www.impactjournals.com/oncotarget/ Oncotarget, 2017, Vol. 8, (No. 8), pp: 14003-14016

Research Paper

# Genome-wide screen identifies a novel prognostic signature for breast cancer survival

**Xuan Y. Mao[1], Matthew J. Lee[1], Jeffrey Zhu[1], Carissa Zhu[1], Sindy M. Law[2], Antoine M. Snijders[1]**

[1]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA

[2]Department of Psychiatry, Weill Institute for Neurosciences, University of California San Francisco, San Francisco, California, USA

***Correspondence to:*** *Antoine M. Snijders,* ***email:*** *AMSnijders@lbl.gov*

***Keywords:*** *breast cancer, prognostic score, relapse-free survival, gene biomarkers*

***Received:*** *August 15, 2016* ***Accepted:*** *December 31, 2016* ***Published:*** *January 21, 2017*

## ABSTRACT

**Large genomic datasets in combination with clinical data can be used as an unbiased tool to identify genes important in patient survival and discover potential therapeutic targets. We used a genome-wide screen to identify 587 genes significantly and robustly deregulated across four independent breast cancer (BC) datasets compared to normal breast tissue. Gene expression of 381 genes was significantly associated with relapse-free survival (RFS) in BC patients. We used a gene co-expression network approach to visualize the genetic architecture in normal breast and BCs. In normal breast tissue, co-expression cliques were identified enriched for cell cycle, gene transcription, cell adhesion, cytoskeletal organization and metabolism. In contrast, in BC, only two major co-expression cliques were identified enriched for cell cycle-related processes or blood vessel development, cell adhesion and mammary gland development processes. Interestingly, gene expression levels of 7 genes were found to be negatively correlated with many cell cycle related genes, highlighting these genes as potential tumor suppressors and novel therapeutic targets. A forward-conditional Cox regression analysis was used to identify a 12-gene signature associated with RFS. A prognostic scoring system was created based on the 12-gene signature. This scoring system robustly predicted BC patient RFS in 60 sampling test sets and was further validated in TCGA and METABRIC BC data. Our integrated study identified a 12-gene prognostic signature that could guide adjuvant therapy for BC patients and includes novel potential molecular targets for therapy.**

## INTRODUCTION

Breast cancer (BC) is the leading female malignancy and the second leading cause of cancer deaths in U.S. women, with tumor metastasis being the underlying cause in most of these breast cancer related deaths [1, 2]. Breast carcinogenesis is a multi-step process in which epithelial cells accumulate genetic alterations, which in a permissive tissue microenvironment progress towards malignancy and may then metastasize to distant organs. Advances in imaging technologies and heightened public awareness of breast cancer have resulted in an increase in the diagnosis of early-stage breast cancer [3–5]. Furthermore, adjuvant systemic therapy has reduced the risk of recurrence and improved overall survival from BC [6]. However, not all patients who receive adjuvant therapy benefit from it and could have been spared the treatment-associated toxicity. Prognostic factors are critical to distinguish patients with poor prognoses, who would benefit from adjuvant therapy, from patients with good prognoses, who may not benefit sufficiently from adjuvant therapy to outweigh the risks associated with treatment [7].

Traditional prognostic factors currently used to guide the use of systemic therapy and predict outcome include tumor size, lymph node involvement, histological grade, age, race, estrogen receptor (ER), progesterone receptor (PR) and epidermal growth factor receptor (HER2) status [8]. However, a critical problem with BC

is the difference in clinical outcome among patients with the same disease. This heterogeneous clinical outcome is manifested by differences in disease susceptibility, progression, treatment response, and relapse, even among individuals with the same apparent histopathological disease. These differences seem to be in part controlled by so-called tumor modifier genes, multiple low-penetrance susceptibility genes that interact with each other and their environment to contribute to the disease process.

Clinical patient survival data, along with genomic datasets can be used to identify genes important in patient survival. Recently, a large gene expression database across normal human tissues became available and which can be used to identify the biological mechanisms underlying different diseases and identify potential novel therapeutic targets [9, 10]. We combined independent BC databases to identify a gene expression signature of differentially expressed genes. Using gene co-expression network analyses, we investigated the genetic architecture of this signature in normal breast tissue. We subsequently identified and validated a 12-gene signature that predicts BC survival.

## RESULTS

### Meta-analysis identified a 587-gene signature frequently deregulated in human breast cancer

We conducted a meta-analysis of genes consistently deregulated in human BCs. We collected gene transcript data from normal and tumor breast tissues represented by four independent gene expression data sets totaling 160 invasive ductal carcinomas and 191 normal breast tissues (Figure 1A) [11–15]. The significant differential expression of genes was assessed by a fold change cutoff of 1.5 and adjusted p-value<0.01 (Supplementary Table 1). This resulted in a gene signature of 795 probe IDs (590 down-regulated and 205 up-regulated) represented by 587

A.

| Human Breast Tumor Data | Invasive ductal carcinoma samples | Normal breast samples | PMID |
|---|---|---|---|
| GSE3744 | 40 | 7 | 16473279; 20400965 |
| GSE10780 | 42 | 143 | 19266279 |
| GSE21422 | 5 | 5 | 21314937 |
| GSE29044 | 73 | 36 | 23704896 |
| **Total** | 160 | 191 | |

B.

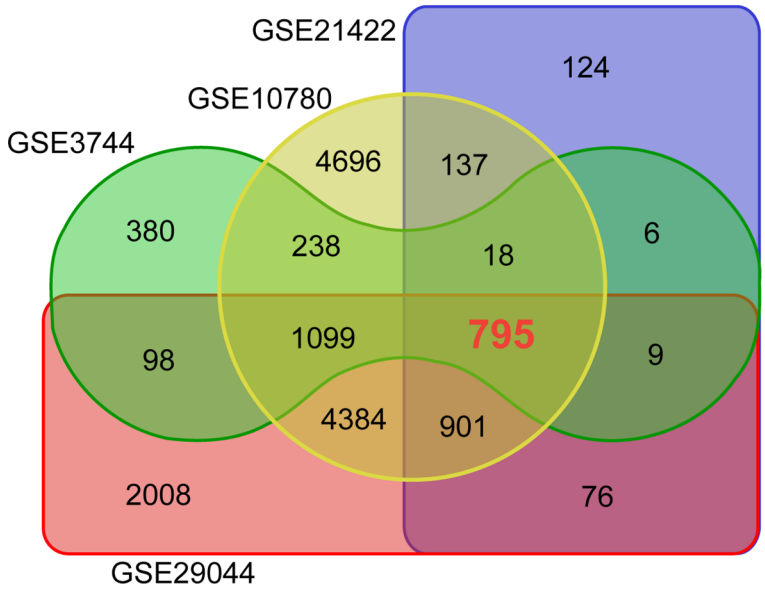



**Figure 1: The human breast tissue data sets used in this study. A.** Four independent gene transcript data sets containing invasive ductal carcinoma and normal breast tissue samples were used. **B.** Differential expression of tumor versus normal using a fold-change cut-off of 1.5 and adjusted p-value 0.01 identified the 795 common probe ID set.

unique genes, for which the direction of expression was consistent across all four datasets (Figure 1B and Figure 2; Supplementary Table 2).

### 381 genes significantly associated with relapse-free survival in breast cancer patients

To investigate whether any of the 587 common deregulated genes were associated with relapse-free survival (RFS), we evaluated the prognostic value for each individual gene in a large public clinical microarray database using the Kaplan-Meier plotter (http://kmplot.com/) (Figure 2) [16]. The BC patient cohort was divided into two equal groups based on median expression for each gene and compared by a Kaplan-Meier survival analysis. In addition, the hazard ratio with a 95% confidence interval and logrank p-value was calculated to evaluate the prognostic significance of each gene for

