Identification of a Biomarker Panel for Early Detection of Lung Cancer Patients.

Bethany Geary[1,2], Michael J. Walker[2], Joseph T. Snow[2,3], David C. H. Lee[1], Maria Pernemalm[4], Saeedeh Maleki-Dizaji[2], Narges Azadbakht[2], Sophia Apostolidou[5], Julie Barnes[6], Piotr Krysiak[7], Rajesh Shah[7], Richard Booton[8], Caroline Dive[9,10], Philip A Crosbie[1,5,8], Anthony D. Whetton[1,3,10,*].

[1] Stoller Biomarker Discovery Centre, Institute of Cancer Sciences, Faculty of Medical and Human Sciences, University of Manchester, Manchester, United Kingdom
[2] Stem cell and leukaemia proteomics laboratory, Institute of Cancer Sciences, Faculty of Medical and Human Sciences, University of Manchester, Manchester, United Kingdom
[3] Department of Earth Sciences, University of Oxford, Oxford, United Kingdom.
[4] Science for Life Laboratory, Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden.
[5] Gynaecological Cancer Research Centre, Department of Women's Cancer, Institute for Women's Health, University College London, London, United Kingdom
[6] Abcodia, Cambourne, Cambridgeshire, United Kingdom
[7] Department of Thoracic Surgery, Wythenshawe Hospital, Manchester University NHS Foundation Trust, United Kingdom
[8] North West Lung Centre, Wythenshawe Hospital, Manchester University NHS Foundation Trust, United Kingdom
[9] Clinical and Experimental Pharmacology Group, Cancer Research UK Manchester Institute, University of Manchester, Manchester, United Kingdom
[10] Cancer Research UK Lung Cancer Centre of Excellence

[*] Corresponding author: Anthony D. Whetton tony.whetton@manchester.ac.uk +44 (0)161 2756267

Abstract

Lung cancer is the most common cause of cancer related mortality worldwide, characterised by late clinical presentation (49-53% of patients are diagnosed at stage IV) and consequently poor outcomes. One challenge in identifying biomarkers of early disease is the collection of samples from patients prior to symptomatic presentation. We used blood collected during surgical resection of lung tumours in an iTRAQ isobaric tagging experiment to identify proteins effluxing from tumours into pulmonary veins. 40 proteins were identified as having an increased abundance in the vein draining from the tumour compared to “healthy” pulmonary veins. These protein markers were then assessed in a second cohort that utilised the mass spectrometry (MS) technique: Sequential window acquisition of all theoretical fragment ion spectra (SWATH) MS. SWATH-MS was used to measure proteins in serum samples taken from 25 patients <50 months prior to and at lung cancer diagnosis and 25 matched controls. The SWATH-MS analysis alone produced an 11 protein marker panel. A machine learning classification model was generated that could discriminate patient samples from patients within 12 months of lung cancer diagnosis and control samples. The model was evaluated as having a mean AUC of 0.89, with an accuracy of 0.89. This panel was combined with the SWATH-MS data from one of the markers from the first cohort to create a 12 protein panel. The proteome signature developed for lung cancer risk can now be developed on further cohorts.

Keywords:

Biomarker, Lung Cancer, SRM, SWATH, Proteomics, Early Detection

Introduction

In the UK, lung cancer is responsible for one in five of all cancer deaths and has one of the poorest prognoses of all malignant diseases with 75% of patients dying within the first year and less than 10% surviving for five years. Such a prognosis is attributed to the late clinical presentation of the majority of patients; 72-76% are diagnosed at a late stage (stage III or IV)[1–3]. The National Lung Screening Trial (NLST) showed for the first time a significant reduction (20%) in lung cancer specific mortality by screening asymptomatic at risk patients with low dose computed tomography (LDCT)[4,5]. The UK lung cancer screening trial (UKLS), utilising advances in nodule management, achieved an 85% early stage detection rate (stage I/II) with curative action possible for 90% of those cases with a false positive rate of 3.6%. The development of a biomarker that could identify disease would be a valuable adjunct for early detection strategies. The importance of early detection is underlined by the strong association between lung cancer stage and survival. Patients who present with advanced lung cancer (stage IV) have a life expectancy that is usually only measured in months[6], whereas early stage disease is eminently curable. However, even in patients with stage I disease who have had curative intent surgery survival is correlated with tumour size, with 5-year survival post lobectomy ranging from 92% for 21mm tumours to 81% for 30mm tumours[6].

Blood sampling is both minimally invasive and readily available to most clinicians, qualities that make serum a good matrix for a clinical screening assay, it is however a challenging matrix for proteomic analysis. Proteins in serum have a high dynamic range with the tissue derived proteins, which are potentially biomarkers of value, present at a low abundance. Previous studies have used tissue for the discovery of potential markers as a way of circumnavigating this challenge. However, the elevation of a protein in the tumour/neighbouring tissue doesn't necessitate that it will be altered in the serum. This can lead to a high degree of putative biomarker failure in proposed clinical decision usage. Circulating tumour cell (CTC) number is an independent prognostic biomarker for overall survival in both non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC)[7,8] but as an early diagnosis tool, CTC research is in its infancy with newer technologies emerging capable of rare cell analysis where single CTC candidates can be sequenced to confirm tumour origin[9]. Circulating tumour DNA also has potential as an early detection biomarker for cancer when specificity and sensitivity for identified lung cancer mutations are applied but here false positives may also be a challenge in ageing populations[10–12]. As stated, an additional challenge of identifying biomarkers for early detection of cancer is the prospective

collection of high quality samples from patients prior to symptomatic presentation. Lung cancer in particular has common co-morbidities such as Chronic Obstructive Pulmonary Disease (COPD) or infection. These co-morbidities cause protein expression changes which can complicate proteomic analysis.

We took two approaches to discover biomarkers for early detection of lung cancer. In the lung resection serum study, we utilised a serum sample set taken just prior to lung tumour resection to look for candidate biomarkers in blood using a previously developed global discovery proteomic workflow[13,14]. Blood was collected from the pulmonary vein draining the tumour bearing lobe (C), from the pulmonary artery (A) and pulmonary vein from the non-cancerous lobe (N). Blood was taken from the tumour draining pulmonary vein as this was where the highest concentration of tumour related biomarkers would be. Blood was taken from the non-cancer draining pulmonary vein (N) as the blood had been through a non-cancer lung and could potentially act as a control for any benign lung disease without the cancer. Blood was taken from the pulmonary artery (A) as this was the point in the circulatory system furthest away from the tumour draining pulmonary vein and should therefore have the lowest level of tumour related biomarkers.

In the prospectively collected serum study, the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) was an ovarian cancer screening trial of post-menopausal women that recruited 206,638 volunteers. Blood samples were collected during the trial over the course of 10 years. This sample resource offered us an unparalleled opportunity to explore blood borne markers prior to the development of disease in a longitudinal study design. Although set up to discover early detection biomarkers for Ovarian Cancer, 142 women in the study developed lung cancer providing a precious resource to test our hypothesis that subjects who develop lung cancer will have a different protein signature in serum compared to controls prior to the development of their clinically apparent disease. In our research on ovarian cancer, we saw modulation of 90 potential biomarkers for the early detection of ovarian cancer using mass spectrometry[15,16]. Here we have used this strength of mass spectrometry in multiplex assays to rapidly assess potential biomarker profiles linked to early lung cancer in peripheral serum samples.

Methods

Isobaric tagging for relative and absolute quantitation (iTRAQ) analysis

Blood samples were obtained during surgical resection of squamous cell carcinoma of the lung. Serum was prepared from 6 patients from three sites: from the pulmonary artery (A) a pulmonary vein draining from a tumour (C), and a clean pulmonary vein (non-cancer draining) (N). The three samples from distinct sites of collection from one person were then compared using an isobaric tagging based proteomic workflow. A small volume of serum from each sample was taken and pooled to create a control serum sample. Samples were depleted/buffer exchanged using a MARS-14 column (Agilent Technologies, Cheadle, UK). After digestion, all of the samples and two pooled controls were then separated into 3 sets for 8-plex iTRAQ labelling (Figure 1). Labelling was performed according to the manufacturer's instructions (AB Sciex, Warrington). Labelled samples were fractionated and analysed by mass spectrometry as described previously[14]. Pooled control samples were created from all of the plasma samples from the 6 patients and then used as technical replicates in each 8 channel experiment. The duplicates of the pooled control sample in each iTRAQ set were used to monitor the variability and define a statistically significant change [17]. The pooled control samples used the iTRAQ channels 119 and 121 in all 3 8-plex experiments. Protein $\log_2$ fold changes were analysed and corrected for with regards to technical, within-person and inter-person variation as described previously[13,17].

Mass spectrometry data files were searched using ProteinPilot (version 5). Peptide matches were searched with the paragon engine in thorough ID mode. A maximum of 2 missed cleavages were allowed. Carbamidomethylation of cysteine residues was a fixed modification and methionine oxidation was a variable modification. Searches were made against the Uniprot SwissProt/trEMBL human database (downloaded on 04/07/2018). A 1% FDR cut-off was applied to the searched data. Statistical tests, plots and downstream analysis was performed using the R language (version 3.4.1). Venn diagrams were generated using the eulerr R package (version 4.1.0).

Sample collection for SWATH analysis

For the SWATH analysis, 150 samples from 50 patients over 3 time points were acquired from the UK Collaborative Trial on Ovarian Cancer Screening (UKCTOCS; ISRCTN Number 22488978) trial blood

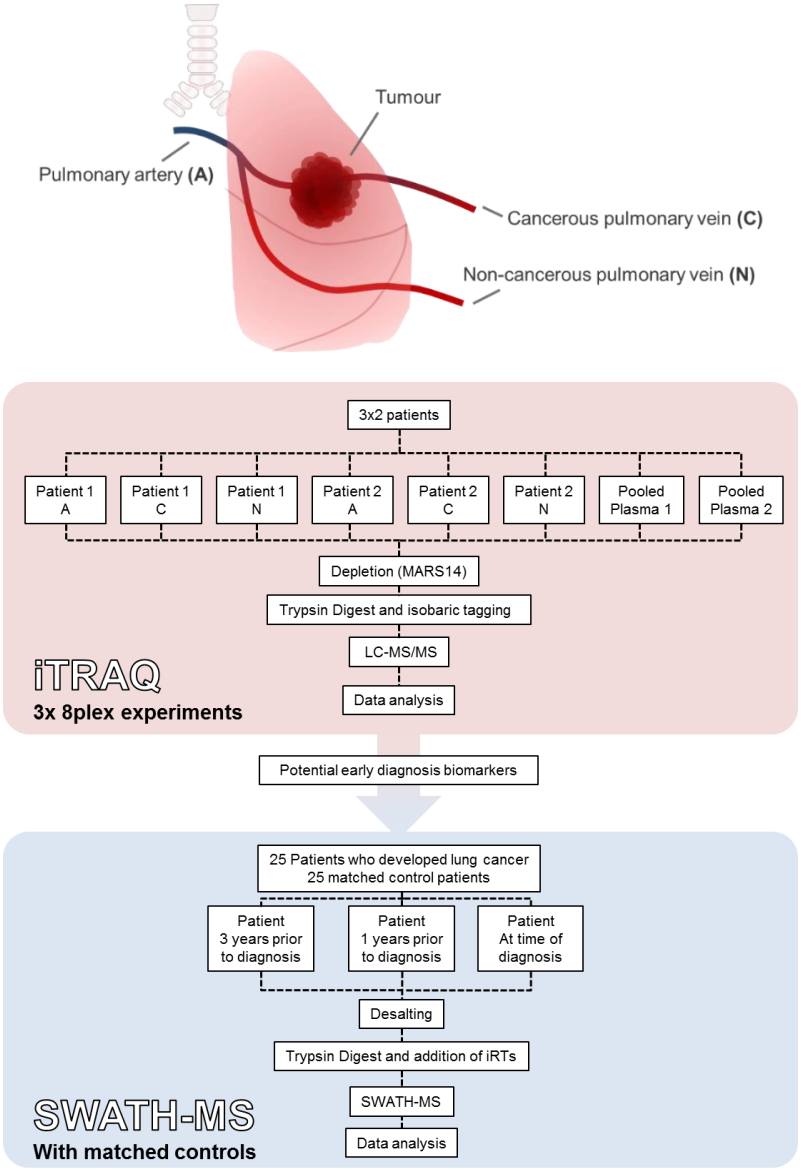


Figure 1. Schematic representation of the overall study design. In the lung resection serum study, blood samples were taken from 6 patients during surgical resection of early stage squamous cell carcinoma. The samples were taken from three different locations; one from the pulmonary artery (A) and two from different pulmonary veins (C and N). Samples for the prospectively collected serum study were taken from 25 patients that were part of the UKCTOCS trial. Samples were collected at three different time points prior to lung cancer diagnosis (0-12 months, 12-24 months and 35-60 months). Samples from case patients were collected in the years preceding a lung cancer diagnosis, controls were taken over three years with no cancer diagnosis for 3 years after the last sample collection.

