University of Glasgow

Husi, H., Skipworth, R. J.E., Cronshaw, A., Stephens, N. A., Wackerhage, H., Greig, C., Fearon, K. C.H., and Ross, J. A. (2015) Programmed cell death 6 interacting protein (PDCD6IP) and Rabenosyn-5 (ZFYVE20) are potential urinary biomarkers for upper gastrointestinal cancer. Proteomics Clinical Applications, 9(5-6), pp. 586-596.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

This is the peer reviewed version of the following article: Husi, H., Skipworth, R. J.E., Cronshaw, A., Stephens, N. A., Wackerhage, H., Greig, C., Fearon, K. C.H., and Ross, J. A. (2015) Programmed cell death 6 interacting protein (PDCD6IP) and Rabenosyn-5 (ZFYVE20) are potential urinary biomarkers for upper gastrointestinal cancer. Proteomics Clinical Applications, 9(5-6), pp. 586-596, which has been published in final form at http://dx.doi.org/10.1002/prca.201400111. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

http://eprints.gla.ac.uk/128436/

Deposited on: 16 November 2016

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk

Programmed cell death 6-interacting protein (PDCD6IP) and Rabenosyn-5 (ZFYVE20) are potential urinary biomarkers for upper gastrointestinal cancer

Holger Husi[1,2], Richard J.E. Skipworth[2], Andrew Cronshaw[3], Nathan A. Stephens[2], Henning Wackerhage[4], Carolyn Greig[2,5], Kenneth C. H. Fearon[2,6], James A. Ross[2,6]

[1] Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, UK

[2] School of Clinical Sciences, University of Edinburgh, Edinburgh, UK

[3] School of Biological Sciences, University of Edinburgh, Edinburgh, UK

[4] School of Medical Sciences, University of Aberdeen, Aberdeen AB25 2ZD, UK

[5] School of Sport, Exercise and Rehabilitation Sciences, University of Birmingham, Birmingham B15 2TT, UK

[6] joint senior authors

Corresponding author:

Professor James A. Ross

or Richard Skipworth

University of Edinburgh

School of Clinical Sciences

Chancellors Building

49 Little France Crescent

Edinburgh , EH16 4SB, UK

E-mail: j.a.ross@ed.ac.uk

or rskipworth@hotmail.com

FAX: +44-131-242-6520

Abbreviations: BPS, Biomarker Pattern Software; emPAI, exponentially modified protein abundance index; uGI, upper gastro-intestinal; OGJ, oesophagogastric junction; PCDMS,

ProteinChip Data Manager software; SPA, sinapinic acid; CRP, C-reactive protein.

Keywords: mass spectrometry / Upper gastrointestinal cancer / Urine biomarker

Word count of text: 3,958; Word count of text and legends: 6,267.

Abstract

Purpose - Cancer of the upper digestive tract (uGI) is a major contributor to cancer-related death worldwide. Due to a rise in occurrence, together with poor survival rates and a lack of diagnostic or prognostic clinical assays, there is a clear need to establish molecular biomarkers.

Experimental design - Initial assessment was performed using urines from 60 control and 60 uGI cancer patients by MS to establish a peak-pattern or fingerprint model, which was validated by a further set of 59 samples.

Results - We detected 86 cluster-peaks by MS above frequency- and detection-thresholds. Statistical testing and model building resulted in a peak-profiling model of 5 relevant peaks with an 88% overall sensitivity and 91% specificity, and overall correctness of 90%. High resolution MS from 40 samples in the 2-10 kDa range resulted in 646 identified proteins, and pattern matching identified 4 of the 5 model-peaks within significant parameters, namely Programmed cell death 6-interacting protein (PDCD6IP/Alix/AIP1), Rabenosyn-5 (ZFYVE20), Protein S100A8 and Protein S100A9.

Conclusions and clinical relevance - We demonstrate that MS analysis of human urine can identify lead biomarker candidates in uGI cancers which makes this technique potentially useful in defining and consolidating biomarker patterns for uGI cancer screening.

1. Introduction

Malignant tumours of the upper gastro-intestinal (uGI) tract are a significant cause of cancer-related death worldwide. Gastric cancer is the fourth most common cancer globally (approximately 8% of all cancer cases) and the second most common cause of cancer-related death [1]. Oesophageal cancer is the sixth most common cancer worldwide [2]. It affects twice as many males than females, amounts to approximately 4% of all cancer cases and 5% of all cancer related deaths, and has a 5-year survival rate of only 13% [3]. In comparison, pancreatic cancer has one of the lowest survival rates of all (16% at one year following diagnosis which drops to 3% by 5 years). In Western populations, pancreatic cancer contributes 3% of all diagnosed cancer cases and 5% of all cancer-related deaths [4].

The incidence of gastric cancer has been falling continually over the last few decades [5], but the incidence rates of pancreatic and oesophageal cancers globally have shown a worrying increase [6, 7]. The development of various therapeutic strategies, based either on improved surgical techniques or neo-adjuvant/adjuvant chemotherapy has improved survival rates but only in the order of a few percent [8-10]. One of the main reasons for this is that most cases are diagnosed at an advanced stage. As a consequence, it is of utmost importance to develop methodologies to diagnose the onset of malignancy as soon as possible, preferably prior to the development of clinical symptoms.

Currently, there is a clear lack of accepted clinical molecular markers which can be used reliably to assess for the presence of malignancy [11]. The most successful and clinically widely used assay for pancreatic cancer is serum biomarker CA19.9. However, its use is limited by poor sensitivity in symptomatic patients (only 80%), false negative

results of 5-10% in the Lewis negative phenotype [12], and increased false positive rates of 10-60% in the presence of obstructive jaundice [13]. CA19.9 can not be used for screening asymptomatic populations [14], and its specificity to diagnose pancreatic cancer in symptomatic patients ranges from 82 to 90% [15]. Results using commercial sources of CA19.9 assays are not interchangeable, and show very poor usability for other tumour types such as gastric cancer, with detection rates of 3 to 38% [16]. A more suitable pan-cancer biomarker was identified as CA215, which, like CA19.9, consists of carbohydrate-associated epitopes of proteins, but the positive predictive values were shown to be very low, ranging from 38 to 83% across a number of common cancer types [17]. Recent studies have focused on micro RNA, where elevated quantities of miR-222 in diseased tissue could be associated with pancreatic cancer and poor prognosis [18]. A potential complementary set of biomarkers for gastric cancer was recently identified as V-crk avian sarcoma virus CT10 oncogene homolog-like (CRKL) and miR-126 [19], and legumain (LGMN) was described to be associated with poor prognosis and metastasis [20]. Oesophageal cancer biomarkers were identified as phospholipase A2 group IIA (PLA2G2A) [21], salivary miRs [22], and HCLS1-associated protein X-1 (HAX1) [23] amongst others. Potential markers for a wide range of epithelial tumours include S100A6 (breast [24], bowel [25], gastric [26], hepatocellular [27], ovarian [28], thyroid [29], and uGI cancer [30]), and the protease inhibitor inter-alpha-trypsin inhibitor heavy chain H4 (ITIH4) (breast [31, 32], ovarian [33], prostate [34, 35], pancreatic [36, 37], lung [38], colon [39] and bladder cancer [34]).

