356:217-26.

SUPPLEMENTARY APPENDIX

to

"The prognostic role of a gene signature from tumorigenic breast cancer cells."

by

Rui Liu, Xinhao Wang, Grace Y. Chen, Piero Dalerba, Austin Gurney, Timothy Hoey, Gavin Sherlock, John Lewicki, Kerby Shedden and Michael F. Clarke

Supplementary MATERIALS AND METHODS

Isolation of tumorigenic breast cancer cells (TBCC), non-tumorigenic breast cancer cells (NTCC) and normal breast epithelium (NBE). $ESA^{+}CD44^{+}CD24^{-/low}Lin^{-}$ TBCC and $CD24^{+}Lin^{-}$ NTCC were isolated by fluorescence activated cell sorting (FACS) from either primary malignant pleural effusions (n = 3) or human breast tumors grown as solid xenografts in immunodeficient NOD-SCID mice (n = 3). Origin and characteristics of the breast cancer tissue samples used in this study are listed in ***Suppl. Table 1***. All primary tissue samples were collected under a protocol approved by the University of Michigan institutional review board (IRB) between 2000 and 2003. Xenograft implantation and TBCC purification procedures were performed as previously described[1]. Normal breast epithelium (NBE) tissue was obtained from reduction mammoplasties. Tissue samples were processed within an hour after breast reduction surgery. The tissue was cut into small pieces, and the pieces were then minced with a blade to yield 2~3 mm^3 pieces. After washing with HBSS twice, minced tissue was dissociated with 200U/ml of type III collagenase (Worthington biochemical corp., Lakewood, NJ) in Medium 199 at 37°C for about 2 hrs. During the incubation, tissue was pipetted every 30 min. Dissociation was stopped by adding 5% FBS and cells were diluted with Medium-199, then filtered sequentially through a sterile 100-µm nylon mesh and a 40-µm cell strainer (Falcon) to obtain a single cell suspension. Cells were then washed twice with HBSS + 2% HICS (heat-inactivated calf serum). Staining of cells for flow cytometry followed the procedure described previously[1]. In brief, except anti-human ESA-FITC was from Biomeda corparation (Foster City, CA), all the other antibodies including anti-human CD10-PE, anti-CD45-PE-Cy5, -CD31, -CD64, and -CD140b as well as Streptavidin-PE-Cy5 conjugate were all purchased from BD pharmingen (San Diego, CA). Dead cells were eliminated by using 7-AAD (7-aminoactinomycin D, Molecular probes) at 1 µg/ml of final concentration. Flow cytometry was performed on a FACSVantage (Becton Dickinson). ESA^+ and/or $CD10^+$, Lin- cells were sorted twice, and the purity of double sorted cells was >95%.

RNA amplification and microarray analysis. For each sample, 10K to 35K fresh TBCC or NBE cells were sorted into eppendorf tubes containing 0.9ml of Trizol, up to a total volume of 1ml. Then, 20µg glycogen (Roche Diagnostic Corporation, Indianapolis, IN) and 10µg of linear acrylamide (Ambion, Austin, TX) were added to each sample, and total RNA was extracted following the manufacturer's protocol. RNA was treated with 1.5µl of RNase-free DNase I (2U/ml, Ambion, Austin, TX) in the presence of 1.5µl RNase inhibitor (10U/ml, Invitrogen) in 35µl reaction volume at 37°C for 20mins. To maximize purity, RNA was re-extracted again in Trizol and subjected to a second round of purification. RNA was quantified with the RiboGreen RNA Quantitation Kit (Molecular Probes Inc., Eugene, Oregon), and 100ng of total RNA were used in the first reverse transcription of two consecutive rounds of linear amplification with the protocol previously described by Baugh *et al.*[2] and Iwashita *et al.* [3]. Biotinylated cRNA was fragmented according to the Affymetrix technical manual. 15μg of cRNA from each sample was hybridized per chip to Affymetrix HG_U133 A or B chips. Hybridization, washing and scanning were done following the manufacturer's instruction.

Real-time PCR. Differences in gene expression levels identified by microarray experiments were validated by Taqman real-time PCR (Applied Biosystems, Foster City, California, USA) on three TBCC samples and one NBE sample. Briefly, cDNA was prepared from total RNA using the random hexamer method of reverse transcription, according to the manufacturer's instructions (Applied Biosystems). In a custom 384 well low density array format, 96 target gene assays were incorporated for two samples (two assays were used for a single gene in each sample). Fourteen genes were selected from the 186-gene signature. B2M (beta-2 microglobulin) was used as endogenous control. Real-time PCR was performed in triplicate using the comparative C_T method on the ABI Prism 7900HT Sequence Detection System according to the manufacturer's instructions. For each loading pool (48 PCR reactions), cDNA synthesized from 40 ng total RNA of each sample was used as template and reaction volume per spot is 2 μl.

Microarray data normalization. To obtain an expression measure for a given probe set, the mismatch hybridization values were subtracted from the perfect match values, and the average of the middle 50% of these differences was used as the expression measure for the probe set. A quantile normalization procedure was then applied to adjust for differences in the probe intensity distribution across different chips. Specifically, we applied a monotone linear spline to each chip that mapped quantiles 0.01 up to 0.99 (in increments of 0.01) exactly to the corresponding quantiles of a standard chip. The transform log2 [200 + max(X;0)] was then applied.

