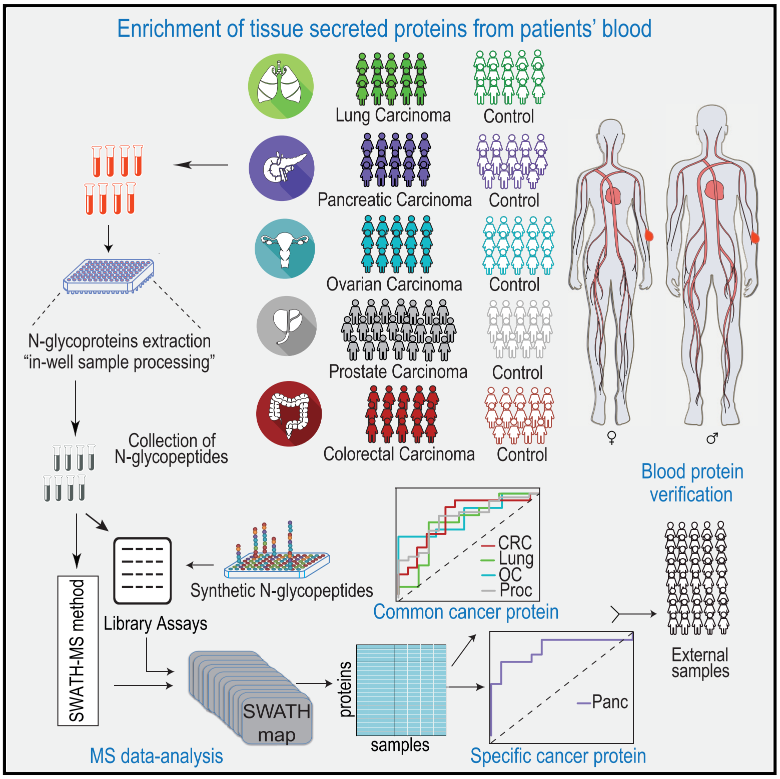
Cell Reports

Resource

Similarities and Differences of Blood N-Glycoproteins in Five Solid Carcinomas at Localized Clinical Stage Analyzed by SWATH-MS

Graphical Abstract



Highlights

- SWATH-MS glycoproteomics enables large-scale plasma analysis across multiple cancers
- Localized-stage carcinomas display common blood changes related to platelet proteins
- Both common and cancer-specific markers stratify subjects with localized cancer
- Data resource of hundreds of cancer-associated N-linked glycoproteins readily available

Authors

Tatjana Sajic, Yansheng Liu,
Eirini Arvaniti, ..., Silke Gillessen,
Manfred Claassen, Ruedi Aebersold

Correspondence

sajic@imsb.biol.ethz.ch (T.S.),
aebersold@imsb.biol.ethz.ch (R.A.)

In Brief

Sajic et al. perform a multi-tumor plasma proteomic study in which they enrich and analyze tissue-secreted plasma glycoproteins to examine blood protein changes in early-stage localized cancers. They demonstrate that many proteins secreted upon platelet activation are changed in several tissue carcinomas, whereas others have changes specific to a single carcinoma type.

Sajic et al., 2018, Cell Reports *23*, 2819–2831
May 29, 2018 © 2018 The Author(s).
https://doi.org/10.1016/j.celrep.2018.04.114

CellPress

Cell Reports
Resource

Similarities and Differences of Blood N-Glycoproteins in Five Solid Carcinomas at Localized Clinical Stage Analyzed by SWATH-MS

Tatjana Sajic,[1,15,16,*] Yansheng Liu,[2,15] Eirini Arvaniti,[1,14] Silvia Surinova,[3] Evan G. Williams,[1] Ralph Schiess,[4] Ruth Hüttenhain,[5] Atul Sethi,[6] Sheng Pan,[7] Teresa A. Brentnall,[8] Ru Chen,[8] Peter Blattmann,[1] Betty Friedrich,[1,14] Emma Niméus,[9] Susanne Malander,[10] Aurelius Omlin,[11,12] Silke Gillessen,[11,12] Manfred Claassen,[1] and Ruedi Aebersold[1,13,*]

[1]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, 8093 Zurich, Switzerland
[2]Department of Pharmacology, Cancer Biology Institute, Yale University School of Medicine, West Haven, CT 06516, USA
[3]UCL Cancer Institute, University College London, London, UK
[4]ProteoMediX AG, 8952 Schlieren, Switzerland
[5]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, CA 94158, USA
[6]Department of Biomedicine, University of Basel/University Hospital Basel, and Swiss Institute of Bioinformatics, Basel, Switzerland
[7]The Brown Foundation Institute of Molecular Medicine, The University of Texas Health Science Center at Houston, 1825 Pressler, Houston, TX 77030, USA
[8]Department of Medicine, University of Washington, Seattle, WA 98195, USA
[9]Department of Clinical Sciences Lund, Surgery, Oncology and Pathology, Lund University, and Skåne University Hospital, Department of Surgery, Lund, Sweden
[10]Department of Clinical Sciences Lund, Oncology and Pathology, Lund University, and Skåne University Hospital, Department of Oncology, Lund, Sweden
[11]Department of Oncology and Hematology, Kantonsspital St. Gallen, St. Gallen, Switzerland
[12]Department of Medical Oncology, Inselspital, Bern University Hospital, University of Bern, 3010 Bern, Switzerland
[13]Faculty of Science, University of Zurich, 8057 Zurich, Switzerland
[14]PhD Program in Systems Biology, University of Zurich and ETH Zurich, Zurich, Switzerland
[15]These authors contributed equally
[16]Lead Contact
*Correspondence: sajic@imsb.biol.ethz.ch (T.S.), aebersold@imsb.biol.ethz.ch (R.A.)

https://doi.org/10.1016/j.celrep.2018.04.114

SUMMARY

Cancer is mostly incurable when diagnosed at a metastatic stage, making its early detection via blood proteins of immense clinical interest. Proteomic changes in tumor tissue may lead to changes detectable in the protein composition of circulating blood plasma. Using a proteomic workflow combining N-glycosite enrichment and SWATH mass spectrometry, we generate a data resource of 284 blood samples derived from patients with different types of localized-stage carcinomas and from matched controls. We observe whether the changes in the patient's plasma are specific to a particular carcinoma or represent a generic signature of proteins modified uniformly in a common, systemic response to many cancers. A quantitative comparison of the resulting N-glycosite profiles discovers that proteins related to blood platelets are common to several cancers (e.g., THBS1), whereas others are highly cancer-type specific. Available proteomics data, including a SWATH library to study N-glycoproteins, will facilitate follow-up biomarker research into early cancer detection.

INTRODUCTION

Carcinomas, the most common types of cancer, develop from epithelial cells in a wide range of tissues. When detected at an advanced stage, carcinomas usually have a poor prognosis, making their early detection (e.g., via protein biomarkers) a clinically important priority. Compared with tissue sections and other clinical diagnostic approaches, early cancer detection's primary interest lies in human blood because blood plasma is easily accessible and hypothesized to contain proteins secreted or shed from tissues that reflect homeostatic changes associated with most cancers (Cima et al., 2011; Surinova et al., 2015).

Many malignancies share dysregulations in a range of molecular pathways that lead to similar systematic disorders and common responses to therapy (Hanahan and Weinberg, 2011). Historically, most biomarker discovery studies reported a list or signature of molecules significantly changed between cohorts of control samples and the specific carcinoma in question. Only in October 2012 did The Cancer Genome Atlas (TCGA) Research Network begin systematically analyzing commonalities in perturbed genomic profiles across different tissue tumors (Weinstein et al., 2013). Since then, several large-scale studies have analyzed clinical samples representing various cancer types by genomic techniques (Hoadley et al., 2014; Kandoth et al., 2013; Uhlen et al., 2017), resulting in the illustration of genetic commonalities, differences, and emergent themes across various tumor lineages.

In recent years, significant technical improvements have advanced mass spectrometry (MS) to a stage where large numbers of proteins can be quantified reproducibly and accurately across large sample cohorts (Guo et al., 2015; Liu et al., 2015), meaning that proteomic techniques have matured to a level where challenging biological and clinical questions can now be addressed.

To date, plasma biomarker projects have aimed principally at detecting different-abundance proteins in a cohort of case and control samples collected within the same tumor entity. The question as to what extent detected protein markers are specific to a particular type of cancer or shared between different cancers remains generally unanswered. No proteomic investigation has looked for plasma biomarkers for different tissue cancers within the same study. To address these open questions, we analyzed blood samples of patients with one of five carcinoma types: colorectal cancer (CRC), pancreatic cancer (Panc), lung cancer (Lung), prostate cancer (Proc), and ovarian cancer (OC), samples that were collected from patients whose tumor was still localized. From each sample, we selectively isolated N-glycosylated peptides to increase the coverage of low-abundant tissue-secreted proteins via the reduction of sample complexity. Analyzing the resulting peptides in their de-glycosylated form (Zhang et al., 2003) by reproducible sequential window acquisition of all theoretical mass spectra (SWATH)-MS (Gillet et al., 2012), we generated a digital representation of each plasma or serum sample that could be queried for the presence and quantity of specific peptides using a targeted data analysis (Röst et al., 2014).

The results of this cross-tumor study at the plasma proteomic level insightfully reveal that early carcinomas display ‘‘specific biomarkers’’ for individual carcinoma types as well as ‘‘common biomarkers’’ of cancer-related blood changes and that the majority of common carcinoma markers detected were angiogenesis-regulatory proteins released from activated blood platelets. These markers appear to be ‘‘sensitive’’ to early carcinoma existence, whereas specific markers were demonstrably associated with a range of reported oncogenes. Our study indicates that the state-of-the-art blood proteomics of several tissue carcinomas at a time can now empirically prove that a certain fraction of the blood proteome follows very similar expression changes in cancer patients. A data resource of hundreds of N-glycoproteins consistently measured in the blood of subjects with various solid tumors and a spectral library of thousands of N-linked glycopeptides for measuring cancer-relevant glycoproteomes are now fully available to the cancer research community.

RESULTS

Changes in Blood Plasma Proteins Associated with Five Localized-Stage Carcinomas

The precise definition of and distinctions between tumor stages is a particular challenge for cross-tumor studies because tumor progression and the stage at which tumors are diagnosed differ notably between different organ or tissue malignancies (McPhail et al., 2015; Siegel et al., 2012). For instance, more than 50% of prostate carcinomas are diagnosed as localized disease. In contrast, pancreatic adenocarcinoma, often characterized by an asymptomatic early stage (Yachida et al., 2010), is incidentally discovered at an early stage in only around 10% of cases (Majumder et al., 2015).

To minimize the confounding effects of advanced stage IV tumors (metastatic disease outside of regional lymph nodes), blood samples collected for this study were derived from patients with localized stages of CRC, Lung, Proc, Panc, and OC (Table S1). All cancer samples were consecutively recruited as diagnosed, and patients did not receive any type of treatment. For each particular cancer type, the control samples were collected together with cancer samples in the same local hospital and were matched with respect to age and gender for each individual cohort (Tables S1); i.e., each cancer type had its own specific matched control group. This precludes the control and cancer samples from being processed arbitrarily (i.e., different protocols, sample storage, etc.) (STAR Methods; Table S5). The controls consisted of healthy individuals with no known history of cancer. In particular, the Proc control cohort included male individuals with benign prostate disease. To extend the MS detection to lower-abundance plasma proteins, we selectively enriched and analyzed N-glycosylated peptides by their de-glycosylated form. N-glycosylated proteins are frequently released or secreted from tissue into the circulation (Schiess et al., 2009; Surinova et al., 2015), and it has been demonstrated that, in the MS analysis of such enriched samples, proteins can be robustly quantified to the low nanogram per milliliter concentration range (Liu et al., 2013). We prepared the samples in large scale by solid-phase extraction of N-glycoproteins and by multi-well plate-based sample processing, as described previously (Cima et al., 2011; Surinova et al., 2015). In the discovery phase (cross-tumor analysis), we processed 162 blood samples, of which 155 were individual samples and 7 were experimental replicates of randomly chosen cancer or control samples from among the five cohorts. In two independent validation cohorts, we measured an additional 129 samples and 3 experimental replicates. Overall, 294 samples from 284 distinct subjects and 10 experimental replicates were prepared and analyzed (Figure 1A).

