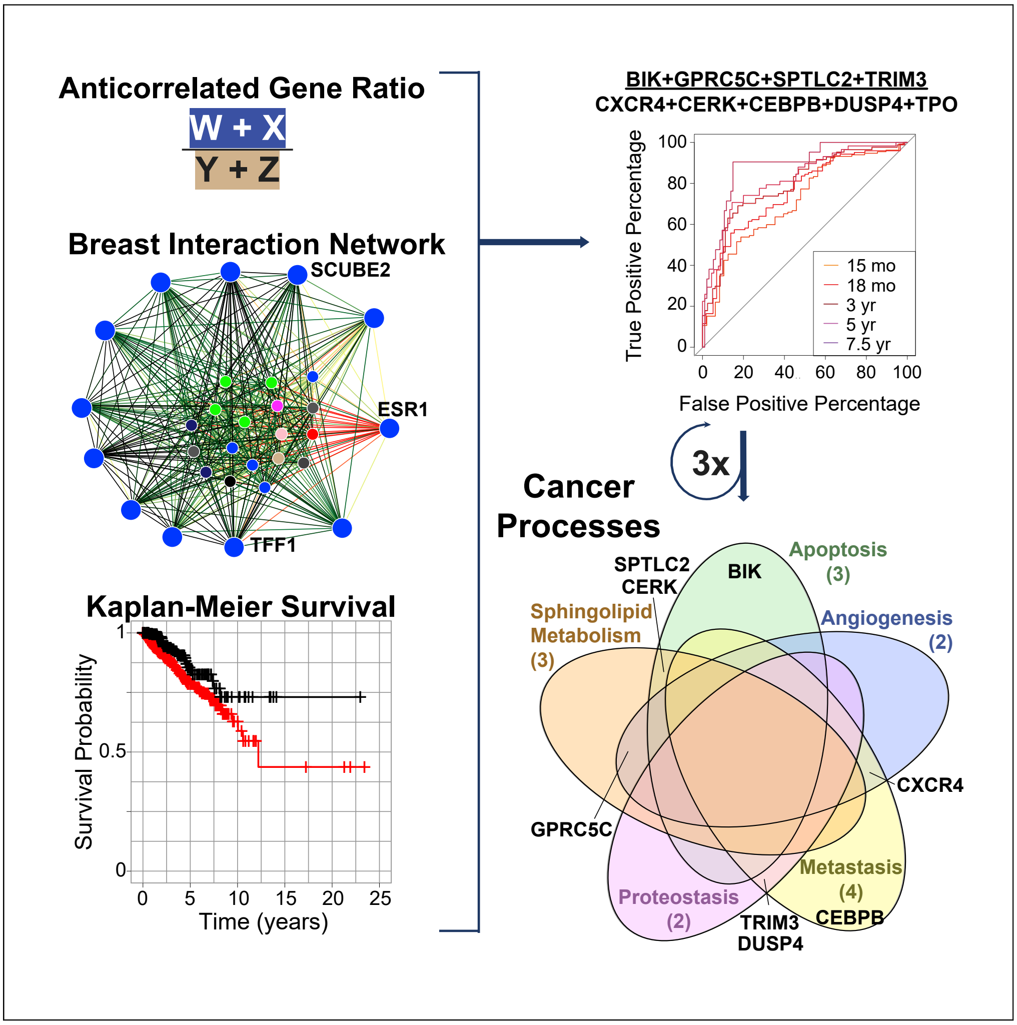
iScience

CellPress
OPEN ACCESS

Article

A network approach reveals driver genes associated with survival of patients with triple-negative breast cancer



Courtney D. Dill, Eric B. Dammer, Ti'ara L. Griffen, Nicholas T. Seyfried, James W. Lillard, Jr.

jlillard@msm.edu

Highlights

WGCNA identifies coexpression modules predicted to drive TNBC patient survival

Module hubs and DEGs reveal transcriptional changes associated with high survival

Nine genes act synergistically to influence TNBC progression, relapse, and survival

These genes' levels represent reversible changes in TNBC hallmark cancer processes

Dill et al., iScience *24*, 102451
May 21, 2021 © 2021 The Authors.
https://doi.org/10.1016/j.isci.2021.102451

iScience

CellPress OPEN ACCESS

Article

A network approach reveals driver genes associated with survival of patients with triple-negative breast cancer

Courtney D. Dill,[1] Eric B. Dammer,[2,3] Ti'ara L. Griffen,[1] Nicholas T. Seyfried,[2,3,4] and James W. Lillard, Jr.[1,5,*]

SUMMARY

We aimed to identify triple-negative breast cancer (TNBC) drivers that regulate survival time as predictive signatures that improve TNBC prognostication. Breast cancer (BrCa) transcriptomic tumor biopsies were analyzed, identifying network communities enriched with TNBC-specific differentially expressed genes (DEGs) and correlated strongly to TNBC status. Two anticorrelated modules correlated strongly to TNBC subtype and survival. Querying module-specific hubs and DEGs revealed transcriptional changes associated with high survival. Transcripts were nominated as biomarkers and tested as combinatoric ratios using receiver operator characteristic (ROC) analysis to assess survival prediction. ROC test rounds integrated genes with established interactions to hubs and DEGs of key modules, improving prediction. Finally, we tested whether integration of literature-derived genes for implicated hallmark cancer processes could improve prediction of survival. Complementary coexpression, differential expression, genetic interaction, and survival stratification integrated by ROC optimization uncovered a panel of "linchpin survival genes" predictive of patient survival, representing gene interactions in hallmark cancer processes.

INTRODUCTION

Triple-negative breast cancer (TNBC) accounts for 10 to 20% of all invasive breast cancer (BrCa) cases and lacks estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2 (HER2) responsiveness. TNBC is an aggressive cancer associated with poor prognosis relative to non-TNBC, that is, higher and earlier risk of relapse and recurrence, as well as poor survival. The 5-year overall and disease-free survival rates are 62.1% and 57.5% for TNBC versus 80.8% and 75.3% for non-TNBC, respectively ($p < 0.001$) (Goncalves et al., 2018).

Owing to loss of receptor signaling, TNBC does not respond to hormone receptor or HER2-directed therapies. Instead, standard therapy consists of a combination of chemotherapeutic drugs demonstrating marginal efficacy. The ineffective nature of current TNBC therapies suggests that the genetic and molecular patterns associated with TNBC progression are complex with interconnected dysregulation. Thus, there is a need to understand the molecules and mechanisms associated with survival of patients with TNBC at a systems level, which will allow for better prediction of gene targets in TNBC and improve TNBC prognostication.

In line with known poor TNBC prognosis, this subtype exhibits enhanced evasion of apoptosis, angiogenesis, and metastasis, among hallmark cancer processes as defined by Hanahan and Weinberg (Hanahan and Weinberg, 2011). Although there have been advances in our understanding of the biology of primary breast tumors relating to these and other hallmark processes, our knowledge of how and why patients with TNBC experience poor prognoses compared with patients without TNBC is limited and warrants a circumspect unbiased analysis.

Using an integrative systems biology approach, we performed a cross-platform meta-analysis of large numbers of BrCa transcriptomes curated in The Cancer Genome Atlas (TCGA; N = 777) and GEO DataSets (N = 1,234) to assess the correlation between gene transcript expression in coexpressed transcript modules to sample traits relevant to TNBC. The network structure was leveraged to nominate genes whose transcript levels in combination are tied to survival in either generalized BrCa or specifically TNBC, nominating

[1] Department of Microbiology, Biochemistry, and Immunology, Morehouse School of Medicine, 720 Westview Dr SW, HG 341B, Atlanta, GA 30310, USA

[2] Center for Neurodegenerative Disease, Emory University School of Medicine, Atlanta, GA 30322, USA

[3] Department of Biochemistry, Emory University School of Medicine, Atlanta, GA 30322, USA

[4] Department of Neurology, Emory University School of Medicine, Atlanta, GA 30322, USA

[5] Lead contact

*Correspondence: jlillard@msm.edu

https://doi.org/10.1016/j.isci.2021.102451

genes to predict patient survival. Genes were then combined in a ratio with choice of numerator or denominator set by membership in anticorrelated modules of interest. The best of the nominated gene combinations we ranked by receiver operator characteristic (ROC) analysis implicate both known and potentially unappreciated genes in hallmark dysregulated cancer processes.

Overall, this study successfully leveraged the coexpression network of the various BrCa subtypes in large transcriptomic cohorts including patients with TNBC, overlapped this structure with differential expression, selected candidate transcripts influencing survival overall in BrCa, and then homed-in on TNBC-specific linchpin survival genes. Selection of genes implicated in the gene interaction network of breast tissue also contributed to the optimized survival gene list. Mining of the prognosis indicators nominated here for testing in future work is warranted to discover the complex molecular and functional interactions among hallmark cancer process genes driving mortality in patients with TNBC, with TNBC subtype-specific treatment outcomes.

RESULTS

Overview of systems-biology-leveraged survival gene discovery workflow

The phenotype of less severe BrCa subtypes, such as Luminal A, Luminal B, and HER2-enriched types are well characterized, which has led to the discovery of drug therapies that have increased survival time for these patient groups. However, the molecules and mechanisms responsible for the more lethal phenotype of TNBC involve more complex interactions and presently, a targeted therapy for TNBC does not exist. This unmet need fostered our development of a nontraditional pipeline to define and survey the systems biology landscape of BrCa and TNBC for identifying TNBC targets (Figure 1). The literature for BrCa of the past >20 years is strewn with reports of prognostic indicators which validate but are not part of an integrated framework of understanding. Here, we leverage differences in transcript levels unique to TNBC in the context of systems biology: **(Function 1)** a coexpression network encompassing all BrCa subtypes' RNA quantitation (N = 777; TCGA RNA-seq cohort—input **(input A)**; a breakdown of clinical traits and tumor staging for the TCGA BRCA cohort is provided in Tables S1–S3**. (Function 2)** Genetic interactions specific to breast tissue from the genome-scale integrated analysis of gene networks in tissues (GIANT) analysis framework (Wong et al., 2018) also were integrated into the nomination of genes that can predict survival and therefore are most likely to affect it. The more traditional **(function 3)** Kaplan-Meier (KM) survival analysis focused on coexpression network hubs of key TNBC-associated modules intersecting with **(function 4)** differentially expressed genes in BrCa subtypes. The KM significant genes were nominated transcripts for **(function 5)** ROC survival prediction that were moderately predictive in both BrCa in general and TNBC specifically as assessed by predictor area under the curve (AUC) in an independent meta-analysis of 1,234 array-measured BrCa transcriptomes—**(input B)**. More than 100,000 combinations of genes encompassing a small tract of the total network landscape of relevant differentially expressed gene combinations were tested by an additional round of ROC to improve survival prediction specifically in TNBC, where the genes ranked best by AUC shifted when focusing only on the 180 TNBC cases in the array metanalysis. The resulting top ranked prognostic indicator specific to TNBC was finally improved once more by identification of the hallmark cancer processes implicated by the genes already found though our unbiased nomination process, and a final network-informed addition of known gene products from literature **(input C)** regarding those hallmark cancer processes highly relevant to BrCa (Figure 1, Venn diagram at bottom left). We then validated the array data (**input B**) ROC test result using ROC analysis of the same genes found in TCGA input data (**input A**). A total of 111,385 combinations of genes were evaluated using area under the ROC curve (AUC) analysis. The final ratio of gene abundances significantly outperformed prediction of survival of patients with TNBC compared with patients without TNBC, identified as those best improving survival prediction of patients with TNBC. An extensive literature search demonstrated final gene indicators are associated with hallmark cancer processes. Our final indicator identified sphingosine/ceramide balance, balance of proapoptotic and antiapoptotic signaling, proteostasis, angiogenesis, and metastasis as processes regulated by our indicator genes. Importantly, while our results are informed by known gene signatures, 7 of 9 of the survival-linked gene products, which we nominate, were arrived at via unbiased exploration of the network landscape.

Identification of coexpressed BrCa biopsy transcript communities and initial selection of a subset of modules associated with TNBC.

Coexpression simplifies biomarker panel selection and enables gene equivalence determinations for consolidation of such panels, including ones for molecular and other BrCa subtypes (Wirapati et al.,

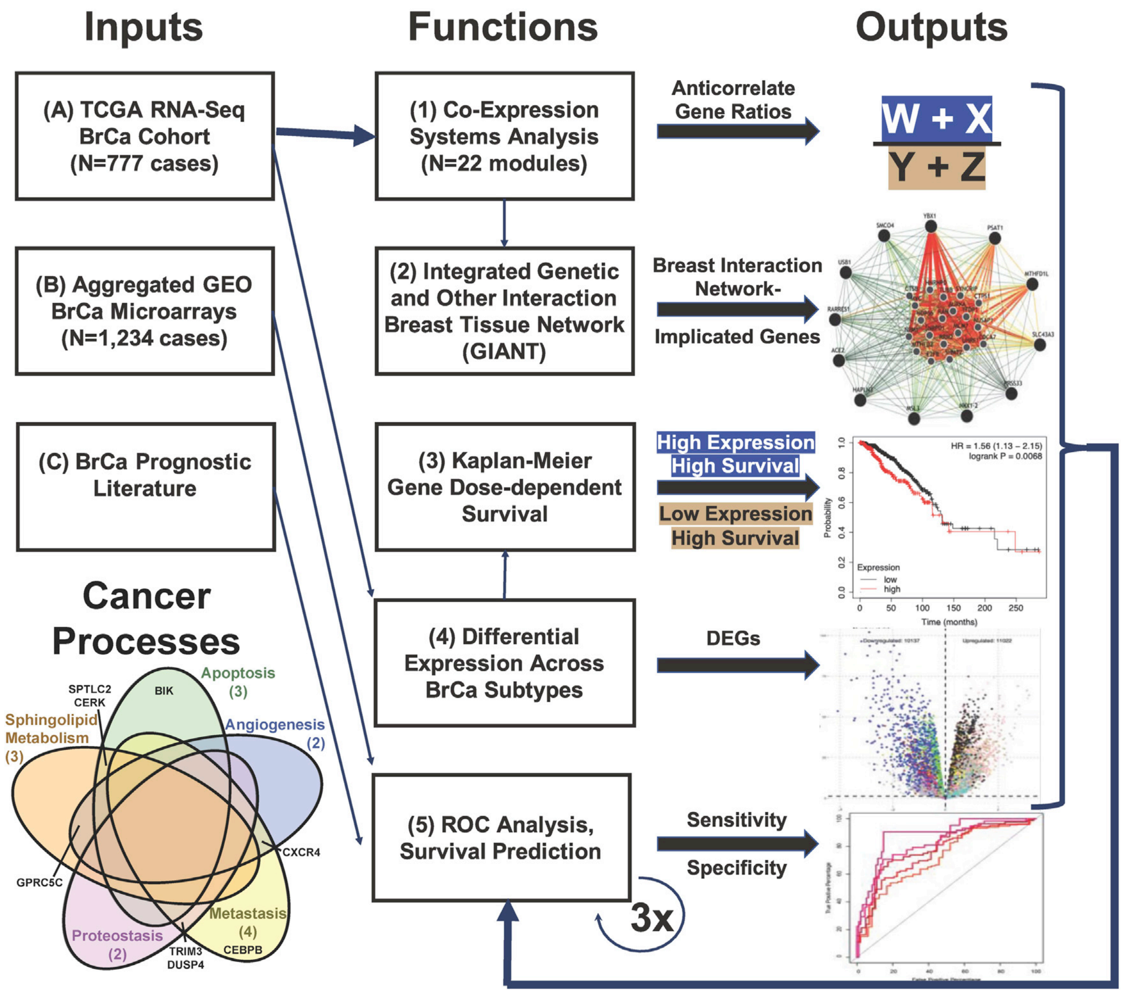

Inputs
Functions
Outputs
(A) TCGA RNA-Seq BrCa Cohort (N=777 cases)
(B) Aggregated GEO BrCa Microarrays (N=1,234 cases)
(C) BrCa Prognostic Literature
(1) Co-Expression Systems Analysis (N=22 modules)
(2) Integrated Genetic and Other Interaction Breast Tissue Network (GIANT)
(3) Kaplan-Meier Gene Dose-dependent Survival
(4) Differential Expression Across BrCa Subtypes
(5) ROC Analysis, Survival Prediction
3x
Anticorrelate Gene Ratios
W + X
Y + Z
Breast Interaction Network-
Implicated Genes
YBX1
SMCO4
PSAT1
USB1
MTHFD1L
RARRES1
SLC43A3
ACE2
PRSS33
HAPLN3
MSL3
NKX1-2
High Expression
High Survival
Low Expression
High Survival
HR = 1.56 (1.13 – 2.15)
logrank P = 0.0068
Probability
Expression
low
high
Time (months)
DEGs
Sensitivity
Specificity
Cancer Processes
Apoptosis (3)
BIK
SPTLC2
CERK
Sphingolipid Metabolism (3)
Angiogenesis (2)
CXCR4
GPRC5C
Metastasis (4)
CEBPB
Proteostasis (2)
TRIM3
DUSP4

Figure 1. Overview of analysis workflow for this study

As described in results section 1. See also Figures S4 and Tables S1–S3.

2008). After cleanup, batch correction with normalization, and outlier removal, the BrCa transcriptome containing 31,338 gene products across 773 nonoutlier BrCa (including 92 TNBC) tumor samples was clustered into coexpressed communities (modules) of gene transcripts using WGCNA with a 1-minus topological overlap matrix metric based on pairwise gene transcript correlations in an adjacency matrix calculated with biweight midcorrelation (bicor) for robust correlation, as described in methods. Bicor is a built-in correlation feature of WGCNA based on median provided as an alternative to Pearson correlation which is based on mean. Bicor was used, as opposed to Pearson correlation, to provide robust correlations with less weight given to outlier measures (Oldham et al., 2008; Langfelder and Horvath, 2012). A total of 22 network coexpression modules and corresponding quantitative module eigengenes (MEs) (for modules numbered by their size rank from largest to smallest) were identified: M1 to M22. The $\log_2$ relative FPKM/central tendency weighted first principal component equivalent to each of the 22 MEs is provided in Table S4. Table S5 provides the complete list of gene transcripts referenced by module membership and Pearson correlation to each of the 22 MEs (kME), where these correlations correspond to their weighted contribution to the ME in which they fall.

