Proteogenomic characterization of 2002 human cancers reveals pan-cancer molecular subtypes and associated pathways

Yiqun Zhang1,8, Fengju Chen1,8, Darshan S. Chandrashekar2,3,8, Sooryanarayana Varambally2,3,4 & Chad J. Creighton1,5,6,7✉

Mass-spectrometry-based proteomic data on human tumors—combined with corresponding multi-omics data—present opportunities for systematic and pan-cancer proteogenomic analyses. Here, we assemble a compendium dataset of proteomics data of 2002 primary tumors from 14 cancer types and 17 studies. Protein expression of genes broadly correlates with corresponding mRNA levels or copy number alterations (CNAs) across tumors, but with notable exceptions. Based on unsupervised clustering, tumors separate into 11 distinct proteome-based subtypes spanning multiple tissue-based cancer types. Two subtypes are enriched for brain tumors, one subtype associating with MYC, Wnt, and Hippo pathways and high CNA burden, and another subtype associating with metabolic pathways and low CNA burden. Somatic alteration of genes in a pathway associates with higher pathway activity as inferred by proteome or transcriptome data. A substantial fraction of cancers shows high MYC pathway activity without MYC copy gain but with mutations in genes with noncanonical roles in MYC. Our proteogenomics survey reveals the interplay between genome and proteome across tumor lineages.
The central dogma of molecular biology states that DNA makes RNA, and RNA makes protein. At the same time, many regulatory mechanisms occur after mRNAs are manufactured, and steady-state transcript abundances only partially predict protein abundances\textsuperscript{12-14}. Recent technological advances in mass spectrometry have allowed for large-scale surveys of the cancer proteome. The Clinical Proteomic Tumor Analysis Consortium (CPTAC) is a National Cancer Institute initiative to accelerate the understanding of the molecular basis of cancer through the application of large-scale, mass spectrometry-based proteomics. CPTAC has carried out individual studies for a range of tissue-based cancer types\textsuperscript{15-19}, integrating proteome with genome, collectively known as proteogenomics, while efforts outside of CPTAC have similarly studied additional cancer types\textsuperscript{16-18}. In these recent studies, proteogenomic analyses carried out within specific cancer types have included protein-level tumor versus normal adjacent tissue comparisons, cataloging the functional consequences of somatic mutation and copy number alteration (CNA) on the proteome, and defining tumor molecular subtypes and associated pathways and immune cell infiltrates. CPTAC and other studies have collectively made mass spectrometry-based proteomic data on over 2000 human tumors available in the public domain, with corresponding data at the molecular levels of mRNA, copy number, and small mutation. These data present an opportunity for systematic and pan-cancer analyses of the entire cohort of tumors with proteomic data, including defining molecular subtypes and associated pathways relevant to multiple cancer types. A combined large pan-cancer cohort of tumor proteomic profiles would allow for proteogenomic analyses to identify commonalities, differences, and emergent themes across tumor lineages\textsuperscript{20}. In recent studies, we carried out pan-cancer proteomic studies to define molecular subtypes\textsuperscript{21} and protein correlates of more aggressive disease\textsuperscript{22}, each of these studies involving a more limited set of tumors and cancer types.

In this work, we comprehensively analyze proteomics data and corresponding multi-omics data on 2002 primary tumors from 14 different tissue-based cancer types. We explore protein-level versus corresponding mRNA-level associations, noting associations found in the transcriptome but not the proteome and vice versa. We define pan-cancer, proteome-based subtypes that cut across tumor lineages. Finally, we explore the interactions between the cancer proteome and somatic DNA-level alteration of cancer-associated pathways across tumors.

Results

Protein-level correlations. We assembled a compendium dataset of mass spectrometry-based proteomics data of 2002 primary tumors from 17 individual studies\textsuperscript{3-19} (Supplementary Data 1). The cancer types represented in this proteomic compendium dataset included breast (n = 230 tumors with proteomics data), colorectal (n = 187), gastric (n = 80), glioblastoma (n = 100), head and neck (n = 108), liver (n = 165), lung adenocarcinoma (n = 111), lung squamous (n = 110), ovarian (n = 269), pancreatic (n = 137), pediatric brain (n = 219), prostate (n = 76), renal (n = 110), and uterine (n = 100). For most of these tumors, corresponding multi-omics data were available for mRNA (n = 1899 out of the 2002 tumors), DNA somatic small mutation (single nucleotide variants (SNVs), and insertions/deletions, i.e., indels; n = 1698), and DNA somatic CNA (n = 1837). For the proteomic and transcriptomic compendium datasets, we normalized expression values within each cancer type, whereby neither tissue-dominant differences nor inter-laboratory batch effects would drive the downstream analyses\textsuperscript{21,23-25}. The total proteomics compendium dataset consisted of 15,439 genes with proteins measured in at least one tumor, including 10,129 genes with proteins represented in half of the tumors for at least seven cancer types profiled. The phospho-protein compendium consisted of 199,284 phospho-protein features involving 11,671 genes, 5419 phospho-proteins represented in half of the tumors for at least seven cancer types profiled. To facilitate access by biomedical researchers, we integrated the above proteomic data with the UALCAN data portal\textsuperscript{21,22,26}, allowing users to query proteins of interest for comparisons of interest (http://ualcan.path.uab.edu/).

Overall, protein expression of genes broadly correlated with the corresponding mRNA levels or CNA status across tumors, but with notable exceptions (Supplementary Data 2 and 3). Across 1899 tumors and 10,129 genes, tumors displayed a median gene-wise protein versus mRNA correlation r value of 0.40 (Fig. 1a, b), with 97.1% of genes having significant positive correlations (Pearson correlation $p < 0.01$). For specific functional categories of genes, protein-mRNA correlations tended to be higher or lower. For example, mRNA levels of genes involved\textsuperscript{27} in ribosome, oxidative phosphorylation, electron transport chain, and humoral immune response pathways tended to correlate poorly with protein expression across tumors (Fig. 1c), consistent with previous observations in individual cancer types\textsuperscript{6,12,15}. Based on 9744 genes with available data, gene-level correlations between protein and CNA were broadly positive. However, protein versus CNA correlations tended to be lower than the corresponding mRNA versus CNA correlations (Fig. 1d). For the 13 cancer types examined, median gene-wise protein versus CNA Pearson correlation r values ranged from 0.04 (prostate) to 0.21 (lung squamous) across the 9744 genes, while mRNA versus CNA correlation r values ranged from 0.12 to 0.41 ($p < 1E^{-7}$, paired t-test, comparing median gene-wise correlation by cancer type for protein versus mRNA). Taking a set of 756 genes with significant positive correlations between protein and CNA but not between mRNA and CNA ($p < 0.001$ and $p > 0.05$, respectively), these genes were enriched for many gene categories, including “protein-containing complex,” “mitochondrial,” “protein complex,” “ribosomal subunit,” “translation,” “electron transport chain,” “DNA repair,” and “regulation of cell cycle process” (Fig. 1d). For individual genes of interest (e.g., genes involving the PI3K/AKT/mTOR pathway, Fig. 1e), copy loss and indel mutations could be associated with lower protein expression.

