Tissue biomarkers of breast cancer and their association with conventional pathologic features

Running Title: Protein biomarkers in breast cancer tissue

Liping Chung[1,†], Sabah Shibli[2,†], Katrina Moore[2], Elisabeth E Elder[3], Frances M Boyle[4], Deborah J Marsh[1], and Robert C Baxter[1,*]

[1]Hormones and Cancer Division, Kolling Institute of Medical Research, University of Sydney, Royal North Shore Hospital, NSW 2065 Australia; [2]Department of Breast Endocrine Surgery, Royal North Shore Hospital, NSW 2065 Australia; [3]Westmead Breast Cancer Institute, Westmead Hospital, NSW 2145 Australia and [4]Patricia Ritchie Centre for Cancer Care and Research, Mater Hospital, NSW 2060 Australia

[†] These authors contributed equally

Corresponding author: Robert C Baxter, DSc

Kolling Institute of Medical Research, Royal North Shore Hospital

St Leonards, NSW 2065, Australia

Ph: +61 2 9926 4730 Fax: +61 2 9926 8484

Email: robert.baxter@sydney.edu.au

Abstract

Background. Tissue protein expression profiling has the potential to detect new biomarkers to improve breast cancer (BC) diagnosis, staging, and prognostication. This study aimed to identify tissue proteins that differentiate breast cancer tissue from healthy breast tissue using protein-chip mass spectrometry and to examine associations with conventional pathological features.

Methods. To develop a training model, 82 BC and 82 adjacent unaffected tissue (AT) samples were analysed on cation-exchange protein chips by time-of-flight mass spectrometry. For validation, 89 independent BC and AT sample pairs were analysed.

Results. From the protein peaks that were differentially expressed between BC and AT by univariate analysis, binary logistic regression yielded two peaks that together classified BC and AT with a ROC area-under-the-curve of 0.92. Two proteins, ubiquitin and S100P (in a novel truncated form), were identified by liquid chromatography-tandem mass spectrometry and validated by immunoblotting and reactive-surface protein-chip immunocapture. The combined marker panel was positively associated with high histologic grade, larger tumour size, lymphovascular invasion, ER and PR positivity, and HER2 overexpression, suggesting that it may be associated with a HER2-enriched molecular subtype of breast cancer.

Conclusion. This independently validated protein panel may be valuable in the classification and prognostication of breast cancer patients.

Key words: Breast cancer diagnosis, mass spectrometry, proteomics, tissue biomarkers

Introduction

Breast cancer is the most frequently diagnosed cancer, and the leading cause of cancer death, in women worldwide (Jemal *et al*, 2011), with the lifetime risk of developing breast cancer estimated to be 1 in 8 in Western countries (Feuer *et al*, 1993). Patient survival has increased steadily over recent decades, attributable in part to advances in both mammographic screening (Kopans, 2011) and adjuvant systemic treatment protocols (Peto *et al*, 2012). Whereas pathological features such as tumour size, node positivity, hormone receptor positivity and HER2 overexpression have been used to guide clinicians' prescription of adjuvant therapy, true personalised medicine requires the development of better biomarkers of risk and response to therapy.

Gene expression profiling is emerging as a tool for classifying breast cancers, guiding therapy, and predicting treatment responses (Cheang *et al*, 2008; Haas *et al*, 2011). However genome and transcriptome analyses alone provide only a partial picture since alternative splicing of mRNA, combined with more than 100 unique post-translational protein modifications, mean that each gene may give rise to multiple protein species (Banks *et al*, 2000).

Analysing the proteome may provide a more dynamic reflection of the impact of the cell's genetic program on its immediate environment (Aebersold *et al*, 2005). Cancer proteomics encompasses the identification and quantitative analysis of differentially expressed proteins relative to healthy tissue counterparts at different stages of disease. Proteomic technologies can also be used to identify markers for cancer diagnosis, to

monitor disease progression and efficacy of therapy, and to identify new therapeutic targets (Srinivas *et al*, 2001).

Surface-enhanced laser desorption/ionisation time-of-flight (SELDI-TOF) mass spectrometry (MS) is a high-throughput proteomic method that involves solid-phase extraction of subsets of the proteome prior to analysis by TOF MS (Callesen *et al*, 2008). It has the ability to rapidly analyse hundreds of samples, essential for obtaining biologically and statistically relevant data in medical proteomic research. A recent review of protein profiling studies of breast cancer demonstrates that, despite a considerable diversity among these studies, there is a pattern of conformity developing, with increasing numbers of studies reporting similar peaks in protein profiles (Galvao *et al*, 2011). This suggests convergence to a set of common discriminatory peaks for breast cancer, with reproducibility across different clinical studies.

In this study we have employed SELDI-TOF MS to discover tissue biomarkers of breast cancer, and validate them on an independent sample set. We have used two immunological methods to verify the identified proteins. Finally, the expression levels of these proteins have been associated with clinical pathological variables in order to explore their potential value in breast cancer classification and prognosis.

Materials and Methods

Patient Samples

The study involved 404 patient samples comprised of 202 pairs: breast tumour tissue

(BC) and adjacent unaffected breast tissue (AT) from each subject. For the discovery phase, 102 sample pairs were obtained from the Kolling Institute Breast Tumour Bank at Royal North Shore Hospital, Sydney Australia. For independent validation, 100 sample pairs were provided by the Australian Breast Cancer Tissue Bank, Sydney Australia. All breast tissue samples were collected at the day of surgery with prior informed consent, and the study was approved by the Human Research Ethics Committee of the Northern Sydney Central Coast Area Health Service, Sydney, Australia. At the time of surgical resection, tissues were immediately taken to a pathologist, who sampled both the tumour itself and adjacent tissue of normal appearance. Both samples were snap-frozen in liquid nitrogen within 20 minutes of resection and stored at -80°C. Estrogen receptor (ER) and progesterone receptor (PR) were scored as either negative or positive by immunohistochemistry, using rabbit monoclonal SP1 (Biocare Medical, Concord, CA) and mouse monoclonal Clone PgR636 (Dako, Carpinteria, CA), respectively. HER2 status was defined as positive or negative by immunohistochemistry using the HercepTest (Dako). Any equivocal result using this test was confirmed by FISH.

Tissue Preparation

Approximately 20 mg of each tissue sample (BC or AT) was prepared for proteomic analysis by grinding with a mortar and pestle while frozen in liquid nitrogen, then solubilising in 10 volumes of lysis buffer (9.5 M urea, 2% 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS), 1% dithiothreitol). Lysates were added to a QiaShredder spin column (Qiagen, Hilden, Germany) and centrifuged (12,000 rpm, 5 min) to remove insoluble material. Samples were applied to

weak cation-exchange (CM10) protein chips (Bio-Rad Laboratories, Hercules, CA) for immediate analysis as described below, or aliquotted and stored at -80°C for future analysis. The protein concentration of each extract was determined by BCA Protein Assay (Thermo Scientific, Rockford, IL).

Preparation of protein chips

CM10 protein chips were pre-equilibrated twice with 5 µl of binding buffer (50 mM sodium acetate pH 6.0) for 5 min. Protein extracts were diluted 1:5 with binding buffer and 5 µl of each diluted extract was pipetted onto the chips. All samples were run in duplicate. Chips were then incubated with shaking using a MicroMix 5 (settings: form 20, amplitude 4), (EURO/DPC Instrument Systems, Flanders, NJ) for 90 min at room temperature. Each spot was treated with 2 x 1 µl of 50% cyano-4-hydroxycinnamic acid in 50% acetonitrile containing 0.5% trifluoroacetic acid (TFA), and air-dried.

Generation of MS profiles

Protein profiles were initially obtained using a PBSIIc protein chip reader (Bio-Rad Laboratories, Hercules, CA), and in the latter part of the study, a SELDI Enterprise Edition protein chip reader (Bio-Rad). Mass spectra were generated for each sample in the mass/charge (*m/z*) range of 1,000-30,000 with a laser intensity setting of 175 (arbitrary units). The laser was optimised for data collection between 1,000 and 15,000 *m/z*, with detector sensitivity set at 8. Peaks below 1,000 *m/z* were deflected away from the detector. Data were averaged from 328 spectra evenly distributed across each spot. Mean values from duplicate spectra for each sample were used in all subsequent analyses. The *m/z* value for each peak was determined using external calibration with

protein standards including bovine insulin (5734.51 Da), equine cytochrome C (12362 Da), equine apomyoglobulin (16952.3 Da), and bovine carbonic anhydrase (29023.70 Da) (Sigma-Aldrich, St. Louis, MO). After calibration, spectra were baseline-subtracted and normalised using the total ion current between 1500 and 15,000 *m/z*. Spectra that required a normalisation factor >2 were repeated, and if the high normalisation factor persisted, these data were discarded. Peak detection was initially performed using Biomarker Wizard Version 3.2.2 (Bio-Rad Laboratories, Hercules, CA) on all peaks with signal/noise ratio ≥5 and present in at least 10% of all spectra. Subsequently all MS spectra were exported to ProteinChip Data Manger v4.1 used with the ProteinChip SELDI System Enterprise Edition (Bio-Rad) to refine the combined data analysis.