The communities clustered and indicated with module colors below the dendrogram also displayed consistency among individual transcripts that correlate within each module positively (red in gene-level heatmap) or negatively (blue) with factors influencing BrCa diagnosis (Figure 2A). However, modules are considered as a weighted average of gene members in the ME calculation, and each ME (or simply, the module each summarizes) can be correlated to available traits to select modules with significant trait correlations of

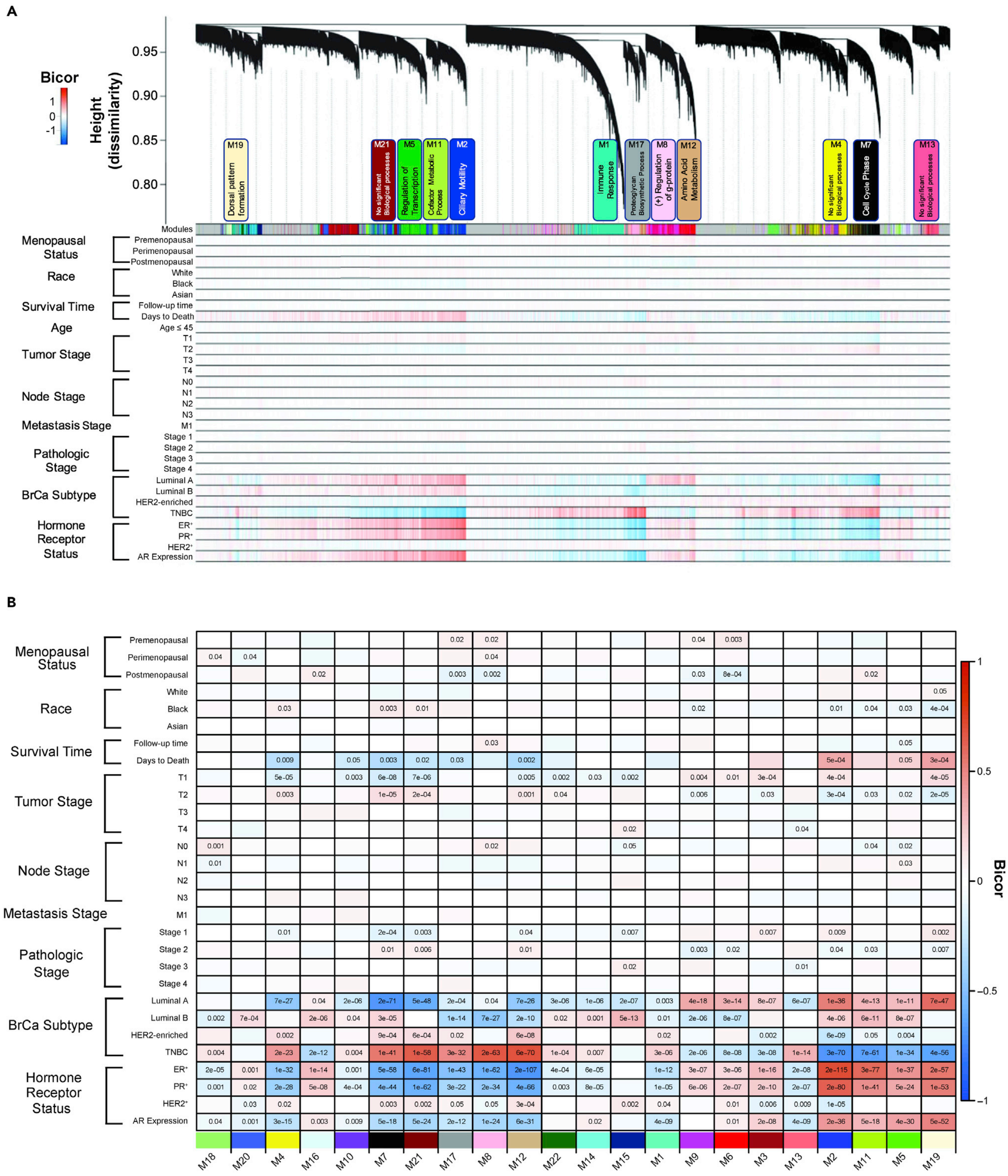


Figure 2. Coexpression network analysis of BrCa transcripts identified modules enriched with markers of biological processes and module-trait correlations reveal module communities relevant to BrCa subtype

(A) In the first row below the dendrogram, each colored vertical streak represents a gene with membership in the module of that color, which contains a group of highly coexpressed transcripts. A total of 22 modules were identified. A Kruskal-Wallis test among the four subtype-specific case groups (Luminal A,

Figure 2. *Continued*

Luminal B, HER2-enriched, and TNBC) with significance set to 0.05, and/or TNBC binary trait bicor revealed 12 modules (M12, M21, M8, M17, M7, M2, M11, M5, M19, M4, M13, and M1) were significantly associated with the TNBC subtype.

(B) Heatmap of module trait relationships. Overlaid numbers in panel B are Student's p values for bicor significance of trait correlation to the module eigengenes. Module-Trait bicor color scale (−1, blue; 0, white; +1 red) indicates modules with significant Student's p cluster together, in particular into M2-like and M12-like clusters. See also Data S1 and S2. See also Tables S4 and S5.

interest, alleviating multiple testing greatly. We chose to focus on modules with significant correlation to the TNBC subtype binary trait to identify gene communities relevant to TNBC biology. First, to identify subtype-associated modules, Kruskal-Wallis ANOVA $p < 0.05$ indicating significant overall separation among the four BrCa subtypes (Luminal A, Luminal B, HER2-enriched, and TNBC) revealed 12 modules (M4, M7, M21, M17, M8, M12, M1, M13, M2, M11, M5, M19) (Figure 2B). The box plots for these modules display the relative expression within each module. With the exception of M1, M4, and M17, the eigengene values for TNBC were qualitatively different from the other receptor-positive subtypes (Figure 3). In addition, a Wilcoxon rank-sum test indicated significant difference between TNBC and non-TNBC (Luminal A, Luminal B, and HER2-enriched) cases. Importantly, the intersection of this twelve-module list with the list of MEs significantly correlated to the TNBC subtype binary trait was complete, with significance of correlation ranging from $3.0\text{x}10^{-6}$ (M1) to $3.0\text{x}10^{-70}$ (M2) (Figure 2B). These modules were biologically coherent, over-representing the ontologies listed in Figure 2A. Gene ontology enrichment analysis was performed on the aforementioned modules correlated to the TNBC subtype trait via GO-Elite (Zambon et al., 2012), identifying coherent biology of coexpressed transcript modules. A list of the top five biological processes with false discovery rate (FDR)-adjusted p values for each of these modules is in Table 1.

Box plots for the top positively correlated module to TNBC (M12, bicor = 0.69 and $p = 6.0\text{x}10^{-70}$; amino acid metabolism) and the top negatively correlated module (M2, bicor = −0.69 and $p = 3.0\text{x}10^{-70}$; ciliary motility) in Figure 3 show starkly higher and lower expression in TNBC relative to receptor-positive subtypes in these two opposing modules, respectively. These two modules, as represented by their top hub kME values for each other, indicate that M2 and M12 are also the most strongly anticorrelated to each other (Table S5, red versus green scale for positive versus negative Pearson correlations of hubs in M2, positive for kME_{blue} and negative for kME_{tan}, and in M12, *vice versa*).

Finally, module trait relationships were assessed for confounding variables by regressing out variables, for example, age and race, before recalculation of the eigengenes (Data S2). Significant correlation remained between the 12 modules and TNBC after regression of selected traits. In previous BrCa racial disparity studies, major gene expression differences have been observed between Black and white patients, but after adjusting for proportional differences in molecular subtypes, such differences were significantly reduced or nullified (Liu et al., 2018). There is indeed a lack of race, menopause (which we did not regress), and age correlation of any significance in Data S2, page 3 and particularly page 4. This supports a conclusion that subtype differences are not confounded by differences in these traits. Analysis was performed on the unregressed data set because these traits were particularly not well correlated before regression, Figures 2B and S2, page 4. Notably, we chose to focus on BrCa subtypes independent of stage because subtype was a major driver of the network structure, as indicated by the large proportion of modules with significant correlation to the binary trait (TNBC/non-TNBC). Late-stage TNBC was underrepresented in our cohort, which reduced the statistical power to determine correlation.

Module relatedness, correlation to molecular BrCa traits, and known indicators of TNBC

Module relatedness was determined via clustering based on correlation distance metric (Figure 4A). Three clusters were identified. The first cluster at the left contained five modules: M18, M20, M4, M16, and M10. The second cluster included five modules: M7, M21, M17, M8, and M12, and the third cluster contained nine modules: M1, M9, M6, M3, M13, M2, M11, M5, and M19; of which, the last 4 were highly similar in their trait correlations and relatedness to each other. The heatmap in Figure 4B displays the total 22 module communities in module-related order. All available quantitative and binary-coded clinical traits were correlated to MEs pairwise for samples which had the traits specified. This greatly reduces the multiple testing problem from 31,338 individual genes to 22 modules and presents modules that represent transcript network structure, with some modules strongly correlated to traits of the BrCa cases, making them candidates for harboring key drivers of molecular causality, either upstream or downstream of the highly correlated traits of interest such as subtype (Luminal A, Luminal B, HER2-enriched, TNBC), hormone-receptor-reported

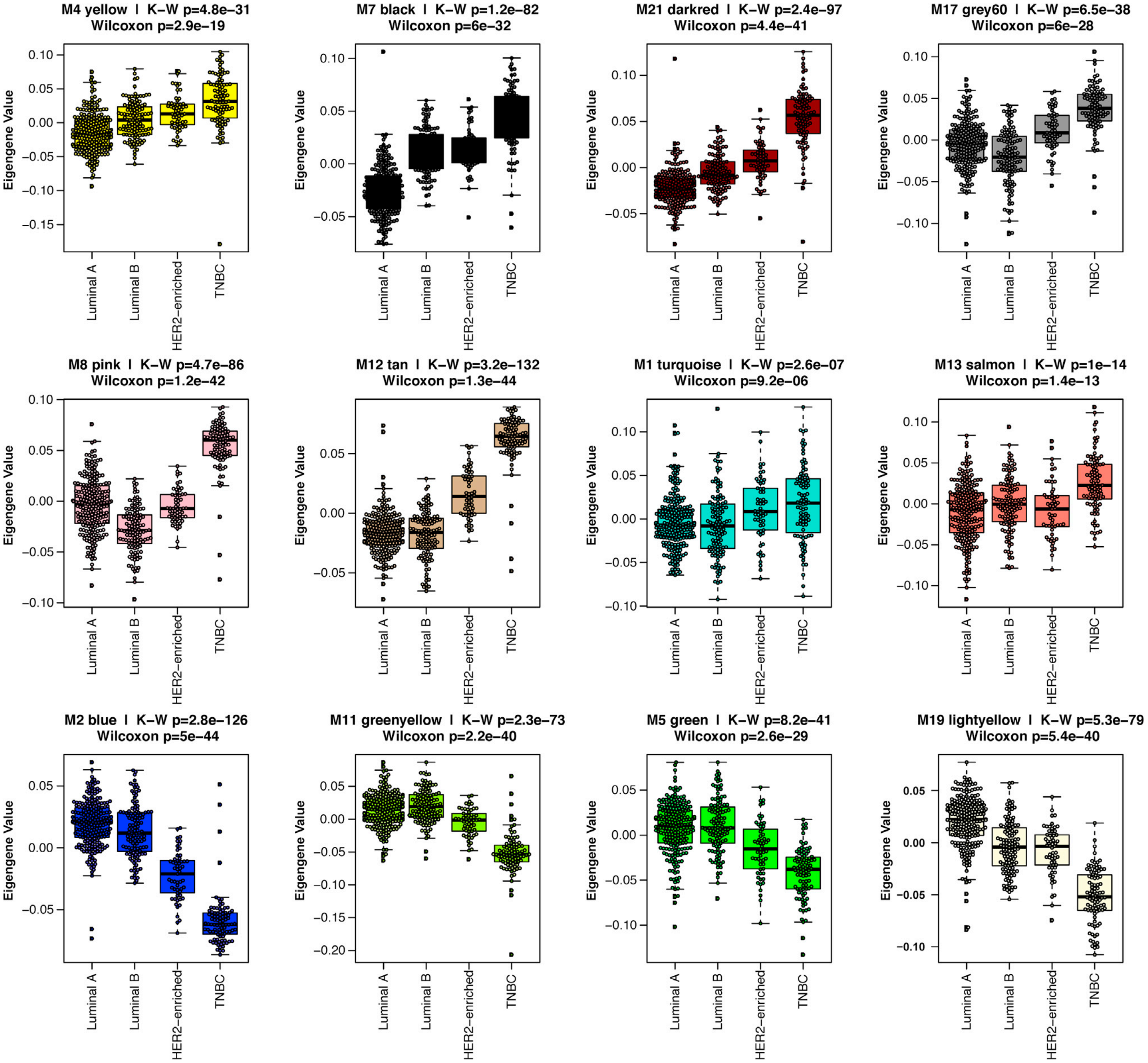


Figure 3. Differences in eigengene values for M2-and M12 -like modules across 4 BrCa subtypes

Box plots displaying module eigengene values among the four BrCa subtypes are displayed with ANOVA Kruskal-Wallis values $p < 0.05$. A Wilcoxon rank-sum test assessed significant difference between TNBC and non-TNBC (Luminal A, Luminal B, and HER2-enriched) cases. Significance was also set to $p < 0.05$. For box plots, N = 493. See also Data S1 and S2.

status (ER$^+$, PR$^+$, HER2-enriched, AR levels), and survival time (days to death for individuals succumbing to BrCa) (Figure 2B and Data S1, page 4).

We used module correlation (bicor) to a TNBC binary sample trait, to assess subtype differences. Bicor ranging from −1 to +1 represents eigengene (composite weighted within-module gene) expression correlation independently to each of the traits, but separate correlations are made to molecular traits, such as PR$^+$ and ER$^+$ expression, on a sample-by-sample basis. The heatmap in Data S1, page 4 displays p value significance for bicor between the 22 modules and the final BrCa factors that had strong correlation to days to death, BrCa subtype, and hormone receptor status. Our goal was to identify modules that both positively and negatively correlate to TNBC binary status and patient survival so we could in turn identify the relative expression pattern or phenotype of TNBC and non-TNBC. These anticorrelated genes were

Table 1. Ontology enrichment of modules with correlation to TNBC versus receptor positive BrCa status

Module and *subtype ANOVA p*	Top biological processes	GO-Elite FET FDR	
Yellow (M4) p = 4.8E-31	No significant biological processes associated	NA	
Black (M7) p = 1.2E-82	Cell cycle phase	7.80E-87	M12-like
	Cell cycle	2.84E-93	
	DNA metabolic process	1.04E-49	
	Cell division	7.12E-41	
	Regulation of cell cycle process	5.99E-34	
Darkred (M21) p = 2.4E-97	No significant biological processes associated	NA	
Grey60 (M17) p = 6.5E-38	Proteoglycan biosynthetic process	5.53E-02	
	Actin filament-based process	6.01E-04	
	Cell junction organization	4.17E-02	
	Skeletal system development	4.42E-02	
	Cellular component movement	1.88E-02	
Pink (M8) p = 4.7E-86	Positive regulation of g-protein	1.98E-04	
	Hemidesmosome assembly	7.24E-04	
	Tissue development	1.35E-07	
	Positive regulation of neuroblast proliferation	7.74E-03	
	Positive regulation of endothelial cell migration	2.92E-03	
Tan (M12) p = 3.2E-132	Cellular modified amino acid metabolic process	2.96E-02	
Turquoise (M1) p = 2.6E-07	Immune Response	3.22E-161	
	Regulation of immune response	6.95E-91	
	Positive regulation of immune response	5.12E-73	
	Defense response	2.67E-73	
	Response of lymphocyte activation	4.93E-53	
Salmon (M13) p = 1.0E-14	No significant biological processes associated	NA	
Blue (M2) p = 2.8E-126	Ciliary or flagellar motility	3.24E-02	M2-like
	GPI anchor metabolic process	3.76E-01	
	Phototransduction	7.77E-01	
	Positive regulation of glucose import	8.41E-01	
	Branched chain family amino acid metabolic process	9.33E-01	
Greenyellow (M11) p = 2.3E-73	Cofactor metabolic process	1.95E-02	
	Oxidation-reduction process	4.67E-02	
	Lipid metabolic process	9.73E-02	
	Thioester metabolic process	7.81E-01	
	Cofactor metabolic process	1.95E-02	
Green (M5) p = 8.2E-41	Regulation of transcription, DNA-dependent	4.48E-16	
	Androgen receptor signaling pathway	3.10E-03	
	Protein modification by small protein	1.55E-04	
	conjugation or removal	3.96E-04	
	Organelle organization	3.03E-02	
	Nuclear-transcribed mRNA poly(A) tail shortening		
Lightyellow (M19) p = 5.3E-79	Dorsal/ventral pattern formation	1.69E-03	
	Forebrain development	1.53E-01	
	Epithelial tube morphogenesis	3.21E-01	
	Specification of symmetry	3.21E-01	
	Regulation of embryonic development	4.00E-01	

A one-tailed Fisher's exact test with FDR correction detected significant overlap between GO lists of gene symbols and members of each module.

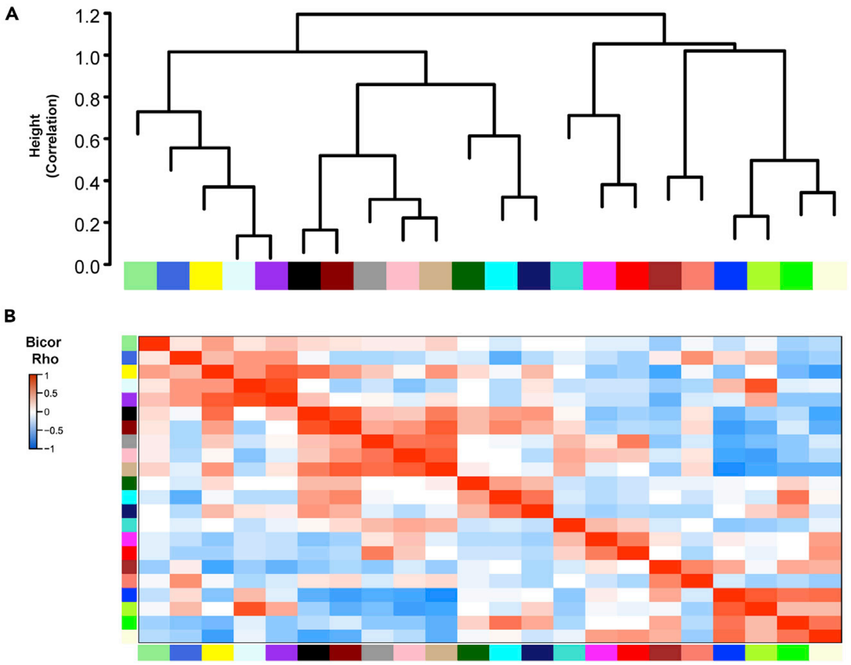


Figure 4. Module relatedness clustering

(A) Clustering dendrogram based on ME correlation.

(B) Heatmap of 22 module communities organized by module related order.

used to predict survival of patients during later analyses. The block of 5 modules on the left (M7, M21, M17, M8, M1) significantly and positively correlated to TNBC binary status and they are positively correlated to each other as related patterns across all cases in the network. This block of 5 modules also negatively correlated to survival time suggesting there are likely key drivers in the module that could be influencing TNBC survival. The block of 4 modules on the right (M2, M11, M5, M19) negatively correlated to TNBC and positively correlated to days to death suggesting these genes are also likely key drivers in the module that could be influencing TNBC survival.

Ultimately, we chose to focus on the M12 and M2 modules because they were prototypic, (anti-) correlated modules among seven other module communities (M7, M21, M17, M8, M11, M5, M19) that significantly correlated to TNBC binary status and patient survival time.

Four non-TNBC modules (M2, M11, M5, and M19) negatively correlated with the TNBC subtype trait with significance of correlation $p < 1\text{x}10^{-25}$, and five TNBC modules (M7, M21, M17, M8, and M12) were just as significantly positively correlated. Three modules (M1, M4, and M13) were not closely related and were not included for further analysis. Secondary analysis could focus on these three unrelated modules. Notably, M7, M21, M17, M8, and M12 (higher expression specific to TNBC) and M2, M11, M5, and M19 (lower in TNBC) moderate-to-high significance of correlation was consistent across all four subtype-specific binary traits, and correlation itself was along a negative-to-positive (or *vice versa*) continuum across the four subtypes ordered by increasing severe/poor prognosis (Data S1, page 4). The trait for days to death also significantly correlated in opposing directions to M12 and M2, respectively (Data S1, pages 16 and 25). Hormone and HER2 receptor status across all BrCa subtypes also were the strongest and most significantly correlated to M2 and M12, in opposing directions (Data S1, page 4). A similar heatmap of all 22 modules correlated to the full set of available quantitative or binary traits can be found in Figure 2B.