Broad Coverage and Quantification of Blood Plasma N-Glycoproteins by Targeted MS

To consistently quantify the same N-glycoproteins across all individuals, we utilized large-scale, reproducible sample preparation followed by SWATH-MS. SWATH-MS is a next-generation, massively parallel targeting technique for proteomics measurements. The identified peptides are quantified by label-free quantification. This approach consists of a data-independent acquisition (DIA) strategy where fragment ion spectra of all peptides contained in a sample are recorded, followed by *in silico* data analysis, where sets of targeted peptides are assigned to protein signatures and quantified (Figure 1A; Guo et al., 2015). To search the thus acquired MS data, we established, to date, the largest spectral library of plasma N-glycopeptides by fractionated shotgun analysis, containing high-quality MS2 information for 4,347 N-glycosylated peptides corresponding to 1,151 plasma glycoproteins. This library essentially maximizes the detection coverage of the cancer plasma glycoproteomes (Figure 1A).

Using the configured N-glycoprotein spectral library and OpenSWATH workflow (Röst et al., 2014; Figure 1A), we identified 1,444 distinct N-glycopeptides (272 UniProt-annotated

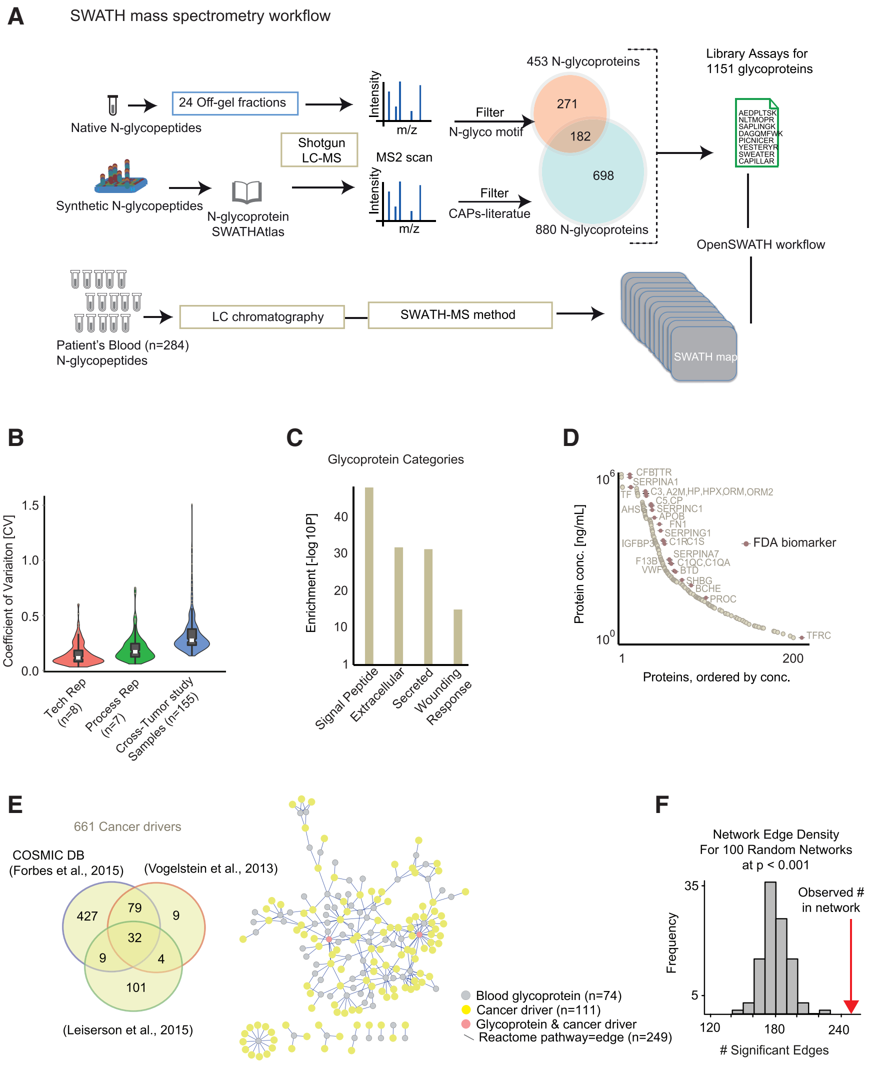


Figure 1. Consistent N-Glycoprotein Profiling in 284 Blood Samples

(A) Schematic representation of the SWATH-MS workflow and SWATH assay library generation from native and synthetic glycopeptides of cancer-associated proteins (CAPs).

(B) Protein CV violin plots corresponding to all clinical samples (blue) and technical (red) and whole-process replicates (green). n corresponds to the number of samples in the cross-tumor dataset.

(C) DAVID functional annotation (FDR < 0.05) of measured glycoproteins. The plasma proteome reference set (Farrah et al., 2011) is used as the background proteome for enrichment analysis.

(D) Protein concentrations are estimated based on the plasma proteome reference set. Pink circles represent FDA-approved biomarkers.

(E) Reactome network view of 74 blood glycoproteins significantly connected to 111 oncogene drivers.

(F) Distribution of the number of edges created using a "switching algorithm" and 100 random networks in the Reactome. The red arrow marks the edge number observed between 74 glycoproteins and 111 cancer drivers—well above that randomly expected.

See also Figure S1.

glycoproteins) at a controlled false discovery rate (FDR) of 1% in initial cross-tumor analysis. To perform objective proteomics comparison between five carcinomas, we filtered these data to select the 1,360 glycopeptides that were quantified in at least two-thirds of the clinical samples, which collapsed into 203 glycoproteins (Table S6). Of these reported values, 88.4% were directly determined from the corresponding signal intensities. The remaining 11.6% were aligned and requantified with the transfer of identification confidence (TRIC) algorithm (Röst et al., 2016). After data acquisition, systematic batch artifacts were corrected at the level of high-quality peptide fragment ion signals (STAR Methods). The reproducibility was nearly perfect for the MS technical replicates (square root of coefficient of determination [R] = 0.96–0.98) and was also excellent between experimental replicates (R = 0.92–0.94) (Figures S1A and S1B).

Specifically, the calculated coefficients of variation (CVs) of protein abundance between all individual biological samples was, on average, 35%, whereas the corresponding values were 16% and 19% for technical and experimental replicates, respectively (Figure 1B), suggesting that biological variation can be revealed by our method. DAVID functional enrichment analysis (Huang et al., 2009) of the 203 glycoproteins confirmed that secreted, extracellular proteins with signal peptides were significantly enriched (Figure 1C). The estimated dynamic range of protein abundance covered ~7 orders of magnitude (Figure 1D). Notably, among our 203 glycoproteins, 30 have already been used as plasma biomarkers for various diseases as approved by the Food and Drug Administration (FDA) (Anderson, 2010; Figure 1D). Taken together, these analyses indicate that these data provide a high-quality, comprehensive dataset for cross-tumor analysis of human plasma/serum proteins.

Circulating Blood Glycoproteins Linked to Diverse Cancer Drivers

To further explore whether the consequences of common genomic lesions across different tumors were apparent in the generated protein profiles, we explored the relationship between the list of known cancer genes and the measured blood glycoproteins using the Reactome signaling pathway database (Liu et al., 2014; Matthews et al., 2009). Specifically, we asked whether the regulated glycoproteins identified in this study were significantly associated with cancer drivers based on their functional implication in common signaling pathways documented in Reactome. We therefore combined information from the Catalogue of Somatic Mutations in Cancer (COSMIC)

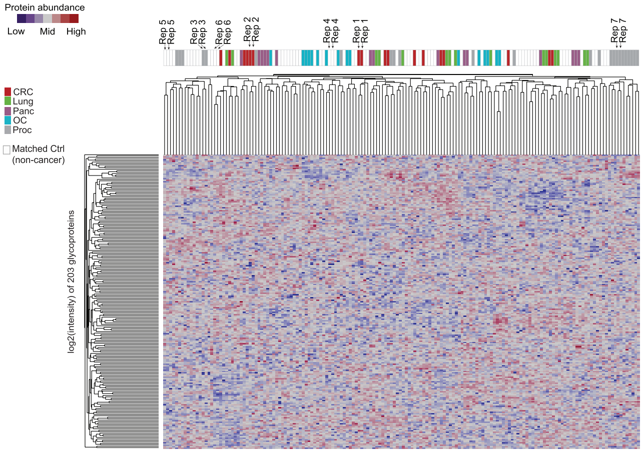


Figure 2. Unsupervised Clustering Based on 203 Glycoproteins in the Cross-Tumor Dataset

Arrows mark the paired samples of seven experimental replicates (Rep1–7) across 155 subjects. For further details, see STAR Methods.

census database (Forbes et al., 2015), the Vogelstein oncogene list (Vogelstein et al., 2013), and the recently published HotNet2 pan-cancer (Leiserson et al., 2015; Figure 1E). This analysis yielded 661 genes as reported cancer drivers (Figure 1E). We found that 74 glycoproteins quantified in our study were strongly interconnected with 118 of 661 cancer drivers in the functional interaction network (249 interaction network edges; Figure 1E, right; Table S7). The degree of the interconnection is significantly higher than expected by chance in random networks ($p < 0.001$; Figure 1F). The results therefore establish the datasets as a valuable resource to link blood glycoprotein abundance changes to different types of tissue cancers.

Global N-Glycoprotein Expression Patterns of Five Carcinomas

To explore the variation in N-glycoprotein patterns between the different groups of cancer patients in a cross-tumor study, we used hierarchical cluster analysis to group the samples according to the quantitative protein profiles. As an initial quality check, we performed unsupervised clustering and observed that, among all 162 measured blood samples, the seven paired samples for whole-process replicates (i.e., 7 experimental replicates) always clustered directly adjacent to one another, indicating high experimental reproducibility (Figure 2). The patterns generated from the 162 blood samples revealed two main clusters and several small sub-clusters. Interestingly, neither the main clusters nor the sub-clusters of blood samples were driven by tumor tissue type, and only modest clustering was observed for Proc (Figure 2). Neither the cancer samples nor their controls clustered according to original cohort or batch processing, thus indicating both significant individual biological variations and tumor tissue heterogeneity. Overall, the data indicate that hierarchical clustering based on the obtained plasma proteome profiles of the tested cancer cases does not distinguish tumor tissue origin at the early, localized disease stage.

Blood Protein Expression across Five Carcinoma Cohorts

We next perceived protein significance and fold change (FC) obtained by MSstat analysis (log2-transformed FC; Table S6; Choi et al., 2014) for each glycoprotein between all cancer types and their respective controls ($FC_{Cancer/Ctrl}$). The $FC_{Cancer/Ctrl}$ distributions indicate that most blood glycoproteins harbor small expression changes in early carcinomas (Figure 3A). To assess the FC variation related to the total procedure, we spiked to each blood sample an equal amount of the bovine N-glycoprotein fetuin-B before glycoprotein enrichment, and, based on its successive measurements, we confirmed a minimal method variation (Figure S1C). The maximal FCs observed were 2.7 and 2.5 (i.e., original-scale FCs) for the proteins Angiopoietin-related protein 7 (ANGPTL7) and Thrombospondin-1 (THBS1) in Proc and CRC respectively, whereas approximately 90% of the FC data were within the range of ±1.2-fold (i.e., corresponding to $\log_2 FC = |0.26|$; Figure 3A). Consequently, we defined differential expression as the subset of proteins that changed with a FC above $|1.2|$ and statistical cutoff of the nominal p value below 0.05. This permissive statistical cutoff in the initial analysis was used to maximize the detection of common glycoprotein changes between five independent tissue cancers. Several proteins were shared between multiple cancer cohorts, particularly THBS1, which was differentially regulated in all tissue carcinomas except in the pancreas (Figures 3B–3F; Table S1).