There is a current trend to use a combination of several markers to assess tumour growth for diagnostic and prognostic purposes. An improvement in CA19.9 applicability

for pancreatic cancer detection involves the combination of several additional biomarkers, such as regenerating islet-derived 1 beta (REG1B), syncollin (SYCN), anterior gradient homolog 2 protein (AGR2), and lysyl oxidase-like 2 (LOXL2) [40]. An extension of this approach involves the multiplexing of many different targets simultaneously using antibody-arrays [41, 42] or quantum dot technologies [43].

Initial studies using transcriptomic and genomic array screens of tissue biopsies have shown promise in defining potential biomarkers and response to personalised chemotherapy (reviewed in e.g. [44-46]). However, these studies require invasive methods to obtain tissue, and currently remain at the experimental phase. For this reason, a focus on non-invasive sampling techniques may be beneficial. One medium of choice is serum, but since proteolytic fragmentation and degradation of biomolecules by endogeneous proteases can occur during or after sample collection [47, 48], this approach poses obstacles that might be difficult to standardise. Population heterogeneity and genetic variation can also have an unexpected impact on the implementation of biomarker screening assays, and therefore there is a need to simplify the screening medium. Urine, which contains approximately 4000 to 5000 proteins [49-51] (compared with at least 10000 in blood [52]), offers a (usually) unlimited quantity of a viable alternative. Urine is also relatively stable in protein/peptide composition and fragmentation state [53, 54], even after prolonged storage [55].

In this study, we have used a combination of mass spectrometric techniques to analyse urine samples from a cohort of uGI cancer patients and non-cancer controls, in order to define peptide fingerprint profiling patterns and then perform independent protein expression profiling. Both sets of results were subsequently combined using

bioinformatics tools to uncover significantly altered potential biomarkers, of which several were then validated by an independent technique. Additionally, we assessed whether there was a correlation of these markers with a plasma marker of inflammation, namely C-reactive protein (CRP). Similar procedures were employed in a previous study [30], where a cation-exchanger was used to clean and enrich urinary proteins and peptides (CM10 chip). This approach, in essence, focussed on medium to high abundance molecules present in human urine. In the present study, we focussed on low-abundance molecules using metal chelate affinity chromatography (IMAC30 [Cu^{2+}-complexed] chip).

2. Materials and methods

The Methods were carried out according to our protocols published previously [30].

2.1. Materials. Buffers, gels and SELDI chips were from Bio-Rad (Hemel Hempstead, UK), and all other chemicals were obtained from Sigma-Aldrich (Gillingham, UK), unless stated in the text.

2.2. Sample collection. Urine samples were obtained from 179 participants (86 upper GI patients and 93 non-cancer controls). Non-cancer controls were normal healthy subjects without any known inflammatory, neoplastic or renal conditions. Participants were aged between 43-83 years for the cancer group and between 19-86 years for controls. One third of the cancer patients had pancreatic cancer; approximately one third had oesophageal cancer; approximately one sixth had cancer of the oesophagogastric junction (OGJ); and approximately one sixth had gastric cancer. Summary participant demographics are shown in Table 1 and full details are provided in Supplementary Table 1. Random morning urine samples were either supplied by the participants of the study (controls) or collected prior to surgery in the operating theatre (cancer patients). All urine samples were stored at -40◦C. Long-term storage of samples (>1 month) was at -80◦C.
All procedures were approved by the local research ethics committee, and the study conformed to the standards set by the Declaration of Helsinki. Written informed consent was obtained.

2.3. SELDI-TOF-MS. IMAC30 SELDI-chips were prepared for sample application according to the manufacturer's recommendations, as previously reported [30]. Briefly, IMAC30 chips were loaded with 0.1 M $CuSO_4$, washed with water, neutralised with 0.1 M NaHAc pH4 and washed with water, followed by two washes with binding buffer (0.1 M $NaHPO_4$, 0.5 M NaCl), and then processed in a bioprocessor-assembly by incubating 0.1 ml urine and 0.1 ml binding buffer for 1 hour at room temperature with vigorous shaking, followed by three washes with 0.2 ml binding buffer for 5 minutes at room temperature with vigorous shaking, and two washes with 0.2 ml water at room temperature with vigorous shaking. The chips were removed from the bioprocessor assembly, air-dried and 2 times 1 μl energy-absorbing matrix (SPA, in 50% ACN, 0.5% TFA) was added. Air-dried chips were analysed in a PCS4000 SELDI-TOF instrument (Bio-Rad, Hemel Hempstead, UK) by measuring the 1000-25000 Da range with a laser setting of 2.5 μJ. Spectra were exported as '.xml' files. The SELDI instrument was calibrated using the ProteinChip All-In-one peptide standard (Bio-Rad, Hemel Hempstead, UK). Source voltage was 25000 V, and detector voltage was 2946 V. Quality control and consistency were ensured by using one random pool of urines on one spot per chip. Spectra of the full analysis (179 cohort samples, 26 quality control samples, 410 spectra in total) were recorded in two large batches to minimize instrument variability and drift. Spectral alignments of all quality controls ensured consistency of all spectra.

2.4. Data processing. ProteinChip Data Manager Software (PCDMS) version 4.1 with integrated Biomarker Wizard cluster analysis (Bio-Rad, Hemel Hempstead, UK) was used for analysis. SELDI-TOF-MS traces were split into control and cancer groups. The

baseline was subtracted from individual m/z traces and profiles were normalised using total ion current, followed by identification of peak clusters using the cluster analysis tool. In the first pass, peaks were selected where the signal to noise (S/N) ratio was >5, and valley depth was >3, and in the second pass, S/N >2 and valley depth >2. The cluster mass window was set to 0.2% of mass. Clustered peaks were only included if they occurred in at least 10% of all spectra. The resulting p-values, ROC areas, average and median m/z values, and intensities of the clustered peaks were exported and saved as '.csv' files and used for model building. Heat-maps using Pearson's correlation and principle component analysis (PCA) plots were calculated to assess global group divisions (i.e. cancer and control). A two-sample t-test was used to compare mean normalized intensities between the case and control groups. The p-value was set at 0.05 to be statistically significant.

