BMJ Oncology

Novel proteomics-based plasma test for early detection of multiple cancers in the general population

Bogdan Budnik, Hossein Amirkhani, Mohammad H Forouzanfar, Ashkan Afshin

To cite: Budnik B, Amirkhani H, Forouzanfar MH, *et al*. Novel proteomics-based plasma test for early detection of multiple cancers in the general population. *BMJ Oncology* 2024;**3**:e000073. doi:10.1136/bmjonc-2023-000073

Received 13 April 2023
Accepted 17 October 2023

ABSTRACT

Objective Early detection of cancer is crucial for reducing the global burden of cancer, but effective screening tests for many cancers do not exist. This study aimed to develop a novel proteome-based multi-cancer screening test that can detect early-stage cancers with high accuracy.

Methods and analysis We collected plasma samples from 440 individuals, healthy and diagnosed with 18 early-stage solid tumours. Using proximity extension assay, we measured more than 3000 high-abundance and low-abundance proteins in each sample. Then, using a multi-step statistical approach, we identified a limited set of sex-specific proteins that could detect early-stage cancers and their tissue of origin with high accuracy.

Results Our sex-specific cancer detection panels consisting of 10 proteins showed high accuracy for both males (area under the curve (AUC): 0.98, 95% CI 0.96, 1) and females (AUC: 0.983, 95% CI 0.95, 1.00). At stage I and at the specificity of 99%, our panels were able to identify 93% (95% CI 79%, 100%) of cancers among males and 84% (95% CI 68%, 100%) of cancers among females. Our sex-specific localisation panels consisted of 150 proteins and were able to identify the tissue of origin of most cancers in more than 80% of cases. The analysis of the plasma concentrations of proteins selected showed that almost all the proteins were in the low-concentration part of the human plasma proteome.

Conclusion The proteome-based screening test showed promising performance compared with other technologies and could be a starting point for developing a new generation of screening tests for the early detection of cancer.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Early detection of cancer can enhance patient outcomes significantly. Recent efforts to achieve this have primarily focused on imaging techniques and liquid biopsy approaches such as circulating tumour DNA testing. Proteomics, the large-scale study of proteins, represents a potential avenue for early cancer detection. However, the use of plasma proteins as biomarkers for early cancer detection has been challenging due to the complexity of the proteome and the lack of sensitivity in detecting low-abundance proteins.

WHAT THIS STUDY ADDS

⇒ This study demonstrates the potential utility of a plasma proteome-based test for the early detection of 18 solid tumours, representing all major human organs of origin. We found that a limited set of plasma proteins could differentiate cancer samples from normal ones, and even distinguish between different types of cancers with high accuracy. We also found that the most useful biomarkers for early-stage cancer detection were proteins present in low concentrations in the plasma proteome. Moreover, the study provides evidence that cancer protein signatures are likely sex-specific. These findings pave the way for a cost-effective, highly accurate, multi-cancer screening test that can be implemented on a population-wide scale.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The findings of this study can have important implications for cancer screening policies. The developed proteome-based diagnostic test outperforms existing technologies, providing a more efficient approach for early cancer detection. This could reshape screening guidelines, making this plasma test a standard part of routine check-ups. Moreover, the identification of low-abundance proteins and sex-specific protein signatures as sensitive biomarkers opens new avenues of research in proteomics and cancer biology. Further validation in larger population cohorts is needed to establish the reliability and generalisability of our findings. Ultimately, the implementation of such a test in healthcare systems could greatly reduce both health and financial burdens associated with cancer.

INTRODUCTION

Cancer is a leading cause of mortality globally, accounting for one in every six deaths.[1] In the absence of established risk factors for many cancers, early detection and early treatment remain the cornerstone of clinical and public health strategies for reducing the global burden of cancer and saving lives. However, currently, no effective test exists for the early detection of many cancers. Nearly 60% of cancer-related deaths are due to cancers for which no screening test exists.[2] Additionally, existing screening tests (ie, colonoscopy, CT scan, mammography, pap test) have major

Linked

► http://dx.doi.org/10.1136/bmjonc-2023-000184

Check for updates

Novelna Inc, Palo Alto, California, USA

Correspondence to
Dr Ashkan Afshin; ashkan@novelna.com and Dr Bogdan Budnik; bbudnik@novelna.com

limitations, including invasiveness, high cost and low accuracy for early stages.

Liquid biopsy, the analysis of biomarkers in non-solid specimens, has emerged as a promising approach for developing novel biomarkers.[3] In recent years, efforts have been made to develop a genomics-based liquid biopsy test for screening multiple cancers at once.[4 5] For example, a blood test has been developed to identify the presence of over 50 cancers based on their methylation signatures in cell-free DNA.[6] However, these genomics-based multi-cancer tests have shown low sensitivity for early-stage cancers (<50%).[7] Additionally, they are too expensive (US$>500) to be covered by most insurance companies and incorporated as routine screening tests in the healthcare system.

Protein biomarkers in the blood have the potential to be used for early detection and ongoing monitoring of diseases, but the current options lack sensitivity and specificity.[8] For example, prostate specific antigen is the most common biomarker for prostate cancer screening.[9] There are several protein biomarkers used to guide breast cancer treatment including oestrogen receptor, progesterone receptor, human epidermal growth factor receptor 2 and Ki-67. In addition, cancer antigen 15-3 (CA 15-3) and cancer antigen 27.29 (CA 27.29) are sometimes used to monitor response to breast cancer treatment. Cancer antigen 125 (CA 125) is the most widely used biomarker to monitor response to treatment in women diagnosed with ovarian cancer. Carcinoembryonic antigen (CEA) is often used to monitor colorectal cancer, especially in people who have already been treated. Another biomarker called carbohydrate antigen 19-9 (CA 19-9) can also be elevated in colorectal cancer. CA 19-9 is the most common biomarker for pancreatic cancer. However, it is not specific to pancreatic cancer and can be elevated in other gastrointestinal cancers and conditions. Alpha-fetoprotein is a biomarker sometimes used in the diagnosis and monitoring of liver cancer. Progastrin-releasing peptide is often used as a biomarker for small-cell lung cancer. Other potential biomarkers like CYFRA 21-1 and CEA are used for non-small cell lung cancer.

In this paper, we explore the potential use of plasma proteins as biomarkers for solid tumours (excluding melanoma) in specific organs and the need to search for biomarkers in the depths of the proteome that are currently undetectable. We discuss the lower sensitivity of current protein assays compared with nucleic acid detection methods and the need to be able to detect very small amounts of proteins to identify early stages of cancer growth through liquid biopsy.

