A Novel Prognostic Two-Gene Signature for Triple Negative Breast Cancer

Mansour A Alsaleem [1,2], Graham Ball [3], Michael S Toss [1], Sara Raafat [1], Mohammed Aleskandarany [1,4], Chitra Joseph [1], Angela Ogden [5], Shristi Bhattarai [5], Padmashree C G Rida [5], Francesca Khani [6], Melissa Davis [7], Olivier Elemento [8], Ritu Aneja [5], Ian O Ellis [1], Andrew Green [1], Nigel P Mongan [9,10] and Emad Rakha [1,4,11].

[1] Nottingham Breast Cancer Research Centre, Division of Cancer and Stem Cells, School of Medicine, University of Nottingham, Nottingham, UK.

[2] Faculty of Applied Medical Sciences, Onizah Community College, Qassim University, Qassim, Saudi Arabia.

[3] John van Geest Cancer Research Centre, Nottingham Trent University, Nottingham, UK

[4] Faculty of Medicine, Menoufyia University, Shebin El Kom, Egypt.

[5] Department of Biology, Georgia State University, Atlanta, GA, USA.

[6] Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, New York, NY, USA

[7] Department of Genetics, Franklin College of Arts and Sciences, University of Georgia, Athens, GA, USA.

[8] Institute for Computational Biomedicine, Department of Physiology and Biophysics, Weill Cornell Medicine of Cornell University, New York, NY, USA

[9] University of Nottingham Biodiscovery Institute, Faculty of Medicine and Health Sciences, School of Veterinary Medicine and Science, University of Nottingham, Nottingham, UK.

[10] Department of Pharmacology, Weill Cornell Medicine, New York, NY, USA.

[11] Department of Histopathology, Division of Cancer and Stem Cells, School of Medicine, The University of Nottingham and Nottingham University Hospitals NHS Trust, Nottingham City Hospital, Nottingham, UK.

Corresponding author:

Professor Emad Rakha

Department of Histopathology, Division of Cancer and Stem Cells, School of Medicine, The University of Nottingham and Nottingham University Hospitals NHS Trust, Nottingham City Hospital, Nottingham, NG5 1PB, UK

Email: Emad.Rakha@nottingham.ac.uk

Keywords: triple negative breast cancer, TNBC, prognostic gene signature, ANN, ACSM4, SPDYC, NGS

Running Title: Prognostic stratification of triple negative breast cancer

ABSTRACT

The absence of a robust risk stratification tool for triple negative breast cancer (TNBC) underlies imprecise and non-selective treatment of these patients with cytotoxic chemotherapy. This study aimed to interrogate transcriptomes of TNBC resected samples using next generation sequencing to identify novel biomarkers associated with disease outcomes. A subset of cases (n=112) from a large, well-characterized cohort of primary TNBC (n=333) were subjected to RNA-sequencing. Reads were aligned to the human reference genome (GRCH38.83) using the STAR aligner and gene expression quantified using HTSEQ. We identified genes associated with distant metastasis-free survival and breast cancer-specific survival by applying supervised artificial neural network analysis with gene selection to the RNA-sequencing data. The prognostic ability of these genes was validated using the Breast Cancer Gene-Expression Miner v4. 0 and Genotype 2 outcome datasets. Multivariate Cox regression analysis identified a prognostic gene signature that was independently associated with poor prognosis. Finally, we corroborated our results from the two-gene prognostic signature by their protein expression using immunohistochemistry. Artificial neural network identified two gene panels that strongly predicted distant metastasis-free survival and breast cancer-specific survival. Univariate Cox regression analysis of 21 genes common to both panels revealed that the expression level of eight genes was independently associated with poor prognosis ($p<0.05$). Adjusting for clinicopathological factors including patient's age, grade, nodal stage, tumor size, and lymphovascular invasion using multivariate Cox regression analysis yielded a two-gene prognostic signature (*ACSM4* and *SPDYC*) which was associated with poor prognosis ($p<0.05$) independent of other prognostic variables. We validated the protein expression of these two genes, and it was significantly associated with patient outcome in both independent and combined manner ($p<0.05$). Our study identifies a prognostic gene signature that can predict prognosis in TNBC patients and could potentially be used to guide the clinical management of TNBC patients.

BACKGROUND

Breast cancer (BC) is a heterogeneous disease with variations in morphological features, molecular profiles, and therapy responses [1]. Triple negative breast cancer (TNBC), defined by the absence of expression of Estrogen Receptor, Progesterone Receptor and Human Epidermal Growth Factor 2, comprises 15%-30% of BC, and presents considerable challenges with regard to clinical management due to lack of targeted therapies [2,3]. Moreover, TNBC often has an unfavorable prognosis with increased probability of early metastasis, disease recurrence, and shorter overall survival [4,5]. Although TNBC generally displays aggressive behavior, patient outcomes can vary considerably. Around 23% of early-diagnosed TNBC patients remain disease free for more than five years while death within five years of diagnosis is inevitable for almost all metastatic TNBC patients [6–8]. Therefore, the complexity, molecular variability, and unpredictability of TNBC behavior warrants further investigation [9]. The biological heterogeneity of TNBCs has provided an impetus to develop tools for prognostic stratification, however, there are inconsistent results owing to a small cohort of patients, gene expression datasets obtained from different gene expression platforms and the use of microarray versus quantitative reverse transcriptase polymerase chain reaction which also makes head-to-head comparison challenging [10,11].

Various multigene prognostic tests are available for estrogen receptor-positive tumors for patient risk stratification and to guide therapy choice, whereas in estrogen receptor -negative tumors, and specifically TNBC tumors with a higher proliferation rate, these multigene signatures provide no clinical value [12]. Lehmann et al, used gene expression profiles to classify TNBCs into six molecular subtypes: Basal-like 1 and 2, Mesenchymal, Mesenchymal Stem-like, Immunomodulatory, and Luminal Androgen Receptor [13]. Burstein et al proposed an alternative gene expression classification for TNBC categorizing the tumor into four TNBC molecular subtypes: Luminal Androgen Receptor, Mesenchymal, Basal like immune

suppressed, and Basal like immune activated [14]. However, distant metastasis-free survival (DMFS) analysis showed poor prognosis for TNBCs regardless of their molecular profile subtype [15]. Therefore, there is an urgent unmet need for clinically validated prognostic markers that can predict outcomes for TNBC patients [15].

Unbiased omics technologies, including Next Generation Sequencing (NGS), are expected to lead a paradigm shift for precision medicine from a pathological microscopy-based diagnosis to gene signature-based diagnosis, prognosis, and treatment approaches [16]. NGS enables transcriptomic profiling of TNBC and identification of genomic alterations such as copy number changes, insertions, deletions and mutations; consequently, studies exploring inter-tumor heterogeneity in different types of tumors are now possible [17,18].

For successful NGS analysis, clinical samples must be maintained in conditions that would allow for DNA and RNA preservation and subsequent extraction. At present, most clinical samples are processed and archived as formalin-fixed, paraffin-embedded (FFPE) tissue samples in which the DNA and RNA necessary for NGS analysis is often fragmented [19]. However, FFPE tissue samples, if processed and stored properly, have been shown to preserve sufficient DNA and RNA material for extraction for NGS analysis [20]. The present study utilizes NGS transcriptomic analysis of a large cohort of TNBC FFPE tissue samples and aims to identify a molecular prognostic signature predicting risk for poor outcomes in TNBC.

METHODS

Nottingham TNBC Cohort

A retrospective well-characterized series of primary invasive TNBC (n=333) samples obtained from patients presented to Nottingham City Hospital, UK between 1987 to 2006, was included in this study. Clinicopathological data, including patient age at diagnosis, tumor size, tumor grade, nodal stage, lymphovascular invasion (LVI), and Nottingham Prognostic Index were collected from patients' medical records. The mean patient age was 48 years (range 27-69) and tumor sizes in diameter at the time of presentation ranged from 0.25 – 8.00 cm (1.5-2.8 cm within the interquartile), with a mean tumor size of 2.2 cm. Patients received a combination of treatment options including: surgery, radiation and chemotherapy according to standard protocols [21]. Outcome data including BC-specific survival (BCSS) and distant metastasis-free survival (DMFS) were available and prospectively maintained. BCSS was defined as the time (in months) from the primary surgical treatment to the time of death from BC, while DMFS was defined as the duration (in months) from the time of primary surgery to the first occurrence of distant metastasis. Estrogen Receptor, Progesterone Receptor and Human Epidermal Growth Factor 2 status of primary tumors were determined at the time of primary diagnosis from full-face sections of resected tumors according to published guidelines [22], (See Supplementary (A) for full details).

Transcriptomic Analysis

RNA sequencing was performed on representative FFPE tissue of an in house TNBC cohort (n=112) which had also been assessed histopathologically for tumor burden, (See Supplementary (A) for full details). Artificial Neural Network (ANN) database mining approach was used to build a classifier using the RNA-sequence matrices and identify genes associated with disease outcomes (DMFS and BCSS). In ANN, learning rates and momentum

were set at 0.1 and 0.5, respectively [23]. Each tumor sample had 39,684 corresponding genes. The input codes were "0" if patients showed neither evidence of metastasis (DMFS) nor death from BC (BCSS) within five years, and "1" if metastasis or death due to BC was evident in the first five years after diagnosis. Although BCSS is the ultimate endpoint of cancer outcome, DMFS was chosen as an end point based on the high likelihood of TNBC patients being diagnosed with distant metastases within five years of diagnosis [8]. Prior to ANN testing, a Monte-Carlo cross validation procedure was applied to avoid data over-fitting and false discovery. Documentation of such approach has proven to outperform the commonly used leave-one-out cross validation [24]. The input data were randomly divided into three subsets; 60% for training, 20% for validation to ensure model performance during the training process, and 20% for blind testing of the original model [25]. Genes identification by the forward stepwise approach using ANN was performed as described previously [26]. Based upon the distribution of performance on aforementioned model, ANN generated two panels of genes, representing the top 1% of the RNA sequence matrices that significantly predicted DMFS and BCSS, respectively. Genes common to both the DMFS and BCSS panels were identified using the Venny 2.0 online tool [27]. Receiver operating characteristics curves were generated to assess the predictive value of the differentially expressed gene panel presenting the sensitivity and specificity of the tested model (Supplementary (B) Figure 1).

Pathway Analysis

The online publicly available web-based gene set analysis tool, Webgestalt, (http://www.webgestalt.org/option.php) was used to identify differentially regulated canonical pathways using the overrepresentation enrichment analysis (ORA). The pathway analysis was based on the top 200 ranked genes predicting DMFS and BCSS. The reference gene list was set to the "genome_protein_coding". The ratio of observed versus expected number of genes

in the category was recorded for each significant category using the enrichment ratio (R) scores using Panther pathway database [28].

Prognostic Gene Signature Score

In compliance with the Reporting Recommendations for Tumor Marker Prognostic Studies criteria (REMARK), the associations between the expression of genes in our 21-gene panel, common to both the DMFS and BCSS gene prediction panels identified by ANN, and DMFS or BCSS were evaluated both individually, as well as after adjusting for standard prognostic variables [29,30]. Thus, DMFS and BCSS probabilities were individually computed on our gene panel using Kaplan-Meier testing model. Additionally, multivariate Cox regression analysis was used to calculate the estimate effect size [i.e., Hazard ratio (HR), along with 95% confidence interval (CI)] of the genes that were statistically significant in univariate Kaplan-Meier testing model for both DMFS and BCSS, which included the genes and standard prognostic variables, regardless of the statistical significance of standard prognostic variables in univariate analysis. The genes which showed significant prognostic impact independently in multivariate Cox regression analysis were further examined in a combined multivariate Cox regression analysis to identify a signature with a minimum number of genes that showed the most significant association with DMFS and BCSS.

External Validation of Transcriptomic Data

For independent validation of the results, the prognostic value of the two-gene signature predictors of DMFS and BCSS were evaluated using the Breast Cancer Gene-Expression Miner v4. 0 (Bc-GenExMiner) database which includes RNA-sequence expression data from 4713 BC patients, including 254 TNBC patients [31]. These genes were also interrogated through the Genotype 2 outcome tool (http://www.g-2-o.com), a web-based server utilizing NGS and gene chip data of 6,697 breast cancer patients including 612 TNBC patients with

outcome data. Computed receiver operating characteristics values were used to generate the transcriptomic fingerprint for mutational status from The Cancer Genome Atlas RNA-sequence and NGS mutation data. The average expression of significant genes was designated as a metagene for a given genotype. By employing gene chip data, associations between the expression of the metagene and patient outcomes were computed by multivariate Cox regression and Kaplan-Meier survival analysis [32].

Immunohistochemistry (IHC)

Assessment of the protein expression of the identified two-gene prognostic signature was performed using rabbit anti-SPDYC (NBP1-80832, lot # R36476, Novous Biological, UK) and rabbit anti-ACSM4 (PA5-62082, lot # R59771, Thermofisher, UK) antibodies on tissue microarrays prepared for the IHC cohort, (See Supplementary (A) for full details).

Statistical Analysis

IBM SPSS 24.0 (Chicago, IL, USA) software was used for statistical analysis. For dichotomization of mRNA expression and protein expression levels of different genes, the X-tile bioinformatics version 3.6.1 (Yale University, USA) was utilized with DMFS as an endpoint. Cox proportional hazard models were used for multivariate analysis model adjusting for patients age, tumor grade, nodal stage, tumor size, and LVI status as covariates to adjust for potential confounding influence of these variables on associations between the tested genes and the outcomes of interest. Spearman's Rho test was used to evaluate correlations between continuous variables of the transcriptomic and protein expression data whereas the chi-square test was performed to analyze relationships between categorical variables. A p-value of <0.05 was deemed significant, (See Supplementary (A) for full details).

RESULTS

Gene Selection

To build a classifier panel for outcome prediction in TNBC, ANN analysis of the RNA-sequence matrices data of the transcriptomic cohort was performed and genes were ranked based on relationships between their expression and clinical outcomes in terms of DMFS and BCSS. The top ranked genes predicting DMFS (DMFS genes panel) and those predicting BCSS (BCSS genes panel) were investigated to determine the most statistically enriched pathways, (Supplementary (A) Table 2 & Supplementary (C) for full details).

Using the Venny tool, we identified a total of 21 genes that were common to both the DMFS and BCSS ANN panels. The 21-gene panel predicted patients' DMFS and BCSS with 92% sensitivity and 94% specificity (Supplementary (B) Figure 2). The probability of finding a gene by random chance in the top 200 was 0.03, whereas the probability of randomly finding the 21 genes collectively was $6.2x10^{-33}$, (Supplementary (B) Figure 3).

Univariate Kaplan–Meier survival analysis showed that elevated expression of some genes was significantly associated with shorter DMFS and BCSS, whereas elevated expressions of other genes showed statistically significant association with longer DMFS and BCSS (Supplementary (A) Table 3 & Supplementary (B) Figures 4 A-D). Multivariate Cox regression analysis models incorporating patient's age, tumor grade, nodal stage, tumor size, and LVI status revealed that eight of the 21 genes were independent predictors of DMFS and BCSS, (Supplementary (A) Table 4 A-D).

Prognostic Two-Gene Signature

The prognostic gene signature was identified after statistically distilling the eight genes in a multivariate Cox regression analysis to identify a signature with a minimum number of genes

that show most significant association with BCSS and DMFS. The analysis revealed two genes *ACSM4* and *SPDYC* that most significantly and independently predicted both DMFS and BCSS (*ACSM4;* DMFS: p=0.015, 95% CI=1.21-6.13, HR=2.72 : BCSS: p=0.004, 95% CI=1.44-6.83, HR=3.14), and (*SPDYC* ; DMFS: p=0.012, 95% CI=1.23-5.45, HR=2.59 :BCSS: p=0.016, 95% CI=1.18-5.09, HR=2.45) (Supplementary (A) Table 5). There was no linear association between the mRNA expression of *ACSM4* and *SPDYC.* To investigate the prognostic value of the two-gene signature, a linear prognostic score was generated using the sum of the product of normalized expression levels of these two genes and their respective regression coefficients, as follows:

The prognostic two-gene signature score ∑= (*ACSM4* normalized expression * *ACSM4* expression β-value) + (*SPDYC* normalized expression * *SPDYC* expression β-value) (Table1).

Using X-tile cut-off generator, patients with higher mRNA expression score of the prognostic two-gene signature had worse outcome in terms of shorter DMFS and BCSS when compared with those with lower mRNA expression score (Figure 1). Cox regression analysis confirmed that the prognostic two-gene signature harbors significant prognostic value in terms of predicting shorter DMFS and BCSS independent of patient age, tumor grade, nodal stage, tumor size, and LVI status (Table 2).

External Validation of Genomic Findings

Using the Bc-GenExMiner tool to analyze publicly available RNA-sequencing data, we observed that higher expression of *SPDYC* was significantly associated with worse prognosis in the whole/unselective cohorts of BC (n=4308, p<0.0001) [31]. Validating genes expressions on the restricted TNBC cohort (n=254), revealed a similar trend of poor prognosis (p=0.006) [31]. Moreover, the integration of our proposed prognostic two-gene signature in the public domain Genotype 2 outcome, using the median of each gene expression in the

whole/unselective cohorts of BC (n=4029), indicated that higher expression of *ACSM4* and *SPDYC* were associated with worse prognosis (both p<0.001). More importantly in the context of this study, the prognostic value of the two-gene signature (*ACSM4 and SPDYC*) were significantly associated with poorer outcome when examined in the TNBC subtype cohort alone (n=612, p<0.001) [32] (Figure 2).

Immunohistochemistry of the Prognostic Two-Gene Signature

The morphological assessment of the tissue samples revealed cytoplasmic expression for both proteins; ACSM4 (H-score range 5-295) and SPDYC (H-score range 5-290) (Supplementary (B) Figure 5).

Univariate survival analysis revealed that higher expression of ACSM4 and SPDYC was significantly associated with patients' poor outcomes (DMFS; p<0.001, BCSS; p=0.009 for ACSM4) and (DMFS and BCSS, both p=0.004 for SPDYC) (Figure 3), which is concordant with the findings obtained from transcriptomic data.

Multivariate Cox regression analysis showed that SPDYC protein expression was an independent prognostic factor regardless of patient age, tumor grade, nodal stage, tumor size, and LVI status for DMFS (p=0.015, 95% CI =1.17 - 4.74, HR=2.365) and BCSS (p=0.015, 95% CI =1.18- 4.78, HR=2.377). Likewise, multivariate Cox regression analysis showed that ACSM4 protein expression was a significant independent prognostic factor for DMFS (p=0.002, 95% CI=1.35- 3.89, HR= 2.267), but not in BCSS (p=0.057, 95% CI=0.98- 2.93 , HR= 1.698) (Table 3 A & B).

In a combined multivariate Cox regression analysis, SPDYC protein expression was an independent prognostic factor that predicted shorter DMFS and BCSS (DMFS: p=0.03, 95% CI=1.07-5.86, HR=2.50: BCSS: p=0.03, 95% CI=1.08-5.96 HR=2.54), regardless of patient age, tumor grade, nodal stage, tumor size, and LVI status. ACSM4 protein expression also was observed to be an independent prognostic factor, associated with shorter DMFS (p=0.003, 95%

CI =1.01-3.20, HR=1.83), regardless of patient age, tumor grade, nodal stage, tumor size, and LVI status, but not with BCSS (p=0.27, 95% CI=0.76-2.56 , HR=1.40) (Table 4). Correspondingly, we observed a significant positive linear association between ACSM4 and SPDYC protein expression (r=0.29, p<0.001), signifying that these proteins might be synergistically driving TNBC disease progression (Figure 4). Furthermore, using only cases that were informative for both biomarkers, a linear prognostic score was generated using Cox proportional hazard analysis to test whether dual expression of SPDYC and ACSM4 proteins was associated with worse outcome. The equation generated used the sum of the product of the quantitative H-score and their respective regression coefficient as follows:

Protein expression prognostic score: $\sum$= (ACSM4 H-score * ACSM4 H-score β value) + (SPDYC H-score * SPDYC H-score β value) (Table 5).

This protein expression prognostic score was then dichotomized using X-tile software to determine the optimal score to classify patients into high and low risk groups using DMFS as an end point. In the 257 investigated cases, the scores ranged from 15.43-365.05 with high protein expression risk scores (score > 170) observed in 159/257 (62%) cases.

When testing the association between the prognostic score and outcome, univariate analysis demonstrated that cases with higher protein expression score had a significantly shorter DMFS (p=0.02) but not BCSS (p=0.06) (Figure 5). Multivariate Cox regression analysis model demonstrated that protein expression prognostic score was an independent prognostic factor for DMFS (p=0.03, 95% CI=1.04- 3.32 , HR=1.83) independent of patient age, tumor grade, nodal stage, tumor size, and LVI status, but not for BCSS (p=0.07, 95% CI=0.94-2.96, HR=1.83) (Table 6).

Finally, when we stratified our cohort based on chemotherapy treatment, the 10-year DMFS of patients who were not offered chemotherapy (n=83) and showed low expression of ACSM4 was 84% compared to 44% of those with high expression and the difference was statistically

significant (p=0.005). However, those with low expression of SPDYC had 83% 10-year DMFS compared to 70% in those with high expression but the difference was not statistically significant (p=0.209). Similarly, with the prognostic two gene signature, the 10-year DMFS of patients with low expression was 84% compared to 69% of those with high expression (p=0.309).

Testing the performance of the prognostic two-gene at the transcriptomic and protein Levels

The prognostic signature at the mRNA level captured 58% sensitivity, 69% specificity, 54% positive predictive value, 72% negative predictive value, and 64% accuracy in dichotomizing distant metastasis outcome of TNBC patients. In comparison, the prognostic signature at the protein level showed 73% sensitivity, 42% specificity, 30% positive predictive value, 82% negative predictive value, and 50% accuracy in dichotomizing distant metastasis outcome of TNBC patients (Supplementary (A) Table 6).

DISCUSSION

Molecular classification of BC provides opportunities for enhanced personalized therapy [33]. In TNBC, conventional prognostic factors such as age, tumor size, tumor grade, and lymph node status have limited risk-predictive influence as these tumors are mostly of higher grade with increased chances of recurrence and metastasis [1]. Therefore, deciphering genomic profiles of TNBC using advanced techniques is an unmet need. Moreover, the utilization of ANN to mine the transcriptomic profile of TNBC in order to identify genes associated with clinical outcome is a promising approach to stratify patients for risk prediction [34].

