Computational and Structural Biotechnology Journal 21 (2023) 365–377

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

PCLassoLog: A protein complex-based, group Lasso-logistic model for cancer classification and risk protein complex discovery

Wei Wang [a], Haiyan Yuan [a], Junwei Han [b,*], Wei Liu [a,*]

[a] College of Science, Heilongjiang Institute of Technology, Harbin 150050, China
[b] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China

ARTICLE INFO

Article history:
Received 10 September 2022
Received in revised form 2 December 2022
Accepted 3 December 2022
Available online 6 December 2022

Keywords:
Protein complex
Deep proteomic data
Group Lasso
Logistic model
Cancer classification

ABSTRACT

Risk gene identification has attracted much attention in the past two decades. Since most genes need to be translated into proteins and cooperate with other proteins to form protein complexes to carry out cellular functions, which significantly extends the functional diversity of individual proteins, revealing the molecular mechanism of cancer from a comprehensive perspective needs to shift from identifying individual risk genes toward identifying risk protein complexes. Here, we embed protein complexes into the regularized learning framework and propose a protein complex-based, group Lasso-logistic model (PCLassoLog) to discover risk protein complexes. Experiments on deep proteomic data of two cancer types show that PCLassoLog yields superior predictive performance on independent datasets. More importantly, PCLassoLog identifies risk protein complexes that not only contain individual risk proteins but also incorporate close partners that synergize with them. Furthermore, selection probabilities are calculated and two other protein complex-based models are proposed to complement PCLassoLog in identifying reliable risk protein complexes. Based on PCLassoLog, a pan-cancer analysis is performed to identify risk protein complexes in 12 cancer types. Finally, PCLassoLog is used to discover risk protein complexes associated with gene mutation. We implement all protein complex-based models as an R package PCLassoReg, which may serve as an effective tool to discover risk protein complexes in various contexts.

1. Introduction

Identifying risk genes is the first step in revealing the molecular mechanism of cancer, finding drug targets, and developing novel therapeutic strategies. Since the advent of high-throughput omics data, a plethora of computational approaches have been developed to identify risk genes [1–6]. However, since most genes need to be translated into proteins and cooperate with other proteins to form protein complexes or functional modules to carry out cellular functions, although risk genes provide potential targets for cancer research, they are inherently insufficient in revealing the underlying molecular mechanisms and guiding the development of therapeutic strategies from a comprehensive perspective. Protein complexes are key molecular entities that integrate multiple gene products to perform cellular functions [7]. They are basic representatives of functional modules. Data from single cell organisms provide evidence that>50 % or even 80 % of proteins work in protein complexes (complexome) [7–9]. Protein complexes play critical roles in an array of biological processes, including protein synthesis, signaling and cellular degradation processes [7]. Given the vital functions of these macromolecular machines, identifying risk protein complexes in tumor samples is fundamental to our understanding of cancer biology. Furthermore, the sophisticated organization of individual proteins into macromolecular assemblies significantly extends the functional diversity of individual proteins and allows cells to acquire novel functionalities that are beyond the performance of individual proteins [10–12]. Thus, revealing the molecular mechanism of cancer and developing new therapeutic strategies from a comprehensive perspective need to shift from identifying individual risk genes toward identifying risk protein complexes.

The study of protein complex is becoming an important means to reveal the molecular mechanism and guide drug development. To explore the molecular details of how severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infects cells and develop drugs and vaccines against COVID-19, Krogan and co-workers

* Corresponding authors.
E-mail addresses: hanjunwei@ems.hrbmu.edu.cn (J. Han), liuwei@hljit.edu.cn (W. Liu).

https://doi.org/10.1016/j.csbj.2022.12.005

characterized SARS-CoV-2-host protein complexes formed by the physical interactions between 26 SARS-CoV-2 proteins and human proteins using affinity-purification mass spectrometry, and identified 66 druggable human proteins targeted by 69 compounds, which may lead to a therapeutic regimen to treat COVID-19 [13]. For human cancer, the role of some protein complexes in cancer has been revealed, such as the E-cadherin/catenin adhesion complex in the development and progression of cancer [14], the TWIST/Mi2/NuRD protein complex in cancer metastasis [15], the oncogenic Tcf/beta-catenin protein complex [16] and the shelterin complex [17]. Potential therapeutic strategies that target protein complexes rather than individual proteins are being revealed. For example, the CAV1-GLUT3 complex has been reported to be targeted by Atorvastatin to suppress tumor growth in non-small cell lung cancer [18]. However, studying key protein complexes in cancer at the system level lacks reliable targets. To date, few methods have been proposed to identify risk protein complexes on a large scale. In our previous study, we proposed a protein complex-based prognostic model, PCLasso [19], which shows superior prognostic performance than those based on individual genes. PCLasso uses survival outcome as the dependent variable and embeds protein complexes into the group Lasso-Cox model to construct prognostic models, which identify risk protein complexes associated with survival outcomes. However, PCLasso is unable to deal with the classification problem and identify the risk protein complexes that distinguish tumors from non-tumors, which may play important roles in the occurrence and development of cancer.

In this study, we embed protein complexes into classification models under the regularized learning framework, and propose a protein complex-based, group Lasso-logistic (PCLassoLog) model toward accurate cancer classification and risk protein complex discovery. PCLassoLog is an extension of PCLasso tuned for protein complex-based classification problems. We apply PCLassoLog to the classification of lung adenocarcinoma (LUAD) and hepatocellular cancer (HCC) patients and prove that PCLassoLog has superior predictive performance than the Lasso-logistic model based on individual genes and is able to identify cancer-related risk protein complexes. In addition, we calculate selection probabilities and implement two other protein complex-based group selection models to complement PCLassoLog in identifying reliable risk protein complexes. Finally, PCLassoLog is used to discover risk protein complexes associated with cancer development in 12 cancer types and those associated with gene mutation in LUAD. These risk protein complexes may provide potential targets at the protein complex level for cancer research and guide the development of new therapeutic strategies.

2. Materials and methods

2.1. Datasets

We collected deep protein expression datasets (1424 tumors and 1060 non-tumors) for 12 cancer types including LUAD [20,21], HCC [22], clear cell renal cell carcinoma (ccRCC) [23], colon adenocarcinoma (COAD) [24], endometrial carcinoma (EC) [25], esophageal squamous cell carcinoma (ESCC) [26], glioblastoma (GBM) [27], gastric cancer (GC) [28], head-and-neck squamous cell carcinoma (HNSCC) [29], lung squamous cell carcinoma (LSCC) [30], ovarian cancer (OV) [31], and pancreatic ductal adenocarcinoma (PDAC) [32] from recently published proteomics studies (Table 1). The two protein expression datasets for LUAD were obtained from the studies of Gillette et al. [20] and Xu et al. [21], and were named LUAD.Gillette.Prot and LUAD.Xu.Prot, respectively. We also obtained mRNA expression datasets of the same samples in these two studies and named them LUAD.Gillette.mRNA and LUAD.Xu.mRNA, respectively. In addition, we obtained three independent mRNA expression datasets for LUAD from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/): GSE10072 [33], GSE19804 [34], and GSE19188 [35], denoted as LUAD-valid1, LUAD-valid2, and LUAD-valid3, respectively. For HCC, we obtained five independent mRNA expression datasets (GSE54236 [36], GSE76427 [37], GSE64041 [38], GSE47197, and GSE14520 [39]) from the GEO database for verification, which were denoted as HCC-valid1, HCC-valid2, HCC-valid3, HCC-valid4, and HCC-valid5, respectively. In total, the mRNA expression data of 909 tumors and 794 non-tumors were obtained (Table 1). Data processing and normalization were provided in **Supplementary Text**.

2.2. Protein complexes

The protein complexes were obtained from the CORUM database (Release 3.0) [7]. We chose to download the core set as it is a reduced dataset that is essentially free of redundant entries. The core set contains 3512 mammalian protein complexes, from which 2417 human protein complexes composed of 3420 unique proteins were selected for downstream analysis.

2.3. Lasso-logistic model

Assume that we have a protein expression matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n] \in R^{n\times p}$ and a vector of response variables $\boldsymbol{Y} = [y_1, y_2, \ldots, y_n]^T \in R^n$, where n is the number of samples and p is the number of proteins. The i-th sample can be denoted as $(\boldsymbol{x}_i, y_i)$, where $\boldsymbol{x}_i \in R^p$ is its expression value vector and $y_i \in \{0, 1\}$ is a binary response variable. Without loss of generality, we assume that $y_i = 1$ means that the sample is a tumor sample, and $y_i = 0$ means it is a non-tumor sample. Linear logistic regression models the condition probability $P(y = 1|\boldsymbol{x})$ by

$$\log\left(\frac{P(y=1|\boldsymbol{x})}{1-P(y=1|\boldsymbol{x})}\right) = \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta} \tag{1}$$

where β_0 is the intercept and $\boldsymbol{\beta} = \left[\beta_1, \beta_2, \ldots, \beta_p\right]^T \in R^p$ is the parameter vector. Then we have

$$P(y=1|\boldsymbol{x}) = \frac{e^{\beta_0+\boldsymbol{x}^T\boldsymbol{\beta}}}{1+e^{\beta_0+\boldsymbol{x}^T\boldsymbol{\beta}}} \tag{2}$$

For logistic regression model, the likelihood function is:

$$L(\beta_0, \boldsymbol{\beta}) = \prod_{i=1}^{n}[P(y_i=1|\boldsymbol{x}_i)]^{y_i}[1-P(y_i=1|\boldsymbol{x}_i)]^{1-y_i} \tag{3}$$

Substituting (2) into (3) and taking the logarithm, we obtain the log-likelihood function:

$$LL(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n}\left[y_i(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}) - \log\left(1+e^{\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)\right] \tag{4}$$

The regression coefficients $\beta_0, \boldsymbol{\beta}$ can be estimated by maximizing the log-likelihood function (4). This is equivalent to solving the following problem:

$$\min_{\beta_0,\boldsymbol{\beta}}\left\{-\frac{1}{n}\sum_{i=1}^{n}\left[y_i(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}) - \log\left(1+e^{\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)\right]\right\} \tag{5}$$

In the high-dimensional setting ($p \gg n$), directly solving the problem (5) is ill-posed. The Lasso-logistic model can effectively solve this problem by including a regularization term:

$$\min_{\beta_0,\boldsymbol{\beta}}\left\{-\frac{1}{n}\sum_{i=1}^{n}\left[y_i(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}) - \log\left(1+e^{\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)\right] + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{6}$$

Table 1
Details about the datasets used for the 12 cancer types.

