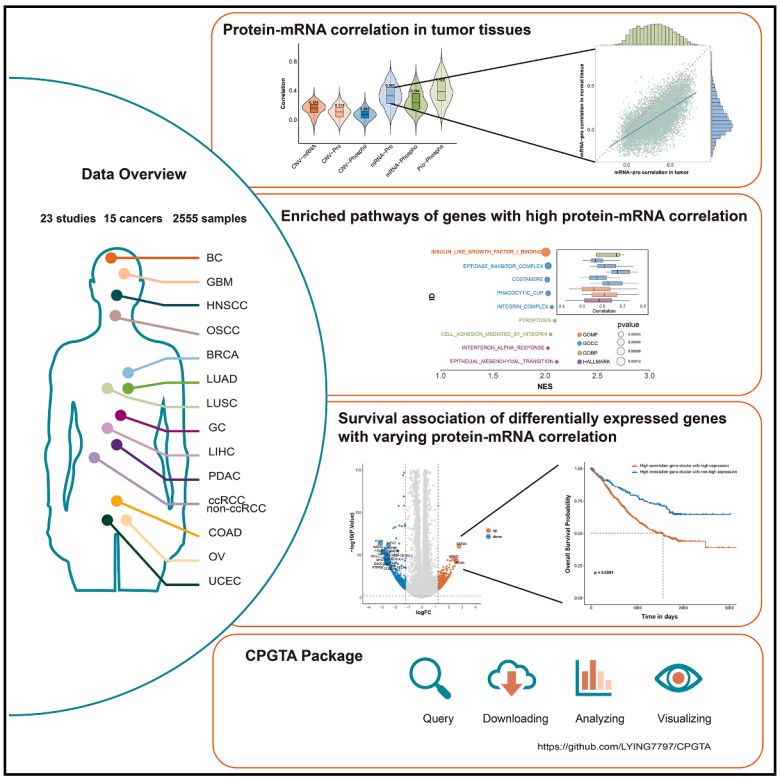
iScience

Article

A data-driven pan-cancer proteogenomic analysis reveals the characteristics of human cancer protein expression

Graphical abstract



Highlights

- Large-scale pan-cancer matched multi-omics resource of 2,555 tumors across 15 cancer types
- Systematic assessment of protein-mRNA concordance across key tumor biological processes
- Differentially expressed proteins with high mRNA concordance link to adverse clinical outcomes
- CPGTA R package enables streamlined and integrative pan-cancer multi-omics analysis

Authors

Ying Li, Xin Liu, Zhen Wang, Qiong Wu, Chulin Sha, Xiaolin Li

Correspondence

shachulin@gmail.com (C.S.), xiaolinli@ieee.org (X.L.)

In brief

Computational bioinformatics; Systems biology; Omics

Li et al., 2026, iScience *29*, 114500
January 16, 2026
https://doi.org/10.1016/j.isci.2025.114500

CellPress

iScience

CellPress
OPEN ACCESS

Article

A data-driven pan-cancer proteogenomic analysis reveals the characteristics of human cancer protein expression

Ying Li,[1,2,3] Xin Liu,[1] Zhen Wang,[1] Qiong Wu,[1] Chulin Sha,[1,*] and Xiaolin Li[1,2,4,5,*]
[1]Hangzhou Institute of Medicine, Chinese Academy of Sciences, Hangzhou, Zhejiang 310018, China
[2]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, Zhejiang 310024, China
[3]University of the Chinese Academy of Sciences, No.1 Yanqihu East Rd, Huairou District, Beijing 101408, P.R. China
[4]Institute of AIX, College of Statistics and Data Science, School of Information and Electronic Engineering, Sussex Artificial Intelligence Institute, Zhejiang Gongshang University, Hangzhou, Zhejiang 314423, China
[5]Lead contact
*Correspondence: shachulin@gmail.com (C.S.), xiaolinli@ieee.org (X.L.)
https://doi.org/10.1016/j.isci.2025.114500

SUMMARY

Genomics and epigenomics outline potential cellular changes, while proteomics reflects actual molecular events. To systematically bridge the molecular hierarchies and validate their functional interplay, we established the most comprehensive pan-cancer paired multi-omics resource to date, systematically integrating proteomic, transcriptomic, and genomic data from both tumor and adjacent normal tissues spanning 15 cancer types (2,555 tumor samples), thereby enabling a robust cross-omics exploration. Analysis revealed that tumor tissues exhibit higher correlation between transcriptomic and proteomic expression levels compared to normal tissues. Key tumor development pathways exhibited strong mRNA-protein correlations. Genes with high mRNA-protein correlation and high expression were associated with lower survival rates, highlighting potential therapeutic targets. We developed a comprehensive tool, the CPGTA R package, based on reintegrated datasets that facilitates multi-omics data integration and reanalysis. Our research enhances cancer molecular characterization while providing insights into mechanisms underlying cancer progression and treatment resistance.

INTRODUCTION

Continuous advancements in sequencing technologies and informatics over the past two decades, following the introduction of a high-quality human reference genome, have firmly established genomics as a cornerstone in cancer research.[1] Moreover, the proliferation of large-scale omics data resources, such as The Cancer Genome Atlas and the International Cancer Genome Consortium, has significantly facilitated pan-cancer multi-omics research in recent years.

The Pan-Cancer Atlas, through its comprehensive interrogation of more than 11,000 tumor samples spanning 33 prevalent cancer types, has defined cancer drivers[2,3] and elucidated the cells of origin[4] and molecular features[5–8] across various cancer types. Additionally, it has identified commonalities in the tumor microenvironments (TMEs)[9] among different cancer types and explored the clinical impacts of these pan-cancer studies.[10] These works focused predominantly on genomic data resources, have yielded profound insights into the molecular underpinnings of cancer, thereby contributing substantially to the field of cancer genomics. However, an exclusive focus on genomics has its limitations. Observations of genomic alterations often presume downstream phenotypes that may not be empirically validated.[11] Moreover, aberrations at the protein and phosphorylation levels frequently occur post-transcription or post-translation,[12,13] underscoring the complexity of cancer biology beyond genomic alterations.

Concurrently, proteomics has evolved as a complementary counterpart and extension to genomics. Mass spectrometry-based proteomics techniques can provide a global profile of protein expression and post-translational modifications. The inception of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) has accelerated proteogenomic research into cancer by integrating a comprehensive array of omics data, with a primary focus on proteomic resources, including whole-exome and whole-genome sequencing, epigenomic, transcriptomic, and phosphoproteomic, thereby facilitating a deeper understanding of the molecular mechanisms of cancer. Li et al. undertook a multi-omics pan-cancer analysis to assess the impact of oncogenic drivers on the functional states of cancer, revealing potential new therapeutic avenues.[14] Geffen et al. delineated the regulatory patterns that are both shared and distinct across different cancer types from the perspective of proteins and post-translational modifications (PTMs), underscoring the pivotal role of PTMs in cancer progression.[15]

Currently available data resources integrating CPTAC-derived proteomics data include CancerProteome,[16] a specialized platform that aggregates mass spectrometry-based proteomics data across 29 cancer types to functionally analyze

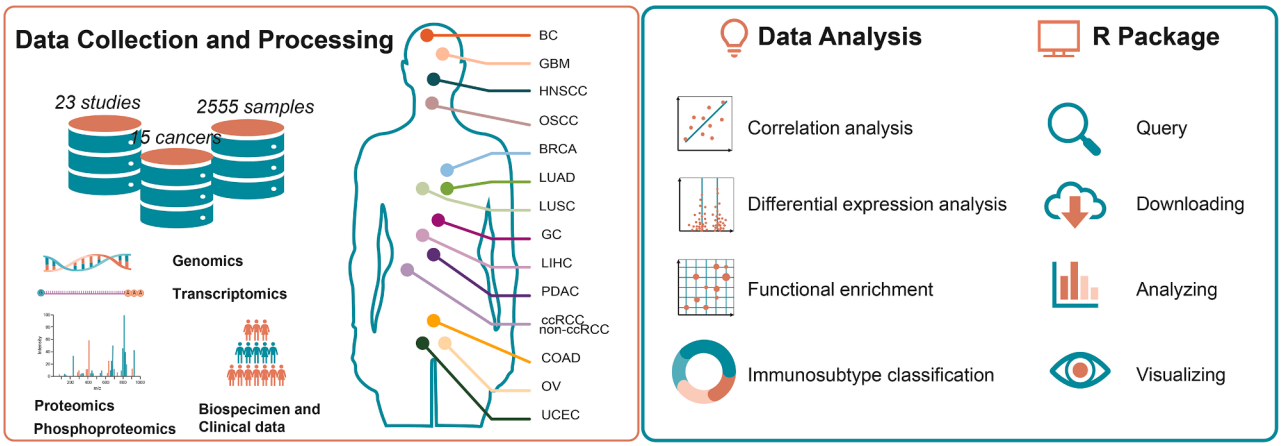


Figure 1. Overview of the 15 cancer types analyzed, the multi-omics datasets, and the key functionalities of the CPTGA package

protein/microprotein abundances and PTM profiles, though its focus remains primarily on proteomic and phosphoproteomic data with limited inclusion of other omics data types. In contrast, LinkedOmicsKB[17] is a multi-omics knowledge base utilizing CPTAC pan-cancer data (1,000+ tumors spanning ten cancer types) to enable cross-omics exploration of genes, mutations, phosphorylation sites, and phenotypes through precomputed analytics and interactive visualization tools; however, its dataset is exclusively derived from the CPTAC initiative, resulting in a relatively limited sample size.

Our study addresses these gaps by establishing the most comprehensive pan-cancer multi-omics resource to date, spanning 15 cancer types and comprising 2,555 tumor samples (sourced from CPTAC and external studies) with some matched adjacent normal tissues. All integrated omics data were processed and standardized through a unified analytical pipeline. This cohort features quadra-modal molecular profiling—genomic, transcriptomic, proteomic, and phosphoproteomic characterizations.

Leveraging our systematically reintegrated quantitative proteomic datasets, we aim to construct a pan-cancer global protein landscape, thereby enabling comprehensive exploration of multi-omics correlations and functional insights. Building upon this foundation, we conducted differential expression analysis at the pan-cancer proteomic level, where the proteogenomic correlations of differentially expressed proteins are widely distributed. Notably, highly correlated, differentially upregulated proteins demonstrate significantly greater impact on clinical outcomes. Given the relatively low mRNA-protein correlations observed in tumors, we subsequently compared mRNA immunophenotyping with proteomic phenotyping results at the pan-cancer level and specific cancer, revealing that immunophenotyping based on the transcriptome may exhibit superior prognostic predictive value. To facilitate multi-omics data integration and advance pan-cancer research, we developed the R package CPGTA, a comprehensive tool for cancer studies based on reintegrated datasets. Using CPGTA, we conducted a systematic analysis of lung squamous cell carcinoma (LUSC) through mutation characterization ("cptsnv"), differential expression analysis ("cptda"), cross-cancer profiling ("cptgev"), transcriptome-proteome correlation analysis ("cptca"), and survival assessment ("cptsa"), validating keratin family markers (KRT5, KRT6A/B/C, and KRT14) in LUSC. These markers showed significant upregulation in LUSC with stronger mRNA-protein level correlation in tumor tissues. CPGTA provides a powerful resource for validation and discovery, offering opportunities for integrative multi-omics investigations in precision cancer medicine.