Among the group of 74 glycoproteins that we found to be directly interconnected with 118 cancer drivers (Figure 1E), 25 proteins were significantly differentially regulated in at least one of the carcinomas (Table S7). We highlighted these proteins and their respective cancer drivers in the network for each cancer (Figures 3G–3K). These plots allowed us to visually compare the "driver hubs" that emerged in each of the five functional networks. Comparative analysis of the perturbed networks and their consequences on glycoprotein abundance showed that three known oncogenes—TP53, the gene with the highest mutation frequency across tumors (Petitjean et al., 2007); the key tumor suppressor gene adenomatous polyposis coli (APC; Aoki and Taketo, 2007); and a tumor progression cell surface proteoglycan, syndecan 4 (SDC4; Beauvais and Rapraeger, 2004)—were found to be emphasized in all five networks of regulated plasma proteins (Figures 3G–3K). One of the known mutated genes in Panc, SMAD family member 4 (SMAD4), also known as deleted in Panc, locus 4 (DPC4) (Majumder et al., 2015), was interconnected with apolipoprotein B (APOB), which significantly changed in pancreatic and CRC blood (Figure 3G). In another case, the known Lung driver PIK3CA (Scheffler et al., 2015) was highlighted because of its connection to interleukin 1 receptor accessory protein (ILI1RAP), upregulated in the plasma of Lung (Figures 3H and S2A). These results suggest that different oncogenic mutations might influence the blood protein abundance of functionally related proteins as an outcome window demonstrating various cancer diseases.

Shared Glycoprotein Signatures across Carcinomas Reveal a Common Role for Platelets

We hypothesize that two types of blood protein changes exist in the plasma/serum of tumor patients—proteins that are specific

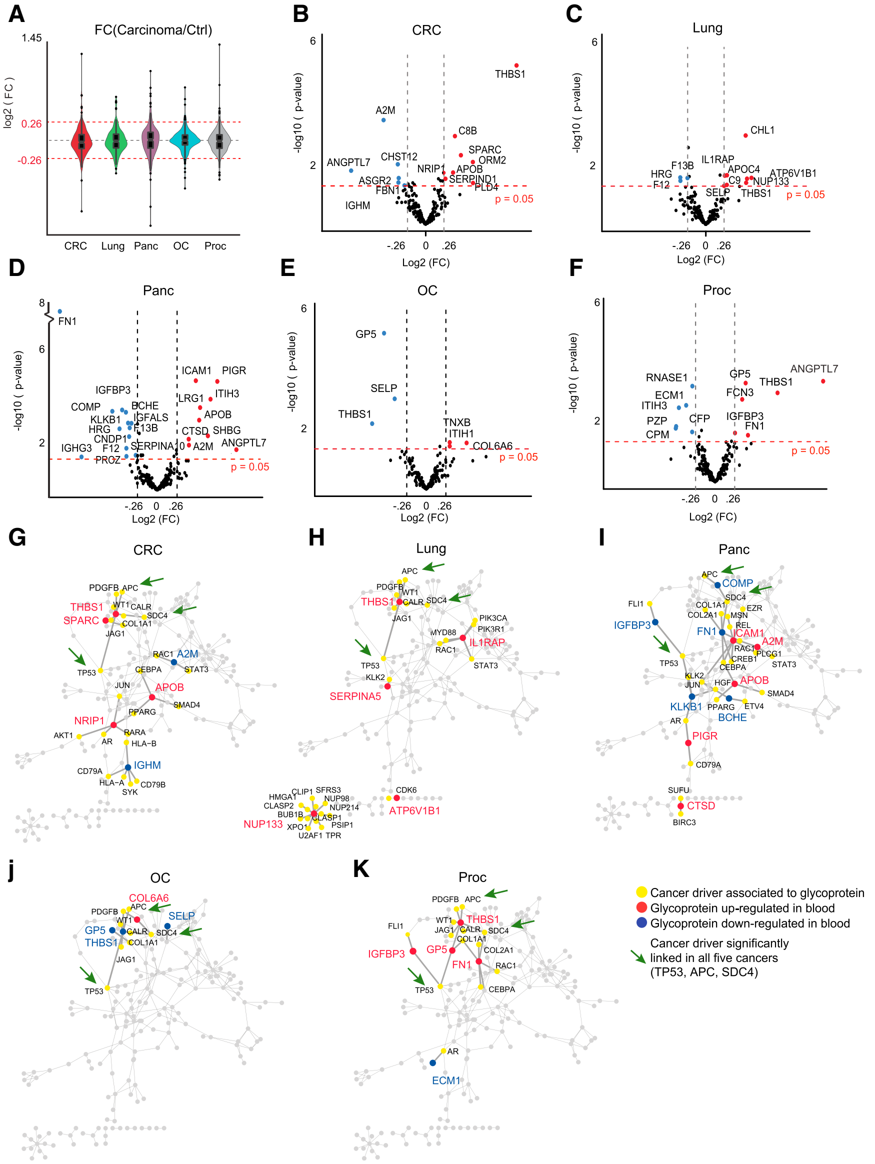


Figure 3. Variation and Functional Links between Glycoproteins and Oncogenes

(A) Violin plots representing distributions of all calculated glycoprotein $FC_{Cancer/Ctrl}$ across five cancer types.

(B–F) Five volcano plots corresponding to five carcinomas (n = 29–36 subjects per cohort)–colorectal cancer (CRC; B), lung cancer (Lung; C), pancreatic cancer (Panc; D), ovarian cancer (OC; E), and prostate cancer (Proc; F)–visualized with differentially expressed proteins ($p < 0.05$); downregulated proteins are presented as blue circles and upregulated ones as red circles (FC > |±1.2|).

(G–K) Functional relation networks of 118 cancer drivers and 74 glycoproteins for all five carcinomas: CRC (G), Lung (H), Panc (I), OC (J), and Proc (K). Glycoproteins are highlighted as up- or downregulated (red or blue circles, respectively) when significantly changed in cancer cohorts. Cancer drivers related to blood-regulated proteins in cancer cohorts are highlighted in yellow. Green arrows indicate common cancer drivers: p53, SDC4, and APC.

to a certain cancer and those common to many malignancies. To test this hypothesis, we compared the lists of differentially abundant proteins for the five cancer cohorts (Figure 4A). The protein THBS1 significantly changed expression in the blood of four of five carcinomas—CRC, Lung, Proc, and OC (Figures 4A and 4B). Somewhat surprisingly, in ovarian carcinoma, THBS1 was downregulated (Figure 4B), indicating that common blood proteins modulated across carcinomas can nevertheless have different responses depending on the tissue of origin. The same results were observed for other estrogen-positive tumors, such as breast cancer (Suh et al., 2012). ANGPTL7 showed expression changes in three of the cancer cohorts (the CRC, pancreatic, and prostate cohorts; Figures 4A and S2B), whereas 10 further glycoproteins were significantly altered in at least two of the five cancers (Table S8).

Intriguingly, three of the top pathways that appear as most relevant for the protein set significantly altered in more than one carcinoma type were associated with blood platelets; 9 of 12 proteins were involved in platelet activation, signaling, and aggregation (Figure 4C; Table S8) or were secreted from activated platelets (Wijten et al., 2013; Figures 4A and S2C). Among these is THBS1, a protein differentially regulated in multiple cancers (Figure 4B). Previous findings suggest that proteins that covary in terms of expression levels over different conditions tend to be involved in the same biological modules (Foster et al., 2006). Interestingly, among 9 common proteins associated with platelets, 4 positively and significantly correlate with THBS1 (Spearman's rank correlation coefficients (rho) ≥ 0.3, $p < 0.01$) in cancer patients (Figure S2C). The four proteins with the highest positive correlation with THBS1 expression (rho > 0.4, $p < 1.0e-4$) in the cancer blood were platelet glycoprotein V (GP5), tissue inhibitor of metalloproteinases 1 (TIMP1), multimerin 1 (MMRN1), and osteonectin (SPARC) (Figure 4D). The plasma levels of these proteins followed regulatory patterns similar to THBS1 for all five cancer cohorts, including Panc and OC (Figure 4E). Also, remarkably, these proteins are directly linked to platelet function or structure (Table S8) and their abundance levels in the plasma of cancer patients were also correlated (Figure 4D, left). In contrast to THBS1, the common cancer protein ANGPTL7 is not related to platelet function and did not show any correlation with other platelet proteins in terms of their abundance levels (Figure 4D, right). This suggests that altered platelet

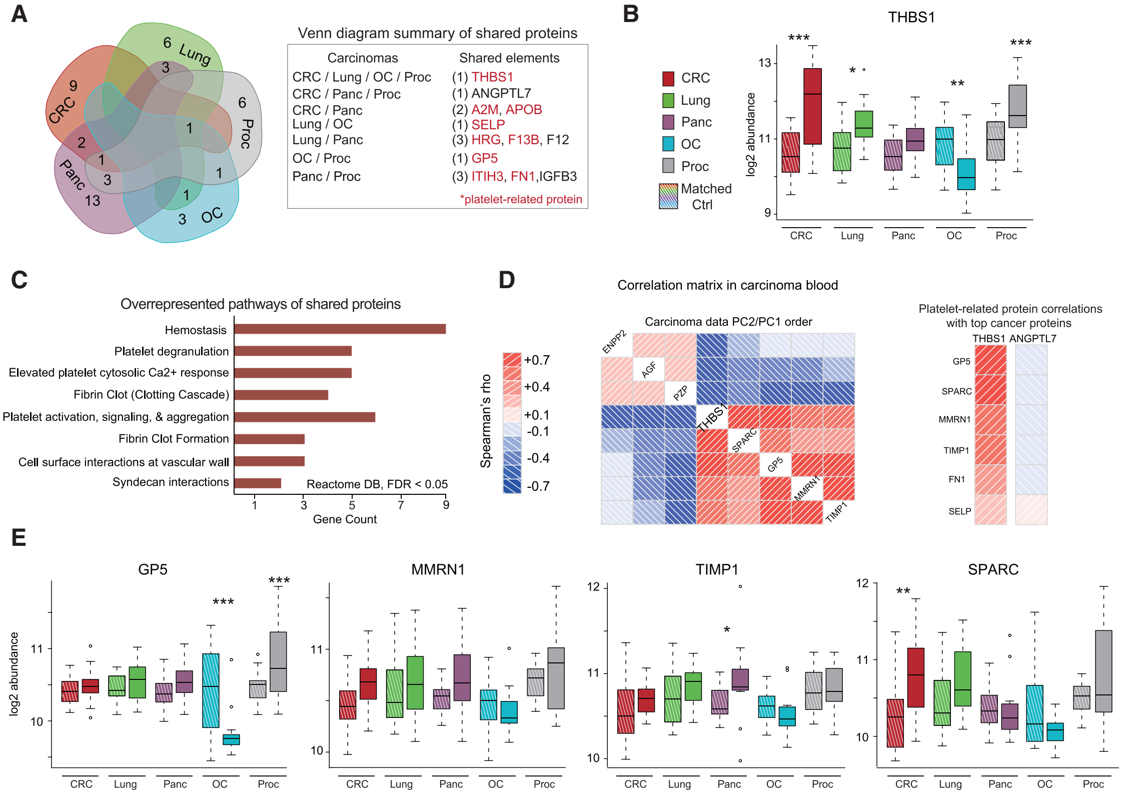


Figure 4. Blood Protein Changes in Five Carcinomas

(A) Venn diagram and its summary table of shared significantly regulated proteins between cancers.

(B) THBS1 $\log_2$ protein abundance in five cancer cohorts and their respective controls (p value cutoffs: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$). THBS1 measurements in five non-cancerous groups were remarkably stable.

(C) Overrepresented biological processes (FDR < 0.05) in the shared proteins set revealed by statistical analysis based on the Reactome pathway database (Matthews et al., 2009).

(D) Proteins that correlate with THBS1 in cancer samples based on Spearman's rank correlation coefficients (rho) (criteria: Spearman rho > |0.4|, $p < 0.0001$) calculated between THBS1 concentrations and all other quantified glycoproteins. Correlations between the top two cancer proteins and platelet-related proteins are shown by color and intensity of shading as presented in the legend.

(E) Log_2 protein abundance of four proteins that positively covary with THBS1 (rho > 0.4, $p < 1.0e{-}05$). Boxplots indicating median line (for n = 14–22 subjects) and the maximum and minimum of the given data (upper and lower hinges, respectively).

