Cell Press
OPEN ACCESS

Perspective

Proteogenomic data and resources for pan-cancer analysis

Yize Li,[1,2,27] Yongchao Dou,[3,4,27] Felipe Da Veiga Leprevost,[5,27] Yifat Geffen,[6,27] Anna P. Calinawan,[7,27] François Aguet,[6] Yo Akiyama,[6] Shankara Anand,[6] Chet Birger,[6] Song Cao,[1,2] Rekha Chaudhary,[8] Padmini Chilappagari,[8] Marcin Cieslik,[9] Antonio Colaprico,[10,11] Daniel Cui Zhou,[1,2] Corbin Day,[12] Marcin J. Domagalski,[8] Myvizhi Esai Selvan,[7] David Fenyö,[13,14] Steven M. Foltz,[1,2] Alicia Francis,[8] Tania Gonzalez-Robles,[13,14,15] Zeynep H. Gümüş,[7] David Heiman,[6] Michael Holck,[8] Runyu Hong,[13,14] Yingwei Hu,[16] Eric J. Jaehnig,[3,4] Jiayi Ji,[17] Wen Jiang,[3,4] Lizabeth Katsnelson,[13,14] Karen A. Ketchum,[8] Robert J. Klein,[7] Jonathan T. Lei,[3,4] Wen-Wei Liang,[1,2] Yuxing Liao,[3,4] Caleb M. Lindgren,[12] Weiping Ma,[7] Lei Ma,[8] Michael J. MacCoss,[18] Fernanda Martins Rodrigues,[1,2] Wilson McKerrow,[13,14] Ngoc Nguyen,[8] Robert Oldroyd,[12] Alexander Pilozzi,[8] Pietro Pugliese,[19] Boris Reva,[7] Paul Rudnick,[20] Kelly V. Ruggles,[13,15] Dmitry Rykunov,[7] Sara R. Savage,[3,4] Michael Schnaubelt,[16] Tobias Schraink,[13,14,15] Zhiao Shi,[3,4] Deepak Singhal,[8] Xiaoyu Song,[17] Erik Storrs,[1,2] Nadezhda V. Terekhanova,[1,2] Ratna R. Thangudu,[8] Mathangi Thiagarajan,[21] Liang-Bo Wang,[1,2] Joshua M. Wang,[13,14] Ying Wang,[13,14] Bo Wen,[3,4] Yige Wu,[1,2] Matthew A. Wyczalkowski,[1,2] Yi Xin,[8] Lijun Yao,[1,2] Xinpei Yi,[3,4] Hui Zhang,[16] Qing Zhang,[6] Maya Zuhl,[8] Gad Getz,[6,22,23] Li Ding,[1,2,24,25] Alexey I. Nesvizhskii,[5] Pei Wang,[7] Ana I. Robles,[26,*] Bing Zhang,[3,4,*] Samuel H. Payne,[12,*] and Clinical Proteomic Tumor Analysis Consortium

[1]Department of Medicine, Washington University in St. Louis, St. Louis, MO 63130, USA
[2]McDonnell Genome Institute, Washington University in St. Louis, St. Louis, MO 63130, USA
[3]Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX 77030, USA
[4]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
[5]Department of Pathology, University of Michigan, Ann Arbor, MI 48109, USA
[6]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA
[7]Department of Genetic and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[8]ICF, Rockville, MD 20850, USA
[9]Department of Computational Medicine & Bioinformatics, Department of Pathology, University of Michigan Medical School, Ann Arbor, MI 48109, USA
[10]Department of Public Health Sciences, University of Miami Miller School of Medicine, Miami, FL 33136, USA
[11]Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL 33136, USA
[12]Department of Biology, Brigham Young University, Provo, UT 84602, USA
[13]Institute for Systems Genetics, NYU Grossman School of Medicine, New York, NY 10016, USA
[14]Department of Biochemistry and Molecular Pharmacology, NYU Grossman School of Medicine, New York, NY 10016, USA
[15]Department of Medicine, NYU Grossman School of Medicine, New York, NY 10016, USA
[16]Department of Pathology, Johns Hopkins University, Baltimore, MD 21231, USA
[17]Tisch Cancer Institute and Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[18]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[19]Department of Sciences and Technologies, University of Sannio, Benevento 82100, Italy
[20]Spectragen Informatics, Bainbridge Island, WA 98110, USA
[21]Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA
[22]Cancer Center and Department of Pathology, Mass. General Hospital, Boston, MA 02114, USA
[23]Harvard Medical School, Boston, MA 02115, USA
[24]Siteman Cancer Center, Washington University in St. Louis, St. Louis, MO 63130, USA
[25]Department of Genetics, Washington University in St. Louis, St. Louis, MO 63130, USA
[26]Office of Cancer Clinical Proteomics Research, National Cancer Institute, Rockville, MD 20850, USA
[27]These authors contributed equally
*Correspondence: roblesa@mail.nih.gov (A.I.R.), bing.zhang@bcm.edu (B.Z.), sam_payne@byu.edu (S.H.P.)

https://doi.org/10.1016/j.ccell.2023.06.009

SUMMARY

The National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC) investigates tumors from a proteogenomic perspective, creating rich multi-omics datasets connecting genomic aberrations to cancer phenotypes. To facilitate pan-cancer investigations, we have generated harmonized genomic, transcriptomic, proteomic, and clinical data for >1000 tumors in 10 cohorts to create a cohesive and powerful dataset for scientific discovery. We outline efforts by the CPTAC pan-cancer working group in data harmonization, data dissemination, and computational resources for aiding biological discoveries. We also discuss challenges for multi-omics data integration and analysis, specifically the unique challenges of working with both nucleotide sequencing and mass spectrometry proteomics data.

INTRODUCTION

Comprehensive molecular profiling is radically changing cancer research. Genomic catalogs of tens of thousands of tumors generated by The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) add immense depth to our understanding of mutations that drive tumorigenesis.[1] As sequencing on individual tumor cohorts are published, the next wave of manuscripts from these consortia examines patterns across cancer types to elucidate the context-dependent nature of mutations and their impacts.[2] One limitation of these sequencing-centric efforts is the paucity of data for proteins and their modifications. A few select proteins were monitored through antibody-based approaches such as reverse phase protein arrays (RPPA), but broad and unbiased proteomics data were not generated. As proteins represent the primary molecules responsible for metabolism, signaling, and structure, comprehensive and quantitative protein measurements are an essential part of phenotypic characterization. To connect genotype to phenotype, a true proteogenomic approach is needed.[3]

Proteogenomics analysis is a powerful method for discovering the next generation of precision treatments for cancer as it explicitly links genomic mutations to their impact on cellular physiology.[4–6] Early work by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) demonstrated extensive proteome coverage with TCGA samples,[7] but also identified that sample collection protocols for TCGA allowed significant ischemia prior to tissue freezing. Thus the phosphorylation data measured in these tumors represented a mix of cancer-related and ischemia-related signaling.[8] As aberrant cellular signaling is an important hallmark of cancer dysfunction and ischemia activates several of the same pathways (e.g., MAPK signaling and apoptosis), it is necessary to create proteogenomic data from freshly acquired tumors with protocols designed to avoid ischemic artifacts.[9,10]

The CPTAC dataset currently includes 10 cancer cohorts of prospectively collected tumors analyzed with genomics, transcriptomics, proteomics, and phosphoproteomics (Figure 1). Molecular classifications derived from these primary data types are also available, e.g., HLA typing, immune cell decomposition, and ancestry prediction. Other protein post-translational modification (PTM) data such as acetylomics and glycoproteomics were generated for select cancer types. Standard clinical/demographic data and histology images have also been made available. Distributions of sex, age, tumor grade, tumor stage, smoking history, and recurrence status are illustrated in Figure 2. Detailed information of sample provenance is given in Tables S1 and S2. In the original publications investigating a single cancer cohort,[11–20] data were processed and analyzed by disease-specific working groups using customized genomics and proteomics data analysis pipelines. Therefore, to enable pan-cancer integrative analysis, and for consistency and reproducibility, we created a compendium of datasets where all proteogenomic data have been re-processed and harmonized.

Concurrent with this manuscript detailing the data processing and dissemination, CPTAC investigators have pursued biologically motivated pan-cancer analyses to illuminate mechanisms of cancer development. Pan-cancer investigation of protein post-translational modifications (PTMs) identified a subset of tumors with significant changes to cellular regulation, including dysregulated DNA repair, altered metabolic regulation associated with immune response, and patterns of acetylation that affect kinase specificity.[21] An integration of somatic driver mutations and proteomics data across tumor types resolves distinct cancer hallmark patterns.[22] Analysis groups continue to conduct thematic studies using the pan-cancer dataset described here, according to five identified themes: oncogenic drivers and pathways; DNA damage response; cell of origin; tumor microenvironment and immunotherapy; and clinical imaging, biomarkers, and actionable targets.

CPTAC datasets are generated as a resource for cancer research, and community-driven re-analysis is a positive and anticipated outcome from the program. Indeed, numerous groups have already begun re-examining the data.[23,24] They powerfully use proteogenomic data to reveal new molecular subtypes,[25–27] prognostic markers,[28–30] novel protein variants from alternative splicing and RNA editing,[31–33] and extensive post-translational regulation for protein complexes.[34,35] To facilitate an increased data reuse and serve the broad audience of cancer data stakeholders, we present our computational methodology for data harmonization and multiple dissemination mechanisms to share both the raw and processed data.

National Cancer Institute Data Commons

The Genomic Data Commons (GDC, https://portal.gdc.cancer.gov) and Proteomic Data Commons (PDC, https://pdc.cancer.gov) are National Cancer Institute (NCI) Cloud resources that coordinate storage and analysis of genomics and proteomics data for cancer research. The proteogenomic data generated by the CPTAC program is publicly disseminated through GDC and PDC, which host raw and processed data according to their in-house pipelines. As components of NCI Cloud resource, the GDC and PDC are fully integrated with other NCI Research Data Commons resources, e.g., the Cancer Imaging Archive (TCIA, https://www.cancerimagingarchive.net/), facilitating cloud-based analysis of proteomic, genomic, and imaging data. Driven primarily by the CPTAC projects, PDC organizes the data through a robust data model to maintain consistency and integrity of both data and associated metadata, and provides an interface to filter, query, search, and visualize proteogenomic data. A direct link to the harmonized data tables stored at the Proteome Data Commons is https://pdc.cancer.gov/pdc/cptac-pancancer.

Finally, in addition to thematic repositories, NCI's Cancer Research Data Commons contains a data type-agnostic resource, the Cancer Data Service (CDS). CPTAC has placed the processed and curated data files into the Cancer Data Service (CDS; https://dataservice.datacommons.cancer.gov/). The CPTAC data stored in the CDS includes all the harmonized proteogenomic data for our pan-cancer analyses, including mutation calls, RNA and protein quantification tables, clinical and demographic data, and derived molecular data such as HLA typing, immune cell decomposition, and ancestry prediction. The CPTAC pan-cancer data hosted in CDS is controlled data. Access to controlled access data on CDS is through the NCI data access policies approved, dbGaP compiled whitelists. Users can access the data for analysis with a queryable web

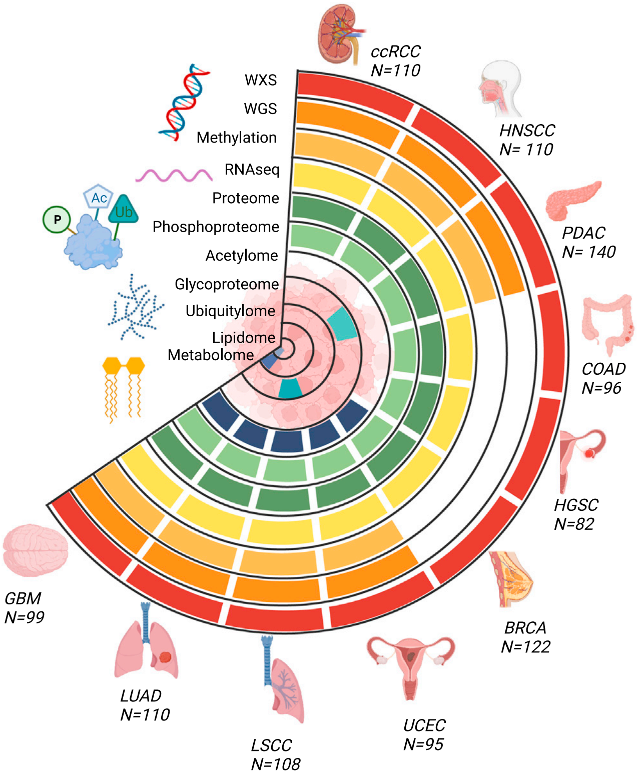


Figure 1. Tumor types and data types of the CPTAC pan-cancer dataset

Overview of the available molecular data types for the CPTAC pan-cancer cohort (n = 1072, see Table S1 for list of excluded cases and reasons for exclusion from the original datasets). Whole exome, whole genome, transcriptome, proteome, and phosphoproteome data are available for all ten cancer types. Normal samples are available for a subset of tumor types, see Tables S1 and S2.

portal through the Seven Bridges Cancer Genomics Cloud with dbGaP Study Accession, phs001287.v16.p6.