In the current study, a discovery phase and two validation phases were implemented. The in-house transcriptomic TNBC cohort was used for the discovery phase for ANN analysis. Whereas the protein expression and publicly available external transcriptomic BC data were used for the validation phases of findings. More importantly, regardless of the statistical differences in the distribution of clinicopathological parameters between transcriptomic and IHC cohorts, our gene signature showed statistical association with outcome both at transcriptomic and protein expression level. Our study supports the utility of applying ANN to integrate distinct clinical and molecular data to find novel prognostic biomarkers associated with TNBC poor outcome.

Our study employed ANN for the analysis of our transcriptomic cohort to discover novel prognostic genes associated with outcome in TNBC. ANN is a powerful tool for the analysis of complex data, overcoming high background noise, and thus identifying the influence of many interacting factors [35]. ANN analysis, unlike conventional statistical approaches such as hierarchical clustering, linear regression, and principal component analysis, is not limited by linear functionality; thus, identification of biological relationships between biomarkers and clinical outcomes is improved [24]. Furthermore, unlike conventional statistical techniques

used in the medical diagnostic and prognostic approaches, ANN can produce greater accuracy model than its counterparts [36]. Therefore, it is highly suitable for the identification of potential key genes driving TNBC outcomes. ANN modelling uses a supervised learning approach, a multi-layer perception architecture with a sigmoid transfer function, where weights are updated by a back propagation algorithm [37].

In this study, ANN analysis identified the top ranked genes predicting DMFS and BCSS. We then employed a web-based tool to identify the signaling pathways significantly enriched in the significant top ranked gene panels. For instance, TNBC patients frequently harbor higher expression of the epidermal growth factor receptor EGFR; however, studies have failed to establish significant benefit from EGFR-targeted therapies or tyrosine kinase inhibitors, suggesting the need to therapeutically target other pathways in these tumors [38,39]. Moreover, the significance and over-activation of pathways such as; P38 MAPK , the PDGF, and the RAS pathways in BC metastatic sites and their association with DMFS and BCSS in TNBC have been previously documented [40–42]. Additionally, the 21 gene panel generated by ANN analysis that was strongly associated with both DMFS and BCSS in TNBC included several novel and potentially targetable biomarkers in TNBC outcome. For instance, higher expression of *DOCK10* (also known as dedicator of cytokeratin-10/ZIZ3) [43], has been previously identified as an indicator of poor prognosis in TNBC patients and as a predictor of distant metastasis [44]. In our transcriptomic cohort, *DOCK10* emerged as a significant prognostic marker of BCSS and DMFS however, it was not significantly prognostic in multivariate Cox regression analysis. We also found that high expression of *BICC1*, an RNA binding protein, a negative regulator of the WNT signaling pathway with potential involvement in regulating gene expression during embryonic development [45], was associated with DMFS but not with BCSS; thus, it was not included in the final signature.

In our study, we distilled the initial 21 gene panel down to eight genes that when tested individually for their prognostic value, were significantly associated with both DMFS and BCSS using univariate and multivariate analysis after adjusting for the potentially confounding variables. These genes are implicated in pro-oncogenic pathways in BC. *PPL* (also known as Periplakin) is a part of the cornified envelop in keratinocytes and desmosomes with intermediate filaments. *PPL* can act in the PKB/AKT-mediated signaling pathway [46]. In TNBC, silencing *PPL* decreased cell migration and invasion [47]. *SPDYC* is a member of the speedy/Ringo cyclin-dependent kinase (CDK) family with known functions in cell cycle transitions and progression [48]. *SPDYC* plays an important role in activating both *CDK1* and *CDK2* expression [49]. CDK2 high expression has been previously described to be associated with shorter survival in metastatic melanoma cases and endocrine resistance in SKBR3-HER2 positive BC cell lines [50,51]. Furthermore, down regulation of *CDK1* has been found to increase synthetic lethality of TNBC cell lines if accompanied with *c-Myc* high expression [52]. However, *SPDYC* role in BC is still undefined [48]. *ACSM4* encodes a protein with known functions in the conjugation of carboxylic acids and in fatty acid beta oxidation. Interestingly, upregulation of metabolic pathways has been found to interact with cellular transcriptomic and proteomics of both CD4 and CD8 T cells in HIV disease [53]. Although *ACSM4* has been shown to have a role in AIDS progression, there are no reports with its role in BC [54,55]. We have previously reported a strong correlation between tumor infiltrating lymphocytes and TNBC outcome [56]. However, our current analysis did not identify known inflammation and immune response related genes associated with outcome in the TNBC 21 gene panel. Future studies should therefore seek to identify novel mechanisms contributing to aberrant inflammatory and immune response pathways involved in tumor infiltrating lymphocytes in TNBC. Furthermore, genes such as *AC020931.1*, *DCTN1-AS1, RP11-29H23.5, PAXBP1-AS1,*

and *RPS10P18* require further investigation to decipher their role and function in BC progression.

The original hypothesis underpinning this study was that a gene expression signature would more accurately predict both DMFS and BCSS in TNBC than a single gene. Multivariate Cox regression analysis enabled us to further filter the set of eight genes to a prognostic two-gene signature (*ACSM4* and *SPDYC*) showing strong association with both DMFS and BCSS. We tested whether immunohistochemical assessment of the protein expression of the ACSM4 and SPDYC genes could be used to predict patient outcomes. Our study confirmed that protein expression had independent prognostic significance in TNBCs and showed strong statistical association with worse outcomes (i.e., shorter DMFS and BCSS). These genes when combined in a linear score, successfully stratified TNBC patients into high- and low-risk subgroups; in the former group, which is at a higher risk of developing distant metastasis, could benefit from greater vigilance and more aggressive treatment regimens. We have validated our ANN investigation and RNA-sequencing results by studying protein expression which showed that a prognostic score derived from the immunohistochemical evaluation of the two biomarkers could significantly predict distant metastasis, and thus support personalized prognostic evaluation and guiding treatment choices to improve disease outcomes.

In this study, the prognostic value of the two-gene signature at the mRNA level yielded 58% sensitivity, and 64% accuracy in dichotomizing distant metastasis outcome of TNBC patients. By contrast, at the protein level, our proposed two-gene signature demonstrated 73% sensitivity, and 50% accuracy in dichotomizing distant metastasis outcome of TNBC patients. Our proposed two-gene signature showed promising accuracy and sensitivity results in predicting the risk of distant metastasis in TNBC patients, which is even more important as presently TNBC patients solely rely on chemotherapy treatment. Moreover, those patients who

are deemed at high risk of distant metastasis may benefit from the stratification for an improved treatment decision.

Furthermore, our proposed two-gene signature is only based on two genes (*ACSM4* and *SPDYC*), unlike other commercially available prognostic assays including those designed for ER-positive tumors [57]. Our prognostic gene signature may be amenable to the development of affordable molecular tests based on quantitative reverse transcriptase polymerase chain reaction as the sensitivity, specificity, and accuracy of our two-gene signature is proved to be much stronger at the mRNA level. The prognostic gene signature might be suitable for use in routine clinical practice because the proposed two-gene signature has prognostic value in dichotomizing TNBC patients and may provide important information for treatment decisions.

The mainstay of TNBC treatment is cytotoxic chemotherapy [58]. However, chemotherapy decision for metastatic TNBC patients are given based on a combination of aspects relates to the disease and patient physical characteristics (i.e., tumor burden, patient age, co-morbidities, prior treatments received in the adjuvant setting, and patient preference) [59]. Despite the interesting finding of this study and the significant difference in the survival of patients who were not offered chemotherapy based on the expression of ACSM4 (with worse outcome of patients with high expression), the 10-year DMFS of patients with low expression (84%) may not justify recommendation for omission of chemotherapy in those patients. However, to make such a recommendation, a clinical trial utilizing a sufficiently large number of TNBC patients may be warranted to determine whether TNBC patients with low ACSM4 expression can avoid chemotherapy without worse outcome.

A challenge of applying the NGS technique to deciphering the molecular characteristics of TNBC tumors includes access to the technology and the integrity of tumor samples to guarantee sufficient tumor RNA extraction [60]. Variation in sample quality and preparation

may negatively influence the outputs of NGS analysis and therefore must be carefully controlled. In addition, NGS analysis must consider intrinsic tumor heterogeneity between patients. Samples used in this study were processed in a strictly standardized procedure implemented in Nottingham University Hospitals with immediate sample fixation following surgery, with standard protocols optimized to preserve tissue architecture, subcellular details and importantly the integrity of biologic materials including proteins, DNA, and RNA. Nonetheless, our retrospective study was limited to a single center using an in-house transcriptomic and protein expression cohort for this investigation. However, the public domain data used in this study supports the value of both *ACMS4* and *SPDYC* high expression conferring poor prognosis for BC patients, especially those diagnosed with TNBC molecular subtype. Hence, further external validation is strongly recommended.

Conclusion

Personalized medicine seeks to stratify BC patients to ensure optimal treatment and thus, improved patient outcomes. Our study has identified a two-gene signature that stratifies TNBC patients into high and low risk groups for developing distant metastasis, which can potentially guide clinical decision-making. The robust methods used herein to identify our prognostic gene signature followed by validation of the findings at the protein expression level, suggest that this promising two-gene signature provides avenues for further *in vitro* functional investigation and for new drug development for TNBC patients who are in dire need of effective therapeutic options.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The RNA sequencing was supported by institutional funds to RA from Georgia State University. The IHC was supported by the PhD scholarship funded by the Saudi Arabia Ministry of Education Qassim University.

Conflict of Interest

The authors have no conflicts of interest to declare.

Ethical Approval and Consent to Participate

This work obtained ethics approval by the North West – Greater Manchester Central Research Ethics Committee under the title; Nottingham Health Science Biobank (NHSB), reference number 15/NW/0685. Informed consent was obtained from all individuals prior to surgery to use their tissue materials in research. All samples used in this study were pseudo-anonymized and collected prior to 2006 and stored in compliance with the UK Human Tissue Act.

Availability of Data and Materials

The authors confirm the data that has been used in this work is available on reasonable request.

Authors' Contributions

MAA, MA, and GB participated in its conception, design, experimentation, analysis, interpretation, and manuscript drafting. MAA and MS conducted the immunohistochemical studies and participated in the analysis and interpretation. MS and MA helped with pathology review and manuscript drafting. SR, and CJ helped in immune-histochemical analysis and interpretation. AO, SB, PR, FK, MD, OE, RA, IE, AG and NM participated in interpretation and manuscript drafting. ER conceived and supervised the study, participated in its design,

interpretation, and analysis, including drafting. All authors contributed to drafting and reviewing the manuscript and approved the submitted and final version.

Acknowledgements

Mansour A Alsaleem is supported and funded by Qassim University, Kingdom of Saudi Arabia. We express thanks to Innovate UK for funding (ISCF bid Ref 18181), and the Nottingham Health Science Biobank and BC Now Tissue Bank for the provision of tissue samples.

References:

1 Rakha EA, Chan S. Metastatic triple-negative breast cancer. Clin Oncol (R Coll Radiol) 2011;23:587–600.

2 Ahn SG, Kim SJ, Kim C, Jeong J. Molecular Classification of Triple-Negative Breast Cancer. J Breast Cancer 2016;19:223.

3 Liedtke C, Bernemann C, Kiesel L, Rody A. Genomic profiling in triple-negative breast cancer. Breast Care (Basel) 2013;8:408–413.

4 Khalifeh IM, Albarracin C, Diaz LK, Symmans FW, Edgerton ME, Hwang RF, et al. Clinical, Histopathologic, and Immunohistochemical Features of Microglandular Adenosis and Transition Into In Situ and Invasive Carcinoma. Am J Surg Pathol 2008;32:544–552.

5 Stead LA, Lash TL, Sobieraj JE, Chi DD, Westrup JL, Charlot M, et al. Triple-negative breast cancers are increased in black women regardless of age or body mass index. Breast Cancer Res 2009;11:R18.

6 Haffty BG, Yang Q, Reiss M, Kearney T, Higgins SA, Weidhaas J, et al. Locoregional Relapse and Distant Metastasis in Conservatively Managed Triple Negative Early-Stage Breast Cancer. J Clin Oncol 2006;24:5652–5657.

7 Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, Sawka CA, et al. Triple-Negative Breast Cancer: Clinical Features and Patterns of Recurrence. Clin Cancer Res 2007;13:4429–4434.

8 Foulkes WD, Smith IE, Reis-Filho JS. Triple-Negative Breast Cancer. N Engl J Med 2010;363:1938–1948.

9 Yam C, Mani SA, Moulder SL. Targeting the Molecular Subtypes of Triple Negative Breast Cancer: Understanding the Diversity to Progress the Field. Oncologist 2017;:theoncologist.2017-0095.

10 Alizadeh AA, Ross DT, Perou CM, van de Rijn M. Towards a novel classification of human malignancies based on gene expression patterns. J Pathol 2001;195:41–52.

11 Katagiri T, Yoshimaru T, Matsuo T, Kiyotani K, Miyoshi Y, Tanahashi T, et al. Molecular features of triple negative breast cancer cells by genome-wide gene expression profiling analysis. Int J Oncol 2012;42:478–506.

12 Győrffy B, Hatzis C, Sanft T, Hofstatter E, Aktas B, Pusztai L. Multigene prognostic tests in breast cancer: past, present, future. Breast Cancer Res 2015;17:11.

13 Lehmann BD, Jovanović B, Chen X, Estrada M V., Johnson KN, Shyr Y, et al. Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. PLoS One 2016;11:e0157368.

14 Burstein MD, Tsimelzon A, Poage GM, Covington KR, Contreras A, Fuqua SAW, et al. Comprehensive Genomic Analysis Identifies Novel Subtypes and Targets of Triple-Negative Breast Cancer. Clin Cancer Res 2015;21:1688–1698.

15 Ménard S. Heterogeneity of triple-negative breast carcinomas. Oncologie 2012;14:28–30.

16 Nagahashi M, Wakai T, Shimada Y, Ichikawa H, Kameyama H, Kobayashi T, et al. Genomic landscape of colorectal cancer in Japan: clinical implications of comprehensive genomic sequencing for precision medicine. Genome Med 2016;8:136.

17 Lips EH, Michaut M, Hoogstraat M, Mulder L, Besselink NJM, Koudijs MJ, et al.

Next generation sequencing of triple negative breast cancer to find predictors for chemotherapy response. Breast Cancer Res 2015;17:134.

18 Desmedt C, Voet T, Sotiriou C, Campbell PJ. Next-generation sequencing in breast cancer: first take home messages. Curr Opin Oncol 2012;24:597–604.

19 Endrullat C, Glökler J, Franke P, Frohme M. Standardization and quality management in next-generation sequencing. Appl Transl Genomics 2016;10:2–9.

20 McDonough SJ, Bhagwate A, Sun Z, Wang C, Zschunke M, Gorman JA, et al. Use of FFPE-derived DNA in next generation sequencing: DNA extraction methods. PLoS One 2019;14:e0211400.

21 Wahba HA, El-Hadaad HA. Current approaches in treatment of triple-negative breast cancer. Cancer Biol Med 2015;12:106–116.

22 Muftah AA, Aleskandarany MA, Al-Kaabi MM, Sonbul SN, Diez-Rodriguez M, Nolan CC, et al. Ki67 expression in invasive breast cancer: the use of tissue microarrays compared with whole tissue sections. Breast Cancer Res Treat 2017;164:341–348.

23 Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323:533–536.

24 Lancashire LJ, Powe DG, Reis-Filho JS, Rakha E, Lemetre C, Weigelt B, et al. A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks. Breast Cancer Res Treat 2010;120:83–93.

25 Picard RR, Cook RD. Cross-Validation of Regression Models. J Am Stat Assoc 1984;79:575–583.

26 Xu Q-S, Liang Y-Z, Du Y-P. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. J Chemom 2004;18:112–120.

27 Oliveros JC. VENNY. An interactive tool for comparing lists with Venn Diagrams [Internet]. [cited 20 June 2019]. Available from: http://bioinfogp.cnb.csic.es/tools/venny/.

28 Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res 2017;45:W130–W137.

29 Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): Explanation and Elaboration. PLoS Med 2012;9:e1001216.

30 McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM, et al. REporting recommendations for tumour MARKer prognostic studies (REMARK). Br J Cancer 2005;93:387–391.

31 Jézéquel P, Campone M, Gouraud W, Guérin-Charbonnel C, Leux C, Ricolleau G, et al. Bc-GenExMiner: An easy-to-use online platform for gene prognostic analyses in breast cancer. Breast Cancer Res Treat 2012;131:765–775.

32 Pongor L, Kormos M, Hatzis C, Pusztai L, Szabó A, Gyorffy B. A genome-wide approach to link genotype to clinical outcome by utilizing next generation sequencing and gene chip data of 6,697 breast cancer patients. Genome Med 2015;7:104.

33 van de Vijver MJ, He YD, van ’t Veer LJ, Dai H, Hart AAM, Voskuil DW, et al. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. N Engl J Med

2002;347:1999–2009.

34 Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkänen L, Joensuu H. Artificial neural networks applied to survival prediction in breast cancer. Oncology 1999;57:281–286.

35 Ball G, Mian S, Holding F, Allibone RO, Lowe J, Ali S, et al. An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. Bioinformatics 2002;18:395–404.

36 Tafeit E, Reibnegger G. Artificial Neural Networks in Laboratory Medicine and Medical Outcome Prediction. Clin Chem Lab Med 1999;37:845–853.

37 Abdel-Fatah TMA, Agarwal D, Liu D-X, Russell R, Rueda OM, Liu K, et al. SPAG5 as a prognostic biomarker and chemotherapy sensitivity predictor in breast cancer: a retrospective, integrated genomic, transcriptomic, and protein analysis. Lancet Oncol 2016;17:1004–1018.

38 Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest 2011;121:2750–2767.

39 Costa R, Shah AN, Santa-Maria CA, Cruz MR, Mahalingam D, Carneiro BA, et al. Targeting Epidermal Growth Factor Receptor in triple negative breast cancer: New discoveries and practical insights for drug development. Cancer Treat Rev 2017;53:111–119.

40 Forte L, Turdo F, Ghirelli C, Aiello P, Casalini P, Iorio MV, et al. The PDGFRβ/ERK1/2 pathway regulates CDCP1 expression in triple-negative breast

cancer. BMC Cancer 2018;18:586.

41 Fan Y, Li M, Ma K, Hu Y, Jing J, Shi Y, et al. Dual-target MDM2/MDMX inhibitor increases the sensitization of doxorubicin and inhibits migration and invasion abilities of triple-negative breast cancer cells through activation of TAB1/TAK1/p38 MAPK pathway. Cancer Biol Ther 2019;20:617–632.

42 Adeyinka A, Nui Y, Cherlet T, Snell L, Watson PH, Murphy LC. Activated mitogen-activated protein kinase expression during human breast tumorigenesis and breast cancer progression. Clin Cancer Res 2002;8:1747–1753.

43 Ruiz-Lafuente N, Alcaraz-García M-J, García-Serna A-M, Sebastián-Ruiz S, Moya-Quiles M-R, García-Alonso A-M, et al. Dock10, a Cdc42 and Rac1 GEF, induces loss of elongation, filopodia, and ruffles in cervical cancer epithelial HeLa cells. Biol Open 2015;4:627–635.

44 Westcott JM, Prechtl AM, Maine EA, Dang TT, Esparza MA, Sun H, et al. An epigenetically distinct breast cancer cell subpopulation promotes collective invasion. J Clin Invest 2015;125:1927–1943.

45 Kraus MR-C, Clauin S, Pfister Y, Di Maïo M, Ulinski T, Constam D, et al. Two mutations in human BICC1 resulting in Wnt pathway hyperactivity associated with cystic renal dysplasia. Hum Mutat 2012;33:86–90.

46 Ruhrberg C, Hajibagheri MA, Parry DA, Watt FM. Periplakin, a novel component of cornified envelopes and desmosomes that belongs to the plakin family and forms complexes with envoplakin. J Cell Biol 1997;139:1835–1849.

47 Choi YK, Woo S-M, Cho S-G, Moon HE, Yun YJ, Kim JW, et al. Brain-metastatic triple-negative breast cancer cells regain growth ability by altering gene expression

patterns. Cancer Genomics Proteomics;10:265–275.

48 Cheng A, Solomon MJ. Speedy/Ringo C regulates S and G $_2$ phase progression in human cells. Cell Cycle 2008;7:3037–3047.

49 Mourón S, De Cárcer G, Seco E, Fernández-Miranda G, Malumbres M, Nebreda AR. RINGO C is required to sustain the spindle-assembly checkpoint. J Cell Sci 2010;123:2586–2595.

50 Bogunovic D, O'Neill DW, Belitskaya-Levy I, Vacic V, Yu Y Lo, Adams S, et al. Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival. Proc Natl Acad Sci U S A 2009;106:20429–20434.

51 Karavasilis V, Reid A, Sinha R, de Bono JS. Cancer drug resistance. In: Cancer Drug Design and Discovery. : Elsevier Inc., 2008. p. 405–423.

52 Liu Y, Zhu YH, Mao CQ, Dou S, Shen S, Tan Z Bin, et al. Triple negative breast cancer therapy with CDK1 siRNA delivered by cationic lipid assisted PEG-PLA nanoparticles. J Control Release 2014;192:114–121.

53 Wu JQ, Dwyer DE, Dyer WB, Yang YH, Wang B, Saksena NK. Genome-wide analysis of primary CD4+ and CD8+ T cell transcriptomes shows evidence for a network of enriched pathways associated with HIV disease. Retrovirology 2011;8:18.

54 Guzmán-Fulgencio M, Jiménez JL, Jiménez-Sousa MA, Bellón JM, García-Álvarez M, Soriano V, et al. ACSM4 polymorphisms are associated with rapid AIDS progression in HIV-infected patients. J Acquir Immune Defic Syndr 2014;65:27–32.

55 Hendrickson SL, Lautenberger JA, Chinn LW, Malasky M, Sezgin E, Kingsley LA, et

al. Genetic variants in nuclear-encoded mitochondrial genes influence AIDS progression. PLoS One 2010;5:e12862.

56 Althobiti M, Aleskandarany MA, Joseph C, Toss M, Mongan N, Diez-Rodriguez M, et al. Heterogeneity of tumour-infiltrating lymphocytes in breast cancer and its prognostic significance. Histopathology 2018;73:887–896.

57 Gyanchandani R, Lin Y, Lin H-M, Cooper K, Normolle DP, Brufsky A, et al. Intratumor Heterogeneity Affects Gene Expression Profile Test Prognostic Risk Stratification in Early Breast Cancer. Clin Cancer Res 2016;22:5362–5369.