Generation of the "Invasiveness Gene Signature" (IGS). A list of genes showing significant differential expression between TBCC and NBE was generated using following criteria: 1) the difference of average signal intensity TBCC and NBE must be 2 fold or more; 2) if up-regulated, average signal intensity in TBCC samples must be larger than 300, compared to an approximate mean of 1000 for each array, and average P value (present or absent P value) must be smaller than 0.01; 3) if down-regulated, criteria identical to those described above applied to the expression values from NBE samples (average signal intensity in NBE > 300, and $p<0.01$). Genes included in the list were further selected based on the p-value of a t-test comparing differences in expression values between TBCC and NBE samples: genes with a p-value ≥ 0.005 were eliminated, leaving a list of 186 differentially expressed genes. Average expression levels of genes in the 6 TBCC samples were divided by average expression level of those in NBE samples, and resulted ratios were transformed into log values and used as expression levels for genes in gene signature. For cases where expression values from 3 paired primary TBCC or non-tumorigenic cancer cell (NTCC) samples were used for gene signature, the log2 ratios of average expression levels in 3 tumorigenic cancer cell or non-tumorigenic cancer cell samples divided by those in normal breast samples were used as expression levels for genes in gene signature. False discovery rate (FDR) was controlled using the Benjamini-Hochberg procedure[4, 5] applied to the p-values of corresponding t-tests. Using the above criteria, the FDR is less than 5% for the genes in the list. For Affymetrix arrays, latest annotation files (4/2005) were downloaded from Affymetrix web site and used for all further analysis. For Rosetta/NKI oligonucleotide array, Oligo sequences were downloaded from Rosetta website and a blast search was performed between oligos and sequences from NCBI Genes database to annotate the array. Array elements from different array platforms were mapped to each other by gene symbols.

Patient datasets transformation. Patient datasets used in this study were downloaded from the two following web-sites: http://microarray-pubs.stanford.edu/wound_NKI/ and http://microarray-pubs.stanford.edu/wound. For the Affymetrix data in the downloaded patient datasets, signal intensity values for each probe were transformed into log ratios by dividing each value by the average intensity of the probe across all samples within each dataset. Probes with an average signal intensity value smaller than 20 across all samples within the dataset were filtered out. If the signal intensity value for a probe in a given dataset was less than 20, then it was converted to 20. For the three datasets related to early lung cancer, prostate cancer and medulloblastoma, if a signal intensity value was larger than 16000, then it was converted to 16000, following the same convention used by other authors [6].

Generation of receiver operating characteristic (ROC) curves. For the generation of receiver operating characteristic (ROC) curves, the correlation coefficient of each patient to the gene signature was used as the diagnostic value and the curve was generated using the GraphPad Prism software, version 4.03 (GraphPad Software Inc., San Diego, CA).

Supplementary RESULTS

Validation of the differential expression of IGS genes. Taqman real time PCR was applied in a custom 384 well low density array format, 96 target gene assays were incorporated for two samples (two assays were used for a single gene in each sample). Fourteen genes were selected from the breast tumorigenic gene signature. B2M (beta-2 microglobulin) is used as endogenous control. Real-time PCR was performed in duplicate arrays for each sample using the comparative CT method.

To validate the differential expression of these 186 genes, we randomly selected 14 genes and performed real time PCR. Due to the limited availability of primary breast tumor samples, we were able to perform real-time PCR validation in only three TBCC samples from xenografts and one NBE sample. Eight genes (AMMECR1, PLP2, MAPK14, HS2ST1, KDELR3, PDE8A, ISGF3G and GAPD) were up-regulated in the gene signature, while the other 6 genes (IRX3, CEBPD, BCL2, GOLGIN-67, MGP, LTF) were down-regulated. As shown in ***Suppl. Table 2***, in two of the three TBCC samples (T1 and T2) all 14 genes displayed expression patterns consistent with those observed by microarray analysis. In the third TBCC sample (T3), expression levels were consistent with those from microarray data in 8 out of 14 tested genes. Overall, these results showed that gene expression patterns observed by microarray analysis could be confirmed on individual tumor samples by real-time PCR, thus indicating that the IGS can be considered as a reliable depiction of the transcriptional differences between TBCC and NBE in our sample set.

Functional annotation of the IGS. The IGS contains a list of 186 genes which is substantially different from other previously published breast cancer gene signatures. The overall annotation of the 186 genes in the IGS is summarized in **Table 1**. Based on the Gene-Ontology (GO) terms to which these genes are annotated, the biological role of 47 genes in the 186-gene signature is unknown. The remaining 139 genes are classified into more than 20 categories. To evaluate whether the signature was enriched in genes that are coordinately involved in specific biochemical pathways or cellular functions, we performed an advanced annotation study using the Affymetrix GO Mining Tool. With a Bonferroni corrected Chi-square p-value <0.05, the enriched genes were grouped into up-regulated and down-regulated genes, which were further annotated under three major categories (***see also the supplementary file "Gene Annotation of the IGS")***. Among the up-regulated genes, we identified three groups with well-documented functions:

1) ***genes involved in the "Positive regulation of IkB/NFkB cascade"***: ECOP, CASP8 and TICAM2. ECOP (EGFR Co-amplified and Over-expressed Protein) is a novel protein that up-regulates NF*k*B transcriptional activity and promotes cellular resistance to apoptotic challenge [7]. Overexpression of TICAM2 (also called TIRP) is reported to activate NF*k*B and potentiate IL-1 receptor-mediated NF*k*B activation [8].

CASP8-dependent activation of NF*k*B initiates proliferation of resting T cells [9]. These three genes are all involved in NF*k*B activation. The activation of I*k*B/NF*k*B has been observed in multiple human cancers and is thought to promote tumorigenesis, mainly by protecting transformed cells from apoptosis[10, 11]. I*k*B/NF*k*B has also been reported to mediate invasion/metastasis in human cancer cells[12, 13]. Up-regulation of this group of genes in TBCC is consistent with the concept that the NF*k*B pathway is likely to play an important role in tumorigenesis.

2) ***genes involved in "Receptor signaling protein serine/threonine kinase activity"*** : MAPK14 and STK39. Both of them are components of the RAS/MAPK pathway known to be involved in cell proliferation and growth. Their increased expression in TBCC may reflect aberrant activation of the RAS pathway, a biological hallmark of multiple cancers.