Overall, these module correlations to known molecular determinants of BrCa severity and survival implicate these twelve modules in mechanisms of BrCa that may not be fully appreciated. Fully consistent with quantitation of the key BrCa receptors correlating positively to M2, notable members of M2 include its hub, one of two classic nuclear hormone receptors for estrogen (ESR1 gene for ERα, $\text{kME}_{\text{blue}} = 0.83$) and the nuclear hormone receptor for progesterone ($\text{kME}_{\text{blue}} = 0.70$). HER2 receptor (ERBB2) was not assigned to any

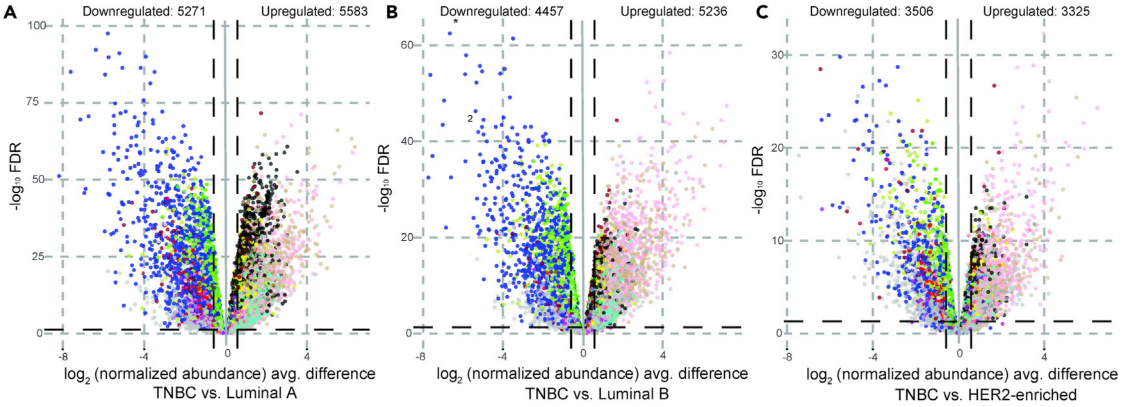


Figure 5. Volcano plots of differentially expressed genes (DEGs) in TNBC

TNBC case group was compared with receptor-positive case groups Luminal A (A), Luminal B (B), and HER2-enriched BrCa (C). Log_2 (fold change) for each comparison versus Benjamini-Hochberg FDR is plotted, and gene transcripts are colored by module membership (no module: light gray). DEGs were counted for >50% change (vertical cutoff lines at $x = \pm 0.58$) and FDR<5% in each individual comparison. See also Tables S6–S8.

module, with no absolute value of kME greater than 0.30, though it was weakly correlated best with a kME to M2 of 0.27, and the related receptor ERBB4 has blue (M2) membership with kME 0.70. Therefore, M2 and its strongest anticorrelate M12 are the top module candidates harboring genes with potential as baseline-measured prognostic if not also mechanistic indicators for TNBC. Thus, we termed M2, M11, M5, and M19, M2-like and M12, M7, M21, M17, and M8, M12-like.

Survival analysis using M2- and M12-like differentially expressed genes

An unpaired two-tailed t test followed by Benjamini-Hochberg FDR estimation was conducted to compare differential expression among TNBC tumor subtype cases (N = 92) and the three receptor-positive subtypes: Luminal A (N = 226), Luminal B (N = 118), and HER2-enriched (N = 57). Statistical significance for counting differentially expressed genes was set to FDR<0.05. Volcano plots in Figure 5 report the number of genes upregulated and downregulated as well as differentially expressed between TNBC and each of the non-TNBC (Luminal A, Luminal B, and HER2-enriched) tumor groups. A complete list of differentially expressed genes (DEGs) ($p < 0.0001$) between TNBC and each non-TNBC tumor group can be found in Table S6, filtered for consistency of directional change and significance in all three pairwise comparisons of non-TNBC cases to the TNBC group to further control false positives (N = 1,518). The top 20 downregulated and upregulated genes for each of the four M2-like and five M12-like modules linked to TNBC are given in Tables S7 and S8. Table S7 has rank based on FDR for the Luminal A *versus* TNBC comparison, and Table S8 shows rank based on the comparison FDR for HER2-enriched *versus* TNBC subtypes.

After differential expression analysis, KM survival analysis (Gyorffy et al., 2010) nominated high- and low-expressed gene subsets that significantly distinguished patients dichotomized by survival propensity in the selected two groups of opposing module communities, namely, the aggressive TNBC modules (M7, M21, M17, M8, M12) and the non-TNBC modules (M2, M11, M5, and M19). Nine TNBC module MEs were analyzed to test the association of low *versus* high eigengene value with progression free interval (PFI) time (Figure 6), which is analogous to relapse-free survival (RFS) (Liu et al., 2018), in all BrCa subtypes of the TGCA RNA-seq data. A difference in PFI, higher survival with lower expression, was seen in all M12-like modules and a difference in PFI, higher survival with higher expression, was seen in all M2-like modules. However, M8 and M11 exhibited an association trend in the opposing direction respectively different from trends for the other M12-and M2-like modules. Although we performed the analysis across all 773 individuals, not just the 92 TNBC cases, this result suggests that gene coexpression in these two eigengenes is compensatory and not exacerbating TNBC poor prognosis. Hazard ratios (HRs) with 95% confidence intervals (CIs) confirm four of the five M12-like modules have relative differences in expression of coexpressed genes that coincides with the significant risk for the patients to relapse. HRs confirm three of the four M2-like modules have relative decreases in expression of coexpressed genes coinciding with significantly lower risk of relapse.

Following eigengene survival association, the top 20 TNBC DEGs in the M12 (upregulated) and M2 (downregulated) modules as ranked by log_2-fold change in Table 2 were also analyzed by KM analysis to test the

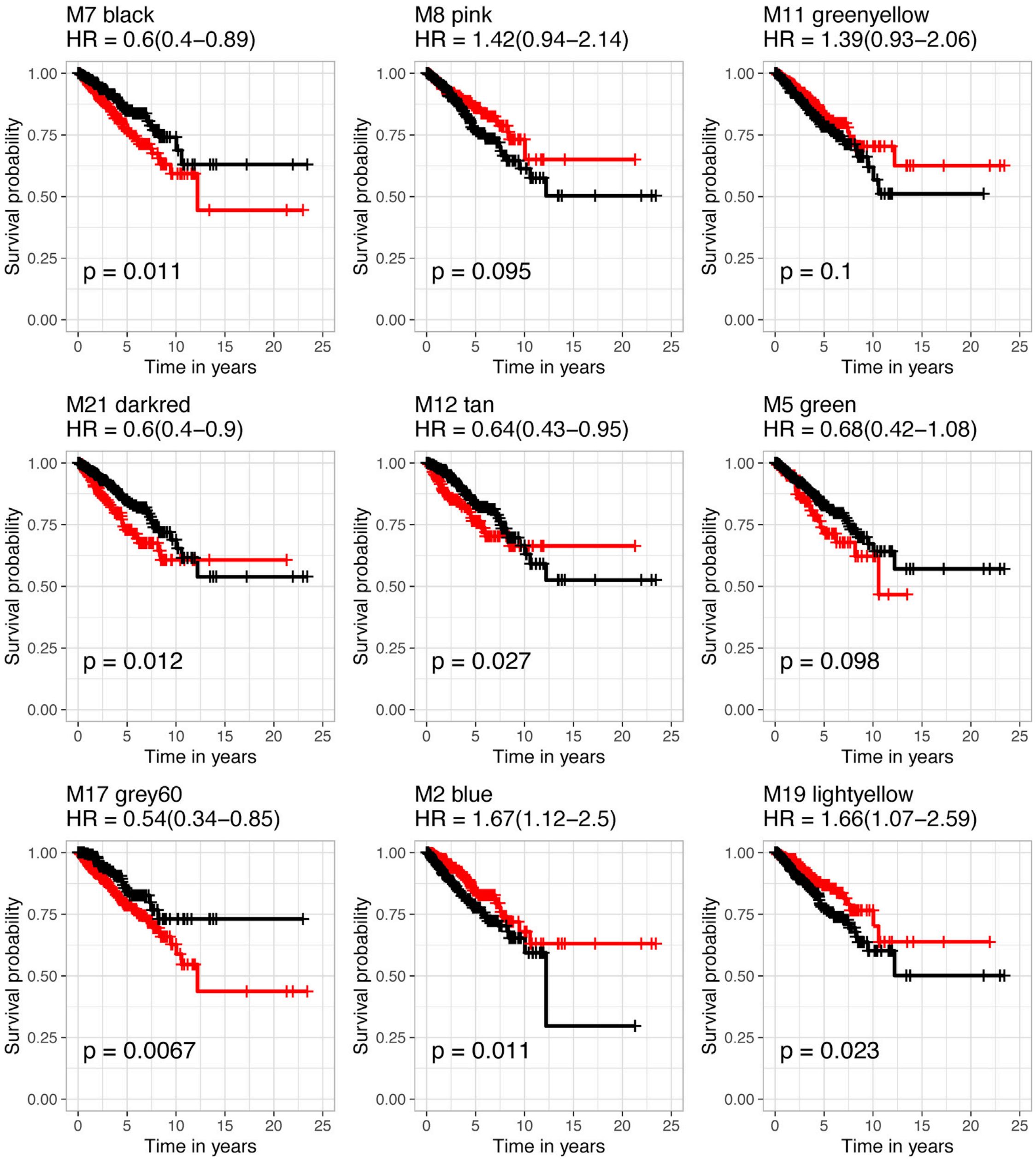


Figure 6. Survival analysis reveals association between M12-and M2-like module eigengene values and progression free interval time PFI in breast cancer

Hazard ratios confirmed low expression of M12-like eigengene values, except M8, resulted in a significantly high likelihood of relapse, while high expression of M2-like modules, except M11, resulted in significantly less likelihood of relapse (left). The red line in each Kaplan-Meier plot represents survival of cases in the higher expression tier and the black line represents survival in cases with lower expression. A log rank (Mantel-Cox) test determined p values and hazard ratio (HR) scores. See also Figure S1 and Table 2.

association of low *versus* high single-gene expression with RFS in all BrCa subtypes of an independently curated RNA-seq meta-analysis (Gyorffy et al., 2010). A difference in RFS, higher survival with lower expression, was seen in the M12-specific DEG hub gene *PSAT1* (p = 0.0042; HR = 1.87, CI = 1.21–2.89). A hub gene is a gene with a connectivity score kME = 0.6 or higher. *PSAT1* (kME 0.81) is the number 1 hub gene in the M12 module. In the opposing direction—higher survival with high expression—of M2 genes *TFF1* (p = 0.011; HR = 0.49, CI = 0.28–2.86) and *SCUBE2* (p = 0.00045; HR = 0.47, CI = 0.3–0.72) (Figure S1). *TFF1* is the top DEG in the M2 module with 256-fold higher expression in TNBC than HER2-enriched BrCa. *SCUBE2*

Table 2. Relapse-free survival analysis using BrCa RNA-Seq (N = 1,090) for Kaplan-Meier plots

DEGs TNBC versus Luminal A[b]							DEGs TNBC versus Luminal A[b]						
M2							M12						
				Kaplan-Meier							Kaplan-Meier		
Gene	WGCNA kMEblue	FDR	Log$_2$ diff	Log rank P	Direction[a]	Selected	Gene	WGCNA kMEtan	FDR	Log$_2$ diff	Log rank P	Direction[a]	Selected
TFF1	0.55	7.13E-52	-8.18	0.011	+	Yes	HORMAD1	0.54	2.46E-61	6.35	0.4	ns	
AGR3	0.61	9.16E-86	−7.61	0.0056	+	no probe	**ART3**	0.67	4.74E-64	6.19	**0.0035**	–	Yes
TFF3	0.56	2.87E-70	−7.14	0.034	+		LINC01956	0.64	5.79E-55	5.59	NA	NA	
SRARP	0.54	1.11E-47	−6.88	NA	NA		FOXCUT	0.78	9.44E-67	5.50	NA	NA	
AGR2	0.60	2.63E-71	−6.73	0.014	+	Yes	GABBR2	0.63	4.14E-42	5.23	0.069	ns	
ESR1[c]	0.83	6.05E-93	−6.38	0.01	+	**	ZIC1	0.52	2.77E-37	4.99	0.097	ns	
GP2	0.48	6.38E-34	−5.94	0.0061	+	Yes	NKX1-2	0.75	1.06E-38	4.88	NA	NA	
CT62	0.69	6.94E-85	−5.92	0.0062	+	no probe	PRSS33	0.57	2.54E-23	4.68	**0.0035**	–	no probe
LINC00504	0.74	3.06E-98	−5.79	NA	NA		ENSG00000 179066	0.60	5.26E-55	4.38	NA	NA	
FOXA1	0.47	1.59E-90	−5.75	0.091	ns		ENSG00000 248538	0.63	3.82E-35	4.38	NA	NA	
POTEKP	0.66	8.69E-46	−5.70	NA	NA		GFRA3	0.52	1.33E-44	4.31	0.22	ns	
ENSG00000240800	0.53	4.35E-33	−5.62	NA	NA		NDUFB4P11	0.70	1.12E-40	4.24	NA	NA	
ABCC8	0.71	4.47E-49	−5.51	0.00056	+	Yes	SLC26A9	0.65	6.60E-39	4.12	0.061	ns	
ENSG00000235584	0.62	2.03E-33	−5.50	NA	NA		**FZD9**	0.64	5.05E-61	4.11	**0.0027**	–	Yes
TNRC18P1	0.67	3.82E-59	−5.50	NA	NA		**PSAT1**	0.81	2.39E-59	4.09	**0.0042**	–	Yes
SLC44A4	0.40	1.92E-71	−5.50	0.017	+		CASC8	0.64	4.83E-49	4.09	NA	NA	
ERBB4	0.70	2.00E-68	−5.47	0.008	+ (<5 yrs)	Yes	LINC01198	0.68	4.54E-28	4.09	0.29	ns	
SCUBE2	0.75	1.09E-65	−5.47	0.00045	+	Yes	**OPRK1**[d]	0.60	1.45E-27	3.95	0.35	ns	Yes***
TTC6	0.62	1.63E-75	−5.43	4.30E-05	+	no probe	ABCA13	0.57	3.08E-29	3.82	0.3	ns	
LINC02568	0.65	3.90E-61	−5.28	NA	NA		**MARCO**	0.64	8.72E-34	3.81	**0.011**	–	Yes
							YBX1	0.73	3.18E-53	1.42	0.12	ns	
Total M2 numerator genes (bold): 6							Total M12 Denominator Genes (bold): 5						

KMplot.com-curated data for BrCa RNA-seq (Gyorffy et al., 2010) was used to perform survival analysis of gene transcript-dose effects on RFS in BrCa (all subtypes). Genes were nominated by Luminal A versus TNBC DEG FDR rankings (Table S7) of 20 or less with M2 (left) or M12 (right) membership. Genes with significant dose effect (log rank p value < 0.015) were nominated for inclusion in the subsequent first round of ROC analysis. See also Figures 6 and S1.

[a] Minus, low expression, high survival; plus, high expression, high survival.

[b] Consistent and significant for all 3 TNBC pairwise comparisons.

[c] Estrogen receptor excluded as known driver.

[d] OPRK1 K-M log rank p = 0.00056 for array data (–).

(kME = 0.75) is the top differentially expressed hub gene in the M2 module. As expected, expression of *PSAT1* was highest in the TNBC subtype sample group, and expression of *TFF1* and *SCUBE2* was lowest for the TNBC group, each reaching a two-group Wilcoxon rank-sum test $p < 2.2 \times 10^{-16}$. Based on their association with survival, we selected 11 significant gene transcripts (6 from M2 and 5 from M12) as a seed list of potential high-performing BrCa prognostic signature genes (see the following text). Of the 20 M2-like DEGs, 10 genes have HRs that confirm patients are significantly less likely to relapse, whereas five of 21 M12-like DEGS have HRs that confirm patients are significantly more likely to relapse.

Interaction analysis of M2-and M12-like modules confirms driver nodes

The coexpression networks of top ranked genes (kME = 0.6 or higher) for the M12-and M2-like modules were analyzed for mammary-gland-specific genetic and other interactions, which were assessed using the GIANT reference database (Wong et al., 2018). Other interactions between survival-linked genes for the top anticorrelated modules, M12 (*PSAT1* and *PRSS33*) and M2 (*TFF1* and *SCUBE2*) were also assessed. Target pair interaction scores for all M12- and M2-like top ranked hub genes were obtained via an adapted interface to the GIANT v2 database (Greene et al., 2015) and are reported in Table S9.

Modules with higher connectivity breast-specific mammary interactions (M7, M21, M17, M8, M12, and M2) are displayed in Figure 7. The M12-like module, M7, with significant associations to biological processes such as the cell cycle, was the only M12-like module that had strong interactions between all ten M8 hub genes and GIANT breast-specific mammary implicated genes, including a validated BrCa biomarker, *BRCA1*. In addition, *EGFR* is higher expressed in TNBC versus non-TNBC cases and is a hub of M8 as indicated in both the kME table (Table S5, kMEpink = +0.841, ranked fourth in the module) and Figure 7. However, in Figure 6, dichotomization of the M8 eigengene into low- and high-expressing cases and comparison of progression-free interval as a proxy for survival in these two groups does not reach significance in the data analyzed. Overall, our results support that *EGFR* mRNA is elevated in TNBC, which is congruent with literature (Nielsen et al., 2004), but in the data examined as a whole across the 4 different BrCa subtypes, *EGFR* dichotomized high and low relative expression does not separate low from high progression-free interval or in summary no significant association between *EGFR* and our survival outcome, progression-free survival.

Three hub genes of M12, *PSAT1* (kME = 0.81), YBX1 (kME = 0.72), and *MTHFD1L* (kME = 0.72), are strongly interconnected in the M12 hub-derived network of breast tissue (Figure 7). *PSAT1* has the highest M12 kME and has significant mammary specific interactions, *YBX1* has the strongest edges in the network, and *MTHFD1L* had fewer strong interactions. Moreover, *PRSS33*, a nonhub M12 member (kME = 0.57) and significantly associated with BrCa survival, did not contribute to the nomination by GIANT of mammary network-implicated genes, having no significant mammary specific interactions among hubs of M12 or their most connected genetic interactors.