bank. The UKCTOCS trial design has been detailed previously [18]. Briefly, UKCTOCS is a 13 centre randomised controlled trial investigating the impact of ovarian cancer screening on disease mortality. During the trial 202,638 post-menopausal women aged 50-74 at recruitment (2001-2005) were randomly assigned (2:1:1 ratio) to routine care (control; n=101,359) or annual screening using serum cancer antigen 125 (CA125) (multimodal screening, n=50 640) or transvaginal ultrasound (n=50,639). Participants had no active malignancy at recruitment. All participants were linked using their National Health Service number to national cancer and death registry electronic health records as well as Hospital Episode Statistics (those resident in England) and the Myocardial Ischaemia National Audit Project (MINAP). In addition, women were sent two follow-up questionnaires, the most recent in 2014. All women provided a blood sample at recruitment with women randomised to the multimodal screening group (n=50,640) continuing to donate serum annually for up to 11 years from randomisation. Sample collection stopped at the end of screening in December 2011. The SWATH analysis study follows the PRoBE (Prospective specimen collection with retrospective-blinded-evaluation) design wherein biological specimens and clinical data are collected prospectively from a cohort that represents the target population, outcome status is ascertained on follow up and nested case control studies are undertaken within the cohort with blinded assay of specimens [19].

Case samples were obtained from patients that were diagnosed with lung cancer during the trial or age matched controls. Controls were sourced from the trial and were matched to the lung cancer patients with the following matching criteria:

- Age at sample taken will be within 5 years
- Lifetime smoking status
- Regional centre for collection
- The time to spin to be within 4 hours of that of the matched case
- The time interval between longitudinal samples to be within 1 year of that indicated in the case
- Sample collection date to be within ± 2 years, if possible

Of the patients 25 developed lung cancer and 25 were matched controls also from the UKCTOCS trial. Three patients from the lung cancer group were COPD positive at the time of sampling with 4 patients developing COPD after sampling. 21 of the case volunteers had ever smoker status with the remaining 4 never smokers. The types of lung cancer was most commonly adenocarcinoma with 5 other groups of cancer type diagnosed, the stages of lung cancer at diagnosis were between stage I

and stage IV, with the majority of cases being late stage of either stage III or stage IV (Tables 1 & 2).

UKCTOCS was approved by the UK North West Multicentre Research Ethics Committee (North West MREC 00/8/34) with site specific approval from the local regional ethics committees and the Caldicott guardians (data controllers) of the primary care trusts. All women gave written consent for use of samples and data in ethically approved secondary studies. The subset of samples used for the present study has been approved by the Yorkshire & The Humber - Sheffield Research Ethics Committee (REC Ref 15/YH/0044).

The following sample preparation steps were performed within 96-well format plates. Crude serum samples were mixed thoroughly before 10µL was taken and diluted to 50uL with 50mM Ammonium Bicarbonate. Diluted serum was filtered and desalted using 7K Zeba™ Spin desalting plates (ThermoFisher Scientific, Hemel Hempstead).

Protein digestion and peptide isolation

Protein abundance was assayed using a Bradford assay (Bio-Rad, Watford, UK), then 40µg of protein was reduced with 60mM tris(2,0-carboxyethyl) phosphine (TCEP) at 60^0C for 60 minutes. Alkylation was performed using 10mM iodoacetamide over 30 minutes (room temperature, dark) and digestion completed with 20:1 trypsin (Promega, Southampton, UK) at 37^0C overnight. Digested peptide was purified using Waters (Wilmslow UK) SepPak C18 columns.

Sequential window acquisition of all theoretical fragment ion spectra (SWATH) analysis

SWATH-MS analysis was performed on a 6600 TripleTOF mass spectrometer (Sciex, Warrington, UK) coupled to a Dionex Ultimate 3000 HPLC (Dionex, Thermo, UK). Peptides were separated through an Acclaim PepMap 100 C18 column. Buffer A comprised of 2% Acetonitrile, 0.1% Formic acid, 98% Water. Buffer B comprised of 80% Acetonitrile, 0.1 % Formic acid and 20% Water. Citric acid (20mM) with 0.1% v/v ACN and 0.1% v/v FA was used as sample buffer. Peptides were loaded at 5 µl/min for 10 min prior to being eluted over a 120-minute gradient at 0.3 µl/min.

Table 1. Summary of cancer types within the cases

Cancer type	Frequency
Adenocarcinoma	15
Carcinoid tumour	1
Large cell carcinoma	1
Mucinous adenocarcinoma	1
Non-small cell lung carcinoma	2
Squamous cell carcinoma	5

Table 2. Summary of cancer stage at diagnosis within the cases

Stage at diagnosis	Frequency
Stage I	2
Early stage I/II	2
Stage II	5
Stage III	2
Stage III/IV	2
Stage IV	10
Unable to stage	2

Mass spectrometry data files were searched using openSWATH (version 2.0.0). Peptide matches were scored using pyProphet (version 0.18.3) and the search results aligned using the feature alignment script from MSproteomicstools. Statistical tests and downstream analysis was performed using the R language (version 3.4.1). Throughout the analysis the random seed was set to 500. Resultant data was transformed and normalised using the Bioconductor (release 3.5) packages SWATH2Stats and MSstats. SWATH-MS runs with a transition level FDR of greater than 0.3 were excluded from the analysis. Mann-Whitney tests were performed to determine statistical significance between conditions.

Pathway and ontology analysis

Pathway and enrichment analysis and statistics, including significance testing and multiple test correction, was performed using IPA (Ingenuity® Systems). Enrichment analysis for the iTRAQ analysis used a custom background comprised of all the proteins observed in the iTRAQ experiment. Similarly, the enrichment analysis for the SWATH dataset used a custom background of all proteins observed in the SWATH analysis. Changes with a *p* value of <0.05 were considered significant.

Results

Protein levels obtained from blood vessels just prior to surgical resection

We have previously established and verified an isobaric tagging proteomic workflow with low technical variation which allows discovery experiments to be appropriately powered with six patients or more[14]. Using this workflow samples have been analysed from serum resulted in more than 1000 proteins being relatively quantified. Proteins identified cover more than 6 orders of magnitude of known concentration allowing observation of proteins like carcinoembryonic antigen (CEA, a tumour marker for multiple cancers). To identify candidate biomarkers for early detection of lung cancer, in the lung resection serum study we initially collected serum from six lung cancer patients taken just prior to surgical resection of early stage squamous cell carcinoma (Stage I-IIIA). The protein levels were compared using three isobaric tagging experiments with peptides fractionated by two-dimensional liquid chromatography. This resulted in 1,044 proteins being identified at < 1% false discovery rate (FDR) at protein level.

Statistically reliable quantitative information from isobaric labelling was available for more than 800 proteins identified in the three different serum samples across the six patients, including proteins classified as tissue leakage proteins like CA-125, KRT5, KIT and MET[20–22]. The three different blood

sources were compared to pooled control samples enabling the analysis of proteomic differences between the test samples (Figure 2). iTRAQ ratio values where the p value was above the threshold were filtered out. Significance (p <0.05) is the p value determined by ProteinPilot for each iTRAQ ratio pair. 6 proteins had a higher mean abundance in the serum draining from the cancerous lobe compared to the pool (Table 3). The relative levels of proteins were compared between the pulmonary veins draining the cancerous (C) or non-cancerous (N) lobes (Figure 3). When compared to the protein abundance levels in the pulmonary artery (A), 40 proteins were observed to have a $\log_2$ fold change greater than 1.2 in the cancer draining vein (C/A) while additionally showing a lower fold change in the non-cancer draining vein (N/A) (Table 4).

The coagulation system has strong links to cancer progression through tumour angiogenesis, and it is 'activated' in most cancer patients[23]. 3 proteins of 40 seen as changing in the (C/A) vs (N/A) samples were part of the complement and coagulation cascades. The 52 remaining proteins included members of the proteolysis or the immune response processes. 10 of the proteins had associations with lung disease and cancer. Analysis through web-GESTALT identified DAPK1, APOA1 and CA2 as all having associations with Adenocarcinoma, whilst KIT, MPO and SCGB1A1 as associated with lung disease and HNRNPA1, KRT5, MDH2, TKT and MRC2 as linked with breast cancer.

One of the differentially expressed proteins, myeloperoxidase (MPO) is a marker for non-malignant inflammatory lung disease and was used here to verify the large scale discovery proteomics as a reliable immunoassay was available (Figure 4). MPO is the most abundant protein in neutrophils, which are recruited to the lung in response to stimuli such as cigarette smoke[24]. This recruitment promotes the local release of MPO which converts several pro-carcinogens to their active forms[25]. The ELISA results agreed with the proteomics showing a significant increase in the levels of MPO (p < 0.05). Further validation was performed with SWATH mass spectrometry (see below).

