A retrospective study of SPAG5 as a novel potentially actionable oncogene, prognostic biomarker and chemotherapy sensitivity predictor: - an integrated genomic, transcriptomic and protein analysis of 10,000 Breast Cancers

Tarek M.A. Abdel-Fatah*[1], Devika Agarwal[2], Dong-Xu Liu[3,4], Roslin Russell[5], Oscar M Rueda[5], Karen Liu[3], Bing Xu[3], Paul M. Moseley[1], Andrew R. Green[6], Prof Alan G. Pockley[2], Prof Robert C. Rees[2], Prof Carlos Caldas[5], Prof Ian O. Ellis[6], Prof Graham R. Ball*[2], Prof Stephen Y.T. Chan*[1]

[1]Clinical Oncology Department, Nottingham University Hospitals NHS Trust, Nottingham NG5 1PB, UK

Professor S.Y.T. Chan DM, T.M.A. Abdel-Fatah PhD, P.M. Moseley BSc (Hons)

[2]John van Geest Cancer Research Centre, School of Science and Technology, Nottingham Trent University, Clifton campus, Nottingham NG11 8NS, UK

Professor R.C. Rees PhD, Professor A.G. Pockley PhD, Professor G.R. Ball PhD, D. Agarwal MSc

[3]Liggins Institute, the University of Auckland, Auckland 1142, New Zealand

D-X Liu PhD

[4]The Institute of Genetics and Cytology, Northeast Normal University, Changchun, China

D-X Liu PhD

[5]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge, UK

Professor C. Caldas FMedSci, R. Russell PhD, O.M. Rueda PhD

[6]Division of Cancer and Stem Cells, School of Medicine, University of Nottingham and Nottingham University Hospitals NHS Trust, Nottingham NG5 1PB, UK

Professor I.O. Ellis FRCPath, A.R. Green PhD

Running title: *SPAG5* is novel actionable oncogene that predicts survival benefit from anthracycline therapy in breast cancer patients

*Authors contributed equally

Corresponding author:

Professor Stephen YT Chan

Clinical Oncology Department

Nottingham University Hospitals NHS Trust

Nottingham NG5 1PB, U.K.

Telephone: +44(0) 115 969 1169

Fax: +44(0)115 962 8047

E-Mail: steve.chan@nuh.nhs.uk

The corresponding author confirms that he has full access to all the data in the study and has final responsibility for the decision to submit the manuscript for publication.

Research in context

Evidence before this study

Advances in molecular biology have generated a huge amount of data, which has then been used to generate multigene-profiles for guiding chemotherapy treatment. Unfortunately, almost all of these approaches face common issues, such as insufficiently high levels of evidence, the over-fitting of computational models, false discovery rates, and the lack of a potential biological mechanism to support their use as accurate predictors of the therapeutic response. Furthermore, they do not offer a significant improvement in predictive accuracy over the well-established pathological parameters or the cheaper, conventional immunohistochemistry approach, and may not be available for logistical or financial reasons. Given that, the majority of the prognostic power of these assays comes from genes that are related to cell proliferation, a St Gallen International Expert Consensus recently recommended the use of proliferation markers/profiles when choosing the appropriate systemic-treatment. However, the best molecular-marker/test that should be used continues to be debated. There was an urgent need to conduct a multi-dimensional clinical studies using multi-procedures and be validated in a large number of patients to identify a biomarker that could drive proliferation, and could be used clinically to stratify BC patients' for the best outcome. In addition there was a need to apply novel statistical approaches that are not limited by the linear functionality of conventional statistical approaches and can provide high sensitivity and specificity with excellent validity to prevent false discovery and over-fitting with sufficient power.

Added value of this study

To the best of our knowledge this is the first multi-dimensional study, with more than 10,000 patients, to report on the clinicopathological utilities of SPAG5 in breast cancer. Our findings suggest that: **1)** Amplification/gain of the *SPAG5* locus at Ch17q11·2 occurs in 10-20% of all breast cancers; **2)** The *SPAG5*-gene copy number aberrations (CNAs) and its transcript and protein are associated with poor clinical outcome and adverse clinicopathological features, including *TP53*-mutation, PAM50-LumB, and PAM50-HER2; and **3)** Both high expression of *SPAG5* mRNA transcript and protein are independent predictors for response to chemotherapy.

Implications of all the available evidence

Our findings have the potential to introduce an accurate predictive biomarker for chemotherapy response, which would facilitate the tailoring of treatments to individual patients with breast cancer. This work may lead to the development of novel strategies for more effectively managing and treating a subtype of breast cancer.

Abstract =300 words

Background:

Although the use of proliferation markers/profiles has been recommended when choosing the appropriate systemic-treatment for breast cancer (BC), the best molecular-marker/test that should be used needs to be identified.

Methods:

To identify factors that drive proliferation and its associated features in BC an artificial neural network (ANN) based integrative data-mining methodology was applied to three cohorts [(Nottingham-discovery (ND), Uppsala and METABRIC (Molecular Taxonomy of Breast Cancer International Consortium)]. The most prominent genes in the resulting interactome-map were then identified. Given that *SPAG5* was associated with many features of proliferation, featured prominently in the interactome-map and has a fundamental role in mitotic-progression,, we hypothesized that it could be a better indicator of proliferation activity. (BC). Subsequently to test if it could provide a more accurate guide for the delivery of systemic therapies in BC, we investigated the clinico-pathological utility of *SPAG5:* gene copy number aberrations (CNAs); mRNA and protein expression, in over 10,000 BCs. Integrated analysis of *SPAG5*-gene CNAs, transcript and protein expression was conducted in the ND cohort (n=171) and validated in the METABRIC cohort (n=1980). In addition, the associations of *SPAG5* CNAs, transcript and/or protein with BC specific survival (BCSS), disease free survival (DFS) and/or distant relapse free survival (DRFS) were analysed in multiple cohorts including Uppsala (n=249), METABRIC, three-untreated lymph node (LN) negative cohorts (n=684), a combined multicentre clinical data set (n=5439), Nottingham historical early-stage-primary BC (Nottingham-HES-BC; n=1650), Nottingham oestrogen receptor (ER) negative BC (n=697), Nottingham anthracycline-Neoadjuvant-chemotherapy (Nottingham-AC-Neo-ACT; n=200), and MD Anderson Cancer Centre Taxane/anthracycline (MDACC-T/AC-Neo-ACT; n=508) cohorts. The association of SPAG5 transcript and protein expression with pathological response rate (pCR) were also tested in [MDACC-T/AC-Neo-ACT (n=508) and the phase II trial NCT00455533; n=253)] and [Nottingham-AC-Neo-ACT (n=200)] cohorts; respectively.

Findings:

SPAG5 gene gain/amplification at the Ch17q11·2 locus was found in 10.4% of BC (206/1980 (; METABRIC) and was reported in 19·4% of PAM50-HER2 (46/237) and 17·8% of PAM50-LumB (87/488). *SPAG5*-CNA gain/amplification and high *SPAG5*-transcript and SPAG5-protein were associated with increased risk of death from BC [Uppsala; (HR (CI 95%): 1·50 (1·18-1·92); p=0·00010, METABRIC; (HR (CI 95%): 1·68 (1·40-2·01) p<0·0001), and Nottingham-HSE-BC; (HR (CI 95%): 1·68 (1·32-2·12), p<0·0001); respectively]. Multivariable Cox regression models, including other validated-prognostic factors, (Uppsala: age, size, LN-stage, genomic grade index (GGI), ER, *TP53* mutation and *MKi*67; METABRIC: age, size, LN-stage, histologic-grade, ER, progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), hormone-therapy, chemotherapy, interaction term of *SPAG5* and both chemotherapy[y and hormonotherapy; Desmedt-untreated LN- cohort: ER, Nottingham prognostic index (NPI), 76-gene prognostic signature (Veridex) and Adjuvant-Online (AOL); Nottingham-HES-BC: menopausal status, size, LN- stage, histologic-grade, ER, PR, HER2, ki67, hormone-therapy, chemotherapy, interaction term of SPAG5 and both chemotherapy[y and hormonotherapy), showed that high *SPAG5*-transcript and high SPAG5-protein were associated with shorter BCSS

[Uppsala: (HR (CI 95%): 1·62 (1·03-2·53) p=0·036); METABRIC: (HR (CI 95%): 1·27 (1·02-1·58) p=0·034); Desmedt-untreated LN- cohort: (HR (CI 95%): 2·34 (1·24-4·42) p=0·0090), and Nottingham-HES-BC (HR (CI 95%): 1·73 (1·23-2·46) p=0·0020); respectively].

In ER-negative-BC with high SPAG5-protein, administration of anthracycline-adjuvant-chemotherapy had reduced the risk of death by 60% compared to chemotherapy-naive (HR (95% CI): 0·37 (0·20-0·60); p=0·0010). A multivariable Cox regression analysis, which included other validated prognostic factors for chemotherapy (e.g., menopausal status, size, lymph node stage, histologic grade, ER, PR, HER2, Bcl2, chemotherapy, interaction term of SPAG5 and both chemotherapy[y), revealed that *SPAG5*-transcript+ was independently associated with decreased risk of DRFS after receiving Taxane/anthracycline-Neo-ACT [MDACC-T/AC-Neo-ACT: (HR (CI 95%): 0·68 (0·48-0·97); p=0·0070)].

In multivariable logistic regression analysis, both *SPAG5*-transcript+ and SPAG5-protein+ and were independent predictors for higher pCR after combination-cytotoxic chemotherapy [MDACC-T/AC-Neo-ACT: (OR (95% CI) 1·71 (1·07-2·74); p=0·024), and Nottingham-AC-Neo-AC: (OR (95% CI): 8·75 (2·42-31); p=0·0010); respectively].