M2 was the module with significant breast-specific mammary gland interactions. Modules with lower connectivity, breast-specific mammary gland interactions are displayed in Figure S2. The M2 hub-derived breast interaction network is highly regulated by one gene, *ESR1* (Figure 7, bottom right panel), with kME = 0.82. *ESR1* has significant mammary-specific interactions with 13 genes including most notably, *BRCA1*, a validated risk gene for BrCa. Moreover, while the nonhub *TFF1* (kME = 0.54) and *SCUBE2* (kME = 0.75 but rank 34 among M2 hubs) integrated from the differentially expressed genes have a significant association of high expression with high RFS, neither was among those strongly connected within the tissue-specific hub-defined network. Interestingly, repeating GIANT with the same M2 hubs minus estrogen receptor (*ESR1*) implicated an entirely different set of interactors among the remaining hubs, with more balanced connectivity (*data not shown*). Notably, the M2 hub-derived GIANT network, *CEBPB* had high M2 hub connectivity but belonged to M12, thereby bridging the two anticorrelated modules. Such genes, known as bottleneck nodes, constitute an important conduit for the interchange of information between the different gene modules which they bridge, enriched as a class for genes essential for survival (Yu et al., 2007), and more often successfully targeted by drugs (Yao and Rzhetsky, 2008), elevating the importance of inclusion of this gene in subsequent analysis.

ROC of survival prediction for integration of marker candidates, testing, and validation of a prognostic TNBC-specific biomarker ratio panel

These nominated transcripts based on RNA-seq, WGCNA, and KM analyses were assembled into prognostic indicator ratios, tested, and optimized by rounds of ROC analysis in an independent 1,234 array

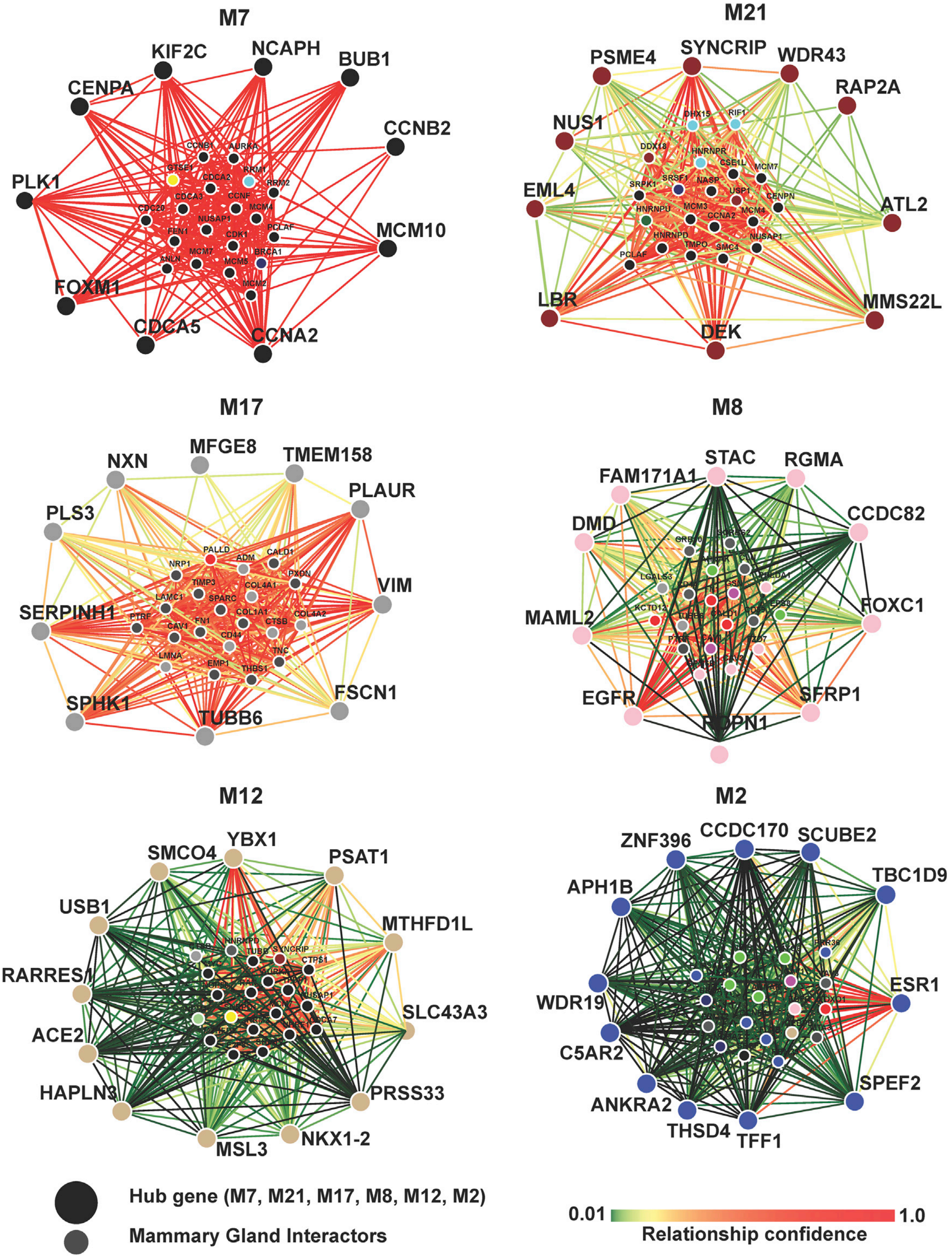
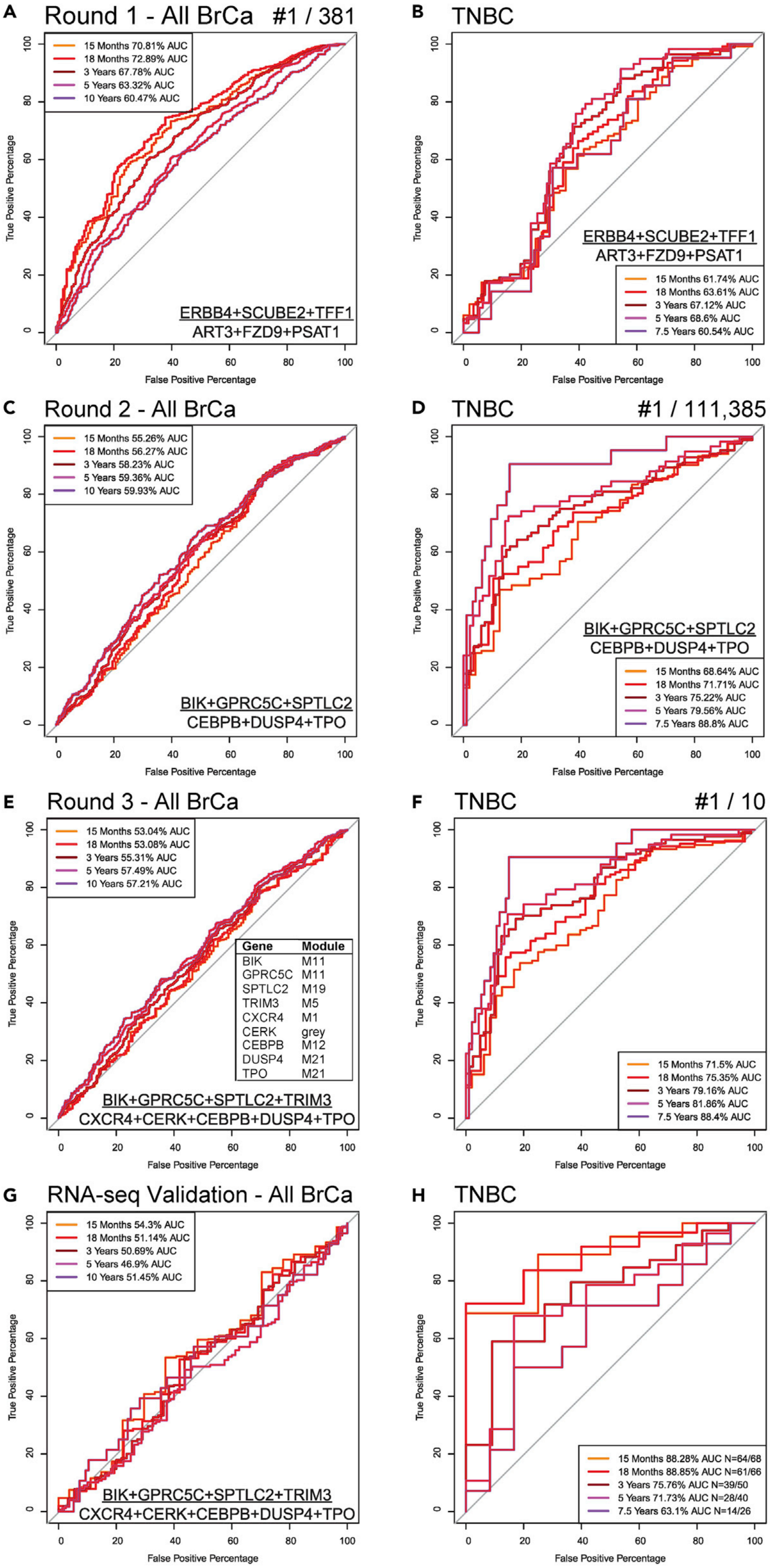


Figure 7. Genome-scale integrated analysis of gene networks in tissues (GIANT) identifies higher connectivity breast-tissue-specific interactions between survival-linked genes and module hub-implicated genes as well as hubs in M12-like and M2-like modules

The heatmap scale of edge (interaction) colors represents the confidence of the predicted interactions based on mammary-gland-specific interactions. High confidence interactions are represented by red edges (see scale). Input for GIANT was restricted to the top ten hub genes for each module, plus nonhub survival-linked genes such as *PRSS33* (M12). Small nodes in the center of each network are implicated by the GIANT database network as the best connected to the hubs specified in mammary tissue, colored by their module membership. See also Figure S2 and Table S9.



Gene	Module
BIK	M11
GPRC5C	M11
SPTLC2	M19
TRIM3	M5
CXCR4	M1
CERK	grey
CEBPB	M12
DUSP4	M21
TPO	M21

Figure 8. Top-performing survival predictors by ROC AUC

ROC curves predicting RFS were generated on combinatorial ratios of M2(-like)/M12(-like) nominated genes, first using microarray-measured normalized transcript abundances in N = 1,234 BrCa cases with RFS traits. In round 1, transcripts

Figure 8. *Continued*

consisting mainly of M2 and M12 hubs nominated by Kaplan-Meier distinction of RFS (Table 2) via their levels in independent RNA-Seq were tested in 381 combinatorial ratios for RFS prediction using all microarray BrCa cases (top performer shown in A), and only TNBC cases for the same normalized transcript abundance ratio (B). Round 2 added nominations of transcripts from M2-like and M12-like DEGs, as well as *CEBPB* and other genes from the breast tissue gene interaction network, producing the top ranked ROC curve with the best AUC(s) obtained using only TNBC microarrays, shown first for all BrCa (C), or TNBC only (D). Round 3 ROC tested whether additions of known BrCa genes in pathway(s) implicated by round 2 top predictor genes might further improve AUC, and the top performer in TNBC was plotted for all BrCa (E) and TNBC specific cases (F). Finally, the same ratio was calculated using the 773 RNA-Seq TCGA case abundances used for network building, normalized like the array data, and validated by ROC analysis for all BrCa (G) and TNBC (H). See also Figure S3 and Tables S10–S12.

meta-analysis of BrCa (Györffy and Schäfer, 2009), including 180 TNBC cases (Karn et al., 2011) as a data set (traits provided in Table S10). Currently, treatment data for these patients is not available, thus survival is irrespective of treatment. Follow-up duration for this cohort of patients was 10 years for all BrCa subtypes and 7.5 years specifically for patients with TNBC. Our goal was to use this independent aggregate data from an orthogonal measurement technique to test combinatorial gene expression ratios made up of candidate genes discovered as described in aforementioned results, based on the framework of combining M2-and M12-like genes of anticorrelated coexpression transcriptome network modules identified in our curated RNA-seq meta-analysis. We hypothesized that negative correlation of gene expression profiles with high correlation to disease traits including survival could drive discriminatory difference in a ratio of genes (a) with numerator genes having expression positively correlated with survival, divided by (b) denominator genes that synergize or act together negatively to impact mechanisms of survival.

The 11 DEGs highlighted with the most significant dose-effects on survival in KM plots (Table 2), available for testing in the array data, were nominated for a first round of combinatorial tests for their potential to contribute to a multigene equal-weight ratio for prediction of BrCa survival. M2 (or M2-like) module member relative abundance was divided by M12 (or M12-like) module member abundance within sample, so that opposing changes within a sample would amplify sensitivity for survival prediction, and specificity of the prediction across many samples could be tested by the ROC of survival prediction for that ratio, given known survival status of each patient during extended follow-up. If a TNBC sample expressed M2 transcripts with a role in survival at elevated levels, this might counter the general TNBC-associated direction of change and poor TNBC prognosis. In addition, if those transcripts are mechanistically involved in slowing cancer progression, they might improve survival across BrCa cases in general. Therefore, high relative expression of M2 genes should represent high survival odds, while low relative expression of M12 genes should represent higher survival odds in BrCa in general. Therefore, a larger ratio predicts better survival prognosis in this model. The complete set of 11 genes combined and 380 additional combinations covering all possible 1:1, 2:2, and 3:3 gene ratios were tested for prediction of RFS using AUC as the readout; all combinations were >50% AUC on average across calculations for 5 specific time points in BrCa (Table S11), supporting the use of M2/M12(-like) gene ratios for survival prediction. The best combination for the full cohort of 1,234 BrCa cases of mixed subtype (Figure 8A) had a maximum AUC of 72.9% for prediction of survival at 18 months, though Table S6 demonstrates that many combinations were similar. The top ratio was (*ERBB4* + *SCUBE2* + *TFF1*/*ART3* + *FZD9* + *PSAT1*), outperforming the naive ratio of all 11 transcripts by only 1.1%. Prediction of survival in the subset of 180 TNBC array cases was highest for this combination at 5 years, but only with a 68.6% AUC (Figures 8B), 4.7% higher than the 11-gene ratio. Interestingly, the 2-gene combination *ERBB4*/*FZD9* was the best 5-year TNBC survival predictor among the 381 combinations, with a 76.0% AUC. This first round of ROC tests suggests different combinations of genes play more of a role in TNBC specifically than in BrCa in general.

We expanded our test to consider transcript abundance ratios of genes implicated by genetic interactions, and M2-or M12-like module members are also highly ranked DEGs in TNBC, in particular compared with the next most severe BrCa subtype, HER2-enriched (Table S8). Of the 184 genes in the table, for the numerator we selected the top 14 M11 genes, the top 5 from M19 plus a lower ranked M19 DEG, *NCAM2*, as well as *ERBB4* and *SCUBE2* from ROC round 1. Two additional genes implicated by the interaction network of M2 hubs that were also DEGs (*AHNAK* of M5 and *VAV3* of M2) were included as well, representing 24 total candidate numerator genes from M2-like modules. For the round 2 ROC indicator ratio denominator candidates, we selected *CEBPB* of M12 implicated by GIANT as connected to the M2 hub network and *CEBPG* of M7 as a negative control not implicated by our integrated analysis, but still in an M12-like module. DEGs implicated by the M12 hub interaction network were *CDCA7*, *CTPS1*, and *MCM7*. *ART3*, *FZD9*, and *PSAT1* carried over from the first round, and the top 12 M8 DEGs plus the top 14 M21 DEGs in Table S8 rounded out a total of 34 candidates in the M12-like genes of the denominator.

We expected that the 111,385 complete combinations (devised again as 1:1, 2:2, and 3:3 gene ratios), would be enriched for higher AUCs in prediction of TNBC-specific survival compared to overall BrCa because HER2-enriched versus TNBC DEGs (Table S8) are included in addition to the top nominations from survival analysis (Table 2) and the tissue interaction networks (Figure 7). This was indeed the case (Table S12), with the average AUC of the top 10 combinations (all 3:3 ratios) at 79.0% for predicting survival at 5 years, compared with 70.8% in round 1. On the other hand, in round 2, the best all-subtype BrCa prognostication ranked by maximum overall BrCa predictor AUC at any time point but averaged for the top 10 at the 5-year time point was <64.0%, though better at earlier time points, not exceeding 73.1% for any combination at any time point. Moreover, the best performing TNBC predictor obtained by this relatively unbiased survey of candidates performed poorly for overall BrCa (AUC_{max} < 60.0%, Figure 8C), while exceeding 88% for TNBC 7.5-year survival and 79.6% at 5 years (Figure 8D). The top-ranked ratio consisted of markers distinct from round 1, despite 5 of 6 of the round 1 top performing ratio genes being tested: (*BIK* + *GPRC5C* + *SPTLC2*/*CEBPB* + *DUSP4* + *TPO*). These 6 genes were enriched among most of the top 100 performing combinations (Table S12), suggesting robustness of the approach.

To perform a third and final round of ROC tests, we decided to examine peer-reviewed literature for the aforementioned 6 genes and determine logical candidates to augment the ratio by their known involvement in both the biological pathways already implicated and with known roles in BrCa if not TNBC survival prognosis. We are thereby integrating verified biological knowledge into the largely unbiased process of gene nomination and testing, which improves prospects of reproducibility and general applicability of our biomarker ratio in principle. We also nominated *TRIM3*, found in 46 of the top 100 TNBC predictors, but only appearing in 22.6% of all round 2 combinations; it is found in place of *GPRC5C* as the only change in the second-best round 2 TNBC survival predictor, which outperforms 5-year survival prediction for TNBC of the top indicator at 83.3% AUC versus <80.0% (Table S12).

Sphingolipid pathway genes related to BrCa (implicated by *SPTLC2*) in the literature included long-chain ceramide synthase *CERS4*/*LASS4* (connected to round 1 candidate *SCUBE2* by opposing its role in SHH shedding (Gencer et al., 2017; Tsai et al., 2009), where this developmental morphogen being reexpressed in BrCa has been specifically implicated in its migration and invasiveness (Riaz et al., 2019). In addition, ceramide kinase *CERK* was implicated in BrCa metastatic migration (Schwalm et al., 2020). Chemokine receptor signaling was of interest, and we found *CXCR4* and *CXCR5* comentioned with *GPRC5C* in a review of vascular development dysregulation relevant to cancer (De Francesco et al., 2017). Angiogenesis is a requirement for larger tumor growth. We chose to test *CXCR4* because of its higher kME among CXCRs in the weakly M12-like M1, which was the only module harboring this class of genes including *CXCR5* (Table S5). Adding *CERS4*/*LASS4* to the numerator did not improve test ROC AUC (*data not shown*), but the remaining new genes (numerator: *TRIM3*; denominator: *CXCR4* and *CERK*) in addition to members of the round 2 top-performing ratio did raise average AUC for the 5 time points tested in the top TNBC predictor from 76.8% to 79.3%, indicating that literature-informed improvements to the predictor are possible. The final 9-gene prognostic indicator was ineffective for prediction of overall BrCa survival (Figure 8E), compared with TNBC (Figure 8F), for which it was highly specific. Significance of improvement in ROC curves was calculated for relevant pairwise comparisons, along with predictor significance of AUC and 95% CIs, and these are provided in Table 3. Citations across cancer literature were mined by gene symbol, submitting genes from the round 1 top performing ratio and all candidates for rounds 2 and 3 using OncoScore (Piazza et al., 2017) to determine known relevance to cancer, and we found that most of the genes were represented with a significant score (Figure S3).

To validate our round 3 ROC test of the 9-gene predictor, we calculated the 9-gene ratio in the TCGA RNA-seq quantified samples. Data were subjected to sample-wise normalization identical to the array normalization performed. ROC curves were similar to the ones produced from array-measured samples for both BrCa RFS prediction (Figure 8G) and for TNBC (Figure 8H), where this ratio was only effective for prediction of survival in TNBC. Thus, our ratio predictor of survival determined in array data from independent BrCa cases with gene nominations based on the network built on RNA-seq data validates in the RNA-seq data, indicating reproducibility and robustness of the predictor.