METHODS

Study design

We collected plasma samples from 440 patients diagnosed with 18 distinct types of cancer, as well as from healthy individuals (online supplemental figure S1). We focused on early-stage common solid tumours where the early detection followed by medical/surgical treatment can significantly enhance the survival of the patients. Then, we used Olink's proximity extension assay (PEA) technology to measure proteins in the plasma samples. Using the machine learning approaches, we identified protein biomarkers for detection of cancer as well as identification of the site of origin of the cancer.

Sample collection

The EDTA plasma samples used in this analysis were provided by the Ukraine Association of Biobank (UAB).[10] The samples were collected from mostly asymptomatic patients who have gone under routine medical check-up and were identified to have early stage tumours. All the patients included in this study were treatment-naïve, and their plasma samples were collected before tumour removal or any other form of treatment. The normal samples were collected from healthy blood donors. A summary of the number and type of samples is outlined in table 1.

The UAB was established in 2017 with the goal of growing and developing the biobank network across Ukraine. The network encompasses biobanks from the main Institutes of Medical Sciences in the country and operates under the guidance of the ESBB, ISBER and NCI guidelines. The UAB has established policies and necessary documents such as the Patient Consent Policy, Patient Information Sheet, Biobank Consent Form and Sample Application Form to ensure ethical and transparent operations in medical institutions and hospitals.

UAB places a strong emphasis on maintaining the highest ethical standards in its operations and has implemented strict bioethical policies and state-of-the-art procedures. The anonymity of donors is protected, and they are fully informed about the purpose of medical research and the potential benefits of scientific discoveries through legal consent documentation. All specimens are obtained with the fully informed and signed consent of donors.

Every specimen collected by the UAB is processed following a rigorous Standard Operating Procedure. Quality control is conducted on all collected tissue samples, and the specimens are handled with utmost care and precision from the time of excision to storage and shipment to ensure that the customers receive only the highest quality specimens.

Protein measurement

The protein levels in plasma were determined by using the Olink Explore 3072 technology. The investigators performing the assays were blind to the patients' diagnosis. A full description of Olink's PEA technology has been reported elsewhere.[11–13] The list of the proteins measured, and their characteristics are provided in online supplemental table S1. The protein measurement for all samples other than females' normal plasma samples were performed in the same run. The plasma proteins in females' normal samples were measured separately

Table 1 Characteristics of the samples included in the study

Cancer	Subtype	Number	Sex (% male)	Age (mean/SD)	Stage (N)		
					I	II	III
Bladder	Urothelial carcinoma	22	55	63.1 (12.8)	8	14	–
Breast	Invasive ductal carcinoma	22	0	46.3 (5.4)	–	–	22
Brain	Astrocytoma	22	27	47.1 (8.7)	–	–	–
Cervical	Cervix uteri carcinoma	22	0	50.1 (8.1)	22	–	–
Colorectal	Adenocarcinoma	22	41	57.8 (5.9)	–	16	6
Kidney	Kidney renal papillary cell carcinoma	22	23	57.5 (5.2)	9	13	–
Liver	Hepatocellular carcinoma	22	41	56.7 (7.4)	–	15	5
Lung	Small and non-small cell	44	45	57.6 (6.3)	–	22	22
Oesophagus	Squamous cell carcinoma	22	59	61.4 (8.9)	–	22	–
Osteosarcoma	Steoblastic osteosarcoma	22	23	40.4 (6)	–	22	–
Ovarian	Epithelial	22	0	53.7 (6.6)	22	–	–
Pancreas	Ductal adenocarcinoma	22	55	62.7 (6.1)	–	22	–
Prostate	Adenocarcinoma	22	100	67.4 (8)	7	15	–
Stomach	Adenocarcinoma	22	55	69.2 (5.8)	13	9	–
Testis	Seminoma	22	100	52.2 (7.1)	–	22	–
Thyroid	Medullary thyroid cancer	22	36	52.6 (7)	–	22	–
Uterus	Endometrial cancer	22	0	52 (5.6)	17	5	–
Normal	–	22	100	59.6 (7.2)	–	–	–
Normal	–	22	0	50.2 (6)	–	–	–

and standardised to other samples using Olink's bridging protocol.

Briefly, this technology employs antibody-based detection to assess the levels of 3072 target proteins in plasma. The antibodies were each conjugated with two complementary probes and divided into four separate 384-plex panels. Each panel included three control assays for quality control purposes (interleukin 6 (IL-6), interleukin-8 (CXCL8) and tumour necrosis factor). The process began with an overnight incubation to allow the conjugated antibodies to bind to the target proteins in the samples. This was followed by an extension and preamplification step, where the hybridisation and extension of the complementary probes took place. The extended DNA was then amplified through PCR and indexed to prepare the libraries, which were sequenced using the Illumina NovaSeq platform. The counts obtained from the sequencing were subject to a quality control and normalisation procedure, which involved the use of internal controls to reduce intra-assay variability. These included an incubation control consisting of a non-human antigen, an extension control consisting of a unique pair of probes, and an amplification control consisting of a double-stranded DNA sequence. Additionally, external controls such as a negative control (buffer sample) and plate controls (pool of plasma) were used to determine the limit of detection and adjust levels between plates, respectively. Finally, two known samples were used as sample controls to calculate the precision of the measurements. After quality control and normalisation, the data was provided in a Normalised Protein eXpression (NPX) unit, which is on a log2 scale and indicates a high protein level with a high NPX value.

The analytical performance of Olink's panels has been carefully validated for sensitivity, dynamic range, specificity, precision and scalability (online supplemental table S2). Analytical measuring range was defined by the lower limit of quantification and upper limit of quantification and reported in pg/mL. The high dose hook effect (a state of antigen excess relative to the reagent antibodies resulting in falsely lower values) has also been determined for each analyte.

All assays have been thoroughly validated for precision (repeatability and reproducibility). Intra-assay variation (within-run) has been calculated as the mean CV for six individual samples, within each of seven separate runs during the validation studies. Inter-assay variation (between-runs) was calculated as the mean CV, for the same six individual samples, among seven separate runs during the validation studies.