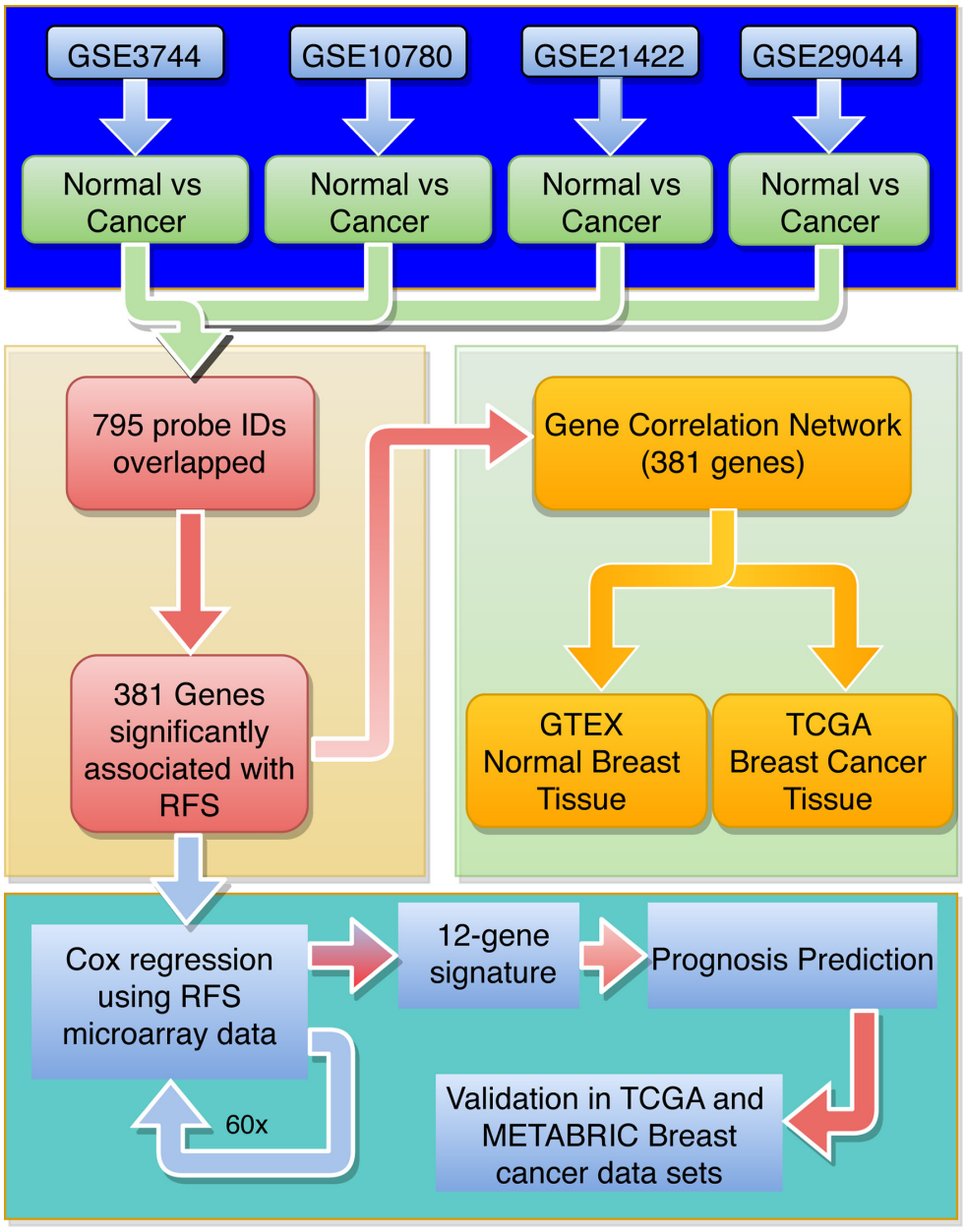



**Figure 2: Flow diagram for identifying and validating a prognostic biomarker panel for breast cancer.** The 795 robustly deregulated probe IDs were identified using 4 breast tumor microarray data sets (blue). To identify individual genes associated with relapse-free survival (RFS), Kaplan Meier survival analysis was run on the overlapping IDs (yellow). A gene expression correlation network approach was used to identify cliques of functionally related genes (green). Cox regression was run on 60 random tumor samples for 381 genes significantly associated with RFS (turquoise) to generate the 12-gene signature. The 12-gene signature was used to generate a prognosis scoring system, which was validated using the TCGA and METABRIC BC data sets.

RFS. This analysis identified 381 genes significantly associated with RFS (p-value<6.3E-05; Figure 3, Table 1 and Supplementary Table 3); 249 genes had a hazard ratio < 1 (higher gene expression associated with good prognosis) and 133 genes had a hazard ratio > 1 (higher gene expression associated with poor prognosis).

## Genes that predict prognosis are enriched for microenvironment- and cell cycle-related biological processes

To reveal the biological functions enriched in the 381-gene set associated with RFS, we performed Gene Ontology analysis separately on the 249 genes that exhibited a HR<1 and 133 genes with HR>1. The 249-gene signature (HR<1) was significantly enriched for tissue microenvironment related processes including cell adhesion (adj. p-value=6E-04), cell migration (adj. p-value=2.74E-05), wound healing (adj. p-value=3.1E-03), and vasculature development (adj. p-value=4.13E-05) (Supplementary Table 4). On the other hand, the 133-gene signature (HR>1) was strongly enriched for cell cycle related processes (adj. p-value=5.33E-51) (Supplementary Table 4). This strong dichotomy between RFS genes with HR<1 - associated with tumor processes enriched for tissue microenvironment-related biological functions (e.g. vasculature, wound healing, cell migration) - and RFS genes with HR>1 - almost exclusively associated with cell cycle related processes - prompted us to further investigate the genetic architecture of these genes in normal breast tissues and BCs.

## Gene co-expression network analysis visualizes the genetic architecture of RFS associated genes in normal breast and breast cancer

Since gene sets that are correlated in expression across tissue samples often share a common function, co-expression network analysis has been used to identify clusters of genes with common biological functionality important in normal or tumor tissues. We used data obtained from the GTEX database of 214 normal human breast tissues and the TCGA database of 1100 BC samples to reveal the genetic architecture of RFS associated genes

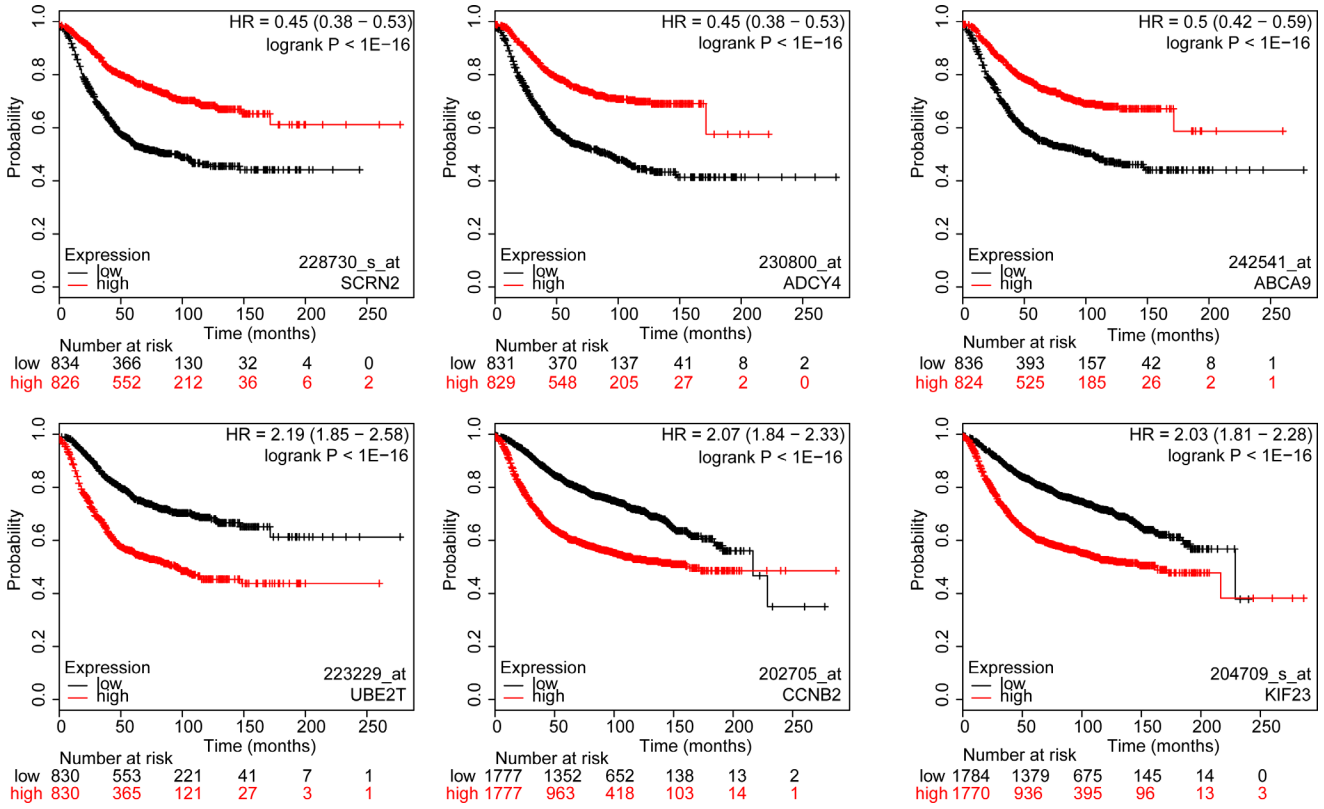



**Figure 3: Kaplan-Meier survival curves for breast cancer patients according to tumor expression of genes with highest and lowest hazard ratios.** The breast cancer patient cohort was divided into two equal groups based on median expression for each gene and compared by a Kaplan-Meier survival analysis. The estimate of the hazard ratio (HR) and log-rank p-value of the curve comparison between the groups is shown. Top three genes with the lowest HR values (top row): SCRN2, ADCY4 and ABCA9. Top three genes with the highest HR values (bottom row): UBE2T, CCNB2 and KIF23. Low and high risks indicated in black and red, respectively.