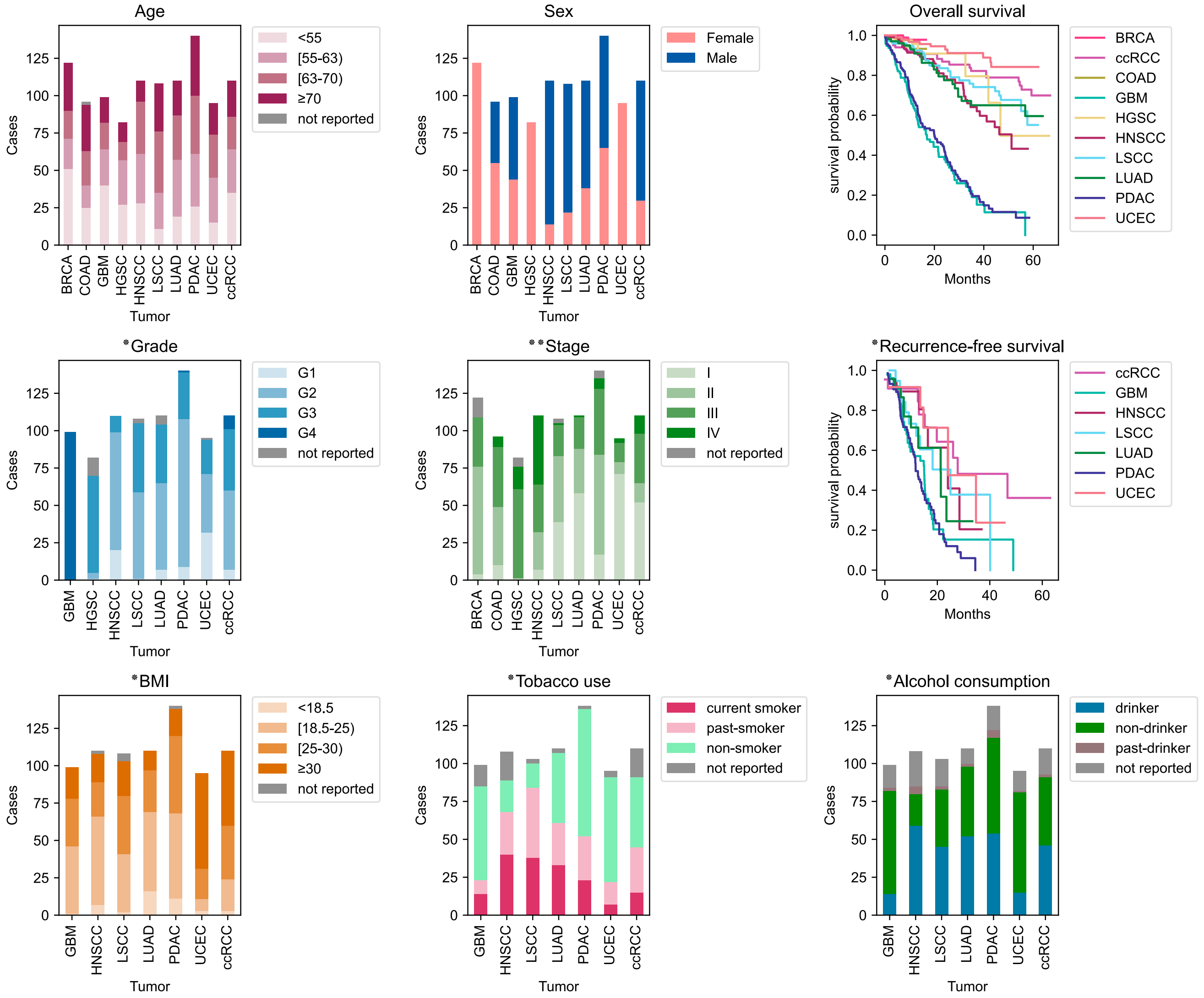
Data from multiple pipelines

Proteomic and genomic data analysis methods are continually evolving, and a variety of software tools exist for processing raw data into variant calls and quantifications (e.g., RNA or protein abundance matrices) that can be used for downstream analyses. As CPTAC consists of multiple groups with expertise in each data type, we have often analyzed data with multiple pipelines. Applying different tools to the same set of data may lead to different results and sometimes different conclusions. Therefore, benchmarking is important for tool assessment and selection. For somatic mutation calling, results from the ICGC-TCGA DREAM Somatic Mutation Calling Challenge show that different algorithms have characteristic error profiles, and an ensemble of pipelines always outperforms the best individual pipeline.[36] Based on this observation, and leveraging our team members' experience from the Multi-Center Mutation Calling in Multiple Cancers (MC3) project,[37] somatic mutation calling in our harmonized dataset was based on integrated results from the Broad Institute and Washington University in St. Louis pipelines, which each included multiple algorithms. RNA-seq data processing pipelines are now relatively mature with much overlap between widely used pipelines (e.g., https://nf-co.re/rnaseq). The major difference between the three pipelines used in this project is that the pipeline from Baylor College of Medicine includes circular RNAs in addition to linear RNAs. Quantifications for the vast majority of genes are not affected by circular RNAs and show very high correlation among the three pipelines. To compare different pipelines for proteomics data quantification, we have developed OmicsEV,[38] which uses more than a dozen evaluation metrics to comprehensively assess data depth, data normalization, batch effect, biological signal, platform reproducibility, and multi-omics concordance. Among the publicly available tools used by the CPTAC centers, the FragPipe pipeline usually provides higher data depth while maintaining similar or better performance for other metrics. Using three deep learning-derived features as evaluation metrics (predicted phosphosite probability, absolute retention time [RT] difference between observed and predicted RTs, and Pearson's correlation coefficient between observed and predicted spectra), we further found that FragPipe achieved higher sensitivity and quality for phosphopeptide identification and phosphosite localization compared with the other tested pipelines.[39] Based on these evaluation results, we provide one non-redundant, harmonized version with data across all cancer types and omics data types (see Baylor College of Medicine [BCM] pipeline for pan-cancer multi-omics data harmonization in Data S1 for details). However, we would like to emphasize that benchmarking is usually complicated by the lack of absolute ground truth, and thus more efforts should be put toward this important but challenging task. We have, therefore, also included results from multiple data processing pipelines in the data compendium. Users are encouraged to read the method description associated with each pipeline; explicit details can be found in the Data S1.

Programmatic Data Access

Simplifying data access can significantly remove barriers to community use and improve transparency and reproducibility. Therefore, CPTAC has created a software package that streams final quantitative data tables directly into a programming environment as dataframe variables (Figure 3). The Python application programming interface (API),[40] which originally streamed data from the individual cancer type publications, has been updated to provide access to the harmonized pan-cancer datasets described previously. Because data are streamed in native *pandas* dataframes, they are easily integrated with common machine learning and visualization packages such as SciKit-learn, PyTorch, Plotly, Seaborn, etc. Additionally, access to this API is also straightforward within R using the *reticulate* package for Python/R interconversion.

Computational APIs also extend the utility of CPTAC proteogenomic data by connecting them to other large public datasets.[41] We have recently expanded our popular R/Bioconductor tool, TCGAbiolinks,[42] to stream CPTAC pan-cancer data. In addition to leveraging the numerous software tools available within Bioconductor, TCGAbiolinks facilitates access to molecular data from TCGA, GENIE, MET500, GTEx, GEO, and IHEC. With TCGAbiolinks internal functions to harmonize data from diverse consortia, end-users can explore and validate hypotheses on a comprehensive library of reference datasets using sharable and reproducible codes.[43] See http://bioconductor.



*Not reported for some or all CPTAC2 cohorts (BRCA, COAD, UCEC)

**Stage not reported for GBM

Figure 2. Demographics of the CPTAC dataset

Distributions of selected clinical features among the pan-cancer cohort illustrated in Figure 1. Age is stratified by quartiles. Grade information is not available for BRCA and COAD cohorts. Stage information is not available for the GBM cohort. BMI, tobacco use, and alcohol use data are not available for BRCA, COAD, and HGSC cohorts. For survival plots, time starts at diagnosis. Additional clinical features, such as race and ethnicity, are available for exploration on the ProTrack pan-cancer sample dashboard.

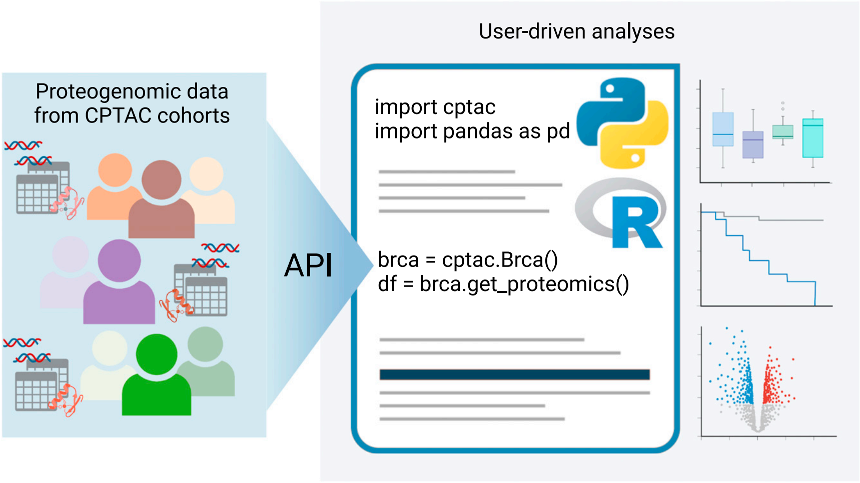
org/packages/release/bioc/html/TCGAbiolinks.html for tutorials and instructions.

Web portals for data visualization and analysis

CPTAC teams have created several web portals for visualization and exploration of pan-cancer proteogenomics data (Figure 4). Each of these websites draws from the data compendium the appropriate datasets for pan-cancer analyses.

PepQuery

Cancer genomic studies have identified many genomic aberrations that may give rise to abnormal proteins, which are promising candidates for cancer biomarkers, drug targets, and neoantigens. Validation of their expression at the protein level is a critical step toward the clinical translation of these findings. PepQuery (http://www.pepquery.org) allows quick and easy proteomic validation of genomic aberrations, such as single nucleotide variants (SNVs), insertions and deletions (INDELs), RNA editing sites, novel junctions, fusions, and novel transcription regions, using MS/MS data.[44,45] We have recently introduced a new data indexing algorithm to improve the search speed and have expanded the dataset collection in the PepQuery web server to include MS/MS data from all 10 CPTAC studies, which increased the total number of MS/MS spectra to more than one billion.[46] Through the PepQuery web server and a mirror site at PDC (https://pdc.cancer.gov), users can directly query CPTAC and other MS/MS data with a novel peptide or DNA sequence of interest to look for supporting peptide spectrum matches (PSMs). For each PSM, annotated



```
import cptac
import pandas as pd

brca = cptac.Brca()
df = brca.get_proteomics()
```

Figure 3. Streaming data with APIs
Programmatic access to CPTAC proteogenomic data across all cohorts is provided by both Python and R API.

spectra are provided for manual evaluation. Moreover, the stand-alone version and the implementation of PepQuery in the Galaxy Proteomics platform (https://proteomics.usegalaxy.eu/) support batch analysis and user-provided MS/MS data, and the identification results can be visualized using PDV.[47]

LinkedOmics and LinkedOmicsKB

LinkedOmics (http://www.linkedomics.org) is a data analysis portal that allows the characterization of any clinical or molecular feature of interest (e.g., survival, BRAF_V600E mutation, miR200c expression, or CHEK2-S422 phosphorylation) using cancer multi-omics data from TCGA and CPTAC.[48] We now provide the pan-cancer harmonized datasets described in this paper for all CPTAC cohorts in LinkedOmics. For each CPTAC study, the database stores data for >500,000 attributes including clinical attributes, mutations at site and gene levels, copy number alterations at region and gene levels, methylations at site and gene levels, mRNA expression, miRNA expression, protein expression, and PTM at site and protein levels. Using three analytical modules, including LinkFinder, LinkCompare, and LinkInterpretor, these data can be mined to reveal the consequences of genetic aberrations, characterize functions of genes and PTMs, and uncover molecular basis of cancer phenotypes.

The on-the-fly, user-defined data queries in LinkedOmics provide a high level of flexibility for analyzing CPTAC data, but performing data analysis on-the-fly is time consuming, and integrating and co-visualizing results from multiple cancer types and multiple omics data types remain challenging. To address these challenges, we further developed LinkedOmicsKB, a new knowledge portal that makes precomputed results for individual genes and phenotypes readily available through a single query. All results for a query gene or phenotype are presented on a single page with user-friendly visualization to facilitate easy comprehension. The knowledge portal is available at https://kb.linkedomics.org.

PTMcosmos

PTMcosmos is an interactive web portal designed to catalog and visualize PTMs in humans. As a key regulator of protein activity, PTMs play an essential role in our understanding of cancer and dysregulated cellular states. The PTM sites detected across all CPTAC studies were harmonized using protein sequences from UniProt's reviewed proteome, allowing for the integration of extensive annotations from many established databases including the UniProt Knowledge Base, PhosphoSitePlus, and protein 3D structures. In total, we harmonized 210,112 PTM sites and annotated them with 11,265 publications. Additionally, to investigate the relationship between genetic alterations found in cancer and PTMs that are in close spatial proximity, we included cancer somatic mutations detected in the samples of CPTAC and TCGA. Finally, we developed interactive visualization tools to allow researchers to explore the existing literature on a PTM site, the difference in abundance between tumor and normal samples, and the PTM-mutation clusters on protein structures. PTMcosmos portal is publicly available at https://ptmcosmos.wustl.edu/.

ProTrackPath: pan-cancer portal

We have developed a web application for accessing pathway enrichment scores across the pan-cancer cohorts. While previous ProTrack applications allow users to visualize normalized raw data for individual cancers,[49–51] the ProTrackPath pan-cancer portal presents pathway enrichment scores across cancer types, calculated with a single sample gene set enrichment analysis (ssGSEA).[52] The user specifies a pathway database such as Hallmark,[53] KEGG,[54] or Reactome,[55] then selects a set of pathways to visualize. An interactive heatmap is then generated, which users can customize by sorting according to any given track or toggling categorical variables on and off. Additionally, the portal includes a sample dashboard view, which allows for viewing clinical characteristics. This allows users to explore the distributions of the cancer types along with various demographic and clinical features as bar graphs. Users can filter samples by toggling features in each bar graph's interactive legend, and then populate the heatmap with their custom-generated cohort. The portal is available to the public at http://pancan.cptac-data-view.org/.