Statistical analysis

Our approach involved two main steps. In the first step, we searched for a limited number of proteins that could identify any cancer in its early stages. In the second step, we classified each type of cancer against the others to find a cancer-specific signature for localisation (ie, tissue of origin). We conducted each step separately for male

and female samples and calculated a probability-based score at each step for a person's protein array sample. Then, we classified the person based on the calculated score. As with other biomarker studies, our experiment size was relatively small.[14] Therefore, we used a multi-step approach to define a minimal set of proteins (also known as the 'best set) that could classify samples from a stable model and be generalisable. To select the features forming the minimum set, we apply logistic regression with L1 penalty to 100 bootstrap samples of the original dataset. The proteins exhibiting the highest number of non-zero coefficients are selected. This allowed us to identify the proteins with stronger association and best statistical significance for (a) differentiating all cancers from healthy individuals (pan-cancer analysis) and (b) differentiating each cancer from other solid tumours (localisation analysis). The use of the L1 penalty ensured the sparsity of the selected biomarkers, preventing the simultaneous selection of correlated biomarkers. We performed a leave-one-out validation to evaluate the performance of the model with selected features. We used the area under the curve (AUC) of the receiver operating characteristic curve as a statistical measure to assess the performance of the biomarker panels. For cancer localisation, the proteins were evaluated to identify the site of origin and/or subtype of the cancer for each cancer versus other cancers in the dataset. The top cancer specific proteins with the strongest associations with the target cancers were selected iteratively, which resulted in different size minimum sets for the cancers with the overall highest average performance. Then, for a given sample, the predicted probabilities of it belonging to different cancers were calculated using the selected proteins, and the cancer with the highest probability score was determined as the predicted cancer.

We transformed each problem into a binary classification one. For the pan-cancer detection problem, the positive class encompassed all cancer samples, and the control class included normal samples. For the localisation of each cancer, the samples of that specific cancer were considered positive cases, while other cancer samples were considered negative cases.

We employed logistic regression with an L1 penalty to identify a sparse set of informative biomarkers. Each protein was standardised to zero mean and unit SD to ensure comparable coefficients. The strength of the L1 penalty was determined using stratified fivefold cross-validation on a logarithmic scale ranging from 1e-4 to 1e4. The desired number of biomarkers was then selected based on the magnitude of their absolute coefficients. The pre-processing and modelling were carried out using the Scikit-learn Python library.[15]

After identifying the biomarkers, linear classifiers were trained on the training data using the selected biomarkers. In the pan-cancer detection component, the predicted probability for each sample is compared with a threshold to determine whether it is normal or cancerous. The threshold is set to attain the desired level of specificity. The positive samples from the detection component then proceed to the localisation component, which consists of one model for each type of cancer. The sample is assigned to the class with the highest predicted probability.

For data visualisation, we used matplotlib (V.3.6.0),[16] seaborn (V.0.12.0)[17] and plotly (V.5.10.0).[18] Pre-processing and modelling were performed in Python (V.3.9.13), using scikit-learn (V.1.1.2),[15] pandas (V.1.4.4),[19] numpy (V.1.23.3)[20] and statsmodels (V.0.13.2).[21]

After biomarkers selection and to maximise the use of data in our training models, we used the leave-one-out method as our evaluation method. In this approach, the number of folds equals the number of instances in the data set. We applied the learning algorithm once for each instance, using all other instances as a training set and using the selected instance as a single-item test set.

PATIENT AND PUBLIC INVOLVEMENT

Considering that this was an initial proof-of-concept study, patients and the public were not directly involved in its design. However, they will play a critical role in the dissemination of these preliminary results and in the subsequent validation phase of the research.

RESULTS

Effect of protein correlation and sex

Out of the 3072 proteins that were analysed, 287 did not pass the quality measurements and were excluded from further analysis. The abundance of proteins showed varying correlation levels, with positive correlations potentially being a result of shared biological pathways. Many protein pairs displayed high positive correlation, as indicated by the correlation matrix (figure 1). However, highly correlated proteins could make the analysis unstable, despite being related to cancer. Additionally, different biological pathways contribute to the initiation and progression of cancer, which highlights the need to capture diverse cancer types through different pathways and potentially less correlated proteins.[22] Therefore, we designed our analysis to select the most informative proteins from internally correlated sets, allowing us to classify a larger number of patients with higher accuracy.

We found that the protein-cancer association varied significantly between males and females (figure 1). A simple comparison between cancer and normal samples for men and women showed a poor correlation across protein profiles. For the male cohort, 80% of proteins with a p value below 0.05 had no significant difference in females (online supplemental table S2). Similarly, 83% of proteins in females with a p value of less than 0.05 showed no difference in males. When considering a more stringent p value threshold of 0.001, these proportions increase to 97.8% for males and 99.1% for females. An analysis of volcano plots showed minimal overlap between top 100 proteins selected based on their differential expression and their p value. An examination of

