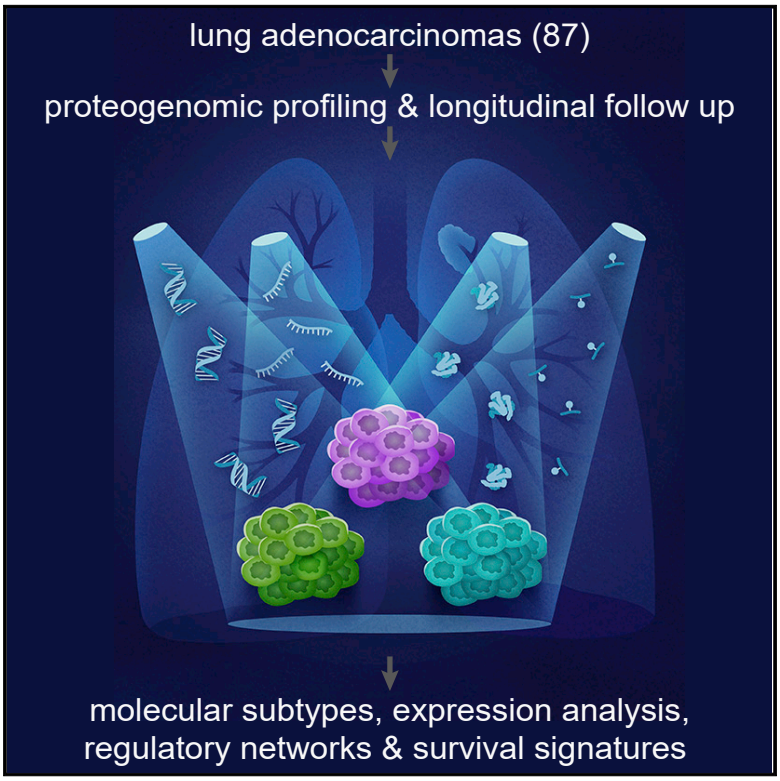
Cell Reports
Medicine

Article

Proteogenomic analysis of lung adenocarcinoma reveals tumor heterogeneity, survival determinants, and therapeutically relevant pathways

Graphical abstract



Highlights

- Lung adenocarcinoma has three subtypes defined by genome alteration profiles
- Tumors with greater immune content have reduced RNA:protein correlation
- Protein and RNA signatures predicting survival of patients are identified
- Phosphoproteomic networks identify potential therapeutic vulnerabilities among subtypes

Authors

Anthony R. Soltis, Nicholas W. Bateman, Jianfang Liu, ..., Christopher A. Moskaluk, Robert F. Browning, Jr., Matthew D. Wilkerson

Correspondence

craig.shriver@usuhs.edu (C.D.S.), cam5p@virginia.edu (C.A.M.), robert.f.browning2.civ@health.mil (R.F.B.), matthew.wilkerson@usuhs.edu (M.D.W.)

In brief

Soltis et al. report a proteogenomic characterization of lung adenocarcinoma from the United States, expanding the disease's etiology with a structurally altered subtype and connecting expression correlations between RNA and protein to tumor immune content. Integrative analysis reveals signatures of patient survival and targets for therapeutic intervention among molecular subtypes.

Soltis et al., 2022, Cell Reports Medicine 3, 100819
November 15, 2022
https://doi.org/10.1016/j.xcrm.2022.100819

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Proteogenomic analysis of lung adenocarcinoma reveals tumor heterogeneity, survival determinants, and therapeutically relevant pathways

Anthony R. Soltis,[1,2] Nicholas W. Bateman,[2,3,4] Jianfang Liu,[5] Trinh Nguyen,[6] Teri J. Franks,[7] Xijun Zhang,[1,2] Clifton L. Dalgard,[1] Coralie Viollet,[1,2] Stella Somiari,[5] Chunhua Yan,[6] Karen Zeman,[8,9] William J. Skinner,[8,9] Jerry S.H. Lee,[2,9,10,11] Harvey B. Pollard,[1] Clesson Turner,[9] Emanuel F. Petricoin,[12] Daoud Meerzaman,[6] Thomas P. Conrads,[3,4] Hai Hu,[5,9] APOLLO Research Network, Craig D. Shriver,[9,*] Christopher A. Moskaluk,[13,*] Robert F. Browning, Jr.,[8,9,*] and Matthew D. Wilkerson[1,14,*]

[1]The American Genome Center, Collaborative Health Initiative Research Program, Department of Anatomy, Physiology and Genetics, Uniformed Services University of the Health Sciences, Bethesda, MD 20814, USA
[2]Henry M. Jackson Foundation for Military Medicine, Bethesda, MD 20817, USA
[3]Women's Health Integrated Research Center, Inova Women's Service Line, Inova Fairfax Medical Campus, Falls Church, VA 22042, USA
[4]Gynecologic Cancer Center of Excellence, Murtha Cancer Center Research Program, Uniformed Services University of the Health Sciences, Bethesda, MD 20814, USA
[5]Chan Soon-Shiong Institute of Molecular Medicine at Windber, Windber, PA 15963, USA
[6]Center for Biomedical Informatics and Information Technology, National Cancer Institute, Rockville, MD 20850, USA
[7]Pulmonary and Mediastinal Pathology, Department of Defense, Joint Pathology Center, Silver Spring, MD 20910, USA
[8]Department of Medicine, Walter Reed National Military Medical Center, Bethesda, MD 20814, USA
[9]John P. Murtha Cancer Center Research Program, Uniformed Services University, Bethesda, MD 20814, USA
[10]Lawrence J. Ellison Institute for Transformative Medicine, Los Angeles, CA 90064, USA
[11]Departments of Medicine, Chemical Engineering and Material Sciences, University of Southern California, Los Angeles, CA 90007, USA
[12]Center for Applied Proteomics and Molecular Medicine, George Mason University, Manassas, VA 20110, USA
[13]Department of Pathology, University of Virginia, Charlottesville, VA 22908, USA
[14]Lead contact
*Correspondence: craig.shriver@usuhs.edu (C.D.S.), cam5p@virginia.edu (C.A.M.), robert.f.browning2.civ@health.mil (R.F.B.), matthew.wilkerson@usuhs.edu (M.D.W.)
https://doi.org/10.1016/j.xcrm.2022.100819

SUMMARY

We present a deep proteogenomic profiling study of 87 lung adenocarcinoma (LUAD) tumors from the United States, integrating whole-genome sequencing, transcriptome sequencing, proteomics and phosphoproteomics by mass spectrometry, and reverse-phase protein arrays. We identify three subtypes from somatic genome signature analysis, including a transition-high subtype enriched with never smokers, a transversion-high subtype enriched with current smokers, and a structurally altered subtype enriched with former smokers, *TP53* alterations, and genome-wide structural alterations. We show that within-tumor correlations of RNA and protein expression associate with tumor purity and immune cell profiles. We detect and independently validate expression signatures of RNA and protein that predict patient survival. Additionally, among co-measured genes, we found that protein expression is more often associated with patient survival than RNA. Finally, integrative analysis characterizes three expression subtypes with divergent mutations, proteomic regulatory networks, and therapeutic vulnerabilities. This proteogenomic characterization provides a foundation for molecularly informed medicine in LUAD.

INTRODUCTION

Lung adenocarcinoma (LUAD) is a leading cause of cancer deaths in the United States[1] despite advances in therapeutics targeting somatically altered genes and immune checkpoints. A major challenge in diagnosing and treating individuals with LUAD is the vast morphological and molecular heterogeneity within and among tumors.[2–4] Several national and international molecular profiling efforts have cataloged a diversity of somatic DNA alterations in LUAD, including driver gene mutations, copy-number alterations, and fusion genes,[5–7] as well as molecular subtypes defined by RNA expression.[6,8,9] The established RNA expression subtypes of LUAD (terminal respiratory unit [TRU], proximal proliferative [PP], and proximal inflammatory [PI]) have distinct clinical outcomes, therapeutic responses, and underlying mutations.[6,8] Despite these advances, it remains challenging to predict clinical outcomes for all individuals with LUAD based on clinical or molecular characteristics.[10] In addition, many LUAD tumors do not possess a molecular alteration currently indicated for targeted therapy.[3]

Table 1. Characteristics of cohort and proteogenomic data types

Patient and tumor features	Statistic	Summary
Year of sample collection	range	2012–2018
Race	white (%)	92
	black (%)	8
Gender	female (%)	48.3
Age at diagnosis	mean ± SD (years)	66.3 ± 10.2
Smoking history	current (%)	23
	former (%)	59.8
	never (%)	17.2
Pack years	mean ± SD (years)	40.4 ± 25.8
Stage	I (%)	52.9
	II (%)	27.6
	III (%)	17.2
	IV (%)	0
	N/A (%)	2.3
Tumor grade	1 (%)	0
	2 (%)	39.1
	3 (%)	60.9
Histological subtype	acinar (%)	33.3
	papillary (%)	18.4
	solid (%)	36.8
	others (%)	11.5
Distant metastasis	(%)	37.9
Lost to follow up before 3 years	(%)	4.5
Median follow up time (months)		50
Proteogenomic data		**Platform**
Germline whole-genome sequencing		Illumina[a] 39× mean coverage
Tumor whole-genome sequencing		Illumina[a] 116× mean coverage
RNA sequencing		Illumina[a,b] 14,374 transcripts
MS proteomics		Orbitrap MS[b] 7,614 proteins
MS phosphoproteomics		Orbitrap MS[b] 10,093 phosphosites
RPPA		Aushon 2470 arrayer[b] 307 antibodies

Summary statistics displayed for patients, tumors, and protogenomic platforms. There are no biological or technical replicates in this article's tables or figures. RNA-seq transcript count refers to protein-coding genes with minimal RNA expression, at least 2 transcripts per million. Some platforms have different subsets available by closed and open access. See also Table S1.

[a]Data repository availability is by closed access.

[b]Data repository availability is by open access.

Recent analyses employing shotgun mass spectrometry (MS)-based proteomics have elucidated new translational and post-translational layers of tumor biology across several tumor types that were not observable through genomics alone.[11] The joint characterization of tumor proteomics with genomics and transcriptomics enables proteogenomic analysis, which may enhance our understanding of the molecular mechanisms that drive tumor phenotypes, identify proteome-specific markers of outcome, and identify novel treatment paradigms. Initial proteogenomic studies in LUAD, including the NCI's Clinical Proteomic Tumor Analysis Consortium (CPTAC), have described the broad proteogenomic landscape of LUAD, including proteomic changes related to mutated genes and protein signaling networks.[12–16] However, these studies have also reported a wide range in correlation of relative abundances of RNA expression to protein expression across genes (correlation values: 0.14, 0.17, 0.28, 0.34, 0.53).[12–16] This range indicates that uncertainty remains in the relationship of protein and RNA levels in bulk LUAD tumors, which could be a consequence of differences in sample preparation, molecular platforms, or RNA and/or protein degradation across cohorts. Validation of gene-wise RNA and protein expression relationships in independent proteogenomic LUAD cohorts would reduce this uncertainty. To date, proteogenomic studies in LUAD have primarily included tumors from outside the United States with high rates of non smoking and EGFR mutation,[12–14] which is an important but incomplete segment of the disease. Additionally, clinical follow-up data in the published cohorts have been limited, and there are few independently validated proteomic markers of clinical outcome in LUAD,[17] compared with the large number available by RNA expression.[10]

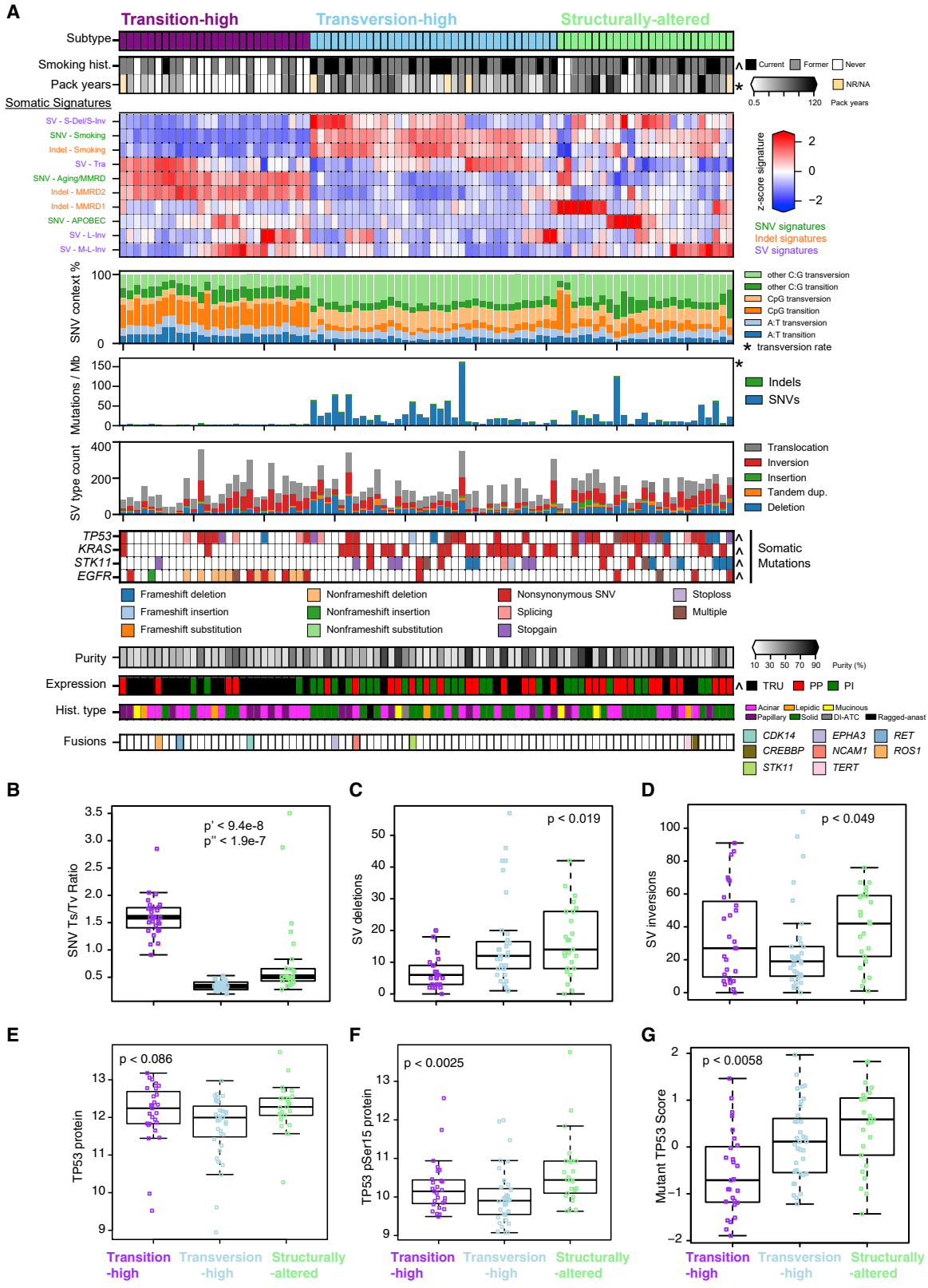
It is clinically important to characterize LUAD molecular etiology so that diagnostics and therapeutic interventions can be individualized. To address this aim, we, the Applied Proteogenomic OrganizationaL Learning and Outcomes (APOLLO) research network,[18,19] performed deep proteogenomic profiling of LUAD from a cohort of individuals in the United States unselected for tobacco use. These data were then comprehensively analyzed to identify LUAD's major proteogenomic alterations and subtypes, possible therapeutic vulnerabilities, and molecular discriminants of outcome.

RESULTS

Tumor collection and analysis strategy

Eighty-seven patients with LUAD were selected and acquired from the Lung Cancer Biospecimen Resource Network (https://lungbio.sites.virginia.edu/) with individual consent and institutional review board approval. LUAD samples were primary tumors that had been surgically resected for curative intent between 2012 and 2018. Of these, 80% were stage I or II, and 83% were from patients who smoked (Tables 1 and S1). Tumor histological subtypes were assigned by expert review of matched formalin-fixed, paraffin-embedded (FFPE) sections, revealing three main histologic subtypes: acinar, papillary, and solid (Figure S1). Tumor tissues were then analyzed by five molecular profiling assays: whole-genome sequencing (WGS), RNA sequencing (RNA-seq), MS-based proteomics

CellPress
OPEN ACCESS



(legend on next page)

and phosphoproteomics, and reverse-phase protein arrays (RPPAs) (Figure S2). Matched normal tissues were analyzed by DNA WGS. Our analysis strategy involved systematic interrogation of each assay platform to identify molecular alterations and subtypes. This was followed by integrated proteogenomic analyses to characterize subtypes, comparatively analyze RNA and protein expression, and identify molecular discriminants of patient survival.

Somatic genome signature subtypes link molecular etiologies with smoking histories

LUAD whole genomes displayed a wide range in tumor mutational burden (TMB) and structural variants (SVs) (TMB: 0.35–176 mutations per megabase; SV range: 14–245; Figure 1A). To identify common patterns among these somatic alterations, we applied a multi-modal correlated topic modeling framework[20] to jointly determine signatures from the frequencies of single-nucleotide variant (SNV) base changes in their tri-nucleotide contexts, short insertion or deletion (indel) compositions, sizes, and genomic contexts, as well as SV types and lengths (Figure S3). This analysis revealed three SNV, three indel, and four SV signatures, several of which are associated with known etiologies for specific mutational processes (Figures S3B–S3D). The three SNV signatures represent established substitution profiles associated with LUAD tumors:[5,6,21] an aging signature characterized by C>T mutations in the NCpG context, a smoking signature comprising C>A transversions, and an APOBEC cytidine deaminase activity signature comprising C>T and C>G mutations in TCN contexts (Figure S3B). Among indel signatures, one was similar to the COSMIC signatures ID5 and ID3, the latter of which is associated with tobacco smoking (Figure S3C). The other two indel signatures (MMRD1 and MMDR2) both resemble DNA replication/repair slippage and have thymine insertions at long homopolymers, with the MMRD1 signature also having cytosine and thymine deletions at long homopolymers (Figure S3C). The four SV signatures were distinguished by long (>10 Mb) inversions, short (1–10 kb) deletions and inversions, medium (100 kb–10 Mb) inversions, and high interchromosomal translocation frequencies (Figure S3D).