**Table 1: Twelve-gene prognostic gene signature**

| Gene symbol | Gene name | Affymetrix ID | Hazard Ratio | p-value |
|---|---|---|---|---|
| EPS15 | Epidermal growth factor receptor substrate 15 | 217886_at | 0.73 | 9.30E-08 |
| MELK | Maternal Embryonic Leucine Zipper Kinase | 204825_at | 1.89 | 1.00E-16 |
| NUF2 | NDC80 Kinetochore Complex Component | 223381_at | 1.63 | 2.30E-09 |
| RNASEH2A | Ribonuclease H2 Subunit A | 203022_at | 1.56 | 1.90E-14 |
| S100P | S100 Calcium Binding Protein P | 204351_at | 1.45 | 2.50E-10 |
| THYN1 | Thymocyte Nuclear Protein 1 | 218491_s_at | 0.76 | 2.70E-06 |
| TIMM17A | Translocase Of Inner Mitochondrial Membrane 17 Homolog A | 201821_s_at | 1.55 | 3.70E-14 |
| TSC1 | Tuberous Sclerosis 1 | 209390_at | 0.74 | 4.00E-07 |
| USP47 | Ubiquitin Specific Peptidase 47 | 223119_s_at | 0.65 | 2.40E-07 |
| ZBTB16 | Zinc finger and BTB domain containing 16 | 205883_at | 0.6 | 1.00E-16 |
| PLPP1 | Phospholipid Phosphatase 1 | 209147_s_at | 0.77 | 4.10E-06 |
| PLEKHH2 | Pleckstrin Homology, MyTH4 And FERM Domain Containing H2 | 227148_at | 0.59 | 1.70E-10 |

in normal and tumor breast tissue (Figure 2). We first calculated correlation coefficients of 381 genes associated with RFS across 214 normal human breast tissues and 1100 breast cancer samples (Figure 4A). We then constructed a gene expression correlation network where nodes represented individual gene and edges connecting genes represented a correlation in their expression (Figure 4B, $R \geq |0.6|$; p-value<8E-08). In normal breast tissue, three main co-expression cliques were identified (Figure 4B). One clique was highly enriched for genes involved in cell cycle and mitosis, and whose genes all had a hazard ratio for RFS > 1 (Figure 4B, 4D). The remaining two co-expression cliques contained predominantly genes with a hazard ratio for RFS <1. One clique was enriched for genes involved in transcriptional regulation and cell adhesion, while the other clique was generally involved in cytoskeleton organization and metabolic processes. Interestingly, while expression levels of genes within each clique were predominantly positively correlated, expression levels of genes between these two cliques were negatively correlated (Figure 4C). The cell cycle clique is connected to these two cliques through EZH2, MCM2 and MAD2L1.

A similar co-expression correlation analysis using TCGA data revealed two main co-expression cliques (Figure 5A, 5B). Similar to normal breast tissue, one clique was highly enriched for genes involved in cell cycle and mitosis, all of which had a hazard ratio for RFS > 1 (Figure 5B). The remaining clique contained genes with a hazard ratio for RFS <1 and was enriched for blood vessel development, cell adhesion and mammary gland development. These two co-expression cliques were negatively correlated through 7 genes: CREBRF, DIXDC1, AHNAK, CYBRD1, NOSTRIN, TNS2 and TNFSF12 (Figure 5C). Given the negative correlation with cell cycle related genes, these 7 genes could mediate negative regulation of cell growth and are potential therapeutic targets.

## A 12-gene prognostic signature predicts breast cancer patient survival

Using the 381-gene set associated with RFS we developed a gene signature that accurately predicts BC patient survival (Figure 2). We created 60 training sets through randomly selecting 300 patients each time from the BC gene expression dataset GSE6532, which has RFS information of 393 patients. The residual 93 patients from all 60 training sets formed the 60 test sets. We then performed Cox regression analysis on all 60 training sets to simultaneously assess the importance of the genes within the 381-gene in the RFS. The genes that recurred in at least half of the training sets were included in our final 12-gene signature: EPS15, MELK, NUF2, PLEKHH2, PLPP1, RNASEH2A, S100P, THYN1, TIMM17A, TSC1, USP47, ZBTB16 (Table 1). The average beta-value (Cox regression coefficient) of each of the 12 genes was calculated and used as a weighting factor for the expression value of each gene. A prognostic score was estimated for each patient: gene expression values were multiplied by their respective beta-value and the prognostic score was determined as the sum of resulting weighted gene expression values. The patients were ranked by their prognostic score,

divided into two equal sized cohorts based on the median score, and Kaplan-Meier analysis was performed to determine differences in RFS between two cohorts. Using the mean beta values developed in the training set, prognostic scores were calculated for all patients in the 60 test sets. Patients were again ranked on their prognostic score and divided into two cohorts based on the average prognostic-score cut-point in the 60 training