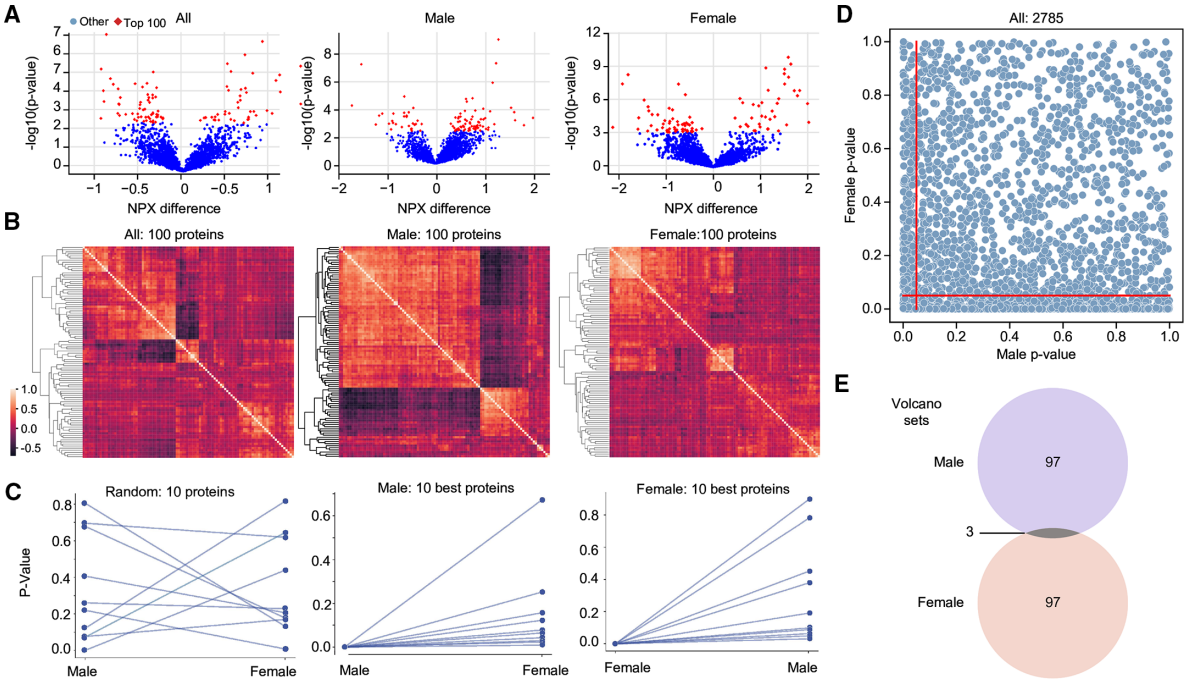

Figure 1 The difference between protein biomarkers in males and females. The volcano plots showing differential abundances of proteins among both sexes, males and females. Top 100 proteins based on their p value highlighted in red (A). The correlation matrix of the top 100 proteins based on p value among males, females and both sexes (B). Ranking of the top 10 proteins based on p values in males and females as well as the ranking of 10 randomly selected proteins in males and females (C). The scatter plot of p-values for each protein among males and females (D). The Venn diagram of top 100 proteins based on volcano plots in males and females (E). NPX, Normalised Protein eXpression.

the top 10 proteins based on p values revealed variations in the order of protein importance between males and females. In males, the 10 proteins with the lowest p values differentiating cancer from normal were IGFBP2, CKB, CEACAM16, MMP1, ENPP5, ELN, ITIH3, WNT9A, KRT19 and SSC4D, collectively averaging a p value of less than 0.0001. Conversely, these same proteins exhibit an average p value of 0.15 in females. On the other hand, in females, the 10 proteins with the lowest p values (C3, SEMA4C, NAPRT, GBP2, MAEA, SEPTIN9, PSMD5, IL1A, CLTA and CNST) had an average p value of less than 0.00001 for females and 0.31 for males.

Optimal number of proteins for cancer detection and localisation

Finding the best sex-specific sets to identify a cancer was performed in two steps. At the first step, we detected the protein signature of any cancer (pan-cancer/any cancer classifier) to classify any cancer from normal, followed by the second step, identifying the tissue of origin of cancer and cancer subtypes (ie, small cell and non-small cell cancers of lung, and cervical and endometrial cancers of uterine).

For each step, we tested several numbers of proteins and expected to see an increasing performance of the model by AUC with more proteins to capture different cancer populations. The model performance increased quickly with adding a few more proteins but after 10 proteins in any cancer model, no more improvement in AUC was observed (figure 2). A similar phenomenon was observed in the cancer localisation step. Since the performance of different predictive models for specific cancers plateaued at different numbers, we employed an algorithm to pick the proteins-cancer pairs for additional proteins if improved overall AUC more. For the localisation panels, the highest performance was observed at the level of 150 proteins. As expected, different numbers of proteins were picked as the best set by the algorithm. Moreover, some proteins were helpful in localising more than one cancer.

Sex-specific cancer detection panel

Our final detection panels for males and females each consisted of 10 proteins that were differentially expressed among normal and cancer plasma samples (figure 2). The distribution of proteins included in the detection panels in males and females and their corresponding AUC have been shown in online supplemental figure S2 and online supplemental table S3, respectively. Each protein of the panel alone had a low to medium detection accuracy but when assessed in combination with other proteins as a panel they achieved a very high accuracy in detection of early-stage cancers (figure 3). The proteins in the panel showed low to medium correlation indicating each protein contributing new information and presenting a different pathway to the panel.

Overall, at the detection step, our protein panels showed high sensitivity and specificity among males and females (figure 3). At the specificity of 99%, the overall sensitivity of our test was 90% (CI 84%, 96%) among males and

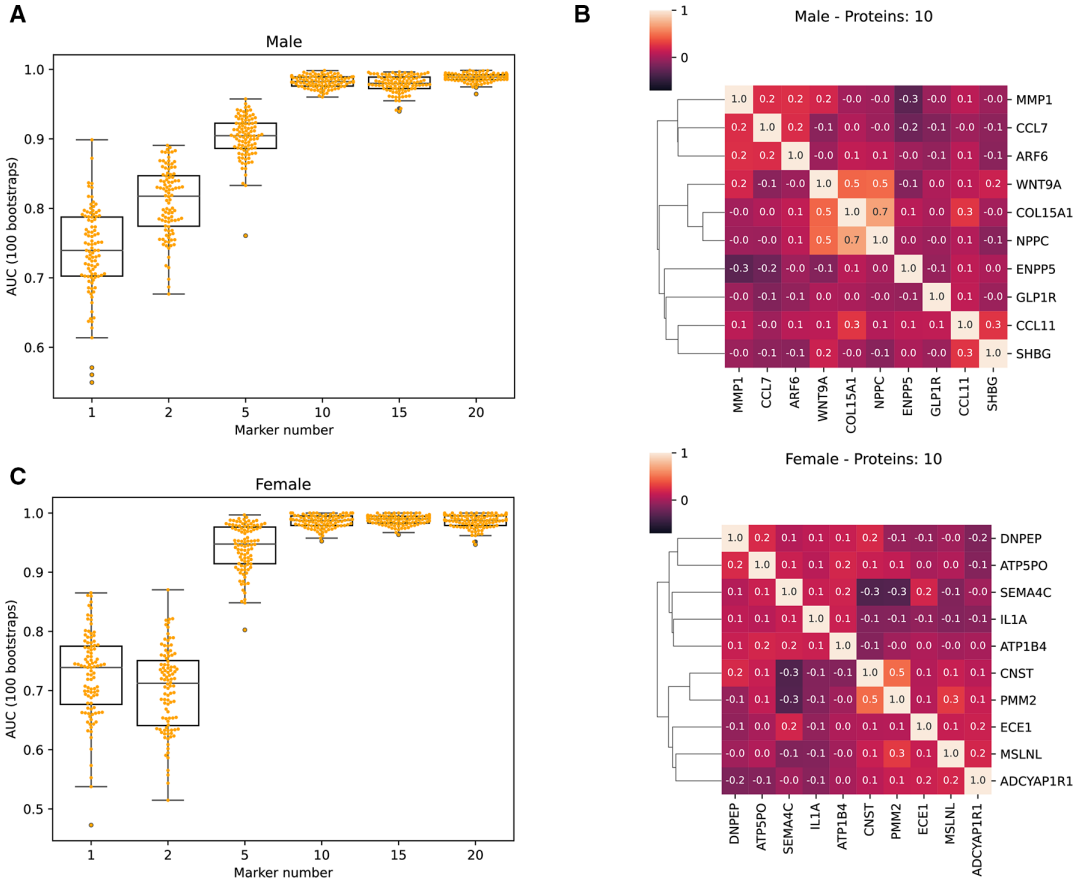

Figure 2 Selection of cancer detection panel. The relationship between the number of proteins included in the protein panel and performance of the panel (A). The correlation between proteins included the detection panel by sex (B). AUC, area under the curve.

