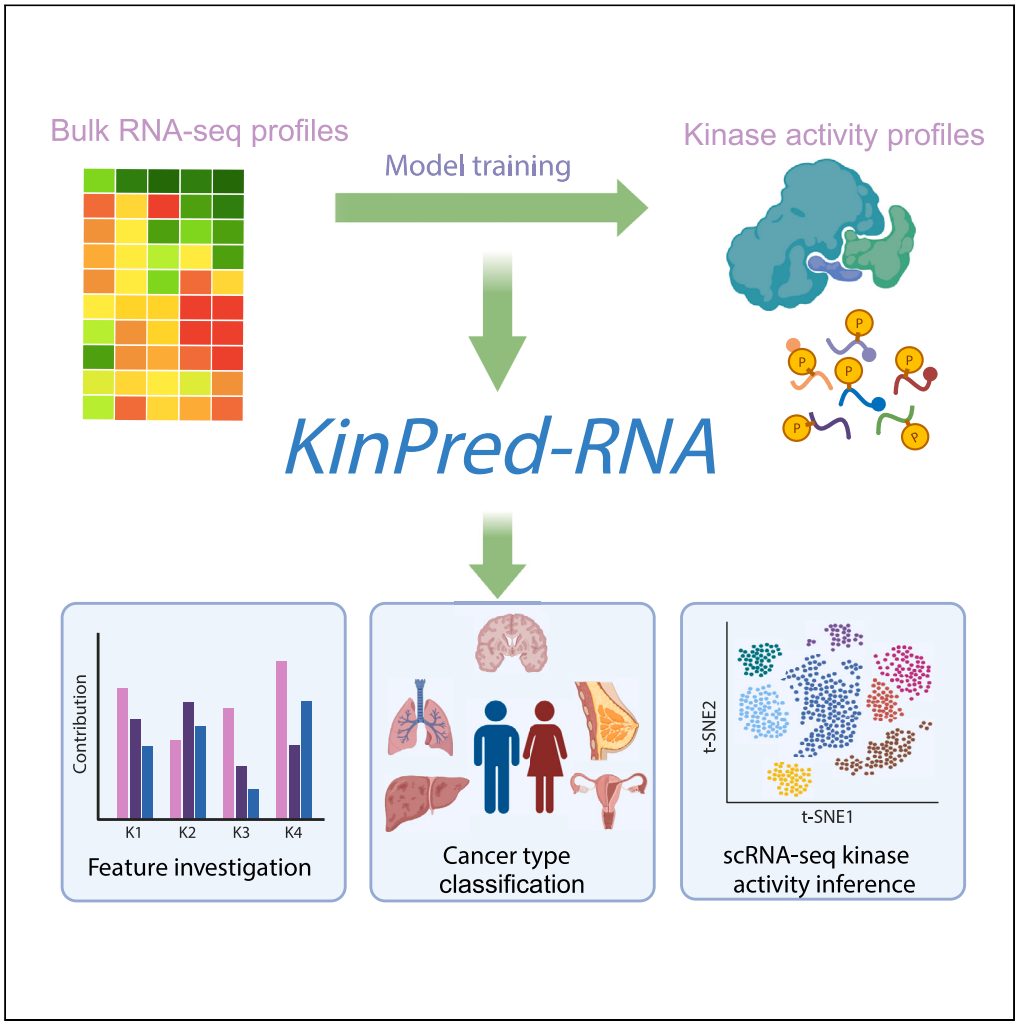
iScience

CellPress
OPEN ACCESS

Article

KinPred-RNA—kinase activity inference and cancer type classification using machine learning on RNA-seq data



Yuntian Zhang, Lantian Yao, Chia-Ru Chung, ..., Wenyang Zhang, Yuxuan Pang, Tzong-Yi Lee

leetzongyi@nycu.edu.tw

Highlights

Provide a cost-effective tool for predicting kinase activities from bulk RNA-seq data

Outperform other regression models in terms of accuracy and robustness

Suitable to analysis of kinase activities across multiple cancer types

Link kinase activity to biological functions and advance cancer identification

Zhang et al., iScience 27, 109333
April 19, 2024
https://doi.org/10.1016/j.isci.2024.109333

iScience

CellPress
OPEN ACCESS

Article

KinPred-RNA—kinase activity inference and cancer type classification using machine learning on RNA-seq data

Yuntian Zhang,[1,2,9] Lantian Yao,[3,4,9] Chia-Ru Chung,[5] Yixian Huang,[1,2] Shangfu Li,[1] Wenyang Zhang,[2] Yuxuan Pang,[6] and Tzong-Yi Lee[7,8,10,*]

SUMMARY

Kinases as important enzymes can transfer phosphate groups from high-energy and phosphate-donating molecules to specific substrates and play essential roles in various cellular processes. Existing algorithms for kinase activity from phosphorylated proteomics data are often costly, requiring valuable samples. Moreover, methods to extract kinase activities from bulk RNA sequencing data remain undeveloped. In this study, we propose a computational framework KinPred-RNA to derive kinase activities from bulk RNA-sequencing data in cancer samples. KinPred-RNA framework, using the extreme gradient boosting (XGBoost) regression model, outperforms random forest regression, multiple linear regression, and support vector machine regression models in predicting kinase activities from cancer-related RNA sequencing data. Efficient gene signatures from the LINCS-L1000 dataset were used as inputs for KinPred-RNA. The results highlight its potential to be related to biological function. In conclusion, KinPred RNA constitutes a significant advance in cancer research by potentially facilitating the identification of cancer.

INTRODUCTION

Kinases, enzymes that catalyze phosphate group transfer, are critical in cancer pathogenesis.[1,2] They are the focus of extensive research as potential drug targets, particularly with multi-targeted receptor tyrosine kinase (RTK) inhibitors for cancer treatment.[2,3] As of 2020, 52 small-molecule protein kinase inhibitors have received approval from the United States Food and Drug Administration (FDA),[4] yet the exploration of many kinases remains limited.[5] Therefore, understanding kinase activity profiles in cancer tissues is fundamental to cancer treatment.

Traditional protein kinase activity measurement methods, involving radioactivity and 32P incorporation,[6,7] face challenges such as limitations in simultaneous kinase measurement and health risks.[8,9] With recent technological advances, non-radioactive methods based on fluorescent or luminescent peptide substrates are gaining popularity. These methods have the advantage of providing a more natural environment for studying kinase activity, such as the ability to perform studies in cell lysates and live cells. However, measuring a wide range of kinase activities is still expensive and challenging.

The high costs of non-radioactive kinase activity measurement have prompted a shift toward high-throughput sequencing technology. Proteomics sequencing data have been used to calculate the corresponding kinase activities.[10] In this context, Crowl et al. developed an algorithm called KSTAR to predict patient-specific kinase activities from phosphoproteomics data.[10] Their results demonstrated the potential of computational tools for integrating multiple data types to identify cancer biomarkers and therapeutic targets. Computational tools for the prediction of kinase inhibitor resistance and selectivity, such as those of Lo et al.[11] and Yang et al.,[12] demonstrate the utility of these activities. In addition, deep learning models have been proposed for kinase phosphorylation prediction, such as the generic deep convolutional neural network framework called NetPhosPan by Fenoy et al.[13] and the deep learning model EMBER by Kirchoff et al.[14] These models accurately predict kinase activities and phosphorylation events using machine learning algorithms.

Effective biomarkers for diagnosing and treating cancer have been sought for many years. Despite significant advances in non-radioactive methods for the measurement of kinase activity, studies have not yet explored the correlation between kinase activity and RNA-seq

[1]Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen 518172, China
[2]School of Medicine, The Chinese University of Hong Kong, Shenzhen 518172, China
[3]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China
[4]Kobilka Institute of Innovative Drug Discovery, School of Medicine, The Chinese University of Hong Kong, Shenzhen 518172, China
[5]Department of Computer Science and Information Engineering, National Central University, Taoyuan 320953, Taiwan
[6]Division of Health Medical Intelligence, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo, Japan
[7]Institute of Bioinformatics and Systems Biology, National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan
[8]Center for Intelligent Drug Systems and Smart Bio-devices (IDS^{2}B), National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan
[9]These authors contributed equally
[10]Lead contact
*Correspondence: leetzongyi@nycu.edu.tw
https://doi.org/10.1016/j.isci.2024.109333

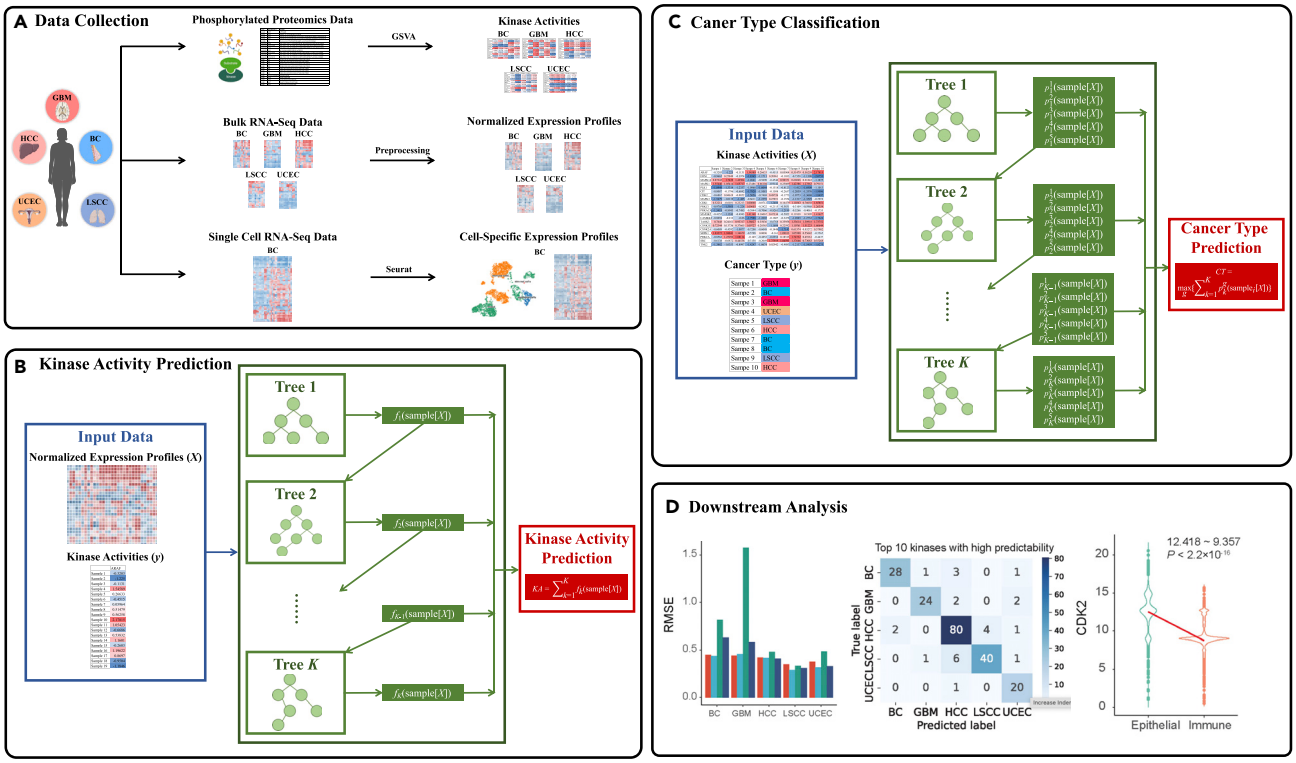


Figure 1. The overall workflow of this study

(A) Data collection and preprocessing: To ensure data quality, phosphorylated proteomics data, bulk RNA-seq data, and scRNA-seq data were collected from five cancer types, including breast cancer (BC), glioblastoma multiforme (GBM), hepatocellular carcinoma (HCC), lung squamous cell carcinoma (LSCC), and uterine corpus endometrial carcinoma (UCEC). KESA, preprocessing, and Seurat were used to preprocess the three types of data, respectively, to standardize the data for downstream analysis.