To determine if these somatic genome signatures might identify LUAD subtypes with coordinated mutational processes, we clustered tumors by their signature profiles and identified three signature subtypes (Figure 1A). We designated these subtypes as transition-high, transversion-high, and structurally altered. The transition-high subtype was defined by high aging SNV and MMRD2 indel signatures and had the greatest SNV transition/transversion ratio (Figure 1B), the most never smokers, the most tumors with acinar histology, and a very low TMB (median 2.3). The transversion-high subtype was defined by the greatest levels of the smoking SNV and indel signatures and had the greatest enrichment of current smokers and the highest TMB (median 20.7). The structurally altered subtype was defined by the MMRD1 indel and the medium-long inversion signatures. The structurally altered signature subtype had the greatest enrichment of former smokers, a high TMB (median 14.7), and intermediate levels of the smoking SNV and indel signatures. Looking further into tumor-wise SV burden, the structurally altered subtype also had the most structural deletions and inversions among these subtypes (Figures 1C and 1D). Genome-wide somatic copy-number alterations resembled published LUAD profiles[6,8] and did not associate with the signature subtypes (Figure S3E).

We then interrogated tumor whole genomes for significantly mutated genes (Figure S4A), revealing significant enrichments between the signature subtypes (Figure 1A). *EGFR* somatic mutations were enriched in the transition-high subtype (Fisher's exact $p < 0.05$), while *KRAS* and *STK11* somatic mutations were enriched in the transversion-high subtype ($p < 0.05$), similar to earlier studies.[6] In contrast, *TP53* somatic mutations were most frequent in the structurally altered subtype ($p < 0.05$), suggesting a causal relationship with this subtype's high structural deletion and inversion events. While *TP53* RNA and protein expression were unchanged among the subtypes (Figures S3F and 1E), the structurally altered subtype displayed the greatest TP53 pSer15 levels ($p < 0.0025$), which is a residue phosphorylated in response to DNA damage,[22] consistent with this subtype's high SV burden (Figure 1F). Concordantly, the structurally altered subtype exhibited the greatest expression of a mutant TP53 pan-cancer RNA signature[23] (Figure 1G). The pairing of *TP53* mutation and structural deletion elevation is consistent with observations from another recent LUAD cohort.[24]

Within the non-coding somatic genome, we detected recurrently mutated regulatory regions (Figure S4A), some of which were identified as somatic quantitative trait loci with *cis* genes. Among these was a regulatory element that associated with reduced RNA expression of the surfactants *SFTPD*, *SFTPA1*, and *SFTPA2* (Figure S4B). Clustering of somatic SV break points identified significantly recurrent events within the *STK11* gene locus (Figure S4C). These alterations did not associate with the somatic signature subtypes.

Characterization of RNA and protein correlations among tumors

We hypothesized that comparative analysis of protein expression versus RNA expression may reveal differential

Figure 1. Subtyping LUAD by whole-genome somatic signatures

(A) Clustering of LUAD tumors by somatic single-nucleotide variant (SNV), insertion or deletion (indel), and structural variant (SV) signatures. Columns are tumors, and rows are somatic signature values or patient/tumor features (n = 87). Patients and tumor features are tested for association with somatic signature subtypes: ANOVA (*$p < 0.05$); χ^2 test (^$p < 0.05$). Di-atc and ragged-anast refer to dispersed intra-alveolar tumor cells pattern and ragged-anastomosing glands pattern, respectively.

(B–G) Transition/transversion ratios, SV deletions, SV inversions, TP53 and phospho-TP53 expression from RPPA, and mutant *TP53* RNA expression scores compared across somatic signature subtypes. Boxplot lines indicates 25%, 50%, and 75% percentiles, and points are tumors with horizontal jitter added for visualization. p' refers to Wilcoxon rank-sum test on structurally altered versus transition subtype. p'' refers to Wilcoxon rank sum on structurally altered versus transversion subtype. p refers to Wilcoxon rank-sum test on structurally altered versus other subtypes.

See also Figures S1–S4 and Table S1.

CellPress
OPEN ACCESS

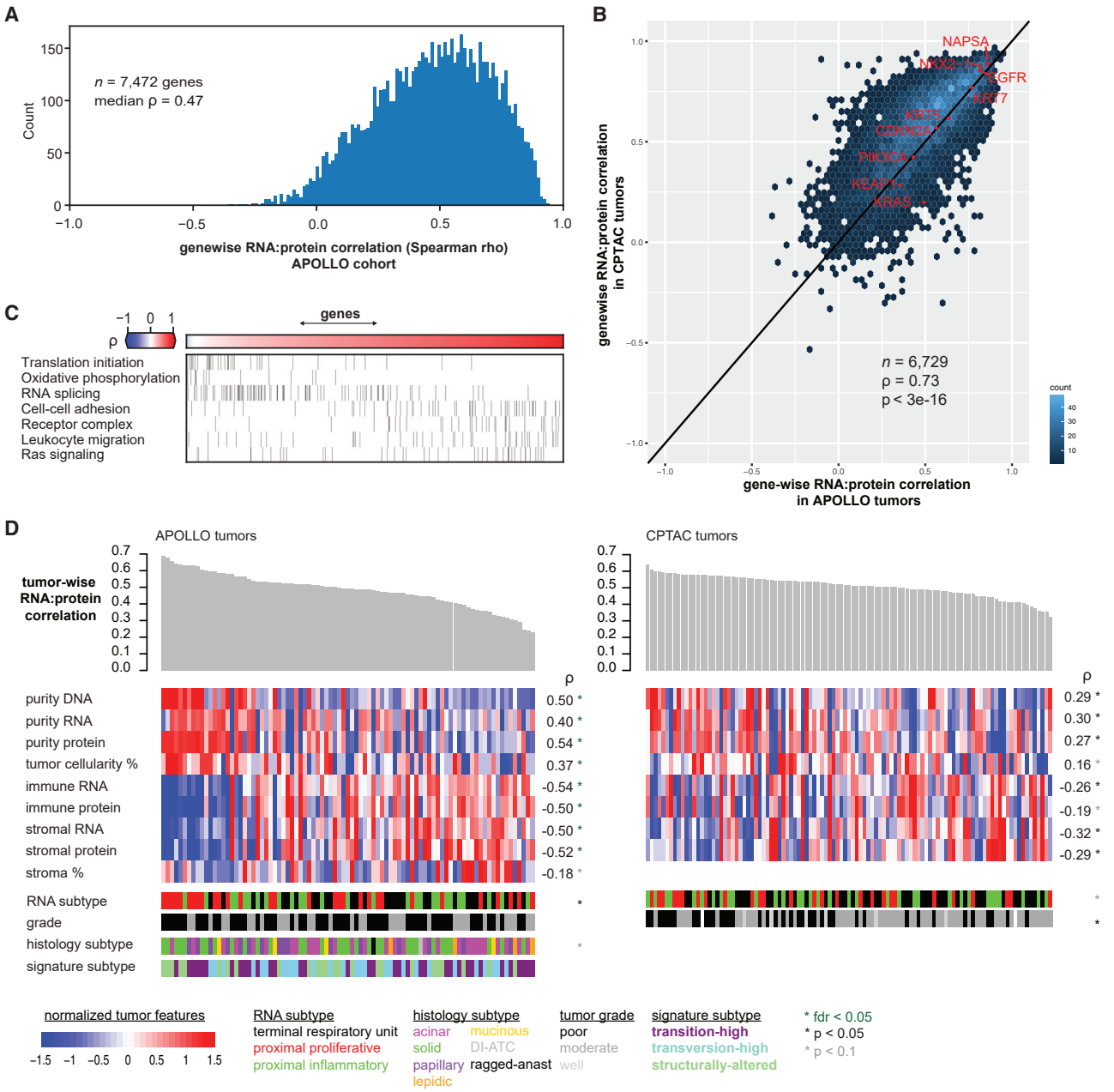


Figure 2. Correlation of gene-wise and tumor-wise RNA and protein expression

(A) Gene-wise RNA and protein expression correlations in the APOLLO cohort: 87 tumors and 7,472 co-detected genes.

(B) Gene-wise RNA and protein correlation comparison between APOLLO and CPTAC cohorts: 106 tumors, over 6,729 common, expressed genes between the cohorts. p refers to Spearman correlation test.

(C) Pathway enrichments according to gene-wise RNA and protein expression correlation in the APOLLO cohort.

(D) Tumor-wise RNA:protein expression correlation in APOLLO (n = 87) and in CPTAC cohorts (n = 105). Columns indicate individual tumors. Rows are molecular features except for manual slide review features of tumor cellularity percentage, stroma percentage, and grade. Tumor features tested for association with tumor-wise RNA:protein correlation by Spearman correlation tests for continuous variables and by Kruskal-Wallis tests for categorical variables.

See also Figure S5 and Table S1.

post-transcriptional regulation across genes and across LUAD tumors. To compare RNA expression and protein expression among all 7,472 co-detected genes, we calculated gene-wise RNA:protein correlations across tumors. The median gene-wise correlation was 0.47 with 84% of genes having statistically significant positive correlation (Figure 2A). The APOLLO cohort's median gene-wise correlation was very similar to the LUAD CPTAC cohort[13] but much larger than other recent studies in

LUAD with values of 0.14–0.34.[12–16] This range of correlations may be the result of different protocols used in earlier studies. We compared our gene-wise RNA:protein correlations with the CPTAC cohort and found a strong and significant positive correlation of gene-wise RNA:protein correlation values between independent LUAD cohorts (Spearman's correlation test, ρ = 0.73, p < 3e−16; Figure 2B). In addition, these correlations persisted throughout tertile strata of RNA and protein expression, indicating that these distributions are only modestly influenced by absolute abundances (ρ range 0.64–0.78, p << 0.001) (Figure S5). Markers of LUAD differentiation, including *NKX2-1*, *NAPSA*, and *KRT7*, were among genes with high gene-wise RNA:protein correlation in both cohorts. Different biological pathways were enriched across the range of gene-wise RNA to protein correlations in the APOLLO cohort, similar to other tumor types.[25,26] Highly correlated genes were enriched in cell adhesion and RAS signaling, and poorly correlated genes were enriched in translation initiation and oxidative phosphorylation (Figure 2C).

We then compared correlations between RNA and protein expression using all genes within individual LUAD tumors, called tumor-wise RNA:protein correlations. We found a range of tumor-wise RNA:protein correlations across the APOLLO cohort (ρ range 0.23–0.69), indicating substantial inter-tumor heterogeneity (Figure 2D). We then compared tumor-wise RNA:protein correlations with molecular properties of cellular heterogeneity and found a positive association with tumor purity estimated from DNA WGS (Spearman correlation test, ρ = 0.51, p < 8e−7) as well as with tumor cellularity estimates from histological review (ρ = 0.37, p < 4e−4). In contrast, tumor-wise RNA:protein correlations were negatively correlated with immune and stromal cell RNA expression scores (ρ = −0.54 and −0.50, respectively; p < 2e−6 on each) and immune and stromal cell protein expression scores (ρ = −0.50 and −0.52, respectively; p < 2e−6 on each). Tumors with the lowest tumor-wise RNA:protein correlations had elevated percentages of stroma from histological review in some cases, although they were not significant overall (p < 0.087).

To determine if these tumor-wise characteristics are generalizable in LUAD, we analyzed tumor-wise RNA:protein correlations in the CPTAC cohort (Figure 2D) by the same method and identified a similar range across tumors (ρ = 0.33–0.64; Table S2A) as in the APOLLO cohort. Again in the CPTAC cohort, tumor-wise RNA:protein correlation was positively correlated with tumor purity based on WGS somatic mutation signal (Spearman correlation test, ρ = 0.29, p < 0.0027). Tumor-wise RNA:protein correlations negatively correlated with RNA immune score, RNA stroma score, and protein stroma score (ρ = −0.26, −0.32, and −0.29, respectively, each p < 0.01) and trended significantly with protein immune score (ρ = −0.19, p = 0.056). In the CPTAC cohort, we also detected a significant association of poorly differentiated tumors with greater RNA:protein correlations (p < 0.002). Therefore, we discovered and validated that immune-enriched LUAD tumors have greater variability between their RNA and protein levels compared with highly pure tumors.

Transcript and protein determinants of patient survival

We identified genes with RNA expression or protein expression associated with patient overall survival (OS) and with metastasis-free survival (MFS) (Tables S2B and S2C). Focusing on genes with RNA and protein co-expression, we identified genes with RNA or protein nominally associated with OS (9 RNAs and 16 proteins, Wald test p < 0.001), but these were not significant after multiple testing correction (false discovery rate [FDR] > 0.25). With MFS, we identified significant associations for 560 "survival proteins" and 155 "survival RNAs" (Wald test, FDR < 0.25). Between all co-expressed proteins and RNAs, MFS hazard ratios were significantly correlated (Figure 3A; Spearman's correlation test ρ = 0.48, p < 3e−16). This correlation of hazard ratios was larger than a similar analysis performed in prostate cancer (ρ = 0.25).[27]

Combining survival proteins and their corresponding log hazard ratios as weights into a protein survival signature (and similarly for an RNA survival signature), we found that the aggregate expression of these proteins or these RNAs strongly discriminated patients by OS and MFS (Figure 3B; OS: protein signature Wald test, p < 1.8e−5, RNA signature p < 5.9e−5; MFS: protein signature p < 2.9e−10; RNA signature, p < 8.5e−9). Additionally restricting to 66 survival RNA-proteins, protein, and RNA signatures significantly predicted survival (OS: protein signature, p < 7.2e−5; RNA signature, p < 7.3e−5; MFS: protein signature, p < 7.5e−9; RNA signature, p < 1.2e−8; Table S2). All expression signatures remained significantly associated after controlling for additional covariates of tumor stage, histological subtype, gender, and adjuvant treatment (Figure S6A; Wald test, p < 0.05). We then compared gene-wise RNA:protein correlations among survival gene sets and found striking, significant differences (Kruskal-Wallis test, p < 7e−13; Figure 3C). Survival RNA proteins had the greatest RNA:protein correlations, followed by survival RNAs and then survival proteins.

To validate these signatures, we then applied the same survival signatures to the RNA and proteomics datasets of the CPTAC cohort, which had a median follow-up time of 15.6 months (Figure 3D). Our survival protein signature was significantly associated with patient OS (Wald test, p < 0.033) and trended significantly with patient MFS (p < 0.057) (Table S2A). The survival protein signature remained significantly associated with OS when including tumor stage as a covariate (p < 0.021). The survival RNA signatures significantly associated with patient OS (p < 0.0078) and MFS (p < 0.0043). Both associations remained significant when including tumor stage as a covariate (p < 0.05). Restricting to the survival RNA proteins, both RNA and protein signatures associated with OS (protein p < 0.011, RNA p < 0.0045) and MFS (protein p < 0.016, RNA p < 0.0022), all of which remain significant with tumor stage as a covariate (p < 0.05). Validating the APOLLO cohort, survival gene sets had significantly different gene-wise RNA:protein correlation trends in the CPTAC cohort, with survival RNA-proteins genes having the greatest RNA:protein correlation, followed by survival RNAs, and then survival proteins (Kruskal-Wallis test, p < 2e−12; Figure 3E). Therefore, OS and MFS in patients with LUAD can be predicted by signatures of protein expression and by RNA expression across independent cohorts.

By phosphoprotein expression, we identified a smaller number of survival phosphoproteins associated with MFS (MS-based proteomics n = 96 and RPPA n = 28). Seventy-nine of the MS-based survival phosphoproteins were positively

CellPress
OPEN ACCESS

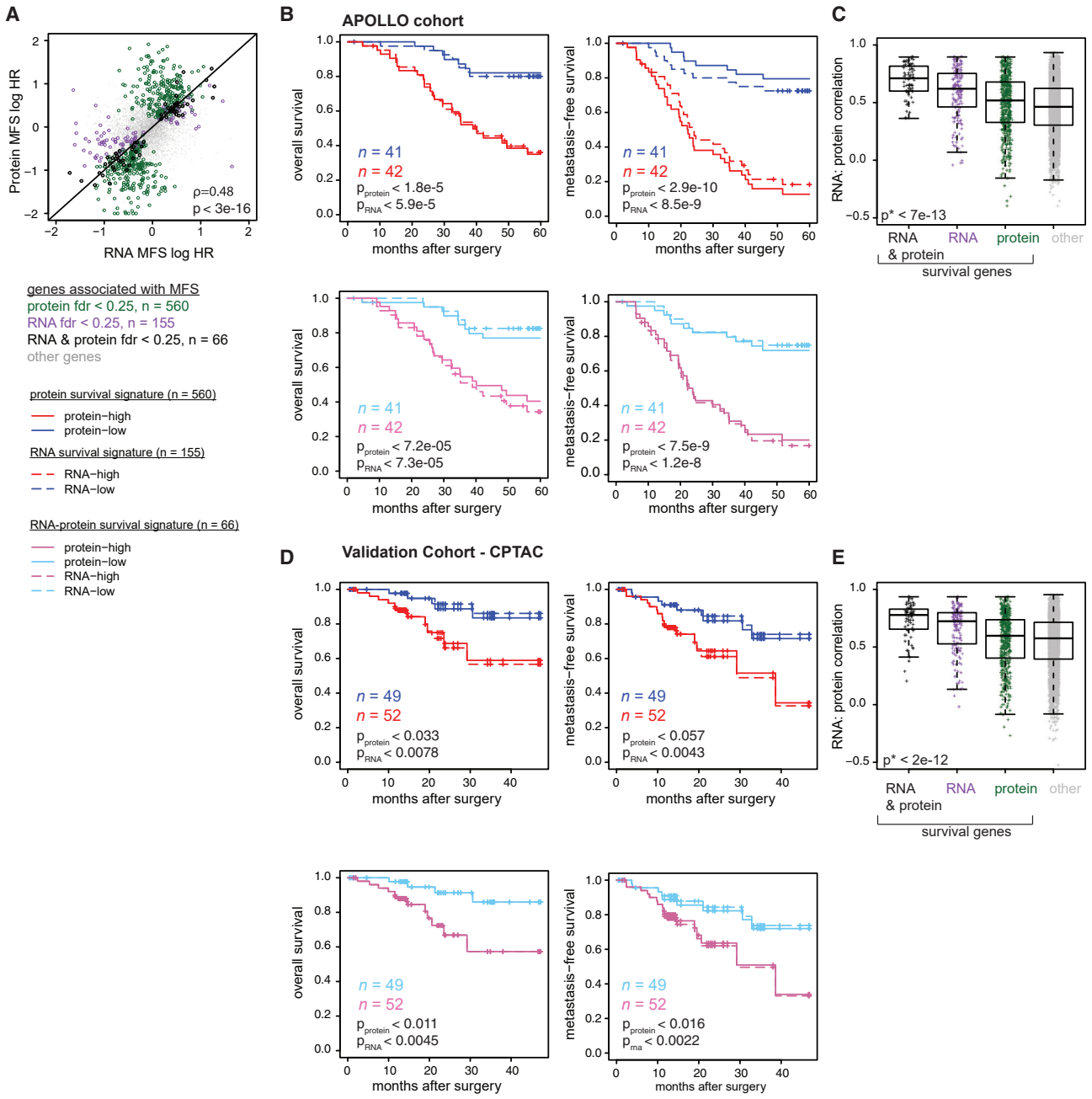


Figure 3. RNA and protein expression determinants of patient survival

(A) Comparison of log hazard ratios between RNA expression and protein expression on matched genes in APOLLO cohort (n = 87). Points outside the axis scale (less than 2 or greater than 2) are plotted as −2 and 2, respectively. ρ and p refer to the Spearman rank correlation coefficient and p value, respectively. "Other" refers to genes not associated with survival.