3) ***genes involved in "Methyltransferase Activity"***: ICMT, ATIC, DNMT3A and METTL2. ICMT methylates the carboxyl-terminal isoprenylcysteine of CAAX proteins (e.g., Ras and Rho proteins). In the case of the Ras proteins, carboxyl methylation is important for targeting the proteins to the plasma membrane. Bergo and his colleagues found ICMT was required for oncogenic transformation induced by K-Ras and B-RAF[14]. DNMT3A catalyzes CpG methylation on genomic DNA. The role of DNA methylation in silencing of tumor suppressor genes and cancer transformation is a subject of intense investigation. DNMT3A and ICMT were about 2-fold elevated in TBCC, implicating their potential involvement in oncogenesis. METTL2/METL was studied in regulating NOTCH receptor activation in Drosophila[15]. This gene was 3.5-fold elevated in TBCC suggesting that the Notch pathway, which has been shown to play a role in stem cell self-renewal, might be dysregulated in TBCC[16].

Interestingly, despite several of the genes up-regulated in the IGS are known to play key roles in oncogenesis, the majority of the 186 genes included in the IGS remain poorly characterized. Further in-depth study of these genes may help elucidate additional mechanisms in the development of breast cancer and shed new light on the biology of TBCC.

Comparison of IGS and WR prognostic powers. Comparison of IGS and WR gene signatures was based on the NKI breast cancer patient dataset, since this same dataset was utilized also by the authors who first described the WR signature[17]. The NKI dataset and all patient information related to the WR signature was downloaded from the original authors' web-site (http://microarray-pubs.stanford.edu/wound_NKI/). Individual patient information was mainly retrieved from the "Clinical_Data_Supplement" *Microsoft Excel* file, which contains both correlation values to the WR signature (under column "Corr_CSR_activated") and patient stratification in "activated" and "quiescent" categories (under column "WS_all_Original"). Correlation values were used in receiver operating characteristic (ROC) curves and univariate/multivariate Cox survival analyses. Patient stratification in "activated" and "quiescent" categories was used for Kaplan-Meier survival curves.

To compare IGS and WR signature prognostic power, we first analyzed the ROC curves of both IGS and WR signatures in predicting tumor metastasis at 5 years. The test was performed on a group of 262 breast cancer patients, selected from the 295 patients included in the NKI database based on the availability of up to 5 years complete follow-up information. We found that the ROC curves of the IGS and WR signature have very similar area under the curve (AUC) values (***Suppl. Fig. 3***), indicating that IGS and WR signatures have similar prognostic potential. We also performed a univariate Cox survival analysis using the IGS and the WR signatures on the same group of 262 tumors, which showed that the two signatures were associated with very similar prognostic power: hazard ratio (HR) per 0.1 correlation = 1.5 ($p = 2.4*10^{-7}$) and 1.4 ($p = 2.9*10^{-7}$), respectively ***(Suppl. Table 3a***).

REFERENCES:

1. Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, Clarke MF. Prospective identification of tumorigenic breast cancer cells. Proc Natl Acad Sci U S A 2003;100(7):3983-8.

2. Baugh LR, Hill AA, Brown EL, Hunter CP. Quantitative analysis of mRNA amplification by in vitro transcription. Nucleic Acids Res 2001;29(5):E29.

3. Iwashita T, Kruger GM, Pardal R, Kiel MJ, Morrison SJ. Hirschsprung disease is linked to defects in neural crest stem cell function. Science 2003;301(5635):972-6.

4. Klipper-Aurbach Y, Wasserman M, Braunspiegel-Weintrob N, et al. Mathematical formulae for the prediction of the residual beta cell function during the first two years of disease in children and adolescents with insulin-dependent diabetes mellitus. Med Hypotheses 1995;45(5):486-90.

5. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003;19(3):368-75.

6. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. Nat Genet 2003;33(1):49-54.

7. Park S, James CD. ECop (EGFR-coamplified and overexpressed protein), a novel protein, regulates NF-kappaB transcriptional activity and associated apoptotic response in an IkappaBalpha-dependent manner. Oncogene 2005;24(15):2495-502.

8. Bin LH, Xu LG, Shu HB. TIRP, a novel Toll/interleukin-1 receptor (TIR) domain-containing adapter protein involved in TIR signaling. J Biol Chem 2003;278(27):24526-32.

9. Dohrman A, Kataoka T, Cuenin S, Russell JQ, Tschopp J, Budd RC. Cellular FLIP (long form) regulates CD8+ T cell activation through caspase-8-dependent NF-kappa B activation. J Immunol 2005;174(9):5270-8.

10. Cao Y, Karin M. NF-kappaB in mammary gland development and breast cancer. J Mammary Gland Biol Neoplasia 2003;8(2):215-23.

11. Greten FR, Karin M. The IKK/NF-kappaB activation pathway-a target for prevention and treatment of cancer. Cancer Lett 2004;206(2):193-9.

12. Huber MA, Azoitei N, Baumann B, et al. NF-kappaB is essential for epithelial-mesenchymal transition and metastasis in a model of breast cancer progression. J Clin Invest 2004;114(4):569-81.

13. Shishodia S, Aggarwal BB. Nuclear factor-kappaB activation mediates cellular transformation, proliferation, invasion angiogenesis and metastasis of cancer. Cancer Treat Res 2004;119:139-73.

14. Bergo MO, Gavino BJ, Hong C, et al. Inactivation of Icmt inhibits transformation by oncogenic K-Ras and B-Raf. J Clin Invest 2004;113(4):539-50.

15. Zhang SX, Guo Y, Boulianne GL. Identification of a novel family of putative methyltransferases that interact with human and Drosophila presenilins. Gene 2001;280(1-2):135-44.

16. Callahan R, Egan SE. Notch signaling in mammary development and oncogenesis. J Mammary Gland Biol Neoplasia 2004;9(2):145-63.