(B) Kinase activity prediction: normalized bulk RNA-seq gene expression profiles and XGBoost regression model were used to construct the kinase activity prediction model.

(C) Cancer type classification: XGBoost was also used to develop a classification model to differentiate the different cancer types.

(D) Downstream analysis: after model development, performance evaluation, and analysis of predicted kinase activity profiles for scRNA-seq data, we conducted further investigations to evaluate model performance.

expression levels in cancer. In addition, there are no studies that have investigated the potential of specific kinase activities as biomarkers for the differentiation of different types of cancer.[15,16] This study aims to bridge this gap with KinPred-RNA, an interpretable machine learning model for predicting kinase activities from bulk RNA-seq and phosphorylated proteomics data in various cancers. In addition, high-resolution profiles of kinase activities of individual cells based on their RNA-seq profiles were obtained using single-cell RNA-seq (scRNA-seq) data from invasive breast cancer tissues. The use of KinPred-RNA can predict kinase activity, providing insight into how computational methods can identify potential kinases involved in tumorigenesis and other developmental processes that reflect cell heterogeneity.[17–19] In short, this study would demonstrate the immense potential of high-throughput sequencing and computational tools in identifying cancer biomarkers and therapeutic targets. These findings would represent a major step forward in the quest for more effective cancer diagnosis and treatment strategies and highlight the need for continued research in this critical area.

To achieve our research goals, we adopted a structured, multi-step workflow, as shown in Figure 1. Our first step was to collect and preprocess data from five specific cancer types: breast cancer (BC), glioblastoma multiforme (GBM), hepatocellular carcinoma (HCC), lung squamous cell carcinoma (LSCC), and uterine corpus endometrial carcinoma (UCEC). These data comprised phosphorylated proteomics, bulk RNA-seq, and scRNA-seq datasets. The next step was to apply various regression techniques to predict kinase activity. These techniques included extreme gradient boosting (XGBoost) regression, random forest (RF) regression, multiple linear regression, and support vector machine (SVM) regression.[15–17] In the final step, we rigorously evaluated the developed models through analysis of the predicted kinase activity profiles, especially for scRNA-seq data. To ensure the reliability of our models, we carried out additional investigations and validations, providing a path for future research discoveries. For transparency and reproducibility, all analysis was conducted in Python version 3.8.1, and statistical testing was conducted in the R computing environment (version 4.0.0). BioRender and the ggplot package in R were used to generate visual representations such as the flowchart and other graphics.

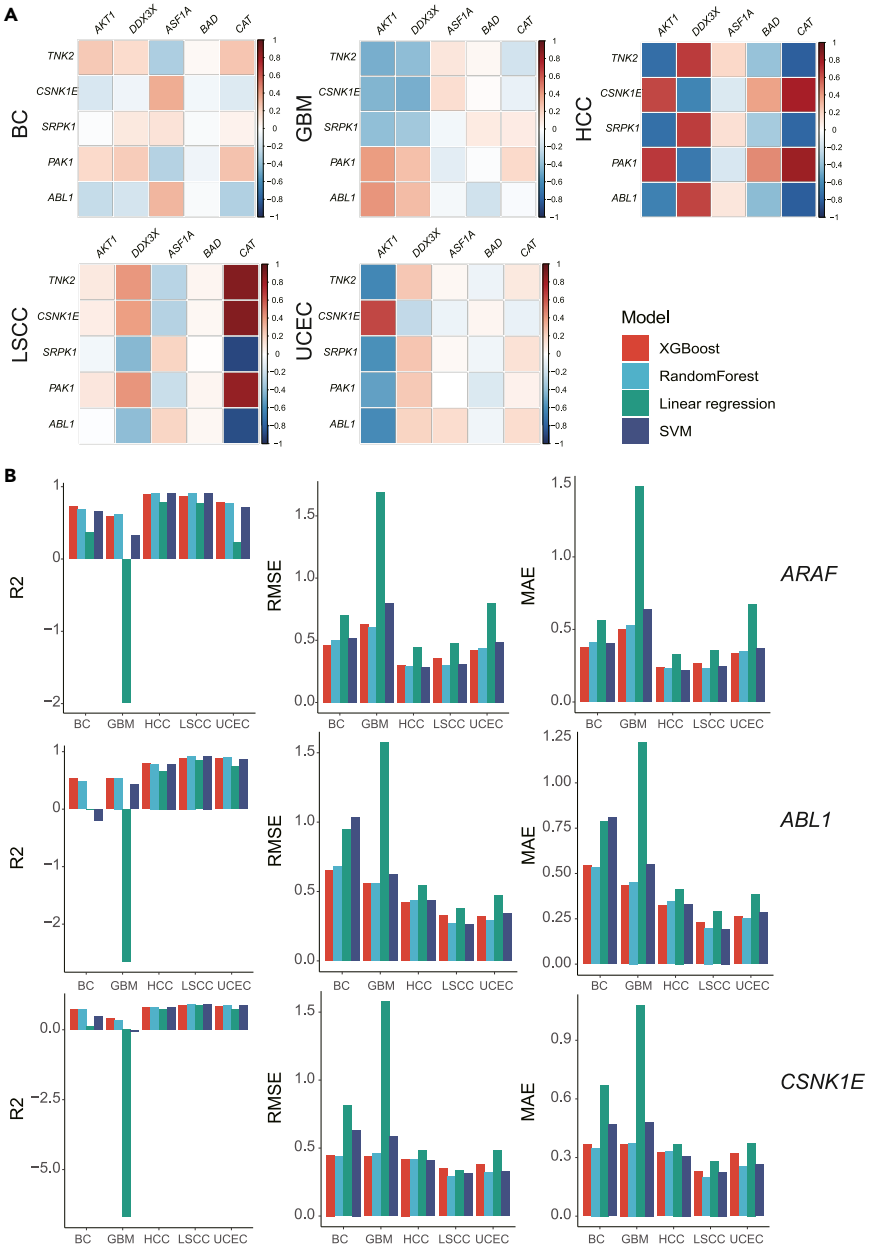


Figure 2. Non-linear regression model links gene expression and kinase activity

(A) Pearson's correlation coefficient analysis between kinase activity and gene expression profiles of corresponding substrates for different cancer types. The figure shows the correlation between the level of kinase activity (y axis) and gene expression profiles (x axis) for different cancer types, highlighting the extremely varied relationship between kinase activity and gene expression.

(B) Benchmark results of XGBoost model for kinase activity of ARAF, ABL1, and CSNK1E with random forest regression, multiple linear regression, and SVM regression, revealing XGBoost model's better performance and stability within five cancer types. The three kinases were reported to be associated with multiple types of tumor growth including breast invasive cancer and colorectal cancer.

RESULTS

Relationships between kinase activities and gene expression of substrates

Pearson's correlation coefficients (PCC) were used to evaluate the relationship between kinase activities and gene expression of substrates. The KSEA algorithm was applied to our previous kinase substrate database to calculate kinase activities. The results shown in Figure 2A revealed that under different cancer conditions, the majority of kinase activities did not show a strong correlation with the gene expression of substrates. TNK2 kinase activity, for instance, correlated differently with CAT gene expression across cancer types: negatively in HCC

(PCC = −0.8) and positively in LSCC (PCC = 0.89). Similar trends were observed for other kinase-substrate pairs, such as the ARAF-CAT pair in HCC, where the PCC was −0.81, and in LSCC, where the PCC was −0.87.

However, certain kinase/substrate pairs, such as ABL1 and CAT gene expression, consistently showed negative correlations in various cancers, including HCC and LSCC. Despite these findings, it is important to note that the majority of the kinase/substrate pairs did not show obvious associations in terms of PCCs, and the kinase activity did not consistently affect the gene expression of the substrates under different cancer conditions.

These findings highlight the complex relationship between kinase activities and the expression of substrate genes, which can vary significantly between cancer types. Therefore, instead of solely depending on the gene expression of individual substrates, it is crucial to use machine learning techniques to estimate particular kinase activities. This would enable a better comprehension of the intricate connections between kinase activity and gene expression, as well as their variations in diverse types of cancer.

Performance of machine learning algorithms on predicting kinase activities

Figure 3 illustrates the performance of various predictive models for three kinases across five cancer types, elucidating the intricate nexus between kinase activities and cancer. Each of these cancers would have unique characteristics that could affect the accuracy of predicting the kinase activity. Four machine learning algorithms, XGBoost regression, RF regression, multiple linear regression, and SVM regression were benchmarked to evaluate the performance of the prediction models.

ARAF, ABL1, and CSNK1E were chosen due to their calculable kinase activities from phosphorylated proteomics data across all five cancer types, and their known association with tumorigenesis. In terms of selection constrains, ARAF, ABL1 and CSNK1E achieved good performance in the KinPred-RNA model (higher R^2, lower RMSE and mean absolute error [MAE]). All the three kinases were also reported to be associated with cancer growth and invasion. It was reported that ARAF kinase could negatively regulate ERBB3-AKT signaling and hence suppress tumor metastasis.[20] ABL1 was the first oncogene to be associated with the development of leukemias.[21] CSNK1E was covered to be a potential marker for prognosis in colorectal cancer.[22] That is why the three kinases were chosen as the benchmark kinases.

The XGBoost and RF regression models performed better than multiple linear regression and SVM regression models in predicting kinase activity across all five cancer types. The XGBoost model displays more consistent performance across cancer types (HCC, LSCC, GBM, BC, and UCEC) with smaller variations in R^2, RMSE, and MAE for the three kinases (ARAF, ABL1, and CSNK1E), compared to the RF model. This finding was particularly noteworthy given the heterogeneity of these cancers, highlighting the importance of accurately predicting kinase activities across different cancers. That is why XGBoost was chosen as the basic model of KinPred-RNA. Particularly, the model performed exceptionally well, accurately predicting over 50% of kinase activities with an R^2 value greater than 0.5, in HCC and LSCC (Figure 3A). Accurate kinase activity prediction could facilitate the development of targeted therapies, significantly impacting cancer diagnosis and treatment. The relatively smaller sample sizes could explain the weaker performance of the model in GBM and BC. The numbers of BC and GBM individuals (including both bulk RNA-seq datasets and corresponding phosphoproteomics data) are 122 and 108 which are ranked the relative lower two cancer types among the five cancer types (202 in LSCC, 115 in UCEC, and 318 in HCC). This variability could be attributed to the complex relationships between kinase activity and cancer specificity, alongside batch effects in samples. Therefore, further research is needed to better understand the complex interactions between kinases and cancer types, which may ultimately lead to improved diagnosis and treatment of cancer. Another key finding is the positive correlation between the proportion of kinases with an R^2 above 0.5 and the sample size in each cancer type dataset, indicating that larger cohorts improve the performance of KinPred-RNA (Figure 3B; PCC = 0.61).

