Next generation pan-cancer blood proteome profiling using proximity extension assay

Mathias Uhlen (mathias.uhlen@scilifelab.se)
Science for Life Laboratory, Department of Protein Science, KTH Royal Institute of Technology
https://orcid.org/0000-0002-4858-8056

María Bueno Álvez
Science for Life Laboratory, Department of Protein Science, KTH Royal Institute of Technology
https://orcid.org/0000-0002-2669-7796

Fredrik Edfors
Stanford University https://orcid.org/0000-0002-0017-7987

Kalle von Feilitzen
Royal Institute of Technology

Martin Zwahlen
Science for Life Laboratory, Department of Protein Science, KTH Royal Institute of Technology

adil mardinoglu
KTH

Per-Henrik Edqvist
Uppsala University https://orcid.org/0000-0002-8330-0134

Tobias Sjöblom
Uppsala University https://orcid.org/0000-0001-6668-4140

Emma Lundin
Department of Immunology, Genetics and Pathology, Uppsala University

Natallia Rameika
Department of Immunology, Genetics and Pathology, Uppsala University

Tomas Axelsson
Department of Medical Sciences, Uppsala University

Mikael Åberg
Department of Medical Sciences, Clinical Chemistry and SciLifeLab, Uppsala University, Uppsala, Sweden

Jessica Nordlund
European Infrastructure for Translational Medicine; Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University https://orcid.org/0000-0001-8699-9959

Wen Zhong
Royal Institute of Technology https://orcid.org/0000-0002-7422-6104

Max Karlsson

Science for Life Laboratory, Department of Protein Science, KTH Royal Institute of Technology
https://orcid.org/0000-0002-7000-4416

Ulf Gyllensten
Uppsala University https://orcid.org/0000-0002-6316-3355

Fredrik Pontén
Department of Immunology, Genetics and Pathology, Uppsala University https://orcid.org/0000-0003-0703-3940

Linn Fagerberg
Department of Protein Science, Science for Life Laboratory, KTH-Royal Institute of Technology
https://orcid.org/0000-0003-0198-7137

Article

Keywords:

Posted Date: November 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2025767/v1

Version of Record: A version of this preprint was published at Nature Communications on July 18th, 2023. See the published version at https://doi.org/10.1038/s41467-023-39765-y.

Abstract

Cancer is a highly heterogeneous disease in need of accurate and non-invasive diagnostic tools. Here, we describe a novel strategy to explore the proteome signature by comprehensive analysis of protein levels using a pan-cancer approach of patients representing the major cancer types. Plasma profiles of 1,463 proteins from more than 1,400 cancer patients representing altogether 12 common cancer types were measured in minute amounts of blood plasma collected at the time of diagnosis and before treatment. AI-based disease prediction models allowed for the identification of a set of proteins associated with each of the analyzed cancers. By combining the results from all cancer types, a panel of proteins suitable for the identification of all individual cancer types was defined. The results are presented in a new open access Human Disease Blood Atlas. The implication for cancer precision medicine of next generation plasma profiling is discussed.

Introduction

A comprehensive characterization of blood proteome profiles in cancer patients can contribute to a better understanding of the disease etiology, resulting in earlier diagnosis, risk stratification and better monitoring of the different cancer subtypes. Cancer Precision Medicine aims to enable high-resolution individualized diagnosis by the use of molecular tools such as genomics, proteomics and metabolomics, with subsequent optimized treatment and monitoring of cancer patients. Of particular importance is the possibility to identify cancers early, allowing initiation of treatment and thereby improving patient outcome by avoiding tumor progression, metastasis, and emergence of treatment resistant tumors. When cancers are detected at an earlier stage, treatment is more effective and survival is drastically improved (1). As an example, according to US-based statistics (2), the five-year survival for breast cancer is 99% when detected at an early stage (localized), whereas survival decreases to only 30% when detected at later stages (metastasized). Similarly, the corresponding survival for ovarian cancer is 93% at early stage and 31% when detected at later stage (2). Based on this, several population screening programs have been initiated to identify cancer before symptoms arise, including screening for prostate cancer using PSA protein level (3), colorectal cancer by detecting blood in feces (4) and breast cancer using mammography (5). However, most population screening tests yield a relatively high number of false positives, causing an unwanted psychological burden for the individual and costly validation before a diagnosis can be confirmed. Moreover, population screening programs are still lacking for the majority of cancers and there is therefore a large need for a single screening test that could detect different forms of cancer at an early stage.

The main focus of Cancer Precision Medicine in the past decade has been to use genomics, involving next generation sequencing to explore the genetic make-up of individual cancers. Huge efforts have been made to gain genetic insight into tumors from patients, including The Cancer Genome Atlas (TCGA) (6, 7); the International Cancer Genome Consortium (ICGC) (8); and the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium (9). Although invaluable insights regarding the biology of individual cancers have been gained by these efforts, the genomics information has not led to substantial changes in therapeutic

regimes or facilitated screening for cancer in the population. Therefore, a move towards a multi-omics analysis has been suggested (10), including functional analysis and alternative assay platforms, such as proteomics using either dissected tumor biopsies or non-invasive body fluids (11).

An interesting approach in Cancer Precision Medicine is thus to use protein profiling to allow for liquid biopsy assays from blood. However, the staggering dynamic range in concentrations of blood proteins spanning at least ten orders of magnitude, with concentrations as low as pg/ml for cytokines, makes multiplex analysis involving even a handful of protein targets difficult. This has hampered the development of multiplex blood protein assays during the last few decades. This situation has now changed with the recent development of high-throughput platforms for sensitive proteomics assays in blood, such as Somascan (12) and Proximity Extension Assay (PEA) (13). These platforms allow thousands of target proteins to be analyzed simultaneously using a few microliters of blood with sensitivity to detect and quantify proteins present in low femtomolar amounts. This means that even proteins well below the detection level for mass spectrometry can now be accurately quantified and used for population screening.

Here, we describe a novel strategy for pan-cancer analysis in which the plasma profiles of patients with different types of cancer are compared to find cancer-specific signatures that can distinguish each type of cancer from other cancer types. This is in contrast to the standard procedure for cancer biomarker discovery, in which patients with a specific cancer are compared with healthy controls. Next Generation Blood Profiling (14), combining the antibody-based PEA with next generation sequencing, has been used to quantify protein concentrations in multiple cancer types. Samples from more than 1,400 cancer patients from a standardized biobank collection have been analyzed, along with a wealth of clinical meta data (15). Altogether, 12 cancer types including the most prevalent types such as colorectal-, breast-, lung- and prostate cancer, have been studied. AI-based prediction models were used to identify a panel of proteins associated with each of the analyzed cancers, with the primary aim to find protein panels with the ability to detect cancer signatures from blood plasma, and to further distinguish the type of cancer. The resulting panels have been further assessed to distinguish cancer patients from healthy individuals, as well as to distinguish patients with early disease.

Results

Description of the study workflow

The plasma proteome of 1,477 cancer patients and 74 healthy individuals were characterized using the Olink Explore 1536 Proximity Extension Assay (PEA) technology, allowing the quantification of 1,463 proteins using less than 3 microliters of plasma (13). Aiming to identify a plasma proteome signature for each of the cancers, we devised a workflow based on AI-based prediction models and differential protein expression analysis (Fig. 1). First, the results from all analyzed proteins (1,463 proteins) were used as a predictor of disease outcome to identify proteins reflecting disease status of each cancer sample. Next, differential expression analysis allowed us to select a subset of proteins that were up-regulated in one

cancer type compared to the other cancer types. Combining both results, we selected a set of relevant upregulated proteins for each cancer type and subsequently investigated whether a multiclassification model based on the selected proteins was able to distinguish the precise cancer type of a patient. We further validated the potential of the selected biomarkers by building cancer prediction models classifying each cancer from a healthy cohort, and finally confirmed that the classification allows for accurate identification of early-stage cancer patients.