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

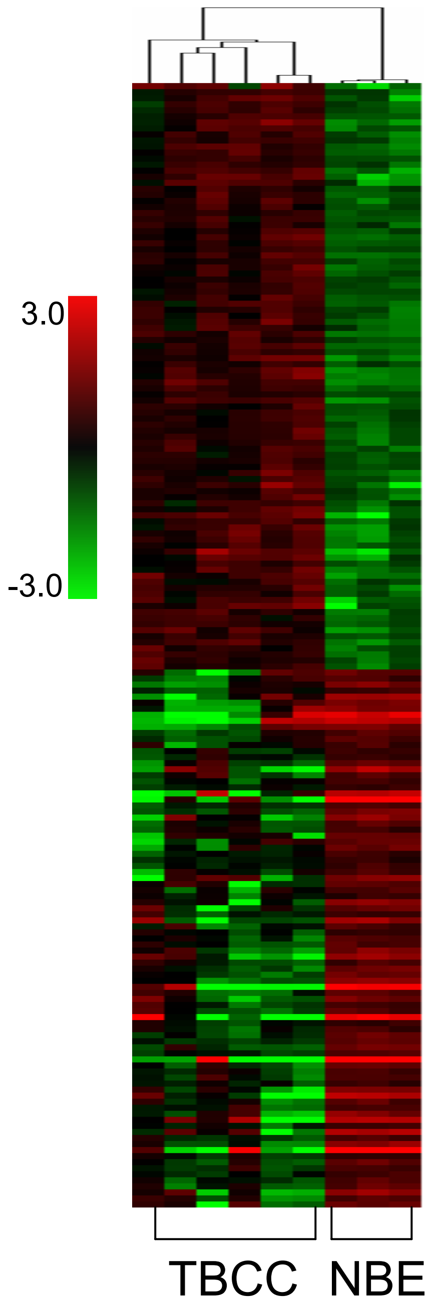
Suppl. Table 1. Origin of breast cancer tissue samples utilized in the study.

Sample	Origin	Source[a]	Histology	Patient stage[b]	ER status
T1	malignant pleural effusion	mouse xenograft (p2)	infiltrating ductal carcinoma	IV	+
T2	primary tumor	mouse xenograft (p2)	adenocarcinoma	$pT_XN_0M_0$[c]	+
T3	malignant pleural effusion	mouse xenograft (p2)	invasive lobular carcinoma	IV	+
T4	malignant pleural effusion	primary tissue	metastatic breast carcinoma	IV	-
T5	malignant pleural effusion	primary tissue	invasive lobular carcinoma	IV	+
T6	malignant pleural effusion	primary tissue	metastatic breast carcinoma	IV	+

[a] The $CD44^+/CD24^{-/low}$ cancer cell subset was isolated either from primary malignant pleural fluids directly harvested from patients undergoing thoracentesis (*primary tissue*) or from breast cancer tissues grown as solid tumor xenografts in immunodeficient NOD-SCID mice (*mouse xenograft*). In the case of solid tissue xenografts, $CD44^+/CD24^{-/low}$ cancer cells were purified from the second *in vivo* serial passage (*p2*).

[b] According to the 6th American Joint Committee on Cancer (AJCC) staging system for breast cancer (2002).

[c] This patient was free of lymphnode (N_0) and distant site (M_0) metastases. The exact pathological characteristics of the primary tumor are not available (pT_X), because the patient underwent neoadjuvant chemotherapy before surgery.

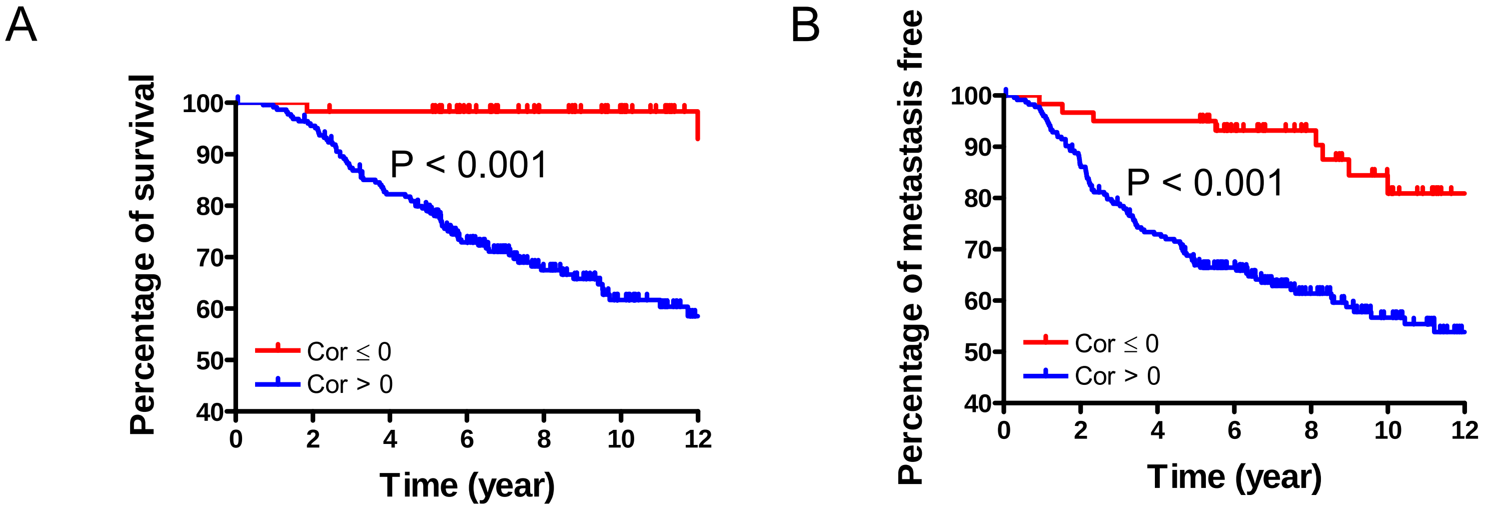


Suppl. Fig. 1. Cluster analysis of 6 tumorigenic breast cancer cells (TBCC) and 3 normal breast epithelium (NBE) samples based on the 186 genes from the Invasiveness Gene Signature (IGS).