(B) Overall and metastasis-free survival in APOLLO cohort (n = 83 with follow up), with high and low referring to a 50th percentile split on the respective signature score. p refers to Cox proportional hazards Wald test of the continuous signature score. Top panels are signatures based on survival proteins and survival RNAs, and bottom panels are signatures based on survival RNA proteins.

(C and E) Gene-wise RNA:protein correlation across survival gene sets compared by Kruskal-Wallis tests, p*.

(D) CPTAC cohort survival following same layout as (B). 50th percentile split for visualization was based on entire cohort, n = 106, and plotted for those with follow up, n = 101.

See also Figure S6 and Table S2.

correlated with corresponding MS-based global proteins (Spearman correlation test, $p < 0.05$).

Proteogenomic subtyping of LUAD

To subtype LUADs by genome-wide expression, we first applied the RNA expression subtype predictor[8] to assign tumors to the TRU, PI, and PP subtypes.[6,8] Applying the predictor to RNA expression or global protein expression resulted in highly similar subtype assignments (Fisher's exact test, $p < 8e{-}19$) (Figure S6B). We also performed unsupervised clustering on the cohort's RNA and global protein expression, which also resulted in significantly associated subtype assignments to the RNA subtype predictions ($p < 1.8e\text{-}8$ in both cases) (Figures S6C and S6D). Furthermore, multi-omic clustering on joint RNA and protein expression also revealed significantly associated subtype-assignments (Figure S6E). Therefore, the LUAD tumor subtypes (TRU, PI, and PP) are a robust stratification across both RNA expression and protein expression. Unsupervised phosphoproteomic cluster assignments associated with RNA subtype predictions, primarily for PP and TRU. To standardize with prior studies,[6,8,28,29] we utilized expression subtype assignments based on the RNA subtype predictor applied to RNA-seq throughout the current study.

The expression subtypes were enriched with distinct histological subtypes—TRU with acinar and PI with solid, corroborating earlier cohorts[6,8] (Figure 4A). The expression subtypes overexpressed their canonical marker genes[8] by RNA expression and by protein expression. For example, the TRU subtype overexpressed surfactant protein C (*SFTPC*) and thyroid transcription factor 1 (*NKX2-1*; also known as *TTF1*). The PP subtype overexpressed thymine DNA glycosylase (*TDG*) and glutathione peroxidase 2 (*GPX2*). The PI subtype overexpressed the immune cell markers cluster of differentiation 163 (*CD163*) and vascular cell adhesion protein 1 (*VCAM1*). Proteogenomic pathway analysis simultaneously integrating RNA, protein, and phosphoprotein expression data identified distinct overexpressed pathways among the expression subtypes (Figure 4B). The TRU subtype overexpressed protein secretion and developmental pathways of adipogenesis and myogenesis. The PI overexpressed inflammatory and interferon-γ signaling pathways. The "immune-cold" PP subtype overexpressed proliferation-related pathways. With few exceptions, the subtype-specific signals were consistent across RNA, protein, and phosphoprotein expression, indicating that distinct transcriptional process of the subtypes carry through translation and post-translational regulation.

Three somatically mutated driver genes were associated with the subtypes, *STK11* and *KEAP1* mutations in PP, and *EGFR* mutations in TRU (Figure 4A). By transcript and protein expression, *EGFR*, *KEAP1*, and *STK11* (via STK11 pSer30) were also over- or underexpressed in their respective subtype (Kruskal-Wallis on enriched subtype versus others, $p < 0.05$). Genome-wide TMB was greatest in PI compared with other subtypes (medians: PI: 27.8, PP: 13.3, TRU: 3.2, Kruskal-Wallis $p < 6e{-}6$) (Figure 4A). The proportions of high TMB tumors defined by a clinically used threshold of 10 mutations per Mb[30] were also significantly different among the subtypes (PI: 78%, PP: 67%, TRU: 23%, Fisher's exact $p < 5e{-}5$). By immune cell scores, the subtypes showed significant differences ($p < 5e{-}7$) with PI and TRU leading PP, validating an earlier report[28] (Figure 4A). Tumor purity was greatest in PP ($p < 0.001$). Additionally, we discovered that the subtypes display different tumor-wise RNA:protein correlations led by PP (PP: 0.52, PI: 0.49, TRU: 0.47, Kruskal-Wallis $p < 0.05$) (Figures 2 and 4A). High tumor purity and tumor-wise RNA:protein correlation in PP may be partially explained by reduced immune cell presence in these tumors.

Comparing tumor subtypes by patient survival outcomes, we found that the expression subtypes and histological subtypes significantly associated with MFS (Figure 4C), while somatic genome signature subtypes did not associate with OS or MFS (Figure S6F). Therefore, the intrinsic biology captured in LUAD expression subtypes are a determinant of MFS, expanding on prior reports on OS.[6,8,10] Forty-five of the 66 survival RNA proteins were also differentially expressed among the subtypes, indicating shared underlying biology.

Integrative network modeling of LUAD subtypes

To identify potential therapeutic vulnerabilities of the RNA subtypes, we used integrative network modeling to describe subtype-specific proteogenomic signals in the context of known molecular associations. First, we inferred kinase activities across tumors using phosphoproteomic expression and known kinase-substrate interactions, which were then used to identify kinase activity enrichment scores for each subtype (Figure 5A). Re-examination of these enrichments with phosphoresidue abundances corrected for global protein expression distinguished direct kinase activity changes from enrichments partially explained by changes in substrate availability (Figure S7A). To infer subtype-specific transcription factor (TF) activities, we identified TF motif matches in known LUAD regulatory elements[31] and compared these against proximal (−50/+10 kb) gene expression levels corrected for *cis* copy-number alterations (Figure S7B). These protein kinase and TF enrichments were then integrated with mutated and copy-number-altered genes, phosphorylation sites, global proteins, and enriched pathways into subtype-specific network models via known kinase-substrate, protein-protein, and protein-pathway interactions.

The PI network is characterized by molecular interactions that drive interferon γ (IFN-γ) signaling and inflammation (Figure 5B). Activated protein kinase C delta (PRKCD) downstream of the IFN-γ receptor phosphorylates a variety of targets in PI, including S727 of STAT1, which is necessary for STAT1's transcriptional activity.[32] Increased STAT1 transcription, global protein expression, and TF activity further supported its activation in PI tumors. STAT1 drives both immunosurveillance, consistent with observed increases in HLA protein and *CIITA* RNA expression, and immunosuppression, indicated by enhanced RNA expression of the immune inhibitory receptor *CD274* (PD-L1) (Figure 6). The PI subtype also significantly overexpresses the PD-L1 protein, detected using two widely used research based anti-PDL1 antibody clones (E1L3N and CAL10) and the SP-142 clone, which is used as an FDA-approved companion diagnostic for atezolizumab (Figure 6). All three anti-PDL1 antibodies have undergone extensive validation testing for clinical-tissue-based analysis of PD-L1 levels.[33,34] Increases in IFN regulatory factor (IRF) TF activities were additionally supported by enhanced

CellPress
OPEN ACCESS

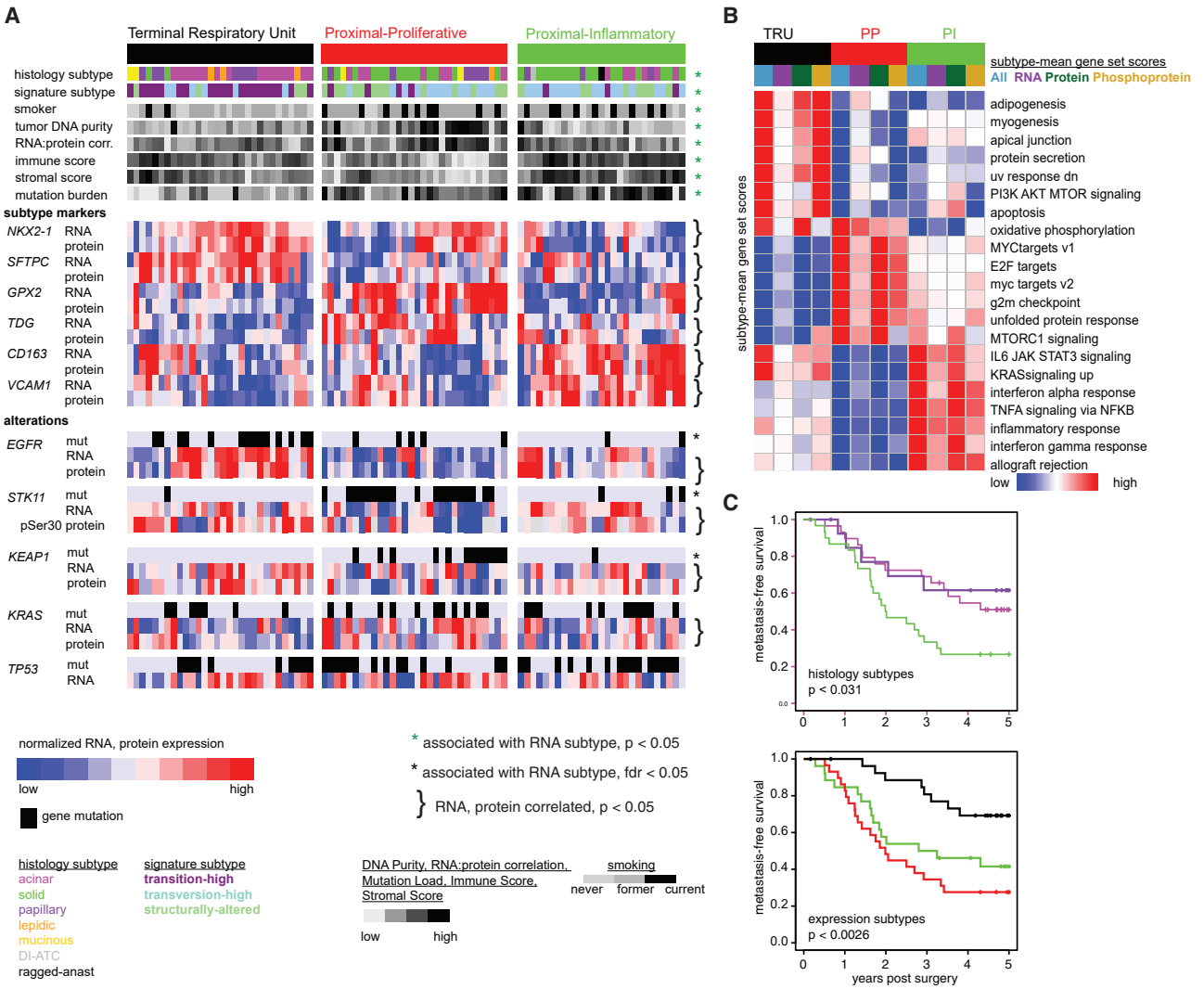


Figure 4. Molecular subtype characteristics and survival outcomes

(A) RNA expression subtypes. Tumors (n = 87) appear in columns and clinical and genomic features in rows. Protein refers to MS proteomics expression. RNA refers to RNA-seq expression. mut refers to non-silent gene mutations. Immune and stromal scores refer to RNA-based ESTIMATE scores. Continuous features analyzed by Kruskal-Wallis tests. Categorical features analyzed by Fisher's exact tests. RNA and protein expression compared by Spearman correlation tests.

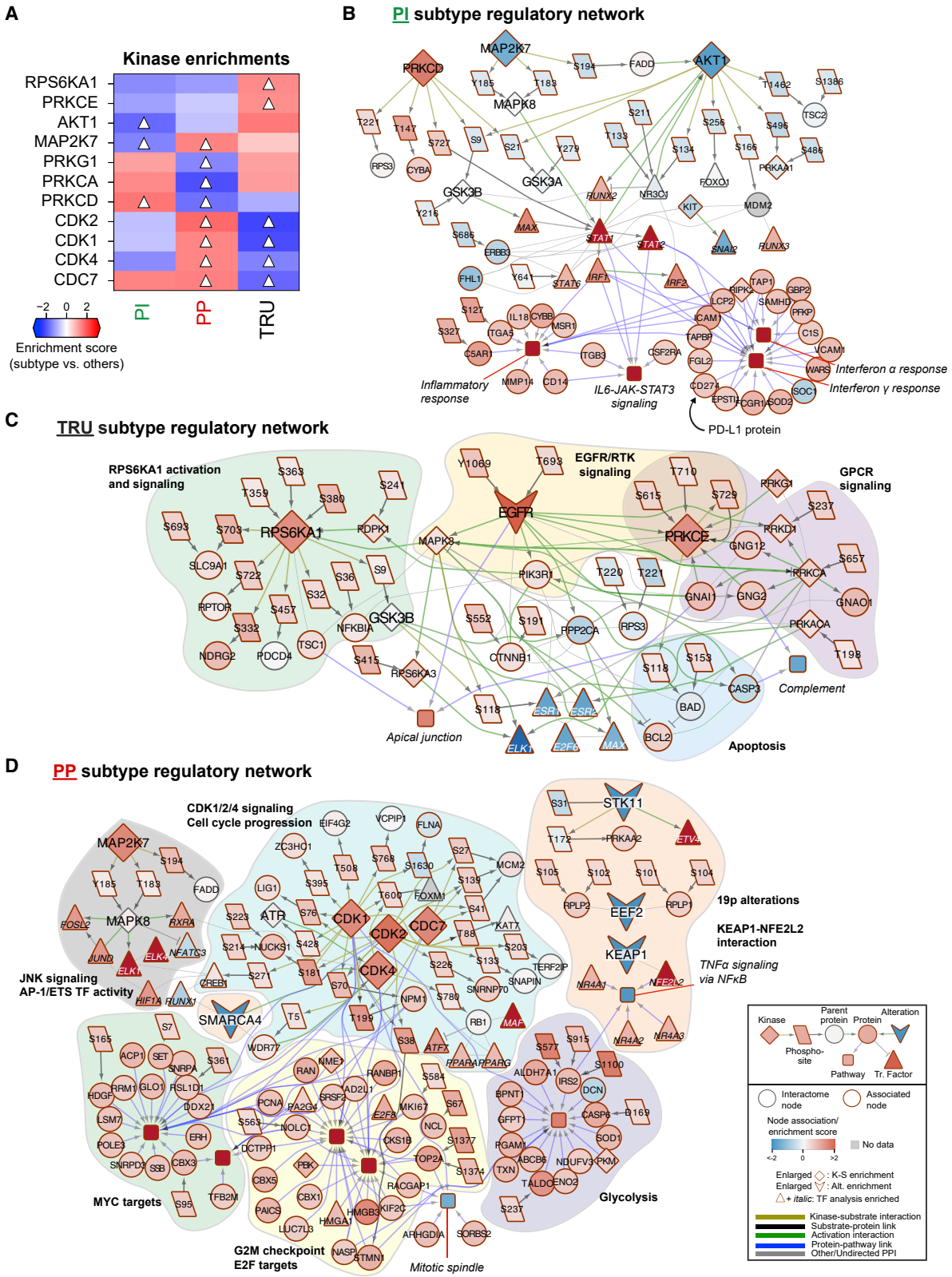
(B) Proteogenomic expression analysis of RNA expression subtypes. Columns indicate molecular enrichment (RNA, protein, phosphoprotein by subtype), and rows indicate gene sets. Phosphoprotein expression is from combined RPPA and MS platforms.

(C) Survival outcomes of RNA expression subtypes and histological subtypes, analyzed by log rank tests (p).

See also Figure S6.

IFN-γ and inflammatory signaling in PI. *CTLA4* transcript expression, which predominantly derives from $CD4^+/CD8^+$ T cells and regulatory T cells in lung tumors,[35] was also increased in PI tumors (Figure 6). Given elevated immune infiltration, high TMB, and enhanced IFN-γ signaling, coupled with increased PD-L1 protein and *CTLA4* RNA expression, the PI subtype may encapsulate the subset of tumors most likely to respond to immune checkpoint inhibitors.