2.5. Model building and validation. Clustered peak lists were analysed with the Biomarker Pattern Software (BPS) (Bio-Rad, Hemel Hempstead, UK). m/z versus intensity matrices were analysed using decision tree-analysis, selecting the standard error rule of minimum cost-tree regardless of size, and the method used was Gini. V-fold testing was set to 1000. 60 cancer samples and 60 control samples were randomly chosen and used as the learning and testing dataset. The remainder of 59 samples was used as the validation dataset for blind-testing. Sensitivity was defined as the probability of predicting cancer cases, and the specificity was defined as the probability of predicting control samples.

2.6. Peak isolation and identification. Peaks observed in the IMAC30 chip-type (see Supplementary Table 2) that showed marked expression differences between control and cancer samples, and were branching points in the model (see Table 2) were further investigated. 0.5 ml urine from positive or negative samples in relation to specific peaks was added to 30 μl Cu^{2+}-loaded IMAC30 spin column resin (Bio-Rad, Hemel Hempstead, UK) and 0.75 ml binding buffer (0.1 M $NaHPO_4$, 0.5 M NaCl) and incubated for one hour at room temperature under constant agitation. Unbound material was removed and the resin washed four times with 0.3 ml binding buffer. Bound material was separated by electrophoresis on a 16.5% Tris-Tricine gel (Bio-Rad, Hemel Hempstead, UK), and gel bands in the region of 2-10 kDa were excised after Coomassie staining (BioSafe Coomassie, Bio-Rad, Hemel Hempstead, UK). Positive and negative samples were chosen based on the presence or absence of a specific m/z peak to be identified from SELDI-TOF-MS analysis. Proteins and peptides from gel bands were digested in situ with trypsin, the resulting peptides eluted with ACN, and analysed by LC-MS/MS [30]. Fragmentation spectra were then processed by Xcalibur and BioWorks software (Thermo Fisher Scientific, Loughborough, UK) and submitted to the Mascot search engine (Matrix Science, London, UK) using UniProt/SwissProt as the reference database. Mascot search parameters were: enzyme specificity trypsin; maximum missed cleavage 1; fixed modifications cysteine carbamidomethylation; variable modification methionine oxidation; precursor mass tolerance +/-3Da; fragment ion mass tolerance +/- 0.4 Da. Only Mascot hits with a false discovery rate <0.05 were taken into consideration.

2.7. Mascot-SELDI matrix matching. Observed proteins with at least two peptide matches from the LC-MS/MS analysis were then further analysed by pattern matching based on SELDI-TOF-MS measured expression levels of peaks of interest (expected abundance in selected samples). This was done using software written in-house, which compares observed protein expression patterns in a pre-defined set of samples (LC-MS/MS results) against a matrix of peak patterns (SELDI-TOF clustered peak intensities, where estimated peaks are set to null) in the same set of samples. The scoring is based on sensitivity (percent observed over expected) and specificity (percent not observed over not expected), and results are presented in descending order of cumulative scores. The distribution of identified peptides within a protein, as well as calculated molecular mass of identified proteins, were also used to assess whether breakdown products were likely to account for mass variances between the expected mass and the molecular weight of the full length protein.

2.8. Validation of Mascot search results. Cross-validation of identified peaks was done by Western blotting of raw urine samples (20 μl per well) using standard protocols [30]. Antibodies used were goat-anti-human Rabenosyn-5 (N-20) (SantaCruz Biotechnology, Santa Cruz, CA, U.S.A.) (used at a dilution of 1:200); mouse-anti-human PDCD6IP (ab56932) (AbCam, Cambridge, UK) (used at a dilution of 1:1000); rabbit-anti-human serum albumin (Sigma-Aldrich, Gillingham, UK) (used at a dilution of 1:1000); and peroxidase-coupled secondary antibodies were from Upstate (Lake Placid, NY, U.S.A.) (used at a dilution of 1:5000). Detection of signals was by chemiluminescence using ECL Western blotting reagents (Thermo Fisher Scientific, Cramlington, UK).

2.9 Measurement of plasma CRP. Plasma CRP was assayed using automated methods on an Olympus AU2700 analyser (Olympus Diagnostica GmbH [Irish Branch], Lismeehan, Ireland), in the Department of Clinical Chemistry, Royal Infirmary of Edinburgh (fully accredited by Clinical Pathology Accreditation [UK] Ltd).

3. Results

We collected mass spectra in the m/z range of 1000 to 25000 for the 179 patient urine samples. The full peak list used in PCDMS data analysis is shown in Supplementary Table 2. We could identify 86 cluster peaks (with threshold settings as described in the Materials and Methods section), of which 42 peak clusters showed p-values <0.05, indicating a potentially significant difference in expression levels for a particular protein or peptide cluster. Analysis using peak clustering and group distribution demonstrated that both control and uGI cancer groups shared a general overlap in PCA for both the IMAC30 and CM10 [30] chip-types (see Supplementary Figure 1), but we distinct enough to allow a degree of separation in heat-map plotting using Pearson's correlation.

Decision-tree modelling using the Biomarker Pattern software (BPS) of peak clusters (see Supplementary Table 3 for raw input data and group dividers) of 60 random samples of each cohort was validated by the remainder of the entire cohort (33 control and 26 cancer samples). The decision tree model is shown in Figure 1A, and in more detail in Supplementary Figure 2. The validation data-set showed a sensitivity of 85% and a specificity of 79% with an overall correctness of 81%, and application of the derived model to the entire cohort showed a sensitivity of 88%, specificity of 91% and an overall correctness of 90%. Statistical analysis of the five m/z cluster peaks implicated in the model is shown in Table 2. Two of the candidate targets at m/z 3589 and 13387 displayed down-regulation in cancer, whereas m/z 2577, 5646 and 7477 peak clusters showed a cancer-associated up-regulation. All potential markers were statistically

significant ($p<0.05$). Expression analysis using box-and-whisker-plots of these five lead candidates is shown in Figure 1, panels B to F.