MS data analysis

Data analysis was designed in three stages. For initial discovery, biomarker panels were developed on the training data set using 102 BC and AT sample pairs. Cluster analysis was performed using Biomarker Wizard version 3.2.2 (Bio-Rad). Univariate analysis of individual peaks was performed by Mann-Whitney U test using SPSS (Version 18.0, SPSS Inc., Chicago, IL). All protein peaks that significantly discriminated BC from AT at $p<0.001$ were then subjected to multivariate analysis using forward and reverse binary logistic regression (SPSS) to develop the training model. The discriminatory power of each putative marker was further described using receiver operating characteristic (ROC) area-under-the-curve (AUC) analysis. To test protein panels that were best able to discriminate BC from AT, 10-fold internal cross-validation was used as previously described (Ambroise & McLachlan, 2002; Scarlett *et al*, 2006). External validation was carried using an independent set of 100 paired BC and AT samples.

After external validation, to consolidate and unify the initial discovery and validation data a further analysis was performed on the combined data sets. This coincided with the acquisition of new peak cluster analysis software, ProteinChip Data Manager Version 4.1 (Bio-Rad). Similar to the initial discovery phase, both univariate analysis using nonparametric statistics and multivariate analysis using binary logistic regression were applied, confirming a final two-protein marker panel and allowing calculation of overall estimates of sensitivity and specificity, accuracy and ROC AUC values. The final stage of data analysis was to re-evaluate the two-protein panel on the separate training and validation sets to ensure consistency between the findings from the new and original software packages. In this re-testing, all common peaks obtained from the combined dataset study were used for each regression analysis to achieve classification of tumour samples separately in the training and validation sets.

Protein identification

For purification of the putative biomarkers, tissue lysates were fractionated using a cation-exchange resin (Mustang S, Pall Corp., Ann Arbor, MI) with stepwise pH elution from pH 4 to pH 9 in a 96-well filter plate format (AcroPrep™, Pall) as previously described (Chung *et al*, 2009). Proteins of interest in the eluates were monitored by SELDI-TOF MS on normal-phase (NP20) chips. Fractions containing an approximately 8.5 kDa putative biomarker were further purified using reverse phase liquid chromatography (LC) on a 250 x 4.6 mm Jupiter 5 μm 300-Å C18 column (Phenomenex, Lane Cove, Australia), eluted with a 35-min linear gradient from 15-60% acetonitrile in 0.1% TFA at 1.5 ml/min, followed by separation on 12% SDS-PAGE

detected with SYPRO ruby protein stain (Invitrogen, Eugene, OR). Protein bands of interest were excised from the gel and analysed using both nanoLC-ESI-MS/MS and MALDI-TOF peptide mass fingerprinting by the Australian Proteome Analysis Facility (Macquarie University and University of New South Wales, Sydney, Australia). The protein peak at 9.2 kDa was purified and identified in a similar manner.

Immunological validation of protein markers

To detect ubiquitin and S100P by western blotting, BC and AT tissue extracts were separated by 12% SDS-PAGE and transferred to PVDF membranes (Bio-Rad). Membranes were blocked for 1 h at room temperature with 5% skim milk. Ubiquitin was detected by incubating the transferred membranes for 2 h at room temperature on a shaking platform with anti-human ubiquitin monoclonal antibody (R&D Systems, Minneapolis, MN) at 1:500 dilution in 5% skim milk. For S100P western blotting, samples were concentrated 5-fold by centrifugal ultrafiltration with 3-kDa MW cut-off (Nanosep 3K Omega, Pall Corp., Ann Arbor, MI) prior to electrophoresis. This was necessary to increase detection sensitivity. Concentrated samples were separated and transferred, and membranes blocked, as described above, and S100P was detected by incubating overnight at 4°C with rabbit anti-human antibody (Invitrogen) at 1:500 dilution in 5% skim milk. Secondary antibody, peroxidase-linked anti-rabbit IgG (1:2000) was added for 1 h at room temperature and the protein bands were visualised by enhanced chemiluminescence using the SuperSignal® West Pico Luminol/Enhancer solution (Thermo Scientific, Rockford, IL). Western blot data were imaged using the LAS 3000 imaging system (Fujifilm, Stamford CT) and the images were analysed with Multi-Gauge version 3.0 software (Fujifilm). The quantitative data were normalised to

the loading control of β-actin, and analysed using the Wilcoxon signed rank test (SPSS).

To confirm the identity of the *m/z* 8558 protein peak by protein chip immunocapture, pre-activated RS100 protein chips (Bio-Rad) were pre-coupled with 2 μg of monoclonal anti-human ubiquitin antibody (R&D) in 50 mM $NaHCO_3$ buffer (pH 9.2) at 4ºC. The spots were washed with 50 μM BSA to block the remaining active sites. Tissue lysates were diluted 1:5 in buffer containing 50% human serum in 0.1% Triton X-100 in PBS, spotted onto RS100 protein chips, and incubated 2 h at room temperature on a shaker to achieve optimal binding. After washing with PBS, all spots were rinsed by 50 mM Tris-HCl, 1 M urea, 0.1% CHAPS, 0.5 M NaCl pH 7.2. After further washing in 5 mM HEPES pH 7.2, the spots were coated with 2 x 1 μl of 50% sinapinic acid in 50% acetonitrile, 0.5% TFA and air-dried. The chips were then analysed on the SELDI-TOF MS. A His-tagged recombinant ubiquitin standard (10.6 kDa; R&D) was used as a control. The *m/z* 9226 protein peak was similarly verified using RS100 protein chips to confirm its identity as S100P. Prior to protein chip preparation, all tissue extracts were pre-concentrated as described above for Western blotting. RS100 protein chips were pre-coupled with 2 μg of rabbit anti human S100P antibody (Invitrogen) in 50 mM $NaHCO_3$ buffer (pH 9.2) at 4ºC. The samples were then treated and analysed as described above. His-tagged recombinant S100P (12.6 kDa; Novus Biologicals, Littleton, CO) was used as a control.

Statistical analysis of clinical features

The association between levels of the two protein markers, individually and in combination, and tumour pathologic variables (tumour size, histological grade,

lymphovascular invasion, lymph node involvement, estrogen and progesterone receptor status, and HER2 expression) was examined using the Mann-Whitney U test (SPSS). Subgroup analyses were also performed, in which lymph node negative (n=84) or positive (n=85) groups were analysed separately. Significance was set at $P<0.05$.

Results

Patient characteristics

A total of 202 pairs of tissue samples were used in this study, generating 808 spectra, of which 684 (duplicate spectra on 171 pairs of samples) were subjected to full analysis. Of the 102 pairs of samples selected for the training stage, 82 pairs were fully analysed. Eight of the remaining 20 pairs were excluded on clinico-pathologic grounds: 4 had DCIS, 2 had neo-adjuvant treatment and 2 had recurrent tumours; and 12 sample pairs were excluded when their mass spectra did not meet normalisation criteria. For the validation set of 100 samples pairs, 89 pairs of the subjects were analysed. Seven sample pairs were excluded on clinico-pathologic grounds: 4 had neo-adjuvant therapy, 1 had metastatic disease and 2 had recurrent disease; 3 sample pairs were lost during preparation; and 1 pair was excluded when the mass spectra did not meet normalisation criteria. The median age for the patients included in the training and validation sets was 60 (range 28-92) and 58 (range 27-85), respectively. The clinical pathologic characteristics of the tumours including histologic type and grade, size, presence of lymphovascular invasion (LVI), hormone receptor (estrogen receptor (ER), progesterone receptor (PR)) and human epidermal growth factor receptor 2 (HER-2) status, as well as lymph node status are presented in Table 1.

Selection of protein biomarker panel by MS based protein profiling

The training set sample pairs (BC and AT) were subjected to MS analysis in duplicate to identify putative protein biomarkers that could distinguish tumour from unaffected tissue. The 82 sample pairs whose spectra were amenable to normalisation yielded 328 spectra, from which 53 common peaks were determined by clustering analysis. Of these, 14 peaks (*m/z* 1337, 1705, 1842, 2033, 3790, 3804, 8346, 8548, 8599, 9205, 9239, 9292, 9641, 12220) were significantly differentially expressed ($P<0.005$, Mann-Whitney test). These individual putative biomarkers had ROC-AUC values ranging from 0.70 to 0.84. The 14 peaks were tested in forward and reverse binary logistic regression analysis with 10-fold cross-validation. This produced a final panel of 3 peaks (*m/z* 1842, 8599, 9292) that classified BC and AT, with ROC AUC of 0.87 as shown in Figure 1A (curve Ti).

Independent validation

The three putative biomarkers were tested using an independent validation set of 100 sample pairs, of which 89 pairs of spectra (in duplicate, 356 spectra) could be analysed after normalisation. For the validation set, 57 common protein peaks were determined by clustering analysis. Testing the 3-protein panel derived from the training set on the independent sample set of 89 BC and 89 AT samples gave a ROC AUC of 0.91 (Figure 1A, curve Vi). The sensitivity and specificity were 80.9% and 91%, respectively, and overall accuracy was 90%.