To identify key features in the kinase activity prediction model and improve interpretability, we used the feature importance function of the XGBoost model. In Figure 3C, we selected ARAF, ABL1, and CSNK1E as examples to demonstrate the main contributing genes to the three kinases (Table 1). Each type of cancer was characterized by specific genes that contribute to the XGBoost model. For example, the top genes which contribute to ABL1's activity prediction model among the five cancer types (BC, GBM, HCC, LSCC, and UCEC) are AURKB, ETS1, PGM1, UBE2C, and SNCA. These findings underscore the impact of cancer-specific genes on kinase phosphorylation. AURKB was reported to be a potential prognostic indicator in early basal-like breast cancer.[23] ETS1 was used to be reported to be a key factor in tumor angiogenesis in GBM.[24] PGM1 encodes an enzyme involved in glycogen metabolism. The downregulation of PGM1 in HCC was previously known to be associated with the malignancy of HCC.[25] It was observed that UBE2C could promote squamous cell carcinoma and the expression of UBE2C is higher in cancer tissue samples than adjacent normal tissues.[26] SNCA was previously known for its role in Parkinson's disease, but recent evidence suggests that it may also be involved in tumorigenesis, and its downregulation has been associated with a better prognosis in LUAD.[27] Our results here may indicate the potential influence of specific genes to several kinases' phosphorylation levels and this influence could be related to tumorigenesis.

Performance of machine learning algorithms on identification of cancer types

To illustrate the potential of kinase activity-based differentiation of cancer types, we ranked kinases in cancer tissues by their predictability (Table S6). The predictive ability was evaluated using the R^2 derived from the prediction models developed in this study. Figure 3A visualizes the proportions of kinases with high prediction results ($R^2 > 0.3$ or 0.5). HCC and LSCC achieve the best prediction results with 53% and 59% kinases performing high predictability ($R^2 > 0.5$). Yet, only 11% of kinases in GBM demonstrate good predictability ($R^2 > 0.3$) when utilizing the bulk RNA-seq dataset. To further investigate the predictive power of these kinases, the 10 highest and lowest predictable kinases were

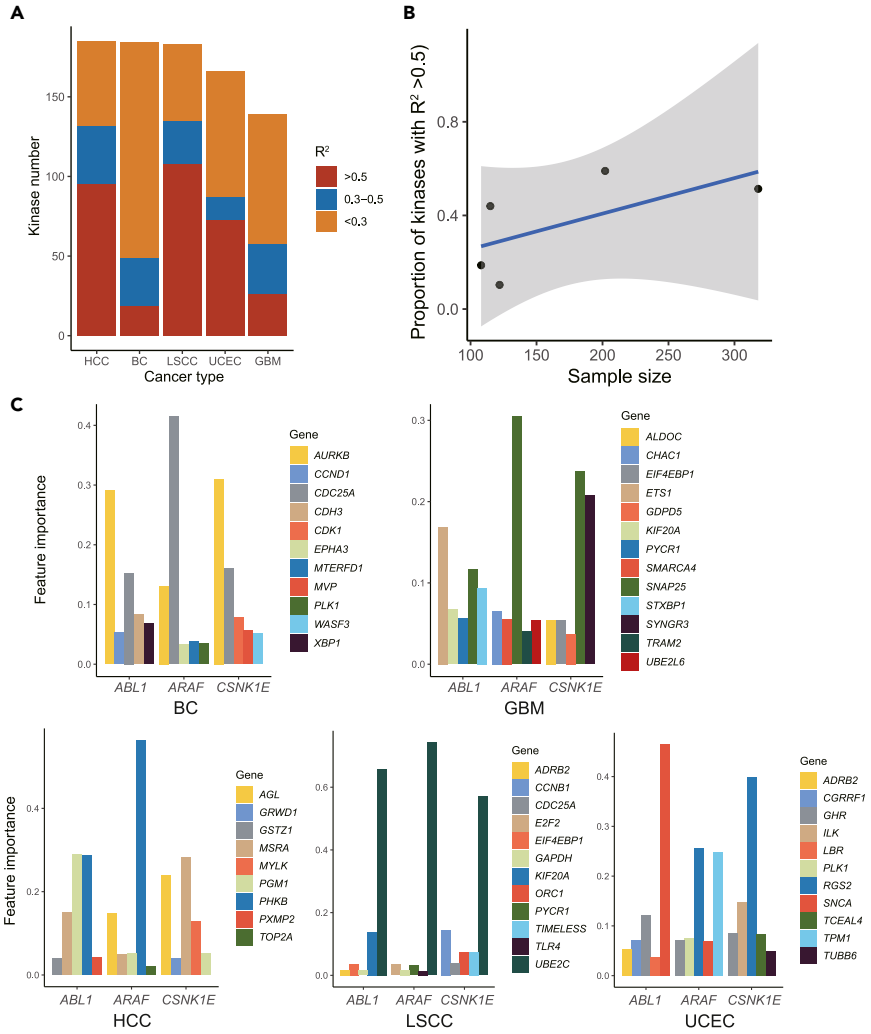


Figure 3. Kinase activity prediction results and corresponding feature importance

(A) R^2 distribution results of kinase activity prediction model for five cancer types based on KinPred-RNA model. Kinases are classified into three groups according to their predictability (R^2). Dark yellow represents kinases with R^2 smaller than 0.3. Dark blue represents kinases with R^2 between 0.3 and 0.5. Dark red represents kinases with R^2 larger than 0.5.

(B) Linear relationship between the proportion of kinases with R^2 larger than 0.5 and sample size of each cancer type datasets. Each point indicates one cancer type. (Pearson's r = 0.61) The dark area is the confidential interval.

(C) Feature importance produced by KinPred-RNA model for kinases ABL1, ARAF, and CSNK1E, indicating cancer-type specific characteristics of contributing genes among each cell type.

iteratively selected. The top 10 kinases with the highest predictability were identified and shown in Table 2, while the bottom 10 kinases represented those with the lowest predictability. These 10 kinases were used as input features for the XGBoost multi-classification model to predict the corresponding cancer type of each sample.

In particular, using the top 10 kinases as inputs, the model attained a micro F1-score of 0.885 in cancer type differentiation. The confusion matrix shown in Figure 4A illustrates the classification accuracy. In contrast, the model achieved only 0.627 micro F1-score when using the 10 least predicted kinases as input features (Figure 4B). Apart from that, when we utilized kinases ranked 31–40 and kinases ranked 91–100, the metrics also corresponds to our hypothesis (micro F1-score = 0.783; micro F1-score = 0.682). This result further confirmed the effectiveness of the ranking method used for evaluating kinase predictability. The classification accuracies achieved with different numbers of kinases as input features were compared to evaluate the performance of the model. Figure 4C shows ROC curves for the top 10 kinases used in five binary classification models, each differentiating one specific cancer type from four others. Each plot includes the benchmark results of XGBoost, RF, logistic regression, and SVM. As we can see from the five plots, XGBoost and RF performed best than logistic regression and SVM in the binary classification tasks. The area under the receiver operating characteristic curve (AUC) for XGBoost binary classification model in BC, GBM, HCC, LSCC, and UCEC is 0.937, 0.908, 0.919, 0.894, and 0.923, respectively. F1-score, accuracy, Mathew's correlation coefficient

Table 1. Feature importance of LINCS-L1000 genes contributing to KinPred-RNA model of ARAF, ABL1, and CSNK1E's kinase activity

Cancer	Kinase	Gene	Importance	Cancer	Kinase	Gene	Importance
BC	ARAF	CDC25A	0.4147	GBM	ARAF	SNAP25	0.3047
	ABL1	AURKB	0.2905		ABL1	ETS1	0.1687
	CSNK1E	AURKB	0.3099		CSNK1E	SNAP25	0.2369
	ARAF	AURKB	0.1310		ARAF	CHAC1	0.0656
	ABL1	CDC25A	0.1529		ABL1	SNAP25	0.1164
	CSNK1E	CDC25A	0.1602		CSNK1E	SYNGR3	0.2073
	ARAF	MTERFD1	0.0378		ARAF	SMARCA4	0.0555
	ABL1	CDH3	0.0836		ABL1	STXBP1	0.0940
	CSNK1E	CDK1	0.0782		CSNK1E	ALDOC	0.0544
	ARAF	PLK1	0.0350		ARAF	UBE2L6	0.0536
	ABL1	XBP1	0.0681		ABL1	KIF20A	0.0671
	CSNK1E	MVP	0.0565		CSNK1E	EIF4EBP1	0.0538
	ARAF	EPHA3	0.0332		ARAF	TRAM2	0.0402
	ABL1	CCND1	0.0532		ABL1	PYCR1	0.0564
	CSNK1E	WASF3	0.0516		CSNK1E	GDPD5	0.0372
HCC	ARAF	PHKB	0.5638	LSCC	ARAF	UBE2C	0.7429
	ABL1	PGM1	0.2892		ABL1	UBE2C	0.6572
	CSNK1E	MSRA	0.2810		CSNK1E	UBE2C	0.5706
	ARAF	AGL	0.1478		ARAF	E2F2	0.0345
	ABL1	PHKB	0.2859		ABL1	KIF20A	0.1376
	CSNK1E	AGL	0.2387		CSNK1E	CCNB1	0.1446
	ARAF	PGM1	0.0505		ARAF	PYCR1	0.0314
	ABL1	MSRA	0.1501		ABL1	EIF4EBP1	0.0366
	CSNK1E	MYLK	0.1296		CSNK1E	TIMELESS	0.0735
	ARAF	MSRA	0.0499		ARAF	GAPDH	0.0168
	ABL1	PXMP2	0.0419		ABL1	ADRB2	0.0159
	CSNK1E	PGM1	0.0509		CSNK1E	ORC1	0.0731
	ARAF	TOP2A	0.0204		ARAF	TLR4	0.0142
	ABL1	GSTZ1	0.0390		ABL1	GAPDH	0.0152
	CSNK1E	GRWD1	0.0396		CSNK1E	CDC25A	0.0380
UCEC	ARAF	RGS2	0.2565	UCEC	CSNK1E	GHR	0.0858
	ABL1	SNCA	0.4648		ARAF	GHR	0.0715
	CSNK1E	RGS2	0.3986		ABL1	ADRB2	0.0527
	ARAF	TPM1	0.2484		CSNK1E	TCEAL4	0.0834
	ABL1	GHR	0.1222		ARAF	SNCA	0.0686
	CSNK1E	ILK	0.1469		ABL1	LBR	0.0380
	ARAF	PLK1	0.0763		CSNK1E	TUBB6	0.0487
	ABL1	CGRRF1	0.0718				

(MCC), AUC of the four models for five cancer types can be found in Table 3. These results confirm the robustness of the model in predicting kinase activities in different cancer types and highlight the importance of selecting kinases with the highest predictability as input features to achieve the highest accuracy rate in classifying specific cancer types.

