

Identification of Plasma LGALS3BP via Proteomics and Development of a Non-invasive Diagnostic Model for Early-Stage Lung Adenocarcinoma

Rong Gong

First Affiliated Hospital of Kunming Medical University

Rongya Hu

First Affiliated Hospital of Kunming Medical University

Qiuyue Xu

First Affiliated Hospital of Kunming Medical University

Xiao Wang

First Affiliated Hospital of Kunming Medical University

Hui Zhang

First Affiliated Hospital of Kunming Medical University

Leiyang Dai

First Affiliated Hospital of Kunming Medical University

Chenglu He

First Affiliated Hospital of Kunming Medical University

Junxi Pan

First Affiliated Hospital of Kunming Medical University

Yong Duan

First Affiliated Hospital of Kunming Medical University

Fujun Zhang

zhangfujun@yduy.cn

First Affiliated Hospital of Kunming Medical University

Article

Keywords: Lung adenocarcinoma, Benign pulmonary nodules, LGALS3BP, Plasma biomarker, Proteomics, Diagnostic model

Posted Date: August 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10421610/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Distinguishing early-stage lung adenocarcinoma (LUAD) from benign pulmonary nodules remains challenging. This study aimed to identify plasma biomarkers and develop a non-invasive diagnostic model. Label-free quantitative proteomics screened tumor tissues and plasma, identifying LGALS3BP as a candidate. Validation involved RT-qPCR and ELISA in independent cohorts. A diagnostic model incorporating plasma LGALS3BP, CEA, family history, and nodule characteristics was developed (n = 138) and validated (n = 47). LGALS3BP was significantly upregulated in LUAD tissues and plasma. The combined model achieved an AUC of 0.893 in the training cohort, significantly outperforming the baseline model (NRI = 0.185, P = 0.019). Validation confirmed robust performance (AUC = 0.856). The model demonstrated high clinical utility, effectively stratifying risk with a positive predictive value of 91.3% in the intermediate-to-high risk group. In conclusion, plasma LGALS3BP is a novel biomarker for LUAD. The developed model integrating LGALS3BP with clinical factors provides a robust tool for differentiating malignant from benign nodules, potentially improving risk stratification and minimizing unnecessary invasive interventions.

Introduction

Lung cancer remains the leading cause of cancer-related mortality worldwide^[1], with lung adenocarcinoma (LUAD) being the most prevalent subtype^[2]. While low-dose computed tomography screening has improved the detection of pulmonary nodules, accurately distinguishing between benign and malignant nodules remains a significant clinical challenge^[3]. The current lack of highly specific non-invasive diagnostic methods has resulted in two major clinical problems: (1) a considerable proportion of patients with benign pulmonary nodules (BPN) are subjected to unnecessary invasive procedures, such as biopsies or surgery, due to false-positive diagnoses; and (2) some early-stage LUAD cases are frequently missed or diagnosed with delay, allowing for disease progression and ultimately contributing to poor patient outcomes. Conventional tumor markers have proven inadequate for this purpose, often lacking the requisite sensitivity and specificity for reliable early detection^[4]. To address this, we employed label-free quantitative proteomics to identify robust candidates, screening Galectin-3-binding protein (LGALS3BP) as a top target. Recognizing the clinical advantages of nucleic acid detection, we further developed a multi-parameter diagnostic model incorporating *LGALS3BP* expression levels for the non-invasive differentiation of early-stage LUAD from BPN.

Methods

Study Design and Participants

The study design consisted of an initial phase utilizing archived samples for proteomic profiling and candidate biomarker screening, and a subsequent phase focused on the development and validation of plasma-based diagnostic models. The entire study was conducted at the First Affiliated Hospital of Kunming Medical University. The enrollment process for the plasma analysis phase is outlined in Fig. 1 (Fig. 1).

Patient Recruitment and Sample Collection

For tissue proteomics analysis, LUAD patients who underwent surgical treatment between January and April 2022 were enrolled. Inclusion criteria were: (1) no prior anti-tumor therapy including radiotherapy, chemotherapy,

targeted therapy, or immunotherapy; (2) histopathologically confirmed LUAD; (3) stage I disease according to the 8th edition AJCC staging system^[5]. Exclusion criteria included: (1) previous history of lung tumors or other malignancies; (2) inability to participate in follow-up. For plasma analysis, patients with pulmonary nodules who underwent surgical treatment or bronchoscopic biopsy were enrolled between January and December 2024. Inclusion criteria were: (1) imaging features suggestive of malignancy requiring surgical intervention or biopsy; (2) no prior disease-related treatment. Exclusion criteria included: (1) insufficient plasma volume (< 1.5 mL); (2) incomplete clinical data; (3) non-adenocarcinoma histology; (4) stage II or higher disease according to the 8th edition AJCC staging system; (5) previous history of lung tumors or other malignancies. Patients were classified into LUAD or BPN groups based on final histopathological diagnosis.

Clinical Data Collection

Comprehensive clinical data were collected including demographics (age, gender), smoking history, family history of lung cancer, serum tumor markers, and imaging features. Thin-section CT images were reviewed to characterize nodule density as pure ground-glass nodule (pGGN), solid nodule, or part-solid nodule. High-risk imaging features including lobulation, spiculation, vacuole sign/air bronchogram sign, cavity formation, pleural retraction sign, and vascular convergence sign were systematically recorded.

Label-free Quantitative Proteomics

Quantitative proteomic analysis was conducted by PTM Biolabs, Inc. (Hangzhou, China). Differentially expressed proteins were identified using thresholds of $|\log_2FC| > 0.585$ and $P < 0.05$.

Expression validation based on public databases

To ensure the reliability of tissue-derived biomarkers, we referred to the proteomic data integrated within The Human Protein Atlas (HPA) ^[6]. We utilized the Clinical Proteomic Tumor Analysis Consortium (CPTAC) LUAD dataset^[7] available through the HPA portal to quantify protein abundance in tumor versus normal tissues. Furthermore, to validate the expression differences at the transcriptional level, we leveraged the GEPIA2^[8] web server, a well-established platform for RNA-seq analysis, to analyze differential expression based on The Cancer Genome Atlas (TCGA) data. Additionally, the concordance between transcriptomic (TCGA) and proteomic (CPTAC) levels was evaluated to confirm expression consistency.

Validation of Consistency between Plasma mRNA and Protein Expression

To validate the consistency between mRNA and protein levels, we performed an orthogonal analysis using an independent cohort. Plasma *LGALS3BP* mRNA levels were quantified by RT-qPCR (Takara, Dalian, China), while protein concentrations were measured by ELISA (Proteintech, Wuhan, China). Specific primer pairs for *LGALS3BP* were selected from the pre-designed library of Sangon Biotech (Shanghai, China) (sequences listed in Table 1). The specificity of the amplification was further confirmed by Sanger sequencing. The correlation between mRNA

and protein expression was evaluated using Pearson's or Spearman's rank correlation coefficients, depending on the normality of the data distribution.

Table 1
Primer sequences used for RT-qPCR

Gene	Direction	Sequence (5' – 3')
<i>LGALS3BP</i>	Forward	GATCCAGTCGTAGTCAGTGCTT
	Reverse	AGACCCTCGTCCAAGACTCTTA
<i>β-actin</i>	Forward	CCTGGCACCCAGCACAAT
	Reverse	GGGCCGGACTCGTCATAC

Diagnostic Model Construction and Validation

Cohort Division

Among the patients in the plasma cohort, 185 patients (124 LUAD and 61 BPN) was stratified randomly divided into a training cohort (n = 138, 70%) and a validation cohort (n = 47, 30%) using the random number generator in SPSS 25.0 software. The demographic and clinical characteristics were well-balanced between the two cohorts.

Variable Selection and Model Development

In the training cohort, baseline characteristics were compared between the LUAD and BPN groups to identify candidate predictors. These predictors included age, gender, smoking history, family history of cancer, nodule characteristics, high-risk imaging features, and levels of carcinoembryonic antigen (CEA), cytokeratin 19 fragment antigen 21 – 1 (CYFRA21-1), Squamous Cell Carcinoma Antigen (SCC), and plasma *LGALS3BP* mRNA levels. The inclusion of *LGALS3BP* mRNA levels was justified by strong correlation with protein expression in prior consistency verification ($\rho = 0.833$, $P < 0.001$) and its cost-effectiveness. Variables with $P < 0.1$ in the group comparisons were subsequently included in the multivariate logistic regression model using the forward likelihood ratio (LR) stepwise method. Nodule characteristics were treated as a categorical variable with pGGN set as the reference category.

Model Performance Assessment

Model performance was evaluated in terms of discrimination, calibration, and clinical utility. Discrimination was assessed using ROC curves, reporting the area under the curve (AUC) with 95% confidence intervals (CIs). The optimal cut-off value was determined using the Youden index, and sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV) were calculated for both cohorts. Calibration was visually inspected using calibration plots with bias-corrected curves (1,000 bootstrap resamples) and quantified via the Hosmer–Lemeshow test. Clinical utility was validated through Decision Curve Analysis (DCA) and Clinical Impact Curves (CIC).