Cancer types	Source / GEO accession NO.	Quantitative method / Platform	Datasets	Expression profiles	NO. of tumors	NO. of non-tumors	NO. of proteins/ mRNAs
LUAD	Gillette et al. [20]	TMT 10-plex	LUAD.Gillette.Prot	Protein	110	101	7136
	Xu et al. [21]	LFQ	LUAD.Xu.Prot	Protein	103	103	6594
	Gillette et al. [20]	HiSeq4000	LUAD.Gillette.mRNA	mRNA	110	101	17,679
	GSE140343 [21]	GPL20795	LUAD.Xu.mRNA	mRNA	51	49	15,340
	GSE10072	GPL96	LUAD-valid1	mRNA	58	49	12,754
	GSE19804	GPL570	LUAD-valid2	mRNA	60	60	21,298
	GSE19188	GPL570	LUAD-valid3	mRNA	91	65	21,298
HCC	Gao et al. [22]	TMT 11-plex	HCC	Protein	159	159	6478
	GSE54236	GPL6480	HCC-valid1	mRNA	78	77	19,595
	GSE76427	GPL10558	HCC-valid2	mRNA	115	52	34,693
	GSE64041	GPL6244	HCC-valid3	mRNA	60	60	23,307
	GSE47197	GPL16699	HCC-valid4	mRNA	61	61	16,390
	GSE14520	GPL3921	HCC-valid5	mRNA	225	220	12,742
ccRCC	Clark et al. [23]	TMT 10-plex	ccRCC	Protein	110	84	7150
COAD	Vasaikar et al. [24]	TMT 10-plex	COAD	Protein	100	97	4376
EC	Dou et al. [25]	TMT 10-plex	EC	Protein	95	49	7908
ESCC	Liu et al. [26]	TMT 11-plex	ESCC	Protein	124	124	6461
GBM	Wang et al. [27]	TMT 11-plex	GBM	Protein	99	10	8828
GC	Ge et al. [28]	LFQ	GC	Protein	84	84	5439
HNSCC	Huang et al. [29]	TMT 11-plex	HNSCC	Protein	109	63	7515
LSCC	Satpathy et al. [30]	TMT 11-plex	LSCC	Protein	108	99	8218
OV	Hu et al. [31]	iTRAQ	OV	Protein	83	20	7599
PDAC	Cao et al. [32]	TMT 11-plex	PDAC	Protein	140	67	5755

TMT: tandem mass tags-based isobaric labeling; LFQ: label-free quantification; iTRAQ: isobaric tag for relative and absolute quantitation; LUAD: lung adenocarcinoma; HCC: hepatocellular carcinoma; ccRCC: clear cell renal cell carcinoma; COAD: colon adenocarcinoma; EC: endometrial carcinoma; ESCC: esophageal squamous cell carcinoma; GBM: glioblastoma; GC: gastric cancer; HNSCC: head-and-neck squamous cell carcinoma; LSCC: lung squamous cell carcinoma; OV: ovarian cancer; PDAC: pancreatic ductal adenocarcinoma.

where $\lambda > 0$ is a control parameter and $|\cdot|$ is the l_1 norm. It generates a sparse solution $\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)$ with only a few nonzero coefficients β_ks, corresponding to the proteins selected. Then the probability that a new sample is a tumor sample can be estimated by:

$$P(y=1|\boldsymbol{x}) = \frac{e^{\hat{\beta}_0+\boldsymbol{x}^T\hat{\boldsymbol{\beta}}}}{1+e^{\hat{\beta}_0+\boldsymbol{x}^T\hat{\boldsymbol{\beta}}}} \tag{7}$$

where $\boldsymbol{x}$ is the protein expression vector of the new sample. Given a threshold *th*, the new sample is predicted to be a tumor sample if $P(y=1|\boldsymbol{x}) \geqslant th$ and a non-tumor sample otherwise. In this study, th = 0.5 was used. As important predictors, proteins with nonzero coefficients can be considered as potential risk proteins for further investigation.

2.4. Protein complex-based, group Lasso-logistic model

Note that the Lasso-logistic model is similar to the Lasso-Cox model [19] except for the log likelihood function (the first term in equation (6)). We use the same strategy as PCLasso to integrate protein complexes into the regularized learning framework (6) and propose a protein complex-based, group Lasso-logistic model (PCLassoLog) to predict the class of samples and identify risk protein complexes. The group Lasso-logistic model for nonoverlapping groups can be formulated as follows [40]:

$$\min_{\beta_0,\boldsymbol{\beta}}\left\{-\frac{1}{n}\sum_{i=1}^{n}\left[y_i\left(\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}\right)-\log\left(1+e^{\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)\right]+\lambda\sum_{k=1}^{K}\sqrt{|G_k|}\,\|\,\boldsymbol{\beta}_{G_k}\,\|\right\} \tag{8}$$

where $\|\cdot\|$ is the Euclidean norm or l_2 norm, $|G_k|$ is the size of the k-th group, and $\boldsymbol{\beta}_{G_k} \in R^{|G_k|}$ is a coefficient vector of the k-th group (Fig. S1**A**). PCLassoLog deals with the overlap problem of protein complexes by using a latent group Lasso method [19,41,42]:

$$\min_{\beta_0,\boldsymbol{\beta}}\left\{-\frac{1}{n}\sum_{i=1}^{n}\left[y_i\left(\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}\right)-\log\left(1+e^{\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)\right]+\lambda\sum_{k=1}^{K}\sqrt{|G_k|}\,\|\,\boldsymbol{\gamma}_k\,\|\right\}$$
$$s.t.\quad \boldsymbol{\beta}=\sum_{k=1}^{K}\boldsymbol{\gamma}_k \tag{9}$$

where equation $\boldsymbol{\beta}=\sum_{k=1}^{K}\boldsymbol{\gamma}_k$ means that $\boldsymbol{\beta}$ is decomposed into the sum of K latent vectors $\boldsymbol{\gamma}_k=\left(\gamma_{k1},\gamma_{k2},...,\gamma_{kp}\right)^T\in R^p, k=1,2,...,K$, whose supports correspond to the proteins contained in each protein complex, i.e., $\mathrm{supp}(\boldsymbol{\gamma}_k)\subset G_k$ (Fig. S1**B**). PCLassoLog is solved in the same way as PCLasso [19] (see **Supplementary Text**).

2.5. Other group selection methods

In addition to the Lasso penalty, we investigated other two penalties, namely smoothly clipped absolute deviation penalty (SCAD) [43] and minimax concave penalty (MCP) [44]. The two penalties add an additional parameter a for relaxing the penalties, and are defined for $\lambda > 0$ as follows:

$$p_{\lambda,a}^{SCAD}(|z|)=\begin{cases}\lambda|z|, & if\ |z|\leqslant\lambda\\ \frac{-|z|^2+2a\lambda|z|-\lambda^2}{2(a-1)} & if\ \lambda<|z|\leqslant a\lambda\\ \frac{(a+1)\lambda^2}{2} & if\ |z|>a\lambda\end{cases} \tag{10}$$

$$p_{\lambda,a}^{MCP}(|z|)=\begin{cases}\lambda|z|-\frac{|z|^2}{2a} & if\ |z|\leqslant a\lambda\\ \frac{a\lambda^2}{2} & if\ |z|>a\lambda\end{cases} \tag{11}$$

where $a > 2$ for SCAD and $a > 1$ for MCP.

Among the three penalties, the Lasso is the largest, followed by SCAD, and MCP is the smallest (**Figure S2**). The protein complex-based, group SCAD-logistic model (PCSCAD) and group MCP-logistic model (PCMCP) are defined by replacing the Lasso penalty in PCLassoLog with SCAD and MCP penalties, respectively (see **Supplementary Text**).

2.6. Model evaluation and selection probability

The area under the ROC curve (AUC) and classification accuracy were used to evaluate classification performance. To obtain an unbiased evaluation, we adopted a resampling strategy. Let I_m be the m-th random subsample of $\{1,2,...,n\}$ of size $\lfloor 2n/3 \rfloor$ without replacement, where $\lfloor x \rfloor$ is the largest integer not greater than x,$m = 1,2,...,M$. Each I_m was used as a training set to train the model. Then the AUCs and classification accuracies of each model on the corresponding test set $\{1,2,...,n\}\{I_m$ and independent validation datasets were estimated. All M AUCs and accuracies were used to evaluate the overall predictive performance of the model. For each training set I_m, cross-validation was used to determine the optimal λ value, denoted as λ_m. For the PCSCAD and PCMCP model, a series of a-values of 4, 6, 8, ..., 500 were investigated.

To identify reliable risk protein complexes, we define the selection probability (SP) of the k-th protein complex according to the stability selection theory [45–47] as follow:

$$SP(k) = \frac{1}{M}\#\left\{m : k \in \hat{S}^{\lambda_m}(I_m)\right\}$$

where $\hat{S}^{\lambda_m}(I_m) \subset \{1,2,...,K\}$ denotes the protein complexes selected by PCLassoLog when applied to subsample I_m. Selection probabilities give a confidence measure for risk protein complexes (see **Supplementary Text**).

3. Results

3.1. Overview of the PCLassoLog model

The PCLassoLog model is a classification model that selects important predictors at the protein complex level to achieve accurate classification and identify risk protein complexes. The PCLassoLog model has three inputs: a protein expression matrix, a vector of binary response variables, and a number of known protein complexes (Fig. 1). Considering that proteins usually function by forming protein complexes, PCLassoLog regards proteins belonging to the same protein complex as a group and constructs a group Lasso penalty (l_1/l_2 penalty) based on the sum (i.e. l_1 norm) of the l_2 norms of the regression coefficients of the group members to perform feature selection at the group level [41]. With the group Lasso penalty, PCLassoLog trains the logistic regression model and obtains a sparse solution at the protein complex level, that is, the proteins belonging to a protein complex are either wholly included or wholly excluded from the model. PCLassoLog outputs a prediction model and a small set of protein complexes included in the model, which are referred to as risk protein complexes (Fig. 1). For further details, see Materials and Methods.

3.2. PCLassoLog yields superior predictive performance based on protein complexes

To evaluate the classification performance of PCLassoLog and its ability to identify risk protein complexes, we constructed the PCLassoLog model based on three protein expression datasets: LUAD.Gillette.Prot, LUAD.Xu.Prot, and HCC (Table 1). For the convenience of description, we will refer to the experiments based on the three datasets as the "LUAD.Gillette.Prot", "LUAD.Xu.Prot", and "HCC" cases, respectively. For each case, we used a training set to construct the PCLassoLog model and used a test set and multiple independent validation sets to evaluate classification accuracy and robustness. The samples in the test set are similar to those in the training set, while the samples of the validation sets are completely independent of the training set, and different platforms may be used, which could better verify the robustness of PCLassoLog (Table 1). The AUCs and classification accuracies were calculated to evaluate classification performance. We also performed the individual protein-based Lasso-logistic model following the same procedure and compared the classification performance of the two models.

3.2.1. LUAD.Gillette.Prot

Two independent protein expression datasets for LUAD were collected. We first trained the PCLassoLog model using the LUAD.Gillette.Prot dataset. The LUAD.Gillette.Prot dataset was randomly split into a training set (LUAD.Gillette.Prot–train, n = 140) and a test set (LUAD.Gillette.Prot–test, n = 71), keeping the ratio of the number of tumors to non-tumors in the test set the same as that in the training set. The PCLassoLog model trained on the LUAD.Gillette.Prot–train dataset (see **Supplementary Text**) contained 17 protein complexes composed of 34 proteins (Fig. 2**A-C**; **Table S1**). Many proteins belonging to these 17 protein complexes have been reported to play critical role in tumorigenicity or promote the proliferation, metastasis, and invasion of LUAD, such as ILK [48], PARVA [49], AQP4 [50], and PRKCI [51]. The Lasso-logistic model selected 25 individual proteins (Fig. S3**A**; **Table S2**). Some proteins identified by the Lasso-logistic model were shared by PCLassoLog, but without the information at the protein complex level. Both models achieved perfect classification on the LUAD.Gillette.Prot-test dataset, and were almost perfect on the independent protein expression dataset LUAD.Xu.Prot, with the AUC of PCLassoLog slightly larger (Fig. 2**D**).