Investigations for applying kinase activities prediction models on scRNA-seq data

After demonstrating the effectiveness of KinPred-RNA with bulk RNA-seq datasets, we assessed its robustness using two separate scRNA-seq datasets. First, we tested our kinase model in invasive breast cancer tissue samples. We utilized scRNA-seq data from GSE180286, encompassing samples from five primary breast cancer donors, as provided by Xu et al.[28] Figure 5A shows the UMAP[29] plot of primary breast cancer

Table 2. The top 10 kinases ranked according to their R^2 values for five different cancer types based on KinPred-RNA (LINCS-L1000 gene signatures + XGBoost)

Kinase	BC		Kinase	GBM		Kinase	HCC		Kinase	LSCC		Kinase	UCEC	
MAPK3	R^2	0.7862	MAP4K5	R^2	0.7340	CIT	R^2	0.9336	ERN1	R^2	0.9104	MAP4K5	R^2	0.9308
	RMSE	0.4402		RMSE	0.3833		RMSE	0.2540		RMSE	0.3114		RMSE	0.2404
	MAE	0.3491		MAE	0.3141		MAE	0.1728		MAE	0.2066		MAE	0.1849
MAPK1	R^2	0.7408	PRKCI	R^2	0.7045	CDK1	R^2	0.9293	IRAK1	R^2	0.9058	MAP3K1	R^2	0.9032
	RMSE	0.5375		RMSE	0.3920		RMSE	0.2479		RMSE	0.3173		RMSE	0.2852
	MAE	0.4574		MAE	0.3069		MAE	0.2001		MAE	0.2210		MAE	0.2193
CSNK1E	R^2	0.7374	SRPK1	R^2	0.6560	TAOK3	R^2	0.9289	SRPK3	R^2	0.9036	IRAK1	R^2	0.9019
	RMSE	0.4487		RMSE	0.3945		RMSE	0.2636		RMSE	0.3240		RMSE	0.2891
	MAE	0.3681		MAE	0.3297		MAE	0.1931		MAE	0.2478		MAE	0.2403
ABL1	R^2	0.7319	ABL2	R^2	0.6423	TNK2	R^2	0.9272	MAPK14	R^2	0.9018	MAPK3	R^2	0.8972
	RMSE	0.4601		RMSE	0.4665		RMSE	0.2603		RMSE	0.3223		RMSE	0.2866
	MAE	0.3770		MAE	0.3821		MAE	0.1912		MAE	0.2432		MAE	0.2064
ERN1	R^2	0.7267	CDK2	R^2	0.5996	CAMKK2	R^2	0.9250	CSNK2A1	R^2	0.8988	CIT	R^2	0.8967
	RMSE	0.4712		RMSE	0.5396		RMSE	0.2706		RMSE	0.3267		RMSE	0.2807
	MAE	0.3664		MAE	0.4313		MAE	0.1954		MAE	0.2290		MAE	0.2367
CDK2	R^2	0.6577	KIT	R^2	0.5931	IRAK1	R^2	0.9177	MAP3K1	R^2	0.8907	CAMKK2	R^2	0.8857
	RMSE	0.4874		RMSE	0.4429		RMSE	0.2901		RMSE	0.3350		RMSE	0.2996
	MAE	0.4053		MAE	0.3430		MAE	0.2198		MAE	0.2399		MAE	0.2346
CDK1	R^2	0.6396	ABL1	R^2	0.5819	PIK3CA	R^2	0.9141	CAMKK2	R^2	0.8887	PLK1	R^2	0.8818
	RMSE	0.5724		RMSE	0.6341		RMSE	0.2854		RMSE	0.3471		RMSE	0.3073
	MAE	0.4873		MAE	0.5011		MAE	0.2126		MAE	0.2207		MAE	0.2343
MAPK13	R^2	0.5416	CAMK2G	R^2	0.5774	MAP4K5	R^2	0.9104	TAOK3	R^2	0.8881	TNK2	R^2	0.8816
	RMSE	0.6154		RMSE	0.6329		RMSE	0.2845		RMSE	0.3502		RMSE	0.3049
	MAE	0.4907		MAE	0.4973		MAE	0.2187		MAE	0.2125		MAE	0.2556
ARAF	R^2	0.5281	LATS1	R^2	0.5759	CDK2	R^2	0.9072	ARAF	R^2	0.8881	ARAF	R^2	0.8805
	RMSE	0.6487		RMSE	0.4937		RMSE	0.2987		RMSE	0.3247		RMSE	0.3188
	MAE	0.5428		MAE	0.4156		MAE	0.2376		MAE	0.2302		MAE	0.2614
PAK1	R^2	0.5272	CAMKK2	R^2	0.5600	PAK1	R^2	0.8985	PLK1	R^2	0.8858	MAP4K4	R^2	0.8798
	RMSE	0.5986		RMSE	0.4486		RMSE	0.2996		RMSE	0.3210		RMSE	0.3221
	MAE	0.4931		MAE	0.3422		MAE	0.2356		MAE	0.2241		MAE	0.2414

The R^2 values represent the goodness of fit of the developed prediction models for each kinase in each cancer type. RMSE means root-mean-square error and MAE means mean absolute error.

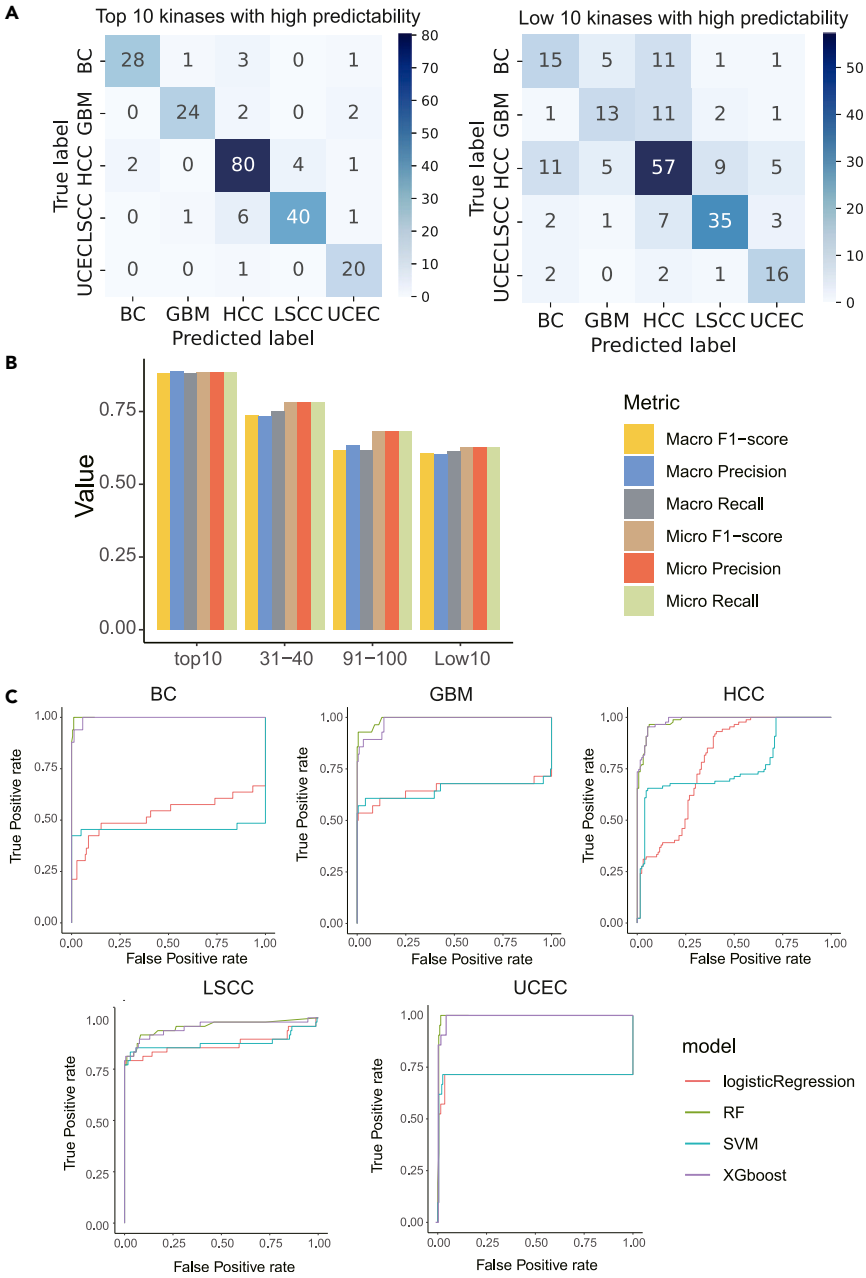


Figure 4. Cancer type classification ability of high predictable kinases' activity as input features based on XGBoost classification model

(A) Confusion matrix of XGBoost multi-classification model with ten highest predictable kinases' activity as the input features and with ten lowest predictable kinases' activity as the input features, indicating better performance of higher predictable kinases' activity as the input features in classifying cancer types.

(B) Classification performances of XGBoost multi-classification model with combinations of kinases with different predictability (ranked according R^2).