Gene Symbol	FC in RT-PCR (TBCC vs NBE)			FC in Microarray (TBCC vs NBE)
	T1	T2	T3	
AMMECR1	7.08	4.02	**1.15**	5.66
PLP2	7.34	5.18	**0.89**	4.29
MAPK14	3.82	3.15	**0.41**	4.00
HS2ST1	7.72	3.89	**0.97**	4.00
KDELR3	8.46	5.37	**0.52**	4.00
PDE8A	6.62	5.92	**0.51**	4.00
ISGF3G	4.26	4.15	2.48	3.73
GAPD	22.48	11.02	6.06	2.14
IRX3	0.45	0.01	0.22	0.35
CEBPD	0.50	0.38	0.26	0.29
BCL2	0.02	0.00	0.03	0.23
GOLGIN-67	0.00	0.03	0.05	0.18
MGP	0.01	0.00	0.13	0.15
LTF	0.00	0.00	0.00	0.09

FC, fold change; RT-PCR, reverse transcription and real time polymerase chain reaction; TBCC, tumorigenic breast cancer cells; NBE, normal breast epithelium;

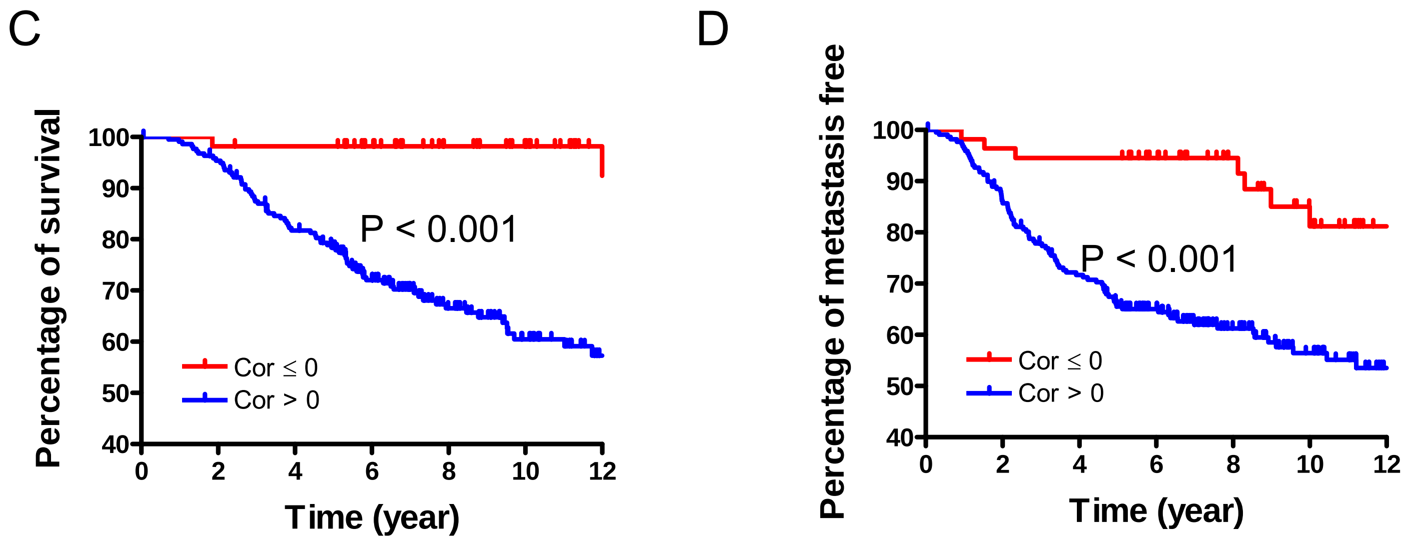
Suppl. Table 2. Validation of the IGS by quantitative RT-PCR. A list of 14 genes was randomly selected from the 186 genes included in the IGS. The differential expression of these 14 genes between the CD44^{+}/CD24$^{-/low}$ cancer cell subset (i.e. TBCC) of three breast tumors and normal breast epithelium cells (NBE) was measured by real time PCR. The average fold change (FC) in differential expression levels of the same genes as measured by microarray analysis is listed as reference.



No. At Risk							
Cor ≤ 0:	60	59	59	48	35	26	13
Cor > 0:	224	211	178	135	86	56	28

No. At Risk							
Cor ≤ 0:	60	58	57	46	33	23	11
Cor > 0:	224	192	159	123	78	51	27

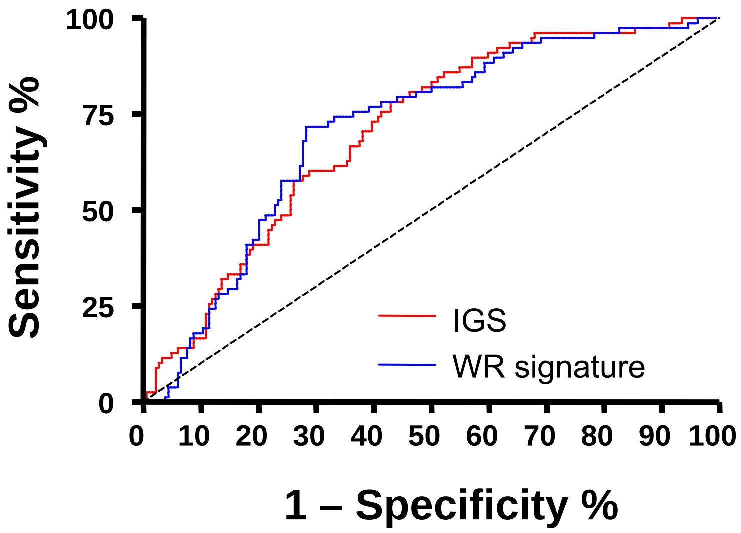
Suppl. Figure 2, A-B. The IGS adds prognostic information within the group of high-risk early breast cancer patients identified by NIH consensus criteria. Patients were classified into two groups based on their positive (Cor > 0) or negative (Cor ≤ 0) correlation value to the IGS, and their survival was analyzed by Kaplan-Meier curves. Within high-risk patients as identified by NIH criteria, the IGS was able to stratify patients in two groups with substantially different overall (A) and metastasis-free survival (B).