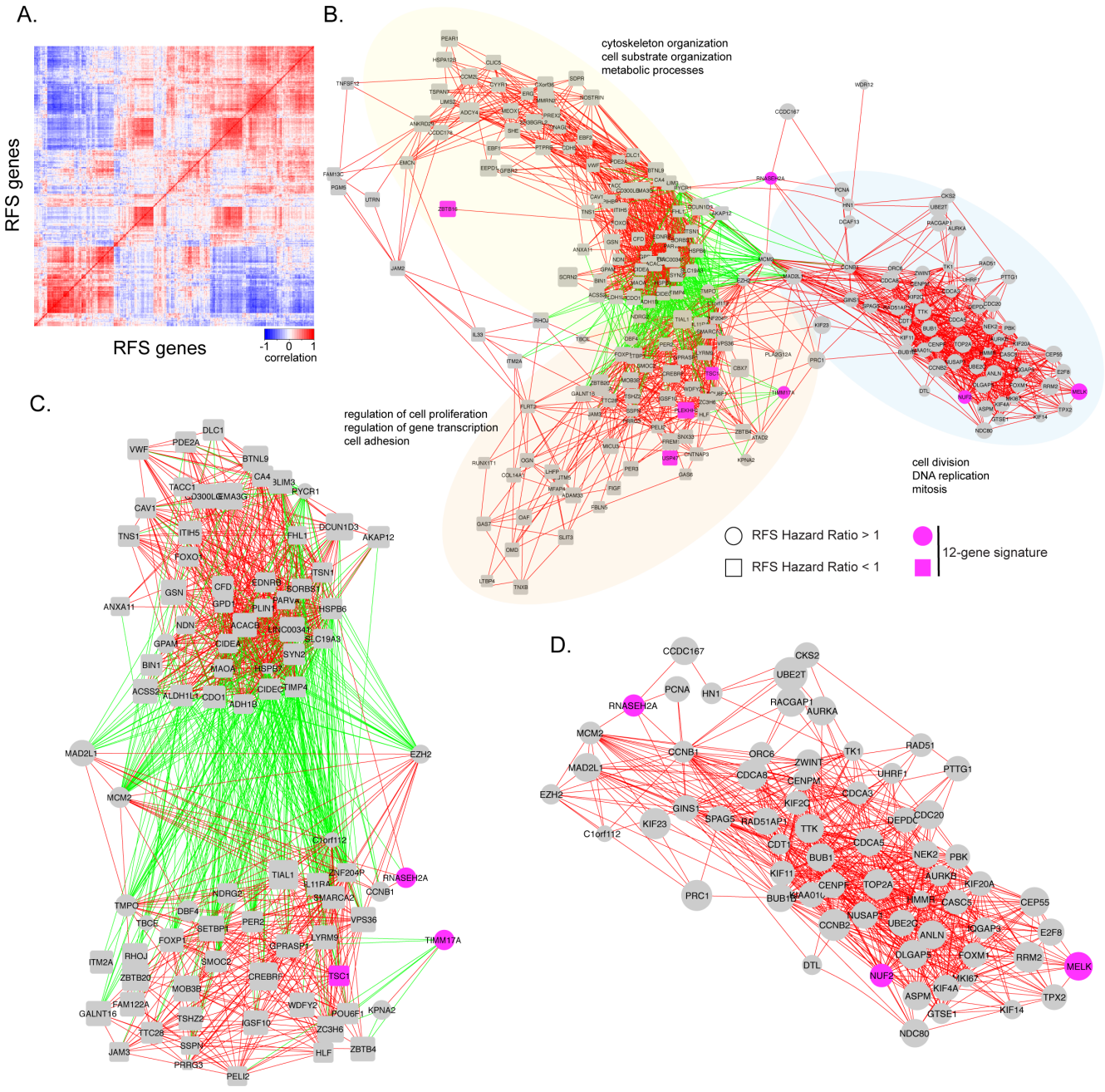



**Figure 4: Visual representation of correlations in gene expression in normal human breast tissue samples. A.** The heat map shows the correlation in gene expression between normal breast tissue samples obtained from GTEX. Positive correlations are indicated in red, while negative correlations are indicated in blue. **B.** Gene expression correlation network of RFS significant genes in normal breast tissue samples. Individual genes are indicated as nodes. Red edges indicate a positive correlation in gene expression ($r \geq 0.6$) between two genes. Green edges indicate a negative correlation in gene expression between two genes ($r \leq -0.6$). Labels indicate significant biological enrichment (adjusted p-value<0.05). Pink colored genes are present in the 12-gene prognostic signature. Three major functional cliques were separated based on gene-ontology. Clique 1 (yellow): cytoskeleton organization, cell substrate organization, and metabolic processes. Clique 2 (orange): regulation of cell proliferation, regulation of gene transcription, and cell adhesion. Clique 3 (blue): ell division, DNA replication, and mitosis. Genes with hazard ratio for RFS >1 are indicated as circles and those with HR<1 as squares. **C.** Enlargement of negative correlations and the genes associated with them. **D.** Enlarged cell division, DNA replication, and mitosis clique.

sets. Kaplan-Meier analysis was performed and a log-rank test was used to determine if there was a significant difference in RFS between two cohorts. The hazard ratio was calculated for each of the 60 test sets. In only 2 out of 60 (3.3%) test sets, the hazard ratio confidence interval crossed "1" (Figure 6A).

## Validation of 12-gene prognostic signature

We then tested our 12-gene prognostic signature in an independent set of 1100 BC patients obtained from the TCGA database. Prognostic scores for all 1100 patients were calculated and patients were ranked based on their

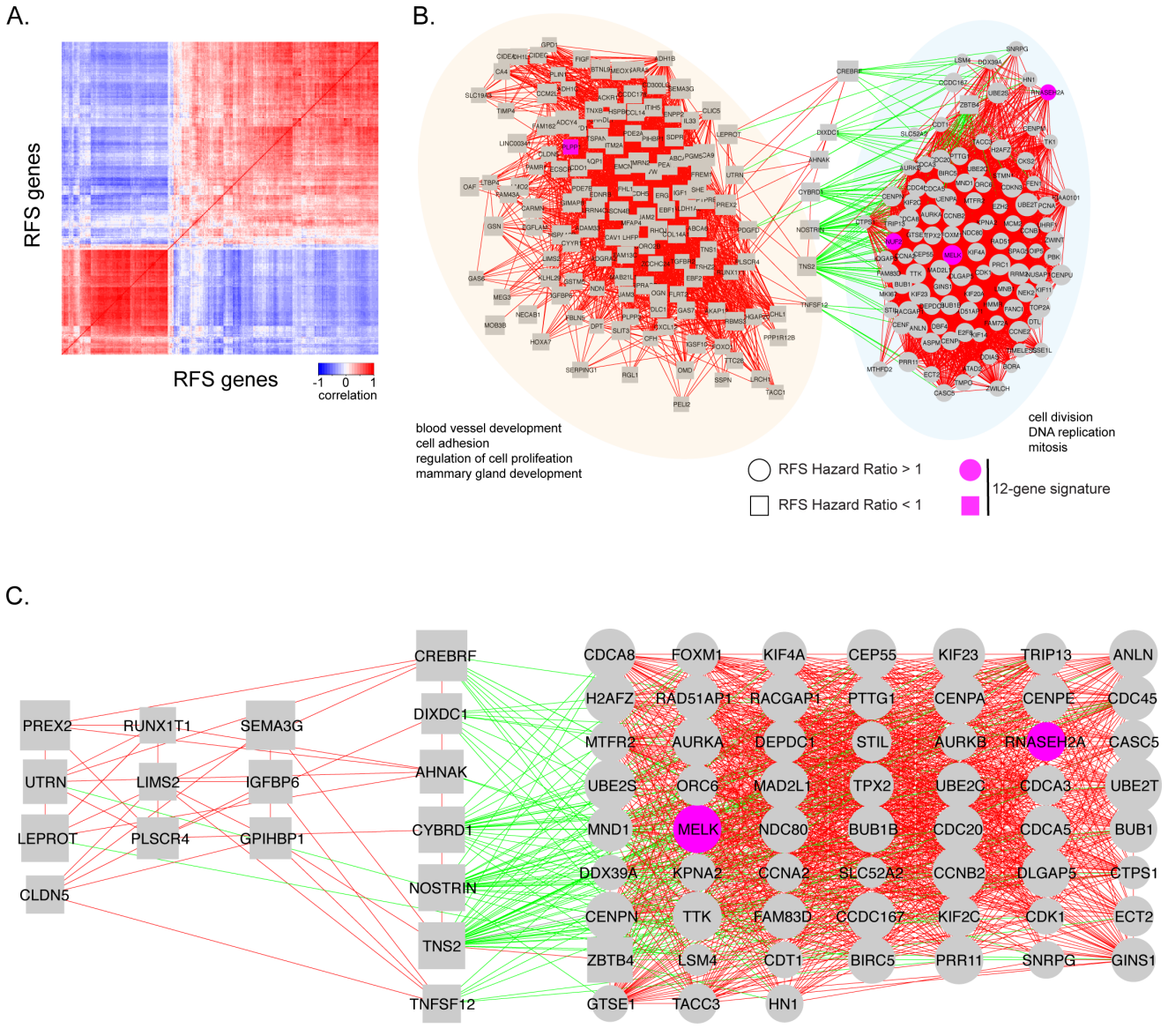
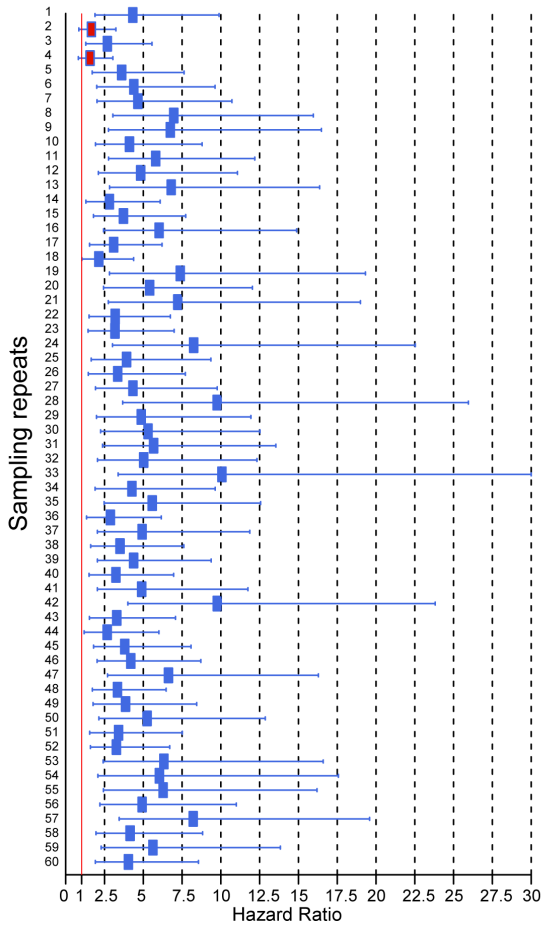



**Figure 5: Visual representation of correlations in gene expression in breast cancer samples. A.** The heat map shows the correlation in gene expression between breast cancer samples obtained from TCGA. Positive correlations are indicated in red, while negative correlations are indicated in blue. **B.** Gene expression correlation network of RFS significant genes in breast cancer samples. Individual genes are indicated as nodes. Red edges indicate a positive correlation in gene expression ($r \geq 0.6$) between two genes. Green edges indicate a negative correlation in gene expression between two genes ($r \leq -0.6$). Labels indicate significant biological enrichment (adjusted p-value<0.05). Pink colored genes are present in the 12-gene prognostic signature. Two major functional cliques were separated based on gene-ontology. Clique 1 (orange): blood vessel development, cell adhesion, regulation of cell proliferation and mammary gland development. Clique 2 (blue): cell division, DNA replication, and mitosis. Genes with hazard ratio for RFS >1 are indicated as circles and those with HR<1 as squares. **C.** Correlation network with negatively correlated genes and its association with cell division, DNA replication, and mitosis genes, as well as some blood vessel development, cell adhesion, regulation of cell proliferation, and mammary gland development genes.