(C) ROC curves of four binary classification models including XGBoost, random forest, logistic regression, and support vector machine for tasks including BC vs. others; GBM vs. others; HCC vs. others; LSCC vs. others; UCEC vs. others.

tissues. Total cells were grouped into 14 clusters. PTPRC and EPCAM, recognized as marker genes for immune and epithelial cells respectively,[30,31] were used for cell type identification. EPCAM was reported to be an indicator for epithelial status of primary and systemic tumor cells. Here, PTPRC shows high expression level in clusters 3, 4, 5, 7, 10, and 12 while EPCAM shows high expression level in clusters 2, 8, 9, and 11 (Figure 5B). Thus, we annotated clusters 3, 4, 5, 7, 10, and 12 as immune cells and clusters 2, 8, 9, and 11 as epithelial cells (Figure 5C). We applied KinPred-RNA model based on bulk RNA-seq datasets previously to the breast cancer scRNA-seq model. The results of the kinase activity for each cell recapitulate specific kinases that could be associated with tumorigenesis. For example, Polo-like kinase 1 (PLK1), a serine/threonine protein kinase, was upregulated in breast cancer compared to normal.[32] KinPred-RNA identifies PLK1 to be differentially

Table 3. Benchmarking results of using top 10 kinases with highest predictability (highest R^2) for binary classifying cancer types based on four models

Cancer type	metric	XGBoost	RF	Logistic Regression	SVM
BC	F1-score	0.9765	0.9661	0.7890	0.8368
	Accuracy	0.9770	0.9677	0.8525	0.8756
	MCC	0.9086	0.8712	0.1607	0.3982
	AUC	0.9367	0.8939	0.5152	0.5909
GBM	F1-score	0.9714	0.9760	0.8512	0.9063
	Accuracy	0.9724	0.9770	0.8894	0.9217
	MCC	0.8723	0.8946	0.3560	0.6004
	AUC	0.9081	0.9107	0.5714	0.6964
HCC	F1-score	0.9260	0.9256	0.6529	0.6707
	Accuracy	0.9263	0.9263	0.6774	0.7097
	MCC	0.8460	0.8467	0.3003	0.3976
	AUC	0.9195	0.9156	0.6281	0.6474
LSCC	F1-score	0.9389	0.9428	0.9468	0.9468
	Accuracy	0.9401	0.9447	0.9493	0.9493
	MCC	0.8214	0.8347	0.8507	0.8507
	AUC	0.8944	0.8899	0.8854	0.8854
UCEC	F1-score	0.9767	0.9860	0.9367	0.9430
	Accuracy	0.9770	0.9862	0.9447	0.9493
	MCC	0.8656	0.9195	0.6349	0.6710
	AUC	0.9235	0.9498	0.7355	0.7594

The R^2 values represent the goodness of fit of the developed prediction models for each kinase in each cancer type.
Note: AUC: area under the receiver operating characteristic curve; MCC: Matthew's correlation coefficient.

activated between epithelial cells and immune cells (mean expression 2.147–1.810; p value $<2.2 \times 10^{-16}$; Figure 5D). Additionally, PAK1 was known to be a breast cancer oncogene and the activation of PAK1 could lead to malignant transformation.[33] Our results reveal that PAK1 activity is differentially expressed between epithelial cells and immune cells (mean expression 3.905–1.966; p value $<2.2 \times 10^{-16}$). CDK2 was reported to be required for breast cancer through low molecular weight isoform of cyclin E and CDK2 inhibitor together with CDK4/6 inhibitor could be used for the treatment of breast cancers.[34,35] We found CDK2 was differentially activated between epithelial cells and immune cells (mean expression 12.418–9.357; p value $<2.2 \times 10^{-16}$). Overall, these findings indicate KinPred-RNA recovers previous findings in breast cancer samples.

On the other hand, lung cancer tissues were used to evaluate the efficacy of KinPred RNA. We collected primary lung adenocarcinomas scRNA-seq datasets (GSE131907).[36] The scRNA datasets of Lung cancer sample 8th and normal sample 8th were used to build kinase activity profiles through KinPred-RNA model obtained previously. t-SNE plot showing the two samples were visualized in Figure 5E.[37] ROCK1 and ROCK2 are protein serine/threonine kinases and the key modulators in tumor cell invasion. It was reported that ROCK1 and ROCK2 are required for non-small cell lung cancer anchorage-independent growth[38]; and suppression of ROCK1 and ROCK2 was sufficient to impair this type of growth. Here, KinPred-RNA reports UBE2C as the main contributing genes to the ROCK1 and ROCK2 activity prediction, highlighting cancer type specific factors contributing to kinase activity prediction (Figure 5F). The differential expression of ROCK1 and ROCK2 activities between lung cancer and normal samples was confirmed by our analysis (ROCK1 mean expression: 0.220–0.215; p value $<2.2 \times 10^{-16}$; ROCK2 mean expression: 7.457–6.071; p value $<2.2 \times 10^{-16}$; Figure 5G). These findings suggest that KinPred-RNA could accurately infer the relative kinase activities in the cell-type specific levels. Our tools could implicate the possibility of utilizing pan-cancer scRNA-seq datasets to understand cell-type specific activities comprehensively.

DISCUSSION

Given the limited knowledge about many kinases, understanding kinase activity profiles in cancer tissues is crucial. To address this issue, we developed KinPred-RNA, a machine-learning model that uses bulk RNA-seq data to predict kinase activities in various cancers. Our model was used to distinguish cancer types based on kinase activities and to estimate these profiles using single-cell RNA data. When predicting kinase activity from cancer tissue-derived bulk RNA-seq datasets, KinPred-RNA outperformed baseline models. To further validate our model, we generated a list of kinases ranked by their predictability, revealing those that were most reliable in classifying multiple cancer types. Our kinase activity prediction models identified kinases with significantly different activities between cell groups from scRNA-seq samples of invasive breast cancer. Our results demonstrate the reliability of KinPred-RNA and its potential downstream applications. Our approach, utilizing bulk RNA-seq

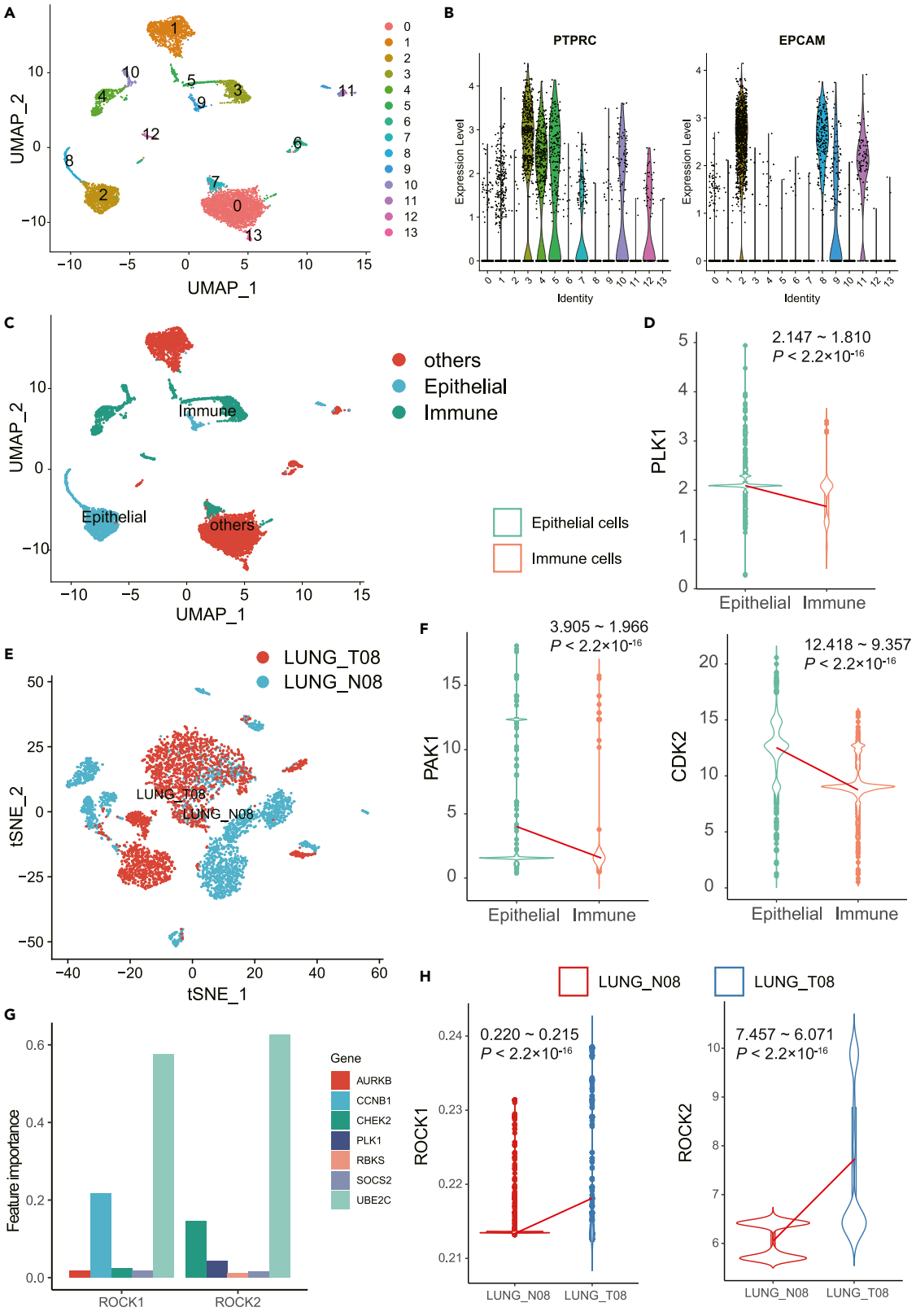

A
UMAP_2
UMAP_1
0
1
2
3
4
5
6
7
8
9
10
11
12
13
B
PTPRC
EPCAM
Expression Level
Identity
C
others
Epithelial
Immune
Epithelial cells
Immune cells
D
2.147 ~ 1.810
P < 2.2×10-16
PLK1
Epithelial
Immune
E
LUNG_T08
LUNG_N08
tSNE_2
tSNE_1
F
3.905 ~ 1.966
P < 2.2×10-16
PAK1
12.418 ~ 9.357
P < 2.2×10-16
CDK2
G
Feature importance
Gene
AURKB
CCNB1
CHEK2
PLK1
RBKS
SOCS2
UBE2C
ROCK1
ROCK2
H
LUNG_N08
LUNG_T08
0.220 ~ 0.215
P < 2.2×10-16
ROCK1
7.457 ~ 6.071
P < 2.2×10-16
ROCK2

Figure 5. Application of KinPred-RNA model to pan-cancer scRNA-seq datasets

(A) UMAP plot of a breast invasive scRNA-seq dataset (GSE180286) clustered into 14 groups.

(B) Relative log-transformed transcriptional expression level of PTPRC and EPCAM among 14 clusters. PTPRC is used as the marker gene for immune cells and EPCAM is used as the marker gene for epithelial cells.

(C) Annotated results of UMAP plot indicated in (A). Immune cells group and epithelial cells group are labeled.

(D) Violin plot of PLK1 predicted kinase activity among immune cells and epithelial cells. PLK1, mean level 2.147 ~ 1.810, $p < 2.2 \times 10^{-16}$.

(E) t-SNE plot of a primary lung adenocarcinomas scRNA-seq dataset containing LUNG 08 tumor sample donor and LUNG 08 normal sample donor (GSE131907).

(F) Violin plot of PAK1 and CDK2 predicted kinase activity among immune cells and epithelial cells. PAK1, mean level 3.905 ~ 1.966, $p < 2.2 \times 10^{-16}$; CDK2, mean level 12.418 ~ 9.357, $p < 2.2 \times 10^{-16}$.

(G) Relative feature importance of genes contributing to KinPred-RNA model of kinases ROCK1 and ROCK2. Gene UBE2C was observed to be the main contribution of the model.