No. At Risk							
Cor ≤ 0:	55	54	53	44	33	24	12
Cor > 0:	218	205	172	130	85	53	28

No. At Risk							
Cor ≤ 0:	55	53	52	43	32	22	10
Cor > 0:	218	186	134	117	77	51	26

Suppl. Figure 2, C-D. The IGS adds prognostic information within the group of high-risk early breast cancer patients identified by St. Gallen consensus criteria. Patients were classified into two groups based on their positive (Cor > 0) or negative (Cor ≤ 0) correlation value to the IGS, and their survival was analyzed by Kaplan-Meier curves. Within high-risk patients as identified by St. Gallen criteria, the IGS was able to stratify patients in two groups with substantially different overall (C) and metastasis-free survival (D).

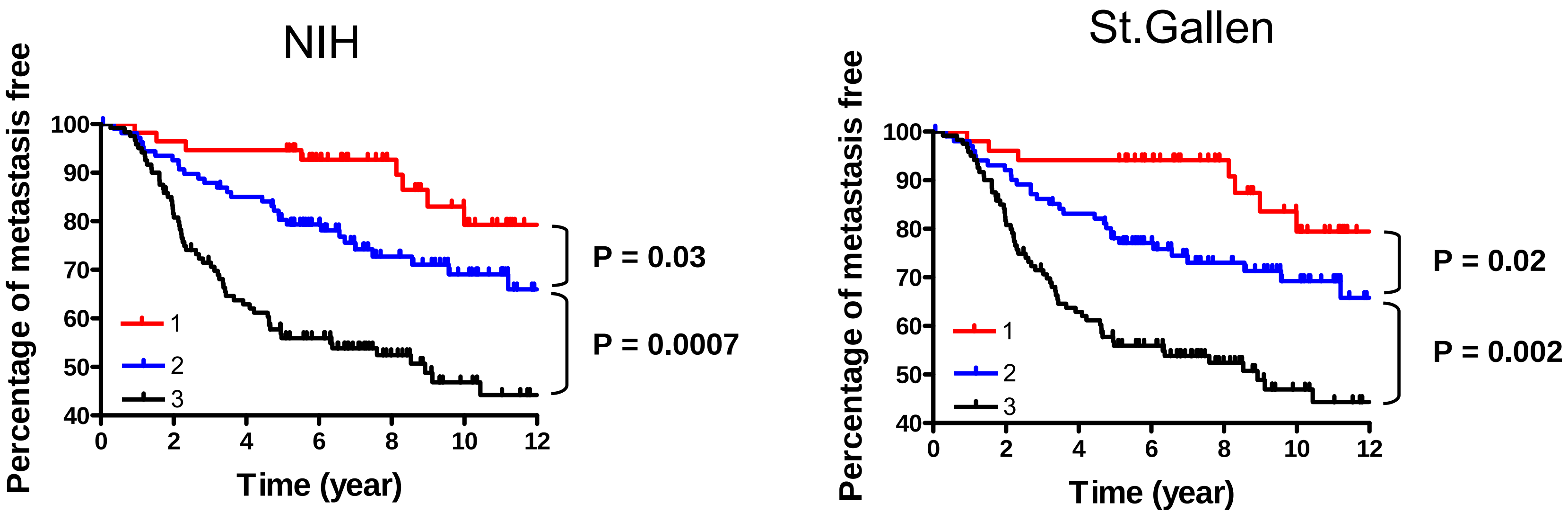


Suppl. Fig. 3. Receiver operating characteristic (ROC) curve comparison of the "invasiveness gene signature" (IGS) and the "wound response" (WR) gene signature for prediction of 5-year metastasis-free survival. The analysis was performed on a group of 262 breast cancer patients selected from the 295 patients included in the NKI database, based on the availability of complete 5-year follow-up records**.** The area under the curve (AUC) is 0.71 ($P < 0.001$) for both the IGS and the WR signature.

a) Univariate analysis	Risk of Metastasis	
	HR (95% CI, per 0.1 correlation)	P value
IGS	1.5 (1.3 – 1.7)	**$2.4*10^{-7}$**
WR	1.4 (1.2 – 1.6)	**$2.9*10^{-7}$**

b) Multivariate analysis	Risk of Metastasis	
	HR (95% CI, per 0.1 correlation)	P value
IGS	1.3 (1.1 – 1.5)	**0.001**
WR	1.2 (1.1 – 1.4)	**0.003**
WR + IGS combination	n.a.	**$2.4*10^{-8}$**

Suppl Table 3. Univariate (a) and multivariate (b) Cox proportional survival analysis of the IGS and WR gene signatures as predictors of metastasis-free survival. The analysis was performed on 262 breast cancer patients of the NKI dataset, based on the availability of up to 5 years complete follow-up information. The correlation value to the IGS or the WR signature was modeled as a continuous variable. HR = hazard ratio, CI = confidence interval, n.a. = not applicable. Parameters found to be statistically significant ($p < 0.05$) were shown in bold.