We then evaluated whether we inadvertently biased our model-prediction towards a specific cancer type, and could show that two major cancer types in our cohort (oesophagus and pancreas) were distributed equally (see Supplementary Figure 3). Additionally, we could not detect an age-dependency or age-covariance of any of the peaks of interest. These m/z peaks were then further investigated in order to elucidate their molecular composition. Protein and peptides from 40 urine samples were then enriched by IMAC30 batch chromatography, followed by peptide gel electrophoresis, gel band excisions and tryptic digest. Mascot searching after LC-MS/MS analysis resulted in 646 positively identified proteins in the 2-10 kDa range with peptide counts ≥ 2 each and Mascot scores ≥ 16. This expression pattern was then compared to the expected expression pattern derived from SELDI-TOF-MS measured peak intensity values in the same 40 samples in order to associate the m/z peaks with identified proteins. The scores were calculated as a percentage of the expected pattern in the Mascot-identified protein list compared to the measured pattern of peaks found by SELDI-TOF above base-line (sensitivity), thereby setting all estimated peaks as null values, which were used to calculate the specificity. Table 3 lists all molecules found by this approach.

The m/z 13387 peak cluster expression profile could not be positively matched to proteins and protein fragments found by gel-excision, LC-MS/MS and Mascot searching. This is most likely due to low expression levels in the sample even after IMAC-chromatographic enrichment, which was also evident by the low intensity values observed by SELDI mass spectrometry (Table 2). The expression pattern of the m/z 2577

peak cluster matched the pattern of both glial fibrillary acidic protein and Rabenosyn-5; m/z 3589 matched S100A8; m/z 5646 matched with neuron navigator 1, 2, 3, and programmed cell death 6-interacting protein (PDCD6IP, ALIX); and m/z 7477 matched S100A9 (Table 3). A detailed list of these molecules, including peptide sequences is supplied in Supplementary Table 4. Both Rabenosyn-5 and PDCD6IP were cross-validated by Western blot analysis (Figure 2), and both molecules showed a very good correlation to the expected cluster-peak pattern observed by SELDI analysis, thereby validating both the predictor model and the Mascot identification, as well as the SELDI peak clustering of these molecules.

Furthermore, we evaluated whether plasma levels of C-reactive protein (CRP) correlated with any of the proposed cancer-associated SELDI-TOF m/z peaks. Plotting the CRP levels of 122 patient samples against the measured intensity levels of 9 individual m/z cluster peaks implicated as potential biomarkers shows no correlation between any of the m/z clusters and CRP levels (Figure 3). The same result was obtained previously for m/z cluster peak lead candidates for the CM10 chip-type [30] (Figure 3).

4. Discussion

Urine is arguably the best biological substrate for biomarker development in the clinical setting, as it is relatively stable and easily obtainable in large quantities through non-invasive sampling. Previous studies have found SELDI-TOF-MS to be ideally suited for urine analysis, as it combines speed and high throughput with relatively low cost [56]. The main disadvantage of SELDI-TOF-MS is the medium resolution of the spectra obtained. However, this technique is able to resolve peaks in the 1000-25000 Da range from spectra with less than 500 peaks. We found that the IMAC30 chip, and previously the CM10 chip [30], are both useful chip-types for the analysis of human urine, and we were able to generate models based on the full analysis of 179 uGI cancer and non-cancer samples. Using the tree-analysis method, we established a statistical model with an overall sensitivity of 88% and specificities of 91% across the entire datasets. Using expression pattern matching, we could assign several proteins identified in urine to our proposed biomarkers. These sensitivities and specificities are superior to those exhibited by established serum uGI cancer markers, such as CA 19-9 used in the diagnosis of pancreatic cancer (sensitivity of about 79-81% and a specificity of about 82-90%) [57]. Larger prospective studies are required to investigate whether theses statistical models can be used effectively in clinical practice.

GFAP is a marker for glioblastoma, where serum levels are substantially elevated compared to healthy controls in whom the molecule is practically undetectable [58]. Our observed up-regulation of GFAP in urine from uGI cancer patients would indicate that this protein is not restricted to glioblastoma, but might be a more general marker of malignant disease. We could also detect a down-regulation in uGI cancer samples for

S100A8. It was reported that the serum levels of S100A8 were marginally reduced in cancer patients [59], which is similar to our observed down-regulation of the S100A8 associated m/z 3589 cluster. The presence of members of the S100 gene family, including S100A9, in our screen is not entirely surprising. The up-regulation of this protein in cancer has already been documented in our previous study and was validated using Western blotting [30]. Neuron navigators 1 to 3, which are cytoskeletal regulators that track microtubules, belong to a molecular class that can reorganize the cytoskeleton and induce neurite outgrowth and axonal elongation [60]. Neuron navigator 2 was found to be a retinoic-acid response-element in neuroblastoma [61], and Neuron navigator 3 was reported to be a cancer-associated molecule that contributes to the pathogenesis of some basal and squamous cell cancers [62]. However, based on our peptide sequence identification, it is likely that we observed only one of these three molecules, but due to sequence similarities we cannot distinguish between them in our analysis.

Importantly, we were able to validate two of our potential biomarkers through Western blotting, namely Rabenosyn-5 and PDCD6IP. Rabenosyn-5 (ZFYVE20) is a Rab GTPase that was reported to be involved in endosomal trafficking, including a role in the lysosomal trafficking of cathepsin D from the Golgi apparatus to lysosomes [63]. Programmed cell death 6-interacting protein (PDCD6IP), also termed AIP1 or Alix, has a described role in actin filament bundling and endosomal sorting, and mediates extracellular integrin-mediated cell adhesions [64]. Our measured up-regulation, both by MS and by Western blotting, could be an indicator of impending metastasis within the cancer patients.

It is noteworthy that all of the postulated peak markers in the present study were fragments, unlike the proteins we identified in the CM10 chip-type based cancer dataset [30]. This phenomenon could be due to chromatography-resin properties, where smaller polypeptides were favoured over larger ones, or biochemical/biological reasons, such as fragmentation states of low abundance molecules.

Systemic inflammation (as evidenced by an elevated plasma CRP level [65, 66]) is a common finding in patients with cancer (where it plays a role in the aetiology of cachexia and cancer-associated anorexia), and can be a confounding factor in biomarker discovery [67]. However, others have failed to find such an association [68]. In this study, we could not detect any correlation between biomarkers, cancer types and CRP, a robust serum marker of systemic inflammation.

In conclusion, we could demonstrate several potential urinary uGI cancer markers using mass spectrometry techniques, bioinformatics data processing, and Western blotting. Most of these markers were up-regulated (rather than down-regulated) in the urinary proteome in association with cancer. Further work is required before such biomarkers can be utilised in the clinical setting. Firstly, larger prospective studies, with the possible inclusion of anti-cancer intervention, will help determine biomarker validity. Secondly, the development of rapid, high-throughput assays (such as ELISA) for confirmed biomarkers will aid the day-to-day employment of biomarker measurement in the clinical setting. It is also worth mentioning that a simultaneous screen of several markers, as opposed to decision-making using a single biomarker, will undoubtedly expand the usability, accuracy, sensitivity and specificity in clinical diagnosis of uGI cancers as has been shown in assessing pancreatic cancer [69]. Lastly, further studies to

explain the mechanistic roles of these biomarkers in uGI cancer would improve our understanding of the disease and potentially yield new therapeutic targets.