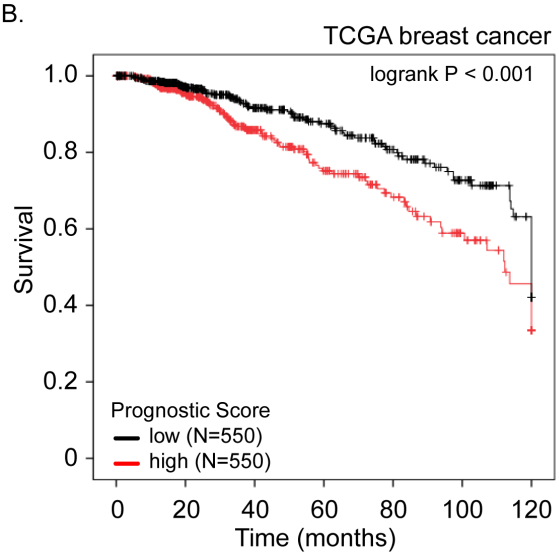
score and divided into two equal sized cohorts. Kaplan-Meier analysis revealed a significant difference between the two patient cohorts. Patients with a high prognostic score had a significantly shorter overall survival compared to patients with a low prognostic score (Figure 6B; p<0.001). To determine if our prognostic score was independent of age at diagnosis, tumor stage, estrogen- and progesterone-receptor status, we ran multivariate Cox regression force-entry with these factors including the prognostics scores as covariates. We found that prognostic score, age at diagnosis and tumor stages III and IV (compared to stage I) were significantly associated with overall survival (Figure 6C). These data confirmed that our prognostic score has clinical validity independent of tumor stage and age at diagnosis (p-value=0.007, HR=2.1, 95% CI:1.2-3.7) (Figure 6C).

We further validated our 12-gene prognostic signature in a second independent breast cancer dataset consisting of 1980 BC patients and containing data for individual breast cancer molecular subtypes (METABRIC; [17, 18]). Prognostic scores for all 1980 patients were calculated as described above for the TCGA cohort and patients were ranked based on their score and divided into two equal sized cohorts. Kaplan-Meier analysis revealed a significant difference between the two patient cohorts (Figure 7A; p=1.01E-17). To address the interaction of our signature with breast cancer molecular subtypes we stratified our patient cohort by molecular subtype (based on PAM50; [19]) and used Kaplan-Meier analysis to investigate differences in survival between the low and high prognostic score cohorts. We found that higher prognostic score was significantly associated with shorter

A.




B.




C.

| Factors | Hazard Ratio | 95% CI | Significance |
|---|---|---|---|
| Tumor stage II vs I | 1.616 | 0.93-2.82 | .091 |
| Tumor stage III vs I | 3.341 | 1.87-5.98 | .000 |
| Tumor stage IV vs I | 10.939 | 5.31-22.55 | .000 |
| Estrogen receptor status | 1.633 | 0.94-2.85 | .084 |
| Progesterone receptor status | 1.294 | 0.77-2.16 | .325 |
| Prognostic score | 2.123 | 1.23-3.66 | .007 |
| Age at diagnosis | 1.030 | 1.02-1.04 | .000 |

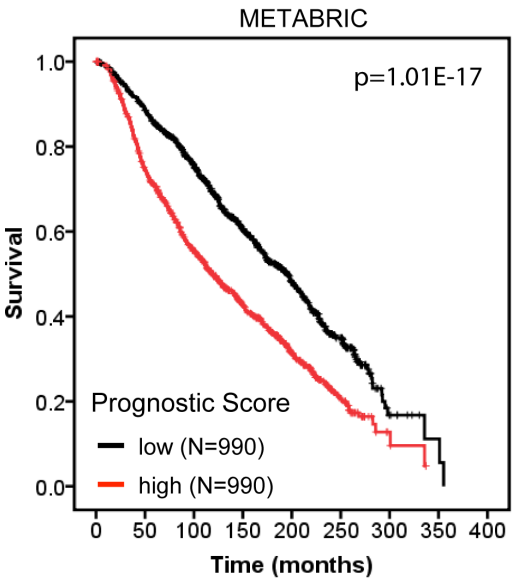
**Figure 6: A 12-gene signature predicts breast cancer patient prognosis. A.** For each of 60 test sets the hazard ratio and the 95% confidence interval was calculated using a Cox model based on the prognostic score with groups as covariates, and subsequently plotted in a forest-plot diagram. The red line indicates a HR value of 1, or the null hypothesis. The two red boxes indicate the insignificant trials (confidence interval included HR value of 1) **B.** Kaplan-Meier overall survival curve for breast cancer patients according to prognostic score using the 12-gene signature. The BC patient cohort was divided into two equal groups based on the prognostic score. The log-rank p-value of the curve comparison between the groups is shown. **C.** The hazard ratio and the 95% confidence interval was calculated using a Cox model based on tumor stage (I-IV), estrogen receptor and progesterone receptor status, age at diagnosis and prognostic score as covariates.

survival in "normal-like", "luminal A" and "HER2" subtypes (Figure 7B). To determine, in this data set, if our prognostic score was independent of age at diagnosis, tumor grade, estrogen- and progesterone-receptor status and molecular subtype (PAM50) we ran multivariate Cox regression force-entry with these factors including the prognostics scores as covariates. We further confirmed that our prognostic signature has clinical validity independent of age at diagnosis, estrogen receptor status, tumor grade and molecular subtype.

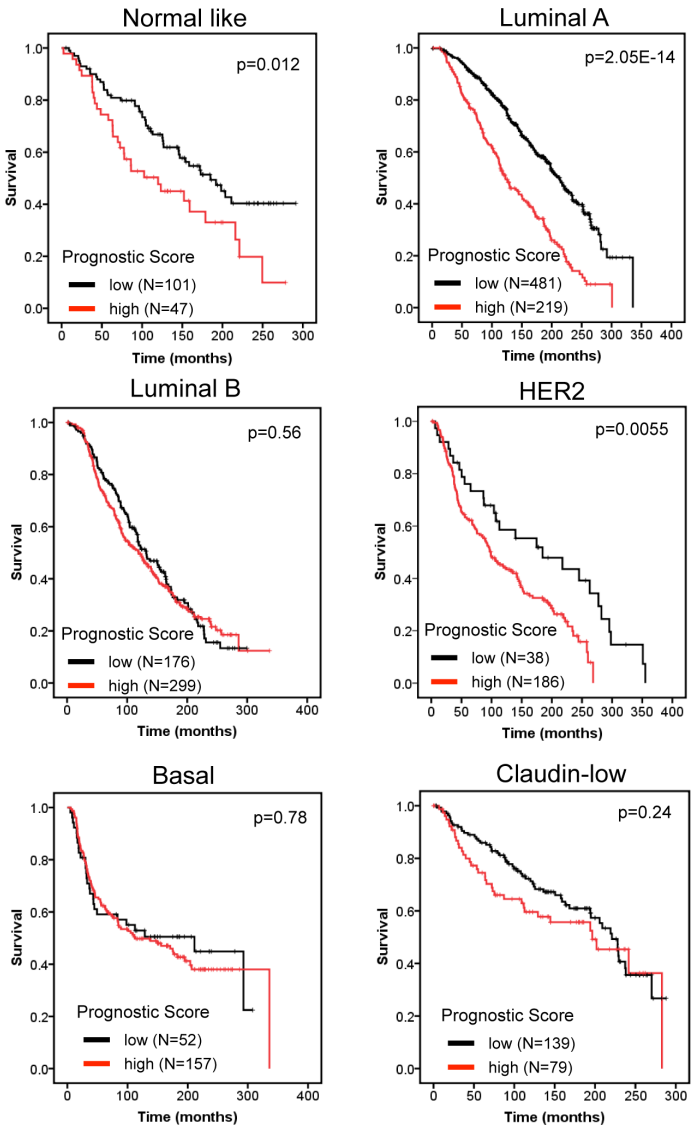
## DISCUSSION

Selecting patients who would most likely benefit from adjuvant systemic therapy is important considering the associated risks of treatment; the development of prognostic biomarkers is useful in this regard. While it remains difficult to identify good targets for the development of targeted therapies, cancer genome analysis has shown great promise in identifying key aberrations in tumor growth and survival pathways that could serve

A.




B.