Proteomic signatures associated with higher tumor grade could yield molecular clues as to processes underlying more aggressive tumors. To identify proteomic correlates of grade, we followed an approach previously demonstrated in a limited dataset of 558 tumors from five cancer types\textsuperscript{22}. In this present study, we found on the order of hundreds of total proteins and mRNAs differentially expressed with higher grade ($p < 0.01$, Pearson correlation) for each of ten cancer types (head and neck, liver, lung adenocarcinoma, lung squamous, ovarian, pancreas, pediatric glioma, prostate, renal, uterine), involving 1265 tumors for which grade information was available (Fig. S1a, b). Significant overlapping gene features were between the total protein signatures and the mRNA signatures (Fig. S1c). Each cancer type showed a proteomic and mRNA signature of grade distinctive from the other cancer types. At the same time, differential expression patterns involving a subset of genes were also shared across multiple cancer types (Fig. 2a), with 1936 proteins being significant ($p < 0.01$) for any two or more cancer types. Differentially expressed proteins and mRNAs were enriched for specific pathways\textsuperscript{28} (Figs. 2b and S1d), with many pathways significant for multiple cancer types. The set of top enriched pathways uncovered here overlapped highly with the pathways we uncovered previously\textsuperscript{22} based on five of the nine cancer types represented in our compendium. This finding suggests a core set of pathways associated with more aggressive...
tumors, even as the proteomic grade correlates differ by cancer type. Additional pathways in this study include alpha 6 beta 4 signaling (liver, pancreas) and oxidative damage (prostate, uterine). One example pathway of interest enriched within proteins increased with higher tumor grade, significant ($p < 0.05$, one-sided Fisher's exact test) for four out of ten cancer types—lung adenocarcinoma, ovarian, renal, and uterine—was type II interferon signaling (Fig. 2c), involving JAK–STAT-signaling and
Fig. 1 Correlations of proteomic abundances with transcriptomics, copy number alterations, and somatic mutations. **a** Heat map of gene-wise Pearson correlations of mRNA and protein expression (10,129 genes, with protein data for >6 individual cancer types) across all tumors studied (n = 1899 with available protein and mRNA data) as well by individual cancer type. Red, significant positive correlation. **b** Histogram of gene-wise Pearson correlations of mRNA and protein expression in 1899 tumors, based on 10,129 genes. **c** Annotated gene sets (by Gene Ontology, or GO) across 10,129 genes ranked by Pearson gene-wise correlation between protein and mRNA. Enrichment or anti-enrichment patterns are significant with p < 0.0001, FDR < 0.0001 by GSEA71. **d** Heat maps of gene-wise Pearson correlations of both mRNA expression and copy number alteration (CNA) and protein expression and CNA (9744 genes, out of the 10,129 with combined mRNA/CNA data for at least one cancer type), by individual cancer type. Red, significant positive correlation. **e** Out of 9744 genes, 756 were significantly positively correlated between expression and CNA for mRNA (Pearson p < 0.001) but not for protein (p > 0.05) for at least seven cancer types. Selected enriched GO terms for these genes are shown. p values by one-sided Fisher's exact test. “Expected,” estimated by probability. **f** Box plots of expression of selected proteins (PTEN, STK11, TSC1, TSC2) by somatic alteration class (“hom.”, homozygous; “het.”, heterozygous; “SNV,” Single Nucleotide Variant; “indel,” insertion/deletion). p values by t-test. Box plots represent 5% (lower whisker), 25% (lower box), 50% (median), 75% (upper box), and 95% (upper whisker). n = biologically independently tumors.

antiviral and growth-inhibitory effects. Another enriched pathway of interest—significant (p < 0.05) for liver, renal, and uterine cancer types—was oncostatin M cytokine signaling (Fig. 3d), known to play a significant role in inflammation, autoimmunity, and cancers29.

De novo proteome-based subtypes. Molecular subtypes can provide insights into the pathways and processes appearing deregulated within tumor subsets21,23. The 2002 tumors in the proteomics compendium dataset separated into 11 distinct proteome-based pan-cancer subtypes based on unsupervised clustering (Fig. S2). These subtypes, s1 through s11, each spanned multiple tissue-based cancer types (Fig. 3a). Notably, the s11 subtype was specific to brain tumors, spanning glioblastomas and pediatric brain tumors. Previously, ten proteome-based pan-cancer subtypes, referred to as k1 through k10, were identified based on 532 tumors and five tissue-based cancer types21. Within these 532 tumors, there were significant patterns of overlap between the s1–s11 subtype assignments and the previous k1–k10 subtype assignments (Fig. 3a). Specifically, the correspondence between the current versus the previous subtyping included s1 to k1 (previously associated with the proteasome complex and ubiquitin21), s2 to k2 (associated with T cells and the immune response), s3 to k4 (associated with basal-like breast cancer), s4 to k5 (associated with epithelium and oxidative phosphorylation and TCA cycle pathways), s5 to k6 (associated with tumor stroma), s6 to k7 (associated with tumor stroma and collagen VI), s8 to k9 (associated with hemoglobin complex), and s9 to k10 (associated with endoplasmic reticulum and steroid biosynthesis pathway). The s7, s10, and s11 subtypes did not strongly correspond to the previous k1–k10 subtypes. The 11 proteome-based subtypes of the present study were each characterized by widespread differential expression patterns at the levels of both total protein and phospho-protein (Fig. 3b and Supplementary Data 4). Of the 1073 proteins for which total levels best distinguished the subtypes, 225 proteins had a drug target association by DrugBank30 (Supplementary Data 5).

Previously compiled gene annotations27 and gene signatures21,23 helped characterize the proteome-based subtypes represented by the compendium dataset (Supplementary Data 5). Within the top differentially expressed proteins underscoring each proteome-based subtype, specific gene categories (by Gene Ontology, or GO, annotation) were over-represented (Fig. 3c), consistent with the correspondences noted above with the previous k1–k10 subtyping21. In addition, subtype s7 involved “axon guidance” and “frizzled binding” genes; subtype s10 involved “DNA repair” and “chromatin organization” genes; and subtype s11 involved “synapse,” “dendrite,” and “axon” genes. As applied to the proteomic data (Fig. 3d), gene signatures of immune cell types indicated the presence of T cells and higher expression of immune checkpoint pathway genes within s2 tumors. In contrast, s5 and s6 tumors showed signatures of B cells, eosinophils, and mast cells and higher expression of complement pathway genes. The above observations would reinforce the notion of different immune response pathways being active within different tumor subsets21. Within s4 and s11 tumors, proteomic signatures of metabolism indicated higher activation of pathways involving fatty acid metabolism, glycolysis and gluconeogenesis, pentose phosphate, TCA cycle, and oxidative phosphorylation, while s1, s6, and s8 tumors each showed higher signature levels of some but not all of the above pathways (Fig. 3d). On average, protein expression of matrix metalloproteinases (MMPs) was higher in s5, s6, and s8 tumors, while protein expression of collagen VI family was higher in s6 and s8 tumors (Fig. 3d). Canonical markers of neuroendocrine tumors, previously found over-expressed in ~4% of tumors in The Cancer Genome Atlas (TCGA) cohort23, were higher on average in s6 and s11 tumors, the latter comprised of all brain tumors (Fig. 3d).

We examined proteomic datasets external to our proteomic compendium dataset for evidence of the manifestation of our proteome-based pan-cancer subtypes. Using a protein-based classifier developed from the proteomics compendium dataset, we classified 7694 TCGA tumors with Reverse-Phase Protein Array (RPPA) data according to s1–s11 proteome-based pan-cancer subtypes (Fig. 4a). The RPPA platform would be antibody-based and independent of the mass spectrometry platform. We had previously classified the TCGA RPPA profiles according to k1 through k10 subtypes21, and the relationships between the s1–s11 and k1–k10 subtype classifications in the TCGA RPPA dataset mirrored the relationships observed in the proteomic compendium dataset (Figs. 4b and 3a). Also consistent with the proteomic compendium results, TCGA s3 tumors were highly enriched for basal-like breast cancer (p < 1E−35, one-sided Fisher's exact test), TCGA s11 tumors were highly enriched for brain tumors (TCGA GBM and LGG projects, p < 1E−50), and TCGA s10 tumors were moderately enriched for brain tumors (p < 0.005). Similarly, we classified 375 cancer cell lines with mass spectrometry-based proteomic data according to s1–s11 subtypes (Fig. 4c and Supplementary Data 4). Consistent with previous results23, not all differential patterns observable in human tumor proteomic data appeared as strong in the cell line proteomic data, particularly regarding immune-associated or stroma-associated subtypes, attributable to various factors including growth conditions of cell lines lacking tumor microenvironmental effects. For 301 of the 375 cell lines, CRISPR knockout screens globally assessed gene essentiality31. In taking the global correlation between differential protein expression profile versus gene essentiality scoring profile for each cell line, s3 and s10 cell lines had consistent negative correlations in contrast to the other subtypes (Fig. 4d). This observation indicated that s3 and s10 cell lines (and, by extension, their tumor counterparts) tended to express essential genes highly (Fig. 4e).
The s10 and s11 subtypes (consisting of 350 and 441 tumors, respectively) involved brain tumors, with s10 enriched for pediatric brain tumors and s11 consisting entirely of brain tumors (pediatric and adult glioblastoma). These subtypes did not strongly correspond to any of the previous proteome-based subtypes, as the previous data did not involve brain tumors.