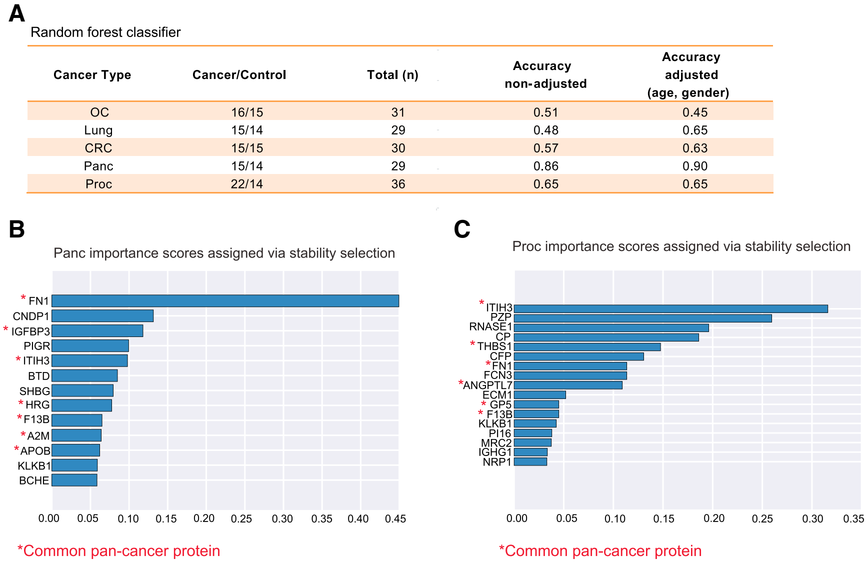
See also Figure S2.

protein expression, exemplified by THBS1, a major blood regulator of angiogenesis (Jiménez et al., 2000), is a common feature in cancerous states.

Random Forest-Based Classification Analysis of Cancer Types in the Cross-Tumor Dataset

Next we sought to define predictive blood signatures for each individual cancer type within the cross-tumor dataset. As a first step, we performed random forest-based classification analysis (Breiman, 2001), treating each carcinoma type separately; i.e., patient samples from a specific cancer type were compared with the corresponding controls. To account for the demographic characteristics, the random forest analysis was repeated, including covariates such as age and gender, where applicable (adjusted model). Although the incidence of other confounders (e.g., smoking, obesity) can conceivably be found sporadically in both the control and cancer subjects, we were unable to test these additional confounders in a more systematic way. Based on random forest analysis, the highest prediction accuracy was demonstrated in cases of pancreatic and prostate carcinoma (86% and 65% out-of-bag accuracy, respectively; non-adjusted model; Figure 5A). The results of this adjusted analysis for the pancreatic and prostate cohorts were very similar to the non-adjusted analysis results (Figure 5A). By contrast, colorectal, lung, and ovarian carcinomas could not be sufficiently separated from their respective controls at this stage (out-of-bag accuracies = 57%, 48%, and 51%, respectively; Figure 5A), indicating the relative difficulty of early disease prediction using the plasma proteome for these cancer types, a difficulty that could be ascribed to the larger individual variation of the cancers in question.

To further examine potential cancer-related functions of blood proteins identified in our cross-tumor study, we performed feature selection via stability selection analysis (Meinshausen



Random forest classifier

Cancer Type	Cancer/Control	Total (n)	Accuracy non-adjusted	Accuracy adjusted (age, gender)
OC	16/15	31	0.51	0.45
Lung	15/14	29	0.48	0.65
CRC	15/15	30	0.57	0.63
Panc	15/14	29	0.86	0.90
Proc	22/14	36	0.65	0.65

Figure 5. Random Forest-Based Classification Analysis of Cancer Types in the Cross-Tumor Dataset

(A) "Out-of-bag" accuracy estimates for non-adjusted and age- and gender-adjusted models of random forest classifiers provided. N corresponds to the number of subjects per cancer cohort.

(B and C) List of top-scoring proteins generated via stability selection analysis for pancreatic (B) and prostate (C) carcinoma cohorts. Prioritized as top-scoring predictors are all proteins with importance scores above a threshold, set as 1/10 of the highest importance score assigned within each cohort.

See also Figure S3 for colorectal, lung, and ovarian carcinoma cohorts.

and Buhlmann, 2010), including age and gender as covariates. The importance scores of top predictor proteins are shown for well-separated Panc and Proc (Figures 5B and 5C) and for the other three cancer types (Figures S3A–S3C). A high importance score was assigned to age (covariate = "years of life") in 4 of the 5 cancer cohorts in initial cross-tumor analysis (Figure S3D).We observed that roughly half of the proteins selected via stability selection analysis are common to all cancer types (Figures 5 and S3), whereas the other half are cancer type-exclusive. The primary shared proteins include fibronectin (FN1), inter-alpha-trypsin inhibitor heavy chain H3 (ITIH3), THBS1, and ANGPTL7. By contrast, carnosine dipeptidase 1 (CNDP1), polymeric immunoglobulin receptor (PIGR), and butyrylcholinesterase (BCHE) in pancreas and RNase A (RNASE1), pregnancy zone protein (PZP), and peptidase inhibitor 16 (PI16) in Proc are examples of tissue-specific markers (Figure 5C).

Common and Specific Blood Proteins in Cancer

To take advantage of the cross-tumor study design, we visualized and compared the individual stratification ability via receiver operating characteristics (ROC) analysis—respective areas under the ROC curves (AUC)—for FN1 and THBS1 as examples of shared cancer proteins and PIGR and PZP as examples of tissue-specific markers (Panc and Proc; Figures 6A–6D). The boxplots of PIGR and FN1 for Panc and PZP and THBS1 for Proc illustrate the protein distribution against matched controls (Figures 6B and 6D; see also Figure 4B for THBS1). THBS1, as mentioned above, changed expression in 4 of 5 cancers and provided almost equally good separation in Proc, OC, CRC, and Lung carcinomas (AUC values; Figures 6C), whereas the pancreatic candidate marker FN1 was equally capable to distinguish lung and prostate carcinomas (Figure 6A). In contrast, PIGR changed expression solely in the pancreas (Figure 6B, right) and yielded a strong stratification capability for Panc exclusively (Figure 6A, right). PZP, a hormone-sensitive protein, demonstrated significant downregulation in the serum of Proc patients and stratified them reasonably well against benign prostate disease (Figure 6C, right). PZP also stratified OC cancer in female subjects and a had tendency for elevated plasma levels in these patients (Figure 6C, right). However, PZP is a quite specific to prostate malignancies because, in reality, these hormone-related cancers (i.e., OC and Proc) cannot be confounded. Interestingly, PZP and SHBG have previously been reported in a smaller cohort of blood samples as highly abundant proteins in women versus men (Geyer et al., 2016). Notably, these findings were consistent in our data, and female plasma indeed contained higher levels of these two proteins (Figure S4). The proteins common to carcinomas, interestingly, tend to represent the higher protein FC in the cancer plasma compared with tumor-specific (e.g., Panc: FN1 = |2.5 ± 0.2| versus PIGR = |1.7 ± 0.2|; Proc: THBS1 = |1.8 ± 0.3| versus PZP = |1.5 ± 0.2|). Nevertheless, the difference of the respective AUC values for FN1 (AUC = 0.938) versus PIGR (AUC = 0.844) in pancreatic and THBS1 (AUC = 0.755) versus PZP (AUC = 0.773) in Proc were not statistically significant (bootstrap-based p = 0.262 and 0.787, respectively; Figures 6A–6C, right), suggesting that the specific maker candidates can achieve a similar predictive power than common markers even when the FCs are smaller. We hereby confirm that the markers common to carcinomas generally represent the highest protein FC in the cancer plasma and effective stratification abilities but that specific markers are necessary to make any blood cancer signature tissue type specific.

Common and Cancer-Specific Blood Proteins in the Independent Cohorts of Pancreatic and Prostate Carcinoma

To verify the putative glycoprotein markers for Proc and Panc discovered by the cross-tumor analysis either as tissue-specific or common cancer changes, we collected two additional, independent cohorts of localized pancreatic (n = 45) and prostate (n = 84) carcinoma (Table S5). Likewise, we extracted glycoproteins from 132 plasma samples (i.e., 129 clinical samples and 3 identical plasma replicates), measured them again by SWATH-MS, and analyzed them for protein statistical significance in an

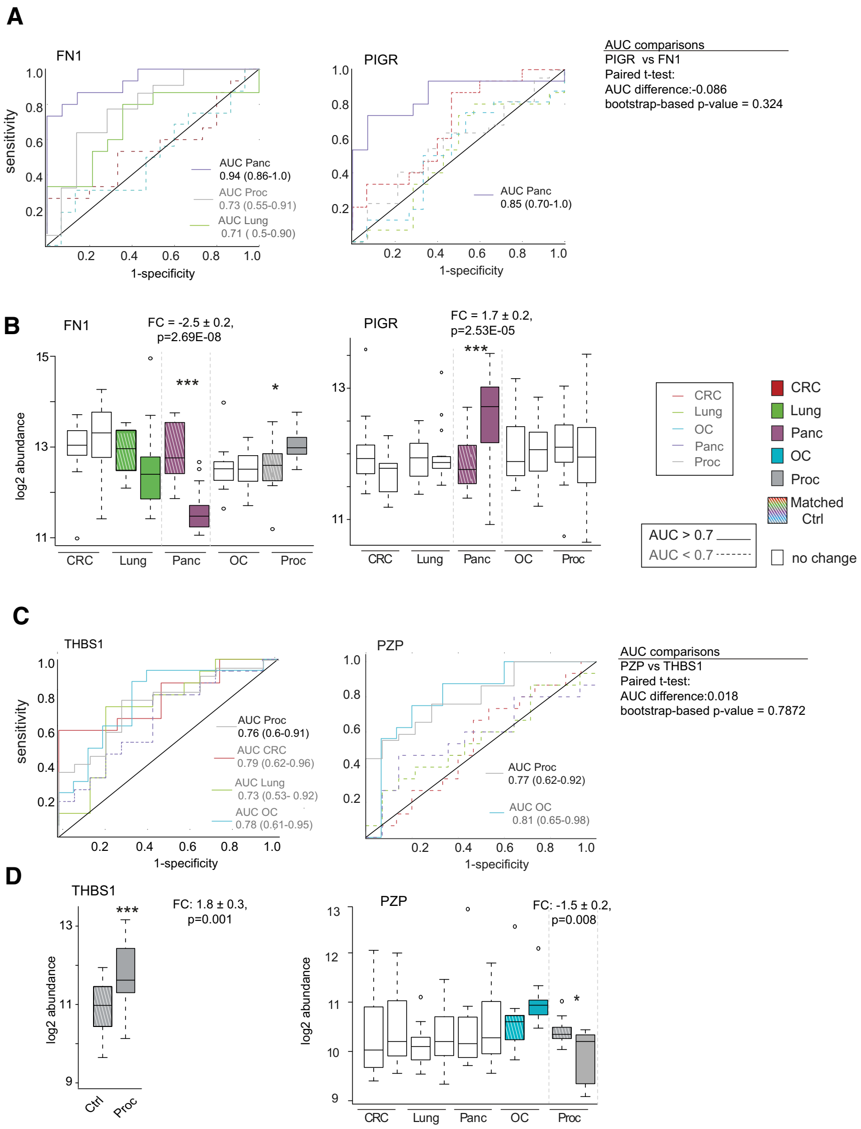


Figure 6. Common and Specific Predictors in the Plasma Proteome of Pancreatic and Prostate Cancer

(A–D) Individual proteins selected as common (FN1 and THBS1) or specific (PIGR and PZP) cancer markers in pancreatic (A) and prostate (C) carcinoma cohorts visualized by ROC curves. The confidence interval (CI) of 95% for each ROC curve is presented in parenthesis. All traces below the AUC cutoff value of 0.7 are presented with dashed lines. Respective AUCs were compared by using a bootstrap test for two correlated ROC curves. Boxplots were generated from log2 abundance of the respective proteins: FN1 and PIGR for Panc (B) and THBS1 and PZP for Proc (D) cohort. Fold change and p values for each respective marker were obtained from MS stats analysis. Line and plot colors correspond to the respective cancers. Boxplots indicating median line (for n = 14–22 subjects) and the maximum and minimum of the given data (upper and lower hinges, respectively).

independent experiment (Table S2). In total, we identified and functionally annotated 356 plasma glycoproteins, 272 of which were identified in the cross-tumor discovery cohort and 294 and 243 in the prostate and pancreatic validation cohorts, respectively (Table S6).