Re-analysis of combined data sets

To increase the statistical power of the training and validation analyses and confirm the results using a newer software version, we combined the data sets into a single analysis of all 171 breast tissue sample pairs. Using new clustering analysis software, ProteinChip Data Manager Version 4, we found 28 peaks common to all spectra in the *m/z* range of 2,500 to 15,000. Peaks of lower mass were excluded from this analysis because the putative marker at *m/z* 1842 had been determined by LC-MS/MS to be non-peptide in nature (data not shown). By univariate analysis (Mann-Whitney), the significant peaks ($p<0.001$) were selected with the additional criterion that individual ROC-AUC was at least 0.80, summarized in Table 2. Multivariate analysis using binary logistic regression again confirmed the two protein markers at *m/z* 8558 and 9226. The difference in *m/z* values from those determined in the initial training set analysis (*m/z* 8599, 9292) is larger than expected and may be attributable to the fact that they are averaged from 684 spectra (171 sample pairs in duplicate) rather than 328 spectra (82 sample pairs in duplicate), re-calibration of standard curves between the initial and subsequent analyses, the use of different analysis software, and the relative mass-inaccuracy of this technique. Both protein peaks were elevated in BC tissue relative to AT. The sensitivity and specificity for the binary classification using the combined 2-marker panel were 77.2% and 88.9% respectively, with a ROC AUC value of 0.92 (Figure 1B, curve C).

Re-testing of initial training and validation sets

For final confirmation of the potential 2-marker panel, it was re-tested on the original separate training and validation sets. The sensitivity and specificity of the classification for breast tissue biopsy samples were 73.2% and 87.8% respectively in the training set,

compared to 80.9% and 91% in the validation set. Their corresponding ROC AUC values were 0.86 (curve Tr) and 0.91 (curve Vr) for the training and validation sets, respectively (Figure 1C).

Together, these results suggest that two protein biomarkers in combination provide efficient discrimination between breast cancer tissue and healthy tissue. Figure 1(D-F) demonstrates the performance of the two protein peaks of *m/z* 8558 and 9226 alone and in combination. By paired sample t-test a significant difference between BC and AT groups was found for each protein tested separately (Figure 1D and E, n=171, $p<0.001$). For the two-protein combined panel, the mean value was 3.3-fold increased in BC compared to AT samples (Figure 1F, n=171, $p<0.001$).

Identification and verification of putative biomarkers

Both proteins of *m/z* 8558 and 9226, retained by weak cation-exchange protein chips, were significantly increased in breast cancer tissue. For identification, initial purification was carried out using cation-exchange followed by reversed-phase HPLC. Eluted fractions were pooled and fractionated by SDS-PAGE, and bands around 8 kDa were excised for final identification by LC-MS/MS. Ubiquitin was identified from 6 peptides (two overlapping) giving 72% sequence coverage. The calculated mass of monomeric ubiquitin (8560 Da) was in good agreement with the consensus mass obtained experimentally with SELDI (*m/z* 8558). Similarly, analysis of the marker of approximately 9.2 kDa identified it as a fragment or variant of S100P (10,400 Da) from two peptides giving 24% sequence coverage relative to full-length S100P (Supplementary Figure S1). Notably, the two peptides found in this study were identical

to those previously used to identify S100P in a MALDI-MS study of proteins upregulated in colorectal cancer (Lam *et al*, 2010).

Immunological verification of the two protein identities was performed using both Western blotting and protein chip immunocapture. For ubiquitin, Western blot confirmed differential expression of this protein between BC and AT tissue extracts. Figure 2A shows that for BC and AT samples from four randomly selected patients, relative overexpression of ubiquitin in the cancer samples was observed. When quantitated and analysed for 8 randomly selected sample pairs, the increase in ubiquitin in BC was significant (Figure 2B, p=0.017, Wilcoxon signed rank test). The identity of this protein as ubiquitin was also verified by immunocapture on RS100 protein chips (Figure 2C). The *m/z* 8558 peak, captured by immobilised ubiquitin antibody and displayed by SELDI-TOF MS, was increased in two BC samples in panels (ii) and (iv) compared to their corresponding AT samples in panels (i) and (iii), and absent when the capture antibody was nonimmune IgG (panel (vi)). Panel (v) shows His-tagged recombinant ubiquitin (10.6 kDa) as a control.

Similarly, the expression of S100P was also examined by Western blot in 8 random sets of BC and AT samples. Figure 2D shows the Western blot data for four pairs, indicating variable levels of this protein between patients, with upregulation in BC samples. When quantitated and analysed for all 8 sample pairs, the increase in immunoreactive S100P in BC was significant (Figure 2E, p=0.012, Wilcoxon signed rank test). By immunocapture using the same S100P antibody immobilised on RS100 protein chips, an apparently truncated form (*m/z* 9226) of S100P protein was observed, similar to that

found in the discovery program using CM10 cation-exchange chips. This peak was more abundant in BC samples (panels (ii) and (iv)) than in the corresponding AT samples (panels (i) and (iii)), and absent when the capture antibody was nonimmune IgG (panel (vi)) (Figure 2F). Panel (v) shows His-tagged recombinant S100P (12.6 kDa) as a control.

To further confirm the identity of the 9.22 kDa protein as a short form of S100P associated with breast cancer, we also isolated this protein from cell lysates prepared from MCF-7 breast cancer cells. As shown in Supplementary Figure S2(A-C), this protein could be immunoprecipitated from MCF-7 lysates using three different S100P antibodies (rabbit monoclonal, mouse polyclonal and rabbit polyclonal). Together with the S100P sequence data (Supplementary Figure S1), this unequivocally confirms its relationship to S100P. Also visible in the immunoprecipitates was a smaller peak of 10.48 kDa, presumably representing full-length S100P. The 9.22 kDa form could be separated from the full-length protein by further purification on reverse-phase HPLC (Supplementary Figure S2D).

Association of two protein biomarkers and their combination with prognostic variables

To investigate the potential prognostic value of ubiquitin and S100P separately and in combination in breast cancer, we initially examined the association of each protein with variables including tumour stage, nodal stage, histologic type and grade, hormone receptor (ER, PR) and HER2 status, and lymphovascular invasion (LVI). As shown in Table 3, significant positive associations were seen between expression of the short

form of S100P and tumour size, higher grade, LVI, lymph node involvement, hormone receptor positive status, and HER2 overexpression, whereas for ubiquitin a significant association was only seen with tumour size, grade, and HER2. When analysed together (Table 3), the combined panel was significantly associated with tumour histologic grade, size, and LVI, and also with ER-positive (ER+) and PR-positive (PR+) status and HER2 overexpression (Figure 3).

Since levels of the short form of S100P showed stronger associations than ubiquitin with each of the pathological indicators (except for grade), and appeared to point to an ER/PR+, HER2 overexpressing phenotype (possibly corresponding to a "HER2-enriched" molecular subtype (Reis-Filho & Pusztai, 2011)), we undertook further analysis of its relationship to these prognostic features. When examined separately for ER- and ER+ tumours, high S100P expression in both groups was equally associated with tumour size and the presence of LVI (not shown). However, the association between S100P and lymph node involvement was only significant for ER- tumours ($p=0.010$). In contrast, the association between S100P and HER2 overexpression was only significant for ER+ tumours ($p=0.004$), supporting the concept that a high S100P level might be associated with a hormone receptor-positive, HER-2 enriched molecular subtype.

When examined separately for lymph node-negative and positive tumours, the positive association between ubiquitin, the short form of S100P, or the combined panel and LVI, ER+ status, and PR+ status was entirely attributable to the lymph node-positive tumours. A significant relationship between the combined panel and HER2

overexpression was also confined to the lymph node-positive tumours (Supplementary Table S1). This subanalysis again points to a link between high expression of the short form of S100P in breast tumours, and an ER/PR positive, HER2 overexpressing phenotype that has been associated with markers of poor patient outcome without treatment. However, because sample numbers are low in some subanalyses, these interpretations should be regarded as preliminary.

Discussion

We have used SELDI-TOF MS to discover two proteins that, in combination, show high discrimination between breast cancer and healthy breast tissue samples. A limitation of the protocol was that no microdissection was used, so that tissue samples could have contained heterogeneous cell types. Despite this technical limitation, a robust panel of two putative breast cancer biomarkers was discovered, and verified on an independent sample set. After purification, the proteins were identified by LC-MS/MS as ubiquitin and a truncated form of the S100-family member, S100P.

SELDI-TOF MS has been used previously to discover tissue biomarkers in various cancers, although the majority of such studies in breast cancer have examined serum rather than tumour tissue. Included among proteins previously identified from breast tumour tissue lysates are albumin fragments (Gast *et al*, 2009) and complement C3a (Zhang *et al*, 2012), both presumably derived from the circulation. Tissue proteomic profiling using SELDI-TOF MS has also yielded peak clusters that can contribute to the classification of breast tumours into molecular subtypes (Brozkova *et al*, 2008; Goncalves *et al*, 2008) that resemble the luminal A and B, basal, and HER2-like

subtypes defined by gene expression analysis (Reis-Filho & Pusztai, 2011).