C.

| Factors | Hazard Ratio | 95% CI | Significance |
|---|---|---|---|
| Age at diagnosis | 1.037 | 1.03-1.04 | 0.000 |
| Estrogen receptor status | .689 | 0.55-0.87 | 0.001 |
| Progesterone receptor status | .914 | 0.79-1.1 | 0.233 |
| Grade | | | 0.010 |
| PAM50 | | | 0.002 |
| Prognostic score | 1.355 | 1.18-1.56 | 0.000 |

**Figure 7: The 12-gene signature predicts overall survival independent of clinical factors and molecular subtypes.** **A.** Kaplan-Meier overall survival curve for breast cancer patients according to prognostic score using the 12-gene signature. The BC patient cohort was divided into two equal groups based on the prognostic score. The log-rank p-value of the curve comparison between the groups is shown. **B.** Kaplan-Meier overall survival curve for breast cancer patients stratified by molecular subtype. The log-rank p-value of the curve comparison between the groups is shown. **C.** The hazard ratio and the 95% confidence interval was calculated using a Cox model based on tumor grade, estrogen receptor and progesterone receptor status, age at diagnosis, molecular subtype (PAM50) and prognostic score as covariates.

as prognostic biomarkers and targets for therapeutic intervention. We created a 12-gene prognostic scoring system, which robustly predicted BC patients' RFS in independent breast cancer data sets. Our gene signature could guide adjuvant therapy for breast cancer patients and includes novel potential molecular targets for therapy. Genes in our signature did not overlap with existing gene signatures that predict breast cancer outcome and metastasis [20–22]. Multiple reasons can explain the lack of overlap between these signatures, including differences in sample size and data sets, clinical phenotypes and methods of signature development. Also, we have shown using co-expression network analysis that functionally related genes often strongly correlate in expression. Even though different signatures select different genes, they may still originate from co-expression cliques representing the same biological function. For example, the Oncotype DX gene signature, which is prognostic of breast cancer recurrence, consists of 16 cancer genes. Five of these genes were also included in our analysis (MKi67, STK15, BIRC5, CCNB1 and MMP11), but were not selected in our final gene signature. However, MKi67, STK15, BIRC5 and CCNB1 were all part of the same strongly interconnected and cell-cycle enriched co-expression clique. Our analysis selected NUF2, MELK and RNASEH2A from the same clique, however, given the strong correlations in expression, any one of the highly connected genes is likely to perform equally well. Using multivariate Cox regression with our 12-gene signature and the Oncotype DX 16-gene signature, we determined that our 12-gene signature was independent ($p<0.005$; HR=2.4, 95% CI:1.7-3.4), but equally important as the Oncotype DX gene signature ($p<0.005$; HR=2.2, 95% CI: 1.3-3.7). Another important variable associated with breast cancer survival is molecular subtype. Using a cohort of 1980 breast cancer patients with approximately 30 years of follow-up we determined that our signature could predict breast cancer patient survival for "normal-like", "luminal-A" and "HER2" subtypes, but not "luminal-B", "basal" and "claudin-low" subtypes. We should point out that in our analysis patients were stratified into two equally sized cohorts based on the median prognostic score and then further stratified by molecular subtype. This resulted in unequally sized cohorts for each subtype, which could potentially have confounded our analysis. To test this, we generated equally sized cohorts based on prognostic score for each individual subtype. We first stratified patients by molecular subtype and then further stratified patients inside each subtype by the median of the prognostic score. This analysis revealed similar observations as presented in Figure 7B confirming that our results are not confounded by unequally sized cohorts within different score groups. Future studies are granted to investigate whether our prognostic score can predict sensitivity to radiation- and/or chemotherapy.

The majority of the genes in our signature have previously been associated with cancer progression and patient outcome. MELK, NUF2 and ZBTB16 play important roles in cell cycle-related processes. Loss of ZBTB16 expression has been reported in a number of different tumor types including prostate cancer, non-small cell lung cancer, melanoma [23–25]. Overexpression of MELK, a serine/threonine kinase implicated in embryogenesis and cell cycle control has been identified in numerous human cancer types including breast, prostate, brain, colorectal and gastric cancer [26–30]. In BC, overexpression of MELK correlated with poor prognosis, whereas knockdown decreased proliferation [28, 30, 31]. NUF2 is part of a conserved protein complex associated with the centromere and plays a regulatory role in chromosomal segregation. Down regulation of NUF2 in pancreatic cancer cell lines inhibited tumor growth and enhanced apoptosis [32] whereas upregulation of NUF2 in colon cancer cells promoted tumorigenicity [33]. Overexpression of EPS15, which plays a role in terminating growth factor signaling, was shown to be a favorable prognostic factor in BC [34, 35]. Our signature also included the inner mitochondrial membrane protein TIMM17A. Decreased expression of TIMM17A reduced the aggressiveness of BC cells and TIMM17A expression was significantly associated with BC survival [36–38]. PLEKHH2 and TSC1 are involved in cell adhesion and actin dynamics. Loss of TSC1 was shown to result in the deregulation of cell motility and adhesion [39]. A polymorphic variant of TSC1 was associated with delayed age at diagnosis of ER-positive ductal carcinomas [40]. Also, TSC1, in coordination with TSC2, inhibits MTOR, which promotes cell growth and cell cycle progression [41]. PLPP1 degrades lysophosphatidate and is often down-regulated in tumor types. Using syngeneic and xenograft mouse models showed that PLPP1 overexpression in BC cells decreased tumor growth and the metastasis [42]. S100P is overexpressed in a variety of human tumor types [43]. S100P transcription is influenced by a number of signaling molecules including progesterone, androgens, glucocorticoids, BMP4 and IL6 and through interactions with a various proteins integrates and regulates multiple signaling pathways involved in degradation of extracellular matrix, invasion and metastasis and tumorigenesis (reviewed in [44]).

The role of PLEKHH2, USP47 and THYN1 has not been extensively studied in cancer progression. PLEKHH2 protein was enriched in renal glomerular podocytes, and shown to interact with focal adhesion proteins and actin to stabilize the actin cytoskeleton [45]. USP47 plays an important role in base-excision repair and the maintenance of genome integrity [46]. Depletion of USP47 induced accumulation of Cdc25A and decreased cell survival [47]. However, our results indicate that patients with high breast tumor expression of USP47 have significantly

better relapse-free survival compared to patients with low breast tumor expression of USP47 (HR=0.65; p-value=2.40E-07). Thus, the exact role of USP47 in BC has yet to be determined. The role of THYN1 in BC is currently unknown, however, downregulation of THYN1 has been correlated with the induction of apoptosis in a specific B-cell lymphoma cell line [48].

Our gene co-expression network analysis identified a number of potential therapeutic targets. We found that 7 genes CREBRF, DIXDC1, AHNAK, CYBRD1, NOSTRIN, TNS2 and TNFSF12 were negatively correlated with the strongly interconnected cell cycle and mitosis clique. Indeed, a number of these genes have been identified as candidate tumor suppressor genes including CREBRF, DIXDC1, AHNAK and TNS2 [49–52]. Furthermore, NOSTRIN was found to be a potential negative regulator of disease aggressiveness in pancreatic cancer and CYBRD1 was identified as part of an iron regulatory gene signature that predicts outcome in BC [53, 54]. TNFSF12 (TWEAK) can promote cell death in tumor cell lines under certain conditions [55–57], and may also activate local macrophages to inhibit tumor progression [58]. The negative correlation of these 7 genes with the cell cycle enriched gene co-expression clique was observed in the co-expression network of breast tumor samples, but not the normal breast tissue co-expression network. This suggests that a therapeutic approach that increases expression of one or more of these 7 genes could collapse the tumor cell cycle machinery, while sparing adverse effects in healthy tissue.

In summary, we have generated a prognostic scoring system and 12-gene signature that is prognostic of BC patient relapse-free survival. Furthermore, using co-expression network analysis, we investigated the genetic architecture of RFS associated genes in normal and tumor tissues and identified 7 potential therapeutic targets that could be developed to target the tumor cell cycle machinery. Our analysis pipeline could furthermore be applied to other tumor types.

## MATERIALS AND METHODS

### Data sets used in this study

Gene transcript data of normal and tumor breast tissues was obtained from NCBI GEO accession numbers: GSE3744 (40 invasive ductal carcinoma samples and 7 normal breast samples), GSE10780 (42 invasive ductal carcinoma samples and 143 normal breast samples), GSE21422 (5 invasive ductal carcinoma samples and 5 normal breast samples) and GSE29044 (72 invasive ductal carcinoma samples and 36 normal breast samples). Normal breast gene transcript data used for generating gene expression correlation networks was obtained from GTEX (http://www.gtexportal.org/home/datasets) using the RPKM normalized gene transcript counts table [9, 10].

### Statistical analysis

GEO2R was used to calculate the differential expression of tumor versus normal using a fold-change cutoff of 1.5 and adjusted p-value 0.01. Association of differentially expressed genes and relapse-free survival in breast cancer patients was assessed using Kaplan-Meier plotter (http://kmplot.com) including KM survival analysis, hazard ratio with 95% confidence interval and logrank p-value for each gene using all available patients (not restricted to any clinical parameters such as grade, PR status, etc) [16].

