BMC Cancer

RESEARCH

Open Access

Association of pre-diagnosis plasma proteomic contexture with overall survival of early- and late-stage colon cancer patients

Shun Li[1,2], Hao Wang[1,2], Xiao-Qian Xu[1,2], Wei-Ming Li[1,2], Hong You[3], Ji-Dong Jia[3], You-Wen He[4] and Yuan-Yuan Kong[1,2*]

Abstract

Background The pre-diagnosis plasma proteomic contexture of colon cancer patients may reflect host immune and biological conditions and potentially associate with survival outcomes. We aimed to characterize pre-diagnosis proteomic contextures in colon cancer patients and determine potential association with overall survival of the patients.

Methods Baseline plasma samples collected at an average of 7.90 years before diagnosis from colon cancer patients in the UK Biobank cohort were analyzed using Olink proteomics technology. Cox-regression analysis was applied to identify distinct pre-diagnosis proteomic contextures and determine their association with survival outcomes.

Results In early-stage colon cancer, a 10-protein pre-diagnosis profile was identified, involving biological processes of extracellular matrix remodeling and immune evasion through deregulation of innate immune activation. Increased activity in these pathways before diagnosis was associated with poor survival outcomes. In late-stage cases, an 8-protein pre-diagnosis profile was linked to pathways involving in cell adhesion, angiogenesis, and pro-inflammatory response. Similarly, heightened activity in these pathways prior to diagnosis correlated with worse survival. When combined with two demographic factors age and sex, these proteomic profiles demonstrated strong predictive associations with survival outcomes at multiple time points post-diagnosis. The area under the receiver operating characteristic curve values were 0.85, 0.82, and 0.89 for early-stage cancer at 1, 5, and 10 years, respectively, and 0.71, 0.72, and 0.79 for late-stage cancer over the same periods.

Conclusions Biological processes like extracellular matrix remodeling and pro-inflammatory response are active well before diagnosis and may play a critical role in shaping colon cancer progression.

Keywords Pre-diagnosis proteomic contexture, Overall survival, Colon cancer, Association, Super and poor survivors

*Correspondence:
Yuan-Yuan Kong
kongyy@ccmu.edu.cn
Full list of author information is available at the end of the article

Introduction

Colon cancer typically progresses over a lengthy period, evolving from benign polyps through a series of genetic and epigenetic alterations that span 10 to 15 years [1]. This extended latency offers an opportunity for baseline, pre-diagnosis proteomic markers to reveal biological changes that may be associated with cancer progression and survival outcomes long before clinical diagnosis [2]. UK Biobank (UKBB), with its extensive baseline plasma samples (53,014 samples for nearly 3000 proteins) and long-term follow-up, provides an invaluable resource for exploring these pre-diagnosis proteomic signatures [3]. The UKBB's large-scale, community-based, longitudinal dataset allows for robust analyses across diverse outcomes, strengthening the predictive power of potential biomarkers for broader applications.

The host's biological predispositions, as reflected by its plasma proteomic profile, may play a pivotal role in shaping the trajectory of colon cancer development long before clinical diagnosis [4]. To capture the multifaceted biological processes that occur long before clinical detection, we introduce the concept of an individual's pre-diagnosis proteomic contexture. This term refers to a complex network of plasma proteins involved in critical biological pathways, offering a snapshot of the host's immune status, physiological conditions, and overall biological resilience or vulnerability. We hypothesize that these proteomic contextures likely experience significant shifts during the pre-diagnostic phase of colon cancer, reflecting the individual's capacity to resist or succumb to disease progression. As such, they serve as a unique lens through which to examine the biological 'fate' of patients, closely tied to cancer progression and long-term survival outcomes.

Despite advances in identifying pre-diagnosis biomarkers linked to colon cancer risk [5, 6], few studies have examined the association of pre-diagnosis proteomic contextures with long-term survival outcomes [7]. In this study, we explored these associations by first examining two contrasting survival outcomes—"super survivors" who outlive typical prognostic expectations, and "poor survivors" who experience rapid disease progression despite comparable staging at diagnosis. This approach leverages these survival extremes to provide an opportunity to uncover unique pre-diagnosis proteomic contextures that may reflect underlying resilience or vulnerability of host immunity to aggressive cancer biology. By taking advantage of the extensive baseline pre-diagnosis plasma proteomic data from the UKBB cohort at recruitment, we aim to characterize pre-diagnosis proteomic contextures associated with survival in both early- and late-stage colon cancer.

Methods

This study utilized data from the UK Biobank, a large-scale biomedical resource that supports research by linking comprehensive baseline health data with long-term health outcomes across a diverse population cohort. Ethical approval for the study was granted by the North West Multi-Centre Research Ethics Committee under UK Biobank application number 129053.

This study, based on data from the UK Biobank, leveraged integrated health records and proteomic data to investigate pre-diagnosis protein contextures associated with long-term survival in colon cancer patients. We selected participants diagnosed with colon cancer after recruitment as our study cohort and examined proteomic differences between two distinct survival groups: super survivors and poor survivors, in both early and late stages. Differentially expressed proteins (DEPs) were identified and subjected to enrichment analysis, with pre-diagnosis protein signatures constructed to predict survival outcomes in both stages. The model's performance was then tested in the entire cohort of colon cancer patients. Subgroup analyses were performed to assess the model's effectiveness by sex and age, and sensitivity analyses explored potential variations in model performance based on tumor location and metastasis sites.

Definitions of colon cancer and metastasis

Colon cancer and metastasis status were identified in the UK Biobank cohort using the International Classification of Diseases, 10 th revision (ICD-10) codes from hospital records:

- Colon Cancer: Defined by ICD-10 code C18, including subcategories C18.0-C18.9, covering malignant neoplasms of the colon.
- Metastasis: Defined by codes C77, C78, and C79 for secondary malignant neoplasms (of lymph nodes, of respiratory and digestive organs, of other and unspecified sites, respectively). Specifically, C78.0 refers to secondary malignant neoplasm of the lung, C78.6 refers to secondary malignant neoplasm of retroperitoneum and peritoneum, and C78.7 refers to secondary malignant neoplasm of the liver and intrahepatic bile duct.
- Tumor location: Determined based on ICD-10 codes: right-sided tumors (C18.0–C18.4) and left-sided tumors (C18.5–C18.7).

Cohort selection

The cohort of colon cancer patients were identified from the UK Biobank cohort using ICD-10 codes.

- Inclusion criteria: Confirmed diagnosis of colon cancer, defined by at least one documented ICD-10 code for colon cancer (C18, including subcategories C18.0-C18.9).
- Exclusion criteria:

  a) A history of colon cancer prior to recruitment or diagnosis within half-year after recruitment.
  b) Absence of plasma proteomics data collected at the time of recruitment.

Classification of super and poor survivors

To capture meaningful differences in survival-related proteomic expression, we used a data-driven approach to classify patients into super and poor survivor groups, reflecting the natural distribution of survival times within the cohort. Rather than applying arbitrary whole-number thresholds, we derived the cut-off points based on observed survival distributions (from diagnosis to all-cause mortality), ensuring biologically and clinically meaningful groupings (Fig. 1B). These classifications focus on survival extremes to identify distinct proteomic patterns potentially that may influence survival trajectories.

- Super survivors were defined as patients whose survival exceeded the upper two-thirds percentile of the cohort's survival distribution (> 6.08 years post-diagnosis, Fig. 1B). By deriving cut-offs from the cohort, this data-driven approach avoids variations in survival times across different populations, ensuring that the classification accurately reflects meaningful survival differences within the studied group. Although 6.1% of these patients eventually succumbed to the disease, their prolonged survival warranted this classification, as it may offer insights into biological factors contributing to survival.
- Poor survivors were defined as patients with survival time below the median of all deceased patients within the cohort (< 1.17 years post-diagnosis, Fig. 1B). This approach captures individuals with markedly limited survival, enabling the identification of proteomic signatures associated with aggressive disease and poor prognosis.