Acknowledgements

We thank J. Black for technical assistance with Western blotting. Funding of this work was provided by the University of Edinburgh.

Conflict of Interest Statement

The authors have declared no conflict of interest.

Role of authors

HH did the sample preparations, SELDI measurements, data analysis, Mascot searches, software design and coding, and co-wrote the manuscript, NS,CG and HW organised patient recruitment, sampling and clinical analysis, AC did the LC-MS/MS measurements, RS and JR co-wrote the manuscript, JR, KF, HH and RS devised the study and JR and KF supervised the research.

References

1. Ferlay, J., Shin, H. R., Bray, F., Forman, D. *et al.* Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int. J. Cancer* **2010**, 127 (12), 2893-2917.

2. Shahbaz Sarwar, C. M., Luketich, J. D., Landreneau, R. J., Abbas, G., Esophageal cancer: an update. *Int. J. Surg.* **2010**, 8 (6), 417-422.

3. Zhang, Y., Epidemiology of esophageal cancer. *World J. Gastroenterol.* **2013**, 19 (34), 5598-5606.

4. Gungor, C., Hofmann, B. T., Wolters-Eisfeld, G., Bockhorn, M., Pancreatic Cancer. *Br. J. Pharmacol.* **2013**.

5. Hu, Q. D., Zhang, Q., Chen, W., Bai, X. L., Liang, T. B., Human development index is associated with mortality-to-incidence ratios of gastrointestinal cancers. *World J. Gastroenterol.* **2013**, 19 (32), 5261-5270.

6. Rastogi, T., Devesa, S., Mangtani, P., Mathew, A. *et al.* Cancer incidence rates among South Asians in four geographic regions: India, Singapore, UK and US. *Int. J. Epidemiol.* **2008**, 37 (1), 147-160.

7. Clemons, N., Phillips, W., Lord, R. V., Signaling pathways in the molecular pathogenesis of adenocarcinomas of the esophagus and gastresophageal junction. *Cancer Biol. Ther.* **2013**, 14 (9).

8. Matsumoto, S., Takayama, T., Wakatsuki, K., Enomoto, K. *et al.* Predicting early cancer-related deaths after curative esophagectomy for esophageal cancer. *Am. Surg.* **2013**, 79 (5), 528-533.

9. Taylor, M. D., LaPar, D. J., Davis, J. P., Isbell, J. M. *et al.* Induction chemoradiotherapy and surgery for esophageal cancer: survival benefit with downstaging. *Ann. Thorac. Surg.* **2013**, 96 (1), 225-230.

10. Hennies, S., Hermann, R. M., Gaedcke, J., Grade, M. *et al.* Increasing toxicity during neoadjuvant radiochemotherapy as positive prognostic factor for patients with esophageal carcinoma. *Dis. Esophagus.* **2013**.

11. Fry, L. C., Monkemuller, K., Malfertheiner, P., Molecular markers of pancreatic cancer: development and clinical relevance. *Langenbecks Arch. Surg.* **2008**, 393 (6), 883-890.

12. Ballehaninna, U. K., Chamberlain, R. S., Serum CA 19-9 as a Biomarker for Pancreatic Cancer-A Comprehensive Review. *Indian J. Surg. Oncol.* **2011**, 2 (2), 88-100.

13. Singh, S., Tang, S. J., Sreenarasimhaiah, J., Lara, L. F., Siddiqui, A., The clinical utility and limitations of serum carbohydrate antigen (CA19-9) as a diagnostic tool

for pancreatic cancer and cholangiocarcinoma. *Dig. Dis. Sci.* **2011**, 56 (8), 2491-2496.

14. Kim, J. E., Lee, K. T., Lee, J. K., Paik, S. W. *et al.* Clinical usefulness of carbohydrate antigen 19-9 as a screening test for pancreatic cancer in an asymptomatic population. *J. Gastroenterol. Hepatol.* **2004**, 19 (2), 182-186.

15. Zhang, S., Wang, Y. M., Sun, C. D., Lu, Y., Wu, L. Q., Clinical value of serum CA19-9 levels in evaluating resectability of pancreatic carcinoma. *World J. Gastroenterol.* **2008**, 14 (23), 3750-3753.

16. Passerini, R., Cassatella, M. C., Boveri, S., Salvatici, M. *et al.* The pitfalls of CA19-9: routine testing and comparison of two automated immunoassays in a reference oncology center. *Am. J. Clin. Pathol.* **2012**, 138 (2), 281-287.

17. Lee, G., Ge, B., Huang, T. K., Zheng, G. *et al.* Positive identification of CA215 pan cancer biomarker from serum specimens of cancer patients. *Cancer Biomark.* **2010**, 6 (2), 111-117.

18. Lee, C., He, H., Jiang, Y., Di, Y. *et al.* Elevated expression of tumor miR-222 in pancreatic cancer is associated with Ki67 and poor prognosis. *Med. Oncol.* **2013**, 30 (4), 700.

19. Wang, J., Chen, X., Li, P., Su, L. *et al.* CRKL promotes cell proliferation in gastric cancer and is negatively regulated by miR-126. *Chem. Biol. Interact.* **2013**.

20. Guo, P., Zhu, Z., Sun, Z., Wang, Z. *et al.* Expression of legumain correlates with prognosis and metastasis in gastric carcinoma. *PLoS One* **2013**, 8 (9), e73090.

21. Ren, P., Zhang, J. G., Xiu, L., Yu, Z. T., Clinical significance of phospholipase A2 group IIA (PLA2G2A) expression in primary resected esophageal squamous cell carcinoma. *Eur. Rev. Med. Pharmacol. Sci.* **2013**, 17 (6), 752-757.

22. Xie, Z., Chen, G., Zhang, X., Li, D. *et al.* Salivary microRNAs as promising biomarkers for detection of esophageal cancer. *PLoS One* **2013**, 8 (4), e57502.

23. Li, M., Tang, Y., Zang, W., Xuan, X. *et al.* Analysis of HAX-1 gene expression in esophageal squamous cell carcinoma. *Diagn. Pathol.* **2013**, 8, 47.