58 Isakoff SJ. Triple-negative breast cancer: Role of specific chemotherapy agents. Cancer J. 2010;16:53–61.

59 Biganzoli L, Cufer T, Bruning P, Coleman R, Duchateau L, Calvert AH, et al. Doxorubicin and paclitaxel versus doxorubicin and cyclophosphamide as first-line chemotherapy in metastatic breast cancer: The European Organization for Research and Treatment of Cancer 10961 Multicenter Phase III Trial. J Clin Oncol 2002;20:3114–3121.

60 de Abreu FB, Peterson JD, Amos CI, Wells WA, Tsongalis GJ. Effective quality management practices in routine clinical next-generation sequencing. Clin Chem Lab Med 2016;54:761–771.

Figures' titles and legends:

Figure 1

Title: Univariate Kaplan-Meier survival analysis of the prognostic two gene signature predicting Breast Cancer Specific Survival (BCSS) and Distant Metastasis-Free Survival (DMFS) (Transcriptomic cohort, n=112)

Legend: Univariate Kaplan Meier survival analyses to test associations between prognostic two gene signature at the transcriptomic level and clinical outcomes (Significant P-values are bolded HR: Hazard ratio).

Figure 2

Title: Univariate Kaplan Meier survival analysis of our proposed combinatorial two gene signature predicting overall Survival (public domain datasets)

Legend *:* To validate our findings, we utilized the Breast Cancer Gene-Expression Miner v4.0 (bc-GenExMiner v4.0) datasets which includes 5861 breast cancer patients & Genotype 2 outcome public portal, A genome-wide approach to link genotype to clinical outcome by utilizing next generation sequencing and gene chip data of 6,697 breast cancer patients. A) In the Breast Cancer Gene-Expression Miner data portal, high *SPDYC* mRNA expression confers a poor prognosis in the whole (i.e. unselected cohorts) of Breast cancer patients (n=4308, p value<0.0001). B) In the Breast Cancer Gene-Expression Miner data portal, high *SPDYC* mRNA expression confers poor prognosis in the Triple Negative Breast Cancer patients (n=254, p value=0.006). C) In the Genotype 2 outcome public portal, high *ACSM4* mRNA expression confers a poor prognosis outcome in the whole (i.e. unselected cohorts) of Breast cancer patients (n=4029, p value<0.0001). D) In the Genotype 2 outcome public portal, high *SPDYC* mRNA expression confers a poor prognosis outcome in the whole (i.e. unselected cohorts) of Breast cancer patients (n=4029, p value<0.0001). E) In the

Genotype 2 outcome public portal, high *SPDYC & ACSM4* mRNA expression confers a poor prognosis outcome in Triple Negative Breast Cancer patients (n=612, p value<0.0001).

** The data portal used to obtain the Kaplan Meier plot integrates the somatic mutations in the gene and computes the combined transcriptional fingerprint of the mutation(s) using Receiver operating characteristics analysis of breast cancer RNA-seq data and uses the top up and down metagenes to estimate patient survival using Cox regression analysis on gene chip data. An important element is estimation of the transcriptional signature for each somatic mutation, which is carried out by Receiver operating characteristics analysis on the mutation and RNA-seq data.

Figure 3

Title: Univariate Kaplan Meier survival analysis for ACSM4 and SPDYC protein expression for association with Breast Cancer Specific Survival (BCSS) and Distant Metastasis-Free Survival (DMFS) (IHC Cohort, n=333)

Legend: Univariate Kaplan Meier survival analyses to test associations between the ACSM4 and SPDYC protein expression and clinical outcomes (Significant P-values are bolded
HR: Hazard ratio)

Figure 4

Title: Violin plots demonstrating a positive correlation between protein expressions of SPDYC and ACSM4 (Correlation Coefficient, r=0.29, p=0.00001) (IHC Cohort, n=333).

Figure 5

Title: Univariate Kaplan Meier survival analysis of the protein expression of the two gene signature score predicting Breast Cancer Specific Survival (BCSS) and Distant Metastasis-Free Survival (DMFS) (IHC Cohort, n=333)

Legend: Univariate Kaplan Meier survival analyses to test associations between the two gene prognostic signature protein expression and clinical outcome (Significant P-values are bolded HR: Hazard ratio)

Supplementary (A)

Nottingham TNBC Cohort

A subset of this cohort (n=112) was used for NGS analysis and was referred as "transcriptome cohort", whereas the "IHC cohort" (n=333) was used to validate the identified biomarkers at the protein level using immunohistochemistry (IHC). Selection of the transcriptome cohort was based upon availability of FFPE tissue block with sufficient tumor burden. Patients included in the transcriptomic cohort (n=112) were treated at Nottingham University Hospitals between 1987 and 1998. The IHC patient's cohort (n=333) were treated at Nottingham University Hospitals between 1987 and 2006. All patients were managed in a standard manner, where all patients underwent a mastectomy or wide local excision, as decided by disease characteristics or patient choice. Adjuvant therapy was based on Nottingham Prognostic Index (NPI) and Estrogen Receptor status. Prior to the year 2000, not all Estrogen Receptor negative BC patients were offered chemotherapy. Human Epidermal Growth Factor 2 status was not assessed in routine practice or influenced treatment decision during that time. The transcriptome patient's cohort showed 51% of patients received the classical cyclophosphamide, methotrexate and 5-fluorouracil (CMF) adjuvant chemotherapy and 66% received radiotherapy. The patient's cohort (n=333) showed 44% of patients received adjuvant CMF chemotherapy, 29% of patients received 5-flourouracil, epirubicin, cyclophosphamide (FEC) adjuvant chemotherapy, and 20% received doxorubicin, cyclophosphamide (AC) adjuvant chemotherapy. Furthermore, 71% of the IHC cohort received radiotherapy treatment. Distribution of clinicopathological parameters between NGS transcriptomic cohort and the TNBC IHC cohort were tested using Chi-square test (Supplementary Table1).

Transcriptomic Analysis

Invasive tumor cells were macro-dissected from unstained tissue sections from multiple tumor blocks of each case where tumor cellularity (i.e. tumor burden) was at least 50% of the tissue section area were macro-dissected. Hematoxylin and Eosin sections of tumor blocks were microscopically assessed for invasive tumor burden and to guide tumor macro-dissection. Four 10μm unstained tissue sections were used per case. Macro-dissected tissues were deparaffinized, rehydrated, and centrifuged to remove excess ethanol. RNA was extracted using the Omega Mag-Bind XP formalin fixed paraffin embedded RNA isolation kit (Omega, M2595-01) and Kingfisher Flex magnetic particle separator (ThermoFisher) as per manufacturer's instructions. RNA was measured with a Nanodrop 2000c spectrophotometer (Thermo Scientific). First strand cDNA synthesis was performed on approximately 100 ng RNA at 25°C for 10 min, 42°C for 15 min, and 70°C for 15 minutes using random hexamers and ProtoScript II Reverse Transcriptase (New England BioLabs, Ipswich, MA). Second strand synthesis and RNA sequencing libraries were prepared using the Illumina TruSeq RNA access library kit (Illumina, RS-301-2002) and sequenced on an Illumina HiSeq 2500 using PE75 run chemistry. The targeted read count was 60M total reads per sample. Sequencing was performed at the Emory Integrated Genomics Core Facility, Emory University, Atlanta, USA. Raw FastQ sequence reads files were quality assessed and adapter processed using the trim galore wrapper for Fastqc and Cutadapt reads with phred scores >30 retained. The resultant quality trimmed reads were aligned to the hg38 (GRCh38.83) build of the human genome using the STAR aligner [1]. Transcript abundance quantification were performed using HTSEQ [2]. Normalization of genes was determined using transcripts per kilobase million (TPM) [3]. Only one sample per patient was included in downstream analyses by random selection.

Immunohistochemistry (IHC)

To assess the expression in normal breast tissue and evaluate the degree of heterogeneity of expression in tumor tissue, full-face BC tissue sections from 15 cases representative of different BC molecular subtypes, histological grades and stages, were subjected to IHC staining and interpretation before IHC staining and scoring was performed on tissue microarrays (TMAs). TMAs were prepared from tumor samples utilizing 0.6 mm cores, a single sample of each patient representative of the histological grades and stages was included using The TMA Grand Master® (3D HISTECH®, Budapest, Hungary), as previously described [4].

Briefly, xylene was used to deparaffinize tissue slides followed by rehydration through three changes of alcohol. Subsequently, heat-induced citrate antigen retrieval of epitopes was performed (pH 6.0) for 20 minutes using a microwave oven (Whirpool JT359 Jet Chef 1000W). IHC was conducted using the Novolink Max Polymer Detection system (Leica, Newcastle, UK). Novolink peroxidase blocking buffer was applied to slides for 5 minutes to block the endogenous peroxidase activity followed by three washes with Tris-Buffered Saline (TBS, pH 7.6). Protein blocking buffer was applied for 5 minutes to the slides, followed by another three TBS washes. Incubation of the primary antibodies (dilutions were 1:50 for SPDYC and 1:100 for ACSM4) was done for 16 hours at 4°C, followed with three TBS wash for each antibody. Then, incubation for 30 minutes with post primary blocking buffer, and another 30 minutes for the polymer buffer with three TBS wash interval for all. Finally, the 3,3′-diaminobenzidine (DAB) chromogen was applied to the slides for 5 minutes followed by three TBS washes. The slides were then counterstained with Novolink hematoxylin for 6 minutes, dehydrated and cover slipped. Importantly, the specificity of antibodies used in the experiment was confirmed through peptide blocking experiments. Anti SPDYC was neutralized with recombinant human protein antigen (NBP1-80832PEP, lot # PR03260, Novous biological, UK) using 1:25 dilution incubated for 16 hours at 4 °C. Whereas Anti

ACSM4 was neutralized with recombinant human protein antigen (NBP2-14696PEP, lot # 00004251, Novous biological, UK) using 1:50 dilution incubated for 16 hours at 4 °C following the same staining protocol previously described.

4 μm TMA Immunostained sections were digitally scanned at 20X magnification using a NanoZoomer machine (Hamamatsu Photonics, Welwyn Garden City, UK). Morphological evaluation of the cytoplasmic immunoreactivity was assessed using the H-score method based on the intensity of protein expression (0 = negative, 1 = weak, 2 = moderate, 3 = strong) and percentage of stained cells (0–100) as previously reported [5]. TMA cores were considered scorable if invasive tumor cells represented >15% of the total TMA core area. In addition, 25% of TMA sections were scored by two scorers (who were blinded to each other's scores as well as to the clinical data for the samples) to assess inter-observer concordance. The intra-class correlation co-efficient for SPDYC and ACSM4 were 0.77 and 0.80, respectively, indicating substantial concordance between scorers. Moreover, the discordant cases were re-scored by the both observers and a consensus score was agreed upon and assigned.

Statistical analysis

The results of the combinatorial prognostic two gene signature expression, both at the mRNA and protein levels, were analyzed vis-à-vis distant metastasis with respect to sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy. Sensitivity was calculated based on the percentage of recorded histopathologically proven BC tumors interpreted as positive for distant metastasis, defined as (Number of true positive distant metastasis cases harboring high expression of the combinatorial prognostic two-gene signature) × 100/(Number of true positive distant metastasis cases harboring high expression of the combinatorial prognostic two-gene signature + Number of false negative distant metastasis cases harboring low expression of the combinatorial prognostic two-gene signature), where true positive were the cases correctly recorded as distant metastasis positive, and false

negative were cases diagnosed as positive for distant metastasis but harbored low expressions of the combinatorial prognostic two-gene signature on both mRNA and protein levels. While specificity was calculated based on the percentage of recorded histopathologically proven BC tumors interpreted as negative for distant metastasis, defined as (Number of true negative distant metastasis cases harboring low expression of the combinatorial prognostic two-gene signature) × 100/(Number of true negative distant metastasis cases harboring low expression of the combinatorial prognostic two-gene signature + Number of false positive distant metastasis cases harboring high expression of the combinatorial prognostic two-gene signature) where true negative is a case correctly recorded as a distant metastasis negative, and false positive a case recorded as distant metastasis negative but harbored high expressions of the combinatorial prognostic two gene signature on both mRNA and protein levels [6].

To evaluate the precision rate of this prognostic index in predicting distant metastasis, the positive predictive value [the proportion of subjects with distant metastasis who are correctly diagnosed. PPV = true positive/true positive + false positive] and the negative predictive value [proportion of subjects without distant metastasis who are correctly diagnosed. NPV= true negative/true negative + false negative] were calculated [7]. Finally, to investigate the accuracy of the overall ability of the expressions of the combinatorial prognostic two-gene signature to correctly classify cases as high or low risk for TNBC metastatic disease, defined as (number of cases true positive for distant metastasis harboring high expressions of the combinatorial prognostic two-gene signature + true negative cases for distant metastasis harboring low expression of the combinatorial prognostic two-gene signature) × 100/(total number of patients who underwent scanning) were calculated [8].

Supplementary (A) Table 1: Patients Cohorts

Clinicopathological Parameter	Transcriptomic cohort (n=112) No (%)	IHC Cohort (n=333) No (%)	P-value	Clinicopathological Parameter	Transcriptomic cohort (n=112) No (%)	IHC Cohort (n=333) No (%)	P-value
Age (years)				**Vascular Invasion**			
< 50	62 (56)	154 (46)	**0.025**	**Yes**	42 (38)	89 (27)	**0.005**
≥ 50	50 (44)	179 (54)		**No**	70 (62)	244 (73)	
Tumor Size (cm)				**Distant Metastasis**			
< 2	32 (28)	144 (43)	**< 0.001**	**Yes**	46 (41)	91 (28)	**0.001**
≥ 2	80 (72)	189 (57)		**No**	64 (59)	241 (72)	
Grade				**Breast Cancer Specific Survival**			
1	1 (1)	4 (1)		**Alive**	61 (55)	216 (65)	
2	6 (6)	24 (7)	0.469	**Died of BC**	42 (38)	89 (27)	**0.033**
3	104 (93)	305 (92)		**Died of other causes**	9 (7)	27 (8)	
Nodal Stage				**Chemotherapy**			
1	68 (61)	223 (67)		**Yes**	57 (51)	257 (65)	
2	31 (27)	77 (23)	0.270	**No**	44(40)	123 (31)	**0.005**
3	13 (12)	32 (10)		**No treatment**	11 (10)	14 (4)	
Nottingham Prognostic Index				**Radiotherapy**			
Good Prognosis	3 (3)	96 (29)		**Yes**	74 (66)	278 (71)	
Moderate Prognosis	76 (68)	194 (58)	**0.037**	**No**	28 (25)	97 (25)	0.668
Poor Prognosis	33 (29)	42 (13)		**No treatment**	10 (9)	19 (4)	

Supplementary (A) Table 2: Pathway analysis of genes strongly predictive of both Breast Cancer Specific Survival (BCSS) and Distant Metastasis-Free Survival (DMFS) (Transcriptomic Cohort, n=112)

Description	Gene Set	Gene Set Size	Overlap	Significance P-value	Enrichment Ratio	Gene Symbol	Gene Name
JAK/STAT signaling pathway	P00038	15	1	0.130	7.1937	*STAT1*	signal transducer and activator of transcription 1
p38 MAPK pathway	P05918	34	2	**0.038**	6.3473	*MAP3K4*	mitogen-activated protein kinase 4
						MAP2K6	mitogen-activated protein kinase 6
Ras Pathway	P04393	70	3	**0.025**	4.6245	*STAT1*	signal transducer and activator of transcription 1
						MAP2K6	mitogen-activated protein kinase 6
						MAP3K4	mitogen-activated protein kinase 4
EGF receptor signaling pathway	P00018	115	4	**0.019**	3.7532	*STAT1*	signal transducer and activator of transcription 1
						MAP3K4	mitogen-activated protein kinase 4
						MAP2K6	mitogen-activated protein kinase 6
						EGFR	epidermal growth factor receptor
Cadherin signaling pathway	P00012	153	4	**0.048**	2.821	*PCDHGC5*	protocadherin gamma subfamily C, 5
						PCDHB1	protocadherin beta 1
						FZD5	frizzled class receptor 5
						EGFR	epidermal growth factor receptor
PDGF signaling pathway	P00047	125	3	0.105	2.5897	*MAP3K4*	mitogen-activated protein kinase 4
						STAT1	signal transducer and activator of transcription 1
						ETV3	ETS variant 3

Table legend: The top 200 ranked genes predicting DMFS (DMFS genes panel) and those predicting BCSS (BCSS genes panel) were investigated to determine the most statistically enriched pathways in our DMFS and BCSS gene lists by conducting Panther enrichment pathway analysis using Webgestalt. Since the aforementioned biologically important pathways showed statistically significant enrichment, they merited more in-depth evaluation, nevertheless, they reinforced the power of our discovered classifier panels in predicting TNBC outcome.

Supplementary (A) Table 3: Univariate analyses for the 21 genes identified by Venny tool that potentially predictive of both Breast Cancer Specific Survival (BCSS) and Distant Metastasis-Free Survival (DMFS) (Transcriptomic Cohort, n=112)

Gene ID	Breast Cancer specific Survival			Distant Metastasis Free Survival		
	Frequency		P-value	Frequency		P-value
	Below-cut-point Expression (N)	Above cut-point Expression (N)		Below-cut-point Expression (N)	Above cut-point Expression (N)	
AC020931.1	74	38	**0.037**	69	37	**0.012**
AC079305.10	50	62	**0.02**	47	59	**0.001**
AC084809.2	95	17	**0.002**	90	16	**0.012**
ACSM4	80	32	**< 0.001**	75	31	**< 0.001**
MEX3A	90	22	**0.015**	87	19	**0.038**
NDUFA4L2	72	40	**0.036**	66	40	**0.010**
PAXBP1-AS1	84	28	**0.033**	81	25	**0.020**
BICC1	72	40	0.160	68	38	**0.048**
CCDC54	71	24	0.330	68	38	0.115
DCTN1-AS1	91	21	**0.047**	85	21	**0.006**
DOCK10	91	21	**0.020**	86	20	**0.037**
GTF3C6	72	40	0.242	68	38	0.076
HARBI1	94	18	0.113	90	16	**0.017**
PPL	94	18	**0.002**	89	17	**< 0.001**
RP1129H23.5	82	30	**0.002**	77	29	**0.001**
RP11409C19.2	72	40	0.180	70	36	0.068
RPS10P18	95	17	**0.002**	89	17	**0.003**
RPS3AP47	91	21	0.588	86	20	0.509
SNORD99	90	22	0.099	84	22	0.075
SPDYC	90	22	**0.014**	86	20	**0.018**
SRP72P2	91	21	0.381	20	86	0.175
Total	**112**			**106**		

- Significant P-values are bolded

Supplementary (A) Table 4 (A): Multivariate Cox regression analysis for genes potentially associated with both Breast Cancer Specific Survival (BCSS) and Distant Metastasis-Free Survival (DMFS) (Transcriptomic Cohort, n=112)

Covariates	Breast Cancer-Specific Survival				Distant Metastasis-Free Survival			
	P-value	Hazard Ratio	95% CI		P-value	Hazard Ratio	95% CI	
			Lower	Upper			Lower	Upper
Age	0.061	1.836	0.971	3.472	0.047	1.920	1.009	3.653
Grade	0.452	1.712	0.421	6.954	0.441	1.743	0.424	7.157
Nodal Stage	0.116	1.582	0.893	2.801	0.173	1.480	0.842	2.600
Tumor Size	0.258	0.675	0.342	1.333	0.137	0.601	0.307	1.176
Vascular Invasion	0.454	1.365	0.604	3.083	0.550	1.277	0.573	2.842
AC020931.1	0.103	0.528	0.245	1.137	**0.043**	0.437	0.196	0.976
Age	0.113	1.682	0.884	3.210	0.101	1.727	0.899	3.315
Grade	0.362	1.914	0.475	7.714	0.341	1.986	0.484	8.150
Nodal Stage	0.102	1.595	0.911	2.793	0.148	1.500	0.866	2.595
Tumor Size	0.351	0.724	0.367	1.427	0.179	0.635	0.327	1.232
Vascular Invasion	0.388	1.413	0.645	3.097	0.416	1.374	0.639	2.953
AC079305.10	**0.046**	1.983	1.012	3.887	**0.004**	2.871	1.394	5.912
Age	0.225	1.512	0.775	2.951	0.184	1.583	0.804	3.117
Grade	0.230	2.364	0.580	9.629	0.219	2.400	0.594	9.700
Nodal Stage	0.077	1.693	0.945	3.034	0.094	1.631	0.920	2.890
Tumor Size	0.239	0.663	0.335	1.314	0.117	0.583	0.296	1.146
Vascular Invasion	0.668	1.203	0.517	2.795	0.723	1.160	0.511	2.637
AC084809.2	**0.047**	2.184	1.010	4.721	0.131	1.848	0.832	4.104
Age	**0.040**	1.969	1.030	3.763	0.048	1.919	1.005	3.664
Grade	0.176	2.436	0.671	8.841	0.174	2.466	0.672	9.049
Nodal Stage	**0.015**	2.135	1.161	3.927	0.027	1.957	1.080	3.545
Tumor Size	0.165	0.612	0.306	1.224	0.097	0.560	0.282	1.110
Vascular Invasion	0.839	0.914	0.382	2.183	0.796	0.893	0.381	2.096
ACSM4	**<0.001**	3.749	1.964	7.158	**0.001**	3.154	1.645	6.046

- Significant P-values are bolded CI: Confidence interval

Supplementary (A) Table 4(B): Multivariate Cox regression analysis of genes potentially associated with both Breast Cancer Specific Survival (BCSS) and Distant Metastasis-Free Survival (DMFS) (Transcriptomic Cohort, n=112)