Our work promises to augment the understanding of the molecular characterization of cancer and offer new insights into the potential mechanisms driving cancer progression and treatment resistance.

RESULTS

Data profile

To investigate the regulation of protein activities of cancer, we compiled, organized, and standardized multi-omics datasets provided by the CPTAC consortium and other cohorts. These datasets encompass cancer patient samples with matched gene copy number variation, mRNA expression, protein abundance, phosphorylation and clinical data from 15 cancer types: brain cancer,[18] breast carcinoma (BRCA),[19,20] clear cell renal cell carcinoma (ccRCC),[21] colorectal adenocarcinoma (COAD),[22] glioblastoma,[23–25] head and neck squamous cell carcinoma (HNSCC),[26] lung adenocarcinoma,[27–29] LUSC,[30] ovarian high-grade serous carcinoma (OV),[31,32] pancreatic ductal adenocarcinoma (PDAC),[33,34] gastric cancer (GC),[35] hepatocellular carcinoma,[36] uterine corpus endometrial carcinoma,[37,38] non-clear cell renal cell carcinoma (non-ccRCC),[39] and oral squamous cell carcinoma.[40] In summary, the assembled data consists of 2,555 tumor samples have available data for proteomics data used in this study. For each analysis, we used all samples having data for the types of omics measurement required, resulting in small variations of the total number of samples used. The pan-cancer-level proteome dataset involves 15 types of cancer, comprising 2,555 samples, with 15,316 proteins detected. Among these, 10,208 proteins were identified in seven or more cancer types, and 5,418 proteins were detected in all 15 cancer types (Figure 1). The multi-omics data utilized in this study is detailed in Table S1.

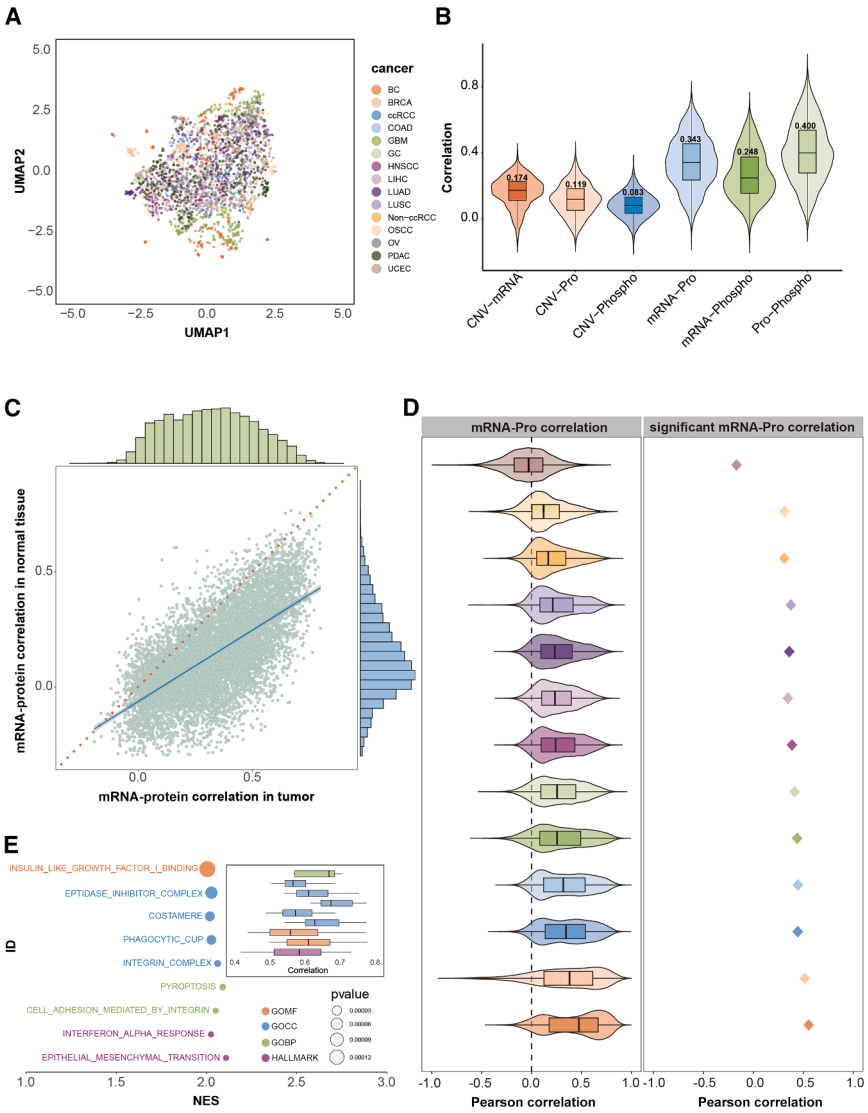


Figure 2. Multi-omics correlation analysis reveals enhanced mRNA-protein correlation in tumor tissues

(A) UMAP plots of normalized proteomics data originating from different cancer types.

(B) The Pearson's correlation between different omics in pan-cancer.

(C) The mRNA-protein level Pearson's correlation between tumor and normal tissues in pan-cancer level.

(D) The mRNA-protein level Pearson's correlation in tumors of different cancers and the median of significant correlations ($p < 0.05$).

(E) Pan-cancer level: GSEA enrichment analysis for genes sorted by mRNA-protein level Pearson's correlation in tumor tissues (colored font); the mRNA-protein level Pearson's correlation distribution of genes involved in the corresponding enriched pathways (upper right corner) (NES >2, p <0.05).

Multi-omics correlation analysis reveals characteristic patterns of transcriptome-proteome expression in tumor tissues

We conducted a comprehensive investigation of the correlations among multi-omics data at the pan-cancer level. UMAP[41] visualization of normalized pan-cancer proteomic and transcriptomic data demonstrated effective integration of most cancer types from various studies, indicating the normalization approach efficiently mitigated batch effects (Figures 2A, 3A, S1A, and S1B). Pairwise correlation analyses were performed on multi-omics data from 23 studies covering 15 cancer types. Figure 2B illustrates a low to non-existent correlation of copy number variation (CNV) with other omics data in tumor tissues, with the highest correlation observed between the proteomic and the phosphoproteomic. This likely arises because phosphoproteomic expression is dependent on proteome levels. The median correlation between the transcriptomic and proteomic across pan-cancer tissues is 0.343, aligning with findings from previous studies.[42–44] We further validated the correlation between proteomic and transcriptomic expression in cancer cell lines, and the results were highly consistent with the conclusions obtained at the tumor tissue level (Figure S1C). Subsequently, we investigated whether the mutation status of 299 key driver genes[45] would affect the correlation between proteomic and transcriptomic expression. The experimental results showed that samples carrying mutant driver genes exhibited a trend toward stronger correlation in mRNA-protein expression of the corresponding gene compared to samples with wild-type driver genes, although the difference was not statistically significant (Mann-Whitney U test p value = 0.427, Figure S1D). At the individual gene level, the mRNA-protein correlations were almost identical between mutant and wild-type genes, showing no obvious differences (Table S2). Furthermore, through in-depth analysis of existing data, we discovered that certain cancer-specific highly expressed markers displayed different mRNA-protein expression correlation patterns in normal tissues versus tumor tissues. Notably, these tumor-specific highly expressed markers demonstrated stronger correlation between mRNA-protein expression in tumor tissues (Figure S1E).

We investigated patterns in the correlation between transcriptomic and proteomic in tumor versus normal tissues across cancer types. Across pan-cancer and within specific cancer types, tumor tissues generally exhibit a higher correlation than normal tissues (Figures 2C, 2D, and 3B). This observation suggests that tumor tissues may exhibit diminished control over post-transcriptional regulation, potentially affecting gene expression and contributing to tumor growth.[42] However, within specific cancers, there are certain differences in the correlation

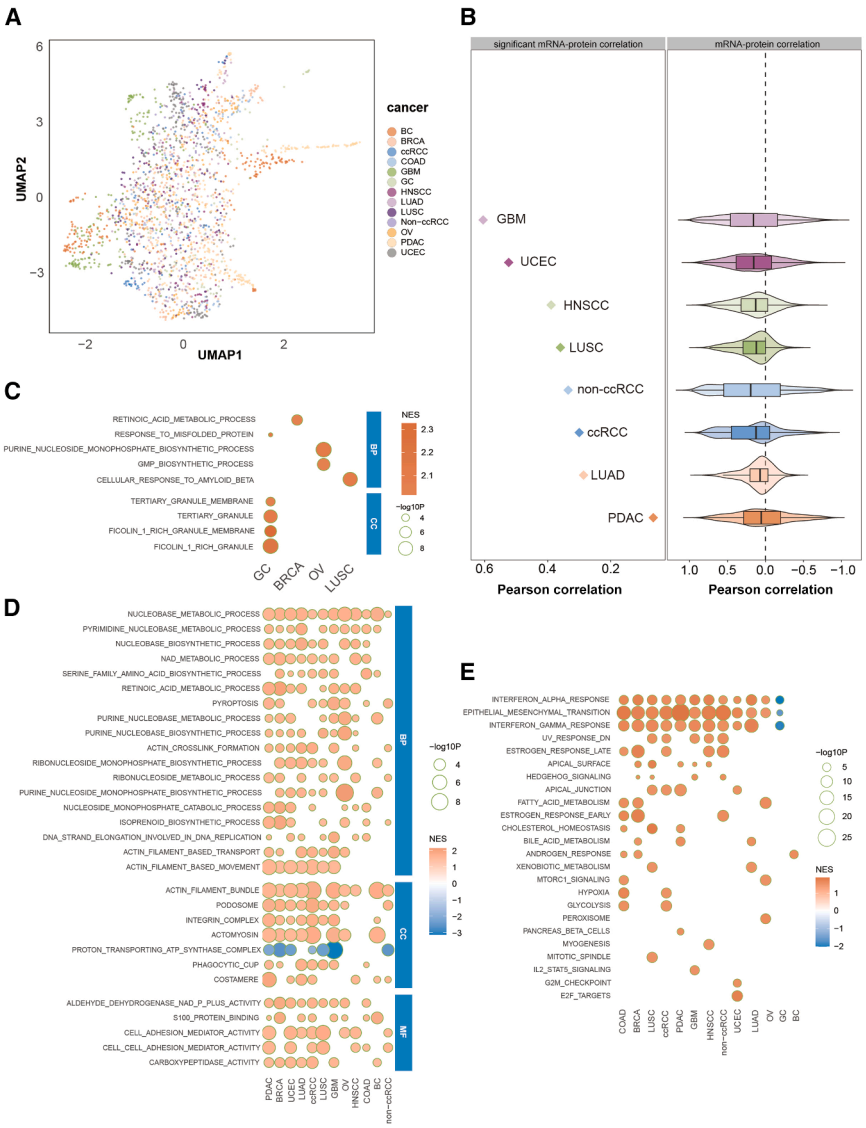


Figure 3. Multi-omics correlation analysis reveals characteristic patterns of transcriptomic-proteomic expression in tumor tissues

(A) UMAP plots of normalized transcriptomics data originating from different cancer types.