Interpretation:

SPAG5 is a novel amplified gene on Ch17q11.2 in PAM50-LumB and PAM-HER2 BC, and its transcript and protein products are independent prognostic and predictive biomarkers, with potential clinical utility as a biomarker for combination cytotoxic chemotherapy sensitivity, especially in ER- BC.

Funding:

Nottingham Hospitals Charity and the John and Lucille van Geest Foundation.

Words=3000

Introduction

Approximately 1·68 million women are diagnosed with BC worldwide annually, with over 500,000 dying of the disease (~1,400 per day).[1] Despite continuing success, the delivery of effective precision medicine requires: 1) the discovery of novel therapeutic targets in subgroups of BC, and 2) improvements in the efficacy of treatments by identifying stratification biomarkers that can predict an individual patient's response to a particular therapy.[2] Although chemotherapy is offered to approximately 60% of patients with BC,[3] either alone or in combination with other targeted-therapies, a meta-analysis of 123 randomized trials involving over 100,000 patients has concluded that chemotherapy reduces recurrence and mortality rates by only 20-33%.[4] Although a St Gallen International Expert Consensus recently recommended the use of proliferation markers/profiles when choosing the appropriate systemic-treatment, the best molecular-marker/test that should be used continues to be debated.[5]

The main aim of the current study was to identify a biomarker that could drive proliferation and could be used to stratify BC patients' outcome. To achieve this, we decided to apply an artificial neural network (ANN) algorithm [6] to three gene expression datasets, and use factors that are directly and indirectly related to proliferation, defined as clinical class questions, to train it. The most prominent genes in the resulting interactome-map would then be developed and the best followed up, through an integrated analysis at the levels of copy number aberrations (CNAs), mRNA transcript and protein, in order to assess the clinico-pathological implications and utilities in a combined total of over 10,000 patients. Here we present the results of our ANN analysis, and the gene *SPAG5* (Sperm-associated antigen 5), which featured prominently in the interactome -map of proliferation and had a great impact on patients' survival. Given that SPAG5 has a fundamental role in the function and dynamic regulation of mitotic spindles, and in mitotic progression and chromosome segregation fidelity, [7] we hypothesized that SPAG5 could be a better measurement of proliferation activity and provide a more accurate guide for the delivery of systemic therapies in BC.

Patients and Methods

Study design and cohorts

Study design, patient's cohorts and demographics used in this study are summarized in Fig.1 and appendix p1-3.

All patients completed written informed consented, as per hospital standard of care, for excess tumour tissue to be used in research. The study was approved by the Institutional Review Board or Independent Ethics Committee and the Hospital Research and Innovations Department at all participating sites. Tumour Marker Prognostic Studies (REMARK) criteria, as recommended by McShane *et al*, [8] were followed throughout this study.

I) Identification of proliferation drivers and validation the prognostic function of *SPAG5*-CNAs, transcript and protein expression in BC

A) Discovery cohort: Nottingham discovery (ND) cohort (n=171)

The ANN modelling-based data mining approach to identify factors that drive proliferation and its associated features in BC was explored in the ND cohort, consisting of a set of 171 stage I and II invasive BC with a median follow-up of 180 months (IQR 143-194), previously described by our group in several molecular profiling studies.[9] This cohort has also been used for exploring the integrated analysis of *SPAG5* CNAs, transcript, and protein expression.

B) Test cohort: Uppsala (n=249)

The ANN modelling-based data mining approach and the clinicopathological significance of *SPAG5* gene expression were tested in the Uppsala cohort composed of 315 women representing 65% of all BCs resected in Uppsala County, Sweden (1987-1989) with a median follow-up of 126 months (IQR 119-134).[10] Gene expression data were available for only 249 patients.

C) Validation cohorts:

1) METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) cohort (n=1980)

The ANN modelling-based data mining approach and integrated *SPAG5*-CNA and *SPAG5*-trancript analysis was validated using the METABRIC cohort; a set of 1980 BCs with median follow-up of 109 months (IQR 62-155).[11] In this cohort, oestrogen receptor (ER) positive and/or lymph-node (LN) negative patients did not receive adjuvant chemotherapy (ACT), whereas ER-negative and/or LN-positive patients received ACT. Additionally, none of the human epidermal growth factor receptor-2 (HER2)-overexpression patients received trastuzumab.

2) Untreated lymph-node negative BC cohorts:

The prognostic significance of *SPAG5*-mRNA expression was assessed in three publically available datasets of LN-negative BC (n=684). These patients did not receive any adjuvant systemic therapy, thereby allowing the effect of SPAG5-

transcript on the natural history of the disease to be observed. These datasets were described in previous publications by Wang *et al* [12] (n=286), Desmedt *et al* [13] (n=196), and Schmidt *et al* [14] (n=200). The median follow up of each cohort is summarized in appendix p1-3.

3) Multi-Centre Combined cohorts (MCC; n=5439)

We evaluated the prognostic utility of *SPAG5*-mRNA expression in a large combined BC cohort which was sourced from 36 publically-available, global datasets (n=5439) using the online bc-GenExMiner program (http://bcgenex.centregauducheau.fr). [15] A list of all the datasets, with references, is summarized in appendix p4-6.

4) Nottingham Historical early stage BC cohort (Nottingham-HES-BC; n=1650)

The clinicopathological significance and prognosis of SPAG5 protein expression was also validated in a cohort of BC (n=1650; age>71 years) [16] whose tissues were suitable for SPAG5 immunohistochemistry (IHC). These patients were diagnosed and treated uniformly between 1986 and 1999 at the Nottingham City Hospital (NCH), Nottingham, UK. Patients within the good prognosis group (Nottingham Prognostic Index (NPI) <3·4) did not receive systemic adjuvant therapy. Pre-menopausal patients within the moderate and poor prognosis groups were candidates for CMF chemotherapy (cyclophosphamide 750 mg m^{-2}, methotrexate 50 mg m^{-2} and 5-fluorouracil 1 g m^{-2}, on day 1 of a 21-day cycle.). Conversely, postmenopausal ER-positive patients with moderate or poor NPI were offered hormonal therapy, whereas ER-negative patients received CMF chemotherapy. Clinical data were maintained on a prospective basis with a median follow-up of 143 months (IQR 114-174). [16]. The median follow up of subgroups is summarized in appendix p1-3.

II) The clinical significance of SPAG5-protein and *SPAG5*-transcript expression in the context of currently used chemotherapy in BC

In order to evaluate the value of *SPAG5* transcript and protein expression as a biomarker in the context of current combination cytotoxic chemotherapy, we further evaluated the clinical significance of SPAG5 protein and transcript expression in adjuvant and neo-adjuvant chemotherapy settings.

1. Nottingham early stage ER-negative BC adjuvant chemotherapy cohort (Nottingham-ER-negative ; n=697)

To evaluate the survival benefit of SPAG5-protein expression, we analysed its expression in a consecutive series of 697 early stage ER-negative BC's who had been diagnosed and managed at NCH between 1999 and 2007. This series included: 1) The ER-negative BC patients of Nottingham historical early stage BC cohort (n=332) who were managed before 2000 and treated either with no chemotherapy or with adjuvant CMF; and 2) the new ER-negative early stage BC patients (n=365) who were managed after 2000 and received either no chemotherapy or anthracycline-based-ACT (AC-ACT). [17] The median follow up of different treatment subgroups is summarized in appendix p1-3.

2. Nottingham anthracycline based Neo-Adjuvant Chemotherapy cohort (Nottingham AC-Neo-ACT; n=200)

The relationship between SPAG5-protein expression and response to chemotherapy was evaluated by investigating its expression in pair-matched pre-chemotherapy core biopsies and post-chemotherapy surgical specimens, from 200 female patients with locally-advanced primary BC (LAP-BC) (stage IIIA-C) that had been treated with anthracycline-based Neo-ACT (AC-Neo-ACT) [18] at NCH between 1996 and 2012. Sixty three percent of patients (127/200) received six cycles of an anthracycline-based therapy (FEC: 5-fluorouracil (5-FU) 500 mg m^{-2}, epirubicin 75–100 mg m^{-2}, cyclophosphamide 500 mg m^{-2}, on day 1 of a 21 day cycle), whereas 37% of patients received FEC plus Taxane (73/200). All patients underwent mastectomy or breast-conserving surgery and axillary dissection, followed by adjuvant radiation therapy. Patients with ER-positive BCs were offered 5 years of adjuvant endocrine therapy. The median follow-up time was 67 months (IRQ 27-81).

University of Texas MD Anderson Cancer Centre-Taxane/Anthracycline-based neo-adjuvant chemotherapy cohort (MDACC-T/AC-Neo-ACT; n=508)

The relationship between *SPAG5*-transcript expression and response to chemotherapy was evaluated using MDACC-T/AC-Neo-ACT cohort in which patients were selected for newly diagnosed HER2- negative BC and treated with sequential taxane and anthracycline–based neo-adjuvant chemotherapy (then endocrine adjuvant therapy if ER-positive). Details of patients' characteristics have been previously reported. [19] The median follow-up time was 38 months (IRQ 26-53).

3. **Multicentre phase II AC-Neo-ACT clinical trial cohort (NCT00455533; n=253)**

The relationship between *SPAG5*-transcript and the response to AC-Neo-ACT was validated using a randomised, open-label, multicentre, phase II clinical trial (NCT00455533) in which women with early stage BC (T2–3, N0–3, M0, tumour size 2·0 cm) have received AC-Neo-ACT regimens (cyclophosphamide plus doxorubicin (AC), followed by ixabepilone or paclitaxel). Full details of the study design and the patient characteristics have been described previously. [20] Out of 295 patients enrolled into the trial, 253 patients had available gene expression and pCR data.