Statistical Analysis

All statistical analyses were performed using IBM SPSS Statistics (version 25.0; IBM Corp., Armonk, NY, USA), GraphPad Prism software (version 9.5.0; GraphPad Software, San Diego, CA, USA) and R software (version 4.3.3;

R Foundation for Statistical Computing). Continuous variables are presented as mean $\pm$ standard deviation (SD) or median with interquartile range (IQR), and categorical variables as frequencies and percentages. Normality was assessed using the Shapiro-Wilk test. Group differences were evaluated using the independent Student's t-test or Mann-Whitney U test for continuous variables, and the Chi-square (χ^2) or Fisher's exact test for categorical variables. For proteomic data, *P*-values were adjusted using the Benjamini-Hochberg procedure to control the false discovery rate (FDR). Functional enrichment and logistic regression-based nomogram construction were conducted using the *clusterProfiler* and *rms* packages in R, respectively. A two-tailed *P* < 0.05 (or adjusted *P* < 0.05 for proteomics) was considered statistically significant.

Results

Tissue Proteomics Reveals Extensive Protein Dysregulation in Early-Stage LUAD

Label-free quantitative proteomics identified 4,926 proteins in LUAD tumor tissues (n = 10) and 4,554 proteins in normal adjacent tissues (NATs, n = 5). We identified 1,188 differentially expressed proteins, including 761 upregulated and 427 downregulated proteins in LUAD tissues compared to NATs (Fig. 2A-B). Hierarchical clustering analysis clearly segregated tumor samples from normal tissues based on differentially expressed protein profiles. The volcano plot revealed that upregulated proteins were predominant, suggesting widespread activation of pro-oncogenic signals and protein expression reprogramming in LUAD tissues (Fig. 2C).

GO enrichment analysis indicated that differentially expressed proteins were markedly enriched in biological processes involving monosaccharide/hexose metabolism, nucleotide-sugar metabolism and adhesion-dependent cell spreading, implying glycolytic reprogramming and membrane structural remodeling as core features of LUAD (Supplementary, Fig. S1C). At the cellular component level, these proteins were mainly located in secretory granules, actin cytoskeleton, cell-substrate junctions, extracellular matrix and endoplasmic reticulum chaperone complexes (Fig. S1D). Molecular functions were enriched in actin binding, adhesion molecule binding, monosaccharide binding, as well as antioxidant and multiple oxidoreductase activities (Fig. S1B).

KEGG enrichment uncovered prominent enrichment in three categories of pathways: metabolic cascades including pentose and glucuronate interconversions, folate biosynthesis and glycolysis/gluconeogenesis; oncogenic signaling such as pathways in cancer and the PI3K-Akt pathway; and metastasis-associated pathways including ECM-receptor interaction and focal adhesion (Fig. S1A).

Plasma Proteomics Identifies Circulating Protein Alterations in LUAD Patients

Plasma proteomic analysis comparing LUAD patients (n = 6) and BPN patients (n = 4) identified 334 differentially expressed proteins, with 108 upregulated and 226 downregulated in LUAD plasma (Fig. 3A-C). The predominant downregulation of plasma proteins may reflect tumor-mediated consumption of host proteins, impaired hepatic protein synthesis, or systemic immune dysregulation associated with malignancy.

GO enrichment analysis indicated that plasma differentially expressed proteins were enriched in biological processes governing protein metabolism, phosphorylation, MAPK cascade, kinase activity and immunoglobulin-mediated immunity (Fig. S2B). For cellular components, these proteins mainly resided at the cell leading edge,

microtubule organizing center, cytoplasmic plasma membrane and immunoglobulin complex (Fig. S2C). Molecular functions were concentrated in phospholipid binding, GTPase regulatory activity, chaperone binding and immunoglobulin receptor binding (Fig. S2D).

KEGG enrichment identified enriched pathways grouped into cancer-associated cascades (proteoglycans in cancer, microRNAs in cancer, focal adhesion, apoptosis) and multiple key signaling axes including Ras, Rap1, cAMP, sphingolipid and phospholipase D signaling pathways (Fig. S2A).

Screening for Consistently Upregulated Proteins in LUAD Tissues and Plasma

To identify plasma biomarkers for early-stage LUAD diagnosis, an integrative analysis of tissue and plasma proteomics data was performed. Proteins that were consistently upregulated in both LUAD tissues and plasma were selected as candidate biomarkers with potential clinical utility. This analysis identified 11 proteins showing consistent expression patterns: CTSC, GLB1, CTSA, SEC16A, MIF, SERPINA1, DCXR, ASS1, LGALS3BP, LCP1, and BAX (Table 2).

Table 2						
Candidate biomarkers identified by consistent expression patterns in tissue and plasma proteomics						
Protein accession	Gene name	Tissue Proteomics		Plasma Proteomics		Trend consistenc
		(Tumors vs. NATs)		(LUADs vs. BPNs)		
Fold Change		P-value	Fold Change	P-value		
P53634	CTSC	3.657	0.003722344	3.597527528	0.000691155	Yes
P16278	GLB1	2.768	0.020536054	2.318447952	0.008006277	Yes
P10619	CTSA	2.581	0.020149999	1.916451405	0.008867449	Yes
O15027	SEC16A	1.652	0.019252753	19.2572112	0.011995563	Yes
P14174	MIF	5.195	7.94E-07	5.754508846	0.020642336	Yes
P01009	SERPINA1	3.201	0.017348947	19.2572112	0.011995563	Yes
Q7Z4W1	DCXR	1.762	0.032932	1.927564219	0.024353873	Yes
P00966	ASS1	2.523	0.003946638	3.485346008	0.029173359	Yes
Q08380	LGALS3BP	1.897	0.037038299	37.54379839	0.031786816	Yes
P13796	LCP1	3.634	3.65E-05	3.902405954	0.040654777	Yes
Q07812	BAX	3.339	0.000258164	3.72201442	0.044630408	Yes

Database Validation of Candidate Biomarkers

Validation using the Human Protein Atlas revealed that 6 of the 11 candidate proteins (GLB1, SEC16A, SERPINA1, LCP1, LGALS3BP, and DCXR) were significantly differentially expressed between LUAD tumors and

NATs. Notably, *GLB1*, *SEC16A*, and *LGALS3BP* showed the most significant upregulation in tumor tissues (Fig. 4A).

To assess concordance between mRNA and protein expression, we analyzed TCGA RNA-sequencing data. *LGALS3BP* mRNA levels were significantly elevated in LUAD tissues compared to paired normal tissues ($P < 0.001$, Fig. 4B), consistent with proteomic findings. In contrast, *GLB1* and *SEC16A* showed no significant differences at the transcriptional level.

Based on these validation results and considering the consistency between mRNA and protein expression, *LGALS3BP* was selected as the priority candidate for subsequent clinical validation.

ELISA and RT-qPCR Validation of Plasma *LGALS3BP*

We performed orthogonal verification in an independent cohort of 42 patients (27 LUAD and 15 BPN). Spearman's rank correlation analysis revealed a strong and statistically significant positive correlation between plasma *LGALS3BP* mRNA levels and protein concentrations in the overall cohort ($\rho = 0.833$, $P < 0.001$; Fig. 5). This high concordance supports the use of *LGALS3BP* mRNA as a reliable surrogate for protein expression in subsequent analyses.

Baseline Clinical Characteristics and Expression of Plasma *LGALS3BP* mRNA in the Study Cohort

The complete cohort comprised 185 patients, including 124 LUAD and 61 BPN patients. The cohort was randomly divided into a training cohort ($n = 138$, 92 LUAD and 46 BPN) and a validation cohort ($n = 47$, 32 LUAD and 15 BPN). In the complete cohort, the median relative expression of plasma *LGALS3BP* mRNA in the BPN group was 1.01 (IQR: 0.59–1.69), whereas it was significantly upregulated in the LUAD group [median: 3.16 (IQR: 1.48–6.42); $P < 0.001$] (Fig. 6). Accordingly, univariate analysis in the training cohort identified six potential predictors with $P < 0.1$ (Table 3), which were subsequently included in the multivariate analysis.

Table 3
Clinical
characteristics

Variables were compared between LUAD and BPN groups using Student's t-test for continuous variables with normal distribution, Mann-Whitney U test for non-normally distributed continuous variables, and Chi-square test or Fisher's exact test for categorical variables.

Multivariate logistic regression analysis and combined prediction model construction

Multivariate logistic regression was performed to identify independent predictors for early-stage LUAD (Table 4). Multicollinearity diagnostics confirmed model stability, with Variance Inflation Factors (VIF) ranging from 1.048 to 1.093 and tolerance values between 0.915 and 0.954, indicating no significant collinearity. Ultimately, four variables with $P < 0.05$ were identified as independent predictors and included in the Combined Prediction Model. The final logistic regression equation was as follows: $\text{Logit}(P) = -1.935 + 0.657 \times \text{CEA} + 0.601 \times \text{LGALS3BP} + 2.398 \times \text{Family history} - 1.420 \times \text{Solid nodule} + 2.272 \times \text{Part-solid nodule}$. A nomogram was subsequently constructed to provide an intuitive visualization tool for clinical risk assessment. This nomogram allows

clinicians to estimate the individual probability of LUAD by summing the points corresponding to each predictor's value (Fig. 7).