TRU tumors are broadly characterized by overactive EGFR signaling (Figures 2 and 5). Our network captures activation of ERK (MAPK3 or ERK1) and PI3K-PDK1 (PIK3R1 or p85 and PDPK1) by EGFR as well as downstream activation of RPS6KA1 (aka RSK1 or P90RSK1), which promotes cell proliferation and inhibition of apoptosis.[36,37] Enhanced RPS6KA1 kinase activity in TRU tumors is supported in our data by increased phosphorylation of its own residues, including S380, and by several substrate sites, including GSK3B S9, RPTOR S722, NFKBIA (IκBα) S32, and PDCD4 S457. PRKCE activity is also enhanced in TRU. PRKCE has been classified as an oncoprotein due to its anti-apoptotic cellular functions,[38] including inhibitory S118 phosphorylation of pro-apoptotic BAD (Figures 5C and 6), which is also elevated in TRU. Additionally, increased global protein expression and phosphorylation of several G protein-coupled receptor (GPCR) molecules, PKA (PRKACA global and



(legend on next page)

T198), and PKCα (PRKCA global and S657) were associated with this subtype. Taken together, the TRU subtype captures tumors with marked growth factor signaling that may be most responsive to EGFR inhibitors[39] and to compounds directed at other members of these signaling cascades (e.g., PRKCE[40] or RPS6KA1[37]).

The PP subtype network is characterized by enhanced cell-cycle and glycolytic biological processes in an immune-cold microenvironment[41] (Figures 2, 4, 5D, and 6). Pronounced cyclin-dependent kinase (CDK) activities (CDKs 1, 2, and 4 and CDC7) were implicated by enhanced phosphorylation of several target residues, including S780 on RB1 by CDK4, which disrupts inhibition of E2F TFs that drive cell-cycle progression[42] (Figure 6). Several of these regulatory states were not implicated by global abundance changes of their respective parent proteins. MAP2K7 kinase activity was also significantly enhanced in PP, reflected by increased phosphorylation at T183/Y185 of MAPK8 and S194 of FADD, the latter of which is associated with G2/M cell-cycle regulation[43] and poor prognosis in LUAD.[44] Additionally, global protein and phosphoprotein expression of several proliferation markers were increased in PP, including TOP2A, MKI67, IRS2, and HDGF. Somatic alterations in *STK11* and/or *KEAP1* encompass 26 of 30 PP tumors, while *EEF2* and *SMARCA4* alterations are also significantly associated with PP. Inactivation of SMARCA4 can be synthetic lethal with CDK4;[45] thus, SMARCA4-altered PP tumors with high CDK4 activity may be responsive to CDK4/6 inhibitor therapies. Metabolic reprogramming was also indicated in PP by upregulation of proteins involved in glycolysis and glutaminolysis, which is consistent with cellular responses to *STK11* loss mediated by HIF-1α in conjunction with enhanced cellular stress and reactive oxygen species (ROS).[46] Indeed, coincident *STK11-KEAP1* alterations were frequently observed in PP tumors (Figure 6) as KEAP1 inactivation promotes NFE2L2 activity and antioxidant gene expression[47] (Figures S7C–S7F). While the majority of PP tumors do not possess targetable oncogene mutations, recent studies have demonstrated therapeutic vulnerabilities aimed at PP metabolism, including glutaminase inhibition in *STK11-KEAP1-KRAS* mutants[48] and stearoyl-coenzyme A (CoA) desaturase (SCD) inhibition, which is upregulated in PP tumors along with other ferroptosis-protective molecules (Figure 6), in combination with ferroptosis inducers in *STK11-KEAP1* co-mutants regardless of *KRAS* status.[49]

DISCUSSION

Here, we report a large-scale proteogenomic characterization of LUAD from a United States population unselected for tobacco use. Using six molecular profiling technologies, we measured four layers of LUAD biology: genome, transcriptome, proteome, and phosphoproteome. Our systematic analysis of these data identified tumor subtypes, alterations, signaling patterns, and markers of survival. The detail and comprehensive prospective longitudinal clinical data for this United States cohort is distinctive among the published cohorts for LUAD proteogenomic studies. This is in contrast to the most recent major proteogenomic research in LUAD by Gillette et al. through the NCI's CPTAC, which had less longitudinal clinical follow up and represented LUAD tumors from around the world, with less than a third of the tumors from the United States. Both studies offer complementary insights into LUAD, but our study leverages the United States' standard of care and treatment protocols as well as common exposures for a stronger survival analysis and ultimate translation to medical practice. Although normal adjacent tissue matched with tumor specimens were not part of our protocol, as was included in CPTAC LUAD, we did comprehensive correlative analysis with their data to successfully validate our work.

By somatic genome signature analysis, we identified three subtypes with coordinated molecular etiologies and tobacco use. The transition-high and transversion-high signature subtypes represent never and current smokers and correspond to tumor groups described in prior LUAD cohorts.[5–7,21] The structurally altered subtype, however, reveals a bifurcation of smokers (former versus current) by a distinct pathway of LUAD mutagenesis, structural genome disorganization, and *TP53* alterations. Potentially in the structurally altered subtype, tobacco mutagens produced a moderate transversion signature and a *TP53* mutation, individuals quit smoking, and over time, structural alterations accumulated due to inhibition of DNA repair checkpoints by the TP53-null phenotype to produce tumors despite the lack of continued direct DNA damage by tobacco mutagens. In transversion-high tumors, the continued smoke exposure may have produced a stronger transversion signature and additional sequence mutations such as in *KRAS*, which then led to the development of the tumors. The structurally altered subtype may have broad implications in the spectrum of challenges with LUAD from identifying potential prevention pathway targets for these former smokers, broadening screening to include a wider range of former smokers, and guiding earlier and more aggressive therapies such as *TP53*-directed T cell-based therapy.[50]

Our comparative expression analysis of RNA and protein revealed two interesting LUAD characteristics. First, we found that gene-wise RNA:protein expression correlation is highly

Figure 5. Proteogenomic network characterization of subtypes

(A) Kinase enrichments based on known kinase-substrate links to measured MS-based proteomics and RPPA phosphoresidues from APOLLO cohort (n = 87). Triangles indicate significant kinase enrichments in either PI, TRU, or PP subtypes (combined FDR < 0.01).

(B–D) Regulatory networks of subtypes. Box to right indicates network layout (top) and node/edge shape, size, and color schemes (bottom): node shapes indicate molecule types or pathways; red outlines identify nodes significantly associated with the subtype (gray otherwise); blue-to-red shading indicates node association/enrichment with subtype (gray denotes no measured data); enlarged diamonds and "vee" shapes indicate enriched kinases and mutated genes, respectively; red outlined triangles with italic text labels indicate TFs identified from TF enrichment analysis; and edge color represents types of protein-protein or protein-pathway links. Values and data sources for each network node are listed in Table S3.

See also Figure S7.

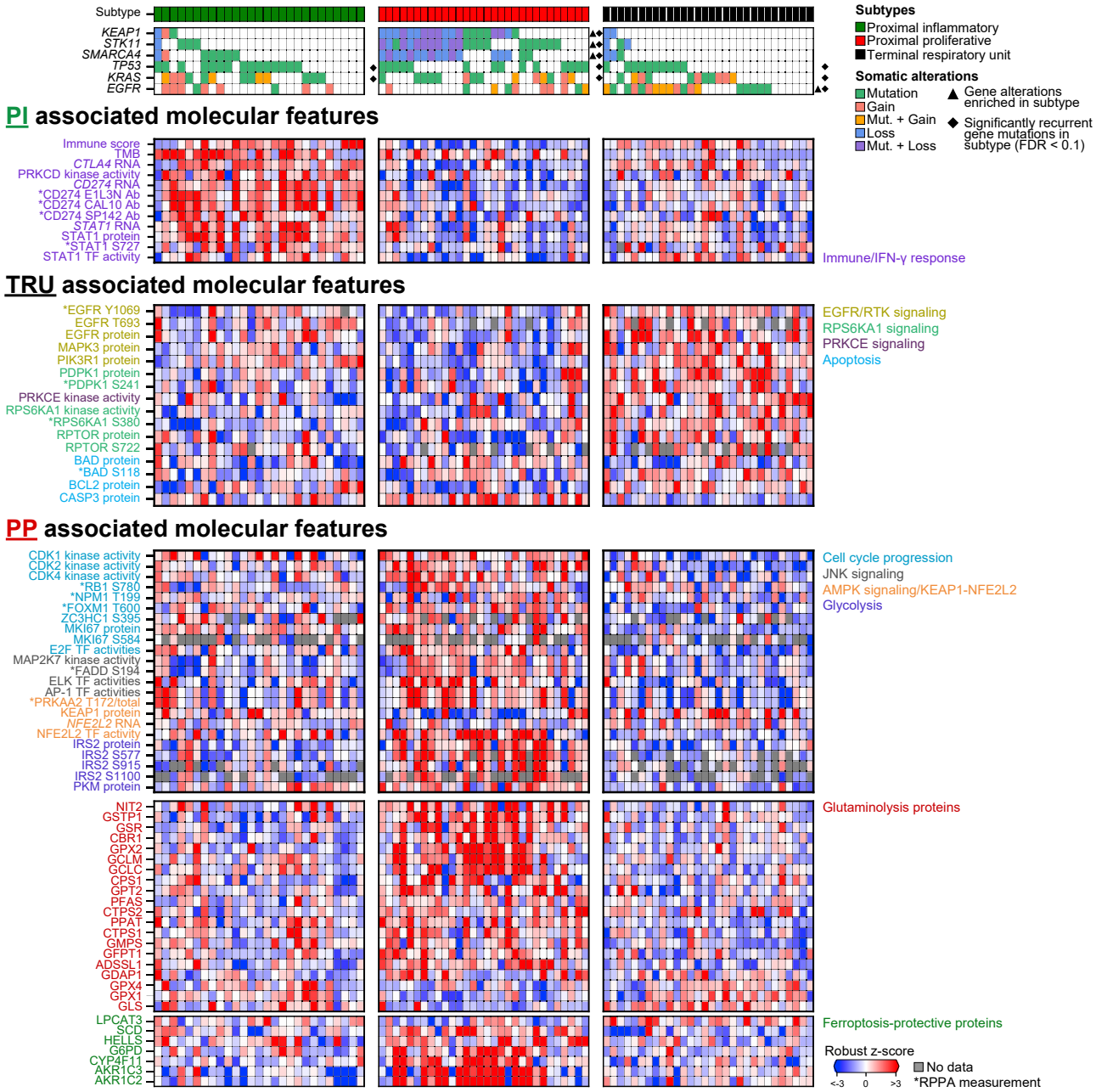


Figure 6. Proteogenomic features associated with subtype networks

Individual features associated with LUAD subtypes and networks (related to Figure 5). Sample-wise somatic alterations in *KEAP1, STK11, SMARCA4*, *TP53*, *KRAS*, and *EGFR* with black triangles to the right indicating significant enrichment of molecular alterations in the given subtype (Fisher's exact test $p < 0.05$) and black diamonds indicating significantly recurrent somatic mutations in the subtype (MutEnricher FDR < 0.1). Additional panels display select individual molecular features associated with the subtypes (see STAR Methods for molecular type statistics). Asterisks indicate feature measurement from the RPPA platform.

consistent across two independent LUAD cohorts. This finding offers expanded capabilities for biomarker development. Second, we found, and validated in an independent cohort, that tumor-wise correlation of RNA and protein expression varies according to immune cellular heterogeneity in LUAD. LUADs with high tumor-wise RNA:protein correlation have a high grade of differentiation similar to clear cell renal carcinoma,[25] have high purity, and have low immune cell content and tend to be in the PP subtype. In contrast, LUADs with a low tumor-wise RNA:protein correlation have low grade, low purity, and high immune cell content. Increased immune cell heterogeneity may present a greater diversity of pathway expression and

post-transcriptional regulation reducing RNA:protein correlations in the tumor.

The APOLLO cohort's long clinical follow-up data and comprehensive proteogenomic data provided an opportunity to identify biomarkers for LUAD clinical outcome. Among co-measured genes, we found that survival proteins outnumber survival RNAs and that survival markers significant at both RNA and protein levels were more correlated with one another than unique survival transcripts or proteins. Supporting this phenomenon in a different tumor type, a recent report in gastric carcinoma found greater gene-wise RNA:protein correlation in RNA survival genes than in non-survival genes.[26] RNA and protein expression signatures based on these gene sets significantly predicted survival outcomes in both this APOLLO cohort and an independent CPTAC cohort. Protein expression markers may have enhanced performance in clinically available FFPE specimens and small-volume biopsies.[51] This understanding of the correlation of RNA and protein expression with survival has significant implications in individual patient prognosis with LUAD as well as the development of future prognostic testing for LUAD.

Our proteogenomic analysis identified potential vulnerability targets for the LUAD RNA subtypes, providing a roadmap for future studies. The PI subtype had a concentration of immune features such as IFN-γ signaling, PD-L1 expression, and high TMB. The current reliance on a single analyte such as PD-L1 expression[52] or TMB[30] as a biomarker for LUAD's responsiveness to immunotherapy may be enhanced using the PI subtype to offer added predicted capacity. TRU harbors enhanced EGFR signaling and kinase activity from PRKCE and RPS6KA1, which are potential therapeutic targets. The PP subtype's proteogenomic profile suggested that CDK inhibitors and glutaminase inhibitors may be beneficial. Taken together, our comprehensive results may lead to advances in LUAD precision medicine, such as molecularly informed clinical trials or improved molecular diagnostics.

Early diagnosis for LUAD is limited to radiographic CT screening and biopsy when tumors are large enough (6–8 mm or greater). The comprehensive proteogenomic understanding and signatures described in this study may aid early LUAD diagnosis with small samples or liquid biopsies to guide precision treatments. Already, the established LUAD subtypes offer prognostic value; however, prognosis must be coupled with a treatment to improve patient outcomes. Prospective clinical studies that test our understanding of these subtypes is the next step. Clinical trials are needed with PI tumors and checkpoint inhibitors as well as with PP tumors, which include KRAS-negative tumors that have been refractory to targeted therapy until recently. We already have precedent for this concept with the TRU subtype of LUAD, which highly correlates with the EGFR-positive tumors that clinically respond well to *EGFR* inhibitors. Being able to identify LUAD and the subtypes with small amounts of proteins or RNA will be a leap forward in our fight against lung cancer. As prevention is the ultimate goal for all cancers, understanding the different molecular subtypes of LUAD combined with contributing environmental and clinical factors may define specific pathways that could be therapeutically targeted to halt the development of malignancy in pre-cancerous and early-stage LUAD tumors.

Limitations of the study

This study did not measure normal adjacent lung tissue, so we were not able to compare RNA or protein expression between tumors and normal lung tissue. Adjuvant and neo-adjuvant treatment data were limited in this cohort, so we were not able to assess patient response to the predicted therapeutic vulnerabilities described in the article. Future proteogenomic studies of larger cohorts with treatment response may address these limitations.

CONSORTIA

The members of the APOLLO Research Network for this study include Rebecca Blackwell, Gauthaman Sukumar, Dagmar Bacikova, Camille Alba, Elisa McGrath, Sraavya Polisetti, Meila Tuck, Alden Chiu, Gabe Peterson, Caroline Larson, Leonid Kvecher, Brenda Deyarmin, Jennifer Kane, Katie Miller, Kelly A. Conrads, Brian L. Hood, Sasha C. Makohon-Moore, Tamara S. Abulez, Elisa Baldelli, Mariaelena Pierobon, Qing-rong Chen, Henry Rodriguez, Sean E. Hanlon, Anthony R. Soltis, Nicholas W. Bateman, Jianfang Liu, Trinh Nguyen, Teri J. Franks, Xijun Zhang, Clifton L. Dalgard, Coralie Viollet, Stella Somiari, Chunhua Yan, Karen Zeman, William J. Skinner, Jerry S. H. Lee, Harvey B. Pollard, Clesson Turner, Emanuel F. Petricoin, Daoud Meerzaman, Thomas P. Conrads, Hai Hu, Craig D. Shriver, Christopher A. Moskaluk, Robert F. Browning, Jr., and Matthew D. Wilkerson.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Specimens and clinical data
  - CPTAC cohort
- METHOD DETAILS
  - Tumor histological evaluation
  - Tissue sectioning and utilization
  - DNA sample handling and library preparation
  - DNA library clustering and whole genome sequencing
  - RNA library preparation and sequencing
  - Proteomics specimen preparation
  - Mass spectrometry-based proteomics
  - Reverse phase protein array (RPPA)
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Quantitative data processing pipeline for MS global and phosphoproteome analyses
  - MS variant peptide identification from patient-specific proteogenomic databases
  - Differential global and phosphoproteomics
  - Germline and somatic variant calling and sample concordance

- Germline variant pathogenicity analysis
- Somatic variant filtration and annotation
- DNA copy number segments and gene-level copy number values
- RNA-seq alignment, quantification, and differential expression analysis
- Sample identity matching
- RNA-protein correlations
- Somatic mutation recurrency analysis
- Somatic signatures analysis
- Somatic quantitative trait loci (QTL) analysis
- Tumor expression subtyping
- TP53 mutation signature
- Cell type estimation
- Gene ontology and other pathway enrichment analyses
- Multi-omics gene set analysis
- Survival analysis
- Phosphoprotein and kinase enrichment analysis
- Transcription factor enrichment analysis
- Integrative network modeling

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2022.100819.

ACKNOWLEDGMENTS

We acknowledge the LCBRN and anonymous patients. We thank Sofia Echelmeyer for the graphical abstract artwork. Funding for this study was provided in part by USUHS #HU00012120002 to C.D.S., NIH/NHLBI IAA-A-HL-14-001 to H.B.P., and DOD/CDMRP #W81XWH-10-1-081 to C.A.M. The views expressed in this manuscript are solely of the authors and do not reflect the official policy of the Departments of Army/Navy/Air Force, Department of Defense, USUHS, HJF, or the United States government.