Of the two breast cancer-associated proteins identified in this study, ubiquitin is a small protein of 76 amino acids that is involved both in apoptotic signalling (Vucic *et al*, 2011) and transcriptional regulation (Hammond-Martel *et al*, 2011). Although monomeric ubiquitin has been identified in several previous biomarker studies in breast cancer, its exact relationship to disease status is unclear. In a SELDI-TOF MS study of breast cancer cell lines, we previously discovered ubiquitin as a strongly downregulated protein following treatment with chemotherapeutic drugs (Leong *et al*, 2007). Another SELDI analysis found the combination of a high ubiquitin level and low ferritin light chain level to be a positive prognostic marker in node negative breast cancer (Ricolleau *et al*, 2006). In contrast, SELDI was also used to show that a protein of similar mass (not identified as ubiquitin), was a significant predictive factor for axillary lymph node metastasis (Nakagawa *et al*, 2006). In a MALDI MS analysis of microdissected cells from invasive breast cancer and healthy (reduction mammoplasty) tissue, ubiquitin was one of a cluster of proteins with increased expression in the cancer tissue (Sanders *et al*, 2008).

Several E3 ubiquitin ligases are regarded as tumour suppressors in breast cancer and are either mutated or downregulated; in contrast, some others are regarded as oncogenes and are overexpressed (Chen *et al*, 2006). Among key downregulated or mutated E3 ligases are BRCA1 and Siah1, involved in DNA repair and transcriptional regulation, among other functions. E3 ligases downregulated in cancer are involved in both monoubiquitination (Hahn *et al*, 2012) and polyubiquitination (Wen *et al*, 2010), and low expression of the E3 ligase Siah1 is associated with poorer disease-free survival in

women with breast cancer (Confalonieri *et al*, 2009). It may be speculated that the increased level of monomeric ubiquitin that we observed associated with larger tumours, higher grade, and HER2 overexpression, but not with other pathological markers (Table 3), reflects a decrease in the activity of some key ubiquitin ligase complexes. Interestingly, a component of the Siah1 ubiquitination complex, calcyclin-binding protein/Siah1-interacting protein (CacyBP/SIP), has increased expression in breast cancer tissue compared to adjacent unaffected breast tissue, and is associated with markers of poor prognosis (Wang *et al*, 2010). CacyBP/SIP is a documented binding partner of S100P (Filipek *et al*, 2002), raising the possibility that the disruption of ubiquitination pathways in breast cancer might be involved in the increased levels of both of the cancer-related biomarkers discovered in our study.

In contrast to the relatively weak associations observed between elevated ubiquitin levels and tumour size, higher grade, and HER2 overexpression, a high level of the novel short form of S100P was positively associated with larger tumours, higher grade, LVI, lymph node involvement, ER/PR positivity, and HER2 overexpression. Of the two identified biomarkers, S100P made the stronger contribution towards the association of the combined panel towards each of these pathological features apart from tumour grade. Since the association between S100P and HER2 overexpression was exclusive to the ER+ tumours ($p=0.004$), and absent in the ER- subgroup, a high tissue S100P level may point to a group of tumours with high ER/PR positive status, HER2 overexpression, and – given the association with size, grade and LVI – relatively poor outcome, although our study did not include actual outcome variables. This corresponds most closely to the "HER-enriched" breast cancer subtype (Reis-Filho & Pusztai, 2011; Slamon *et al*, 1987), and suggests that S100P might have potential both in the

classification of breast cancer, and possibly as a target for therapy.

S100P is a member of the calcium-binding S100 protein family that contain a characteristic structural domain known as the EF hand motif (Marenholz *et al*, 2004). There are at least 24 homologous S100 proteins with similar subcellular localisation, but differing in expression pattern and function (Marenholz *et al*, 2004). S100 proteins are low molecular weight (10-12 kDa) acidic proteins that exist as intracellular or secreted homo- or hetero-dimers with composition depending on the abundance of individual family members and the cellular context (Santamaria-Kisiel *et al*, 2006). Although the factors that regulate S100P have not been studied extensively, DNA microarray studies have included S100P among panels of genes upregulated by estradiol (Terasaka *et al*, 2004), progesterone (Bray *et al*, 2005) and HER2 overexpression (Mackay *et al*, 2003). These preliminary gene expression reports are consistent with the clinical associations we observed between high S100P levels and ER/PR positive and HER2 overexpressing tumours.

S100P has been associated with the progression of several types of cancer including pancreatic, prostate, colorectal and breast, through effects on tumour growth and metastasis (Jiang *et al*, 2011; Lam *et al*, 2010). At least some of its effects have been shown to be mediated through extracellular interaction with RAGE (receptor for activated glycation end products) (Arumugam *et al*, 2004). Several studies of pancreatic cancer-related molecular profiles have identified S100P as a significantly elevated gene (Crnogorac-Jurcevic *et al*, 2003; Logsdon *et al*, 2003), whose upregulation is an early event in the development of pancreatic cancer (Whiteman *et al*, 2007). In breast cancer, S100P was linked to immortalisation of breast epithelial cells *in vitro* and both tumour

progression (Guerreiro Da Silva *et al*, 2000; Schor *et al*, 2006) and early relapse (Barraclough *et al*, 2010) in patients. Survival of breast cancer patients with S100P positive carcinomas was significantly worse than those negative for S100P (Barraclough *et al*, 2009; Wang *et al*, 2006). S100P was also prominent among genes over-expressed in primary breast cancer cells from high-grade tumours (Dairkee *et al*, 2009). In contrast, gastric cancers that stain positive for S100P are associated with a better patient outcome that those that are negative for S100P (Jia *et al*, 2009).

The S100P form detected in our study by MS on cation-exchange chips, and confirmed by MS after selective binding to immobilised S100P antibody, appeared at a *m/z* value of 9226. This contrasts with the expected size of mature S100P which contains 95 amino acids and has a molecular mass of 10.4 kDa, suggesting that the observed S100P species detected by MS is a previously unreported truncated form of this protein. An amino-terminally truncated form of S100P, termed migration-inducing gene 9 protein or MIG9, has been reported in GenBank (Protein Accession No. AAS00487.1), described as an alternatively spliced product. The predicted protein is identical to S100P[8-95] except for an isoleucine to methionine substitution at S100P residue 12 (MIG9 residue 5), and has a predicted molecular mass of 9.64 kDa. If the true translation start site is methionine-5, the predicted molecular mass would be 9.21 kDa and could explain our observed peak on SELDI-TOF MS. Importantly, it is unlikely that the many immuno-histochemical studies that have measured S100P distribution in patient tissues could distinguish between S100P and these truncated forms. Mass spectrometry would be the optimal method for this identification. We have therefore identified for the first time a novel isoform of S100P that is associated with pathologic markers in breast cancer.

In conclusion, this study has discovered two protein biomarkers, ubiquitin and S100P — the latter as a novel truncated isoform — that, in combination, provide high discrimination between breast cancer tissue and healthy breast tissue. Correlation with clinical pathologic variables demonstrated that high values for the two-protein panel were associated with high histologic grade and tumour size, presence of lymphovascular invasion, ER and PR positive status, and HER2 overexpression. We propose that this independently validated protein biomarker panel may indicate a HER2-enriched breast cancer subtype with poor prognosis, and that measurement of S100P, in particular, may be valuable both in the classification of breast cancer and as a possible target for treatment.

Acknowledgements

This work was supported by Project Grant #632558 to R.C.B., K.M., D.J.M. and F.M.B. from the National Health & Medical Research Council, Australia. D.J.M. is the recipient of an ARC Future Fellowship and CINSW Fellowship, Australia. We thank Prof Ross Smith for helpful discussions. We also thank the Kolling Institute Breast Tumour Bank at Royal North Shore Hospital, and Australian Breast Cancer Tumour Bank at Westmead Hospital, Sydney Australia for their support in providing patient information and breast tissue samples. This research has been facilitated by access to the Australia Proteome Analysis Facility (APAF) funded by the National Collaborative Research Infrastructure Strategy (NCRIS).

