Received: 3 November 2022 | Revised: 21 December 2022 | Accepted: 13 January 2023

DOI: 10.1002/prca.202200093

Proteomics Clinical Applications

RESEARCH ARTICLE

Serum proteomics profiling identifies a preliminary signature for the diagnosis of early-stage lung cancer

Roberto Gasparri[1] | Roberta Noberini[2] | Alessandro Cuomo[2] | Avinash Yadav[2] | Davide Tricarico[3,4] | Carola Salvetto[4] | Patrick Maisonneuve[5] | Valentina Caminiti[1] | Giulia Sedda[1] | Angela Sabalic[1] | Tiziana Bonaldi[2,6] | Lorenzo Spaggiari[1,6]

[1]Department of Thoracic Surgery, IEO, European Institute of Oncology IRCCS, Milan, Italy

[2]Department of Experimental Oncology, IEO European Institute of Oncology IRCCS, Milan, Italy

[3]AITEM Artificial Intelligence Technologies Multipurpose s.r.l., Turin, Italy

[4]Department of Mathematics "G. Peano", University of Turin, Turin, Italy

[5]Division of Epidemiology and Biostatistics, IEO, European Institute of Oncology IRCCS, Milan, Italy

[6]Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy

Correspondence
Roberto Gasparri, Department of Thoracic Surgery, IEO, European Institute of Oncology IRCCS, Via Giuseppe Ripamonti 435, 20141 Milan, Italy.
Email: Roberto.gasparri@ieo.it

Tiziana Bonaldi, Department of Experimental Oncology, IEO European Institute of Oncology IRCCS, Via Adamello 16, 20139 Milano MI, Italy.
Email: tiziana.bonaldi@ieo.it

Funding information
Associazione Italiana per la Ricerca sul Cancro, Grant/Award Number: IG-2018-21834; Horizon 2020 Framework Programme, Grant/Award Number: 823839 (EPIC-XS); Ministero della Salute, Grant/Award Numbers: 5X1000, Ricerca Corrente

Abstract

Purpose: Lung cancer is the most common cause of death from cancer worldwide, largely due to late diagnosis. Thus, there is an urgent need to develop new approaches to improve the detection of early-stage lung cancer, which would greatly improve patient survival.

Experimental Design: The quantitative protein expression profiles of microvesicles isolated from the sera from 46 lung cancer patients and 41 high-risk non-cancer subjects were obtained using a mass spectrometry method based on a peptide library matching approach.

Results: We identified 33 differentially expressed proteins that allow discriminating the two groups. We also built a machine learning model based on serum protein expression profiles that can correctly classify the majority of lung cancer cases and that highlighted a decrease in the levels of Arysulfatase A (ARSA) as the most discriminating factor found in tumors.

Conclusions and Clinical Relevance: Our study identified a preliminary, non-invasive protein signature able to discriminate with high specificity and selectivity early-stage lung cancer patients from high-risk healthy subjects. These results provide the basis for

Abbreviations: HPLC, high performance liquid chromatography; LFQ, label-free quantification; LR, linear regression; MV, microvesicle; RF, random forest; SVM, support vector machine.

Roberto Gasparri, Roberta Noberini and Alessandro Cuomo contributed equally to this work.

future validation studies for the development of a non-invasive diagnostic tool for lung cancer.

KEYWORDS
biomarker, early diagnosis, lung cancer, mass spectrometry, machine learning, serum

1 | INTRODUCTION

Lung cancer is the most common cause of death from cancer worldwide (WHO, 2019). Although the 5-year survival rate is 54% for cases diagnosed when the disease is still localized, only 15% of lung cancers are identified at early stages (American Cancer Society, 2018). Low-dose computed tomography has been shown to provide a 20% reduction in lung cancer mortality, and has been proposed as an annual screening to test the high-risk population. However, the high false-positive rate, costs and risks associated with radiation limit its use in the clinical context [1]. Thus, there is the urgent need to develop new approaches for the detection of early-stage lung cancer, which would greatly improve patient survival.

The last decade has witnessed an increasing interest in the design of new technologies for the identification of biomarkers suitable to screen the asymptomatic at-risk population, which should be reproducible, non-invasive and cost-effective. In this scenario, body fluids, such as blood (plasma or serum), exhaled breath and urine represent ideal clinical samples to be analyzed, because of their minimally invasive accessibility and availability. Proteins represent particularly interesting biomarkers, because they are relatively stable and are the biological endpoint responsible for most cellular functions. Currently, the technology of choice for the systematic and quantitative profiling of proteins in complex biological matrixes is mass spectrometry (MS)-based proteomics, which has emerged as a powerful tool in biomarker discovery [2]. Various studies attempted the identifications of protein biomarkers for lung cancer early diagnosis by proteomics approaches in accessible fluids, such as blood and urines (reviewed in [3, 4]). Although several potential blood biomarkers were identified, the majority was not specific for lung cancer, but was rather linked to either inflammation or metabolic alterations, or found to be dysregulated also in other cancer types. In addition, many of the serum proteomics studies conducted so far suffered from the lack of rigorous patient enrolment criteria and of standardized protocols for sample collection, processing and analysis, which may have hindered the discovery of lung cancer specific biomarkers. More promising results were obtained in some recent urine proteomics studies, where biomarkers/signatures distinguishing lung cancer from healthy subject, or from other tumor types, were found [5].

While blood is the most widely used and the most convenient source of patient samples for biomarker discovery, the comprehensive analysis of plasma/serum proteomes is a challenging task [6]. This is due to the high complexity and extreme dynamic range of the proteins present in plasma, as well as the presence of a few proteins at very high concentrations. The presence of such very abundant proteins masks the detection of less abundant "tissue leakage" proteins that could represent potential biomarkers. Different strategies can be employed to overcome these challenges, including depletion of highly abundant proteins and biochemical fractionation. An alternative promising approach involves the isolation of plasma-derived extracellular vesicles, which are shed from cells all over the body and can be released into the bloodstream [7]. In recent years, there has been a growing interest in the proteomic profiling of extracellular vesicles as the source of biomarkers and as mediators of disease mechanisms [7]. Extracellular vesicles include exosomes, microvesicles (MVs), apoptotic bodies, and apoptotic microparticles, which are characterized by different sizes and cellular origin. MVs are large vesicles (100 nm−1 μm) that protrude from the plasma membrane. They are found not only in blood, but also in urines and other biological fluids. MVs carry a "signature" of the protein content of the cells they originate from and the secretion of MVs from cancer cells contributes to angiogenesis, metastasis, tumor formation, and disease progression [8]. Hence, they are an attractive source of biomarkers.