Next, we verified the classification performance of PCLassoLog on more independent datasets. Due to the lack of protein expression data at present, we tried to use mRNA expression data as a proxy. We first investigated the mRNA expression data in the LUAD.Gillette.mRNA dataset, which were measured on the same samples as the LUAD.Gillette.Prot dataset. Correlation analysis based on the common 1037 protein complexes found that 99.5 %

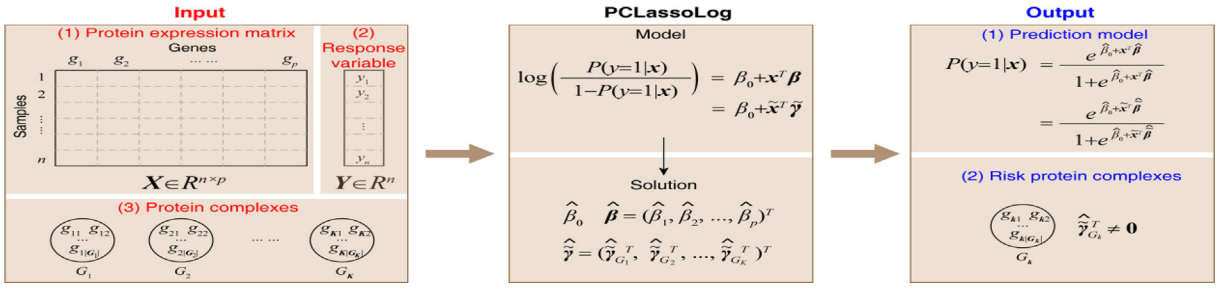

Fig. 1. Workflow of the PCLassoLog model. The PCLassoLog model takes a protein expression matrix, a binary response variable vector, and a collection of protein complexes as input. It estimates the correlation between protein expression and the response variable, and performs feature selection at the protein complex level. PCLassoLog outputs a prediction model for the classification of new patients and risk protein complexes of the disease.

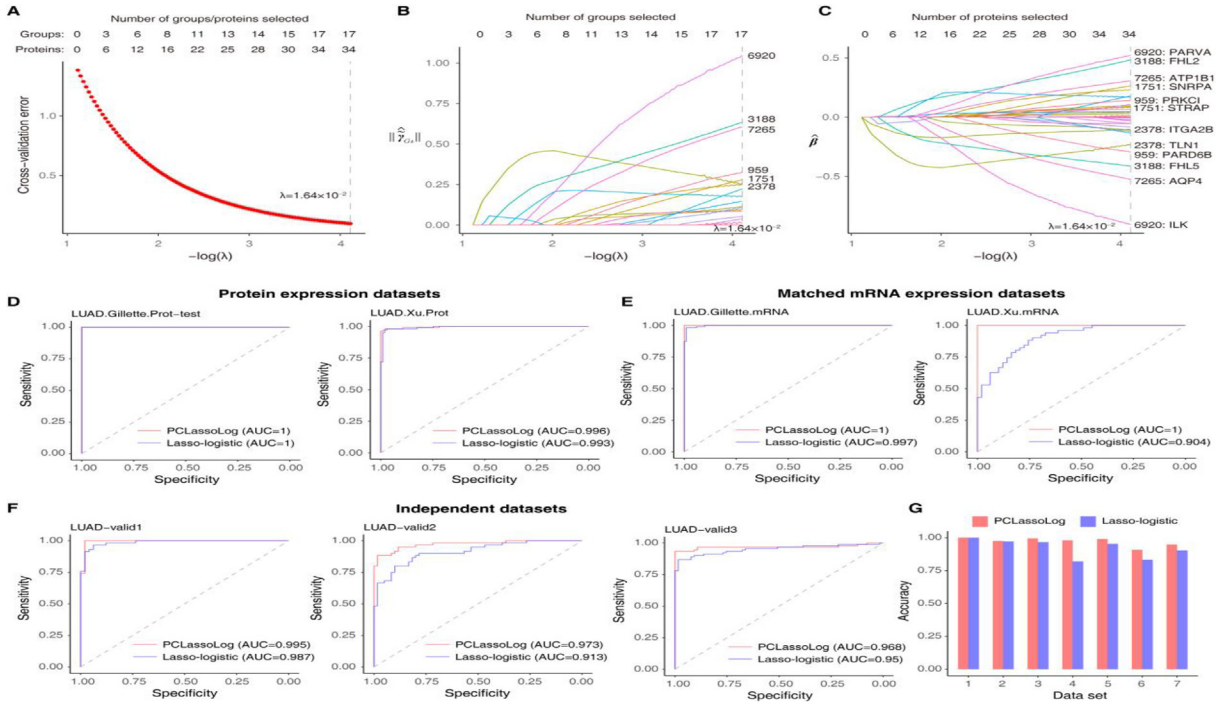

Fig. 2. Classification of LUAD patients based on models constructed on the LUAD.Gillette.Prot dataset. (A) Plot of cross-validation errors. Each point represents the cross-validation error of a λ value, with error bars to give a confidence interval for the cross-validation errors. The vertical bar indicates the λ value corresponding to the final model we selected. The top of the plot gives the size of each model at the group (protein complex) level and the individual protein level. (B) Norms of the coefficient vectors of the protein complexes. The complex IDs of the six protein complexes with the largest norm are shown on the right. (C) Coefficient paths of individual proteins. The proteins corresponding to the protein complexes in (B) are shown on the right. Proteins belonging to the same protein complex were selected into the model at the same time. (D-F) ROC curves of the PCLassoLog and Lasso-logistic models on two protein expression datasets (D), two matched mRNA expression datasets (E), and three independent mRNA expression datasets (F). (G) Classification accuracies of the PCLassoLog and Lasso-logistic models on the seven datasets. 1: LUAD.Gillette.Prot-test; 2: LUAD.Xu.Prot; 3: LUAD.Gillette.mRNA; 4: LUAD.Xu.mRNA; 5: LUAD-valid1; 6: LUAD-valid2; 7: LUAD-valid3.

sample-wise protein-mRNA pairs and 83.3 % gene-wise protein-mRNA pairs show significant positive Spearman correlations (BH adjusted p value < 0.05) (Fig. S4**A**). Among the 34 proteins contained in PCLassoLog, 31 (91.2 %) proteins have significant positive correlations (**Table S1**), such as ILK, PARVA, FHL2, and FHL5 (Fig. S4**-B-4C**), suggesting that mRNA expression data may be a feasible proxy for protein expression data. Indeed, both PCLassoLog and Lasso-logistic achieved almost perfect classification on the LUAD.Gillette.mRNA dataset (Fig. 2**E**). For other four mRNA datasets, the AUCs obtained by PCLassoLog are larger than those of the Lasso-logistic model, as well as classification accuracies (Fig. 2**E-G**).

3.2.2. LUAD.Xu.Prot

We next constructed a PCLassoLog model with 18 protein complexes / 36 proteins (Fig. S5**A-C**; **Table S3**) and a Lasso-logistic model with 27 proteins (Fig. S3**B**; **Table S4**) based on the LUAD.Xu.Prot dataset following the same procedure as above (see **Supplementary Text**). PCLassoLog and Lasso-logistic yielded comparable AUCs on the LUAD.Xu.Prot−test dataset and the independent protein expression dataset LUAD.Gillette.Prot (Fig. S5**D**). Considering the significant positive correlation of protein-mRNA pairs between LUAD.Xu.Prot and LUAD.Xu.mRNA datasets (Fig. S4**D-F**; **Table S3**), we further evaluated the performance of PCLassoLog on the mRNA expression datasets. PCLassoLog obtained larger AUCs than those of the Lasso-logistic model on all the two matched mRNA expression datasets and three independent datasets (Fig. S5**E-F**). The classification accuracies showed the same trend (Fig. S5**G**).

3.2.3. HCC

We further constructed a PCLassoLog model with 18 protein complexes / 35 proteins (**Figure S6A-C**; **Table S5**) and a Lasso-logistic model with 24 proteins (Fig. S3**C**; **Table S6**) based on the HCC dataset (see **Supplementary Text**). Both PCLassoLog and Lasso-logistic models achieved perfect classification on the HCC−test dataset (**Figure S6D**). The AUCs of PCLassoLog on the five independent datasets were all larger than those of the Lasso-logistic model (**Figure S6D**), as well as the classification accuracies (**Figure S6E**), which further confirms the superior predictive performance of PCLassoLog.

3.3. PCLassoLog produces discriminative and robust features at the level of protein complexes

Next, we looked into the PCLassoLog and Lasso-logistic models, and tried to explore the reasons why PCLassoLog could outperform the Lasso-logistic model by comparing the features selected by the two models. For the two LUAD protein expression datasets, the two models share seven (**Table S1**) and eight (**Table S3**) proteins, respectively. The difference is that Lasso-logistic only selects individual proteins, while PCLassoLog selects all proteins that are present in the protein complex containing it. We focused on these shared proteins for comparison. For the "LUAD.Gillette.Prot" case, the seven shared proteins belong to six protein complexes. To compare the discriminative ability and robustness of features, we calculated the *t*-statistics of the six protein complexes (see **Supplementary Text** for the definition of the *t*-statistic of the protein complex) and their member proteins in all the eight datasets (Fig. 3**A**). The larger absolute value of *t*-statistic reflects stronger discriminative ability. Proteins that were consistently up-regulated (e.g., FHL2; unpaired two-sided *t*-test, *t*-statistic > 0, p < 0.05; points colored in red) or down-regulated (e.g., FHL5 and TLN1; unpaired two-sided *t*-test, *t*-statistic < 0, p < 0.05; points col-