(H) Violin plot of ROCK1, ROCK2 predicted kinase activity among tumor cells and normal cells. ROCK1, mean level 0.220 ~ 0.215, $p < 2.2 \times 10^{-16}$; ROCK2, mean level 7.457 ~ 6.071, $p < 2.2 \times 10^{-16}$.

datasets from human tissues for kinase activity prediction, is rare and novel. Our method bridges the gap between bulk RNA-seq data and kinase activities without requiring prior access to phosphoproteomics data from cancer tissues, as done in previous research such as the KSTAR algorithm by Crowl et al.[10] Our kinase activity prediction model could be used to construct kinase activity profiles for different cancer types. In brief, our study provides valuable insights into the complex and poorly understood field of kinase activity profiling in cancer.

The identification of cancer types has been a major concern in the bioinformatics community for a long time. In the past, as shown by Lu et al.,[39] researchers have used miRNA profiles to classify cancers. Our study introduces an innovative approach for cancer type identification by analyzing kinase activities. We validated this approach by achieving high classification accuracy, demonstrating that highly predictable kinases based on bulk RNA-seq data can be used to accurately classify specific cancer types. scRNA sequencing technology has revolutionized our ability to study biological processes in detail. To identify differentially expressed kinases that can be used to distinguish between different cell groups, our research takes advantage of this cutting-edge technology. Through the construction of kinase activity profiles for invasive breast cancer at the single cell level, we identified kinases with potential as clinical markers for cancer diagnosis and treatment. Further experimental validation of our findings could lead to a significant improvement in the health care system.

Researchers have thoroughly examined kinase functions, identifying multiple ligand-receptor pairs influencing phosphorylation. Attributing kinase phosphorylation to a limited number of ligand bindings oversimplifies its regulation and ignores the complexity of multiple influencing factors. The lack of paired bulk RNA-seq and proteomics datasets, crucial for linking mRNA levels with kinase activity, presents a significant challenge. Furthermore, tissue samples possess inherent complexity and scRNA-seq data are sparse, exacerbating the already challenging situation. These difficulties have always hindered the ability to draw definite inferences. To address data sparsity, utilizing bulk RNA-seq data have proven significantly effective. Furthermore, the inclusion of LINCS-L1000 gene signatures can estimate the remaining 80% of unmeasured transcripts, which helps to substantially reduce the noise generated by excessive features in the data. Given these advancements, KinPred-RNA emerges as a promising tool. Its interpretive power and capacity to model non-linear relationships enable effective linking of mRNA expression with pan-cancer kinase activity. Understanding the intricate relationship between mRNA levels and kinase functional mechanisms is a major step forward with this groundbreaking approach.

KinPred-RNA outperformed other methods in predicting kinase activity in the five studied cancer types. It accurately estimated more than 50% of kinase activities with R-squared greater than 0.5 in HCC and LSCC tissues. This model has enormous potential as a powerful tool for the construction of kinase activity profiles of specific cancer types, as well as for the identification of abnormal kinases that contribute to the pathogenesis of specific cancer diseases. Furthermore, our model could rapidly estimate the functions of specific kinases in cancer and help to predict clinical prognosis. In the future, to improve the applicability of our model and to gain a broader understanding of the functions of specific kinases in the pathogenesis of cancer diseases, we aim to include more cancer tissue samples. KinPred-RNA will play an important role in advancing the field of bioinformatics and promoting the improvement of healthcare in the fight against cancer.

Limitations of the study

Although our research has demonstrated the utility of our model in the prediction of specific kinase activities and the classification of cancer types, it is not without limitations. A major limitation of our model is that its performance varies across cancer types. While our model has proven effective in predicting kinase activity in HCC and LSCC, its accuracy in GBM is comparatively lower. Moreover, *in vivo* and *in vitro* experiments are crucial to validate the functions of kinases identified by our model. These experiments would be essential for the confirmation of the potential clinical significance of our findings and for the identification of the most promising targets for therapeutic intervention. Nevertheless, we are confident in the potential of our model, which combines LINCS-L1000 gene signatures and bulk RNA-seq data, to accurately predict kinase activity profiles of specific cancer types.

We also recognized the challenge posed by the lack of corresponding phosphorylated proteomics and bulk RNA-seq data in developing our model. This challenge could be addressed with more research focusing on pan-cancer gene expression and phosphorylated proteomics. Additionally, the complex landscape of pan-cancer samples is another challenge since various cancer type and patients' personal conditions may have a great impact on the training of KinPred-RNA. Conquering batch effects during model training is another critical issue, especially considering the diverse nature of cancer types and patient conditions. To address this issue, we applied a 0–1 transformation to the gene expression levels within the same type of cancer. Despite the challenges, KinPred-RNA is a promising tool for the advancement of kinase phosphorylation research.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Data collection and preprocessing
  - Development of kinase activities prediction models
  - Classifying five types of cancer using kinase activities
  - Analysis of kinase activity profiles predicted by KinPred-RNA for breast cancer tissues and lung cancer tissues
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109333.

ACKNOWLEDGMENTS

This work was supported by the National Science and Technology Council (NSTC112-2321-B-A49-016, 112-2740-B-400-005, and 113-2222-E-008-001-MY2) and The National Health Research Institutes (NHRI-EX113-11320BI), Taiwan. This work was also financially supported by the Center for Intelligent Drug Systems and Smart Bio-devices (IDS^{2}B) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project and Yushan Young Fellow Program (112C1N084C) by the Ministry of Education (MOE) in Taiwan.

AUTHOR CONTRIBUTIONS

Y.Z. and S.L. carried out the data collection and preprocessing. C.-R.C. and Y.Z. participated in the data analyses, model construction, and drafted the manuscript. Y.Z., L.Y., C.-R.C., and T.-Y.L. participated in the design of the study and performed the draft revision. Y.Z., L.Y., C.-R.C., W.Z., Y.H., Y.P., and T.-Y.L. conceived of the study, and participated in the design and helped to revise the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 26, 2023
Revised: December 7, 2023
Accepted: February 21, 2024
Published: February 28, 2024

REFERENCES

1. Wang, C., Xu, H., Lin, S., Deng, W., Zhou, J., Zhang, Y., Shi, Y., Peng, D., and Xue, Y. (2020). GPS 5.0: An Update on the Prediction of Kinase-specific Phosphorylation Sites in Proteins. Dev. Reprod. Biol. *18*, 72–80. https://doi.org/10.1016/j.gpb.2020.01.001.
2. Paul, M.K., and Mukhopadhyay, A.K. (2004). Tyrosine kinase - Role and significance in Cancer. Int. J. Med. Sci. *1*, 101–115. https://doi.org/10.7150/ijms.1.101.
3. Zwick, E., Bange, J., and Ullrich, A. (2002). Receptor tyrosine kinases as targets for anticancer drugs. Trends Mol. Med. *8*, 17–23. https://doi.org/10.1016/S1471-4914(01)02217-1.
4. Roskoski, R., Jr. (2020). Properties of FDA-approved small molecule protein kinase inhibitors: A 2020 update. Pharmacol. Res. *152*, 104609. https://doi.org/10.1016/j.phrs.2019.104609.
5. Hasinoff, B.B., and Patel, D. (2010). The lack of target specificity of small molecule anticancer kinase inhibitors is correlated with their ability to damage myocytes *in vitro*. Toxicol. Appl. Pharmacol. *249*, 132–139. https://doi.org/10.1016/j.taap.2010.08.026.
6. Casnellie, J.E., and Krebs, E.G. (1984). The use of synthetic peptides for defining the specificity of tyrosine protein kinases. Adv. Enzyme Regul. *22*, 501–515. https://doi.org/10.1016/0065-2571(84)90028-1.
7. Casnellie, J.E. (1991). Assay of protein kinases using peptides with basic residues for phosphocellulose binding. Methods Enzymol. *200*, 115–120. https://doi.org/10.1016/0076-6879(91)00133-h.
8. Wang, Y., and Ma, H. (2015). Protein kinase profiling assays: a technology review. Drug Discov. Today Technol. *18*, 1–8. https://doi.org/10.1016/j.ddtec.2015.10.007.
9. González-Vera, J.A. (2012). Probing the kinome in real time with fluorescent peptides. Chem. Soc. Rev. *41*, 1652–1664. https://doi.org/10.1039/c1cs15198c.
10. Crowl, S., Jordan, B.T., Ahmed, H., Ma, C.X., and Naegle, K.M. (2022). KSTAR: An algorithm to predict patient-specific kinase activities from phosphoproteomic data. Nat. Commun. *13*, 4283. https://doi.org/10.1038/s41467-022-32017-5.
11. Lo, Y.C., Liu, T., Morrissey, K.M., Kakiuchi-Kiyota, S., Johnson, A.R., Broccatelli, F., Zhong, Y., Joshi, A., and Altman, R.B. (2019). Computational analysis of kinase inhibitor selectivity using structural knowledge. Bioinformatics *35*, 235–242. https://doi.org/10.1093/bioinformatics/bty582.
12. Yang, Z.Y., Ye, Z.F., Xiao, Y.J., Hsieh, C.Y., and Zhang, S.Y. (2022). SPLDExtraTrees: robust machine learning approach for predicting kinase inhibitor resistance. Brief. Bioinform. *23*, bbac050. https://doi.org/10.1093/bib/bbac050.
13. Fenoy, E., Izarzugaza, J.M.G., Jurtz, V., Brunak, S., and Nielsen, M. (2019). A generic deep convolutional neural network

framework for prediction of receptor-ligand interactions-NetPhosPan: application to kinase phosphorylation prediction. Bioinformatics *35*, 1098–1107. https://doi.org/10.1093/bioinformatics/bty715.