In this study, we applied proteomics technologies combined with machine learning to identify protein biomarkers/signatures able to distinguish early stage lung cancer patients from healthy subjects, by profiling the proteomes of serum MVs.

2 | MATERIAL AND METHODS

2.1 | Study design and patient selection

This is a prospective, single-center, case control study. The protocol was approved by the local Ethics Committee on October 2017 (N: R650/17-IEO 532). The results of the study are based on 87 subjects (Table 1, and Table S1). Lung cancer patients (n = 46) were subjects aged between 50 and 80 years, with a diagnosis of T1N0 lung cancer, with no previous medical treatment and no malignancies within the previous 5 years. Controls subjects (n = 41) were healthy at-risk subjects (heavy smokers or subjects with non-cancer related pulmonary diseases), aged between 50 and 80 years, with no malignancies within the previous 5 years and with a recent negative Chest X-ray or Ct scan. At the time of registration, the subjects were duly informed and signed the study-specific informed consent. For each participant, clinical-demographic information was collected such as: comorbid conditions (cardiac, respiratory, or metabolic disorders), tumor characteristics (histology, stage and size of the nodule,

TABLE 1 Clinical profiles and demographics of healthy controls and lung cancer patients

		All	Lung cancer cases	Controls
Number of patients		87	46 (52.8%)	41 (47.1%)
Age	Median (range)	64 (50-80)	65 (51-80)	64 (50-78)
Sex	Men	45	22	23
	Women	42	24	18
Smoker	No	12	7	5
	Ex	40	20	20
	Yes	35	19	16
TNM 8	Stage I		33 (71.7%)	
	Stage II		6 (13%)	
	Stage III		7 (15%)	

etc.) and lifestyle information, with special emphasis on tobacco smoke exposure.

2.2 | MV isolation

Blood samples (at least 15 mL) were collected by standard phlebotomy, discarding the first 3 mL of blood to prevent contamination by skin. The serum was prepared by leaving the blood in the tubes for 3 h at room temperature to allow blood clotting, followed by centrifugation at 1000 x *g* for 10 min at room temperature. The serum was removed immediately after centrifugation and stored at −80°C. For the isolation of MVs, 1 mL of serum was centrifuged at 4000 rpm for 20 min at 4°C to remove cellular debris. The supernatant was diluted 1:2 in ice-cold PBS and centrifuged at 20,000 x *g* at 4°C for 1 h. The pellet was washed twice with ice-cold PBS and centrifuged at 20,000 x *g* at 4°C for 1 h. Although the published microvesicle isolation protocol described only one wash [12], preliminary tests showed that an additional wash provided cleaner MV preparations (not shown). The microvesicle pellets were stored at −80°C until use.

2.3 | Peptide library generation

To generate a reference spectral library for MS acquisition, we selected five tumor samples representative of the sample cohort in terms of gender, tumor stage, grade, and comorbidities, to obtain the most comprehensive list of reference proteins/peptides. The MVs isolated from the sera of the five patients were pooled and half of the amount was digested using the PreOmics iST sample preparation kit and fractionated though the Pierce High pH Reversed-Phase Peptide Fractionation Kit (Thermo Fisher Scientific), obtaining eight fractions. The other half of the MV pool was separated by SDS page. The lane was cut into eight gel bands that were in-gel digested as previously described [9]. Proteins extracted from the tissue biopsies from the same patients were processed similarly.

Clinical significance

Lung cancer is by far the leading cause of cancer death worldwide, with 94.500 woman's death and 183.100 man's deaths estimated for 2018 in the EU, with a 5-year overall survival rate of 18%. This worst-case scenario is supported by the persistent delay of clinical diagnosis in most lung cancer patients, which are detected in advanced stages of the disease. Thus, there is an urgent need to develop new approaches to improve the detection of early-stage lung cancer, which would greatly improve patient survival. In this study, we profiled 87 human serum samples from healthy, high-risk donors and lung cancer patient. To overcome the challenges related with the analysis of serum proteomes, we chose to analyze circulating microvesicles, identifying a list of candidate non-invasive markers for future validation.

2.4 | LC-MS/MS analysis

Sera (1 µl) and MV pellets obtained from 1 mL of serum were digested using the PreOmics iST sample preparation kit, following the manufacturer's guidelines. Approximately 1/10 of the digested peptides was analyzed. Peptide mixtures were separated by reversed-phase chromatography on an EASY-nLC 1200 ultra high-performance liquid chromatography (UHPLC) system through an EASY-Spray column (Thermo Fisher Scientific), 25-cm long (inner diameter 75 µm, PepMap C18, 2 µm particles), which was connected online to a Q Exactive HF (Thermo Fisher Scientific) instrument through an EASY-Spray Ion Source (Thermo Fisher Scientific). Both for library and study samples, the purified peptides were loaded in buffer A (0.1% formic acid in water) at a constant pressure of 980 Bar. They were separated through the following gradient: 64 min of 3% to 30% of buffer B (0.1% formic acid, 80% acetonitrile), 10 min 30% to 60% of buffer B, 1 min 60% to 95% buffer B, at a constant flow rate of 250 nl/min. The column temperature was kept at 45°C under EASY-Spray oven control. For the library samples the mass spectrometer was operated in a "top-15" data-dependent acquisition (DDA) mode. MS spectra were collected in the Orbitrap mass analyzer at a 60,000 resolution (200 m/z) within a range of 300 to 1650 m/z with an automatic gain control (AGC) target of 3e6 and a maximum ion injection time of 20 ms. The 15 most intense ions from the full scan were sequentially fragmented with an isolation width of 1.4 m/z, following higher-energy collisional dissociation (HCD) with a normalized collision energy (NCE) of 28%. The resolution used for MS/MS spectra collection in the Orbitrap was 15,000 at 200 m/z with an AGC target of 1e5 and a maximum ion injection time of 80 ms. Precursor dynamic exclusion was enabled with a duration of 20s. The study samples were analyzed using the BoxCar acquisition mode [10] and the mass spectrometer was operated under MaxQuant.Live control, with default parameters [11, 12].