Procedures

1- The ANN modelling-based data mining approach

To identify factors that could drive proliferation and its associated features in BC, a number of factors that are directly and indirectly related to proliferation, defined as clinical class questions (e.g. histological-grade; mitotic index (MI); *MKi67*; *TOP2A*; *KIF2C*; *BIRC5* and 5-year-survival), were analysed by applying an ANN modelling-based data mining approach in three gene expression array transcriptomic datasets, that included the ND, Uppsala and METABRIC cohorts. The ANNs have been selected to data mine the clinical data sets identified in this study as they have previously been shown to be able to identify biomarkers, with high sensitivity and specificity that predict clinical features with excellent validity for unseen data sets. [6] In additions, ANNs unlike conventional statistical approaches (such as hierarchical clustering, principal components analysis or linear regression) are not limited by linear functionality; this provides improved representation of biological features. The ranked orders of genes, produced in this way were compared across multiple proliferation related clinical class questions within a given dataset. The top 100 ranked genes for predicting each clinical class question, based on minimum average route mean squared error, were compared and commonalities identified at the probe level. Further comparisons were then made for the same clinical class questions in the other datasets in order to determine a consensus list of gene probes across all of the features and data sets. The strongest 100 integrated interactions were selected for visualisation in Cytoscape (Version 3.1.1, The Cytoscape Consortium; San-

Francisco, USA). [21] Further details of the ANN approach is presented in appendix p7-10.

2- *SPAG5* CNAs

CNAs at the *SPAG5* locus on chromosome 17q11.2 were retrieved from both high resolution (<100 kb) oligonucleotide microarrays, comparative genomic hybridization (aCGH; ND cohort), and Affymetrix SNP 6.0 platform profiling (METBRIC cohort), that has been previously described by our group. [9, 11] The oligonucleotide array data can be access at (http://www.ncbi.nlm.nih.gov/geo/; series accession number-GSE8757) whereas SNP data are available through the European Genotype Archive (http://www.ebi.ac.uk/ega/page.php) under accession Number: EGAS00000000082). An additional analysis considered a set of 85 individuals of European ancestry for whom genotyping was performed on non-cancerous tissue and gene expression values from matched normal tissue were available.[11]

3- *SPAG5* and *MKi67* gene expression

SPAG5 and *MKi67* mRNA expression data were retrieved and analysed in the following cohorts:- ND [using Agilent gene expression arrays at (http://www.ebi.ac.uk/miamexpress/ with accession number E-TABM-576), Uppsala [using Affymetrix U133A&B Gene-Chips microarray profiling data at (http://www.ncbi.nlm.nih.gov/geo/) with series accession number (GSE4922)], and METABRIC [using Illumina HT-12 v3 platform (Bead Arrays)[11] data at (http://www.ebi.ac.uk/ega/page.php) under accession Number (EGAS00000000082)]. In addition, the *SPAG5* and *MKi67* mRNA expression data has been retrieved for three publically available datasets of LN-negative BC in which patients did not receive any adjuvant systemic therapy: Wang *et al*[12] (accession number: GSE2034; n=286), Desmedt *et al*[13] (accession number: GSE7390; n=196), and Schmidt *et al*[14] (accession number: GSE11121; n=200). For the MCC cohort, details of the gene expression data processing, normalization and the statistical tests have been described previously. [15] In this cohort, gene expression data were converted to a common scale (median equal to 0 and standard deviation equal to 1) in order to merge all of the studies data and create combined cohorts (for more details see appendix p8). [22] The gene expression data for the MDACC-T/AC-Neo-ACT cohort and the phase II clinical trial (NCT00455533) has been downloaded using accession number GSE25066 and GSE41998; respectively.

4- Immunohistochemistry (IHC) staining of SPAG5 *and* Ki67

The ND, Nottingham-HES-BC, Nottingham-ER- and Nottingham-AC-Neo-ACT cohorts were IHC profiled for SPAG5, Ki67 and other biological parameters. Tissue microarrays (TMAs), as described in detail in appendix p8-12 have been used for IHC profiling of SPAG5 in all cohorts except in Nottingham-AC-Neo-ACT where full-face sections of core biopsies have been used.

Determination of the cut-offs

The median in each cohort was used as cut-off between low and high expression gene/protein expression

Outcomes:

The clinicopathological and biomarkers associations: The clinicopathological and molecular characteristics of *SPAG5* transcript were determined in the Uppsala, METABRIC, MCC and MDACC-T/AC-Neo-ACT cohorts. *SPAG5*–CNA moleculo-pathological associations were analysed in METABRIC cohort. The associations between SPAG5 protein expression and clinicopathological parameters, as well as prognostic biomarkers, were analysed in the Nottingham-HES-BC, the Nottingham-ER-negative and the Nottingham-AC-Neo-ACT cohorts. The clinicopathological parameters including mainly: tumour size, LN- stage, histological grade, genomic grade index (GGI), *TP53* mutation, intrinsic molecular subclasses, PAM50, HER2 amplification/overexpression, hormone receptors, Ki67, mitotic index, Bcl2 and other biological biomarkers.

Breast cancer specific survival (BCSS): *SPAG5* transcript expression association with BCSS was explored in the ND cohort and validated in Uppsala, METABRIC and the untreated LN-negative Desmedt *et al* cohorts. *SPAG5*-CNAs association with BCSS was tested in METABRIC cohort whereas the association between SPAG5 protein expression and BCSS was analysed in the ND cohort, Nottingham-HES-BC cohort and Nottingham-ER-negative cohorts.

Disease free survival (DFS): *SPAG5* transcript expression association with DFS was examined in untreated LN-negative cohorts (Wang *et al* and Desmedt *et al*), MCC and Nottingham-AC-Neo-ACT cohorts.

Distant relapse free survival (DRFS): *SPAG5* transcript expression association with DRFS were determined in untreated LN-negative Schmidt *et al* and Desmedt *et al* cohorts. Furthermore, to test *SPAG5* transcript expression as a biomarker for outcome after neo-adjuvant combination cytotoxic chemotherapy, their association with DRFS has been analysed in the MDACC-T/AC-Neo-ACT cohort.

Pathological complete response (pCR) and residual cancer burden (RCB): To evaluate SPAG5 protein and transcript expression as a predictive biomarker for response to combination cytotoxic chemotherapy, their associations with both pCR and RCB [23] have been analysed in the Nottingham-AC-Neo-ACT, the MDACC-

T/AC-Neo-ACT, and the phase II AC-Neo-ACT clinical trial cohort (NCT00455533); respectively. The pCR was defined as the absence of any residual invasive carcinoma at both the primary site and in axillary LNs.

Statistical analysis were performed using STATISTICA (Stat Soft Ltd, Tulsa, USA) and SPSS (version 17, Chicago, USA) by the authors (TAF, GRB) who were blinded to the clinical data. The Chi-square test was used for testing associations between categorical variables, and a multivariable Cox model was fitted to the data using survival time as the endpoint. All tests were two-sided with a 95% CI and a p value of <0·050 was considered to be indicative of statistical significance. Multiple-testing correction was applied to all p-values using the Bonferroni method. The range of corrections were (5 - 48,803) across the different analyses. Gene-dosage levels to gene expression were evaluated using the Jonckheere's trend test in order to evaluate the significance of the correlation between CNAs and aberrant gene-expression. Pearson correlations between mRNA expression log intensity values and SPAG5 protein expression (H-score) were used to determine whether mRNA expression levels correlated with protein levels. See appendix p8-9 for details.

Power analysis and false discovery correction

Power analysis for the ANN model was conducted using a logistic regression power model (of which ANNs are an extension with a greater power), using G*Power 3.1.9 software (Heinrich Heine University of Dusseldorf, Dusseldorf, Germany). [24] To determine sample size, an alpha of 0·050, a power of 0·80, an effect size (odd ratio = 1·72) and two-tailed test, were chosen for binary questions or classes (e.g., low vs., high expression). Based on the assumptions of the power model, the desired sample size is 88 (44 in each low and high class). The use of a Monti Carlo cross validation (MCCV) strategy was further used to prevent false discovery, over-fitting and to increase the power of the algorithm used (see appendix p7 for detail). By repeatedly testing on an unseen data set and stopping accordingly, over-fitting is prevented. False discovery is further reduced in this study by parallel analysis on multiple questions in multiple datasets. With each separate analysis reducing the probability that a gene could be discovered by random chance, and yet still be a common result across multiple analyses, of separate datasets.

The probability (p) of the 30 genes occurring as common in the top 100 out of the whole expression array for the three cohorts for a minimum of 4 proliferation-related factors = $1 \cdot 43x10^{-31}$ (see the calculation in appendix p7).

A retrospective power analysis was conducted to determine the confidence in the calculated hazard ratio and associated p value for 10 year survival and to ascertain how applicable the result would be to a global population. The power of study was determined using PASS software (NCSS, version 13, Utah, USA).

Results

Our ANN analysis in three cohorts (ND, Uppsala and METABRIC cohorts) identified the top 100 ranked genes that predict most of the proliferation-related features (appendix p13-27). We chose to further study the clinicopathological implication of *SPAG5* because it was found to be among 30 common gene-probes that were predictive across most of the proliferation features and datasets, and it features prominently in the interactome maps (appendix p25-27). In addition, in a small set of BC, investigators found that *SPAG5* transcript was a mong few genes that were associated with poor prognosis in ER-positive BC.[25] Because Ki67 has been used by many investigators as a marker for proliferation when choosing the appropriate systemic-treatment, subsequently we chose to be used it as a control in our study.