Proteins high in s11 tumors were highly enriched for brain tissue-specific genes by GTEX (Methods). By key somatic alteration events, protein and phospho-protein features, and pathway-associated proteomic signatures, s10 tumors showed increased alteration of the MYC, Wnt, and Hippo pathways compared to the rest of the tumors (Figs. 3d and 5a). Brain tumors of the
s10 subtype appeared molecularly distinct from s11 brain tumors. For example, s10 brain tumors showed lower expression of proteins involved in metabolic pathways, while s11 tumors showed higher average expression for these same pathways (Fig. 5b). In terms of previously identified glioblastoma subtypes at the protein level, s10 glioblastomas associated with nmf3 classical tumors, while s11 glioblastomas associated with nmf1 proneural tumors (Fig. 5c). Concerning pediatric brain tumor histologic types, s10 tumors were associated with ependymomas and medulloblastomas, while s11 tumors were associated with...
gangliogliomas and low-grade gliomas (Fig. 5c). Our pediatric brain tumors were divided primarily among s2, s5, s10, and s11 subtypes (Fig. 3a). In pediatric brain tumors from the CBTRUS cohort14,32, the above subtypes involved differences in patient survival, both among the patients with tumors in the proteomic compendium and those with tumors having RNA-seq data only (Fig. 5d).

The proteome-based subtypes also showed differences from each other in overall CNA burden. The s10 and s3 subtypes had the highest CNA burden on average, and the s1 and s1 subtype had the lowest CNA burden (Fig. 6a). A high rate of CNA within a given tumor may result in an altered molecular profile due to the extensive DNA damage involv33. Across all tumors, we computed the correlation between the expression of each gene and CNA burden at both protein and mRNA levels. We observed most of the significant gene-level associations at both the protein and mRNA levels. However, some genes were significantly correlated with CNA burden at the protein level but not at the mRNA level or vice versa (Fig. 6b). Genes increasing with higher CNA burden at both protein and mRNA levels involved chromatin, histone modification, transcription factor binding, cell cycle checkpoint, DNA repair, and methyltransferase complex (Fig. 6c). Genes increasing with higher CNA burden at the protein level but not at the mRNA level involved nuclear hormone receptor binding, histone deacetylation, and DNA repair. Genes increasing with higher CNA burden at the mRNA level but not at the protein level involved the electron transport chain, the NADH dehydrogenase complex, gluconeogenesis, and the nucleosome. Consistent with previous studies examining molecular correlates of overall structural variation burden across tumors22,23, tumors with high CNA burden showed both high proteome-based signature scoring for DNA damage response pathways and low scoring for immune cell infiltrates, though with some genes in these signatures being significant by protein but not mRNA (Fig. 6d).

Pathway-level somatic alterations. We hypothesized that somatic alteration of well-characterized oncogenic or tumor suppressor pathways would be reflected in the cancer proteome, in terms of the downstream effects of altered pathway signaling. Across the entire proteomic compendium tumor dataset, assessment of genes within pathways demonstrated a high number of somatic alterations (small mutation or CNA) involving chromatin modification (73.4% of 1597 tumors involving 12 cancer types with small mutation and CNA data available), p53/Rb-related (73.1%), SWI/SNF complex (70.0%), PI3K/AKT/mTOR (69.2%), Receptor Tyrosine Kinase signaling (RTK, 55.0%), Wnt/beta-catenin (51.8%), MYC/MYCN (39.6%), NRF2 (8.2%), and Hippo signaling (36.3%) (Fig. 7a). The above pathways were altered in different ways involving different genes in different cancer types (Fig. S3). Expected associations of cancer type with somatic pathway alteration were observed, including MYC amplification in breast cancer; Wnt pathway alterations via APC mutation in colorectal cancer; NRF2 pathway alteration in squamous lung cancer; KRAS mutations in colorectal, lung adenocarcinoma, and pancreatic cancers; TP53 mutations in ovarian, lung squamous, and head and neck cancers; and mTOR pathway alterations via VHL mutation in renal cancer. As a means to integrate proteomic with somatic DNA alteration data, gene expression signatures for TP53, k-ras, MTOR, Wnt/beta-catenin, MYC, Nrf2/Keap1, and YAP1/ Hippo—based on data from experimental models—were applied to both the protein and the mRNA expression profiles of the tumors. We scored each tumor sample profile for each of the above pathways, with a higher signature score indicating a greater level of pathway activation in cells. For each pathway considered, relative levels of the corresponding signature—both at protein and mRNA levels—were significantly different between somatically altered versus unaltered tumors for that pathway, with the differences being in the anticipated direction (Fig. 3c). The one notable exception to the above was the p53 signature, where the mRNA-based but not the protein-based signature scoring was significantly associated with p53/Rb-related DNA alterations. The p53 signature consisted of 27 canonical p53 transcriptional targets23 represented in the proteomic compendium dataset. At the mRNA level, 17 of the 27 p53 signature genes had a significant negative expression correlation (p < 0.05) with TP53 mutation status, while at the protein level, only 9 of the 27 genes had a similar negative correlation.

In addition to entire gene signatures, specific proteins could show differential levels in according to somatic pathway alteration. For example, we could define proteomic signatures associated with mutant TP53 or mutant KRAS (Fig. 8a), representing two well-established cancer genes for which small mutation is the primary mode of pathway deregulation. Interestingly, a substantial number of proteins differentially expressed with TP53 mutation were not reflected at the mRNA level (Fig. 8b). Out of 2752 proteins differential with TP53 mutation (false discovery rate, or FDR < 10%), 1134 (41%) were not similarly altered significantly (p < 0.05) at the mRNA level (Supplementary Data 2 and 3). Similarly, out of 5414 genes with mRNA differential with TP53 mutation (FDR < 10%), 2272 (42%) were not similarly altered at the protein level. TP53 mutation associates with higher overall mutational burdens in cancer32, and most of the proteomic signature of TP53 mutation overlapped with the proteomic signature of CNA burden described above; at the same time, 638 (23%) of the 2752 TP53 mutation-associated genes (FDR < 10%) were not significant in the same direction (p < 0.05) in the CNA burden analysis (Supplementary Data 2). The KRAS mutant gene signatures involved 199 proteins and 580 mRNAs (FDR < 10%), with 109 of the significant proteins showing the same trend (p < 0.05) at the mRNA level. Genes with expression higher with TP53 mutation at both the protein and mRNA levels included genes involved in cell division, while genes higher specifically at the protein level involved SWI/SNF
complex, and genes higher specifically at the mRNA level involved the proteasome, NF-kappaB signaling, and Wnt pathway (Fig. 8c). Genes with expression higher or lower with KRAS mutation at the mRNA level included genes involved in cell adhesion or humoral immune response, respectively (Fig. 8c). Individual genes of interest associated with TP53 or KRAS mutation at the protein but not mRNA levels would have known functional roles with the respective pathways (Supplementary Data 2 and 3). For example, protein levels for p53 and HEATR1 were higher with TP53 mutation (Fig. 8d), where depletion of
HEATR1 would lead to impaired proliferation and induction of
p53-dependent cell cycle arrest34. Proteins higher with KRAS
mutation included INSR (insulin receptor)—where insulin can
promote invasion and migration of KRAS mutant cell35—and
ERBIN (erb2 interacting protein)—which can regulate KRAS
mutant-induced tumorigenesis36.