2.5 | Data analysis

MS raw files were processed with MaxQuant version 1.6.0.15, and the extracted MS/MS spectra were matched by the Andromeda search engine against tryptic peptides (maximum of two missed cleavages) derived from human reference proteomes (Uniprot UP000005640, 80,027 entries). The search included cysteine carbamidomethylation as a fixed modification and methionine oxidation and acetylation of the protein N-terminus as variable modifications. Required minimum peptide length was seven amino acids and maximum mass tolerances were 4.5 p.p.m. for precursor ions after nonlinear recalibration and 20 p.p.m. for fragment ions. MaxLFQ was performed separately in parameter groups with a minimum ratio count of 1 [13]. If applicable, peptide identifications were transferred between samples by 'match between runs' within a 0.7 min window after retention time alignment. Identifications in the library were stringently filtered for a FDR $<$ 1% at both peptide spectrum match and protein group levels. The proteinGroups MaxQuant output file is included as Dataset S1.

Data analysis was performed using the Perseus platform [14] and the DEP software (Bioconductor version 3.13, https://bioconductor.org/packages/release/bioc/html/DEP.html). The proteinGroups output table from MaxQuant was filtered for "reverse", "only identified by site", "contaminants", and at least 70% data completeness in each group (healthy, tumor). The resulting data matrix contained 933 protein groups. Missing values were then replaced by random numbers drown from a normal distribution, assuming that these values belonged to low intensity spectrum of the distribution (down shift = 1.8 width = 0.3). To determine significantly changing proteins between disease and control condition, a two-sample Student's t-test was used. The original P-value was than corrected for an FDR of 0.05 by the Benjamini–Hochberg method. Results were filtered to have both a significant FDR-corrected *p*-value and a minimum log2-fold change of $\pm$ 1 (Flagged as TRUE in table S). Raw and normalized LFQ values for all the quantitated proteins are reported in Table S2.

2.6 | Machine learning model development

The construction of a ML tumor prediction model was performed using the software Python and was composed by three phases: data preparation, design and testing. During data preparation, testing and training set pairs to be used in the next phases were generated. Because of the data set size, a 10-times iterated 5-fold stratified cross validation was used. During each of the 10 iterations the data set was shuffled and split in five roughly equallysized folds. Each fold was in turn used as validation set for testing, and the remaining folds were used for training and hyper-parameter tuning, resulting in a total of 50 test-train sets pairs. In the design phase, the best combination of feature selection and classification algorithms, and related hyper-parameter settings, were identified by testing Support Vector Machine (SVM), linear regression (LR) and Random forest (RF) 24 algorithms. Feature selection algorithms were set to determine the 10 most discriminant proteins, while the remaining hyper-parameters optimization was performed by grid search approach. For each of the test-train set pair, 3 out of 4 folds in the training set were used to train each combination, evaluating the performance on the remaining one (optimization fold), of the classification algorithm fitted on the selected features. In this phase, the samples in the test set were excluded to avoid data leakage. The results obtained in each iteration were merged and the best performer was selected as the final solution. Finally, during testing phase, the performance of selected combination was assessed. For each of the test-train set pair, the solution was trained on the training set and evaluated on the test set. Results over the pairs were merged by average and 95% confidence intervals were computed by mean of quantiles. The selected proteins and their relative importance were tracked at each step, to enable post-analysis on the most important discriminant features. With the scope of facilitating the adoption of the proposed method in the clinical setting, we evaluated further reducing the number of proteins required for classification. The contribution of each feature was measured by testing models with different sets of selected proteins and measuring the resulting performance change (i.e., the amount of performance gained/lost when a certain protein is included or not).

3 | RESULTS

Method implementation for serum proteome profiling. The discovery of MS-based biomarkers from plasma/serum is limited by the extreme dynamic range of this biological fluid, which spans over 11 orders of magnitude. Therefore, obtaining a deep coverage of the serum proteome typically requires biochemical fractionation approaches to reduce serum complexity. In this study, we adopted a fractionation strategy involving the isolation of MVs, which requires simple high-speed centrifugation steps [12]. This approach is cost-effective (since MV isolation does not require any expensive equipment or commercial kits) and can be applied to large cohorts of samples. In addition, the MV proteome may be enriched in proteins derived from the tumor tissue, which may represent useful biomarkers, compared with fractionated whole serum. To verify the performance of the MS-based analysis of MVs, we compared the results obtained from whole sera and of MVs isolated using a protocol adapted from [12] (Figure 1A). By profiling the sera from 10 subjects, of which five healthy subjects and five lung cancer tumor patients, we quantified 220 to 260 proteins from whole sera and 500 to 800 proteins from MVs (Figure 1B), confirming the better performance of MV analysis. To further improve serum proteome coverage, we employed the BoxCar method, a "match-library" strategy that can mitigate the problem of the extreme dynamic range in body fluids analysis by transferring peptide identifications between samples (more precisely from a library of previously acquired experimental spectra to the clinical samples), leading to deeper proteome coverage [10]. In this pilot experiment, we generated a reference library composed by the MS spectra corresponding to the a proteome of approximately 1500 proteins, by processing MVs obtained from five lung cancer patients, which were pooled and fractionated by HpH. By applying the Box Car approach to the small cohort described above,