Gain/amplification at the *SPAG5* locus (17q11·2) occurred in 16% (26/171) and 10·4% (206/1980) of BC, in the ND and METABRIC cohorts respectively. *SPAG5*-gain/amplification was more common in high-grade, PAM50-HER2, and PAM50-LumB. A strong correlation between *SPAG5*-CNA and *SPAG5*-transcript expression was apparent (ND cohort: Spearman-correlation r=0·81; Bonferroni-adjusted-p=0·010) and METABRIC: Spearman-correlation r=0·87; Bonferroni-adjusted-p<0·0001). ER-negative and ER-positive BC exhibited a higher level of *SPAG5*-transcript (correlation-coefficient=0·19; Bonferroni-adjusted-p<0·0001 and correlation-coefficient=0·37; Bonferroni-adjusted-p<0·0001; respectively) compared to normal individuals (n=85). However, the level of *SPAG5*-transcripts in ER-negative disease was higher than that in ER-positive disease (correlation-coefficient=0·18; Bonferroni-adjusted-p<0·0001). Furthermore, the PAM50-LumB, PAM50-Basal and PAM50-HER2+ BC-subclasses exhibited higher levels of *SPAG5*-transcripts than PAM-50-normal-like, PAM50-Lum-A disease, and normal tissue (all-adjusted-p<0·0001; Figure appendix p28).

As a continuous and categorical variable, compared to low *SPAG5*-transcript expression, high SPAG5-transcript level (>median) was associated with high-grade *TP53*- mutation, and HER2 gain/amplification. In the METABRIC study, 10-novel-prognostic biological subgroups have been identified by the joint clustering of CNA and gene expression data (integrative-clusters (intClust)). [11] Herein, *SPAG5*-gain/amplification was shown to be associated with intClust-1, 5, and 6 (all-p<0·0001) whereas high *SPAG5*- was associated with intClust-1, 5, 9, and 10;

appendix p29-32. Furthermore, high *SPAG5*-transcript expression (>median) was associated with other molecular parameter/indices/subclasses that predict higher probability of response to Neo-ACT: RCB-0/I,[23] genomic-chemo-sensitivity predictor, [19] genomic-excellent-pathologic-response predictor, [19] 96-gene-genomic-grade index (GGI), [26] diagonal linear discrimination analysis of 30-gene signature (DLDA30), [27] and PAM-50-gene signature [28] (all p<0·0001, appendix p33).

Additionally, there was a strong correlation between *SPAG5*-transcript and SPAG5-protein expression (Pearson-correlation (r=0·75); Bonferroni-adjusted-p=0·001). In the Nottingham-HES-BC cohort, 20% (272/1368) of patients showed high SPAG5-protein (H-score≥10) that was associated with aggressive phenotypes including HER2+ (p=0·030), Luminal-B (ER+/HER-/high-Ki67), an absence of hormone receptors, and *TP53*-mutation (appendix p34-37). In the Nottingham-ER-negative cohort, high SPAG5-protein (H-score>10) was observed in 51% (355/697) and was associated with lympho-vascular-invasion, high-grade, and high-ki67 (all p<0·0001; appendix p38-41). In Nottingham AC-Neo-ACT locally advanced BC cohort, high SPAG5-protein (H-score>10) was observed in 25·0% (50/200) of pre-chemotherapy core biopsies and was associated with high-grade, Luminal-B (ER+/HER2-/high Ki67), ER-/HER-negative, and *TP53*-mutation (all adjusted p<0·0001). Among different cohorts neither *SPAG5* transcript nor protein was associated with LN stage or disease clinical stage.

SPAG5-gain/amplification was associated with shorter BCSS than *SPAG5*-normal/loss in all patients (HR (CI 95%):1·50 (1·18-1·92); p=0·0016) and the ER-positive subgroup (HR (CI 95%): 1·55 (1·18-2·04); p=0·00020), but not in ER-negative tumours (HR (CI 95%): 1·58 (0·97-2·56), p=0·065) (Fig.2A-C); METABRIC cohort.

As continuous variables, high *SPAG5*-transcript expression levels was associated with shorter BCSS than low *SPAG5*-transcript [ND cohort: (HR (CI 95%):1·50 (0·98-2·32); p=0·065), Uppsala cohort: (HR (CI 95%): 1·99 (1·44-2·76); p<0·0001) and METBRIC cohort: (HR (CI 95%): 1·89 (1·55-2·31); p<0·0001)]. As a categorical variable high *SPAG5*-transcript (>median), was associated with shorter BCSS than low *SPAG5*-transcript [Uppsala: (HR (CI 95%): 1·98 (1·29-3·04); p=0·0020), and METABRIC: (HR (CI 95%): 1·68 (1·40-2·01); p<0·0001, Fig2D). High *SPAG5*-transcript was associated with shorter BCSS than low *SPAG5*-transcript in ER-positive sub-groups but not in ER-negative tumours (Fig.2E-F). Also, in the low-risk BC (NPI≤3·4), LN-negative, as well as LN-positive (METABRIC), high *SPAG5* transcript was associated with shorter BCSS than low *SPAG5* transcript (Figure appendix 42). In Uppsala cohort with 249 cases (124 in the high *SPAG5* transcript group) achieved a power of 83% to detect a hazard ratio (HR) of 1·98 , when the 10 years survival rate for high and low *SPAG5* transcript are 53% and 71%; respectively and p value <0·050. Similarly for the METABRIC cohort transcript expression analysis, a power model using a two-sided log-rank test with an overall sample size of 1950 subjects (970 in the high *SPAG5*-transcript groups) achieved a power in excess of ≥99·9% to detect a hazard ratio (HR) of 1·68, when the proportion surviving (BCSS) to 10 years in the high *SPAG5* and low *SPAG5* are 78% and 66%; respectively, with p <0·0001.

In untreated LN-negative cohorts high *SPAG5* transcript (>median) was associated with shorter DSF, DRFS and BCSS than low *SPAG5* transcript (Fig.2G-I]. In the untreated LN-negative-BC cohorts: Wang *et al* (n=286; 143 cases with high *SPAG5*), Schmitt *et al* (n=200; 100 cases with high *SPAG5*), and Desmedt *et al* (n=198; 99 cases with high *SPAG5*), the retrospective power of each to detect HRs of 1·3, 1·4, and 1·99, at ten years of DFS, DRFS and BCSS was 82%, 84%, and 98%; respectively, with p<0·050.

In the Uppsala cohort multivariable Cox regression analysis including patient age, LN-stage, tumour-size, GGI, ER status, *TP53* mutation and *MKi67*-transcript, revealed that high *SPAG5*-transcript and LN-stages were independently associated with increased-risk of death (Table-1A). Similarly in the METABRIC cohort, a multivariable Cox regression model which included patient age, tumour size, grade, LN-stage, HER2, ER, PR, hormone-therapy, and chemotherapy, demonstrated that high *SPAG5*-transcript was independently associated with shorter BCSS (Table-1B). Furthermore, multivariable Cox regression models showed that high *SPAG5*-transcript was associated with clinical outcome independently of both PAM50 and intClust prognostic subclasses (Table 1C-D). In addition, in the untreated LN-negative Desmedt *et al* cohort, high *SPAG5*-transcript was associated with shorter BCSS after adjustment for ER status and other prognostic signatures/indices such as 76-gene prognostic signature (Veridex)[13], Adjuvant-Online (AOL) and the NPI (Table 1-E).

In the MCC cohort high *SPAG5*-transcript (>median) was associated with increased risk of relapse compared to low *SPAG5*-transcript expression in all patients and LN-negative, LN-positive, and ER-positive BC subgroups, but not in the ER-negative-subgroup (Figure appendix p43). In MCC, with an overall sample size of 5439 (2711 in high *SPAG5*), a ≥99·9% power to detect a HR of 1·68 for DFS, with p-value <0·0010, was achieved. In the MCC cohort, multivariable Cox regression models confirmed that the high *SPAG5*-ranscript is an independent poor prognostic factor after controlling for NPI (HR (CI 95%): 1·19 (1·09-1·30); p=0·00020), AOL (HR (CI 95%): 1·18 (1·03-1·35); p=0·017), and 72-proliferation-gene-signature [29] (HR (CI

95%): 1·18 (1·10-1·27); p<0·0001). Univariate analysis showed that high *MKi67* transcript expression was associated with a higher risk of relapse compared to low *MKi67* expression. However, multivariable Cox regression models revealed that *MKi67* transcript expression was not an independent prognostic factor for BC after controlling for NPI (HR (CI 95%): 1·09 (1·00-1·20); p=0·060) and AOL (HR (CI 95%): 0·93 (0·83-1·05); p=0·26).

Similarly high SPAG5-protein expression was associated with shorter BCSS than low SPAG5-protien expression (ND: HR (CI 95%): 1·06 (1·02-1·09), p=0·0010, and Nottingham-HES-BC cohorts: HR (CI 95%): 1·68 (1·32-2·12), p<0·0001; Fig.3A]. SPAG5-protein+ was also associated with increased-risk of death in ER-positive subgroups (Fig.3B), but not in ER-negative subgroups (Fig.3C). In the low-risk (NPI<3·4), LN-negative, as well as LN-positive subgroups high SPAG5 protein expression was associated with shorter BCSS than low SPAG5 (Figure appendix p44). For the ND and Nottingham HES BC cohort, with an overall sample size of 128 and 1342 subjects (24 and 273 cases in the high SPAG5-protein subgroups), 80% and 99·0% powers to detect a HR of 1.10 and 1·68 is achieved with p-value <0·050, when the proportion surviving in the SPAG5+ subgroup at ten year of BCSS is 60% and 63%; respectively.

Multivariable Cox regression analysis reveals that high SPAG5-protein was independently associated with a poorer BCSS at 10 years compared to low SPAG5 expression, after adjustment for adjuvant hormone-therapy and chemotherapy, grade, size, LN-stage, HE2, ER, PR, age, Ki67 and interaction-terms (SPAG5* chemotherapy and SPAG5* hormone-therapy); Table 1F.