AUTHOR CONTRIBUTIONS

M.D.W., C.A.M., R.F.B., and C.D.S. led this APOLLO study. M.D.W. coordinated overall data analysis. C.A.M., J.L., H.H., W.J.S., M.D.W., K.Z., and R.F.B. performed clinical analysis. A.R.S., M.D.W., X.Z., C.V., H.B.P., and C.L.D. performed DNA WGS analysis. X.Z., C.T., and M.D.W. performed germline DNA analysis. A.R.S., C.L.D., and M.D.W. performed RNA-seq analysis. N.W.B., A.R.S., T.P.C., E.F.P., H.B.P., and M.D.W. performed proteomics and proteogenomics analysis. C.A.M. and T.J.F. performed pathology analysis. M.D.W., J.L., and H.H. performed survival analysis. T.N., A.R.S., C.Y., M.D.W., and D.M. performed pathway analysis. A.R.S. performed network analysis. C.D.S. and J.S.H.L. coordinated APOLLO study design. M.D.W. and A.R.S wrote the manuscript, which all authors reviewed.

DECLARATION OF INTERESTS

M.D.W., R.F.B., and C.D.S. are inventors for a provisional patent application related to findings reported in this manuscript. J.S.H.L. serves as Chief Science and Innovation Officer for Ellison Institute, LLC (paid); board of trustee for Health and Environmental Institute, Inc. (unpaid, travel support); and scientific advisory board for AtlasXomics, Inc., and ATOM, Inc. (unpaid, travel support).

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location.

Received: April 19, 2022
Revised: May 9, 2022
Accepted: October 18, 2022
Published: November 15, 2022

REFERENCES

1. Lin, J., Kamamia, C., Brown, D., Shao, S., McGlynn, K.A., Nations, J.A., Carter, C.A., Shriver, C.D., and Zhu, K. (2018). Survival among lung cancer patients in the U.S. Military Health system: a comparison with the SEER population. Cancer Epidemiol. Biomarkers Prev. *27*, 673–679. https://doi.org/10.1158/1055-9965.EPI-17-0822.
2. Grilley-Olson, J.E., Hayes, D.N., Moore, D.T., Leslie, K.O., Wilkerson, M.D., Qaqish, B.F., Hayward, M.C., Cabanski, C.R., Yin, X., Socinski, M.A., et al. (2013). Validation of interobserver agreement in lung cancer assessment: hematoxylin-eosin diagnostic reproducibility for non-small cell lung cancer: the 2004 World Health Organization classification and therapeutically relevant subsets. Arch. Pathol. Lab Med. *137*, 32–40. https://doi.org/10.5858/arpa.2012-0033-OA.
3. Skoulidis, F., and Heymach, J.V. (2019). Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy. Nat. Rev. Cancer *19*, 495–509. https://doi.org/10.1038/s41568-019-0179-8.
4. de Sousa, V.M.L., and Carvalho, L. (2018). Heterogeneity in lung cancer. Pathobiology *85*, 96–107. https://doi.org/10.1159/000487440.
5. Imielinski, M., Berger, A.H., Hammerman, P.S., Hernandez, B., Pugh, T.J., Hodis, E., Cho, J., Suh, J., Capelletti, M., Sivachenko, A., et al. (2012). Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell *150*, 1107–1120. https://doi.org/10.1016/j.cell.2012.08.029.
6. Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543–550. https://doi.org/10.1038/nature13385.
7. Lee, J.J.K., Park, S., Park, H., Kim, S., Lee, J., Lee, J., Youk, J., Yi, K., An, Y., Park, I.K., et al. (2019). Tracing oncogene rearrangements in the mutational history of lung adenocarcinoma. Cell *177*, 1842–1857.e21. https://doi.org/10.1016/j.cell.2019.05.013.
8. Wilkerson, M.D., Yin, X., Walter, V., Zhao, N., Cabanski, C.R., Hayward, M.C., Miller, C.R., Socinski, M.A., Parsons, A.M., Thorne, L.B., et al. (2012). Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. PLoS One *7*, e36530. https://doi.org/10.1371/journal.pone.0036530.
9. Yatabe, Y., Kosaka, T., Takahashi, T., and Mitsudomi, T. (2005). EGFR mutation is specific for terminal respiratory unit type adenocarcinoma. Am. J. Surg. Pathol. *29*, 633–639. https://doi.org/10.1097/01.pas.0000157935.28066.35.
10. Ringnér, M., and Staaf, J. (2016). Consensus of gene expression phenotypes and prognostic risk predictors in primary lung adenocarcinoma. Oncotarget *7*, 52957–52973. https://doi.org/10.18632/oncotarget.10641.
11. 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. https://doi.org/10.1038/s41571-018-0135-7.
12. 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, e217. https://doi.org/10.1016/j.cell.2020.05.043.

13. 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, e235. https://doi.org/10.1016/j.cell.2020.06.013.

14. 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, e217. https://doi.org/10.1016/j.cell.2020.06.012.

15. Sharpnack, M.F., Ranbaduge, N., Srivastava, A., Cerciello, F., Codreanu, S.G., Liebler, D.C., Mascaux, C., Miles, W.O., Morris, R., McDermott, J.E., et al. (2018). Proteogenomic analysis of surgically resected lung adenocarcinoma. J. Thorac. Oncol. *13*, 1519–1529. https://doi.org/10.1016/j.jtho.2018.06.025.

16. Stewart, P.A., Parapatics, K., Welsh, E.A., Müller, A.C., Cao, H., Fang, B., Koomen, J.M., Eschrich, S.A., Bennett, K.L., and Haura, E.B. (2015). A pilot proteogenomic study with data integration identifies MCT1 and GLUT1 as prognostic markers in lung adenocarcinoma. PLoS One *10*, e0142162. https://doi.org/10.1371/journal.pone.0142162.

17. Gasparri, R., Sedda, G., Noberini, R., Bonaldi, T., and Spaggiari, L. (2020). Clinical application of mass spectrometry-based proteomics in lung cancer early diagnosis. Proteomics. Clin. Appl. *14*, e1900138. https://doi.org/10.1002/prca.201900138.

18. Lee, J.S.H., Darcy, K.M., Hu, H., Casablanca, Y., Conrads, T.P., Dalgard, C.L., Freymann, J.B., Hanlon, S.E., Huang, G.D., Kvecher, L., et al. (2019). From discovery to practice and survivorship: building a national real-world data learning healthcare framework for military and veteran cancer patients. Clin. Pharmacol. Ther. *106*, 52–57. https://doi.org/10.1002/cpt.1425.

19. Fiore, L.D., Rodriguez, H., and Shriver, C.D. (2017). Collaboration to accelerate proteogenomics cancer care: the department of veterans affairs, department of Defense, and the national cancer institute's applied proteogenomics OrganizationaL learning and outcomes (APOLLO) network. Clin. Pharmacol. Ther. *101*, 619–621. https://doi.org/10.1002/cpt.658.

20. Funnell, T., Zhang, A.W., Grewal, D., McKinney, S., Bashashati, A., Wang, Y.K., and Shah, S.P. (2019). Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. PLoS Comput. Biol. *15*, e1006799. https://doi.org/10.1371/journal.pcbi.1006799.

21. 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. https://doi.org/10.1038/s41586-020-1943-3.

22. Lakin, N.D., and Jackson, S.P. (1999). Regulation of p53 in response to DNA damage. Oncogene *18*, 7644–7655. https://doi.org/10.1038/sj.onc.1203015.

23. Donehower, L.A., Soussi, T., Korkut, A., Liu, Y., Schultz, A., Cardenas, M., Li, X., Babur, O., Hsu, T.K., Lichtarge, O., et al. (2019). Integrated analysis of TP53 gene and pathway alterations in the cancer genome Atlas. Cell Rep. *28*, 3010. https://doi.org/10.1016/j.celrep.2019.08.061.

24. Carrot-Zhang, J., Yao, X., Devarakonda, S., Deshpande, A., Damrauer, J.S., Silva, T.C., Wong, C.K., Choi, H.Y., Felau, I., Robertson, A.G., et al. (2021). Whole-genome characterization of lung adenocarcinomas lacking alterations in the RTK/RAS/RAF pathway. Cell Rep. *34*, 108784. https://doi.org/10.1016/j.celrep.2021.108784.

25. 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. https://doi.org/10.1016/j.cell.2019.12.026.

26. Mun, D.G., Bhin, J., Kim, S., Kim, H., Jung, J.H., Jung, Y., Jang, Y.E., Park, J.M., Kim, H., Jung, Y., et al. (2019). Proteogenomic characterization of human early-onset gastric cancer. Cancer Cell *35*, 111–124.e10. https://doi.org/10.1016/j.ccell.2018.12.003.

27. Sinha, A., Huang, V., Livingstone, J., Wang, J., Fox, N.S., Kurganovs, N., Ignatchenko, V., Fritsch, K., Donmez, N., Heisler, L.E., et al. (2019). The proteogenomic landscape of curable prostate cancer. Cancer Cell *35*, 414–427.e6. https://doi.org/10.1016/j.ccell.2019.02.005.

28. Faruki, H., Mayhew, G.M., Serody, J.S., Hayes, D.N., Perou, C.M., and Lai-Goldman, M. (2017). Lung adenocarcinoma and squamous cell carcinoma gene expression subtypes demonstrate significant differences in tumor immune landscape. J. Thorac. Oncol. *12*, 943–953. https://doi.org/10.1016/j.jtho.2017.03.010.

29. Ringnér, M., Jönsson, G., and Staaf, J. (2016). Prognostic and chemotherapy predictive value of gene-expression phenotypes in primary lung adenocarcinoma. Clin. Cancer Res. *22*, 218–229. https://doi.org/10.1158/1078-0432.CCR-15-0529.

30. Marabelle, A., Fakih, M., Lopez, J., Shah, M., Shapira-Frommer, R., Nakagawa, K., Chung, H.C., Kindler, H.L., Lopez-Martin, J.A., Miller, W.H., Jr., et al. (2020). Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. Lancet Oncol. *21*, 1353–1365. https://doi.org/10.1016/S1470-2045(20)30445-9.

31. Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W., et al. (2018). The chromatin accessibility landscape of primary human cancers. Science *362*, eaav1898. https://doi.org/10.1126/science.aav1898.

32. Sadzak, I., Schiff, M., Gattermeier, I., Glinitzer, R., Sauer, I., Saalmüller, A., Yang, E., Schaljo, B., and Kovarik, P. (2008). Recruitment of Stat1 to chromatin is required for interferon-induced serine phosphorylation of Stat1 transactivation domain. Proc. Natl. Acad. Sci. USA *105*, 8944–8949. https://doi.org/10.1073/pnas.0801794105.

33. Parra, E.R., Villalobos, P., Mino, B., and Rodriguez-Canales, J. (2018). Comparison of different antibody clones for immunohistochemistry detection of programmed cell death ligand 1 (PD-L1) on non-small cell lung carcinoma. Appl. Immunohistochem. Mol. Morphol. *26*, 83–93. https://doi.org/10.1097/PAI.0000000000000531.

34. Karnik, T., Kimler, B.F., Fan, F., and Tawfik, O. (2018). PD-L1 in breast cancer: comparative analysis of 3 different antibodies. Hum. Pathol. *72*, 28–34. https://doi.org/10.1016/j.humpath.2017.08.010.

35. Gentles, A.J., Hui, A.B.Y., Feng, W., Azizi, A., Nair, R.V., Bouchard, G., Knowles, D.A., Yu, A., Jeong, Y., Bejnood, A., et al. (2020). A human lung tumor microenvironment interactome identifies clinically relevant cell-type cross-talk. Genome Biol. *21*, 107. https://doi.org/10.1186/s13059-020-02019-x.

36. Anjum, R., and Blenis, J. (2008). The RSK family of kinases: emerging roles in cellular signalling. Nat. Rev. Mol. Cell Biol. *9*, 747–758. https://doi.org/10.1038/nrm2509.

37. Poomakkoth, N., Issa, A., Abdulrahman, N., Abdelaziz, S.G., and Mraiche, F. (2016). p90 ribosomal S6 kinase: a potential therapeutic target in lung cancer. J. Transl. Med. *14*, 14. https://doi.org/10.1186/s12967-016-0768-1.

38. Basu, A., and Sivaprasad, U. (2007). Protein kinase Cepsilon makes the life and death decision. Cell. Signal. *19*, 1633–1642. https://doi.org/10.1016/j.cellsig.2007.04.008.

39. Liu, T.C., Jin, X., Wang, Y., and Wang, K. (2017). Role of epidermal growth factor receptor in lung cancer and targeted therapies. Am. J. Cancer Res. *7*, 187–202.

40. Astsaturov, I., Ratushny, V., Sukhanova, A., Einarson, M.B., Bagnyukova, T., Zhou, Y., Devarajan, K., Silverman, J.S., Tikhmyanova, N., Skobeleva, N., et al. (2010). Synthetic lethal screen of an EGFR-centered network to improve targeted therapies. Sci. Signal. *3*, ra67. https://doi.org/10.1126/scisignal.2001083.

41. Lizotte, P.H., Ivanova, E.V., Awad, M.M., Jones, R.E., Keogh, L., Liu, H., Dries, R., Almonte, C., Herter-Sprie, G.S., Santos, A., et al. (2016). Multiparametric profiling of non-small-cell lung cancers reveals distinct immunophenotypes. JCI Insight *1*, e89014. https://doi.org/10.1172/jci.insight.89014.

42. Macdonald, J.I., and Dick, F.A. (2012). Posttranslational modifications of the retinoblastoma tumor suppressor protein as determinants of function. Genes Cancer *3*, 619–633. https://doi.org/10.1177/1947601912473305.

43. Scaffidi, C., Volkland, J., Blomberg, I., Hoffmann, I., Krammer, P.H., and Peter, M.E. (2000). Phosphorylation of FADD/MORT1 at serine 194 and association with a 70-kDa cell cycle-regulated protein kinase. J. Immunol. *164*, 1236–1242. https://doi.org/10.4049/jimmunol.164.3.1236.

44. Chen, G., Bhojani, M.S., Heaford, A.C., Chang, D.C., Laxman, B., Thomas, D.G., Griffin, L.B., Yu, J., Coppola, J.M., Giordano, T.J., et al. (2005). Phosphorylated FADD induces NF-kappaB, perturbs cell cycle, and is associated with poor outcome in lung adenocarcinomas. Proc. Natl. Acad. Sci. USA *102*, 12507–12512. https://doi.org/10.1073/pnas.0500397102.

45. Xue, Y., Meehan, B., Fu, Z., Wang, X.Q.D., Fiset, P.O., Rieker, R., Levins, C., Kong, T., Zhu, X., Morin, G., et al. (2019). SMARCA4 loss is synthetic lethal with CDK4/6 inhibition in non-small cell lung cancer. Nat. Commun. *10*, 557. https://doi.org/10.1038/s41467-019-08380-1.

46. Faubert, B., Vincent, E.E., Griss, T., Samborska, B., Izreig, S., Svensson, R.U., Mamer, O.A., Avizonis, D., Shackelford, D.B., Shaw, R.J., and Jones, R.G. (2014). Loss of the tumor suppressor LKB1 promotes metabolic reprogramming of cancer cells via HIF-1alpha. Proc. Natl. Acad. Sci. USA *111*, 2554–2559. https://doi.org/10.1073/pnas.1312570111.

47. Taguchi, K., and Yamamoto, M. (2017). The KEAP1-NRF2 system in cancer. Front. Oncol. *7*, 85. https://doi.org/10.3389/fonc.2017.00085.

48. Galan-Cobo, A., Sitthideatphaiboon, P., Qu, X., Poteete, A., Pisegna, M.A., Tong, P., Chen, P.H., Boroughs, L.K., Rodriguez, M.L.M., Zhang, W., et al. (2019). LKB1 and KEAP1/NRF2 pathways cooperatively promote metabolic reprogramming with enhanced glutamine dependence in KRAS-mutant lung adenocarcinoma. Cancer Res. *79*, 3251–3267. https://doi.org/10.1158/0008-5472.CAN-18-3527.

49. Wohlhieter, C.A., Richards, A.L., Uddin, F., Hulton, C.H., Quintanal-Villalonga, À., Martin, A., de Stanchina, E., Bhanot, U., Asher, M., Shah, N.S., et al. (2020). Concurrent mutations in STK11 and KEAP1 promote ferroptosis protection and SCD1 dependence in lung cancer. Cell Rep. *33*, 108444. https://doi.org/10.1016/j.celrep.2020.108444.

50. Hsiue, E.H.C., Wright, K.M., Douglass, J., Hwang, M.S., Mog, B.J., Pearlman, A.H., Paul, S., DiNapoli, S.R., Konig, M.F., Wang, Q., et al. (2021). Targeting a neoantigen derived from a common TP53 mutation. Science *371*, eabc8697. https://doi.org/10.1126/science.abc8697.

51. Hood, B.L., Darfler, M.M., Guiel, T.G., Furusato, B., Lucas, D.A., Ringeisen, B.R., Sesterhenn, I.A., Conrads, T.P., Veenstra, T.D., and Krizman, D.B. (2005). Proteomic analysis of formalin-fixed prostate cancer tissue. Mol. Cell. Proteomics *4*, 1741–1753. https://doi.org/10.1074/mcp.M500102-MCP200.

52. Herbst, R.S., Baas, P., Kim, D.W., Felip, E., Pérez-Gracia, J.L., Han, J.Y., Molina, J., Kim, J.H., Arvis, C.D., Ahn, M.J., et al. (2016). Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. Lancet *387*, 1540–1550. https://doi.org/10.1016/S0140-6736(15)01281-7.

53. Kheradpour, P., and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. *42*, 2976–2987. https://doi.org/10.1093/nar/gkt1249.

54. Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M., and Stamatoyannopoulos, J.A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc. Natl. Acad. Sci. USA *107*, 139–144. https://doi.org/10.1073/pnas.0912402107.

55. Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773. https://doi.org/10.1093/nar/gky955.

56. Fornes, O., Castro-Mondragon, J.A., Khan, A., van der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranašić, D., et al. (2020). Jaspar 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *48*, D87–D92. https://doi.org/10.1093/nar/gkz1001.

57. 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. https://doi.org/10.1073/pnas.0506580102.

58. Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. *43*, D512–D520. https://doi.org/10.1093/nar/gku1267.

59. Wu, G., Feng, X., and Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. Genome Biol. *11*, R53. https://doi.org/10.1186/gb-2010-11-5-r53.

60. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164. https://doi.org/10.1093/nar/gkq603.

61. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457. https://doi.org/10.1038/nmeth.3337.

62. Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573. https://doi.org/10.1093/bioinformatics/btq170.