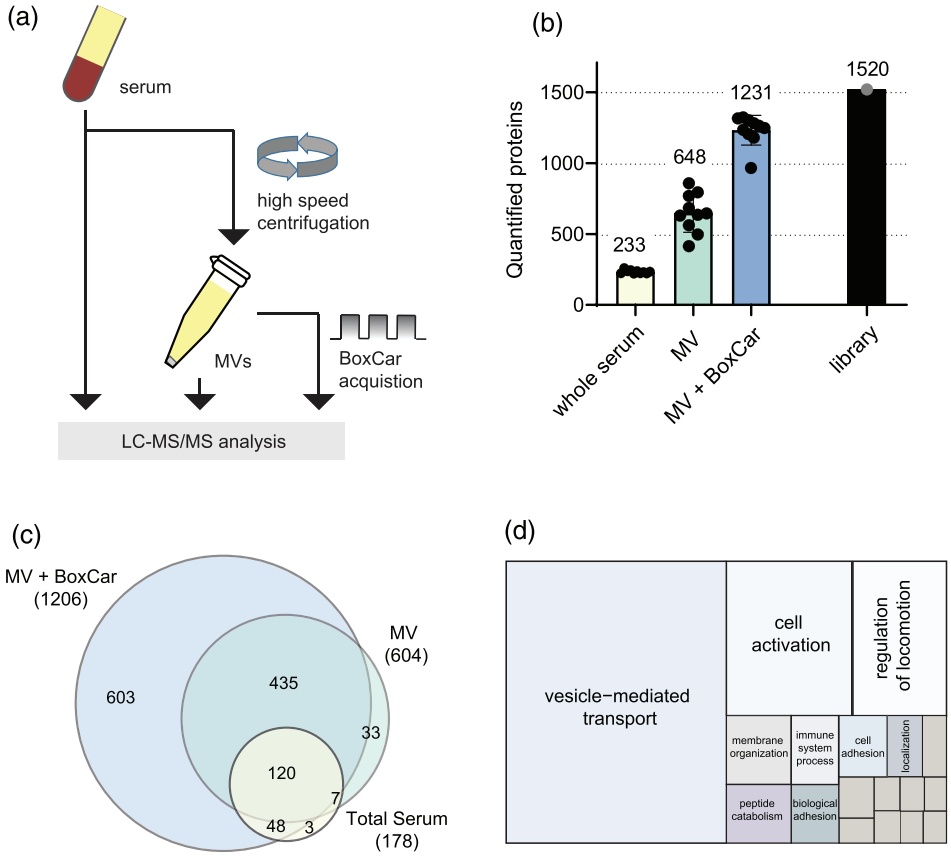

FIGURE 1 **Method implementation for serum proteome profiling. (A)** Experimental scheme for the implementation of a MS-based strategy for the analysis of serum proteomes. Whole sera or microvesicles (MVs) obtained by high speed centrifugation, also in combination with the BoxCar acquisition approach, were analyzed. **(B)** Number of proteins quantified by analyzing whole sera and microvesicles (with/without BoxCar acquisition). **(C)** Venn diagram showing the overlap among the proteins shown in (B). **(D)** GO biological process enriched terms in microvesicles compared with total serum. The most represented terms are indicated. The other terms include: phosphorous metabolism; multicellular organismal process; behaviour; receptor metabolism; response to stimulus; drug metabolism; signaling; fatty acid derivative metabolism; viral entry into host cell

we were able to improve the number of quantified proteins to 1000 - 1300. This list of proteins comprises most of the proteins quantified from whole serum (Figure 1C). Importantly, a gene ontology analysis of the proteins detected only in the MV proteomes highlighted vesicle-mediated transport, and membrane organization as the most enriched biological process terms, confirming the presence of vesicle in our preparation (Figure 1D). Furthermore, other terms, such as cell activation, were present, which may reveal biological processes altered in the tumor and useful as potential biomarkers.

Serum proteome profiling of lung cancer sera. With the aim of identifying circulating biomarkers for the diagnosis of early-stage lung cancer, we analyzed 87 sera samples, which included 46 tumor samples and 41 healthy samples. Healthy samples comprised high risk subjects (86% of the subjects were smokers or heavy smokers), as well as subject with other non-cancer diseases, such as asthma, chronic obstructive pulmonary disease, obesity and cardiovascular diseases (Table S1). Because our main goal is the identification of biomarkers for early diagnosis, most of the tumor samples (72%) were early stage tumors. To improve the number of quantified proteins, we used a library-match approach [10], where identifications can be transferred from samples that are abundant, fractionated and representative of the sample analyzed, to the sample under investigation. In this study, we acquired a reference library generated by fractionating both the MVs and the whole protein extracts from tissue biopsies from five lung cancer patients representative of the sample cohort, which contained 2565 identified proteins (Figure 2A). The choice of using also tissue proteins for library construction is dictated by the idea that MVs contain proteins derived from tumor cells, which may be useful as biomarkers. The MaxQuant software suite was utilized for protein identification and label-free quantification (LFQ). On average, we quantified 973 proteins in each serum sample. The differential expression analysis was performed with the DEP package from Bioconductor in the R statistical framework. The LFQ intensities from MaxQuant protein report were filtered to retain only those genes which had at least 70% valid values in at least one condition. This resulted in a data matrix of 933 genes across 87 serum samples. The missing values in the matrix were imputed from a left shifted gaussian distribution. Thirty three proteins showing a significant differential expression were identified. The corresponding