85% (CI 76%, 100%) among females. We also observed high accuracy across all stages of cancer among males and females (online supplemental table S4). At stage I and at the specificity of 99%, our panel was able to identify 93% (CI 77%, 100%) of cancers among males and 84% (CI 68%, 100%) of cancers among females. The performance of the panels varied across cancers. Overall, some cancers were easier to detect (eg, cancers of kidney in males and colon in females). On the other hand, the detection of cancers like bladder cancer in females and thyroid cancer

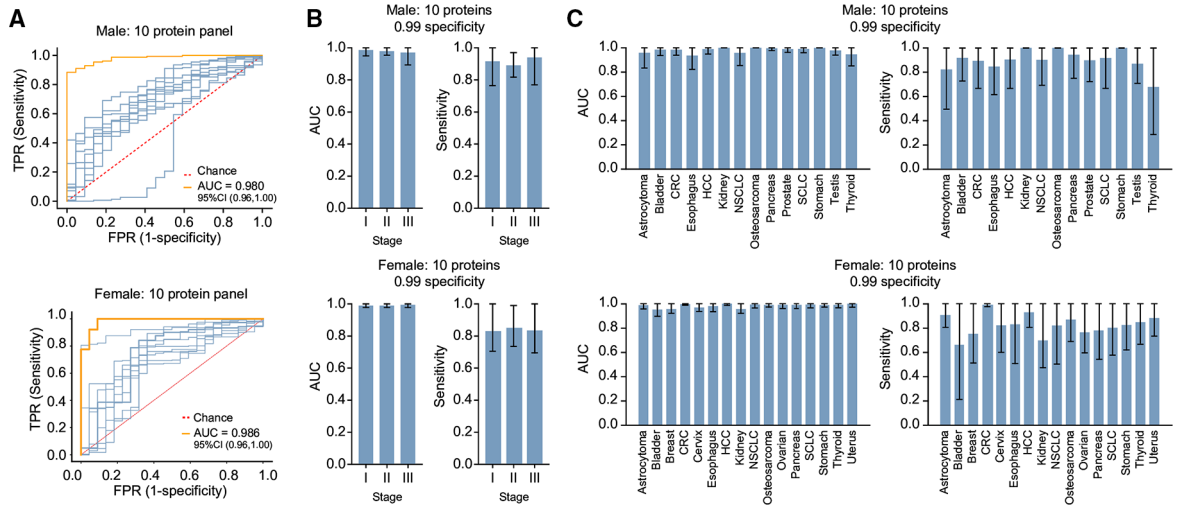

Figure 3 Performance of the detection panel. ROC curve (receiver operating characteristic curve) of detection panel for males and females; the blue lines represent individual ROC curves of each of the 10 proteins included in the panel (A). Sensitivity of the detection panel at the specificity of 99% for each stage of the cancer for males and females (B). Sensitivity of detection panel at the specificity of 99% by stage for males and females (C).

in males were more challenging but achieved with relatively high accuracy after optimisation.

To further evaluate the performance of the cancer set to detect any cancer, we excluded one cancer at a time and evaluated the fitted model on other cancers by the algorithm above on the excluded (unseen) cancer. Our analysis showed an acceptable performance with an AUC of, on average, above 0.9 for an unseen cancer (online supplemental figure S4). The cancer-out model has better performance on unseen cancers of pancreas and liver cancer but a lower performance to detect unseen thyroid cancer and astrocytoma. This shows that the protein set correlates very well with the cancer and serves as a cancer detection test. A plausible baseline cancer detection signature will enable efficient expansion to other cancers with a limited number of new samples added to the training set.

Sex-specific cancer localisation panel

Our localisation panels consisted of 150 proteins. Each sample was fed into separate cancer vs other cancers and the prediction probability for that cancer was calculated. The top two highest probability was used to identify the tissue of origin. The number of proteins allocated to each cancer was selected in a way that optimised the overall performance of the test. In males, the highest number of proteins was allocated to bladder cancer and the lowest number of proteins were allocated to liver cancer (online supplemental figure S5). In females, the highest and the lowest number of proteins were allocated to ovarian cancer and bladder cancer, respectively (online supplemental figure S5). Some proteins were used for differentiation of more than one cancer. For example, in the male localisation panel, CHP1, NT5C1A, PADI4, IL1A, OBP2B, OFD1, PITHD1, TFAP2A, TRIM40, CCL28, LRP2BP, AGR2 were selected in the localisation panels of the three cancers. Similarly in females, proteins associated with more than one cancer were MAPT, PAEP, PER3, SEMA4C, CRYM, FYB1, CRISP2, TREH, GRAP2, CES2, CRACR2A, CKB, GFAP, GH1, SOD3, TMPRSS15, DPP10 (online supplemental table S5).

The top1 and top2 localisation accuracy of the tests for males were 81% (95% CI 74%, 87%) and 89% (95% CI 84%, 94%), respectively, and 67% (95% CI 60%, 74%) and 84% (CI 79%, 89%) for females (figure 4). We evaluated the overall performance of the test by its ability to correctly classify the samples at the detection and localisation level and the test showed an accuracy of 75% for males and 54% for females (online supplemental figure S6).

Role of downregulated and low-concentration proteins

Through our analysis of 18 cancers and normal plasma samples, we found that only a few cancers can be uniquely identified by up-regulated proteins, which are typically preferred as biomarkers. We discovered that many cancers showed much higher specificity using downregulated proteins rather than just upregulated proteins. As the number of cancers included in a single pan-cancer test increases, it will be crucial to have both types of regulation biomarkers to achieve high cancer specificity among many different cancers.