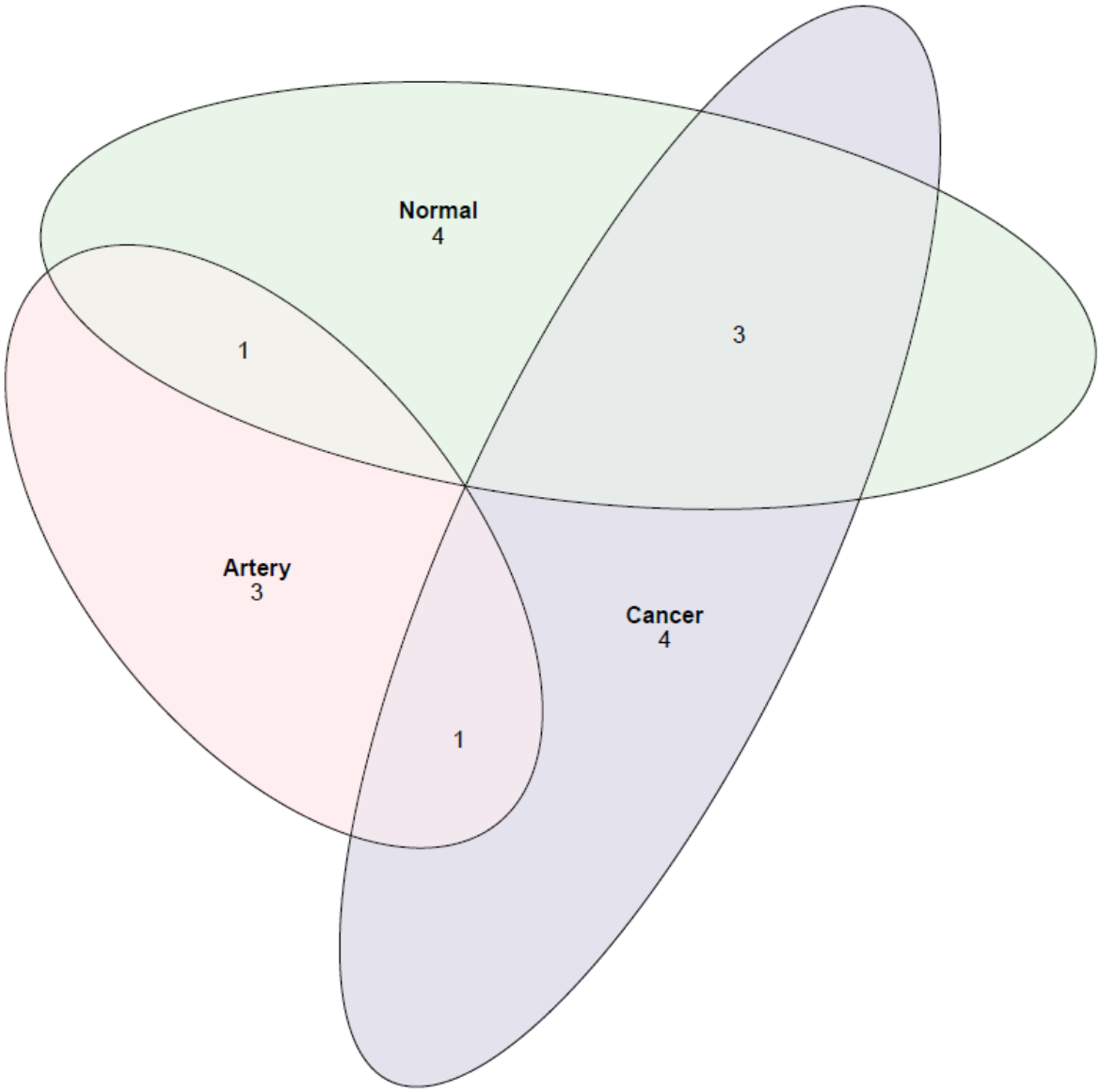


Figure 2. Venn diagram displaying the overlap in proteins that were found to be upregulated in the different serum sources. Two iTRAQ labels were reserved for pooled samples. This could allow for a comparison between the different patient specific bio-fluids, 6 proteins were upregulated in the cancer draining vein (C) compared to the pool. 3 of these proteins were only seen as upregulated in the cancer draining vein (C) compared to the other sources.

Table 3. Mean log_2 iTRAQ ratios of proteins showing a difference greater than the 95% confidence interval in the majority of patients between the cancer draining vein (C) and the control pooled sample from all patients.

Uniprot entry name	Protein names	Log_2 iTRAQ ratio
AL1A1_HUMAN	Retinal dehydrogenase 1	1.000356
APOC4_HUMAN	Apolipoprotein C-IV	0.935885
CO3_HUMAN	Complement C3	0.247728
FA8_HUMAN	Coagulation factor VIII	0.848364
IGHM_HUMAN	Immunoglobulin heavy constant mu	0.929923
K1C14_HUMAN	Keratin, type I cytoskeletal 14	0.991677
K2C1_HUMAN	Keratin, type II cytoskeletal 1	0.843437
K2C6C_HUMAN	Keratin, type II cytoskeletal 6C	1.360037

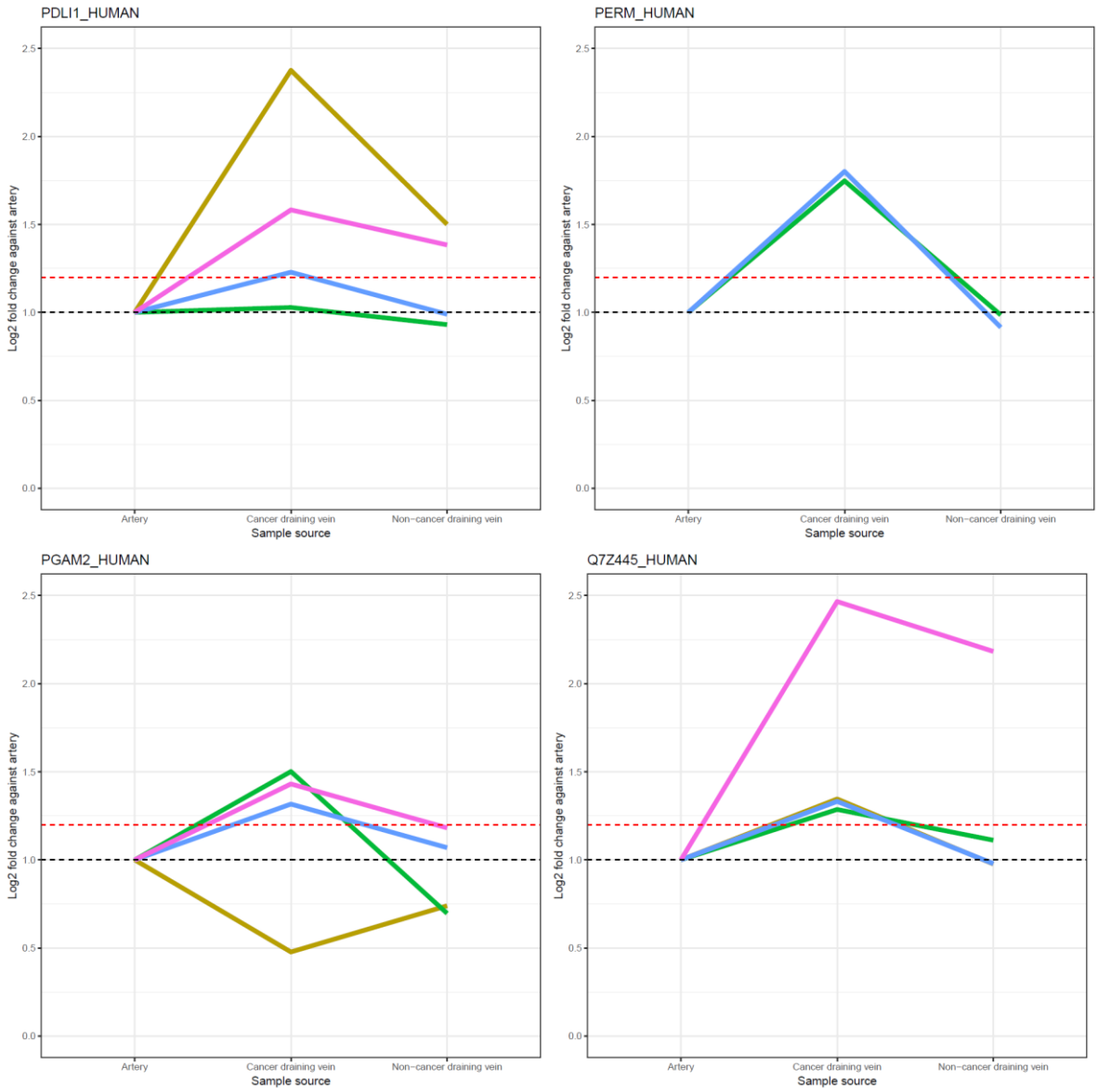


Figure 3. Isobaric tagging ratio values for four proteins showing signs of elevation in tumour draining serum samples relative to arterial abundance levels. Across the three isobaric tagging experiments more than 800 proteins were quantified from the three sources of serum. Fold changes between the cancer draining vein and artery (C/A) and between the non-cancer draining vein and artery (N/A). 40 proteins that had a fold change greater than 1.2 in the cancer draining vein (C/A) and a lower fold change in the non-cancer draining vein (N/A) were identified. Each solid line represents an individual patient. The red dashed line indicates the 1.2 fold change mark. Each isobaric tagging experiment contained the complete set of samples for two patients. Some proteins were not identified in all of the experiments, and as a result proteins were only quantified in either 2, 4 or 6 of the patients.

Table 4. Mean fold change of the 40 proteins that showed a greater than 1.2 mean log_2 fold change between the cancerous vein (C) and the pulmonary artery (A) in the majority of patients.

Uniprot entry name	Cancerous vein / Artery (C/A)	Non-cancerous vein / Artery (N/A)
ADH1G_HUMAN	1.25	1.07
APOA1_HUMAN	2.18	1.95
CAH2_HUMAN	1.54	0.95
CREL1_HUMAN	1.21	1.06
CRIS2_HUMAN	1.31	1.15
D6RHJ6_HUMAN	1.62	1.33
DAPK1_HUMAN	1.34	0.84
E9PFB0_HUMAN	1.24	0.83
FHR3_HUMAN	1.61	1.46
FHR4_HUMAN	3.61	2.45
HBA_HUMAN	1.41	0.79
IMP3_HUMAN	5.88	1.32
K1C14_HUMAN	1.39	0.82
K2C5_HUMAN	2.04	0.99
K2C6C_HUMAN	2.26	0.61
KIT_HUMAN	1.28	0.91
LIRA2_HUMAN	1.64	1.35
MDHM_HUMAN	1.35	1.03
MRC2_HUMAN	1.59	1.17
NECT1_HUMAN	1.64	1.26
O60420_HUMAN	1.21	0.88
PDLI1_HUMAN	1.55	1.20
PERM_HUMAN	1.77	0.95
PGAM2_HUMAN	1.42	0.98

PLF4_HUMAN	1.38	0.91
PNPH_HUMAN	1.27	0.82
Q53R15_HUMAN	1.38	1.22
Q5T4G3_HUMAN	2.46	1.46
Q6FHV6_HUMAN	2.38	0.80
Q7Z445_HUMAN	1.61	1.31
Q86TT2_HUMAN	1.26	1.05
Q9UMV1_HUMAN	1.56	1.43
ROA1_HUMAN	1.74	1.26
S6OS1_HUMAN	2.99	2.48
SBSN_HUMAN	1.26	1.18
SMG7_HUMAN	1.58	0.91
TKT_HUMAN	1.20	0.97
UB2L3_HUMAN	1.67	1.24
UTER_HUMAN	1.30	1.11
ZN639_HUMAN	1.55	0.81

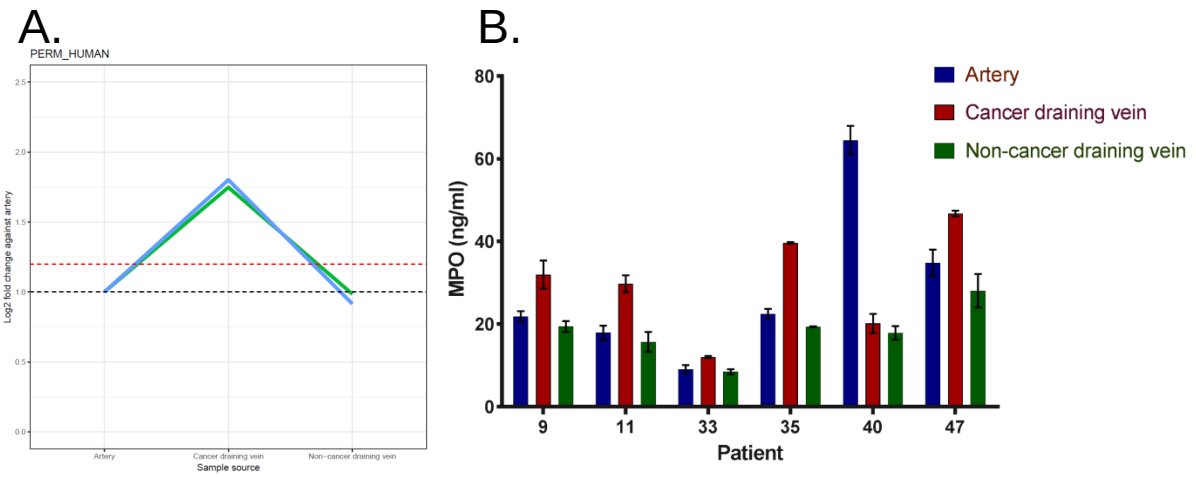


Figure 4. Myeloperoxidase abundance levels. (A) iTRAQ data for Myeloperoxidase from each sample source from patients 35 and 47. **(B)** ELISA results of myeloperoxidase levels in different blood samples taken from the different vessels during surgical resection of the lung showed a significant increase in myeloperoxidase in the blood vessel draining from a tumour containing lobe. Each sample was measured in triplicate technical replicates. Error bars depict the standard error of the mean.