NGlycositeAtlas portal

N-Linked glycosylation is one of the most abundant protein modifications and is highly relevant to disease progression in cancer.[56] With the advances in experimental and computational approaches, glycoproteomics has provided comprehensive characterization of glycosite-specific glycosylation of glycoproteins and valuable insights into their biological functions in cancer.[57–61] However, there is still a lack of the integration of large-scale characterization of glycoproteomic data from different cancer types for pan-cancer research. We identified intact N-linked glycopeptides (see Data S1) to create a database resource termed N-GlycositeAtlas 2.0, which contains more than 90,629 intact N-linked glycopeptides (representing 5,665 N-linked glycosite-containing peptides) of over 2,000 glycoproteins from CPTAC

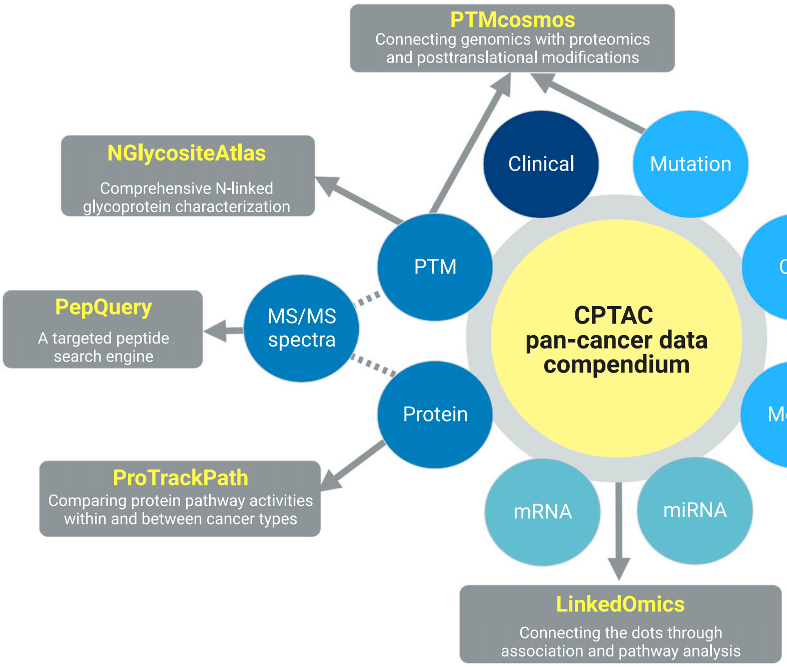

PTMcosmos
Connecting genomics with proteomics and posttranslational modifications
NGlycositeAtlas
Comprehensive N-linked glycoprotein characterization
PepQuery
A targeted peptide search engine
ProTrackPath
Comparing protein pathway activities within and between cancer types
LinkedOmics
Connecting the dots through association and pathway analysis
CPTAC pan-cancer data compendium
Clinical
Mutation
CNV
Methyl
miRNA
mRNA
Protein
MS/MS spectra
PTM

Figure 4. Web portals to CPTAC data

Multiple websites present CPTAC's proteogenomic data for visual exploration.

data. The NGlycositeAtlas database and consensus MS/MS spectra are available at https://www.biomarkercenter.org/nglycositeatlas.

ANALYTICAL CHALLENGES FOR PAN-CANCER MULTI-OMICS

With the rapid development of molecular measurement technologies, cancer datasets have become multi-modal. CPTAC has created rich proteogenomic datasets that measure DNA, RNA, and protein molecules within tumors and adjacent normal tissues (NATs). This diversity of data catalogs a comprehensive map of cellular state, providing researchers the opportunity to understand the subtle regulatory interplay between DNA mutation events that give rise to dysregulated signaling networks and the ultimate cellular phenotype. This large and comprehensive dataset presents several challenges in data integration and interpretation. In this section, we outline several important considerations for the re-use and re-analysis of proteogenomic data.

The first challenge in a proteogenomic dataset is to ensure that identifiers are harmonized. The following examples demonstrate the challenge. Many genes have multiple protein isoforms due to alternative splicing, including a noted change in splicing patterns in cancers.[62–64] Each isoform may have a unique function and combining all data into a single ''gene level'' measurement could obscure these differences. Suppose that mRNA data identifies two distinct transcripts. The transcriptomics data table, therefore, reports two database identifiers each with a separate quantitative value. If the proteomics data do not identify peptides that differentiate the two isoforms, which protein identifier should be used? To which transcript data should the protein abundance be compared? As orthogonal data types, proteomics and transcriptomics frequently identify different isoforms. This situation is equally complex when integrating PTMs, mutations, or epigenetics. If a phosphorylation or a coding mutation is observed, which protein isoform should it be associated with? Which transcript/protein should be used in comparison with methylation data? Mapping PTMs and coding mutations to different protein isoforms will make it difficult to study the impact of somatic mutations on PTMs. Thus, for a large multi-omics harmonization task such as presented here, we recommend careful consideration and transparency in reporting analytical methods. As potential solutions to mitigate the aforementioned challenges, we suggest the following: (1) using the same versions of genome assembly and gene annotation for the processing of data from all omics platforms and all cancer types; (2) reporting gene-level quantification when isoform level analysis is unrealistic; and (3) applying a consistent and transparent rule for representative isoform selection when representative isoform selection is needed but the data are isoform agnostic, e.g., phosphosite localization annotation.

A second challenge is embracing the full proteogenomic landscape as the molecular characterization of cells and tissues becomes more complete. We emphasize that each data type provides unique value and helps to clarify complex phenotypes. For example, the proteome and the transcriptome are distinct, and each provides a meaningful view of cellular processes. A rich body of research demonstrates that the mRNA and protein abundances frequently have a poorer correlation than expected,[65–69] a consequence of both translational and post-translational regulation.[70–73] As cancers are often characterized by regulatory dysfunction, exploring the source of this dysfunction can be best understood by combining transcriptomics and proteomics.[74] Similarly, the consequence of somatic mutation in kinases is best observed by combining genomics and phosphoproteomics. Indeed, many biological hypotheses can be best addressed by a fruitful combination of data types. To understand the consequence of genomic copy number variation, Gonçalves et al., combined genomics and proteomics and discovered widespread post-transcriptional attenuation in protein abundance mitigating the impact of gene amplification, especially to preserve stoichiometry in protein complexes.[34] The search for novel amino acid variants[75] and cancer neo-antigens[76–78] is inherently a proteogenomic investigation, as is the discovery of tumor-specific splice isoforms[79,80] and fusion proteins.[81] Combining all the proteogenomics levels into a single analysis is challenging, but the non-negative matrix factorization (NMF) methodology is frequently used for integrative clustering to highlight the unique contribution of each data type.[82]

Despite the great effort to harmonize the multi-omics datasets across different cancer studies, we want to emphasize that ''batch'' effects between different cancer types could still remain in the pan-cancer datasets due to both technical factors, as omics experiments of different cancer types were carried out by different labs and/or using different platforms, and biological factors, as different organs and cancer types have intrinsically different biology. Thus, when analyzing the pan-cancer data, one needs to carefully adjust for these batch effects across different cancer types. For example, when fitting a regression model to study the dependence of molecular abundances on other attributes, one can include cancer-type indicators as covariates to account for cancer-type specific mean values of molecules. Other analysis techniques, such as meta-analysis framework, could also be used to perform pan-cancer level inferences.

Finally, we focus on a challenge specific to PTMs. In the CPTAC data, we report quantitative measurement of phosphorylation and selected datasets also have data for acetylation and glycosylation. Although missing values are a regular part of all omics data, they are more pronounced in PTM data. One place where this is particularly problematic is pan-cancer analysis. If a PTM site is well quantified in one cancer type (e.g., EGFR tyrosine 1172), it may have many missing values in another, which would complicate a pan-cancer comparison of protein activation. One might be tempted to roll together all PTMs in a protein into a single measurement - e.g., the average phosphorylation state of EGFR. However, we advise against this, as PTMs at each site in a protein can be functionally independent and may not correlate across samples. Both experimental and computational approaches are being developed to improve PTM peptide identification, which will help alleviate the missing value problem in PTM proteomics.[83]

Conclusion

Pan-cancer proteogenomic data analysis requires a consistent dataset processed with a unified pipeline across all samples. Several groups have created proteogenomic datasets on cancer cohorts, exploring diverse genetic backgrounds for common cancers,[84–87] pediatric tumors,[50] or understudied tumor types.[88,89] For pan-cancer analyses it is important that individual datasets follow similar SOPs and process data in a consistent manner. Therefore, we have re-processed the data from CPTAC's 10 cancer cohorts to create a pan-cancer proteogenomic dataset. We presented the description of methods used to create this data compendium, methods of data access, as well as key considerations for pan-cancer multi-omics data analysis. This resource has been used within CPTAC for biological discoveries under various themes. We hope this also serves as a resource for the broader cancer research community to advance cancer diagnosis and treatment.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.ccell.2023.06.009.

ACKNOWLEDGMENTS

The Clinical Proteomic Tumor Analysis Consortium (CPTAC) is supported by the National Cancer Institute of the National Institutes of Health under award numbers U24CA210955, U24CA210985, U24CA210986, U24CA210954, U24CA210967, U24CA210972, U24CA210979, U24CA210993, U01CA214114, U01CA214116, and U01CA214125.

Additional funding support was provided by NIH awards R33CA263705, T32CA203690, T32GM136542 and Leidos Biomed contract 20X042F01/TO01 and the Simmons Center for Cancer Research. Figure 1 representing the data overview for this manuscript was created using BioRender.com. This project has been funded in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. 75N91019D00024, Task Order No. 7591029F00029, and Contract No. 75N91019D00024, Task Order 75N91020F00029. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the U.S. Government.

AUTHOR CONTRIBUTIONS

Study Conception & Design: G.G., L.D., A.I.N., P.W., A.I.R., B.Z., and S.H.P.

Formal Analysis: Y. Li, Y.D., F.D.V.L., and Y.G.

Visualization: A.P.C.,Y.G, Y.H., Y. Liao, B.R., S.S., and X.Y.

Data Curation: Y. Li, Y.D., F.D.V.L., Y.G., A.P.C., F.A., Y.A., S.A., C.B., S.C., R.C., P.C., M.C., A.C., D.C.Z., C.D., M.E.S., D.F., S.M.F., A.F., T.G., Z.H.G., D.H., M.H., R.H., Y.H., E.J.J., J.J., W.J., L.K., K.K., R.J.K., J.L., W.L., Y. Liao, C.M.L., W. Ma, L.M., M.J.M., F.M.R., W. McKerrow, N.N., R.O., A.P., P.P., B.R., P.R., K.V.R., D.R., S.S., M.S., T.S., Z.S., D.S., X.S., E.S., N.V.T., R.R.T., M.T., L.W., J.M.W., Y. Wang, B.W., Y. Wu, M.A.W., Y.X., L.Y., X.Y., H.Z., Q.Z., M.Z., G.G., L.D., A.I.N., P.W., A.I.R., B.Z., and S.H.P.

Writing – Original Drafts: Y. Li, Y.D., F.D.V.L., Y.G., A.P.C., A.C., Y.H., L.D., P.W., B.Z., and S.H.P.

Writing – Review & Editing: B.Z. and S.H.P.

Supervision: D.F., K.V.R., H.Z., G.G., L.D., A.I.N., P.W., A.I.R., B.Z., and S.H.P.

DECLARATION OF INTERESTS

F.A. is an inventor on a patent application related to SignatureAnalyzer-GPU filed by the Broad Institute and is an employee and shareholder of Illumina Inc. since 8 November 2021.

REFERENCES

1. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93.

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.

3. Alfaro, J.A., Sinha, A., Kislinger, T., and Boutros, P.C. (2014). Onco-proteogenomics: cancer proteomics joins forces with genomics. Nat. Methods *11*, 1107–1113.

4. Mani, D.R., Krug, K., Zhang, B., Satpathy, S., Clauser, K.R., Ding, L., Ellis, M., Gillette, M.A., and Carr, S.A. (2022). Cancer proteogenomics: current impact and future prospects. Nat. Rev. Cancer *22*, 298–313.

5. Rodriguez, H., Zenklusen, J.C., Staudt, L.M., Doroshow, J.H., and Lowy, D.R. (2021). The next horizon in precision oncology: proteogenomics to inform cancer diagnosis and treatment. Cell *184*, 1661–1670.

6. Zhang, B., Whiteaker, J.R., Hoofnagle, A.N., Baird, G.S., Rodland, K.D., and Paulovich, A.G. (2019). Clinical potential of mass spectrometry-based proteogenomics. Nat. Rev. Clin. Oncol. *16*, 256–268.

7. Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S., et al. (2014). Proteogenomic characterization of human colon and rectal cancer. Nature *513*, 382–387.

8. Mertins, P., Yang, F., Liu, T., Mani, D.R., Petyuk, V.A., Gillette, M.A., Clauser, K.R., Qiao, J.W., Gritsenko, M.A., Moore, R.J., et al. (2014). Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. Mol. Cell. Proteomics *13*, 1690–1704.

9. 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*, 561–577.e22.

10. 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.

11. 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. (2020). Integrated proteogenomic characterization of clear cell renal cell carcinoma. Cell *180*, 207.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. Geffen, Y., Anand, S., Akiyama, Y., Yaron, T.M., Song, Y., Johnson, J.L., Govindan, A., Özgün, B., Li, Y., Huntsman, E., et al. Clinical Proteomic Tumor Analysis Consortium (2023). Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation. Cell *186*. Published online August 14, 2023.

22. 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. Clinical Proteomic Tumor Analysis Consortium (2023). Pan-cancer proteogenomics connects oncogenic drivers to functional states. Cell *186*. Published online August 14, 2023.

23. Wu, P., Heins, Z.J., Muller, J.T., Katsnelson, L., de Bruijn, I., Abeshouse, A.A., Schultz, N., Fenyö, D., and Gao, J. (2019). Integration and analysis of CPTAC proteomics data in the context of cancer genomics in the cBioPortal. Mol. Cell. Proteomics *18*, 1893–1898.

24. Zhan, X., Cheng, J., Huang, Z., Han, Z., Helm, B., Liu, X., Zhang, J., Wang, T.-F., Ni, D., and Huang, K. (2019). Correlation analysis of histopathology and proteogenomics data for breast cancer. Mol. Cell. Proteomics *18*, S37–S51.

25. Chen, F., Chandrashekar, D.S., Varambally, S., and Creighton, C.J. (2019). Pan-cancer molecular subtypes revealed by mass-spectrometry-based proteomic characterization of more than 500 human cancers. Nat. Commun. *10*, 5679.

26. Tong, M., Yu, C., Zhan, D., Zhang, M., Zhen, B., Zhu, W., Wang, Y., Wu, C., He, F., Qin, J., and Li, T. (2019). Molecular subtyping of cancer and nomination of kinase candidates for inhibition with phosphoproteomics: reanalysis of CPTAC ovarian cancer. EBioMedicine *40*, 305–317.

27. 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.

28. Huang, W., Chen, J., Weng, W., Xiang, Y., Shi, H., and Shan, Y. (2020). Development of cancer prognostic signature based on pan-cancer proteomics. Bioengineered *11*, 1368–1381.

29. Zhao, J., Cheng, M., Gai, J., Zhang, R., Du, T., and Li, Q. (2020). SPOCK2 serves as a potential prognostic marker and correlates with immune infiltration in lung adenocarcinoma. Front. Genet. *11*, 588499.

30. Wu, Z.-H., and Yang, D.-L. (2020). Identification of a protein signature for predicting overall survival of hepatocellular carcinoma: a study based on data mining. BMC Cancer *20*, 720.