This approach allowed for a detailed comparison of proteomic profiles associated with both extended and limited survival outcomes in colon cancer, facilitating insights into potential biological factors linked to survival outcomes.

Proteomic data collection

Plasma proteomic data were obtained from the UK Biobank's April 2023 release, encompassing approximately 55,000 participants, and generated through the Pharma Proteomics Project [8]. Multiplexed proteomic assays were performed on baseline plasma samples using the Olink Explore 3072 platform, which utilizes dual barcoded antibody technology for semi-quantitative readouts of 2,923 proteins across 8 panels. After quality control measures, 2,699 proteins with less than 30% missing values in the cohort were retained for analysis. The proteomic data in the UK Biobank was pre-normalized using inverse-rank normalized approaches, ensuring comparability across samples.

Statistical analysis

Continuous variables were summarized as means with standard deviations (SD) for normally distributed data or medians with interquartile ranges (IQR) for data with non-normal distribution. Group comparisons were conducted using the Student's t-test for normally distributed variables and the Mann–Whitney U-test for non-normally distributed data. Categorical variables were expressed as frequencies and percentages and were compared using Chi-square or Fisher's exact test, as appropriate.

DEPs were identified using unadjusted Cox regression models to account for survival time and normalized proteomic datasets. A DEP was defined as a protein with a log2 fold change outside the range of − 1 to 1 and an unadjusted p-value < 0.05.

To identify biological pathways associated with survival, we performed Gene Ontology (GO) enrichment analysis for metastatic and non-metastatic colon cancer groups. Using the "clusterProfiler" package in R, proteins were ranked by fold-change in expression between super and poor survivors. GO terms were derived from annotated biological processes in reference gene sets, facilitating hierarchical classification of gene functions and pathway insights relevant to survival.

The time from sample collection to diagnosis was calculated for each individual as the time difference between the date of baseline plasma sample collection and the date of colon cancer diagnosis. For survival analysis, univariable and multivariable Cox proportional hazards regression models were used to assess the significance of survival differences between super and poor survivors, and the predictive value of selected protein biomarkers alongside clinical covariates, respectively. Hazard ratios

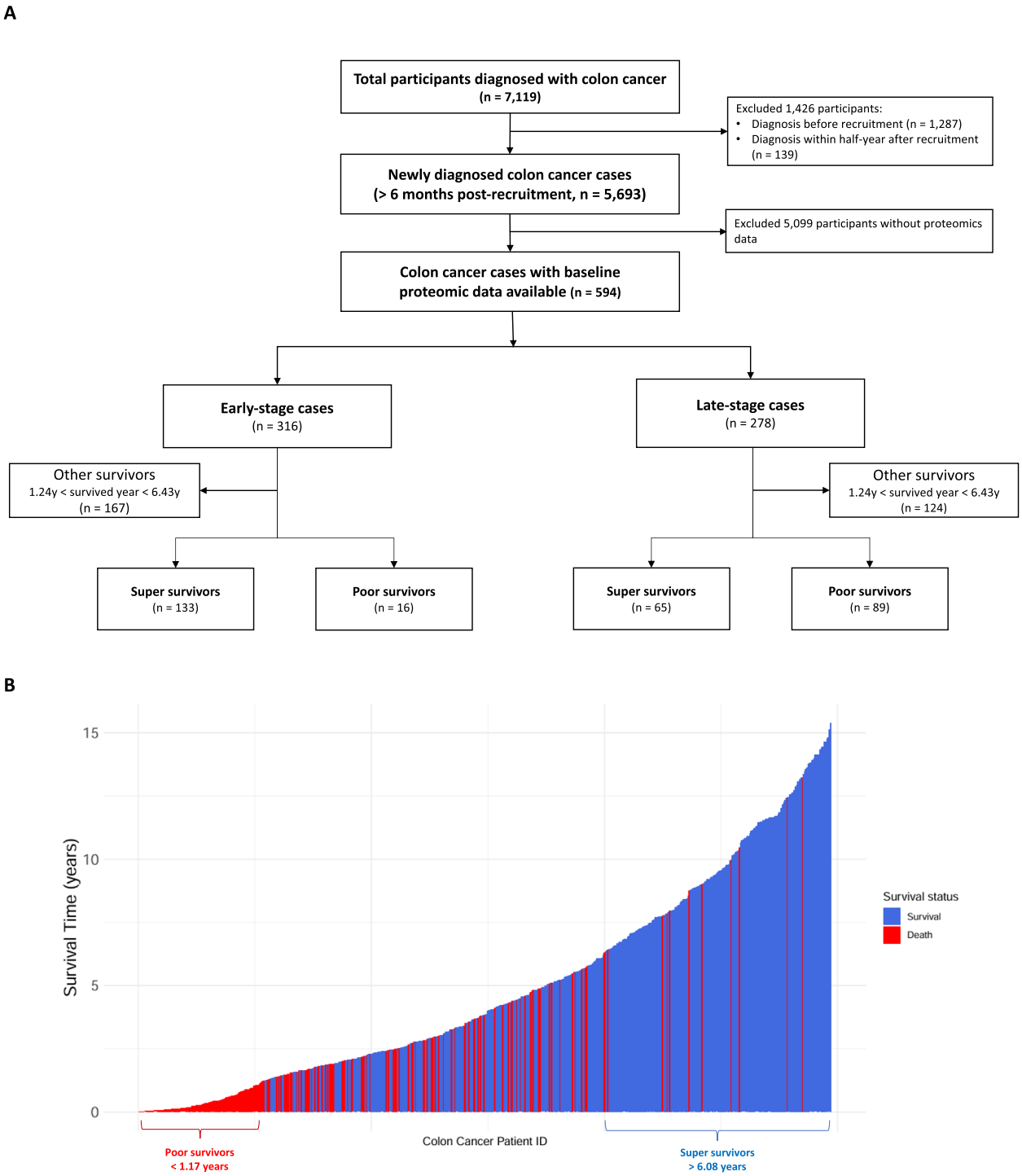

Fig. 1 **A** Flow chart of the study. **B**. Survival time distribution from colon cancer diagnosis to death or last follow-up. Each vertical bar represents an individual diagnosed with colon cancer, illustrating the time from diagnosis to either death (red) or the last follow-up point (blue). Blue bars indicate participants who survived until the study's end, while red bars denote those who succumbed to the disease during the follow-up period. Poor survivors referred to the patients who survived less than the median survival time of deceased patients (1.17 years), while super survivors described patients who lived beyond two-thirds of the observed survival time (6.08 years)

(HRs) with 95% confidence intervals (CIs) were calculated to quantify the impact of each biomarker on survival risk.

To handle missing values in the proteomic dataset, predictive mean matching was used for data imputation. Lasso regression was applied to select significant proteins for inclusion in predictive models. Time-dependent receiver operating characteristic (ROC) curves were calculated using the timeROC package to evaluate the predictive performance of proteomic contextures at 1-year (short-term), 5-year (medium-term), and 10-year (long-term) survival intervals across super survivors, poor survivors, and the broader colon cancer cohort. Model accuracy was assessed by the area under the ROC curve (AUROC) with 95% bootstrap confidence intervals. The optimal cut-off for proteomic contextures was determined by Youden Index using SurvivalROC package.

Subgroup analyses were conducted to assess model performance stratified by sex and age at diagnosis. Age at diagnosis was categorized using a cut-off of 70 years as the median diagnosis age for colon cancer [9]. Sensitivity analysis was conducted for left-sided only and right-sided only colon cancer cases, as well as for three common metastasis sites: liver, lung, and retroperitoneum/peritoneum.

All analyses were performed in R version 4.2.3 (R Foundation for Statistical Computing), and statistical significance was defined as a two-tailed P-value < 0.05.