24. McKiernan, E., McDermott, E. W., Evoy, D., Crown, J., Duffy, M. J., The role of S100 genes in breast cancer progression. *Tumour. Biol.* **2011**, 32 (3), 441-450.

25. Melle, C., Ernst, G., Schimmel, B., Bleul, A., von, E. F., Colon-derived liver metastasis, colorectal carcinoma, and hepatocellular carcinoma can be discriminated by the Ca(2+)-binding proteins S100A6 and S100A11. *PLoS One* **2008**, 3 (12), e3767.

26. Wang, X. H., Zhang, L. H., Zhong, X. Y., Xing, X. F. *et al.* S100A6 overexpression is associated with poor prognosis and is epigenetically up-regulated in gastric cancer. *Am. J. Pathol.* **2010**, 177 (2), 586-597.

27. Hua, Z., Chen, J., Sun, B., Zhao, G. *et al.* Specific expression of osteopontin and S100A6 in hepatocellular carcinoma. *Surgery* **2011**, 149 (6), 783-791.

28. Wei, B. R., Hoover, S. B., Ross, M. M., Zhou, W. *et al.* Serum S100A6 concentration predicts peritoneal tumor burden in mice with epithelial ovarian cancer and is associated with advanced stage in patients. *PLoS One* **2009**, 4 (10), e7670.

29. Sofiadis, A., Dinets, A., Orre, L. M., Branca, R. M. *et al.* Proteomic study of thyroid tumors reveals frequent up-regulation of the Ca2+ -binding protein S100A6 in papillary thyroid carcinoma. *Thyroid* **2010**, 20 (10), 1067-1076.

30. Husi, H., Stephens, N., Cronshaw, A., MacDonald, A. *et al.* Proteomic analysis of urinary upper gastrointestinal cancer markers. *Proteomics Clin. Appl.* **2011**, 5 (5-6), 289-299.

31. van, d. B., I, Sparidans, R. W., van Winden, A. W., Gast, M. C. *et al.* The absolute quantification of eight inter-alpha-trypsin inhibitor heavy chain 4 (ITIH4)-derived peptides in serum from breast cancer patients. *Proteomics Clin. Appl.* **2010**, 4 (12), 931-939.

32. van Winden, A. W., van, d. B., I, Gast, M. C., Engwegen, J. Y. *et al.* Serum degradome markers for the detection of breast cancer. *J. Proteome Res.* **2010**, 9 (8), 3781-3788.

33. Zhang, Z., Bast, R. C., Jr., Yu, Y., Li, J. *et al.* Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. *Cancer Res.* **2004**, 64 (16), 5882-5890.

34. Villanueva, J., Shaffer, D. R., Philip, J., Chaparro, C. A. *et al.* Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J. Clin. Invest* **2006**, 116 (1), 271-284.

35. Jayapalan, J. J., Ng, K. L., Shuib, A. S., Razack, A. H., Hashim, O. H., Urine of patients with early prostate cancer contains lower levels of light chain fragments of inter-alpha-trypsin inhibitor and saposin B but increased expression of an inter-alpha-trypsin inhibitor heavy chain 4 fragment. *Electrophoresis* **2013**, 34 (11), 1663-1669.

36. Song, J., Patel, M., Rosenzweig, C. N., Chan-Li, Y. *et al.* Quantification of fragments of human serum inter-alpha-trypsin inhibitor heavy chain 4 by a surface-enhanced laser desorption/ionization-based immunoassay. *Clin. Chem.* **2006**, 52 (6), 1045-1053.

37. Chung, J. C., Oh, M. J., Choi, S. H., Bae, C. D., Proteomic analysis to identify biomarker proteins in pancreatic ductal adenocarcinoma. *ANZ. J. Surg.* **2008**, 78 (4), 245-251.

38. Zelvyte, I., Wallmark, A., Piitulainen, E., Westin, U., Janciauskiene, S., Increased plasma levels of serine proteinase inhibitors in lung cancer patients. *Anticancer Res.* **2004**, 24 (1), 241-247.

39. Hamm, A., Veeck, J., Bektas, N., Wild, P. J. *et al.* Frequent expression loss of Inter-alpha-trypsin inhibitor heavy chain (ITIH) genes in multiple human solid tumors: a systematic expression analysis. *BMC. Cancer* **2008**, 8, 25.

40. Makawita, S., Dimitromanolakis, A., Soosaipillai, A., Soleas, I. *et al.* Validation of four candidate pancreatic cancer serological biomarkers that improve the performance of CA19.9. *BMC. Cancer* **2013**, 13 (1), 404.

41. Sanchez-Carbayo, M., Antibody microarrays as tools for biomarker discovery. *Methods Mol. Biol.* **2011**, 785, 159-182.

42. Tanzer, M., Liebl, M., Quante, M., Molecular biomarkers in esophageal, gastric, and colorectal adenocarcinoma. *Pharmacol. Ther.* **2013**.

43. Lee, K. H., Galloway, J. F., Park, J., Dvoracek, C. M. *et al.* Quantitative molecular profiling of biomarkers for pancreatic cancer with functionalized quantum dots. *Nanomedicine.* **2012**, 8 (7), 1043-1051.

44. Brennan, D. J., Kelly, C., Rexhepaj, E., Dervan, P. A. *et al.* Contribution of DNA and tissue microarray technology to the identification and validation of biomarkers and personalised medicine in breast cancer. *Cancer Genomics Proteomics* **2007**, 4 (3), 121-134.

45. Baron, A., Moore, P. S., Scarpa, A., DNA array/microarrays in oncological research with focus on pancreatic cancer. *Adv. Clin. Path.* **2001**, 5 (4), 115-120.

46. Honda, K., Ono, M., Shitashige, M., Masuda, M. *et al.* Proteomic approaches to the discovery of cancer biomarkers for early detection and personalized medicine. *Jpn. J. Clin. Oncol.* **2013**, 43 (2), 103-109.

47. Koomen, J. M., Li, D., Xiao, L. C., Liu, T. C. *et al.* Direct tandem mass spectrometry reveals limitations in protein profiling experiments for plasma biomarker discovery. *J. Proteome Res.* **2005**, 4 (3), 972-981.

48. Banks, R. E., Stanley, A. J., Cairns, D. A., Barrett, J. H. *et al.* Influences of blood sample processing on low-molecular-weight proteome identified by surface-enhanced laser desorption/ionization mass spectrometry. *Clin. Chem.* **2005**, 51 (9), 1637-1649.

49. Adachi, J., Kumar, C., Zhang, Y., Olsen, J. V., Mann, M., The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins. *Genome Biol.* **2006**, 7 (9), R80.