Table 3. ROC analysis Statistics for top performing transcript abundance ratios

Round	Survival predictor	15-month TNBC	18-month TNBC	3-year TNBC	5-year TNBC	7.5-year TNBC
1	ERBB4+NCAM2+SCUBE2ART3+FZD9+PSAT1	p=0.0023 CI 0.5399-0.7373	p=0.00026 CI 0.5681-0.7522	p=1.5e-05 CI 0.6042-0.7723	p=2.1e-05 CI 0.6158-0.7838	p=0.024 CI 0.5267-0.7499
2	BIK+GPRC5C+SPTLC2CEBPB+DUSP4+TPO	p=6.7e-05 CI 0.6025-0.7703	p=1.3e-06 CI 0.641-0.7932	p=1.1e-08 CI 0.6769-0.8275	p=6.9e-10 CI 0.7168-0.8744	p=1.5e-08 CI 0.8058-0.9703
3	BIK+GPRC5C+SPTLC2+TRIM3 CXCR4+CERK+CEBPB+DUSP4+TPO	p=5.3e-06 CI 0.6311-0.7989	p=2.0e-08 CI 0.6791-0.828	p=5.1e-11 CI 0.7224-0.8608	p=3.3e-11 CI 0.7483-0.8889	p=2.1e-08 CI 0.8079-0.9601
	Rounds compared for improvement (2,000 bootstrap p)					
	Round 3 versus Round 1	0.13	0.059	**0.032**	**0.016**	**0.00013**
	Round 2 versus Round 1	0.23	0.17	0.13	**0.044**	**0.00018**
	Round 3 versus Round 2	0.32	0.26	0.22	0.33	0.53

Round	Survival predictor	15-month BrCa	18-month BrCa	3-year BrCa	5-year BrCa	10-year BrCa
1	ERBB4+NCAM2+SCUBE2 ART3+FZD9+PSAT1	p=4.6e-16 CI 0.6665-0.756	p=9.8e-22 CI 0.6901-0.7711	p=6.3e-20 CI 0.6432-0.7147	p=3.9e-14 CI 0.6023-0.6712	p=5.9e-08 CI 0.5722-0.6535
2	BIK+GPRC5C+SPTLC2CEBPB+DUSP4+TPO	p=0.023 CI 0.4973-0.6078	p=0.0048 CI 0.5124-0.6131	p=1.5e-05 CI 0.5425-0.622	p=1.6e-07 CI 0.5578-0.6293	p=1.6e-06 CI 0.5587-0.6398
3	BIK+GPRC5C+SPTLC2+TRIM3 CXCR4+CERK+CEBPB+DUSP4+TPO	p=0.12 CI 0.4776-0.5833	p=0.10 CI 0.4821-0.5796	p=0.0036 CI 0.5136-0.5926	p=2.1e-05 CI 0.5390-0.6107	p=0.00036 CI 0.5309-0.6134
	Rounds compared for improvement (2,000 bootstrap p)					
	Round 1 versus Round 2	**5.40E-06**	**1.10E-07**	**3.30E-04**	**0.044**	0.32
	Round 1 versus Round 3	**1.60E-07**	**2.30E-10**	**2.10E-06**	**0.0063**	0.081

Significance and 95% CI are provided for each of the top predictors, in ROC analysis of either TNBC or all BrCa at the five selected timepoints. Comparisons of top predictor ROC curves testing for significance of improvement between rounds of ROC tests are also provided, with significant comparison p values bolded. See also Figure 8 and Tables S10–S12.

DISCUSSION

Complex interaction of hallmark cancer process genes affects cancer survival

A systems biology pipeline (Figure 1) was used to integrate large-scale BrCa transcriptomes, to assess RNA coexpression and to identify gene products influencing survival of patients with BrCa, with emphasis on TNBC, a heterogeneous BrCa subtype with the worst prognosis and fewest specific treatments. While the classical subtype according to representative markers ER, PR, and HER2 for TNBC are well known, the genomic characteristics of TNBC are more complex than expected. Among previous classification studies, six intrinsic subtypes of TNBC have been identified (basal-like [BL], androgen receptor [AR], mesenchymal [M], and immune). These intrinsic TNBC subtypes include two BL subtypes, one with increased cellular proliferation and response to DNA damage response (BL1, 10%) and another with high growth factor signaling with myoepithelial markers (BL2, 20%); two M subtypes associated with cell differentiation and growth factor signaling (M, 20% and MSL, 10%, respectively); an immunomodulatory (IM, 20%) type enriched with immune cell processes; and a luminal androgen subtype characterized by hormone signaling mediated by androgen receptor (LAR, 10%) (Lehmann et al., 2011, 2016; Burstein et al., 2015).

WGCNA identified prototypic module communities with opposing correlation to coordinated gene expression changes in TNBC, which were a rich source of survival driver genes. We homed-in on hubs of these modules having higher connectivity in breast-tissue-specific networks, which also implicated interesting characteristic genes known to drive BrCa including a bottleneck, *CEBPB*. Many DEGs from these modules were highly associated with survival by KM survival analysis. All the aforementioned results were fed into tests of combinatorial gene expression ratio by ROC analysis (Figure 8). We found that initially promising genes linked to BrCa survival and based on some of the most apparent DEGs (by fold change) also fitting the coexpressed, anticorrelated framework of selection criteria underperformed in predicting survival in TNBC versus BrCa in general. However, adding to the nomination list additional genes from neighboring, yet distinct M2-like and M12-like transcript communities, biasing these choices by different criteria for top-ranking DEGs with very low FDR from the comparison of TNBC to HER2-enriched BrCa subtypes, resulted in better performance for prediction of TNBC survival among the top-ranked ROC curves. Among the genes performing best, hallmark cancer processes were implicated as discussed in the following text. Therefore, to further augment performance and to cement reproducibility of our predictor ratio, we mined the literature for definitive work demonstrating the involvement of pathway-related genes on BrCa, if not TNBC survival, and prediction AUC for the top-performing ratios indeed increased.

The hallmark processes of cancer we identified as relevant to TNBC via our relatively unbiased nomination and testing for predictive linkage between gene expression and survival include genes such as *BIK*, encoding a protein that balances proapoptotic and antiapoptotic signaling roles, and key sphingolipid metabolism enzymes that can catalyze interconversion of a lipid class that functions with anticancer propensity into lipid species that promote cancer, each having effects on cellular motility, differentiation, apoptosis, inflammation/immunity, and angiogenesis. These lipids, and their sphingolipid pathway enzymes, are appreciated to be dysregulated in numerous cancers with potential as drug targets (Ryland et al., 2011). Interestingly, the initial hit in the pathway, serine palmitoyl transferase long-chain base complex subunit *SPTLC2* has a low cancer-literature-relevance score (Figure S3), but catalyzes the rate-limiting step controlling flux of molecules into the sphingolipid pool subject to transformation to downstream products in subsequent steps of the pathway, such as ceramide synthesis by *CERS4*, or ceramide-1-phosphate production by *CERK*, which switch the active function of their products relative to substrates, sphingosine, and ceramide, respectively.

Angiogenesis-specific gene function was also implicated by the M2-like transcript abundance of *GPRC5C*, identified as a gene suppressed by estrogen receptor activation and capable of slowing MCF-7 growth (Yamaga et al., 2014), and a close paralog of *GPRC5* family genes *GPRC5A* and *GPRC5B*, with more established roles in cancer to date (Acquafreda et al., 2009; De Francesco et al., 2017; Hirabayashi and Kim, 2020), including dysregulation of ceramide production in a *GPRC5B*-deficient model (Kim et al., 2018). *GPRC5C* is only yet established to function in maintaining relatively higher blood pH, which may have an effect on endothelial proliferation capacity (Faes et al., 2016). *CXCR4*, nominated from literature comention with *GPRC5C* (De Francesco et al., 2017), has roles in immunity and a tumor microenvironment in addition to angiogenesis.

Interestingly, elevated *CXCR4* promoting a conducive tumor microenvironment through upregulation of stress activated kinase signaling is consistent with a weak M12-like signature in our network and with better survival when expressed less abundantly (Chatterjee et al., 2014). *MAPK1* and 2, *p38*, *JNK*, and other stress-activated kinases are also regulated by *DUSP4* (Mazumdar et al., 2016), which we took to be M12-like based on module assignment to M21. But on examination, it is not well-assigned owing to discordance of dissimilarity (directly used in WGCNA for assignment) and kME, which implicates it as an M19 (M2-like) member (kME = 0.54). Indeed, *DUSP4* was reported to be downregulated in TNBC (Mazumdar et al., 2016). We subsequently tested ratios with *DUSP4* and 1/*DUSP4* in the numerator for improving ROC AUC to no avail (*data not shown*). A potential explanation for the positive *DUSP4* contribution to the predictive ratio when offside in the denominator, compared with where its network membership would suggest *DUSP4* has a more complex biological role. The loss in TNBC of *DUSP4* could be a downstream effect of other expression changes in TNBC that has negative correlation to patient survival but with no role affecting survival or dual roles including one that compensates for the other in TNBC. Elucidation of these possibilities requires further study. Regardless, both *DUSP4* (Saigusa et al., 2013) and *TRIM3* (Huang et al., 2017) on opposite sides of our 9-gene ratio are suppressors of metastasis of hepatic cancer, suggesting tissue-specific if not pan-cancer roles in survival.

CEBPB, the high-information-load bottleneck node between M12 and M2, promotes inflammation-mediated metastasis in BrCa (Kurzejamska et al., 2014). The aforementioned 8 genes of the ROC round 3 survival predictor ratio are shown in Figure 1 (Venn diagram at lower left) with their overlapping roles in 5 hallmark cancer-related processes indicated. The remaining gene of the 9, thyroid peroxidase (*TPO*), has a less established role in cancer with a low cancer-literature-relevance score (Piazza et al., 2017) of 22.0, but it has been shown that autoantibodies to *TPO* predict BrCa risk (Tosovic et al., 2012). In total, TNBC survival prediction capability of our 9-gene equal-weight ratio was bolstered by selection of DEGs within two anticorrelated sets of modules of the network tied to BrCa survival and specific shifts by subtype. The ability of these genes to both encode the survival risk of patients and integrate hallmark cancer processes in TNBC nominates them as linchpins of TNBC survival. We further showed that the BrCa network is robust, demonstrating that the nomination framework applied from the RNA-seq-derived network to predictions in array data validates in RNA-seq data.

Focus on DEGs of only M12 and M2 misses linchpin TNBC survival genes, but finds genes with well-studied roles in BrCa survival

All the gene ratios tested in round 1 performed better (or similar) for all BrCa compared with TNBC survival prediction. Moreover, the previously nominated linchpin survival genes of top-performing predictor combinations refined in ROC analysis rounds 2 and 3 in our pipeline outperformed genes in the naively informed round 1 predictor. This is attributed to more refined parameters for DEGs in later rounds compared with DEGs only in M2 and M12 and with the largest fold changes (of the TNBC versus Luminal A comparison) in the data set for round 1. Top DEGs by fold change (upregulated consistently in TNBC compared with each non-TNBC group) included *PSAT1*, an M12 hub (Table S6). *PSAT1* encodes phosphoserine aminotransferase. Selective loss of *PSAT1* suppresses migration, invasion, and experimental metastasis in TNBC (Metcalf et al., 2020). *PSAT1* catalyzes serine biosynthesis, where serine is required for several anabolic processes, such as protein, nucleic acid, and lipid synthesis, including for the initial step of *de novo* sphingosine production. Because metabolic processes are reprogrammed in cancer to promote growth and proliferation, it is not surprising that modified amino acid metabolism was overrepresented as an M12-specific biological process (Table 1). Another enzyme of the serine synthesis pathway found in M12 was *PHGDH* (kME_{tan} = 0.63), encoding 3-phosphoglycerate dehydrogenase. Overexpression of *PHGDH* has been reported in BrCa tumors and cell lines (Mullarky et al., 2016). Notably in TNBC, amplification and overexpression of *PHGDH* is associated with aggressive disease (Locasale et al., 2011; Possemato et al., 2011; Pollari et al., 2011). Mullarky et al. demonstrated that BrCa cell lines intrinsically overexpressing *PHGDH* are uniquely sensitive to its knockdown, whereas others are insensitive, suggesting that *PHGDH* inhibitors may have cancer cell survival in their crosshairs.

M2 downregulation in TNBC (Figure 3) was evident for M2 genes *TFF1* and *SCUBE2*, and their transcript levels also predict BrCa survival (Figure S1 and Table 2), as previously reported for *TFF1*, with established roles in the inhibition of proliferation, migration, and invasion of BrCa cells *in vivo* (Yi et al., 2020). *TFF1* is a secreted protein normally expressed in gastrointestinal mucosa, with transcriptional regulation by *ESR1* and 2 (Pelden et al., 2013). It is well-studied in relation to cancer (Figure S3, *TFF1* OncoScore = 80) and has been proposed as a biomarker for breast and other cancers (Wang et al., 2018; Yi et al., 2020; Yusufu et al., 2019; Klett et al., 2018). *TFF1* gene products were elevated in ER^+ and PR^+ BrCa (Yi et al., 2020; El-nagdy et al., 2018), whereas they are downregulated in TNBC (Yi et al., 2020). Indeed, *TFF1* levels correlate with *ESR1* and other transcription factors of M2, namely *GATA3*, *FOXA1*, and *MYB* (Table S5), supporting *TFF1* regulation by *ESR1* in our transcriptomic coexpression network. Given the aforemnentioned information, loss of M2 hub *ESR1* would be a driver of *TFF1* downregulation in TNBC and of increased cancer aggression. Enhancing expression or activity of *ESR1* or other M2 transcription factors could be a therapeutic approach in TNBC.

SCUBE2 is a lipid-binding protein and coreceptor for *VEGFR2* to mediate angiogenesis. Guan et al. reported high *SCUBE2* expression associated with increased disease-free interval and good prognosis in BrCa, also finding ethnicity-specific differences in the expression of *SCUBE2* (Guan et al., 2020). *SCUBE2* is downregulated in invasive BrCa but overexpressed by breast cancer stem cells (BCSCs) (Chen et al., 2018). Possibly, the true prognostic value of *SCUBE2* is not observed in tissue biopsies because BCSCs are rare cells compared with differentiated cells with lower expression of *SCUBE2*. TNBC has a higher proportion of BCSCs compared with other BrCa subtypes (Honeth et al., 2008), promoting progression through proliferation, migration to metastatic sites, and therapy resistance. *SCUBE2*-expressing TNBC BCSCs also increase NOTCH signaling, epithelial to mesenchymal transition, and chemoresistance. Patients with TNBC with African ancestry have a higher prevalence of tumors expressing stem cell markers such as CD44, which elevates the expression of *SCUBE2* (Jiagge et al., 2018), thus *SCUBE2* could serve as a therapeutic target for African-American patients with TNBC. Our study and those cited support *SCUBE2* and *TFF1* links to estrogen signaling, but further *in vitro* studies are necessary to determine how these genes are connected in TNBC signaling (Fernandez-Ramires et al., 2009; Smith et al., 2008). Given the additional role of *SCUBE2* in SHH signaling (Tsai et al., 2009) and SHH signaling downregulation by ceramide (Gencer et al., 2017), a closer look at interactions of *SCUBE2* with ceramides and linchpin survival genes affecting sphingolipid balance is also warranted.

We noticed that M12 hub genes *PSAT1*, *YBX1*, and *MTHFD1L* with high confidence interactions to the nucleic acid and DNA metabolism module M7 (Figure 7; *M7 nodes, black*), plus *PHGDH*, have been comentioned in previous work finding gross dysregulation of the proteome in model cell lines for *HPRT1* deficiency (Dammer et al., 2015). Hypergeometric overlap testing was performed for overrepresentation of 1,055 *HPRT1* deficiency-dysregulated gene products among M2-like or M12-like module hubs (kME$\geq$0.70, n = 552) in the BrCa network, relative to the hubs of other modules (n = 1,209). Of 88 *HPRT1*-linked hub proteins, 38 were hubs in the 9 BrCa network modules of core interest to TNBC (p = 0.0066; OR = 1.79), indicating significant enrichment within hubs of our TNBC survival-linked network communities. Twenty of the 38 were M7 hubs, and M7 harbors *HPRT1* as a bottleneck (kME_{black} = 0.561; kME_{yellow} = 0.556). As an M12-like module, the M7 DNA synthesis module is elevated in TNBC, consistent with the study by Sedano, et al., which found *HPRT1* elevation particularly in TNBC, and association with poor BrCa outcomes (Sedano et al., 2020). The unusual genetic interaction network of *HPRT1*, as seen in both proteomics and RNA, makes this region of the network a rich hunting ground for mechanisms of cancer progression. We found that connectivity among the M12-like community of modules is relevant to TNBC survival, which supported our decision to expand our search over the network landscape to all M2-like and M12-like modules when we homed-in on TNBC linchpin survival genes.

Overall, ROC analysis helped us establish a predictor of progression free survival in BrCa without literature or other selection bias. We discuss how the unique combination of implicated pathways relating to BrCa survival overlap with known hallmark pathways of cancer. We predict that therapies targeting multiple of the 9 implicated targets and their respective pathways via modulation of expression (or function) of these key driver genes will improve treatment outcomes through synergistic effects on mechanistically distinct molecular pathways influencing BrCa progression, relapse, and survival.

Limitations of the study

The identification of candidate biomarker genes from high throughput abundance data for the use of therapeutic prediction is subject to defects by way of analysis parameters. Often, the result is unreproducible by complementary methods, limiting the power of translating these gene lists into clinical biomarkers and therapeutic targets. The workflow used in this study addressed this problem by integrating systems biology approaches to discover nominated genes and unbiasedly testing their combinations, representative of biological interaction possibilities, arriving at multigene survival indicators well-supported by prior biological knowledge, also demonstrating that existing literature can augment the predictivity of a nominated biomarker list. Text mining in future integrative systems biology analysis pipelines such as ours, amenable to automation, would be valuable.

We sampled a small region of a vast network landscape. Ideally, all hubs, hub interactors, bottlenecks, and DEGs of M2-and M12-like modules would have been tested for potential to contribute to the combinatorial predictive biomarker. Although logical as the choice of genes likely to differentiate patient survival, examining the top 20 or fewer DEGs of the M2(-like) and M12(-like) modules for survival and ROC analyses probably introduced bias toward better-studied genes, while leaving many details of the network landscape unexplored. To optimize predictors of TNBC survival, distinct subsets of DEGs were nominated, but changing the nomination criteria will allow for application of the analysis pipeline to other subtypes or BrCa in general. The computational demands of ROC testing of all gene combinations in an expanded landscape search would explode owing to the number of potential combinations to test. With only 24 numerator and 34 denominator candidates, round 2 of ROC analysis already tested more than100,000 combinations. This also could be viewed as a multiple testing problem, despite the large number of biological measurements involved. Machine learning might be used to optimize prescreen of small-number combinations before assembling them into more complex gene ratios. Indeed, machine learning approaches for BrCa and TNBC prognostic marker panel definition are gaining traction (Alsaleem et al., 2020). Weights for gene contributions to combinatorial predictors are often inferred by linear regression (Liu et al., 2019). We considered all combinations with equal gene weight after sample-wise normalization, which affected underlying biological functional interactions driving the ratio's sensitivity and specificity. Future work may explore the benefits of integrating both unweighted and weighted combinatorial predictors.