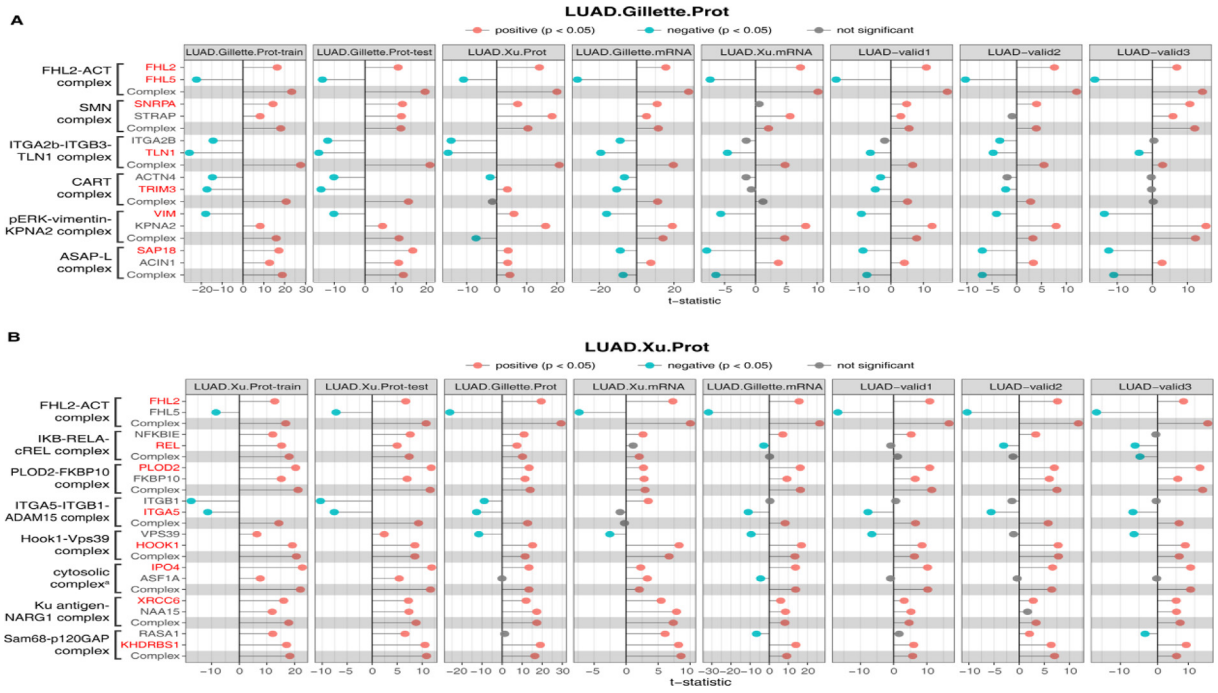

Fig. 3. Comparison of the protein complexes selected by PCLassoLog and the proteins selected by Lasso-logistic. (A) The "LUAD.Gillette.Prot" case. (B) The "LUAD.Xu.Prot" case. The proteins shown in red were selected by both models. The x-axis represents *t*-statistics. *P* values were calculated by unpaired two-sided Student's *t*-test. Red points indicate that the expression of these proteins is significantly up-regulated in tumor samples, while cyan points indicate significant down-regulation (BH adjusted *p* value < 0.05). Grey means no significant difference. For each protein complex, "Complex" represents a pseudo-protein constructed by a linear combination of proteins contained in this protein complex, where the coefficients of the linear combination are obtained from the PCLassoLog model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ored in cyan) across all datasets help the prediction models achieve robust predictive performance on independent datasets. In contrast, proteins with inconsistent expression levels across datasets (e.g., VIM) may reduce the robustness of prediction models, especially the Lasso-logistic model based on individual proteins.

The features at the protein complex level tend to be more discriminative. For example, for the "LUAD.Gillette.Prot" case, the *t*-statistic of the FHL2-ACT complex increased by 42.4 %, 81.3 %, 41.2 %, 78.3 %, 38.8 %, 59.0 %, 58.9 %, and 104.6 % in the eight datasets compared with the *t*-statistic of FHL2 (Fig. 3**A**). For the "LUAD.Xu.Prot" case, the *t*-statistic of the Ku antigen-NARG1 complex increased by 11.2 %, 21.0 %, 44.7 %, 34.8 %, 37.3 %, 46.6 %, 20.4 %, and 21.0 % in the eight datasets compared with the *t*-statistic of XRCC6 (Fig. 3**B**). Other examples include SMN complex vs SNRPA, ITGA2b-ITGB3-TLN1 complex vs TLN1 for the "LUAD.Gillette.Prot" case (Fig. 3**A**), and PLOD2-FKBP10 complex vs PLOD2 for the "LUAD.Xu.Prot" case (Fig. 3**B**). Furthermore, the features at the protein complex level tend to be more robust. For example, the SMN complex was consistently up-regulated across all eight datasets, but SNRPA did not show a significant up-regulation in the LUAD.Xu.mRNA dataset (Fig. 3**A**). These two trends may lead to the superiority of the PCLassoLog model based on protein complexes over the Lasso-logistic model based on individual proteins.

3.4. PCLassoLog produces better overall predictive performance

To show that the superior predictive performance of PCLassoLog does not depend on the specific division of the dataset, we performed 100 random divisions on the LUAD.Gillette.Prot, LUAD.Xu.Prot, and HCC datasets, respectively. For each division, we used the same strategy as the above experiments to construct classification models and evaluate predictive performance. We trained 100 models for each of these three datasets (**Figure S7**). Then the resulting 100 AUCs and 100 classification accuracies obtained on each dataset were used to evaluate the overall predictive performance.

For the "LUAD.Gillette.Prot" case, except for the LUAD.Gillette.Prot-test dataset, the AUCs and classification accuracies obtained by PCLassoLog on the other six datasets are significantly higher than those of the Lasso-logistic model (Fig. 4**A**). For the "LUAD.Xu.Prot" case, the predictive performance of PCLassoLog on the two protein expression datasets and two matched mRNA expression datasets is comparable to that of the Lasso-logistic model, while the AUCs and classification accuracies obtained on the three independent datasets are again significantly higher than those of the Lasso-logistic model (Fig. 4**B**). For the "HCC" case, except for the HCC-test dataset, the AUCs and classification accuracies of PCLassoLog on the five independent datasets are significantly higher than those of the Lasso-logistic model (Fig. 4**C**), further confirming the better prediction performance of PCLassoLog.

Next, we investigated the robustness of PCLassoLog in identifying risk protein complexes. From the 100 random divisions, PCLassoLog identified 69 and 98 risk protein complexes in the two datasets, respectively (**Table S7 and S8**). Of these, 29 (21 %) protein complexes were identified in both datasets (Fig. 4**D**). At the individual protein level, 53 (23 %) proteins were identified in both datasets. The overlap is almost twice that of the Lasso-Logistic model, where only 33 (12 %) proteins are shared between the two datasets (Fig. 4**D** and **Table S9**). This indicates that PCLassoLog has stronger robustness in identifying risk protein complexes than the traditional Lasso-logistic model in identifying risk proteins.

3.5. Comparison with other protein complex-based methods

We next compared PCLassoLog with PCSCAD and PCMCP, which were constructed following the same procedure as described above

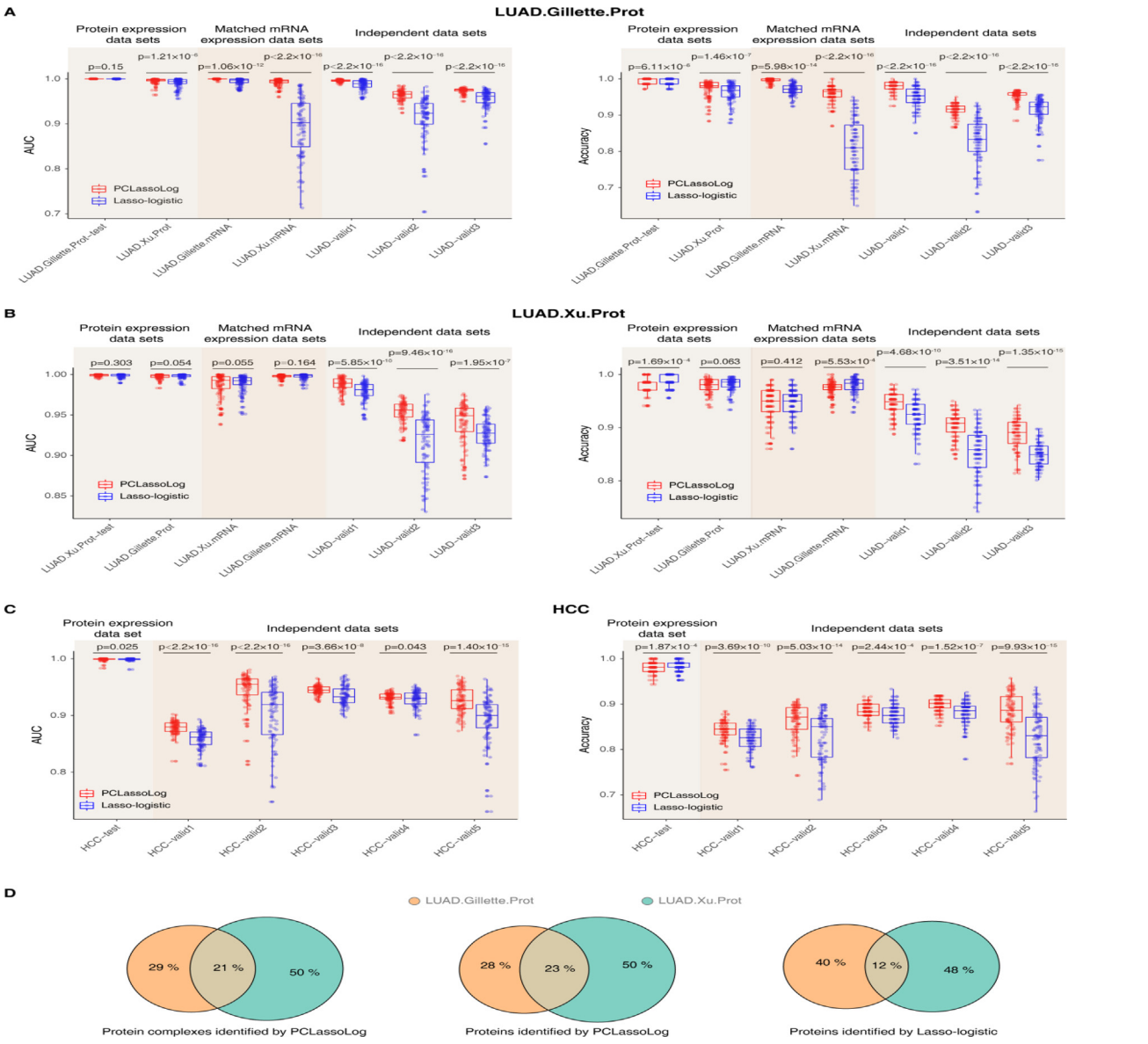

Fig. 4. Comparison of overall prediction performance between PCLassoLog and Lasso-logistic. (A-C) The prediction performance of PCLassoLog and Lasso-logistic models in the "LUAD.Gillette.Prot" case (A), the "LUAD.Xu.Prot" case (B), and the "HCC" case (C). Left: AUC; Right: classification accuracy. For each case, 100 models are constructed based on the training set. Each box is plotted based on 100 AUCs/accuracies, which are calculated using the prediction results of the 100 models. The middle bar represents the median, and the box represents the interquartile range; bars extend to 1.5 × the interquartile range. *P* values are calculated by the two-sided Wilcoxon rank-sum test. (D) Comparison of risk protein complexes and proteins identified in the LUAD.Gillette.Prot and LUAD.Xu.Prot datasets.

for evaluating the overall predictive performance of PCLassoLog (see **Supplementary Text**). The number of protein complexes selected by the PCMCP model is less than that of PCSCAD, and PCSCAD is less than that of PCLassoLog (**Figure S8A, S9A, S10A**), which is consistent with the successive reduction of penalties for large coefficients of these three models (**Figure S2**). PCLassoLog penalizes large coefficients the most, allowing more protein complexes to enter the model to minimize training errors. Indeed, when a is small, the average cross-validation errors of PCSCAD and PCMCP are greater than those of PCLassoLog in all the three datasets (**Figure S11**). As $a \to \infty$, the number of protein complexes included in the PCSCAD and PCMCP models and the cross-validation errors become close to those of PCLassoLog (**Figure S8A, S9A, S10A, S11**), which is consistent with their penalty function approaching that of PCLassoLog as $a \to \infty$. At the individual protein level, PCSCAD and PCMCP contain fewer proteins than Lasso-logistic when a is small, and more than Lasso-logistic model as $a \to \infty$ (**Figure S8B, S9B, S10B**).