Results

A total of 7,119 colon cancer cases were identified in the cohort using ICD-10 codes, from which 5,693 new diagnoses were included after excluding 1,426 cases for pre-existing or near-term (diagnosis within half-year after recruitment) diagnoses. Of the 5,693 patients, 594 had pre-diagnosis plasma proteomic data available, analyzed using the Olink platform (Fig. 1A). These samples were collected at recruitment, with a mean time of 7.90 years from sample collection to diagnosis. The cohort was divided by metastasis status and ICD-10 stage into two groups: early-stage (stage I-II) cases without any metastasis (n = 278) and late-stage (stage III-IV) cases with metastasis (n = 316). Early-stage cases were diagnosed at an older average age (mean age at diagnosis: 70.34 years) compared to late-stage cases (mean age: 67.90 years, Table 1). Additionally, early-stage cases exhibited longer survival durations from diagnosis to all-cause mortality (mean survival: 5.99 years for early-stage vs. 3.74 years for late-stage). To identify distinct contexture associating with overall survival, we first selected patients in each group by survival duration into "poor" (surviving < 1.17 years) and "super" survivors (surviving > 6.08 years; Fig. 1B).

Baseline characteristics and survival patterns

Early- and late-stage groups differed markedly in demographic and survival patterns. In the early-stage cohort, 133 super survivors and 16 poor survivors were identified (Table 2). The mean ages at recruitment for super and poor survivors were 60.97 and 64.25 years, respectively (P = 0.06). Super survivors had a younger mean age at diagnosis of 66.56 years, compared with the 74.81 years observed in poor survivors (P < 0.001). As expected, early-stage super survivors exhibited a significantly longer survival duration post-diagnosis (9.89 years) than poor survivors (0.44 years). Given similar ages at recruitment, super survivors in early-stage cancer had a shorter interval from baseline to diagnosis (5.04 years) than poor survivors (10.09 years).

Among late-stage cases, 65 super survivors and 89 poor survivors were identified, with similar mean ages at recruitment (59.57 years for super survivors and 60.84 years for poor survivors, P = 0.16, Table 2). While super survivors were again younger at diagnosis (64.94 years) than poor survivors (69.01 years), the age gap was narrower than in the early-stage group. Similarly, although the interval from baseline to diagnosis was shorter for super survivors (4.88 years) than for poor survivors (7.67 years), the difference was less pronounced than that observed in early-stage cases (Table 2). Compared with poor survivors, super survivors who lived more than 10 years were diagnosed at a significantly younger age (mean 63.11 vs. 69.01 years) and more frequently exhibited lymph node-only metastases (50.0% vs. 6.7%, Supplementary Materials: Table S1). In contrast, liver metastases were significantly more prevalent among poor survivors (70.8% vs. 32.1%, P = 0.001, Supplementary Materials: Table S1).

In both early- and late-stage groups, other factors, including body mass index (BMI), alcohol consumption, and smoking status, showed no significant differences between super and poor survivors (Table 2). Together, these findings highlight potential age- and stage-dependent influences on survival, with diagnostic timing appearing particularly relevant in early-stage cases.

Pre-diagnosis proteomic expression differences between super and poor survivors

Proteomic analysis revealed distinct pre-diagnosis expression contextures associated with survival outcomes across early and late stages. In early-stage colon cancer, 143 pre-diagnosis DEPs were identified between super and poor survivors, with 120 DEPs upregulated and 23 downregulated in poor survivors prior to diagnosis (Fig. 2A, Supplementary materials: Table S2, Fig.S1). The highest fold changes were noted in insulin receptor (INSR, 9.70 log2 fold change) and interleukin- 20

Table 1 Baseline characteristics of early- and late-stage colon cancer cohorts in the UK Biobank

Baseline characteristic[a]	Early stage (316)	Late stage (278)	*P* value
Age at recruitment (year)	61.27 (6.41)	60.26 (6.78)	0.06
Age at diagnosis (year)	70.34 (7.40)	67.90 (7.32)	< 0.001
Sex			
Male	183 (57.9)	145 (52.2)	0.18
Female	133 (42.1)	133 (47.8)	
Ethnic (White %)	304 (96.5)	263 (96.0)	0.91
BMI	27.98 (4.77)	27.87 (4.61)	0.77
Year from baseline to diagnosis	8.56 (3.79)	7.15 (3.69)	< 0.001
Year from diagnosis to death	5.99 (3.89)	3.74 (3.85)	< 0.001
Alcohol drinker status			
Never	19 (6.0)	11 (4.0)	0.54
Previous	13 (4.1)	12 (4.3)	
Current	284 (89.9)	252 (90.6)	
Smoking			
Never	146 (46.2)	149 (53.6)	0.12
Previous	139 (44.0)	99 (35.6)	
Current	31 (9.8)	26 (9.4)	
Survival status			
Alive	278 (88.0)	106 (38.1)	< 0.001
Death	38 (12.0)	172 (61.9)	
Stage[b]			
III	-	63 (22.7)	-
IV	-	215 (77.3)	
Metastasis type			
Lymph node metastasis only	-	63 (22.7)	-
Distant metastasis only	-	120 (43.2)	
Both types of metastases	-	95 (34.2)	
Metastasis site			
Liver and intrahepatic bile duct		153 (55.0)	-
Lung		83 (29.8)	
Retroperitoneum and peritoneum		80 (28.8)	

[a] N (%) or Mean (SD). *BMI* body-mass index

[b] Stage information is not available in the UK Biobank, only lymph node metastasis was used as proxy for stage III, and distant metastasis was used as proxy for stage IV

receptor A (IL20RA, − 4.21 log2 fold change), suggesting a pre-diagnosis proteomic contexture that may drive aggressive cancer progression in early stages.

In late-stage colon cancer, only 80 pre-diagnosis DEPs were identified, with 75 upregulated and five downregulated in poor survivors before diagnosis (Fig. 2B, Supplementary materials: Table S3, Fig.S1). Although fewer DEPs were detected, the pre-diagnosis proteomic shifts in late-stage cancer reflected survival-linked processes. The most upregulated protein, A1BG (1.96 log2 fold change), and downregulated protein, ITGB6 (−2.05 log2 fold change) in poor survivors, also suggest a distinct pre-diagnosis protein contexture associated with late-stage disease progression. This stage-specific variation in DEPs highlights how the pre-diagnosis proteomic contexture shifts as the disease advances and survival dynamics evolve.

GO enrichment and biological mechanisms

GO enrichment analysis further elucidated survival-associated pathways in each stage. In early-stage cases, significantly enriched terms related to extracellular matrix (ECM) integrity and cell signaling were prominent (the most enriched terms *collagen-containing extracellular matrix*, Fig. 3A). The elevated pre-diagnosis expression of ECM-related proteins, such as ADAMTS1, ADAMTS4, DDR1, and ACTA2, in poor survivors indicates that ECM remodeling processes

Table 2 Baseline characteristics of super and poor survivors of early- and late-stage colon cancer in the UK Biobank cohort

Baseline characteristic[a]	Early stage			Late stage		
	Super survivors (133)	Poor survivors (16)	*P* value	Super survivors (65)	Poor survivors (89)	*P* value
Age at recruitment (year)	60.97 (6.63)	64.25 (5.21)	0.06	59.57 (6.58)	60.84 (6.21)	0.22
Age at diagnosis (year)	66.56 (7.22)	74.81 (5.68)	< 0.001	64.94 (6.56)	69.01 (7.32)	< 0.001
Sex						
Male	73 (54.9)	11 (68.8)	0.43	39 (60.0)	39 (43.8)	0.07
Female	60 (45.1)	5 (31.2)		26 (40.0)	50 (56.2)	
Ethnic (White %)	125 (94.7)	16 (100.0)	0.75	61 (95.3)	85 (95.5)	1.00
BMI	27.65 (4.26)	30.01 (6.22)	0.06	27.30 (4.60)	28.37 (4.96)	0.18
Year from baseline to diagnosis	5.04 (2.51)	10.09 (3.29)	< 0.001	4.88 (2.46)	7.67 (3.60)	< 0.001
Year from diagnosis to death	9.89 (2.45)	0.44 (0.36)	< 0.001	9.77 (2.61)	0.37 (0.32)	< 0.001
Alcohol drinker status						
Never	11 (8.3)	2 (12.5)	0.75	1 (1.5)	5 (5.6)	0.40
Previous	5 (3.8)	1 (6.2)		5 (7.7)	5 (5.6)	
Current	117 (88.0)	13 (81.2)		58 (89.2)	79 (88.8)	
Smoking						
Never	65 (48.9)	9 (56.2)	0.86	35 (53.8)	53 (59.6)	0.74
Previous	58 (43.6)	6 (37.5)		24 (36.9)	28 (31.5)	
Current	10 (7.5)	1 (6.2)		5 (7.7)	8 (9.0)	
Survival status						
Alive	128 (96.2)	0 (0.0)	< 0.001	58 (89.2)	0 (0.0)	< 0.001
Death	5 (3.8)	16 (100.0)		7 (10.8)	89 (100.0)	