Our ROC analysis validating array-based results in RNA-seq data used corrected and complete traits for the TCGA samples (Liu et al., 2018). Nonetheless, the statistical power of the validation ROC analysis was limited by an unusual imbalance of outcomes in the cases with sufficient follow-up time. Twenty-three of the 92 TCGA TNBC cases were tumor-free before 15 months, at their last follow-up. This highlights the possible impact of therapies on candidate gene linkage to survival outcomes, hiding markers of naive BrCa prognosis already targeted by therapy. Because we chose to focus on survival of patients with

TNBC, the biomarkers we have found are likely poorly engaged by current treatments. Biomarkers relying on bulk tissue biopsy also suffer from an inability to distinguish rarer cell type contributions to gene product abundances and prognosis. Integrated systems biology pipelines examining single-cell RNA-seq data therefore hold promise to distinguish BCSC and other rare but important cell-type contributions to BrCa mortality.

Conclusion

This study successfully leverages coexpression and interaction networks of invasive tumors of the four BrCa subtypes in large transcriptomic cohorts, overlapping this structure with differential expression, nominating candidate biomarker genes tested for combinatorial functional interactions that influence survival of patients with TNBC. It also demonstrates that hub status is insufficient to assign genes as key drivers of causality and directionality in WGCNA networks. The combinations performing best comprised linchpin survival genes, representing hallmark cancer processes that involve sphingolipid metabolism, regulation of apoptosis, proteostasis, angiogenesis, and metastasis propensity. Therefore, our top survival-related genes from ROC ranking are generally already well-established, but their specific combination here implicates unappreciated functional interactions in BrCa remaining to be fully explored. Thus, our network serves as a resource for the research community. The networks identified for BrCa also remain to be leveraged by systems pharmacology, which may focus on therapies targeting the functions of gene combinations that normalize the profile of entire survival-associated network modules. Finally, this analysis pipeline holds promise for broader application to other disease-specific tissue-level transcriptomic and proteomic networks.

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Dr. James W Lillard, Jr (jlillard@msm.edu).

Materials availability

Materials used or generated in this study will be available upon reasonable request, and a material transfer agreement may be required.

Data and code availability

All data are available from the corresponding author upon reasonable request.

METHODS

All methods can be found in the accompanying transparent methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102451.

ACKNOWLEDGMENTS

Dr. Kathleen Carter's critical reading of this manuscript was indispensable. We thank TCGA for use of BRCA project data, and all patients that donated samples for transcriptomics.

This study is funded in part by the National Cancer Institute of the National Institutes of Health U54CA118638 and P30CA138292, the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R25GM058268, the Gates Millennium Scholarship Program, the Winship Cancer Institute, and the Morehouse School of Medicine.

AUTHOR CONTRIBUTIONS

C.D.D: Conceptualized and designed, acquired data, analyzed and interpreted data, and prepared manuscript. E.D: analyzed and interpreted data, and prepared manuscript. T. L. G: analyzed and interpreted data, and prepared manuscript. N.T.S.: provided code and resources. J.W.L: Conceptualized and designed and interpreted data, and prepared manuscript. All authors approved the final manuscript.

DECLARATION OF INTERESTS

Drs. James W. Lillard Jr., Eric B. Dammer, and Nick Seyfried report no competing interests related to this manuscript. Courtney D. Dill and Ti'ara L. Griffin also report no competing interests related to this manuscript.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in science. One or more of the authors of this paper received support from a program designed to increase minority representation in science. The author list of this paper includes contributors from the location where the research was conducted who participated in the data collection, design, analysis, and/or interpretation of the work.

Received: November 19, 2020
Revised: March 8, 2021
Accepted: April 15, 2021
Published: May 21, 2021

REFERENCES

Acquafreda, T., Soprano, K.J., and Soprano, D.R. (2009). GPRC5A: a potential tumor suppressor and oncogene. Cancer Biol. Ther. *8*, 963–965.

Alsaleem, M.A., Ball, G., Toss, M.S., Raafat, S., Aleskandarany, M., Joseph, C., Ogden, A., Bhattarai, S., Rida, P.C.G., Khani, F., et al. (2020). A novel prognostic two-gene signature for triple negative breast cancer. Mod. Pathol. *33*, 2208–2220.

Burstein, M.D., Tsimelzon, A., Poage, G.M., Covington, K.R., Contreras, A., Fuqua, S.A., Savage, M.I., Osborne, C.K., Hilsenbeck, S.G., and Chang, J.C. (2015). Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. Clin. Cancer Res. *21*, 1688–1698.

Chatterjee, S., Behnam Azad, B., and Nimmagadda, S. (2014). The intricate role of CXCR4 in cancer. Adv. Cancer Res. *124*, 31–82.

Chen, J.H., Kuo, K.T., Bamodu, O.A., Lin, Y.C., Yang, R.B., Yeh, C.T., and Chao, T.Y. (2018). Upregulated SCUBE2 expression in breast cancer stem cells enhances triple negative breast cancer aggression through modulation of notch signaling and epithelial-to-mesenchymal transition. Exp. Cell Res. *370*, 444–453.

Dammer, E.B., Göttle, M., Duong, D.M., Hanfelt, J., Seyfried, N.T., and Jinnah, H.A. (2015). Consequences of impaired purine recycling on the proteome in a cellular model of Lesch–Nyhan disease. Mol. Genet. Metab. *114*, 570–579.

De Francesco, E.M., Sotgia, F., Clarke, R.B., Lisanti, M.P., and Maggiolini, M. (2017). G protein-coupled receptors at the crossroad between physiologic and pathologic angiogenesis: old paradigms and emerging concepts. Int. J. Mol. Sci. *18*, 2713.

Elnagdy, M.H., Farouk, O., Seleem, A.K., and Nada, H.A. (2018). TFF1 and TFF3 mRNAs are higher in blood from breast cancer patients with metastatic disease than those without. J. Oncol. *2018*, 4793498.

Faes, S., Uldry, E., Planche, A., Santoro, T., Pythoud, C., Demartines, N., and Dormond, O. (2016). Acidic pH reduces VEGF-mediated endothelial cell responses by downregulation of VEGFR-2; relevance for anti-angiogenic therapies. Oncotarget *7*, 86026–86038.

Fernandez-Ramires, R., Sole, X., De Cecco, L., Llort, G., Cazorla, A., Bonifaci, N., Garcia, M.J., Caldes, T., Blanco, I., Gariboldi, M., et al. (2009). Gene expression profiling integrated into network modelling reveals heterogeneity in the mechanisms of BRCA1 tumorigenesis. Br. J. Cancer *101*, 1469–1480.

Gencer, S., Oleinik, N., Kim, J., Panneer Selvam, S., De Palma, R., Dany, M., Nganga, R., Thomas, R.J., Senkal, C.E., Howe, P.H., et al. (2017). TGF-β receptor I/II trafficking and signaling at primary cilia are inhibited by ceramide to attenuate cell migration and tumor metastasis. Sci. Signal. *10*, eaam7464.

Goncalves, H., Jr., Guerra, M.R., Duarte Cintra, J.R., Fayer, V.A., Brum, I.V., and Bustamante Teixeira, M.T. (2018). Survival study of triple-negative and non-triple-negative breast cancer in a Brazilian cohort. Clin. Med. Insights Oncol. *12*, 1179554918790563.

Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet. *47*, 569–576.

Guan, X., Cai, M., Du, Y., Yang, E., Ji, J., and Wu, J. (2020). CVCDAP: an integrated platform for molecular and clinical analysis of cancer virtual cohorts. Nucleic Acids Res. *48*, W463–W471.

Gyorffy, B., Lanczky, A., Eklund, A., Denkert, C., Budczies, J., Li, Q., and Szallasi, Z. (2010). 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 Res. Treat. *123*, 725–731.

Györffy, B., and Schäfer, R. (2009). Meta-analysis of gene expression profiles related to relapse-free survival in 1,079 breast cancer patients. Breast Cancer Res. Treat. *118*, 433–441.

Hanahan, D., and Weinberg, Robert a. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Hirabayashi, Y., and Kim, Y.-J. (2020). Roles of GPRC5 family proteins: focusing on GPRC5B and lipid-mediated signalling. J. Biochem. *167*, 541–547.

Honeth, G., Bendahl, P.O., Ringner, M., Saal, L.H., Gruvberger-Saal, S.K., Lovgren, K., Grabau, D., Ferno, M., Borg, A., and Hegardt, C. (2008). The CD44+/CD24- phenotype is enriched in basal-like breast tumors. Breast Cancer Res. *10*, R53.

Huang, X.-Q., Zhang, X.-F., Xia, J.-H., Chao, J., Pan, Q.-Z., Zhao, J.-J., Zhou, Z.-Q., Chen, C.-L., Tang, Y., Weng, D.-S., et al. (2017). Tripartite motif-containing 3 (TRIM3) inhibits tumor growth and metastasis of liver cancer. Chin. J. Cancer *36*, 77.

Jiagge, E., Chitale, D., and Newman, L.A. (2018). Triple-negative breast cancer, stem cells, and african ancestry. Am. J. Pathol. *188*, 271–279.

Karn, T., Pusztai, L., Holtrich, U., Iwamoto, T., Shiang, C.Y., Schmidt, M., Müller, V., Solbach, C., Gaetje, R., Hanker, L., et al. (2011). Homogeneous datasets of triple negative breast cancers enable the identification of novel prognostic and predictive signatures. PLoS One *6*, e28403.

Kim, Y.-J., Greimel, P., and Hirabayashi, Y. (2018). GPRC5B-Mediated sphingomyelin synthase 2 phosphorylation plays a critical role in insulin resistance. iScience *8*, 250–266.

Klett, H., Fuellgraf, H., Levit-Zerdoun, E., Hussung, S., Kowar, S., Kusters, S., Bronsert, P., Werner, M., Wittel, U., Fritsch, R., et al. (2018). Identification and validation of a diagnostic and prognostic multi-gene biomarker panel for pancreatic ductal adenocarcinoma. Front. Genet. *9*, 108.

Kurzejamska, E., Johansson, J., Jirström, K., Prakash, V., Ananthaseshan, S., Boon, L., Fuxe, J., and Religa, P. (2014). C/EBPβ expression is an independent predictor of overall survival in breast cancer patients by MHCII/CD4-dependent mechanism of metastasis formation. Oncogenesis *3*, e125.

Langfelder, P., and Horvath, S. (2012). Fast R functions for robust correlations and hierarchical clustering. J. Stat. Softw. *46*, i11.

Lehmann, B.D., Bauer, J.A., Chen, X., Sanders, M.E., Chakravarthy, A.B., Shyr, Y., and Pietenpol, J.A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J. Clin. Invest. *121*, 2750–2767.

Lehmann, B.D., Jovanović, B., Chen, X., Estrada, M.V., Johnson, K.N., Shyr, Y., Moses, H.L., Sanders, M.E., and Pietenpol, J.A. (2016). Refinement of triple-negative breast cancer molecular subtypes: implications for neoadjuvant chemotherapy selection. PLoS One *11*, e0157368.

Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell *173*, 400–416.e11.

Liu, L., Chen, Z., Shi, W., Liu, H., and Pang, W. (2019). Breast cancer survival prediction using seven prognostic biomarker genes. Oncol. Lett. *18*, 2907–2916.

Locasale, J.W., Grassian, A.R., Melman, T., Lyssiotis, C.A., Mattaini, K.R., Bass, A.J., Heffron, G., Metallo, C.M., Muranen, T., Sharfi, H., et al. (2011). Phosphoglycerate dehydrogenase diverts glycolytic flux and contributes to oncogenesis. Nat. Genet. *43*, 869–874.

Mazumdar, A., Poage, G.M., Shepherd, J., Tsimelzon, A., Hartman, Z.C., Den Hollander, P., Hill, J., Zhang, Y., Chang, J., Hilsenbeck, S.G., et al. (2016). Analysis of phosphatases in ER-negative breast cancers identifies DUSP4 as a critical regulator of growth and invasion. Breast Cancer Res. Treat. *158*, 441–454.

Metcalf, S., Dougherty, S., Kruer, T., Hasan, N., Biyik-Sit, R., Reynolds, L., and Clem, B.F. (2020). Selective loss of phosphoserine aminotransferase 1 (PSAT1) suppresses migration, invasion, and experimental metastasis in triple negative breast cancer. Clin. Exp. Metastasis *37*, 187–197.

Mullarky, E., Lucki, N.C., Beheshti Zavareh, R., Anglin, J.L., Gomes, A.P., Nicolay, B.N., Wong, J.C.Y., Christen, S., Takahashi, H., Singh, P.K., et al. (2016). Identification of a small molecule inhibitor of 3-phosphoglycerate dehydrogenase to target serine biosynthesis in cancers. Proc. Natl. Acad. Sci. U S A *113*, 1778–1783.

Nielsen, T.O., Hsu, F.D., Jensen, K., Cheang, M., Karaca, G., Hu, Z., Hernandez-Boussard, T., Livasy, C., Cowan, D., Dressler, L., et al. (2004). Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. Clin. Cancer Res. *10*, 5367–5374.

Oldham, M.C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., and Geschwind, D.H. (2008). Functional organization of the transcriptome in human brain. Nat. Neurosci. *11*, 1271–1282.

Pelden, S., Insawang, T., Thuwajit, C., and Thuwajit, P. (2013). The trefoil factor 1 (TFF1) protein involved in doxorubicininduced apoptosis resistance is upregulated by estrogen in breast cancer cells. Oncol. Rep. *30*, 1518–1526.

Piazza, R., Ramazzotti, D., Spinelli, R., Pirola, A., De Sano, L., Ferrari, P., Magistroni, V., Cordani, N., Sharma, N., and Gambacorti-Passerini, C. (2017). OncoScore: a novel, Internet-based tool to assess the oncogenic potential of genes. Sci. Rep. *7*, 46290.

Pollari, S., Käkönen, S.M., Edgren, H., Wolf, M., Kohonen, P., Sara, H., Guise, T., Nees, M., and Kallioniemi, O. (2011). Enhanced serine production by bone metastatic breast cancer cells stimulates osteoclastogenesis. Breast Cancer Res. Treat. *125*, 421–430.

Possemato, R., Marks, K.M., Shaul, Y.D., Pacold, M.E., Kim, D., Birsoy, K., Sethumadhavan, S., Woo, H.K., Jang, H.G., Jha, A.K., et al. (2011). Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. Nature *476*, 346–350.

Riaz, S.K., Ke, Y., Wang, F., Kayani, M.A., and Malik, M.F.A. (2019). Influence of SHH/GLI1 axis on EMT mediated migration and invasion of breast cancer cells. Sci. Rep. *9*, 6620.

Ryland, L.K., Fox, T.E., Liu, X., Loughran, T.P., and Kester, M. (2011). Dysregulation of sphingolipid metabolism in cancer. Cancer Biol. Ther. *11*, 138–149.

Saigusa, S., Inoue, Y., Tanaka, K., Toiyama, Y., Okugawa, Y., Shimura, T., Hiro, J., Uchida, K., Mohri, Y., and Kusunoki, M. (2013). Decreased expression of DUSP4 is associated with liver and lung metastases in colorectal cancer. Med. Oncol. *30*, 620.

Schwalm, S., Erhardt, M., Römer, I., Pfeilschifter, J., Zangemeister-Wittke, U., and Huwiler, A. (2020). Ceramide kinase is upregulated in metastatic breast cancer cells and contributes to migration and invasion by activation of PI 3-kinase and akt. Int. J. Mol. Sci. *21*, 1396.

Sedano, M.J., Ramos, E.I., Choudhari, R., Harrison, A.L., Subramani, R., Lakshmanaswamy, R., Zilaie, M., and Gadad, S.S. (2020). Hypoxanthine phosphoribosyl transferase 1 is upregulated, predicts clinical outcome and controls gene expression in breast cancer. Cancers (Basel) *12*, 1522.

Smith, D.D., Saetrom, P., Snove, O., Jr., Lundberg, C., Rivas, G.E., Glackin, C., and Larson, G.P. (2008). Meta-analysis of breast cancer microarray studies in conjunction with conserved cis-elements suggest patterns for coordinate regulation. BMC Bioinformatics *9*, 63.

Tosovic, A., Becker, C., Bondeson, A.-G., Bondeson, L., Ericsson, U.-B., Malm, J., and Manjer, J. (2012). Prospectively measured thyroid hormones and thyroid peroxidase antibodies in relation to breast cancer risk. Int. J. Cancer *131*, 2126–2133.

Tsai, M.-T., Cheng, C.-J., Lin, Y.-C., Chen, C.-C., Wu, A.-R., Wu, M.-T., Hsu, C.-C., and Yang, R.-B. (2009). Isolation and characterization of a secreted, cell-surface glycoprotein SCUBE2 from humans. Biochem. J. *422*, 119–128.

Wang, W., Li, Z., Wang, J., Du, M., Li, B., Zhang, L., Li, Q., Xu, J., Wang, L., Li, F., et al. (2018). A functional polymorphism in TFF1 promoter is associated with the risk and prognosis of gastric cancer. Int. J. Cancer *142*, 1805–1816.

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

Wong, A.K., Krishnan, A., and Troyanskaya, O.G. (2018). Giant 2.0: genome-scale integrated analysis of gene networks in tissues. Nucleic Acids Res. *46*, W65–W70.

Yamaga, R., Ikeda, K., Boele, J., Horie-Inoue, K., Takayama, K.-I., Urano, T., Kaida, K., Carninci, P., Kawai, J., Hayashizaki, Y., et al. (2014). Systemic identification of estrogen-regulated genes in breast cancer cells through cap analysis of gene expression mapping. Biochem. Biophys. Res. Commun. *447*, 531–536.

Yao, L., and Rzhetsky, A. (2008). Quantitative systems-level determinants of human genes targeted by successful drugs. Genome Res. *18*, 206–213.

Yi, J., Ren, L., Li, D., Wu, J., Li, W., Du, G., and Wang, J. (2020). Trefoil factor 1 (TFF1) is a potential prognostic biomarker with functional significance in breast cancers. Biomed. Pharmacother. *124*, 109827.

Yu, H., Kim, P.M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007). The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput. Biol. *3*, e59.

Yusufu, A., Shayimu, P., Tuerdi, R., Fang, C., Wang, F., and Wang, H. (2019). TFF3 and TFF1 expression levels are elevated in colorectal cancer and promote the malignant behavior of colon cancer by activating the EMT process. Int. J. Oncol. *55*, 789–804.