In the Nottingham-ER-negative cohort, high SPAG5-protein was associated with decreased risk of death from BC (HR (95% CI): 0·85 (0·78-0·94); p=0·0010) (Fig.3D) compared to low SPAG5-protein expression. However, a subgroup analysis of ACT-naïve cases showed that patients with high and low SPAG5-protein expression exhibited similar BCSS (HR (95% CI): 0·90 (0·63-1·27); p=0·54), whereas in the subgroup that received ACT; high SPAG5-protein exhibited lower risk of death that low SPAG5-protein level (HR(95% CI): 0·41 (0·26-0·64); p<0·0001) (Figure appendix p45). In ER-negative BC with high SPAG5-protein, administration of anthracycline-ACT had reduced the risk of death by 60% compared to chemotherapy-naive (HR (95% CI): 0·37 (0·20-0·60); p=0·0010) (Fig.3E). Meanwhile administration of anthracycline-ACT had no impact on tumours with ER-negative /low SPAG5-protein phenotype (Fig.3F). A multivariable Cox regression model confirms that SPAG5 was a predictive marker and that the interaction-term between SPAG5-protein and the administration of anthracycline-based adjuvant chemotherapy was a significant predictor of BCSS (Table-1G).

In the MDACC-T/AC-Neo-ACT cohort, after receiving combination cytotoxic chemotherapy, there was a marginally shorter DRFS in those patients with high *SPAG5*-transcript tumours compared to those with low *SPAG5* (HR (CI 95%): 1·3 (0·92-1·95); p=0·12; appendix p43). In those patients that did not achieve pCR, high *SPAG5*-transcript was significantly associated with shorter DRFS than those with low *SPAG5*-transcript (HR (CI 95%): 1·74 (1·17-2·52); p=0·0070; appendix p46). A multivariable Cox regression analysis which included other prognostic factors for chemotherapy, namely genomic-chemo-sensitivity predictor, GGI, DLDA30, PAM-50-genes, American joint committee of cancer (AJCC) stages, and *MKi67* transcript, revealed that high *SPAG5* was independently associated with decreased risk of distant relapse after receiving Neo-ACT (HR (CI 95%): 0·68 (0·48-0·97); p=0·0070; Table1H).

To validate our previous observation, we investigated the relationship of *SPAG5* transcript expression and response to combination cytotoxic chemotherapy in the MDACC-T/AC-Neo-ACT cohort (n=508) in which 488 cases had pCR data were available. Of them, 20% (99/488) achieved pCR. As a continuous variable, high *SPAG5*-transcript levels were associated with higher pCR compared to low *SPAG5*-transcript (OR (CI 95%): 2·6 (1·8-3·9); p<0·0001). As a categorical variable, high *SPAG5*-transcript (>median) was association with higher pCR; 29% (70/246) vs 12% (29/242) for low *SPAG5*-transcript (OR (95% CI): 2·90 (1·80-4·70), p<0·0001). Multivariable logistic regression analysis including parameter/indices/subclasses that associated with higher pCR as well as: AJCC clinical stage, histological grade, ER, PR, and patient age, demonstrated that high *SPAG5*-transcript was an independent predictor for higher pCR (Table-2A).

We further validated our results in a multicentre phase II anthracycline-based-Neo-ACT clinical trial cohort (NCT00455533) [20] in which 27% (69/253) and 34% (86/253) of patients achieved pCR and RCB-0/RCB-1 rates, respectively. As a continuous variable, *SPAG5*-transcript expression was associated with a marginally higher incidence of pCR and RCB-0/RCB-1 ((OR (CI 95%): 1·33 (0·98-1·79); p=0·065) and (OR (CI 95%): 1·29 (0·98-1·71); p=0·075); respectively). Using the median to categorize *SPAG5*-transcript expression into (high) and (low); high *SPAG5*-transcript was associated with higher pCR and RCB-0/RCB-1 rates ((OR (CI 95%): 1·99 (1·13-3·45); p=0·016) and (OR (CI 95%): 1.97 (1·16-3·34); p=0·010), respectively) compared to low *SPAG5*-transcript. In a multivariable logistic regression model which included ER, PR, HER2, tumour size, menopausal status, and *MKi67*, and *SPAG5* transcript expression, *SPAG5*-transcript was significantly associated with RCB-0/RCB-1 (Table 2B).

Similar to transcriptomic findings, patients with high SPAG5-protein (H-score>10) disease prior to chemotherapy, who received AC-Neo-ACT, exhibited similar 5-year DFS following surgery (HR (95% CI): 1·1 (0·90-1·30); p=0·40) to those with low SPAG5-protein- disease (appendix p43). Importantly, patients with high SPAG5-protein expression in the residual tumour specimen after receiving AC-Neo-ACT were at a higher risk of relapse (HR (95% CI): 2·2 (1·2-4·2); p=0·010) compared to those with low SPAG5-protein residual tumours, at the 5-year follow-up (Figure appendix p46). In the Nottingham-AC-Neo-ACT cohort, 14·5% (29/200) of patients had achieved pCR and 40% (20/50) of patients with high SPAG5-protein BC achieved pCR compared to 6% (9/150) of those with low SPAG5-protein disease (OR (CI 95%): 10·8 (4·5-26·29); p<0·0001; appendix p43). Furthermore, 37% (18/49) of BC that exhibited high SPAG5-protein disease became negative for SPAG5-protein after receiving AC-Neo-ACT (McNemar-test; p=0·0040). Multivariable logistic regression analysis revealed that high SPAG5-protein was an independent predictor of pCR, whereas Ki67 was not, after controlling for age, taxane, grade, AJCC stage, ER, HER2, Ki67, Bcl2, and TOP2A (Table-2C).

Discussion

To our knowledge this is the first multi-dimensional study to report on the clinico-pathological utilities of *SPAG5* in BC in more than 10,000 patients. Our findings suggest that: **1)** Amplification/gain of the *SPAG5* locus at Ch17q11·2 occurred in 10-20% of BC, **2)** The *SPAG5*-gene-CNA and its transcript and protein were associated with poor clinical outcome and adverse clinicopathological features, including *TP53*-mutation, PAM50-LumB, and PAM50-Her2, **3)** Both *SPAG5* transcript+ and protein+ are independent predictors for response to chemotherapy.

Recent advances in molecular biology have generated a huge amount of data, which has then been used to generate multigene-profiles for guiding chemotherapy treatment. Unfortunately, almost all of these approaches face common issues such as insufficiently high levels of evidence, the over-fitting of computational models, false discovery rates, [30] and the lack of a potential biological mechanism to support their use as predictors of therapeutic response. Furthermore, they do not offer a significant improvement in predictive accuracy over the well-established pathological parameters or the cheaper, conventional immunohistochemistry approach, and may not be available for logistical or financial reasons. [31] In fact, the majority of the prognostic power of these assays comes from genes that are related to cell proliferation. The data presented herein are significant as the prognostic and predictive capacities of SPAG5 have been shown to be independent of many of these multigene tests and Ki67. Furthermore, our integrated network inference bioinformatics analysis has revealed that *MKi67* was less influential on other proliferation factors, and lacked the centrality of other probes.

In agreement with the results of data mining the Oncomine-microarray database, we found BC, like most human cancers, exhibited a higher level of SPAG5-transcript expression compared to normal tissue (appendix 47-53), which in turn is associated with poor clinical outcome (appendix 54-56), especially in ER+ BC. [25] In agreement with a previous study, we have reported a high level of Ch17q11·2 amplifications in HER2+ and ER+ BC, [32] which is the locus of SPAG5. Recently, duplication of CEP17 has been proposed as a marker of chromosomal instability, spindle assembly checkpoint deregulation, and it has been linked to anthracycline-sensitivity *in vitro* and to clinical outcome of AC-ACT. [33] Likewise, given that SPAG5 has an essential role in the progression of the cell cycle during the mitotic phase, SPAG5 dysregulation could contribute to chromosome instability and aneuploidy, both of which are hallmarks of malignant cells and could confer vulnerability on the cancer cell. Given that drugs such as the anthracyclines and taxane, which interfere with the normal progression of mitosis, belong to the most successful chemotherapeutic compounds that are currently used for anti-cancer treatment, SPAG5 could be a molecular target on which the development of "next generation anti-mitotic drugs" could be based. Recent studies in cervical cancer [34, 35] reported *SPAG5* to be up-regulated, and demonstrated that the down-regulation of *SPAG5* inhibited cell proliferation/growth, increased apoptosis and hindered cell migration and invasion. [35] Furthermore, it is possible that "anti-*SPAG5* agents" could sensitize resistant BC cells to current treatment regimens.

The potential clinical significance of our results primarily relates to the identification of BC patients who are likely to benefit from anthracycline-based chemotherapy. Validating our results in a randomized-prospective Neo-ACT trial would allow patients whose tumour response would be poor to be spared from enduring the unnecessary risk of cardiac toxicity, when other more effective agents can be used. Although the mechanism linking *SPAG5*+ and response to anthracycline is unknown and further investigation is warranted, it could be due to the accumulation of DNA damage, abnormal mitoses, and subsequent mitotic catastrophe.[36]

In summary, our findings have the potential to introduce an accurate predictive biomarker for chemotherapy response, which would facilitate the effective tailoring of BC treatment. This work may lead to the development of novel therapeutic strategies for treating a subtype of BC, thereby increasing the chance of cure from BC.

Decleration of Interests

Tarek M.A. Abdel-Fatah, Graham R. Ball and Stephen Y.T. Chan are named inventors on a PCT patent application which is jointly held by the NHS Trust and Nottingham Trent University (US patent publication number 14/404,163 published on 1st June 2012).

Graham R Ball is named on a patent held by Nottingham Trent University (PCT/GB2009/051412, US (Granted) 8788444, EP (Pending) 09796034.8 which covers the Artificial Neural Network algorithms utilised.