The pan-cancer cohort

In this study, we have characterized the plasma proteome of a pan-cancer cohort from the Uppsala-Umeå Comprehensive Cancer Consortium (U-CAN) biobank (15), comprising 1,477 patients from twelve cancer types, including acute myeloid leukemia (n = 50), chronic lymphocytic leukemia (n = 48), diffuse large B-cell lymphoma (n = 55), myeloma (n = 38), colorectal cancer (n = 221), lung cancer (n = 268), glioma (n = 145), breast cancer (n = 152), cervical cancer (n = 102), endometrial cancer (n = 101), ovarian cancer (n = 134) and prostate cancer (n = 163). Plasma samples were collected at the time of diagnosis and before treatment were initiated. Clinical meta data regarding age, sex, diagnosis, and cancer stage or grade was available for the cancer samples (available in **Suppl. data 1**). For each cancer, the age distribution is shown in Fig. 2a.

Identification of cancer-specific proteins

The plasma protein levels of 1,463 protein targets were determined for each of the 1,477 patients to generate more than 2 million data points representing individual plasma protein levels across all the 12 cancer types. The main aim was not to identify a single protein marker to uniquely identify each cancer, but instead to identify a protein signature from each of the cancers to aid in a pan-cancer identification. An initial analysis revealed several upregulated and downregulated proteins in specific cancer types as exemplified in Fig. 2b. Some of these potential biomarkers are cancer-specific, such as Fms related receptor tyrosine kinase 3 (FLT3) in acute myeloid leukemia (ALL) and SLAM family member 7 (SLAMF7) in myeloma, while others are found to be elevated in two or more cancers, such as lymphocyte antigen 9 (LY9) with higher expression in both chronic lymphocytic leukemia (CLL) and myeloma. Interestingly, the B lymphocyte antigen receptor CD79b molecule (CD79B) exhibits elevated plasma levels in all four immune cell related cancers.

Pan-cancer prediction models

To identify proteins relevant for each cancer type, a disease prediction model was built for each cancer type (n = 12), respectively, using all measured proteins (n = 1,463) and 70% of the cancer patients as the training set. The control group in each model was composed of all the other cancer samples and was subsampled to include a similar number of patients to the modelled cancer. We compared the results obtained from two classification algorithms, random forest (RF) and a regularized generalized linear model (glmnet), both of which give an estimation of the overall importance of each protein to the model (range 0-100%). In **Fig. S1a**, a heatmap visualization shows the importance score for the 486 proteins that scored high (> 25% importance) in at least one of the cancer types by glmnet. We observed that some

of the cancer models have a higher number of proteins with high importance scores, suggesting that more proteins are relevant for a correct classification. Moreover, several proteins scored high (> 25% importance) in more than one cancer, as shown in the network visualization revealing relationships between the potential biomarkers in the different cancer types (**Fig. S1b**).

Since two prediction models were used for the analysis (random forest and glmnet), we compared the importance scores for each of the 1,463 protein targets in each of the cancer types given by the two models (Fig. 2c). For some cancers (e.g. glioma), one protein is given the highest score by both models with considerably lower scores for the other proteins (< 50%), while in other cancers there is a continuum of importance scores that can be mainly consistent (e.g. in myeloma) or somewhat less consistent (e.g. in endometrial cancer). The importance scores for each protein across the 12 cancer types using both models are found in **Suppl. data 2**. In general, the two models predict the same proteins with similar importance, but the random forest models tend to be more conservative overall, consistently selecting fewer important proteins.

Evaluation of cancer-specific prediction models

The performance of the cancer prediction models was evaluated using the 30% of the data excluded from the model training. In Fig. 3a, the classification of all patients in the test cohort using both prediction algorithms were performed, scoring each plasma sample for the probability to come from a specific cancer type. We found that both models can separate samples from all the specific cancers, with particular high confidence for three of the immune cell related cancers: AML, CLL and myeloma, all having area under the curve (AUC) (16) of 0.99-1 (**Fig. S2a**). For most other cancers, the models correctly predict the cancer origin of most of the patients, although some overlap was observed, as exemplified by colorectal-, lung- and prostate cancer. The two prediction models yield similar results, with no significant difference in AUC for most cancers (DeLong test p-value > 0.05). However, since the glmnet results yielded a significantly higher AUC for the classification of lung, colorectal, breast and cervical cancers (**Fig. S2a**), we decided to continue the exploratory analysis using the glmnet algorithm.

Characterization of differentially expressed proteins across the cancers

To further investigate the cancer-specific proteome profiles, differential expression analyses were performed in a setting where each cancer was compared to all other cancers. For the male and female cancers, only samples with the same sex were compared. The up- and down-regulated proteins in each cancer are summarized in the volcano plots in **Fig. S3a**. The results showed a distinct differentially expressed proteome for each of the cancers. The differential analysis identified, as expected, similar potential biomarkers as the prediction models, exemplified by FLT3 for AML and CXCL17 and FKBP1 for lung cancer. However, some of the proteins, such as EPO in AML, show high significance in the differential analysis, but were not ranked high by either the random forest and the glmnet prediction models.

In Fig. 3b, the number of upregulated proteins that are shared by different cancer types is shown for the complete set of proteins analyzed. As expected, there were a large number of upregulated proteins shared by the four immune cell related cancers (AML, CLL, lymphoma and myeloma), in many cases consisting of proteins related to immune-related functions. However, most overlapping proteins were observed for lung and colorectal cancer. This observation might reflect common features between these two cancer types, such as adenocarcinoma origin and a high fraction of high-grade tumors with likely similar host inflammatory response. A functional gene ontology (GO) analysis was also performed on the up-regulated proteins for each of the cancer types (**Fig. S3b**). As expected, the up-regulated proteins in the immune cell related cancers (AML, CLL and lymphoma) are related to immune processes, while breast, endometrial and prostate cancer had an over-representation of cell adhesion proteins and both lung and colorectal cancer had an over-representation of apoptotic-related proteins.

Selection of a panel with cancer-specific proteins

Combining the previous results, we sought to identify a panel of proteins based on the ranking from the glmnet models and relevant to each of the analyzed cancers. We only included proteins identified as upregulated by differential expression analysis, aiming to include at least three proteins per cancer and all proteins with more than 50% overall importance as indicated by the cancer prediction models. Based on these criteria, a panel of 83 proteins capturing cancer-specific profiles was selected (Fig. 4a) and all individual proteins are listed in **Suppl. data 3** along with the results from the disease prediction models and differential expression. Lung- and prostate cancer contributed to the largest number of proteins in the panel, 18 and 14, respectively, whereas only three protein targets each were selected for AML, glioma, myeloma and ovarian cancer.

In Fig. 4b, the average plasma levels of the 83 selected protein members of the panel are visualized across all cancer types. Most of the selected proteins had a higher level in only one cancer, while some had high protein levels in multiple cancers. For example, CXADR like membrane protein (CLM), selected to identify endometrial cancer, also showed elevated plasma levels in myeloma patients. In **Fig S4a**, the plasma levels of the most important protein for each cancer are shown across all cancer patients. Only two of the proteins were given a high importance score (> 50%) by the prediction model in more than one cancer. Both FKB prolyl isomerase 1B (FKBP1B) and peroxiredoxin 5 (PRDX5) had higher plasma levels in lung- and colorectal cancer as compared to all the other cancers (**Fig. S4b**), and were also selected independently by the models for both of these cancer types (Fig. 4a). Interestingly, FKBP1B is involved in immunoregulation and protein folding and has previously been linked to colorectal cancer (17) but not to lung cancer. Similarly, PRDX5 has an antioxidant function in normal and inflammatory conditions and although not previously suggested to be involved in cancers, several other proteins of the peroxiredoxin family have been linked to lung and colorectal cancers in transcriptomics analysis of cancer cell lines (18, 19). In **Fig. S5**, a UMAP visualization based on profiles from the 83 panel proteins for all samples shows the relationship within and between samples from different cancer patients. The plasma levels of the panel proteins resulted in a clear separation of the immune cell related cancer patients in distinct regions,

as well as glioma, a good separation of ovarian and cervical cancer patients, and a large overlap of colorectal and lung cancer patients.

To explore the tissue origin of each protein marker, a transcriptomic analysis comparing relative expression levels across was performed across 36 major tissue types (**Fig. S6**) using data from the Human Protein Atlas (20, 21). Out of the 83 protein markers, 74 are expressed in the healthy tissue corresponding to the origin of respective cancer. Thus, the markers for the immune-related cancers are highly expressed in healthy lymphoid tissues, the glioma markers in the healthy brain and some of the lung cancer markers in healthy lung. Interestingly, several other markers are instead expressed and secreted by the liver (22), synthesizing many plasma proteins, and several other markers are shown to be expressed across many tissues.