31. Kahles, A., Lehmann, K.-V., Toussaint, N.C., Hüser, M., Stark, S.G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C.; 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.

32. Peng, X., Xu, X., Wang, Y., Hawke, D.H., Yu, S., Han, L., Zhou, Z., Mojumdar, K., Jeong, K.J., Labrie, M., et al. (2018). A-to-I RNA editing contributes to proteomic diversity in cancer. Cancer Cell *33*, 817–828.e7.

33. Prakash, A., Taylor, L., Varkey, M., Hoxie, N., Mohammed, Y., Goo, Y.A., Peterman, S., Moghekar, A., Yuan, Y., Glaros, T., et al. (2021). Reinspection of a Clinical Proteomics Tumor Analysis Consortium (CPTAC) dataset with cloud computing reveals abundant post-translational modifications and protein sequence variants. Cancers *13*, 5034.

34. Gonçalves, E., Fragoulis, A., Garcia-Alonso, L., Cramer, T., Saez-Rodriguez, J., and Beltrao, P. (2017). Widespread post-transcriptional attenuation of genomic copy-number variation in cancer. Cell Syst. *5*, 386–398.e4.

35. Ryan, C.J., Kennedy, S., Bajrami, I., Matallanas, D., and Lord, C.J. (2017). A Compendium of co-regulated protein complexes in breast cancer reveals collateral loss events. Cell Syst. *5*, 399–409.e5.

36. Ewing, A.D., Houlahan, K.E., Hu, Y., Ellrott, K., Caloian, C., Yamaguchi, T.N., Bare, J.C., P'ng, C., Waggott, D., Sabelnykova, V.Y., et al. (2015). Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat. Methods *12*, 623–630.

37. Ellrott, K., Bailey, M.H., Saksena, G., Covington, K.R., Kandoth, C., Stewart, C., Hess, J., Ma, S., Chiotti, K.E., McLellan, M., et al. (2018). Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell Syst. *6*, 271–281.e7.

38. Wen, B., Jaehnig, E.J., and Zhang, B. (2022). OmicsEV: a tool for comprehensive quality evaluation of omics data tables. Bioinformatics *38*, 5463–5465. btac698.

39. Jiang, W., Wen, B., Li, K., Zeng, W.-F., da Veiga Leprevost, F., Moon, J., Petyuk, V.A., Edwards, N.J., Liu, T., Nesvizhskii, A.I., and Zhang, B. (2021). Deep-learning-derived evaluation metrics enable effective benchmarking of computational tools for phosphopeptide identification. Mol. Cell. Proteomics *20*, 100171.

40. Lindgren, C.M., Adams, D.W., Kimball, B., Boekweg, H., Tayler, S., Pugh, S.L., and Payne, S.H. (2021). Simplified and unified access to cancer proteogenomic data. J. Proteome Res. *20*, 1902–1910.

41. Colaprico, A., Olsen, C., Bailey, M.H., Odom, G.J., Terkelsen, T., Silva, T.C., Olsen, A.V., Cantini, L., Zinovyev, A., Barillot, E., et al. (2020). Interpreting pathways to discover cancer driver genes with Moonlight. Nat. Commun. *11*, 69.

42. Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I., et al. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. *44*, e71.

43. Lehmann, B.D., Colaprico, A., Silva, T.C., Chen, J., An, H., Ban, Y., Huang, H., Wang, L., James, J.L., Balko, J.M., et al. (2021). Multi-omics analysis identifies therapeutic vulnerabilities in triple-negative breast cancer subtypes. Nat. Commun. *12*, 6276.

44. Wen, B., Li, K., Zhang, Y., and Zhang, B. (2020). Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. Nat. Commun. *11*, 1759.

45. Wen, B., Wang, X., and Zhang, B. (2019). PepQuery enables fast, accurate, and convenient proteomic validation of novel genomic alterations. Genome Res. *29*, 485–493.

46. Wen, B., and Zhang, B. (2023). PepQuery2 democratizes public MS proteomics data for rapid peptide searching. Nat. Commun. *14*, 2213.

47. Li, K., Vaudel, M., Zhang, B., Ren, Y., and Wen, B. (2019). PDV: an integrative proteomics data viewer. Bioinformatics *35*, 1249–1251.

48. Vasaikar, S.V., Straub, P., Wang, J., and Zhang, B. (2018). LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res. *46*, D956–D963.

49. Calinawan, A.P., Song, X., Ji, J., Dhanasekaran, S.M., Petralia, F., Wang, P., and Reva, B. (2020). ProTrack: an interactive multi-omics data browser for proteogenomic studies. Proteomics *20*, e1900359.

50. 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.

51. Huang, D., Chowdhury, S., Wang, H., Savage, S.R., Ivey, R.G., Kennedy, J.J., Whiteaker, J.R., Lin, C., Hou, X., Oberg, A.L., et al. (2021). Multiomic analysis identifies CPT1A as a potential therapeutic target in platinum-refractory, high-grade serous ovarian cancer. Cell Rep. Med. *2*, 100471.

52. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

53. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425.

54. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45*, D353–D361.

55. Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., et al. (2020). The reactome pathway knowledgebase. Nucleic Acids Res. *48*, D498–D503.

56. Pinho, S.S., and Reis, C.A. (2015). Glycosylation in cancer: mechanisms and clinical implications. Nat. Rev. Cancer *15*, 540–555.

57. Dong, M., Lih, T.M., Chen, S.-Y., Cho, K.-C., Eguez, R.V., Höti, N., Zhou, Y., Yang, W., Mangold, L., Chan, D.W., et al. (2020). Urinary glycoproteins associated with aggressive prostate cancer. Theranostics *10*, 11892–11907.

58. Hu, Y., Pan, J., Shah, P., Ao, M., Thomas, S.N., Liu, Y., Chen, L., Schnaubelt, M., Clark, D.J., Rodriguez, H., et al. (2020). Integrated proteomic and glycoproteomic characterization of human high-grade serous ovarian carcinoma. Cell Rep. *33*, 108276.

59. Pan, J., Hu, Y., Sun, S., Chen, L., Schnaubelt, M., Clark, D., Ao, M., Zhang, Z., Chan, D., Qian, J., and Zhang, H. (2020). Glycoproteomics-based signatures for tumor subtyping and clinical outcome prediction of high-grade serous ovarian cancer. Nat. Commun. *11*, 6139.

60. Tabang, D.N., Ford, M., and Li, L. (2021). Recent advances in mass spectrometry-based glycomic and glycoproteomic studies of pancreatic diseases. Front. Chem. *9*, 707387.

61. Zhang, Y., Jiao, J., Yang, P., and Lu, H. (2014). Mass spectrometry-based N-glycoproteomics for cancer biomarker discovery. Clin. Proteomics *11*, 18.

62. Climente-González, H., Porta-Pardo, E., Godzik, A., and Eyras, E. (2017). The functional impact of alternative splicing in cancer. Cell Rep. *20*, 2215–2226.

63. Venables, J.P. (2004). Aberrant and alternative splicing in cancer. Cancer Res. *64*, 7647–7654.

64. Venables, J.P., Klinck, R., Koh, C., Gervais-Bird, J., Bramard, A., Inkel, L., Durand, M., Couture, S., Froehlich, U., Lapointe, E., et al. (2009). Cancer-associated regulation of alternative splicing. Nat. Struct. Mol. Biol. *16*, 670–676.

65. Fortelny, N., Overall, C.M., Pavlidis, P., and Freue, G.V.C. (2017). Can we predict protein from mRNA levels? Nature *547*, E19–E20.

66. McManus, J., Cheng, Z., and Vogel, C. (2015). Next-generation analysis of gene expression regulation–comparing the roles of synthesis and degradation. Mol. Biosyst. *11*, 2680–2689.

67. Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol. Syst. Biol. *7*, 548.

68. Payne, S.H. (2015). The utility of protein and mRNA correlation. Trends Biochem. Sci. *40*, 1–3.

69. Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. *13*, 227–232.

70. Aviner, R., Shenoy, A., Elroy-Stein, O., and Geiger, T. (2015). Uncovering hidden layers of cell cycle regulation through integrative multi-omic analysis. PLoS Genet. *11*, e1005554.

71. Cai, Y., Yu, X., Hu, S., and Yu, J. (2009). A brief review on the mechanisms of miRNA regulation. Dev. Reprod. Biol. *7*, 147–154.

72. Grzmil, M., and Hemmings, B.A. (2012). Translation regulation as a therapeutic target in cancer. Cancer Res. *72*, 3891–3900.

73. He, R.-Z., Luo, D.-X., and Mo, Y.-Y. (2019). Emerging roles of lncRNAs in the post-transcriptional regulation in cancer. Genes Dis. *6*, 6–15.

74. Tang, W., Zhou, M., Dorsey, T.H., Prieto, D.A., Wang, X.W., Ruppin, E., Veenstra, T.D., and Ambs, S. (2018). Integrated proteotranscriptomics of breast cancer reveals globally increased protein-mRNA concordance associated with subtypes and survival. Genome Med. *10*, 94.

75. Da Cunha, L.M., Terrematte, P., Fiuza, T.D.S., Silva, V.L.D., Kroll, J.E., De Souza, S.J., and De Souza, G.A. (2022). dbPepVar: a novel cancer proteogenomics database. IEEE Access *10*, 90982–90994.

76. Cleyle, J., Hardy, M.-P., Minati, R., Courcelles, M., Durette, C., Lanoix, J., Laverdure, J.-P., Vincent, K., Perreault, C., and Thibault, P. (2022). Immunopeptidomic analyses of colorectal cancers with and without microsatellite instability. Mol. Cell. Proteomics *21*, 100228.

77. Polyakova, A., Kuznetsova, K., and Moshkovskii, S. (2015). Proteogenomics meets cancer immunology: mass spectrometric discovery and analysis of neoantigens. Expert Rev. Proteomics *12*, 533–541.

78. Xiang, R., Ma, L., Yang, M., Zheng, Z., Chen, X., Jia, F., Xie, F., Zhou, Y., Li, F., Wu, K., and Zhu, Y. (2021). Increased expression of peptides from non-coding genes in cancer proteomics datasets suggests potential tumor neoantigens. Commun. Biol. *4*, 496.

79. Miller, R.M., Jordan, B.T., Mehlferber, M.M., Jeffery, E.D., Chatzipantsiou, C., Kaur, S., Millikin, R.J., Dai, Y., Tiberi, S., Castaldi, P.J., et al. (2022). Enhanced protein isoform characterization through long-read proteogenomics. Genome Biol. *23*, 69.

80. Hatakeyama, K., Ohshima, K., Fukuda, Y., Ogura, S.I., Terashima, M., Yamaguchi, K., and Mochizuki, T. (2011). Identification of a novel protein isoform derived from cancer-related splicing variants using combined analysis of transcriptome and proteome. Proteomics *11*, 2275–2282.

81. Kim, C.-Y., Na, K., Park, S., Jeong, S.-K., Cho, J.-Y., Shin, H., Lee, M.J., Han, G., and Paik, Y.-K. (2019). FusionPro, a versatile proteogenomic tool for identification of novel fusion transcripts and their potential translation products in cancer cells. Mol. Cell. Proteomics *18*, 1651–1668.

82. Mani, D.R., Maynard, M., Kothadia, R., Krug, K., Christianson, K.E., Heiman, D., Clauser, K.R., Birger, C., Getz, G., and Carr, S.A. (2021). PANOPLY: a cloud-based platform for automated and reproducible proteogenomic data analysis. Nat. Methods *18*, 580–582.

83. Bekker-Jensen, D.B., Bernhardt, O.M., Hogrebe, A., Martinez-Val, A., Verbeke, L., Gandhi, T., Kelstrup, C.D., Reiter, L., and Olsen, J.V. (2020). Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. Nat. Commun. *11*, 787.

84. 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.

85. Lehtiö, J., Arslan, T., Siavelis, I., Pan, Y., Socciarelli, F., Berkovska, O., Umer, H.M., Mermelekas, G., Pirmoradian, M., Jönsson, M., et al. (2021). Proteogenomics of non-small cell lung cancer reveals molecular subtypes associated with specific therapeutic targets and immune evasion mechanisms. Nat. Cancer *2*, 1224–1242.

86. Xu, J.-Y., Zhang, C., Wang, X., Zhai, L., Ma, Y., Mao, Y., Qian, K., Sun, C., Liu, Z., Jiang, S., et al. (2020). Integrative proteomic characterization of human lung adenocarcinoma. Cell *182*, 245–261.e17.

87. Qu, Y., Feng, J., Wu, X., Bai, L., Xu, W., Zhu, L., Liu, Y., Xu, F., Zhang, X., Yang, G., et al. (2022). A proteogenomic analysis of clear cell renal cell carcinoma in a Chinese population. Nat. Commun. *13*, 2052.

88. Shi, X., Sun, Y., Shen, C., Zhang, Y., Shi, R., Zhang, F., Liao, T., Lv, G., Zhu, Z., Jiao, L., et al. (2022). Integrated proteogenomic characterization of medullary thyroid carcinoma. Cell Discov. *8*, 120.

89. Dong, L., Lu, D., Chen, R., Lin, Y., Zhu, H., Zhang, Z., Cai, S., Cui, P., Song, G., Rao, D., et al. (2022). Proteogenomic characterization identifies clinically relevant subgroups of intrahepatic cholangiocarcinoma. Cancer Cell *40*, 70–87.e15.