Phospho-proteins are a component of pathway signaling, and,
across the proteomic compendium, specific phospho-proteins of
interest could be associated with pathway-level somatic alteration
or pathway signature across tumors. Overall, phospho-proteins
tended to be significantly positively correlated with the
corresponding total protein, as expected (Supplementary Data 6).
We considered a set of 106 phospho-proteins representing 25
genes involved in the pathway-level somatic alteration classes
(from Fig. 7a), which phospho-protein features had sufficient data
(>50% of tumors for seven or more cancer types). Many other
phospho-proteins of interest, e.g., canonical pathway members of
PI3K/AKT/mTOR37, were not detected in enough tumor
numbers across the proteomic compendium. On the other hand,
58 of the 106 phospho-proteins represented a phosphosite not
cataloged in the Human Protein Reference Database (HPRD)38.
Of the 106 phospho-proteins, 36 significantly correlated (p < 0.05,
Pearson correlation) with the corresponding pathway-level DNA
somatic mutation class (Fig. 9a, 15 of these representing uncatalogued phosphosites), and 24 significantly correlated with the
corresponding pathway-level gene signature (Fig. 9b, 10 of these representing uncatalogued phosphosites), nine of which
significantly correlated with pathway-level mutation class. Of the
nine phospho-proteins, four, including PI3K/AKT/mTOR-related
PTEN:s294 and RTK-related ERBB2:t671 and ERBB2:s968, were not
found in HPRD (Fig. 9c).

A wide range of pathway gene signature levels was evident
within somatically altered and unaltered groups for a given
pathway. In addition to biological and technical noise that would
be inherent in the data, there is the possibility that alterations in
other genes may help drive a given pathway, even if that gene may
not have a well-established or canonical role in that pathway37.
In our proteomic compendium dataset, we examined the set of
tumors with high pathway signature scoring but with no
canonical mutations (see Methods). We aimed to identify any
enrichment patterns for somatic small mutation events involving
a set of 190 genes significantly mutated in pan-cancer studies39,40
(Fig. 10a). We evaluated each of the MYC/MYCN, Hippo,
mTOR, NRF2, p53/Rb-related, RTK, and Wnt/beta-catenin
pathways in this way. We also carried out this analysis across
10,224 tumors in TCGA pan-cancer dataset (with signature
scoring based on mRNA data). We looked for overlap in the
respective results between the proteomic compendium and
TCGA. In the proteomic compendium, 41 genes were signifi-
cantly enriched for mutation events within pathway-activated
tumors with no canonical mutations (p < 0.01, one-sided Fisher’s
exact test), while 45 genes were significantly enriched for TCGA
datasets (p < 0.001). Between the two tumor cohorts, we observed
significant overlap between the respective gene-level associations
(p < 5E–8, one-sided Fisher’s exact test), involving 12 genes, all of
which were enriched for mutation events in tumors with high
MYC signature (Fig. 10b). For each dataset, somatic mutations in
any of the 12 genes would account for a substantial number of
tumors in addition to tumors with MYC or MYCN copy gain
alone (Fig. 10c), on the order of an additional 14–18% of tumors
surveyed. For each dataset, mutations in each of the 12 were
associated with elevated MYC pathway signature scoring, like
what we had observed above regarding MYC and MYCN copy
gain or amplification (Fig. 10d). For several of the 12 genes,
previous experimental studies have demonstrated a role for loss of
the gene leading to increased MYC expression or activity, including
ARID1A41, ATM42, EP30043, PTEN44, RB145, SMARCA446, and
ZFHX347.

Discussion

Across a large pan-cancer cohort of primary tumors, we observed
general correspondence between protein and mRNA differential
expression patterns, but with notable exceptions. Genes with
proteins having less correlation across tumors with the corre-
sponding mRNA or CNA patterns included genes involved in
ribosomes, translation, electron transport chain, humoral
immune response, DNA repair, or cell cycle regulation. Genes with
expression increasing with higher CNA burden at only the
protein level included genes involved in nuclear hormone
receptor binding, histone modification, and DNA repair. Path-
ways associated with high-grade cancers at the protein but not the
mRNA level included type II interferon and Oncostatin M sig-
naling. A gene signature of p53 transcriptional targets applied to
mRNA data showed an expected association with p53 pathway-
associated mutations, while the signature applied to protein data
did not show a similar result. Widespread differential protein
expression patterns could be associated with somatic mutation of
TP53 or KRAS, with some patterns not reflected at the mRNA
level. A lack of correlation or association represents a negative
result, inherently difficult to prove using statistics, and technical
issues regarding the data platforms might conceivably be at play.
At the same time, the fact that the sets of genes involved in the
overall differences observed between protein and mRNA are
significantly enriched for specific functional groups and anno-
tated pathways would suggest that true biology is involved here.
Our study reinforces the notion that cancers should be compre-
hesively surveyed at the protein level, where expression profiling
on tumors has historically been mostly limited to the transcript level.
The proteome-based molecular subtypes identified using our tumor cohort reinforce the notion of pan-cancer molecular classes cutting across tumor lineages and cancer types\(^2\). Most of the proteome-based subtypes identified previously in a more limited cohort\(^3\) were re-discovered here by unsupervised analyses of a larger cohort with additional cancer types. We again observed multiple distinct subtypes involving the immune system, one involving the adaptive immune response and T-cell activation, and others associated with the humoral immune response and complement pathway. We also again observed tumor stroma-associated subtypes, one involving collagen VI network. This present study extends the previously reported subtype-related findings to additional cancer types. Importantly, we could also identify additional proteome-based subtypes, including one enriched for brain tumors and associated with MYC, Wnt, and Hippo pathways and with high CNA burden, and another consisting entirely of brain tumors and associated with increased expression of genes in metabolic pathways and with low CNA burden. As shown here, proteome-based subtyping can provide insights into the pathways and processes appearing deregulated within tumor subsets, suggesting therapeutic opportunities. Differences in subtyping assignments between the present study and previous studies could involve differences in the respective cohorts, including represented cancer types. As more human tumors and additional cancer types are profiled by mass spectrometry-based proteomics, additional subtypes may be uncovered and explored in future studies.

We found that somatic alterations of cancer-associated pathways are reflected in the cancer proteome, whereby tumors with somatic alteration involving genes in a pathway tend to show higher levels of protein-based signature scoring for that pathway. Our study took advantage of the large number of tumors represented in our proteomic compendium dataset, whereby we could carry out tumor subset analyses involving sparse mutation events. As cataloged previously, the gene signatures represent the downstream effects of altered pathway signaling, e.g., the upregulation of transcriptional targets in a cell line or mouse model. The proteomic and genomic data integration represents orthogonal information pointing to a common tumor subset deregulated for a given pathway. Our collective knowledge of molecular pathways has been largely derived from experimental models. The signature analyses allow us to explore experimentally inferred cause-and-effect relationships in the human disease setting, whereby these relationships manifest in tumors as significant correlations. Phospho-proteins can denote pathway signaling, though current mass spectrometry-based datasets may not capture all phospho-protein features of interest regarding a pathway. However, post-translational modifications, including phosphorylation events, may be detected by mass spectrometry. Pathway activity, as measured using gene signatures applied to proteome data, reflects known mutations or copy alteration in most but not all tumors examined, suggesting additional, unexplained, or underappreciated mechanisms of pathway activation. In the case of the MYC pathway, we identified several genes, many of which would have an underappreciated role in the MYC pathway, for which somatic mutation involved higher pathway activation. Through proteogenomics, additional members or connections may be incorporated into the standard pathway model.