There are no further conflicts of interests to disclose by the authors.

Acknowledgements

We would like to acknowledge the Nottingham Trent University Vice Chancellor Bursary Scheme for supporting Devika Agarwal’s PhD programme.

In addition we would like to acknowledge the support Professor Chan and his research group have received from the Nottingham Hospitals Charity.

We would also like to acknowledge the New Zealand Breast Cancer Foundation and the Breast Cancer Cure charities for supporting Dr Dongxu Liu’s research programme.

The supporters of this work played no role in the study design, data collection, data analysis, data interpretation, writing of the manuscript, or in the decision to submit the paper for publication.

Authors' Contributions

S.Y.T.C., T.M.A.A-F., and G.R.B. provided intellectual input, conceptual framework, and designed the study. S.Y.T.C., T.M.A.A-F., D-X.L., D.A., R.R., O.M.R., K.L., B.X., P.M.M., A.R.G., A.G.P., R.C.R., C.C., I.O.E., and G.R.B., were each involved in drafting the manuscript, and took part in critically reviewing it for publication. T.M.A.A-F, D.A, and G.R.B. performed the statistical-analysis, gene expression analysis, and Artificial Neural Network modelling. R.R., O.R., and C.C. provided SPAG5 gene copy number aberrations data, gene expression data and performed the statistical analysis for the METABRIC cohort. S.Y.T.C., T.M.A.A-F., D.A., G.R.B., D-X.L., and I.O.E analysed and interpreted the data. P.M.M. carried out the immunohistochemistry staining. T.M.A.A-F. undertook the pathological assessment of experimental slides. P.M.M, T.M.A.A-F, A.R.G., and R.R. conducted collection and management of patient data.

Abbreviations

AC = anthracycline combination

AC-ACT = anthracycline combination adjuvant chemotherapy

ACT = adjuvant Taxane chemotherapy (Docetaxel 75 to 100mgm-2 every 3 weeks)

ANN = artificial neural network

AOL = Adjuvant-Online

AR = androgen receptor

ASCO = American society of clinical oncology

BC = breast cancer

BCSS = breast cancer specific survival

CAP = college of American pathologists

CI = confidence interval

CNA = copy number aberrations

CMF = cyclophosphamide 750 mg m^{-2}, methotrexate 50 mg m^{-2} and 5-fluorouracil 1 g m^{-2}, on day 1 of a 21-day cycle.

DFS = disease-free survival

DM = distant metastasis

DRFS = distant metastasis-free survival

ER = oestrogen receptor

FEC/FAC = 5-Fluorouracil (5-FU) 500 mg m^{-2}, Epirubicin 75–100 mg m^{-2}, Cyclophosphamide 500 mg m^{-2}, on day 1 of a 21-day cycle.

FFPE = formalin-fixed paraffin embedded

HER2 = human epidermal growth factor receptor 2

HR = hazard ratio

HPA = human protein atlas

IHC = immunohistochemistry

LAP-BC = locally-advanced primary breast cancer

LN = lymph node

MDACC-T/AC-Neo-ACT = University of Texas MD Anderson Cancer Centre-Taxane/Anthracycline-based neo-adjuvant chemotherapy cohort

MCC = Multicentre combined cohort

MCCV = Monti Carlo cross validation

MI = mitotic index

ND = Nottingham discovery cohort

Neo-ACT = neo-adjuvant chemotherapy

Nottingham-AC-Neo-ACT = Nottingham anthracycline-Neoadjuvant-chemotherapy cohort

Nottingham-ER- = Nottingham early stage ER- BC cohort

Nottingham-HES-BC = Nottingham historical early-stage-primary BC cohort

NPI = Nottingham prognostic index

NT = Nottingham series

pCR = pathological complete response

PR = progesterone receptor

TMA = tissue microarray

TNBC = triple negative breast cancer

Tables

Table-1: Multivariable Cox regression models analysis in different breast cancer cohorts.

Variables	**HR**	**95·0% CI**		**P value**
		Lower	**Upper**	
A. Multivariable Cox regression model analysis for breast cancer specific survival in the Uppsala test cohort (*SPAG5* transcript) (n=249)				
***SPAG5* mRNA (high)**	1·62	1·03	2·53	**0·036***
***MKI67* mRNA (high)**	0.991	0.486	1.71	0·77
Lymph node status (positive)	1·61	1·01	2·57	**0·050***
96-gene genomic grade index (GGI) [26]				0·34
G1	1			
G2a	0·94	0·50	1·79	
G2b	1·77	0·82	3·96	
G3	1·73	0·76	3·97	
Age at diagnosis**	1·01	0·99	1·03	0·16
Tumour size (continuous) (mm)	1·09	0·95	1·24	0·21
Oestrogen receptor (positive)	1·43	0·76	2·71	0·27
***TP53* mutation**	1·07	0·62	1·86	0·80
B. Multivariable Cox regression model (1) analysis for breast cancer specific survival in the METABRIC cohort (*SPAG5* transcript) (n=1980)				
***SPAG5* mRNA (high)**	1·27	1·02	1·58	**0·034***
Lymph node (LN) stage				**<0·0001***
Negative	1·00			
1-3 positive LNs	1·68	1·31	2·16	
>3 positive LNs	3·42	2·59	4.52	
Histologic grade				**0·017***
Low	1·00			
Intermediate	1·79	1·08	2·95	
High	2·05	1·23	3·39	
Size	1·01	1.007	1·015	**<0·0001***
Age at diagnosis**	1·01	1·002	1·02	**0·015***
HER2	1·50	1·18	1·91	**0·0010***
Progesterone receptor (positive)	0·77	0·62	0·96	**0·020***
Oestrogen receptor (positive)	1·06	0·78	1·45	0·70
Hormone therapy	1·23	0·82	1·02	0·12
Chemotherapy	1·31	0·96	1·78	0·090
Hormone therapy*SPAG5	0·62	0·41	0·93	**0·021***
Chemotherapy*SPAG5	0·84	0·55	1·28	0·42
C. Multivariable Cox regression model (2) analysis for breast cancer specific survival in the METABRIC cohort (*SPAG5* transcript) (n=1980)				
***SPAG5* mRNA (high)**	1·31	1·04	1·65	**0·020***
PAM-50 Molecular subclasses [28]				**<0·0001***
PAM50-LumA	1			
PAM50-LumB	2·13	1·62	2·80	

PAM50-HER2 PAM50-Basallike PAM50-Normal	2·34 1·89 1·45	1·72 1·38 1·01	3·18 2·59 2·08	
Hormone therapy	1·31	1·06	1·60	**0·010***
Chemotherapy	1·31	1·66	2·59	**<0·0001***
Hormone therapy*_SPAG5_	0·57	0·38	0·84	**0·0050***
Chemotherapy*_SPAG5_	1·18	0·78	1·78	0·43
D. Multivariable Cox regression model (3) analysis for breast cancer specific survival in METABRIC cohort (*SPAG5* transcript) (n=1980)				
***SPAG5* mRNA (high)**	1·33	1·06	1·67	**0·014***
Integrated Clusters (IntClust)[11] intClust.1 intClust.2 intClust.3 intClust.4 intClust.5 intClust.6 intClust.7 intClust.8 intClust.9 intClust.10	 1 1·47 0·38 0·69 1·58 1·13 0·58 0·65 1·08 0·75	 0·92 0·24 0·46 1·09 0·70 0·37 0·44 0·72 0·50	 2·34 0·61 1·03 2·30 1·81 0·93 0·97 1·63 1·13	**<0·0001***
Hormone therapy	1·23	1·003	1·50	**0·047***
Chemotherapy	2·02	1·62	2·51	**<0·0001***
Hormone therapy*_SPAG5_	0·53	0·36	0·77	**0·020***
Chemotherapy*_SPAG5_	1·18	0·78	1·78	0·66
E. Multivariable Cox regression analysis of *SPAG5* transcript in untreated Lymph node negative “Desmedt cohort”				
***SPAG5* mRNA (high)**	2·34	1·24	4·42	**0·0090***
Oestrogen receptor (positive)	0·67	0·38	1·22	0·19
NPI	1·74	0·712	4·23	0·22
Adjuvant-online (AOL)	0·76	0·30	1·94	0·56
76-gene prognostic signature (Veridex) [13]	1·52	0·75	3·06	0·24
F. Multivariable Cox regression analysis of SPAG5 protein for breast cancer-specific survival at 20 years follow-up in Nottingham historical early stage breast cancer cohort (n=1650)				
SPAG5 protein expression (positive)	1·73	1·23	2·46	**0·0020***
Tumour size (continuous) (mm)	1·18	1·07	1·31	**0·0010***
Lymph node (LN) status Negative Positive	 1 1·95	 1·51	 2·52	**<0.0001***
Histological grade Low/intermediate High	 1 1·83	 1·33	 2·50	**0·00020***
Oestrogen receptor (positive)	1·20	0·82	1·74	0·350
HER2 overexpression (positive)	1·60	1·16	2·52	**0·0040***
Progesterone receptor status (positive)	0·66	0·47	0·92	**0·015***
Ki67 (positive)	1·44	1·03	2·01	**0·034***
Chemotherapy status (CMF)	1·55	1·13	2·17	**0·010***
Hormone therapy (yes)	1·31	0·99	1·73	0·059