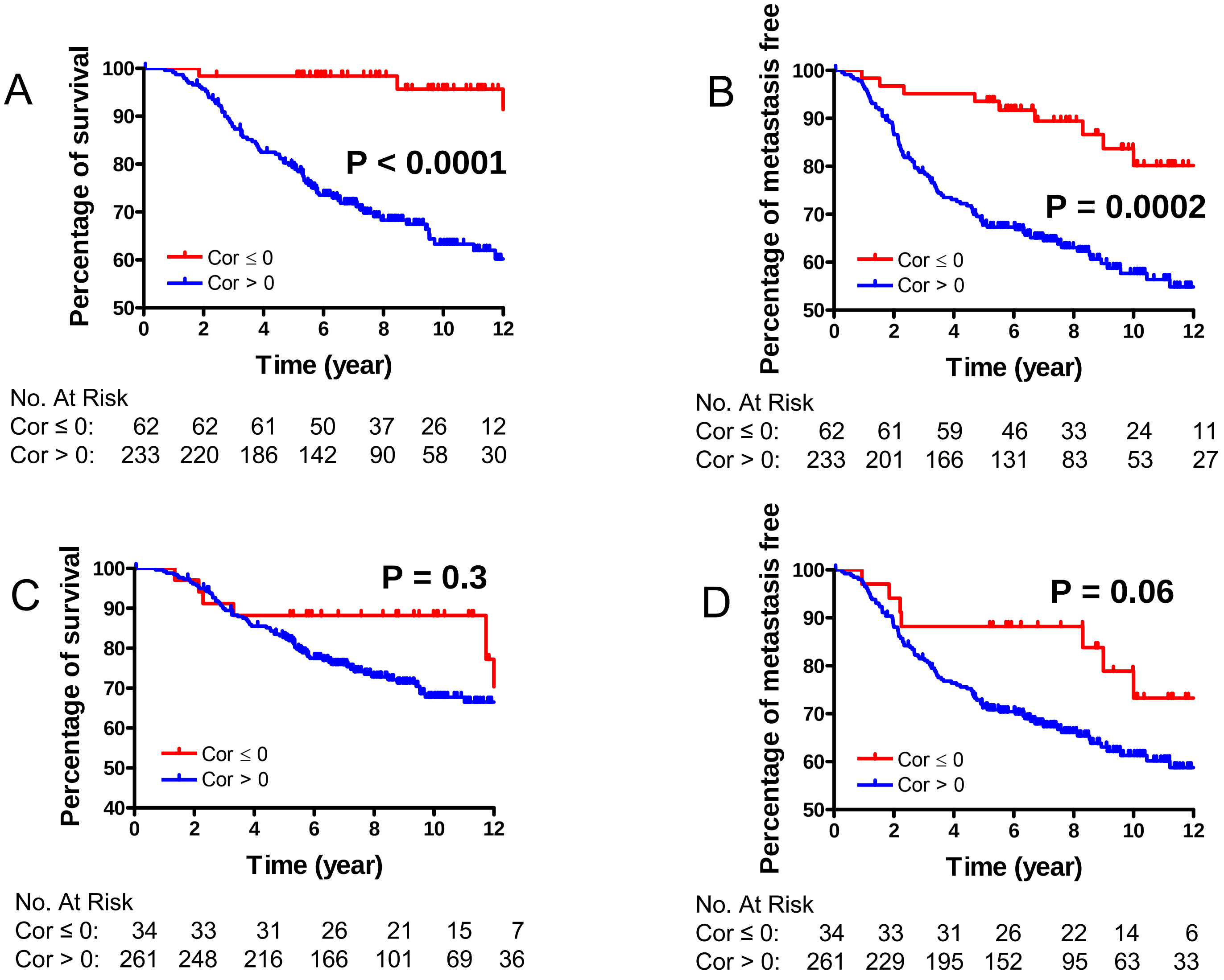
No. At Risk

1:	56	54	53	43	30	21	10
2:	108	99	90	69	46	33	17
3:	120	97	72	58	34	21	13

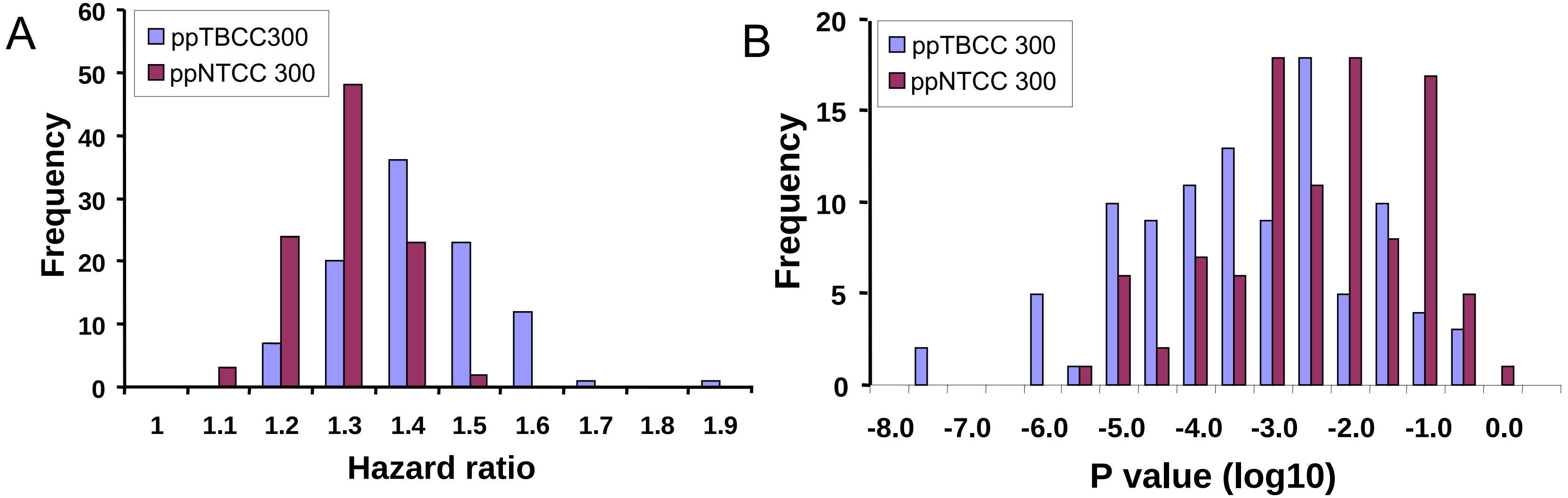
No. At Risk

1:	51	49	48	41	28	19	9
2:	102	93	83	63	44	32	16
3:	120	97	73	58	35	21	13

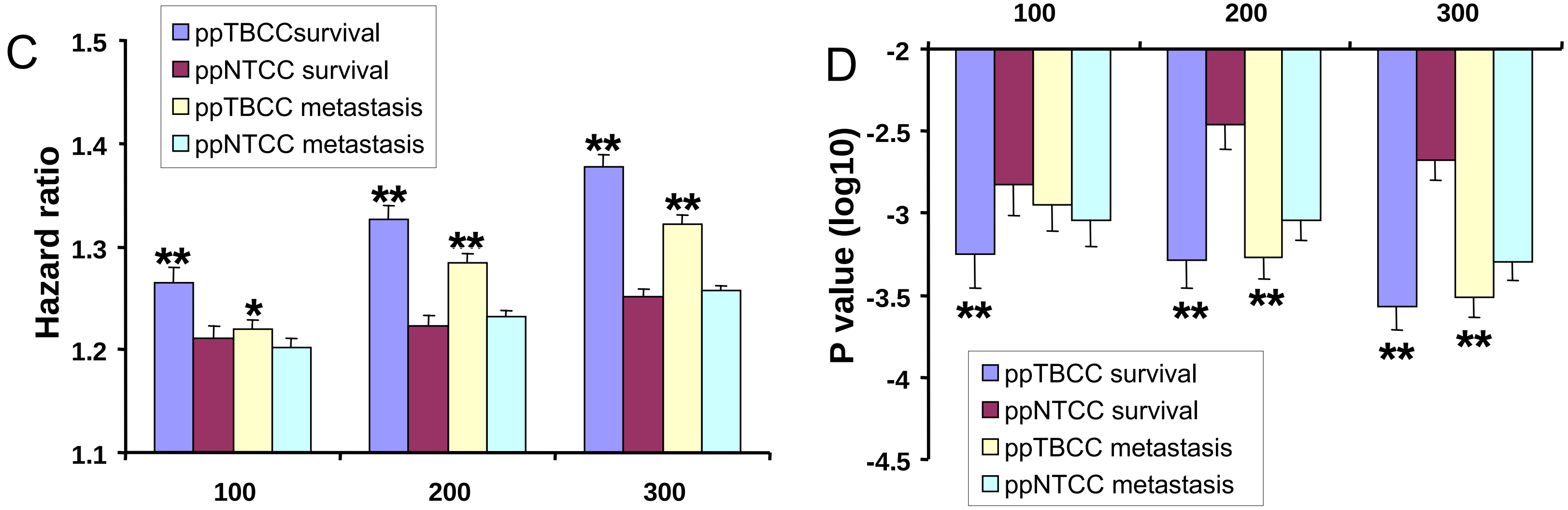
Suppl. Fig. 4. Combined use of IGS and WR signatures for prognosis prediction in "high risk early breast cancer" patients, according to either NIH (a) or St. Gallen (b) consensus guidelines. Patients were classified into three groups: Group 1 (low-risk), patients with both good-prognosis "quiescent" WR and IGS- (Cor ≤ 0) signatures; Group 2 (intermediate-risk), patients with either poor-prognosis "activated" WR or IGS+ (Cor > 0) signatures; Group 3 (high-risk), patients with both poor-prognosis "activated" WR and IGS+ (Cor>0) signatures. Survival of different patient groups was analyzed by Kaplan-Meier curves. Differences in metastasis-free survival are statistically significant ($p < 0.05$) for all group comparisons.



Suppl. Fig. 5. Based on the 186 genes included in the IGS, the gene-expression profile of 3 paired primary breast tumorigenic cancer cells (ppTBCC; A, B) is superior in prognostic power to that of the corresponding 3 paired primary non-tumorigenic cancer cell samples (ppNTCC; C, D). Pearson correlations were calculated using the gene expression values of the 186 genes included in the IGS, comparing the values of each tumor to the average values of 3 ppTBCC (A, B) and of 3 ppNTCC (C, D), respectively. Tumors were divided into two groups based on their correlation values (Cor > 0, Cor ≤0). Kaplan-Meier survival curves were created to compare the two groups of tumors using overall (A, C) or metastasis-free survival (B, D) as clinical end-point.