Network functions of proteins draining from cancerous lobes

Direct comparisons between the sample sources allowed for in depth gene ontology and pathway analysis to be conducted. 50 proteins showed significant (p = <0.05) change between the cancer draining vein (C) and the non-cancer draining vein (N) (Supplemental table S1). The top associated network functions with these 50 proteins were RNA damage and repair, inflammatory response and cell cycle processes. Cancer, cell death and survival network functions also showed enrichment. Cancer had a significant enrichment (p = 0.0489) in the disease and disorder based analysis.

There were 29 proteins that had an increased abundance in the cancer draining vein (C) compared to the pulmonary artery (A) (Supplemental table S2). Pathway analysis revealed similar results to the (C) vs (N) comparison with significant (p = 0.0483) enrichment of “cancer related proteins” and “cell death and survival and cancer related” network functions also showing high levels of enrichment. Of itself this data on proteins egressing from lung tumours prior to resection gives information on the tumour and its environment (see Discussion).

SWATH-MS quantitative capability for validation

SWATH-MS as a Data Independent Acquisition (DIA) method requires reference libraries for protein identification and relative quantification across a number of samples. The spectral reference library can be manipulated to suit the specific analytes to be examined. Therefore we constructed a reference library consisting of a plasma protein library [26] enhanced via use of transition information of peptides from changing proteins in the isobaric tagging experiment. In this way, we cover as many proteins as possible without elevating the false discovery rate [27].

Prospective specimen collection with retrospective blinded evaluation and markers of risk of lung detection: comparison to venous blood from lung tumours.

In the prospectively collected serum study, SWATH-MS was performed on samples from 50 volunteers enrolled in the UKCTOCS trial, where longitudinal serum sample collection was performed (with 3 samples per patient). Data was analysed to identify early markers of lung cancer. This allowed the investigation of the relationship between the serum samples from both studies.

SWATH-MS analysis identified 249 proteins using an m-score cut-off of $1_{x10^{-11}}$ allowing for a calculated protein FDR of 5.4% from a total of 4,285 unique proteotypic peptides. Protein intensities were

normalised with MSstats using the ‘equalise medians’ method. Features with less than 30% coverage across all the samples were then removed. The difference in protein intensities between samples from patients within 1 year of lung cancer diagnosis and samples from control patients was assessed using a Mann-Whitney U test. This identified 65 proteins (Table 5) as significantly differentially expressed ($p \leq 0.05$).

Within the prospectively collected serum study, through sequential testing of different combinations of these 65 proteins, an optimum set of proteins was determined that maximised the separation in hierarchical clustering and principal component analysis (PCA) distance between the ‘within 1 year of cancer diagnosis group’ and the ‘>1 year prior to diagnosis and age matched no lung cancer control groups’. This resulted in a set of proteins (Marked by a ✓ in Table 5) that were analysed via PCA and hierarchical clustering analysis (Figures 5 and 6). This set of proteins included an apolipoprotein, immune response proteins and a coagulation related protein (Alpha-1-antitrypsin). The proteins detected that were highly changed in the cancer draining pulmonary vein from the lung resection serum study did not show the same change in the patients within 1 year of a lung cancer diagnosis in prospectively collected serum study SWATH-MS analysis using serum samples collected from a peripheral vein.

From the lung resection serum study, the list of 40 proteins showing an elevated abundance level in the cancer draining vein compared to the pulmonary artery was analysed in the SWATH-MS dataset. Of those 40 proteins, 12 were detected and identified with sufficient quality to pass through the statistical filtering. Of these 12, there was 1 protein that showed similar differential abundance in the patient samples within 12 months of a lung cancer diagnosis: keratin, type 1 cytoskeletal 14. This protein, effluxing from resectable lung tumours, may contribute to a biomarker signature that indicates presence of early stage disease. We included it in an analysis of group discrimination making a 12 protein signature to determine if specificity, sensitivity or other features were improved. The only improvement seen was in sensitivity outlined below.

We then broadened out timelines and assessed whether SWATH-MS and machine learning generated algorithms could discriminate between the prospectively collected serum study samples acquired 36 months prior to a lung cancer diagnosis. Comparisons between patient samples from within 36 months

Table 5. 58 Proteins with significant changes ($p \leq 0.05$) between the within 1 year of lung cancer diagnosis and the greater than 1-year diagnosis groups as determined with SWATH-MS. No false discovery was calculated for these changes.

Uniprot entry name	Protein name	Log_2 fold change	*p* value	**Included in panel**
PSMD1_HUMAN	26S proteasome non-ATPase regulatory subunit 1	-0.044	0.005	
RLA2_HUMAN	60S acidic ribosomal protein P2	-0.023	0.047	
6PGD_HUMAN	6-phosphogluconate dehydrogenase, decarboxylating	-0.042	0.042	
A1AT_HUMAN	Alpha-1-antitrypsin	0.019	0.011	✓
ACY1_HUMAN	Aminoacylase-1	-0.031	0.009	
ATPO_HUMAN	ATP synthase subunit O, mitochondrial	0.002	0.048	
CALM_HUMAN	Calmodulin	-0.025	0.007	
CK5P3_HUMAN	CDK5 regulatory subunit-associated protein 3	-0.050	0.004	
CERU_HUMAN	Ceruloplasmin	-0.173	0.026	
F13B_HUMAN	Coagulation factor XIII B chain	0.021	0.038	
C1R_HUMAN	Complement C1r subcomponent	-0.046	0.012	
C1RL_HUMAN	Complement C1r subcomponent-like protein	0.032	0.002	
C1S_HUMAN	Complement C1s subcomponent	0.014	0.012	
CO4A_HUMAN	Complement C4-A	0.080	< 0.001	✓
CO9_HUMAN	Complement component C9	0.052	0.002	✓
FHR2_HUMAN	Complement factor H-related protein 2	-0.006	0.018	
SERA_HUMAN	D-3-phosphoglycerate dehydrogenase	-0.028	0.004	
P5CS_HUMAN	Delta-1-pyrroline-5-carboxylate synthase	-0.013	0.011	
DENR_HUMAN	Density-regulated protein	-0.055	0.002	✓
EPIPL_HUMAN	Epiplakin	-0.020	0.044	

ECHD1_HUMAN	Ethylmalonyl-CoA decarboxylase	0.044	0.004	✓
FSCN1_HUMAN	Fascin	-0.025	0.016	
GTF2I_HUMAN	General transcription factor II-I	-0.015	0.027	✓
GFAP_HUMAN	Glial fibrillary acidic protein	-0.030	0.018	
G6PI_HUMAN	Glucose-6-phosphate isomerase	-0.039	0.023	✓
GPX3_HUMAN	Glutathione peroxidase 3	-0.037	0.024	
GSTP1_HUMAN	Glutathione S-transferase P	-0.059	0.008	
SYG_HUMAN	Glycine--tRNA ligase	-0.051	0.027	
GLOD4_HUMAN	Glyoxalase domain-containing protein 4	-0.032	0.021	
HPT_HUMAN	Haptoglobin	0.022	0.004	
HNRPD_HUMAN	Heterogeneous nuclear ribonucleoprotein D0	-0.059	0.008	
HRG_HUMAN	Histidine-rich glycoprotein	0.001	0.033	
GLO2_HUMAN	Hydroxyacylglutathione hydrolase, mitochondrial	0.014	0.030	
IGHA1_HUMAN	Immunoglobulin heavy constant alpha 1	0.025	0.047	
IGHM_HUMAN	Immunoglobulin heavy constant mu	-0.068	0.032	
HV107_HUMAN	Immunoglobulin heavy variable 1-46	-0.151	0.042	
HV303_HUMAN	Immunoglobulin heavy variable 3-23	-0.047	0.045	
KV101_HUMAN	Immunoglobulin kappa variable 1D-33	0.009	0.001	✓
LV301_HUMAN	Immunoglobulin lambda variable 3-1	0.041	< 0.001	
IPO7_HUMAN	Importin-7	-0.015	0.034	
ITA6_HUMAN	Integrin alpha-6	-0.050	0.030	✓
AQR_HUMAN	Intron-binding protein aquarius	-0.056	< 0.001	✓
K1C14_HUMAN	Keratin, type I cytoskeletal 14	-0.041	0.026	
K2C8_HUMAN	Keratin, type II cytoskeletal 8	-0.050	0.025	
LAMC1_HUMAN	Laminin subunit gamma-1	-0.037	0.008	
A2GL_HUMAN	Leucine-rich alpha-2-glycoprotein	0.049	0.006	

LDHA_HUMAN	L-lactate dehydrogenase A chain	0.007	0.006	
LAMP2_HUMAN	Lysosome-associated membrane glycoprotein 2	-0.055	0.006	
MDHC_HUMAN	Malate dehydrogenase, cytoplasmic	0.045	0.047	
MTA2_HUMAN	Metastasis-associated protein MTA2	-0.012	0.033	
GANAB_HUMAN	Neutral alpha-glucosidase AB	-0.053	0.023	
NAMPT_HUMAN	Nicotinamide phosphoribosyltransferase	-0.013	0.029	
NLTP_HUMAN	Non-specific lipid-transfer protein	-0.017	0.010	✓
NASP_HUMAN	Nuclear autoantigenic sperm protein	-0.013	0.004	
NUP50_HUMAN	Nuclear pore complex protein Nup50	0.027	0.030	
PLMN_HUMAN	Plasminogen	0.027	0.044	
GALT2_HUMAN	Polypeptide N-acetylgalactosaminyltransferase 2	-0.026	0.025	
PTGR1_HUMAN	Prostaglandin reductase 1	-0.035	0.009	
PSB4_HUMAN	Proteasome subunit beta type-4	-0.061	0.029	
PDIA1_HUMAN	Protein disulfide-isomerase	-0.013	0.005	
STXB2_HUMAN	Syntaxin-binding protein 2	-0.016	0.043	
TLN1_HUMAN	Talin-1	-0.015	0.014	
TFR1_HUMAN	Transferrin receptor protein 1	-0.044	0.006	
UGPA_HUMAN	UTP--glucose-1-phosphate uridylyltransferase	-0.042	0.041	
XRCC6_HUMAN	X-ray repair cross-complementing protein 6	-0.005	0.037	

