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].

[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] Cancer Biology and Translational Research, Faculty of Medicine and Health Sciences, University of Nottingham, Nottingham, UK.

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

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

Background: 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 (NGS) to identify novel biomarkers associated with disease outcomes.

Methods: A subset of cases (n=112) from a large, diverse and well-characterised cohort of primary TNBCs (n=333) were subjected to RNA-sequencing (60M total reads/sample) and analyzed using the Illumina HiSeq 2500 platform. We identified genes associated with distant metastasis-free survival (DMFS) and breast cancer-specific survival (BCSS) by combining the application of supervised artificial neuronal network (ANN) 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.

Results: ANN identified two gene panels that strongly predicted DMFS and BCSS. 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 ($p<0.05$).

Conclusion: 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 (ERα), Progesterone Receptor (PR) and Human Epidermal Growth Factor 2 (HER2), 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 unfavourable 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 (RT-PCR), which also makes head-to-head comparison challenging [10,11].

Various multigene prognostic tests are available for ER-positive tumors for patient risk stratification and to guide therapy choice, whereas in ER-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 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 samples and aims to identify a molecular prognostic signature predicting risk for poor outcomes in TNBC.

METHODS

Nottingham TNBC Cohort

A retrospective well-characterised 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 (NPI) 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 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. ER, PR, and HER2 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 (ROC) 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 ROC 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

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 (TMAs) 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 utilised 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). In addition, a weak positive linear association between the mRNA expression of *ACSM4* and *SPDYC* (r = 0.036, p=0.710) was identified, suggesting that these genes might be collaboratively promoting disease progression. 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 harbours 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 dichotomised 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 (PPV), 72% negative predictive value (NPV), and 64% accuracy in dichotomising distant metastasis outcome of TNBC patients. In comparison, the prognostic signature at the protein level showed 73% sensitivity, 42% specificity, 30% PPV, 82% NPV, and 50% accuracy in dichotomising distant metastasis outcome of TNBC patients (Supplementary (A) Table 6).

DISCUSSION

Molecular classification of BC provides opportunities for enhanced personalised 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 signalling pathways significantly enriched in the significant top ranked gene panels. For instance, TNBC patients frequently harbour higher expression of 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 signalling 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* is a part of the cornified envelop in keratinocytes and desmosomes with intermediate filaments. *PPL* can act in the PKB/AKT-mediated signalling 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* over 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 (TILs) 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 TILs 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 signature of genes 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 RT-PCR 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 over 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 the NGS technique in 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 centre 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, future external validation is strongly recommended.

Conclusion

Personalised medicine seeks to stratify BC patients ensuring 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 grants to RA from the National Cancer Institute (R01 CA169127, U01 CA179671, and R03 CA188527). 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.

Figure legends:

Figure 1 legend: Univariate Kaplan Meier survival analyses to test associations between prognostic two gene signature at the transcriptomic level and clinical outcomes. (Transcriptomic Cohort, n=112)

Figure 2 legend*:* To validate our findings, we utilised 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 utilising 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).

Figure 3 legend: Univariate Kaplan Meier survival analyses to test associations between the ACSM4 and SPDYC protein expression and clinical outcomes (IHC Cohort, n=333)

Figure 4**:** 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 legend: Univariate Kaplan Meier survival analyses to test associations between the two gene prognostic signature protein expression and clinical outcome

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, et al. Molecular Classification of Triple-Negative Breast Cancer. J Breast Cancer 2016;19:223.

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

4 Khalifeh IM, Albarracin C, Diaz LK, 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, 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, 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, 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, et al. 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, 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, et al. Multigene prognostic tests in breast cancer: past, present, future. Breast Cancer Res 2015;17:11.

13 Lehmann BD, Jovanović B, Chen X, 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, 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, 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, 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, et al. Next-generation sequencing in breast cancer:

first take home messages. Curr Opin Oncol 2012;24:597–604.

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

20 McDonough SJ, Bhagwate A, Sun Z, 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, 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, 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 (2007-2015) VA interactive tool for comparing lists with V diagrams. http://bioinfogp. cnb. csic. es/tools/venny/index. htm. Venny 2.1.0 [Internet]. [cited 20

June 2019]. Available from: http://bioinfogp.cnb.csic.es/tools/venny/.

28 Wang J, Vasaikar S, Shi Z, et al. 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, et al. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): Explanation and Elaboration. PLoS Med 2012;9:e1001216.

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

31 Jézéquel P, Campone M, Gouraud W, Charbonnel C, Leux C, Ricolleau G CL bc-G an easy-to-use online platform for gene prognostic analyses in breast cancer. BCRT 2012; 131: 765-775. bc-GenExMiner [Internet]. [cited 30 November 2017]. Available from: http://bcgenex.centregauducheau.fr/BC-GEM/GEM-Citation.php.

32 Pongor L, Kormos M, Hatzis C, et al. 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. 2011. doi:10.1186/s13073-015-0228-1.

33 van de Vijver MJ, He YD, van ’t Veer LJ, 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, et al. Artificial neural networks applied to survival prediction in breast cancer. Oncology 1999;57:281–286.

35 Ball G, Mian S, Holding F, 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, 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, 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, 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, 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, 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, et al. 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, 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, 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, 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, et al. 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, 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, et al. 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, 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, et al. Cancer drug resistance. In: Cancer Drug Design and Discovery. : Elsevier Inc., 2008. p. 405–423.

52 Liu Y, Zhu YH, Mao CQ, 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, et al. 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, 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, et al. Genetic variants in nuclear-encoded mitochondrial genes influence AIDS progression. PLoS One 2010;5:e12862.

56 Althobiti M, Aleskandarany MA, Joseph C, 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, 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, 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, et al. Effective quality management practices in routine clinical next-generation sequencing. Clin Chem Lab Med 2016;54:761–771.