Covariates	Breast Cancer specific Survival				Distant Metastasis Free Survival			
	Significance	Hazard Ratio	95% CI		Significance	Hazard Ratio	95% CI	
			Lower	Upper			Lower	Upper
Age	0.306	1.421	0.725	2.784	0.198	1.557	0.794	3.053
Grade	0.423	1.762	0.440	7.049	0.318	2.029	0.507	8.129
Nodal Stage	0.043	1.808	1.018	3.211	0.070	1.689	0.958	2.977
Tumor Size	0.332	0.715	0.363	1.408	0.145	0.607	0.311	1.188
Vascular Invasion	0.437	1.366	0.623	2.994	0.577	1.247	0.574	2.7111
MEX3A	**0.032**	2.208	1.072	4.546	0.102	1.857	0.884	3.904
Age	0.082	1.751	0.932	3.293	0.077	1.769	0.940	3.329
Grade	0.330	1.937	0.512	7.328	0.320	1.954	0.521	7.330
Nodal Stage	0.141	1.540	0.867	2.738	0.191	1.455	0.830	2.552
Tumor Size	0.384	0.734	0.366	1.472	0.267	0.678	0.341	1.347
Vascular Invasion	0.450	1.358	0.614	3.004	0.523	1.289	0.591	2.811
NDUFA4L2	0.181	1.569	0.811	3.033	0.068	1.846	0.955	3.569
Age	**0.039**	1.960	1.033	3.720	0.041	1.949	1.028	3.695
Grade	0.275	2.098	0.555	7.924	0.255	2.169	0.572	8.217
Nodal Stage	0.164	1.524	0.842	2.756	0.196	1.469	0.820	2.630
Tumor Size	0.199	0.639	0.332	1.266	0.099	0.566	0.288	1.114
Vascular Invasion	0.440	1.397	0.598	3.264	0.495	1.339	0.579	3.097
PAXBP1-AS1	**0.049**	0.411	0.170	0.996	**0.036**	0.362	0.140	0.936
Age	**0.028**	2.081	1.083	3.999	**0.019**	2.185	1.135	4.208
Grade	0.229	2.360	0.583	9.551	0.188	2.581	0.630	10.579
Nodal Stage	**0.035**	1.904	1.047	3.465	**0.034**	1.920	1.052	3.506
Tumor Size	0.260	0.675	0.341	1.337	0.111	0.577	0.294	1.134
Vascular Invasion	0.480	1.346	0.591	3.066	0.510	1.315	0.582	2.970
DCTN1-AS1	**0.003**	3.061	1.452	6.454	**< 0.001**	3.874	1.873	8.013

- Significant P-values are bolded CI: Confidence interval

Supplementary (A) Table 4(C): Multivariate Cox regression analysis of genes potentially associated with both Breast Cancer Specific Survival (BCSS) and Distant Metastasis-Free Survival (DMFS) (Transcriptomic Cohort, n=112)

Covariates	Breast Cancer specific Survival				Distant Metastasis Free Survival			
	Significance	Hazard Ratio	95% CI		Significance	Hazard Ratio	95% CI	
			Lower	Upper			Lower	Upper
Age	**0.044**	1.920	1.019	3.618	**0.050**	1.881	0.999	3.543
Grade	0.214	2.293	0.619	8.497	0.203	2.364	0.629	8.885
Nodal Stage	0.272	1.399	0.768	2.550	0.236	1.426	0.793	2.564
Tumor Size	0.204	0.646	0.328	1.269	0.104	0.575	0.295	1.121
Vascular Invasion	0.313	1.573	0.652	3.794	0.462	1.376	0.588	3.217
DOCK10	**0.031**	2.215	1.076	4.560	0.075	1.939	0.935	4.025
Age	0.097	1.740	0.905	3.344	0.110	1.712	0.885	3.313
Grade	0.276	2.190	0.535	8.971	0.241	2.333	0.565	9.626
Nodal Stage	0.198	1.459	0.821	2.595	0.259	1.385	0.787	2.437
Tumor Size	0.301	0.692	0.344	1.390	1.77	0.619	0.309	1.242
Vascular Invasion	0.225	1.687	0.725	3.928	0.266	1.595	0.700	3.637
PPL	**0.004**	2.961	1.424	6.157	**0.004**	2.966	1.414	6.220
Age	0.064	1.845	0.964	3.532	**0.043**	1.977	1.022	3.825
Grade	0.347	1.956	0.483	7.914	0.270	2.212	0.539	9.079
Nodal Stage	0.140	1.538	0.869	2.722	0.207	1.440	0.817	2.538
Tumor Size	0.215	0.645	0.323	1.293.	0.072	0.531	0.267	1.058
Vascular Invasion	0.279	1.581	0.689	3.627	0.336	1.492	0.660	3.374
RP1129H23.5	**0.001**	2.832	1.500	5.346	**0.001**	2.975	1.569	5.641
Age	0.024	2.119	1.102	4.078	0.027	2.088	1.085	4.016
Grade	0.360	1.903	0.480	7.541	0.309	2.053	0.513	8.207
Nodal Stage	0.048	1.685	1.005	2.826	0.087	1.566	0.938	2.617
Tumor Size	0.189	0.663	0.320	1.252	0.088	0.556	0.283	1.092
Vascular Invasion	0.181	1.663	0.789	3.503	0.240	1.558	0.755	3.263
RPS10P18	**< 0.001**	3.941	1.852	8.389	**0.001**	3.399	1.611	7.171

- Significant P-values are bolded CI: Confidence interval

Supplementary (A) Table 4 (D): Multivariate Cox regression analysis of genes potentially associated with both Breast Cancer Specific Survival (BCSS) and Distant Metastasis-Free Survival (DMFS) (Transcriptomic Cohort, n=112)

Covariates	Breast Cancer specific Survival				Distant Metastasis Free Survival			
	Significance	Hazard Ratio	95% CI		Significance	Hazard Ratio	95% CI	
			Lower	Upper			Lower	Upper
Age	**0.046**	1.907	1.012	3.596	0.050	1.886	1.000	3.557
Grade	0.209	2.279	0.631	8.232	0.190	2.378	0.650	8.695
Nodal Stage	0.125	1.561	0.884	2.756	0.129	1.545	0.881	2.710
Tumor Size	0.140	0.596	0.300	1.185	0.066	0.529	0.269	1.042
Vascular Invasion	0.247	1.666	0.702	3.953	0.383	1.452	0.628	3.356
SPDYC	**0.003**	2.892	1.453	5.758	**0.006**	2.646	1.317	5.317

- Significant P-values are bolded CI: Confidence interval

Supplementary (A) Table 5: Combined multivariate Cox regression analysis of genes potentially associated with both Breast Cancer-Specific Survival and Distant Metastasis-Free Survival (Transcriptomic Cohort, n=112)

Covariates	Breast Cancer-Specific Survival				Distant Metastasis-Free Survival			
	P-value	Hazard Ratio	95% CI		P-value	Hazard Ratio	95% CI	
			Lower	Upper			Lower	Upper
Age	**0.043**	2.053	1.023	4.119	**0.046**	2.033	1.013	4.079
Grade	0.203	2.331	0.633	8.582	0.208	2.433	0.610	9.703
Nodal Stage	0.181	1.523	0.822	2.819	0.208	1.483	0.803	2.739
Tumor Size	0.202	0.616	0.293	1.296	0.153	0.588	0.283	1.219
Vascular Invasion	0.215	1.813	0.708	4.642	0.253	1.697	0.685	4.206
ACSM4	**0.004**	3.143	1.445	6.834	**0.015**	2.727	1.213	6.135
SPDYC	**0.016**	2.455	1.183	5.096	**0.012**	2.594	1.234	5.450
DCTN1-AS1	0.954	1.032	0.358	2.976	0.304	1.721	0.611	4.849
PPL	0.883	1.079	0.395	2.947	0.926	0.950	0.317	2.841
RP1129H23.5	0.565	1.268	0.566	2.840	0.724	1.158	0.513	2.612
RPS10P18	0.093	2.134	0.882	5.160	0.251	1.716	0.682	4.316
AC079305.10	0.078	2.027	0.924	4.444	0.007	3.014	1.348	6.740
PAXBP1-AS1	0.083	0.432	0.168	1.114	0.051	0.366	0.133	1.005

- Significant P-values are bolded CI: Confidence interval

Supplementary *(A)* Table 6: Distribution of cases within transcriptomic and IHC cohorts to test the prognostic index with reference to clinical evidence of distant metastasis.

Prognostic index of the two gene signature at mRNA Level	**Distant Metastasis**			
	Expression level	**Yes**	**No**	**Total**
	Above cut-point Expression	**25 TP**	21 FP	46
	Below-cut-point Expression	18 FN	**46 TN**	64
Total		43	67	110
Prognostic index of the two gene signature at protein level	**Distant Metastasis**			
	Expression level	**Yes**	**No**	**Total**
	Above cut-point Expression	**48 TP**	111 FP	159
	Below-cut-point Expression	18 FN	**80 TN**	98
Total		66	191	257

- TP = true positive, FP = false positive, TN = true negative, and FN = false negative.
- TP = true positive, FP = false positive, TN = true negative, and FN = false negative.
- Sensitivity = TP * 100/(TP+FN).
- Specificity = TN * 100/(TN +FP).
- PPV, the proportion of subjects with distant metastasis who are correctly diagnosed; PPV = TP* 100/(TP + FP).
- NPV, proportion of subjects without distant metastasis who are correctly diagnosed; NPV= TN * 100 / (TN + FN).
- Accuracy = TP + TN * 100/Total number of cases.

Supplementary (B)

Supplementary (B) Figure 1: Flowchart of Analytical and Experimental Methodology of the Study

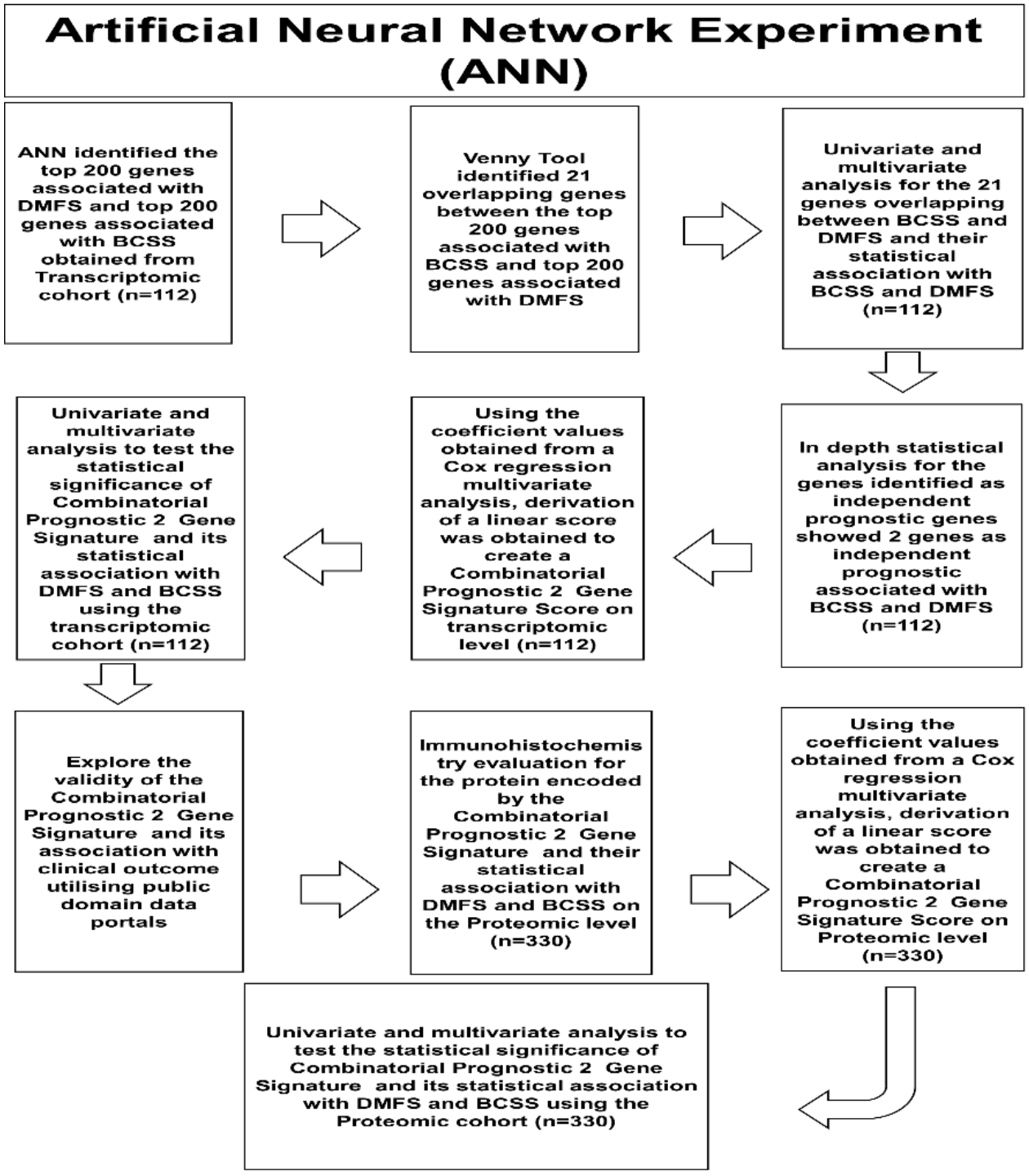
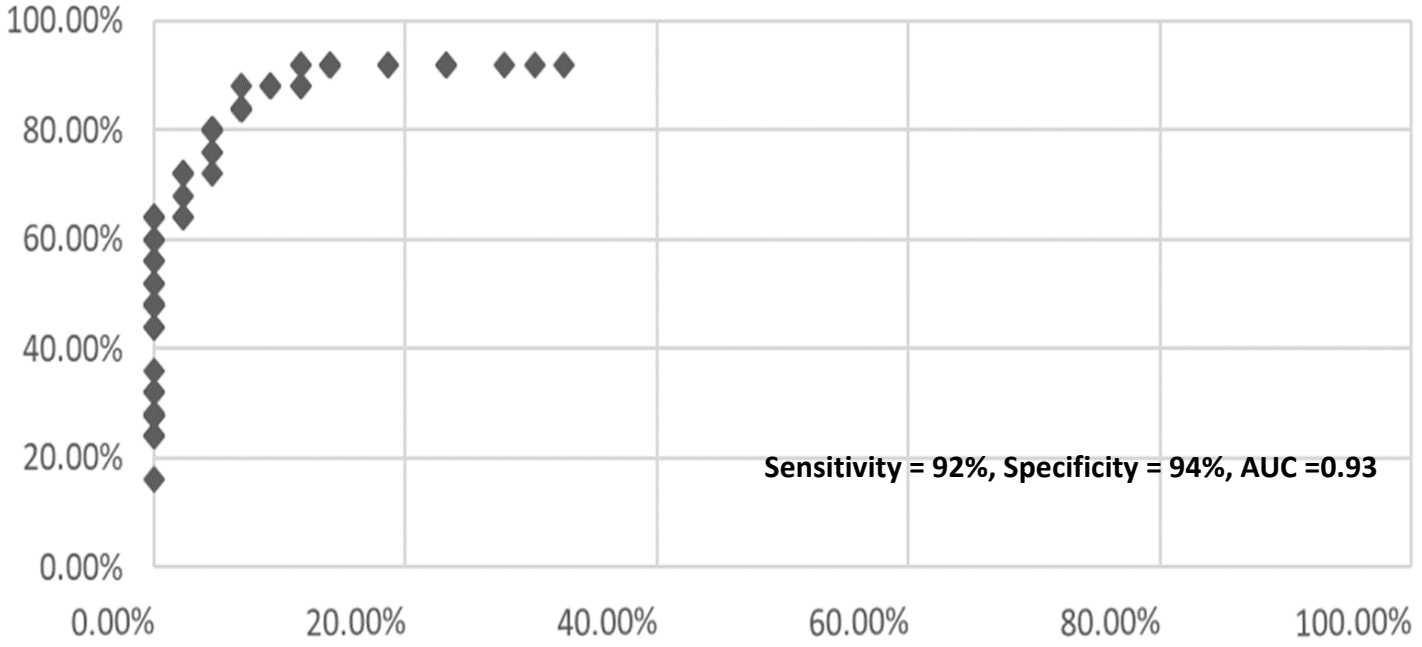


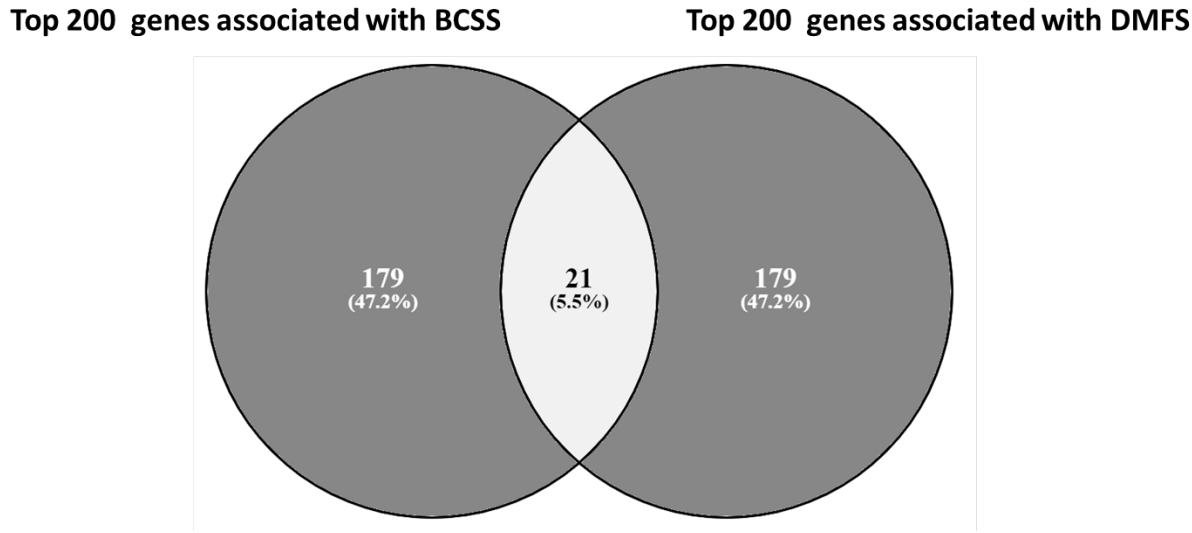
Figure legend: The layout of steps used in the conducted experiment utilizing the mRNA cohort of 112 TNBC cases to identify genes associated with clinical outcome.
DMFS: Distant Metastasis Free Survival, BCSS: Breast Cancer Specific Survival

Supplementary (B) Figure 2: Receiver Operative Characteristic curve: Receiver Operative Characteristic curve depicting the sensitivity and specificity of the predicted 21 gene panel associated with both Breast cancer-specific survival (BCSS) and Distant metastasis-free survival (DMFS) (Transcriptomic cohort, n=112)

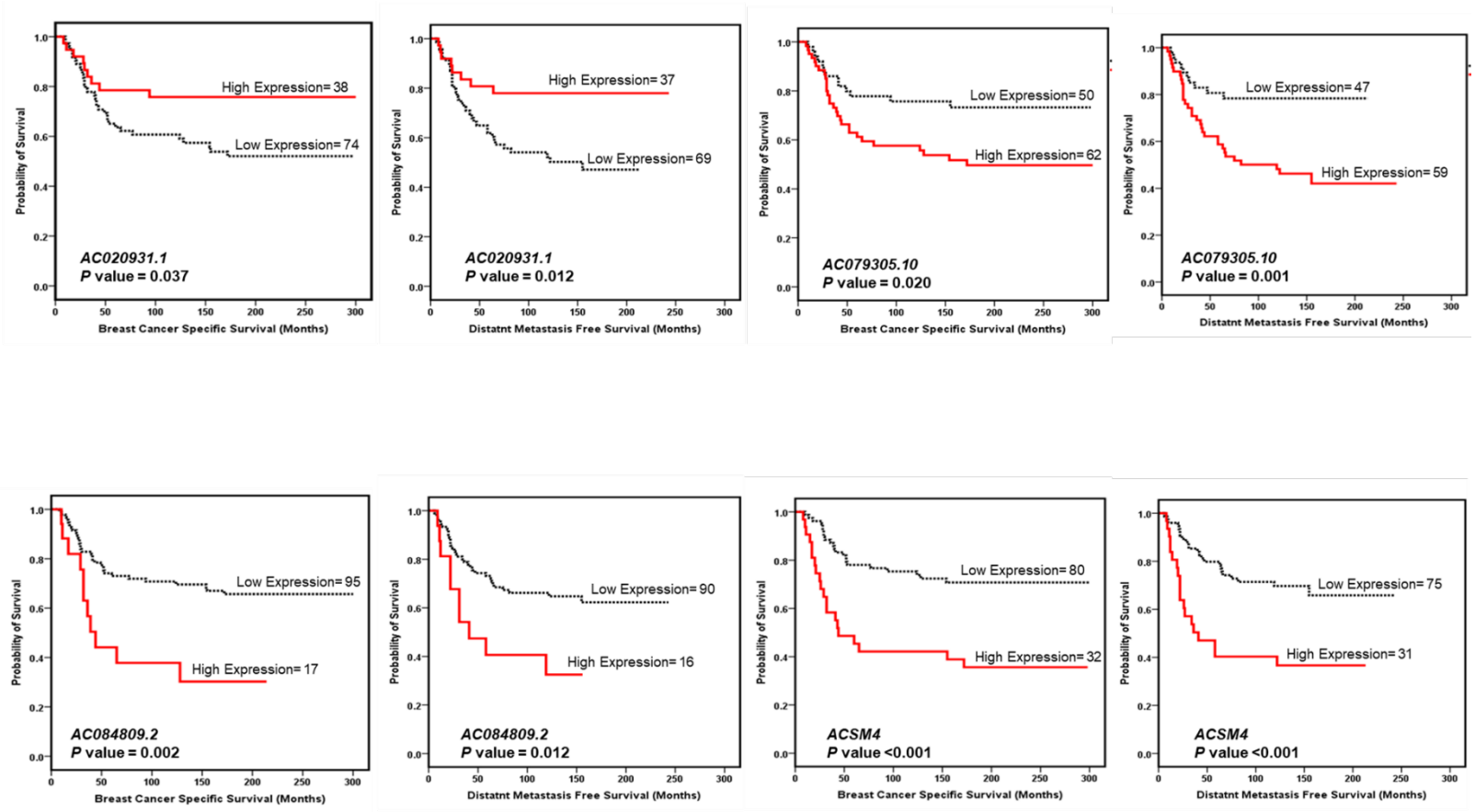


- Area Under the Curve (AUC)