Table 4
Multivariate logistic regression analysis of independent risk factors for early-stage LUAD using LR stepwise method.

Family history of lung cancer	B	P-value	OR	95% CI
	2.398	0.038	11.260	1.148–105.51
CEA	0.657	0.019	1.720	1.115–3.337
LGALS3BP mRNA	0.601	0.000	1.733	1.334–2.494
Nodule type				
pGGN	–	–	1.00 (Ref.)	–
Part-solid nodule	2.272	0.043	9.702	1.074–87.678
Solid nodule	-1.420	0.008	0.242	0.085–0.686

Diagnostic Performance of the Combined Model for Early-Stage LUAD

Performance in the training cohort

The combined prediction model demonstrated robust statistical performance in the training cohort. The likelihood ratio test revealed statistically significant differences ($\chi^2 = 75.130$, $P < 0.001$), and Nagelkerke's R^2 reached 0.583, indicating that the model accounted for a considerable proportion of variance in the diagnosis of early-stage LUAD.

In terms of discriminative ability, the Combined Prediction Model yielded an AUC of 0.893 (95% CI: 0.842–0.944, Fig. 8A), which was significantly superior to single predictors, including *LGALS3BP* mRNA (AUC = 0.806, 95% CI: 0.733–0.880), nodule features (AUC = 0.644, 95% CI: 0.550–0.743), family history (AUC = 0.603, 95% CI: 0.555–0.651), and CEA (AUC = 0.601, 95% CI: 0.501–0.701). Compared to the baseline clinical model incorporating only CEA, family history, and nodule characteristics (AUC = 0.803, 95% CI: 0.731–0.880), the addition of *LGALS3BP* significantly improved diagnostic performance. The net reclassification improvement (NRI) was 0.185 (95% CI: 0.030–0.339, $P = 0.019$), and the integrated discrimination improvement (IDI) was 0.195 (95% CI: 0.128–0.262, $P < 0.001$), demonstrating the substantial incremental value of *LGALS3BP* in risk stratification.

At the optimal cut-off value of 0.631 (corresponding to the maximum Youden index of 0.696), the model achieved a sensitivity of 78.3%, specificity of 91.3%, overall accuracy of 82.6%, PPV of 94.7%, and NPV of 67.7%.

Calibration analysis demonstrated good agreement between predicted probabilities and observed outcomes (Fig. 8B), which was further supported by a non-significant Hosmer–Lemeshow test ($\chi^2 = 8.219$, $P = 0.412$). Furthermore, DCA demonstrated that across a broad threshold probability range (roughly 5%–85%), the combined model achieved remarkably higher net clinical benefit than the “treat-all” and “treat-none” strategies (Fig. 8C). The CIC illustrated that rising risk thresholds reduced false-positive counts while retaining a high true-

positive detection rate for LUAD. At moderate risk thresholds, the model achieved a favorable balance between sensitivity and specificity, supporting its capability to streamline clinical practice and improve diagnostic discrimination between early-stage LUAD and BPN (Fig. 8D).

Performance in the validation cohort

In the validation cohort, the model maintained robust diagnostic performance with an AUC of 0.856 (95% CI: 0.741–0.971, Fig. 9A). At the same cut-off of 0.631, the model achieved a sensitivity of 81.3%, specificity of 86.7%, and an overall accuracy of 83.0%. Good calibration was confirmed by the Hosmer-Lemeshow test ($\chi^2 = 6.428, P = 0.599$). While the calibration curve largely aligned with the diagonal reference line (Fig. 9B), minor discrepancies were noted, with the curve dipping slightly below the diagonal between probabilities of 0.25 and 0.5, and rising slightly above it when probabilities exceeded 0.75. DCA demonstrated consistent net clinical benefit across threshold probabilities of 10%–80%, supporting the model's clinical utility for pulmonary nodule management (Fig. 9C).

Clinical Utility Assessment in the Thoracic Cohort

Based on DCA-derived net benefits, patients were stratified into three risk tiers (Table 5). In the low-risk group (< 40%), the observed malignancy rate was 42.4% (14/33). Subgroup analysis of the false-negative cases indicated that pGGNs accounted for 50% (4/8) of these misclassifications. In the intermediate-to-high risk group (40%–85%), the model demonstrated a positive predictive value (PPV) of 91.3% (21/23) for LUAD. In the very high-risk category (> 85%), the model exhibited high accuracy with only one misclassification among the predicted LUAD cases.

Table 5
Classification performance of the model across different predicted probability thresholds.

Predicted Probability	Predicted Class	Ground Truth: BPN	Ground Truth: LUAD	Total
< 40%	BPN	19	14	33
	LUAD	0	0	0
40%–85%	BPN	13	8	21
	LUAD	2	21	23
> 85%	BPN	0	0	0
	LUAD	1	76	77

Discussion

Principal Findings and Clinical Implications

By integrating tissue and plasma proteomics, this study identifies LGALS3BP as a novel, highly specific plasma biomarker for early-stage LUAD. We further develop a combined diagnostic model incorporating *LGALS3BP* mRNA with clinical variables, which demonstrates robust performance in distinguishing malignant from benign

pulmonary nodules. The consistent net benefit observed in decision curve analyses across cohorts highlights the model's clinical utility in real-world practice.

LGALS3BP: A Novel Biomarker with Biological Plausibility

The biological functions of LGALS3BP align closely with the pathogenesis of tumor^[9]. It promotes tumor invasion by interacting with galectin-3 and integrins, regulating focal adhesion and cytoskeletal organization—processes significantly enriched in our tissue proteomics (Fig. S1D). Furthermore, its activation of the PI3K-AKT pathway (Fig. S1A) underscores its role in driving proliferation^[10]. Beyond invasion, LGALS3BP may contribute to immune evasion by binding to CD33-Siglecs on myeloid cells^{[11],[12]}. This is reflected systemically in the enrichment of complement and coagulation cascades in plasma (Fig. S2A). Additionally, aberrant glycosylation enhances LGALS3BP's oncogenic functions^[13]. The observed reprogramming of glycosylation-related metabolism in our tissue data (Fig. S1C) may directly drive these altered glycan patterns in early-stage disease. Notably, although LGALS3BP lacks a classical signal peptide, its enrichment in vesicular components (Fig. S1D) supports exosome-mediated secretion^[14]. The strong correlation between plasma LGALS3BP mRNA and protein levels ($\rho = 0.833$) indicates that tumor-derived EVs carry both transcripts and proteins reflecting the parent tumor's profile. This concordance validates qPCR-based mRNA detection as a cost-effective, high-throughput alternative to protein assays, facilitating clinical translation.

Comparison with Existing Diagnostic Approaches

Traditional Tumor Markers

Conventional markers such as CEA and CYFRA 21 – 1 exhibit limited sensitivity for early-stage disease. In contrast, *LGALS3BP* alone achieved an AUC of 0.806, and its integration into the combined model significantly boosted performance to an AUC of 0.893 with 91.3% specificity, demonstrating clear superiority.

Clinical Prediction Models

Although the Brock model^[15] shows excellent performance (AUCs: 0.960–0.970), it lacks variables specific to unique environmental exposures, such as the coal-derived carcinogens prevalent in Xuanwei and Fuyuan, China^[16]. In this specific high-risk setting characterized by pronounced familial aggregation, our model explicitly integrates *family history of cancer*—a critical variable capturing hereditary susceptibility—alongside the molecular biomarker LGALS3BP. By combining epidemiological risk factors with tumor-associated molecular signatures, our model offers superior biological relevance and predictive precision for populations with distinct gene-environment backgrounds.

Recent studies have explored circulating biomarkers for lung nodule diagnosis, challenges regarding model robustness remain. For instance, a previously reported 3-miRNA signature demonstrated satisfactory performance in its training cohort (AUC = 0.920), yet its specificity dropped significantly to 45.0%–54.2% in independent validation cohorts^[17]. This substantial discrepancy suggests potential overfitting and limited generalizability across different populations.

Similarly, Jia et al. reported a six-protein panel (APOA4, CD14, PFN1, APOB, PLA2G7, and IGFBP2) identified through plasma proteomic profiling. While their model demonstrated high sensitivity (91.4%–100%) across different cohorts, it consistently suffered from low specificity (35%–54.2%) in independent validations^[18]. Our model achieves a more balanced profile. This equilibrium is clinically critical, particularly our high positive predictive value (94.7%), which provides the confidence needed to justify invasive interventions. Furthermore, by integrating a single plasma biomarker with routine clinical variables, our approach offers a simpler, cost-effective alternative to complex multi-marker assays or next-generation sequencing platforms.