Our study results provide a framework for understanding the molecular landscape of cancers at the proteome level. The associated datasets and gene-level associations represent a resource for the research community, including helping to identify gene candidates for functional studies. Beyond the additional tumors and cancer types represented in our proteomic compendium of 2022 tumors, the pan-cancer datasets compiled here involve multiple levels of molecular data in addition to the proteome,
including mRNA-level and DNA-level data on the same set of tumors. In contrast, due in part to the timing of data releases, our first pan-cancer proteomic tumor subtyping study focused mainly on proteomic data in CPTAC tumors, while our recent study examining tumor grade correlates incorporated both protein and mRNA data but no somatic DNA-level data on CPTAC tumors.

Our present proteogenomics survey enabled us to explore the interplay between genome and proteome, including aspects not represented in the transcriptome. When previously surveying proteomic correlates of tumor grade, we identified specific protein kinases having functional impact in vitro in uterine endometrial cancer cells, which provides a template for other researchers to utilize the gene-level associations provided in the present study. In addition to correlates of tumor grade, proteomic correlates of somatic alteration of pathways, including TP53 or KRAS mutation, may be of interest for further study. Regarding the pan-cancer proteome-based subtypes, a number of these appear manifested in cancer cell lines, and additional data, such as DepMap gene dependency data, may be leveraged to help select targets of interest to examine in these cell lines. RPPA profiling data can classify tumors according to subtype, where RPPA antibody-based features may also lend themselves to relevant immunohistochemistry studies. For all cancer types studied, we have added the proteomic datasets to the user-friendly UALCAN data portal, facilitating differential analyses by protein, giving the research community ready access to our results.

Methods

Proteomic datasets. We assembled a compendium dataset of mass spectrometry-based proteomics data of primary tumors from 14 cancer types and 17 individual
ARTICLE

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-022-30342-3 | www.nature.com/naturecommunications

studies3–19 (Supplementary Data 1). Most of these previous studies were led by either CPTAC or the International Cancer Proteogenome Consortium (ICPC). The cancer types represented in the proteomics compendium dataset were the following: Breast Invasive Carcinoma (n = 230 tumors with proteomics data)14, Colorectal Adenocarcinoma (n = 187)15,6, Gastric Cancer (n = 80)16, Glioblastoma (n = 100)17, Head and Neck Squamous Cell Carcinoma (n = 108)18, Hepatocellular Carcinoma (n = 165)19, Lung Adenocarcinoma (n = 111)20, Lung Squamous Cell Carcinoma (n = 110)20, Ovarian Serous Cystadenocarcinoma (n = 209)21, Pancreatic Ductal Adenocarcinoma (n = 137)22, Pediatric Brain Tumors (n = 219)23, Prostate Adenocarcinoma (n = 76)24, Renal Cell Carcinoma (n = 110)25, and Uterine Corpus Endometrial Carcinoma (n = 100)26. The 2002 tumors in the compendium represented 1982 patients, with the pediatric brain tumor dataset involving 219 tumors from 199 patients. The above studies analyzed the tumors using global proteomic and phosphoproteomic profiling by liquid chromatography-tandem mass spectrometry (LC-MS/MS). We obtained processed protein expression data from the CPTAC Data Portal27 or the associated publications’ supplementary tables. Proteomic data, as provided by the CPTAC Data Portal and related publications, were processed at the gene level rather than at the protein isoform level; as a simplification, we did not consider different isoforms for the same protein in the present study.

For each study, taking the expression values provided in the associated data table, we normalized proteomic data for downstream analyses in the following manner and as previously described21,22. First, within each proteomic profile, we normalized log2 expression values to standard deviations from the median. Next, we normalized expression values across samples to standard deviations from the median. Similarly, we separately normalized both total protein and phospho-proteins for a given cancer type and dataset. For datasets where two different data centers generated values on the same tumors, we averaged normalized values from the respective data centers in instances of duplicate samples for the same tumor sample. As intended, by normalizing expression within each cancer type and within each proteomic dataset, neither tissue-dominant differences nor inter-laboratory batch effects would drive the downstream analysis results, including unsupervised subtype discovery. For the compendium dataset of total proteins, a total of 15,439 unique genes by Entrez identifier were represented in at least seven cancer types.

Transcriptomic datasets. Of the 2002 human tumors with proteomics data, 1899 had corresponding RNA-seq data. For CPTAC projects utilizing tumors from TCGA, we obtained TCGA data RNA-seq data from the Broad Institute’s Firehose...
**Fig. 8** Proteomic and transcriptomic signatures associated with somatic mutation of TP53 or KRAS. a Differential expression heat maps of top total protein features correlated with small somatic mutation of TP53 (left, FDR < 10%) or with somatic hotspot mutation of KRAS (right, FDR < 10%) across all cancers studied (n = 1440 tumors with protein and mutation data). SD standard deviation from the median. b Across TP53 and KRAS comparisons, heat map representation of differential statistics for protein and mRNA features correlated with mutation status (FDR < 10%). c Taking the genes from b, gene sets were defined according to mutation association by protein versus mRNA (using p < 0.05 with the 5414 genes significant with FDR < 10% for any comparison). For each mutation signature analysis, genes significant (up in mutant versus down) for both protein and mRNA or for one molecular level but not the other were assessed for represented categories by GO, with selected enriched categories represented here. p values by one-sided Fisher’s exact test. d For selected genes (TP53, HEATR1, INSR, ERBIN), box plots compare TP53 or KRAS mutant versus wild-type tumors for relative levels of protein versus mRNA, as indicated. p values by t-test. Box plots represent 5% (lower whisker), 25% (lower box), 50% (median), 75% (upper box), and 95% (upper whisker). n = biologically independent tumors.
Similarly to the CPTAC copy data. For the prostate cancer dataset, gene-level copy amplification (for all genes), then thresholded to values approximating homozygous deletion lute copy data according to ploidy (dividing gene copy value by average copy value level absolute copy data (0, 1, 2, 3, 4, 5) for the other CPTAC projects from the SNP array-based CNA the original study). For CPTAC projects utilizing tumors from TCGA, we obtained data were available for the ICPC Gastic Cancer project (as these were not a part of proteomics data, 1837 had corresponding gene-level CNA data. No gene-level CNA Copy number alteration (CNA) datasets data were available for the non-CPTAC projects from links or accession numbers found.

**Fig. 9 Phospho-proteins associated with pathway alteration.** a For 106 phospho-proteins (with data for >50% of samples for seven or more cancer types) with genes involved in curated pathway sets^3,25,57,68, the Pearson correlation with each pathway-level DNA mutation class from Fig. 7a was computed. Shown are 36 of the 106 phospho-proteins for which a significant association (p < 0.05) with its corresponding pathway mutation class was found. b For the 106 phospho-proteins, the Pearson correlation with each pathway gene signature from Fig. 7b was computed. Shown are 24 phospho-proteins for which a significant association (p < 0.05) with its corresponding pathway signature was found. Nine phospho-proteins are shared between a and b results. c For selected phospho-proteins PTEN:s294, ERBB2:t671, and ERBB2:s968, associations with mTOR pathway or RTK pathway mutation class (left) and association with mTOR pathway or RTK pathway signature (right) is represented. p values by t-test (box plots) or Pearson correlation (scatter plots). Box plots represent 5% (lower whisker), 25% (lower box), 50% (median), 75% (upper box), and 95% (upper whisker). n = biologically independent tumors.

data portal [https://gdc.broadinstitute.org], and we obtained RNA-seq data for the other CPTAC projects from the Genome Data Commons [https://gdc.cancer.gov/]. We obtained RNA-seq data on CBTN pediatric brain tumors through the public project on the Kids First Data Resource Portal and Cavatica [https://cbtn.org]. We obtained data for the non-CPTAC projects from links or accession numbers provided with the associated publications. We normalized expression values across samples to standard deviations from the median within each cancer type and dataset, as we carried out above for proteomic data.