Our evaluation of the plasma concentrations of proteins that were selected showed that the great majority of proteins were in the low abundance group. Of the ten proteins included in the male panel, seven required no dilution and only three required some level of dilution to be detected reliable by the proximity technology: ENPP5 (10-fold dilution), COL15A1 (100-fold dilution) and SHBG (1000-fold dilution). Similarly, only one of the ten proteins included in the female panel required some level of dilution: CNST (10-fold dilution). This highlights the importance of low-concentration proteins to see precancerous states and early stages where the tumour has little systemic impact and generated footprints.

DISCUSSIONS

In this study, we showed that a measurement of limited set of plasma proteins could classify cancer samples from normal and differentiate different cancers. This finding is the foundation for a multi-cancer screening test for the early detection of 18 solid tumours that cover all major human organs of origin for such cancers at the earliest stage of their development with high accuracy. It is important to diagnose cancer at very early stages where curative treatments are achievable with surgery and available treatments. Additionally, for the first time to our knowledge, we found compelling evidence that the cancer protein signatures are most likely sex-specific for all cancers.

Our study also showed that biological signals for early-stage cancers are much more evident in the low-concentration part of the human plasma proteome. It was also promising to observe that a set of proteins could differentiate all cancers from normal and sensitive to detect unseen cancers.

In our study, we analysed a range of proteins found in classical cancer pathways. However, we discovered that only a very small number of these proteins could be used as biomarkers for early-stage cancer. In contrast, many proteins that were effective biomarkers for early-stage cancer were found at low concentrations across the entire plasma proteome. This finding may be due to the fact that most of our knowledge about the role of proteins in cancer pathways comes from studies of transcriptome at the tissue level in advanced stages of cancer, and the expression of proteins at the mRNA and protein levels do not always correspond. In addition, the concentration of proteins in tissue and plasma may not be strongly correlated. Finally, our samples were mainly from early-stage cancers, where classic cancer pathways may not be highly active. This finding has major implication for developing the next generation of diagnostics highlighting the role low-abundance protein in early detection of disease.

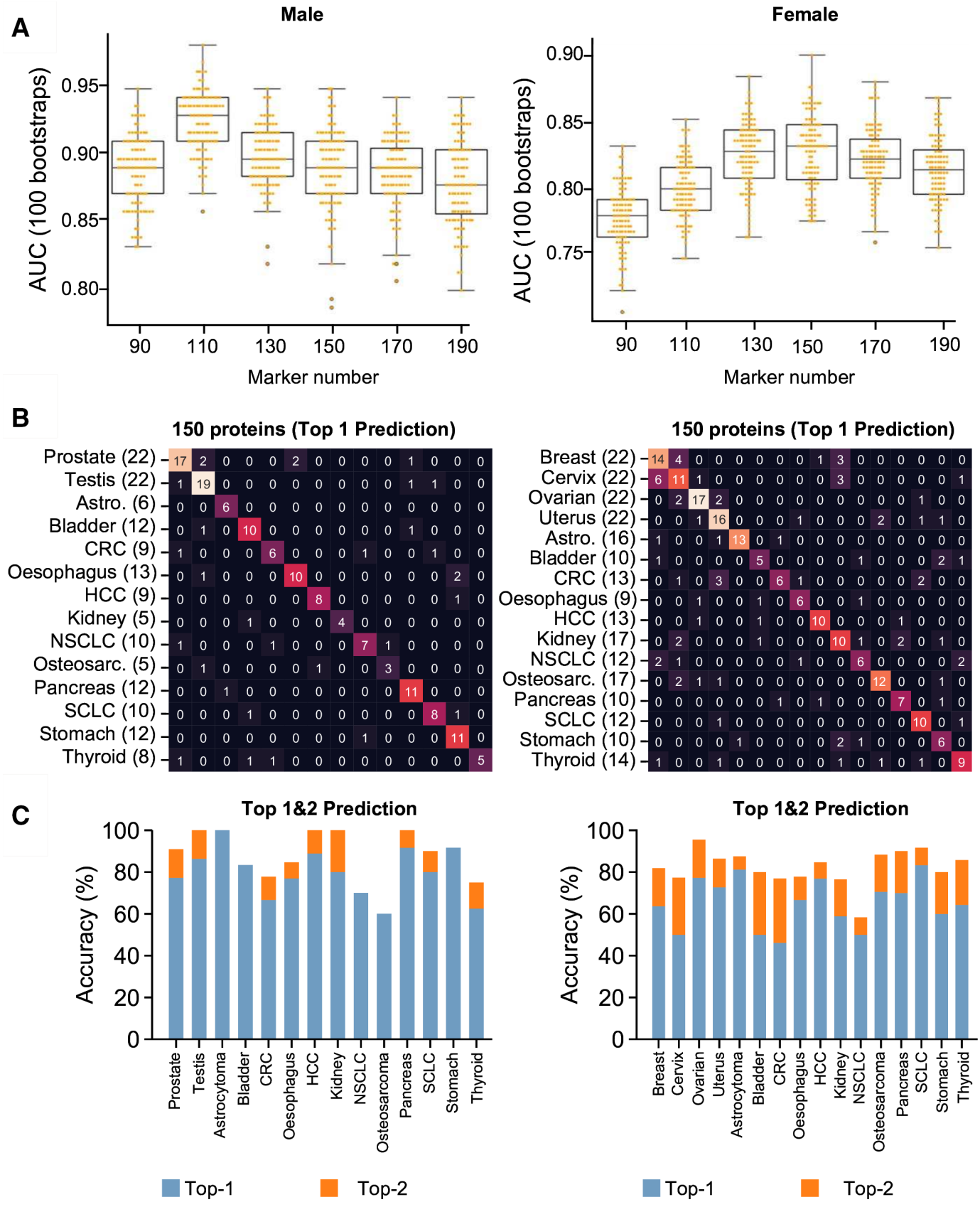

Figure 4 The selection and performance of the localisation panels across cancers. The relationship between the number of proteins included in the localisation panel and the overall performance of the panel (A). The confusion matrix showing the performance of the of the test in correctly identifying the source of the cancer in the first prediction (B) and the bar chart of the overall performance of the localisation panel (C). AUC, area under the curve.