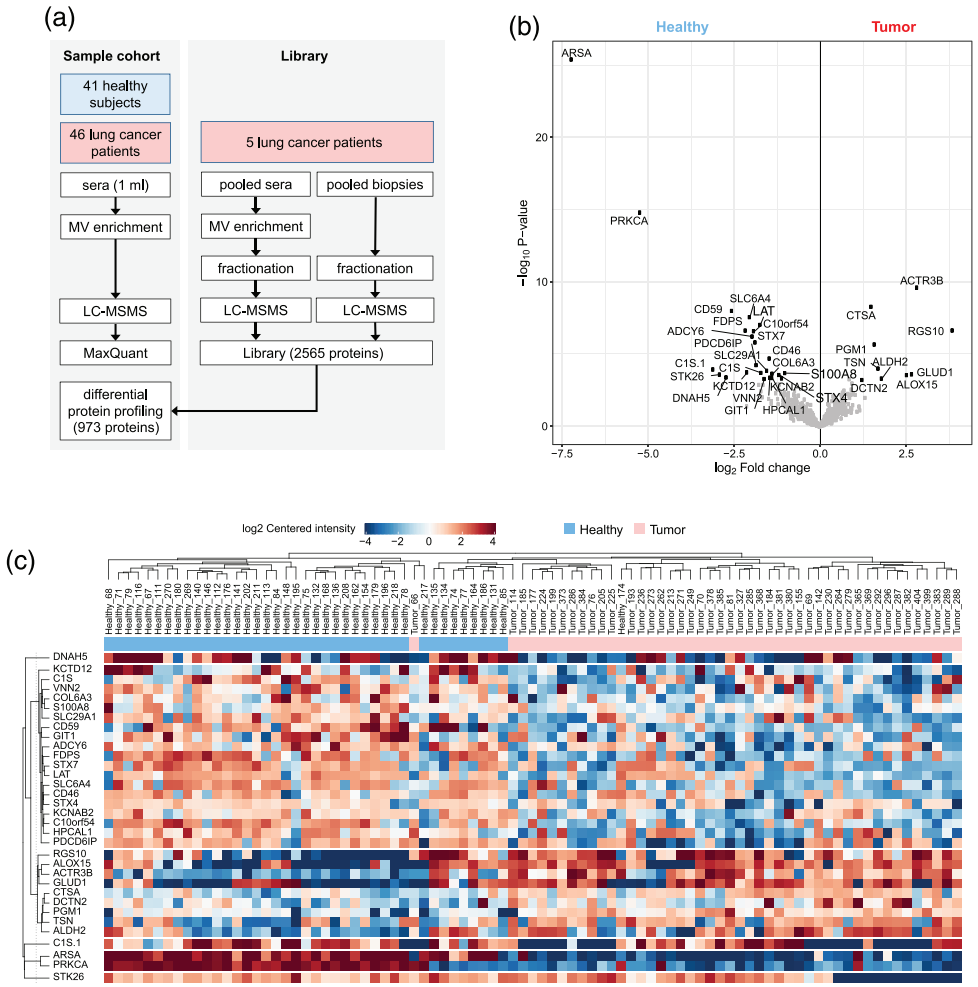

FIGURE 2 **Serum proteome profiling of lung cancer sera.** **(A)** Experimental workflow illustrating the match-library strategy used for the acquisition of the serum proteome of healthy subjects and lung cancer patients. The average number of proteins quantified from each microvesicle samples is indicated. **(B)** Volcano Plot showing 33 DEGs originating from the comparison of the sera of healthy and tumor subjects. Dots represent proteins/genes, distributed by fold change along the x axis and by p-value along the y axis. Proteins/genes with a statistically significant differential expression (adjusted p value < 0.05, log2(fold change) > 1) are shown in black. The right panel shows the up-regulated DEGs, the left panel shows the down-regulated DEGs in tumor compared to healthy samples. **(C)** Clustered heatmap of 87 samples and 33 DEGs. Red represents up-regulated genes and blue represents down-regulated genes. C1S: A0A087X232; C1S.1: C9IZP8

differential expressed genes (DEGs) are shown in the volcano plot in Figure 2B and their intensities in the heatmap in Figure 2C. Among down-regulated proteins, we identified interesting candidates, such as Arylsulfatase A (ARSA) and Protein Kinase α-type (PRKCA). Changes in the activity and level of ARSA in tumor patients were reported by other studies [15, 16] conducted on the serum of lung cancer patients. PRKCA is a protein kinase that is involved in the positive and negative regulation of a number of biological processes, by directly phosphorylating targets such as RAF1, BCL2, CSPG4, TNNT2/CTNT, or activating signaling cascades involving MAPK1/3 (ERK1/2) and RAP1GAP [17]. Aberrant high expression of PRKCA levels have been also found in lung adenocarcinoma patients, especially those with Epidermal Growth Factor Receptor (EGFR) mutations [18].

Machine learning model development and biomarker panel selection. With the goal to identify a combinatorial predictor, we tested different machine learning algorithms evaluating their performances based on the accuracy of the prediction and other parameters, like sensitivity, specificity and area under the curve (see Figure 3A-B and section 2.6 of Material and methods for the description of the model development). During the design phase, a combination of RF, as feature selection algorithm, and SVM with linear kernel, as classification model, emerged as the best performer, with an overall accuracy of 94.28% (95% C.I.: 87% to 100%). Similar performances were also obtained by combinations of LR and SVM with linear kernel for both the feature selection and classification algorithms. The performance obtained during the testing phase were similar to the ones obtained in the design phase (Table 2, Figure 3C). All the top 10 proteins present in the top three algorithm combinations were also present among the 33 DEGs identified by standard analyses. Four proteins were present in the top 10 positions of all the machine learning algorithms: ARSA, PRKCA, ACTR3B, and CD59. ARSA and PRKCA were respectively the first and second most relevant ones in every list (Figure 3D), with ARSA con-