[a] N (%) or Mean (SD). *BMI* body-mass index

were actively underway even before the clinical onset of cancer (Fig. 4A). Moreover, markers of immune evasion through deregulation of innate immune activation (IL15), metabolic dysregulation (INSR, APOH), and genomic instability (TP53) were upregulated in poor survivors long prior to disease diagnosis (Fig. 4A), linking early ECM alterations, immune evasion, metabolic shifts, and genomic instability to survival outcomes.

In contrast, late-stage cancer GO enrichment underscored processes associated with cell adhesion and active inflammatory response initiated long before diagnosis (Fig. 3B). Elevated expression of EDIL3 indicated an active angiogenesis in poor survivors prior to diagnosis, while higher expression of TNFSF8 and PPP3R1 in poor survivors suggested pre-diagnosis pro-inflammatory state (Fig. 4B). These findings suggest that survival differences in both stages were driven by pre-diagnosis active ECM remodeling or tumor cell adhesion. While early-stage survival differences may be additionally driven by metabolic factors and genomic instability, survival in late-stage disease may further hinge on angiogenesis and inflammation.

Pre-diagnosis proteomic contexture associated with survival outcomes in all patients

Building on our proteomic findings, we identified pre-diagnosis proteomic contextures that were strongly associated with survival outcomes in early- and late-stage colon cancer, reflecting distinct biological processes active before clinical diagnosis. For early-stage colon cancer, among 12 DEPs identified within the top five GO terms, a 10 pre-diagnosis proteomic contexture (ACTA2, ADAMTS1, ADAMTS4, APOH, CCN3, DDR1, DPT, IL15, INSR, TP53; Supplementary material: Table S4) demonstrated the highest AUROC. Specifically, all 10 pre-diagnosis proteins except IL15 were higher in super survivors than poor survivors (Fig. 4A). This contexture demonstrated strong associations with survival outcomes over 1, 5, and 10 years, with AUROCs of 0.93, 0.92, and 0.91, respectively, in super and poor survivors. Incorporating age at diagnosis and sex further improved the AUROCs to 0.95, 0.96, and 0.97 (Fig. 5A), as these factors themselves demonstrated significant prognostic value (Supplementary Materials: Fig.S2). These associations remained robust across all 316 early-stage cases, with AUROCs of 0.86, 0.73, and 0.78, respectively. Similarly, when combined age and sex, AUROCs reached 0.85, 0.82,

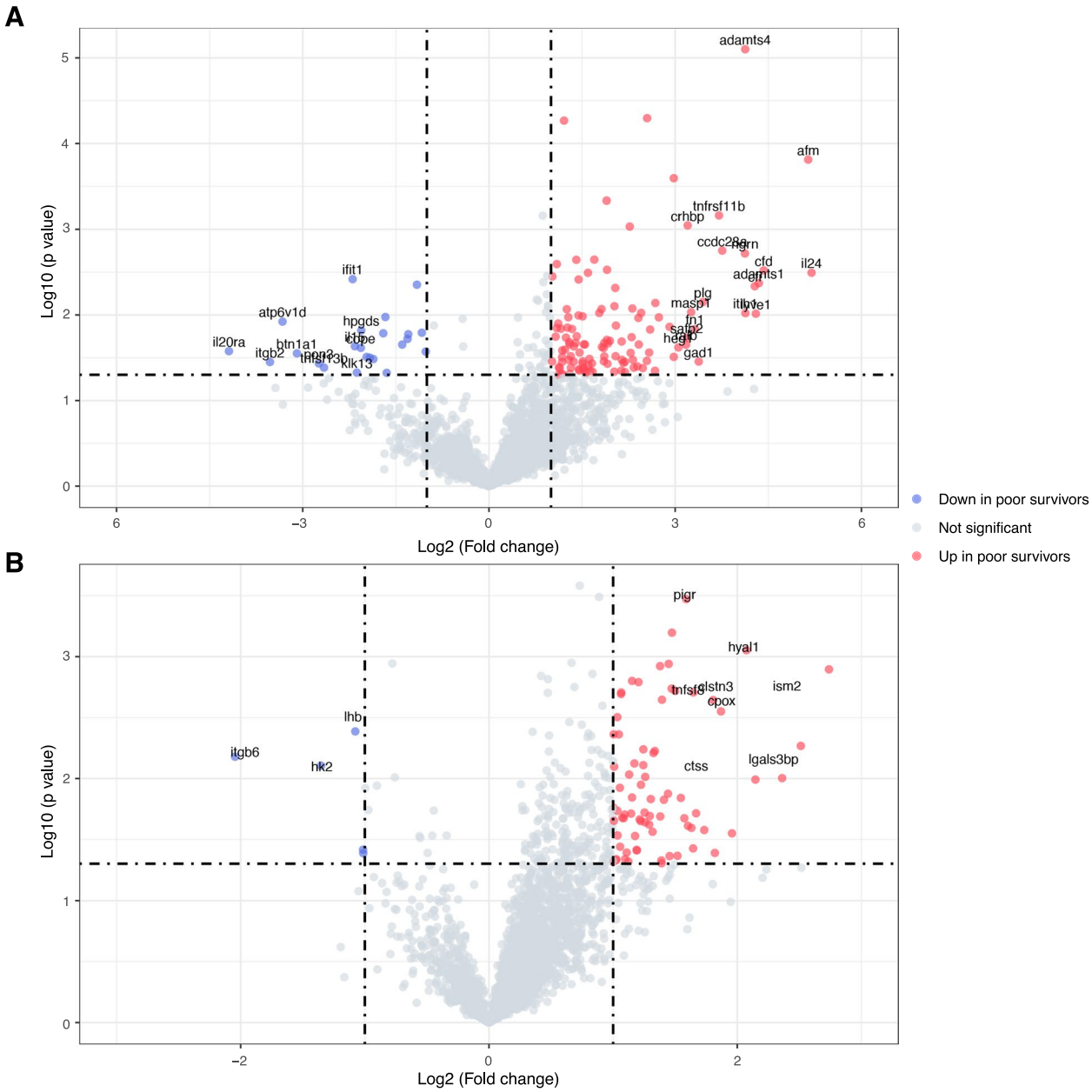

Fig. 2 Volcano plot of baseline plasma proteins in super and poor survivors among early- and late-stage colon cancer patients. **A**. Early-stage colon cancer **B**. Late-stage colon cancer |Log2 Fold change|≥ 1 and *p* value < 0.05 were applied. Grey represents not-significant change in expression. Red represents up-regulation in poor survivors. Blue represents down-regulation in poor survivors

and 0.89, respectively (Fig. 5B). For subgroup analysis, in early-stage cases, AUROCs were consistently higher for women compared to men and for younger patients compared to older patients (Supplementary Materials: Fig.S3-S4). Among 140 left-sided only early-stage cases, AUROCs were 0.84, 0.80, 0.89, respectively; in 152 right-sided only early-stage cases, AUROCs were 0.74, 0.82, 0.85, respectively (Supplementary Materials: Fig.S5), suggesting no significant difference between these two groups.