Cancer Cell, Volume *41*

Supplemental information

Proteogenomic data and resources

for pan-cancer analysis

Yize Li, Yongchao Dou, Felipe Da Veiga Leprevost, Yifat Geffen, Anna P. Calinawan, François Aguet, Yo Akiyama, Shankara Anand, Chet Birger, Song Cao, Rekha Chaudhary, Padmini Chilappagari, Marcin Cieslik, Antonio Colaprico, Daniel Cui Zhou, Corbin Day, Marcin J. Domagalski, Myvizhi Esai Selvan, David Fenyö, Steven M. Foltz, Alicia Francis, Tania Gonzalez-Robles, Zeynep H. Gümüş, David Heiman, Michael Holck, Runyu Hong, Yingwei Hu, Eric J. Jaehnig, Jiayi Ji, Wen Jiang, Lizabeth Katsnelson, Karen A. Ketchum, Robert J. Klein, Jonathan T. Lei, Wen-Wei Liang, Yuxing Liao, Caleb M. Lindgren, Weiping Ma, Lei Ma, Michael J. MacCoss, Fernanda Martins Rodrigues, Wilson McKerrow, Ngoc Nguyen, Robert Oldroyd, Alexander Pilozzi, Pietro Pugliese, Boris Reva, Paul Rudnick, Kelly V. Ruggles, Dmitry Rykunov, Sara R. Savage, Michael Schnaubelt, Tobias Schraink, Zhiao Shi, Deepak Singhal, Xiaoyu Song, Erik Storrs, Nadezhda V. Terekhanova, Ratna R. Thangudu, Mathangi Thiagarajan, Liang-Bo Wang, Joshua M. Wang, Ying Wang, Bo Wen, Yige Wu, Matthew A. Wyczalkowski, Yi Xin, Lijun Yao, Xinpei Yi, Hui Zhang, Qing Zhang, Maya Zuhl, Gad Getz, Li Ding, Alexey I. Nesvizhskii, Pei Wang, Ana I. Robles, Bing Zhang, Samuel H. Payne, and Clinical Proteomic Tumor Analysis Consortium

Supplemental Data File for - Proteogenomic Data and Resources for Pan-Cancer Analysis

Key Resources Table

Software and Algorithms		
bam-readcount v0.8	McDonnell Genome Institute	https://github.com/genome/bam-readcount
BWA v0.7.17-r1188	(Li and Durbin, 2009)[1]	http://bio-bwa.sourceforge.net/
CharGer v.0.5.4	(Scott et al., 2019)[2]	https://github.com/ding-lab/CharGer/tree/v0.5.4
COCOON	Li Ding Lab	https://github.com/ding-lab/COCOONS
WES hg38 characterization pipeline	Getz Lab	https://terra.bio/
bam-readcount v0.8	McDonnell Genome Institute	https://github.com/genome/bam-readcount
GATK4's CalculateContamination	GATK	https://gatk.broadinstitute.org/hc/en-us/articles/360036888972-CalculateContamination
GATK4 Picard tools	GATK	https://github.com/broadinstitute/picard
GATK4 Funcotator	GATK	https://gatk.broadinstitute.org/hc/en-us/articles/360037224432-Funcotator
DeTiN	(Taylor-Weiner et al., 2018)[3]	https://github.com/getzlab/deTiN
Spectrum Mill	Karl R. Clauser, Steven Carr Lab	https://proteomics.broadinstitute.org/
GISTIC2.0	(Mermel et al., 2011)[4]	ftp://ftp.broadinstitute.org/pub/GISTIC2.0/GISTIC_2_0_23.tar

		.gz
MutSig2CV	(Lawrence et al., 2013)[5]	https://github.com/getzlab/MutSig2CV
ConsensusClusterPlus v1.48.0	(Wilkerson and Hayes, 2010)[6]	https://bioconductor.org/packages/ConsensusClusterPlus/
COSMIC Mutational Signatures v3	(Alexandrov et al., 2020)[7]	https://cancer.sanger.ac.uk/cosmic/signatures/
DEPO	(Sun et al., 2018)[8]	http://dinglab.wustl.edu/depo
EricScript v0.5.5	(Benelli et al., 2012)[9]	https://sites.google.com/site/bioericscript/
germlinewrapper v1.1	Li Ding Lab	https://github.com/ding-lab/germlinewrapper
HTSeq v0.11.2	(Anders et al., 2015)[10]	https://github.com/simon-anders/htseq
INTEGRATE v0.2.6	(Zhang et al., 2016)[11]	https://sourceforge.net/projects/integrate-fusion/
Manta v1.6.0	(Chen et al., 2016)[12]	https://github.com/Illumina/manta
MuTect v1.1.7	(Cibulskis et al., 2013)[13]	https://github.com/broadinstitute/mutect
Pindel v0.2.5	(Ye et al., 2009)[14]	https://github.com/genome/pindel
Python v3.7	Python Software Foundation	https://www.python.org/
R v3.6	R Development Core Team	https://www.R-project.org
R-rollup	(Polpitiya et al., 2008)[15]	https://omics.pnl.gov/software/danter
Samtools v1.2	(Li et al., 2009)[16]	https://www.htslib.org/
SignatureAnalyzer	(Alexandrov et al., 2020)[7]	https://github.com/broadinstitute/getzlab-SignatureAnalyzer
somaticwrapper v1.3 and v1.5	Li Ding Lab	https://github.com/ding-lab/somaticwrapper
STAR-Fusion v1.5.0	(Haas et al., 2019)[17]	https://github.com/STAR-Fusion/STAR-Fusion
Strelka v2.9.2	(Kim et al., 2018)[18]	https://github.com/Illumina/strelka
UpSetR	(Conway et al., 2017)[19]	https://github.com/hms-dbmi/UpSetR/
VarScan v2.3.8	(Koboldt et al., 2012)[20]	https://dkoboldt.github.io/varscan/
xCell v1.2	(Aran et al., 2017)[21]	http://xcell.ucsf.edu/
CIBERSORTx	(Newman et al., 2019)[22]	https://cibersortx.stanford.edu/

BIC-seq2	(Xi et al., 2016)[23]	http://compbio.med.harvard.edu/BIC-seq/
Methylation analysis	Li Ding Lab	https://github.com/ding-lab/cptac_methylation
Ancestry prediction	Li Ding Lab	https://github.com/ding-lab/ancestry
MSI prediction	Li Ding Lab	https://github.com/ding-lab/msisensor
NetMHC4	(Andreatta and Nielsen, 2016)[24]	https://services.healthtech.dtu.dk/service.php?NetMHC-4.0
DNAScope	(Freed et al., 2017)[25]	https://www.biorxiv.org/content/10.1101/115717v2
EIGENSOFT	(Patterson et al., 2006)[26]	https://www.hsph.harvard.edu/alkes-price/software/
RNA-SeQC 2.3.6	(Graubert et al., 2021)[27]	https://github.com/getzlab/rnaseqc
miRNA quantification	Li Ding Lab	https://github.com/ding-lab/CPTAC_miRNA
MSFragger	(Kong et al., 2017)[28]	https://msfragger.nesvilab.org/
Philosopher	(da Veiga Leprevost et al., 2020)[29]	https://github.com/Nesvilab/philosopher
TMT-Integrator	(Djomehri et al., 2020)[30]	http://tmt-integrator.nesvilab.org/
Combat	(Johnson et al., 2007)[31]	http://biosun1.harvard.edu/complab/batch/
DreamAI	(Ma et al., 2020)[32]	https://www.biorxiv.org/content/10.1101/2020.07.21.214205v2
GPQuest	(Hu et al., 2020)[33]	https://www.biomarkercenter.org/gpquest
Glycositeatlas	(Sun et al., 2019)[34]	http://nglycositeatlas.biomarkercenter.org
GlycomeDB	(Ranzinger et al., 2011)[35]	http://www.glycome-db.org
OmicsOne	(Zhang et al., 2021)[36]	https://github.com/huizhanglab-jhu/Omics%20One

Data and Code Availability

As detailed in the main manuscript there are multiple methods of data dissemination depending on the user's preference for raw or processed data. Code for the various analyses is also available, as noted below in the subsections describing each analysis. In brief, raw and processed proteomics data can be accessed via Proteomic Data Commons (PDC) at https://pdc.cancer.gov. A direct link to all harmonized proteomics data tables is –

https://pdc.cancer.gov/pdc/cptac-pancancer. Raw genomic and transcriptomic data files can be accessed via the Genomic Data Commons (GDC) Data Portal at https://portal.gdc.cancer.gov. All processed CPTAC pan-cancer data can be accessed via the Cancer Data Service (CDS). The CPTAC pan-cancer data hosted in CDS is controlled data. Access to controlled access data on CDS is through the NCI DAC approved, dbGaP compiled whitelists. Users can access the data for analysis through the Seven Bridges Cancer Genomics Cloud (SB-CGC), accessible with a queryable web portal through the Seven Bridges Cancer Genomic Cloud with dbGaP Study Accession, phs001287.v16.p6. Data tables harmonized by the BCM pipeline can be accessed at https://kb.linkedomics.org/download.

Sample Collection

Prospective biospecimen collection (tumor, germline blood and adjacent normal samples where feasible) followed a tumor type specific protocol and standard operating procedures (SOPs), where sample collection, qualification and processing were optimized for both genomics and proteomics [37–46]. CPTAC samples were collected by 30+ tissue source sites from both domestic and international locations and processed by a central biospecimen core resource. The samples were pathology qualified by a general pathologist and later reconfirmed by a disease-specific expert pathologist through histopathology image review and immunohistochemistry assays where applicable.

A data compendium

The CPTAC dataset includes genomics, transcriptomics, proteomics, PTM-proteomics and clinical/demographic data. Each data type can be processed several ways to produce different outputs. For example, Whole Genome Sequencing (WGS) can be processed to identify somatic mutations, or with different algorithms it could be processed to identify copy number variations. Additionally, different pipelines can process data towards a similar goal, e.g., somatic mutations can be identified via numerous distinct tools and pipelines.

In the descriptions below, a set of software analysis pipelines are listed, organized by category (e.g., genomics) and then the specific resulting data type (e.g. somatic mutation). We also note instances where multiple pipelines have been used to generate the same data type. In these situations, we include a description of each pipeline. In the data release, the different versions of the data types are often referenced by the institution that produced them, e.g., the BCM transcriptomics pipeline versus WashU transcriptomics pipeline.

Clinical and demographic data

Clinical information was downloaded from CPTAC Data Coordinating Center (DCC). Columns with different titles which have the same meaning were unified. For the Breast, Colon and Ovarian tumors, consent age was converted from months to years. Numbers of examined para-aortic and other lymph nodes were combined for UCEC, as well as numbers of pelvic, para-aortic and other lymph nodes positive for tumor. Follow-up information was processed to extract earliest time for recurrence and latest time for survival, as well as measure of success of outcome, ecog performance status score, and karnofsky performance status score at date of last contact or death. Two versions of both overall and progression-free survival intervals were calculated, one pair employed the date of initial pathological diagnosis as the index date; the other pair was calculated from the date of sample collection. Earliest of the events of new tumor after initial diagnosis, tumor progression, and tumor recurrence was selected as the recurrence event for GBM. Note on the type of the event was added. Data on additional surgery for loco-regional new tumors and for metastasis were merged. Values for sex, race, ethnicity, vital status, tumor focality, type of new tumor, pathological stage, and histologic grade were standardized. Major characteristics of the cohort are shown in Figure 2 of the main manuscript.

Genomics

Somatic Mutation Calling (pipeline from Washington University in St Louis, WashU)

Somatic mutations were called by the Somaticwrapper pipeline v1.6 (https://github.com/ding-lab/somaticwrapper), which includes four different callers, i.e., Strelka v.2.9.2 (ref [18]), MUTECT v1.1.7 (ref [13]), VarScan v.2.3.8 (ref [20]), and Pindel v.0.2.5 (ref [14]) from WES. We kept the exonic SNVs called by any two callers among MUTECT v1.1.7, VarScan v.2.3.8, and Strelka v.2.9.2 and indels called by any two callers among VarScan v.2.3.8, Strelka v.2.9.2, and Pindel v.0.2.5. For the merged SNVs and indels, we applied a 14X and 8X coverage cutoff for tumor and normal, separately. We also filtered SNVs and indels by a minimal variant allele frequency (VAF) of 0.05 in tumors and a maximal VAF of 0.02 in normal samples. We filtered any SNV, which was within 10bp of an indel found in the same tumor sample. Finally, we rescued the rare mutations with VAF of [0.015, 0.05) in cancer driver genes based on the gene consensus list reported in ref [47].

DNP Calling (pipeline from Washington University in St Louis, WashU)

In step 12 of Somaticwrapper pipeline v1.6 (https://github.com/ding-lab/somaticwrapper), it combined adjacent SNVs into DNP by using COCOON (https://github.com/ding-lab/COCOONS): As input, COCOON takes a MAF file from standard variant calling pipeline. First, it extracts variants within a 2bp window as DNP candidate sets. Next, suppose the corresponding BAM files used for variant calling are available. In that case, it extracts the reads (denoted as n_t) spanning all candidate DNP locations in each variant set, and then counts the number of reads with all the co-occurring variants (denoted as n_c) to calculate co-occurrence rate (r_c=n_c/n_t); If r_c ≥ 0.8, the nearby SNVs will be combined into DNP and it also updates annotation for the DNPs from the same codon based on the transcript and coordinates information in the MAF file.

Somatic mutation calling (pipeline from Broad Institute of MIT and Harvard)

Patient whole exome sequencing (WES) data for matched tumor/normal samples were analyzed using the Getz Lab's production hg38 WES characterization pipeline. The hg38 characterization pipeline runs on the Terra cloud-based analysis platform (https://terra.bio/). This pipeline is the Getz Lab's standard computational workflow, and the analysis steps are organized into five modules: (1) DNA Sequence Data Quality Control (including GATK4's CalculateContamination [version GATK 4.1.4.1] and GATK4 Picard tools [version GATK 4.0.5.1]). (2) Somatic Copy Number Analysis (GATK4 Best Practices Workflow [version GATK 4.1.4.1]). (3) Somatic Variant Discovery, which includes the discovery of single-nucleotide variants (SNVs) and insertions/deletions (indels), using MuTect (ref [13]) and Manta+Strelka v2 (refs [12,18]). Next, deTiN v1.8.9 (ref [3]) is run to account for and rescue tumor-in-normal contamination. The resulting SSNV and indel VCFs are each run through the GATK4 Funcotator (version GATK 4.1.4.1). (4) Post-Discovery Filtering, which employs a collection of filters for removal of alignment artifacts, germline variants, and common sequencing artifacts that occur in normal panels. (5) Merging of adjacent somatic Single Nucleotide Polymorphisms (SNPs) into dinucleotide polymorphisms (DNPs), trinucleotide polymorphisms (TNPs), and Oligo-nucleotide polymorphism (ONPs). (see CPTAC Pan-Can PTM manuscript by the Getz Lab for further details)

Somatic Mutation Callset Harmonization (pipeline from Broad Institute of MIT and Harvard)

The per patient variant calls employed by the CPTAC PanCan working group are derived from the harmonization of variant calls made independently by the Broad and WashU teams. First, we filter calls outside of the inosine chemical erasing (ICE) interval (Genomics Platform at the Broad Institute) and apply a "panel-of-normals" built from aggregating CPTAC and TCGA cohorts for consistency between the two pipelines. In addition, to account for differences in the two mutation callsets we: (i) removed all calls with Variant Allele Frequency (VAF) < 0.05 from both pipelines and rescued only high confident calls, and (ii) long ONPs were collapsed to shorter ONPs by imposing a more stringent merging cerita that requires a 2bp gap length at max. Next, we used Asymtools2 (ref [48]) to identify a sequencing artifact affecting CPTAC2 whole exome sequencing. We then corrected for the sequencing artifact by ranking context-specific mutations by its allelic fraction. Finally, the functional impact of harmonized calls was annotated using GATK Funcotator.