Radiomics and AI-Based Approaches

Despite promising performance (AUCs: 0.85–0.93)^[19], radiomics and AI approaches face limitations regarding strict imaging protocols, "black box" interpretability, and cross-center generalizability^[20]. Conversely, our model utilizes straightforward nodule morphologies accessible to general radiologists. Crucially, the addition of plasma *LGALS3BP* offers orthogonal biological insights, enabling the detection of molecularly aggressive LUADs.

Clinical Decision-Making and Risk Stratification

The risk stratification analysis reveals distinct performance characteristics across probability tiers that inform clinical application.

In the low-risk group (< 40% probability), the observed malignancy rate of 42.4% indicates a suboptimal NPV. This limitation is predominantly driven by pGGNs, which accounted for 50% of false negatives in this stratum. These results underscore that a low model probability should not supersede guideline-based surveillance for pGGNs. Consequently, these patients require continued radiological follow-up, with the model serving as supplementary rather than definitive evidence.

In the intermediate-to-high risk group (40–85% probability), the model demonstrates robust positive predictive value (91.3% in training, 90.9% in surgical cohort), confirming its utility in the diagnostic "gray zone" where clinical and radiological features are equivocal. This performance profile supports using the model to guide decisions toward tissue diagnosis (biopsy or surgical resection) while minimizing false-positive interventions. The high PPV indicates that patients stratified to this tier warrant expedited diagnostic workup.

In the highest risk category (> 85% probability), the model achieved 95.0% accuracy in the training cohort, validating its reliability for prioritizing surgical intervention. However, given the small validation cohort (n = 47), this threshold-specific performance requires confirmation in larger independent samples before being used as a standalone criterion for bypassing biopsy.

Limitations and Challenges

Several limitations warrant consideration. This single-center study (n = 185) limits generalizability across ethnic populations and healthcare settings. The cohort comprised patients with radiologically suspicious nodules referred for invasive evaluation, creating verification bias that enriched disease prevalence (51.4% vs. 3–10% in screening populations^{[21],[22]}). This reflects the model's intended clinical context: patients with indeterminate nodules already triaged for potential invasive workup, rather than unselected screening populations.

Consequently, the model's performance metrics are optimized for this intermediate-to-high pretest probability setting and will require recalibration if applied to low-prevalence screening cohorts.

A notable limitation is the model's reduced sensitivity for pGGNs, which accounted for 50% of false negatives in the low-risk group. Importantly, LGALS3BP expression levels did not differ significantly between pGGN malignancies (median fold-change: 1.99 [IQR: 0.93–5.85]) and part-solid/solid malignancies (2.16 [0.86-4.00], $P = 0.603$), indicating that the reduced sensitivity stems not from lower biomarker secretion in pGGNs. However, the wider interquartile range in the pGGNs group suggests greater biological heterogeneity. This finding suggests that the current model framework, which integrates molecular and clinical features through logistic regression, may inadequately balance these inputs for pGGNs—where indolent imaging appearance (low solid component, smaller size) may inappropriately down-weight strong molecular signals.

Moving forward, key priorities include multi-center prospective validation in ethnically diverse cohorts, head-to-head comparison with complementary biomarker platforms, development of pGGN-specific models or multi-marker panels, technical standardization with regulatory validation, and cost-effectiveness analysis to inform clinical adoption.

Conclusions

This study identifies plasma LGALS3BP as a promising biomarker for early-stage lung adenocarcinoma and develops a clinically interpretable diagnostic model that integrates molecular and clinical features. The model demonstrates robust discriminative ability and high positive predictive value, positioning it as a decision-support tool for patients with radiologically indeterminate pulmonary nodules. Addresses a critical unmet need by providing objective risk stratification in the diagnostic "gray zone," potentially reducing unnecessary invasive procedures while maintaining high sensitivity for malignancy. However, limitations must be acknowledged. The model is not intended for population-wide screening but as a triage tool for patients with diagnostic uncertainty facing decisions about invasive workup.

Declarations

Acknowledgements

Not applicable

Author contributions

Gong Rong performed the bioinformatic analysis, screened candidate biomarkers, validated the predictive model, and wrote the manuscript. Hu Rongya was responsible for the collection of plasma samples and patient clinical information, and participated in the development of the model. Zhang Fujun collected the tissue samples used in this study and provided critical suggestions for the clinical application of the research. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics declarations

The study protocol adhered to the ethical guidelines outlined in the 1975 Declaration of Helsinki. Ethical approval was granted by the Ethics Committee of the First Affiliated Hospital of Kunming Medical University (Approval Nos. 2020L42 for the tissue proteomics component; 2022L43 for the plasma analysis component), and informed consent was obtained from all participants in the study.

Funding

This study was supported by: (1) Priority Union Foundation of Yunnan Provincial Science and Technology Department and Kunming Medical University(202101AY070001-086); (2) Yunnan Province Science and Technology Talent and Platform Program - Clinical Medicine Research Center Special Project (grant NO. 202505AJ310006).

References

1. GBD 2023 Cancer Collaborators. The global, regional, and national burden of cancer, 1990–2023, with forecasts to 2050: a systematic analysis for the Global Burden of Disease Study 2023. *Lancet* **406**, 1565–1586 (2025).
2. Bade, B. C. & Dela Cruz, C. S. Lung cancer 2020: epidemiology, etiology, and prevention. *Clin. Chest. Med.* **41**, 1–24 (2020).
3. Zhong, D. et al. Lung nodule management in low-dose CT screening for lung cancer: lessons from the NELSON Trial. *Radiology*. 313, e240535; (2024). 10.1148/radiol.240535
4. Song, W. A. et al. Utility of squamous cell carcinoma antigen, carcinoembryonic antigen, Cyfra 21 – 1 and neuron specific enolase in lung cancer diagnosis: a prospective study from China. *Chinese medical journal*. 124, 3244–3248 (2011).
5. Amin, M. B. et al. *AJCC Cancer Staging Manual* 8th edn (Springer, 2017).
6. Uhlén, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
7. Gillette, M. A. et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell* **182**, 200–225e35 (2020).
8. Tang, Z. et al. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res.* **47**, W556–W560 (2019).
9. Capone, E., Iacobelli, S. & Sala, G. Role of galectin 3 binding protein in cancer progression: a potential novel therapeutic target. *J. Transl Med.* **19**, 405 (2021).
10. Liu, Y. et al. Aberrant expression of LGALS3BP drives an unfavorable prognosis and more aggressive in HCC via regulating PI3K/AKT signaling. *Tissue cell*. **89**, 102471 (2024).
11. Läubli, H. et al. Lectin galactoside-binding soluble 3 binding protein (LGALS3BP) is a tumor-associated immunomodulatory ligand for CD33-related Siglecs. *J. Biol. Chem.* **289**, 33481–33491 (2014).
12. Stanczak, M. A. et al. Self-associated molecular patterns mediate cancer immune evasion by engaging Siglecs on T cells. *J. Clin. Investig.* **128**, 4912–4923 (2018).
13. Nigjeh, E. N. et al. Spectral library-based glycopeptide analysis-detection of circulating galectin-3 binding protein in pancreatic cancer. *Proteomics. Clin. Appl.* **11**, 9–10 (2017).

14. Song, Y. et al. Plasma exosomes from endometrial cancer patients contain LGALS3BP to promote endometrial cancer progression. *Oncogene*. 40, 633–646 (2021).
15. McWilliams, A. et al. Probability of cancer in pulmonary nodules detected on first screening CT. *N. Engl. J. Med.* **369**, 910–919 (2013).
16. Wong, J. Y. Y. et al. Lung cancer risk by geologic coal deposits: A case-control study of female never-smokers from Xuanwei and Fuyuan, China. *Int. J. Cancer*. **144**, 2918–2927 (2019).
17. Zheng, D. et al. Identification and evaluation of circulating small extracellular vesicle microRNAs as diagnostic biomarkers for patients with indeterminate pulmonary nodules. *Journal of nanobiotechnology*. 20, 172 (2022).
18. Jia, B. et al. An integrated proteomic classifier to distinguish benign from malignant pulmonary nodules. *Clin. Proteomics*. **22**, 11 (2025).
19. Shi, L., Zhao, J., Wei, Z., Wu, H. & Sheng, M. Radiomics in distinguishing between lung adenocarcinoma and lung squamous cell carcinoma: a systematic review and meta-analysis. *Front. Oncol.* **14**, 1381217 (2024).
20. Dykan, I. & Kolotilov, N. Artificial Intellect, Radiomics and Radiogenomics. *Radiation Diagnostics Radiation Therapy*. **12**, 34–44 (2021).
21. Aberle, D. R. et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 365, 395–409 (2011).
22. MacMahon, H. et al. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. *RADIOLOGY* **284**, 228–243 (2017).