For late-stage cases, Lasso regression refined the initial pool of 89 DEPs to 10 candidates, from which an optimized 8-protein contexture (CLSTN3, EDIL3, LGALS3BP, PKD1, PPP3R1, PSMD1, TNFSF8, and YY1; Fig. 4B & Supplementary material: Table S5) was identified, demonstrating the highest AUROC and

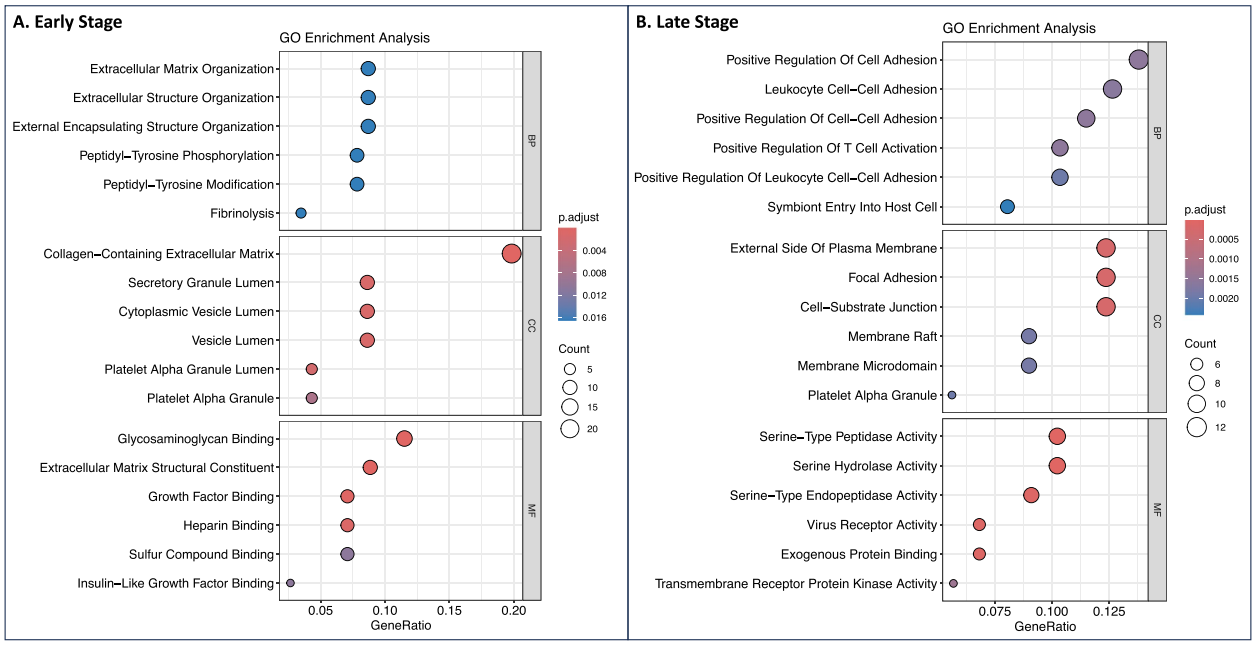

Fig. 3 Gene ontology enrichment analysis for DEPs upregulated in poor survivors in early- and late-stage colon cancer. The dot plot visualizes the top enriched GO terms based on differentially expressed proteins upregulated in poor survivors. Dot size corresponds to the number of associated proteins, while color intensity reflects the level of statistical significance (*p*-value)

highlighting pathways linked to survival dynamics in advanced disease. This contexture showed strong associations with survival outcomes over 1, 5, and 10 years (AUROCs: 0.80, 0.80, and 0.81, respectively) and remained stable with the addition of age and sex, with AUROCs at 0.82, 0.83, and 0.84, respectively (Fig. 5C). Validation across 278 late-stage cases further supported these associations, with AUROCs of 0.70, 0.69, and 0.74, which improved to 0.71, 0.72, and 0.79 upon incorporating demographic factors (Fig. 5D). Notably, the AUROC for 63 stage III cases (1-year: 0.89, 5-year: 0.79, 10-year: 0.86) was higher than that for 215 stage IV cases (1-year: 0.69, 5-year: 0.68, 10-year: 0.82) (Supplementary Materials: Fig.S6). Sensitivity analysis, which excluded deaths from non-cancer causes, showed consistent model performance, with 1-year, 5-year, and 10-year AUCs (adjusted for age and sex) remaining stable at 0.71, 0.72, and 0.79, respectively. In late-stage cases, similar AUROC performance was observed across subgroups, including comparisons between men and women (Supplementary Materials: Fig.S3), younger and older patients (Supplementary Materials: Fig.S4), as well as between left-sided and right-sided tumors (Supplementary Materials: Fig.S5). Furthermore, the model demonstrated comparable performance across various metastasis sites, including 153 cases with liver metastases, 83 cases with lung metastases, and 80 cases with retroperitoneal/peritoneal metastases (Supplementary Materials: Fig.S7).

Discussion

Our study identified the pre-diagnosis proteomic contextures associated with survival outcomes in colon cancer across early and late stages. By examining patients with extreme survival outcomes (i.e., super and poor survivors), we found that pre-existing alterations in proteomic contextures, particularly in ECM remodeling and cell adhesion, were associated with survival trajectories well before clinical diagnosis in both stages. In early-stage cases, we identified a 10-protein pre-diagnosis contexture associated with survival, which includes multiple biological processes such as ECM remodeling and immune evasion. For late-stage cases, an 8-protein pre-diagnosis contexture emerged, linked to processes including cell adhesion, angiogenesis and active inflammatory response. These associations suggest that, long before tumor detection, proteomic contexture changes in the plasma may mirror a host's immunological and biological environment, potentially shaping the fate of cancer progression and long-term survival.

In early-stage colon cancer, we identified a 10-protein pre-diagnosis proteomic contexture (ACTA2, ADAMTS1, ADAMTS4, APOH, CCN3, DDR1, DPT, INSR, IL15, TP53) strongly associated with survival outcomes. These proteins except for IL15 were elevated in