63. Bergmann, E.A., Chen, B.J., Arora, K., Vacic, V., and Zody, M.C. (2016). Conpair: concordance and contamination estimator for matched tumor-normal pairs. Bioinformatics *32*, 3196–3198. https://doi.org/10.1093/bioinformatics/btw389.

64. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504. https://doi.org/10.1101/gr.1239303.

65. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S., and Swanton, C. (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. *17*, 31. https://doi.org/10.1186/s13059-016-0893-4.

66. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

67. Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P.W., Levine, D.A., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun. *4*, 2612. https://doi.org/10.1038/ncomms3612.

68. 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. https://doi.org/10.1186/gb-2011-12-4-r41.

69. Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169. https://doi.org/10.1093/bioinformatics/btu638.

70. Li, Q., and Wang, K. (2017). InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. Am. J. Hum. Genet. *100*, 267–280. https://doi.org/10.1016/j.ajhg.2017.01.004.

71. Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., et al. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. *38*, e178. https://doi.org/10.1093/nar/gkq622.

72. Pin, E., Federici, G., and Petricoin, E.F., 3rd. (2014). Preparation and use of reverse protein microarrays. Curr. Protoc. Protein Sci. *75*, 27.7.1–27.7.29. https://doi.org/10.1002/0471140864.ps2707s75.

73. Meng, C., Basunia, A., Peters, B., Gholami, A.M., Kuster, B., and Culhane, A.C. (2019). MOGSA: integrative single sample gene-set analysis of multiple omics data. Mol. Cell. Proteomics *18*, S153–S168. https://doi.org/10.1074/mcp.TIR118.001251.

74. Soltis, A.R., Dalgard, C.L., Pollard, H.B., and Wilkerson, M.D. (2020). MutEnricher: a flexible toolset for somatic mutation enrichment analysis of tumor whole genomes. BMC Bioinf. *21*, 338. https://doi.org/10.1186/s12859-020-03695-z.

75. Hagberg A, S.D., and Swart, P. (2008). Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference, 11–15.

76. Pedersen, B.S., and Quinlan, A.R. (2017). Who's who? Detecting and resolving sample anomalies in human DNA sequencing studies with Peddy. Am. J. Hum. Genet. *100*, 406–413. https://doi.org/10.1016/j.ajhg.2017.01.017.

77. Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput. Biol. *6*, e1000770. https://doi.org/10.1371/journal.pcbi.1000770.

78. Ruggles, K.V., Tang, Z., Wang, X., Grover, H., Askenazi, M., Teubl, J., Cao, S., McLellan, M.D., Clauser, K.R., Tabb, D.L., et al. (2016). An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. Mol. Cell. Proteomics *15*, 1060–1071. https://doi.org/10.1074/mcp.M115.056226.

79. O'Connell, M.J., and Lock, E.F. (2016). R.JIVE for exploration of multisource molecular data. Bioinformatics *32*, 2877–2879. https://doi.org/10.1093/bioinformatics/btw324.

80. Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics *28*, 2184–2185. https://doi.org/10.1093/bioinformatics/bts356.

81. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

82. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

83. Zhu, Y., Orre, L.M., Johansson, H.J., Huss, M., Boekel, J., Vesterlund, M., Fernandez-Woodbridge, A., Branca, R.M.M., and Lehtiö, J. (2018). Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. Nat. Commun. *9*, 903. https://doi.org/10.1038/s41467-018-03311-y.

84. Gordon, D.B., Nekludova, L., McCallum, S., and Fraenkel, E. (2005). TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. Bioinformatics *21*, 3164–3165. https://doi.org/10.1093/bioinformatics/bti481.

85. Wilkerson, M.D., Cabanski, C.R., Sun, W., Hoadley, K.A., Walter, V., Mose, L.E., Troester, M.A., Hammerman, P.S., Parker, J.S., Perou, C.M., and Hayes, D.N. (2014). Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. Nucleic Acids Res. *42*, e107. https://doi.org/10.1093/nar/gku489.

86. Mukhopadhyay, S., and Katzenstein, A.L.A. (2011). Subclassification of non-small cell lung carcinomas lacking morphologic differentiation on biopsy specimens: utility of an immunohistochemical panel containing TTF-1, napsin A, p63, and CK5/6. Am. J. Surg. Pathol. *35*, 15–25. https://doi.org/10.1097/PAS.0b013e3182036d05.

87. Travis, W.D., Brambilla, E., Noguchi, M., Nicholson, A.G., Geisinger, K.R., Yatabe, Y., Beer, D.G., Powell, C.A., Riely, G.J., Van Schil, P.E., et al. (2011). International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. J. Thorac. Oncol. *6*, 244–285. https://doi.org/10.1097/JTO.0b013e318206a221.

88. International Agency for Research on Cancer (2015). WHO Classification of Tumours of the Lung, Pleura, Thymus and Heart, 4th Edition (World Health Organization).

89. Sica, G., Yoshizawa, A., Sima, C.S., Azzoli, C.G., Downey, R.J., Rusch, V.W., Travis, W.D., and Moreira, A.L. (2010). A grading system of lung adenocarcinomas based on histologic pattern is predictive of disease recurrence in stage I tumors. Am. J. Surg. Pathol. *34*, 1155–1162. https://doi.org/10.1097/PAS.0b013e3181e4ee32.

90. Kadota, K., Nitadori, J.I., Sima, C.S., Ujiie, H., Rizk, N.P., Jones, D.R., Adusumilli, P.S., and Travis, W.D. (2015). Tumor spread through air spaces is an important pattern of invasion and impacts the frequency and location of recurrences after limited resection for small stage I lung adenocarcinomas. J. Thorac. Oncol. *10*, 806–814. https://doi.org/10.1097/JTO.0000000000000486.

91. Lee, S., Zhao, L., Rojas, C., Bateman, N.W., Yao, H., Lara, O.D., Celestino, J., Morgan, M.B., Nguyen, T.V., Conrads, K.A., et al. (2020). Molecular analysis of clinically defined subsets of high-grade serous ovarian cancer. Cell Rep. *31*, 107502. https://doi.org/10.1016/j.celrep.2020.03.066.

92. Vrana, M., Goodling, A., Afkarian, M., and Prasad, B. (2016). An optimized method for protein extraction from OCT-embedded human kidney tissue for protein quantification by LC-MS/MS proteomics. Drug Metab. Dispos. *44*, 1692–1696. https://doi.org/10.1124/dmd.116.071522.

93. Baldelli, E., Calvert, V., Hodge, A., VanMeter, A., Petricoin, E.F., 3rd, and Pierobon, M. (2017). Reverse phase protein microarrays. Methods Mol. Biol. *1606*, 149–169. https://doi.org/10.1007/978-1-4939-6990-6_11.

94. Signore, M., Manganelli, V., and Hodge, A. (2017). Antibody validation by western blotting. Methods Mol. Biol. *1606*, 51–70. https://doi.org/10.1007/978-1-4939-6990-6_4.

95. Käll, L., Canterbury, J.D., Weston, J., Noble, W.S., and MacCoss, M.J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods *4*, 923–925. https://doi.org/10.1038/nmeth1113.

96. Taus, T., Köcher, T., Pichler, P., Paschke, C., Schmidt, A., Henrich, C., and Mechtler, K. (2011). Universal and confident phosphorylation site localization using phosphoRS. J. Proteome Res. *10*, 5354–5362. https://doi.org/10.1021/pr200611n.

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

98. Lazar, C., Gatto, L., Ferro, M., Bruley, C., and Burger, T. (2016). Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. J. Proteome Res. *15*, 1116–1125. https://doi.org/10.1021/acs.jproteome.5b00981.

99. Raczy, C., Petrovski, R., Saunders, C.T., Chorny, I., Kruglyak, S., Margulies, E.H., Chuang, H.Y., Källberg, M., Kumar, S.A., Liao, A., et al. (2013). Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics *29*, 2041–2043. https://doi.org/10.1093/bioinformatics/btt314.

100. Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., and Saunders, C.T. (2018). Strelka2: fast and accurate calling of germline and somatic

variants. Nat. Methods *15*, 591–594. https://doi.org/10.1038/s41592-018-0051-x.

101. 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. https://doi.org/10.1093/bioinformatics/btv710.

102. Roller, E., Ivakhno, S., Lee, S., Royce, T., and Tanner, S. (2016). Canvas: versatile and scalable detection of copy number variants. Bioinformatics *32*, 2375–2377. https://doi.org/10.1093/bioinformatics/btw163.

103. Huang, K.L., Mashl, R.J., Wu, Y., Ritter, D.I., Wang, J., Oh, C., Paczkowska, M., Reynolds, S., Wyczalkowski, M.A., Oak, N., et al. (2018). Pathogenic germline variants in 10, 389 adult cancers. Cell *173*, 355–370.e14. https://doi.org/10.1016/j.cell.2018.03.039.

104. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60, 706 humans. Nature *536*, 285–291. https://doi.org/10.1038/nature19057.

105. Wang, L., Nie, J., Sicotte, H., Li, Y., Eckel-Passow, J.E., Dasari, S., Vedell, P.T., Barman, P., Wang, L., Weinshiboum, R., et al. (2016). Measure transcript integrity using RNA-seq data. BMC Bioinf. *17*, 58. https://doi.org/10.1186/s12859-016-0922-z.

106. Zhang, W., Bojorquez-Gomez, A., Velez, D.O., Xu, G., Sanchez, K.S., Shen, J.P., Chen, K., Licon, K., Melton, C., Olson, K.M., et al. (2018). A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. Nat. Genet. *50*, 613–620. https://doi.org/10.1038/s41588-018-0091-2.

107. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. *7*, 500–507. https://doi.org/10.1038/nprot.2011.457.

108. Donehower, L.A., Soussi, T., Korkut, A., Liu, Y., Schultz, A., Cardenas, M., Li, X., Babur, O., Hsu, T.K., Lichtarge, O., et al. (2019). Integrated analysis of TP53 gene and pathway alterations in the cancer genome Atlas. Cell Rep. *28*, 1370–1384.e5. https://doi.org/10.1016/j.celrep.2019.07.001.

109. Fishbein, L., Leshchiner, I., Walter, V., Danilova, L., Robertson, A.G., Johnson, A.R., Lichtenberg, T.M., Murray, B.A., Ghayee, H.K., Else, T., et al. (2017). Comprehensive molecular characterization of pheochromocytoma and paraganglioma. Cancer Cell *31*, 181–193. https://doi.org/10.1016/j.ccell.2017.01.001.

110. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740. https://doi.org/10.1093/bioinformatics/btr260.

111. 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. https://doi.org/10.1016/j.cels.2015.12.004.

112. Chen, E.G., Wang, P., Lou, H., Wang, Y., Yan, H., Bi, L., Liu, L., Li, B., Snijders, A.M., Mao, J.H., and Hang, B. (2018). A robust gene expression-based prognostic risk score predicts overall survival of lung adenocarcinoma patients. Oncotarget *9*, 6862–6871. https://doi.org/10.18632/oncotarget.23490.

113. Touzet, H., and Varré, J.S. (2007). Efficient and accurate P-value computation for position weight matrices. Algorithms Mol. Biol. *2*, 15. https://doi.org/10.1186/1748-7188-2-15.

114. Soltis, A.R., Kennedy, N.J., Xin, X., Zhou, F., Ficarro, S.B., Yap, Y.S., Matthews, B.J., Lauffenburger, D.A., White, F.M., Marto, J.A., et al. (2017). Hepatic dysfunction caused by consumption of a high-fat diet. Cell Rep. *21*, 3317–3328. https://doi.org/10.1016/j.celrep.2017.11.059.

115. Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The human transcription factors. Cell *175*, 598–599. https://doi.org/10.1016/j.cell.2018.09.045.

116. Woods, A., Johnstone, S.R., Dickerson, K., Leiper, F.C., Fryer, L.G.D., Neumann, D., Schlattner, U., Wallimann, T., Carlson, M., and Carling, D. (2003). LKB1 is the upstream kinase in the AMP-activated protein kinase cascade. Curr. Biol. *13*, 2004–2008. https://doi.org/10.1016/j.cub.2003.10.031.

117. Willows, R., Sanders, M.J., Xiao, B., Patel, B.R., Martin, S.R., Read, J., Wilson, J.R., Hubbard, J., Gamblin, S.J., and Carling, D. (2017). Phosphorylation of AMPK by upstream kinases is required for activity in mammalian cells. Biochem. J. *474*, 3059–3073. https://doi.org/10.1042/BCJ20170458.

118. Brubaker, D.K., Paulo, J.A., Sheth, S., Poulin, E.J., Popow, O., Joughin, B.A., Strasser, S.D., Starchenko, A., Gygi, S.P., Lauffenburger, D.A., and Haigis, K.M. (2019). Proteogenomic network analysis of context-specific KRAS signaling in mouse-to-human cross-species translation. Cell Syst. *9*, 258–270.e6. https://doi.org/10.1016/j.cels.2019.07.006.

119. Su, G., Kuchinsky, A., Morris, J.H., States, D.J., and Meng, F. (2010). GLay: community structure analysis of biological networks. Bioinformatics *26*, 3135–3137. https://doi.org/10.1093/bioinformatics/btq596.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Primary antibody information is listed in Table S1C.	Various	Various
Biological samples		
APOLLO1 cohort (bulk DNA, RNA, protein) information is listed in Table S1A.	Lung Cancer Biospecimen Research Network	N/A
Chemicals, peptides, and recombinant proteins		
Mayer's Hematoxylin Solution	Sigma Aldrich	Cat# MHS32
Eosin Y Solution Aqueous	Sigma Aldrich	Cat# HT110216
Phosphatase Inhibitor Cocktail 3	Sigma Aldrich	Cat# P0044
Phosphatase Inhibitor Cocktail 2	Sigma Aldrich	Cat# P5726
Biotin blocking system	Dako	Cat# X0590
Hydrogen peroxide	Sigma-Aldrich	Cat# 323381
I-block	Invitrogen	Cat# T2015
IRDye680RD Streptavidin fluorescent dye	LI-COR Biosciences	Cat# 926-68079
PBS	Gibco	Cat# 14190136
Reblot Plus Mild Antibody stripping solution	Millipore	Cat# 2502
Serum free protein block	Dako	Cat# X0909
Sypro Ruby Protein Blot Stain	Invitrogen	Cat# S11791
Mass Spec-Compatible Human Protein Extract, Digest	Promega	Cat# V6951
Pierce Peptide Retention Time Calibration Mixture	Pierce	Cat# 88321
Critical commercial assays		
RNeasy Lipid Mini Kit	Qiagen	cat # 74804
QIAamp DNA Mini Kit	Qiagen	cat # 51304
CSA kit	Dako	K1500
Pierce BCA Protein Assay Kit	Thermo Fisher Scientific	Cat# 23225
Soluble Smart Digest Kit	Thermo Fisher Scientific	Cat# 3251711
TMT10plex Isobaric Label Reagent Set plus TMT11-131C Label Reagent	Thermo Fisher Scientific	Cat# A34808
High-Select™ TiO2 Phosphopeptide Enrichment Kit	Thermo Fisher Scientific	Cat# A32993
High-Select™ Fe-NTA Phosphopeptide Enrichment Kit	Thermo Fisher Scientific	Cat# A32992
TruSeq DNA PXR-Dree High Throughput Library Prep Kit (96 Samples)	Illumina	20015963
TruSeq DNA CD Indexes (96 Indexes, 96 Samples)	Illumina	20015949
IDT for Illumina - TruSeq DNA UD Indexes (96 Indexes, 96 Samples)	Illumina	20022370
TruSeq Stranded Total RNA Library Prep Gold (96 Samples)	Illumina	20020599
IDT for Illumina - TruSeq RNA UD Indexes (96 Indexes, 96 Samples)	Illumina	20022371
HiSeq X Ten Reagent Kit v2.5 - 10 pack	Illumina	FC-501-2521
HiSeq 3000/4000 PE Cluster Kit	Illumina	PE-410-1001
HiSeq 3000/4000 SBS Kit (150 cycles)	Illumina	FC-410-1002
Kapa Library Quantification Kits - Complete Kit	Roche	7960298001
NGS Fragment Analysis Kit (35–6000 bp)	Agilent	DNF-486-0500-OB

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
WGS data	This paper	dbGap phs003011.v1.p1 and NCI Genomic Data Commons
RNA sequencing data	This paper	dbGap, phs003011.v1.p1 and NCI Genomic Data Commons
Proteomic data	This paper	NCI proteomic data commons, phs003011.v1.p1, ProteomeXChange # PXD036025
CPTAC LUAD data	Gillette et al.[13]	https://proteomic.datacommons.cancer.gov/, Study ID: PDC000153, File ID 31ba3150-cec4-4749-b06e-443239185938
COSMIC signatures (v 3.0)	Alexandrov et al.[21]	https://cancer.sanger.ac.uk/signatures/
ENCODE motifs	Kheradpour et al.[53]	http://compbio.mit.edu/encode-motifs/
ENCODE Repli-Seq	Hansen et al.[54]	https://www.encodeproject.org/
GENCODE v28	Frankish et al.[55]	https://www.gencodegenes.org/
JASPAR motifs (2020)	Fornes et al.[56]	https://jaspar.genereg.net/
MSigDB (v7.0)	Subramanian et al.[57]	http://www.gsea-msigdb.org/gsea/downloads.jsp
PhosphoSitePlus	Hornbeck et al.[58]	https://www.phosphosite.org/
Reactome Functional Interactome (v071718)	Wu et al.[59]	https://reactome.org/download-data
TCGA ATAC-seq	Corces et al.[31]	https://www.science.org/10.1126/science.aav1898
Software and algorithms		
ANNOVAR, 2017-07-17	Wang et al.[60]	https://annovar.openbioinformatics.org/en/latest/
Break Point Inspector v1.7	https://github.com/hartwigmedical/hmftools	https://github.com/hartwigmedical/hmftools/releases/tag/bpi-v1.7
CIBERSORT, version 1.06	Newman et al.[61]	https://cibersort.stanford.edu/
Canvas	Illumina	version 1.28.0.272 + master
ConsensusClusterPlus, 1.24	Wilkerson et al.[62]	https://bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html
Conpair	Bergmann et al.[63]	https://github.com/nygenome/Conpair
Cytoscape	Shannon et al.[64]	https://cytoscape.org/
deconstructSigs v1.8.0	Rosenthal et al.[65]	https://cran.r-project.org/web/packages/deconstructSigs/index.html
DESeq2 v1.16.0	Love et al.[66]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
ESTIMATE	Yoshihara et al.[67]	https://rdrr.io/rforge/estimate/
GISTIC2	Mermel et al.[68]	https://github.com/broadinstitute/gistic2
GSEA v4.0.3	Subramanian et al.[57]	http://www.gsea-msigdb.org/gsea/index.jsp
HTSeq v0.9.1	Anders et al.[69]	https://htseq.readthedocs.io/en/master/
illumina sequence analysis software, resequencing workflow	Illumina	Resequencing Workflow, version 6.19.1.403 + NSv6
illumina sequence analysis software, tumor normal workflow	Illumina	Tumor Normal Workflow, version 6.9.1.177 + NSv6
Interlap	GitHub	https://brentp.github.io/interlap/
Intervar, 2.0.2 20180118	Li et al.[70]	https://github.com/WGLab/InterVar
Isaac aligner	Illumina	version Isaac-04.17.06.15
Lung Cancer Expression Subtype Centroid Predictor	Wilkerson et al.[8]	https://github.com/mwilkers/lungCancerSubtypes/blob/main/lung_adenocarcinoma_subtypes/wilkerson.2012.LAD.predictor.centroids.csv
Manta	Illumina	version 1.1.1
Mascot	Matrix Science	https://www.matrixscience.com/
Mapsplice, V2.2	Wang et al.[71]	http://www.netlab.uky.edu/p/bioinfo/MapSplice2