(B) The mRNA-protein level Pearson's correlation in normal tissues of different cancers and the median of significant correlations ($p < 0.05$).

(C) GO pathways enriched in certain cancer types (NES >2, p <0.05).

(D) GO pathways enriched in most cancer types (|NES| > 1.5, p <0.05).

(E) HALLMARK pathways enriched across different cancer types (|NES| > 1.5, p <0.05).

between the transcriptomic and proteomic expression level across different cancer types (Table S3). For example, the correlation between mRNA-protein is the poorest in GC, COAD, and PDAC, with median values of-0.03, 0.12, and 0.17, respectively; in other cancer types, the median correlation exceeds 0.2. The highest median correlation, observed in LUSC, is 0.47 (Table S3). These variations may reflect diverse post-transcriptional regulatory behaviors across different tumors.

We next performed enrichment analysis on genes sorted by mRNA-protein correlation to uncover pathway characteristics in the pan-cancer. Pan-cancer level enrichment analysis of genes sorted by mRNA-protein level Pearson's correlation in tumor tissues reveals that genes with high mRNA-protein correlation are primarily enriched in pathways related to Epithelial-Mesenchymal transition (EMT) and cell proliferation, cell adhesion and migration, as well as immune and inflammatory responses (Figure 2E). We found a strong correlation exists between transcriptomic and proteomic expression for genes in pathways closely linked to tumor metastasis and inflammatory responses.

Additionally, the same enrichment analyses were conducted at the specific cancer level to validate findings from the pan-cancer level and identify pathways uniquely enriched across different cancer types. We found that some pathways are distinctly enriched in specific types of cancer. For example, in BRCA, there is a particular enrichment of the RETINOIC ACID METABOLIC PROCESS, while in LUSC, there is a specific enrichment of the CELLULAR RESPONSE TO AMYLOID BETA (Figure 3C). Simultaneously, certain pathways were found to be enriched across almost all cancer types. These pathways were primarily associated with nucleic acid metabolism and biosynthesis, amino acid and fatty acid metabolism, cell adhesion and cytoskeletal dynamics, inflammatory and immune responses, as well as signaling pathways such as EMT, interferon responses, and mTORC1 signaling (Figures 3D and 3E). These findings suggest that key pathways driving the development of different cancers are distinct.

Pan-cancer differential protein expression analysis and the association of transcriptome-proteome correlation with patient prognosis

At the pan-cancer level, proteomic data were employed for T-N (tumor-normal tissue) differential analysis to identify proteins that are highly and specifically expressed in tumor tissues across various cancer types. This differential analysis was conducted using proteomic data from cancer types that had both tissue types available, involving 13 cancer types and selecting genes that appeared in more than 50% of the samples. The differential proteins were identified by the limma method.[46] Notably, several proteins, including ERBB4,[47] ARNT2,[48–50] and MEX3A,[51–54] were identified as being highly expressed across pan-cancer

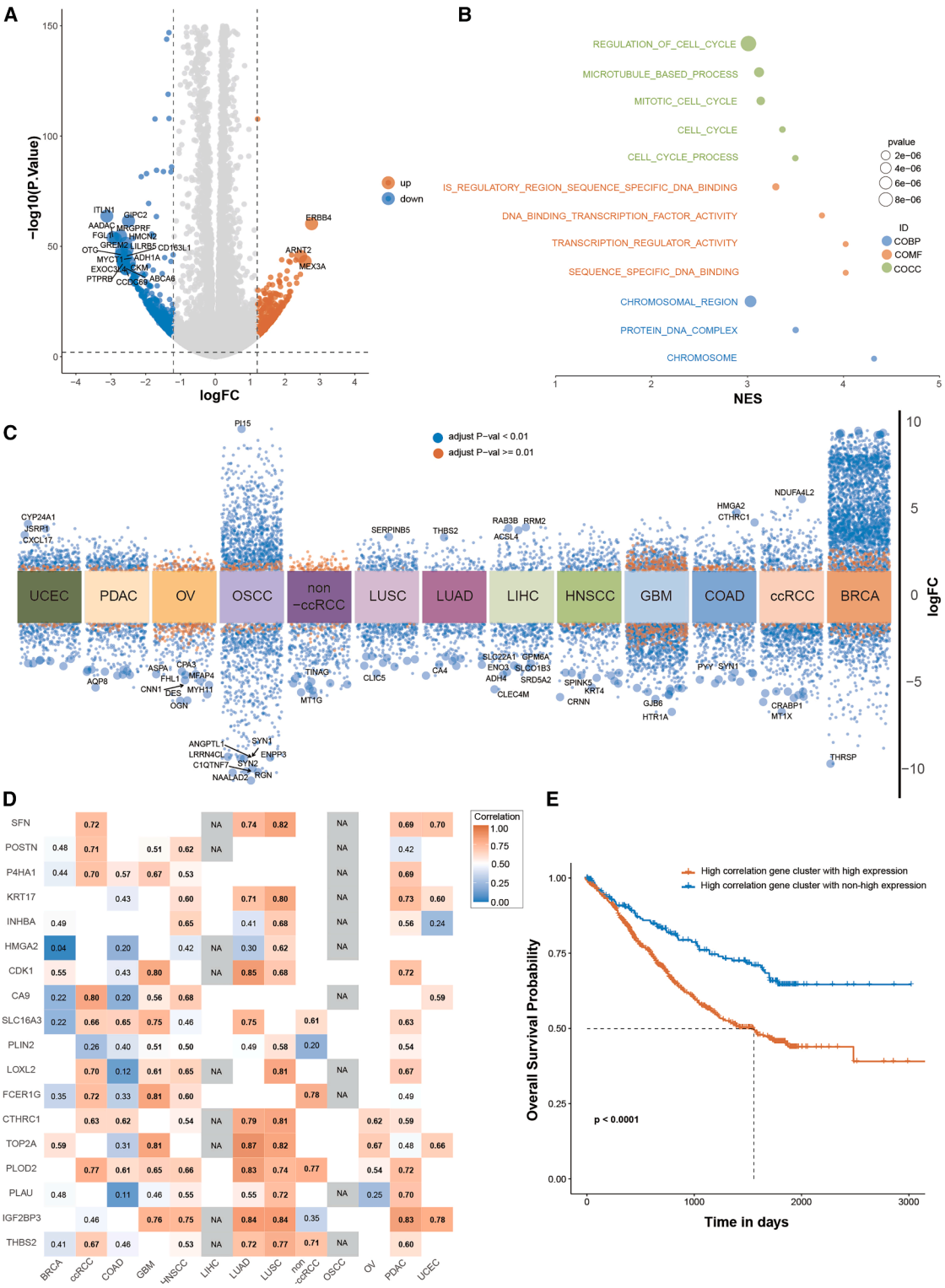


Figure 4. Pan-cancer differential protein expression analysis and the association of mRNA-protein correlation with patient prognosis

(A) Pan-cancer level differential protein analysis (NES >1.2, p <0.05).

(B) Pan-cancer level tumor-specific overexpressed protein GSEA enrichment results.

(legend continued on next page)

contexts (Figure 4A). Subsequently, proteins that are specifically overexpressed at the pan-cancer level underwent pathway enrichment analysis. It was found that these proteins are predominantly enriched in pathways related to cell cycle regulation, transcriptional regulation, and chromosomal stability. These pathways are closely associated with the rapid division, uncontrolled proliferation, and genomic instability characteristic of tumor cells, highlighting their potential roles in cancer progression and development (Figure 4B).

Additionally, differential expression analyses were conducted at the specific cancer level to identify commonly overexpressed proteins (Figure 4C). Furthermore, certain proteins, such as TOP2A,[55–57] CTHRC1,[58–60] THBS2,[61,62] and PLOD2,[63] were found to be upregulated in most cancer types, and this upregulation coincides with the proteins identified as upregulated at the pan-cancer level. These genes exhibited a high correlation between transcriptomic and proteomic expression across the majority of cancer types, suggesting that most tumor-specific highly expressed proteins in cancers exhibit conserved post-transcriptional regulation (Figure 4D). Conversely, proteins specifically downregulated across multiple cancer types displayed a generally lower correlation between transcriptomic and proteomic levels compared to those that were specifically upregulated (Figure 5A).

Next, we used hierarchical clustering to discover similarities in mRNA-protein level correlations across cancer types. A high correlation gene cluster, present in 13 cancer types, was identified (Figure 5B). This gene cluster is primarily enriched in pathways related to the cytoskeleton, immune regulation, and inflammatory response. (Figure 5C). Finally, we investigated whether mRNA-protein level correlations of tumor highly expressed genes at the patient level were associated with variations in overall survival across cancer types. Screening pan-cancer level tumor highly expressed genes from high correlation gene clusters for panel; we found that highly correlated, highly expressed genes were associated with lower and shorter survival (Figure 4E). In contrast, if non-highly correlated genes were selected, no significant association was observed between them (Figure 5D). The above results suggest that differential genes with high mRNA-protein level correlations can better predict patient prognosis.

Multi-omics classification of tumor microenvironment and its prognostic significance

Alexander Bagaev and colleagues have previously classified the TME by analyzing transcriptomes of over 10,000 patients, identifying four conserved TME subtypes across 20 different cancers.[64] Given that the proteome represents a more direct expression form, we extended this classification at the pan-cancer level using both transcriptomic and proteomic data. Our findings indicated that this classification method presents conserved subtype categories in both transcriptome and proteome (Figure 6A). However, the concordance between RNA TME and Pro TME classifications was low, with an overall agreement of 44.98%. The highest concordance was observed for subtype D, reaching 62.8% (Figure 6B). Next, we observed the consistency of the two types of TME in each cancer type, with the highest being non-ccRCC (60.976%) and HNSCC (57.407%) (Figure 6C). From this, we can infer that in non-clear cell renal cell carcinoma, genes related to the immune microenvironment are relatively conserved between the proteome and transcriptome.

To elucidate the clinical relevance of our findings, we focused on the relationship between RNA TME and Pro TME subtypes and their association with survival outcomes across pan-cancer. At the pan-cancer level, the correlation between RNA TME and survival events showed that the IE and IE/F subtypes had better survival outcomes compared to the D and F subtypes, although the survival differences between groups did not reach statistical significance (Figure 6D). In contrast, Pro TME showed no notable differences in survival outcomes across immune subtypes. Further independent analyses of data from different studies yielded consistent conclusions, with RNA TME showing statistically significant survival differences between subtypes in certain cancer types, such as non-ccRCC (Figure 6E). These findings suggest that transcriptome-based immunotyping may provide greater clinical relevance in certain cancer types, offering potentially more precise prognostic and therapeutic insights compared to proteome-based classifications.