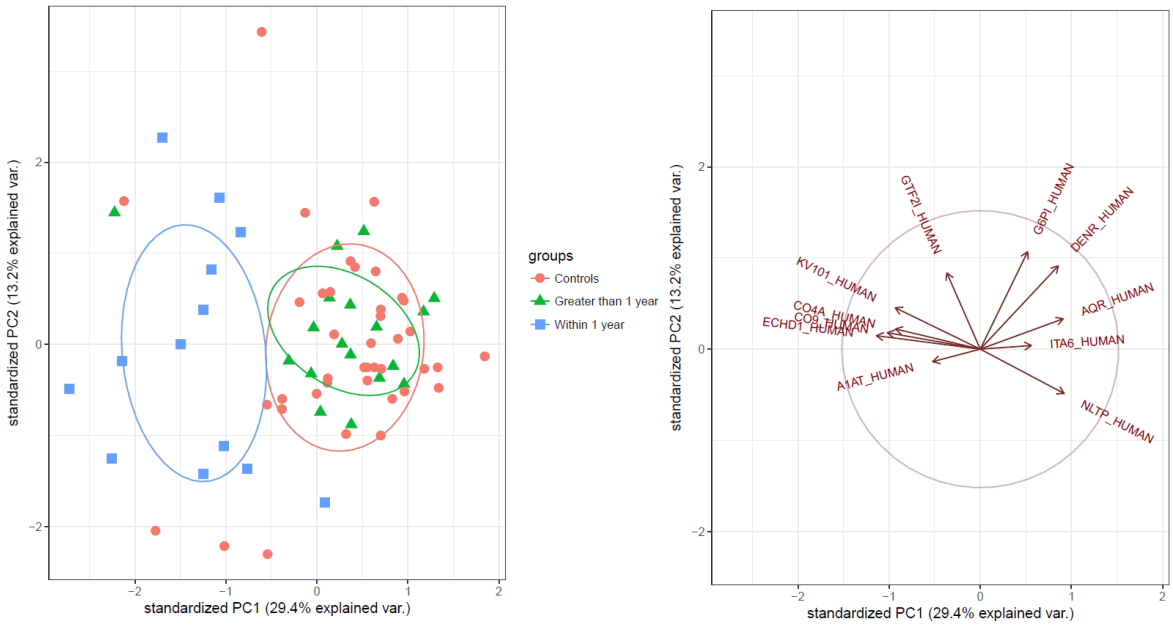


Figure 5 – Analysis of sample separation using an 11 protein signature for separation of those within 1 year of a lung cancer diagnosis and control samples. The PCA plot is constructed with the 11 quantification values for 11 proteins that displayed the greatest discriminatory power between experimental groups. Samples are grouped into those within 1 year prior to cancer diagnosis, those with greater than 1 year prior to diagnosis and the matched controls. The control samples overlapped the samples of the patients with greater than 1 year to diagnosis. The ellipses display a 0.75 probability of plotted position. The PCA plot shows a potential for discrimination between patients' samples close to cancer diagnosis and samples with a longer time before diagnosis. 79 of the samples that did not pass the quality control filtering or contained missing values for the panel proteins were not used in the PCA.

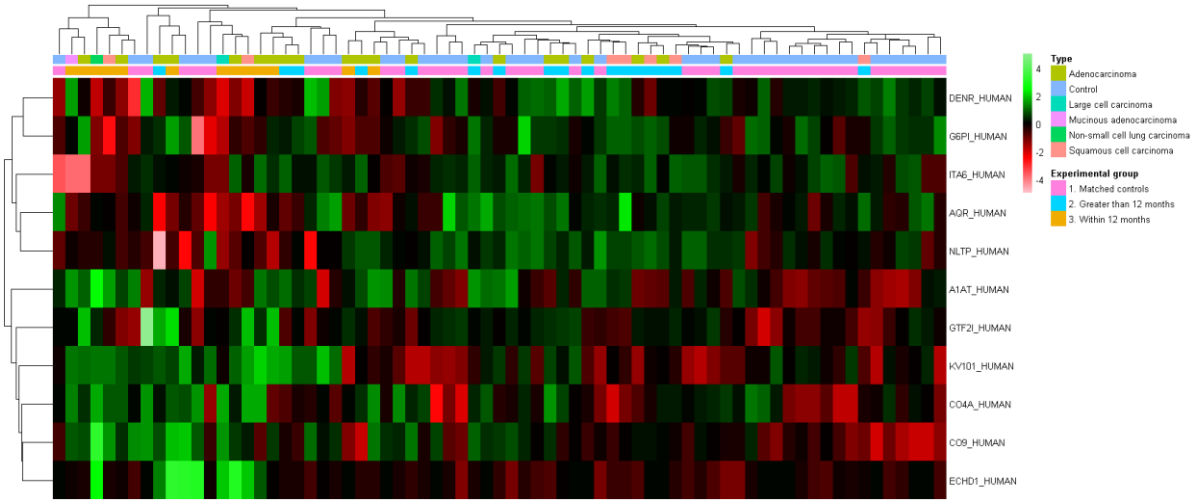


Figure 6 – Hierarchical clustering using protein panel levels of abundance between experimental groups and their discriminatory power. The difference in intensities of 11 proteins between samples from patients close to lung cancer diagnosis and the remaining samples is shown. Samples are grouped into those within 1 year prior to cancer diagnosis, those with greater than 1 year prior to diagnosis and the matched controls. Annotations along the columns show both the time to diagnosis and the Cancer type. Control samples closely matched the samples of the patients with a longer time to diagnosis. The clustering was performed using the Manhattan distance between samples, row scaling was performed on the $\log_2$ data for each protein by the subtraction of the mean from each feature and then dividing by the standard deviation. The 79 samples that did not pass the quality control filtering or contained missing values for the panel proteins were not used in the analysis.

of a lung cancer diagnosis were compared with those with greater than 36 months and matched non-cancer diagnosis controls.

Cancer related pathways and associated proteins

Pathway and network analyses were conducted on the set of 65 proteins from the prospectively collected serum study showing a significant degree of change within one year of lung cancer diagnosis. Significant enrichment of specific pathways was observed in the SWATH dataset that looked for early detection biomarkers at 1 year prior to diagnosis. Pathway analysis found enrichment of the network functions associated with organismal survival, cancer, cell death and survival amongst others (Supplementary table S3). The proteins that were associated with the Cancer network function were ALDH18A1, BAG1, CANX, CD163, CLU, CXCL8, CXCL12, CYBB, ENPP2, GAS7, GFI1, GPX3, HAGH, HRAS, HRG, IGF2BP1, IGHM, IL2, IL13, KRAS, LAMP2 and MDH1. A number of diseases and disorders were seen as enriched, with AQR, C9, EPPK1, F13B, GPI, GPX3, LAMC1, LDHA, MTA2, NAMPT, P4HB and SCP2 all having association with lung adenocarcinoma. Protein abundance levels for each protein from the prospectively collected serum study were plotted over time (Figure 7). Proteins that were observed undergoing a $\log_2$ fold change of 1 or greater were investigated and the ratio of samples from control patients and lung cancer patients was assessed. The isolation of proteins that were observed with a similar fold change in samples from patients that went on to develop lung cancer, several key proteins could be highlighted. Intron-binding protein aquarius (Figure 7A and B) showed a decrease in abundance over time on a patient by patient basis in the prospectively collected serum study lung cancer group with the trend following a linear correlation as determined by linear model analysis ($p = 0.003$). Density regulating protein (Figure 7C and D) also displayed this linear decrease in protein abundance levels in the lung cancer group ($p = 0.011$).

Machine learning verification of models

In order to assess the predictive power of the model, machine learning analysis was employed to determine the probability that protein panels could predict the early diagnosis of lung cancer. The predictive model was created after 1,000 iterations of training and testing. Training and testing datasets, created at random in a 70:30 split, were made using the intensity information of the 20 target proteins. Subset class imbalance was corrected using the SMOTE algorithm from the DMwR R package. The

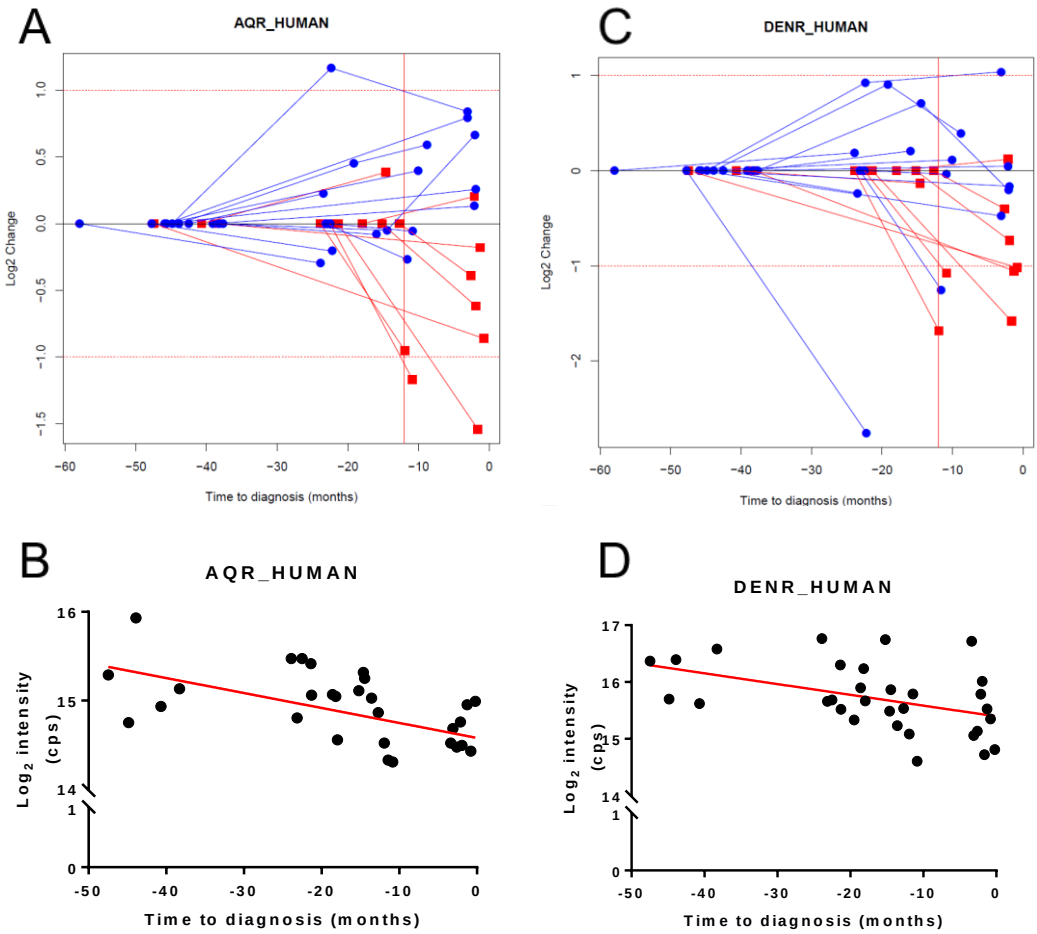


Figure 7 – Specific protein abundance levels: AQR and DENR. (A, B) Log_2 fold change in protein abundance for each patient across the experimental timeline. The time to diagnosis in months is shown on the x-axis with a red vertical line at 12 months. Fold change was calculated based of the first observed time point for each patient. Time points for the controls are shown as the same as the Case patients they are matched with. Points in red are from patients who went on to develop cancer, points in blue are from the control patients. **(C, D)** Linear model analysis showed a significant (p = 0.003 and 0.011 respectively) correlation of protein abundance over time closer to diagnosis.

training set was then used with the RandomForest R package to develop models. The resultant performance of the training model versus the testing set was recorded for each iteration and a receiver operating curves (ROC) were plotted for the early detection of lung cancer one year prior to diagnosis. For this 11 protein panel, the mean area under the curve (AUC) value for the RandomForest analysis was 0.89 with an accuracy of 0.89, a sensitivity of 0.88, and a specificity of 0.89 (Figure 8). We have shown in the lung resection serum study that keratin, type 1 cytoskeletal 14 egresses from lung tumours. The effectiveness of using this protein as a biomarker for early detection of lung cancer was tested by adding it to the discriminatory protein panel to create a combined panel. However, after testing the combined panel using machine learning methods as described above, a mean AUC value of 0.89, mean accuracy of 0.88, mean sensitivity of 0.91 and mean specificity of 0.88 were determined. In other words, there was only a marginal gain in sensitivity.