Classification of the pan-cancer cohort based on the selected protein panel

A multiclass classification based on the glmnet algorithm was used to explore the sensitivity and specificity of the panel proteins to predict a specific cancer and to assess the model's ability to distinguish the different cancer types. Comparative receiver operating characteristic (ROC) analyses were performed for each cancer type in which the specificity/sensitivity measured as AUC was determined for different number of proteins. These analyses included (i) all proteins (n = 1,463), (ii) those selected in the panel (n = 83), (iii) the three most important proteins per cancer and (iv) the single most important protein per cancer. The results (Fig. 5a) show that the panel of 83 proteins can identify the right cancer with both high selectivity and sensitivity with AUC ranging between 0.93 and 1 for all cancer types. The analysis using all proteins gave only slightly better results, while the use of only the top 3 proteins in each cancer gave somewhat less reliable results. The lowest performance scores were obtained when using only the top protein for each of the 12 cancers.

The AUC values for the different protein numbers are summarized for each of the cancers (Fig. 5b), showing the advantage of selecting multiple proteins to identify cancer patients from blood plasma. However, the results also suggest that a panel with only a handful of protein markers can achieve the same prediction reliability as using all proteins. Here, our results demonstrate that a panel of only 83 proteins yields highly promising results (AUC) for simultaneous identification of all 12 cancer types. As shown in Fig. 5c, there is some overlap in the prediction results for some of the cancers, such as lung and colorectal cancer, while for other cancers, such as glioma and immune-related cancers, the samples have a high probability of being correctly classified.

Performance of classification of cancer samples from a healthy cohort

An important question is how well the model based on the pan-cancer study can distinguish cancer patients from healthy individuals. To investigate this, for each of the 12 cancer types, a new cancer prediction model was built but this time including 74 healthy individuals previously studied as part of a

wellness study (23, 24) as the control group instead of all of the other cancers. As described above, each of the cancers contributed to the panel with a different number of proteins (3–18) and these new models were based only on these specific proteins. We again used 70% of the samples as the training set and the remaining 30% to test the performance of the model. The results for four of the cancers are shown in Fig. 6a-d and all cancers in **Fig. S7**. For CLL (Fig. 6a), the model can distinguish cancer patients from healthy controls with total accuracy (AUC = 1). Similarly, the same analysis for colorectal- (Fig. 6b), ovarian- (Fig. 6c) and lung cancer (Fig. 6d), respectively, shows high accuracy with all AUC results above 0.83, demonstrating that the protein panel can distinguish cancer patients from healthy individuals with high accuracy. However, caution is required since the wellness panel was sampled and analyzed in a separate study, thus sample bias can not be ruled out. These results suggest that the protein panel and the prediction model are suitable to identify patients with the analyzed cancer types as well as distinguish cancer patients from healthy individuals (without a cancer diagnosis).

Stratification of patients with cancers of different stages

An important quest in the field of Cancer Precision Medicine is to aid clinicians to indicate the stage of the cancer. For some cancers in this study, a relatively large number of patients had stage data available and therefore we investigated whether the protein panel could stratify patients into stages for these cancer types. In Fig. 6e, we show four examples of proteins where the plasma levels correlate with disease stage, including (i) CD22 used to identify CLL patients; (ii) amphiregulin (AREG) in colorectal cancer patients; (iii) arbhydrolsase domain containing 14B (ABHD14B) in lung cancer patients; and (iv) the ovarian cancer biomarker Progestagen associated endometrial protein (PAEP). These examples demonstrate the possibility to perform stage stratification simply by analysing selected plasma protein levels, but further analyses in additional cohorts are needed to demonstrate the validity of the protein panel for cancer stage stratification.

Classification of early-stage cancer samples

The most important objective in the field of cancer precision medicine is to identify cancer at an early stage to provide successful therapeutic intervention and patient survival. To assess the ability of the protein panel to distinguish early-stage cancer from healthy individuals, we focused on patients with early stage colorectal and lung cancer, where the sample sizes of patients with less advanced disease (stage 1) are relatively large. The performance of the same prediction models trained for these two cancer types were now tested using early stage (stage 1) cancer samples compared to healthy samples. In Fig. 6f **(top)**, the cancer probability score for stage 1 lung cancer patients is compared with the corresponding score for healthy individuals. A clear difference in score is shown for most samples and the AUC-score (Fig. 6f, **bottom**) for separating stage 1 lung cancer patients from healthy individuals is 0.79. Similarly, for the early stage colorectal cancer patients, a clear difference is predicted by the protein panel model (Fig. 6g, **top**), and the AUC-score (Fig. 6g, **bottom**) is 0.78. This highlights the potential of the selected biomarker panel to identify early stage colorectal and lung cancer patients, although more in depth analysis in independent cohorts is warranted.

The open access disease blood profile section of the Human Protein Atlas

A new Human Disease Blood Atlas resource has been created as part of the Blood Protein section of the Human Protein Atlas (www.proteinatlas.org). This new section contains more than 2 million data points representing the individual blood level for the 1,463 target proteins in the 1,477 cancer patients. A discussion about the different proteins important for the prediction model across the different cancers is also included in the new Disease Blood Atlas. The resource is available without restrictions (open access) to allow researchers both from academia and industry to gain this basic information about blood protein profiles from both healthy- and disease conditions.

Discussion

Here, we describe a novel strategy based on next generation plasma profiling to explore the cancer proteome signatures by comprehensively exploring the protein levels in patients representing most major cancer types. The assay platform allows thousands of proteins to be quantitatively analyzed using only a few microliters of blood opening up new opportunities for Precision Cancer Medicine. The plasma levels of each individual protein have been determined for more than 1,400 cancer patients representing 12 different cancer types, and the results are presented in a new open access Human Disease Blood Atlas.

Applying AI-based disease prediction models based on all measured proteins allowed us to identify a set of proteins associated with each of the cancers studied. A prediction model based on a restricted set of 83 upregulated proteins was built to evaluate the accuracy of the classification of pan-cancer samples and the analysis showed that each cancer has a distinct plasma proteome profile. It is interesting to observe the huge increase in prediction performance when using the protein panel (n = 83) as compared to the use of only the most significant protein marker for each cancer (n = 12). This demonstrates the added advantage of using a panel of blood proteins as exemplified by patients with breast cancer for which individual markers are not selective, but the prediction model using multiple proteins yields a much more accurate classification. The panel allowed the stratification of plasma samples from most cancer types with high sensitivity and specificity and it was also able to detect patients with early disease, as exemplified by stage 1 patients in lung and colorectal cancers.

The proteins in the panel include well-known markers for some of the cancers but also proteins with no previous connection to cancer as discussed in the open access Human Disease Blood Atlas. In this context, it is interesting that the cancer specific elevation of the panel proteins in blood plasma could reflect several causes, such as an increase of leakage or secretion from the tumor or surrounding tissue itself, or due to the bodily response to the tumor. Overall, the gene expression on tissue level seems to indicate that most of the markers identified here are already produced in the healthy tissue, and that their elevation in respective cancer may reflect increased leakage or secretion to plasma by the tumor or surrounding tissue. However, a more in-depth analysis is needed to explain the causal relationship between the proteins and the respective cancer types.

It is important to point out that individual variation of protein plasma levels in both healthy- and disease states calls for validation of the findings using an independent assay platform as well as using independent patient cohorts. Since even a highly selective assay used in a population screening still could generate a large number of false positives, when millions of individuals are screened for presence of cancer, it is particularly important to rule out false positives, which could cause considerable and unnecessary stress for the individual. It is thus important to complement the screening with independent assays to validate positive findings. Fortunately, such assays exist for the cancer types analyzed here, such as mammography for breast cancer, blood in feces and colon spectroscopy for colorectal cancer, radiological examination and/or tissue-based analysis of biopsies for many other cancers. This makes it possible to use the pan-cancer blood assay presented here for the initial population screening and subsequent validation of the positive samples with independent and less cost-effective assay platforms.