To investigate the status of individual proteins in the respective validation sets, we first evaluated the predictive performance of the discovered candidates above (Figures 5B and 5C), starting from the top protein with the highest stability selection score. Adjusted (for age and gender) and non-adjusted AUC analysis revealed 8 candidates for pancreas—FN1, CNDP1, ITIH3, PIGR, BTD, BCHE, F13B, and HRG—and 2 for prostate—PZP and THBS1—tissue (Table S3) that remain significantly predictive of cancer. The corresponding p values of AUC scores for individual proteins (DeLong test, comparison with chance level AUC = 50%) were corrected for the number of tests performed via FDR-based correction (Benjamini and Hochberg, 1995). AUC scores with FDR-corrected p < 0.1 in adjusted and non-adjusted analyses were considered significant (Table S3).

To account for demographic confounders in differential expression analysis of top selected candidates, we used a linear regression model that relates individual protein expression levels (response variable) to the group assignment (i.e., control or cancer) and also to confounders such as age and gender (Table S4; Figure 7). Of seven predictive candidates, the final four glycoproteins (FN1, ITIH3, CNDP1, and PIGR) also retained significant differential expression in the plasma of Panc patients for both discovery and validation cohorts when accounting for age and gender (FDR-corrected p < 0.1; Table S4; Figure 7A). For localized Proc, two predictive glycoproteins, PZP and THBS1, remained significantly differentially expressed (FDR-corrected p < 0.1) in the Proc serum of discovery and validation cohorts when accounting for age (dot plots; Figure 7B).

In summary, our additional experiments (Figure 7) confirmed that two pancreatic tissue-specific proteins from cross-tumor analysis, PIGR and CNDP1, classified the localized cancer statistically different from the chance level (i.e., AUC = 50%) with ROC AUC values of 72% and 75% in the validation data (n = 45, FDR-corrected p = 0.02 and p = 0.001, respectively; Figure 7A). These AUC scores were slightly lower compared with those obtained in the initial pancreatic cohort (n = 29, AUC = 85% and 86% for PIGR and CNDP1, respectively; FDR-corrected p < 0.0001; Figure 7A). Two common cancer proteins from the initial Panc panel, FN1 and ITIH3, related to platelets, exhibited slightly higher predictive performance for Panc compared with tissue-specific proteins, with AUC values of

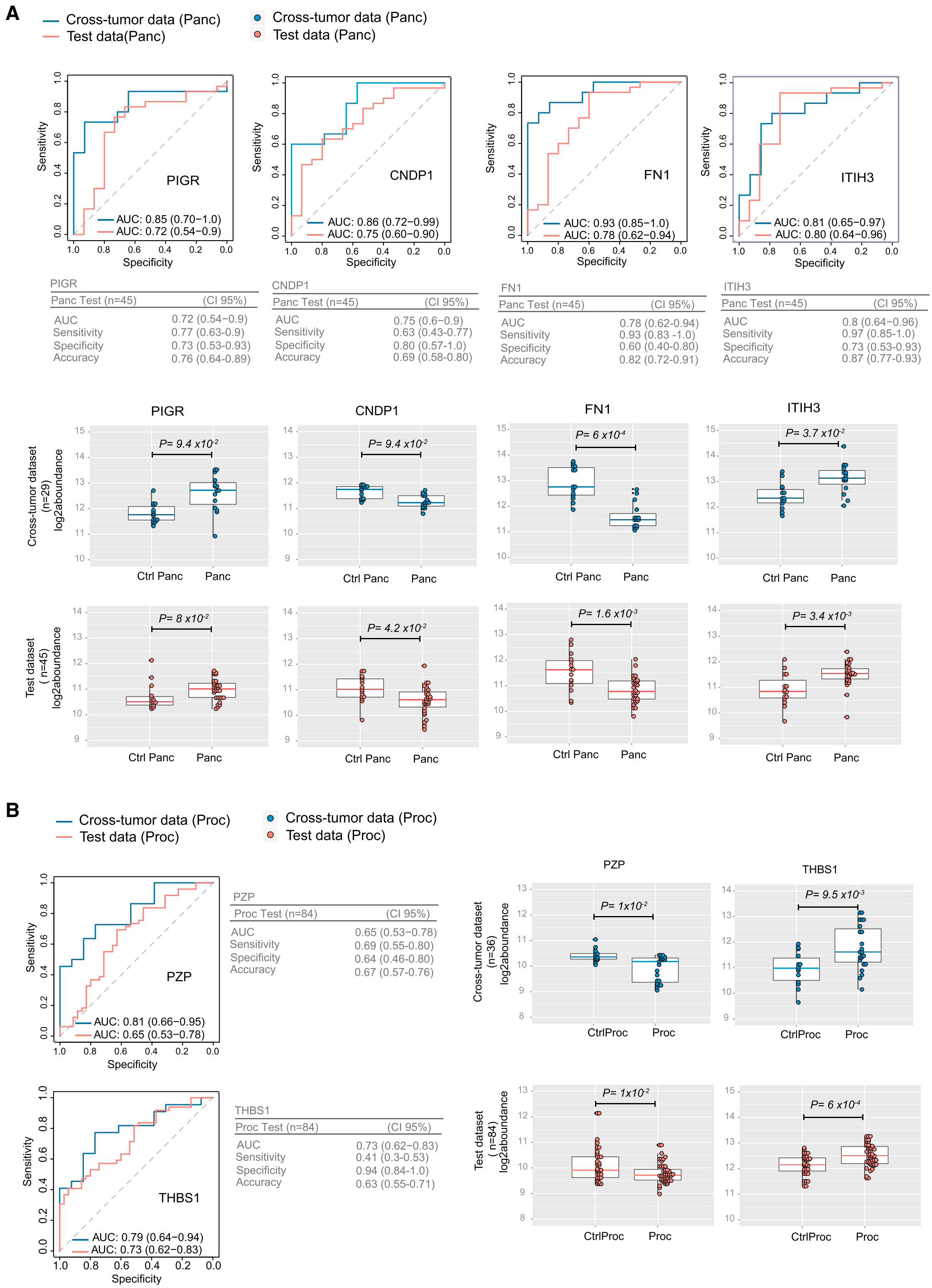
78% and 80%, respectively, and with p values far beyond the chance level of AUC = 50% (FDR-corrected $p < 0.001$). The performance of the hormone-specific marker PZP in the Proc validation data was acceptably higher than would have been expected by chance, with AUC = 65% (FDR-corrected p = 0.03). THBS1, another common cancer protein from cross-tumor analysis, predicted cancer status compared with benign prostate hyperplasia for both the discovery and validation cohorts, with similar AUC values of 73% and 79%, respectively (FDR-corrected $p < 0.001$; Figure 7B).

The combination of proteins, such as the entire composite model of 13 pancreatic or 17 prostate candidates resulting from the above stratification analysis (i.e., a random forest model) could not predict either pancreatic or prostate tumor tissue in additional blood validation cohorts with AUC values as high as in initial discovery cohorts (Figures S5A–S5C). The Panc model of 13 proteins predicted tumors with 71% accuracy, achieving an ROC AUC value of 66% in the test cohort, indicating acceptable overall predictive performance, although at lower levels compared with the discovery cohort (AUC = 0.96, 95% confidence interval [CI] = 0.9–1.0) (Figure S5A). In the validation prostate cohort, the combination of 17 selected proteins predicted the cancer tissue only slightly better than random assignment (AUC = 0.55, 95% CI = 0.43–0.68). A principal-component analysis (PCA) plot based on pancreatic composite model visualized a moderate, respective control-versus-cancer separation in both the training and validation datasets (Figures S5C and S5D).

Interestingly, THBS1, FN1, and ITIH3 are commonly changed cancer platelet proteins identified in the initial cross-tumor analysis whose ''individual'' predictive ability for pancreatic and prostate carcinomas remains stable in the independent populations. Common markers of blood cancerous changes together with tumor type-specific proteins, such as PIGR and CNDP1 for pancreas and PZP for prostate tissue, although confirming their predictive ability and differential expression in the additional cancer subjects, still need to be verified in follow-up experiments against a reference population containing other diseases, such as benign tumors, inflammatory disease, chronic pancreatitis, or diabetes.

DISCUSSION

In the present proteomics study, we analyzed the blood samples of 284 subjects, generating the data resource that allowed us to reveal molecular similarities and differences within the plasma/serum proteome in different human cancers: colorectal, pancreatic, lung, prostate, and ovarian carcinomas. To date, no current blood proteomics investigation combines multi-cancer comparisons within the same analysis; such studies rely on meta-analyses and can be confounded by variable pre-analytical factors, proteomics techniques, and measurement machines (Amess et al., 2013). To avoid experimental bias by using one universal control group for all comparisons, our study design included, for each cancer type, a matched control from each hospital center. To increase the analytical depth of the blood proteome, we subjected the samples to glycoprotein enrichment and acquired proteomics data using reproducible, highly multiplexed SWATH-MS. The entirely independent collection and analysis of external blood samples, totaling 129, allowed us to further test putative biomarker candidates previously discovered by initial cross-tumor analysis of the 155 individuals. Across all samples, we examined whether the molecular systems perturbed in different tumors represent a common, systemic response to cancer or whether the protein biomarkers are specific to each cancer type. Based on our systems-level analysis, although reporting high sensitivity to reflect cancer metabolism, common markers ostensibly are not necessarily specific to individual cancers. We demonstrated that the proteins FN1, THBS1, and ITIH3, involved in platelet activation, signaling, and aggregation, are those that change across different tissues in early cancer and appear to be sensitive to general cancer biology in the blood (Figure 5). THBS1 expression in cancer disease previously delivered somewhat controversial results (Miyata and Sakai, 2013), probably because of its complex function in the process of tissue angiogenesis related to tumor development and staging (Kazerounian et al., 2008). The dysregulated platelet proteome discovered in this cross-tumor study is in line with two recent publications demonstrating that blood-isolated platelet mRNA profiles distinguished accurately between carcinomas (localized or advanced metastatic stage) and healthy individuals (Best et al., 2015). Our study detected carcinoma-specific markers (e.g., CNDP1 and PIGR for pancreatic carcinoma), some of which the literature had previously reported as blood biomarkers. PZP, confirmed here as significantly changed in the prostate serum, has been reported as a protease inhibitor that forms a complex with non-catalytic prostate-specific antigen (PSA), which has been approved for early serum screening of Proc (Christensson et al., 1990). Of the differentially expressed proteins identified in the Panc plasma, PIGR has previously been described as one of the biomarkers overexpressed in the fluid of pancreatic cysts with malignant potential (Park et al., 2015) and has also recently been reported as elevated in the plasma of Panc patients, but at levels not significantly different from those measured in patients with chronic pancreatitis (Sogawa et al., 2016). Although some of our specific and common cancer proteins confirmed their significant ability to predict localized Panc in the validation experiment, our control samples were limited to healthy individuals. This could result in a higher rate of false positive results when testing our predictive candidates against subjects with benign tumors and inflammation disease, such as diabetes or chronic pancreatitis, compared with the results observed in our study. PIGR levels, together with CNDP1, FN1, and ITIH3, were indeed significantly altered in the plasma of localized Panc in the two independent geographic cohorts (i.e., University Hospital Olomouc, Czech Republic, and University of Washington, United States; Figure 7A). Further follow-up experiments will be required to determine whether the pancreas-specific PIGR and CNDP1 plasma protein levels, individually or in combination with other cancer-sensitive markers (i.e., FN1 and ITIH3), can distinguish between chronic pancreatitis and localized pancreatic carcinoma. Two serum candidates, PZP as prostate tissue-specific and THBS1 as a common cancer marker, both confirmed their individual predictive ability for prostate carcinoma when tested against control subjects with benign prostatic disease. Further studies are necessary to characterize