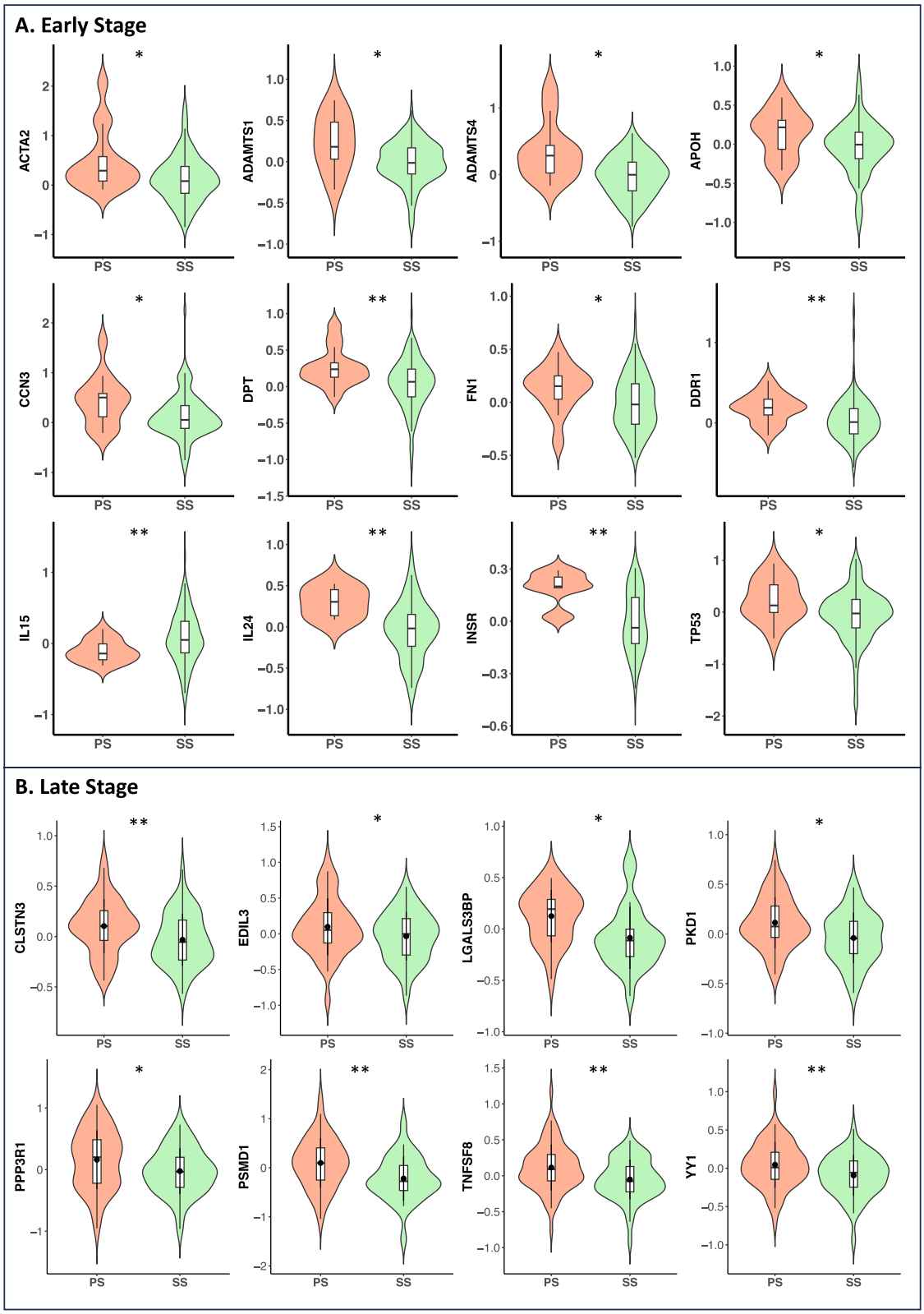

Fig. 4 Violin plot of selected plasma protein levels in poor and super survivors of early- and late-stage colon cancer. The violin plot illustrates the distribution of significant serum protein levels between poor survivors (PS) and super survivors (SS). Wider sections of the plot indicate a higher probability of those protein levels in the corresponding group. Boxplots within the violins highlight the median and interquartile range

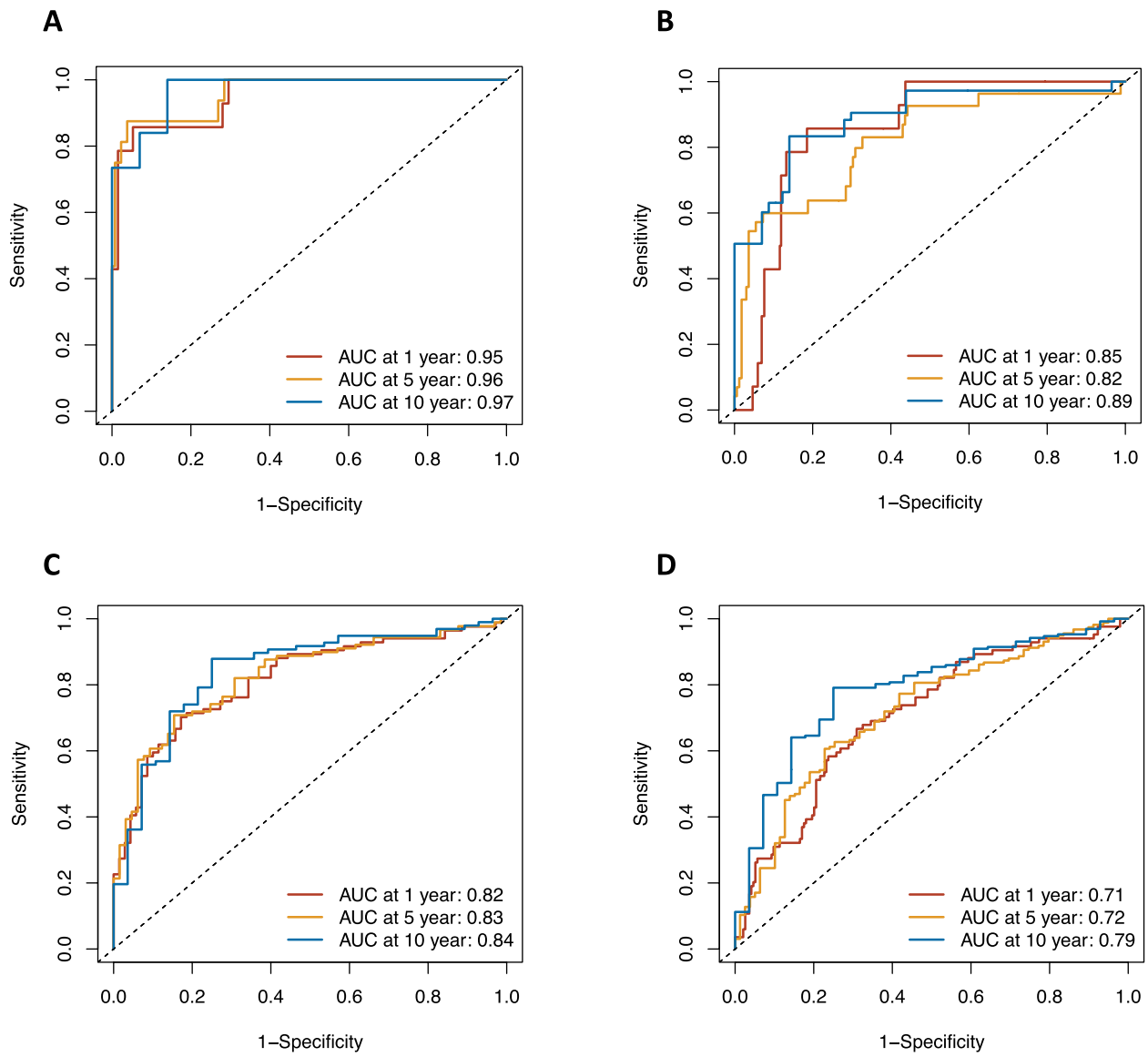

Fig. 5 Predictive model performance based on baseline serum proteins in the early- and late-stage colon cancer. This figure highlights the performance of the most effective predictive model distinguishing between super and poor survivors in early-stage colon cancer based on 10 baseline serum proteins (ACTA2, ADAMTS1, ADAMTS4, APOH, CCN3, DDR1, DPT, IL15, INSR, and TP53), and late-stage colon cancer based on 8 pre-diagnostic plasma proteins (CLSTN3, EDIL3, LGALS3BP, PKD1, PPP3R1, PSMD1, TNFSF8, and YY1). The models all incorporated sex and age at diagnosis as predictors. The upper panels (**A** and **B**) showed models in the early stage. The lower panel (**C** and **D**) showed models in the late stage. Left panels (**A** and **C**) present results for model performance in super and poor survivors, and right panels (**B** and **D**) display validation across patients using imputed data

poor survivors before diagnosis and negatively associated with survival. Notably, previous studies linked elevated levels of biomarkers such as ACTA2, ADAMTS4, APOH, and DDR1 in colon cancer tissues to poor overall survival [10–13]. Importantly, our findings revealed that these biomarkers were already upregulated years before diagnosis, suggesting that the activation of tumor-supportive processes initiated well before clinical detection. Specifically, proteins identified in the early-stage contexture, such as ACTA2 and AAMTS4, play pivotal roles in ECM remodeling, a critical process that facilitates tumor cell detachment, invasion, and metastasis in colon cancer [14–16]. The overexpression of ACTA2 in colon cancer tissues enhances ECM stiffness, facilitating tumor cell invasion and contributing to metastasis and poor prognosis [14]. ADAMTS1 is involved in ECM remodeling

by degrading various components of the matrix [17], linking to enhanced metastatic potential and poorer survival in patients with colon cancer [18]. ECM remodeling is especially critical and a prerequisite for invasion and metastasis for colon cancer progression, serving as potential therapeutic targets [19, 20]. Our findings suggest that active ECM remodeling may initiate approximately 8 years before clinical detection, highlighting the potential for early intervention to disrupt these tumor-supportive processes [21]. Conversely, we found that IL15, a key protein for immune cell activation, was associated with improved survival. The down-regulated levels of IL15 in poor survivors indicated weakened immune surveillance that could accelerate tumor onset and progression [22, 23]. Increasing studies additionally demonstrated therapeutic potential of IL15 as target for colon cancer [24, 25]. Altogether, the pre-diagnosis proteomic contexture reflects a spectrum of tumor-supportive predispositions before clinical diagnosis, particularly ECM remodeling and immune evasion, shaping overall survival for early-stage colon cancer.