The proteome-based diagnostic test showed promising performance compared with other technologies such as circulating tumour DNA tests[23] by significantly outperforming existing multi-cancer screening tests in detecting cancer across all stages (I, II, III) and among all types of cancers. At the specificity of 99% and in stage I of cancer, our test had a sensitivity that was much greater than Galleri[24] and CancerSEEK[25] tests. Additionally, our study demonstrated ability of our 'best-test to achieve much higher accuracy in identifying the tissue of origin of cancers in each sample in comparison to other tests. At the cancer-specific level, all our best-tests were more accurate than other available screening tests. Among the four screening tests that have received the highest recommendation (level A) from the US Preventive Services Task Force (colonoscopy for colon cancer, pap test for cervical cancer, mammography for breast cancer and low-dose CT scan for lung cancer), only colonoscopy and low-dose CT scan had an accuracy of above 90% for cancer detection. However, the sensitivity of our test for detecting early-stage cancer was still higher than the sensitivity of these tests.

Over the past decade, mRNA large-scale sequencing has provided a comprehensive view of gene expression in specific tissues, revealing the proteins that are present in different organs of the human body. The Human Protein Atlas is a useful resource for understanding mRNA and protein expression in multiple healthy tissues.[26] However, it is important to note that tissues are typically composed of complex assemblies of distinct cells that may have different functions and developmental histories. Increasing amounts of information about RNA and protein expression in specific cell types is now becoming available for the individual cells that make up tissues and organs. A challenge in using protein detection for liquid biopsy is that cancer-specific protein biomarkers may be present at ultra-low levels in the blood.[9] This is because proteins that are present at high concentrations in the blood of healthy individuals are unlikely to be significantly increased in patients with early stages of the disease or at early recurrence. The long history of plasma proteome analysis by mass spectrometry show that even proteome coverage was increased from several hundred proteins 30 years ago to more than 5000 proteins based on latest development in chromatography separation technique and data independent acquisition type of acquisition.[27] Still the major problem of cheap and reproducible sample preparation protocols and reliably measuring proteins after first thousands of most abundant proteins prevent development of early stage multi-cancer test by mass spectrometry at acceptable price per sample and general population scale. Thus, assays with greater sensitivity for biomarker proteins that are normally present at very low or undetectable levels in the blood may enable the detection of cancer at an earlier stage of the disease or even at premalignant stages. Our test is based on sensitive proximity assays that require the simultaneous binding of three separate antibodies. This ability to analyse plasma proteome profiles deeply and consistently allowed us to focus our attention on very low-abundant proteins, which we found to be the most precise and accurate biomarkers of early stages for all the cancers studied in our study. Advancing the PEA technology to measure ultra-low protein concentrations will provide better opportunities to detect and classify cancers at a very early stage and even at the precancerous stage.[28]

Our new generation protein-based plasma test has shown high sensitivity in detecting a variety of early-stage tumours in asymptomatic patients, making it a strong candidate for use as a population-wide screening tool that is not currently achievable with existing tests or techniques. Its high specificity can help alleviate concerns about causing harm to patients, and its low cost allows for widespread implementation. To be suitable for large-scale use, a screening test must have high sensitivity and the ability to reduce mortality and morbidity, as well as acceptable for healthcare system cost. In the case of cancer screening, it is also essential for the test to have high specificity to avoid causing undue harm to patients. Our test exhibits these desirable qualities, making it a promising option for cancer screening. We expect that the combination of lower cost and higher accuracy in our test will facilitate its integration into the healthcare system and eventual inclusion in routine annual check-ups. Early detection of cancer has the potential to greatly reduce the societal burden of both health and financial costs. In fact, implementing such interventions can not only be cost-effective but can also result in cost savings for society.

In our study, notable sex-specific differences in cancer detection emerged, necessitating a deeper exploration. The types of cancers in our pan-cancer screening inherently differ between males and females due to the presence of sex-specific cancers like prostate or ovarian cancer. Beyond this, certain proteins are exclusive or more predominant in one sex, affecting detection accuracy. Additionally, the overall distribution and abundance of proteins vary between males and females, with the relationships between these proteins also being sex-specific. Recognising these inherent biological differences highlights the potential benefits of employing gender-tailored biomarker panels, which might enhance detection accuracy. This approach underscores the significance of personalised medicine in contemporary oncology, ensuring diagnostics are attuned to the unique biological signatures of each gender.

Our approach has major strengths, including the total number of proteins measured and accuracy of such measurements across all measured proteins down to very low abundant proteins, the focus on early-stage tumours, the number of studied cancers that represents all major organs of unmet needs included in the study.

Limitations should also be considered in the interpretation of our study findings. The size of the cohort used in this study was relatively small. While we aimed to capture a diverse range of cancers, the limited sample size may restrict the generalisability of our results to

larger populations. Therefore, it is important to validate our test in larger population cohorts to ensure its robustness and reliability across different demographic groups and geographical regions. Another limitation of our study was the lack of comprehensive information on key patients' comorbidities (eg, diabetes, hypertension, obesity). Due to the retrospective nature of the study and the limitations in data availability, these variables were not captured for all cancer subtypes and normal individuals in our dataset, and we were not able to assess their effect on the performance of our panel. Comorbid conditions may introduce additional variability and complexity to the analysis. The influence of these comorbidities on the performance of our test and its accuracy in detecting early-stage cancers needs further investigation. Validation in a cohort with a more diverse range of comorbidities will provide a more comprehensive understanding of the test's performance in real-world clinical settings. Another limitation of our study is the uneven stage distribution of cancer cases, as highlighted in table 1. Our emphasis on sourcing treatment-naive samples, primarily from patients diagnosed during routine check-ups, inadvertently led to an over-representation of specific disease stages. Notably, some cancers had only stage II or III representation, potentially missing insights from other stages. This approach, while offering a unique subset of samples from the biobank, might not comprehensively represent the broader population of cancer patients across all stages. Furthermore, while our study focused on a comprehensive set of proteins and achieved accurate measurements across a wide range of proteins, there may still be limitations in terms of proteome coverage. Despite advancements in proteomic techniques, the complete coverage of the plasma proteome remains a challenge. Certain low-abundance proteins may not have been captured in our analysis. Improvements in proteomic technologies and sample preparation protocols are required to enhance the sensitivity and coverage of the proteome, allowing for a more comprehensive assessment of biomarkers for early-stage cancer detection. Additionally, our study mainly focused on early-stage cancers, where classic cancer pathways may not be highly active. This narrow focus may limit the generalisability of our findings to advanced stages or metastatic cancers. Future studies should aim to explore the performance of our test across a broader spectrum of cancer stages and subtypes to evaluate its effectiveness in different clinical scenarios. Finally, while our test demonstrated promising performance compared with existing technologies and screening tests, it is essential to emphasise the need for independent validation in an external cohort. Validation in an independent population will provide further evidence of the test's accuracy, sensitivity and specificity. It will also help assess its performance in diverse patient populations, accounting for variations in genetic backgrounds, environmental factors and healthcare settings. Robust validation studies are necessary to establish the clinical utility and reliability of our test before its widespread implementation in routine cancer screening programmes.