PIGR

Panc Test (n=45)	(CI 95%)
AUC	0.72 (0.54–0.9)
Sensitivity	0.77 (0.63-0.9)
Specificity	0.73 (0.53-0.93)
Accuracy	0.76 (0.64-0.89)

CNDP1

Panc Test (n=45)	(CI 95%)
AUC	0.75 (0.6−0.9)
Sensitivity	0.63 (0.43-0.77)
Specificity	0.80 (0.57-1.0)
Accuracy	0.69 (0.58-0.80)

FN1

Panc Test (n=45)	(CI 95%)
AUC	0.78 (0.62-0.94)
Sensitivity	0.93 (0.83 -1.0)
Specificity	0.60 (0.40-0.80)
Accuracy	0.82 (0.72-0.91)

ITIH3

Panc Test (n=45)	(CI 95%)
AUC	0.8 (0.64−0.96)
Sensitivity	0.97 (0.85-1.0)
Specificity	0.73 (0.53-0.93)
Accuracy	0.87 (0.77-0.93)

PZP

Proc Test (n=84)	(CI 95%)
AUC	0.65 (0.53−0.78)
Sensitivity	0.69 (0.55-0.80)
Specificity	0.64 (0.46-0.80)
Accuracy	0.67 (0.57-0.76)

THBS1

Proc Test (n=84)	(CI 95%)
AUC	0.73 (0.62−0.83)
Sensitivity	0.41 (0.3-0.53)
Specificity	0.94 (0.84-1.0)
Accuracy	0.63 (0.55-0.71)

Figure 7. Performance of Individual Proteins in Independent Validation Cohorts of Prostate and Pancreatic Carcinoma

(A and B) ROC curves for individual markers corresponding to age and gender non-adjusted analyses in the discovery and test datasets of pancreatic (A) and prostate (B) cancer cohorts. Test cohort summary statistics (i.e., specificity, sensitivity, and accuracy at 95% CI) of individual proteins were calculated at an

(legend continued on next page)

blood protein biomarkers of a broader range of cancers and to evaluate their potential clinical utility to provide some indication of cancer type in combination with tissue-specific markers. In contrast to studies of advanced tumor stages, wherein blood signatures could reflect organ-specific malignancy, this study aligns the real challenge facing proteomics methods whose aim is to detect early-stage cancer disease in the blood of asymptomatic subjects. This study also indicates that emerging proteomics technologies, such as SWATH-MS, that provide comprehensive, accurate protein measurements can enable efficient strategies for early and non-invasive cancer detection. Until recently, the technological challenge of measuring blood proteins, many of which are low in abundance and only observed under specific conditions, has limited proteomics analysis across large-scale clinical studies to the measurement of fewer than a hundred proteins in a maximum of a few dozen samples. Combining SWATH-MS and N-glycosite enrichment provides more comprehensive plasma analysis than previous proteomics studies, allowing thousands of glycopeptides to be reproducibly monitored over hundreds of clinical samples.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Clinical Samples
- METHOD DETAILS
  - ○ Isolation of de-N-glycopeptides from blood samples
- SWATH-MS DATA GENERATION
  - ○ N-glycoprotein SWATH-Assay Library
  - ○ SWATH-MS Measurement and Data Processing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ SWATH-MS data analysis
  - ○ Bioinformatics Data Analyses
  - ○ Functional Network Analysis
  - ○ Data Stratification in Cross-Tumor dataset
  - ○ Data validation on independent cancer cohorts
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures and eight tables and can be found with this article online at https://doi.org/10.1016/j.celrep.2018.04.114.

ACKNOWLEDGMENTS

We thank C. C. Koh and L. Gillet for MS technical assistance and A. Van Drogen for fruitful discussions. We also thank Prof. Olga Vitek at Northeastern University (Boston, MA, USA) for critical discussion of the project. P.B. was supported by the Swiss SystemsX.ch Initiative, evaluated by the Swiss National Science Foundation (TPdF 2013/134). Y.L. was supported by his start-up funds from Yale Cancer Biology Institute. E.N. was supported by the Mrs. Berta Kamprad Foundation, BioCARE, the Swedish Breast Cancer Association, and NHS (National Health Services, ALF). The work carried out in this study was supported by the Swiss National Science Foundation (grants SNSF 31003A_130530 and SNSF 31003A_166435 to R.A).

AUTHOR CONTRIBUTIONS

T.S. and Y.L. designed and oversaw the experiments and wrote the manuscript. T.S. performed data analysis and prepared the figures. E.A. performed the stratification analysis. S.S., R.S., R.H., T.A.B., S.P., R.C., S.M., and E.N. helped with clinical samples and critical reading of the manuscript. E.G.W. edited the manuscript and figures. A.S. and P.B. provided bioinformatics expertise. B.F. helped with the glycocapture experiment. S.G. and A.O. helped with clinical samples and design of the study. M.C. supervised the stratification analysis. R.A. envisioned, designed, and supervised the study. All authors edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 28, 2016
Revised: March 30, 2018
Accepted: April 26, 2018
Published: May 29, 2018

REFERENCES

Amess, B., Kluge, W., Schwarz, E., Haenisch, F., Alsaif, M., Yolken, R.H., Leweke, F.M., Guest, P.C., and Bahn, S. (2013). Application of meta-analysis methods for identifying proteomic expression level differences. Proteomics *13*, 2072–2076.

Anderson, N.L. (2010). The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. Clin. Chem. *56*, 177–185.

Aoki, K., and Taketo, M.M. (2007). Adenomatous polyposis coli (APC): a multi-functional tumor suppressor gene. J. Cell Sci. *120*, 3327–3335.

Beauvais, D.M., and Rapraeger, A.C. (2004). Syndecans in tumor cell adhesion and signaling. Reprod. Biol. Endocrinol. *2*, 3.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Series B Stat. Methodol. *57*, 289–300.

Best, M.G., Sol, N., Kooi, I., Tannous, J., Westerman, B.A., Rustenburg, F., Schellen, P., Verschueren, H., Post, E., Koster, J., et al. (2015). RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics. Cancer Cell *28*, 666–676.

Blattmann, P., Heusel, M., and Aebersold, R. (2016). SWATH2stats: An R/Bioconductor Package to Process and Convert Quantitative SWATH-MS Proteomics Data for Downstream Analysis Tools. PLoS ONE *11*, e0153160.

Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., et al. (2012). A cross-platform toolkit for mass spectrometry and proteomics. Nat. Biotechnol. *30*, 918–920.

Choi, M., Chang, C.Y., Clough, T., Broudy, D., Killeen, T., MacLean, B., and Vitek, O. (2014). MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics *30*, 2524–2526.

optimal cutoff point as determined by Youden's index (Youden, 1950). For further details, see Tables S3 and S4. Additionally, shown are dot plots corresponding to individual protein expression changes (FDR-corrected $p < 0.1$) in the discovery and validation cohorts of two respective carcinomas. N corresponds to the number of subjects per cancer cohort. Corrected p values indicate whether the regression coefficient for group assignment (control or cancer) is significantly different from zero when accounting for confounders such as subject age and gender. See also Figure S5.

Christensson, A., Laurell, C.B., and Lilja, H. (1990). Enzymatic activity of prostate-specific antigen and its reactions with extracellular serine proteinase inhibitors. Eur. J. Biochem. *194*, 755–763.

Cima, I., Schiess, R., Wild, P., Kaelin, M., Schüffler, P., Lange, V., Picotti, P., Ossola, R., Templeton, A., Schubert, O., et al. (2011). Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer. Proc. Natl. Acad. Sci. USA *108*, 3342–3347.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *95*, 14863–14868.

Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods *4*, 207–214.

Farrah, T., Deutsch, E.W., Omenn, G.S., Campbell, D.S., Sun, Z., Bletz, J.A., Mallick, P., Katz, J.E., Malmstrom, J., Ossola, R., et al. (2011). A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. Mol. Cell. Proteomics *10*, M110.006353.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., et al. (2015). COSMIC: exploring the world’s knowledge of somatic mutations in human cancer. Nucleic Acids Res. *43*, D805–D811.

Foster, L.J., de Hoog, C.L., Zhang, Y., Zhang, Y., Xie, X., Mootha, V.K., and Mann, M. (2006). A mammalian organelle map by protein correlation profiling. Cell *125*, 187–199.

Frank, R. (2002). The SPOT-synthesis technique. Synthetic peptide arrays on membrane supports–principles and applications. J. Immunol. Methods *267*, 13–26.

Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. Am. Stat. *56*, 316–324.

Geyer, P.E., Kulak, N.A., Pichler, G., Holdt, L.M., Teupser, D., and Mann, M. (2016). Plasma Proteome Profiling to Assess Human Health and Disease. Cell Syst. *2*, 185–195.

Gillet, L.C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol. Cell. Proteomics *11*, O111.016717.

Guo, T., Kouvonen, P., Koh, C.C., Gillet, L.C., Wolski, W.E., Röst, H.L., Rosenberger, G., Collins, B.C., Blum, L.C., Gillessen, S., et al. (2015). Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. Nat. Med. *21*, 407–413.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., Niu, B., McLellan, M.D., Uzunangelov, V., et al.; Cancer Genome Atlas Research Network (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell *158*, 929–944.

Huang, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Jiménez, B., Volpert, O.V., Crawford, S.E., Febbraio, M., Silverstein, R.L., and Bouck, N. (2000). Signals leading to apoptosis-dependent inhibition of neovascularization by thrombospondin-1. Nat. Med. *6*, 41–48.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.

Kazerounian, S., Yee, K.O., and Lawler, J. (2008). Thrombospondins in cancer. Cell. Mol. Life Sci. *65*, 700–712.

Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. *74*, 5383–5392.

Kunszt, P., Blum, L., Hullár, B., Schmid, E., Srebniak, A., Wolski, W., Rinn, B., Elmer, F.-J., Ramakrishnan, C., Quandt, A., and Malmström, L. (2015). iPortal: the swiss grid proteomics portal: Requirements and new features based on experience and usability considerations. Concurr. Comput. *27*, 433–445.

Leiserson, M.D., Vandin, F., Wu, H.T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat. Genet. *47*, 106–114.

Liu, Y., Hüttenhain, R., Surinova, S., Gillet, L.C., Mouritsen, J., Brunner, R., Navarro, P., and Aebersold, R. (2013). Quantitative measurements of N-linked glycoproteins in human plasma by SWATH-MS. Proteomics *13*, 1247–1256.

Liu, Y., Chen, J., Sethi, A., Li, Q.K., Chen, L., Collins, B., Gillet, L.C., Wollscheid, B., Zhang, H., and Aebersold, R. (2014). Glycoproteomic analysis of prostate cancer tissues by SWATH mass spectrometry discovers N-acylethanolamine acid amidase and protein tyrosine kinase 7 as signatures for tumor aggressiveness. Mol. Cell. Proteomics *13*, 1753–1768.

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. (2015). Quantitative variability of 342 plasma proteins in a human twin population. Mol. Syst. Biol. *11*, 786.

Majumder, S., Chari, S.T., and Ahlquist, D.A. (2015). Molecular detection of pancreatic neoplasia: Current status and future promise. World J. Gastroenterol. *21*, 11387–11395.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., et al. (2009). Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res. *37*, D619–D622.

McPhail, S., Johnson, S., Greenberg, D., Peake, M., and Rous, B. (2015). Stage at diagnosis and early mortality from cancer in England. Br. J. Cancer *112* (*Suppl 1*), S108–S115.

Meinshausen, N., and Buhlmann, P. (2010). Stability selection. J. R. Stat. Soc. Series B Stat. Methodol. *72*, 417–473.

Miyata, Y., and Sakai, H. (2013). Thrombospondin-1 in urological cancer: pathological role, clinical significance, and therapeutic prospects. Int. J. Mol. Sci. *14*, 12249–12272.

Park, J., Yun, H.S., Lee, K.H., Lee, K.T., Lee, J.K., and Lee, S.Y. (2015). Discovery and Validation of Biomarkers That Distinguish Mucinous and Nonmucinous Pancreatic Cysts. Cancer Res. *75*, 3227–3235.