50. Gonzales, P. A., Pisitkun, T., Hoffert, J. D., Tchapyjnikov, D. *et al.* Large-scale proteomics and phosphoproteomics of urinary exosomes. *J. Am. Soc. Nephrol.* **2009**, 20 (2), 363-379.

51. Kentsis, A., Monigatti, F., Dorff, K., Campagne, F. *et al.* Urine proteomics for profiling of human disease using high accuracy mass spectrometry. *Proteomics Clin. Appl.* **2009**, 3 (9), 1052-1061.

52. Jacobs, J. M., Adkins, J. N., Qian, W. J., Liu, T. *et al.* Utilizing human blood plasma for proteomic biomarker discovery. *J. Proteome Res.* **2005**, 4 (4), 1073-1085.

53. Lee, R. S., Monigatti, F., Briscoe, A. C., Waldon, Z. *et al.* Optimizing sample handling for urinary proteomics. *J. Proteome Res.* **2008**, 7 (9), 4022-4030.

54. Havanapan, P. O., Thongboonkerd, V., Are protease inhibitors required for gel-based proteomics of kidney and urine? *J. Proteome Res.* **2009**, 8 (6), 3109-3117.

55. Nicol, J. L., Hoy, W. E., Su, Q., Atkins, R. C., Polkinghorne, K. R., Reproducibility of urinary albumin assays by immunonephelometry after long-term storage at -70 degrees C. *Am. J. Kidney Dis.* **2011**, 58 (4), 685-687.

56. Caffrey, R. E., A review of experimental design best practices for proteomics based biomarker discovery: focus on SELDI-TOF. *Methods Mol. Biol.* **2010**, 641, 167-183.

57. Shah, U. A., Saif, M. W., Tumor markers in pancreatic cancer: 2013. *JOP.* **2013**, 14 (4), 318-321.

58. Lyubimova, N. V., Toms, M. G., Popova, E. E., Bondarenko, Y. V. *et al.* Neurospecific proteins in the serum of patients with brain tumors. *Bull. Exp. Biol. Med.* **2011**, 150 (6), 732-734.

59. Driemel, O., Escher, N., Ernst, G., Melle, C., von, E. F., S100A8 cellular distribution in normal epithelium, hyperplasia, dysplasia and squamous cell carcinoma and its concentration in serum. *Anal. Quant. Cytol. Histol.* **2010**, 32 (4), 219-224.

60. van, H. J., Draegestein, K., Keijzer, N., Abrahams, J. P. *et al.* Mammalian Navigators are microtubule plus-end tracking proteins that can reorganize the cytoskeleton to induce neurite-like extensions. *Cell Motil. Cytoskeleton* **2009**, 66 (10), 824-838.

61. Muley, P. D., McNeill, E. M., Marzinke, M. A., Knobel, K. M. *et al.* The atRA-responsive gene neuron navigator 2 functions in neurite outgrowth and axonal elongation. *Dev. Neurobiol.* **2008**, 68 (13), 1441-1453.

62. Maliniemi, P., Carlsson, E., Kaukola, A., Ovaska, K. *et al.* NAV3 copy number changes and target genes in basal and squamous cell cancers. *Exp. Dermatol.* **2011**, 20 (11), 926-931.

63. Navaroli, D. M., Bellve, K. D., Standley, C., Lifshitz, L. M. *et al.* Rabenosyn-5 defines the fate of the transferrin receptor following clathrin-mediated endocytosis. *Proc. Natl. Acad. Sci. U. S. A* **2012**, 109 (8), E471-E480.

64. Pan, S., Wang, R., Zhou, X., Corvera, J. *et al.* Extracellular Alix regulates integrin-mediated cell adhesions and extracellular matrix assembly. *EMBO J.* **2008**, 27 (15), 2077-2090.

65. Gronn, M., Slordahl, S. H., Skrede, S., Lie, S. O., C-reactive protein as an indicator of infection in the immunosuppressed child. *Eur. J. Pediatr.* **1986**, 145 (1-2), 18-21.

66. Toriola, A. T., Grankvist, K., Agborsangaya, C. B., Lukanova, A. *et al.* Changes in pre-diagnostic serum C-reactive protein concentrations and ovarian cancer risk: a longitudinal study. *Ann. Oncol.* **2011**, 22 (8), 1916-1921.

67. Chechlinska, M., Kowalewska, M., Nowak, R., Systemic inflammation as a confounding factor in cancer biomarker discovery and validation. *Nat. Rev. Cancer* **2010**, 10 (1), 2-3.

68. Van, H. M., Jungner, I., Walldius, G., Garmo, H. *et al.* Risk of prostate cancer is not associated with levels of C-reactive protein and other commonly used markers of inflammation. *Int. J. Cancer* **2011**, 129 (6), 1485-1492.

69. Chang, S. T., Zahn, J. M., Horecka, J., Kunz, P. L. *et al.* Identification of a biomarker panel using a multiplex proximity ligation assay improves accuracy of pancreatic cancer diagnosis. *J. Transl. Med.* **2009**, 7, 105.

Table and Figure Legends:

Table 1. Demography of the cohort used in this study. Urine specimens from 179 study participants were analysed (86 cancer patients and 93 non-cancer controls). Data are presented as finite numbers, except for the participants' ages, which are presented as means with standard deviations in parentheses.

Table 2. Expression profiles of potential biomarker peaks identified in decision tree-analysis models and cluster analysis. Peaks of interest were analysed by m/z clustering with a window of 0.2% of mass in 179 samples. Both median and average m/z and intensities with standard errors and coefficients of variation (%CV) are shown. P-values were calculated using the Mann-Whitney test, and the receiver-operating characteristic (ROC) area with cancer as the positive group. The average and median fold change was calculated using intensity values of individual peak clusters from the cancer and control groups. Frequency % describes the likelihood of a particular peak to be found in a specific group by peak-clustering using a S/N of 5. Sensitivity and specificity values were calculated by computational decision-tree modelling using the peak in question as the sole predictor, and the SELDI normalised cut-off values are the associated mass spectrometric intensity values where the reported sensitivity and specificity levels are met. The model score is derived from the tree-analysis modelling of the entire cohort and describes the importance of a peak within the model.

Table 3. List of potential biomarkers. LC-MS/MS and Mascot-search identified protein expression in the 2-10 kDa range were compared to the expected expression pattern found by IMAC30 chip-based SELDI screens in the same 40 urine samples. Mascot-SELDI matrix matching scores were calculated as percentages for sensitivity (protein found in samples which contain the target peak) and specificity (protein not found in samples which do not contain the target peak). Protein ID was the Swiss-Prot identifier. Mascot scores, number of peptides found, emPAI, % sequence coverage, and the expected % sequence coverage, based on molecular mass and the m/z, are listed.