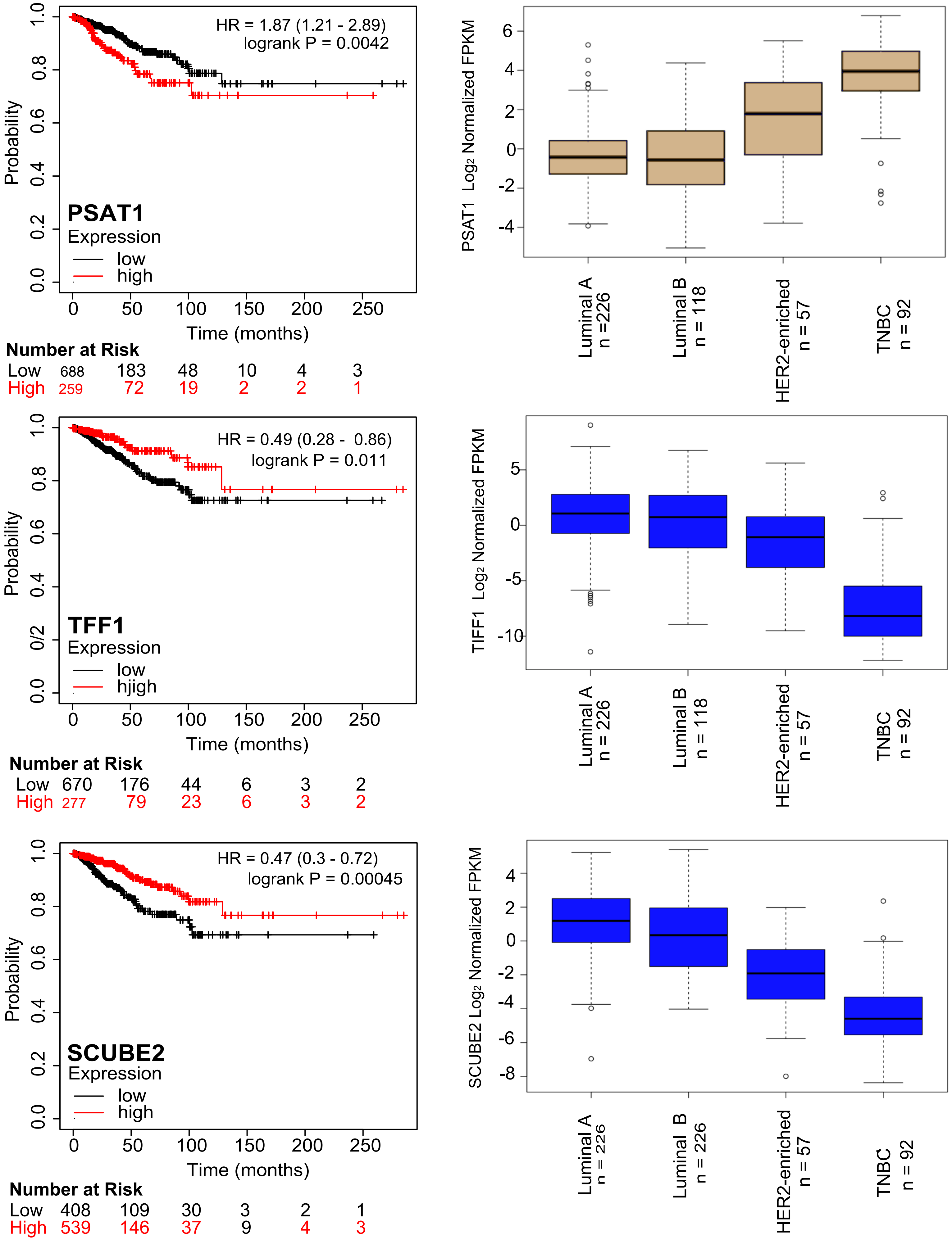
Zambon, A.C., Gaj, S., Ho, I., Hanspers, K., Vranizan, K., Evelo, C.T., Conklin, B.R., Pico, A.R., and Salomonis, N. (2012). GO-Elite: a flexible solution for pathway and ontology over-representation. Bioinformatics *28*, 2209–2210.

iScience, Volume *24*

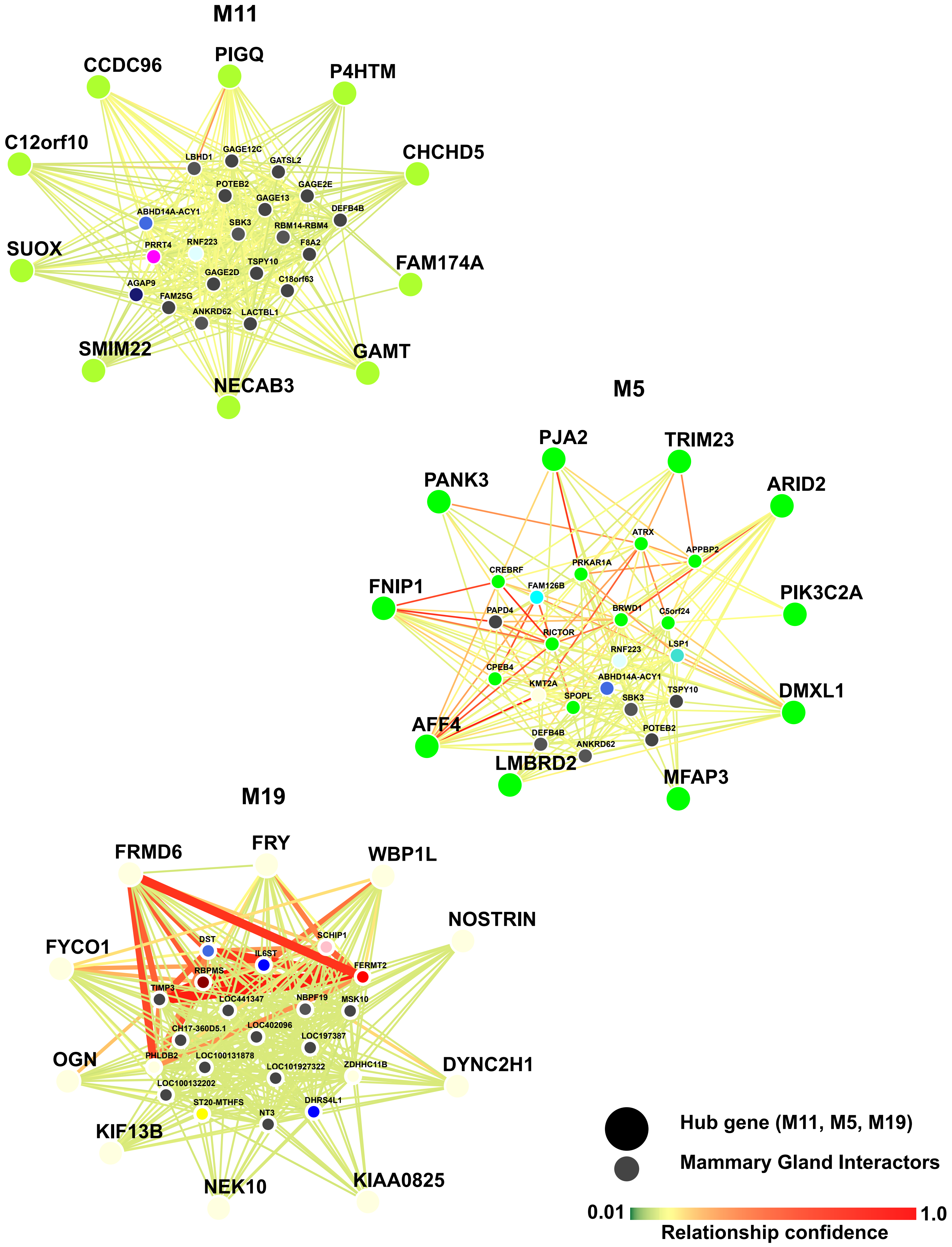
Supplemental information

A network approach reveals driver genes associated with survival of patients with triple-negative breast cancer

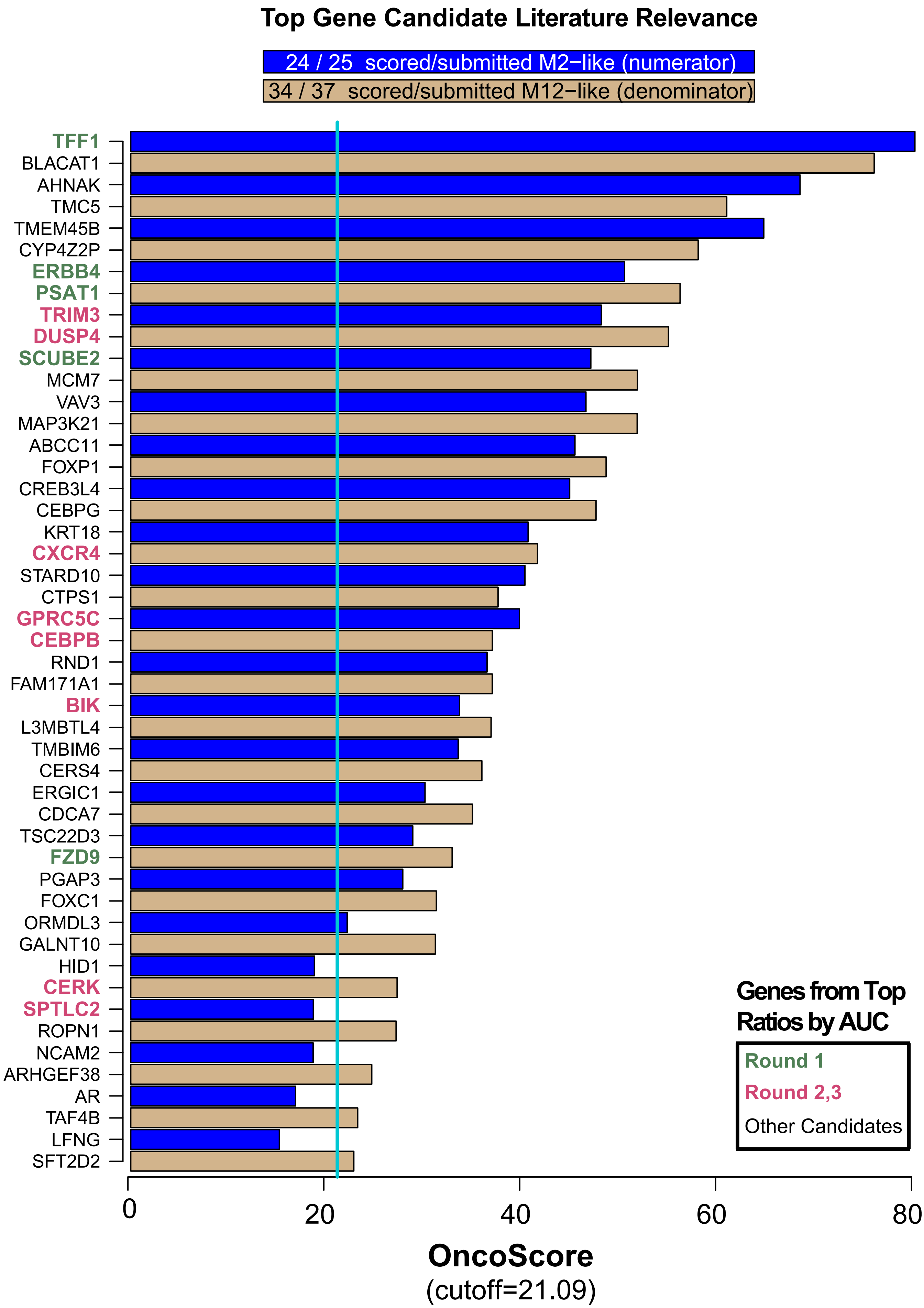
Courtney D. Dill, Eric B. Dammer, Ti'ara L. Griffen, Nicholas T. Seyfried, and James W. Lillard Jr.


HR = 1.87 (1.21 - 2.89)
logrank P = 0.0042
Probability
1.0
0.8
0.6
0.4
0.2
0.0
PSAT1
Expression
low
high
0
50
100
150
200
250
Time (months)
Number at Risk
Low 688 183 48 10 4 3
High 259 72 19 2 2 1
PSAT1 Log2 Normalized FPKM
6
4
2
0
-2
-4
Luminal A n =226
Luminal B n = 118
HER2-enriched n = 57
TNBC n = 92
HR = 0.49 (0.28 - 0.86)
logrank P = 0.011
Probability
1.0
0.8
0.6
0.4
0/2
0.0
TFF1
Expression
low
hjigh
0
50
100
150
200
250
Time (months)
Number at Risk
Low 670 176 44 6 3 2
High 277 79 23 6 3 2
TIFF1 Log2 Normalized FPKM
5
0
-5
-10
Luminal A n = 226
Luminal B n = 118
HER2-enriched n = 57
TNBC n = 92
HR = 0.47 (0.3 - 0.72)
logrank P = 0.00045
Probability
1.0
0.8
0.6
0.4
0.2
0.0
SCUBE2
Expression
low
high
0
50
100
150
200
250
Time (months)
Number at Risk
Low 408 109 30 3 2 1
High 539 146 37 9 4 3
SCUBE2 Log2 Normalized FPKM
4
2
0
-2
-4
-6
-8
Luminal A n = 226
Luminal B n = 226
HER2-enriched n = 57
TNBC n = 92

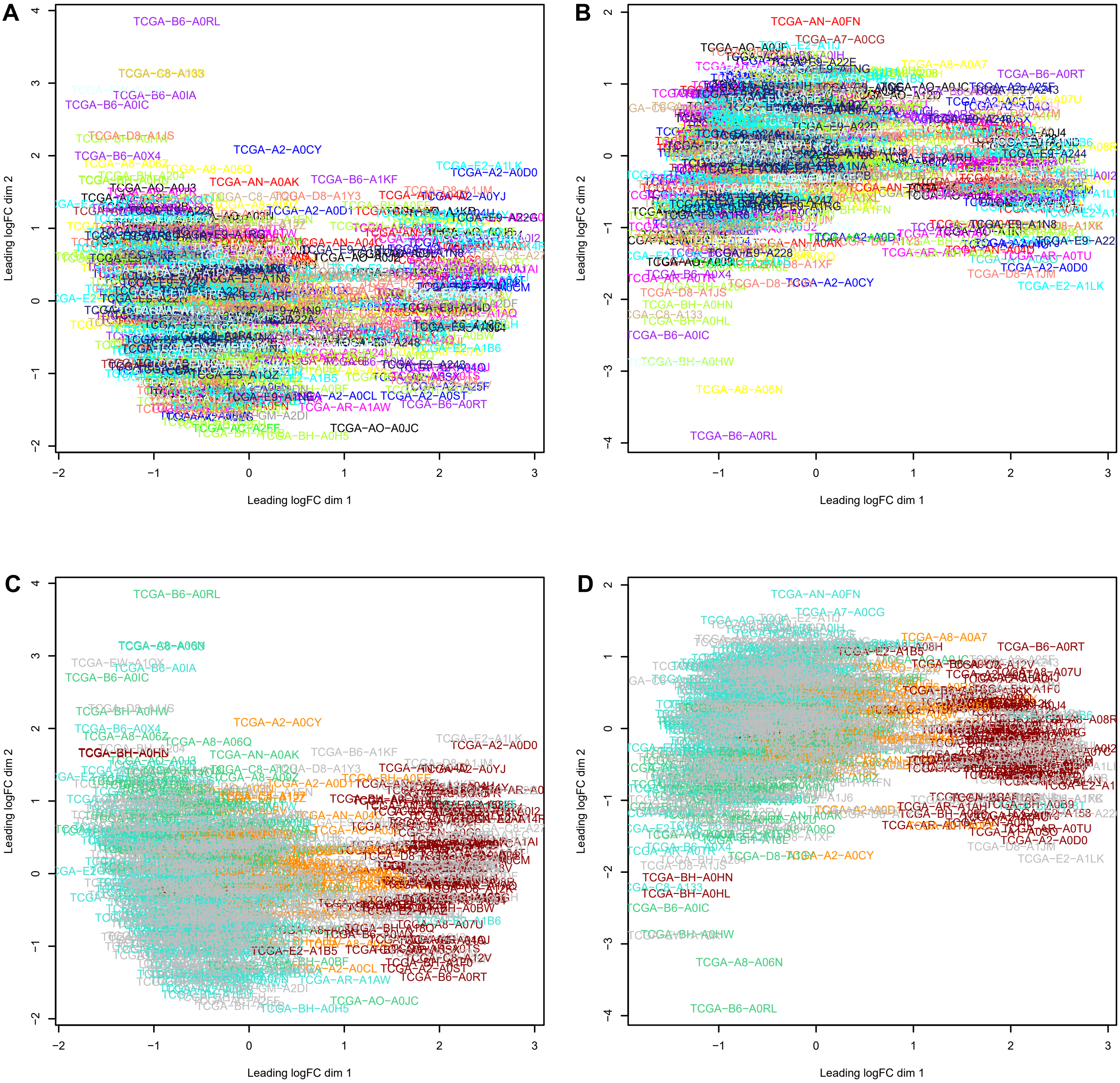
Supplementary Figure 1, Related to Figure 6 and Table 2. Survival and gene expression analysis reveal association between PSAT1, TFF1, SCUBE2, and RFS in BrCa. Hazard ratios confirm low expression of PSAT1 resulted in significant high likelihood of relapse, while high expression of TFF1 and SCUBE2 resulted in significant less likelihood of relapse (left). The red line in each Kaplan-Meier plot represents survival of cases in the higher expression tier and the black line represents survival in cases with lower expression. A log-rank (Mantel-Cox) test determined *p*-values and hazard ratio (HR) scores. Survival results were validated by measuring mRNA gene expression reflected, as $\log_2$ FPKM, via box and whisker plots (right). Gene expression of PSAT1 mRNA was highest among the TNBC subtype biopsy group compared to non-TNBC, while TFF1 and SCUBE2 were lowest in TNBC. A two group, Wilcoxon rank-sum, test determined *p*-values.


M11
PIGQ
CCDC96
P4HTM
C12orf10
CHCHD5
SUOX
FAM174A
SMIM22
GAMT
NECAB3
LBHD1
GAGE12C
GATSL2
POTEB2
GAGE2E
GAGE13
DEFB4B
ABHD14A-ACY1
SBK3
RBM14-RBM4
PRRT4
RNF223
F8A2
TSPY10
GAGE2D
C18orf63
AGAP9
FAM25G
ANKRD62
LACTBL1
M5
PJA2
TRIM23
PANK3
ARID2
FNIP1
PIK3C2A
DMXL1
AFF4
LMBRD2
MFAP3
ATRX
APPBP2
CREBRF
PRKAR1A
FAM126B
PAPD4
BRWD1
C5orf24
RICTOR
RNF223
LSP1
CPEB4
ABHD14A-ACY1
KMT2A
TSPY10
SPOPL
SBK3
DEFB4B
POTEB2
ANKRD62
M19
FRMD6
FRY
WBP1L
NOSTRIN
FYCO1
DST
SCHIP1
IL6ST
FERMT2
RBPMS
TIMP3
LOC441347
NBPF19
MSK10
CH17-360D5.1
LOC402096
LOC197387
OGN
PHLDB2
LOC100131878
LOC101927322
ZDHHC11B
DYNC2H1
LOC100132202
ST20-MTHFS
DHRS4L1
NT3
KIF13B
NEK10
KIAA0825
Hub gene (M11, M5, M19)
Mammary Gland Interactors
0.01
1.0
Relationship confidence

Supplementary Figure 2, Related to Figure 7. Genome-scale integrated analysis of gene networks in tissues (GIANT) identifies lower connectivity breast tissue-specific interactions between survival-linked genes and module hub-implicated genes as well as hubs in M2-like modules (M11, M5, M19). The heatmap scale represents the confidence of the predicted interactions based on mammary gland specific interactions. High confidence interactions are represented by red edges (see scale). Input for GIANT was restricted to the top ten hub genes (kME = 0.6 or higher) for each module. Small nodes in the center of each network are implicated by the GIANT database network as the best connected to the hubs specified in mammary tissue, colored by their module membership.