CONCLUSIONS

In summary, this study serves as a proof of concept for the potential utility of proteomic analysis in the early detection of various cancers. By analysing a comprehensive set of proteomics data and developing cancer-specific protein signatures, we have demonstrated the feasibility and potential performance of this approach for early-stage diagnosis. The findings highlight the importance of considering sex-specific protein profiles and down-regulated proteins as sensitive biomarkers in the early detection of cancers. It is important to note that this study represents an initial exploration into the field of proteomics-based cancer detection, and further validation in larger population cohorts is necessary to establish the reliability and generalisability of our findings. Nonetheless, these results provide a foundation for future research and emphasise the potential of proteomic analysis in revolutionising cancer diagnosis at the population level.

Twitter Ashkan Afshin @aafshinmd

Contributors All authors contributed to drafting the overall structure and flow of the manuscript. Subsequently, authors contributed subsections based on their domain of expertise. AA integrated subsections, incorporated comments and performed additional revisions. AA, as the guarantor, holds full responsibility for the overall content, ensuring the integrity of the finished work and the conduct of the study. AA had complete access to the data and maintained control over the decision to publish.

Funding This research work was financially supported by Novelna Inc.

Competing interests The authors declare they are employed by or hold shares in Novelna Inc.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants but ethical approval for this study was not required as per the guidelines and regulations of the Ukraine Association of Biobank Institutional Review Board (IRB). Deidentified patient samples were obtained from the Ukrainian Biobank, which operates under strict ethical and regulatory protocols. Therefore, due the to the anonymous and non-identifiable nature of the data and exploratory nature of the analysis, the ethical approval was not required. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer-reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iD
Ashkan Afshin http://orcid.org/0000-0001-6239-8809

REFERENCES

1 Tran KB, Lang JJ, Compton K. The global burden of cancer attributable to risk factors, 2010-19: a systematic analysis for the global burden of disease study 2019. *Lancet* 2022;400:563–91.
2 National Cancer Institute. Crunching numbers: what cancer screening statistics really tell us 2018. 2023 Available: https://www.cancer.gov/about-cancer/screening/research/what-screening-statistics-mean
3 Geyer PE, Kulak NA, Pichler G, *et al*. Plasma proteome profiling to assess human health and disease. *Cell Syst* 2016;2:185–95.
4 Marrugo-Ramírez J, Mir M, Samitier J. Blood-based cancer biomarkers in liquid biopsy: a promising non-invasive alternative to tissue biopsy. *Int J Mol Sci* 2018;19:2877.
5 Ding Z, Wang N, Ji N, *et al*. Proteomics technologies for cancer liquid biopsies. *Mol Cancer* 2022;21:53.
6 Hasenleithner SO, Speicher MR. How to detect cancer early using cell-free DNA. *Cancer Cell* 2022;40:1464–6.
7 Loomans-Kropp HA, Umar A, Minasian LM, *et al*. Multi-cancer early detection tests: current progress and future perspectives. *Cancer Epidemiol Biomarkers Prev* 2022;31:512–4.
8 Hackshaw A, Cohen SS, Reichert H, *et al*. Estimating the population health impact of a multi-cancer early detection genomic blood test to complement existing screening in the US and UK. *Br J Cancer* 2021;125:1432–42.
9 Landegren U, Hammond M. Cancer diagnostics based on plasma protein biomarkers: hard times but great expectations. *Mol Oncol* 2021;15:1715–26.
10 Ukraine Association of Biobank. 2022. Available: http://ukrainebiobank.com
11 Fredriksson S, Gullberg M, Jarvius J, *et al*. Protein detection using proximity-dependent DNA ligation assays. *Nat Biotechnol* 2002;20:473–7.
12 Lundberg M, Eriksson A, Tran B, *et al*. Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood. *Nucleic Acids Res* 2011;39:e102.
13 Darmanis S, Nong RY, Hammond M, *et al*. Sensitive plasma protein analysis by microparticle-based proximity ligation assays. *Mol Cell Proteomics* 2010;9:327–35.
14 Skates SJ, Gillette MA, LaBaer J, *et al*. Statistical design for biospecimen cohort size in proteomics-based biomarker discovery and verification studies. *J Proteome Res* 2013;12:5383–94.
15 Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;12:2825–30.
16 Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng* 2007;9:90–5.
17 Waskom ML. Seaborn: statistical data visualization. *JOSS* 2021;6:3021.
18 Plotly Technologies Inc. Collaborative data science plotly technologies Inc. 2015. Available: https://plot.ly
19 McKinney W. Data structures for statistical computing in python. In. van der Walt S, Millman KJ, eds. 9th Python in Science Conference; Austin, Texas.2010:56–61
20 Harris CR, Millman KJ, van der Walt SJ, *et al*. Array programming with numpy. *Nature* 2020;585:357–62.
21 Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. 9th Python in Science Conference; Austin, Texas.2010:92–6
22 Sever R, Brugge JS. Signal transduction in cancer. *Cold Spring Harb Perspect Med* 2015;5:a006098.
23 Wan JCM, Massie C, Garcia-Corbacho J, *et al*. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer* 2017;17:223–38.
24 Klein EA, Richards D, Cohn A, *et al*. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann Oncol* 2021;32:1167–77.
25 Cohen JD, Li L, Wang Y, *et al*. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 2018;359:926–30.
26 Rozenblatt-Rosen O, Stubbington MJT, Regev A, *et al*. The human cell Atlas: from vision to reality. *Nature* 2017;550:451–3.
27 Robbins JM, Peterson B, Schranner D, *et al*. Author correction: human plasma proteomic profiles indicative of cardiorespiratory fitness. *Nat Metab* 2021;3:1275.
28 Petrera A, von Toerne C, Behler J, *et al*. Multiplatform approach for plasma proteomics: complementarity of olink proximity extension assay technology to mass spectrometry-based protein profiling. *J Proteome Res* 2021;20:751–62.