Chemotherapy*SPAG5	1·65	0·85	3·23	0·14
Hormone therapy *SPAG5	1·95	1·14	3·35	**0·015***
G. Multivariable Cox regression analysis of SPAG5 protein for breast cancer-specific survival at 10 years follow-up in Nottingham early stage oestrogen receptor negative breast cancer cases (n=697)				
Model without interaction terms				
SPAG5 protein expression (positive)	0·68	0·50	0·92	**0·013***
Tumour size (continuous)	1·06	1·02	1·09	**0·0010***
Lymph node (LN) status				**<0·0001***
Negative	1			
Positive	2·60	1·92	3·50	
Histological grade				0·059
Low/intermediate	1			
High	1·67	0·98	2·86	
Menopausal status (post vs pre)	1·34	0·99	1·82	0·060
HER2 overexpression (positive)	0·92	0·64	1·31	0·64
Bcl2 (positive)	0·60	0·40	0·90	**0·013***
Chemotherapy status				
No Chemotherapy	1			
CMF	0·80	0·54	1·18	0·260
Anthracycline	0·61	0·42	0·89	**0·010***
Model with interaction terms				
SPAG5 protein expression (positive)	0·48	0·30	0·76	**0·0020***
Tumour size (continuous)	1·05	1·02	1·09	**0·0030***
Lymph node (LN) status				**<0.0001***
Negative	1			
Positive	2·57	1·90	3·46	
Histological grade				0·066
Low/intermediate	1			
High	1·65	0·97	2·82	
Menopausal status (post vs pre)	1·35	0·99	1·84	0·056
HER2 overexpression (positive)	0·93	0·65	1·34	0·70
Bcl2 (positive)	0·63	0·42	0·94	0·023
Chemotherapy status				
No Chemotherapy	1			
CMF	0·79	0·54	1·16	0·23
Anthracycline	0·59	0·40	0·87	**0·008***
SPAG5*CMF interaction term	0·70	0·32	1·50	0·36
SPAG5*Anthracycline interaction term	0·43	0·20	0·93	**0·032***
H. Multivariable Cox regression analysis for clinical outcome (distant relapse-free survival (DRFS)) in the University of Texas MD Anderson Cancer Centre Taxane/Anthracycline based neo-adjuvant cohort (n=508)				
***SPAG5* transcript expression (positive)**	0·68	0·48	0·97	**0·031***
Chemo-sensitivity prediction signature [19]				
Low vs high	0·49	0·36	0·67	**<0.0001***
96- gene genomic grade index (GG1) [26]				
Grade 1/Grade2a vs Grade 2b/Grade 3-like	0·646	0·323	1·29	0·21
30-gene (DLDA30)[27]	2·09	0·95	4.56	0·065

Low vs High				
PAM-50 Molecular subclasses [28]				**0·042***
PAM50-LumA vs others	0·16	0·04	0·58	**0·0060***
PAM50-LumB vs others	0·24	0·07	0·88	**0·031***
PAM50-HER2 vs others	0·14	0·03	0·66	**0·013***
PAM50-Basal-like vs others	0·28	0·07	1·10	0·068
Clinical AJCC stage				
I/II vs III	2·03	1·38	2·99	**0·00040***
***MKI67* transcript**	1·22	0·71	2·07	0·47

*Statistically significant at p<0·05.

**Age was a continuous value with increments of 5 years.

SPAG5; Sperm-associated antigen, ER; Oestrogen receptor; HER2; Human epidermal growth factor receptor 2, Bcl2; B-cell CLL/lymphoma 2; GG;I genomic grade index, IntClust; Integrated Clusters; DLDA: diagonal linear discrimination analysis, AJCC; American Joint Committee of Cancer.

Table-2: Multivariable logistic regression models analysis for pathological complete response (pCR) or residual cancer burden (RCB) in neo-adjuvant.

A. Multivariable logistic regression models analysis for pathological complete response (pCR) in the Nottingham anthracycline-based neo-adjuvant BC cohort (SPAG5 protein expression)				
Variables	**OR**	**95.0% CI**		**P value**
		Lower	**Upper**	
SPAG5 protein expression (+)	8·75	2·42	31·62	**0·0010***
Ki67 protein expression (+)	2·81	0·77	10·24	0·11
Bcl2 protein expression (+)	0·19	0.05	0·69	**0·010***
TOP2A protein expression (+)	3·81	0·98	14·73	0·053
ER protein expression (+)	0·77	0·42	2·84	0·25
HER2 expression (+)	0·84	0·23	3·12	0·79
Taxane (+)	0·67	0·21	2·21	0·52
Age (continuous)**	1·04	0·98	1·10	0·25
AJCC stage (i/II us III/IV)	0.35	0·109	1·52	0·084
Histological Grade (G1/2 vs G3)	0·417	0·11	1·54	0·18
B. Multivariable logistic regression models analysis for pathological complete response (pCR) in the University of Texas MD Anderson Cancer Centre Taxane/Anthracycline-based neo-adjuvant cohort				
***SPAG5* transcript) (+)**	1·71	1·07	2·74	**0**·024*****
Pathological Complete response (pCR) prediction signature Low vs high	1·17	0·44	3·10	0·75
96- gene genomic grade index (GGI) 26 Grade 1/Grade2 a vs Grade 2 b/Grade 3-like	0·26	0·09	0·78	**0**·016*****
30-gene (DLDA30) 27 Low vs High	1·17	0·44	3·10	0·75
PAM-50 Molecular subclasses 28				**0·042***
PAM50-LumA vs others	0·16	0·04	0·58	0·**0060***
PAM50-LumB vs others	0·24	0·07	0·88	**0·031** *
PAM50-HER2 vs others	0·14	0·03	0·66	**0·013** *
PAM50-Basal-like vs o	0·28	0·07	1·10	0·068
Clinical AJCC stage I/II vs III	0.31	0·76	0·45	**0·012***
Histological grade G1/G2 vs G3	2·37	1·15	4·89	**0·020***
Age of patients (continuous)	0·99	0·96	1·01	0·26
Oestrogen receptor status (+)	0·46	0·21	1·04	0·063
Progesterone receptor (+)	1·09	0·86	1·39	0·47
C. Multivariable logistic regression models analysis for residual cancer burden (RCB) in neo-adjuvant cohorts in Multicentre phase II neo-adjuvant clinical trial cohort (NCT00455533; n=253)				
***SPAG5* transcript) (+)**	1·80	1·02	3·02	**0·044***

Oestrogen receptor status (+)	0·59	0·25	1·36	0·21
Progesterone receptor (+)	0·41	0·02	1·02	**0**·04**2***
HER2 (+)	0·96	0·36	2·62	0·94
Age (≥50 years)	0·40	0·22	0·73	**0·0030***
Size(≥5cm)	0·59	0·32	1·09	0·090

*Statistically significant at p<0·05

**Age was a continuous value with increments of 1 years.

SPAG5; Sperm-associated antigen, ER; Oestrogen receptor; HER2; Human epidermal growth factor receptor 2, TOP2A; Topoisomerase II alpha, Bcl2; B-cell CLL/lymphoma 2; GG;I genomic grade index, IntClust; Integrated Clusters; DLDA: diagonal linear discrimination analysis, AJCC; American Joint Committee of Cancer

References:

1. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, et al. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *Eur J Cancer* 2013; **49**: 1374-403.
2. Diamandis M, White NM, Yousef GM. Personalized medicine: marking a new epoch in cancer patient management. *Mol Cancer Res* 2010; **8**: 1175-87.
3. Schmidt M, Victor A, Bratzel D, et al. Long-term outcome prediction by clinicopathological risk classification algorithms in node-negative breast cancer--comparison between Adjuvant!, St Gallen, and a novel risk algorithm used in the prospective randomized Node-Negative-Breast Cancer-3 (NNBC-3) trial. *Ann Oncol* 2009; **20**: 258-64.
4. Peto R, Davies C, Godwin J, et al. Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. *Lancet* 2012; **379**: 432-44.
5. Yerushalmi R, Woods R, Ravdin PM, Hayes MM, Gelmon KA. Ki67 in breast cancer: prognostic and predictive potential. *Lancet Oncol* 2010; **11**: 174-83.
6. 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 and Treat* 2010; **120**: 83-93.
7. Manning AL, Bakhoum SF, Maffini S, Correia-Melo C, Maiato H, Compton DA. CLASP1, astrin and Kif2b form a molecular switch that regulates kinetochore-microtubule dynamics to promote mitotic progression and fidelity. *EMBO J* 2010; **29**: 3531-43.
8. McShane LM, Altman DG, Sauerbrei W, et al. Reporting recommendations for tumor marker prognostic studies (REMARK). *Journal of the National Cancer Institute* 2005; **97**(16): 1180-4.
9. Chin SF, Teschendorff AE, Marioni JC, et al. High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol* 2007; **8**: R215.
10. Ivshina AV, George J, Senko O, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res* 2006; **66**: 10292-301.
11. Curtis C, Shah SP, Chin SF, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 2012; **486**: 346-52.
12. Wang Y, Klijn J, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005; **365**: 671 - 9.
13. Desmedt C, Piette F, Loi S, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clinl Cancer Res* 2007; **13**: 3207-14.
14. Schmidt M, Bohm D, von Torne C, et al. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res* 2008; **68**: 5405 - 13.
15. Jezequel P, Campone M, Roche H, et al. A 38-gene expression signature to predict metastasis risk in node-positive breast cancer after systemic adjuvant

chemotherapy: a genomic substudy of PACS01 clinical trial. *Breast Cancer Res and Treat* 2009; **116**: 509-20.