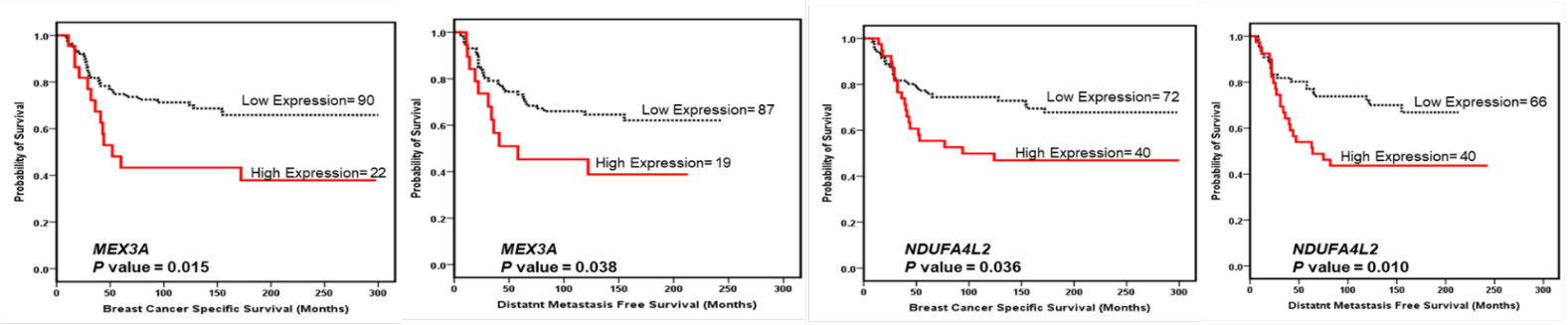
Supplementary (B) Figure 3: The output of Venny diagram tool showing the overlapping genes between the top 200 genes associated with Breast Cancer Specific Survival (BCSS) and the top 200 genes associated with Distant Metastasis Free Survival (DMFS) (Transcriptomic cohort, n=112)

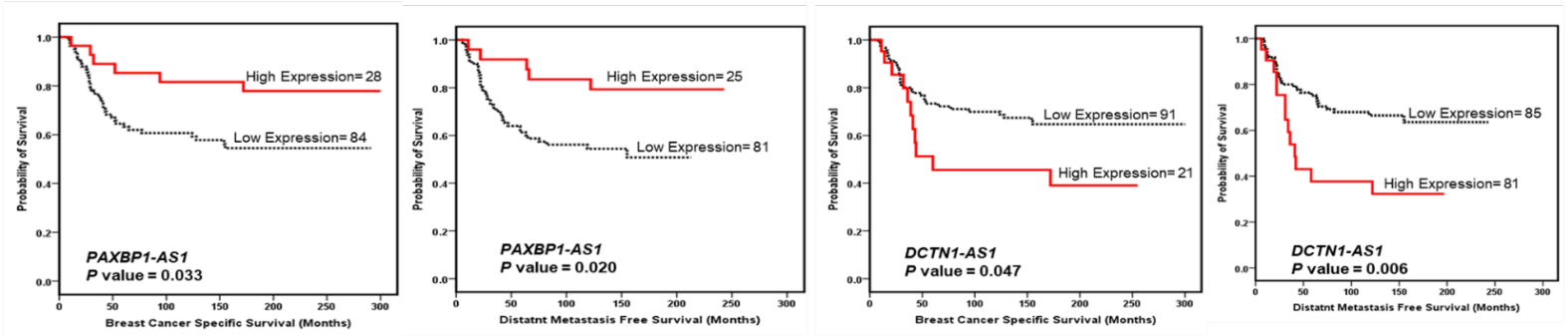


Supplementary (B) Figure 4 A: Univariate Kaplan-Meier survival analyses of the transcripts identified by Artificial Neural Network analysis to be strongly predictive of Breast Cancer-Specific Survival (BCSS) and Distant Metastasis-Free Survival (DMFS) using RNA seq matrices (Transcriptomic cohort, n=112)

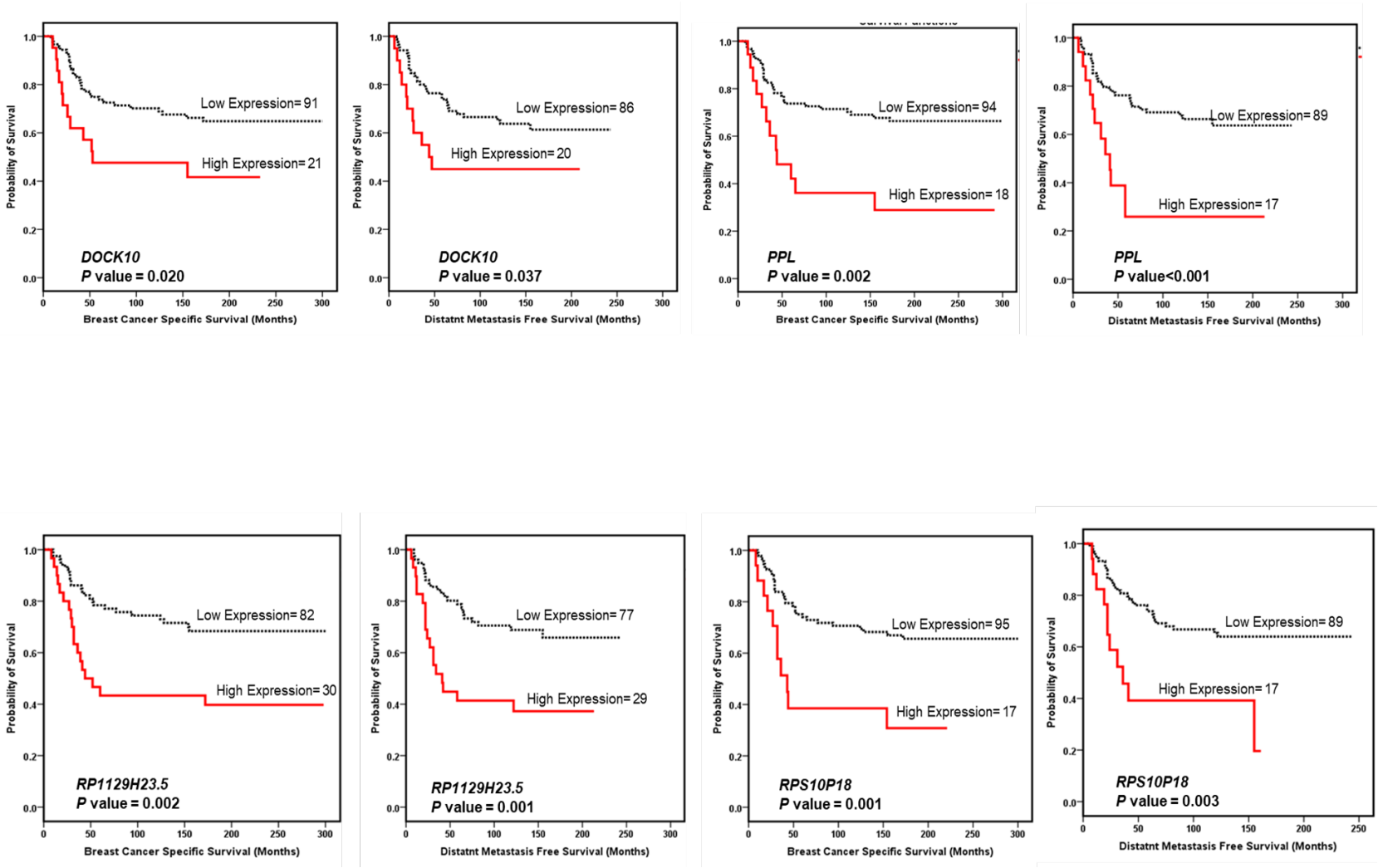


Supplementary (B) Figure 4 B: Univariate Kaplan-Meier survival analyses of the transcripts identified by Artificial Neural Network analysis to be strongly predictive of Breast Cancer-Specific Survival and Distant Metastasis-Free Survival using RNA seq matrices (Transcriptomic cohort, n=112)

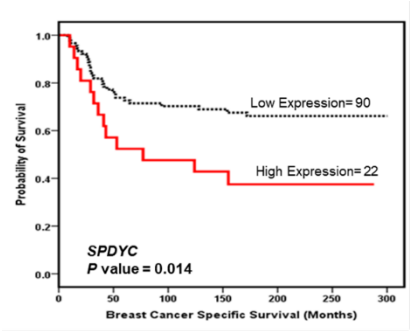


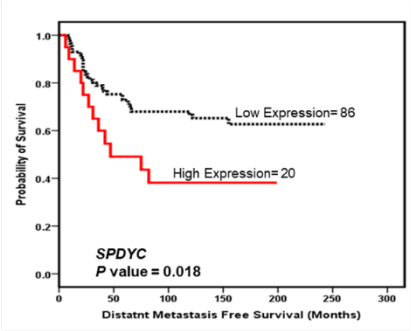


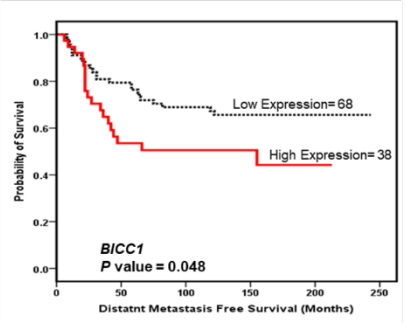
Supplementary (B) Figure 4 C: Univariate Kaplan-Meier survival analyses of the transcripts identified by Artificial Neural Network analysis to be strongly predictive of Breast Cancer-Specific Survival and Distant Metastasis-Free Survival using RNA seq matrices (Transcriptomic cohort, n=112).

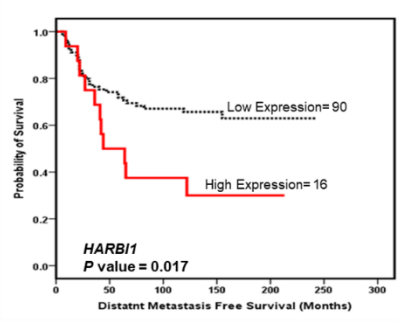


Supplementary (B) Figure 4 D: Univariate Kaplan-Meier survival analyses of the transcripts identified by Artificial Neural Network analysis to be strongly predictive of Breast Cancer-Specific Survival and Distant Metastasis-Free Survival using RNA seq matrices (Transcriptomic cohort, n=112).









Supplementary (B) Figure 5: Immunohistochemical expression of SPDYC and ACSM4 in formalin fixed paraffin embedded tissue microarray cores from TNBC samples (IHC cohort, n=333)

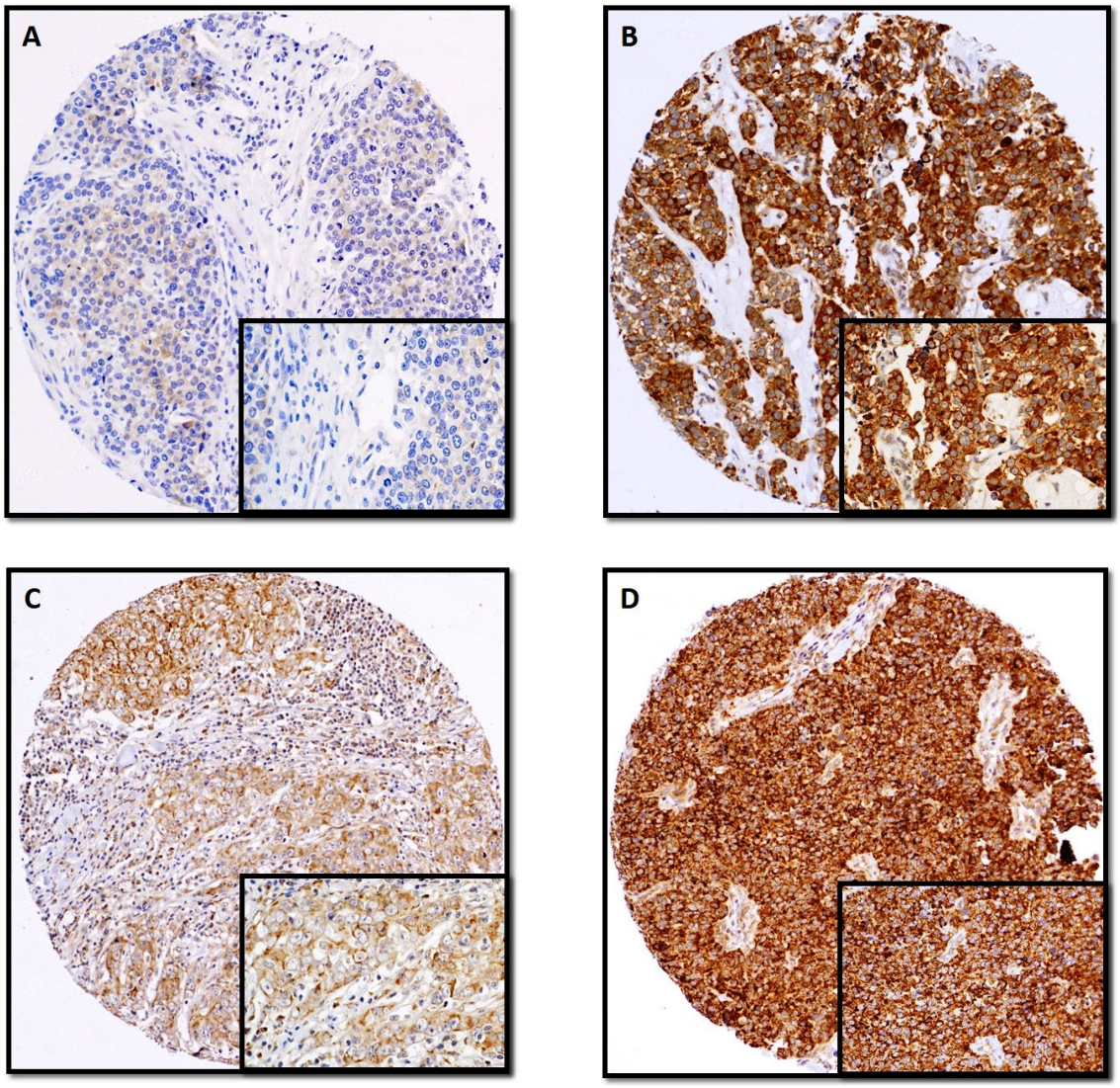


Figure (B) Legend; Immunohistochemical protein expression of SPDYC and ACSM4 in TNBC tissue microarray cores (power 10×, inset high power 40×). A) SPDYC weak expression, B) SPDYC strong expression, C) ACSM4 weak expression, D) ACSM4 strong expression.

Within the proteomic cohort, a significant negative linear correlation was observed between *SPDYC* mRNA (assessed by NGS) and protein level (r= -0.29, p=0.008, 80/112) in contrast to *ACSM4*, which showed a trend towards positive linear association (r=0.08, p=0.52, 65/112). Various factors could have contributed to the inverse correlation between the transcriptomic and proteomic levels of SPDYC. For instance, the improved quantification of the mRNA isoform and differential exon usage for the gene underlying the protein, the depth used in sequencing the cases, and potential considerable difference of proteins in vivo half-lives as compared to the mRNA and vice versa [9].

Supplementary (C)

Supplementary (C) Table A: Genes associated with Distant Metastasis Free Survival (DMFS) identified by Artificial Neural Network analysis (Transcriptomic Cohort, n=112)

Gene Id	initial_alias	description
AANAT	ENSG00000129673	aralkylamine N-acetyltransferase [Source: HGNC Symbol;Acc:HGNC:19]
AC004893.10	ENSG00000238109	ring finger protein 14(RNF14) pseudogene
AC005042.4	ENSG00000204380	PKP4 antisense RNA 1 [Source: HGNC Symbol; Acc:HGNC:52580]
AC007750.5	ENSG00000236841	novel transcript
AC012358.4	ENSG00000227799	pseudogene similar to RIKEN cDNA 2210021J22
AC020931.1	ENSG00000257110	novel transcript
AC064852.4	ENSG00000241409	novel transcript
AC068491.3	ENSG00000223973	platelet-activating factor acetylhydrolase, isoform Ib, alpha subunit 45kDa (PAFAH1B1) pseudogene
AC079305.10	ENSG00000222043	novel transcript
AC079807.2	ENSG00000233230	novel transcript
AC084809.2	ENSG00000226377	novel transcript
AC093724.2	ENSG00000213222	translocase of outer mitochondrial membrane 40 (TOMM40) pseudogene
AC094019.4	ENSG00000236732	ribosomal protein L21 (RPL21) pseudogene
AC098824.6	ENSG00000232202	istidine-rich domain (CHORD)-containing 1 (CHORDC1) pseudogene
ACSM4	ENSG00000215009	acyl-CoA synthetase medium chain family member 4 [Source: HGNC Symbol;Acc:HGNC:32016]
ACSM5	ENSG00000183558	Acyl-CoA Synthetase Medium Chain Family Member 5
ANO1-AS1	ENSG00000254902	ANO1 antisense RNA 1 [Source: HGNC Symbol;Acc:HGNC:40016]
AP001062.7	ENSG00000184441	novel transcript, antisense to C21orf2
AP001189.4	ENSG00000236304	uncharacterized LOC107984360 [Source: NCBI gene;Acc:107984360]
AP004290.1	ENSG00000236583	insulin-like growth factor 2 mRNA binding protein 2 (IGF2BP2) pseudogene
ARMS2	ENSG00000254636	age-related maculopathy susceptibility 2 [Source: HGNC Symbol;Acc:HGNC:32685]
ASIP	ENSG00000101440	agouti signaling protein [Source: HGNC Symbol;Acc:HGNC:745]
ATP6V1G1P4	ENSG00000233346	novel transcript
BANF1P4	ENSG00000223828	barrier to autointegration factor 1 pseudogene 4 [Source: HGNC Symbol;Acc:HGNC:43884]
BICC1	ENSG00000122870	BicC family RNA binding protein 1 [Source: HGNC Symbol;Acc:HGNC:19351]
BRD7P4	ENSG00000218676	bromodomain containing 7 pseudogene 4 [Source: HGNC Symbol;Acc:HGNC:37630]
BRI3P1	ENSG00000225169	brain protein I3 pseudogene 1 [Source: HGNC Symbol;Acc:HGNC:33533]
C14orf1	ENSG00000133935	ergosterol biosynthesis 28 homolog [Source: HGNC Symbol;Acc:HGNC:1187]
C1orf204	ENSG00000188004	small nucleolar RNA host gene 28 [Source: HGNC Symbol;Acc:HGNC:27647]
CBX3P5	ENSG00000257666	chromobox 3 pseudogene 5 [Source: HGNC Symbol;Acc:HGNC:42877]
CCDC54	ENSG00000138483	coiled-coil domain containing 54 [Source: HGNC Symbol;Acc:HGNC:30703]
CCDC6	ENSG00000108091	coiled-coil domain containing 6 [Source: HGNC Symbol;Acc:HGNC:18782]
CEACAM6	ENSG00000086548	carcinoembryonic antigen related cell adhesion molecule 6 [Source: HGNC Symbol;Acc:HGNC:1818]
CEMP1	ENSG00000205923	cementum protein 1 [Source: HGNC Symbol;Acc:HGNC:32553]
CEP164	ENSG00000110274	centrosomal protein 164 [Source: HGNC Symbol;Acc:HGNC:29182]
CEP89	ENSG00000121289	centrosomal protein 89 [Source: HGNC Symbol;Acc:HGNC:25907]
CLUHP2	ENSG00000228947	novel transcript
COL5A1-AS1	ENSG00000204011	COL5A1 antisense RNA 1 [Source: HGNC Symbol;Acc:HGNC:31368]
COL5A2	ENSG00000204262	collagen type V alpha 2 chain [Source: HGNC Symbol;Acc:HGNC:2210]
CRYM	ENSG00000103316	crystallin mu [Source: HGNC Symbol;Acc:HGNC:2418]
CTB-33G10.1	ENSG00000243829	ribosomal protein S9 (RPS9) pseudogene
CTB-47B11.3	ENSG00000248544	novel transcript, antisense to CYFIP2
CTSZ	ENSG00000101160	cathepsin Z [Source: HGNC Symbol;Acc:HGNC:2547]
CUBNP3	ENSG00000235690	cubilin pseudogene 3 [Source: HGNC Symbol;Acc:HGNC:44985]
DBNL	ENSG00000136279	drebrin like [Source: HGNC Symbol;Acc:HGNC:2696]
DCC	ENSG00000187323	DCC netrin 1 receptor [Source: HGNC Symbol;Acc:HGNC:2701]
DCTN1-AS1	ENSG00000237737	DCTN1 antisense RNA 1 [Source: HGNC Symbol;Acc:HGNC:44151]