References

Aebersold R, Anderson L, Caprioli R, Druker B, Hartwell L, Smith R (2005) Perspective: a program to improve protein biomarker discovery for cancer. *J Proteome Res* **4:** 1104-1109

Ambroise C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A* **99:** 6562-6566

Arumugam T, Simeone DM, Schmidt AM, Logsdon CD (2004) S100P stimulates cell proliferation and survival via receptor for activated glycation end products (RAGE). *J Biol Chem* **279:** 5059-5065

Banks RE, Dunn MJ, Hochstrasser DF, Sanchez JC, Blackstock W, Pappin DJ, Selby PJ (2000) Proteomics: new perspectives, new biomedical opportunities. *Lancet* **356:** 1749-1756

Barraclough DL, Platt-Higgins A, de Silva Rudland S, Barraclough R, Winstanley J, West CR, Rudland PS (2009) The metastasis-associated anterior gradient 2 protein is correlated with poor survival of breast cancer patients. *Am J Pathol* **175:** 1848-1857

Barraclough DL, Sewart S, Rudland PS, Shoker BS, Sibson DR, Barraclough R, Davies MP (2010) Microarray analysis of suppression subtracted hybridisation libraries identifies genes associated with breast cancer progression. *Cell Oncol* **32:** 87-99

Bray JD, Jelinsky S, Ghatge R, Bray JA, Tunkey C, Saraf K, Jacobsen BM, Richer JK, Brown EL, Winneker RC, Horwitz KB, Lyttle CR (2005) Quantitative analysis of gene regulation by seven clinically relevant progestins suggests a highly similar mechanism of action through progesterone receptors in T47D breast cancer cells. *J Steroid Biochem Mol Biol* **97:** 328-341

Brozkova K, Budinska E, Bouchal P, Hernychova L, Knoflickova D, Valik D, Vyzula R, Vojtesek B, Nenutil R (2008) Surface-enhanced laser desorption/ionization time-of-

flight proteomic profiling of breast carcinomas identifies clinicopathologically relevant groups of patients similar to previously defined clusters from cDNA expression. *Breast Cancer Res* **10:** R48

Callesen AK, Christensen R, Madsen JS, Vach W, Zapico E, Cold S, Jorgensen PE, Mogensen O, Kruse TA, Jensen ON (2008) Reproducibility of serum protein profiling by systematic assessment using solid-phase extraction and matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrom* **22:** 291-300

Cheang MC, van de Rijn M, Nielsen TO (2008) Gene expression profiling of breast cancer. *Annu Rev Pathol* **3:** 67-97

Chen C, Seth AK, Aplin AE (2006) Genetic and expression aberrations of E3 ubiquitin ligases in human breast cancer. *Mol Cancer Res* **4:** 695-707

Chung L, Nelson AE, Ho KK, Baxter RC (2009) Proteomic profiling of growth hormone-responsive proteins in human peripheral blood leukocytes. *J Clin Endocrinol Metab* **94:** 3038-3043

Confalonieri S, Quarto M, Goisis G, Nuciforo P, Donzelli M, Jodice G, Pelosi G, Viale G, Pece S, Di Fiore PP (2009) Alterations of ubiquitin ligases in human cancer and their association with the natural history of the tumor. *Oncogene* **28:** 2959-2968

Crnogorac-Jurcevic T, Missiaglia E, Blaveri E, Gangeswaran R, Jones M, Terris B, Costello E, Neoptolemos JP, Lemoine NR (2003) Molecular alterations in pancreatic carcinoma: expression profiling shows that dysregulated expression of S100 genes is highly prevalent. *J Pathol* **201:** 63-74

Dairkee SH, Sayeed A, Luciani G, Champion S, Meng Z, Jakkula LR, Feiler HS, Gray JW, Moore DH (2009) Immutable functional attributes of histologic grade revealed by context-independent gene expression in primary breast cancer cells. *Cancer Res* **69:** 7826-7834

Feuer EJ, Wun LM, Boring CC, Flanders WD, Timmel MJ, Tong T (1993) The lifetime risk of developing breast cancer. *J Natl Cancer Inst* **85:** 892-897

Filipek A, Jastrzebska B, Nowotny M, Kuznicki J (2002) CacyBP/SIP, a calcyclin and Siah-1-interacting protein, binds EF-hand proteins of the S100 family. *J Biol Chem* **277:** 28848-28852

Galvao ER, Martins LM, Ibiapina JO, Andrade HM, Monte SJ (2011) Breast cancer proteomics: a review for clinicians. *J Cancer Res Clin Oncol* **137:** 915-925

Gast MC, van Dulken EJ, van Loenen TK, Kingma-Vegter F, Westerga J, Flohil CC, Knol JC, Jimenez CR, van Gils CH, Wessels LF, Schellens JH, Beijnen JH (2009) Detection of breast cancer by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry tissue and serum protein profiling. *Int J Biol Markers* **24:** 130-141

Goncalves A, Charafe-Jauffret E, Bertucci F, Audebert S, Toiron Y, Esterni B, Monville F, Tarpin C, Jacquemier J, Houvenaeghel G, Chabannon C, Extra JM, Viens P, Borg JP, Birnbaum D (2008) Protein profiling of human breast tumor cells identifies novel biomarkers associated with molecular subtypes. *Mol Cell Proteomics* **7:** 1420-1433

Guerreiro Da Silva ID, Hu YF, Russo IH, Ao X, Salicioni AM, Yang X, Russo J (2000) S100P calcium-binding protein overexpression is associated with immortalization of human breast epithelial cells in vitro and early stages of breast cancer development in vivo. *Int J Oncol* **16:** 231-240

Haas JS, Liang SY, Hassett MJ, Shiboski S, Elkin EB, Phillips KA (2011) Gene expression profile testing for breast cancer and the use of chemotherapy, serious adverse effects, and costs of care. *Breast Cancer Res Treat* **130:** 619-626

Hahn MA, Dickson KA, Jackson S, Clarkson A, Gill AJ, Marsh DJ (2012) The tumor suppressor CDC73 interacts with the ring finger proteins RNF20 and RNF40 and is required for the maintenance of histone 2B monoubiquitination. *Hum Mol Genet* **21:** 559-568

Hammond-Martel I, Yu H, Affar EB (2011) Roles of ubiquitin signaling in transcription regulation. *Cell Signal*

Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D (2011) Global cancer statistics. *CA Cancer J Clin* **61:** 69-90

Jia SQ, Niu ZJ, Zhang LH, Zhong XY, Shi T, Du H, Zhang GG, Hu Y, Su XL, Ji JF (2009) Identification of prognosis-related proteins in advanced gastric cancer by mass spectrometry-based comparative proteomics. *J Cancer Res Clin Oncol* **135:** 403-411

Jiang H, Hu H, Tong X, Jiang Q, Zhu H, Zhang S (2011) Calcium-binding protein S100P and cancer: mechanisms and clinical relevance. *J Cancer Res Clin Oncol*

Kopans DB (2011) Just the facts: mammography saves lives with little if any radiation risk to the mature breast. *Health Phys* **101:** 578-582

Lam FF, Jankova L, Dent OF, Molloy MP, Kwun SY, Clarke C, Chapuis P, Robertson G, Beale P, Clarke S, Bokey EL, Chan C (2010) Identification of distinctive protein expression patterns in colorectal adenoma. *Proteomics Clin Appl* **4:** 60-70

Leong S, Christopherson RI, Baxter RC (2007) Profiling of apoptotic changes in human breast cancer cells using SELDI-TOF mass spectrometry. *Cell Physiol Biochem* **20:** 579-590

Logsdon CD, Simeone DM, Binkley C, Arumugam T, Greenson JK, Giordano TJ, Misek DE, Kuick R, Hanash S (2003) Molecular profiling of pancreatic adenocarcinoma and chronic pancreatitis identifies multiple genes differentially

regulated in pancreatic cancer. *Cancer Res* **63:** 2649-2657

Mackay A, Jones C, Dexter T, Silva RL, Bulmer K, Jones A, Simpson P, Harris RA, Jat PS, Neville AM, Reis LF, Lakhani SR, O'Hare MJ (2003) cDNA microarray analysis of genes associated with ERBB2 (HER2/neu) overexpression in human mammary luminal epithelial cells. *Oncogene* **22:** 2680-2688

Marenholz I, Heizmann CW, Fritz G (2004) S100 proteins in mouse and man: from evolution to function and pathology (including an update of the nomenclature). *Biochem Biophys Res Commun* **322:** 1111-1122

Nakagawa T, Huang SK, Martinez SR, Tran AN, Elashoff D, Ye X, Turner RR, Giuliano AE, Hoon DS (2006) Proteomic profiling of primary breast cancer predicts axillary lymph node metastasis. *Cancer Res* **66:** 11825-11830

Peto R, Davies C, Godwin J, Gray R, Pan HC, Clarke M, Cutter D, Darby S, McGale P, Taylor C, Wang YC, Bergh J, Di Leo A, Albain K, Swain S, Piccart M, Pritchard K (2012) Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. *Lancet* **379:** 432-444

Reis-Filho JS, Pusztai L (2011) Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet* **378:** 1812-1823

Ricolleau G, Charbonnel C, Lode L, Loussouarn D, Joalland MP, Bogumil R, Jourdain S, Minvielle S, Campone M, Deporte-Fety R, Campion L, Jezequel P (2006) Surface-enhanced laser desorption/ionization time of flight mass spectrometry protein profiling identifies ubiquitin and ferritin light chain as prognostic biomarkers in node-negative breast cancer tumors. *Proteomics* **6:** 1963-1975

Sanders ME, Dias EC, Xu BJ, Mobley JA, Billheimer D, Roder H, Grigorieva J, Dowsett M, Arteaga CL, Caprioli RM (2008) Differentiating proteomic biomarkers in breast cancer by laser capture microdissection and MALDI MS. *J Proteome Res* **7:** 1500-1507

Santamaria-Kisiel L, Rintala-Dempsey AC, Shaw GS (2006) Calcium-dependent and -independent interactions of the S100 protein family. *Biochem J* **396:** 201-214

Scarlett CJ, Smith RC, Saxby A, Nielsen A, Samra JS, Wilson SR, Baxter RC (2006) Proteomic classification of pancreatic adenocarcinoma tissue using protein chip technology. *Gastroenterology* **130:** 1670-1678

Schor AP, Carvalho FM, Kemp C, Silva ID, Russo J (2006) S100P calcium-binding protein expression is associated with high-risk proliferative lesions of the breast. *Oncol Rep* **15:** 3-6

Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL (1987) Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu

oncogene. *Science* **235:** 177-182

Srinivas PR, Kramer BS, Srivastava S (2001) Trends in biomarker research for cancer detection. *Lancet Oncol* **2:** 698-704

Terasaka S, Aita Y, Inoue A, Hayashi S, Nishigaki M, Aoyagi K, Sasaki H, Wada-Kiyama Y, Sakuma Y, Akaba S, Tanaka J, Sone H, Yonemoto J, Tanji M, Kiyama R (2004) Using a customized DNA microarray for expression profiling of the estrogen-responsive genes to evaluate estrogen activity among natural estrogens and industrial chemicals. *Environ Health Perspect* **112:** 773-781

Vucic D, Dixit VM, Wertz IE (2011) Ubiquitylation in apoptosis: a post-translational modification at the edge of life and death. *Nat Rev Mol Cell Biol* **12:** 439-452

Wang G, Platt-Higgins A, Carroll J, de Silva Rudland S, Winstanley J, Barraclough R, Rudland PS (2006) Induction of metastasis by S100P in a rat mammary model and its association with poor survival of breast cancer patients. *Cancer Res* **66:** 1199-1207

Wang N, Ma Q, Wang Y, Ma G, Zhai H (2010) CacyBP/SIP expression is involved in the clinical progression of breast cancer. *World J Surg* **34:** 2545-2552

Wen YY, Yang ZQ, Song M, Li BL, Yao XH, Chen XL, Zhao J, Lu YY, Zhu JJ, Wang EH (2010) The expression of SIAH1 is downregulated and associated with Bim and apoptosis in human breast cancer tissues and cells. *Mol Carcinog* **49:** 440-449

Whiteman HJ, Weeks ME, Dowen SE, Barry S, Timms JF, Lemoine NR, Crnogorac-Jurcevic T (2007) The role of S100P in the invasion of pancreatic cancer cells is mediated through cytoskeletal changes and regulation of cathepsin D. *Cancer Res* **67:** 8633-8642

Zhang K, Yuan K, Wu H, Li Q, Wang Y, Chen S, Zhang L, Gu H, Fu R (2012) Identification of potential markers related to neoadjuvant chemotherapy sensitivity of breast cancer by SELDI-TOF MS. *Appl Biochem Biotechnol* **166:** 753-763

Table 1. Patient Characteristics

Characteristics	Training Set	Validation Set
No. of patients	82	89
Age (median)	60	58
Histologic type		
Ductal (IDC)	68	76
Lobular (ILC)	10	10
Other	4	3
Histologic grade		
grade 1	7	11
grade 2	32	27
grade 3	43	49
Missing		2
Tumour size		
≤ 2 cm	29	28
≥ 2 cm	53	59
Missing		2
Estrogen receptor		
positive	56	64
negative	25	23
Missing	1	2

Progesterone receptor		
positive	44	54
negative	38	33
Missing	0	2
HER2 overexpression		
positive	15	16
negative	57	63
Missing	10	10
Lymphovascular invasion		
present	34	35
absent	48	54
Lymph node involvement		
positive	42	43
negative	40	44
Missing		2

IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; ER, Estrogen receptor; PR, Progesterone receptor; HER2, Human epidermal growth factor receptor 2

Table 2. Summary of data analysis

Data Analysis Stage	Training	Validation	Combination	Retesting training set	Retesting validation set
No. of patients	82	89	171	82	89
MS profile No.	164	178	342	164	178
ROC-AUC	0.87	0.91	0.92	0.86	0.91
Classification*	Sens 75.6% Spec 91.5 %	Sens 80.9% Spec 91%	Sens 77.2% Spec 88.9 %	Sens 73.2% Spec 87.8 %	Sens 80.9% Spec 90.0 %

* Sensitivity (Sens) and specificity (Spec)

Table 3. Association of two protein markers and their combination with tumour histopathologic variables.

Tumour variables	P Ubiquitin	P S100P	P Combined
Tumour size (T ≤ 2 cm, n=57 *vs.* T > 2 cm, n=112)	0.024	0.009	0.008
Grade (G1, n=18 *vs*. G3, n=92)	.026	.032	.016
LVI (present, n=69 *vs.* absent, n=102)	0.106	0.011	0.044
ER (positive, n=120 *vs.* negative, n=48)	0.059	0.004	0.016
PR (positive, n=98 *vs.* negative, n=71)	0.067	0.006	0.022
HER2 (positive, n=31 *vs.* negative, n=120)	0.033	0.002	0.009
LN* (positive, n=85 *vs.* negative, n=84)	0.315	0.027	0.121
Histologic type (IDC, n=144 *vs.* ILC, n=20)	0.607	0.765	0.708

*LN, Lymph node involvement

Figure Legends

Figure 1. Performance of two protein peaks individually and in combination. (A) The ROC area-under-curve (AUC) after cross-validation was 0.87 (Ti) for the combination of peaks at *m/z* 1842, 8599 and 9292. For the independent validation sample set, the average value of ROC-AUC was 0.91 (Vi). (B) Combination of the discovery and validation sets. The sensitivity and specificity of the combination peaks of *m/z* 8558 and 9226 were 77.2% and 88.9% with a ROC-AUC value of 0.92. (C) Retesting of initial training and validation sets. ROC-AUC values for these tests were 0.86 (Tr) and 0.91 (Vr) for training and validation sets, respectively. (D) Mean peak intensity values ± SEM (Normal *vs.* Cancer) for the marker at *m/z* 8558; (E) Mean values ± SEM for the marker at *m/z* 9226, and (F) Mean values ± SEM for the two markers combined. For the comparisons in panels D-F, n=171, $p<0.001$.

Figure 2. Immunological validation of ubiquitin and S100P. (A) For ubiquitin, four BC and corresponding AT extracts were analysed by immunoblotting, indicating relative upregulation of ubiquitin in some breast cancer patients. β-Actin is shown as a loading control. (B) Densitometric analysis of ubiquitin Western blots of 8 sample pairs. Box plot shows median, upper and lower quartiles; lines show maximum and minimum values. P=0.017, Wilcoxon signed rank test. (C) MS spectra of proteins bound to immobilised mouse anti-ubiquitin antibody. Samples were (i) patient #1 normal tissue, (ii) patient #1 cancer tissue, (iii) patient #2 normal tissue, (iv) patient #2 cancer tissue, (v) recombinant His-tagged ubiquitin, and (vi) patient #2 cancer tissue, mouse IgG control. Arrow indicates the mass of monomeric ubiquitin, *m/z* 8558. N=normal tissue,

C=cancer tissue. (D) For S100P, four BC and corresponding AT extracts were analysed by immunoblotting, indicating relative upregulation of S100P in some breast cancer patients. β-Actin is shown as a loading control. (E) Densitometric analysis of S100P Western blots of 8 sample pairs. Box plot shows median, upper and lower quartiles; lines show maximum and minimum values. P=0.012, Wilcoxon signed rank test. (F) MS spectra of proteins bound to immobilised rabbit anti-S100P antibody. Samples were (i) patient #3 normal tissue, (ii) patient #3 cancer tissue, (iii) patient #4 normal tissue, (iv) patient #4 cancer tissue, (v) recombinant His-tagged S100P, and (vi) patient #4 cancer tissue, rabbit IgG control. Arrow indicates the mass of the S100P form of *m/z* 9226. N=normal tissue, C=cancer tissue.

Figure 3. Association of the combined panel with histo-pathologic variables. Higher expression of the combined panel was significantly associated with higher histologic grade (p=0.016), higher tumour size (p=0.008), and weakly with the presence of LVI (p=0.044). The panel was also relatively increased in tumours that were positive for estrogen receptors (p=0.016), progesterone receptors (p=0.022), and HER2 overexpression (p=0.009). Box plots show median, upper and lower quartiles; lines show maximum and minimum values.