16. Abdel-Fatah TM, Powe DG, Agboola J, et al. The biological, clinical and prognostic implications of p53 transcriptional pathways in breast cancers. *J Pathol* 2010; **220**: 419-34.
17. Abdel-Fatah TM, Perry C, Dickinson P, et al. Bcl2 is an independent prognostic marker of triple negative breast cancer (TNBC) and predicts response to anthracycline combination (ATC) chemotherapy (CT) in adjuvant and neoadjuvant settings. *Ann Oncol* 2013; **24**: 2801-7.
18. Abdel-Fatah TM, McArdle SE, Agarwal D, et al. HAGE in triple negative breast cancer (TNBC) is a novel prognostic, predictive and actionable biomarker: A Transcriptomic and protein expression analysis. *Clinl Cancer Res* 2016 ;**22** : 905-14.
19. Hatzis C, Pusztai L, Valero V, et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* 2011; **305**: 1873-81.
20. Horak CE, Pusztai L, Xing G, et al. Biomarker analysis of neoadjuvant doxorubicin/cyclophosphamide followed by ixabepilone or Paclitaxel in early-stage breast cancer. *Clinl Cancer Res* 2013; **19**: 1587-95.
21. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003; **13**: 2498-504.
22. Shabalin AA, Tjelmeland H, Fan C, Perou CM, Nobel AB. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics* 2008; **24**: 1154-60.
23. Symmans WF, Peintinger F, Hatzis C, et al. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *J Clin Oncol* 2007; **25**: 4414-22.
24. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods* 2007; **39**: 175-91.
25. Buechler S. Low expression of a few genes indicates good prognosis in estrogen receptor positive breast cancer. *BMC Cancer* 2009; **9**: 243.
26. Liedtke C, Hatzis C, Symmans WF, et al. Genomic grade index is associated with response to chemotherapy in patients with breast cancer. *J Clin Oncol* 2009; **27**: 3185-91.
27. Hess KR, Anderson K, Symmans WF, et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J Clin Oncol* 2006; **24**: 4236-44.
28. Parker JS, Mullins M, Cheang MC, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* 2009; **27**: 1160-7.
29. Dexter TJ, Sims D, Mitsopoulos C, et al. Genomic distance entrained clustering and regression modelling highlights interacting genomic regions contributing to proliferation in breast cancer. *BMC Syst Biol* 2010; **4**: 127.
30. Hayes DF, Khoury MJ, Ransohoff D. Why Hasn't Genomic Testing Changed the Landscape in Clinical Oncology? *Am Soc Clinl Oncol Edu Book* 2012: e52-5.
31. Coates AS, Winer EP, Goldhirsch A, et al. -Tailoring therapies-improving the management of early breast cancer: St Gallen International Expert

Consensus on the Primary Therapy of Early Breast Cancer 2015. *Ann Oncol* 2015; **26**: 1533-46.

32. Jonsson G, Staaf J, Vallon-Christersson J, et al. Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. *Breast Cancer Res* 2010; **12**: R42.
33. Bartlett JM, McConkey CC, Munro AF, et al. Predicting Anthracycline Benefit: TOP2A and CEP17-Not Only but Also. *J Clin Oncol* 2015; **33**: 1680-7.
34. Huang L, Zheng M, Zhou QM, et al. Identification of a gene-expression signature for predicting lymph node metastasis in patients with early stage cervical carcinoma. *Cancer* 2011; **117**: 3363-73.
35. Yuan LJ, Li JD, Zhang L, et al. SPAG5 upregulation predicts poor prognosis in cervical cancer patients and alters sensitivity to taxol treatment via the mTOR signaling pathway. *Cell death & disease* 2015; **6**: e1784.
36. Bertheau P, Espie M, Turpin E, et al. TP53 status and response to chemotherapy in breast cancer. *Pathobiology* 2008; **75**: 132-9.

Figure legends

Figure 1: Concept diagram presenting each of the patient cohorts along with a summary of the hypotheses and methodology applied to each.

Figure 2 (A-F): Clinical outcome of *SPAG5*-gene-gain/amplification and transcript in the METABRIC cohort. Kaplan-Meier curves showing the relationship between *SPAG5*-gene-gain/amplification and breast cancer-specific survival (BCSS) in all patients **(A)**, oestrogen receptor positive (ER-positive) **(B)**, and ER- negative subgroups **(C)**. Kaplan-Meier curves are showing the relationship between *SPAG5*-transcript expression and BCSS in all patients **(D)**, oestrogen receptor positive (ER-positive) **(E)**, ER- negative **(F). (G-I):** Relationship between *SPAG5* transcript expression and clinical outcome in untreated lymph node negative breast cancer (BC) cohorts. Kaplan-Meier curves showing the relationship between *SPAG5*-transcript expression and: relapse in Wang *et al* cohort **(G)**, distant relapse in Schmidt *et al* **(H)**, and death from BC in Desmedt *et al* cohort **(I)**. See text for details.

(Homo; loss of both gene alleles, hetero; loss of one copy of the gene), Neu; 2 copies of the gene, Gain; >2 copies of the gene but <6 copies and Amp; amplification ≥6 copy of the gene)

Figure 3 (A-C): Relationship between SPAG5-protein expression and clinical outcome in a large, well-characterized cohort of Nottingham Historical Early Stage BC cohort (n=1650). Kaplan-Meier curves showing the relationship between SPAG5-protein expression and breast cancer-specific survival (BCSS) in all patients **(G)**, oestrogen receptor positive (ER-positive) **(H)** and ER- negative subgroups.

Figure 3 (D-F): Clinical outcome of ER-negative breast cancer stratified according to SPAG5-protein expression and adjuvant chemotherapy treatment status. Kaplan-Meier curves showing the relationship between SPAG5-protein expression and breast cancer specific survival (BCSS) in all ER- cohort cases **(D). E-F:** Kaplan-Meier curves showing the relationship between the adjuvant chemotherapy protocols (no chemotherapy (No-CT), CMF (cyclophosphamide, Methotrexate and 5-Flourouracil) and anthracycline combination therapy) and BCSS in low SPAG5-protein expression (-) **(E)**, and high SPAG5-protein expression (+) **(F)**. See text for details.

Supplementary figure legends:

Supplementary-Fig.S1: Concept diagram summarizing the artificial neural network analysis methodology.

Supplementary-Fig.S2 (A-B): Representative photomicrographs showing SPAG5-protein expression in breast cancer tissue. (A) Low SPAG5-protein expression (-) and **(B)** high Spag5-protein expression (+) in neoplastic cells (magnification x 200)

Supplementary-Fig.S3: Gene interaction maps. Visualization of the top 100 interactions of the common proliferation genes in the Nottingham discovery cohort, showing *SPAG5* as a central hub(Cystoscope software). .

Supplementary-Fig.S4: Gene interaction maps. Visualization of the interaction map of proliferation related factor KIF2C where SPAG5 again holds a prominent position (Cystoscope software).

Supplementary-Fig.5 (A-F): SPAG5-gene copy number aberrations (CNA), as determined by SNP analysis in the METABRIC cohort. The SPAG5-gene-CNA in different histological grades **(A)**, PAM50 molecular breast cancer subtypes: Basal-like (Basal), HER2-enrich (HER2+), luminal A (LumA), Luminal B (LumB) and normal breast like (Normal). **(B)**. Box-and-Whisker plots demonstrating the correlation between SPAG5 transcript expression and both its CNA **(C)**, ER expression **(D)**, molecular subclasses **(E)**, and grade **(F)**. **(G-H):** The relationship between integrative-clusters (Int-Clust1-10) and SPAG5 CNAs **(G)** and transcript **(H)**.

Supplementary-Fig.6 (A-F): Clinical outcome of *SPAG5* transcript in the METABRIC cohort. Kaplan-Meier curves showing the relationship between *SPAG5*-transcript and breast cancer-specific survival (BCSS) in low risk BC [Nottingham Prognostic index (NPI)<3·4] **(A)**, lymph node negative **(B)**, and lymph node positive **(C)** subgroups.

Supplementary-Fig.7 (A-E): Relationship between *SPAG5* transcript levels and clinical outcome in the combined multicentre cohort (MCC, n=5439). The forest plot showing the impact of *SPAG5* transcript on survival in terms of hazard ratio (HR) and a confidence interval (CI) in different cohorts all at once provides a better (visual) insight into the variability of results between studies **(A)**. Kaplan-Meier curves showing the relationship between *SPAG5*-transcript expression and the risk of relapse or death from breast cancer in oestrogen receptor positive (ER-positive) **(B)**, ER- negative subgroups **(C)**, lymph node negative **(D)**, and lymph node positive **(E)**.

Supplementary-Fig.8 (A-C): Clinical outcome of SPAG5 protein in the Nottingham historical early stage breast cancer cohort. Kaplan-Meier curves showing the relationship between *SPAG5*-transcript and breast cancer-specific survival (BCSS) in low risk BC [Nottingham Prognostic index (NPI)<3·4] **(A)**, lymph node negative **(B)**, and lymph node positive **(C)** subgroups.

Supplementary-Fig.9 (A-E): Clinical outcome of Nottingham early stage ER-negative breast cancer stratified according to SPAG5-protein expression and adjuvant chemotherapy treatment status. Kaplan-Meier curves showing the relationship between SPAG5-protein expression and breast cancer specific survival (BCSS) in: chemotherapy naïve patients **(A)**, chemotherapy treated cohort **(B)**, anthracycline naïve patients **(C)**, and anthracycline treated cohort **(D)**.

Supplementary-Fig.10 (A-E): The relationships between SPAG5-protein and *SPAG5*-transcript expression and pathological complete response (pCR) rate and clinical outcome following anthracycline combined neo-adjuvant chemotherapy (AC-Neo-ACT) treatment with or without taxane. Relationship between SPAG5-protein expression and pCR rate in the Nottingham Anthracycline-based neo-adjuvant BC cohort **(A)**. Kaplan-Meier curves illustrating the relationship between the expression of SPAG5-protein expression in core biopsies prior to chemotherapy **(B)** and in surgically removed residual tumour after chemotherapy, with breast cancer specific survival (BCSS) **(C)**, see text for details. (**D-E)**: Kaplan-Meier curves illustrating the relationship between the expression level of *SPAG5*-transcript and distant relapse free survival (DRFS), in all cases **(D)** and in non-pathological response cases with residual disease **(E)**.