We then asked if we could devise a protein panel for early detection of lung cancer for within 36 months to diagnosis. Using the mean values from all of the iterations the performance of each panel was assessed. The 12 protein panel resulted in a mean AUC of 0.80, an accuracy of 0.78, a sensitivity of 0.79 and a specificity of 0.78. This shows that there is discriminatory power in the protein signature we developed even at the 36 month mark prior to diagnosis.

Discussion

Lung cancer is a leading cause of death, with low survival rates as tumour stages progress. Improving the detection of early stage lung cancer would transform patient outcomes. Using a novel approach of sample collection looking for proteins draining out of a tumour we have identified features of the natural history of tumours in situ via the definition of proteins egressing from the microenvironment and tumour cells. This unique approach to quantifying proteins at different stages of the circulatory system allows for measuring directly which proteins are released by the tumour while at the same time controlling for the abundance levels in the circulating blood in the non-cancerous parts of the organ. A subset of 50 proteins were seen with differential abundance between the cancer draining vein (C) and the non-cancerous vein (N). 60 proteins were identified as having a higher abundance level in the cancer draining vein compared to the pooled samples.

Fifty-eight proteins were identified as having an increased level of abundance in the (C) vein compared to the artery while also having no such change in the (N) vein compared to the pulmonary artery (A). Of

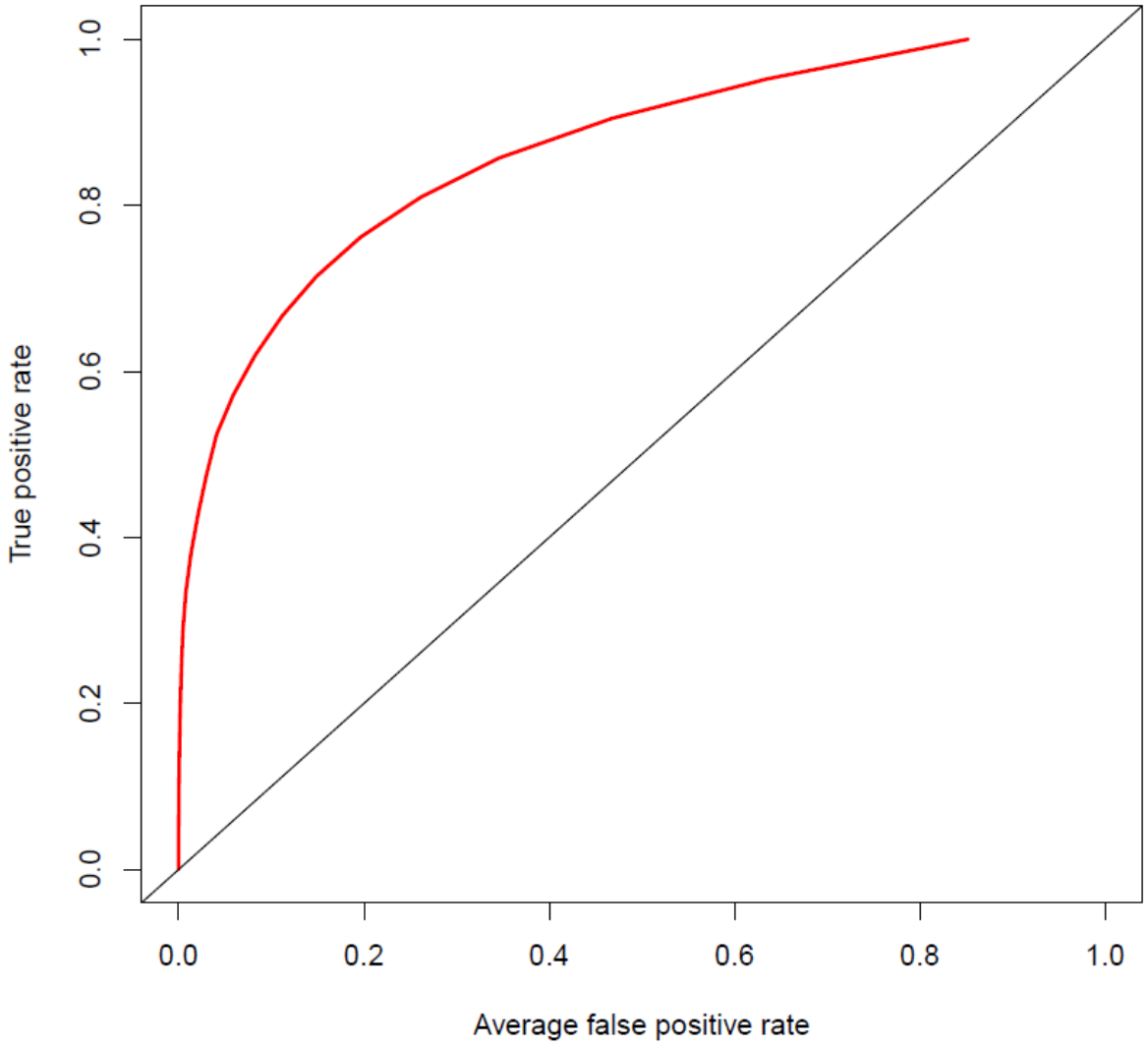


Figure 8: Receiver operating curve for the RandomForest analysis. Using the dataset from the combined protein panel of interest, 1,000 iterations of machine learning testing was conducted. The dataset was split 70:30 at random into a training set and testing set. The training set was balanced to even out the classes using the SMOTE function of the DMwR R package. A model was created using RandomForest and then tested on the testing set. This process was repeated 1,000 times, with the results of each iteration saved and averaged. The mean area under the curve value for the RandomForest analysis was 0.89 with an accuracy of 0.89, a sensitivity of 0.88 and a specificity of 0.89. As with the PCA and heatmap analysis, the samples that were removed due to quality control filtering or missing values for the panel proteins were not used in the randomForest analysis.

the proteins seen to change: circulating apolipoprotein levels have been strongly associated with cancer[28]; CA2 (Carbonic anhydrase 2) has been observed in cancerous blood vessels but not healthy vessels[29]; MET (Hepatocyte growth factor) is a well characterised oncogene, the abundance levels in blood plasma/serum of which have been linked with metastasis, survival and therapy response[30]; Platelet factor 4 (PF4), has been well characterised as a cancer promoting endocrine factor, the over-abundance of which causes adenocarcinogenesis[31]; Phosphoglycerate mutase 2 has also been seen to promote tumour growth and cell proliferation[32]. Plasma purine nucleoside phosphorylase (PNP) levels have not been widely investigated, however higher levels of PNP were observed in a range of cancer types in the circulating blood[33]. PNP has also been linked with pancreatic adenocarcinoma[34]. With many of the differentially expressed proteins identified having previously been observed as having metastatic and angiogenic properties, the analysis of protein abundance levels in the blood egressing from a tumour provides insights into the tumour microenvironment and the surrounding blood vessels. This study helps to identify proteins leaving the site of a tumour but these proteins are not necessarily useful as biomarkers for early detection of lung cancer. Of course, lung cancer resection is too late to be detecting biomarkers but we set up validation approaches to determine if these or any other protein commonly found in serum, could be used in an early detection algorithm.

To compare tumour egress proteins with available serum samples that enable early detection biomarkers to be assessed we used a nested case-control serum sample set from post-menopausal women participating in the UKCTOCS trial and who went on to develop lung cancer and age matched healthy controls. A SWATH reference library inclusive of the proteins egressing the tumour site was constructed as described previous in the methods section. But the SWATH-MS approach allowed a wider study search for a panel of proteins that allowed discrimination between patients close to a cancer diagnosis and those further from a diagnosis and controls.

In the analysis of the 58 prospective proteins identified in the lung resection serum study using the SWATH-MS data acquired from the serum from the prospectively collected serum study, only two of those candidate proteins showed a similar abundance pattern. With the other candidate proteins showing a non-significant change closer to diagnosis. A potential cause for this could be due to the large tumour mass not secreting proteins that are indicative of early stage lung cancer risk. Additionally as the lung resection serum study was not conducted against cancer free controls the comparison between the blood leaving a tumour and the blood of a patient prior to developing lung cancer is limited.

The difference in lung cancer stage and morphology between the lung resection serum study and the prospectively collected serum study also provides a limitation in the comparison between the approaches. This analysis reveals that the study of proteins effluxing from early stage tumours is not necessarily an ideal approach for the determination of blood borne biomarkers for early detection of lung cancer. This disparity in detection may be a contributing factor towards the low level of translation of biomarkers from identification to clinical tests. Despite the vast majority of potential proteins showing no correlation with the prospective samples, these proteins could still provide a use in increasing the understanding of the tumour microenvironment and the link between tumours and proteins in the circulatory system.

Using our informatics approach, the application of SWATH-MS on longitudinal samples was capable of determining a panel of 11 proteins that could discriminate between the samples from patients within a year of lung cancer diagnosis and both the samples of patients greater than one year to lung cancer diagnosis and the matched controls. Many of the 11 proteins that make up the discriminatory panel have known associations with cancer and tumorigenesis but none of the proteins were observed as changing in the lung resection serum study. Only 2 proteins from that study also showed significant changes in protein abundance between the 'within 1 year of lung cancer diagnosis' samples and the other samples. Keratin 14 has previously been observed to regulate metastasis in breast cancer[35] and has been proposed as a marker for differential diagnosis of squamous cell carcinomas[36]. The complement cascade has an intricate relationship with cancer. The potential for components of the complement cascade to be used as biomarkers is still seen as controversial despite multiple studies linking changes in complement component abundance and various cancers[37]. The 11 protein panel does not have a link with lung cancer literature and may not have been detected and chosen in more traditional proteomic approaches. In the 11 protein panel: Density-regulated protein has been observed to interact with the oncogene MCT-1 and is upregulated in breast cancer[38,39]; Integrins have long been associated with cancer with roles in tumour survival and metastasis[40] with integrin alpha 6 having observed links with breast cancer[41]; Alpha 1 antitrypsin deficient patients having been observed with higher risk of lung cancer[42–44]; Ethylmalonyl‐CoA decarboxylase (ECHDC1) has been shown to display increased abundance in particular lines of bladder cancer cells[31] and silencing of the gene resulted in significantly inhibited proliferation of bladder cancer cells; Glucose-6-phosphate isomerase (GPI) has been previously seen to promote tumour growth and metastasis. With higher GPI abundance levels

correlating with worse survival rates[31]. Evaluation of the panel with regard to tumour stage found that there was little to no difference in discriminatory performance on the different stages of tumour. When the 11 protein panel is extended to a 12 protein panel with the protein that egressed from lung tumours there is a minor change in the discriminatory power of the model.