Figure 1

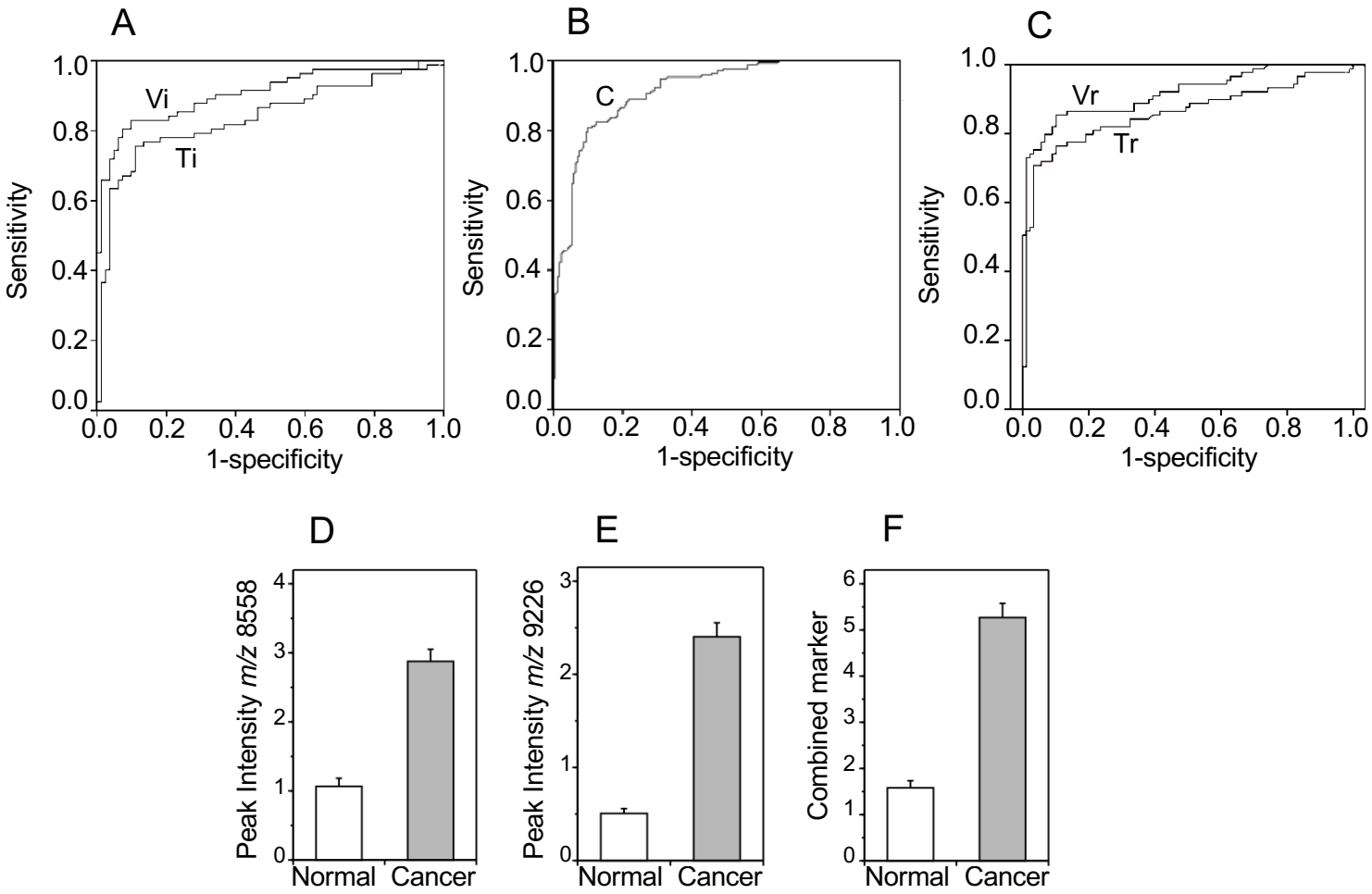

A
B
C
Vi
Ti
C
Vr
Tr
Sensitivity
1-specificity
0.0
0.2
0.4
0.6
0.8
1.0
D
E
F
Peak Intensity m/z 8558
Peak Intensity m/z 9226
Combined marker
Normal
Cancer

Figure 2

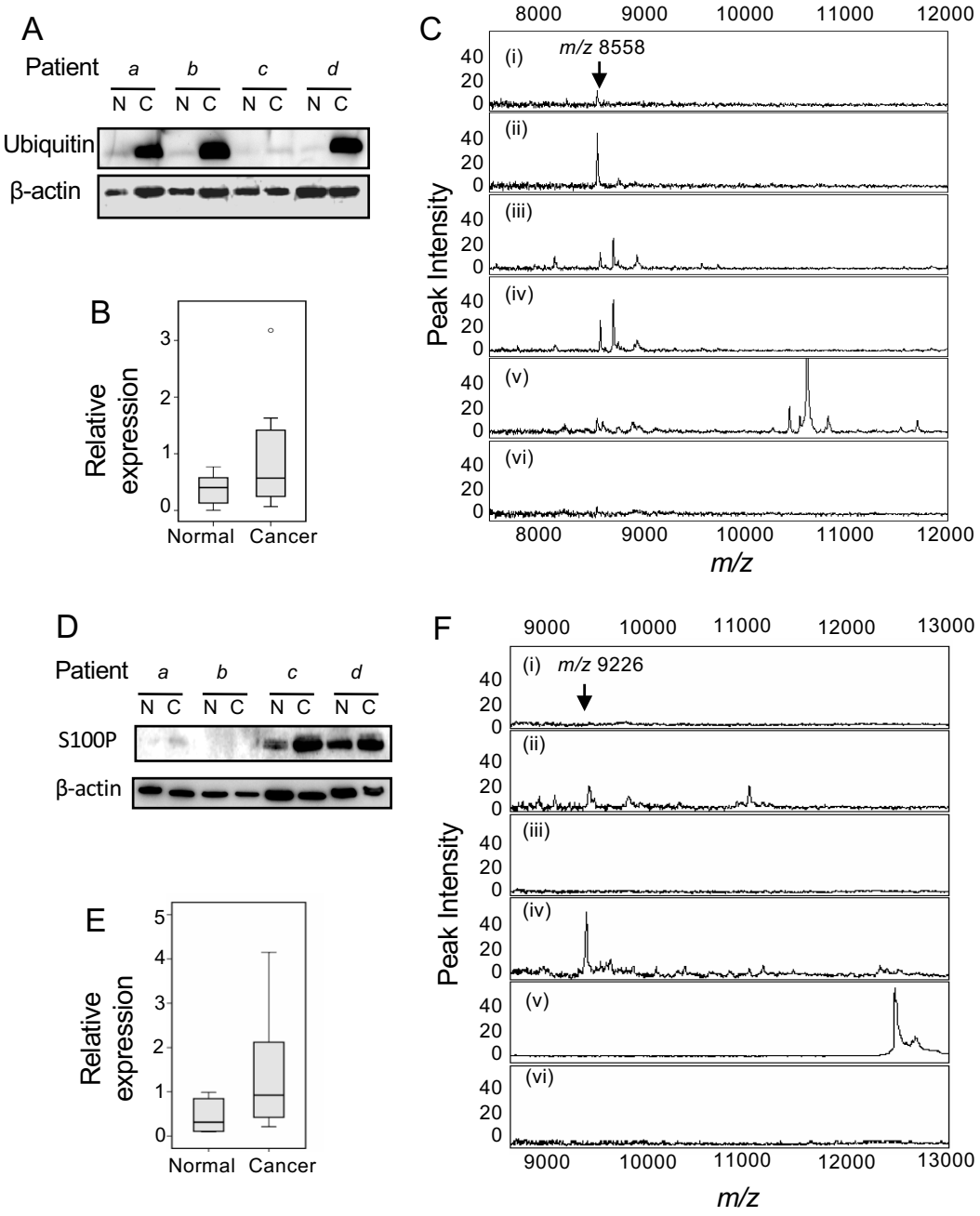
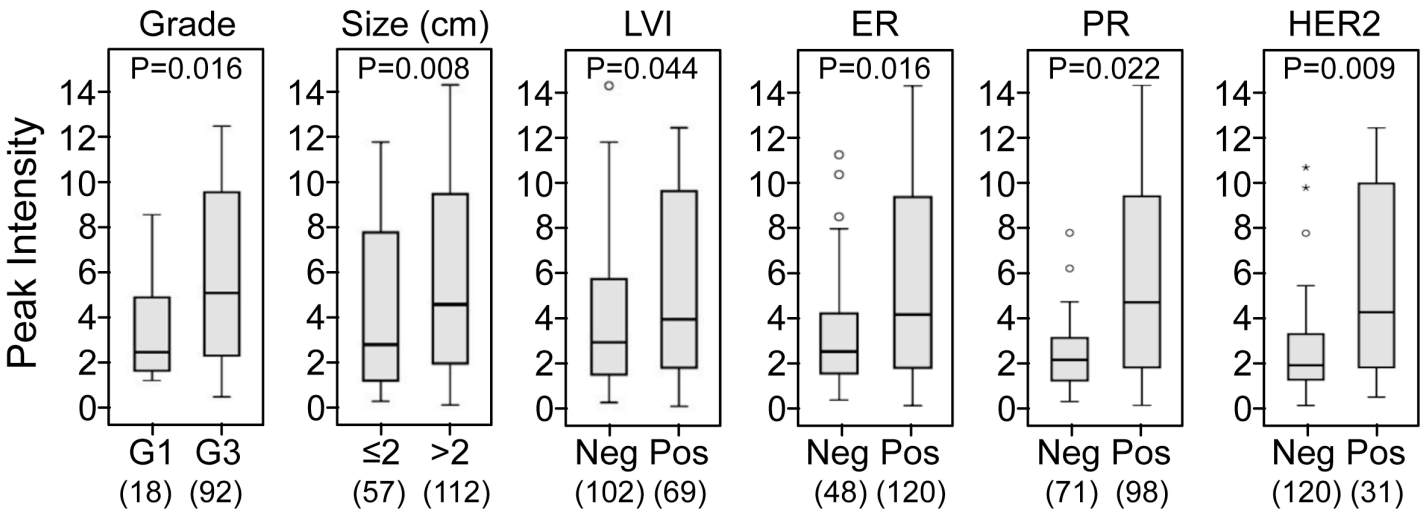