It is also important that the protein panel presented here should be validated in additional cohorts to confirm the validity in each of the cancer types. For example, in two earlier studies of blood from glioma patients (25, 26), only a few upregulated proteins were found and none of these were significantly upregulated here. This demonstrates the importance of several independent studies before establishing a pan-cancer protein panel. Of particular importance is validation in a large background of non-diseased individuals to establish the breadth of false positives. It is also desirable to have the results validated by independent technical platforms, such as sandwich (27) or Somascan (12) assays.

Here, we analyzed patient plasma from 12 of the major cancer types. However, it is of course interesting to expand the analysis to add other frequent and important cancers to the pan-cancer strategy, such as liver, kidney and pancreas cancers. Similarly, it is also valuable to compare the cancer profiles reported here with plasma profiles from patients having other diseases. Our aim in the near future is to be able to report such studies as part of the open access Human Disease Blood Atlas resource for patients in the field of cardiovascular, autoimmune, neurological and infectious disease, respectively. It is also interesting to add more protein targets to the analysis and such larger panels are now available for exploration by both the PEA (13) technology, which currently can analyze 3000 targets, and the Somascan platform (12), including 7000 targets.In summary, we describe a novel strategy for exploration of protein profiles in blood to allow simultaneous identification of each of 12 common cancers using only a minute amount (few microliters) of blood. This open access Human Blood Disease resource allows researchers to explore the cancer profiles of individual proteins across most of the major cancer types. Since the assay can be combined with simple sample collection formats, such as dried blood spots, the results open up the possibility for a cost-effective pan-cancer population diagnosis using a panel of proteins to identify most of the common cancers in a single assay. It is tempting to speculate that such population screening could be organized to allow the discovery of cancers much earlier and thus help clinicians to start treatment of cancer patients at earlier stages, as compared to today. However, to make this feasible, it is of outmost importance to have available secondary independent assays to validate the initial screening to identify and rule out patients with wrongly diagnosed cancer (false positives).

Declarations

Acknowledgements

We thank the entire staff of the Human Protein Atlas program and the Science for Life Laboratory (SciLifeLab) for their valuable contributions. We thank Per Eriksson and Lena Beckman for analysis of the OLink data and Camilla Jysky and Lina Dahlberg for collection of clinical samples. We are grateful to the various cancer cohort chairs, including Malin Enblad (CR), Oscar Simonson (LC), Karin Glimskär-Stålberg (Gyn), Martin Högberg (Leukemi), Gunilla Enblad (Lymphoma), Henrik Lindman (Breast), Göran Hesselager (Glioma) and Michael Häggman (Prostate). This work was supported by WCPR grant from Knut and Alice Wallenberg Foundation, the SciLifeLab & Wallenberg Data Driven Life Science Program (grant: KAW 2020.0239) and Swedish Research Council Grant 2020-06175. The computations and data handling were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

Author contributions

MU conceived and designed the study. FE, PE, TS and FP collected and contributed data to the study. MBA, MK, FE, AM, WZ, LF and MU performed the data analysis. EL, NR, TA, MÅ, JN and UG processed the samples and performed the PEA analysis. KvF and MZ created the database portal. MU and MBA drafted the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Data availability statement

The protein levels aggregated per cancer type for each protein are available for download on the Human Protein Atlas resource download page (https://proteinatlas.org/about/download), and all protein levels are presented on the individual protein summary pages of the new Human Disease Blood Atlas.

Code availability statement

All code necessary for the data analysis and visualization is available on request.

References

1. D. Crosby *et al.*, Early detection of cancer. Science **375**, eaay9040 (2022).
2. Surveillance Epidemiology and End Results (SEER) Program.
3. D. Ilic *et al.*, Prostate cancer screening with prostate-specific antigen (PSA) test: a systematic review and meta-analysis. BMJ **362**, k3519 (2018).
4. U. Ladabaum, J. A. Dominitz, C. Kahi, R. E. Schoen, Strategies for Colorectal Cancer Screening. Gastroenterology **158**, 418–432 (2020).
5. A. Yala *et al.*, Optimizing risk-based breast cancer screening policies with reinforcement learning. Nat Med **28**, 136–143 (2022).
6. N. Cancer Genome Atlas Research *et al.*, The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113–1120 (2013).
7. C. Hutter, J. C. Zenklusen, The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell **173**, 283–285 (2018).
8. J. Zhang *et al.*, The International Cancer Genome Consortium Data Portal. Nat Biotechnol **37**, 367–369 (2019).
9. Icgc Tcga Pan-Cancer Analysis of Whole Genomes Consortium, Pan-cancer analysis of whole genomes. Nature **578**, 82–93 (2020).
10. A. A. Friedman, A. Letai, D. E. Fisher, K. T. Flaherty, Precision medicine for cancer with next-generation functional diagnostics. Nat Rev Cancer **15**, 747–756 (2015).
11. R. Akbani *et al.*, A pan-cancer proteomic perspective on The Cancer Genome Atlas. Nat Commun **5**, 3887 (2014).
12. L. Gold, J. J. Walker, S. K. Wilcox, S. Williams, Advances in human proteomics at high scale with the SOMAscan proteomics platform. N Biotechnol **29**, 543–549 (2012).
13. L. Wik *et al.*, Proximity Extension Assay in Combination with Next-Generation Sequencing for High-throughput Proteome-wide Analysis. Mol Cell Proteomics **20**, 100168 (2021).
14. W. Zhong *et al.*, Next generation plasma proteome profiling to monitor health and disease. Nat Commun **12**, 2493 (2021).
15. B. Glimelius *et al.*, U-CAN: a prospective longitudinal collection of biomaterials and clinical information from adult cancer patients in Sweden. Acta Oncol **57**, 187–194 (2018).
16. T. Fawcett, An introduction to ROC analysis. Pattern Recognition Letters **27**, 861–874 (2006).
17. L. Wang, X. Jiang, X. Zhang, P. Shu, Prognostic implications of an autophagy-based signature in colorectal cancer. Medicine (Baltimore) **100**, e25148 (2021).
18. M. K. Kim *et al.*, Patients with ERCC1-negative locally advanced esophageal cancers may benefit from preoperative chemoradiotherapy. Clin Cancer Res **14**, 4225–4231 (2008).
19. W. Lu *et al.*, Peroxiredoxin 2 is upregulated in colorectal cancer and contributes to colorectal cancer cells' survival by protecting cells from oxidative stress. Mol Cell Biochem **387**, 261–270 (2014).
20. M. Uhlen *et al.*, Tissue-based map of the human proteome. Science **347**, 1260419 (2015).

21. M. Uhlen *et al.*, A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. Science **366**, (2019).
22. M. Uhlen *et al.*, The human secretome. Sci Signal **12**, (2019).
23. G. Bergstrom *et al.*, The Swedish CArdioPulmonary BioImage Study: objectives and design. J Intern Med **278**, 645–659 (2015).
24. A. Tebani *et al.*, Integration of molecular profiles in a longitudinal wellness profiling cohort. Nat Commun **11**, 4487 (2020).
25. C. B. Holst *et al.*, Plasma IL-8 and ICOSLG as prognostic biomarkers in glioblastoma. Neurooncol Adv **3**, vdab072 (2021).
26. H. Jaksch-Bogensperger *et al.*, Proseek single-plex protein assay kit system to detect sAxl and Gas6 in serological material of brain tumor patients. Biotechnol Rep (Amst) **18**, e00252 (2018).
27. E. Engvall, P. Perlmann, Enzyme-linked immunosorbent assay, Elisa. 3. Quantitation of specific antibodies by enzyme-labeled anti-immunoglobulin in antigen-coated tubes. J Immunol **109**, 129–135 (1972).

Online Methods

The pan-cancer study cohort

Plasma samples from cancer patients were obtained from the U-CAN biobank which collects samples from consenting patients diagnosed at the Akademiska hospital in Uppsala as part of the clinical routine and with a high degree of standardization (*1*). Plasma samples were obtained from treatment-naïve patients taken around the time of their diagnosis. Plasma was prepared from whole blood by centrifugation at 2.400 g for seven minutes at room temperature, after which the plasma was aliquoted into several 220 µl vials and immediately frozen for long-term storage at -80°C. Exclusion criteria included any concurrent or previous cancer within the last five years, and arm-to-freezer time exceeding 360 minutes. Diagnosis, stage, age, gender and other variables were obtained from the U-CAN database and the patient's clinical records. The study was approved by the Swedish Ethical Review Authority (EPM dnr 2019-00222). The research was in line with donor consents in U-CAN (28631533, EPN Uppsala 2010-198 with amendments).