CPGTA: A pan-cancer proteogenomics database for comprehensive multi-omics analysis

To facilitate public access and reuse of the data, we have consolidated all processed multi-omics data, along with associated analyses, into a comprehensive R package, named CPGTA. This package is designed to enable efficient download, query, and utilization of multi-omics data. CPGTA primarily comprises five functions.

1. cptod (clinical proteogenomic tumor multi-omics data download): this function allows users to download the integrated datasets, ensuring easy access to comprehensive multi-omics data, including CNV, transcriptomics, proteomics, phosphoproteomics, clinical data, and biospecimen.
2. cptca (clinical proteogenomic tumor correlation analysis): this tool facilitates the exploration of correlations within the multi-omics data, providing insights into the interdependencies between different data types, such as the relationship between transcriptomic expression and proteomic levels.
3. cptda (clinical proteogenomic tumor differential analysis): users can perform differential analysis to identify significant changes between conditions or groups within the data. This function also generates differential volcano plots, which visually highlight the most significant molecular alterations.

(C) Specific cancer type multi-group differential analysis (NES >3, p <0.05).
(D) Heatmap of mRNA-protein level correlation for differentially upregulated proteins shared in specific cancers (NA indicates missing correlation data, blank indicates no correlation data for that gene.
(E) Overall survival pan cancer stratified by high correlation gene clusters with high expression group and non-high expression group.

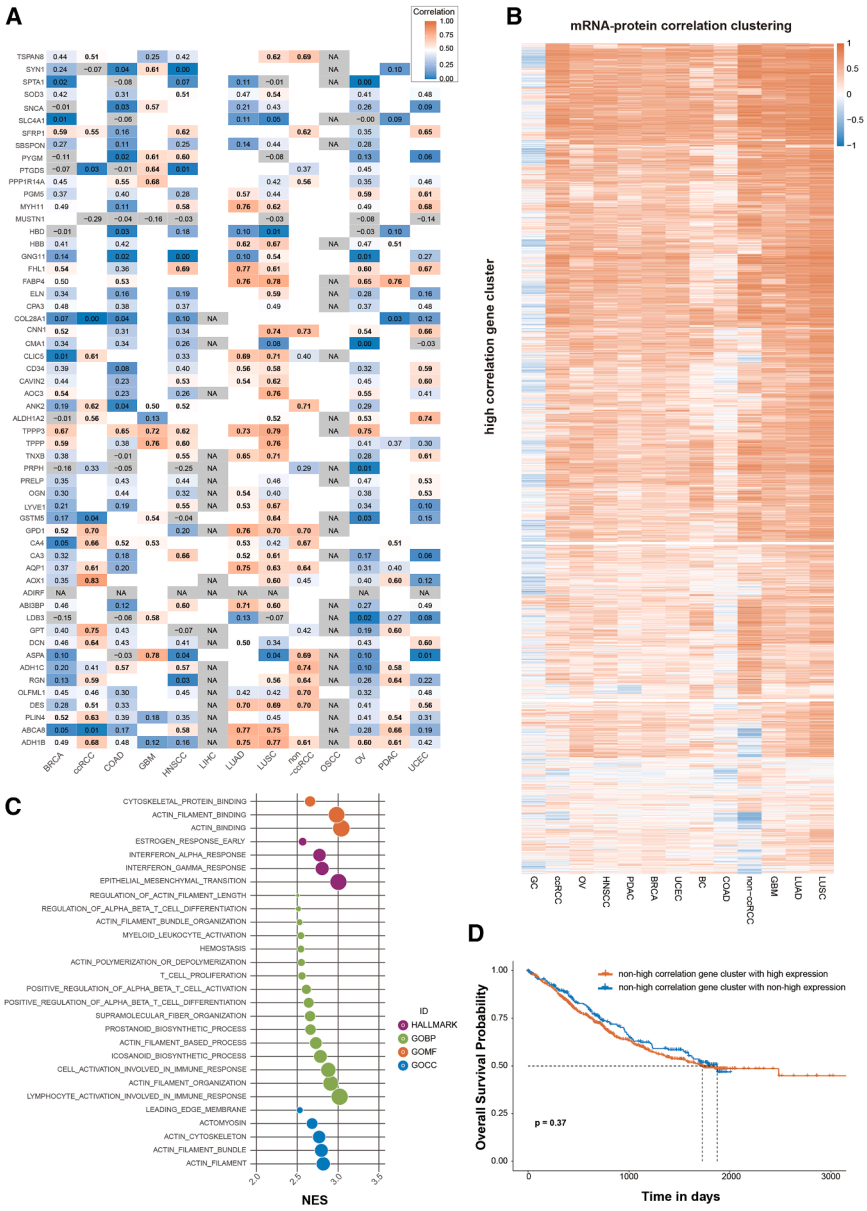


Figure 5. The functional and prognostic significance of high correlation gene clusters

(A) Heatmap of mRNA-protein level correlation for differentially downregulated proteins shared in specific cancers (NA indicates missing correlation data, blank indicates no correlation data for that gene).

(B) Hierarchical clustering of mRNA-protein gene-level correlations across 13 cancer types.

(C) Gene in high correlation gene cluster GSEA enrichment results (NES >2.5, p <0.05).

(D) Overall survival pan cancer stratified by non-high correlation gene clusters with high expression group and non-high expression group.

sualizations, such as mutation frequency plots and heatmaps, to help uncover potential mutation-based biomarkers for patient stratification and therapeutic insights.

7. cptgev (clinical proteogenomic tumor gene expression violin plot): this function generates comprehensive violin plots to visualize and compare expression levels of user-specified genes between tumor and normal tissues across multiple cancer types. It supports visualization of both transcriptomic and proteomic data, enabling direct comparison of gene expression patterns at RNA or protein levels. The function allows researchers to easily identify differential expression signatures of target genes across various cancer types and their corresponding normal tissues.

The CPGTA package integrates multiple functionalities enabling researchers and clinicians to conduct comprehensive multi-omics analyses that advance tumor biology understanding and personalized therapeutic strategies.

4. cptic (clinical proteogenomic tumor immunosubtype classification): this function classifies tumor samples based on their immunological profiles, aiding in the understanding of the TME.
5. cptsa (clinical proteogenomic tumor survival analysis): this function allows for the analysis of survival data, correlating clinical outcomes with molecular profiles. It can generate survival curves to visualize the relationship between molecular markers and patient prognosis, helping to identify potential biomarkers for patient stratification.
6. cptsnv (clinical proteogenomic tumor single-nucleotide variations [SNV] analysis): this function enables the analysis of somatic SNVs in cancer. It supports the identification of key mutations by integrating data from multiple PDC identifiers or cancer types. The function generates vi-

Using LUSC as a demonstration dataset, we illustrated the performance of the CPGTA package in resource-based validation and discovery. The ‘‘cptod’’ function retrieved diverse LUSC data (proteomics, transcriptomics, mutation, phosphoproteomics, and clinical data). Mutation analysis via ‘‘cptsnv’’ identified SNPs as predominant mutations, with C>T and C>A being most common—potentially linked to smoking-induced DNA damage. Higher mutation burden samples showed potential immunotherapy sensitivity (Figure 7A). TP53 and TTN emerged as the most frequently mutated driver genes (Figure 7B), while co-occurrence analysis revealed potential cooperative gene pairs in tumor progression (Figure 7C). We conducted systematic validation and extended analysis of known LUSC specific markers, including keratin family members KRT5, KRT6A, KRT6B, KRT6C, and KRT14. The

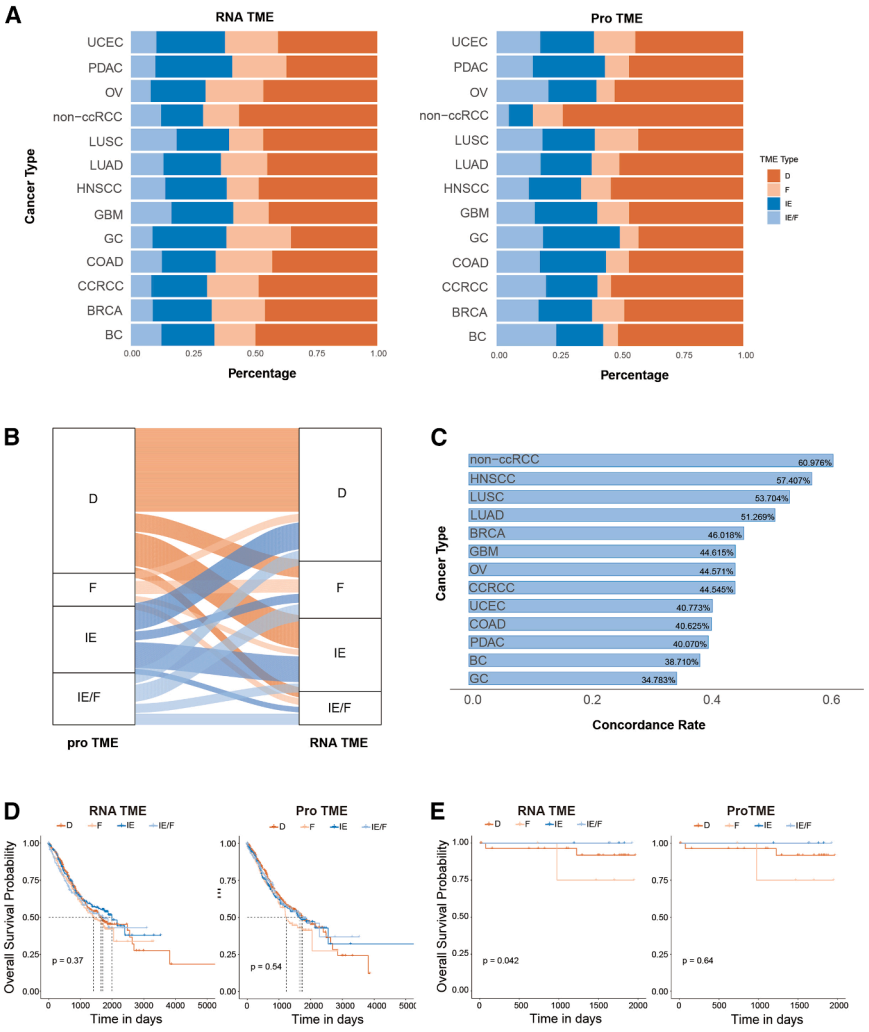


Figure 6. Comparison of transcriptomic and proteomic tumor microenvironment classifications

(A) Bar graphs depicting segregation of 13 cancer types into the four RNA TME and pro TME subtypes.