Gene ontology enrichment analysis was performed using the web-based gene set analysis toolkit (adjusted $p<0.05$ was used as a threshold for significance) (http://bioinfo.vanderbilt.edu/webgestalt/) [59, 60].

### Gene co-expression network construction

Gene expression Spearman correlation coefficients were calculated in "R" for 795 probes (587 genes) that were differentially expressed between breast tumor and normal tissues samples. A gene network was generated where nodes represent individual genes and edges connecting nodes were drawn when the correlation coefficient exceeded $|R|\geq 0.6$ (adjusted p-value$\leq$7.911E-08). The gene co-expression network was visualized using Cytoscape 3.1.1. (http://www.cytoscape.org).

### Prognostic gene signature

BC microarray data (GSE6532), describing RFS status and gene expression for our 357-gene panel, was collected for 393 patients. Sixty random samplings of 300 patients were extracted from this dataset and used as training sets to identify a biomarker panel associated with RFS. The residual 93 patients from each sample were used as test sets to validate the prognostic significance of the biomarker panel. A forward-conditional Cox regression using all 357 genes as covariates was performed using SPSS on each of the training sets in order to identify the biomarker panel. The results of each test were recorded and the genes that appeared in more than half of the training sets were included in our biomarker panel.

Cox regression was repeated on all 60 training sets using our 12-gene signature as covariates using the forced-entry (enter) method to obtain the beta values (coefficient) for each biomarker. The resulting 60 beta values of each biomarker were averaged to estimate the true beta value of each gene. A prognostic scoring system was created based on this formula:

$$\sum_{i=1}^{12} (gene\ i\ b)\ x\ (gene\ i\ expression\ level)$$

The patients were ranked by their prognostic score and divided into two equal sized cohorts. Kaplan-Meier plots were constructed and a long-rank test was used to determine differences among relapse free survival.

Prognostic scores for each of the test set samples were then calculated using the same set of mean beta values developed in the training set. Patients were ranked based on their prognostic score and divided into two cohorts based on the average prognostic-score cut-point in the training sets. Kaplan-Meier plots were constructed and a log-rank test was used to determine differences among RFS.

To further validate our biomarker panel, mRNA expression levels (normalized RNA-seq mRNA expression z-scores) for our 12-gene signature were obtained from cBioPortal for 1100 breast cancer samples (TCGA; http://www.cbioportal.org/data_sets.jsp) [61, 62] and for 1980 breast cancer samples (METABRIC) [17, 18]. New beta values for each of the twelve biomarkers were obtained using Cox regression. Prognostic scores were calculated and patients were ranked based on their score and divided into two equal sized cohorts. Kaplan-Meier analysis and a log-rank test were used to determine differences in survival.

## ACKNOWLEDGMENTS

A.M.S. was supported by the Low Dose Scientific Focus Area, Office of Biological and Environmental Research, U.S. Department of Energy under Contract No. DE AC02-05CH11231.

## CONFLICTS OF INTEREST

The authors declare that they have no competing interests.

## REFERENCES

1. DeSantis C, Ma J, Bryan L, Jemal A. Breast cancer statistics, 2013. CA Cancer J Clin. 2014; 64:52-62.
2. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. CA Cancer J Clin. 2014; 64:9-29.
3. Holloway CM, Easson A, Escallon J, Leong WL, Quan ML, Reedjik M, Wright FC, McCready DR. Technology as a force for improved diagnosis and treatment of breast disease. Canadian journal of surgery. 2010; 53:268-277.
4. Duffy SW, Lynge E, Jonsson H, Ayyaz S, Olsen AH. Complexities in the estimation of overdiagnosis in breast cancer screening. British journal of cancer. 2008; 99:1176-1178.
5. Glass AG, Lacey JV, Jr., Carreon JD, Hoover RN. Breast cancer incidence, 1980-2006: combined roles of menopausal hormone therapy, screening mammography, and estrogen receptor status. Journal of the National Cancer Institute. 2007; 99:1152-1161.
6. Anampa J, Makower D, Sparano JA. Progress in adjuvant chemotherapy for breast cancer: an overview. BMC medicine. 2015; 13:195.
7. Chew HK. Adjuvant therapy for breast cancer: who should get what? The Western journal of medicine. 2001; 174:284-287.
8. Cianfrocca M, Goldstein LJ. Prognostic and predictive factors in early-stage breast cancer. The oncologist. 2004; 9:606-616.
9. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nature genetics. 2013; 45:580-585.
10. Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, Johnson R, Segre AV, Djebali S, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015; 348:660-665.
11. Alimonti A, Carracedo A, Clohessy JG, Trotman LC, Nardella C, Egia A, Salmena L, Sampieri K, Haveman WJ, Brogi E, Richardson AL, Zhang J, Pandolfi PP. Subtle variations in Pten dose determine cancer susceptibility. Nature genetics. 2010; 42:454-458.
12. Richardson AL, Wang ZC, De Nicolo A, Lu X, Brown M, Miron A, Liao X, Iglehart JD, Livingston DM, Ganesan S. X chromosomal abnormalities in basal-like human breast cancer. Cancer cell. 2006; 9:121-132.
13. Colak D, Nofal A, Albakheet A, Nirmal M, Jeprel H, Eldali A, Al-Tweigeri T, Tulbah A, Ajarim D, Malik OA, Inan MS, Kaya N, Park BH, Bin Amer SM. Age-specific gene expression signatures for breast tumors and cross-species conserved potential cancer progression markers in young women. PloS one. 2013; 8:e63204.
14. Kretschmer C, Sterner-Kock A, Siedentopf F, Schoenegg W, Schlag PM, Kemmner W. Identification of early molecular markers for breast cancer. Molecular cancer. 2011; 10:15.
15. Chen DT, Nasir A, Culhane A, Venkataramu C, Fulp W, Rubio R, Wang T, Agrawal D, McCarthy SM, Gruidl M, Bloom G, Anderson T, White J, Quackenbush J, Yeatman T. Proliferative genes dominate malignancy-risk gene signature in histologically-normal breast tissue. Breast cancer research and treatment. 2010; 119:335-346.
16. Gyorffy B, Lanczky A, Eklund AC, Denkert C, Budczies J, Li Q, Szallasi Z. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. Breast cancer research and treatment. 2010; 123:725-731.
17. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Graf S, Ha G, Haffari G, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012; 486:346-352.

18. Pereira B, Chin SF, Rueda OM, Vollan HK, Provenzano E, Bardwell HA, Pugh M, Jones L, Russell R, Sammut SJ, Tsui DW, Liu B, Dawson SJ, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nature communications. 2016; 7:11479.

19. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology. 2009; 27:1160-1167.

20. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415:530-536.

21. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005; 365:671-679.

22. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. The New England journal of medicine. 2004; 351:2817-2826.

23. Xiao GQ, Unger P, Yang Q, Kinoshita Y, Singh K, McMahon L, Nastiuk K, Sha K, Krolewski J, Burstein D. Loss of PLZF expression in prostate cancer by immunohistochemistry correlates with tumor aggressiveness and metastasis. PloS one. 2015; 10:e0121318.

24. Wang X, Wang L, Guo S, Bao Y, Ma Y, Yan F, Xu K, Xu Z, Jin L, Lu D, Xu J, Wang JC. Hypermethylation reduces expression of tumor-suppressor PLZF and regulates proliferation and apoptosis in non-small-cell lung cancers. FASEB journal. 2013; 27:4194-4203.

25. Brunner G, Reitz M, Schwipper V, Tilkorn H, Lippold A, Biess B, Suter L, Atzpodien J. Increased expression of the tumor suppressor PLZF is a continuous predictor of long-term survival in malignant melanoma patients. Cancer biotherapy & radiopharmaceuticals. 2008; 23:451-459.

26. Du T, Qu Y, Li J, Li H, Su L, Zhou Q, Yan M, Li C, Zhu Z, Liu B. Maternal embryonic leucine zipper kinase enhances gastric cancer progression via the FAK/Paxillin pathway. Molecular cancer. 2014; 13:100.

27. Kuner R, Falth M, Pressinotti NC, Brase JC, Puig SB, Metzger J, Gade S, Schafer G, Bartsch G, Steiner E, Klocker H, Sultmann H. The maternal embryonic leucine zipper kinase (MELK) is upregulated in high-grade prostate cancer. Journal of molecular medicine. 2013; 91:237-248.

28. Pickard MR, Green AR, Ellis IO, Caldas C, Hedge VL, Mourtada-Maarabouni M, Williams GT. Dysregulated expression of Fau and MELK is associated with poor prognosis in breast cancer. Breast cancer research. 2009; 11:R60.