The Wellness healthy cohort

Plasma samples from healthy individuals (39 males and 35 females) were selected from a Swedish SciLifeLab SCAPIS Wellness Profiling (S3WP) study as described previously (*2, 3*). The S3WP program includes longitudinal samples from 101 healthy individuals aged 50-64, recruited from the prospective observational Swedish CArdioPulmonary bioImage Study (SCAPIS). The study was approved by the Ethical Review Board of Goteborg, Sweden (registration number 407-15), and all participants provided written informed consent. The study protocol conforms to the ethical guidelines of the 1975 Declaration of Helsinki.

Measurement of protein levels

The protein levels were measured in plasma using the Olink Explore PEA technology (*4*), which uses antibody-binding capabilities to detect the levels of 1,463 targets in plasma coupled with next-generation sequencing (NGS) readout. A total of 1,472 proteins were targeted using specific antibodies, including 1,463 unique proteins related on inflammation, oncology, cardiometabolic and neurology, as well as controls. Each antibody was conjugated separately with two complementary probes, and distributed in four separate 384-plex panels. Each panel contained three control assays (interleukin-6 (IL6), interleukin-8 (CXCL8), and tumor necrosis factor (TNF) used for quality control (QC). In brief, the PEA workflow started with an overnight incubation to allow the conjugated antibodies to bind to the corresponding proteins in the samples. The incubation was followed with an extension and pre-amplification step when the hybridization and extension of complementary probes takes pace. The extended DNA was then amplified by PCR and further indexed to allow the preparation of libraries, which were then sequenced using Illumina's NovaSeq platform. The counts obtained from the sequencing run were subjected to a quality control and normalization procedure. Here, internal controls introduced at different steps were used to reduce intra-assay variability. These include an incubation control consisting of a non-human antigen measured with the same technology, an extension control consisting of an antibody conjugated to a unique pair of probes which are in proximity and is expected to produce a positive signal, and a control in the amplification step consisting of a double-stranded DNA sequence which is expected to produce a positive signal independent of the amplification step. Additionally, external controls such as negative control (buffer sample) and plate controls (pool of plasma) were used to establish a limit of detection (LOD) and adjust levels between plates, respectively. Finally, two known samples acted as sample controls to calculate the precision of the measurements. After quality control and normalization, the data was provided in an arbitrary Normalized Protein eXpression (NPX) unit, which is on a log2 scale and where a high NPX value can be interpreted as a high protein level.

Disease prediction

The caret R package (v 6.0.90) (*5*) was used to build multivariate classification models for each of the cancer types. First, the cancer data was split in 70% for training purposes and 30% for testing purposes using the createDataPartition() function in caret. The data was imputed with the preProcess() function in caret using the "knnImpute" method.

Multivariate prediction models were built for each the different cancers in two settings: 1) based on all measured proteins (n= 1,463) and having a control group composed of a subset of patients from all other cancers; and 2) based on a selected set of proteins and having healthy patients as a control. In both cases, the cancer prediction model was built on the training set using a 5-fold cross-validation and built-in parameter tuning. The contribution of each protein to the model was retrieved using the varImp() function in the caret package. A multiclass classification model was built using the caret train() function to achieve a simultaneous classification of the 12-cancer types based on all cancer samples in the training set and selected panel of proteins, with 5-fold cross validation strategy and parameter tunning.

ROC analyses

The performance of the prediction models was evaluated using the samples in the testing set, which were not part of the training of any of the models. ROC analyses were performed to assess the sensitivity and specificity of the classification, summarized as AUC scores. The pROC R package (v 1.18.0) was used for binary classifications and multiROC (v 1.1.1) was used for multiclass classification. Statistical significance for differences in AUC were calculated using the DeLong test (*6*) implementation in the pROC package, using paired tests for correlated ROC curves and unpaired test when for independent ROC curves, using p-value of 0.05 as threshold for significance.

Differential expression analysis

The differential protein expression was assessed using a two-sided t-test coupled with Benjamini-Hochberg multiple hypothesis correction (*7*), with a significance threshold of 0.05 for adjusted p-values. The adjusted p-values and difference in average expression per group were summarized in volcano plots for each of the analyzed cancers. Enrichment analysis of up-regulated protein sets were performed using the clusterProfiler package (version 3.18.1) (*8*). The enricher() function in clusterProfiler was used to perform overrepresentation analysis against the biological annotations from Gene Ontology (GO) biological processes (BP) (*9*), with subsequent p-value adjustment using the Benjamini-Hochberg method (*7*) and using adjusted p-value < 0.05 as threshold for significance.

Tissue transcriptomic analysis

Tissue transcriptomes were downloaded from the Human Protein Atlas v21 (*10*) and each protein marker was mapped to its corresponding gene. The normalized transcript per million (nTPM) was used as a quantitative measure of expression level, and an nTPM ≥ 1 was used as a cutoff to call a gene expressed in a given tissue.

Data visualization

Data visualization was performed in R (version 4.0.3) (*11*), using the ggplot2 (version 3.3.5) (*12*), ggbeeswarm (version 0.6.0) (*13*), ggraph (version 2.0.5) (*14*), ggrepel (version 0.9.1) (*15*), ggridges (version 0.5.3) (*16*), ggplotify (version 0.1.0) (*17*), igraph (version 1.2.6) (*18*), pheatmap (version 1.0.12) (*19*), patchwork (version 1.1.1) (*20*), pcaMethods (version 1.82.0) (*21*), tidygraph (version 1.2.0) (*22*), UpSetR (version 1.4.0) (*23*) and uwot (version 0.1.10) (*24*) packages. For the heatmap visualization, data was rescaled to a 0-1 scale and hierarchical clustering was performed using the “ward.D2” method. The limma R package (version 3.46.0) (*25*) was used to correct for batch differences for the comparison between the UCAN and Wellness cohorts, and to correct for sex effects for UMAP visualization of cancer samples. The first 30 components resulting from principal component analysis (PCA) were used as input data for UMAP visualization. The figures were assembled in Affinity designer (v 1.10.0.1127).

Method reference list

1. B. Glimelius *et al.*, U-CAN: a prospective longitudinal collection of biomaterials and clinical information from adult cancer patients in Sweden. *Acta Oncol* **57**, 187-194 (2018).

2. G. Bergstrom *et al.*, The Swedish CArdioPulmonary BioImage Study: objectives and design. *J Intern Med* **278**, 645-659 (2015).

3. A. Tebani *et al.*, Integration of molecular profiles in a longitudinal wellness profiling cohort. *Nat Commun* **11**, 4487 (2020).

4. L. Wik *et al.*, Proximity Extension Assay in Combination with Next-Generation Sequencing for High-throughput Proteome-wide Analysis. *Mol Cell Proteomics* **20**, 100168 (2021).

5. M. Kuhn, Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28**, 1-26 (2008).

6. E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837-845 (1988).

7. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society* **57**, 289-300 (1995).

8. T. Wu *et al.*, clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021).

9. M. Ashburner *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).

10. M. Uhlen *et al.*, Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

11. R Core Team, R: A Language and Environment for Statistical Computing. *MSOR connections* **1**, (2014).

12. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

13. E. Clarke , S. Sherrill-Mix, ggbeeswarm: Categorical Scatter (Violin Point) Plots. (2017).

14. T. L. Pedersen, ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. (2021).

15. K. Slowikowski, ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. (2021).

16. C. O. Wilke, ggridges: Ridgeline Plots in 'ggplot2'. (2021).

17. G. Yu, ggplotify: Convert Plot to 'grob' or 'ggplot' Object. (2021).

18. G. Csardi, T. Nepusz, The igraph software package for complex network research. (2006).

19. R. Kolde, pheatmap: Pretty Heatmaps. (2019).

20. T. L. Pedersen, patchwork: The Composer of Plots. (2020).

21. W. Stacklies, H. Redestig, M. Scholz, D. Walther, J. Selbig, pcaMethods–a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* **23**, 1164-1167 (2007).