(B) The Sankey diagram shows the consistency between pro TME and RNA TME subtypes.

(C) The consistency ratio of pro TME and RNA TME subtypes across 13 cancer types.

(D) Overall survival (OS) analysis of RNA TME and Pro TME subtypes (D, F, IE, and IE/F) at the pan-cancer level.

(E) Overall survival (OS) analysis of RNA TME and pro TME subtypes (D, F, IE, and IE/F) in non-ccRCC cases.

DISCUSSION

We have established the most comprehensive pan-cancer paired multi-omics data resource to date. Through integrative analysis of harmonized multi-omics datasets (n = 2,555 patients across 15 cancer types) derived from CPTAC and other public cohorts, our work systematically delineates intricate associations between multi-omics layers, particularly proteogenomic dynamics, at the pan-cancer scale.

We focused on the correlation between the transcriptome and proteome, revealing a significantly higher correlation in tumors compared to normal tissues, both at the pan-cancer and individual cancer levels. Notably, slight variations were observed in the correlations among different cancer types. Additionally, we observed that specifically highly expressed markers demonstrated a more correlated trend between transcriptomic and proteomic in tumor tissues compared to normal tissues. We also identified that the genes exhibiting high mRNA-protein level correlation in tumors were mainly enriched in EMT, cell adhesion and inflammatory response-related pathways, indicating the conservation of these genes during transcription and translation phases.

We also found that differential proteins at the pan-cancer level could be corroborated with proteins obtained from single cancer differential analysis, and that genes with high expression of proteins in tumors were generally highly correlated in mRNA-protein level. Consequently, we screened these highly correlated genes and conducted survival analyses for cases with available clinical data. We found high expression of upregulated differential proteins in tumors with high correlation was associated with poorer overall survival. This observation could provide new insights into identifying protein markers for prognostic prediction or drug target screening.

We next explored the characterization of pan-cancer immunophenotyping using transcriptomic and proteomic data separately and found that the concordance between the two

"cptda" function confirmed their significantly elevated expression in LUSC tissues, consistent with previous literature reports (Figure 7D). Cross-cancer analysis with "cptgev" showed these keratin markers were upregulated in both LUSC and HNSCC, reflecting shared squamous epithelial molecular characteristics (Figures 7E and S2A). Transcriptome-proteome integration analysis ("cptca") revealed stronger positive correlations between mRNA-Protein in tumor versus normal tissues (Figures 7E and S2A), indicating tighter transcriptional-translational coupling in LUSC. Survival analysis ("cptsa") showed no significant correlation between individual marker expression and patient outcomes (Figure S2B), suggesting these keratins primarily function in tumor differentiation rather than driving progression or treatment response.

This workflow demonstrates CPGTA's effectiveness for cancer multi-omics analysis. Our findings validated these keratin markers for LUSC while revealing their molecular regulation patterns and clinical significance, highlighting CPGTA's utility for integrated validation and advancing precision oncology through comprehensive multi-omics analysis.

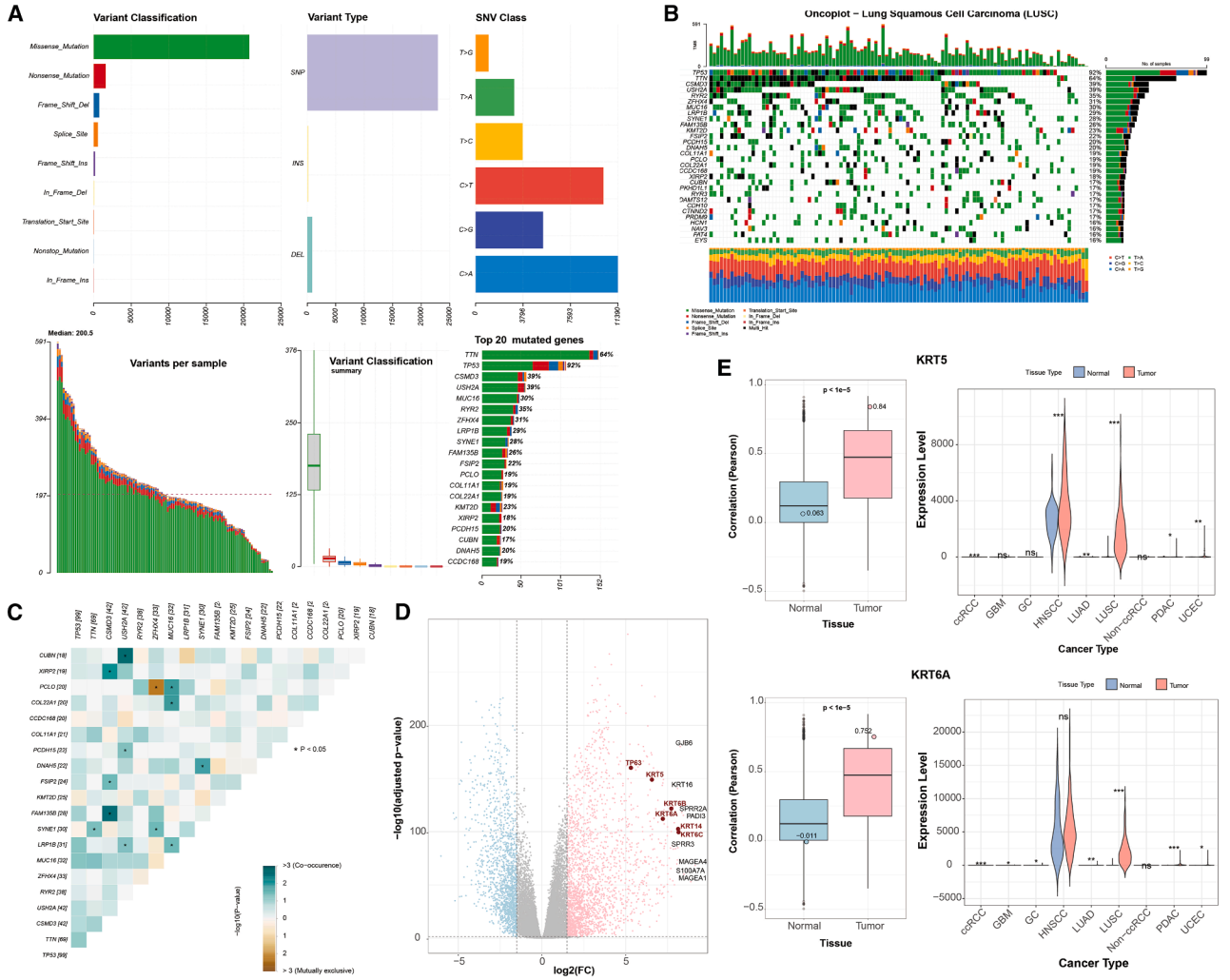


Figure 7. CPGTA package demonstration: multi-omics integration of LUSC mutation spectrum and expression profiles

(A) Summary of mutation annotation format (MAF) data for LUSC. Top left: bar plot showing the distribution of variant classifications. Top middle: bar plot displaying the distribution of variant types, including SNP, INS, and DEL. Top right: bar plot of SNV classes, showing the frequency of nucleotide changes. Bottom left: Histogram of variants per sample. Bottom middle: Boxplot summarizing the variant classification frequencies across samples. Bottom right: Bar plot of the top 20 mutated genes in LUSC.

(B) Oncoplot of LUSC, highlighting the top 30 mutated genes, mutation types, and their frequencies across samples.

(C) Gene co-occurrence and mutual exclusivity heatmap in LUSC. *: $p < 0.05$. The diagonal represents the frequency of mutations in each gene, with the numbers in brackets indicating the total number of samples harboring mutations in the respective gene.

(D) The volcano plot of differentially expressed genes in LUSC, where black text labels indicate the top 10 differentially expressed proteins ranked by Log2(FC) (including KRT14 and KRT6C), and red text highlights proteins of interest.

(E) Top-left: KRT5 transcriptome-proteome correlation in normal/tumor tissues across pan-cancer. Top-right: KRT5 gene expression across cancer types. Bottom-left: KRT6A transcriptome-proteome correlation in normal/tumor tissues across pan-cancer. Bottom-right: KRT6A gene expression across cancer types. ns: not significant, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

was low and varied somewhat across cancer types. Using non-ccRCC as an example, we found that compared to Pro TME, RNA TME may be more clinically relevant, providing more precise prognostic and therapeutic insights.

Finally, we have developed an integrative R package (CPGTA) to explore the pan-cancer analysis of proteogenomic data, in order to provide pan-cancer multi-omics data (genome, transcriptome, proteome, phospho-proteome, clinical data, biospecimen data) and correlation analysis, differential analysis, and immunophenotyping results for the 2,555 cases of 15 cancer types collected so far. We anticipate that CPGTA will serve as a valuable bioinformatic resource for integrative multi-omics investigations in cancer research, facilitating both validation and discovery in precision oncology.

Limitations of the study

Despite providing the most comprehensive pan-cancer paired multi-omics resource to date and systematically characterizing

the complex associations between multi-omic hierarchies, especially between proteome and transcriptome at the pan-cancer level, there are still some limitations in this study. First, while our dataset includes 15 cancer types, certain rare cancers are underrepresented, potentially limiting the generalizability of our findings across all cancer types. Second, the proteomics technology used captures only a subset of the proteome, missing some low-abundance proteins and post-translational modifications that may play crucial roles in cancer biology. Finally, while we identified potential therapeutic targets based on mRNA-protein correlations and survival outcomes, functional validation studies are needed to confirm their clinical relevance and therapeutic potential.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, XiaoLin Li (xiaolinli@ieee.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All data used in this study originate from public databases. Proteomic data, phosphoproteomic data, biospecimen, and clinical data involved in this study are derived from https://proteomic.datacommons.cancer.gov/pdc/. Transcriptomic data and copy number alteration data were obtained from https://portal.gdc.cancer.gov/ or the supplemental information of the source article.
- This study has developed an R package (CPGTA) for the multi-omics data used, which includes functionalities for the multi-omics data utilized in this study, custom correlation analysis, differential analysis, clinical relevance analysis, and more. All data processing and analysis code for this study was performed in R version 4.4.1. All original code has been deposited at GitHub (https://github.com/LYING7797/CPGTA). The CPGTA package and related files are also available on Zenodo (DOI: https://doi.org/10.5281/zenodo.17366523).
- Any additional information required to reanalyze the data reported in this article is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China (grant no. 2022YFC3600902), the National Natural Science Foundation of China (grant no. 32200524), and the Key Research and Development Program of Zhejiang (grant no. 2025C01129). It was also funded by the Zhejiang Leading Innovation and Entrepreneurship Team (grant no. 2022R01006).

AUTHOR CONTRIBUTIONS

Conceptualization, L. Ying and C.S.; data curation, L. Ying; formal analysis, L. Ying and L.X.; investigation, L. Ying; methodology, L. Ying, X.L., and C.S.; software, L. Ying; validation, L. Ying; visualization, L. Ying; writing – original draft, L. Ying; writing – review & editing, L. Ying, C.S., W.Z., and W.Q.; funding acquisition, C.S. and X.L; project administration, C.S. and X.L.; resources, X.L.; supervision, C.S. and X.L.