The PCLassoLog, PCSCAD, PCMCP, and Lasso-logistic model obtained comparable performance on the LUAD.Gillette.Prot-test (**Figure S8C and S12A**), LUAD.Xu.Prot-test (**Figure S9C and S13A**), and HCC-test (**Figure S10C and S14A**) datasets, indicating that all four models could achieve good prediction performance on homogeneous datasets. On the two independent protein expression datasets, LUAD.Xu.Prot (**Figure S8D, S12B**) and LUAD.Gillette.Prot (**Figure S9D, S13B**), PCLassoLog is superior to PCSCAD and PCMCP. This advantage is further observed on almost all mRNA expression datasets, including the five datasets in the "LUAD.Gillette.Prot" case (**Figure S8E-I**, **S12C-G**), the five datasets in the "LUAD.Xu.Prot" case (**Figure S9E-I**, **S13C-G**), and the five datasets in the "HCC" case (**Figure S10D-H**, **S14B-F**). When a is large, the prediction performance of PCSCAD and PCMCP is close to that of

PCLassoLog, which is also consistent with the penalty functions of PCSCAD and PCMCP approaching that of PCLassoLog as $a \to \infty$. Compared with Lasso-logistic model, the prediction performance of PCSCAD and PCMCP is poor when a is small. However, when a is large, the prediction performance of the two models is better than that of Lasso-logistic model (**Figure S8G-I**, **S9D-I**, **S10D**, **S10G-H**, **S12E-G**, **S13B-G, S14B**, and **S14E-F**).

The penalty of PCSCAD and PCMCP for large coefficients increases with the increase of a, which is exemplified by the ITGA2b-ITGB3-TLN1 complex (**Figure S8J**). The SP of ITGA2B-ITGB3-TLN1 complex gradually decreased with the increase of a, and finally stabilized at the SP of the ITGA2b-ITGB3-TLN1 complex in the PCLassoLog model (SP = 0.54) (**Figure S8J**). PCSCAD had the same trend as PCMCP, except that the SP of ITGA2B-ITGB3-TLN1 complex decreased faster, which could be attributed to the penalty of PCSCAD being larger than that of PCMCP (**Figure S2**). This indicates that the penalty strategies of the models are effective.

To compare the robustness of the three protein complex-based models in identifying risk protein complexes, we calculated the Jaccard coefficients of the sets of risk protein complexes identified by the three models on the two LUAD datasets to compare the overlap between them. The Jaccard coefficients of both PCSCAD and PCMCP are smaller than that of PCLassoLog (0.210) (**Figure S8K**). The Jaccard coefficients of PCMCP are not stable, and are slightly smaller than those of PCSCAD for most a values (**Figure S8K**). At the individual protein level, the Jaccard coefficients of all three protein complex-based models are greater than that of the Lasso-logistic model (0.123) (**Figure S8L**), even though the number of risk proteins identified by PCSCAD and PCMCP is less than that of the Lasso-logistic model (**Figure S8B**). This further confirms the robustness of the protein complex-based models in identifying risk protein complexes.

In addition, to estimate the risk of false positive results, we permutated the class labels of the LUAD.Gillette.Prot-train, LUAD.Xu.Prot-train, and HCC-trian datasets, and reconstructed the three protein complex-based models. Results show that the AUCs and accuracies obtained by the three models on all datasets are around 0.5, which is not better than random guessing (**Figure S15**), indicating that the prediction results of the three models are reliable.

3.6. PCLassoLog identifies risk protein complexes

We next examined the top 10 risk protein complexes with the highest SP identified by PCLassoLog in the LUAD and HCC datasets (Table 2-3, **Table S10A-B**). The FHL2-ACT complex was identified as a risk protein complex with a SP of 1 in both the LUAD.Gillette.Prot and LUAD.Xu.Prot datasets (**Table S7-S8**). Both FHL2 and FHL5 in the FHL2-ACT complex belong to the FHL protein family of transcriptional cofactors. Another protein complex related to the FHL protein family, FHL2-FHL3 complex, also obtained a SP of 0.63 (Table 2). FHL proteins could interact with important regulators of cancer development and progression, such as Smad2, Smad3, and Smad4, and suppress tumor cell growth through a TGF-β–like signaling pathway [52]. Numerous studies have reported that FHL2 might act as a cell type specific oncoprotein or tumor suppressor in human cancers [53], such as gastrointestinal cancer [54], glioblastoma [55], and liver cancer [52]. FHL3 is down-regulated in liver and breast cancer patients and could suppress cancer cell growth in these two cancer types [52,56]. Strikingly, FHL2 and FHL3 are significantly up-regulated in all seven LUAD datasets (**Figure S16A-B**), while FHL5 is significantly down-regulated (**Figure S16C**), suggesting that FHL proteins and their synergy may play important roles in the development of LUAD.

Among these 10 risk protein complexes, most of the detected proteins have been reported to play important roles in various cancer types (Table 2). Some of these proteins are related to lung cancer. For example, in the MLC1-Na, K-ATPase-Kir4.1-AQP4-TRPV4-syntrophin complex, both ATP1B1 and AQP4 have been suggested as promising drug targets to combat non-small cell lung cancer [50,57,58]. Inhibiting ATP1B1 expression by siRNA or specific cardenolides treatment has been shown to result in markedly impaired proliferation and migration of lung cancer cells [57]. Inhibiting AQP4 expression could significantly reduce lung cancer cell migration [50]. Some proteins have been shown to be effective in inhibiting tumor proliferation, metastasis or promoting cell apoptosis by being targeted by drugs or miRNAs, such as GAPDH [59] and PDIA3 [60] in the HMGB1-HMGB2-HSC70-ERP60-GAPDH complex, and ITGB1 [61] and JAM2 [62] in the ITGA4-ITGB1-JAM2 complex. For HCC, the top 10 risk protein complexes also contain a large number of proteins related to the development of various cancer types (Table 3). Some of these proteins have been shown to be targets of miRNAs or drugs and exert inhibitory effects in HCC, such as MAT1A [63,64], MAT2B [65–67], and AKAP1 [68]. This indicates that the risk proteins identified by PCLassoLog have high credibility.

Among these risk proteins, a few proteins have been reported to form complexes with other proteins to play important functions in cancer. For example, the physical interaction between risk protein CAV1 and GLUT3 has been reported to be targeted by Atorvastatin to suppress tumor growth in non-small cell lung cancer [18]. MAT2B has been reported to cross talk with HuR and SIRT1 to affect the pro-apoptotic and growth-suppressive effects of liver cancer [67]. However, the role of interplay among the proteins contained in these risk protein complexes in cancer development remains largely unexplored. The risk protein complexes identified by PCLassoLog provide valuable targets for researchers to better study the synergistic effects of proteins at the protein complex level.

3.7. Classification and risk protein complexes identification in pan-cancer

In view of the superior classification performance of PCLassoLog on LUAD and HCC, as well as the ability to identify risk protein complexes, we applied PCLassoLog to the classification and identification of risk protein complexes for 10 additional cancer types whose protein expression has been recently quantified [23,25–28] (Table 1). PCLassoLog achieved median AUC of 1, 1, 1, 0.987, 1, 0.989, 1, 1, 1, and 0.982 on the ccRCC, COAD, EC, ESCC, GBM, GC, HNSCC, LSCC, OV, and PDAC datasets, respectively (Fig. 5**A and S17A-C**). The risk protein complexes and their selection probabilities of the 10 cancer types are provided in **Table S10**.

Twenty-one protein complexes were identified as risk protein complexes in at least five cancer types (Fig. 5**B**). Some of these protein complexes contain well-known cancer-related proteins, such as HSP90-related protein complexes and cell cycle kinase complexes (Fig. 5**C**). Six protein complexes are consistently up-regulated (highlighted in red) in at least five cancer types, such as PLOD2-FKBP10 complex, SRSF9-SRSF6 complex, and FHL2-FHL3 complex (Fig. 5**B-C**). Only one protein complex, the Angiogenin-PRI complex, is consistently down-regulated (highlighted in green) in at least 5 caner types (Fig. 5**B-C**). These protein complexes that are consistently dysregulated in multiple cancer types may be valuable targets for revealing the molecular mechanism of cancers.

In contrast, some protein complexes were identified only in one cancer type. We define a protein complex that has a SP>0.5 in one cancer type but is not identified in other cancer types as a risk protein complex specific to that cancer type. Eight protein complexes are specific to LUAD, such as CART complex and DNA repair complex (Fig. 5**D**). The Methionine adenosyltransferase alpha1 beta-

Table 2
Risk protein complexes identified by PCLassoLog in the LUAD datasets.

NO.	ID	Complex Name	Gene Symbol	SP[a]	Reference (PMID)[c]
1	3188	FHL2-ACT complex	**FHL2**[b]; **FHL5**	1	FHL2: 19139564; 17352216; 17383428; 18615633; 19,139,564
2	5385	GAIT complex	SYNCRIP; **GAPDH**; **EPRS**; RPL13A	1	GAPDH: 25859407; 33029490; 26541605; EPRS: 27612429; 21941282; 33,740,160
3	7265	MLC1-Na, K-ATPase-Kir4.1-AQP4-TRPV4-syntrophin complex	**ATP1B1**; **AQP4**; KCNJ10; MLC1; TRPV4; SNTG1	1	ATP1B1: 17471453; 20460749; AQP4: 21548930; 22372348; 27516192; 22105864; 22808259; 19,112,001
4	280	HMGB1-HMGB2-HSC70-ERP60-GAPDH complex	**GAPDH**; HMGB1; HSPA8; HMGB2; **PDIA3**	0.97	GAPDH: 25859407; 33029490; 26541605; PDIA3: 29228584; 20035634; 26125904; 28101228; 26,004,124
5	2422	ITGA4-ITGB1-JAM2 complex	**ITGB1**; ITGA4; **JAM2**	0.95	ITGB1: 23441154; 28656629; 26509963; 26766915; 26903137; JAM2: 29575013; 27588115; 26,782,073
6	2378	ITGA2b-ITGB3-TLN1 complex	ITGB3; **ITGA2B**; **TLN1**	0.94	ITGA2B: 31523198; 26198048; TLN1: 31068760; 23,722,670
7	1737	SF3b complex	SF3B1; **SF3B2**; SF3B3; SF3B4; PHF5A; DDX42; **SF3B5**; SF3B6	0.94	SF3B2: 31,431,456
8	1085	DNA repair complex NEIL2-PNK-Pol (beta) -LigIII (alpha)-XRCC1	**PDXK**; **POLB**; XRCC1; LIG3; NEIL2	0.9	PDXK: 22854025; 26387143; 32696745; POLB: 19330779; 25561897; 12,126,515
9	5465	IKB (epsilon)-RELA-cREL complex	**NFKBIE**; RELA; **REL**	0.89	NFKBIE: 27670424; REL: 10602468; 939994; 12,430,173
10	5862	CAV1-VDAC1-ESR1 complex	ESR1; **VDAC1**; **CAV1**	0.87	VDAC1: 23233904; 31364685; 22204343; 24781191; 29682501; 21315184; 23663973; 21297950; 27304056; CAV1: 15205342; 31534543; 15692148; 29080835; 30,604,627

Shown are the top 10 risk protein complexes with the highest SP. The full list of risk protein complexes is provided in **Table S10A**. [a]Selection probability; [b]Proteins shown in bold are detected in the LUAD datasets; [c]References related to the detected proteins in LUAD or other cancer types.