22. T. L. Pedersen, tidygraph: A Tidy API for Graph Manipulation. (2020).

23. N. Gehlenborg, UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets. (2019).

24. J. Melville, uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction. (2020).

25. M. E. Ritchie *et al.*, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).

Figures

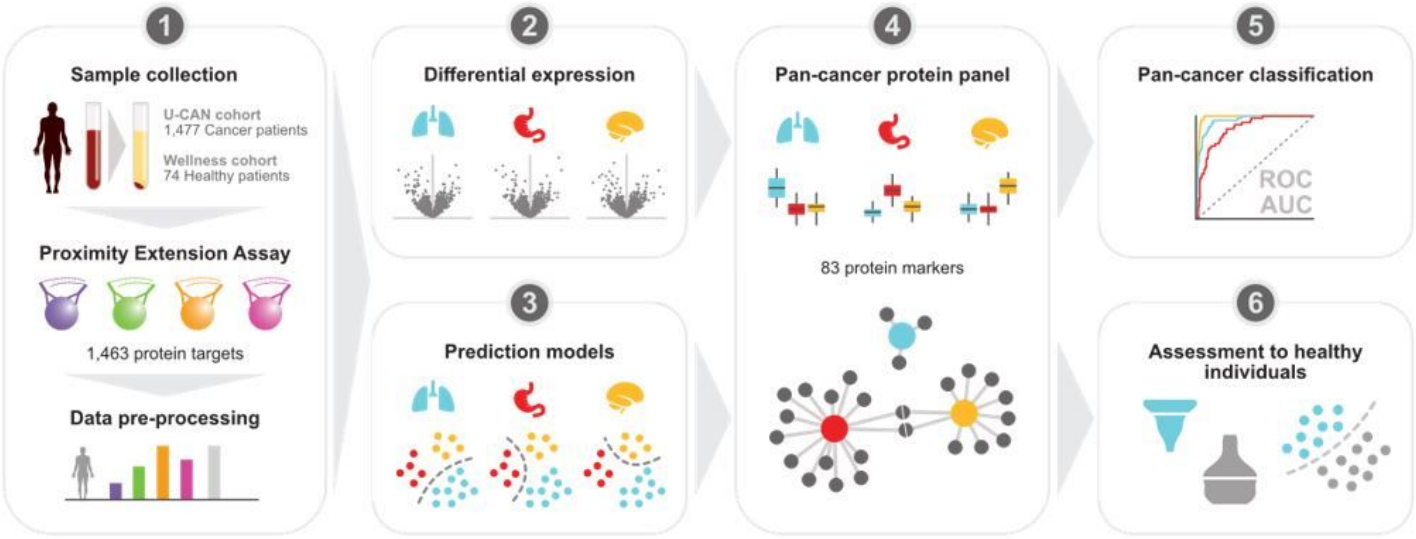

Figure 1

Overview of the pan-cancer study. Schematic representation of the workflow used to identify a pan-cancer biomarker panel for cancer classification. Blood plasma from 1,477 cancer patients and 74 healthy individuals was analyzed using Proximity Extension Assay. Differential expression analysis and prediction models were performed to identify a pan-cancer protein panel. The model for cancer classification was generated using machine learning techniques (70% of the data). The performance of the resulting pan-cancer protein panel was tested against a model test set (30% of the data) and ultimately compared against healthy individuals.

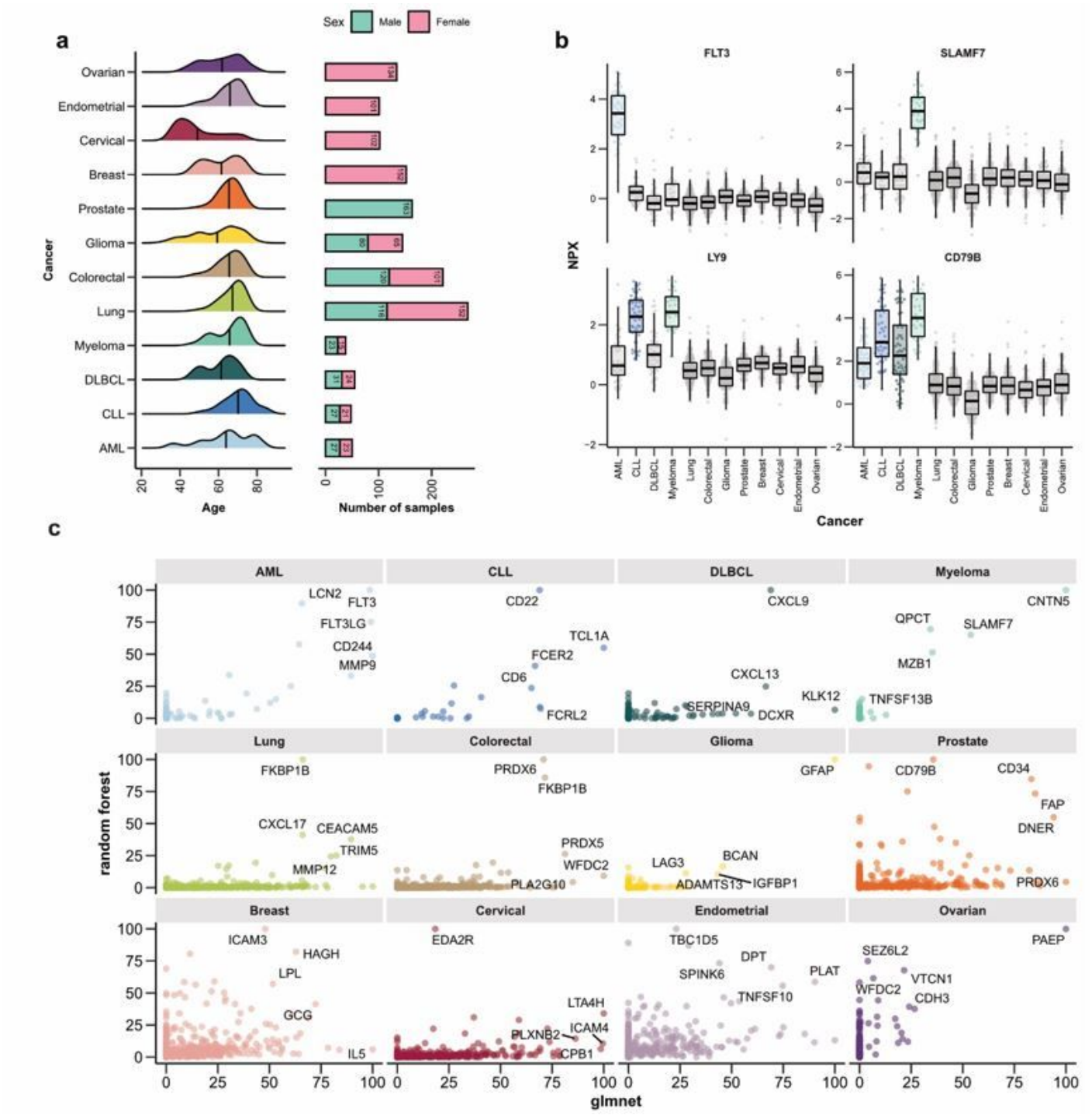

Figure 2

The study cohort and the prediction models to classify each cancer type. a, Age distribution and number of patients included for each cancer in the study. **b**, Examples of protein levels (NPX, y-axis) for four example proteins across the 12 cancer types (x-axis). **c**, Scatter plot between the importance score for two prediction models (random forest (rf) and a regularized generalized linear model (glmnet)).