**Copy number alteration (CNA) datasets.** Of the 2002 human tumors with proteomics data, 1837 had corresponding gene-level DNA data. No gene-level DNA data were available for the ICPC Gastic Cancer Project (as these were not a part of the original study). For CPTAC projects utilizing tumors from TCGA, we obtained SNP array-based CNA “themed” values (−2, −1, 0, 1, 2) from the Broad Institute’s Firehose data portal [https://gdc.broadinstitute.org]. We obtained gene-level absolute copy data (0, 1, 2, 3, 4, 5) for the other CPTAC projects from the Genome Data Commons [https://gdc.cancer.gov/]. We first normalized the absolute copy data according to ploidy (dividing gene copy value by average copy value for all genes), then thresholded to values approximating homozygous deletion (−2); heterozygous deletion (−1), wild-type (0), gain of 1−2 copies (+1), and amplification with at least 3 copies (+2). We obtained gene-level copy data from CBTN pediatric brain tumors from Cavatica [https://cbtn.org] and thresholded similarly to the CPTAC copy data. For the prostate cancer dataset, gene-level copy data based on WGS data was previously generated by the Pan-cancer Analysis of Whole Genomes consortium from a consensus of multiple CNA callers^90. For the hepatocellular carcinoma project, we obtained whole-exome-based gene-level thresholded CNA calls from the National Omics Data Encyclopedia [https://www. biosinfo.org/node/project/detail/OEP0000321].

We computed an overall CNA burden index for each CNA profile, defined as the standard deviation of thresholded CNA values across all genes. The CNA burden index was high in tumors with many copy gain/amplification or loss/deletion events. Across all samples, we computed the Pearson correlation of each protein or mRNA feature (using normalized values) with the CNA burden index. The method of Storey and Tibshirani^92 estimated FDR for genes significantly associated with the overall CNA burden. Using an alternative regression model incorporating cancer type did not improve the overall results, as these did not depend on cancer type differences (as we normalized all gene expression features within each cancer type, removing cancer type-specific differences). In considering proteins versus mRNA features significantly associated with CNA burden, we considered the 10,315 genes for which there were both protein and mRNA data for more than 500 tumors. For gene signatures of DNA damage response pathways, we obtained curated gene sets from ref. 52 and took the average of the normalized gene values.

**Small mutation datasets.** Of the 2002 human tumors with proteomics data, 1698 had corresponding small somatic mutation data (SNVs and indels) by whole-exome or whole-genome sequencing. No small mutation data were available for the ICPC.

---

**Fig. 9 Phospho-proteins associated with pathway alteration.** a For 106 phospho-proteins (with data for >50% of samples for seven or more cancer types) with genes involved in curated pathway sets^3,25,57,68, the Pearson correlation with each pathway-level DNA mutation class from Fig. 7a was computed. Shown are 36 of the 106 phospho-proteins for which a significant association (p < 0.05) with its corresponding pathway mutation class was found. b For the 106 phospho-proteins, the Pearson correlation with each pathway gene signature from Fig. 7b was computed. Shown are 24 phospho-proteins for which a significant association (p < 0.05) with its corresponding pathway signature was found. Nine phospho-proteins are shared between a and b results. c For selected phospho-proteins PTEN:s294, ERBB2:t671, and ERBB2:s968, associations with mTOR pathway or RTK pathway mutation class (left) and association with mTOR pathway or RTK pathway signature (right) is represented. p values by t-test (box plots) or Pearson correlation (scatter plots). Box plots represent 5% (lower whisker), 25% (lower box), 50% (median), 75% (upper box), and 95% (upper whisker). n = biologically independent tumors.

---

**Copy number alteration (CNA) datasets.** Of the 2002 human tumors with proteomics data, 1837 had corresponding gene-level DNA data. No gene-level DNA data were available for the ICPC Gastic Cancer Project (as these were not a part of the original study). For CPTAC projects utilizing tumors from TCGA, we obtained SNP array-based CNA “themed” values (−2, −1, 0, 1, 2) from the Broad Institute’s Firehose data portal [https://gdc.broadinstitute.org]. We obtained gene-level absolute copy data (0, 1, 2, 3, 4, 5) for the other CPTAC projects from the Genome Data Commons [https://gdc.cancer.gov/]. We first normalized the absolute copy data according to ploidy (dividing gene copy value by average copy value for all genes), then thresholded to values approximating homozygous deletion (−2); heterozygous deletion (−1), wild-type (0), gain of 1−2 copies (+1), and amplification with at least 3 copies (+2). We obtained gene-level copy data from CBTN pediatric brain tumors from Cavatica [https://cbtn.org] and thresholded similarly to the CPTAC copy data. For the prostate cancer dataset, gene-level copy data based on WGS data was previously generated by the Pan-cancer Analysis of Whole Genomes consortium from a consensus of multiple CNA callers^90. For the hepatocellular carcinoma project, we obtained whole-exome-based gene-level thresholded CNA calls from the National Omics Data Encyclopedia [https://www. biosinfo.org/node/project/detail/OEP0000321].

We computed an overall CNA burden index for each CNA profile, defined as the standard deviation of thresholded CNA values across all genes. The CNA burden index was high in tumors with many copy gain/amplification or loss/deletion events. Across all samples, we computed the Pearson correlation of each protein or mRNA feature (using normalized values) with the CNA burden index. The method of Storey and Tibshirani^92 estimated FDR for genes significantly associated with the overall CNA burden. Using an alternative regression model incorporating cancer type did not improve the overall results, as these did not depend on cancer type differences (as we normalized all gene expression features within each cancer type, removing cancer type-specific differences). In considering proteins versus mRNA features significantly associated with CNA burden, we considered the 10,315 genes for which there were both protein and mRNA data for more than 500 tumors. For gene signatures of DNA damage response pathways, we obtained curated gene sets from ref. 52 and took the average of the normalized gene values.

**Small mutation datasets.** Of the 2002 human tumors with proteomics data, 1698 had corresponding small somatic mutation data (SNVs and indels) by whole-exome or whole-genome sequencing. No small mutation data were available for the ICPC.
Fig. 10 Noncanonical somatic mutation associations with MYC family alterations. a Schematic of approach to identify somatic mutations (SNVs or indels) associated with a given pathway. Taking the set of tumors with high pathway signature scoring but with no canonical mutations (based on Fig. 7), we consider noncanonical genes for enrichment of mutations. Results are compared between the proteomic compendium and The Cancer Genome Atlas (TCGA), where we observed significant overlap between the respective gene-level associations (p < 5E-8, one-sided Fisher’s exact test). b List of top mutated genes overlapping between proteomic compendium and TCGA results (p < 0.01 and p < 0.001, respectively, one-sided Fisher’s exact test). Results involve 12 genes with mutation events enriched within tumors with high MYC signature but no MYC or MYCN copy gain or amplification. c Across both the proteomic compendium (left) and TCGA (right) tumor datasets, heat maps of MYC/MYCN copy alteration and of somatic mutation (SNV or indel) in genes from b. d Box plots of MYC/MYCN signature scores by alteration class in proteomic compendium (left) and TCGA (right) tumor datasets. p values versus unaligned group (tumors with none of the listed alterations) by t-test. Box plots represent 5% (lower whisker), 25% (lower box), 50% (median), 75% (upper box), and 95% (upper whisker). n = biologically independent tumors. Cancer type represents TCGA project name.
Liver Cancer project. For CPTAC projects utilizing tumors from TCGA, we obtained somatic mutation calls by whole-exome sequencing from the publicly available “MC5” TCGA MAT file [https://www.synapse.org/#!Synapse:syn214102;] variables called by two or more algorithms were used in this study. For the other CPTAC projects, we obtained whole-exome somatic mutation calls from the Genome Data Commons [https://gdc.cancer.gov/]; variables called by two or more algorithms were used in this study. For the prostate cancer dataset, we obtained whole-genome somatic mutation calls from the supplemental of ref. 63. Somatic SNV calls from the ICPC Gastic Cancer project were kindly provided by the authors of the associated study.