Figures

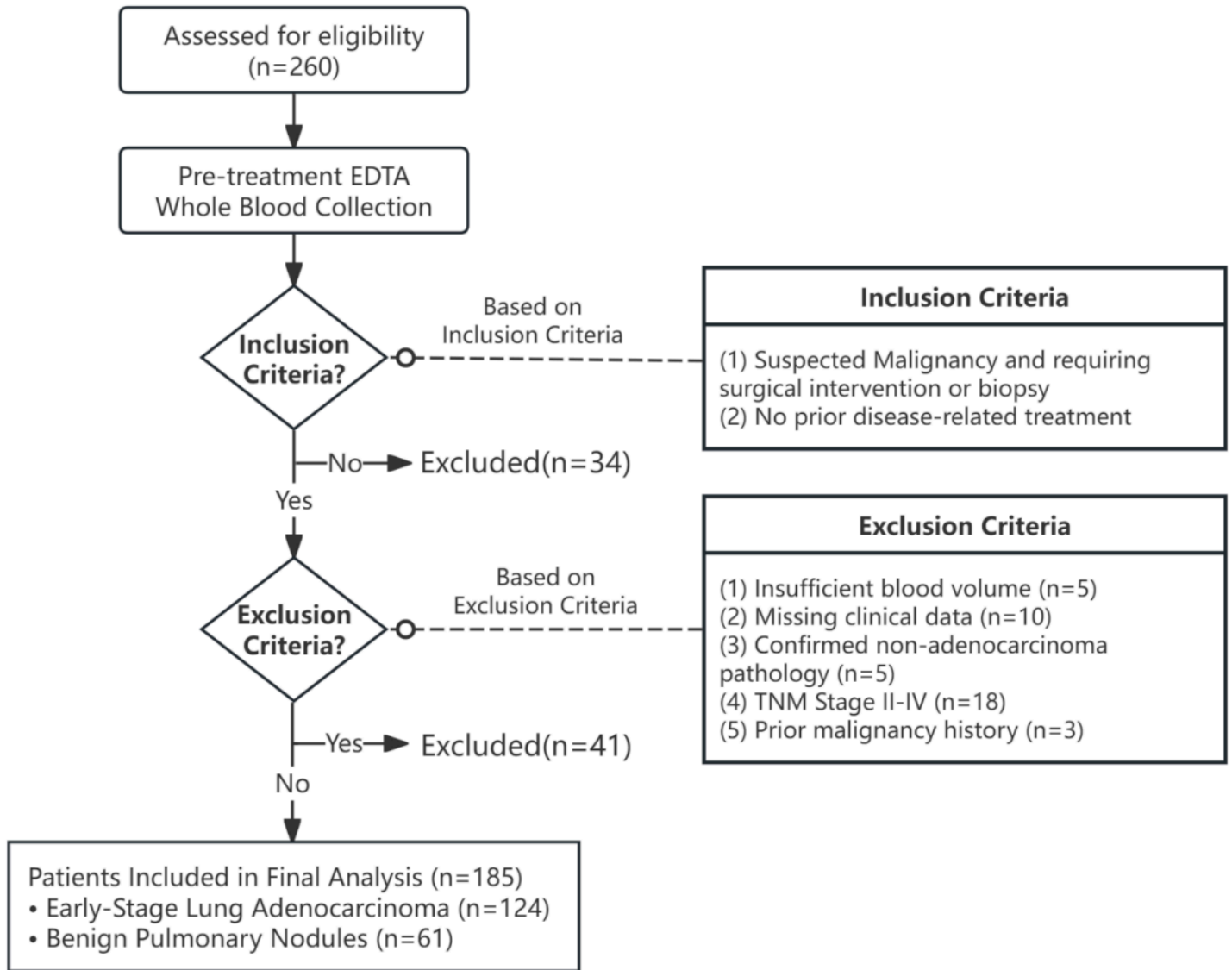


Figure 1

Flowchart of patient selection. The flow diagram illustrates the screening, inclusion, and exclusion process for the study cohort. A total of 260 patients with pulmonary nodules were assessed for eligibility between January and December 2024. After applying the inclusion and exclusion criteria, 185 patients were included in the final analysis, comprising 124 with early-stage LUAD and 61 with BPN.

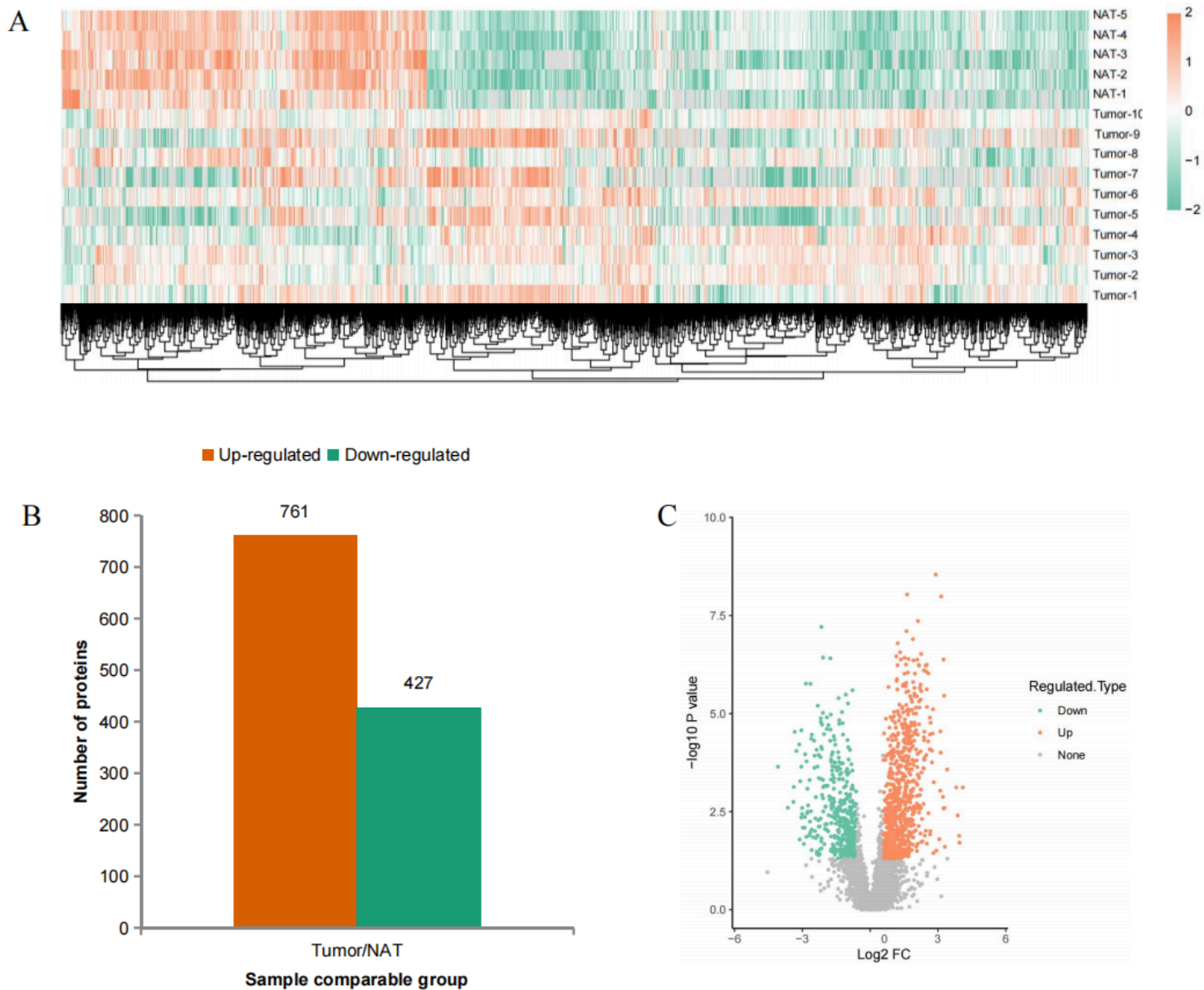


Figure 2

Tissue Proteomics Reveals Extensive Protein Dysregulation in Early-Stage LUAD. A. Hierarchical clustering heatmap of differentially expressed proteins; B. Bar chart showing number of upregulated and downregulated proteins; C. Volcano plot displaying differential protein expression