14. Kirchoff, K.E., and Gomez, S.M. (2022). EMBER: multi-label prediction of kinase-substrate phosphorylation events through deep learning. Bioinformatics *38*, 2119–2126. https://doi.org/10.1093/bioinformatics/btac083.
15. Ma, R., Li, S., Li, W., Yao, L., Huang, H.D., and Lee, T.Y. (2023). KinasePhos 3.0: Redesign and expansion of the prediction on kinase-specific phosphorylation sites. Dev. Reprod. Biol. *21*, 228–241. https://doi.org/10.1016/j.gpb.2022.06.004.
16. Li, Z., Li, S., Luo, M., Jhong, J.H., Li, W., Yao, L., Pang, Y., Wang, Z., Wang, R., Ma, R., et al. (2022). dbPTM in 2022: an updated database for exploring regulatory networks and functional associations of protein post-translational modifications. Nucleic Acids Res. *50*, D471–D479. https://doi.org/10.1093/nar/gkab1017.
17. Huang, K.Y., Hsu, J.B.K., and Lee, T.Y. (2019). Characterization and Identification of Lysine Succinylation Sites based on Deep Learning Method. Sci. Rep. *9*, 16175. https://doi.org/10.1038/s41598-019-52552-4.
18. Su, M.G., Weng, J.T.Y., Hsu, J.B.K., Huang, K.Y., Chi, Y.H., and Lee, T.Y. (2017). Investigation and identification of functional post-translational modification sites associated with drug binding and protein-protein interactions. BMC Syst. Biol. *11*, 132. https://doi.org/10.1186/s12918-017-0506-1.
19. Bretaña, N.A., Lu, C.T., Chiang, C.Y., Su, M.G., Huang, K.Y., Lee, T.Y., and Weng, S.L. (2012). Identifying protein phosphorylation sites with kinase substrate specificity on human viruses. PLoS One *7*, e40694. https://doi.org/10.1371/journal.pone.0040694.
20. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15. https://doi.org/10.1186/s13059-017-1382-0.
21. Greuber, E.K., Smith-Pearson, P., Wang, J., and Pendergast, A.M. (2013). Role of ABL family kinases in cancer: from leukaemia to solid tumours. Nat. Rev. Cancer *13*, 559–571. https://doi.org/10.1038/nrc3563.
22. Tiong, K.L., Chang, K.C., Yeh, K.T., Liu, T.Y., Wu, J.H., Hsieh, P.H., Lin, S.H., Lai, W.Y., Hsu, Y.C., Chen, J.Y., et al. (2014). CSNK1E/CTNNB1 are synthetic lethal to TP53 in colorectal cancer and are markers for prognosis. Neoplasia *16*, 441–450. https://doi.org/10.1016/j.neo.2014.04.007.
23. Yuan, K., Wu, M., Lyu, S., and Li, Y. (2022). Identification of prognostic genes for early basal-like breast cancer with weighted gene co-expression network analysis. Medicine (Baltimore) *101*, e30581. https://doi.org/10.1097/MD.0000000000030581.
24. Tang, J., Li, Y., Liu, B., Liang, W., Hu, S., Shi, M., Zeng, J., Li, M., and Huang, M. (2021). Uncovering a key role of ETS1 on vascular abnormality in glioblastoma. Pathol. Oncol. Res. *27*, 1609997. https://doi.org/10.3389/pore.2021.1609997.
25. Jin, G.Z., Zhang, Y., Cong, W.M., Wu, X., Wang, X., Wu, S., Wang, S., Zhou, W., Yuan, S., Gao, H., et al. (2018). Phosphoglucomutase 1 inhibits hepatocellular carcinoma progression by regulating glucose trafficking. PLoS Biol. *16*, e2006483. https://doi.org/10.1371/journal.pbio.2006483.
26. Liu, P.F., Chen, C.F., Shu, C.W., Chang, H.M., Lee, C.H., Liou, H.H., Ger, L.P., Chen, C.L., and Kang, B.H. (2020). UBE2C is a potential biomarker for tumorigenesis and prognosis in tongue squamous cell carcinoma. Diagnostics (Basel) *10*, 674. https://doi.org/10.3390/diagnostics10090674.
27. Zhang, X., Wu, Z., and Ma, K. (2022). SNCA correlates with immune infiltration and serves as a prognostic biomarker in lung adenocarcinoma. BMC Cancer *22*, 406. https://doi.org/10.1186/s12885-022-09289-7.
28. Zhang, K., Erkan, E.P., Jamalzadeh, S., Dai, J., Andersson, N., Kaipio, K., Lamminen, T., Mansuri, N., Huhtinen, K., Carpén, O., et al. (2022). Longitudinal single-cell RNA-seq analysis reveals stress-promoted chemoresistance in metastatic ovarian cancer. Sci. Adv. *8*, eabm1831. https://doi.org/10.1126/sciadv.abm1831.
29. McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. J. Open Source Softw. *3*, 861. https://doi.org/10.21105/joss.00861.
30. Lee, T.Y., Huang, K.Y., Chuang, C.H., Lee, C.Y., and Chang, T.H. (2020). Incorporating deep learning and multi-omics autoencoding for analysis of lung adenocarcinoma prognostication. Comput. Biol. Chem. *87*, 107277. https://doi.org/10.1016/j.compbiolchem.2020.107277.
31. Lee, T.Y., Hsu, J.B.K., Chang, W.C., Wang, T.Y., Hsu, P.C., and Huang, H.D. (2009). A comprehensive resource for integrating and displaying protein post-translational modifications. BMC Res. Notes *2*, 111. https://doi.org/10.1186/1756-0500-2-111.
32. Liu, Z., Sun, Q., and Wang, X. (2017). PLK1, a potential target for cancer therapy. Transl. Oncol. *10*, 22–32. https://doi.org/10.1016/j.tranon.2016.10.003.
33. Shrestha, Y., Schafer, E.J., Boehm, J.S., Thomas, S.R., He, F., Du, J., Wang, S., Barretina, J., Weir, B.A., Zhao, J.J., et al. (2012). PAK1 is a breast cancer oncogene that coordinately activates MAPK and MET signaling. Oncogene *31*, 3397–3408. https://doi.org/10.1038/onc.2011.515.
34. Akli, S., Van Pelt, C.S., Bui, T., Meijer, L., and Keyomarsi, K. (2011). Cdk2 is required for breast cancer mediated by the low-molecular-weight isoform of cyclin E. Cancer Res. *71*, 3377–3386. https://doi.org/10.1158/0008-5472.CAN-10-4086.
35. Pandey, K., Park, N., Park, K.S., Hur, J., Cho, Y.B., Kang, M., An, H.J., Kim, S., Hwang, S., and Moon, Y.W. (2020). Combined CDK2 and CDK4/6 inhibition overcomes palbociclib resistance in breast cancer by enhancing senescence. Cancers (Basel) *12*, 3566. https://doi.org/10.3390/cancers12123566.
36. Luo, Q., Mo, S., Xue, Y., Zhang, X., Gu, Y., Wu, L., Zhang, J., Sun, L., Liu, M., and Hu, Y. (2021). Novel deep learning-based transcriptome data analysis for drug-drug interaction prediction with an application in diabetes. BMC Bioinf. *22*, 318. https://doi.org/10.1186/s12859-021-04241-1.
37. Van der Maaten, L., and Hinton, G. (2008). Visualizing high-dimensional data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.
38. Vigil, D., Kim, T.Y., Plachco, A., Garton, A.J., Castaldo, L., Pachter, J.A., Dong, H., Chen, X., Tokar, B., Campbell, S.L., and Der, C.J. (2012). ROCK1 and ROCK2 are required for non-small cell lung cancer anchorage-independent growth and invasion. Cancer Res. *72*, 5338–5347. https://doi.org/10.1158/0008-5472.CAN-11-2373.
39. Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., Ferrando, A.A., et al. (2005). MicroRNA expression profiles classify human cancers. Nature *435*, 834–838. https://doi.org/10.1038/nature03702.
40. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
41. Wang, L.B., Karpova, A., Gritsenko, M.A., Kyle, J.E., Cao, S., Li, Y., Rykunov, D., Colaprico, A., Rothstein, J.H., Hong, R., et al. (2021). Proteogenomic and metabolomic characterization of human glioblastoma. Cancer Cell *39*, 509–528.e20. https://doi.org/10.1016/j.ccell.2021.01.006.
42. Krug, K., Jaehnig, E.J., Satpathy, S., Blumenberg, L., Karpova, A., Anurag, M., Miles, G., Mertins, P., Geffen, Y., Tang, L.C., et al. (2020). Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. Cell *183*, 1436–1456.e31. https://doi.org/10.1016/j.cell.2020.10.036.
43. Ng, C.K.Y., Dazert, E., Boldanova, T., Coto-Llerena, M., Nuciforo, S., Ercan, C., Suslov, A., Meier, M.A., Bock, T., Schmidt, A., et al. (2022). Integrative proteogenomic characterization of hepatocellular carcinoma across etiologies and stages. Nat. Commun. *13*, 2436. https://doi.org/10.1038/s41467-022-29960-8.
44. Pan, L., Wang, X., Yang, L., Zhao, L., Zhai, L., Xu, J., Yang, Y., Mao, Y., Cheng, S., Xiao, T., and Tan, M. (2020). Proteomic and Phosphoproteomic Maps of Lung Squamous Cell Carcinoma From Chinese Patients. Front. Oncol. *10*, 963. https://doi.org/10.3389/fonc.2020.00963.
45. Dou, Y., Kawaler, E.A., Cui Zhou, D., Gritsenko, M.A., Huang, C., Blumenberg, L., Karpova, A., Petyuk, V.A., Savage, S.R., Satpathy, S., et al. (2020). Proteogenomic Characterization of Endometrial Carcinoma. Cell *180*, 729–748.e26. https://doi.org/10.1016/j.cell.2020.01.026.
46. Ranjan, G.S.K., Verma, A.K., and Radhika, S. (2019). K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries. In 2019 IEEE 5th International Conference for convergence in technology (I2CT), pp. 1–5.
47. Bai, M., and Sun, C. (2022). Determination of Breast Metabolic Phenotypes and Their Associations With Immunotherapy and Drug-Targeted Therapy: Analysis of Single-Cell and Bulk Sequences. Front. Cell Dev. Biol. *10*, 829029. https://doi.org/10.3389/fcell.2022.829029.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
BC bulk RNA-seq	https://gdc.cancer.gov/	dbGaP: phs000892
GBM bulk RNA-seq	https://gdc.cancer.gov/	dbGaP: phs001287
HCC bulk RNA-seq	https://ega-archive.org/	Access ID: EGAS00001005074
LSCC bulk RNA-seq	http://www.iprox.org/index	Access ID: IPX0001833000
UCEC bulk RNA-seq	https://gdc.cancer.gov/	dbGaP: phs001287
BC phosphorylated proteomics data	https://pdc.cancer.gov/	Access ID: PDC000120
GBM phosphorylated proteomics data	https://pdc.cancer.gov/	Access ID: PDC000205
HCC phosphorylated proteomics data	https://www.ebi.ac.uk/	Access ID: PXD025836
LSCC phosphorylated proteomics data	http://www.iprox.org/index	Access ID: IPX0001833000
UCEC phosphorylated proteomics data	https://pdc.cancer.gov/	Access ID: PDC000441
BC scRNA-seq	https://www.ncbi.nlm.nih.gov/	Access ID: GSE180286
LSCC scRNA-seq	https://www.ncbi.nlm.nih.gov/	Access ID: GSE131907
Software and algorithms		
KinPred-RNA	This paper	https://github.com/tibettiger/kinase_prediction
XGBoost	Chen T et al.[15]	https://xgboost.readthedocs.io/en/stable/
R Seurat package	CRAN	https://cran.r-project.org/web/packages/Seurat/index.html
sklearn	Pedregosa et al.[40]	https://scikit-learn.org/stable/
R ggplot2 package	CRAN	https://cran.r-project.org/web/packages/ggplot2/index.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Prof. Tzong-Yi Lee (leetzongyi@nycu.edu.tw).

Materials availability

No new unique reagents were generated in this study.

Data and code availability

- Pan-cancer bulk RNA-seq data and corresponding phosphorylated proteomics data have been deposited at **the NCI's Genomic Data Commons** (https://gdc.cancer.gov/, https://pdc.cancer.gov/), the European Genome-phenome Archive (https://www.ebi.ac.uk/), iProx (http://www.iprox.org/index) and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- Single-cell RNA-seq data have been deposited at GEO (https://www.ncbi.nlm.nih.gov/) and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- All original code has been deposited at https://github.com/tibettiger/kinase_prediction and is publicly available as of the date of publication.
- Additional information to reanalyze the data reported in this paper is available from the lead contact.