In considering proteins versus mRNA features significantly associated with small molecule and/or immune checkpoint inhibitor responses, we classified each pan-cancer subtype averaged proteins to standard deviations from the median. As a classifier, we used the top set of proteins referring to the same protein, we randomly selected one tumor profile as the reference population. We searched each of the pan-cancer subtype as defined by pan-cancer-based pan-cancer subtype. For the RPPA dataset, we used as a classifier the top set of proteins distinguishing between the pan-cancer subtype. For each pan-cancer subtype, we computed the average normalized value for each protein, based on the centered expression data matrix. We then computed the Pearson correlation between each external profile and each pan-cancer subtype average profile. We assigned each external cancer case to a pan-cancer subtype, based on which subtype profile showed the highest correlation with the given external dataset profile. Supplementary Data 4 provides example calculations in Excel, by which the TCGA and CCLE proteomic profiles are classified according to proteome-based pan-cancer subtype. For the RPPA dataset, we used as a classifier the set of represented total protein features from which a significant association with a particular subtype was observable in the proteomic compendium dataset. Where multiple RPPA features referred to the same protein, we randomly selected one feature.

Gene signature analyses. Our group previously collected gene signatures involving pathways of interest24,25,26. We computed pathway-associated gene expression signature scores (e.g., scores for EMT, hypoxia, KEGG: Glycolysis/Gluconeogenesis, KEGG: Pentose Phosphate pathway, KEGG: Fatty Acid metabolism, KEGG: TCA Cycle, and KEGG: Oxidative Phosphorylation or OXPHOS, k-ras, MYC, Notch, NRF2/KEAP1, PDK3/mTOR, p53, Wnt, and YAP1) and interacted with all genes previously associated with these scores based in the scoring calculation. We used the normalized values within the proteomic and transcriptomic compendium datasets (values normalized to standard deviations from the median within each dataset and cancer type). For NRF2/ KEAP1, hypoxia, Notch, p53, KEGG, and GO term signatures, we computed the average expression of the set of genes within a given signature. For k-ras, MYC, PDK3/mTOR, Wnt, and YAP1 signatures, normalized expression profiles were scored for the above signatures using our t-score metric. We generated gene signature scores of NRF2/KEAP1 pathway as described27, based on four different signatures27. The hypoxia signature was the set of canonical HIF1A target genes28. We generated gene transcription signature scores of p53 pathway, based on four different signatures29. The MYC signature (from data by ref. 60) was from ref. 62. The Settleman k-ras sensitivity signature was from ref. 63. The Wnt signature was taken directly from ref. 64. NOTCH signature was from ref. 61) was from ref. 62. The Settleman k-ras sensitivity signature was from ref. 63. The Wnt signature was taken directly from ref. 64. NOTCH signature was from ref. 61. The Wnt signature was taken directly from ref. 64. NOTCH signature was from ref. 61. The Wnt signature was taken directly from ref. 64. NOTCH signature was from ref. 61. The Wnt signature was taken directly from ref. 64. NOTCH signature was from ref. 61. The Wnt signature was taken directly from ref. 64. NOTCH signature was from ref. 61. The Wnt signature was taken directly from ref. 64. NOTCH signature was from ref. 61. The Wnt signature was taken directly from ref. 64. NOTCH signature was from ref. 61. The Wnt signature was taken directly from ref. 64. NOTCH signature was from ref. 61. The Wnt signature was taken directly from ref. 64. NOTCH signature was from ref. 61.
To computationally infer the infiltration level of specific immune cell types using proteomic data, we used a set of 501 genes specifically over-expressed in one of nine immune cell types from ref. 47. For each Bindera immune cell type signature, we computed the average expression of the set of genes within the given signature. For T cells, we computed a signature score based on the average normalized values of five canonical markers (LCK, CD3E, CD3D, CD3G, CD2). For cytotoxic T cells, we computed a signature score based on canonical markers CD8A and CD8B. For T helper and regulatory T cells, we considered CD4 marker expression. For an immune checkpoint pathway signature, we computed the average of normalized protein values for a set of previously defined genes representing targets for immunotherapy17, including PD达CD1 (PD1), CD247 (CD47), CD137 (CTLA4), CD11C (DCS), CD123 (IL3), TNFRSF9 (CD125), TNFRSF14 (CD134). In addition, sample profiles were scored for average normalized expression of Antigen Presentation MHC class I (APM1) genes (HLA-A/B/C, B2M, TAP1/2, TAPBP) and for average normalized expression of Antigen Presentation MHC class II (APM2) genes.

Pathway-level somatic alteration categories. For the pathway-centric view of somatic alterations across the proteomic compendium (Fig. 7), we focused on key pathways and genes previously annotated across multiple cancer types based on domain knowledge (see Methods)45,52,70, from the 2002 human tumors with proteomic data in our compendium, 1597 tumors had combined whole-exome and CNA data. Key drivers associated with pathway enrichment patterns of small mutations within tumors that showed high pathway-specific expression scoring for a given pathway (SD > 0.5 from the sample median) but with mTOR, NRF2, p53/Rb-related, RTK, Wnt/beta-catenin). Taking the set of tumors with annotation and expression signature scores were available (MYC/MYCN, Hippo, ERBB4, KDM5B, pathway (p53/Rb-related (NATURE COMMUNICATIONS| (2022) 13:2669 | https://doi.org/10.1038/s41467-022-30342-3 | www.nature.com/naturecommunications

Statistical analysis
All p values were two-sided unless otherwise specified. Nominal p values do not involve multiple comparison adjustments, while FDRs involve p values adjusted for multiple gene feature comparisons. We performed all tests using log2-transformed expression values. Visualization using heat maps was performed using both JavaTview (version 1.1.64)69 and matrix2png (version 1.2.1)17. Figures indicate the exact value of p (number of tumors), and the statistical tests used are noted in the figure legends and next to reported p values in the Results section. Box plots represent 5% (lower whisker), 25% (lower box), 50% (median), 75% (upper box), and 95% (upper whisker).

Reporting summary
Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
All data used in this study are publicly available. Proteomics data are available for CPTAC and ICPIC studies at the Proteomic Data Commons (https://pdc.cancer.gov/). For CPTAC studies, transcriptome data, copy number data, and small somatic mutation data are available at the Genomic Data Commons (https://gdc.cancer.gov/). Raw genomic and transcriptomic CPTAC data can be accessed via dbGap Study Accession: phs012877.v17.p5. Raw data may be obtained once authorized access is granted via Data Use Certification (DUC) agreement. Genomic and transcriptomic liver cancer data are available at https://www.bioinfo.org/node/project/detail/OEP000321. Genomic and transcriptomic prostate cancer data are available at found on European Genome-Phenome Archive (EGA), under accession EGA: EGAS00001009900. Genomic and transcriptomic gastric cancer data are available in NCBI SRA (PRJNA505380) and GEO (GEO: GSE122401) databases, respectively. TCGA data RNA-seq data are also available from the Broad Institute’s Firehouse data portal (https://gdac.broadinstitute.org). The TCGA RPPA dataset is available from the TCGA portal (http://tcgaportal.org/tcga/). Cancer Cell Line Encyclopedia (CCLE) datasets are available from the CCLE website (http://www.broadinstitute.org/ccle). For other published studies, molecular data availability information is provided in the associated publication. The compendium datasets of molecular profiles for total protein, phospho-protein, and mRNA—compiled as part of our study—are available through GitHub (https://github.com/chadcreighton/cancer-proteomics-compendium-n2002). Each molecular dataset is uploaded on GitHub as a series of separate project files, using a common protein feature set with the same ordering across files. One can concatenate the files together. Each molecular dataset has a common sample set, allowing one to derive correlations between datasets (e.g., between mRNA and protein). The phospho-protein datasets consist of 5149 phospho-protein features that had available data for >50% of samples in at least seven cancer types. Any remaining data are available within the Article or Supplementary Information. Source data are provided with this paper.