Extending the analysis to a consideration of prediction at a 36 month cut off may have value in a pre-screening strategy. Whilst sensitivity and specificity is decreased compared to the one year prior to diagnosis assessment, the former test may have more value. The potential outcome of using the panel as part of a prospective test or in a wider screening programme would be detection of lung cancer at an earlier stage. Even for late stage lung cancer patients, an extra few months could allow for additional therapy or to increase the chances of being admissible to clinical trials. Positive indication by the panel could alert the patient or health provider to seek chest scans to confirm the diagnosis. In order to fully determine the usefulness of the panel, an investigation into the positive and negative predictive values of the measurements would be required. While the panel currently lacks the validity necessary to be used in screening, the benefits and possibilities are compelling. The tool developed would have application in relatively high risk populations of ever smokers beyond the age of 50. We would envisage that development of a screening tool could precede CT scanning as a means of triage using a small amount of blood that would generally be willingly given. Consent for the small amount needed for such assays is likely to be more easily found than with other assays needing approximately 20 mLs of blood, such as cell based assays[45].

Machine learning based techniques were employed to generate and test statistical models aiming to predict the classes of each sample in the SWATH-MS experiment in respect of developing an understanding of the capability to discriminate as early as possible who will be diagnosed with lung cancer. This analysis was able to characterise the capability of protein abundance as a means to predict the diagnosis of lung cancer within 1 year. The protein panel identified and the statistical models generated from it can be used as part of a training set for further testing on another cohort in future work. The SWATH-MS approach was also able to discriminate between patient samples at 36 months prior to lung cancer diagnosis, however the strength of the models to determine patient condition was reduced in comparison to the separation of samples from within 1 year of lung cancer diagnosis. 79 of the SWATH-MS samples produced either missing values or produced results with too great a protein FDR. In order to reduce the impact of this limitation, the methodology can be improved

by moving to antibody based methods with highly specific antibodies, preferably in a test that can be provided at the point of care. Alternatively, targeted mass spectrometry using optimised conditions and spectral libraries tailored to specifically identify and quantify the proteins in the prospective panel.

A limitation of the informatics is the lack of multiple testing correction, this was due to the number of proteins measured and the discovery nature of the study. Therefore the protein panel would need to be assessed in a verification study. A limitation of the study is that the UKCTOCS trial only had samples from post-menopausal women, this would mean that further validation in men and/or younger women would be needed for further analysis of the panel. Work can now be conducted to analyse further sets of longitudinal samples from the UKCTOCS blood bank using our SWATH-MS approach in order to verify the biomarkers discovered.

The approaches used highlight the capability of SWATH-MS to perform biomarker identification. This research methodology and instrumentation can be further used on a wide range of potential diseases and conditions that can provide valuable insight that may otherwise be missed.

Supporting information

The following supporting information is available free of charge at ACS website http://pubs.acs.org:

Table S1. IPA analysis of proteins from the tumour resection study with higher expression in the cancer draining vein compared to the non-cancer draining vein.

Table S2. IPA analysis of proteins from the tumour resection study with higher expression in the cancer draining vein compared to the artery.

Table S3. IPA analysis of proteins from the prospectively collected serum study with significant differential expression in patients within 1 year of cancer diagnosis

Funding source

Mass spectrometry was supported with equipment grants from Bloodwise and Medical Research Council. This work was supported by the CRUK Manchester Centre award (C5759/A25254). CD and ADW are supported by the NIHR Manchester Biomedical Research Centre. CD is also supported via core funding to the CRUK Manchester Institute and via the CRUK Lung Cancer Centre of Excellence (C5759/A20465). The pulmonary vein study was funded by Roy Castle Lung Cancer Foundation (PC) and the North West Lung Centre Charity (PC). UKCTOCS was funded by the Medical Research Council

(G9901012 and G0801228), Cancer Research UK (C1479/A2884), the Department of Health and with additional support from The Eve Appeal. Senior investigators at UCL supported by the NIHR University College London Hospitals (UCLH) Biomedical Research Centre.

Declaration of interests

The authors have no financial conflicts of interest.

Author contributions

B.G. performed experiments, analysed data, and wrote the manuscript. M.J.W. and J.T.S. performed experiments. D.L., M.P., S.M-D., and N.A. designed and performed data analysis. P.K., R.S., and R.B. performed the collection and banking of serum samples for the lung resection serum study. S.A. and J.B. provided serum samples for the prospectively collected serum study. P.A.C. designed the lung resection serum study. A.D.W. designed and developed the prospectively collected serum study. All authors revised and edited the manuscript.

References

(1) National Cancer Intelligence Network. Stage Breakdown by CCG 2014 http://www.ncin.org.uk/publications/survival_by_stage.

(2) ISDN Scotland. Detect Cancer Early Staging Data http://www.isdscotland.org/Health-Topics/Cancer/Detect-Cancer-Early/.

(3) Northern Ireland Cancer Registry. Incidence by stage 2010-2014 http://www.qub.ac.uk/research-centres/nicr/CancerInformation/official-statistics/.

(4) Team, T. N. L. S. T. R. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *N. Engl. J. Med.* **2011**, *365* (5), 395–409. https://doi.org/10.1056/NEJMoa1102873.

(5) Ferlay, J.; Soerjomataram, I.; Dikshit, R.; Eser, S.; Mathers, C.; Rebelo, M.; Parkin, D. M.; Forman, D.; Bray, F. Cancer Incidence and Mortality Worldwide: Sources, Methods and Major Patterns in GLOBOCAN 2012. *Int. J. Cancer* **2015**, *136* (5), E359–E386. https://doi.org/10.1002/ijc.29210.

(6) Okada, M.; Nishio, W.; Sakamoto, T.; Uchino, K.; Yuki, T.; Nakagawa, A.; Tsubota, N. Effect of Tumor Size on Prognosis in Patients with Non–Small Cell Lung Cancer: The Role of

Segmentectomy as a Type of Lesser Resection. *J. Thorac. Cardiovasc. Surg.* **2005**, *129* (1), 87–93. https://doi.org/10.1016/J.JTCVS.2004.04.030.

(7) Hou, J.-M.; Krebs, M. G.; Lancashire, L.; Sloane, R.; Backen, A.; Swain, R. K.; Priest, L. J. C.; Greystoke, A.; Zhou, C.; Morris, K.; et al. Clinical Significance and Molecular Characteristics of Circulating Tumor Cells and Circulating Tumor Microemboli in Patients With Small-Cell Lung Cancer. *J. Clin. Oncol.* **2012**, *30* (5), 525–532. https://doi.org/10.1200/JCO.2010.33.3716.

(8) Krebs, M. G.; Sloane, R.; Priest, L.; Lancashire, L.; Hou, J.-M.; Greystoke, A.; Ward, T. H.; Ferraldeschi, R.; Hughes, A.; Clack, G.; et al. Evaluation and Prognostic Significance of Circulating Tumor Cells in Patients With Non–Small-Cell Lung Cancer. *J. Clin. Oncol.* **2011**, *29* (12), 1556–1563. https://doi.org/10.1200/JCO.2010.28.7045.

(9) Crosbie, P. A. J.; Shah, R.; Krysiak, P.; Zhou, C.; Morris, K.; Tugwood, J.; Booton, R.; Blackhall, F.; Dive, C. Circulating Tumor Cells Detected in the Tumor-Draining Pulmonary Vein Are Associated with Disease Recurrence after Surgical Resection of NSCLC. *J. Thorac. Oncol.* **2016**, *11* (10), 1793–1797. https://doi.org/10.1016/j.jtho.2016.06.017.

(10) Huang, W.-L.; Chen, Y.-L.; Yang, S.-C.; Ho, C.-L.; Wei, F.; Wong, D. T.; Su, W.-C.; Lin, C.-C. Liquid Biopsy Genotyping in Lung Cancer: Ready for Clinical Utility? *Oncotarget* **2017**, *8* (11), 18590–18608. https://doi.org/10.18632/oncotarget.14613.

(11) Board, R. E.; Williams, V. S.; Knight, L.; Shaw, J.; Greystoke, A.; Ranson, M.; Dive, C.; Blackhall, F. H.; Hughes, A. Isolation and Extraction of Circulating Tumor DNA from Patients with Small Cell Lung Cancer. *Ann. N. Y. Acad. Sci.* **2008**, *1137* (1), 98–107. https://doi.org/10.1196/annals.1448.020.

(12) Krebs, M. G.; Metcalf, R. L.; Carter, L.; Brady, G.; Blackhall, F. H.; Dive, C. Molecular Analysis of Circulating Tumour Cells—Biology and Biomarkers. *Nat. Rev. Clin. Oncol.* **2014**, *11* (3), 129–144. https://doi.org/10.1038/nrclinonc.2013.253.

(13) Zhou, C.; Simpson, K. L.; Lancashire, L. J.; Walker, M. J.; Dawson, M. J.; Unwin, R. D.; Rembielak, A.; Price, P.; West, C.; Dive, C.; et al. Statistical Considerations of Optimal Study Design for Human Plasma Proteomics and Biomarker Discovery. *J. Proteome Res.* **2012**, *11* (4), 2103–2113. https://doi.org/10.1021/pr200636x.

(14) Zhou, C.; Walker, M. J.; Williamson, A. J. K.; Pierce, A.; Berzuini, C.; Dive, C.; Whetton, A. D. A Hierarchical Statistical Modeling Approach to Analyze Proteomic Isobaric Tag for Relative and Absolute Quantitation Data. *Bioinformatics* **2014**, *30* (4), 549–558. https://doi.org/10.1093/bioinformatics/btt722.

(15) Russell, M. R.; Walker, M. J.; Williamson, A. J. K.; Gentry-Maharaj, A.; Ryan, A.; Kalsi, J.; Skates, S.; D'Amato, A.; Dive, C.; Pernemalm, M.; et al. Protein Z: A Putative Novel Biomarker for Early Detection of Ovarian Cancer. *Int. J. Cancer* **2016**, *138* (12), 2984–2992. https://doi.org/10.1002/ijc.30020.

(16) Russell, M. R.; Graham, C.; D'Amato, A.; Gentry-Maharaj, A.; Ryan, A.; Kalsi, J. K.; Ainley, C.; Whetton, A. D.; Menon, U.; Jacobs, I.; et al. A Combined Biomarker Panel Shows Improved Sensitivity for the Early Detection of Ovarian Cancer Allowing the Identification of the Most Aggressive Type II Tumours. *Br. J. Cancer* **2017**, *117* (5), 666–674. https://doi.org/10.1038/bjc.2017.199.

(17) Walker, M. J.; Zhou, C.; Backen, A.; Pernemalm, M.; Williamson, A. J. K.; Priest, L. J. C.; Koh, P.; Faivre-Finn, C.; Blackhall, F. H.; Dive, C.; et al. Discovery and Validation of Predictive Biomarkers of Survival for Non-Small Cell Lung Cancer Patients Undergoing Radical Radiotherapy: Two Proteins With Predictive Value. *EBioMedicine* **2015**, *2* (8), 841–850. https://doi.org/10.1016/j.ebiom.2015.06.013.

(18) Menon, U.; Gentry-Maharaj, A.; Ryan, A.; Sharma, A.; Burnell, M.; Hallett, R.; Lewis, S.; Lopez, A.; Godfrey, K.; Oram, D.; et al. Recruitment to Multicentre Trials--Lessons from UKCTOCS: Descriptive Study. *BMJ* **2008**, *337*, a2079. https://doi.org/10.1136/BMJ.A2079.