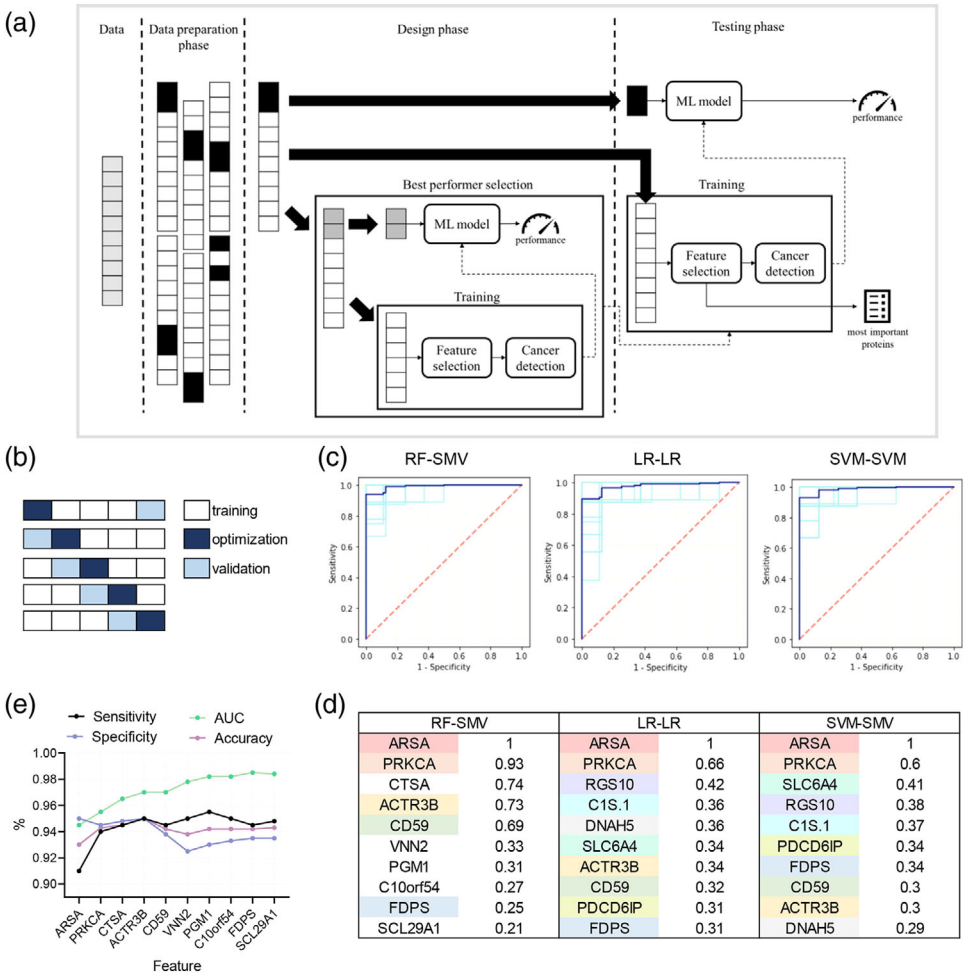

FIGURE 3 **Development of a machine learning approach for LC prediction from serum proteomics data.** (A) Analytical machine learning workflow. All the steps involved in the implementation and the validation of the machine learning model are shown. (B) Splitting process adopted during the data preparation phase. (C) Specificity-sensitivity curves used for AUC calculation for RF-SVM linear, LR-LR, SVM linear-SVM linear combinations. The lines in light-blue represent the ROC curves for each of the 50 splits used in the cross validation, the darker line represents the average. (D) Most relevant proteins isolated through model-based feature selection, sorted by importance in descending order. The weights in the lists are obtained by scaling with a min-max normalization the sums of the weights (or the positive coefficients) obtained for all the 933 proteins at each iteration of the cross validation. (E) Visualization of the evolution of the average performances of the classification with linear SVM assessed by cross validation, adding progressively the ten most important features obtained through the random forest algorithm. Performances refer to the last added protein (which is indicated) in addition to all the proteins which precede it on the feature axis. C1S.1: C9IZP8

tributing for the 90% to 95% to the classification models. To simplify the model using a number of variables manageable in a clinical setting, the number of considered proteins was iteratively diminished, measuring the impact on the predictive performance. This operation confirmed the prominent role of ARSA and PRKCA in the prediction process. Indeed, the algorithm that only elaborate the measure of these two proteins can correctly predict more than 94% of the cases with area under curve > 95% (Figure 3E).

Of the putative markers found in one of the three predictive models, two (CD59 and PDCD6IP) were also found significantly decreased in the urine proteomes of lung cancer patients compared with normal subjects [5], which serves as an external confirmation of our results for these two proteins (Figure 4). Interestingly, these two markers were also decreased compared with patients with benign pulmonary diseases (pneumonia and chronic obstructive pulmonary disease), supporting their potential as specific tumor markers.

4 | DISCUSSION

Lung cancer is a latent disease, which is often asymptomatic or associated with non-specific symptoms. As a consequence, it is often diagnosed at a late stage and is associated with poor prognosis. Although during the last decade new medical approaches have been developed and novel treatments have been implemented into the clinical practice, a rapid, accurate and non-invasive test to detect the neoplasia at a stage when it is still treatable would be highly desirable and has not yet been developed.

TABLE 2 Performance and parameters of the three best classification algorithms

	Random Forest SVM Linear	Logistic Regression Logistic Regression	SVM Linear SVM Linear	Feature selection Classification
Sensitivity	0.9483 (0.76-1)	0.9392 (0.76-1)	0.9506 (0.76-1)	**Performance during the design phase**
Specificity	0.9356 (0.75-1)	0.9414 (0.75-1)	0.9289 (0.75-1)	
AUC	0.9845 (0.92-1)	0.9874 (0.95-1)	0.9854 (0.93-1)	
Accuracy	0.9428 (0.87-1)	0.9406 (0.87-1)	0.9404 (0.87-1)	
Feature selection parameters	max_depth = inf, n_estimators = 50	C = 0.01	C = 0.001	**Hyper-parameter setting**
Classification parameters	C = 0.01	C = 0.0215	C = 0.0022	
Sensitivity	0.9486 (0.78-1)	0.9414 (0.76-1)	0.9556 (0.78-1)	**Performance during the testing phase**
Specificity	0.9356 (0.75-1)	0.9236 (0.75-1)	0.9069 (0.75-1)	
AUC	0.9762 (0.90-1)	0.9864 (0.94-1)	0.9874 (0.95-1)	
Accuracy	0.9429 (0.84-1)	0.9333 (0.86-1)	0.9321 (0.82-1)	