Petitjean, A., Achatz, M.I., Borresen-Dale, A.L., Hainaut, P., and Olivier, M. (2007). TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. Oncogene *26*, 2157–2165.

Röst, H.L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S.M., Schubert, O.T., Wolski, W., Collins, B.C., Malmström, J., Malmström, L., and Aebersold, R. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat. Biotechnol. *32*, 219–223.

Röst, H.L., Liu, Y., D’Agostino, G., Zanella, M., Navarro, P., Rosenberger, G., Collins, B.C., Gillet, L., Testa, G., Malmström, L., and Aebersold, R. (2016). TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. Nat. Methods *13*, 777–783.

Scheffler, M., Bos, M., Gardizi, M., König, K., Michels, S., Fassunke, J., Heydt, C., Künstlinger, H., Ihle, M., Ueckeroth, F., et al. (2015). PIK3CA mutations in non-small cell lung cancer (NSCLC): genetic heterogeneity, prognostic impact and incidence of prior malignancies. Oncotarget *6*, 1315–1326.

Schiess, R., Wollscheid, B., and Aebersold, R. (2009). Targeted proteomic strategy for clinical biomarker discovery. Mol. Oncol. *3*, 33–44.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

Shteynberg, D., Deutsch, E.W., Lam, H., Eng, J.K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R.L., Aebersold, R., and Nesvizhskii, A.I. (2011). iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol. Cell. Proteomics *10*, M111.007690.

Siegel, R., DeSantis, C., Virgo, K., Stein, K., Mariotto, A., Smith, T., Cooper, D., Gansler, T., Lerro, C., Fedewa, S., et al. (2012). Cancer treatment and survivorship statistics, 2012. CA Cancer J. Clin. *62*, 220–241.

Sims, A.H., Smethurst, G.J., Hey, Y., Okoniewski, M.J., Pepper, S.D., Howell, A., Miller, C.J., and Clarke, R.B. (2008). The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis. BMC Med. Genomics *1*, 42.

Sogawa, K., Takano, S., Iida, F., Satoh, M., Tsuchida, S., Kawashima, Y., Yoshitomi, H., Sanda, A., Kodera, Y., Takizawa, H., et al. (2016). Identification of a novel serum biomarker for pancreatic cancer, C4b-binding protein α-chain (C4BPA) by quantitative proteomic analysis using tandem mass tags. Br. J. Cancer *115*, 949–956.

Suh, E.J., Kabir, M.H., Kang, U.B., Lee, J.W., Yu, J., Noh, D.Y., and Lee, C. (2012). Comparative profiling of plasma proteome from breast cancer patients reveals thrombospondin-1 and BRWD3 as serological biomarkers. Exp. Mol. Med. *44*, 36–44.

Surinova, S., Choi, M., Tao, S., Schüffler, P.J., Chang, C.Y., Clough, T., Vysloužil, K., Khoylou, M., Srovnal, J., Liu, Y., et al. (2015). Prediction of colorectal cancer diagnosis based on circulating plasma proteins. EMBO Mol. Med. *7*, 1166–1178.

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., et al. (2017). A pathology atlas of the human cancer transcriptome. Science *357*, eaan2507.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M.; Cancer Genome Atlas Research Network (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Use R!)

Wijten, P., van Holten, T., Woo, L.L., Bleijerveld, O.B., Roest, M., Heck, A.J., and Scholten, A. (2013). High precision platelet releasate definition by quantitative reversed protein profiling–brief report. Arterioscler. Thromb. Vasc. Biol. *33*, 1635–1638.

Wu, G., Feng, X., and Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. Genome Biol. *11*, R53.

Wu, G., Dawson, E., Duong, A., Haw, R., and Stein, L. (2014). ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. F1000Res. *3*, 146.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R.H., Eshleman, J.R., Nowak, M.A., et al. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature *467*, 1114–1117.

Youden, W.J. (1950). Index for rating diagnostic tests. Cancer *3*, 32–35.

Zhang, H., Li, X.J., Martin, D.B., and Aebersold, R. (2003). Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. Nat. Biotechnol. *21*, 660–666.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
Affi-gel Hydrazine resin, 1 l	Bio-Rad	Cat#153-6050
N-glycosidase F, rec. *E. coli*.	Roche Diagnostics AG	Cat#11365193001
PNGaseF (glycerol free), 75,000 units, 500,000 units/ml	BioConcept AG	Cat#P0705L
Fetuin-B	Sigma	Cat#F3004-25MG
Human N-glycoprotein SWATHatlas	(Liu et al., 2014)	JPT Peptide Tech, Berlin, Germany
Trypsin	Promega	Cat#V5113
AB beta-Galactosidase digested	Sigma	Cat#4333606
Critical Commercial Assays		
iRT-kit WR	Biognosys	N/A
Deposited Data		
COSMIS v71	https://cancer.sanger.ac.uk/cosmic	N/A
ProteomeXchange (Proteomics data, raw data files)	http://proteomecentral.proteomexchange.org	PXD004998
SWATH library (TraML format)	http://proteomecentral.proteomexchange.org	PXD004998
REACTOME DB	(Matthews et al., 2009)	N/A
List of cancer drivers	Table S7, excel file	N/A
Proteomics data	Table S5, excel file	N/A
Software and Algorithms		
Analyst TF 1.5.1 software	AB Sciex	N/A
ProteoWizard (version 3.0.3316)	(Chambers et al., 2012)	N/A
OpenMS tool		https://www.openms.de/
Trans-Proteomic Pipeline (TPP v4.6 OCCUPY rev 0, Build 201208211847)	(Keller et al., 2002) (Shteynberg et al., 2011)	N/A
OpenSWATH tool	(Röst et al., 2014)	http://www.openswath.org/en/latest/
TRIC algorithm	(Röst et al., 2016)	N/A
SWATH2stats	(Blattmann et al., 2016)	http://bioconductor.org/packages/release/bioc/html/SWATH2stats.html
MSstats (version MSstats.daily 2.3.5)	(Choi et al., 2014)	http://bioconductor.org/packages/release/bioc/html/MSstats.html
Cluster 3.0 v.1.52	(Eisen et al., 1998)	N/A
Cytoscape	(Shannon et al., 2003)	N/A
Other		
96-Well Sirocco plate	Waters	Cat#186002448
96-Well MACROSpin G10,40-400 μl, ea. Gel Filtration	Nest Group	Cat#SNS S010L
MACROSpin Plate-VydacSilicaC18	Nest Group	Cat#SNS SS18V
3-μm 200 Å Magic C18 AQ resin	MichromBioResources	Cat#2847

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact Dr. Tatjana Sajic (e-mail: sajic@imsb.biol.ethz.ch).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Clinical Samples

In the initial study we used five clinical cancer cohorts. Sample collection and procession were standardized for all five cancer types studied to increase the chance of detecting biological signals. The protocols of the blood collection, processing, and storage within each clinical cohort were approved by the regional ethics committees in each individual health institution before starting patient enrollment and sample collection. The subjects of CR, Lung, and pancreatic localized carcinoma with their respective control subjects were recruited at the University Hospital Olomouc in the Czech Republic, which was approved by the ethics committees of the Medical Faculty at the University Hospital Olomouc and Faculty of Medicine and Dentistry, Palacky University, Olomouc. All individuals have signed an informed consent document.

The blood plasma samples from the above cancer cohorts were prepared according to a plasma protocol described previously (Surinova et al., 2015). In brief, blood was collected into tubes processed with EDTA and was directly centrifuged at 6,067 g for 3 min at 4°C. Plasma was transferred into a fresh tube, frozen at −20°C, and then stored at −80°C.

The samples of epithelial ovarian cancer and their respective control subjects were collected at Skåne University Hospital and ethical permission was approved by the Lund University Ethics Committee. The patients gave written informed consent for participation. All blood samples from the ovarian cancer cohort were drawn into EDTA-coated tubes and were centrifuged at 2000 g, for 10 min and the plasma was stored in −80°C within two hours from sampling.

The Proc cohort was approved by the Ethics Committee of the Canton of St. Gallen, Switzerland. All patients have signed consent for participation. Benign prostate hyperplasia samples were used as the corresponding control samples. Blood serum samples were prepared starting from 8 mL of blood collected in a serum separator tube containing clot activator and gel (Vacutainer, SSTTM II Advance, REF 367953; Becton Dickinson) as described previously (Cima et al., 2011).

The cancer samples of Cross-Tumor data were consecutively recruited as diagnosed at early localized (Stage I, n = 53) or locally advanced clinical stage (Stage II and III, n = 30). The male and female subjects were involved. All patients gave written informed consent for participation. The CRC and Proc samples at the localized stage of disease were partially selected from the previously published clinical studies (Surinova et al., 2015) and (Cima et al., 2011) respectively. The Proc cohort consisted of 22 carcinomas and 14 respective controls, while the other four cancer cohorts consisted of 15 or 16 carcinomas and 14 or 15 respective controls resulting in a total of 155 individuals' blood samples included in this study (Table S1).

In the validation phase we used 129 additional clinical blood samples. Proc cohort (n = 84) was approved by the Ethics Committee of the Canton of St. Gallen, Switzerland. Blood serum samples were prepared in the same way as the samples of the initial training prostate cohort as described above. Panc dataset (n = 45) was approved by the Institutional Review Board at the University of Washington (Seattle, WA).The plasma samples were collected into purple-top tubes (Becton Dickinson, Franklin Lakes, NJ) with EDTA, the potassium salt, as an anticoagulant. The blood was centrifuged at 330 g for 20 min.

Clinico-pathological details of 284 human subjects for cross-tumor analysis and prostate and pancreatic validation experiment are present in the excel Table S5 Patient's characteristics are summarized in Table S1 and Table S2.

METHOD DETAILS

Isolation of de-N-glycopeptides from blood samples

The 294 blood samples were prepared by solid-phase extraction of N glycoproteins as previously described (Cima et al., 2011; Surinova et al., 2015; Zhang et al., 2003), then applied in a high-throughput manner (Surinova et al., 2015). The initial 162 samples were processed in four batches. In each batch, samples were randomized and experimental replicates of the same blood samples were included in parallel. The samples were prepared by using multi-channel well plates (Sirocco plate ref: 186002448 Waters) which enables high-throughput "in-well" sample processing. First, the proteins were treated with sodium periodate solution to oxidize the glycan moieties and then purified by G-10 gel filtration cartridges (Nest Group, Southborough, MA). The samples were conjugated overnight to Affi-gel Hydrazine resin (Bio-Rad) that was loaded in each well of Sirocco plate. With extensive washing procedure we removed unbound proteins from the matrix. Overnight protein digestion (protein/trypsin ratio of 200:1) was performed directly on the plate that contains conjugated glycoproteins on the hydrazine matrix. The next day digested non-glycopeptides are extensively washed out from the matrix. N-linked glycopeptides were released by PNGase F (PNGaseF, N-glycosidase F) enzyme from the hydrazide matrix, subsequently cleaned on MACROSpin Plate-VydacSilicaC18 (ref: Nest Group, Southborough, MA), solubilized in 100 μL of 0.1% aqueous formic acid (FA) with 2% acetonitrile (ACN) and were used for final MS analysis. The reference glycoprotein fetuin-B (Swiss-prot: Q58D62) was spiked in equal 1 pmol/μL amount into the each plasma or serum samples to control intra-experimental variations. N-glycopeptides corresponding to 2 μL of patient blood were used for each MS run.

SWATH-MS DATA GENERATION

N-glycoprotein SWATH-Assay Library

The SWATH assay library was built from the data-dependent acquisition (DDA, also known as shotgun) analysis of synthetic peptides (Liu et al., 2014) and native glycopeptides isolated from natural blood samples. DDA acquisition was performed on

TripleTOF 5600 mass spectrometer equipped with a NanoSpray III source and heated interface (AB Sciex, Concord, Ontario, Canada). The common pool sample of native enriched N-glycopeptides from all the samples in the study was fractionated in 24 off-gel fractions or used directly as non-fractionated sample for downstream MS analysis. In line with native enriched glycopeptides we acquired DDA spectra of synthetic N-glycopeptides previously collected by literature search (Cancer-associated proteins: CAPs) and synthetized using SPOT-synthesis technology (JPT Peptide Tech, Berlin, Germany)(Frank, 2002). To each of the samples indexed retention time (iRT) peptides were added (RT-kit WR, Biognosys) and the peptides samples were injected onto a in-house C18 nanocolumn packed directly in a fused silica PicoTip emitter (New Objective, Woburn, MA, USA) with 3-μm 200 Å Magic C18 AQ resin (Michrom BioResources, Auburn, CA, USA). Reverse phase peptide separation was performed on a NanoLC-Ultra 2D Plus system (Eksigent–AB Sciex, Dublin, CA, USA). The nanoLC gradient was linear from 2 to 35% B (0.1% formic acid in ACN) over 120 min at a flow rate of 300 nl/min and an oven temperature of 70°C.