Source code
Source code is in R (written using version 4.0.3) for identifying de novo pan-cancer proteomic subtypes, using ConsensusClusterPlus method (version 1.59.0), with an example dataset of proteomic expression for the top ~2000 variable genes across 2002 tumors, is available at Github (https://github.com/chadcreighton/cancer-proteomics-compendium-n2002).

Received: 2 December 2021; Accepted: 25 April 2022;
Published online: 13 May 2022

References
1. Vogel, C. & Marcotte, E. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. 13, 227–232 (2012).
2. Varambally, S. et al. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. Cancer Cell 8, 393–406 (2005).
3. Krug, K. et al. Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. Cell 183, 1436–1456 (2020).
4. Mertins, P. et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 534, 55–62 (2016).
5. Nasrallah, S. et al. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. Cell 177, 1035–1049 (2019).
6. Zhang, B. et al. Proteogenomic characterization of human colon and rectal cancer. Nature 513, 382–387 (2014).
7. Wang, L. et al. Proteogenomic and metabolomic characterization of human glioblastoma. Cancer Cell 39, 509–528 (2021).
8. Huang, C. et al. Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. Cancer Cell 39, 361–379 (2021).
9. Gillette, M. et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. Cell 182, 200–225 (2020).
10. Satpathy, S. et al. A proteogenomic portrait of lung squamous cell carcinoma. Cell 178, 4348–4371 (2019).
11. Hu, Y. et al. Integrated proteomic and glycoproteomic characterization of human high-grade serous ovarian carcinoma. Cell Rep. 33, 108276 (2020).
12. Zhang, H. et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. Cell 166, 755–765 (2016).

Oncogene negative enrichment analyses. TCGA datasets were compiled previously, with tumors assigned pathway-associated gene signature scores based on RNA-seq and, as available according to pathway-centric somatic alterations as described above23. For TCGA tumors, we obtained somatic mutation calls from the publicly available “MC3” TCGA MAF file [https://www.synapse.org/#Synapse:syn2144022]; variants called by two or more algorithms were used in this study. Analyses focused on the 10,224 tumors with RNA-seq data.

For both the proteomic compendium and TCGA datasets, we sought to identify enrichment patterns of small mutations within tumors that showed high pathway-associated expression signature but without a canonical alteration identified by pathway annotation. For this analysis, we focused on the 190 genes in both the COSMIC Cancer Consensus Gene List and the set of significantly mutated genes from the TCGA PanCanAtlas project47. We evaluated each pathway for which both mutation annotation and expression signature scores were available (MYC/MYCN, Hippo, mTOR, NR2F2, p53/Rb-related, RTK, Wnt/beta-catenin). Taking the set of tumors with high signature scoring for a given pathway (SD > 0.5 from the sample median) but with no canonical mutations (based on Fig. 7), we looked for statistical enrichment of SNV or indel events (by one-sided Fisher’s exact test) involving any of the 190 genes. We compared results between the proteomic compendium and TCGA, and we selected for downstream analyses the genes with enrichment of mutation events for both datasets (p < 0.01 and p < 0.001, respectively, one-sided Fisher’s exact test).

Code availability
Source code is in R (written using version 4.0.3) for identifying de novo pan-cancer proteomic subtypes, using ConsensusClusterPlus method (version 1.59.0), with an example dataset of proteomic expression for the top ~2000 variable genes across 2002 tumors, is available at Github (https://github.com/chadcreighton/cancer-proteomics-compendium-n2002).

Received: 2 December 2021; Accepted: 25 April 2022; Published online: 13 May 2022

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-022-30342-3 | www.nature.com/naturecommunications

17
13. Cao, L. et al. Proteogenomic characterization of pancreatic ductal adenocarcinoma. Cell 184, 5031–5052 (2021).
14. Petralia, F. et al. Integrated proteogenomic characterization across major histological types of pediatric brain cancer. Cell 183, 1962–1985 (2020).
15. Clark, D. et al. Integrated proteogenomic characterization of clear cell renal cell carcinoma. Cell 179, 964–983 (2019).
16. Dou, Y. et al. Proteogenomic characterization of endometrial carcinoma. Cell Rep. 27, 729–748 (2020).
17. Gao, Q. et al. Integrated proteogenomic characterization of HBV-related hepatocellular carcinoma. Cell 179, 561–577 (2019).
18. Mun, D. et al. Proteogenomic characterization of human early-onset gastric cancer. Cancer Cell 35, 111–124 (2019).
19. Sinha, A. et al. The proteogenomic landscape of curable prostate cancer. Cancer Cell 35, 414–427 (2019).
20. Cancer_Genome_Atlas_Research_Network et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 45, 1113–1120 (2013).
21. Chen, F., Chandrashekar, D., Varambally, S. & Creighton, C. Pan-cancer molecular subtypes revealed by mass-spectrometry-based proteomic characterization of more than 500 human cancers. Nat. Commun. 10, 5679 (2019).
22. Monsivais, D. et al. Mass-spectrometry-based proteomic correlates of grade and stage reveal pathways and kinases associated with aggressive human cancers. Oncogene 40, 2081–2095 (2021).
23. Chen, F. et al. Pan-cancer molecular classes transcending tumor lineage across 32 cancer types, multiple data platforms, and over 10,000 cases. Clin. Cancer Res. 24, 2182–2193 (2018).
24. Akbani, R. et al. A pan-cancer proteomic perspective on The Cancer Genome Atlas. Nat. Commun. 5, 3887 (2014).
25. Chen, F. et al. Pan-urologic cancer genomic subtypes that transcend tissue of origin. Nat. Commun. 8, 199 (2017).
26. Chandrashekar, D., UALCAN: a portal for facilitating tumor subgroup genotype and expression analysis. Bioinformatics 34, 875–881 (2018).
27. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29 (2000).
28. Slenter, D. et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 46, D661–D667 (2018).
29. Massei, A. et al. Oncostatin M: a mysterious cytokine in cancers. Cytokine Growth Factor Rev. 29 (2018).
30. Wang, G. et al. Insulin promotes invasion and migration of KrasG12D-induced skin tumorigenesis. Cancer Cell 31, 820–832.e3 (2017).
31. Peri, S. et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 13, 2363–2371 (2003).
32. Forbes, S. et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 45, D777–D783 (2017).
33. Ding, L. et al. Perspective on oncogenic processes at the end of the beginning of cancer genomics. Cell 173, 305–320 (2018).
34. Wang, S. et al. SWI/SNF component ARID1A restrains pancreatic neoplasia formation. Gut 68, 1259–1270 (2019).
35. Ualcan, P. et al. ATMi promotes apoptosis and suppresses tumorigenesis in response to Myc. Proc. Natl Acad. Sci. USA 103, 1446–1451 (2006).
36. Bahramy, S., Rashi, H., Thimmappaya, R., Navaraj, A. & Thimmappaya, R. Repression of c-Myc and inhibition of G1 exit in cells conditionally overexpressing p300 that is not dependent on its histone acetyltransferase activity. Proc. Natl Acad. Sci. USA 100, 9524–9529 (2003).
37. Ghosh, A., Grigorieva, I., Steele, R., Hoover, R. & Ray, R. PTEN transcriptionally modulates c-myc gene expression in human breast carcinoma cells and is involved in cell growth regulation. Gene 235, 85–91 (1999).

Acknowledgements This work was supported by National Institutes of Health (NIH) grant P30CA118948 (C.J.C.) and by NIH grant U54 CA118948 and Department of Defense (DOD) grant W81XWH-19-1-0588 (S.V.).

Author contributions Conceptualization: C.J.C.; Methodology: C.J.C. and F.C.; Investigation: C.J.C., Y.Z., and F.C.; Formal analysis: C.J.C., Y.Z., F.C., and D.S.C.; Data curation: C.J.C., S.V., and...
Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-30342-3.

Correspondence and requests for materials should be addressed to Chad J. Creighton.

Peer review information Nature Communications thanks Mukesh Kumar, John Minna and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022, corrected publication 2022