DHODH	ENSG00000102967	dihydroorotate dehydrogenase (quinone) [Source: HGNC Symbol;Acc:HGNC:2867]
DNAH3	ENSG00000158486	dynein axonemal heavy chain 3 [Source: HGNC Symbol;Acc:HGNC:2949]
DOCK10	ENSG00000135905	dedicator of cytokinesis 10 [Source: HGNC Symbol;Acc:HGNC:23479]
DSP	ENSG00000096696	desmoplakin [Source:HGNC Symbol;Acc:HGNC:3052]
EIF5B	ENSG00000158417	eukaryotic translation initiation factor 5B [Source: HGNC Symbol;Acc:HGNC:30793]
ERP44	ENSG00000023318	endoplasmic reticulum protein 44 [Source: HGNC Symbol;Acc:HGNC:18311]
EWSAT1	ENSG00000212768	novel transcript
EXOSC3P1	ENSG00000229007	exosome component 3 pseudogene 1 [Source: HGNC Symbol;Acc:HGNC:33989]
FAM138E	ENSG00000248894	novel transcript
FAM196A	ENSG00000188916	inhibitory synaptic factor 2A [Source: HGNC Symbol;Acc:HGNC:33859]
FAM212A	ENSG00000185614	inka box actin regulator 1 [Source: HGNC Symbol;Acc:HGNC:32480]
FGD1	ENSG00000102302	FYVE, RhoGEF and PH domain containing 1 [Source: HGNC Symbol;Acc:HGNC:3663]
FGL2	ENSG00000127951	fibrinogen like 2 [Source: HGNC Symbol;Acc:HGNC:3696]
FOXQ1	ENSG00000164379	forkhead box Q1 [Source: HGNC Symbol;Acc:HGNC:20951]
FUCA2	ENSG00000001036	alpha-L-fucosidase 2 [Source: HGNC Symbol;Acc:HGNC:4008]
GAPDHP70	ENSG00000249489	glyceraldehyde 3 phosphate dehydrogenase pseudogene 70 [Source: HGNC Symbol;Acc:HGNC:4148]
GBP7	ENSG00000213512	guanylate binding protein 7 [Source: HGNC Symbol;Acc:HGNC:29606]
GFOD1-AS1	ENSG00000237786	GFOD1 antisense RNA 1 [Source: HGNC Symbol;Acc:HGNC:40956]
GPD1	ENSG00000167588	glycerol-3-phosphate dehydrogenase 1 [Source: HGNC Symbol;Acc:HGNC:4455]
GPR132	ENSG00000183484	G protein-coupled receptor 132 [Source: HGNC Symbol;Acc:HGNC:17482]
GTF3C6	ENSG00000155130	general transcription factor IIIC subunit 6 [Source: HGNC Symbol;Acc:HGNC:20872]
HARBI1	ENSG00000180423	harbinger transposase derived 1 [Source: HGNC Symbol;Acc:HGNC:26522]
HEPHL1	ENSG00000181333	hephaestin like 1 [Source: HGNC Symbol;Acc:HGNC:30477]
HIGD2A	ENSG00000146066	HIG1 hypoxia inducible domain family member 2A [Source: HGNC Symbol;Acc:HGNC:28311]
HIRIP3	ENSG00000149929	HIRA interacting protein 3 [Source: HGNC Symbol;Acc:HGNC:4917]
HLA-F-AS1	ENSG00000214922	HLA-F antisense RNA 1 [Source: HGNC Symbol;Acc:HGNC:26645]
HNRNPA1P30	ENSG00000233780	heterogeneous nuclear ribonucleoprotein A1 pseudogene 30 [Source: HGNC Symbol;Acc:HGNC:39548]
HNRNPA1P44	ENSG00000249271	heterogeneous nuclear ribonucleoprotein A1 pseudogene 44 [Source: HGNC Symbol;Acc:HGNC:48774]
HSP90B1	ENSG00000166598	heat shock protein 90 beta family member 1 [Source: HGNC Symbol;Acc:HGNC:12028]
IGFL1	ENSG00000188293	IGF like family member 1 [Source: HGNC Symbol;Acc:HGNC:24093]
IGHG3	ENSG00000211897	immunoglobulin heavy constant gamma 3 (G3m marker) [Source: HGNC Symbol;Acc:HGNC:5527]
IPO9-AS1	ENSG00000231871	IPO9 antisense RNA 1 [Source: HGNC Symbol;Acc:HGNC:40892]
IQCC	ENSG00000160051	IQ motif containing C [Source: HGNC Symbol;Acc:HGNC:25545]
KIF11	ENSG00000138160	kinesin family member 11 [Source: HGNC Symbol;Acc:HGNC:6388]
KRT18P57	ENSG00000215867	keratin 18 pseudogene 57 [Source: HGNC Symbol;Acc:HGNC:48884]
LEF1-AS1	ENSG00000232021	LEF1 antisense RNA 1 [Source: HGNC Symbol;Acc:HGNC:40339]
LILRA6	ENSG00000244482	leukocyte immunoglobulin like receptor A6 [Source: HGNC Symbol;Acc:HGNC:15495]
LOXL4	ENSG00000138131	lysyl oxidase like 4 [Source: HGNC Symbol;Acc:HGNC:17171]
LRRC45	ENSG00000169683	leucine rich repeat containing 45 [Source: HGNC Symbol;Acc:HGNC:28302]
MAP3K14	ENSG00000006062	mitogen-activated protein kinase kinase kinase 14 [Source: HGNC Symbol;Acc:HGNC:6853]
MARCKSL1	ENSG00000175130	MARCKS like 1 [Source: HGNC Symbol;Acc:HGNC:7142]
MARS	ENSG00000166986	methionyl-tRNA synthetase [Source: HGNC Symbol;Acc:HGNC:6898]
METTL1	ENSG00000037897	methyltransferase like 1 [Source: HGNC Symbol;Acc:HGNC:7030]
MEX3A	ENSG00000254726	mex-3 RNA binding family member A [Source: HGNC Symbol;Acc:HGNC:33482]
MOGAT1	ENSG00000124003	monoacylglycerol O-acyltransferase 1 [Source: HGNC Symbol;Acc:HGNC:18210]
MSRB3	ENSG00000174100	novel transcript
MT-CO1	ENSG00000198804	mitochondrially encoded cytochrome c oxidase I [Source: HGNC Symbol;Acc:HGNC:7419]

MTRNR2L10	ENSG00000256048	novel transcript
MUC5AC	ENSG00000215182	mucin 5AC, oligomeric mucus/gel-forming [Source: HGNC Symbol;Acc:HGNC:7515]
NANOGP2	ENSG00000228670	Nanog homeobox pseudogene 2 [Source: HGNC Symbol;Acc:HGNC:23100]
NDUFA4L2	ENSG00000185633	NDUFA4, mitochondrial complex associated like 2 [Source: HGNC Symbol;Acc:HGNC:29836]
NIFK-AS1	ENSG00000236859	NIFK antisense RNA 1 [Source: HGNC Symbol;Acc:HGNC:27385]
NOP56P3	ENSG00000257956	NOP56 ribonucleoprotein pseudogene 3 [Source: HGNC Symbol;Acc:HGNC:49801]
NRM	ENSG00000137404	nurim [Source: HGNC Symbol;Acc:HGNC:8003]
NSRP1P1	ENSG00000235614	novel transcript
NUTF2P2	ENSG00000258300	nuclear transport factor 2 pseudogene 2 [Source: HGNC Symbol;Acc:HGNC:19934]
OR10AD1	ENSG00000172640	olfactory receptor family 10 subfamily AD member 1 [Source: HGNC Symbol;Acc:HGNC:14819]
OR13G1	ENSG00000197437	olfactory receptor family 13 subfamily G member 1 [Source: HGNC Symbol;Acc:HGNC:14999]
OR4C46	ENSG00000185928	novel transcript
OTUB2	ENSG00000089723	OTU deubiquitinase, ubiquitin aldehyde binding 2 [Source: HGNC Symbol;Acc:HGNC:20351]
OTUD4P1	ENSG00000118976	OTUD4 pseudogene 1 [Source: HGNC Symbol;Acc:HGNC:33912]
PABPC1L	ENSG00000101104	poly(A) binding protein cytoplasmic 1 like [Source: HGNC Symbol;Acc:HGNC:15797]
PAXBP1-AS1	ENSG00000238197	PAXBP1 antisense RNA 1 [Source: HGNC Symbol;Acc:HGNC:39603]
PCDHGC5	ENSG00000240764	protocadherin gamma subfamily C, 5 [Source: HGNC Symbol;Acc:HGNC:8718]
PCOLCE-AS1	ENSG00000224729	PCOLCE antisense RNA 1 [Source: HGNC Symbol;Acc:HGNC:40430]
PHTF1	ENSG00000116793	putative homeodomain transcription factor 1 [Source: HGNC Symbol;Acc:HGNC:8939]
PIK3CD-AS1	ENSG00000179840	PIK3CD antisense RNA 1 [Source: HGNC Symbol;Acc:HGNC:32346]
PP7080	ENSG00000188242	uncharacterized LOC25845 [Source: NCBI gene;Acc:25845]
PPL	ENSG00000118898	periplakin [Source: HGNC Symbol;Acc:HGNC:9273]
PRPS1	ENSG00000147224	phosphoribosyl pyrophosphate synthetase 1 [Source: HGNC Symbol;Acc:HGNC:9462]
PRR13P1	ENSG00000232824	proline rich 13 pseudogene 1 [Source: HGNC Symbol;Acc:HGNC:50614]
PRSS8	ENSG00000052344	serine protease 8 [Source: HGNC Symbol;Acc:HGNC:9491]
PSMB8-AS1	ENSG00000204261	PSMB8 antisense RNA 1 (head to head) [Source: HGNC Symbol;Acc:HGNC:39758]
RASA4CP	ENSG00000228903	RAS p21 protein activator 4C, pseudogene [Source: HGNC Symbol;Acc:HGNC:44185]
RELL1	ENSG00000181826	RELT like 1 [Source: HGNC Symbol;Acc:HGNC:27379]
RLIMP1	ENSG00000229456	ring finger protein, LIM domain interacting pseudogene 1 [Source: HGNC Symbol;Acc:HGNC:39682]
RNF217	ENSG00000146373	ring finger protein 217 [Source: HGNC Symbol;Acc:HGNC:21487]
RNU6V	ENSG00000206832	RNA, U6 small nuclear variant sequence with SNRPE pseudogene sequence [Source: HGNC Symbol;Acc:HGNC:10230]
RP1-310O13.7	ENSG00000226239	novel transcript
RP11-101O6.2	ENSG00000234937	proteasome (prosome, macropain) 26S subunit, non-ATPase, 7 (Mov34 homolog) (PSMD7) pseudogene
RP11-1379J22.2	ENSG00000244951	novel transcript
RP11-153N17.1	ENSG00000233961	novel transcript
RP11-15B17.1	ENSG00000245322	novel transcript
RP11-178L8.1	ENSG00000243705	ribosomal protein L39 (RPL39) pseudogene
RP11-214D15.2	ENSG00000227175	novel transcript
RP11-228B15.4	ENSG00000225032	novel transcript
RP11-23J18.1	ENSG00000258352	interferon induced transmembrane protein 3 (1-8U) (IFITM3) pseudogene
RP11-254B13.4	ENSG00000234040	ribosomal protein L10 pseudogene 12 [Source :HGNC Symbol;Acc:HGNC:52345]
RP11-281O15.7	ENSG00000253144	pseudogene similar to part of cold shock domain containing E1, RNA-binding CSDE1
RP11-293A21.2	ENSG00000248340	family with sequence similarity 64, member A (FAM64A) pseudogene
RP11-29H23.5	ENSG00000246203	novel pseudogene
RP11-307L3.4	ENSG00000233368	novel transcript
RP11-342F17.1	ENSG00000213755	ribosomal protein L29 (RPL29) pseudogene

RP11-351I21.7	ENSG00000254423	ubiquitin specific peptidase 17-like 2 (USP17L2) pseudogene
RP11-386G21.1	ENSG00000253976	novel transcript
RP11-397E7.4	ENSG00000251411	actin related protein 2/3 complex, subunit 1A, 41kDa (ARPC1A) pseudogene
RP11-409C19.2	ENSG00000253223	PRP3 pre-mRNA processing factor 3 homolog (S. cerevisiae) (PRPF3) pseudogene
RP11-460I13.2	ENSG00000227050	novel transcript
RP11-473M20.5	ENSG00000205890	novel transcript
RP11-473N11.2	ENSG00000256238	SPT16 homolog, facilitates chromatin remodeling subunit pseudogene 1 [Source: HGNC Symbol;Acc:HGNC:31388]
RP11-530C5.1	ENSG00000258048	novel transcript
RP11-603J24.7	ENSG00000237493	RAB13 member RAS oncogene family pseudogene
RP11-6B6.3	ENSG00000236942	GABA(A) receptor-associated protein (GABARAP) pseudogene
RP11-705C15.3	ENSG00000257028	novel transcript
RP11-863P13.4	ENSG00000205037	novel transcript
RP11-867G23.12	ENSG00000254756	novel transcript
RP11-996F15.2	ENSG00000257176	novel transcript
RP13-93L13.1	ENSG00000225461	novel transcript
RP3-521E19.2	ENSG00000257494	novel transcript
RP5-1063M23.1	ENSG00000250770	tetraspanin 11 (TSPAN11) pseudogene
RP5-955M13.4	ENSG00000232358	novel transcript
RPL10AP1	ENSG00000244691	ribosomal protein L10a pseudogene 1 [Source: HGNC Symbol;Acc:HGNC:19813]
RPL5P5	ENSG00000213051	ribosomal protein L5 pseudogene 5 [Source: HGNC Symbol;Acc:HGNC:35564]
RPL7P57	ENSG00000224401	ribosomal protein L7 pseudogene 57 [Source: HGNC Symbol;Acc:HGNC:35901]
RPS10P18	ENSG00000229455	ribosomal protein S10 pseudogene 18 [Source: HGNC Symbol;Acc:HGNC:36239]
RPS15A	ENSG00000134419	ribosomal protein S15a [Source: HGNC Symbol;Acc:HGNC:10389]
RPS15AP12	ENSG00000232134	ribosomal protein S15a pseudogene 12 [Source: HGNC Symbol;Acc:HGNC:36759]
RPS3AP47	ENSG00000205873	novel transcript
SEL1L	ENSG00000071537	SEL1L, ERAD E3 ligase adaptor subunit [Source: HGNC Symbol;Acc:HGNC:10717]
SEMA6A	ENSG00000092421	semaphorin 6A [Source: HGNC Symbol;Acc:HGNC:10738]
SETD9	ENSG00000155542	SET domain containing 9 [Source: HGNC Symbol;Acc:HGNC:28508]
SF3A3P2	ENSG00000254449	splicing factor 3a, subunit 3 pseudogene 2 [Source: HGNC Symbol;Acc:HGNC:23277]
SLC16A13	ENSG00000174327	solute carrier family 16 member 13 [Source: HGNC Symbol;Acc:HGNC:31037]
SLC16A6P1	ENSG00000232457	SLC16A6 pseudogene 1 [Source: HGNC Symbol;Acc:HGNC:48932]
SLC16A8	ENSG00000100156	solute carrier family 16 member 8 [Source: HGNC Symbol;Acc:HGNC:16270]
SLC25A39P1	ENSG00000226148	solute carrier family 25 member 39 pseudogene 1 [Source: HGNC Symbol;Acc:HGNC:43859]
SLC9A7	ENSG00000065923	solute carrier family 9 member A7 [Source: HGNC Symbol;Acc:HGNC:17123]
SNORA65	ENSG00000201302	small nucleolar RNA, H/ACA box 65 [Source: HGNC Symbol;Acc:HGNC:10222]
SNORD45	ENSG00000200706	novel transcript
SNORD99	ENSG00000221539	small nucleolar RNA, C/D box 99 [Source: HGNC Symbol;Acc:HGNC:32762]
SPDYC	ENSG00000204710	speedy/RINGO cell cycle regulator family member C [Source: HGNC Symbol;Acc:HGNC:32681]
SPNS2	ENSG00000183018	sphingolipid transporter 2 [Source: HGNC Symbol;Acc:HGNC:26992]
SRP72P2	ENSG00000188451	signal recognition particle 72 pseudogene 2 [Source: HGNC Symbol;Acc:HGNC:31096]
STARD4-AS1	ENSG00000246859	STARD4 antisense RNA 1 [Source: HGNC Symbol;Acc:HGNC:44117]
SYCP3	ENSG00000139351	synaptonemal complex protein 3 [Source: HGNC Symbol;Acc:HGNC:18130]
TADA1	ENSG00000152382	transcriptional adaptor 1 [Source: HGNC Symbol;Acc:HGNC:30631]
TCTN1	ENSG00000204852	tectonic family member 1 [Source: HGNC Symbol;Acc:HGNC:26113]
TERF2	ENSG00000132604	telomeric repeat binding factor 2 [Source: HGNC Symbol;Acc:HGNC:11729]
TGFB3	ENSG00000119699	transforming growth factor beta 3 [Source: HGNC Symbol;Acc:HGNC:11769]
TIMM22	ENSG00000177370	translocase of inner mitochondrial membrane 22 [Source: HGNC Symbol;Acc:HGNC:17317]
TOP1	ENSG00000198900	DNA topoisomerase I [Source: HGNC Symbol;Acc:HGNC:11986]

TRIM75P	ENSG00000250374	tripartite motif containing 75, pseudogene [Source: HGNC Symbol;Acc:HGNC:32686]
UBE2L2	ENSG00000131982	ubiquitin conjugating enzyme E2 L2 (pseudogene) [Source: HGNC Symbol;Acc:HGNC:12487]
WDR55	ENSG00000120314	WD repeat domain 55 [Source: HGNC Symbol;Acc:HGNC:25971]
WDR88	ENSG00000166359	WD repeat domain 88 [Source: HGNC Symbol;Acc:HGNC:26999]
ZACN	ENSG00000186919	zinc activated ion channel [Source: HGNC Symbol;Acc:HGNC:29504]
ZBED2	ENSG00000177494	zinc finger BED-type containing 2 [Source: HGNC Symbol;Acc:HGNC:20710]
ZBTB2	ENSG00000181472	zinc finger and BTB domain containing 2 [Source: HGNC Symbol;Acc:HGNC:20868]
ZNF181	ENSG00000197841	zinc finger protein 181 [Source: HGNC Symbol;Acc:HGNC:12971]
ZNF213	ENSG00000085644	zinc finger protein 213 [Source: HGNC Symbol;Acc:HGNC:13005]
ZNF354C	ENSG00000177932	zinc finger protein 354C [Source: HGNC Symbol;Acc:HGNC:16736]
ZNF75BP	ENSG00000258212	zinc finger protein 75B, pseudogene [Source: HGNC Symbol;Acc:HGNC:13147]
ZSCAN16	ENSG00000196812	zinc finger and SCAN domain containing 16 [Source: HGNC Symbol;Acc:HGNC:20813]

- Bolded are the transcripts overlapping between DMFS &BCSS identified by Artificial Neural Network analysis.

Supplementary (C) Table B: Genes associated with Breast Cancer Specific Survival (BCSS) identified by Artificial Neural Network analysis (Transcriptomic Cohort, n=112)

Gene ID	initial_alias	description
ABCA4	ENSG00000198691	ATP binding cassette subfamily A member 4 [Source:HGNC Symbol;Acc:HGNC:34]
AC003075.4	ENSG00000237773	novel transcript
AC004951.6	ENSG00000228434	novel transcript
AC020931.1	ENSG00000257110	novel transcript
AC072052.7	ENSG00000231360	novel transcript
AC073834.3	ENSG00000237655	novel transcript, antisense to TTC30A
AC079305.10	ENSG00000222043	novel transcript
AC080125.1	ENSG00000225406	pseudogene similar to part of E74-like factor 2 (ets domain transcription factor) (ELF2)
AC084809.2	ENSG00000226377	novel transcript
ACBD6	ENSG00000230124	acyl-CoA binding domain containing 6 [Source:HGNC Symbol;Acc:HGNC:23339]
ACSM4	ENSG00000215009	acyl-CoA synthetase medium chain family member 4 [Source:HGNC Symbol;Acc:HGNC:32016]
ADAMTS5	ENSG00000154736	ADAM metallopeptidase with thrombospondin type 1 motif 5 [Source:HGNC Symbol;Acc:HGNC:221]
ADORA2B	ENSG00000170425	adenosine A2b receptor [Source:HGNC Symbol;Acc:HGNC:264]
AKAP8L	ENSG00000011243	A-kinase anchoring protein 8 like [Source:HGNC Symbol;Acc:HGNC:29857]
ALG1L5P	ENSG00000226943	asparagine-linked glycosylation 1-like 5, pseudogene [Source:HGNC Symbol;Acc:HGNC:44374]
ANKRD34B	ENSG00000189127	ankyrin repeat domain 34B [Source:HGNC Symbol;Acc:HGNC:33736]
ANO5	ENSG00000171714	anoctamin 5 [Source:HGNC Symbol;Acc:HGNC:27337]
ASXL3	ENSG00000141431	ASXL transcriptional regulator 3 [Source:HGNC Symbol;Acc:HGNC:29357]
ATP1B2	ENSG00000129244	ATPase Na+/K+ transporting subunit beta 2 [Source:HGNC Symbol;Acc:HGNC:805]
B3GNT5	ENSG00000176597	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 5 [Source:HGNC Symbol;Acc:HGNC:15684]
BICC1	ENSG00000122870	BicC family RNA binding protein 1 [Source:HGNC Symbol;Acc:HGNC:19351]
BLVRB	ENSG00000090013	biliverdin reductase B [Source:HGNC Symbol;Acc:HGNC:1063]
BMS1	ENSG00000165733	BMS1, ribosome biogenesis factor [Source:HGNC Symbol;Acc:HGNC:23505]
C11orf24	ENSG00000171067	chromosome 11 open reading frame 24 [Source:HGNC Symbol;Acc:HGNC:1174]
C19orf12	ENSG00000131943	chromosome 19 open reading frame 12 [Source:HGNC Symbol;Acc:HGNC:25443]
C1orf147	ENSG00000162888	chromosome 1 open reading frame 147 [Source:HGNC Symbol;Acc:HGNC:32061]
C1QBP	ENSG00000108561	complement C1q binding protein [Source:HGNC Symbol;Acc:HGNC:1243]
C2orf70	ENSG00000173557	chromosome 2 open reading frame 70 [Source:HGNC Symbol;Acc:HGNC:27938]
CCDC54	ENSG00000138483	coiled-coil domain containing 54 [Source:HGNC Symbol;Acc:HGNC:30703]
CEP350	ENSG00000135837	centrosomal protein 350 [Source:HGNC Symbol;Acc:HGNC:24238]
CGNL1	ENSG00000128849	cingulin like 1 [Source:HGNC Symbol;Acc:HGNC:25931]
CMB9-22P13.1	ENSG00000173727	Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed (FAU) pseudogene
CMYA5	ENSG00000164309	cardiomyopathy associated 5 [Source:HGNC Symbol;Acc:HGNC:14305]
CNTN4-AS1	ENSG00000237990	CNTN4 antisense RNA 1 [Source:HGNC Symbol;Acc:HGNC:39985]
CNTNAP2	ENSG00000174469	contactin associated protein like 2 [Source:HGNC Symbol;Acc:HGNC:13830]
CSDC2	ENSG00000172346	cold shock domain containing C2 [Source:HGNC Symbol;Acc:HGNC:30359]
CTAGE12P	ENSG00000215441	CTAGE family member 12, pseudogene [Source:HGNC Symbol;Acc:HGNC:37297]
CTC-498J12.3	ENSG00000248664	novel transcript
CTD-2026K11.6	ENSG00000203394	novel transcript
CTD-2228K2.2	ENSG00000214278	mitochondrial translational initiation factor 3 (MTIF3) pseudogene
CTD-2290C23.2	ENSG00000241739	ribosomal protein L21 (RPL21) pseudogene
CTNS	ENSG00000040531	cystinosin, lysosomal cystine transporter [Source:HGNC Symbol;Acc:HGNC:2518]
CUBN	ENSG00000107611	cubilin [Source:HGNC Symbol;Acc:HGNC:2548]
CYP1B1-AS1	ENSG00000232973	CYP1B1 antisense RNA 1 [Source:HGNC Symbol;Acc:HGNC:28543]
DCTN1-AS1	ENSG00000237737	DCTN1 antisense RNA 1 [Source:HGNC Symbol;Acc:HGNC:44151]
DDA1	ENSG00000130311	DET1 and DDB1 associated 1 [Source:HGNC Symbol;Acc:HGNC:28360]
DNAH7	ENSG00000118997	dynein axonemal heavy chain 7 [Source:HGNC Symbol;Acc:HGNC:18661]
DNAJC17	ENSG00000104129	DnaJ heat shock protein family (Hsp40) member C17 [Source:HGNC Symbol;Acc:HGNC:25556]