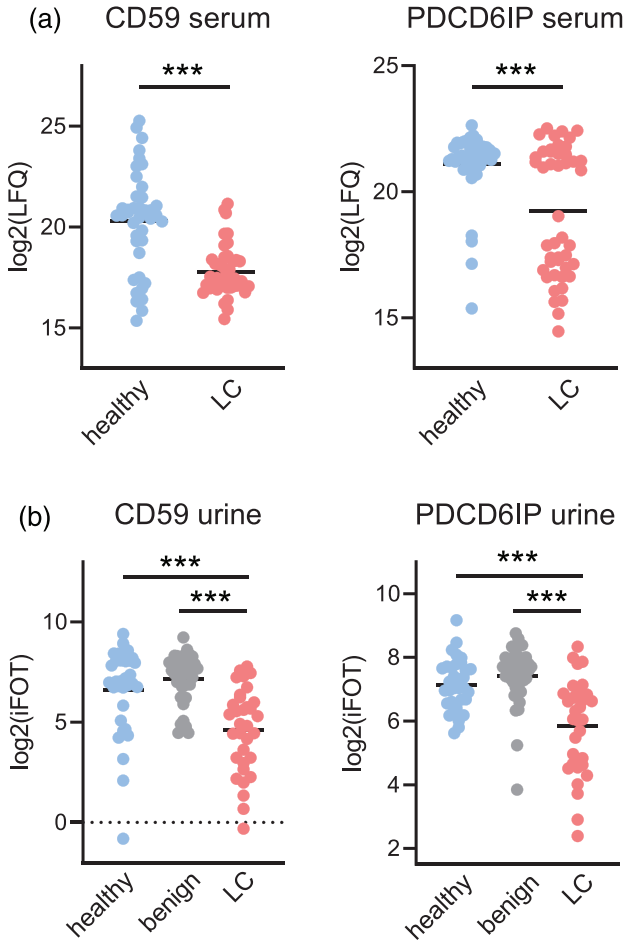

FIGURE 4 **Comparison of two putative markers in serum and urine samples. (A)** Log2(LFQ) values for CD59 and PDCD6IP in our dataset. The statistical analysis was performed as described in the material and methods section. ***: p adjusted < 0.001; **: p adjusted < 0.01. **(B)** The same two proteins were analyzed in [5] in a urine dataset composed of 33 healthy subjects, 40 patients with benign pulmonary conditions (pneumonia, and chronic obstructive pulmonary disease) and 33 patients with lung cancer. ***: $p < 0.001$ by one-way ANOVA followed by Tukey's multiple comparisons test

In this study, we profiled the protein composition of 87 human serum samples from healthy, high-risk donors, and lung cancer patient. To overcome the challenges related with the analysis of serum proteomes, we chose to analyze circulating MVs, identifying a list of candidate markers for future validation. Of note, because no specific step to enrich for tumor-derived MVs was performed, potential biomarkers identified through our approach may derive from the tumor cells, or may originate from other cell populations as a reaction to the growing tumor.

Among the differentially expressed proteins, ARSA showed the highest fold change and discriminant power, and could represent a promising biomarker candidate for the diagnosis of early-stage lung cancer. In addition, to try to further increase the discriminating power, we developed a combinatorial model for cancer early detection, which allows the discrimination of lung cancer patients from control cases with high sensitivity and specificity, and highlights additional candidate markers to be validated. Biochemical assays such as ELISA (enzyme-linked immunosorbent assay) could be employed to validate these changes on a large cohort of patients, and to develop a diagnostic tool easily translatable in a clinical setting.

Furthermore, we envision that the current serum protein signature could be combined with other molecular profiles to achieve the highest possible diagnostic power. For instance, the analysis of volatile organic compounds from breath, by either an electronic nose device or gas chromatography-mass spectrometry, has suggested the existence of a specific fingerprint or "breathprint" able to recognize lung cancer from both healthy donors and other pulmonary diseases [19, 20]. Cancer volatile organic compounds analysis has been also applied to urine samples, generating promising results [21]. In addition, urine proteomics combined with the development of a machine learning model generated a short list of differentially expressed proteins in lung cancer patients compared with healthy subjects [5]. Interestingly, some of the markers found in our serum analysis (CD59 and PDCD6IP) were also found significantly decreased in the urines of lung cancer patients compared with normal subjects, or patients with benign pulmonary diseases.

The results obtained in this study provide promising preliminary evidence of the value of a proteomics serum signature for the diagnosis of early-stage lung cancer, which will need to be validated -beyond this study- in a larger cohort of patients. The development of routine techniques for the quantitation of most relevant protein biomarkers will also be important to allow their application for clinic uses in the future.

5 | ASSOCIATED DATA

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium [22] via the PRIDE partner repository with the dataset identifier PXD034221.

ACKNOWLEDGMENTS

The authors thank Evelyn Oliva Savoia, Alessia Tommasini and Eugenio Graceffo for technical help for sera sample preparation prior to MS analysis. This work was supported partially by the Italian Ministry of Health with Ricerca Corrente and 5 × 1000 funds. Work in TB group is supported by the Italian Association for Cancer Research, grant number IG-2018-21834 (to T.B.) and by EPIC-XS, project number 823839, funded by the Horizon 2020 programme of the European Union.

Open access funding enabled and organized by BIBLIOSAN.

CONFLICTS OF INTEREST

All the authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The database search "proteinGroups" output file and the LFQ data are available as supplemental material.

ORCID

Roberto Gasparri https://orcid.org/0000-0001-7922-7061
Davide Tricarico https://orcid.org/0000-0001-5391-8171
Tiziana Bonaldi https://orcid.org/0000-0003-3556-1265