Figure 1. Decision tree-analysis model and expression profiles of potential biomarker peaks using the IMAC30 chip-type. Cluster-peaks implicated in the tree-analysis model (A) stratifying cancer (blue) and control (red) were plotted according to their normalized intensity values (B to F, y-axis) for peak clusters of m/z 2577 (B), 3589 (C), 5646 (D), 7477 (E), 13387 (F).

Figure 2. Validation of identified proteins by Western blotting. Urine samples were separated by 20% SDS-PAGE and analysed by Western blotting using antibodies specific against Rabenosyn-5, PDCD6IP and albumin. Samples were tested initially for the presence of fragments of Rabenosyn-5 and PDCD6IP in the region of the measured m/z (open triangles) as well as full-length molecules or other breakdown products. Reliable signals in the 35 kDa range for Rabenosyn-5, and 30 kDa for PDCD6IP were then further analysed. Validation and confirmation of LC-MS/MS and Mascot results are shown in

the strip-blots in panels A and C, which show the results of 8 urine samples that were used in LC-MS/MS and subsequent Mascot searches, together with the cluster peak intensity matrices derived from the SELDI analysis (underneath the individual blots). Panel B depicts the analysis of 4 random cancer and 4 control samples, and panel D is the analysis of 8 random cancer samples. Samples were selected based on the IMAC30 SELDI analysis for the presence of a peak cluster at m/z 2577 (Rabenosyn-5) and m/z 5646 (PDCD6IP) (cancer samples) or absence of these m/z peaks (healthy controls). ‘e‘ corresponds to an estimated value, where no peak at the m/z point could be detected above the S/N ratio. The density ratio was calculated based on the densitometric measurements of the Western blot signals for the specified molecule compared with the qualitative loading control of albumin, a common urinary molecule. For cancer locus: C = non-cancer control; D = duodenum; O = oesophagus; OGJ = oesophago-gastric junction; P = pancreas; SB = small bowel.

Figure 3. Relationship of predicted cancer biomarker levels and C-reactive protein levels. CRP levels from 86 cancer patients (blue) and 38 non-cancer controls (red) were compared to normalised and clustered SELDI-MS peak intensities of selected potential biomarker peaks from the IMAC30-based dataset (m/z 2577 and 5646) and the CM10-based dataset (m/z 10234 and 10468). Scatter plot of CRP levels (x-axis, logarithmic scale) against normalised intensity values (y-axis) derived from SELDI mass spectrometry using the IMAC30 (top) or CM10 (bottom) chip-type. The left box plot shows the distribution of CRP values within each participant group.

Supplementary Table legends

Supplementary Table 1. Participant demographics. The particpant demographics are shown, together with notations of groupings used in modeling and the mass spectrometry measured clustered peak intensities of selected significant peaks derived from the statistical analysis of both IMAC30 SELDI chip datasets.

Supplementary Table 2. Peak map and statistical analysis of the entire IMAC30 chip dataset. Spectra specific data, such as sample name and normalization factor, and peak specific data, such as measured m/z and clustered m/z, m/z width at half height of a peak, resolution, normalized intensity, S/N ratio, TOF and TOF at half height of the peak, as well as the peak type based on thresholding (pass 1: S/N min. 5, valley depth min. 3; pass 2: S/N min 2, valley depth min. 2; estimation of missing peaks below threshold), are shown.

Supplementary Table 3. IMAC30 clustered peak table. The table contains sample ID and participant group details, which allows the extraction of peak and intensity values necessary to build the learning, testing and validation files used in model generating and scoring.

Supplementary Table 4. LC-MS/MS identified proteins found in human urine after IMAC-enrichment. Proteins are listed by SwissProt accession numbers, protein names, relative molecular mass (Da), and calculated isoelectric point pI. MS specific data

includes the Mascot identification scores; % sequence coverage based on the identified peptides; total number of matched peptides and spectral counts; and the exponentially modified protein abundance index (emPAI) value. Additionally, the discovery matrix where each individual protein was observed in a sample by LC-MS/MS with the number of identified peptides is included.

Supplementary Table 5. Proteins identified by LC-MS/MS analysis as potential upper GI cancer biomarkers. Protein/peptide peaks of interest from 40 individual urine samples were enriched on IMAC30 resin, separated by gel electrophoresis, and the 2-10 kDa region excised. Tryptic fragments were identified by LC-MS/MS followed by database searching. The UniProt/SwissProt accession name, protein name, mass (Da), and classifier of the 8 potential cancer biomarker entries are listed, together with information such as how many times the protein was found, Mascot scores, number of peptides matched, and the exponentially modified protein abundance index (emPAI) value (averaged across all samples the protein was identified in). Also included are PubMed identifiers (PMID), linked to published articles in which the specified protein was identified in other urine samples, and % sequence coverage of the identified peptides within a given protein sequence. Each individual sequence is also included, and the identified peptide stretches are highlighted in red.

Supplementary Figure Legends

Supplementary Figure 1. Cluster analysis of urine samples. Pearson's correlation analysis of clustered m/z peaks are displayed as a heat-map (A). PCA plots were generated using the ProteinChip Data Manager software. PCA analyses were plotted as component 1 against 2 (B) and 1 against 3 (C). Controls are coloured red, and cancer samples blue. Up-regulated clusters in the heat-map are coloured red, and down-regulated m/z peaks green.

Supplementary Figure 2. Decision tree analysis model of potential upper GI cancer biomarker patterns using IMAC30 SELDI chips, and statistical relevance. 60 control and 60 cancer patient samples were analysed by SELDI-TOF-MS.The measured m/z peaks were clustered with a mass window of 0.2% of mass and used to build decision tree-models to define cancer probabilities. The receiver-operating characteristic (ROC) plots of the model from the learning, validation, and total (all) datasets are shown on the top-right. The sensitivity, specificity and overall correctness were calculated for the IMAC30 model for the learning/testing and validation groups, as well as the entire dataset, and are included on top. N are the number of unique samples.

Supplementary Figure 3. Distribution of cancer types used in the statistical modeling of the learning/testing and validation groups. 86 upper GI cancer samples were randomly distributed into learning/testing (60 samples) and validation (26 samples) groups for model building using the BPS software. The percentage distribution of the

cancer-types within each sub-group is displayed as a pie chart. Numbers around the pie charts refer to the finite number of samples of each cancer types.