DOCK10	ENSG00000135905	dedicator of cytokinesis 10 [Source:HGNC Symbol;Acc:HGNC:23479]
DST	ENSG00000151914	dystonin [Source:HGNC Symbol;Acc:HGNC:1090]
EEF1D	ENSG00000104529	eukaryotic translation elongation factor 1 delta [Source:HGNC Symbol;Acc:HGNC:3211]
EGFR	ENSG00000146648	epidermal growth factor receptor [Source:HGNC Symbol;Acc:HGNC:3236]
ERV3-1	ENSG00000213462	endogenous retrovirus group 3 member 1, envelope [Source:HGNC Symbol;Acc:HGNC:3454]
ETV3	ENSG00000117036	ETS variant 3 [Source:HGNC Symbol;Acc:HGNC:3492]
EXOSC5	ENSG00000077348	exosome component 5 [Source:HGNC Symbol;Acc:HGNC:24662]
F8	ENSG00000185010	coagulation factor VIII [Source:HGNC Symbol;Acc:HGNC:3546]
FAM151A	ENSG00000162391	family with sequence similarity 151 member A [Source:HGNC Symbol;Acc:HGNC:25032]
FAM47C	ENSG00000198173	family with sequence similarity 47 member C [Source:HGNC Symbol;Acc:HGNC:25301]
FAM86KP	ENSG00000163612	family with sequence similarity 86 member K, pseudogene [Source:HGNC Symbol;Acc:HGNC:44098]
FAM86MP	ENSG00000186234	family with sequence similarity 86 member M, pseudogene [Source:HGNC Symbol;Acc:HGNC:44100]
FANCF	ENSG00000183161	FA complementation group F [Source:HGNC Symbol;Acc:HGNC:3587]
FANCG	ENSG00000221834	novel transcript
FCRLB	ENSG00000162746	Fc receptor like B [Source:HGNC Symbol;Acc:HGNC:26431]
FIGN	ENSG00000182263	fidgetin, microtubule severing factor [Source:HGNC Symbol;Acc:HGNC:13285]
FMR1-IT1	ENSG00000236338	novel transcript
FOXD1-AS1	ENSG00000248003	novel transcript
FOXD3	ENSG00000187140	forkhead box D3 [Source:HGNC Symbol;Acc:HGNC:3804]
FZD5	ENSG00000163251	frizzled class receptor 5 [Source:HGNC Symbol;Acc:HGNC:4043]
GIT1	ENSG00000108270	novel transcript
GPR183	ENSG00000169508	G protein-coupled receptor 183 [Source:HGNC Symbol;Acc:HGNC:3128]
GPR33	ENSG00000214943	G protein-coupled receptor 33 (gene/pseudogene) [Source:HGNC Symbol;Acc:HGNC:4489]
GPRIN3	ENSG00000185477	GPRIN family member 3 [Source:HGNC Symbol;Acc:HGNC:27733]
GS1-251I9.3	ENSG00000253358	novel transcript
GTF3C6	ENSG00000155130	general transcription factor IIIC subunit 6 [Source:HGNC Symbol;Acc:HGNC:20872]
GUCY2F	ENSG00000101890	guanylate cyclase 2F, retinal [Source:HGNC Symbol;Acc:HGNC:4691]
HAMP	ENSG00000105697	hepcidin antimicrobial peptide [Source:HGNC Symbol;Acc:HGNC:15598]
HARBI1	ENSG00000180423	harbinger transposase derived 1 [Source:HGNC Symbol;Acc:HGNC:26522]
HDDC2	ENSG00000111906	HD domain containing 2 [Source:HGNC Symbol;Acc:HGNC:21078]
HEPACAM	ENSG00000165478	hepatic and glial cell adhesion molecule [Source:HGNC Symbol;Acc:HGNC:26361]
HEXDC	ENSG00000169660	hexosaminidase D [Source:HGNC Symbol;Acc:HGNC:26307]
IGFBP5	ENSG00000115461	insulin like growth factor binding protein 5 [Source:HGNC Symbol;Acc:HGNC:5474]
IGKV1OR10-1	ENSG00000237592	immunoglobulin kappa variable 1/OR10-1 (pseudogene) [Source:HGNC Symbol;Acc:HGNC:44978]
IKZF5	ENSG00000095574	IKAROS family zinc finger 5 [Source:HGNC Symbol;Acc:HGNC:14283]
ITGA9	ENSG00000144668	integrin subunit alpha 9 [Source:HGNC Symbol;Acc:HGNC:6145]
KIAA0355	ENSG00000166398	KIAA0355 [Source:HGNC Symbol;Acc:HGNC:29016]
KIAA1462	ENSG00000165757	junctional cadherin 5 associated [Source:HGNC Symbol;Acc:HGNC:29283]
KLHL41	ENSG00000239474	kelch like family member 41 [Source:HGNC Symbol;Acc:HGNC:16905]
KRTAP3-3	ENSG00000212899	keratin associated protein 3-3 [Source:HGNC Symbol;Acc:HGNC:18890]
LGALS4	ENSG00000171747	galectin 4 [Source:HGNC Symbol;Acc:HGNC:6565]
LINC01074	ENSG00000227612	novel transcript
LIPE	ENSG00000079435	lipase E, hormone sensitive type [Source:HGNC Symbol;Acc:HGNC:6621]
LRRC3C	ENSG00000204913	leucine rich repeat containing 3C [Source:HGNC Symbol;Acc:HGNC:40034]
MAP2K6	ENSG00000108984	mitogen-activated protein kinase kinase 6 [Source:HGNC Symbol;Acc:HGNC:6846]
MAP3K4	ENSG00000085511	mitogen-activated protein kinase kinase kinase 4 [Source:HGNC Symbol;Acc:HGNC:6856]
MED14	ENSG00000180182	mediator complex subunit 14 [Source:HGNC Symbol;Acc:HGNC:2370]
MED22	ENSG00000148297	mediator complex subunit 22 [Source:HGNC Symbol;Acc:HGNC:11477]
MEX3A	ENSG00000254726	mex-3 RNA binding family member A [Source:HGNC Symbol;Acc:HGNC:33482]
MIR199A1	ENSG00000207752	microRNA 199a-1 [Source:HGNC Symbol;Acc:HGNC:31571]
MMP25	ENSG00000008516	matrix metallopeptidase 25 [Source:HGNC Symbol;Acc:HGNC:14246]
MPND	ENSG00000008382	MPN domain containing [Source:HGNC Symbol;Acc:HGNC:25934]
MRGBP	ENSG00000101189	MRG domain binding protein [Source:HGNC Symbol;Acc:HGNC:15866]
NAPA	ENSG00000105402	NSF attachment protein alpha [Source:HGNC Symbol;Acc:HGNC:7641]
NDUFA4L2	ENSG00000185633	NDUFA4, mitochondrial complex associated like 2 [Source:HGNC Symbol;Acc:HGNC:29836]

NIPSNAP1	ENSG00000184117	nipsnap homolog 1 [Source:HGNC Symbol;Acc:HGNC:7827]
NLRP5	ENSG00000171487	NLR family pyrin domain containing 5 [Source:HGNC Symbol;Acc:HGNC:21269]
NPAP1	ENSG00000185823	nuclear pore associated protein 1 [Source:HGNC Symbol;Acc:HGNC:1190]
ODF2-AS1	ENSG00000225951	ODF2 antisense RNA 1 [Source:HGNC Symbol;Acc:HGNC:49461]
OR10K1	ENSG00000173285	olfactory receptor family 10 subfamily K member 1 [Source:HGNC Symbol;Acc:HGNC:14693]
OR1I1	ENSG00000094661	olfactory receptor family 1 subfamily I member 1 [Source:HGNC Symbol;Acc:HGNC:8207]
OR1S1	ENSG00000172774	novel transcript
OR2A1	ENSG00000221970	olfactory receptor family 2 subfamily A member 1 [Source:HGNC Symbol;Acc:HGNC:8229]
OR7A17	ENSG00000185385	olfactory receptor family 7 subfamily A member 17 [Source:HGNC Symbol;Acc:HGNC:8363]
OR8B4	ENSG00000198657	novel transcript
PAXBP1-AS1	ENSG00000238197	PAXBP1 antisense RNA 1 [Source:HGNC Symbol;Acc:HGNC:39603]
PAXIP1-AS2	ENSG00000214106	PAXIP1 antisense RNA 2 [Source:HGNC Symbol;Acc:HGNC:48958]
PCDHB1	ENSG00000171815	protocadherin beta 1 [Source:HGNC Symbol;Acc:HGNC:8680]
PDIA4	ENSG00000155660	protein disulfide isomerase family A member 4 [Source:HGNC Symbol;Acc:HGNC:30167]
PHOSPHO1	ENSG00000173868	phosphoethanolamine/phosphocholine phosphatase [Source:HGNC Symbol;Acc:HGNC:16815]
POM121L3P	ENSG00000167390	POM121 transmembrane nucleoporin like 3, pseudogene [Source:HGNC Symbol;Acc:HGNC:16440]
PPIAP31	ENSG00000217094	peptidylprolyl isomerase A pseudogene 31 [Source:HGNC Symbol;Acc:HGNC:44962]
PPL	ENSG00000118898	periplakin [Source:HGNC Symbol;Acc:HGNC:9273]
QRSL1P3	ENSG00000257957	QRSL1 pseudogene 3 [Source:HGNC Symbol;Acc:HGNC:43669]
RLBP1	ENSG00000140522	retinaldehyde binding protein 1 [Source:HGNC Symbol;Acc:HGNC:10024]
RNF5	ENSG00000204308	ring finger protein 5 [Source:HGNC Symbol;Acc:HGNC:10068]
ROPN1B	ENSG00000114547	rhophilin associated tail protein 1B [Source:HGNC Symbol;Acc:HGNC:31927]
RP1-274L7.1	ENSG00000229702	novel transcript
RP11-252O18.3	ENSG00000213155	zinc finger, CCHC domain containing 10 (ZCCHC10) pseudogene
RP11-262H14.11	ENSG00000219693	fibroblast growth factor 7 pseudogene 8 [Source:HGNC Symbol;Acc:HGNC:34516]
RP11-286N22.8	ENSG00000256591	novel transcript
RP11-298I3.1	ENSG00000257285	novel transcript
RP11-29H23.5	ENSG00000246203	novel transcript
RP11-409C19.2	ENSG00000253223	PRP3 pre-mRNA processing factor 3 homolog (S. cerevisiae) (PRPF3) pseudogene
RP11-429J17.5	ENSG00000254549	novel transcript
RP11-466G12.2	ENSG00000249758	novel transcript
RP11-542G1.3	ENSG00000251384	novel transcript
RP11-544M22.3	ENSG00000232879	glutaredoxin 5 (GLRX5) pseudogene
RP11-592N21.1	ENSG00000212673	novel transcript
RP11-619A14.2	ENSG00000254933	novel transcript
RP11-798K23.3	ENSG00000251545	novel pseudogene
RP11-967K21.2	ENSG00000255953	novel transcript
RP3-508I15.9	ENSG00000228274	novel transcript, antisense to CBY1
RP4-622L5.7	ENSG00000224066	novel transcript
RPL15P18	ENSG00000228501	ribosomal protein L15 pseudogene 18 [Source:HGNC Symbol;Acc:HGNC:36515]
RPL21P134	ENSG00000233254	ribosomal protein L21 pseudogene 134 [Source:HGNC Symbol;Acc:HGNC:36006]
RPL27A	ENSG00000166441	ribosomal protein L27a [Source:HGNC Symbol;Acc:HGNC:10329]

RPL32P34	ENSG00000239524	ribosomal protein L32 pseudogene 34 [Source:HGNC Symbol;Acc:HGNC:35903]
RPL3P7	ENSG00000225093	ribosomal protein L3 pseudogene 7 [Source:HGNC Symbol;Acc:HGNC:36797]
RPS10P18	ENSG00000229455	ribosomal protein S10 pseudogene 18 [Source:HGNC Symbol;Acc:HGNC:36239]
RPS3AP47	ENSG00000205873	novel transcript
RPS8	ENSG00000142937	ribosomal protein S8 [Source:HGNC Symbol;Acc:HGNC:10441]
SCN10A	ENSG00000185313	sodium voltage-gated channel alpha subunit 10 [Source:HGNC Symbol;Acc:HGNC:10582]
SDPR	ENSG00000168497	caveolae associated protein 2 [Source:HGNC Symbol;Acc:HGNC:10690]
SIPA1L3	ENSG00000105738	signal induced proliferation associated 1 like 3 [Source:HGNC Symbol;Acc:HGNC:23801]
SKP1P1	ENSG00000231234	S-phase kinase associated protein 1 pseudogene 1 [Source:HGNC Symbol;Acc:HGNC:33696]
SLC27A4	ENSG00000167114	solute carrier family 27 member 4 [Source:HGNC Symbol;Acc:HGNC:10998]
SLC39A4	ENSG00000147804	solute carrier family 39 member 4 [Source:HGNC Symbol;Acc:HGNC:17129]
SLC7A9	ENSG00000021488	solute carrier family 7 member 9 [Source:HGNC Symbol;Acc:HGNC:11067]
SMIM5	ENSG00000204323	small integral membrane protein 5 [Source:HGNC Symbol;Acc:HGNC:40030]
SNORD99	ENSG00000221539	small nucleolar RNA, C/D box 99 [Source:HGNC Symbol;Acc:HGNC:32762]
SNRPGP15	ENSG00000224543	small nuclear ribonucleoprotein polypeptide G pseudogene 15 [Source:HGNC Symbol;Acc:HGNC:49371]
SNU13	ENSG00000100138	small nuclear ribonucleoprotein 13 [Source:HGNC Symbol;Acc:HGNC:7819]
SORBS2	ENSG00000154556	sorbin and SH3 domain containing 2 [Source:HGNC Symbol;Acc:HGNC:24098]
SPAG7	ENSG00000091640	sperm associated antigen 7 [Source:HGNC Symbol;Acc:HGNC:11216]
SPATA31A6	ENSG00000185775	SPATA31 subfamily A member 6 [Source:HGNC Symbol;Acc:HGNC:32006]
SPDYC	ENSG00000204710	speedy/RINGO cell cycle regulator family member C [Source:HGNC Symbol;Acc:HGNC:32681]
SPRED1	ENSG00000166068	sprouty related EVH1 domain containing 1 [Source:HGNC Symbol;Acc:HGNC:20249]
SRP72P2	ENSG00000188451	signal recognition particle 72 pseudogene 2 [Source:HGNC Symbol;Acc:HGNC:31096]
ST13P19	ENSG00000228110	ST13, Hsp70 interacting protein pseudogene 19 [Source:HGNC Symbol;Acc:HGNC:38862]
ST6GAL1	ENSG00000073849	ST6 beta-galactoside alpha-2,6-sialyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:10860]
STAT1	ENSG00000115415	signal transducer and activator of transcription 1 [Source:HGNC Symbol;Acc:HGNC:11362]
SYCE1L	ENSG00000205078	synaptonemal complex central element protein 1 like [Source:HGNC Symbol;Acc:HGNC:37236]
SYNGR1	ENSG00000100321	synaptogyrin 1 [Source:HGNC Symbol;Acc:HGNC:11498]
SZT2	ENSG00000198198	SZT2, KICSTOR complex subunit [Source:HGNC Symbol;Acc:HGNC:29040]
TBX1	ENSG00000184058	T-box 1 [Source:HGNC Symbol;Acc:HGNC:11592]
TDRKH	ENSG00000182134	tudor and KH domain containing [Source:HGNC Symbol;Acc:HGNC:11713]
TEX13B	ENSG00000170925	testis expressed 13B [Source:HGNC Symbol;Acc:HGNC:11736]
TMA16P2	ENSG00000232467	translation machinery associated 16 homolog pseudogene 2 [Source:HGNC Symbol;Acc:HGNC:43781]
TMCO1	ENSG00000143183	transmembrane and coiled-coil domains 1 [Source:HGNC Symbol;Acc:HGNC:18188]
TMED10	ENSG00000170348	transmembrane p24 trafficking protein 10 [Source:HGNC Symbol;Acc:HGNC:16998]
TMEM78	ENSG00000177800	transmembrane protein 78 [Source:HGNC Symbol;Acc:HGNC:32307]
TMEM80	ENSG00000177042	transmembrane protein 80 [Source:HGNC Symbol;Acc:HGNC:27453]
TOR1AIP2	ENSG00000169905	torsin 1A interacting protein 2 [Source:HGNC Symbol;Acc:HGNC:24055]
TRBC2	ENSG00000211772	T cell receptor beta constant 2 [Source:HGNC Symbol;Acc:HGNC:12157]
TTPA	ENSG00000137561	alpha tocopherol transfer protein [Source:HGNC Symbol;Acc:HGNC:12404]
TUBA3E	ENSG00000152086	tubulin alpha 3e [Source:HGNC Symbol;Acc:HGNC:20765]
TUBB3P1	ENSG00000220418	tubulin beta 3 class III pseudogene 1 [Source:HGNC Symbol;Acc:HGNC:42339]
U82670.9	ENSG00000229979	high-mobility group nucleosomal binding domain 2 (HMGN2) pseudogene
UPP2	ENSG00000007001	uridine phosphorylase 2 [Source:HGNC Symbol;Acc:HGNC:23061]
URB2	ENSG00000135763	URB2 ribosome biogenesis homolog [Source:HGNC Symbol;Acc:HGNC:28967]
USP51	ENSG00000247746	ubiquitin specific peptidase 51 [Source:HGNC Symbol;Acc:HGNC:23086]
VSIG8	ENSG00000243284	V-set and immunoglobulin domain containing 8 [Source:HGNC Symbol;Acc:HGNC:32063]
XXbac-B444P24.10	ENSG00000161132	proline dehydrogenase (oxidase) 1 (PRODH) pseudogene
YIF1B	ENSG00000167645	Yip1 interacting factor homolog B, membrane trafficking protein [Source:HGNC Symbol;Acc:HGNC:30511]
ZCCHC2	ENSG00000141664	zinc finger CCHC-type containing 2 [Source:HGNC Symbol;Acc:HGNC:22916]
ZMIZ1	ENSG00000108175	zinc finger MIZ-type containing 1 [Source:HGNC Symbol;Acc:HGNC:16493]
ZNF397	ENSG00000186812	zinc finger protein 397 [Source:HGNC Symbol;Acc:HGNC:18818]

ZNF446	ENSG00000083838	zinc finger protein 446 [Source:HGNC Symbol;Acc:HGNC:21036]
ZNF609	ENSG00000180357	zinc finger protein 609 [Source:HGNC Symbol;Acc:HGNC:29003]
ZNF804A	ENSG00000170396	zinc finger protein 804A [Source:HGNC Symbol;Acc:HGNC:21711]
ZSWIM1	ENSG00000168614	novel transcript

References

1 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21.

2 Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 2015;31:166–169.

3 Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics 2010;26:493–500.

4 Abd El-Rehim DM, Ball G, Pinder SE, Rakha E, Paish C, Robertson JFR, et al. High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. Int J Cancer 2005;116:340–350.

5 McCarty KS, Miller LS, Cox EB, Konrath J, McCarty KS. Estrogen receptor analyses. Correlation of biochemical and immunohistochemical methods using monoclonal antireceptor antibodies. Arch Pathol Lab Med 1985;109:716–721.

6 Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ 1994;308:1552.

7 Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ 1994;309:102.

8 Bhargava EK, Rathore PK, Raj A, Meher R, Rana K. Diagnostic Efficacy of Computed Tomography in Detecting Cervical Metastases in Clinically N0 Head and Neck Squamous Cell Carcinoma. Indian J Otolaryngol Head Neck Surg 2016;68:25–29.

9 Liu Y, Beyer A, Aebersold R. Leading Edge Review On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell 2016;165:535–550.