WGS Copy Number Variant Calling (pipeline from Washington University in St Louis, WashU)

We used BIC-seq2 (ref [23]), a read-depth-based CNV calling algorithm to detect somatic copy number variation (CNVs) from the WGS data of tumors. Briefly, BIC-seq2 divides genomic regions into disjoint bins and counts uniquely aligned reads in each bin. Then, it combines neighboring bins into genomic segments with similar copy numbers iteratively based on Bayesian Information Criteria (BIC), a statistical criterion measuring both the fitness and complexity of a statistical model. We used paired-sample CNV calling that takes a pair of samples as input and detects genomic regions with different copy numbers between the two samples. We used a bin size of ∼100 bp and a lambda of 3 (a smoothing parameter for CNV segmentation). We recommend calling segments as copy gain or loss when their log2 copy ratios were larger than 0.2 or smaller than -0.2, respectively (according to the BIC-seq publication).

Genomic data post-processing, GISTIC, and MutSig analysis (pipeline from Broad Institute of MIT and Harvard)

The Genomic Identification of Significant Targets in Cancer (GISTIC2.0) algorithm (ref [4]) was used to identify significantly amplified or deleted focal-level and arm-level events, with Q value <0.25 considered significant (default parameters were used). MutSig2CV was run on the union MAF to evaluate the significance of mutated genes as well as estimate the mutation densities of samples. These results were constrained to genes in the Cancer Gene Census [49], with false discovery rates (q values) recalculated. Genes of q value < 0.1 were declared significant.

Germline SNP and short indel discovery from WGS (pipeline from Broad Institute of MIT and Harvard)
Germline mutations were identified using the GATK4 SNPS + Indels best practice workflow (https://github.com/gatk-workflows/gatk4-rnaseq-germline-snps-indels). This workflow consists of three sub-workflows: (i) Processing-For-Variant-Discovery, (ii) Haplotypecaller-GVCF [50], and (iii) Joint-Discovery. Details regarding the specific Terra workflows used to conduct this analysis can be found in the public workspace (https://app.terra.bio/#workspaces/help-gatk/Germline-SNPs-Indels-GATK4-hg38).

Germline SNP and short indel discovery from WES (pipeline from Washington University in St Louis, WashU)
Germline Variant Calling was performed using germlinewrapper pipeline v.1.1 (https://github.com/ding-lab/germlinewrapper), which implements multiple tools for the detection of germline insertion/deletion (INDELs) variants and single nucleotide variants (SNVs): VarScan (v.2.3.8 (ref [20]), with default parameters, except where: --min-var-freq 0.10 --p-value 0.10, --min-coverage 3 --strand-filter 1) operating on a mpileup stream produced by samtools (v.1.2 with default parameters, except where -q 1 -Q 13); GATK (v.4 (ref [51]) using its haplotype caller in single-sample mode with duplicate and unmapped reads removed and retaining calls with a minimum quality threshold of 10); and Pindel (v. 0.2.5b9 (ref [14]), with default parameters except -m 6, -w 1). We retained germline SNVs called by either GATK or VarScan calls, while we required that indels were called by either pindel or at least two out of the three variant callers. All resulting variants were filtered to have allelic depth ≥ 5 for the alternative allele. We also filtered all INDELs longer than 100 bp.

Structural Variant Calling (pipeline from Washington University in St Louis, WashU)
Structural variants (SVs) were called by Manta v1.6.0 (ref [12]) from WGS tumor and normal paired BAMs. We ran Manta on canonical chromosomes with the default record- and sample-level filters., retaining variants where sample site depth is less than 3x the median chromosome depth near one or both variant breakends, the somatic score is greater than 30, and for small variants (<1000 bases) in the normal sample, the fraction of reads with MAPQ0 around either breakend does not exceed 0.4. It is optimized for the analysis of somatic variation in tumor/normal sample pairs. The paired and split-read evidence were combined during the SV discovery and scoring to improve accuracy. We suggested prioritizing the variants by the number of spanning read pairs that strongly (Q30) support the variants as described by [37].

DNA Methylation Microarray Processing (pipeline from Washington University in St Louis, WashU)
Raw methylation idat files were downloaded from CPTAC DCC and GDC. Beta values of CpG loci were reported after functional normalization, quality check, common SNP filtering, and probe annotation using Li Ding Lab's methylation pipeline v1.1 https://github.com/ding-lab/cptac_methylation.

Gene Level Methylation Data derivation (pipeline from Icahn School of Medicine at Mount Sinai)
Pipeline Advantage/Differentiator: this is one of two pipelines used by CPTAC to generate DNA methylation quantitative data tables. The primary difference between this pipeline and the others is that this pipeline generates gene-level, instead of probe-level, DNA methylation summaries, and it uses the promotor and 5UTR regions, instead of the coding regions, of the genes, which is associated with down-regulation (silencing) of the expression of the genes.

Gene-level DNA methylation data were generated using the beta values of all probes harboring in the islands of promoter and 5 UTR regions of the genes. Samples with missing rate > 10% across all probes were excluded, the mean and median levels were calculated, and the analyses were performed on the discovery cohort of 7 cancer types separately. For 5 cancer types (ccRCC, LUAD, LSCC, HNSCC and PDA) that have DNA methylation data in both NAT and tumor samples, consensus clustering analysis was performed as quality control procedure to evaluate the separation of clusters on NAT vs tumor samples. The misclassified samples were checked for clustering uncertainty and sample labeling.

Ancestry Prediction (pipeline from Washington University in St Louis, WashU)
We used a reference panel of genotypes and clustering based on principal components to identify likely ancestry. We selected 107,765 coding SNPs with a minor allele frequency > 0.02 from the 1000 Genomes Project (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). From this set of loci, we measured the depth and allele counts of each sample using the tool bam-readcount (https://github.com/genome/bam-readcount, version 0.8.0). Genotypes were then called for each sample based on the following criteria: 0/0 if reference count ≥ 8 and alternate count < 4; 0/1 if reference count ≥ 4 and alternate count ≥ 4; 1/1 if reference count < 4 and alternate count ≥ 8; and ./. (missing) otherwise. After filtering markers with missingness > 5%, ~70k markers remain depending on cohort size and makeup. We performed principal component analysis (PCA) on the 1000 Genomes samples to identify the top 20 principal components. We then projected our cohort onto the 20-dimensional space representing the 1000 Genomes data. We then trained a random forest classifier with the 1000 Genomes dataset using these 20 principal components. The 1000 Genomes dataset was split 80/20 for training and validation respectively. On the validation dataset, our classifier achieved 99.2-99.6% accuracy. We then used the fitted classifier to predict the likely ancestry. Each prediction file contains classification probabilities for each sample for five ancestries (EUS - European, EAS - East Asian, SAS - South Asian, AFR - African, AMR - American Admixture) Github: https://github.com/ding-lab/ancestry

MSI Prediction (pipeline from Washington University in St Louis, WashU)

MSI scores were calculated by MSIsensor (https://github.com/ding-lab/msisensor) and interpreted as the percentage of microsatellite sites (with deep enough sequencing coverage) that have a lesion. Samples with an MSIscore > 3.5 are classified as "MSI-High" and the rest will be classified as "MSS." An intermediate class with 1.0 <= score <= 3.5 can be defined as "MSI-Low."

HLA Typing (pipeline from Washington University in St Louis, WashU)
The wild-type protein sequences are obtained from Ensembl database. We constructed different epitope lengths (8-11-mer) from the translated protein sequence. Each sample's HLA type comes from OptiType prediction. We predicted the binding affinity between epitopes and the major histocompatibility complex (MHC) using NetMHC4 (ref [24]). Epitopes with binding affinity ≤ 500nM which are also not present in the wild-type transcript are reported as neoantigens.

WGS Germline Variant Calling (pipeline from University of Michigan and Icahn School of Medicine at Mount Sinai)
We performed germline variant calling using DNAScope [25], implementing a pipeline based on the GATK best-practices and functional equivalence recommendations. Briefly, we first aligned the raw paired-end WGS FASTQ files from 779 CPTAC3 blood derived samples to the latest human genome build GRCh38 (GDC GRCh38.d1.vd1 version) using bwa-mem [52], and performed duplicate marking conformant to Picard specification (compatible with [53]). Next, we called variants using the DNAScope Haplotyper with `--emit_mode gvcf` using default settings, producing one gVCF file per-sample. Next, we genotyped the samples to a set of high quality variants from 2,504 unrelated samples from Phase 3 of the 1000 Genomes Project, which were re-sequenced to high-depth by the New York Genome Center [54]. We used this reference panel for imputation because variant calling was performed directly on GRCh38 assembly with good coverage (~30x), using a compatible GATK-based workflow. This process of extracting genotypes at specified positions is implemented in lopass-genotype.py in the lopass toolkit [https://github.com/mctp/lopass], optimized for the relatively low sequencing depth of these samples (15X-30X). Next, we performed genotype imputation and phasing using GLIMPSE [55], using default settings.

Population Stratification Using Germline WGS Data
To identify the ancestry of the CPTAC3 participants, we performed principal component analysis (PCA) with the 1000 Genomes reference dataset [54]. For this analysis, we removed indels and rare variants (defined by <5% of minor allele frequency, MAF). For the remaining variants with a call rate of at least 0.99, we performed linkage disequilibrium (LD) pruning and PCA with smartpca using the EIGENSOFT software [26]. Finally, we inferred the ancestry of each CPTAC3 participant by visualizing the PCA plot and selecting cutoffs on PCs 1 through 10 corresponding to the five major populations in the 1000 Genomes data (Ad-Mixed American, African, East Asian, European and South Asian).

Transcriptomics

RNA Quantification (pipeline from Washington University in St Louis, WashU)
We obtained the gene-level read count, Fragments Per Kilobase of transcript per Million mapped reads (FPKM), and FPKM Upper Quartile (FPKM-UQ) values by following the GDC's RNA-Seq pipeline (Expression mRNA Pipeline) https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/, except running the quantification tools in the stranded mode. We used HTSeq v0.11.2 (ref [10]) to calculate the gene-level stranded read count (parameters: -r pos -f bam -a 10 -s reverse -t exon -i gene_id -m intersection-nonempty --nonunique=none) using GENCODE v22 (Ensembl v79) annotation downloaded from GDC (gencode.gene.info.v22.tsv). The read count was then converted to FPKM and FPKM-UQ using the same formula described in GDC's Expression mRNA Pipeline documentation.

RNA Quantification (pipeline from Broad institute of MIT and Harvard)
Pipeline Advantage/Differentiator: this pipeline generates multiple levels of expression quantification, from read counts to isoform proportions, as well as extensive QC metrics to facilitate harmonization across different cohorts (e.g., data generated using polyA selection and total RNA protocols from CPTAC 2 and 3, respectively). The pipeline uses methods and parameters that were extensively benchmarked for GTEx [56].

We re-processed all cohorts using the GTEx/TOPMed pipeline described at https://github.com/broadinstitute/gtex-pipeline/blob/master/TOPMed_RNAseq_pipeline.md (ref [56]). All RNA-seq samples were re-aligned to GRCh38 with the GENCODE V34 gene annotation using STAR v2.7.5a. Optical and PCR duplicates were identified with Picard 2.18.17 MarkDuplicates. Quality control and gene-level quantification (in Transcripts per Million [TPM] units) were performed with RNA-SeQC 2.3.6 (ref [27]), and isoform quantification was done with RSEM [57].

RNA quantification and circular RNA prediction (Pipeline from Baylor College of Medicine, BCM)
Pipeline Advantage/Differentiator: this is one of three pipelines used by CPTAC to generate RNA-seq based transcriptomics quantitative data tables. The primary difference between this pipeline and the others is the separation of linear RNA and circular RNA from the total RNAseq data.
RNAseq fastq files were downloaded from GDC and analyzed by the BCM total RNAseq data analysis pipeline using the Hg38 reference genome and GENCODE V34 basic (CHR) annotation. First, CIRI (v2.0.6) was used to call circular RNA with default parameters, and BWA (version 0.7.17-r1188) was used as the mapping tool. The cutoff of supporting reads for circRNA was set to 10. Then we used a pseudo-linear transcript strategy to quantify gene and circular RNA expression. In brief, for each sample, linear transcripts of circular RNAs were extracted and 75bp (read length) from the 3' end was copied to the 5' end. The modified transcripts were called pseudo-linear transcripts. Transcripts of linear genes were also extracted and mixed with pseudo-linear transcripts. RSEM (version 1.3.1) with Bowtie2 (version 2.3.3) as the mapping tool was used to quantify gene and circular RNA expression based on the mixed

transcripts. After quantification, the upper quantiles of coding genes were normalized to 1500 for cross cancer type normalization. The normalized matrix was log2-transformed and separated into gene and circular RNA expression matrices.

miRNA Quantification (pipeline from Washington University in St Louis, WashU)
miRNA-Seq FASTQ files were downloaded from GDC. We reported the mature miRNA and precursor miRNA expression in TPM (Transcripts Per Million) after adapter trimming, quality check, alignment, annotation, reads counting using Li Ding Lab's miRNA pipeline https://github.com/ding-lab/CPTAC_miRNA. The mature miRNA expression was calculated irrespective of its gene of origin by summing the expression from its precursor miRNAs.

RNA Fusion Detection (pipeline from Washington University in St Louis, WashU)
We used three callers, STAR-Fusion v1.5.0 (ref [17]), INTEGRATE v0.2.6 (ref [11]), and EricScript v0.5.5 (ref [9]), to call consensus fusion/chimeric events in our samples. Calls by each tool using tumor and normal RNA-Seq data were then merged into a single file and extensive filtering was done. As STAR-Fusion has higher sensitivity, calls made by this tool with higher supporting evidence (defined by fusion fragments per million total reads, or FFPM > 0.1) were required, or a given fusion must be reported by at least 2 callers. We then removed fusions present in our panel of blacklisted or normal fusions, which included uncharacterized genes, immunoglobulin genes, mitochondrial genes, and others, as well as fusions from the same gene or paralog genes and fusions reported in TCGA normal samples [58], GTEx tissues (reported in STAR-Fusion output), and non-cancer cell studies [59]. Finally, we removed normal fusions from the tumor fusions to curate the final set.