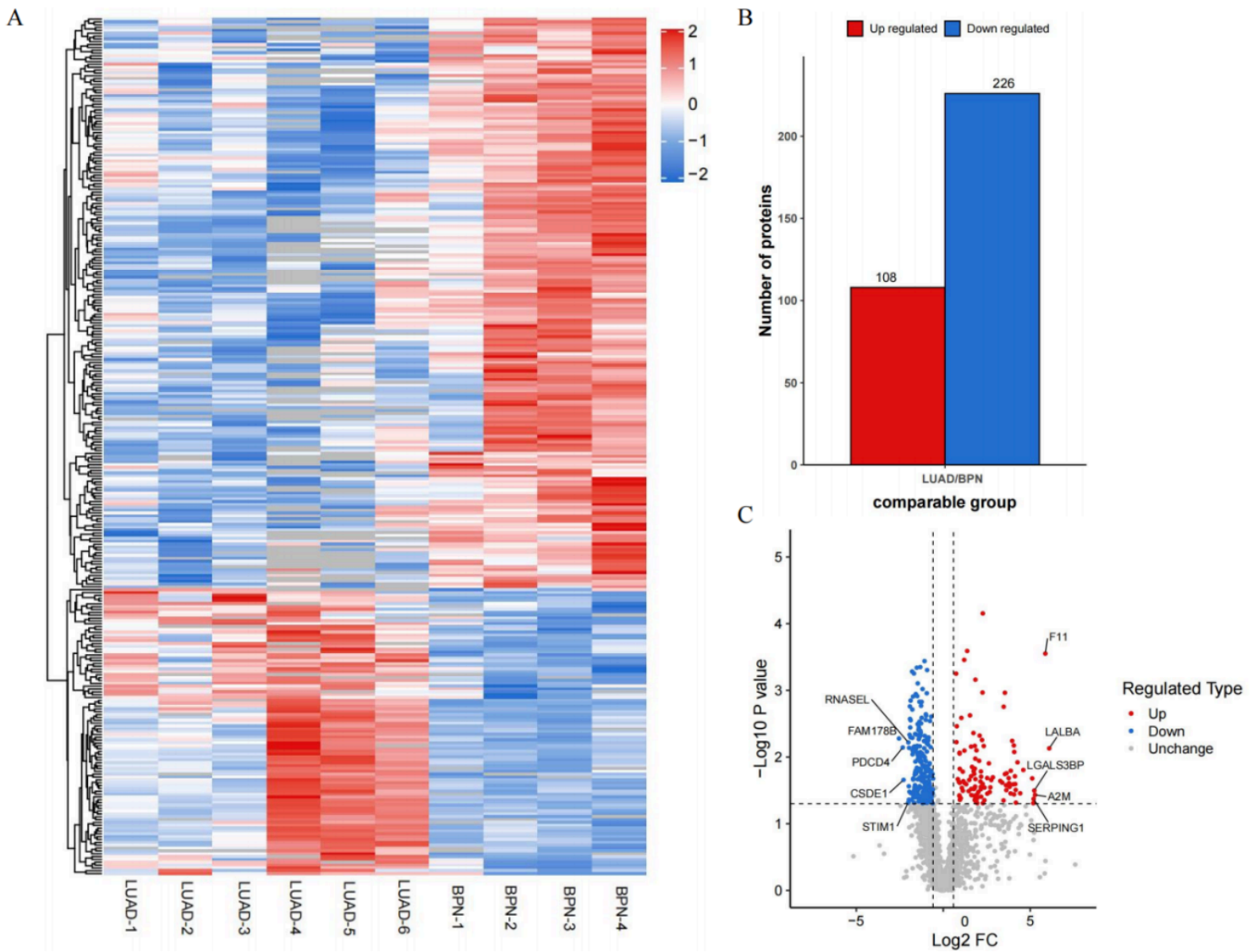


Figure 3

Plasma Proteomics Identifies Circulating Protein Alterations. A. Hierarchical clustering heatmap of plasma differentially expressed proteins; B. Bar chart showing number of upregulated and downregulated plasma proteins; C. Volcano plot of plasma differential protein expression

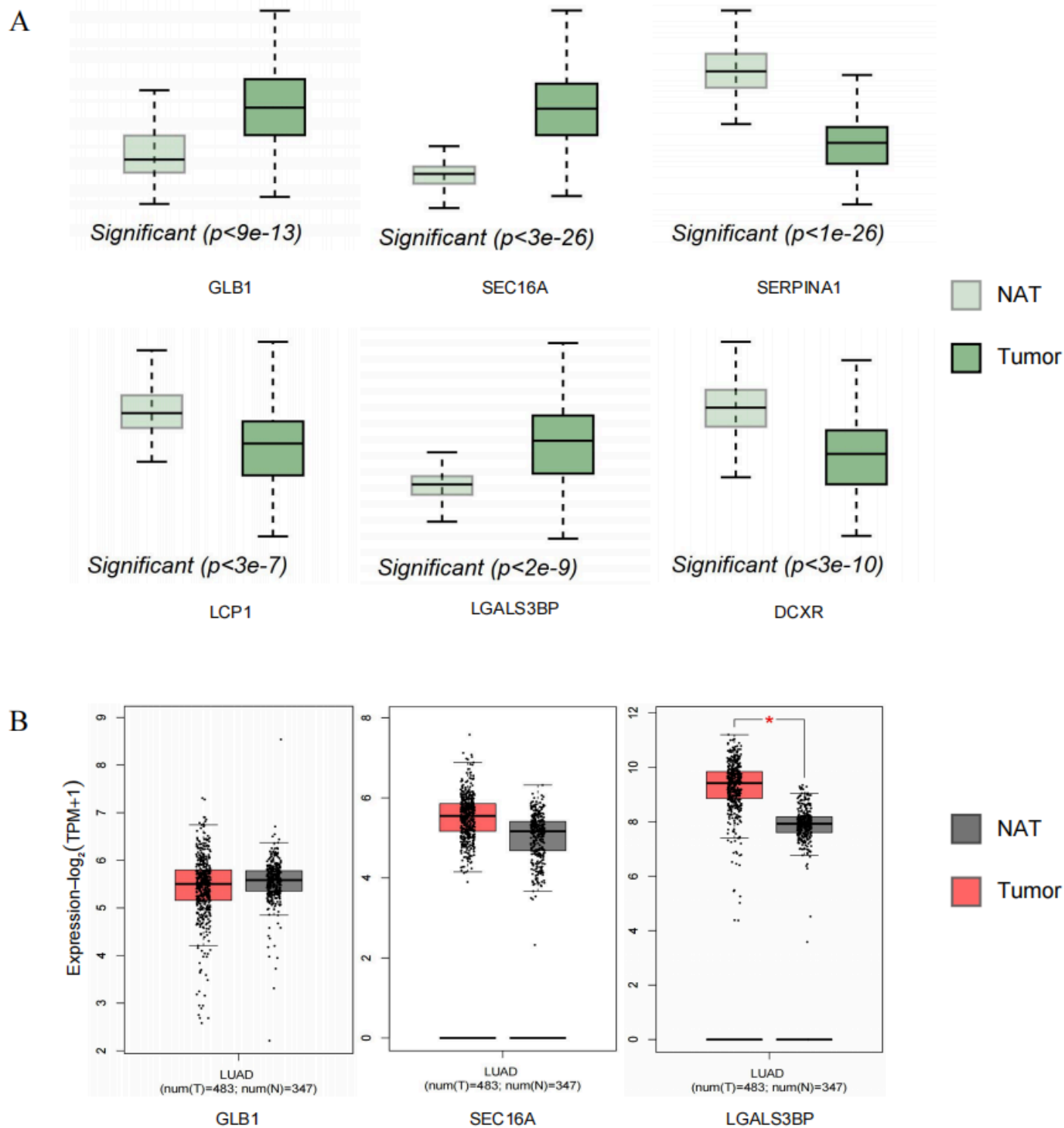


Figure 4

Validation of candidate biomarkers using public databases. A. Protein expression levels of candidate biomarkers in tumors versus NATs based on the CPTAC dataset; B. mRNA expression levels of *GLB1*, *SEC16A*, and *LGALS3BP* in LUAD tissues compared to NATs based on TCGA data. Red asterisks indicate statistical significance ($P < 0.01$). Data sources: The Human Protein Atlas (for A) and GEPIA2 (for B).