The nano-LC and MS instruments were operated by Analyst TF 1.5.1 software (AB Sciex). Electrospray ionization was performed in positive polarity at a voltage of 2.6 kV and was assisted pneumatically by nitrogen (20 psi). Mass spectra and tandem mass spectra were recorded in ''high-sensitivity'' mode over a mass/charge (m/z) range of 50 to 2000 with a resolving power of 30,000 (full width at half maximum [FWHM]). MS/MS spectra acquisition was triggered by DDA mode consisting in a survey scan of 250 ms followed by 20 MS/MS-dependent acquisitions of 50 ms each. MS/MS spectra were generated by collision-induced dissociation (nitrogen) with dynamic collision energy (i.e., rolling collision energy [CE]). DDA selection of the precursor ions was as follows: the 20 most intense ions (threshold of 50 counts), charge state from 2 to 5, isotope exclusion of 4u, and precursor dynamic exclusion of 8 s leading to a maximum total MS duty cycle of 1.15 s. External mass calibration was performed by injecting a 100-fmol solution of β-galactosidase tryptic digest every samples in order to avoid carryover between clinical samples. Raw data files (.wiff) were centroided and converted into mzML format using the ABSciex converter (beta version 2011) and subsequently converted into mzXML using openMS (version 1.8). The converted data files were searched using the search engines X! TANDEM CYCLONE TPP (2011.12.01.1 - LabKey, Insilicos, ISB), Omssa (version 2.1.9), and Comet (version 2013.02, revision 2) against the reviewed canonical Swiss-Prot complete proteome database for human (released Oct 1, 2013) appended with common contaminants and reversed sequence decoys (Elias and Gygi, 2007)(40,951 protein sequences including decoys), fetuin-B (Swiss-prot: Q58D62) bovine protein sequence and iRT peptides sequence. The database search included following criteria: semi-tryptic digestion and allowing up to 2 missed cleavages. Included were 'Carbamidomethyl (C)' as static and 'Deamidated (N); Oxidation (M)' as variable modifications. The mass tolerances were set to 30 ppm for precursor-ions and 0.1 Da for fragment-ions.

The identified peptides were processed and analyzed through the Trans-Proteomic Pipeline (TPP v4.6 OCCUPY rev 0, Build 201208211847) using PeptideProphet (Keller et al., 2002), iProphet (Shteynberg et al., 2011) and ProteinProphet scoring. Peptide identifications were reported at FDR of 0.01. The raw spectral libraries were generated from all valid peptide spectrum matches obtained from native and synthetic peptides filtered for N-glycosylation motive (NXS or NXT; X ≠ P) and converted to TraML format using the OpenMS tool ConvertTSVToTraML (version 1.10.0). Decoy transition groups were generated based on shuffled sequences by the OpenMS tool OpenSwathDecoyGenerator (version 1.10.0) and appended to the final SWATH library in TraML format. The MS assays, constructed from the top six most intense transitions with Q1 range from 400 to 1,200 m/z excluding the precursor SWATH window, were used for targeted data analysis of SWATH maps.

From both fractionated and unfractionated samples we obtained confident MS2 spectra for 453 native N-glycoproteins (2743 N-glycopeptides) (Figure 1). Based on this spectral library we augmented MS2 spectra information by injecting 1604 N-glycopeptides corresponding to 880 glycoproteins that were associated with cancer biology based on databases and literature evidence. DDA data from endogenous and synthetic peptides were then combined at the peptide level. Assays coming from synthetic peptides were only accepted if the corresponding peptides were not identified at the endogenous level in the N-glycoproteome isolated from blood samples. All peptides missing the N-glycosylation motif were removed. Finally, the spectral library was configured to contain high quality MS assays for 4347 N-glycopeptides from 1151 glycoproteins. This library was then converted into the final format by using the OpenMS tool, as described above. The final combined library was optimized for plasma N-glycoprotein identification.

SWATH-MS Measurement and Data Processing

All blood samples were measured on TripleTOF 5600 mass spectrometer operated in SWATH mode as described earlier (Gillet et al., 2012). Reverse phase peptide separation was performed with linear nanoLC gradient as described above. An accumulation time of 100 ms was used for 32 fragment ion spectra of 26 m/z each and for the precursor scans (swaths) acquired at the beginning of each cycle, resulting in a total cycle time of 3.3 s. The swaths were overlapping by 1 m/z and thus cover a range of 400-1200 m/z. The collision energy for each window was determined according to the calculation for a charge 2+ ion centered upon the window with a spread of 15. Raw SWATH data files were converted into the mzXML format using ProteoWizard (version 3.0.3316) (Chambers et al., 2012) and SWATH data analysis was performed using the OpenSWATH tool (Röst et al., 2014) integrated in the iPortal workflow (Kunszt et al., 2015).

QUANTIFICATION AND STATISTICAL ANALYSIS

SWATH-MS data analysis

The OpenSWATH workflow input files consisted of the mzXML files from the SWATH acquired data, the TraML assay library file and the TraML file for iRT peptides. SWATH data were extracted with 50 ppm around the expected mass of the fragment ions and with an extraction window of ± 300 s around the expected retention time after performing alignment of iRT peptides. The runs were subsequently aligned with a target FDR of 0.01 and a maximal FDR of 0.1 for aligned features. In the absence of a confidently identified feature, the peptide and protein intensities were obtained by integration of the respective background signal at the expected peptide retention time (Röst et al., 2016). Next, the recorded feature intensities obtained from automatic OpenSWATH data processing were filtered with functions from the R/Bioconductor package SWATH2stats (Blattmann et al., 2016) to reduce the size of the output data, remove low-quality features, and to only keep the features that were identified in at least 2/3 of data (> 100 MS runs) below m-score cut off of 0.01. This resulted in a list of 1360 glycopeptides achieving peptide FDR below 1%. After data were acquired and analyzed we did find systematic artifacts between the samples prepared in different batches. Thus, in the next step, the batch effect was corrected on the level of high quality fragment ion signals by applying mean centering as described previously(Sims et al., 2008). Fragment ion intensity values were log2 transformed and quintile normalized. Next, we calculated the mean value for all signal intensity recorded in individual batch (batch mean) and subtracted it from each run (fragment ion wise) belonged to individual batch. To preserve the dynamics of the original data intensities, the global mean across all samples was added. Then these corrected fragment intensities were introduced in the R/Bioconductor package MSstats (version MSstats.daily 2.3.5) and converted to relative protein abundances that were used for further statistical data analysis (Choi et al., 2014).

The MSstats output file with relative protein intensities was used as input file for further unsupervised sample clustering. The FCs and p values of all five carcinoma cohorts compared to their respective controls were obtained in five parallel analyses by linear mixed models with expanded scope of biological and technical replication. The raw p values < 0.05 and FCs cutoff ± 1.2 were used as the input for further statistical analysis (Table S6).

Bioinformatics Data Analyses

Hierarchical data clustering analysis was done by a two-dimensional centered heatmap using Cluster 3.0 v.1.52 on the log-transformed, and normalized relative protein intensities (Eisen et al., 1998). City-block distance and average linkage as distance measure were used for clustering. The functional annotation analyses were performed by DAVID (Huang et al., 2009)and Reactome pathway database. Significant (FDR < 0.05) biological modules are represented with –log10 transformation of p value. The volcano plots were obtained directly from the MSstats output. The boxplots or violin plots of total or separate significant proteins in five carcinoma cohorts were plotted using the ggplot2 package (Wickham, 2009) in Rstudio (version 3.0.2). The Spearman correlation between protein pairwise was calculated and visualized by using Corrgram package in Rstudio (Friendly, 2002). Venn diagrams were drawn by an online tool (http://bioinformatics.psb.ugent.be/webtools/Venn/). The results were exported for Cytoscape visualization (Shannon et al., 2003).

Functional Network Analysis

The Reactome Functional Interaction Network (Wu et al., 2010) was used to investigate the functional relationships between regulated glycoproteins and the genes reported as cancer drivers in one of the tree data resources COSMIS v71 (https://cancer.sanger.ac.uk/cosmic), Vogelstein list (Vogelstein et al., 2013) or HotNet analysis (Leiserson et al., 2015)from which a list of 661 altered genes was manually compiled. These cancer genes were found to be significantly mutated in cancer or heavily involved cancer development in the previous studies. The statistical significance of the functional relationships between regulated glycoproteins and altered genes was assessed in Reactome database and its 100 random instances. We used the switching algorithm implemented in the Random Network Plugin for Cytoscape (Wu et al., 2014) and the network graphs were visualized in Cytoscape (Shannon et al., 2003) (Table S7).

Data Stratification in Cross-Tumor dataset

Random Forest analysis

In the cross-tumor study, each carcinoma type was analyzed separately. Patient samples from a specific cancer type were compared with the group of corresponding control samples. For each cancer type, we trained a random forest classifier (ensemble of 500 decision trees) to stratify cancer versus corresponding control. We repeated this analysis with and without including clinical covariates (age, gender) and reported the ''out of bag'' estimate of accuracy for each model.

Feature prioritization via stability selection analysis

Additionally, we performed feature prioritization via stability selection analysis (Meinshausen and Buhlmann, 2010) including age and gender as covariates. We prioritized as top-scoring predictors all proteins with importance scores above a threshold. We used as threshold value the 1/10 of the highest importance score assigned within each carcinoma cohort.

Data validation on independent cancer cohorts

Independent cohorts for Panc and Proc

The discovered diagnostic protein predictors in the initial Panc and Proc cohorts (Figure 5) were evaluated in two independent plasma cohorts of the respective carcinomas, particularly since the random forest models for these two carcinomas displayed the highest ‘‘out-of-bag’’ accuracy in discriminating cancer patients from healthy controls in the discovery cohort. For each carcinoma type, a random forest model (ensemble of 500 decision trees) was fitted on the discovery cohort and subsequently evaluated on both discovery and validation cohorts. The analysis was repeated with and without including confounders (age and gender) as input features, resulting in similar performance results (consistently better when including confounders).

ROC analysis for top predictor proteins

ROC analysis, both non-adjusted and adjusted for age and gender, was performed for all proteins prioritized via stability selection in the Panc and Proc cohorts. In non-adjusted analysis, protein expression values from the discovery and validation cohorts were directly used. In adjusted analysis, a binomial logistic regression model, including confounders, was fitted on the discovery cohort and subsequently applied to both discovery and validation cohorts. p values for the AUC scores of individual proteins were computed via comparison with the chance level AUC = 50% using DeLong’s test for two uncorrelated ROC curves, and subsequently adjusted for the number of tests performed via FDR-based correction (Benjamini and Hochberg, 1995).

Differential expression analysis for top predictor proteins

Differential expression analysis, accounting for age and gender as confounders, was performed for all proteins prioritized via stability selection in the Panc and Proc cohorts. A linear regression model was used to explain protein expression levels conditioning on group assignment (i.e., control or cancer) and confounders such as age and gender. p values were obtained, indicating whether the regression coefficient for the group assignment (control or cancer) is significantly different than zero (Table S4). p values were adjusted for the number of tests performed via FDR-based correction (Benjamini and Hochberg, 1995).

DATA AND SOFTWARE AVAILABILITY

All the raw data of MS measurements, together with the input spectral library are available via the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) with the dataset identifier: PXD004998 (Reviewer account details: Username: reviewer08651@ebi.ac.uk, Password: A19Jz2hl). OpenSWATH related software is available on http://www.openswath.org/en/latest/.