Table 1: Multivariate Cox regression analysis to generate a prognostic score for the two gene signature predicting Breast Cancer Specific Survival (BCSS) and Distant Metastasis-Free Survival (DMFS) (Transcriptomic, n=112)

Covariates	Breast Cancer specific Survival					Distant Metastasis Free Survival				
	(B) value	P-value	Hazard Ratio	95% CI		(B) value	P-value	Hazard Ratio	95% CI	
				Lower	Upper				Lower	Upper
ACSM4	**1.111**	**< 0.001**	3.038	1.653	5.585	**1.065**	**0.001**	2.900	1.570	5.358
SPDYC	**0.745**	**0.026**	2.016	1.092	4.063	**0.833**	**0.016**	2.300	1.166	4.535

Significant P-values are bolded CI: Confidence interval

Table 2: Multivariate Cox regression analysis for the two gene signature predicting Breast Cancer Specific Survival (BCSS) and Distant Metastasis-Free Survival (DMFS) (Transcriptomic Cohort, n=112)

Covariates	**Breast Cancer-Specific Survival**				**Distant Metastasis-Free Survival**			
	P-value	**Hazard Ratio**	**95% CI**		**P-Value**	**Hazard Ratio**	**95% CI**	
			Lower	**Upper**			**Lower**	**Upper**
Age	**0.011**	2.343	1.217	4.513	**0.015**	2.240	1.169	4.294
Tumor Size	0.085	0.540	0.267	1.089	**0.043**	0.490	0.246	0.979
Grade	0.182	2.402	0.663	8.705	0.175	2.454	0.671	8.981
Vascular Invasion	0.694	1.188	1.053	2.805	0.852	1.084	0.466	2.519
Nodal Stage	**0.033**	1.907	1.053	3.454	**0.043**	1.836	1.020	3.307
two gene signature at mRNA level	**< 0.001**	3.891	2.041	7.416	**< 0.001**	3.371	1.780	6.384

- Significant P-values are bolded CI: Confidence interval

Table 3 (A)*:* Multivariate Cox regression analysis for Individual Potential proteins associated with Breast Cancer Specific Survival (BCSS) and Distant Metastasis-Free Survival (DMFS) (IHC Cohort, n=333)

Covariates	Breast Cancer specific Survival				Distant Metastasis Free Survival			
	P-value	Hazard Ratio	95% CI		P-value	Hazard Ratio	95% CI	
			Lower	Upper			Lower	Upper
Age	**0.011**	1.941	1.165	3.234	**0.030**	1.743	1.055	2.879
Tumor Size	0.706	0.908	0.551	1.497	0.619	1.136	0.786	1.879
Grade	0.494	1.267	0.643	2.498	0.385	1.369	0.674	2.782
Vascular Invasion	0.062	1.697	0.975	2.953	0.178	1.466	0.840	2.558
Nodal Stage	0.001	1.794	1.261	1.552	0.027	1.513	1.049	2.180
ACSM4	0.057	1.698	0.983	2.933	**0.002**	2.267	1.350	3.809

- Significant P-values are bolded CI: Confidence interval

Table 3 (B): Multivariate Cox regression analysis for Individual Potential proteins associated with Breast Cancer Specific Survival (BCSS) and Distant Metastasis-Free Survival (DMFS) (IHC Cohort, n=333)

Covariates	Breast Cancer specific Survival				Distant Metastasis Free Survival			
	P-value	Hazard Ratio	95% CI		P-value	Hazard Ratio	95% CI	
			Lower	Upper			Lower	Upper
Age	**0.026**	1.711	1.067	2.745	**0.042**	1.618	1.016	2.576
Tumor Size	0.467	0.839	0.522	1.348	0.922	1.024	0.638	1.644
Grade	0.500	1.292	0.614	2.722	0.422	1.387	0.624	3.079
Vascular Invasion	**0.034**	1.783	1.041	3.044	0.086	1.596	0.935	2.724
Nodal Stage	**0.002**	1.757	1.240	2.489	**0.009**	1.600	1.123	2.280
SPDYC	**0.015**	2.377	1.181	4.783	**0.015**	2.365	1.178	4.748

- Significant P-values are bolded CI: Confidence interval

Table 4: Combined multivariate Cox regression analysis for potential proteins associated with Breast Cancer Specific Survival (BCSS) and Distant Metastasis-Free Survival (DMFS) (IHC Cohort, n=333)

Covariates	Breast Cancer-specific Survival				Distant Metastasis-Free Survival			
	P-value	Hazard Ratio	95% CI		P-value	Hazard Ratio	95% CI	
			Lower	Upper			Lower	Upper
Age	**0.049**	1.733	1.003	2.995	0.105	1.548	0.913	2.624
Tumor Size	0.282	0.741	0.430	1.279	0.942	0.980	0.573	1.678
Grade	0.754	1.128	0.530	2.393	0.641	1.210	0.542	2.700
Vascular Invasion	**0.022**	2.018	1.108	3.676	0.077	1.702	0.945	3.067
Nodal Stage	**0.012**	1.640	1.114	2.414	0.051	1.477	0.988	2.185
ACSM4	0.274	1.402	0.766	2.568	**0.036**	1.826	1.014	3.204
SPDYC	**0.031**	2.545	1.086	5.960	**0.034**	2.508	1.072	5.869

- Significant P-values are bolded CI: Confidence interval

Table 5: Multivariate Cox regression analysis to build prognostic index for the two gene signature predicting Breast Cancer Specific Survival (BCSS) and Distant Metastasis-Free Survival (DMFS) (IHC Cohort, n=333)

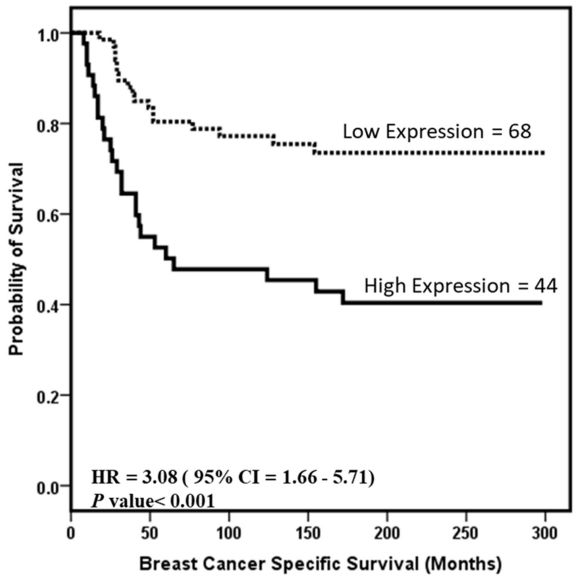
Covariates	Breast Cancer specific Survival					Distant Metastasis Free Survival				
	(B) value	P-value	Hazard Ratio	95.0% CI		(B) value	P-value	Hazard Ratio	95.0% CI	
				Lower	Upper				Lower	Upper
ACSM4	**0.46**	0.129	1.587	0.874	2.883	**0.69**	**0.017**	1.972	1.130	3.440
SPDYC	**1.05**	**0.014**	2.869	1.232	6.678	**1.05**	**0.015**	2.859	1.229	6.651

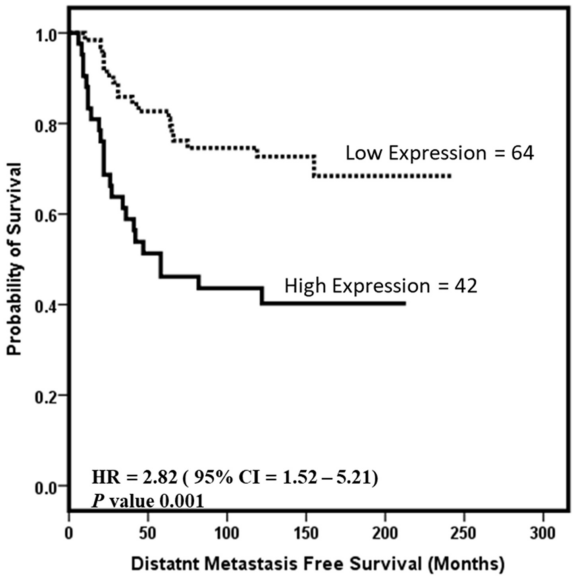
- Significant P-values are bolded CI: Confidence interval

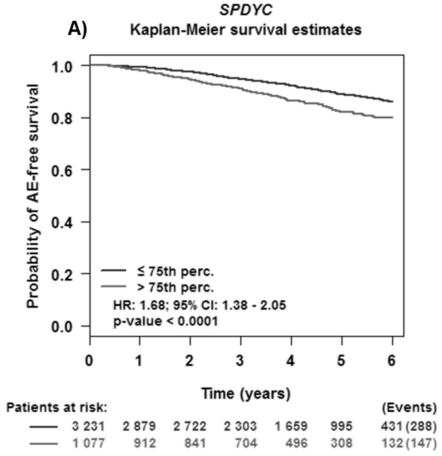
Table 6: Multivariate Cox regression analysis for the protein expression prognostic score predicting Breast Cancer Specific Survival *(BCSS)* and Distant Metastasis-Free Survival *(DMFS)* (IHC Cohort, n=333)

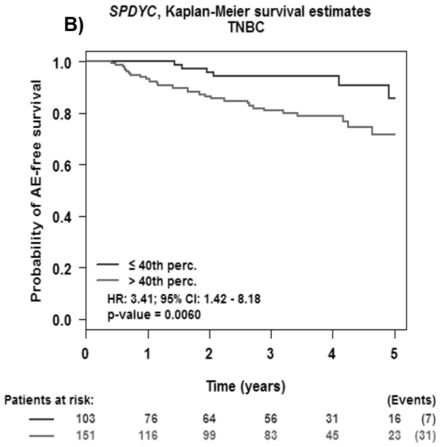
Covariates	**Breast Cancer specific Survival**				**Distant Metastasis Free Survival**			
	Significance	Hazard Ratio	95.0% CI		Significance	Hazard Ratio	95.0% CI	
			Lower	Upper			Lower	Upper
Age	**0.037**	1.782	1.037	3.064	0.080	01.597	0.946	2.698
Tumour Size	0.237	0.721	0.419	1.240	0.796	0.932	0.545	1.592
Grade	0.734	1.134	0.549	2.343	0.607	1.227	0.563	2.675
Vascular Invasion	**0.018**	2.041	1.130	3.684	**0.047**	1.802	1.009	3.219
Nodal Stage	**0.009**	1.677	1.139	2.468	**0.040**	1.509	1.019	2.235
Two gene signature at protein level	0.077	1.637	0.946	2.960	**0.034**	1.867	1.049	3.323

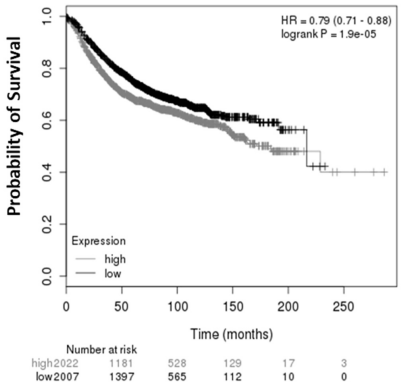
- Significant P-values are bolded CI: Confidence interval

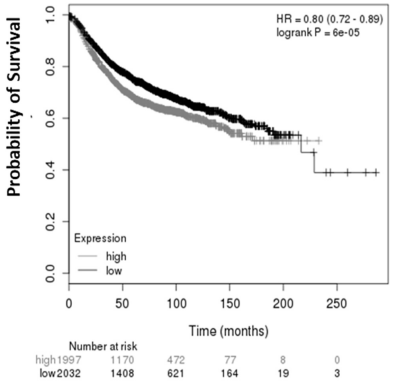

Probability of Survival
1.0
0.8
0.6
0.4
0.2
0.0
Low Expression = 68
High Expression = 44
HR = 3.08 (95% CI = 1.66 - 5.71)
P value< 0.001
0
50
100
150
200
250
300
Breast Cancer Specific Survival (Months)

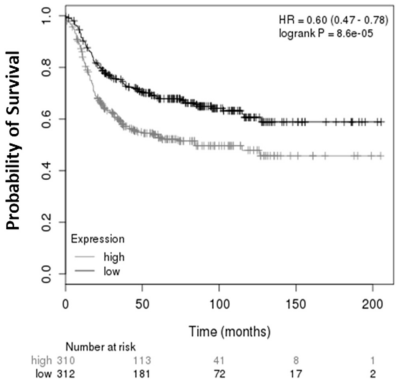

Probability of Survival
1.0
0.8
0.6
0.4
0.2
0.0
Low Expression = 64
High Expression = 42
HR = 2.82 (95% CI = 1.52 – 5.21)
P value 0.001
0
50
100
150
200
250
300
Distatnt Metastasis Free Survival (Months)

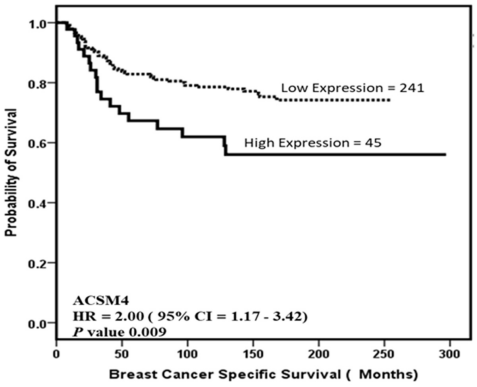

A)
SPDYC
Kaplan-Meier survival estimates
Probability of AE-free survival
1.0
0.8
0.6
0.4
0.2
0.0
0 1 2 3 4 5 6
Time (years)
≤ 75th perc.
> 75th perc.
HR: 1.68; 95% CI: 1.38 - 2.05
p-value < 0.0001
Patients at risk: (Events)
3 231 2 879 2 722 2 303 1 659 995 431 (288)
1 077 912 841 704 496 308 132 (147)

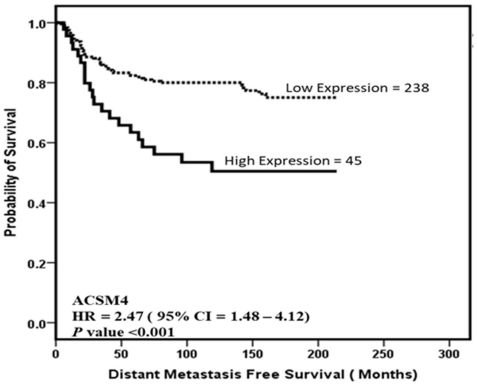

B)
SPDYC, Kaplan-Meier survival estimates
TNBC
Probability of AE-free survival
1.0
0.8
0.6
0.4
0.2
0.0
0 1 2 3 4 5
Time (years)
≤ 40th perc.
> 40th perc.
HR: 3.41; 95% CI: 1.42 - 8.18
p-value = 0.0060
Patients at risk: (Events)
103 76 64 56 31 16 (7)
151 116 99 83 45 23 (31)

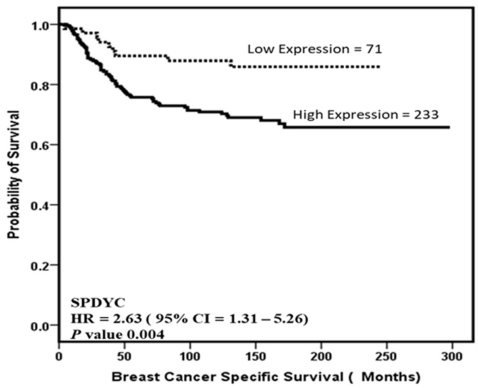

C)
ACSM4 signature (combined)
Probability of Survival
1.0
0.8
0.6
0.4
0.2
0.0
0 50 100 150 200 250
Time (months)
HR = 0.79 (0.71 - 0.88)
logrank P = 1.9e-05
Expression
high
low
Number at risk
high 2022 1181 528 129 17 3
low 2007 1397 565 112 10 0

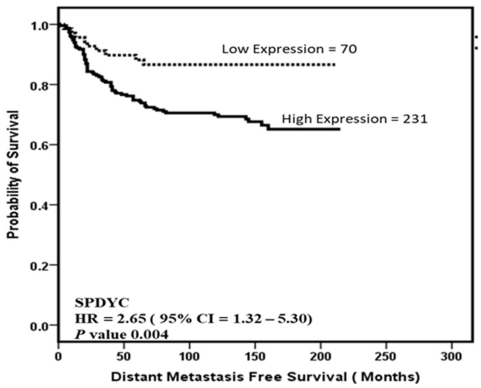

D)
SPDYC signature (combined)
Probability of Survival
1.0
0.8
0.6
0.4
0.2
0.0
0 50 100 150 200 250
Time (months)
HR = 0.80 (0.72 - 0.89)
logrank P = 6e-05
Expression
high
low
Number at risk
high 1997 1170 472 77 8 0
low 2032 1408 621 164 19 3

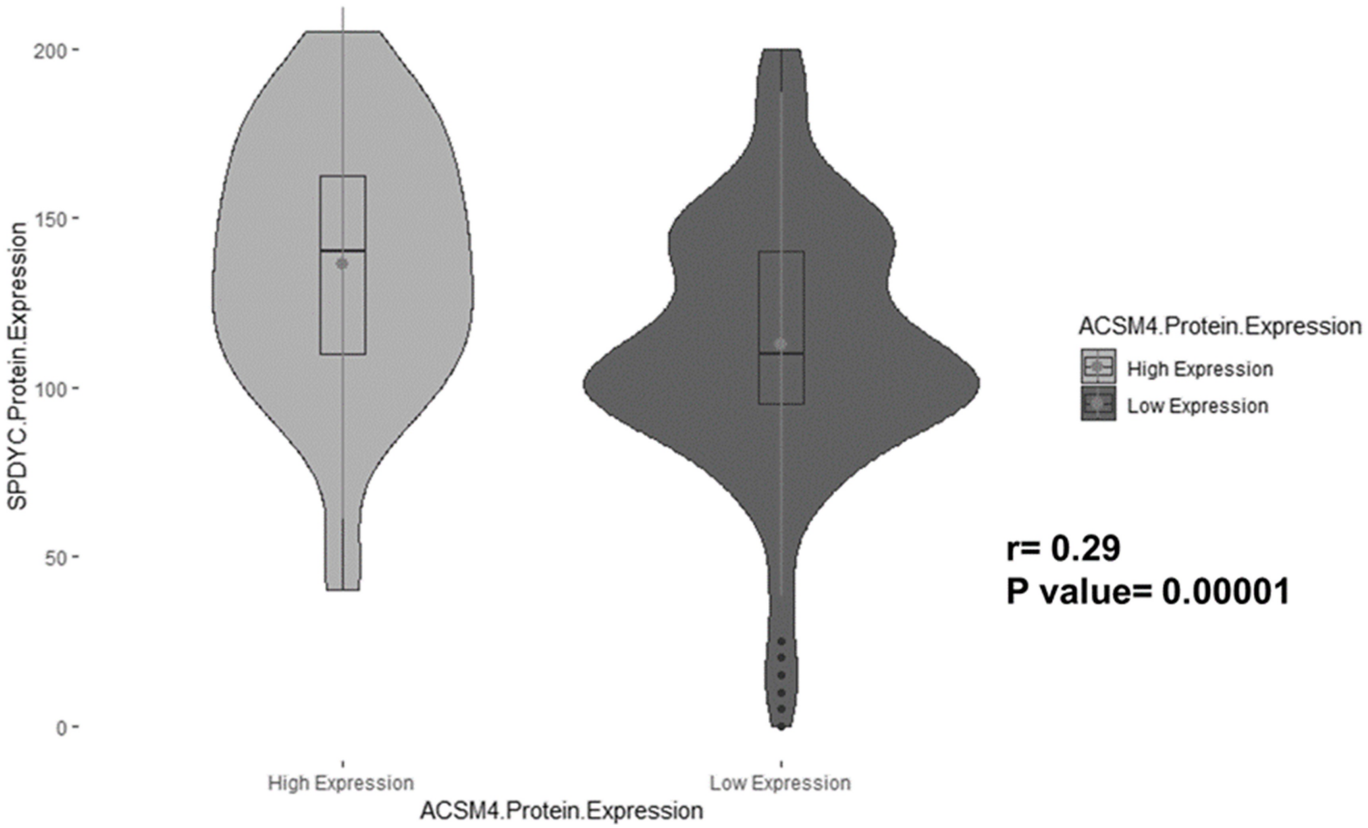

E)
ACSM4+ SPDYC signature (combined) in TNBC
Probability of Survival
1.0
0.8
0.6
0.4
0.2
0.0
0 50 100 150 200
Time (months)
HR = 0.60 (0.47 - 0.78)
logrank P = 8.6e-05
Expression
high
low
Number at risk
high 310 113 41 8 1
low 312 181 72 17 2

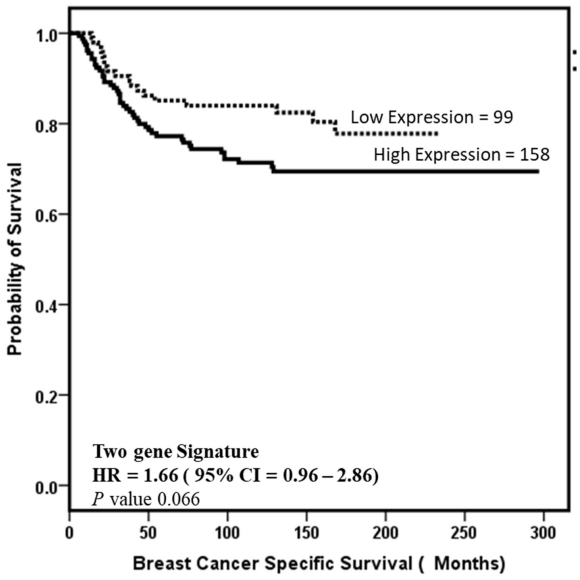

Probability of Survival
1.0
0.8
0.6
0.4
0.2
0.0
Low Expression = 241
High Expression = 45
ACSM4
HR = 2.00 (95% CI = 1.17 - 3.42)
P value 0.009
0
50
100
150
200
250
300
Breast Cancer Specific Survival (Months)

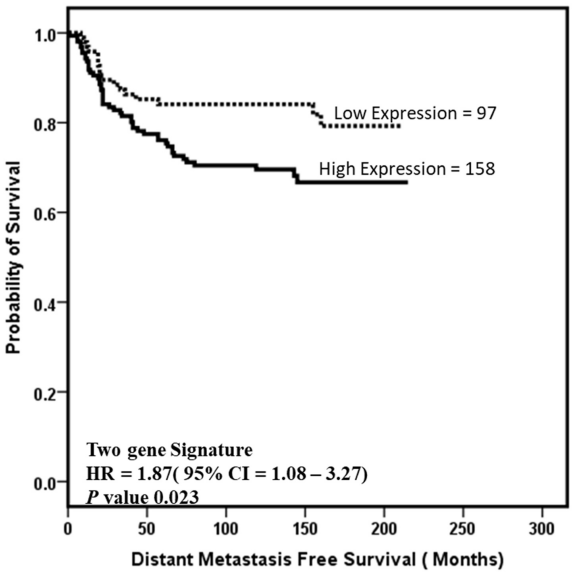

Probability of Survival
1.0
0.8
0.6
0.4
0.2
0.0
Low Expression = 238
High Expression = 45
ACSM4
HR = 2.47 (95% CI = 1.48 – 4.12)
P value <0.001
0
50
100
150
200
250
300
Distant Metastasis Free Survival (Months)


Probability of Survival
1.0
0.8
0.6
0.4
0.2
0.0
Low Expression = 71
High Expression = 233
SPDYC
HR = 2.63 (95% CI = 1.31 – 5.26)
P value 0.004
0
50
100
150
200
250
300
Breast Cancer Specific Survival (Months)


Probability of Survival
1.0
0.8
0.6
0.4
0.2
0.0
Low Expression = 70
High Expression = 231
SPDYC
HR = 2.65 (95% CI = 1.32 – 5.30)
P value 0.004
0
50
100
150
200
250
300
Distant Metastasis Free Survival (Months)


200
150
100
50
0
SPDYC.Protein.Expression
High Expression
Low Expression
ACSM4.Protein.Expression
ACSM4.Protein.Expression
High Expression
Low Expression
r= 0.29
P value= 0.00001


1.0
0.8
0.6
0.4
0.2
0.0
Probability of Survival
Low Expression = 99
High Expression = 158
Two gene Signature
HR = 1.66 (95% CI = 0.96 – 2.86)
P value 0.066
0
50
100
150
200
250
300
Breast Cancer Specific Survival (Months)


1.0
0.8
0.6
0.4
0.2
0.0
Probability of Survival
Low Expression = 97
High Expression = 158
Two gene Signature
HR = 1.87(95% CI = 1.08 – 3.27)
P value 0.023
0
50
100
150
200
250
300
Distant Metastasis Free Survival (Months)