(19) Pepe, M. S.; Feng, Z.; Janes, H.; Bossuyt, P. M.; Potter, J. D. Pivotal Evaluation of the Accuracy of a Biomarker Used for Classification or Prediction: Standards for Study Design. *J. Natl. Cancer Inst.* **2008**, *100* (20), 1432–1438. https://doi.org/10.1093/jnci/djn326.

(20) Anderson, N. L.; Anderson, N. G. The Human Plasma Proteome: History, Character, and Diagnostic Prospects. *Mol. Cell. Proteomics* **2002**, *1* (11), 845–867. https://doi.org/10.1074/MCP.R200007-MCP200.

(21) Wen-Hai Jin, †,‡; Jie Dai, †; Su-Jun Li, †; Qi-Chang Xia, †; Han-Fa Zou, *,‡ and; Rong Zeng*,

†. Human Plasma Proteome Analysis by Multidimensional Chromatography Prefractionation and Linear Ion Trap Mass Spectrometry Identification. **2005**. https://doi.org/10.1021/PR049761H.

(22) Liu, T.; Qian, W.-J.; Gritsenko, M. A.; Xiao, W.; Moldawer, L. L.; Kaushal, A.; Monroe, M. E.; Varnum, S. M.; Moore, R. J.; Purvine, S. O.; et al. High Dynamic Range Characterization of the Trauma Patient Plasma Proteome. *Mol. Cell. Proteomics* **2006**, *5* (10), 1899–1913. https://doi.org/10.1074/mcp.M600068-MCP200.

(23) Nash, G. F.; Walsh, D. C.; Kakkar, A. K. The Role of the Coagulation System in Tumour Angiogenesis. *Lancet. Oncol.* **2001**, *2* (10), 608–613.

(24) Everse, J.; Everse, K. E.; Grisham, M. B. *Peroxidases in Chemistry and Biology*; CRC Press, 1991.

(25) Kiyohara, C.; Yoshimasu, K.; Takayama, K.; Nakanishi, Y. NQO1, MPO, and the Risk of Lung Cancer: A HuGE Review. *Genet. Med.* **2005**, *7* (7), 463–478. https://doi.org/10.1097/01.gim.0000177530.55043.c1.

(26) Liu, Y.; Buil, A.; Collins, B. C.; Gillet, L. C.; Blum, L. C.; Cheng, L.-Y.; Vitek, O.; Mouritsen, J.; Lachance, G.; Spector, T. D.; et al. Quantitative Variability of 342 Plasma Proteins in a Human Twin Population. *Mol. Syst. Biol.* **2015**, *11* (2), 786–786. https://doi.org/10.15252/msb.20145728.

(27) Wu, J. X.; Song, X.; Pascovici, D.; Zaw, T.; Care, N.; Krisp, C.; Molloy, M. P. SWATH Mass Spectrometry Performance Using Extended Peptide MS/MS Assay Libraries. *Mol. Cell. Proteomics* **2016**, *15* (7), 2501–2514. https://doi.org/10.1074/mcp.M115.055558.

(28) Borgquist, S.; Butt, T.; Almgren, P.; Shiffman, D.; Stocks, T.; Orho-Melander, M.; Manjer, J.; Melander, O. Apolipoproteins, Lipids and Risk of Cancer. *Int. J. Cancer* **2016**, *138* (11), 2648–2656. https://doi.org/10.1002/ijc.30013.

(29) Yoshiura, K.; Nakaoka, T.; Nishishita, T.; Sato, K.; Yamamoto, A.; Shimada, S.; Saida, T.; Kawakami, Y.; Takahashi, T. A.; Fukuda, H.; et al. Carbonic Anhydrase II Is a Tumor Vessel Endothelium-Associated Antigen Targeted by Dendritic Cell Therapy. *Clin. Cancer Res.* **2005**, *11* (22), 8201–8207. https://doi.org/10.1158/1078-0432.CCR-05-0816.

(30) Matsumoto, K.; Umitsu, M.; De Silva, D. M.; Roy, A.; Bottaro, D. P. Hepatocyte Growth Factor/MET in Cancer Progression and Biomarker Discovery. *Cancer Sci.* **2017**, *108* (3), 296–307. https://doi.org/10.1111/cas.13156.

(31) Asai, S.; Miura, N.; Sawada, Y.; Noda, T.; Kikugawa, T.; Tanji, N.; Saika, T. Silencing of ECHDC1 Inhibits Growth of Gemcitabine‐resistant Bladder Cancer Cells. *Oncol. Lett.* **2017**, *15* (1), 522–527. https://doi.org/10.3892/ol.2017.7269.

(32) Xu, Y.; Li, F.; Lv, L.; Li, T.; Zhou, X.; Deng, C.-X.; Guan, K.-L.; Lei, Q.-Y.; Xiong, Y. Oxidative Stress Activates SIRT2 to Deacetylate and Stimulate Phosphoglycerate Mutase. *Cancer Res.* **2014**, *74* (13), 3630–3642. https://doi.org/10.1158/0008-5472.CAN-13-3615.

(33) Roberts, E. L. L.; Newton, R. P.; Axford, A. T. Plasma Purine Nucleoside Phosphorylase in Cancer Patients. *Clin. Chim. Acta* **2004**, *344* (1–2), 109–114. https://doi.org/10.1016/J.CCCN.2004.02.008.

(34) Vareed, S. K.; Bhat, V. B.; Thompson, C.; Vasu, V. T.; Fermin, D.; Choi, H.; Creighton, C. J.; Gayatri, S.; Lan, L.; Putluri, N.; et al. Metabolites of Purine Nucleoside Phosphorylase (NP) in Serum Have the Potential to Delineate Pancreatic Adenocarcinoma. *PLoS One* **2011**, *6* (3), e17177. https://doi.org/10.1371/journal.pone.0017177.

(35) Cheung, K. J.; Padmanaban, V.; Silvestri, V.; Schipper, K.; Cohen, J. D.; Fairchild, A. N.; Gorin, M. A.; Verdone, J. E.; Pienta, K. J.; Bader, J. S.; et al. Polyclonal Breast Cancer Metastases Arise from Collective Dissemination of Keratin 14-Expressing Tumor Cell Clusters. *Proc. Natl. Acad. Sci.* **2016**, *113* (7), E854–E863. https://doi.org/10.1073/pnas.1508541113.

(36) Chu, P. G.; Lyda, M. H.; Weiss, L. M. Cytokeratin 14 Expression in Epithelial Neoplasms: A Survey of 435 Cases with Emphasis on Its Value in Differentiating Squamous Cell Carcinomas from Other Epithelial Tumours. *Histopathology* **2001**, *39* (1), 9–16. https://doi.org/10.1046/j.1365-2559.2001.01105.x.

(37) Reis, E. S.; Mastellos, D. C.; Ricklin, D.; Mantovani, A.; Lambris, J. D. Complement in Cancer: Untangling an Intricate Relationship. *Nat. Rev. Immunol.* **2017**, *18* (1), 5–18. https://doi.org/10.1038/nri.2017.97.

(38) Mazan-Mamczarz, K.; Gartenhaus, R. B. Post-Transcriptional Control of the MCT-1-Associated Protein DENR/DRP by RNA-Binding Protein AUF1. *Cancer Genomics Proteomics 4* (3), 233–239.

(39) Oh, J. J.; Grosshans, D. R.; Wong, S. G.; Slamon, D. J. Identification of Differentially Expressed Genes Associated with HER-2/Neu Overexpression in Human Breast Cancer Cells. *Nucleic Acids Res.* **1999**, *27* (20), 4008–4017.

(40) Jin, H.; Varner, J. Integrins: Roles in Cancer Development and as Treatment Targets. *Br. J. Cancer* **2004**, *90* (3), 561–565. https://doi.org/10.1038/sj.bjc.6601576.

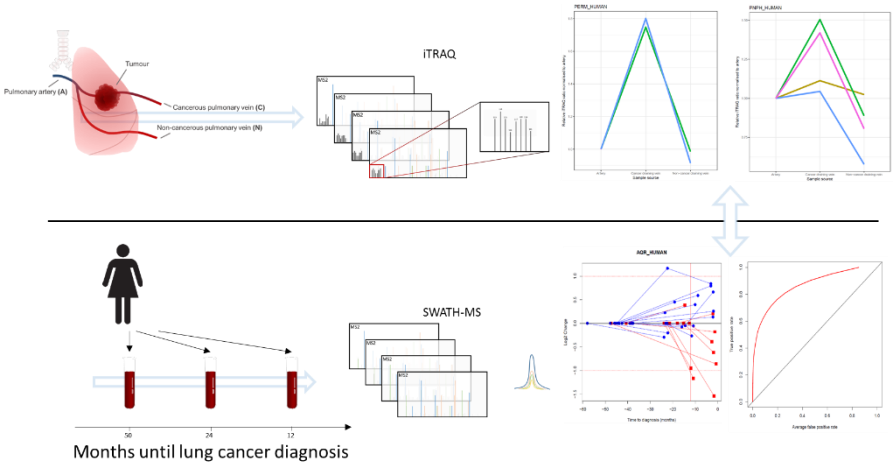
(41) Mariani Costantini, R.; Falcioni, R.; Battista, P.; Zupi, G.; Kennel, S. J.; Colasante, A.; Venturo, I.; Curio, C. G.; Sacchi, A. Integrin (Alpha 6/Beta 4) Expression in Human Lung Cancer as Monitored by Specific Monoclonal Antibodies. *Cancer Res.* **1990**, *50* (18), 6107–6112.

(42) Yang, P.; Sun, Z.; Krowka, M. J.; Aubry, M.-C.; Bamlet, W. R.; Wampfler, J. A.; Thibodeau, S. N.; Katzmann, J. A.; Allen, M. S.; Midthun, D. E.; et al. Alpha1-Antitrypsin Deficiency Carriers, Tobacco Smoke, Chronic Obstructive Pulmonary Disease, and Lung Cancer Risk. *Arch. Intern. Med.* **2008**, *168* (10), 1097. https://doi.org/10.1001/archinte.168.10.1097.

(43) Torres-Durán, M.; Ruano-Ravina, A.; Parente-Lamelas, I.; Abal-Arca, J.; Leiro-Fernández, V.; Montero-Martínez, C.; Pena, C.; Castro-Añón, O.; Golpe-Gómez, A.; González-Barcala, F. J.; et al. Alpha-1 Antitrypsin Deficiency and Lung Cancer Risk. *J. Thorac. Oncol.* **2015**, *10* (9), 1279–1284. https://doi.org/10.1097/JTO.0000000000000609.

(44) Pérez-Holanda, S.; Blanco, I.; Menéndez, M.; Rodrigo, L. Serum Concentration of Alpha-1 Antitrypsin Is Significantly Higher in Colorectal Cancer Patients than in Healthy Controls. *BMC Cancer* **2014**, *14*, 355. https://doi.org/10.1186/1471-2407-14-355.

(45) Quandt, D.; Dieter Zucht, H.; Amann, A.; Wulf-Goldenberg, A.; Borrebaeck, C.; Cannarile, M.; Lambrechts, D.; Oberacher, H.; Garrett, J.; Nayak, T.; et al. Implementing Liquid Biopsies into Clinical Decision Making for Cancer Immunotherapy. *Oncotarget* **2017**, *8* (29), 48507–48520. https://doi.org/10.18632/oncotarget.17397.


Tumour
Pulmonary artery (A)
Cancerous pulmonary vein (C)
Non-cancerous pulmonary vein (N)
iTRAQ
SWATH-MS
Months until lung cancer diagnosis

For TOC only