DECLARATION OF INTERESTS

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 article.

STAR★METHODS

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

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Human tumor samples
  - Cell line models
- METHOD DETAILS
  - Proteomic and phosphoproteomic datasets
  - Transcriptomic datasets
  - Copy number variation (CNV) datasets
  - Mutation datasets
  - Biospecimen and clinical datasets
  - Correlation analysis
  - Gene set enrichment analysis
  - Differential expression analysis
  - Hierarchical clustering and survival analysis
  - Immune subtyping of the tumor microenvironment
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.114500.

Received: March 13, 2025
Revised: June 24, 2025
Accepted: December 16, 2025
Published: December 19, 2025

REFERENCES

1. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. https://doi.org/10.1038/35057062.
2. Ding, L., Bailey, M.H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., Gibbs, D.L., Weerasinghe, A., Huang, K.L., Tokheim, C., et al. (2018). Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. Cell *173*, 305–320.e10. https://doi.org/10.1016/j.cell.2018.03.033.
3. Huang, K.L., Mashl, R.J., Wu, Y., Ritter, D.I., Wang, J., Oh, C., Paczkowska, M., Reynolds, S., Wyczalkowski, M.A., Oak, N., et al. (2018). Pathogenic Germline Variants in 10,389 Adult Cancers. Cell *173*, 355–370.e14. https://doi.org/10.1016/j.cell.2018.03.039.
4. Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V., et al. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell *173*, 291–304.e6. https://doi.org/10.1016/j.cell.2018.03.022.
5. Kahles, A., Lehmann, K.V., Toussaint, N.C., Hüser, M., Stark, S.G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C., and Cancer Genome Atlas Research Network; and Rätsch, G. (2018). Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell *34*, 211–224.e6. https://doi.org/10.1016/j.ccell.2018.07.001.
6. Jayasinghe, R.G., Cao, S., Gao, Q., Wendl, M.C., Vo, N.S., Reynolds, S.M., Zhao, Y., Climente-González, H., Chai, S., Wang, F., et al. (2018). Systematic Analysis of Splice-Site-Creating Mutations in Cancer. Cell Rep. *23*, 270–281.e3. https://doi.org/10.1016/j.celrep.2018.03.052.
7. Taylor, A.M., Shih, J., Ha, G., Gao, G.F., Zhang, X., Berger, A.C., Schumacher, S.E., Wang, C., Hu, H., Liu, J., et al. (2018). Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell *33*, 676–689.e3. https://doi.org/10.1016/j.ccell.2018.03.007.

8. Gao, Q., Liang, W.W., Foltz, S.M., Mutharasu, G., Jayasinghe, R.G., Cao, S., Liao, W.W., Reynolds, S.M., Wyczalkowski, M.A., Yao, L., et al. (2018). Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. Cell Rep. *23*, 227–238.e3. https://doi.org/10.1016/j.celrep.2018.03.050.

9. Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.H., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., et al. (2018). The Immune Landscape of Cancer. Immunity *48*, 812–830.e14. https://doi.org/10.1016/j.immuni.2018.03.023.

10. Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell *173*, 400–416.e11. https://doi.org/10.1016/j.cell.2018.02.052.

11. Ellis, M.J., Gillette, M., Carr, S.A., Paulovich, A.G., Smith, R.D., Rodland, K.K., Townsend, R.R., Kinsinger, C., Mesri, M., Rodriguez, H., et al. (2013). Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. Cancer Discov. *3*, 1108–1112. https://doi.org/10.1158/2159-8290.CD-13-0219.

12. Obernosterer, G., Leuschner, P.J.F., Alenius, M., and Martinez, J. (2006). Post-transcriptional regulation of microRNA expression. Rna *12*, 1161–1167. https://doi.org/10.1261/rna.2322506.

13. Filipowicz, W., Bhattacharyya, S.N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat. Rev. Genet. *9*, 102–114. https://doi.org/10.1038/nrg2290.

14. Li, Y., Porta-Pardo, E., Tokheim, C., Bailey, M.H., Yaron, T.M., Stathias, V., Geffen, Y., Imbach, K.J., Cao, S., Anand, S., et al. (2023). Pan-cancer proteogenomics connects oncogenic drivers to functional states. Cell *186*, 3921–3944.e25. https://doi.org/10.1016/j.cell.2023.07.014.

15. Geffen, Y., Anand, S., Akiyama, Y., Yaron, T.M., Song, Y., Johnson, J.L., Govindan, A., Babur, Ö., Li, Y., Huntsman, E., et al. (2023). Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation. Cell *186*, 3945–3967.e26. https://doi.org/10.1016/j.cell.2023.07.013.

16. Lv, D., Li, D., Cai, Y., Guo, J., Chu, S., Yu, J., Liu, K., Jiang, T., Ding, N., Jin, X., et al. (2024). CancerProteome: a resource to functionally decipher the proteome landscape in cancer. Nucleic Acids Res. *52*, D1155–D1162. https://doi.org/10.1093/nar/gkad824.

17. Liao, Y., Savage, S.R., Dou, Y., Shi, Z., Yi, X., Jiang, W., Lei, J.T., and Zhang, B. (2023). A proteogenomics data-driven knowledge base of human cancer. Cell Syst. *14*, 777–787.e5. https://doi.org/10.1016/j.cels.2023.07.007.

18. Petralia, F., Tignor, N., Reva, B., Koptyra, M., Chowdhury, S., Rykunov, D., Krek, A., Ma, W., Zhu, Y., Ji, J., et al. (2020). Integrated Proteogenomic Characterization across Major Histological Types of Pediatric Brain Cancer. Cell *183*, 1962–1985.e31. https://doi.org/10.1016/j.cell.2020.10.044.

19. Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. Nature *534*, 55–62. https://doi.org/10.1038/nature18003.

20. 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.

21. Clark, D.J., Dhanasekaran, S.M., Petralia, F., Pan, J., Song, X., Hu, Y., da Veiga Leprevost, F., Reva, B., Lih, T.S.M., Chang, H.Y., et al. (2019). Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. Cell *179*, 964–983.e31. https://doi.org/10.1016/j.cell.2019.10.007.

22. Vasaikar, S., Huang, C., Wang, X., Petyuk, V.A., Savage, S.R., Wen, B., Dou, Y., Zhang, Y., Shi, Z., Arshad, O.A., et al. (2019). Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. Cell *177*, 1035–1049.e19. https://doi.org/10.1016/j.cell.2019.03.030.

23. 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.

24. Liu, J., Cao, S., Imbach, K.J., Gritsenko, M.A., Lih, T.S.M., Kyle, J.E., Yaron-Barir, T.M., Binder, Z.A., Li, Y., Strunilin, I., et al. (2024). Multi-scale signaling and tumor evolution in high-grade gliomas. Cancer Cell *42*, 1217–1238.e19. https://doi.org/10.1016/j.ccell.2024.06.004.

25. Kim, K.H., Migliozzi, S., Koo, H., Hong, J.H., Park, S.M., Kim, S., Kwon, H.J., Ha, S., Garofano, L., Oh, Y.T., et al. (2024). Integrated proteogenomic characterization of glioblastoma evolution. Cancer Cell *42*, 358–377.e8. https://doi.org/10.1016/j.ccell.2023.12.015.

26. Huang, C., Chen, L., Savage, S.R., Eguez, R.V., Dou, Y., Li, Y., da Veiga Leprevost, F., Jaehnig, E.J., Lei, J.T., Wen, B., et al. (2021). Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. Cancer Cell *39*, 361–379.e16. https://doi.org/10.1016/j.ccell.2020.12.007.

27. Soltis, A.R., Bateman, N.W., Liu, J., Nguyen, T., Franks, T.J., Zhang, X., Dalgard, C.L., Viollet, C., Somiari, S., Yan, C., et al. (2022). Proteogenomic analysis of lung adenocarcinoma reveals tumor heterogeneity, survival determinants, and therapeutically relevant pathways. Cell Rep. Med. *3*, 100819. https://doi.org/10.1016/j.xcrm.2022.100819.

28. Gillette, M.A., Satpathy, S., Cao, S., Dhanasekaran, S.M., Vasaikar, S.V., Krug, K., Petralia, F., Li, Y., Liang, W.W., Reva, B., et al. (2020). Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. Cell *182*, 200–225.e35. https://doi.org/10.1016/j.cell.2020.06.013.

29. Chen, Y.J., Roumeliotis, T.I., Chang, Y.H., Chen, C.T., Han, C.L., Lin, M.H., Chen, H.W., Chang, G.C., Chang, Y.L., Wu, C.T., et al. (2020). Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression. Cell *182*, 226–244.e17. https://doi.org/10.1016/j.cell.2020.06.012.

30. Satpathy, S., Krug, K., Jean Beltran, P.M., Savage, S.R., Petralia, F., Kumar-Sinha, C., Dou, Y., Reva, B., Kane, M.H., Avanessian, S.C., et al. (2021). A proteogenomic portrait of lung squamous cell carcinoma. Cell *184*, 4348–4371.e40. https://doi.org/10.1016/j.cell.2021.07.016.

31. Zhang, H., Liu, T., Zhang, Z., Payne, S.H., Zhang, B., McDermott, J.E., Zhou, J.Y., Petyuk, V.A., Chen, L., Ray, D., et al. (2016). Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. Cell *166*, 755–765. https://doi.org/10.1016/j.cell.2016.05.069.

32. McDermott, J.E., Arshad, O.A., Petyuk, V.A., Fu, Y., Gritsenko, M.A., Clauss, T.R., Moore, R.J., Schepmoes, A.A., Zhao, R., Monroe, M.E., et al. (2020). Proteogenomic Characterization of Ovarian HGSC Implicates Mitotic Kinases, Replication Stress in Observed Chromosomal Instability. Cell Rep. Med. *1*, 100004. https://doi.org/10.1016/j.xcrm.2020.100004.

33. Cao, L., Huang, C., Cui Zhou, D., Hu, Y., Lih, T.M., Savage, S.R., Krug, K., Clark, D.J., Schnaubelt, M., Chen, L., et al. (2021). Proteogenomic characterization of pancreatic ductal adenocarcinoma. Cell *184*, 5031–5052.e26. https://doi.org/10.1016/j.cell.2021.08.023.

34. Hyeon, D.Y., Nam, D., Han, Y., Kim, D.K., Kim, G., Kim, D., Bae, J., Back, S., Mun, D.G., Madar, I.H., et al. (2023). Proteogenomic landscape of human pancreatic ductal adenocarcinoma in an Asian population reveals tumor cell-enriched and immune-rich subtypes. Nat. Cancer *4*, 290–307. https://doi.org/10.1038/s43018-022-00479-7.