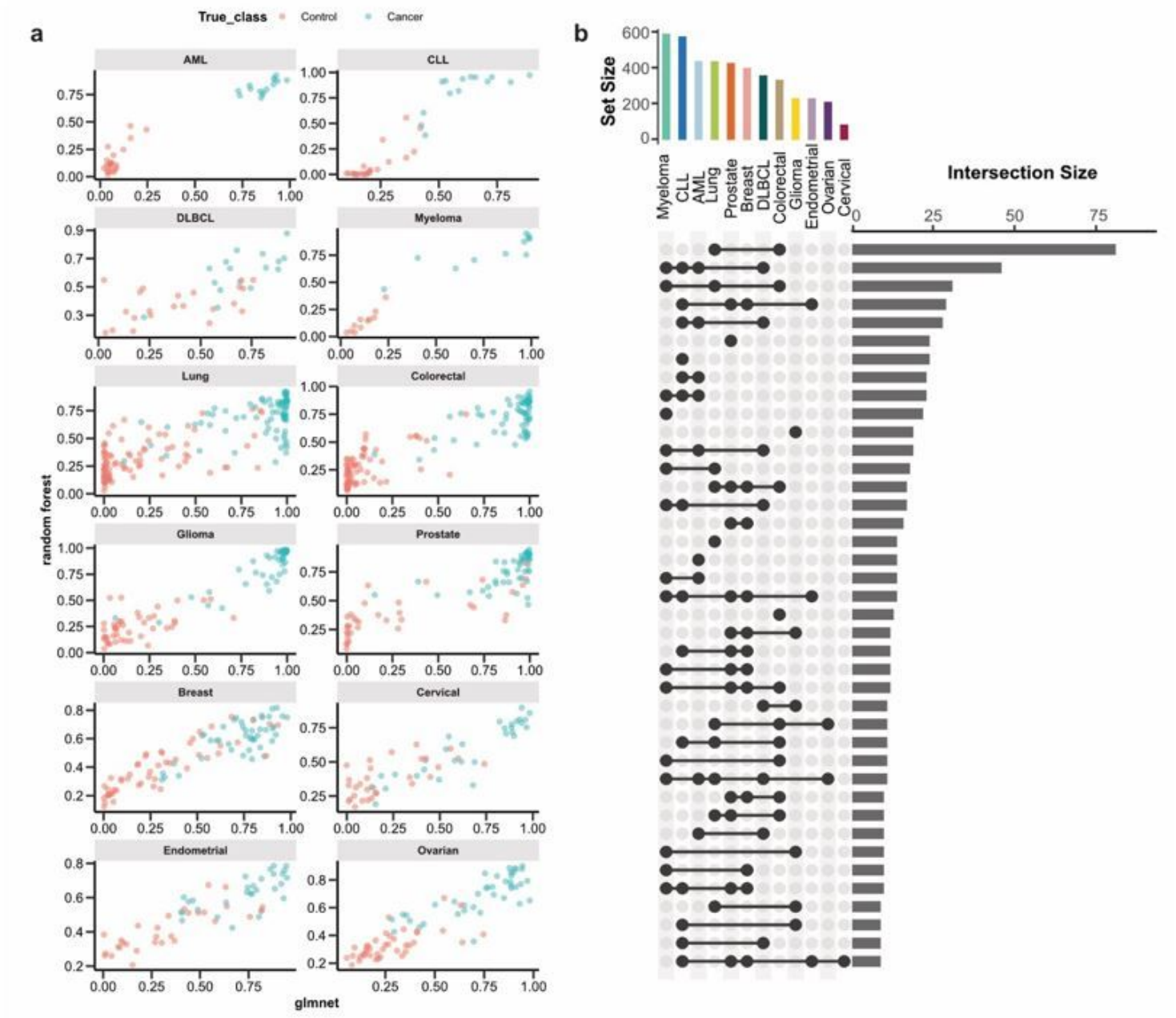

Figure 3

Performance of prediction models and differential expression analysis. a, Comparison of cancer probabilities for samples in the test set for glmnet (n=12) and rf (n=12) cancer models. **b**, Upset plot showing the number of upregulated proteins shared by the different cancer types.

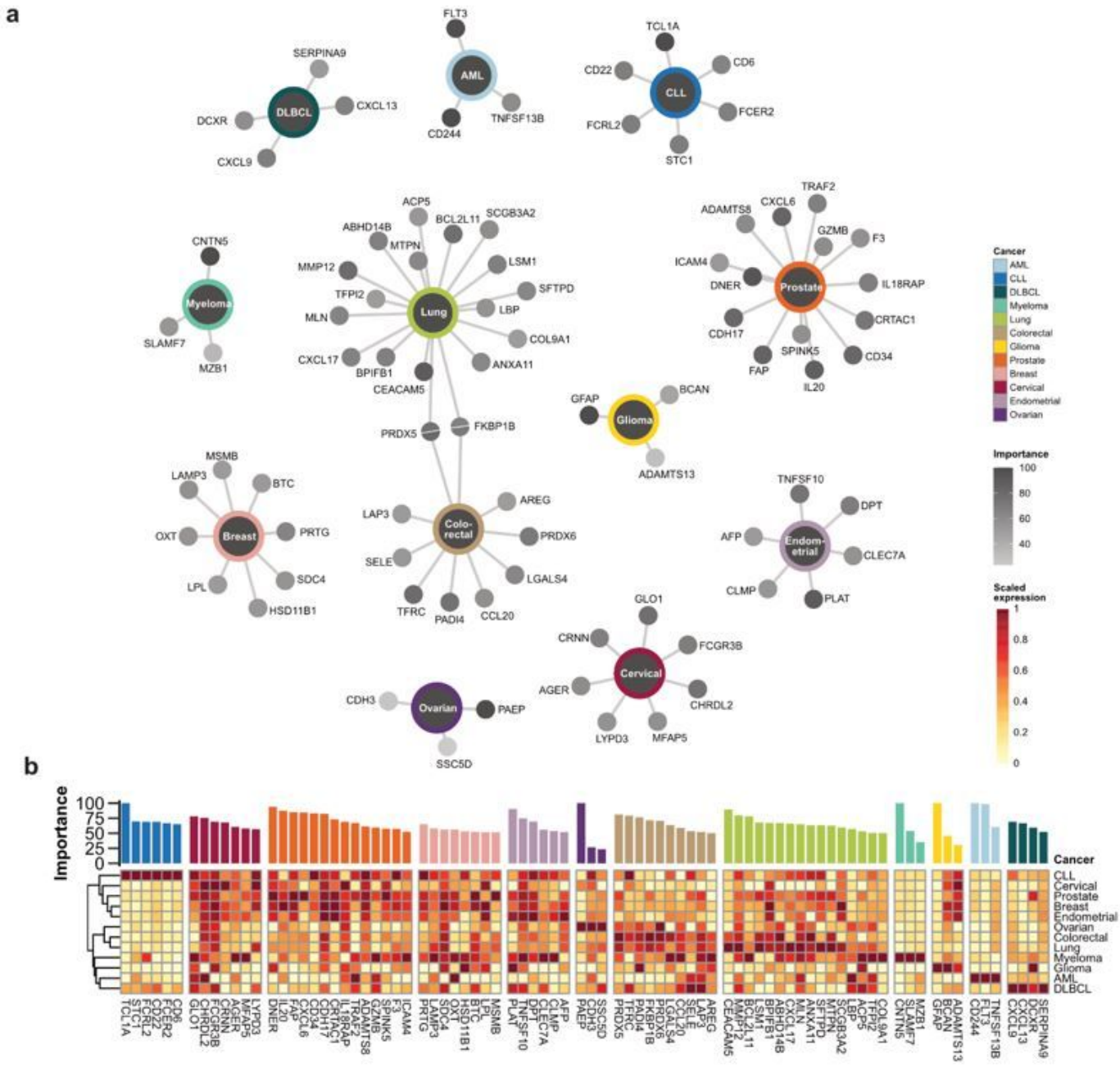

Figure 4

Pan-cancer protein panel. a, Nework visualization of proteins included in the panel. Protein nodes are colored according to the importance score in the specific cancer. **b**, Summarized expression profiles of panel proteins across the cancer types.