In late-stage colon cancer, we identified an 8-protein pre-diagnosis proteomic contexture (CLSTN3, EDIL3, LGALS3BP, PKD1, PPP3R1, PSMD1, TNFSF8, and YY1), with elevated levels associated with poor overall survival. Previous studies have showed CLSTN3, PPP3R1, and YY1 as critical factors in colon cancer detection and progression [26–28]. Again, we further revealed that the change of these biomarkers initiated long before diagnosis. Similar to early-stage findings, proteins involving ECM remodeling and cell adhesion like CLSTN3 play pivotal roles in late-stage colon cancer [29, 30]. In animal model, CLSTN3 increased tumor invasiveness and metastasis by enhancing the ability of cancer cells to detach from the primary tumor and invade surrounding tissues. A proteome-wide Mendelian randomization study prioritized CLSTN3 as a risk factor of colon cancer, where we further revealed that elevated CLSTN3 before diagnosis was associated with poor overall survival [26]. Moreover, EDIL3, an angiogenesis-related protein, was significantly elevated in poor survivors before diagnosis, indicating early activation of tumor vascularization [31]. This early enhancement of blood vessel formation likely fosters tumor growth and progression well before clinical detection, shaping survival trajectories in late-stage colon cancer [32, 33]. PPP3R1 is a regulator of calcium signaling and immune responses. In colon cancer patients, elevated PPP3R1 levels have been linked to the establishment of a pro-inflammatory tumor microenvironment, which can support tumor growth and suppress anti-tumor immune responses, leading to poor survival [27, 34]. Notably, we discovered that YY1, a transcription factor highly expressed in colon cancer tissues, exhibits elevated levels well before diagnosis. YY1 plays a critical role in regulating genes associated with cell proliferation, DNA repair, and tumor budding, and it is widely recognized as a pivotal oncogenic factor throughout the progression of the disease [28]. Its early upregulation underscores its potential as both a therapeutic target and a predictive biomarker in colon cancer tissues [35]. Collectively, the contexture represents critical pathways, including ECM remodeling, angiogenesis, and pro-inflammatory response, actively engaged prior to diagnosis and associated with more aggressive disease behavior in late-stage colon cancer.

Despite these findings, the study has some limitations. First, due to its observational nature, it cannot establish causal relationships between identified proteins and survival outcomes. Further investigation of these proteins' roles in cancer progression and survival would be beneficial to elucidate underlying mechanisms and to strengthen their prognostic utility. Secondly, our analysis was limited to baseline proteomic data, which does not capture potential dynamic changes in biomarker levels over the course of the disease. Future studies should investigate temporal changes in protein expression to provide a more comprehensive view of how proteomic profiles evolve with disease progression and treatment. Third, reliance on ICD-10 codes for diagnosis verification may have resulted in missing cases, as pathology reports were unavailable and imaging data was only accessible for approximately one-fifth of participants within the UK Biobank cohort. Fourth, detailed stage information (stages I-IV) was unavailable in the UK Biobank dataset, and treatment data were incomplete. Metastasis status and metastasis site were used as surrogates for stage classification, which may introduce classification bias or limit finer stratification of early-stage cases. Fifth, the findings in the study lack external validation, underscoring the need for replication in diverse populations to enhance their generalizability.

In conclusion, this study highlights the pre-diagnosis proteomic contexture associated with survival outcomes in early- and late-stage colon cancer, revealing distinct biological pathways that reflect the host's immune and physiological state long before tumor diagnosis. In both stages, aggressive tumor-supportive processes such as ECM remodeling and altered cell adhesion were evident prior to clinical detection, potentially shaping survival trajectories. Early-stage cases exhibited a 10-protein signature indicative of metabolic shifts and immune evasion, while late-stage cases were characterized by an 8-protein signature emphasizing angiogenesis and pro-inflammatory response. Future research should validate these associations in broader populations and further explore the therapeutic implications of these proteomic

patterns, enhancing our understanding of early tumorigenesis and host-tumor interactions.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-025-14099-8.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Acknowledgements
We acknowledged the financial support by the Project of the High-level Public Health Professional Talents of the Beijing Municipal Health Commission under Grant [XUEKEGUGAN-01-018] and the Capital Medical University Cultivation Project (Natural Sciences Category) [PYZ24073]. These fundings did not affect study design, collection, analysis, and interpretation of the data and in the writing of the report.

Authors' contributions
Shun Li was responsible for conducting the study, performing data analysis, and drafting the manuscript. Hao Wang contributed to data collection and analysis. Xiao-Qian Xu provided analytical support. Wei-Ming Li was instrumental in developing and refining the figures. Hong You and Ji-Dong Jia revised the manuscript. You-Wen He contributed to data interpretation and conceptualization of the study. Yuan-Yuan Kong supervised the research and took overall responsibility for the study. All authors approved of the final version of the manuscript.

Funding
This study was supported by the Project of the High-level Public Health Professional Talents of the Beijing Municipal Health Commission under Grant [XUEKEGUGAN-01-018] and the Capital Medical University Cultivation Project (Natural Sciences Category) [PYZ24073].

Data availability
The data that support the findings of this study are available in the UK Biobank at https://www.ukbiobank.ac.uk/ upon application.

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]National Clinical Research Center for Digestive Diseases, State Key Lab of Digestive Health, Beijing Friendship Hospital, Capital Medical University, Beijing, China. [2]Clinical Epidemiology and EBM Unit, Beijing Clinical Research Institute, Beijing, China. [3]Liver Research Center, Beijing Friendship Hospital, Capital Medical University, Beijing, China. [4]Department of Integrative Immunobiology, Duke University School of Medicine, Durham, NC 27710, USA.

Received: 11 December 2024 Accepted: 7 April 2025
Published online: 21 April 2025