35. Mun, D.G., Bhin, J., Kim, S., Kim, H., Jung, J.H., Jung, Y., Jang, Y.E., Park, J.M., Kim, H., Jung, Y., et al. (2019). Proteogenomic Characterization of Human Early-Onset Gastric Cancer. Cancer Cell *35*, 111–124.e10. https://doi.org/10.1016/j.ccell.2018.12.003.

36. Gao, Q., Zhu, H., Dong, L., Shi, W., Chen, R., Song, Z., Huang, C., Li, J., Dong, X., Zhou, Y., et al. (2019). Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. Cell *179*, 1240. https://doi.org/10.1016/j.cell.2019.10.038.

37. 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.

38. Dou, Y., Katsnelson, L., Gritsenko, M.A., Hu, Y., Reva, B., Hong, R., Wang, Y.T., Kolodziejczak, I., Lu, R.J.H., Tsai, C.F., et al. (2023). Proteogenomic insights suggest druggable pathways in endometrial carcinoma. Cancer Cell *41*, 1586–1605.e15. https://doi.org/10.1016/j.ccell.2023.07.007.
39. Li, G.X., Chen, L., Hsiao, Y., Mannan, R., Zhang, Y., Luo, J., Petralia, F., Cho, H., Hosseini, N., Leprevost, F.D.V., et al. (2024). Comprehensive proteogenomic characterization of rare kidney tumors. Cell Rep. Med. *5*, 101547. https://doi.org/10.1016/j.xcrm.2024.101547.
40. Wu, C.S., Li, H.P., Hsieh, C.H., Lin, Y.T., Yi-Feng Chang, I., Chung, A.K., Huang, Y., Ueng, S.H., Hsiao, Y.C., Chien, K.Y., et al. (2025). Integrated multi-omics analyses of oral squamous cell carcinoma reveal precision patient stratification and personalized treatment strategies. Cancer Lett. *614*, 217482. https://doi.org/10.1016/j.canlet.2025.217482.
41. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. J. Open Source Softw. *3*, 861. https://doi.org/10.21105/joss.00861.
42. Ghoshdastider, U., and Sendoel, A. (2023). Exploring the pan-cancer landscape of posttranscriptional regulation. Cell Rep. *42*, 113172. https://doi.org/10.1016/j.celrep.2023.113172.
43. Zhang, Y., Chen, F., Chandrashekar, D.S., Varambally, S., and Creighton, C.J. (2022). Proteogenomic characterization of 2002 human cancers reveals pan-cancer molecular subtypes and associated pathways. Nat. Commun. *13*, 2669. https://doi.org/10.1038/s41467-022-30342-3.
44. Savage, S.R., Yi, X., Lei, J.T., Wen, B., Zhao, H., Liao, Y., Jaehnig, E.J., Somes, L.K., Shafer, P.W., Lee, T.D., et al. (2024). Pan-cancer proteogenomics expands the landscape of therapeutic targets. Cell *187*, 4389–4407.e15. https://doi.org/10.1016/j.cell.2024.05.039.
45. Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell *174*, 1034–1035. https://doi.org/10.1016/j.cell.2018.07.034.
46. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47. https://doi.org/10.1093/nar/gkv007.
47. Lucas, L.M., Dwivedi, V., Senfeld, J.I., Cullum, R.L., Mill, C.P., Piazza, J.T., Bryant, I.N., Cook, L.J., Miller, S.T., Lott, J.H., 4th, et al. (2022). The Yin and Yang of ERBB4: Tumor Suppressor and Oncoprotein. Pharmacol. Rev. *74*, 18–47. https://doi.org/10.1124/pharmrev.121.000381.
48. Jin, D., Wang, N., Xue, Y., Yang, Y., Shi, K., Wu, H., Sheu, J.J.C., Jeong, J.H., Ban, Z., Shen, D., and Yang, L. (2025). Pan-cancer analysis of ARNT2 and its oncogenic role in cervical cancer. J. Gynecol. Oncol. *36*, e113. https://doi.org/10.3802/jgo.2025.36.e113.
49. Jia, Y., Hao, S., Jin, G., Li, H., Ma, X., Zheng, Y., Xiao, D., and Wang, Y. (2019). Overexpression of ARNT2 is associated with decreased cell proliferation and better prognosis in gastric cancer. Mol. Cell. Biochem. *450*, 97–103. https://doi.org/10.1007/s11010-018-3376-y.
50. Yang, B., Yang, E., Liao, H., Wang, Z., Den, Z., and Ren, H. (2015). ARNT2 is downregulated and serves as a potential tumor suppressor gene in non-small cell lung cancer. Tumour Biol. *36*, 2111–2119. https://doi.org/10.1007/s13277-014-2820-1.
51. Xu, J., Chen, S., Hao, T., Liu, G., Zhang, K., Zhang, C., and He, Y. (2024). MEX3A promotes colorectal cancer migration, invasion and EMT via regulating the Wnt/beta-catenin signaling pathway. J. Cancer Res. Clin. Oncol. *150*, 319. https://doi.org/10.1007/s00432-024-05845-9.
52. Lu, Y., Bi, T., Zhou, S., and Guo, M. (2023). MEX3A promotes angiogenesis in colorectal cancer via glycolysis. Libyan J. Med. *18*, 2202446. https://doi.org/10.1080/19932820.2023.2202446.
53. Chen, W., Hu, L., Lu, X., Wang, X., Zhao, C., Guo, C., Li, X., Ding, Y., Zhao, H., Tong, D., et al. (2023). The RNA binding protein MEX3A promotes tumor progression of breast cancer by post-transcriptional regulation of IGFBP4. Breast Cancer Res. Treat. *201*, 353–366. https://doi.org/10.1007/s10549-023-07028-5.
54. Wang, C.K., Chen, T.J., Tan, G.Y.T., Chang, F.P., Sridharan, S., Yu, C.H.A., Chang, Y.H., Chen, Y.J., Cheng, L.T., and Hwang-Verslues, W.W. (2023). MEX3A Mediates p53 Degradation to Suppress Ferroptosis and Facilitate Ovarian Cancer Tumorigenesis. Cancer Res. *83*, 251–263. https://doi.org/10.1158/0008-5472.CAN-22-1159.
55. Zhang, K., Zheng, X., Sun, Y., Feng, X., Wu, X., Liu, W., Gao, C., Yan, Y., Tian, W., and Wang, Y. (2024). TOP2A modulates signaling via the AKT/mTOR pathway to promote ovarian cancer cell proliferation. Cancer Biol. Ther. *25*, 2325126. https://doi.org/10.1080/15384047.2024.2325126.
56. Uuskula-Reimand, L., and Wilson, M.D. (2022). Untangling the roles of TOP2A and TOP2B in transcription and cancer. Sci. Adv. *8*, eadd4920. https://doi.org/10.1126/sciadv.add4920.
57. Wang, T., Lu, J., Wang, R., Cao, W., and Xu, J. (2022). TOP2A promotes proliferation and metastasis of hepatocellular carcinoma regulated by miR-144-3p. J. Cancer *13*, 589–601. https://doi.org/10.7150/jca.64017.
58. Liu, Y., Chen, X., Xu, Y., Yang, T., Wang, H., Wang, Z., Hu, Z., Chen, L., Zhang, Z., and Wu, Y. (2024). CTHRC1 promotes colorectal cancer progression by recruiting tumor-associated macrophages via up-regulation of CCL15. J. Mol. Med. *102*, 81–94. https://doi.org/10.1007/s00109-023-02399-0.
59. Ma, Z., Chao, F., Wang, S., Song, Z., Zhuo, Z., Zhang, J., Xu, G., and Chen, G. (2020). CTHRC1 affects malignant tumor cell behavior and is regulated by miR-30e-5p in human prostate cancer. Biochem. Biophys. Res. Commun. *525*, 418–424. https://doi.org/10.1016/j.bbrc.2020.02.098.
60. Liu, Y.J., Du, J., Li, J., Tan, X.P., and Zhang, Q. (2023). CTHRC1, a novel gene with multiple functions in physiology, disease and solid tumors (Review). Oncol. Lett. *25*, 266. https://doi.org/10.3892/ol.2023.13852.
61. Yang, H., Sun, B., Fan, L., Ma, W., Xu, K., Hall, S.R.R., Wang, Z., Schmid, R.A., Peng, R.W., Marti, T.M., et al. (2022). Multi-scale integrative analyses identify THBS2(+) cancer-associated fibroblasts as a key orchestrator promoting aggressiveness in early-stage lung adenocarcinoma. Theranostics *12*, 3104–3130. https://doi.org/10.7150/thno.69590.
62. Qi, L., Sun, B., Yang, B., and Lu, S. (2023). CEBPB regulates the migration, invasion and EMT of breast cancer cells by inhibiting THBS2 expression and O-fucosylation. Hum. Mol. Genet. *32*, 1850–1863. https://doi.org/10.1093/hmg/ddad022.
63. Lan, J., Zhang, S., Zheng, L., Long, X., Chen, J., Liu, X., Zhou, M., and Zhou, J. (2023). PLOD2 promotes colorectal cancer progression by stabilizing USP15 to activate the AKT/mTOR signaling pathway. Cancer Sci. *114*, 3190–3202. https://doi.org/10.1111/cas.15851.
64. Bagaev, A., Kotlov, N., Nomie, K., Svekolkin, V., Gafurov, A., Isaeva, O., Osokin, N., Kozlov, I., Frenkel, F., Gancharova, O., et al. (2021). Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. Cancer Cell *39*, 845–865.e7. https://doi.org/10.1016/j.ccell.2021.04.014.
65. Goncalves, E., Poulos, R.C., Cai, Z., Barthorpe, S., Manda, S.S., Lucas, N., Beck, A., Bucio-Noble, D., Dausmann, M., Hall, C., et al. (2022). Pan-cancer proteomic map of 949 human cell lines. Cancer Cell *40*, 835–849.e838. https://doi.org/10.1016/j.ccell.2022.06.010.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
proteomics data and phosphoproteomics data	https://proteomic.datacommons.cancer.gov/pdc/	N/A
RNA Seq data	https://portal.gdc.cancer.gov/	N/A
CNV and mutation data	https://portal.gdc.cancer.gov/	N/A
RNA-seq and CNV data of PDC000180	Petralia et al.,[18]	https://pgc-accounts.sbgenomics.com/
RNA-seq data of the PDC000198	Gao et al.,[36]	NODE database: OEP000321
RNA-seq data of PDC000214	Dong et al.,[35]	GEO: GSE122401
CNV data of PDC000116	Vasaikar et al.,[22]	http://linkedomics.org/cptac-colon/
CNV data of PDC000120	Krug et al.,[20]	https://www.linkedomics.org/data_download/CPTAC-BRCA/
RNA Seq data and proteomics data of cancer cell lines	Emanuel Gonçalves et al.,[65]	https://cellmodelpassports.sanger.ac.uk/
Software and algorithms		
R (Version 4.4.1)	https://www.r-project.org/	N/A
CPGTA	This paper	GitHub: https://github.com/LYING7797/CPGTA Zenodo (https://doi.org/10.5281/zenodo.17366523)
UMAP	McInnes et al.,[41]	https://cran.r-project.org/web/packages/umap/index.html
Limma	Ritchie et al.,[46]	https://bioconductor.org/packages/release/bioc/html/limma.html

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Human tumor samples

This study utilized proteomics data from 2,555 primary human tumor samples spanning 15 different cancer types across 23 distinct studies (Table S1). All samples represent primary tumor tissues from cancer patients. Sample demographics including age, sex, gender, ancestry, race, and ethnicity information were obtained from the original studies as available in the Protein Data Commons (PDC) database (https://proteomic.datacommons.cancer.gov/pdc/).