Suppl. Fig. 6. Gene expression profiles generated using the breast tumorigenic cancer cell (TBCC) population are endowed with a superior predictive power (larger hazard ratio and smaller p-value) when compared to those generated from non-tumorigenic cancer cell (NTCC) population. To have a fair comparison between IGS and NTCC gene expression profiles in predicting patient outcome, we analyzed the gene expression profiles of 3 paired primary TBCC (ppTBCC) and NTCC (ppNTCC) samples by comparing with the gene profiles of three normal breast epithelium (NBE) samples. In this analysis, 2910 genes were selected based on a 2-fold difference in expression between ppTBCC and NBE samples (1937 genes) and between ppNTCC and NBE samples (2231 genes). Out of these 2910 genes, 2458 (1642 from ppBCSC and 1884 from ppNTCC) were included in the NKI breast cancer datasets. Starting from this 2458-gene set, which contains both overlapped and distinct genes from the above two lists and includes 154 genes of 186-gene signature, we generated groups of random lists (n = 100 lists/group), each list containing 100, 200, or 300 distinct genes. All these signatures were used in predicting both overall and metastasis-free survival using NKI breast cancer patients datasets. **Panels A and B.** Histogram distribution of the hazard ratio (A) and of its p-value (B) for the group of gene signatures containing 300 genes from ppTBCC and ppNTCC in predicting patient overall survival.



Suppl. Fig. 6, Panel C and D. Average values of hazard ratio (C) and corresponding p-values (D) for groups of 100 random gene signatures containing 100, 200 or 300 genes, generated using expression values from 3 paired primary TBCC (ppTBCC) and NTCC (ppNTCC) samples to predict either overall or metastasis-free survival. *, $P < 0.05$; **, $P < 0.01$.