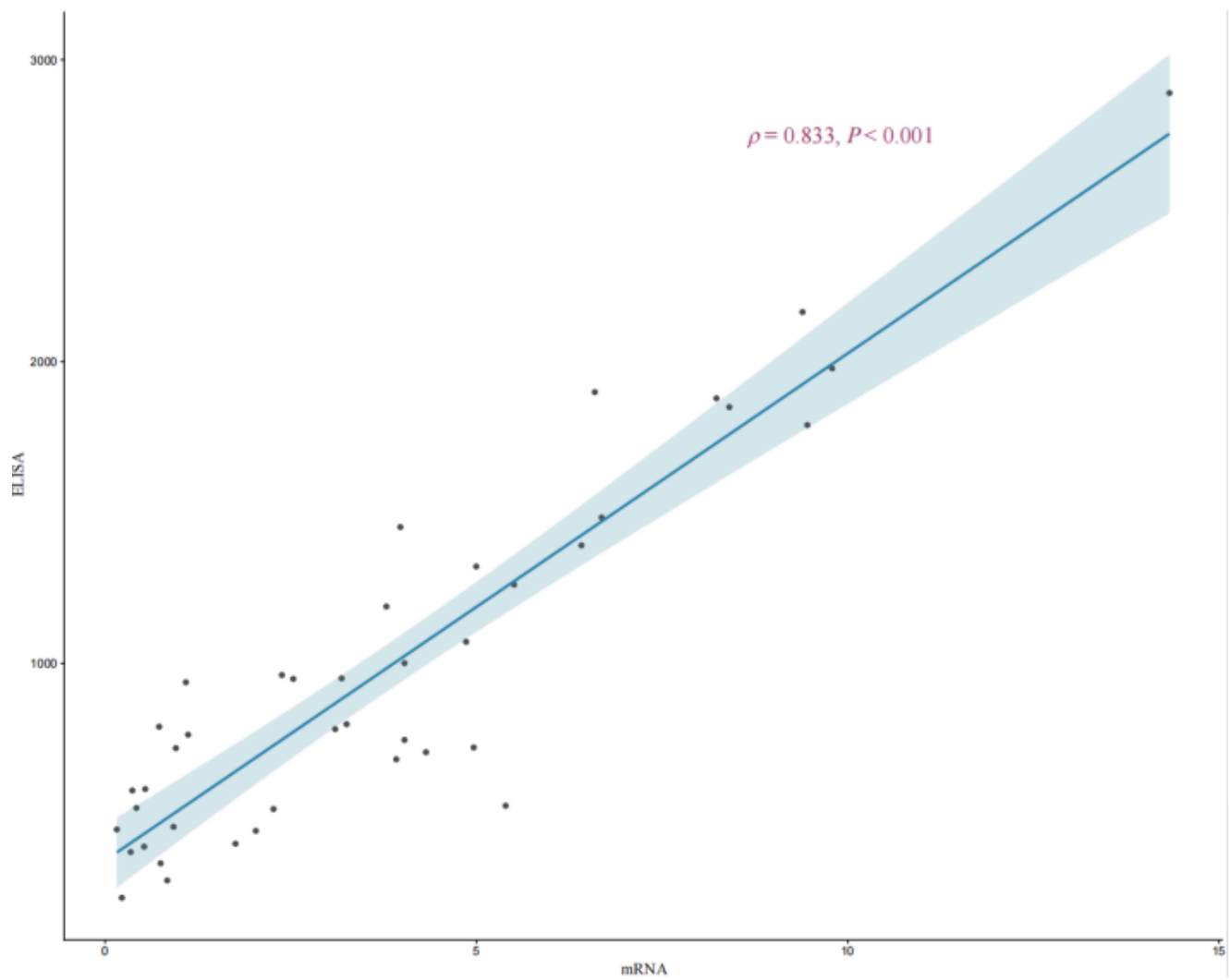


Figure 5

Spearman's rank correlation analysis between plasma LGALS3BP mRNA levels and protein concentrations

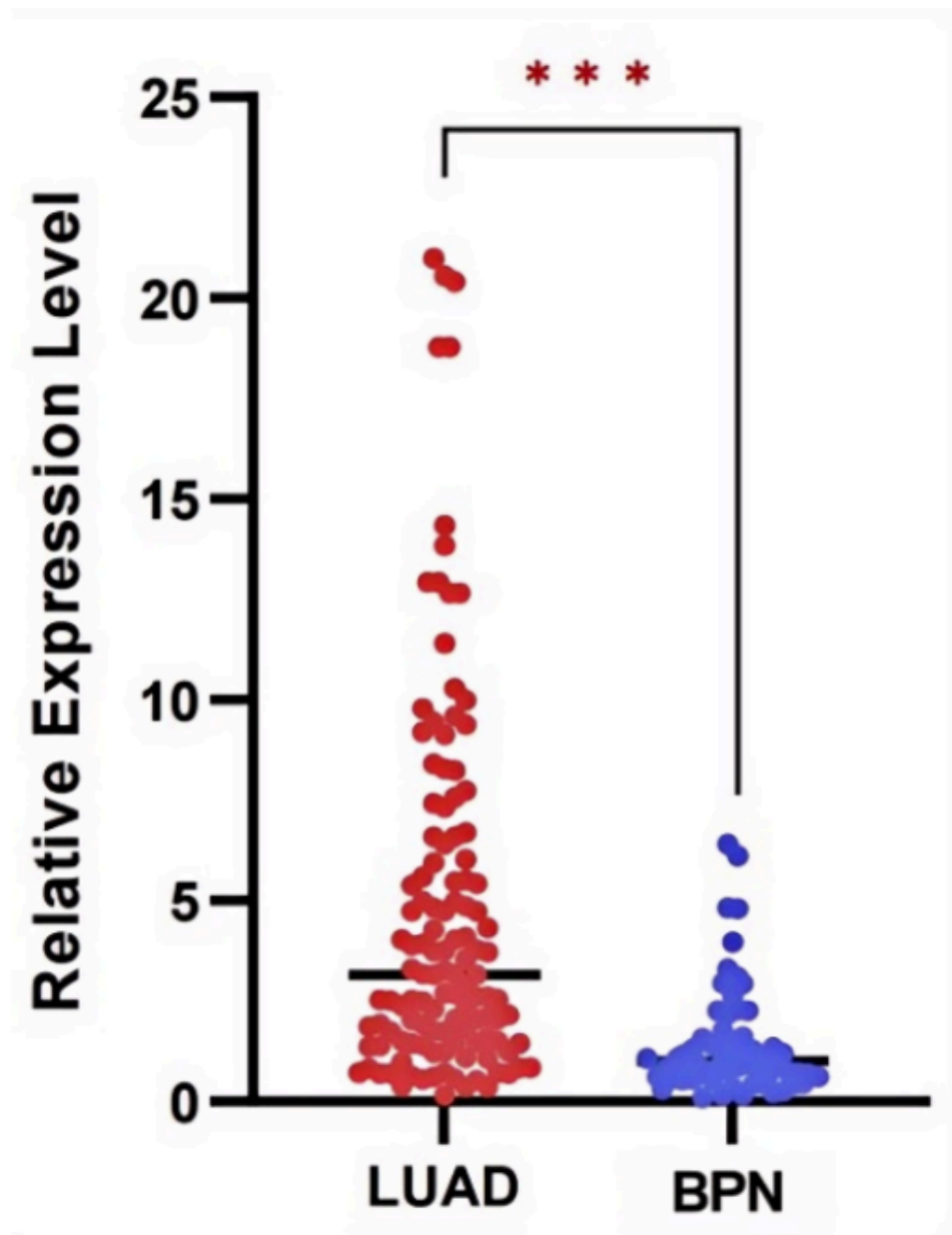


Figure 6

plasma *LGALS3BP* mRNA in complete cohort

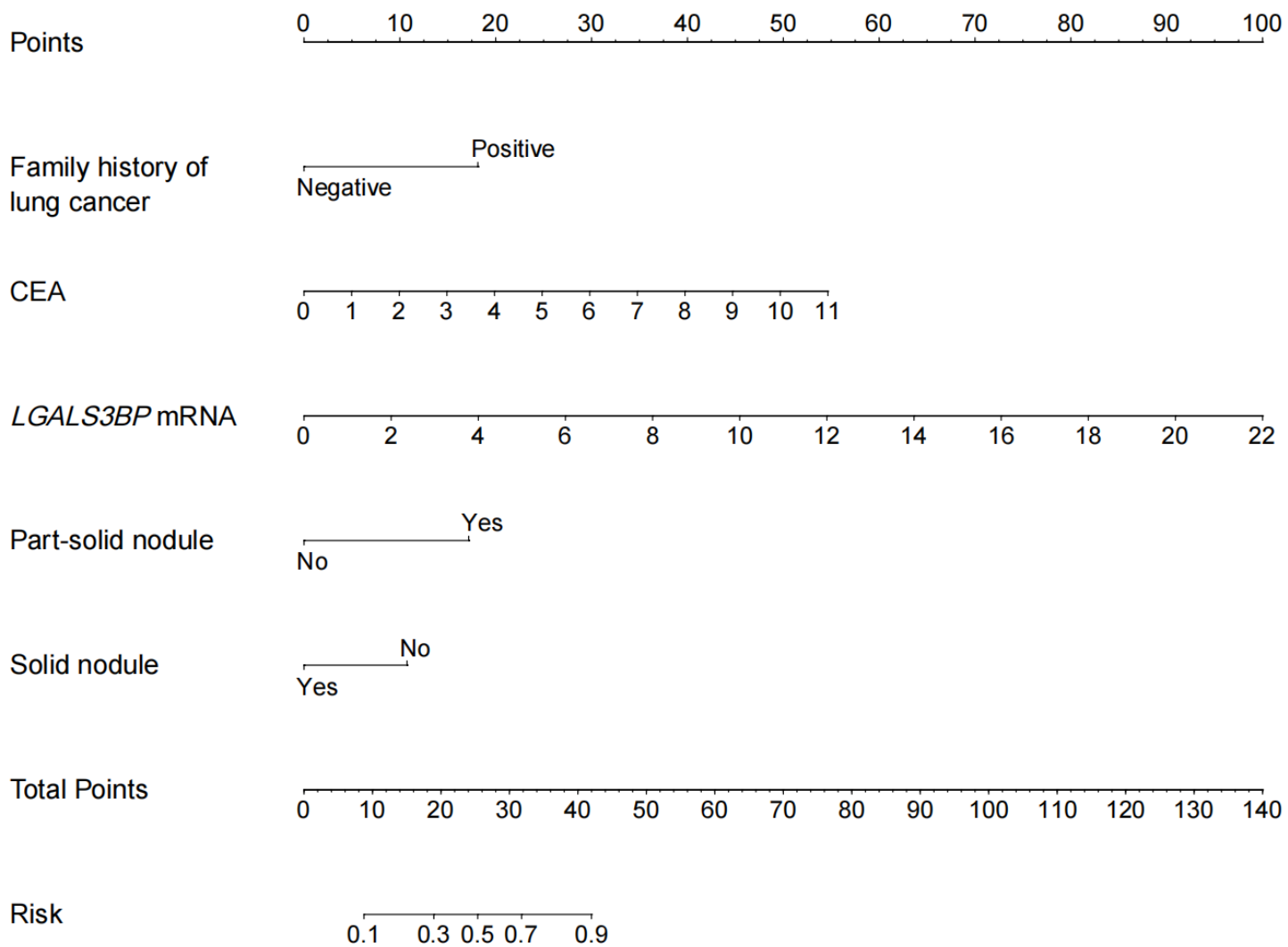


Figure 7

Nomogram incorporating *LGALS3BP* mRNA expression for individualized prediction of early-stage LUAD Diagnosis.