Cell Type Enrichment Deconvolution Using Gene Expression (pipeline from Washington University in St Louis, WashU)
The abundance of each cell type was inferred by the xCell web tool [21], which performed the cell type enrichment analysis from gene expression data for 64 immune and stromal cell types (default xCell signature). xCell is a gene signatures-based method learned from thousands of pure cell types from various sources. We used the FPKM-UQ expression matrix as the input of xCell. xCell generated an immune score per sample that integrates the enrichment scores B cells, CD4+ T-cells, CD8+ T-cells, DC, eosinophils, macrophages, monocytes, mast cells, neutrophils, and NK cells; a micro-environment score which was the sum of the immune score and stroma score. Besides, we applied CIBERSORTx [22] to compute immune cell fractions from bulk gene expression data.

Proteomics

Global proteomics and phosphoproteomics (pipeline from the University of Michigan)
Mass spectrometry data from each individual cohort was downloaded from the DCC website. The data sets were annotated following their metadata information. MzML files were searched using the MSFragger search engine version 3.4 (ref [28]) against a GENCODE34 protein FASTA database appended with an equal number of decoy sequences. The enzyme was set to

stricttrypsin (*cutafter* is KR, *butnotafter* was left blank). The post-processing was done using the Philosopher toolkit version v4.0.1 (ref [29]), and the statistical summarization, and reporting was done using TMT-Integrator [30]. Below are the individual search settings for both the proteome, and phosphoproteome analysis, including the necessary setting to account for individual experimental settings.

General settings for global proteomics. Isotope error was set to (−1/0/1/2/3) for all searches. Cysteine carbamidomethylation (+57.0215) was specified as fixed modification. Methionine oxidation (+15.9949), and protein N-terminal acetylation (+42.01060) were specified as variable modifications. The search was restricted to tryptic peptides, allowing up to two missed cleavage sites. The minimum number of peaks used was set to 15, and the maximum set to 300. The precursor ion tolerance, and the fragment tolerance was set to 20 ppm. Serine TMT labeling (+229.1629), and peptide N-terminal TMT labeling (+229.1629) were specified as fixed modifications. The clear mz range was set to 113.5 - 117.5.

General settings for the Phosphopeptide-enriched samples. The isotope error was set to (0/1/2/3). Cysteine carbamidomethylation (+57.0215), Serine TMT labeling (+229.1629), and peptide N-terminal TMT labeling (+229.1629) were specified as fixed modification. Methionine oxidation (+15.9949), and serine, threonine, tyrosine phosphorylation (+79.9663) were specified as variable modifications. The precursor ion tolerance, and the fragment tolerance was set to 20 ppm. The minimum number of peaks used was set to 15, and the maximum set to 150. PTMProphet was executed by setting static to true, the fragment ppm tolerance set to 15, the PeptideProphet probability threshold set to 0.5, the onions set to b, and the mod list set with STY:79.966331,M:15.9949.

Post-processing. MSFragger output files were processed using PeptideProphet [60] to compute the posterior probability of correct identification for each peptide to spectrum match (PSM). In the phosphopeptide-enriched dataset, PeptideProphet files were additionally processed using PTMProphet [61] to localize the phosphorylation sites. Each data set, from every cohort was then filtered by Philosopher using the sequential method by combining all pep.xml files. Peptides were assigned either as a unique or razor to a single protein with the most peptide evidence using the razor approach. PSMs, peptides, and proteins were filtered to 1% False Discovery Rate (using the best peptide approach for proteins) and applied the picked FDR target-decoy strategy. For each approved PSM, the corresponding precursor ion MS1 intensity was extracted using the Philosopher label-free quantification method, using 10 p.p.m mass tolerance and 0.4 min retention time window for extracted ion chromatogram peak tracing. The quantification of the isobaric tags was done using the Philosopher isobaric quantification method, using a purity threshold of 0.5, a minimum PeptideProphet probability of 0.7, and an m/z tolerance of 20. The bottom 5% PSMs were removed from the final list. The combined protXML file and the individual PSM lists for each data set were further processed using the Philosopher filter command as follows. To generate summary reports on different levels (gene, peptide, and protein for global and phosphopeptide enriched data; additional modification site report for phosphopeptide data), all PSM files were processed together using TMT-Integrator. Each PSM that passed the following criteria were kept for creating integrated reports, including (1) having a TMT label at peptide N-terminus, (2) having non-zero intensity in the reference channel, (3) precursor-ion purity above 50%, (4) summed reported ion intensity (across all channels) not in the lower 5% of all PSMs (2.5% for phosphopeptide enriched data), (5) fully tryptic peptides, and (6) peptide with phosphorylation (for phosphopeptide enriched data). For a peptide with redundant PSMs in the same MS run, only the PSM with the highest summed TMT intensity was kept for later analysis. PSMs mapping to common external contaminant proteins were excluded, and both unique and razor peptides were used for quantification. Next, the reporter ion intensities of each PSM were log2 transformed and normalized by the reference channel intensity (i.e., subtracted log2 reference intensity from those log2 report ion intensities), therefore the intensities were converted into a log2-based ratio (denoted as 'ratios' in the following paragraphs). After converting the intensities to ratios, the PSMs were grouped based on the predefined level (i.e., gene, protein, peptide, and site-level). The interquartile range (IQR) algorithm was then applied to remove the outliers in each PSM group, and the remaining ratios were median centered. The ratios were converted back to abundances using the weighted sum of the MS1 intensities of the top three most intense peptide ions, with the weighting factor (computed for each PSM) taken as the ratio of the reference channel intensity to the summed reporter ion intensity (across all channels). In generating the site-level reports (phosphopeptide-enriched data), sites with PTMProphet computed localization probability equal or greater than 0.75 were considered as confidently localized. Additional details regarding these steps can be found in [37].

Abundance table preprocessing. First, based on the median-aligned intensity tables, we identified outlier TMT multi-plex data points using *Intra TMT-plex T tests*. Specifically, for a given protein (or phosphosite), we compared its abundances in samples from one TMT-plex and that from the rest of the TMT-plexes in the experiment using t-tests. If the measurements of this protein (phosphosite) in one TMT-plex was influenced by artificial/technique factors (e.g. false peptide/protein identification in that TMT experiment output), the corresponding p-value of the T test would be significant. After *Intra TMT-multiplex T tests* were performed for all proteins (phospshosites) across all TMT-plexes of one cancer type experiment, we then applied double-log-transformation to the p-values and identified those falling beyond 4-standard-deviation below the median of all transformed p-values of the cancer type. The corresponding abundance measurements of the outlier protein-TMTplex set were removed (replaced with NA). The numbers of outlier data points removed in each tumor data set were summarized in **Table 1 - ProteinPreprocessing**.

After the outlier removal, for each cancer type, we assessed the batch-effects of TMT-plexes through careful investigation of PC plots for the proteomic (phosphoproteomic) abundance matrices. For data sets with non-ignorable TMT-plex batch effects (**Table 1 - ProteinPreprocessing**), we then applied *Combat* [31]**,** to remove the technical variation across TMT multi-plexes. For some tumor types, such as CCRCC, the TMT-plex design was not balanced in the TMT profiling experiment: i.e. the tumor and normal sample size ratios within each TMT-plex varied in the experiment. We then further adjusted tumor/normal tissue types when performing Combat correction. Additionally, since complete data matrices are needed as input for ComBat, we first performed KNN imputation on the data sets using the 'impute' R package and then we replaced the missing data point back with “NA” in the batch corrected data matrices.

To formally impute missing values, we applied DreamAI (ref [32]; https://github.com/ WangLab-MSSM/DreamAI) on each of the tumor types separately. Imputation was done for the subset of proteins or phosphosites that appeared in at least 50% of samples in each data set.

Table 1 - ProteinPreprocessing

Cancer type	CPTAC2 or CPTAC3	Tumor	Normal	Proteome		Phosphosite	
				outlier TMT	Batch correction (adjustment)	outlier TMT	Batch correction (adjustment)
Breast (BRCA)	II	135		10	None	17	None
Kidney (ccRCC)	III	110	84	72	Combat (T/N)	147	None
Colon (COAD)	II	97	100	15	None	20	None
Brain (GBM)	III	100	10	11	None	1	None
Head & Neck (HNSCC)	III	114	75	21	None	29	None
Lung (LSCC)	III	110	102	8	None	14	None

Lung (LUAD)	III	113	102	22	None	11	None
Ovaries (OV)	II	84	19	20	None	2	Combat
Pancreas (PDAC)	III	145	86	36	None	11	None
Uterine (UCEC)	III	100	49	25	None	3	None

Glycoproteomics (pipeline from Johns Hopkins University)
Mass spectrometry data from each individual cohort was downloaded from the DCC website. The data sets were annotated following their metadata information. MzML files were searched using the GPQuest search engine version 2.1 (refs [62,63]) against a reported glycopeptide database reported in Glycositeatlas [34] appended with an equal number of decoy sequences and an N-linked glycan database, which was collected from the public database of GlycomeDB (http://www.glycome-db.org). The post-processing, statistical summarization, and reporting were done using MS-PyCloud [64]. The consensus spectral library for each glycopeptide was constructed by using SpectraST and data visualization was done by OmicsOne [36]. Below are the individual search settings for glycoproteome analysis, including the necessary setting to account for individual experimental settings.

Searching N-linked glycopeptides using GPQuest. Prior to the database search, ProteoWizard 3.0 was used to convert the .RAW files to .mzML files with the “centroid all scans” option selected. The MS/MS spectra containing the oxonium ions (m/z 204.0966) in the top 10 abundant peaks after removing TMT reporter ions were considered as the potential glycopeptide candidates. Isotope error was set to (−1/0/1/2) for all searches. Cysteine carbamidomethylation (+57.0215) was specified as fixed modification. Methionine oxidation (+15.9949), and protein N-terminal acetylation (+42.01060) were specified as variable modifications. The minimum number of peaks used was set to 15, and the maximum set to 100. The precursor ion tolerance was set to 10 ppm and the fragment tolerance was set to 20 ppm. Lysine TMT labeling (+229.1629), and peptide N-terminal TMT labeling (+229.1629) were specified as fixed modifications. The best hits of all glycopeptide-spectrum matchings (GPSMs) were ranked by the Morpheus scores [65] in descending order, in which those with FDR < 1% and covering > 10% total intensity of each tandem spectrum were reserved as qualified identifications.

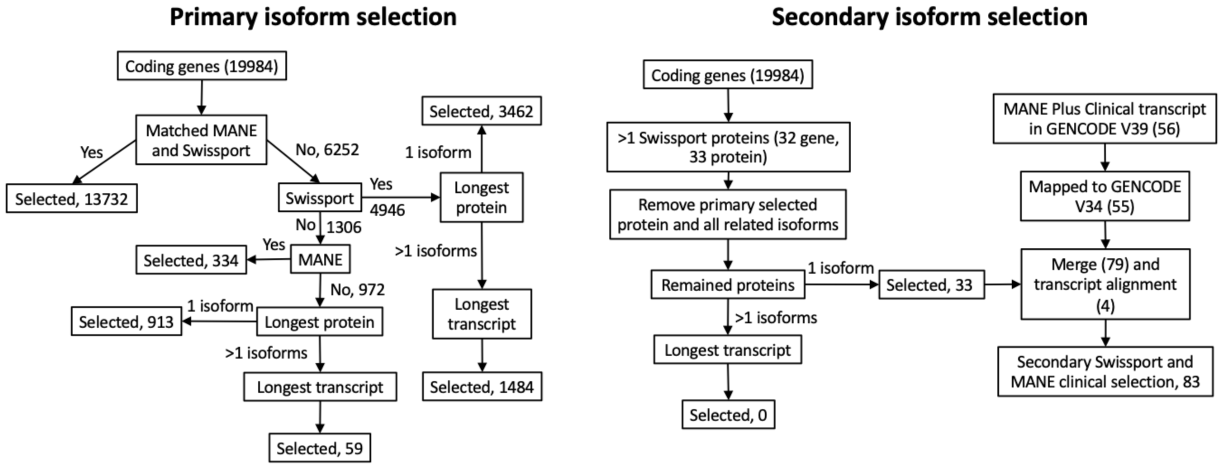
Post-processing using MS-PyCloud. Glycopeptide-spectrum matchings (GPSMs) were filtered based on a user-defined PSM-level false discovery rate (FDR) cutoff and significant GPSMs from all sets of each cohort were grouped to infer the represented proteins parsimoniously using a bipartite graph analysis algorithm adopted in many protein inference tools [66,67]. The final FDRs are then estimated at N-glycopeptide-levels using the reversed decoy search. For isobaric labeled data, the TMT reporter ion intensities are extracted from the mzML file for MS2 scans corresponding to the identified glycopeptides. For TMT data, the TMT reagent lot correction factors were used to adjust the reported TMT intensities for interference between TMT channels. Log2 ratios are calculated at GPSM-level relative to the user-specified reference channel and are then rolled up by the median value to N-glycopeptide-level (intact glycopeptide enriched). Normalized log2 ratio matrices are generated using median normalization (MD norm), and median normalization plus median absolute deviation scaling (MD norm + MAD scaling). Absolute abundances are generated from the log2 ratio matrices by summing the log2 ratios with the median log2 value of the reference channel summed MS2 intensities across all sets for each N-glycopeptide using the same approach used in the proteome and phosphoproteome processing. Additional details regarding these steps can be found in [33,37,62].

BCM pipeline for pan-cancer multi-omics data harmonization

Gene annotation database and representative isoform selection

Using the same versions of genome assembly and gene annotation for the processing of data from all omics platforms and all cancer types is critical for streamlined and accurate downstream pan-cancer multi-omics data integration. The Homo sapiens (human) genome assembly GRCh38 (hg38) and the GENCODE V34 basic (CHR) gene annotation, which were the most updated at the time of data freeze, were selected for the CPTAC pan-cancer data analysis. This annotation includes 19,984 protein coding genes and 40,685 noncoding genes.