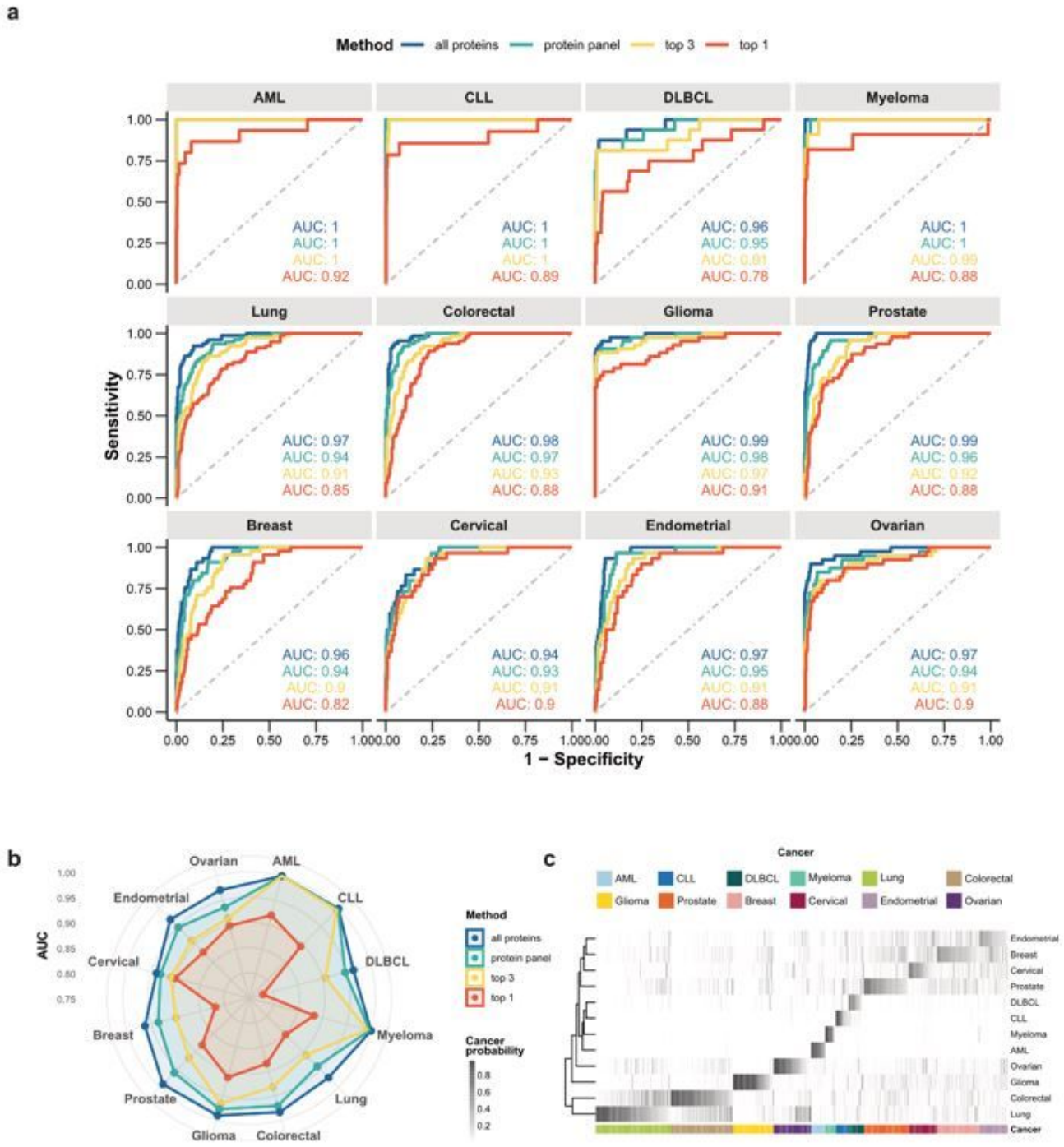

Figure 5

Multiclassification of the pan-cancer test cohort. a, ROC curves for the model for the different cancers based on all proteins (n=1,463), the selected protein panel (n=83) and the top 3 (n=3) and the most important protein (n=1) for each och the 12 cancers. **b**, Summary of the AUC for the different cancers based on different numbers of proteins. **c**, Cancer probabilities for samples in the test set in the pan-cancer classification model.

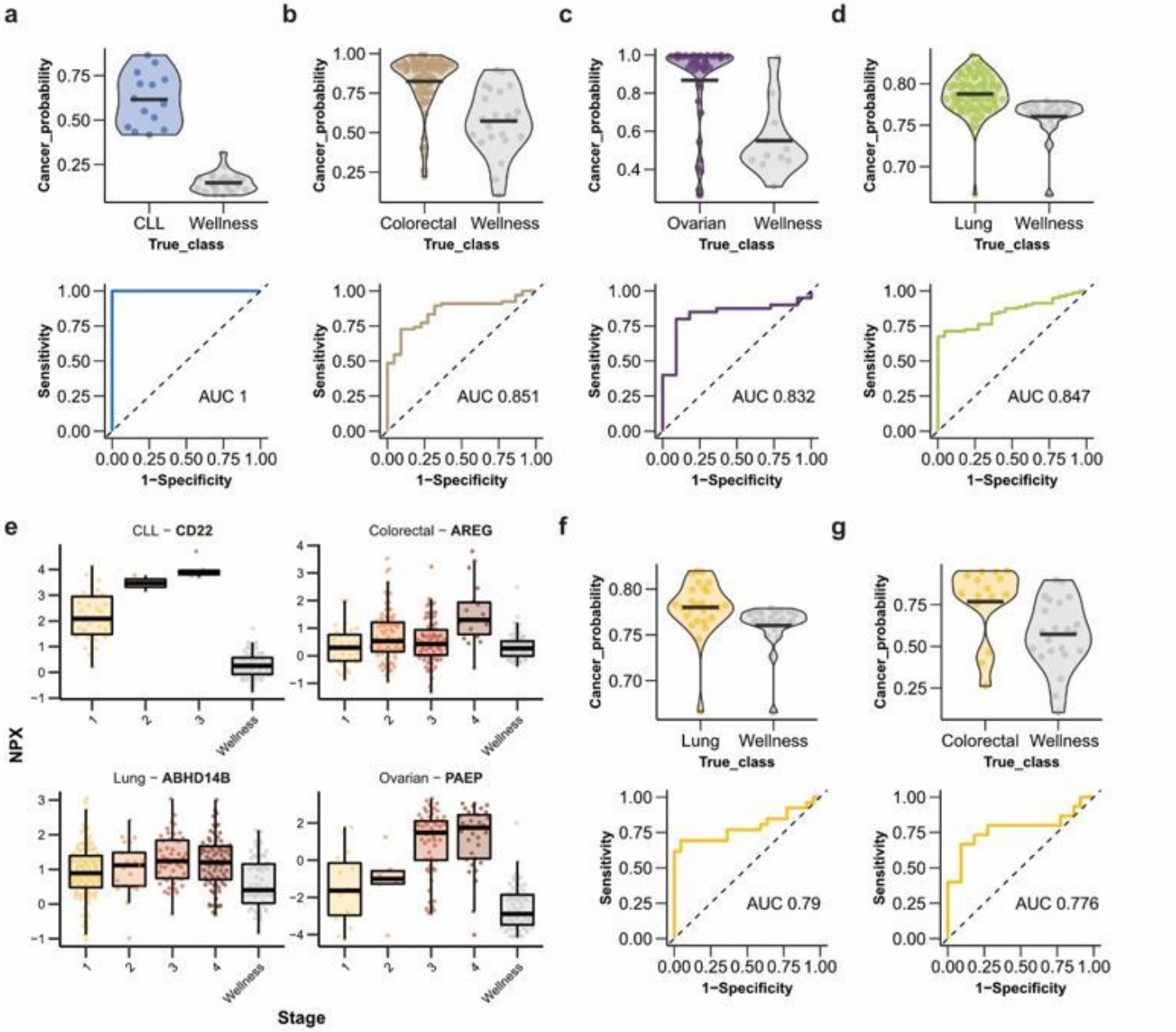

Figure 6

Classification of cancer samples from a healthy cohort based on the selected protein panel. Model results showing the cancer probability for cancer and healthy individuals (top) and the ROC curve with AUC score (bottom) for **a**, CLL, **b**, colorectal cancer, **c**, ovarian cancer, **d**, lung cancer. **e**, Protein levels of four different biomarkers across cancer stages. Model results showing the cancer probability for cancer and healthy individuals (top) and the ROC curve with AUC score (bottom) for **f**, early-stage lung cancer and **g**, early-stage colorectal cancer.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BuenoetalDataS1patientmetadata.xlsx
- BuenoetalDataS2predictionmodels.xlsx
- BuenoetalDataS3proteinpanel.xlsx
- BuenoetalDataS4expressiondata.xlsx
- Buenoetalsupplementarymaterial.docx