References

1. Atkin WS, Morson BC, Cuzick J. Long-term risk of colorectal cancer after excision of rectosigmoid adenomas. N Engl J Med. 1992;326(10):658–62.
2. Shao S, Neely BA, Kao TC, Eckhaus J, Bourgeois J, Brooks J, Jones EE, Drake RR, Zhu K. Proteomic Profiling of Serial Prediagnostic Serum Samples for Early Detection of Colon Cancer in the U.S. Military. Cancer Epidemiol Biomarkers Prev. 2017;26(5):711–8.
3. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3): e1001779.
4. Bertuzzi M, Marelli C, Bagnati R, Colombi A, Fanelli R, Saieva C, Ceroti M, Bendinelli B, Caini S, Airoldi L, et al. Plasma clusterin as a candidate pre-diagnosis marker of colorectal cancer risk in the Florence cohort of the European Prospective Investigation into Cancer and Nutrition: a pilot study. BMC Cancer. 2015;15:56.
5. Harlid S, Gunter MJ, Van Guelpen B. Risk-Predictive and Diagnostic Biomarkers for Colorectal Cancer; a Systematic Review of Studies Using Pre-Diagnostic Blood Samples Collected in Prospective Cohorts and Screening Settings. Cancers (Basel). 2021;13(17):4406.
6. Seum T, Frick C, Cardoso R, Bhardwaj M, Hoffmeister M, Brenner H. Potential of pre-diagnostic metabolomics for colorectal cancer risk assessment or early detection. NPJ Precis Oncol. 2024;8(1):244.
7. Levy JJ, Zavras JP, Veziroglu EM, Nasir-Moin M, Kolling FW, Christensen BC, Salas LA, Barney RE, Palisoul SM, Ren B, et al. Identification of Spatial Proteomic Signatures of Colon Tumor Metastasis: A Digital Spatial Profiling Approach. Am J Pathol. 2023;193(6):778–95.
8. Sun BB, Chiou J, Traylor M, Benner C, Hsu YH, Richardson TG, Surendran P, Mahajan A, Robins C, Vasquez-Grinnell SG, et al. Plasma proteomic associations with genetics and health in the UK Biobank. Nature. 2023;622(7982):329–38.
9. Marley AR, Nan H. Epidemiology of colorectal cancer. Int J Mol Epidemiol Genet. 2016;7(3):105–14.
10. Ito S, Koshino A, Wang C, Otani T, Komura M, Ueki A, Kato S, Takahashi H, Ebi M, Ogasawara N, et al. Characterisation of colorectal cancer by hierarchical clustering analyses for five stroma-related markers. J Pathol Clin Res. 2024;10(4): e12386.
11. Shang XQ, Liu KL, Li Q, Lao YQ, Li NS, Wu J. ADAMTS4 is upregulated in colorectal cancer and could be a useful prognostic indicator of colorectal cancer. Rev Assoc Med Bras. 2020;66(1):42–7.
12. Lu Y, Wang Y, Qiu Y, Xuan W. Analysis of the Relationship between the Expression Level of TTR and APOH and Prognosis in Patients with Colorectal Cancer Metastasis Based on Bioinformatics. Contrast Media Mol Imaging. 2022;2022:1121312.
13. Jeitany M, Leroy C, Tosti P, Lafitte M, Le Guet J, Simon V, Bonenfant D, Robert B, Grillet F, Mollevi C, et al. Inhibition of DDR1-BCR signalling by nilotinib as a new therapeutic strategy for metastatic colorectal cancer. EMBO Mol Med. 2018;10(4):e7918.
14. Lee HW, Park YM, Lee SJ, Cho HJ, Kim D-H, Lee J-I, Kang M-S, Seol HJ, Shim YM, Nam D-H, et al. Alpha-Smooth Muscle Actin (ACTA2) Is Required for Metastatic Potential of Human Lung Adenocarcinoma. Clin Cancer Res. 2013;19(21):5879–89.
15. Andreuzzi E, Capuano A, Poletto E, Pivetta E, Fejza A, Favero A, Doliana R, Cannizzaro R, Spessotto P, Mongiat M. Role of Extracellular Matrix in Gastrointestinal Cancer-Associated Angiogenesis. Int J Mol Sci. 2020;21(10):3686.
16. Chen J, Luo Y, Zhou Y, Qin S, Qiu Y, Cui R, Yu M, Qin J, Zhong M. Promotion of Tumor Growth by ADAMTS4 in Colorectal Cancer: Focused on Macrophages. Cell Physiol Biochem. 2018;46(4):1693–703.
17. Silva SV, Lima MA, Hodgson L, Rodríguez-Manzaneque JC, Freitas VM. ADAMTS-1 has nuclear localization in cells with epithelial origin and leads to decreased cell migration. Exp Cell Res. 2023;433(2): 113852.
18. Lind GE, Kleivi K, Meling GI, Teixeira MR, Thiis-Evensen E, Rognum TO, Lothe RA. ADAMTS1, CRABP1, and NR3C1 identified as epigenetically deregulated genes in colorectal tumorigenesis. Cell Oncol. 2006;28(5–6):259–72.
19. Kim MS, Ha SE, Wu M, Zogg H, Ronkon CF, Lee MY, Ro S. Extracellular Matrix Biomarkers in Colorectal Cancer. Int J Mol Sci. 2021;22(17):9185.
20. Karlsson S, Nyström H. The extracellular matrix in colorectal cancer and its metastatic settling – Alterations and biological implications. Crit Rev Oncol Hematol. 2022;175: 103712.
21. Yuan Z, Li Y, Zhang S, Wang X, Dou H, Yu X, Zhang Z, Yang S, Xiao M. Extracellular matrix remodeling in tumor progression and immune escape: from mechanisms to treatments. Mol Cancer. 2023;22(1):48.
22. Cai M, Huang X, Huang X, Ju D, Zhu YZ, Ye L. Research progress of interleukin-15 in cancer immunotherapy. Front Pharmacol. 2023;14:1184703.
23. Bahri R, Pateras IS, D'Orlando O, Goyeneche-Patino DA, Campbell M, Polansky JK, Sandig H, Papaioannou M, Evangelou K, Foukas PG, et al. IL-15 suppresses colitis-associated colon carcinogenesis by inducing antitumor immunity. Oncoimmunology. 2015;4(9): e1002721.

24. Lei S, Zhang X, Men K, Gao Y, Yang X, Wu S, Duan X, Wei Y, Tong R. Efficient Colorectal Cancer Gene Therapy with IL-15 mRNA Nanoformulation. Mol Pharm. 2020;17(9):3378–91.
25. Thi VAD, Jeon HM, Park SM, Lee H, Kim YS. Cell-Based IL-15:IL-15Rα Secreting Vaccine as an Effective Therapy for CT26 Colon Cancer in Mice. Mol Cells. 2019;42(12):869–83.
26. Sun J, Zhao J, Jiang F, Wang L, Xiao Q, Han F, Chen J, Yuan S, Wei J, Larsson SC, et al. Identification of novel protein biomarkers and drug targets for colorectal cancer by integrating human plasma proteome with genome. Genome Med. 2023;15(1):75.
27. Sun Z, Xia W, Lyu Y, Song Y, Wang M, Zhang R, Sui G, Li Z, Song L, Wu C, et al. Immune-related gene expression signatures in colorectal cancer. Oncol Lett. 2021;22(1):543.
28. Shao Z, Yang W, Meng X, Li M, Hou P, Li Z, Chu S, Zheng J, Bai J. The role of transcription factor Yin Yang-1 in colorectal cancer. Cancer Med. 2023;12(10):11177–90.
29. Oplawski M, Dziobek K, Zmarzły N, Grabarek B, Tomala B, Leśniak E, Adwent I, Januszyk P, Dąbruś D, Boroń D. Evaluation of Changes in the Expression Pattern of EDIL3 in Different Grades of Endometrial Cancer. Curr Pharm Biotechnol. 2019;20(6):483–8.
30. Capone E, Iacobelli S, Sala G. Role of galectin 3 binding protein in cancer progression: a potential novel therapeutic target. J Transl Med. 2021;19(1):405.
31. Aoka Y, Johnson FL, Penta K, Hirata Ki K, Hidai C, Schatzman R, Varner JA, Quertermous T. The embryonic angiogenic factor Del1 accelerates tumor growth by enhancing vascular formation. Microvasc Res. 2002;64(1):148–61.
32. Tabasum S, Thapa D, Giobbie-Hurder A, Weirather JL, Campisi M, Schol PJ, Li X, Li J, Yoon CH, Manos MP, et al. EDIL3 as an Angiogenic Target of Immune Exclusion Following Checkpoint Blockade. Cancer Immunol Res. 2023;11(11):1493–507.
33. Kim N, Ko Y, Shin Y, Park J, Lee AJ, Kim KW, Pyo J. Comprehensive Analysis for Anti-Cancer Target-Indication Prioritization of Placental Growth Factor Inhibitor (PGF) by Use of Omics and Patient Survival Data. Biology (Basel). 2023;12(7):970.
34. Cho SH, Shim HJ, Park MR, Choi JN, Akanda MR, Hwang JE, Bae WK, Lee KH, Sun EG, Chung IJ. Lgals3bp suppresses colon inflammation and tumorigenesis through the downregulation of TAK1-NF-κB signaling. Cell Death Discov. 2021;7(1):65.
35. Shao ZY, Yang WD, Qiu H, He ZH, Lu MR, Shen Q, Ding J, Zheng JN, Bai J. The role of USP7-YY1 interaction in promoting colorectal cancer growth and metastasis. Cell Death Dis. 2024;15(5):347.

Publisher's Note