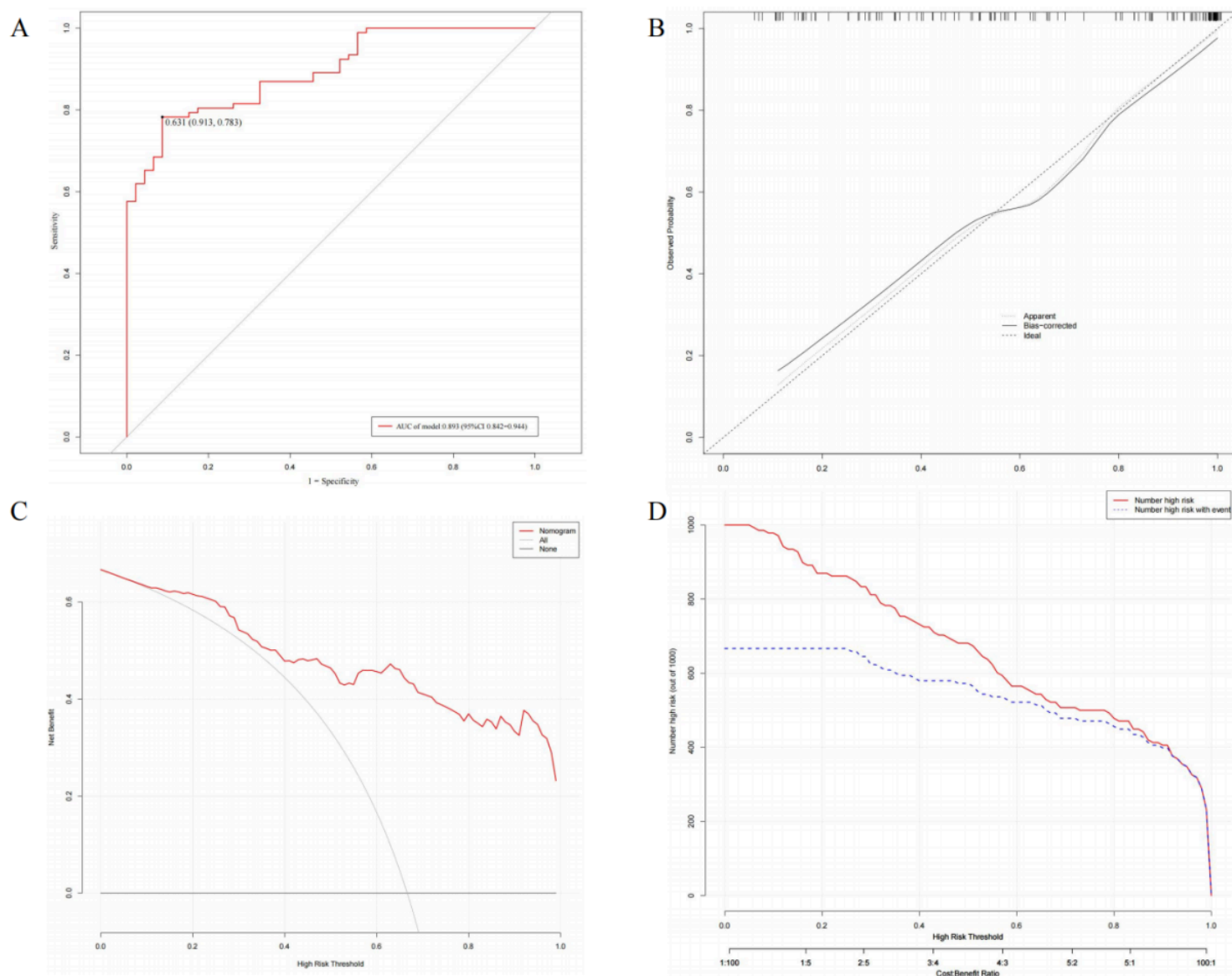


Figure 8

Combined prediction model performance in training cohort. The comprehensive performance evaluation included discrimination, calibration, and clinical utility assessments in the training cohort. A. ROC curve analysis demonstrating the superior diagnostic accuracy of the combined model. B. Calibration plot showing the agreement between predicted probabilities and observed outcomes. C. DCA illustrating the net benefit of the model across different threshold probabilities. D. CIC visualizing the number of true positive cases identified at various risk thresholds.

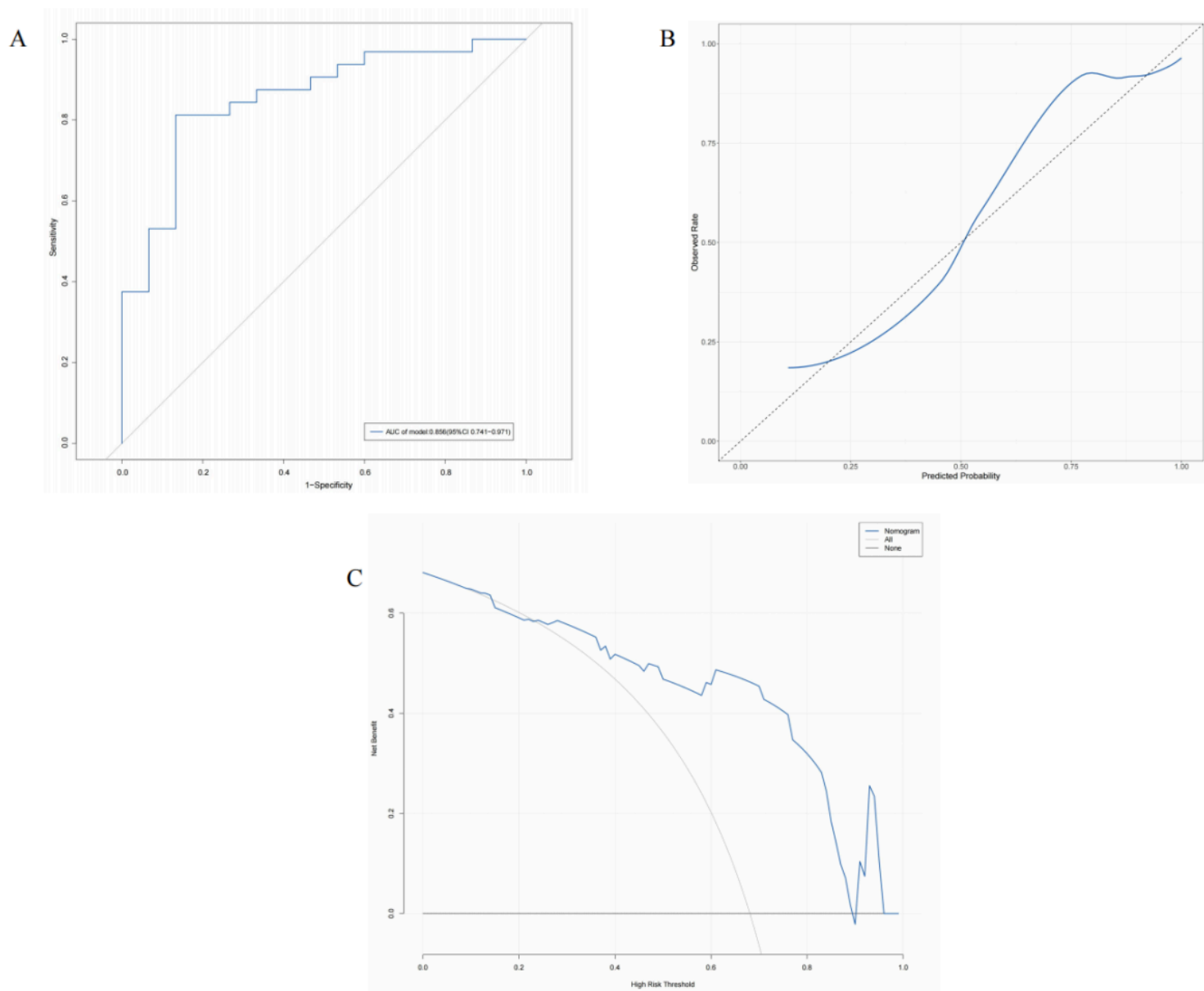


Figure 9

Combined prediction model performance in validation cohort.

Supplementary Files

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

- [SupplementaryMaterials.docx](#)