29. Nakano I, Masterman-Smith M, Saigusa K, Paucar AA, Horvath S, Shoemaker L, Watanabe M, Negro A, Bajpai R, Howes A, Lelievre V, Waschek JA, Lazareff JA, et al. Maternal embryonic leucine zipper kinase is a key regulator of the proliferation of malignant brain tumors, including brain tumor stem cells. Journal of neuroscience research. 2008; 86:48-60.

30. Gray D, Jubb AM, Hogue D, Dowd P, Kljavin N, Yi S, Bai W, Frantz G, Zhang Z, Koeppen H, de Sauvage FJ, Davis DP. Maternal embryonic leucine zipper kinase/murine protein serine-threonine kinase 38 is a promising therapeutic target for multiple cancers. Cancer research. 2005; 65:9751-9761.

31. Wang Y, Lee YM, Baitsch L, Huang A, Xiang Y, Tong H, Lako A, Von T, Choi C, Lim E, Min J, Li L, Stegmeier F, et al. MELK is an oncogenic kinase essential for mitotic progression in basal-like breast cancer cells. eLife. 2014; 3:e01763.

32. Hu P, Shangguan J, Zhang L. Downregulation of NUF2 inhibits tumor growth and induces apoptosis by regulating lncRNA AF339813. International journal of clinical and experimental pathology. 2015; 8:2638-2648.

33. Sugimasa H, Taniue K, Kurimoto A, Takeda Y, Kawasaki Y, Akiyama T. Heterogeneous nuclear ribonucleoprotein K upregulates the kinetochore complex component NUF2 and promotes the tumorigenicity of colon cancer cells. Biochemical and biophysical research communications. 2015; 459:29-35.

34. Dai X, Liu Z, Zhang S. Over-expression of EPS15 is a favorable prognostic factor in breast cancer. Molecular bioSystems. 2015; 11:2978-2985.

35. Amatschek S, Koenig U, Auer H, Steinlein P, Pacher M, Gruenfelder A, Dekan G, Vogl S, Kubista E, Heider KH, Stratowa C, Schreiber M, Sommergruber W. Tissue-wide expression profiling using cDNA subtraction and microarrays to identify tumor-specific genes. Cancer research. 2004; 64:844-856.

36. Yang X, Si Y, Tao T, Martin TA, Cheng S, Yu H, Li J, He J, Jiang WG. The Impact of TIMM17A on Aggressiveness of Human Breast Cancer Cells. Anticancer research. 2016; 36:1237-1241.

37. Salhab M, Patani N, Jiang W, Mokbel K. High TIMM17A expression is associated with adverse pathological and clinical outcomes in human breast cancer. Breast cancer. 2012; 19:153-160.

38. Xu X, Qiao M, Zhang Y, Jiang Y, Wei P, Yao J, Gu B, Wang Y, Lu J, Wang Z, Tang Z, Sun Y, Wu W, Shi Q. Quantitative proteomics study of breast cancer cell lines isolated from

a single patient: discovery of TIMM17A as a marker for breast cancer. Proteomics. 2010; 10:1374-1390.

39. Goncharova E, Goncharov D, Noonan D, Krymskaya VP. TSC2 modulates actin cytoskeleton and focal adhesion through TSC1-binding domain and the Rac1 GTPase. The Journal of cell biology. 2004; 167:1171-1182.

40. Mehta MS, Vazquez A, Kulkarni DA, Kerrigan JE, Atwal G, Metsugi S, Toppmeyer DL, Levine AJ, Hirshfield KM. Polymorphic variants in TSC1 and TSC2 and their association with breast cancer phenotypes. Breast cancer research and treatment. 2011; 125:861-868.

41. Findlay GM, Harrington LS, Lamb RF. TSC1-2 tumour suppressor and regulation of mTOR signalling: linking cell growth and proliferation? Current opinion in genetics & development. 2005; 15:69-76.

42. Tang X, Benesch MG, Dewald J, Zhao YY, Patwardhan N, Santos WL, Curtis JM, McMullen TP, Brindley DN. Lipid phosphate phosphatase-1 expression in cancer cells attenuates tumor growth and metastasis in mice. Journal of lipid research. 2014; 55:2389-2400.

43. Parkkila S, Pan PW, Ward A, Gibadulinova A, Oveckova I, Pastorekova S, Pastorek J, Martinez AR, Helin HO, Isola J. The calcium-binding protein S100P in normal and malignant human tissues. BMC clinical pathology. 2008; 8:2.

44. Prica F, Radon T, Cheng Y, Crnogorac-Jurcevic T. The life and works of S100P - from conception to cancer. American journal of cancer research. 2016; 6:562-576.

45. Perisic L, Lal M, Hulkko J, Hultenby K, Onfelt B, Sun Y, Duner F, Patrakka J, Betsholtz C, Uhlen M, Brismar H, Tryggvason K, Wernerson A, Pikkarainen T. Plekhh2, a novel podocyte protein downregulated in human focal segmental glomerulosclerosis, is involved in matrix adhesion and actin dynamics. Kidney international. 2012; 82:1071-1083.

46. Parsons JL, Dianova, II, Khoronenkova SV, Edelmann MJ, Kessler BM, Dianov GL. USP47 is a deubiquitylating enzyme that regulates base excision repair by controlling steady-state levels of DNA polymerase beta. Molecular cell. 2011; 41:609-615.

47. Peschiaroli A, Skaar JR, Pagano M, Melino G. The ubiquitin-specific protease USP47 is a novel beta-TRCP interactor regulating cell survival. Oncogene. 2010; 29:1384-1393.

48. Jiang XZ, Toyota H, Yoshimoto T, Takada E, Asakura H, Mizuguchi J. Anti-IgM-induced down-regulation of nuclear Thy28 protein expression in Ramos B lymphoma cells. Apoptosis. 2003; 8:509-519.

49. Lee IH, Sohn M, Lim HJ, Yoon S, Oh H, Shin S, Shin JH, Oh SH, Kim J, Lee DK, Noh DY, Bae DS, Seong JK, Bae YS. Ahnak functions as a tumor suppressor via modulation of TGFbeta/Smad signaling pathway. Oncogene. 2014; 33:4675-4684.

50. Goodwin JM, Svensson RU, Lou HJ, Winslow MM, Turk BE, Shaw RJ. An AMPK-independent signaling pathway downstream of the LKB1 tumor suppressor controls Snail1 and metastatic potential. Molecular cell. 2014; 55:436-450.

51. Hong SY, Shih YP, Sun P, Hsieh WJ, Lin WC, Lo SH. Down-regulation of tensin2 enhances tumorigenicity and is associated with a variety of cancers. Oncotarget. 2016; 7:38143-38153. doi: 10.18632/oncotarget.9411.

52. Xue H, Zhang J, Guo X, Wang J, Li J, Gao X, Guo X, Li T, Xu S, Zhang P, Liu Q, Li G. CREBRF is a potent tumor suppressor of glioblastoma by blocking hypoxia-induced autophagy via the CREB3/ATG5 pathway. International journal of oncology. 2016; 49:519-528.

53. Miller LD, Coffman LG, Chou JW, Black MA, Bergh J, D'Agostino R, Jr., Torti SV, Torti FM. An iron regulatory gene signature predicts outcome in breast cancer. Cancer research. 2011; 71:6728-6737.

54. Wang J, Yang S, He P, Schetter A, Gaedcke J, Ghadimi BM, Ried T, Yfantis HG, Lee DH, Gaida MM, Hanna N, Alexander HR, Hussain SP. Endothelial Nitric Oxide Synthase Traffic Inducer (NOSTRIN) is a Negative Regulator of Disease Aggressiveness in Pancreatic Cancer. Clinical cancer research. 2016.

55. Wilson CA, Browning JL. Death of HT29 adenocarcinoma cells induced by TNF family receptor activation is caspase-independent and displays features of both apoptosis and necrosis. Cell death and differentiation. 2002; 9:1321-1333.

56. Nakayama M, Ishidoh K, Kayagaki N, Kojima Y, Yamaguchi N, Nakano H, Kominami E, Okumura K, Yagita H. Multiple pathways of TWEAK-induced cell death. Journal of immunology. 2002; 168:734-743.

57. Schneider P, Schwenzer R, Haas E, Muhlenbeck F, Schubert G, Scheurich P, Tschopp J, Wajant H. TWEAK can induce cell death via endogenous TNF and TNF receptor 1. European journal of immunology. 1999; 29:1785-1792.

58. Gu L, Dai L, Cao C, Zhu J, Ding C, Xu HB, Qiu L, Di W. Functional expression of TWEAK and the receptor Fn14 in human malignant ovarian tumors: possible implication for ovarian tumor intervention. PloS one. 2013; 8:e57436.

59. Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic acids research. 2013; 41:W77-83.

60. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic acids research. 2005; 33:W741-748.

61. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery. 2012; 2:401-404.

62. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science signaling. 2013; 6:pl1.