A
Patient a b c d
N C N C N C N C
Ubiquitin
β-actin
B
Relative expression
Normal Cancer
C
m/z 8558
(i)
(ii)
(iii)
(iv)
(v)
(vi)
Peak Intensity
m/z
D
Patient a b c d
N C N C N C N C
S100P
β-actin
E
Relative expression
Normal Cancer
F
m/z 9226
(i)
(ii)
(iii)
(iv)
(v)
(vi)
Peak Intensity
m/z

Figure 3


Peak Intensity
Grade
P=0.016
G1 G3
(18) (92)
Size (cm)
P=0.008
≤2 >2
(57) (112)
LVI
P=0.044
Neg Pos
(102) (69)
ER
P=0.016
Neg Pos
(48) (120)
PR
P=0.022
Neg Pos
(71) (98)
HER2
P=0.009
Neg Pos
(120) (31)

Supplementary Figure S1

P62988 (UBIQ_HUMAN) MW = 8560, **Ubiquitin**

MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG

Query	Observed	Mr(expt)	Mr(calc)	ppm	Miss	Score	Expect	Rank	Peptide
6	391.1400	780.2655	780.4204	-198.47	0	33	0.027	1	-.MQIFVK.T
138	534.2090	1066.4035	1066.6135	-196.88	0	49	0.00052	1	K.ESTLHLVLR.L
142	541.1877	1080.3609	1080.5451	-170.48	0	34	0.019	1	R.TLSDYNIQK.E
286	508.5109	1522.5109	1522.7740	-172.76	1	62	2.1e-005	1	K.IQDKEGIPPDQQR.L
361	894.3104	1786.6063	1786.9200	-175.53	0	59	4.3e-005	1	K.TITLEVEPSDTIENVK.A
409	662.9076	1985.7011	1986.0521	-176.72	1	79	3.5e-007	1	K.TITLEVEPSDTIENVKAK.I

P25815 (S100P_HUMAN) MW= 10400, **S100P**

MTELETAMGM IIDVFSRYSG SEGSTQTLTK GELKVLMEKE LPGFLQSGKD KDAVDKLLKD LDANGDAQVD FSEFIVFVAA ITSACHKYFE KAGLK

Query	Observed	Mr(expt)	Mr(calc)	ppm	Miss	Score	Expect	Rank	Unique	Peptide
199	538.2917	1074.5688	1074.5710	-1.96	0	29	1.9	1	U	K.ELPGFLQSGK.D
373	679.8246	1357.6346	1357.6361	-1.10	0	85	3.3e-006	1	U	R.YSGSEGSTQTLTK.G

Figure S1. Tryptic peptides (underlined in the full sequence) used to identify ubiquitin and S100P by MASCOT searching.

Supplementary Figure S2

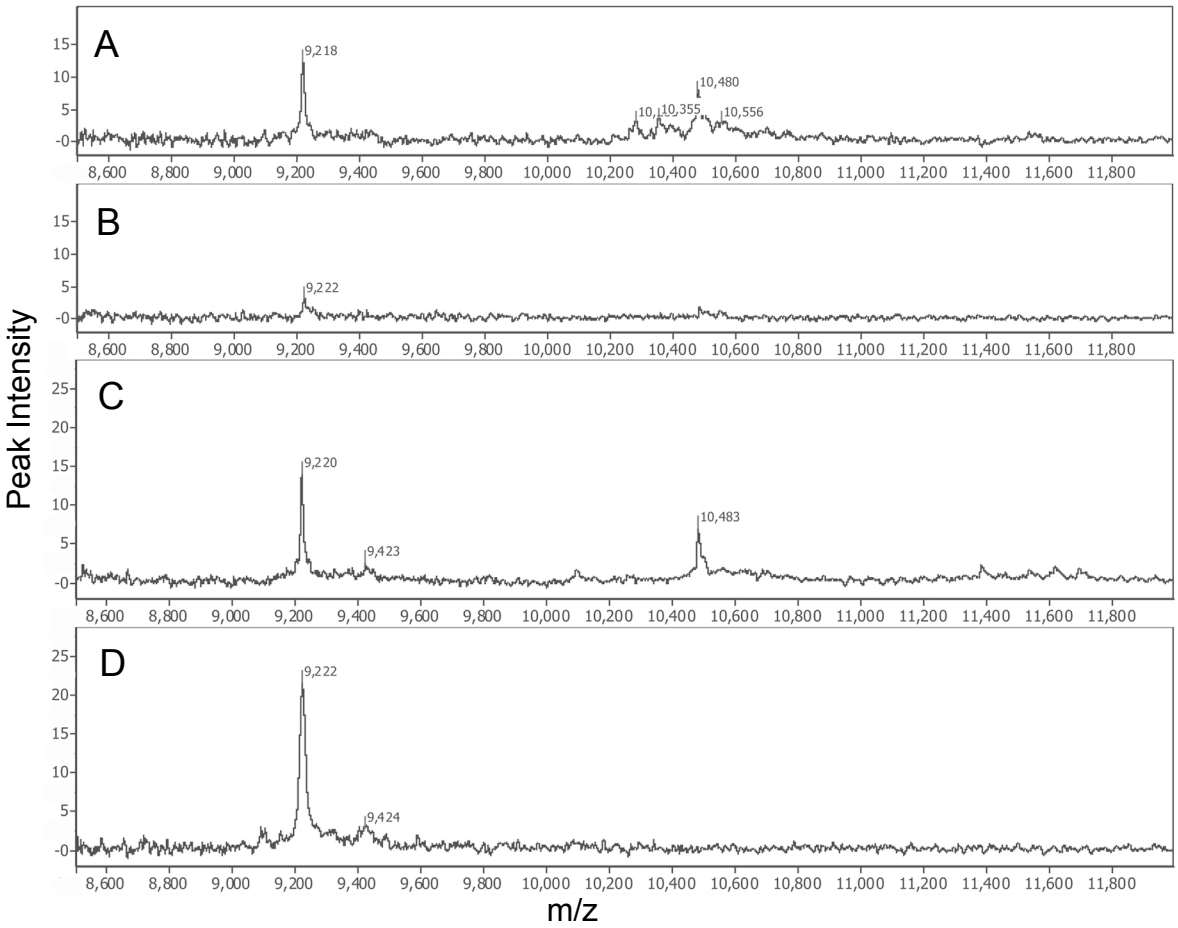


Figure S2. SELDI-TOF MS profiles on normal-phase (NP20) chips, showing protein peaks immunoprecipitated from MCF-7 breast cancer cell lysates using three different S100P antibodies. A: Rabbit monoclonal (Epitomics); B: Mouse polyclonal (Abnova); C: Rabbit polyclonal (Invitrogen). D: Immunoprecipitate from rabbit polyclonal antibody (Invitrogen) after further purification by reverse-phase HPLC.

Methods

Immunoprecipitation. MCF-7 cells were lysed in buffer containing 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, 1 mM EGTA, 1% Triton X-100. To isolate S100P, 25 µl (0.75 mg) of Protein G Dynabeads (Invitrogen) were incubated with S100P antibody (rabbit monoclonal #5263-1, 5.45 µg, Epitomics; MaxPab mouse polyclonal #H0006286-B01P, 5 µg, Abnova; or rabbit polyclonal #18-0046, 4.68 µg, Invitrogen) for 30 min at room temperature with rotation. Following two washes with PBS pH 7.4 containing 0.02% Tween 20, the bound antibody-bead complex was added to 50 µl of cell lysate, mixed by vortexing, and incubated for 2 h at room temperature with rotation. The beads were washed three times with PBS pH 7.4 and bound proteins eluted with 100 mM glycine, pH 2.8. Fractions were monitored by SELDI-TOF MS on normal-phase NP20 protein chips (Bio-Rad).

Reverse-phase HPLC. Immunprecipitated proteins were applied to a 4.6 x 250 mm C18 column (Jupiter, 5 µm, 300 Å) and eluted at 1.5 ml/min with a 30-min gradient from 15% to 60% acetonitrile in 0.1% trifluoroacetic acid.

Supplementary Table S1. Association of two protein markers and their combination with tumor histopathologic variables.

Tumor variables	Lymph Node Negative (n=84)				Lymph Node Positive (n=85)			
	n	P Ubiquitin	P S100P	P Combined	n	P Ubiquitin	P S100P	P Combined
Size ≤2 cm *vs.* >2cm	43 *vs.* 41	0.059	0.013	0.025	13 *vs.* 68	0.057	0.066	0.015
Grade G1 *vs.* G3	10 *vs.* 45	0.775	0.665	0.692	8 *vs.* 47	0.875	0.984	0.338
LVI Present *vs.* Absent	15 *vs.* 69	0.312	0.249	0.404	68 *vs.* 17	0.005	0.001	0.002
ER Positive *vs.* Negative	62 *vs.* 21	0.757	0.173	0.383	58 *vs.* 23	0.011	0.026	0.004
PR Positive *vs.* Negative	54 *vs.* 28	0.284	0.053	0.112	44 *vs.* 23	0.001	<0.001	<0.001
HER2 Positive *vs.* Negative	10 *vs.* 34	0.248	0.054	0.139	22 *vs.* 46	0.064	0.056	0.041