Top Gene Candidate Literature Relevance
24 / 25 scored/submitted M2−like (numerator)
34 / 37 scored/submitted M12−like (denominator)
TFF1
BLACAT1
AHNAK
TMC5
TMEM45B
CYP4Z2P
ERBB4
PSAT1
TRIM3
DUSP4
SCUBE2
MCM7
VAV3
MAP3K21
ABCC11
FOXP1
CREB3L4
CEBPG
KRT18
CXCR4
STARD10
CTPS1
GPRC5C
CEBPB
RND1
FAM171A1
BIK
L3MBTL4
TMBIM6
CERS4
ERGIC1
CDCA7
TSC22D3
FZD9
PGAP3
FOXC1
ORMDL3
GALNT10
HID1
CERK
SPTLC2
ROPN1
NCAM2
ARHGEF38
AR
TAF4B
LFNG
SFT2D2
Genes from Top Ratios by AUC
Round 1
Round 2,3
Other Candidates
0
20
40
60
80
OncoScore
(cutoff=21.09)

Supplementary Figure 3, Related to Figure 8. Scored representation of gene products in the cancer literature. OncoScore (Piazza et al., 2017) was used to determine the relevance of gene products to cancer based on existing literature on Pubmed. The top scoring 24 M2-like (numerator) and 24 M12-like (denominator) gene symbols are shown.


A
B
C
D
Leading logFC dim 1
Leading logFC dim 2
-2
-1
0
1
2
3
4
TCGA-B6-A0RL
TCGA-C8-A133
TCGA-B6-A0IA
TCGA-B6-A0IC
TCGA-D8-A1JS
TCGA-B6-A0X4
TCGA-A2-A0CY
TCGA-A8-A06Q
TCGA-AN-A0AK
TCGA-B6-A1KF
TCGA-E2-A1LK
TCGA-A2-A0D0
TCGA-A2-A0YJ
TCGA-A2-A0D1
TCGA-E3-A1QZ
TCGA-E9-A1NG
TCGA-A2-A0CL
TCGA-A2-A0ST
TCGA-AR-A1AW
TCGA-B6-A0RT
TCGA-AO-A0JC
TCGA-BH-A0H5
TCGA-AC-A2FF
TCGA-AN-A0FN
TCGA-A7-A0CG
TCGA-AO-A0JF
TCGA-E9-A22E
TCGA-A8-A0A7
TCGA-A2-A04Q
TCGA-A2-A0ST
TCGA-A8-A07U
TCGA-E9-A22D
TCGA-E9-A244
TCGA-E9-A247
TCGA-E9-A1RG
TCGA-BH-A1FN
TCGA-E9-A1N8
TCGA-A2-A0D1
TCGA-E9-A22
TCGA-AR-A0TU
TCGA-AN-A04D
TCGA-A2-A0D0
TCGA-D8-A1JM
TCGA-E9-A228
TCGA-D8-A1XF
TCGA-B6-A0X4
TCGA-D8-A1JS
TCGA-BH-A0HN
TCGA-C8-A133
TCGA-BH-A0HL
TCGA-B6-A0IC
TCGA-BH-A0HW
TCGA-A8-A06N
TCGA-EW-A1OX
TCGA-BH-A0HW
TCGA-A8-A06Z
TCGA-BH-A0HN
TCGA-A2-A0J3
TCGA-C8-A12Q
TCGA-D8-A1Y3
TCGA-BH-A0FE
TCGA-A8-A09Z
TCGA-A8-A07O
TCGA-B6-A0WX
TCGA-E2-A1B5
TCGA-BH-A0BF
TCGA-A2-A0CL
TCGA-C8-A12V
TCGA-B6-A1B6
TCGA-E2-A1B5
TCGA-B6-A0RT
TCGA-A2-A08H
TCGA-A2-A25F
TCGA-AR-A1AH
TCGA-BH-A0B9
TCGA-A2-A0SO
TCGA-AR-A0TN
TCGA-A1J6

Supplementary Figure 4, Related to Figure 1. PCA plots of TCGA BRCA dataset without and with zero-FPKM gene filtering (removal of genes with ≥50% zero values), outlier removal (using WGCNA sample connectivity>3SD), and central tendency two-way table median polish transformation (batch site effect removal). Principle component analysis was performed on the TCGA BrCa log2(FPKM) data before (A) and after (B) these steps. Variance of the top 500 contributing genes within the dataset was captured by the limma R package plotMDS function in 2 dimensions, PC1 and PC2. Changes in the range of the y axis post filtering, outlier removal, and transformation were minimal, and batches represented by different colors in panels A and B are not tightly clustered, indicating variation due to batch effect was also minimal. The same plots with samples colored by BrCa subtype (Luminal A, turquoise; Luminal B, seagreen; HER2-enriched, orange; TNBC, darkred) are shown in panels C and D.

Transparent Methods

Data curation and normalization. RNA-Seq data for co-expression analysis in this study was obtained from The Cancer Genome Atlas (TCGA) BRCA project. Breast primary tumor samples were collected from patients with lobular and ductal breast carcinomas with informed consent and IRB approval (Koboldt et al., 2012). FPKM normalized transcript abundance for 777 primary tumors were downloaded from the TCGA portal on or before August 27, 2019. An overview of clinical trait breakdowns collected for 777 case samples is provided (**Supplementary Tables S1- S3**). Level 3 RNA-Seq data was used for this study, which is de-identified and publicly available through TCGA online. Corrected, augmented trait data for all TCGA cases (Liu et al., 2018) was obtained and used for more accurate RFS information.

For array data used in ROC analysis, we curated 1,234 case-sample array-based measurements of transcriptome expression of BrCa of all subtypes selected from meta-analysis studies of BrCa (Györffy and Schäfer, 2009) and TNBC (Karn et al., 2011) as a coherent dataset, made possible by all these arrays sharing a common probe set and design. All data are available as GEO datasets (listed with traits in **Supplementary Table S10**), and as published supplementary data already normalized. MAS5.0-normalized expression data from both meta-analysis studies (Györffy and Schäfer, 2009, Karn et al., 2011) was aligned to use a common three-step sample-wise normalization as described in Karn, *et al* (Karn et al., 2014) prior to merging of the data. 155 TNBC cases from this meta-analysis was used, and a total of 180 TNBC cases were represented. All cases were annotated with relapse occurrence and RFS as necessary traits. Tumor samples identified as TNBC for the TCGA and GEO datasets were confirmed as subtypes of TNBC using the web based TNBCtype tool described by Lehmann BD et al (Chen et al., 2012, Lehmann et al., 2011). Subclassification of TNBC samples can be found in (Supplementary Table S3 and S10).

Data normalization, site effect handling, and removal of outliers. TCGA BrCa RNA-seq FPKM data from multiple collection sites (N=777 baseline case samples with 64,483 gene-wise short read-based quantification) was curated, addressing potential technical, site-specific sources of variance and of noise before gene clustering for co-expression. Only transcripts that had less than 50 percent of values censored were retained. 31,338 genes remained. Site or batch effect handling, normalization, and quality assessment relied on a Tunable Approach for Median Polish of Ratio (TAMPOR). This R-based function implements a two-way table median polish algorithm (Tukey, 1977) and acts on a log-transformed ratio of specific sample signals to the central tendency of a set of case samples common to all batches, which may be biological or technical replicates, or mixed pools of all samples (Johnson et al., 2020).

To capture biological variance of TNBC vs. non-TNBC groups, and to preserve it through normalization, all samples (N =777) for the TCGA-BrCa cohort were processed jointly in the sample-gene transcript matrix. TAMPOR implements user choice (i.e., tunability) for which samples are considered in Eq. 1 to obtain ratio denominators representing central tendency within and across batches for gene- (row)-wise median normalization in both terms of the equation.

[Eq. 1] $$\frac{abundance}{median\ (ALL\ SAMPLEs)_{\ site}} * \frac{grand\ median}{\boldsymbol{median}\left(\{\ \frac{\boldsymbol{abundance}}{\boldsymbol{median\ (ALL\ SAMPLEs)}_{\ \boldsymbol{site}}}\ |\ \boldsymbol{all\ samples\ from\ site}\ \}\right)}$$

Equation 1 applies to each measurement of a given gene transcript across all samples, where the first term represents site-wise median-centered abundance of a specific sample, and the second term is a site-specific normalization factor comprised of the grand median of all site-specific multi-sample (n≥5) medians, divided by the site-specific multi-sample median ratio. Ratios are $\log_2$-transformed, each $\log_2$-transformed ratio is adjusted sample- (column-) wise to a median value of 0, via subtraction of the sample's median $\log_2$-transformed ratio for all transcripts. Ratios are then anti-logged and multiplied by the row-wise abundance mean, extracted at the beginning of the iteration. The process is repeated for a default of 250 iterations or until convergence achieving an absolute value of the Frobenius norm difference from the previous iteration $| (||A||_{F(n-1)} - ||A||_{F(n)}) | < 1.0x10^{-8}$.

TAMPOR-normalized relative gene abundances were $\log_2$-transformed, then checked for network connectivity outliers 3 or more standard deviations from the average z-transformed sample connectivity calculated using the WGCNA R package fundamentalNetworkConcepts function (Oldham et al., 2008). This process is repeated until no outliers are detected. We then used principle component analysis implemented in the limma package plotMDS function to visualize TAMPOR-mediated removal of batch effects after removal of outliers vs. the original data before median polish, with zeroes also censored by $\log_2$-transformation. The samples showed no apparent systematic site effect but did separate by BrCa subtype, particularly TNBC (**Supplementary Figure 4**).

WGCNA for co-expressed gene clustering. The weighted gene co-expression network analysis (WGCNA) R software package was used to assess gene co-expression profiles across breast tumor samples, with the TAMPOR-normalized FPKM abundance matrix as input, and module eigengenes as the key output, which are correlated gene co-expression patterns (Oldham et al., 2008, Langfelder and Horvath, 2012). Eigengenes representing each module are assessed for correlation to sample-specific disease-related and other traits of interest, thereby enabling downstream analyses to focus attention on disease-relevant modules. WGCNA identifies gene clusters by calculating a gene dissimilarity matrix (1-topology overlap matrix) across samples and clustering genes that have similar expression patterns within the patient cohort into modules. As described previously, (Ohandjo et al., 2019) the WGCNA package blockwiseModules function was used for network construction, with module detection features including calls to the WGCNA dynamic tree-cutting algorithm, cutreeHybrid (Oldham et al., 2008, Langfelder and Horvath, 2012). Parameters for the function were as follows: power=6, deepsplit=4, minModuleSize=100, mergeCutHeight=0.15, TOMDenom="mean", corType="bicor", networkType="signed", pamStage=TRUE, pamRespectsDendro=TRUE, maxBlockSize larger than the number of genes being clustered (40000), and reassignThresh=0.05. We used biweight midcorrelation (bicor), as opposed to Pearson correlation, to provide robust correlations with less weight given to outlier measures (Oldham et al., 2008, Langfelder and Horvath, 2012). RNA-seq data often has high dynamic range and therefore often exhibits high variance across samples, making bicor rho and associated Student *p* values ideal for summarizing correlation robustly, a pivotal feature of the analysis.

Similarity between WGCNA module eigengenes (the first principle component of co-expression in a module) was determined by assessing module eigengene relatedness. Eigengenes are the first principal component of each module, overweighting the most highly correlated gene profiles across case samples that contribute to a co-expression network. The WGCNA blockwiseModules function provided an output of eigengene values for each module, further correlated in pairs. The output from this analysis depicted a relatedness dendrogram indicating the relatedness of all modules. This attribute of WGCNA, in addition to module trait correlation, drew attention to the identification of modules of interest.

Differential expression. An unpaired, two-tailed, equal variance hypothesis t-test was conducted to compare differential expression of TNBC subtyped tumor biopsies (N=92) to the non-TNBC group comprised of Luminal A (N=226), Luminal B (N=118), and HER2-enriched (N=57) subtyped tumor biopsies. Tumor samples missing subtype information were excluded from differential expression analysis. The t test is robust to noisy or skewed RNA-seq data and well-accommodates the unequal sample sizes of each group. FDR adjustment was performed using the Benjamini-Hochberg method. Three comparisons were considered to determine differential expression in TNBC: TNBC vs. Luminal A, TNBC vs. Luminal B, and TNBC vs. HER2-enriched subtyped sample abundances. For inclusion in **Supplementary Table S6**, a threshold was set at FDR<0.0001 for each of the above comparisons, and DEGs were only reported if significant and consistent direction of change occurred in all three comparisons.

Gene ontology (GO) enrichment analysis. GO-Elite v.1.2.5, was used to perform GO enrichment analysis on the M12 and M2 modules to identify overall module enrichment of biological functions (Zambon et al., 2012, Young et al., 2010). Version 62 of the Ensembl database (comprised of pre-defined gene lists organized by biological process, molecular function, and cellular component) was selected for standard GO enrichment analysis. Fisher's exact test p, adjusted for false discovery, was used to determine overrepresentation or significant overlap between members of WGCNA modules of interest and pre-defined gene lists (Young et al., 2010). The reference background gene list was the subset of 31,338 genes with symbols in the final cleaned abundance matrix (23,241 symbols).

Survival and gene expression analysis. Survival analysis was performed across 773 individuals to determine M12- and M-2 like eigengene value association to PFI using R packages, survival and survival miner. Generation of survival plots with optimal high/low cutoff for each gene were performed using the cutpoint R package with default parameters and the cutpointr function; cutpointr(data=survTraits, thisME, PFI.1, method = maximize_metric, metric = sum_sens_spec). Additionally, association to RFS for select genes from selected modules, M12 and M2, was determined using KMplotter (Gyorffy et al., 2010). The tool assesses the relationship of 54,000 gene transcript levels individually to RFS in various cancers. Analysis using gene expression paired with patient survival was performed using the Pan Cancer Atlas BrCa dataset (n = 1,090). These data were curated by authors of KMplotter from the Genome Expression Omnibus (GEO), the European Genome-Phenome Archive (EGA), and TCGA. Generation of survival plots was performed with KMplot.com default parameters except for enabling of auto-selection of an optimal high/low cutoff for each gene. Patient samples for the eigengene value and single gene curves were split into two groups (high and low expressers) according to median and mean expression to analyze the prognostic value of a gene. A log-rank (Mantel-Cox) test was used to determine p-values and hazard ratio with 95% confidence intervals for both sets of KM analyses. Box plots of TCGA RNA-seq relative abundance in **Supplementary Figure 1** were graphed using the base R boxplot function applied to TAMPOR-normalized $\log_2$ relative abundance. Significance of TNBC vs. non-TNBC groupwise differences was tested with a Mann-Whitney U test.

Genome-scale integrated analysis of gene networks in tissues (GIANT). GIANT version 2.0 (Wong et al., 2018) was used to examine the breast tissue-specific interaction network of the top ten hubs of the M2- and M12-like modules and selected survival-associated genes. Module memberships of genes predicted to interact with the top hub genes and rankings of the hubs were referenced by kME for each gene transcript's assigned module, as listed in **Supplementary Table S5**. Targets and weights for the top ranked hub genes (kME = 0.6 or higher) were obtained via the Flatiron Institute adapted interface to the GIANT database.

ROC testing of combinatorial transcript ratios as predictors of survival. Sample-wise normalized array probe abundances were collapsed to gene-level measurements keeping probes with maximal variance. Nominated genes represented on the arrays were then combined in all 1:1, 2:2, or 3:3 combinations and unweighted ratios were calculated for each sample. RFS was used as a proxy for survival in general, informing generalized linear model fits of each set of ratio calculations (the predictor) to RFS outcome at five different survival times. Survival is irrespective of treatment received by patients as these records are not available. Follow-up duration for this cohort was 10 years for all BrCa subtypes and 7.5 years specifically for TNBC patients. Binomial variance family with default linkage were specified for the R glm function. Time-dependent binary RFS outcome and the ordered glm-fit predictor were then tested and plot by the roc function of the pROC R package. AUC was used to rank predictor ratios. The verification R package provided functions for calculation of AUC Mann-Whitney U-based p value and 95% CI based on the calculated ROC curves. A one-tailed test for significance of improvement in ROC curve pairs was performed using a bootstrap method (n=2,000 permutations) as implemented in the roc.test function.

Supplemental References

Chen, X., Li, J., Gray, W. H., Lehmann, B. D., Bauer, J. A., Shyr, Y. & Pietenpol, J. A. 2012.

TNBCtype: A Subtyping Tool for Triple-Negative Breast Cancer. *Cancer Inform,* 11**,** 147-56.

Gyorffy, B., Lanczky, A., Eklund, A., Denkert, C., Budczies, J., Li, Q. & Szallasi, Z. 2010. 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,* 123**,** 725-731.

Györffy, B. & Schäfer, R. 2009. Meta-analysis of gene expression profiles related to relapse-free survival in 1,079 breast cancer patients. *Breast Cancer Res Treat,* 118**,** 433-41.

Johnson, E. C. B., Dammer, E. B., Duong, D. M., Ping, L., Zhou, M., Yin, L., Higginbotham, L. A., Guajardo, A., White, B., Troncoso, J. C., et al. 2020. Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. *Nature Medicine,* 26**,** 769-780.

Karn, T., Pusztai, L., Holtrich, U., Iwamoto, T., Shiang, C. Y., Schmidt, M., Müller, V., Solbach, C., Gaetje, R., Hanker, L., et al. 2011. Homogeneous datasets of triple negative breast cancers enable the identification of novel prognostic and predictive signatures. *PloS one,* 6**,** e28403-e28403.

Karn, T., Rody, A., Müller, V., Schmidt, M., Becker, S., Holtrich, U. & Pusztai, L. 2014. Control of dataset bias in combined Affymetrix cohorts of triple negative breast cancer. *Genomics data,* 2**,** 354-356.

Koboldt, D. C., Fulton, R. S., Mclellan, M. D., Schmidt, H., Kalicki-Veizer, J., Mcmichael, J. F., Fulton, L. L., Dooling, D. J., Ding, L., Mardis, E. R., et al. 2012. Comprehensive molecular portraits of human breast tumours. *Nature,* 490**,** 61-70.

Langfelder, P. & Horvath, S. 2012. Fast R Functions for Robust Correlations and Hierarchical Clustering. *J Stat Softw,* 46.

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y. & Pietenpol, J. A. 2011. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of clinical investigation,* 121**,** 2750-2767.

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V., et al. 2018. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell,* 173**,** 400-416.e11.

Ohandjo, A. Q., Liu, Z., Dammer, E. B., Dill, C. D., Griffen, T. L., Carey, K. M., Hinton, D. E., Meller, R. & Lillard, J. W., Jr. 2019. Transcriptome Network Analysis Identifies CXCL13-CXCR5 Signaling Modules in the Prostate Tumor Immune Microenvironment. *Sci Rep,* 9**,** 14963.

Oldham, M. C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S. & Geschwind, D. H. 2008. Functional organization of the transcriptome in human brain. *Nature Neuroscience,* 11**,** 1271-1282.

Piazza, R., Ramazzotti, D., Spinelli, R., Pirola, A., De Sano, L., Ferrari, P., Magistroni, V., Cordani, N., Sharma, N. & Gambacorti-Passerini, C. 2017. OncoScore: a novel, Internet-based tool to assess the oncogenic potential of genes. *Scientific Reports,* 7**,** 46290.

Tukey, J. W. 1977. *Exploratory data analysis*, Reading, Mass.

Wong, A. K., Krishnan, A. & Troyanskaya, O. G. 2018. GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. *Nucleic Acids Research,* 46**,** W65-W70.

Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome biology,* 11**,** R14.

Zambon, A. C., Gaj, S., Ho, I., Hanspers, K., Vranizan, K., Evelo, C. T., Conklin, B. R., Pico, A. R. & Salomonis, N. 2012. GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics,* 28**,** 2209-10.