REFERENCES

1. Kramer, B. S., Berg, C. D., Aberle, D. R., & Prorok, P. C. (2011). Lung cancer screening with low-dose helical CT: results from the National Lung Screening Trial (NLST). *Journal of Medical Screening*, *18*, 109–111. https://doi.org/10.1258/jms.2011.011055
2. Hawkridge, A M., & Muddiman, D C. (2009). Mass Spectrometry–Based Biomarker Discovery: Toward a Global Proteome Index of Individuality. *Annual Review of Analytical Chemistry*, *2*, 265–277. https://doi.org/10.1146/annurev.anchem.1.031207.112942
3. Cho, W. C. S. (2016). Application of proteomics in non-small-cell lung cancer. *Expert Rev. Proteomics*, *13*, 1–4. https://doi.org/10.1586/14789450.2016.1121813
4. Gasparri, R., Sedda, G., Noberini, R., Bonaldi, T., & Spaggiari, L. (2020). Clinical Application of Mass Spectrometry-Based Proteomics in Lung Cancer Early Diagnosis. *Proteomics Clin Appl*, *14*, 1900138. https://doi.org/10.1002/prca.201900138
5. Zhang, C., Leng, W., Sun, C., Lu, T., Chen, Z., Men, X., Wang, Y., Wang, G., Zhen, B., & Qin, J. (2018). Urine Proteome Profiling Predicts Lung Cancer from Control Cases and Other Tumors. *EBioMedicine*, *30*, 120–128. https://doi.org/10.1016/j.ebiom.2018.03.009
6. Geyer, P. E., Holdt, L. M., Teupser, D., & Mann, M. (2017). Revisiting biomarker discovery by plasma proteomics. *Molecular Systems Biology*, *13*, 942. https://doi.org/10.15252/msb.20156297
7. S, E. L. A., Mager, I., Breakefield, X. O., & Wood, M. J. (2013). Extracellular vesicles: biology and emerging therapeutic opportunities. *Nature Reviews Drug Discovery*, *12*, 347–357.
8. Voloshin, T., Fremder, E., & Shaked, Y. (2014). Small but mighty: microparticles as mediators of tumor progression. *Cancer Microenvironment: official journal of the International Cancer Microenvironment Society*, *7*, 11–21. https://doi.org/10.1007/s12307-014-0144-8
9. Piersma, S. R., Warmoes, M. O., De Wit, M., De Reus, I., Knol, J. C., & Jiménez, C. R. (2013). Whole gel processing procedure for GeLC-MS/MS based proteomics. *Proteome Science*, *11*, 17. https://doi.org/10.1186/1477-5956-11-17
10. Meier, F., Geyer, P E., Virreira Winter, S., Cox, J., & Mann, M. (2018). BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nature Methods*, *15*, 440–448. https://doi.org/10.1038/s41592-018-0003-5
11. Hendriks, I. A., Akimov, V., Blagoev, B., & Nielsen, M. L. (2021). MaxQuant.Live Enables Enhanced Selectivity and Identification of Peptides Modified by Endogenous SUMO and Ubiquitin. *Journal of Proteome Research*, *20*, 2042–2055. https://doi.org/10.1021/acs.jproteome.0c00892
12. Wichmann, C., Meier, F., Virreira Winter, S., Brunner, A. D., Cox, J., & Mann, M. (2019). MaxQuant.Live Enables Global Targeting of More Than 25,000 Peptides. *Molecular & Cellular Proteomics*, *18*, 982–994. https://doi.org/10.1074/mcp.TIR118.001131
13. Cox, J., Hein, M Y., Luber, C A., Paron, I., Nagaraj, N., & Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & Cellular Proteomics*, *13*, 2513–2526. https://doi.org/10.1074/mcp.M113.031591
14. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., & Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods*, *13*, 731–740. https://doi.org/10.1038/nmeth.3901
15. Gasa, S., Makita, A., Kameya, T., & Kodama, T. (et al., 1980). Elevated activities and properties of arylsulfatases A and B and B-variant in human lung tumors. *Cancer Research*, *40*, 3804–3809.
16. Laidler, P., Kowalski, D., & Silberring, J. (1991). Arylsulfatase A in serum from patients with cancer of various organs. *Clinica Chimica Acta*, *204*, 69–77. https://doi.org/10.1016/0009-8981(91)90218-2
17. Xu, X., Han, L., Zhao, G., Xue, S., Gao, Y., Xiao, J., Zhang, S., Chen, P., Wu, Z. Y., Ding, J., Hu, R., Wei, B., & Wang, H. (2017). LRCH1 interferes with DOCK8-Cdc42–induced T cell migration and ameliorates experimental autoimmune encephalomyelitis. *Journal of Experimental Medicine*, *214*, 209–226. https://doi.org/10.1084/jem.20160068
18. Salama, M F., Liu, M., Clarke, C J., Espaillat, M. P., Haley, J D., Jin, T., Wang, D., Obeid, L M., & Hannun, Y A. (2019). Correction: PKCα is required for Akt-mTORC1 activation in non-small cell lung carcinoma (NSCLC) with EGFR mutation. *Oncogene*, *38*, 7366. https://doi.org/10.1038/s41388-019-1019-8
19. Gasparri, R., Romano, R., Sedda, G., Borri, A., Petrella, F., Galetta, D., Casiraghi, M., & Spaggiari, L. (2018). Diagnostic biomarkers for lung cancer prevention. *Journal of Breath Research*, *12*, 027111. https://doi.org/10.1088/1752-7163/aa9386
20. Marzorati, D., Mainardi, L., Sedda, G., Gasparri, R., Spaggiari, L., & Cerveri, P. (2019). A review of exhaled breath: a key role in lung cancer diagnosis. *Journal of Breath Research*, *13*, 034001. https://doi.org/10.1088/1752-7163/ab0684
21. Gasparri, R., Capuano, R., Guaglio, A., Caminiti, V., Canini, F., Catini, A., Sedda, G., Paolesse, R., Di Natale, C., & Spaggiari, L. (2022). Volatolomic urinary profile analysis for diagnosis of the early stage of lung cancer. *Journal of Breath Research*, *16*, https://doi.org/10.1088/1752-7163/ac88ec

22. Vizcaíno, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Ríos, D., Dianes, J. A., Sun, Z., Farrah, T., Bandeira, N., Binz, P. A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R. J., Kraus, H. J., Albar, J. P., … Hermjakob, H. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology*, *32*, 223–226. https://doi.org/10.1038/nbt.2839

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Gasparri, R., Noberini, R., Cuomo, A., Yadav, A., Tricarico, D., Salvetto, C., Maisonneuve, P., Caminiti, V., Sedda, G., Sabalic, A., Bonaldi, T., & Spaggiari, L. (2023). Serum proteomics profiling identifies a preliminary signature for the diagnosis of early-stage lung cancer. *PROTEOMICS – Clinical Applications*, e2200093. https://doi.org/10.1002/prca.202200093