METHOD DETAILS

Data collection and preprocessing

The rigorous data collection process for the development of the kinase activity prediction models was critical to this study. To ensure the accuracy and reliability of the model, a large collection of independent cancer samples was obtained from recent research studies,[28,41–45] including breast cancer (BC), glioblastoma multiforme (GBM), hepatocellular carcinoma (HCC), lung squamous cell carcinoma (LSCC), and

uterine corpus endometrial carcinoma (UCEC). For the input RNA-seq dataset, log2 transcripts per kilobase million (TPM) normalized RNA-seq datasets of multiple pan-cancer samples were obtained. Including such diverse cancers allowed for a more comprehensive and nuanced analysis of kinase activity. In addition to the RNA-seq data, the study also included corresponding phosphorylated proteomics datasets to complement the transcriptomic information from previous studies.[28,41–45] Paired bulk RNA-seq data and phosphorylated proteomics datasets were collected in parallel. This comprehensive approach allowed to better understand kinase activities in these five cancers. The kinase activities were then calculated from the phosphorylation proteomics data based on the kinase substrate enrichment method. We downloaded phosphoproteomics data from the National Cancer Institute's Proteomic Data Commons (PDC, https://proteomic.datacommons.cancer.gov/pdc/). Abnormal samples in the data were screened out by quality control analysis. Phosphorylation sites with more than 50% missing values were filtered. K-nearest neighbor (k-NN) imputation was applied to impute the missing values. Then the data was normalized and used for kinase activity analysis. Phosphoproteomics-based single-sample kinase activity was calculated with reference to a single-sample gene function enrichment analysis method. The main steps were to utilize the in-house constructed kinase-substrate dataset with kinase as the name and substrate as the member of the dataset, and utilize the already well-established single-sample gene function enrichment analysis algorithm, plage, to compute the activity of the kinase by R package gene set variation analysis (GSVA).[43] As a valuable resource for future studies, the calculated kinase activity profiles were deposited on GitHub for public access (https://github.com/tibettiger/kinase_prediction/tree/main/data/kinase_activity).

The LINC-L1000 dataset, a set of gene signatures that was reported to represent another 81% non-measured transcripts in the whole genome was used in this study to predict kinase activity in cancer settings.[44] The LINC-L1000 dataset has previously demonstrated efficacy in predicting drug-induced cell viability and drug-drug interactions (DDI),[36] and in this study it outperformed other gene signatures in predicting kinase activity under cancer conditions. To expand the scope of the kinase activity prediction model, scRNA-seq data from invasive breast cancer (GSE180286) and lung adenocarcinomas (GSE131907) was also incorporated into the study.[28,36] The integration of the scRNA-seq data allowed for a more detailed understanding of the cellular mechanisms underlying the initiation and progression of cancer. As an essential resource for future research in this field, the data collection methods used in this study provide a rich and diverse dataset for the development and validation of kinase activity prediction models.

Development of kinase activities prediction models

The development of a reliable tool for the prediction of kinase activity levels in specific cancer types is a critical task, and the process has two major steps. For training the model, 60% of the bulk RNA-seq datasets and corresponding kinase activities' profiles were used as training sets, 20% as testing sets, and 20% as validation sets. In the first step, all of the transcriptomic data is processed to create a refined data set that consists only of the LINCS-L1000 gene signatures that have been shown to be effective in predicting drug-induced cell viability and DDI.[45] Genes not included in this set are excluded to ensure the most accurate prediction of kinase activity. In order to deal with batch effects caused by individual samples, the expression profiles of LINCS-L1000 genes were 0-1 normalized per sample base before used as features.

In the second step, the LINCS-L1000 transformed gene expression dataset is applied to four powerful prediction models, XGBoost regression, RF regression, multiple linear regression, and SVM regression.[15–17] RF regression, multiple linear regression and SVM regression are well-established methods for modeling the relationship between dependent and independent variables, while XGBoost has demonstrated its effectiveness in predicting both continuous and discrete data. Through analysis of the gene expression levels in each sample, XGBoost generates a prediction score for each leaf node of the decision tree. This score is then used to construct multiple weak estimators through numerous iterations, allowing the model to accurately predict kinase activity levels under specific cancer conditions. The prediction of kinase activity is then defined as the sum of the prediction scores of all trees, as shown below.

$$KA = \sum_{k=1}^{K} f_k(\text{sample}_i[\text{GEL}]) \quad \text{(Equation 1)}$$

KA, $f_k(\text{sample}_i[\text{GEL}])$, and K represent kinase activity values, predicted score of k-th decision tree for i-th sample on LINCS-L1000 transformed gene expression values, and number of decision trees, respectively. Then, the prediction score KA of the model can be described as follows during the t-th iteration of the sample.

$$KA(t) = KA(t-1) + f_t(\text{sample}_i[\text{GEL}]) \quad \text{(Equation 2)}$$

To ensure highest accuracy and reduce prediction bias, 5-fold cross-validation was performed using GridSearchCV from Python package scikit-learn version 0.19.1.[46] The adjustable parameters of the XGBoost model, including max_depth and n_estimators, were optimized to achieve the best performance in predicting the corresponding kinase activity levels. The max_depth was set in the range of 4 to 7 and the n_estimators in the range of 100 to 500, allowing an extensive search for the best combination of parameters. This rigorous parameter tuning approach helped to reduce prediction bias and improve the predictive accuracy of the model, making it more effective at identifying kinase activity associated with cancer.[30,31]

A comprehensive approach was used to evaluate the effectiveness of different gene signatures in predicting kinase activity in different cancers. XGBoost regression, RF regression, multiple linear regression, and SVM regression were used to develop models for each type of gene signature. R^2, root mean square error (RMSE), and mean absolute error (MAE) were used as the performance metrics to evaluate the accuracy of the models in predicting kinase activity in five different cancer types: BC, GBM, HCC, LSCC, and UCEC. R^2 represents the proportion of variation in the dependent variable, kinase activities, explained by the independent variable, gene expression levels. It ranges

from negative infinite to 1, where 1 indicates a perfect fit between the predicted and actual values. Therefore, a more accurate prediction of kinase activity is indicated by a higher R^2 value. The R^2 formula used in this study is shown below.

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} \quad \text{(Equation 3)}$$

$$SS_{residual} = \sum_i (y_i - \hat{y}_i)^2 \quad \text{(Equation 4)}$$

$$SS_{total} = \sum_i (y_i - \overline{y})^2 \quad \text{(Equation 5)}$$

$SS_{residual}$ represents the residual sum of squares, SS_{total} represents the total sum of squares, y_i represents the actual number of targets, $\hat{y}_i$ represents the predicted number of targets, and $\overline{y}$ represents the average number of all targets.

RMSE means root mean square error and MAE means mean absolute error. Both the two metrics could be used for measuring the difference between true or predicted values. The RMSE and MAE formula used in this study is shown below.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Yi - Xi)^2} \quad \text{(Equation 6)}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|Yi - Xi|}{n} \quad \text{(Equation 7)}$$

n represents the total number of samples, X_i represents the actual value of the sample, and Y_i represents the predicted value of the sample.

Classifying five types of cancer using kinase activities

To gain a deeper understanding of the differences between the five types of cancer, classification models based on different kinase activities were developed. The models were constructed using the 10 kinase activities that had the highest R^2 values from the prediction models as the input features including ABL1, MAPK3, CDK2, CAMKK2, TAOK3, MAPK1, MAPK5, ERN1, PLK1 and TNK2. In this study, we developed an XGBoost multi-classification model and five binary classification models. The XGBoost model serves the purpose of simultaneously classifying all five types of cancer, to obtain an organized overview. Conversely, each binary model is specifically tailored to differentiate between one type of cancer and the remaining four. For example, one of our models focuses on BC compared to the other types. To assess the effectiveness of our models in binary classification tasks, we compared their performance against four widely utilized machine learning algorithms: XGBoost classification, RF classification, logistic regression, and SVM classification. Through this comparison, we evaluated the relative strengths and weaknesses of our models as compared to established methods in the field. The AUC was used to evaluate the performance of the binary classifier. AUC measures the ability of the model to discriminate between positive and negative samples and is a common metric used to evaluate the performance of classification models. It ranges from 0 to 1, with 1 indicating perfect classification. A higher AUC value indicates better classification performance. F1-score, accuracy and MCC were also utilized to measure the binary classification performance.

All kinases were ranked according to their predictability, as assessed by the R^2 values of the kinase activity prediction model, to identify the most important kinases for cancer classification. The top 10 kinases with the highest predictability were selected as input features for the classifier, while the bottom 10 kinases including PRKACB, MAP2K2, BRAF, BMP2K, PAK4, RET, IKBKB, RAF1, PIM2 and PIM3 were also selected to demonstrate the superiority of the top 10 kinases in classifying specific cancer types. In short, these approaches allowed for a comprehensive evaluation of the effectiveness of different kinases in predicting and classifying specific cancer types.

Analysis of kinase activity profiles predicted by KinPred-RNA for breast cancer tissues and lung cancer tissues

In this study, we applied the KinPred-RNA model to infer cell-specific kinase activity in scRNA-seq datasets, with the goal of evaluating the effectiveness of the KinPred-RNA model, which was originally developed using bulk RNA-seq data. To determine the validity of the model's performance, we focused on two cancer tissues: invasive breast cancer (GSE180286)[47] and lung cancer (GSE131907).[36] The composition of cancer tissues includes a variety of cell types. These include immune cells, epithelial cells, and stromal cells. To accurately identify and label these cell types in breast cancer tissue, we used well-established marker genes such as PTPRC for immune cells and EPCAM for epithelial cells.[30,31] In addition, to improve our understanding of cell composition, we used Seurat[20] to cluster and analyze multiple cell groups within the breast cancer scRNA-seq datasets. Each cell group was manually identified and annotated to ensure the accuracy of cell type labeling. We analyzed scRNA-seq data from both tumor (LUNG_T08) and normal (LUNG_N08) tissue samples of lung cancer tissue to determine differential kinase activation patterns. The scRNA-seq data was aligned with the pretrained KinPred-RNA model through normalization of the data by transforming the log2 transcripts per kilobase million (TPM) gene expression values to a 0-1 scale for each cell. Our analysis identified multiple kinases that displayed varied activation levels between immune and epithelial cells. Notably, kinases linked with tumor invasion exhibited

distinctive activation patterns. This discovery is significant for cancer diagnosis and treatment as it introduces new kinases that may function as biomarkers for discerning between different cell categories within cancerous tissues.

QUANTIFICATION AND STATISTICAL ANALYSIS

Model construction and computations were performed in the Python programming language. The graphic abstract and Figure 1 were generated by Microsoft PowerPoint. R ggplot2 package was used for generating all other plots appearing in this study. scRNA-seq dataset analyses and UMAP plot generating are performed using the standard pipeline of R package Seurat. In the comparison of phosphorylation level means in each group, Student's t test was performed.