Table 3
Risk protein complexes identified by PCLassoLog in the HCC dataset.

NO.	ID	Complex Name	Gene Symbol	SP[a]	Reference[c]
1	2440	ITGA9-ITGB1-ADAM9 complex	**ITGB1**[b]; ADAM9; **ITGA9**	1	ITGB1: 23441154; 28656629; 26509963; 26766915; 26903137; ITGA9: 31008533; 31489579; 26,596,831
2	7197	Methionine adenosyltransferase alpha1 beta-v1	**MAT1A**; **MAT2B**	1	MAT1A: 23665184; 32080887; 22318685; 31496615; 23241961; 24212770; MAT2B: 31493275; 31073374; 23814050; 18,698,677
3	7131	COMMD1-CCDC22-CCDC93-C16orf62 complex	**CCDC22**; **CCDC93**; C16orf62; COMMD1	0.95	—
4	142	CD147-gamma-secretase complex	**BSG**; PSEN1; NCSTN; **APH1A**; PSENEN	0.85	BSG: 31497203; 24264599; APH1A: 30,944,650
5	6998	ST3GAL6-EGFR complex	**EGFR**; **ST3GAL6**	0.81	EGFR: 24212818; 24318021; 25173978; ST3GAL6: 32929335; 25,061,176
6	3525	Tetrameric COG subcomplex	**COG7**; COG8; **COG5**; COG6	0.77	—
7	519	AMY-1-S-AKAP84-RII-beta complex	**PRKAR2B**; **AKAP1**; MYCBP	0.75	PRKAR2B: 29761841; 28008150; AKAP1: 28569781; 33193848; 33,868,472
8	7284	GSK3B-HSP90AA1-PKM2 complex	**HSP90AA1**; PKM; **GSK3B**	0.69	HSP90AA1: 20651736; 17513464; 34226297; 31567483; 30471108; GSK3B: 30144430; 32698955; 30,845,991
9	4158	HSP90-FKBP38-CAM-Ca (2 +) complex	**HSP90AA1**; CALM1; CAL; **FKBP8**	0.68	HSP90AA1: 20651736; 17513464; 34226297; 31567483; 30471108; FKBP8: 30,348,988
10	6782	mitochondrial permeability transition pore (PTP) complex (PPIF-SPG7-VDAC1)	**VDAC1**; **PPIF**; SPG7	0.67	23233904; 31364685; 22204343; 24781191; 29682501; 21315184; 23663973; 21297950; 27304056; PPIF: 33,495,413

Shown are the top 10 risk protein complexes with the highest SP. The full list of risk protein complexes is provided in **Table S10B**. [a]Selection probability; [b]Proteins shown in bold are detected in the HCC datasets; [c]References related to the detected proteins in HCC or other cancer types.

v1 complex is specific to HCC. Its member proteins MAT1A and MAT2B have been used as therapeutic targets for HCC drug development [69,70]. Some of the proteins in these protein complexes have been suggested as biomarkers or therapeutic targets for the corresponding cancer type, such as PIK3C3 (Phosphatidylinositol 3-kinase complex) in GBM [71], and TGFBI (BP-SMAD complex) in GC [72]. These cancer-specific protein complexes deserve further investigation.

3.8. Comparison with survival-related risk protein complexes

We next compared the risk protein complexes identified by PCLassoLog and the survival-related risk protein complexes, which were identified by PCLasso [19]. The main difference is that PCLassoLog uses class labels as dependent variables, while PCLasso uses survival outcomes. In other words, PCLassoLog identifies risk protein complexes on the premise of maximizing classification accuracy, while PCLasso is premised on maximizing prognostic accuracy. Therefore, the risk protein complexes identified by PCLassoLog may be mainly related to the initiation and development of cancer, while the risk protein complexes identified by PCLasso are more likely to be related to cancer progression.

A total of 10 cancer types were analyzed by both PCLassoLog and PCLasso. Only a few protein complexes are identified by both PCLassoLog and PCLasso in LUAD, HCC, ccRCC, EC, GC, HNSCC, and LSCC, respectively (**Table S11**). The risk protein complexes identified by the two models are very different in multiple cancer types, implying that the molecular mechanisms may also change at different stages of cancer development and progression. The common risk protein complexes identified by PCLasso and PCLassoLog may play a role in both cancer development and cancer progression. For example, the PLOD2-FKBP10 complex was identified by both PCLassoLog and PCLasso in LUAD (**Table S11**). PLOD2 has been shown to be elevated in NSCLC cells and related to poor prognosis

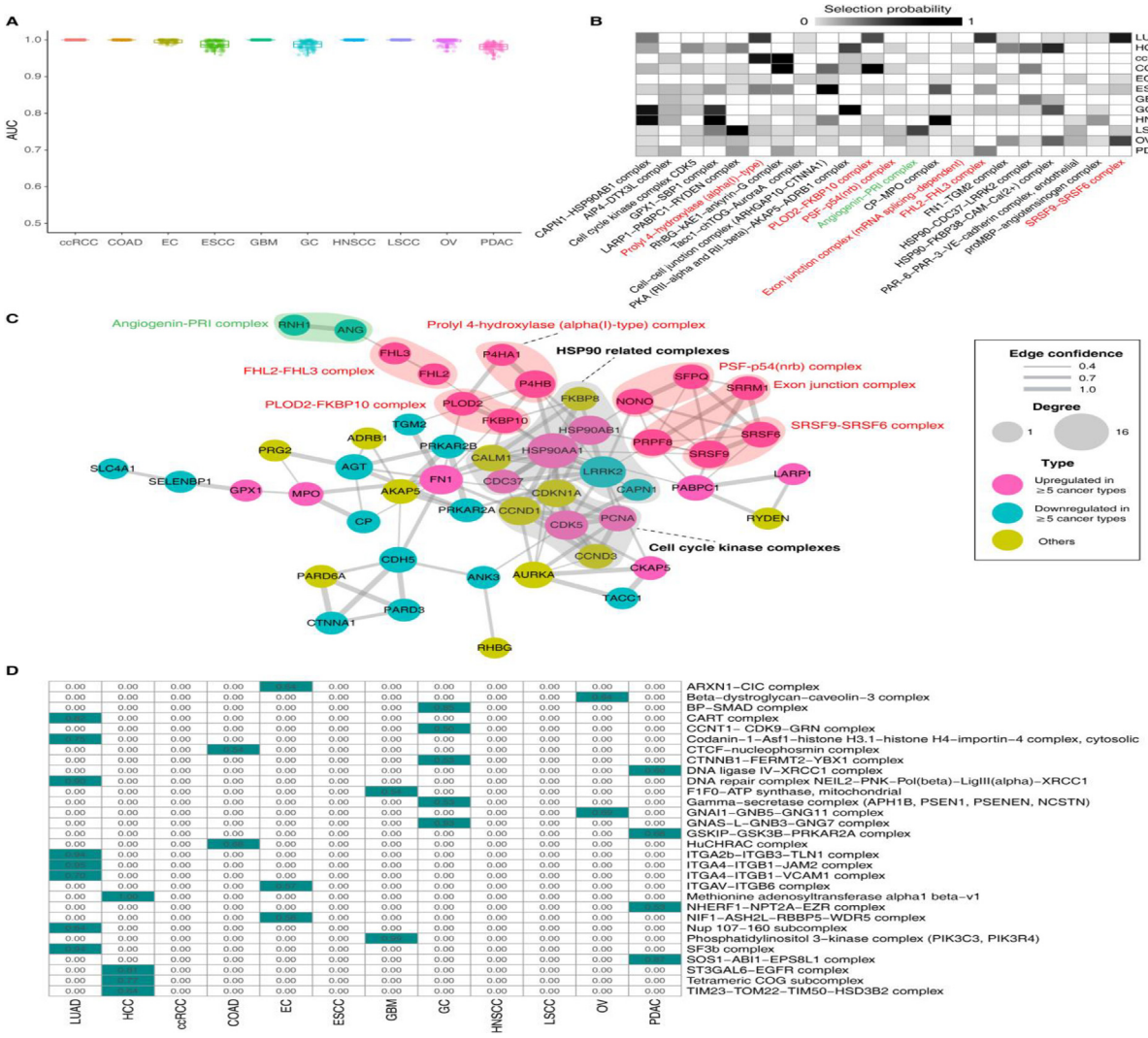

Fig. 5. Pan-cancer analysis of risk protein complexes identified by PCLassoLog. (A) Box plots of AUCs of the PCLassoLog model on the protein expression data of 10 cancer types. For each cancer type, the entire dataset is randomly divided into a training set (2/3) and a test set (1/3). The training set is used to train the model, and the test set is used to evaluate the model. This process is repeated 100 times. Each boxplot represents 100 AUCs obtained on 100 test sets. (B) Risk protein complexes identified by the PCLassoLog model across multiple cancer types. Each row of the matrix represents a type of cancer, and each column represents a protein complex. The color of the (i, j) element of the matrix indicates the SP of the j-th protein complex identified as a risk protein complex of the i-th cancer type. The protein complex shown in red indicates that it is consistently up-regulated in the corresponding cancer types, while green indicates that it is consistently down-regulated. (C) The protein–protein network of risk protein complexes identified in multiple cancer types. The protein–protein interactions were obtained from the STRING database. The width of the line indicates the edge confidence. Pink nodes indicate that the proteins are consistently up-regulated in at least five cancer types, and cyan nodes indicate that the proteins are consistently down-regulated in at least five cancer types. The six consistently up-regulated protein complexes and one consistently down-regulated protein complexes marked in (B) are highlighted in red and green, respectively. HSP90 related complexes and cell cycle kinase complexes are highlighted in grey. (D) Risk protein complexes specific to each cancer type. The values in the rectangular boxes indicate selection probabilities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in LUAD [73]. It promotes NSCLC metastasis directly by enhancing migration and indirectly by inducing collagen reorganization [73]. FKBP10 expression is also negatively correlated with the survival of lung cancer patients. FKBP10 downregulation has been shown to suppress lung cancer growth and cancer stem-like features [74]. Two integrin complexes (ITGAV-ITGB1 complex and ITGB1-NRP1 complex) were identified by both PCLassoLog and PCLasso in GC (**Table S11**). The integrin family regulates a diverse array of cellular functions crucial to the initiation, progression, and metastasis of solid tumours [75]. ITGAV has been shown to be overexpressed in GC and associated with shorter overall survival and disease-free survival. Downregulation of ITGAV could result in suppression of proliferation, migration, and invasion of GC cells [76]. ITGB1 has been shown to be targeted by miR-29c to mediate GC initiation [77] and targeted by miR-29a to mediate GC metastasis [78]. The importance of integrins that affect nearly every step of cancer progression from primary tumour development to metastasis has made them an appealing target for cancer therapy [75,79]. The risk protein complexes identified by PCLassoLog and PCLasso are both important, and they could complement each other to reveal the molecular mechanism of different stages of cancer from a more comprehensive perspective.