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Micro-Vigene	Pin et al.[72]	http://www.vigenetech.com/MicroVigene.htm
MOGSA, version 1.22.1	Meng et al.[73]	https://www.bioconductor.org/packages/release/bioc/html/mogsa.html
MultiModalMuSig	Funnell et al.[20]	https://github.com/shahcompbio/MultiModalMuSig.jl
MutEnricher v1.3.1	Soltis et al.[74]	https://github.com/asoltis/MutEnricher
NetworkX	Hagberg et al.[75]	https://networkx.org/
Pamr	CRAN	https://cran.r-project.org/web/packages/pamr/index.html
Peddy v0.3.0	Pedersen et al.[76]	https://github.com/brentp/peddy
PEER v1.3	Stegle et al.[77]	https://github.com/PMBio/peer/
Proteome Discoverer	Thermo Fisher Scientific Inc.	https://www.thermofisher.com/us/en/home.html
pyQUILTS v3.0	Ruggles et al.[78]	https://github.com/ekawaler/pyQUILTS
R, v3.23 and v4.0	CRAN	https://cran.r-project.org/
r.jive	O'Connell et al.[79]	https://cran.r-project.org/web/packages/r.jive/index.html
RSeQC v2.6.4	Wang et al.[80]	http://rseqc.sourceforge.net/
SAMtools	Li et al.[81]	https://www.htslib.org/
scikit-learn	Pedregosa et al.[82]	https://scikit-learn.org/stable/
SpectrumAI	Zhu et al.[83]	https://github.com/yafeng/SpectrumAI
Strelka	Illumina	version 2.8.0
TAMO	Gordon et al.[84]	http://fraenkel-nsf.csbi.mit.edu/TAMO/
UNCeqR, v0.2	Wilkerson et al.[85]	https://github.com/mwilkers/unceqr
wgsim	GitHub	https://github.com/lh3/wgsim
Other		
PEN Membrane Glass Slides	Leica Microsystems	Cat# 11532918
96 Micro-Tubes in bulk (no caps)	Pressure Biosciences Inc	Cat# MT-96
96 Micro-Caps (150uL) in bulk	Pressure Biosciences Inc	Cat# MC150-96
96 Micro-Pestles in bulk	Pressure Biosciences Inc	Cat# MP-96
9 mm MS Certified Clear Screw Thread Kits	Fisher Scientific	Cat# 03-060-058
Nitrocellulose-coated glass slides (ONCYTE AVID 1- 22 × 51mm NC Pad Per Slide Glass, 25 × 75 × 1mm, Small Dark Blue Box)	Grace Bio-labs	Cat# RD478691-M
ZORBAX Extend 300C18, 2.1 × 12.5 mm, 5 μm, guard cartridge (ZGC)	Agilent	Cat# 821125-932
ZORBAX Extend 300C18, 2.1 × 150 mm, 3.5 μm	Agilent	Cat# 763750-902
EASY-SPRAY C18 2UM 50CM X 75	Fisher Scientific	Cat# ES903
PM100C18 3UM 75UMX20MM NV 2PK	Fisher Scientific	Cat# 164535

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dr. Matthew D. Wilkerson (matthew.wilkerson@usuhs.edu).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

- Data generated in this study (DNA sequencing, RNA sequencing, and proteomic data) are deposited at dbGap under study accession #phs003011.v1.p1, the NCI Cancer Research Data Commons at https://gdc.cancer.gov/about-data/publications/APOLLO-LUAD-2022, and the ProteomeXChange #PXD036025.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact, Matthew Wilkerson (matthew.wilkerson@usuhs.edu), upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Specimens and clinical data

Patients were recruited to donate specimens to the Lung Cancer Biospecimen Resource Network (LCBRN, https://lungbio.sites.virginia.edu) at three institutions: The Medical University of South Carolina, The University of Virginia and Washington University-St. Louis, with informed patient consent and local Institutional Review Board approval. Patients were recruited in Thoracic Surgery clinics by study teams at each institution from those individuals undergoing evaluation for surgical resection of known or suspected primary pulmonary carcinoma. All cases had no evidence of disease and surgical margins were negative. The study teams collected demographic data, including self-reported gender and uniform clinical history, including prior lung cancer procedures, cancer treatments, and exposure data, as well as radiology and pathology reports (Table S1A). Tissue aliquots of tumor and surrounding non-neoplastic tissue were obtained from the surgical resection specimens (see below). Clinical follow up data was obtained from each individual at approximately 6 months intervals following surgery, including treatment data, clinical recurrence of cancer, and survival. Frozen tissue samples of pulmonary adenocarcinoma were obtained from the LCBRN, which had collected the samples using uniform procedures and collection containers. Warm and cold ischemic times prior to aliquot freezing were recorded, and the samples were stored at -80°C in mechanical freezers until analysis. Initial central pathology review was performed by C.A.M. on matched formalin-fixed paraffin-embedded samples taken adjacent to the frozen tissue aliquots, and then on cryostat frozen sections taken from the frozen tissue aliquots. Glandular differentiation was verified for well to moderately differentiated neoplasms. Confirmatory immunohistochemistry (TTF1, napsin, p63, cytokeratin 5/6) was performed on poorly differentiated neoplasms to confirm adenocarcinoma differentiation as per the recommendations of Mukhopadhyay et al.[86] Secondary central pathology review was performed by T.J.F. to provide expert histologic subtyping. Areas of adenocarcinoma were microdissected from the frozen tissue aliquots using the cryostat histologic sections as a guide to maximize tumor cellularity and minimize areas of tissue necrosis. Our study design sought to include 100 cases including up to 50 with recurrence and up to 50 with no recurrence. After tissue and molecular quality control, we had a final cohort of 87 cases.

CPTAC cohort

Sample-level RNA expression, protein expression, gene-level mutation data, specimen and subject data were obtained from supporting tables in Gillette et al.,[13] resulting in 105 cases with RNA, protein, and phenotype data. Clinical follow up data for these specimens were amended with the latest CPTAC clinical data (S046_S056_BI_CPTAC3_LUAD_Discovery_Cohort_Clinical_Data_r2_July2020.xlsx, available from https://proteomic.datacommons.cancer.gov/, Study ID: PDC000153, File ID 31ba3150-cec4-4749-b06e-443239185938). Overall survival and metastasis free survival time intervals were calculated from these data. Tumor DNA purity was calculated from somatic mutations as the median of 2*mutant allele fraction for mutations with a mutant allele fraction <50%. RNA expression subtypes were assigned using the published subtype predictor to gene median centered RNA expression data.

METHOD DETAILS

Tumor histological evaluation

Digitized slides, one per case, were reviewed by a pulmonary pathologist (T.J.F.) who was blinded to the patients' clinical outcomes and molecular analyses. Neoplasms were classified according to the 2015 WHO criteria for the classification of lung adenocarcinomas.[87,88] Invasive adenocarcinomas were classified into the five main subtypes lepidic, acinar, papillary, micropapillary, and solid[87,88] and three less common subtypes cribriform, ragged-anastomosing glands, and dispersed intra-alveolar tumor cells[89] based on the predominant histologic pattern of growth. The predominant subtype was defined by the histologic pattern comprising the highest percentage of the neoplasm after semiquantitative estimation of each pattern in 5% increments. Neoplasms were graded based on the highest grade histologic pattern present, regardless of percent composition, as I, II, or III.[87,89] Variants of adenocarcinoma, including invasive mucinous adenocarcinoma, colloid, fetal, and enteric adenocarcinoma,[87,88] and the presence of spread through air spaces (STAS) were noted.[90]

Tissue sectioning and utilization

Tumor specimens were embedded in optimal cutting temperature (OCT) compound. OCT blocks were then sectioned in an interlacing sequence to obtain material for DNA, RNA, protein and H&E slides. From the top of the block, two sections at 5 microns were

taken for were taken for H&E analysis. A section count of 5-10 was estimated for a tumor, based on the size of the tissue in the OCT block with smaller tumors having more sections to enable sufficient quantity for DNA and RNA isolation. With this count, sections of the tumor were cut at 20 microns for DNA followed by the same number of sections for RNA at 20 microns. For protein analysis, 4 sections (2 sections per slide) cut at 10 microns were placed onto PEN membrane slides. This interlacing sequence (DNA, RNA, protein) was repeated 5 times and included a mid-way H&E and a final H&E. The sequence is listed in Table S1B.

Tumor DNA and RNA sections from this interlacing sequence were pooled for DNA extraction and pooled for RNA extraction. Tumor DNA and RNA was extracted using the Qiagen QIAamp DNA Mini Kit and RNeasy Lipid Mini Kit (Qiagen Germantown, MD) respectively. Whole tumor laser microdissection was performed on the PEN membrane slides before protein analysis, to conduct tissue harvest to represent a bulk tissue representation of the tumor specimen. PEN membrane slides were sequentially utilized from the top of the block until approximately 80mm^2 tissue area was harvested for LMD for MS proteomics and phosphoproteomics. Then, the next PEN membrane sections were utilized sequentially until approximately 15mm^2 tissue area was harvested for RPPA.

Matched normal DNA for germline was isolated from either the buffy coat fraction of blood or non-neoplastic tissue using Qiagen DNeasy Blood & Tissue Kit and the Qiagen QIAamp DNA Mini Kit respectively (Qiagen, Germantown, MD).

DNA sample handling and library preparation

Quality control of input genomic DNA samples was conducted by visual inspection for discoloration and/or presence of precipitants. Genomic DNA quantitation was performed using a fluorescence dye-based assay (PicoGreen dsDNA reagent) and measured by a microplate reader (Molecular Devices SpectraMax Gemini XS) before normalization to 20 ng/uL. Normalized gDNA samples were added into wells of a Covaris 96 microTUBE plate at 55 μL volume and sheared using the Covaris LE220 Focused-ultrasonicator with settings for targeting a peak size of 410 bp (PIP: 450 W, Duty Factor: 18%, Cycles per burst: 200, Time: 60s). Sequencing libraries were generated from 1,000 ng of fragmented DNA using the Illumina TruSeq DNA PCR-Free HT Library Preparation Kit with minor modifications for automation on a Hamilton STAR Liquid Handling System. Adapters for ligation used either TruSeq DNA CD Indexes or IDT for Illumina TruSeq DNA UD Indexes (96 Indexes, 96 Samples). Library size distribution and absence of free adapters and/or adapter dimers was assessed by automated capillary gel-electrophoresis (Advanced Analytical Fragment Analyzer). Library yield and concentration (in nM) was determined by qPCR quantitation using the KAPA qPCR Quantification Kit on a Roche Light Cycler 480 Instrument II.

DNA library clustering and whole genome sequencing

After qPCR quantitation of sequencing libraries, normalization of libraries to 2.2 nM was performed into a working 96 well plate by automation on a Hamilton STAR Liquid Handling System. Libraries were clustered as three lanes within a single flowcell for tumor tissue-derived samples or single lane per single flowcell for germline tissue-derived samples on an Illumina cBot2 using the HiSeq X PE Cluster Kit and a HiSeq X Flow Cell v2.5 before sequencing on an Illumina HiSeq X System with 151 + 7+151 cycle parameters using HiSeq X HD SBS Kit reagents.

RNA library preparation and sequencing

Total RNA sample integrity was assessed using automated capillary electrophoresis on a Fragment Analyzer (Agilent) using the HS RNA Kit (15NT). For all samples with RQN >4.0, a total RNA amount of >100 ng was used as input for library preparation using the TruSeq Stranded total RNA Library Preparation Kit (Illumina). Sequencing libraries were quantified by real-time PCR using the KAPA Library Quantification Complete kit (Roche) and assessed for size distribution and absence of free adapters and adapter dimers on a Fragment Analyzer. Sequencing libraries were pooled and quantified by real-time PCR as above and clustered on a cBot2 (Illumina) using a HiSeq 3000/4000 PE Cluster Kit. Clustered flowcells were sequenced on a HiSeq 3000 System (Illumina) using a HiSeq 3000/4000 SBS Kit (150 cycles) with run conditions generating paired-end reads at 75 bp length.

Proteomics specimen preparation

Collection of lung tumor tissues using whole tumor laser microdissection (LMD), sample digestion, preparation of TMT multiplexes and offline, basic reversed-phase liquid chromatographic (bRPLC) fractionation was performed essentially as previously described.[91] Briefly, whole tissue representations (cancer and stroma combined) were harvested by LMD from fresh-frozen tissue sections (10 μm) on polyethylene naphthalate membrane slides, without any bias for cellular subpopulations. Whole tumor LMD was performed to minimize contamination of tissue samples with OCT mounting medium, a compound known to impact mass spectrometry analysis.[92] The mean tissue area collected per sample was 87 ± 4 mm^2 for MS proteomics and 15 ± 2.4 mm^2 for RPPA. Samples were collected into microcentrifuge tubes containing 50 μL of LC-MS grade water and vacuum dried. The LMD harvested tissue was manually transferred into pressure cycle technology (PCT) Micro-Tubes (Pressure Biosciences, Inc) containing 20 μL 100 mM TEAB/10% acetonitrile, pH 8.0, and subsequently lysed and digested with a heat-stable form of trypsin (SMART Trypsin, ThermoFisher Scientific, Inc.) employing pressure cycling technology with a barocycler (2320EXT Pressure BioSciences, Inc). Peptide digests were transferred to 0.5 mL microcentrifuge tubes, vacuum dried, re-suspended in 100 mM TEAB, pH 8.0 and the peptide concentration of each digest was determined using the bicinchoninic acid assay (BCA assay); the mean peptide yield was 0.67 ± 0.16 μg/mm^2. Thirty micrograms of peptide from each sample, along with a reference sample generated by pooling equivalent amounts of peptide digests from individual patient samples, were aliquoted into a final volume of 100 μL of 100 mM TEAB and labeled

with tandem-mass tag (TMT) isobaric labels (TMT11plex Isobaric Label Reagent Set, ThermoFisher Scientific) according to the manufacturer's protocol. Each TMT-11 multiplex set of samples was loaded onto a C-18 trap column in 10 mM NH_4HCO_3 (pH 8.0) and resolved into 96 fractions through development of a linear gradient of acetonitrile (0.69% acetonitrile/min) on a 1260 Infinity II liquid chromatography system (Agilent). Concatenated fractions (36 pooled samples representing 10% of the entire peptide sample) were generated for global LC-MS/MS analysis. The remaining 90% of peptides were pooled into 12 fractions for serial phosphopeptide enrichment by metal affinity chromatography (TiO2 and Fe-IMAC). Briefly, concatenated peptide fractions were vacuum dried, re-suspended in TiO_2 binding/equilibration buffer and bound to TiO_2 affinity spin columns (High-Select TiO_2 Phosphopeptide Enrichment Kit, Thermo Fisher Scientific, Inc), and sample flow-through and washes were reserved for subsequent enrichment by Fe-NTA (nitrilotriacetic acid) affinity chromatography (High-Select Fe-NTA Phosphopeptide Enrichment Kit, ThermoFisher Scientific, Inc).

Mass spectrometry-based proteomics

Liquid chromatography-tandem mass spectrometry (LC-MS/MS) analyses of TMT-11 multiplexes was performed essentially as previously described.[91] In brief, each concatenated TMT fraction (5 μL, ~600 ng) was loaded on a nanoflow high-performance LC system (EASY-nLC 1200, ThermoFisher Scientific, Inc.) employing a two-column system comprised of a reversed-phase trap column (Acclaim PepMap 100 Å, C-18, 20 mm length, nanoViper Trap column, ThermoFisher Scientific, Inc.) and a heated (50°C) reversed-phase analytical column (Acclaim PepMap RSLC C-18, 2 μm, 100 Å, 75 μm × 500 mm, nanoViper, ThermoFisher Scientific, Inc.) connected online with an Orbitrap mass spectrometer (Q-Exactive HF, ThermoFisher Scientific, Inc.). Peptides were eluted by developing a linear gradient of 2% mobile phase A (2% acetonitrile, 0.1% formic acid) to 32% using mobile phase B (95% acetonitrile, 0.1% formic acid) within 120 min at a constant flow rate of 250 nL/min. High-resolution (R = 60,000 at *m/z* 200) broadband (*m/z* 400–1600) mass spectra (MS) were acquired, from which the top 12 most intense molecular ions in each MS scan were selected for high-energy collisional dissociation (HCD, normalized collision energy of 34%) acquisition in the Orbitrap at high resolution (R = 60,000 at *m/z* 200). Peptide and molecular ions selected for HCD were restricted to z = +2, +3 and +4. The S-Lens RF was set to 60, and both MS1 and MS2 spectra were collected in profile mode. Dynamic exclusion (20 s at a mass tolerance = 10 ppm) was enabled to minimize redundant selection of peptide molecular ions for HCD.