Sample size and allocation: The 2,555 tumor samples were distributed across cancer types as detailed in Table S1. Of these samples, 2,458 samples had phosphoproteomic data; 2,088 samples had corresponding RNA-seq data; 1,739 samples had gene-level copy number variation (CNV) data; and 1,521 samples had mutation data.

This study did not analyze the influence of sex or gender on protein expression patterns, representing a limitation as sex-specific differences in protein expression may exist but were not systematically evaluated.

Cell line models

This study utilized transcriptomic and proteomic data from 942 human cancer cell lines obtained from the Cell Model Passports database (https://cellmodelpassports.sanger.ac.uk/). All cell line data were accessed through the publicly available Cell Model Passports platform with appropriate usage permissions.

METHOD DETAILS

Proteomic and phosphoproteomic datasets

We compiled a comprehensive dataset of mass spectrometry-based proteomics data, encompassing primary tumors from 15 different cancer types across 23 distinct studies (Table S1). All proteomics data and phosphoproteomics data are sourced entirely

from https://proteomic.datacommons.cancer.gov/pdc/.The PDC processes the submitted raw mass spectrometry data through a common data analysis pipeline (CDAP) to produce derived analysis results which can be used to identify patterns of protein and post-translational modifications (PTMs) abundance.

More specifically, in iTRAQ and TMT workflows, peptides are quantified using reporter ions, and these figures are summed up at the gene level. For such experiments, each cell contains a normalized log2 ratio of a gene/protein's value to a control channel from pooled tumor samples. Log2 ratios for visualization are provided for all isotopic labeling studies, including PTMs. Whole proteome studies also offer an unshared_log2_ratio, excluding shared peptide data.

The PDC Common Data Analysis Pipeline generates a protein abundance matrix, rolled up to the gene level, for each study. We utilize this matrix data as the proteomic data source for each study, markedly enhancing data comparability across various research and ensuring that all data adhere to uniform quality control benchmarks.

We then applied a uniform standardization process to all the collected datasets using the following method. For each dataset, we adjusted the log-transformed expression levels to be in terms of standard deviations away from the median. Then, we applied a normalization process to the expression levels across different samples, converting them to standard deviations from the median value.[43] We individually normalized the datasets for total protein and phospho-protein within each dataset.

To normalize the expression levels of proteomic and transcriptomic data from different studies, we used the ComBat function from the sva package (version 3.54.0) in R (version 4.4.1) to adjust for batch effects. The batch variable was defined as the cancer type (i.e., study identifier). ComBat was applied with default parameters (par.prior = TRUE, prior.plots = FALSE) without specifying a model matrix of covariates to preserve, allowing for comprehensive batch effect removal while maintaining biological variation.

Transcriptomic datasets

Out of the 2,555 human tumors with proteomic data, 2,088 had corresponding RNA-seq data. We acquired RNA-seq data for the CPTAC, TCGA, and APOLLO projects from the Genomic Data Commons Data Portal [https://portal.gdc.cancer.gov/]. RNA-Seq data for the other studies (PDC000180, PDC000198, PDC000214) were either downloaded from different platforms or obtained directly from the supplemental materials of scientific articles. All mentioned RNA-Seq data are TPM data, and those without TPM data were converted from counts to TPM.

We normalized the expression values in the samples to the standard deviation from the median for each dataset, just as we did with the proteomic data. For pan-cancer analyses, we further removed batch effects between studies using the ComBat algorithm on the datasets standardized by the above method.

Copy number variation (CNV) datasets

Out of the 2,555 human tumors with proteomic data, 1,739 had corresponding gene-level CNV data. The majority of the CNV data are sourced from the Genomic Data Commons Data Portal [https://portal.gdc.cancer.gov/], while some (PDC000116, PDC000120, PDC000180) are from original study supplements. We normalized the absolute copy number data based on ploidy by dividing each gene's copy number by the average copy number across all genes. Subsequently, we categorized the data into several predefined levels: homologous deletion (−2), heterozygous deletion (−1), wild-type (0), a gain of 1–2 copies (+1), and amplification of at least 5 copies (+2).

Mutation datasets

Out of the 2,555 human tumors with proteomic data, 1,521 had mutation data. We obtained all the mutation data from the Genomic Data Commons Data Portal [https://portal.gdc.cancer.gov/]

Biospecimen and clinical datasets

Biospecimen and clinical datasets each consist of 23 files, corresponding to the PDC (Proteomic Data Commons) identifier for each proteome. All are sourced from https://proteomic.datacommons.cancer.gov/pdc/.

Correlation analysis

For multi-omics correlation study: A Pearson correlation analysis is performed on samples featuring four types of omics data, across 16 studies and 12 cancer types (OV, COAD, BRCA, UCEC, ccRCC, LUAD, BC, GBM, HNSCC, LUSC, PDAC, non-ccRCC). For each omics type (CNV, transcriptomic, proteomics, and phosphoproteomics), standardized data files were merged across studies, and batch effects were removed using the ComBat algorithm. Genes with missing values in more than 50% of the samples were excluded. Phosphoproteomics data were processed by averaging expression values for proteins with multiple phosphorylation sites. Genes and samples common across all four omics datasets were identified, resulting in 3,095 genes and 1,614 samples for downstream analysis. Pairwise Pearson correlation analyses were conducted to evaluate the relationships between different omics layers: CNV vs. Transcriptomic, CNV vs. Proteomics, CNV vs. Phosphoproteomics, Transcriptomic vs. Proteomics, Transcriptomic vs. Phosphoproteomics, Proteomics vs. Phosphoproteomics. For each gene, correlations were computed using the cor.test function in R, and the correlation coefficient, p-value, and test statistic were recorded.

For pan-cancer analysis, the correlation between transcriptomics and proteomics in tumor/normal tissues was evaluated using samples that contained both transcriptomic and proteomic data. Normalized transcriptomic and proteomic data were obtained from the preprocessed files corresponding to each study. Gene expression data across studies were merged, retaining only genes

present in more than 50% of the samples. A total of 10,125 common genes and 904 samples were identified between the transcriptomic and proteomics datasets, resulting in matched tumor and normal datasets for downstream analysis.

Correlation analysis at the specific cancer level is conducted in the same manner.

Gene set enrichment analysis

To identify pathways associated with the correlation between pan cancer mRNA expression and protein abundance, we performed GSEA using the clusterProfiler package in R. Genes were ranked by their correlation coefficients, and GSEA was performed using the Hallmark (h.all.v2023.2.Hs.symbols.gmt) and Gene Ontology (GO) (c5.go.v2023.2.Hs.symbols.gmt) gene sets. Enrichment results were filtered with the criteria: p-value <0.01, adjusted p-value <0.25, and normalized enrichment score (NES) > 2.

Differential expression analysis

We perform differential protein analysis using the limma method. Given the extensive coverage provided by proteomics data, this analysis is conducted exclusively on samples containing both tumor and normal tissue types, covering 13 cancer types (COAD, OV, BRCA, UCEC, ccRCC, LUAD, LIHC, GBM, HNSCC, LUSC, PDAC, non-ccRCC, OSCC). Tumor and normal proteomics datasets were combined across studies. Missing values were imputed using the minimum observed value for each dataset. Batch effects across studies were corrected using the ComBat algorithm from the sva package. Proteomics data from tumor and normal samples were merged into a single matrix containing 1,809 tumor samples and 1,022 normal samples. A grouping variable indicating tumor or normal samples was created, and a design matrix was constructed. Differential protein expression analysis was performed using the limma package. A linear model was fitted to the data, and empirical Bayes statistics were used to compute adjusted p-values. Proteins with |logFC| > 1.2 and p-value <0.05 were considered differentially expressed.

The differential protein analysis method for specific cancer levels is the same as described above.

Hierarchical clustering and survival analysis

Proteomics data from tumor and normal tissues, and transcriptomics data from tumor tissues were obtained for studies with complete data. Batch effects across studies were corrected using the ComBat algorithm from the sva package. Differential protein expression analysis was performed using the limma package, with |logFC| > 1 and p-value <0.05 considered significant.

For each study, the correlation between mRNA-Protein expression was calculated using Pearson correlation. Genes with high correlation across studies were clustered using hierarchical clustering. Three gene clusters were identified, and the cluster with the highest average correlation was selected for further analysis. Differentially upregulated proteins were intersected with genes in the high-correlation cluster. For these genes, tumor expression levels were compared to the mean expression in normal tissues. Patients were grouped into ‘‘High correlation gene cluster with high expression’’ or ‘‘High correlation gene cluster with non-high expression’’ expression groups based on the number of genes with higher tumor expression. Kaplan-Meier survival analysis and log rank tests were performed to evaluate the prognostic significance of these groups.

Immune subtyping of the tumor microenvironment

All preprocessed proteomic and transcriptomic data were integrated separately, and batch effects were removed using the ComBat algorithm from the sva package before performing TME subtyping. The pipeline from https://github.com/BostonGene/MFP was applied to both datasets for immune subtyping. The final output is a table containing the subtyping results for each sample.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were performed using R (version 4.4.1). Proteomics and phosphoproteomics data from 2,555 human tumors across 15 cancer types were processed through the PDC Common Data Analysis Pipeline, with expression levels standardized as standard deviations from the median. Batch effects between studies were corrected using ComBat function from the sva package (version 3.54.0) with default parameters (par.prior = TRUE, prior.plots = FALSE).

For multi-omics correlation analysis, Pearson correlation was performed on 3,095 genes across 1,614 samples using the cor.test function in R. Pan-cancer correlation analysis between transcriptomics and proteomics included 10,125 common genes across 904 samples. Genes with missing values in >50% of samples were excluded from analyses.

Differential protein expression analysis was conducted using the limma package on 1,809 tumor and 1,022 normal samples. Statistical significance was defined as |logFC| > 1.2 and p-value <0.05. Missing values were imputed using the minimum observed value for each dataset.

Gene Set Enrichment Analysis was performed using the clusterProfiler package with significance criteria of p-value <0.01, adjusted p-value <0.25, and normalized enrichment score (NES) > 2.

For survival analysis, patients were stratified based on expression patterns of genes in high-correlation clusters, and Kaplan-Meier survival analysis with log rank tests was used to evaluate prognostic significance.