Some analyses, such as somatic mutation annotation, gene level methylation quantification, phosphosite location annotation, and sequence-based multi-omics data visualization and integration, further benefit from the selection of a representative transcript/protein isoform for each gene locus to ensure standardized reporting and efficient data integration. Swiss-Prot and MANE Select are two major efforts that aim to select one high-quality representative transcript for each protein-coding gene. Leveraging these two resources, we developed a workflow (Supplementary Figure 1) and selected one representative isoform for each protein coding gene (Table S3). Moreover, 32 genes are associated with multiple Swiss-Prot proteins, and all non-representative proteins and their longest transcripts were designated as secondary isoforms for these genes. The MANE Plus Clinical isoforms were added as supplements of MANE Selects when needed for clinical variant reporting. If MANE Plus Clinical and secondary selects are with the same proteins and different transcripts for a gene, the MANE Plus Clinical selected isoform will be used as the secondary selected isoform. For noncoding genes, the longest isoforms were selected as their representative isoforms.



Supplementary Figure 1. Workflow for the selection of representative isoforms for each protein coding gene.

Pan-cancer multi-omics data harmonization workflow

Standardized data processing was applied to all cancer types using the same versions of genome assembly, gene annotation, and representative isoform selection. Somatic mutations reported by the combined Broad and WashU pipelines in MAF format were converted to VCF format and reannotated using ANNOVAR (v2019Oct24) based on the representative isoforms. Segment level copy number variations (CNVs) were called by CopywriteR (v2.20.0) using matched tumor and normal WXS data. Gistic2 was used to generate gene and focal level CNV results based on GENCODE V34 basic (CHR) annotation. Probes of the methylation EPIC array were reannotated using the representative isoforms. Probe level meta values were downloaded from GDC. Gene and isoform level meta values were defined as the median of meta values of probes located in their 5' UTR and promoter regions. RNAseq fastq files were downloaded from GDC and analyzed by the BCM total RNAseq data analysis pipeline using GENCODE V34 basic (CHR) annotation. Arriba (v2.0.0) was used to call gene fusion using RNAseq data and GENCODE V34 basic (CHR) annotation. Gene and phosphosite intensities reported by the Michigan pipeline from the analysis of global and phosphoproteomics data were normalized across cancer types by median centering of the medians of reference intensities of each cancer type. Phosphosite reannotation was performed to ensure consistent interpretation of the data across TMTs and cancer types. Specifically, phosphosites were mapped to the primary and secondary isoforms, and those that could not be mapped exactly to the selected isoforms were discarded. For sites with peptides that matched more than one location on a single protein sequence, the first matching position was selected. Based on the updated site ID, rows with a duplicated ID were discarded after ordering by decreasing number of missing values. The site ID consists Ensembl gene ID, Ensembl protein isoform ID, site position based on the protein isoform sequence, fifteen-mer (+/- 7 amino acids) based on the protein isoform sequence, and a flag for whether the protein is a primary (1) or secondary (2) selected sequence.

To facilitate human understanding of the Ensembl gene IDs, each ID was assigned a unique gene name for display. First, all primary selected isoforms were ordered by presence of a SwissProt ID that mapped to the Ensembl protein ID, transcript ID listed in the MANE plus Clinical annotation, longer CCDS length, longer transcript length, and finally, alphabetic order of the Ensembl gene ID. The first Ensembl gene ID was assigned the gene symbol (e.g., AHRR for ENSG00000286169.1) and all following had the Ensembl gene ID appended (e.g., AHRR_ENSG00000063438 for ENSG00000063438).

References

1. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. 10.1093/bioinformatics/btp324.
2. Scott, A.D., Huang, K.-L., Weerasinghe, A., Mashl, R.J., Gao, Q., Martins Rodrigues, F., Wyczalkowski, M.A., and Ding, L. (2019). CharGer: clinical Characterization of Germline variants. Bioinformatics *35*, 865–867. 10.1093/bioinformatics/bty649.
3. Taylor-Weiner, A., Stewart, C., Giordano, T., Miller, M., Rosenberg, M., Macbeth, A., Lennon, N., Rheinbay, E., Landau, D.-A., Wu, C.J., et al. (2018). DeTiN: overcoming tumor-in-normal contamination. Nat Methods *15*, 531–534. 10.1038/s41592-018-0036-9.
4. Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol *12*, R41. 10.1186/gb-2011-12-4-r41.
5. Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218. 10.1038/nature12213.
6. Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573. 10.1093/bioinformatics/btq170.
7. Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al. (2020). The repertoire of mutational signatures in human cancer. Nature *578*, 94–101. 10.1038/s41586-020-1943-3.
8. Sun, S.Q., Mashl, R.J., Sengupta, S., Scott, A.D., Wang, W., Batra, P., Wang, L.-B., Wyczalkowski, M.A., and Ding, L. (2018). Database of evidence for precision oncology portal. Bioinformatics *34*, 4315–4317. 10.1093/bioinformatics/bty531.
9. Benelli, M., Pescucci, C., Marseglia, G., Severgnini, M., Torricelli, F., and Magi, A. (2012). Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. Bioinformatics *28*, 3232–3239. 10.1093/bioinformatics/bts617.
10. Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169. 10.1093/bioinformatics/btu638.
11. Zhang, J., White, N.M., Schmidt, H.K., Fulton, R.S., Tomlinson, C., Warren, W.C., Wilson, R.K., and Maher, C.A. (2016). INTEGRATE: gene fusion discovery using whole genome and transcriptome data. Genome Res *26*, 108–118. 10.1101/gr.186114.114.
12. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222. 10.1093/bioinformatics/btv710.
13. Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol *31*, 213–219. 10.1038/nbt.2514.
14. Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865–2871. 10.1093/bioinformatics/btp394.
15. Polpitiya, A.D., Qian, W.-J., Jaitly, N., Petyuk, V.A., Adkins, J.N., Camp, D.G., Anderson, G.A., and Smith, R.D. (2008). DAnTE: a statistical tool for quantitative analysis of -omics data. Bioinformatics *24*, 1556–1558. 10.1093/bioinformatics/btn217.
16. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis,

G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. 10.1093/bioinformatics/btp352.

17. Haas, B.J., Dobin, A., Li, B., Stransky, N., Pochet, N., and Regev, A. (2019). Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol *20*, 213. 10.1186/s13059-019-1842-9.
18. Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., et al. (2018). Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods *15*, 591–594. 10.1038/s41592-018-0051-x.
19. Conway, J.R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics *33*, 2938–2940. 10.1093/bioinformatics/btx364.
20. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res *22*, 568–576. 10.1101/gr.129684.111.
21. Aran, D., Hu, Z., and Butte, A.J. (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol *18*, 220. 10.1186/s13059-017-1349-1.
22. Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol *37*, 773–782. 10.1038/s41587-019-0114-2.
23. Xi, R., Lee, S., Xia, Y., Kim, T.-M., and Park, P.J. (2016). Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. Nucleic Acids Res *44*, 6274–6286. 10.1093/nar/gkw491.
24. Andreatta, M., and Nielsen, M. (2016). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics *32*, 511–517. 10.1093/bioinformatics/btv639.
25. Freed, D., Aldana, R., Weber, J.A., and Edwards, J.S. (2017). The Sentieon Genomics Tools - A fast and accurate solution to variant calling from next-generation sequence data (Bioinformatics) 10.1101/115717.
26. Patterson, N., Price, A.L., and Reich, D. (2006). Population Structure and Eigenanalysis. PLoS Genet *2*, e190. 10.1371/journal.pgen.0020190.
27. Graubert, A., Aguet, F., Ravi, A., Ardlie, K.G., and Getz, G. (2021). RNA-SeQC 2: Efficient RNA-seq quality control and quantification for large cohorts. Bioinformatics, btab135. 10.1093/bioinformatics/btab135.
28. Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D., and Nesvizhskii, A.I. (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods *14*, 513–520. 10.1038/nmeth.4256.
29. da Veiga Leprevost, F., Haynes, S.E., Avtonomov, D.M., Chang, H.-Y., Shanmugam, A.K., Mellacheruvu, D., Kong, A.T., and Nesvizhskii, A.I. (2020). Philosopher: a versatile toolkit for shotgun proteomics data analysis. Nat Methods *17*, 869–870. 10.1038/s41592-020-0912-y.
30. Djomehri, S.I., Gonzalez, M.E., da Veiga Leprevost, F., Tekula, S.R., Chang, H.-Y., White, M.J., Cimino-Mathews, A., Burman, B., Basrur, V., Argani, P., et al. (2020). Quantitative proteomic landscape of metaplastic breast carcinoma pathological subtypes and their relationship to triple-negative tumors. Nat Commun *11*, 1723. 10.1038/s41467-020-15283-z.
31. Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127. 10.1093/biostatistics/kxj037.
32. Ma, W., Kim, S., Chowdhury, S., Li, Z., Yang, M., Yoo, S., Petralia, F., Jacobsen, J., Li, J.J., Ge, X., et al. (2020). DreamAI: algorithm for the imputation of proteomics data

(Bioinformatics) 10.1101/2020.07.21.214205.

33. Hu, Y., Pan, J., Shah, P., Ao, M., Thomas, S.N., Liu, Y., Chen, L., Schnaubelt, M., Clark, D.J., Rodriguez, H., et al. (2020). Integrated Proteomic and Glycoproteomic Characterization of Human High-Grade Serous Ovarian Carcinoma. Cell Rep *33*, 108276. 10.1016/j.celrep.2020.108276.
34. Sun, S., Hu, Y., Ao, M., Shah, P., Chen, J., Yang, W., Jia, X., Tian, Y., Thomas, S., and Zhang, H. (2019). N-GlycositeAtlas: a database resource for mass spectrometry-based human N-linked glycoprotein and glycosylation site mapping. Clin Proteomics *16*, 35. 10.1186/s12014-019-9254-0.
35. Ranzinger, R., Herget, S., von der Lieth, C.-W., and Frank, M. (2011). GlycomeDB--a unified database for carbohydrate structures. Nucleic Acids Res *39*, D373-376. 10.1093/nar/gkq1014.
36. Zhang, H., Ao, M., Boja, A., Schnaubelt, M., and Hu, Y. (2021). OmicsOne: associate omics data with phenotypes in one-click. Clin Proteomics *18*, 29. 10.1186/s12014-021-09334-w.
37. 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. (2020). Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. Cell *180*, 207. 10.1016/j.cell.2019.12.026.
38. 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. 10.1016/j.cell.2020.10.036.
39. 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. 10.1016/j.cell.2019.03.030.
40. 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. 10.1016/j.ccell.2021.01.006.
41. 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. 10.1016/j.ccell.2020.12.007.
42. 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. 10.1016/j.cell.2021.07.016.
43. 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. 10.1016/j.cell.2020.06.013.
44. 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. 10.1016/j.xcrm.2020.100004.
45. 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. 10.1016/j.cell.2021.08.023.
46. 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. 10.1016/j.cell.2020.01.026.

47. 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 *173*, 371-385.e18. 10.1016/j.cell.2018.02.060.
48. Haradhvala, N.J., Polak, P., Stojanov, P., Covington, K.R., Shinbrot, E., Hess, J.M., Rheinbay, E., Kim, J., Maruvka, Y.E., Braunstein, L.Z., et al. (2016). Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell *164*, 538–549. 10.1016/j.cell.2015.12.050.
49. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer *18*, 696–705. 10.1038/s41568-018-0060-1.
50. Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol *36*, 983–987. 10.1038/nbt.4235.
51. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297–1303. 10.1101/gr.107524.110.
52. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 10.48550/ARXIV.1303.3997.
53. Regier, A.A., Farjoun, Y., Larson, D.E., Krasheninina, O., Kang, H.M., Howrigan, D.P., Chen, B.-J., Kher, M., Banks, E., Ames, D.C., et al. (2018). Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat Commun *9*, 4038. 10.1038/s41467-018-06159-4.
54. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2021). High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios (Genomics) 10.1101/2021.02.06.430068.
55. Rubinacci, S., Ribeiro, D.M., Hofmeister, R.J., and Delaneau, O. (2021). Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat Genet *53*, 120–126. 10.1038/s41588-020-00756-0.
56. The GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. 10.1126/science.aaz1776.
57. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323. 10.1186/1471-2105-12-323.
58. 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 Reports *23*, 227-238.e3. 10.1016/j.celrep.2018.03.050.
59. Babiceanu, M., Qin, F., Xie, Z., Jia, Y., Lopez, K., Janus, N., Facemire, L., Kumar, S., Pang, Y., Qi, Y., et al. (2016). Recurrent chimeric fusion RNAs in non-cancer tissues and cells. Nucleic Acids Res *44*, 2859–2872. 10.1093/nar/gkw032.
60. Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem *74*, 5383–5392. 10.1021/ac025747h.
61. Shteynberg, D.D., Deutsch, E.W., Campbell, D.S., Hoopmann, M.R., Kusebauch, U., Lee, D., Mendoza, L., Midha, M.K., Sun, Z., Whetton, A.D., et al. (2019). PTMProphet: Fast and Accurate Mass Modification Localization for the Trans-Proteomic Pipeline. J Proteome Res *18*, 4262–4272. 10.1021/acs.jproteome.9b00205.
62. Hu, Y., Shah, P., Clark, D.J., Ao, M., and Zhang, H. (2018). Reanalysis of Global Proteomic

and Phosphoproteomic Data Identified a Large Number of Glycopeptides. Anal Chem *90*, 8065–8071. 10.1021/acs.analchem.8b01137.

63. Mertins, P., Tang, L.C., Krug, K., Clark, D.J., Gritsenko, M.A., Chen, L., Clauser, K.R., Clauss, T.R., Shah, P., Gillette, M.A., et al. (2018). Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography-mass spectrometry. Nat Protoc *13*, 1632–1661. 10.1038/s41596-018-0006-9.
64. Chen, L., Zhang, B., Schnaubelt, M., Shah, P., Aiyetan, P., Chan, D., Zhang, H., and Zhang, Z. (2018). MS-PyCloud: An open-source, cloud computing-based pipeline for LC-MS/MS data analysis (Bioinformatics) 10.1101/320887.
65. Wenger, C.D., and Coon, J.J. (2013). A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. J Proteome Res *12*, 1377–1386. 10.1021/pr301024c.
66. Ma, Z.-Q., Dasari, S., Chambers, M.C., Litton, M.D., Sobecki, S.M., Zimmerman, L.J., Halvey, P.J., Schilling, B., Drake, P.M., Gibson, B.W., et al. (2009). IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. J Proteome Res *8*, 3872–3881. 10.1021/pr900360j.
67. Patro, R., and Kingsford, C. (2013). Predicting protein interactions via parsimonious network history inference. Bioinformatics *29*, i237-246. 10.1093/bioinformatics/btt224.