3.9. Risk protein complexes associated with genomic mutations in LUAD

In addition to risk protein complexes related to cancer development, PCLassoLog can also be used to identify risk protein complexes in other contexts, such as risk protein complex related to genomic mutations. Based on protein expression data from the

studies of Gillette et al. [20], we applied PCLassoLog to classification of LUAD patients with/without TP53, KRAS, STK11, EGFR, KEAP1, RB1, BRAF mutation, and ALK fusion, and obtained median AUC of 0.844, 0.675, 0.862, 0.780, 0.947, 0.686, 0.843, and 0.985, respectively (**Figure S18**). This indicates that protein complexes could accurately predict some mutant phenotypes in LUAD, especially TP53, STK11, KEAP1 mutation, and ALK fusion. Risk protein complexes related to each mutant phenotype were identified (**Table S12**). Among them, risk protein complexes with the largest SP include the sulphiredoxin-peroxiredoxin complex (SP = 1) that related to KEAP1 mutation, the DDX11-Ctf18-RFC complex (SP = 0.99) and MDC1-H2AFX-TP53BP1 complex (SP = 0.92) that related to TP53 mutation, the RhBG-kAE1-ankyrin-G complex (SP = 0.83) that related to ALK fusion.

4. Discussion

In this study, we propose the PCLassoLog model and verify its high classification accuracy and ability to identify reliable risk protein complexes. Using PCLassoLog, we identified risk protein complexes associated with cancer development in 12 cancer types and with gene mutations in LUAD. Pan-cancer analysis revealed risk protein complexes that play important functions in multiple cancer types and risk protein complexes specific to each cancer type.

PCLassoLog embeds protein complexes into the group Lasso-logistic model, and uses mixed l_1 and l_2 penalties to achieve feature selection. At the individual protein level of each protein complex, the l_2 norm fits the regression coefficients to obtain the optimal linear combination to accumulate the weaker discriminant ability of individual proteins and form pseudo-proteins at the protein complex level with strong discriminant ability. The l_1 penalty tends to choose only a few nonzero coefficients. Thus, at the protein complex level, the l_1 norm ensures selection of a few important protein complexes, which are referred to as risk protein complexes. By introducing latent variables, PCLassoLog can elegantly handle the overlap problem of protein complexes. Although many proteins are subunits of more than one protein complex, PCLassoLog produces a sparse solution, which matches the way proteins function in different protein complexes. For a protein belonging to multiple protein complexes, PCLassoLog independently estimates its regression coefficient in different protein complexes. The coefficients of this protein in the unselected protein complexes will be zero. In contrast, the coefficients of this protein in the selected protein complexes will be nonzero and will contribute to the final model. This is consistent with that a protein may belong to multiple protein complexes, but only some of them are related to the development of cancer.

Compared with the Lasso-logistic model based on individual proteins, PCLassoLog selects features with stronger reproducibility at the level of protein complexes across different datasets (Fig. 4**D**). In terms of classification accuracy, although the prediction performance of the Lasso-logistic model on the test sets is comparable to that of PCLassoLog, its prediction performance on the independent datasets decreases dramatically (Fig. 4**A-C**). This may be due to the unstable expression of individual proteins, which is usually caused by tumor heterogeneity or batch effects [19,80]. Protein complexes have been reported to be inherently resistant to batch effects and have the advantage of being more robust than individual proteins as features at the functional level [19,81]. Thus, it is not surprising that PCLassoLog can achieve more robust prediction performance on independent datasets. Both models use cross-validation to select the optimal λ. Accordingly, the Lasso-logistic model included less proteins than PCLassoLog (**Figure S7**). We also investigated a variant of Lasso-logistic model (Lasso-logistic2) which selects the same number of features as PCLassoLog by controlling λ (**Figure S19A**). The prediction performance of the Lasso-logistic2 model is comparable to that of the Lasso-logistic model on the test set, but poor on the independent data sets (**Figure S19B-C**). This further indicates that PCLassoLog has a more robust and accurate performance when the number of protein features is comparable. Compared with the Lasso-logistic model with elastic net penalty (ENet, $\alpha = 0.5$), which also uses both l_1 and l_2 penalties, the predictive performance of PCLassoLog was better in the "LUAD.Gillette.Prot" case and slightly worse in the "LUAD.Xu.Prot" and "HCC" case. However, PCLassoLog has the characteristic of identifying risk protein complexes that ENet does not have. In addition, although PCLassoLog is a linear model, the overall predictive performance is better than that of nonlinear models such as random forest and extreme gradient boosting (see **Supplementary Text** for details) in the three cases (**Figure S20**).

Compared with PCSCAD and PCMCP, PCLassoLog has the highest prediction accuracy on independent datasets (**Figure S8-10**), which could be attributed to it recruiting more protein complexes into the model in order to compensate for its over-shrinkage of large coefficients. However, this may also cause PCLassoLog to include some irrelevant features into the model. Therefore, we calculated the selection probability of each risk protein complex to measure its reliability. In general, it is reliable enough to select risk protein complexes with a probability>0.5 [46,47]. In addition, PCSCAD and PCMCP are implemented to complement PCLassoLog in identifying reliable risk protein complexes. PCSCAD and PCMCP effectively capture important features by reducing the penalty for large coefficients, but at the expense of certain prediction accuracy. The three models could be used in combination to better complete the task of accurate classification and identifying risk protein complexes. Nonetheless, because there is no gold standard for risk protein complexes at present, we cannot use a precision and recall framework to evaluate the reliability of the risk protein complexes discovered, which will be an issue to be further resolved in the future.

Similar to the PCLasso model, the most important contribution of PCLassoLog is the ability to identify risk protein complexes, which may provide valuable targets for researchers to study the synergistic effects of proteins on cancer development. The difference is that PCLasso identifies risk protein complexes associated with cancer progression, whereas PCLassoLog identifies risk protein complexes associated with cancer initiation and development. Moreover, PCLassoLog can be used to identify risk protein complexes in other contexts, such as those associated with genomic mutation (**Table S12**). With risk protein complexes, researchers can discover and understand protein functions and carcinogenic mechanisms from a more comprehensive perspective. Further, new therapeutic strategies could be developed from the perspective of the complex, such as simultaneously inhibiting the expression of proteins in the complex, or blocking their binding, etc. PCLassoLog provides researchers with an effective tool to discover unknown risk protein complexes. Note that one limit of PCLassoLog is that the risk protein complexes identified by PCLassoLog are derived from the known protein complex database CORUM. Since changes in protein interactions during cancer progression may form new protein complexes that play key carcinogenic roles, how to identify such dynamic risk protein complexes is an issue worthy of further research.

We have developed an R package 'PCLassoReg' (https://CRAN.R-project.org/package=PCLassoReg and https://github.com/weiliu123/PCLassoReg), a freely available implementation of all protein complex-based models, including PCLassoLog, PCSCAD, PCMCP, and PCLasso. In addition to the risk protein complexes discovered in this study (**Table S10 and S12**), researchers can use our package to identify risk protein complexes for other cancer types or in other contexts based on their own data.

5. Availability

PCLassoLog is an R package freely available in CRAN (https://CRAN.R-project.org/package=PCLassoReg) and the GitHub repository (https://github.com/weiliu123/PCLassoReg).

Funding

This work was supported by the National Natural Science Foundation of China [82272945; 11901170], the Natural Science Foundation of Heilongjiang Province [LH2021F048], the National Fund cultivation special project [2021GJ09], and the Provincial Echelon Training Program of Heilongjiang Institute of Technology [2020LJ01].

CRediT authorship contribution statement

Wei Wang: Methodology, Software, Writing – original draft. **Haiyan Yuan:** Visualization, Investigation. **Junwei Han:** Conceptualization. **Wei Liu:** Conceptualization, Methodology, Software, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The results shown here are part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.12.005.

References

[1] López-Bigas N, Ouzounis CA. Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res 2004;32:3108–14.
[2] Köhler S, Bauer S, Horn D, et al. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 2008;82:949–58.
[3] Smedley D, Köhler S, Czeschik JC, et al. Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. Bioinformatics 2014;30:3215–22.
[4] Gambin T, Yuan B, Bi W, et al. Identification of novel candidate disease genes from de novo exonic copy number variants. Genome Med 2017;9:83.
[5] Ang JC, Mirzal A, Haron H, et al. Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection. IEEE/ACM Trans Comput Biol Bioinf 2016;13:971–89.
[6] Lazar C, Taminau J, Meganck S, et al. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans Comput Biol Bioinform 2012;9:1106–19.
[7] Giurgiu M, Reinhard J, Brauner B, et al. CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res 2019;47:D559–63.
[8] Güldener U, Münsterkötter M, Kastenmüller G, et al. CYGD: the Comprehensive Yeast Genome Database. Nucleic Acids Res 2005;33:D364–8.
[9] Berggård T, Linse S, James P. Methods for the detection and analysis of protein-protein interactions. Proteomics 2007;7:2833–42.
[10] Ruepp A, Brauner B, Dunger-Kaltenbach I, et al. CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res 2008;36:D646–50.
[11] Zheng J, Chen X, Yang Y, et al. Mass Spectrometry-Based Protein Complex Profiling in Time and Space. Anal Chem 2021;93:598–619.
[12] Bludau I, Aebersold R. Proteomic and interactomic insights into the molecular basis of cell functional diversity. Nat Rev Mol Cell Biol 2020;21:327–40.
[13] Gordon DE, Jang GM, Bouhaddou M, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 2020;583:459–68.
[14] Nollet F, Berx G, van Roy F. The role of the E-cadherin/catenin adhesion complex in the development and progression of cancer. Mol Cell Biol Res Commun 1999;2:77–85.
[15] Fu J, Qin L, He T, et al. The TWIST/Mi2/NuRD protein complex and its essential role in cancer metastasis. Cell Res 2011;21:275–89.
[16] Lepourcelet M, Chen YN, France DS, et al. Small-molecule antagonists of the oncogenic Tcf/beta-catenin protein complex. Cancer Cell 2004;5:91–102.
[17] Luo Z, Liu W, Sun P et al. Pan-cancer analyses reveal regulation and clinical outcome association of the shelterin complex in cancer, *Brief Bioinform* 2021;22:bbaa441.
[18] Ali A, Levantini E, Fhu CW, et al. CAV1 - GLUT3 signaling is important for cellular energy and can be targeted by Atorvastatin in Non-Small Cell Lung Cancer. Theranostics 2019;9:6157–74.
[19] Wang W, Liu W. PCLasso: a protein complex-based, group lasso-Cox model for accurate prognosis and risk protein complex discovery, *Brief Bioinform* 2021;22:bbab212.
[20] Gillette MA, Satpathy S, Cao S, et al. Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. Cell 2020;182:200–225.e235.
[21] Xu JY, Zhang C, Wang X, et al. Integrative Proteomic Characterization of Human Lung Adenocarcinoma. Cell 2020;182:245–261.e217.
[22] Gao Q, Zhu H, Dong L, et al. Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. Cell 2019;179:561–577.e522.
[23] Clark DJ, Dhanasekaran SM, Petralia F, et al. Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. Cell 2019;179:964–983.e931.
[24] Vasaikar S, Huang C, Wang X, et al. Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. Cell 2019;177:1035–1049.e1019.
[25] Dou Y, Kawaler EA, Cui Zhou D, et al. Proteogenomic Characterization of Endometrial Carcinoma. Cell 2020;180:729–748.e726.
[26] Liu W, Xie L, He YH, et al. Large-scale and high-resolution mass spectrometry-based proteomics profiling defines molecular subtypes of esophageal cancer for therapeutic targeting. Nat Commun 2021;12:4961.
[27] Wang LB, Karpova A, Gritsenko MA, et al. Proteogenomic and metabolomic characterization of human glioblastoma. Cancer Cell 2021;39:509–528.e520.
[28] Ge S, Xia X, Ding C, et al. A proteomic landscape of diffuse-type gastric cancer. Nat Commun 2018;9:1012.
[29] Huang C, Chen L, Savage SR, et al. Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. Cancer Cell 2021;39:361–379.e316.
[30] Satpathy S, Krug K, Jean Beltran PM, et al. A proteogenomic portrait of lung squamous cell carcinoma. Cell 2021;184:4348–4371.e4340.
[31] Hu Y, Pan J, Shah P, et al. Integrated Proteomic and Glycoproteomic Characterization of Human High-Grade Serous Ovarian Carcinoma. Cell Rep 2020;33:108276.
[32] Cao L, Huang C, Cui Zhou D, et al. Proteogenomic characterization of pancreatic ductal adenocarcinoma. Cell 2021;184:5031–5052.e5026.
[33] Landi MT, Dracheva T, Rotunno M, et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. PLoS One 2008;3:e1651.
[34] Lu TP, Tsai MH, Lee JM, et al. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. Cancer Epidemiol Biomarkers Prev 2010;19:2590–7.
[35] Hou J, Aerts J, den Hamer B, et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PLoS One 2010;5:e10312.
[36] Villa E, Critelli R, Lei B, et al. Neoangiogenesis-related genes are hallmarks of fast-growing hepatocellular carcinomas and worst survival. Results from a prospective study. Gut 2016;65:861–9.
[37] Grinchuk OV, Yenamandra SP, Iyer R, et al. Tumor-adjacent tissue co-expression profile analysis reveals pro-oncogenic ribosomal gene signature for prognosis of resectable hepatocellular carcinoma. Mol Oncol 2018;12:89–113.
[38] Makowska Z, Boldanova T, Adametz D, et al. Gene expression analysis of biopsy samples reveals critical limitations of transcriptome-based molecular classifications of hepatocellular carcinoma. J Pathol Clin Res 2016;2:80–92.
[39] Roessler S, Jia HL, Budhu A, et al. A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. Cancer Res 2010;70:10202–12.
[40] Ming Y, Yi L. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B (Stat Methodol) 2006;68:49–67.
[41] Jacob L, Obozinski G, Vert JP. Group lasso with overlap and graph lasso. In International Conference on Machine Learning. 2009.
[42] Park H, Niida A, Miyano S, et al. Sparse overlapping group lasso for integrative multi-omics analysis. J Comput Biol 2015;22:73–84.
[43] Fan JQ, Li RZ. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Publ Am Stat Assoc 2001;96:1348–60.
[44] Zhang CH. Nearly unbiased variable selection under minimax concave penalty. Ann Stat 2010;38:894–942.
[45] Alexander DH, Lange K. Stability selection for genome-wide association. Genet Epidemiol 2011;35:722–8.
[46] Sun H, Wang S. Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. Bioinformatics 2012;28:1368–75.
[47] Meinshausen N, Bühlmann P. Stability selection. J R Stat Soc 2010;72:417–73.

[48] Nikou S, Arbi M, Dimitrakopoulos F-I-D, et al. Integrin-linked kinase (ILK) regulates KRAS, IPP complex and Ras suppressor-1 (RSU1) promoting lung adenocarcinoma progression and poor survival. J Mol Histol 2020;51:385–400.
[49] Huang AH, Pan SH, Chang WH, et al. PARVA promotes metastasis by modulating ILK signalling pathway in lung adenocarcinoma. PLoS One 2015;10:e0118530.
[50] Xie Y, Wen X, Jiang Z, et al. Aquaporin 1 and aquaporin 4 are involved in invasion of lung cancer cells. Clin Lab 2012;58:75–80.
[51] Regala RP, Weems C, Jamieson L, et al. Atypical protein kinase Ciota plays a critical role in human lung cancer cell growth and tumorigenicity. J Biol Chem 2005;280:31109–15.
[52] Ding L, Wang Z, Yan J, et al. Human four-and-a-half LIM family members suppress tumor cell growth through a TGF-beta-like signaling pathway. J Clin Invest 2009;119:349–61.
[53] Kleiber K, Strebhardt K, Martin BT. The biological relevance of FHL2 in tumour cells and its role as a putative cancer target. Anticancer Res 2007;27:55–61.
[54] Wang J, Yang Y, Xia HH, et al. Suppression of FHL2 expression induces cell differentiation and inhibits gastric and colon carcinogenesis. Gastroenterology 2007;132:1066–76.
[55] Li M, Wang J, Ng SS, et al. The four-and-a-half-LIM protein 2 (FHL2) is overexpressed in gliomas and associated with oncogenic activities. Glia 2008;56:1328–38.
[56] Niu C, Yan Z, Cheng L, et al. Downregulation and antiproliferative role of FHL3 in breast cancer. IUBMB Life 2011;63:764–71.
[57] Mijatovic T, Roland I, Van Quaquebeke E, et al. The alpha1 subunit of the sodium pump could represent a novel target to combat non-small cell lung cancers. J Pathol 2007;212:170–9.
[58] Warth A, Muley T, Meister M, et al. Loss of aquaporin-4 expression and putative function in non-small cell lung cancer. BMC Cancer 2011;11:161.
[59] Yun J, Mullarky E, Lu C, et al. Vitamin C selectively kills KRAS and BRAF mutant colorectal cancer cells by targeting GAPDH. Science 2015;350:1391–6.
[60] Zhao S, Wen Z, Liu S, et al. MicroRNA-148a inhibits the proliferation and promotes the paclitaxel-induced apoptosis of ovarian cancer cells by targeting PDIA3. Mol Med Rep 2015;12:3923–9.
[61] Laudato S, Patil N, Abba ML et al. P53-induced miR-30e-5p inhibits colorectal cancer invasion and metastasis by targeting ITGA6 and ITGB1 2017;141:1879-1890.
[62] Li GC, Cao XY, Li YN et al. MicroRNA-374b inhibits cervical cancer cell proliferation and induces apoptosis through the p38/ERK signaling pathway by binding to JAM-2 2018;233:7379-7390.
[63] Yang H, Cho ME, Li TW, et al. MicroRNAs regulate methionine adenosyltransferase 1A expression in hepatocellular carcinoma. J Clin Invest 2013;123:285–98.
[64] Li Y, Lu L, Tu J, et al. Reciprocal Regulation Between Forkhead Box M1/NF-κB and Methionine Adenosyltransferase 1A Drives Liver Cancer. Hepatology 2020;72:1682–700.
[65] Wu L, Chen P, Ying J, et al. MAT2B mediates invasion and metastasis by regulating EGFR signaling pathway in hepatocellular carcinoma. Clin Exp Med 2019;19:535–46.
[66] Simile MM, Peitta G, Tomasi ML, et al. MicroRNA-203 impacts on the growth, aggressiveness and prognosis of hepatocellular carcinoma by targeting MAT2A and MAT2B genes. Oncotarget 2019;10:2835–54.
[67] Yang H, Zheng Y, Li TW, et al. Methionine adenosyltransferase 2B, HuR, and sirtuin 1 protein cross-talk impacts on the effect of resveratrol on apoptosis and growth in liver cancer cells. J Biol Chem 2013;288:23161–70.
[68] Du Q, Han J, Gao S, et al. Hypoxia-induced circular RNA hsa_circ_0008450 accelerates hepatocellular cancer progression via the miR-431/AKAP1 axis. Oncol Lett 2020;20:388.
[69] Ramani K, Mato JM, Lu SC. Role of methionine adenosyltransferase genes in hepatocarcinogenesis. Cancers (Basel) 2011;3:1480–97.
[70] Frau M, Feo F, Pascale RM. Pleiotropic effects of methionine adenosyltransferases deregulation as determinants of liver cancer progression and prognosis. J Hepatol 2013;59:830–41.
[71] Xia Q, Xu M, Zhang P, et al. Therapeutic Potential of Autophagy in Glioblastoma Treatment With Phosphoinositide 3-Kinase/Protein Kinase B/Mammalian Target of Rapamycin Signaling Pathway Inhibitors. Front Oncol 2020;10:572904.
[72] Suzuki M, Yokobori T, Gombodorj N, et al. High stromal transforming growth factor β-induced expression is a novel marker of progression and poor prognosis in gastric cancer. J Surg Oncol 2018;118:966–74.
[73] Du H, Chen Y, Hou X, et al. PLOD2 regulated by transcription factor FOXA1 promotes metastasis in NSCLC. Cell Death Dis 2017;8:e3143.
[74] Ramadori G, Ioris RM, Villanyi Z, et al. FKBP10 Regulates Protein Translation to Sustain Lung Cancer Growth. Cell Rep 2020;30:3851–3863.e3856.
[75] Desgrosellier JS, Cheresh DA. Integrins in cancer: biological implications and therapeutic opportunities. Nat Rev Cancer 2010;10:9–22.
[76] Wang H, Chen H, Jiang Z, et al. Integrin subunit alpha V promotes growth, migration, and invasion of gastric cancer cells. Pathol Res Pract 2019;215:152531.
[77] Han TS, Hur K, Xu G, et al. MicroRNA-29c mediates initiation of gastric carcinogenesis by directly targeting ITGB1. Gut 2015;64:203–14.
[78] He B, Xiao YF, Tang B, et al. hTERT mediates gastric cancer metastasis partially through the indirect targeting of ITGB1 by microRNA-29a. Sci Rep 2016;6:21955.
[79] Hamidi H, Ivaska J. Every step of the way: integrins in cancer progression and metastasis. Nat Rev Cancer 2018;18:533–48.
[80] Liu W, Li C, Xu Y, et al. Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. Bioinformatics 2013;29:2169–77.
[81] Goh WW, Wong L. Protein complex-based analysis is resistant to the obfuscating consequences of batch effects –- a case study in clinical proteomics. BMC Genomics 2017;18:142.