Reverse phase protein array (RPPA)

Tissue lysates were kept at −80°C until they were immobilized onto nitrocellulose coated slides (Grace Bio-labs, Bend, OR) using an Aushon 2470 arrayer (Aushon BioSystems, Billerica, MA). Each sample was printed in technical triplicates along with reference standards used for internal quality control/assurance. To estimate the amount of protein in each sample, selected arrays (one in every 15) were stained with Sypro Ruby Protein Blot Stain (Molecular Probes, Eugene, OR) following manufacturing instructions.[72,93] Samples for RPPA analyses exhibited mean protein yields of 1.1 ± 0.24 μg/mm2 tissue area collected.

Prior to antibody staining, the arrays were treated with Reblot Antibody Stripping solution (Chemicon, Temecula, CA) for 15 min at room temperature, washed with PBS and incubated for 4 h in I-block (Tropix, Bedford, MA). Arrays were then probed with 3% hydrogen peroxide, a biotin blocking system (Dako Cytomation, Carpinteria, CA), and an additional serum free protein block (Dako Cytomation, Carpinteria, CA) using an automated system (Dako Cytomation, Carpinteria, CA) as previously descried.[93] Each array was then probed with one antibody targeting an unmodified or a post-translationally modified epitope. Antibodies were validated as previously described.[94] Slides were then probed with a biotinylated secondary antibody matching the species of the primary antibody (anti-rabbit and anti-human, Vector Laboratories, Inc. Burlingame, CA; anti-mouse, CSA; Dako Cytomation Carpinteria, CA). A commercially available tyramide-based avidin/biotin amplification kit (CSA; Dako Cytomation Carpinteria, CA) coupled with the IRDye680RD Streptavidin fluorescent dye (LI-COR Biosciences, Lincoln, NE) was employed to amplify the detection of the signal. Slides were scanned on a laser scanner (TECAN, Mönnedorf, Switzerland) using the 620 and 580 nm wavelength channels for antibodies and total protein slides, respectively. Images were analyzed with a commercially available software (Micro-Vigene 5.1.0.0; Vigenetech, Carlisle, MA) as previously described[72]; this software performs automatic spot finding and subtraction of the local background along with the unspecific binding generated by the secondary antibody. Finally, each sample was normalized to its corresponding amount of protein derived from the Sypro Ruby stained slides and technical replicates were averaged.

RPPA antibody identifiers were mapped to Uniprot protein accessions and HGNC identifiers through manual inspection of commercial antibody names and corresponding human protein entries curated within the UniProt resource (uniprot.org). Pan-specific antibodies were assigned to multiple protein isoform accessions and residues for modified proteins were mapped to curated protein model positions.

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantitative data processing pipeline for MS global and phosphoproteome analyses

Analyses of TMT global and phosphoproteome data were performed as previously described.[91] Briefly, peptide identifications were generated by searching LC-MS/MS data against a publicly available, non-redundant human proteome database (Swiss-Prot, *Homo sapiens*, downloaded 12/0½017) appended with porcine trypsin (Uniprot: P00761) using Mascot (v. 2.6.0, Matrix Science, Inc.) and Proteome Discoverer (v. 2.2.0.388, ThermoFisher Scientific, Inc.) with the following parameters: precursor mass tolerance of 10 ppm, fragment ion tolerance of 0.05 Da, a maximum of two tryptic missed cleavage sites, static modification for TMT reporter ion tags

(229.1629 Da) on N-termini and lysyl residues, and dynamic modifications for oxidation (15.9949 Da) on methionine residues, as well as phosphorylation (79.9663 Da) on seryl, threonyl or tyrosyl residues for phosphoproteome analyses. The resulting peptide spectral matches (PSMs) were filtered using an FDR <1.0% (q-value < 0.01), as determined by the Percolator[95] node of Proteome Discoverer. Phosphoproteome search results were further analyzed by the ptmRS node[96] within Proteome Discoverer as a confidence measure for the post-translational modifications identified.

TMT reporter ion intensities were extracted using Proteome Discoverer at a mass tolerance of 20 ppm, and PSMs lacking a TMT reporter ion signal in TMT channel *m/z* 126 (TMT-126 - the pooled study reference which is a pool of all tumor digests in each sample multiplex), PSMs lacking TMT reporter ion intensity in all TMT channels, or PSMs exhibiting an isolation interference of $\geq$50% were excluded from downstream analyses. Log_2-transformed TMT reporter ion ratios corresponding to individual patient tissue samples were calculated for each PSM against the pooled reference standard (TMT-126). Log_2-transformed PSM abundance distributions were normalized by calculating the mode-centered *Z* score transformation adapted from Mertins P et al.[97] for each channel in the TMT-11 multiplex as follows: normalized PSM (Log_2Ratio) = [PSM (Log_2Ratio) – ModeCenter PSM (Log_2Ratio)/σ PSM (Log_2Ratio). For global protein-level abundance, the abundances of proteins identified by a unique PSM (i.e. in which a PSM maps uniquely to a single protein accession) were determined by calculating the median log_2-transformed abundance ratios of all such PSMs. The abundances of PSMs mapping to multiple proteins (i.e. "multi-mapper" PSMs) were compared to mapped unique protein abundances using a mean-squared-error approach to assign them to unique proteins based on comparative abundance analyses. Multi-mapper PSMs were assigned to the corresponding unique protein accessions exhibiting the smallest difference in relative abundance levels comparatively and candidates not identified by a unique PSM were excluded from downstream analyses. Protein-level abundance was calculated from normalized, median log_2-transformed TMT reporter ion ratio abundances from a minimum of two PSMs corresponding to a single protein accession. Missing abundances for proteins quantified in $\geq$50% of all patient samples were imputed using a k-nearest neighbor (k-NN) strategy (adapted from[97]) using the pamr (Prediction Analysis for Microarrays) R package[98]; proteins quantified in fewer than 50% of samples were not further considered.

The abundances of phosphorylated (phospho)-PSMs were assembled at the level of discrete phosphosites that map to a unique protein using a tiered strategy aimed at defining high- and low-confidence phospho-PSMs. First, TMT reporter ion intensities were processed for phospho-PSMs as described above to calculate normalized, log_2-transformed abundance ratios of phospho-PSMs for a given patient sample. The number and amino acid positions of phosphosites that were identified in the database search for a given phospho-PSM were compared with phosphosite positions predicted by the ptmRS algorithm. A high-confidence phospho-PSM was determined when all phosphosites identified by database search also exhibited >50% probability of being the "best" predicted phosphosite for a given phospho-PSM. A low-confidence phospho-PSM was determined when any phosphosite identified by database search was not predicted as a phosphosite or exhibited <50% probability of being predicted as the "best" phosphosite. Low-confidence phospho-PSM candidates were further prioritized using a tiered strategy in which unique phosphosite variants identified for the same phospho-PSM event were selected based on the highest ptmRS probability score that exhibited the lowest search engine rank in the TMT-11 patient sample plex with the greatest number of total PSMs. Normalized log_2-transformed protein-specific phosphosite abundances were determined by calculating the median abundance of phospho-PSMs exhibiting the same phosphosite as well as methionine oxidation state. Phosphosites quantified redundantly as both low- and high-confidence versions were further filtered to prioritize only high-confidence phosphosites for downstream analyses. For phosphosites co-identified in companion global proteomic data, median log_2-transformed, protein-specific phosphosite abundances were also normalized to the total protein abundance quantified in global proteome analyses.

MS variant peptide identification from patient-specific proteogenomic databases

Patient-specific proteome databases were constructed from germline and somatic tumor mutation calls assembled from whole genome sequencing (WGS) data as well as alternative splice and fusion events assembled from companion total RNA-seq data using QUILTs (https://github.com/ekawaler/pyQUILTS)[78] using the prepare_refseq proteome database workflow against a RefSeq human reference database (downloaded 8/2019, 53,912 entries). MapSplice junctions were retained if they were supported by $\geq$ 10 reads. Search databases were constructed for each TMT-11 multiplex by concatenating the human reference database with patient-specific variant databases corresponding to patient samples analyzed within a given TMT-11 sample multiplex. Protein entries reflecting variants that were redundant across patient samples within a given TMT-11 multiplex were merged into unique database entries using in-house scripts. Global.raw files for a given TMT-11 multiplex were searched using Mascot (Matrix Science) and Proteome Discoverer (ThermoFisher Scientific) software against respective patient-specific databases using identical parameters as described above. Investigation of variant peptides mapping to novel splice junctions were confirmed by BLASTP (nr database for taxid:9606) and showed that all candidates mapped to reviewed or predicted human proteins. Quantified variant peptides (peptides mapping to patient-specific variant, but not reference protein entries) were prioritized for downstream analyses from each patient sample TMT-11 multiplex.

Variant peptide abundances were calculated using identical parameters as described above. Briefly, log_2-transformed TMT reporter ion ratios corresponding to individual patient tissue samples were calculated for each variant PSM against the pooled reference standard and log_2-transformed PSM abundance distributions were normalized by mode-centered *Z* score transformation. We further implemented orthogonal verification strategies to promote high-confidence variant peptide identifications by employing SpectrumAI (https://github.com/yafeng/SpectrumAI).[83] Variant peptides encoding a missense substitution of interest were considered verified if a single tandem MS spectra exhibited fragment ions flanking a substitution of interest was identified in all patient

sample plexes in which a variant peptide was quantified. Variant peptides encoding missense substitutions derived from germline mutation calls were further compared to variant genotypes derived from WGS data with the expectation that variant peptide abundance will be significantly increased (Mann Whitney U $p > 0.05$) in patients homozygous alternate or heterozygous for a given variant versus homozygous reference samples.

Differential global and phosphoproteomics

We identified differentially expressed total proteins and PTMs between tumor groups (e.g. by expression subtypes) using ordinary least-squares regression, controlling for influences of patient gender and ancestry. For PTMs measured by MS, we considered imputed measurements (k-nearest neighbors, at least 50% real measurements) as well as regressions that used the unimputed measurements, requiring a 20% measurement rate in ''case'' tumors and skipping instances with no measurements in ''control'' sets. Unless otherwise noted, proteins and PTMs possessing and FDR q-value < 0.1 were deemed significant.

Germline and somatic variant calling and sample concordance

All WGS samples were initially processed through the Illumina Sequence Analysis Software Resequencing Workflow (version 6.19.1.403 + NSv6). This workflow aligned sequencing reads to the human reference genome (NCBI GRCh38 with decoys) with the Isaac aligner (version Isaac-04.17.06.15)[99] and called germline variants (SNVs and short indels) with Strelka2 (version 2.8.0),[100] structural variants (SVs) with Manta (version 1.1.1),[101] and copy number variants with Canvas (version 1.28.0.272 + master).[102] Initial sample quality features assessed at this stage included total pass fail reads, percent aligned reads, and total coverage depth (target ~30X for germline specimens, ~90X for tumors). We also inferred sample gender and ancestries from DNA evidence and compared these to sample clinical information. We predicted sample gender from chromosome X heterozygous to homozygous variant ratios with a support vector machine (SVM) classifier trained on DNA specimens of known gender. Sample ancestries were inferred using Peddy.[76] Briefly, principal component reduction was performed on genotype calls at specific loci from 2,504 samples in the 1000 Genomes project and an SVM classifier was trained on the resulting first four components, using known ancestries as the training labels. Sample genotype calls at these same loci were then mapped to principal component space and the trained SVM was used to predict underlying ancestries.

We then called somatic variants from matched tumor and normal specimens using the Illumina Sequence Analysis Software Tumor Normal Workflow (version 6.9.1.177 + NSv6), which called somatic SNVs and indels with Strelka2, somatic structural variants with Manta, and somatic copy number alterations with Canvas. Prior to running this workflow, we first verified that matched sample DNA specimens derived from the same individual. For this we ran Conpair[63] on expected matched sample pairs, using the 7,353 autosomal GRCh38 markers provided with the tool. Tumor-normal pair concordances exceeded 95%. We also ran Conpair's contamination estimation tool and found that all final tumor specimens displayed less than 0.5% contamination. In addition to Conpair, we utilized a hierarchical clustering scheme that compared the sample genotypes from the same marker set. This analysis also revealed that all expected sample pairs indeed derived from the same individuals.

Tumor purity was estimated from tumor WGS data by Canvas within the Illumina Tumor Normal Workflow. We identified four cases with 100% tumor purity estimates; subsequent manual review found that these tumors had a normal-like copy number profiles and mean somatic VAFs inconsistent with such high purity estimates. Thus, for these four cases, tumor purity estimates were assigned as two times their mean somatic SNV frequency.

Germline variant pathogenicity analysis

We used InterVar[70] to classify all germline variants. We selected variants called Pathogenic (P) or Likely Pathogenic (LP) and occurring in genes reported as germline mutations by The Cancer Genome Atlas.[103] We further filtered these variants using ExAC,[104] retaining those with an allele frequency less than 1%. Germline mutation calls were then reviewed by a medical geneticist (C.E.T).

Somatic variant filtration and annotation

For somatic SNVs and indels, we retained variants passing all Strelka2 filters (i.e. PASS variants). To further control for potential false positives and/or caller artifacts, we implemented a panel of normals (PON) approach whereby additional somatic variant calls were removed if they were observed in ''pseudo somatic'' call sets derived from the matched normal specimens. For this, we ran the Tumor Normal Workflow on every normal WGS dataset against a synthetic normal sample. This synthetic normal was generated with wgsim (https://github.com/lh3/wgsim) and mimics an ~30X depth WGS dataset with all base calls matching the hg38 reference genome. We ran wgsim with the following parameters to create a set of simulated WGS paired-end FASTQ files with no variants from the reference genome: -N 475,000,000 -d 420 -s 95 -e 0.0–1 150 -2 150 -r 0.0 -R 0.0 -X 0.0. We then ran the Resequencing Workflow on this synthetic dataset and used the resulting outputs as the ''normal'' specimen when running the Tumor Normal Workflow with the true normal WGS samples as the ''tumors''. Passing variants occurring in two or more pseudo somatic samples formed a filter list. These SNV and indel variants were then filtered from tumor somatic variants in the cohort. Finally, we annotated these filtered somatic SNVs and indels with ANNOVAR[60] against GENCODE v28 gene models,[55] to be used throughout the remainder of the study. Tumor mutational burden (TMB) was defined as the sum of SNVs per megabase and indels per megabase.

For somatic copy number alterations (CNAs), we retained segments called by Canvas that passed all caller filters (i.e. PASS variants) for gene-level analysis. For somatic SVs, we ran Break Point Inspector on the Manta calls (https://github.com/hartwigmedical/hmftools/releases/download/bpi-v1-7/bpi-v1.7.jar), which re-examines support for these calls from the tumor and normal sample BAM files directly. Somatic SVs not passing Manta and additional Break Point Inspector filters were discarded.

DNA copy number segments and gene-level copy number values

We used continuous $\log_2$(CN) – 1 (logCN) values as the numerical factors for Canvas segments for several analyses. Here, CN, or the normalized floating point copy number (e.g. 2.0 equals copy number 2), is equal to: (2 * RC)/(DC), where RC is the mean read count per bin reported by Canvas for the segment and DC is the overall diploid coverage value for the sample estimated by Canvas. For autosomes, we used $\log_2$(CN) – 1 values directly; for sex chromosomes, we adjusted the calculations on chromosomes X and Y to $\log_2$(CN) – 0 for male samples, whereas $\log_2$(CN) – 1 was used for chromosome X segments in female samples.

For gene-level copy values, we extracted sample-wise logCN values corresponding to individual gene coordinates. If no Canvas segment overlapped the gene coordinates, logCN was set to 0.0; if a single segment overlapped the gene, this value was used for logCN; otherwise, if more than one segment overlapped the gene coordinates, the segment covering the greatest proportion of the gene was used for logCN.

RNA-seq alignment, quantification, and differential expression analysis

Raw paired-end RNA sequencing reads were aligned to the human reference genome (hg38) using MapSplice[71] (version 2.2.2) with the `-fusion` parameter set. Gene-level read counts against GENCODE (version 28) basic gene models[55] were calculated by HTSeq[69] (version 0.9.1) with the parameters: `-s reverse -t exon -m intersection-nonempty`. For calculating transcripts per million (TPM) values, the average of all transcript lengths corresponding to the same gene were used for gene length factors and only protein-coding genes (identified by the presence of at least one annotated CDS element in its gene model) were considered in the "per million" normalization factor calculations. Read alignment statistics and sample quality features were calculated with SAMtools[81] and RSeQC.[80] Read characteristics (e.g. total reads, mapping percentages, pairing percentages), transcript integrity number (TIN),[105] 5′ to 3′ gene body read coverage slopes, and rRNA content were inspected for sample quality determination.

We used DESeq2[66] (version 1.16.1) to calculate differential gene expression between sample groups (e.g. by subtypes). We generally considered genes to be differentially expressed if their cohort mean TPM expression levels were greater than one, if their estimated adjusted p values were less than or equal to 0.1, and if the absolute values of their $\log_2$ fold-changes following shrinkage (i.e. using the lfcShrink function in DESeq2) were greater than or equal to 0.322.

Sample identity matching

As done for DNA WGS tumor-normal pairs (see *Germline and somatic variant calling and sample concordance*), we conducted analyses to assure appropriate sample concordance between RNA-seq and proteomic datasets with WGS specimens. To assess concordance between RNA-seq and DNA WGS datasets, we extracted genotypes at the ~7,000 loci selected by Conpair.[63] DNA WGS genotypes were extracted from sample VCF files produced by the Illumina workflow and RNA-seq sample genotypes were calculated with UNCeqR.[85] For all pairwise sets of genotypes, we computed distances based on the counts of matching genotype calls and used hierarchical clustering (Euclidean distance and average linkage) to identify sample groups. All normal WGS, tumor WGS, and tumor RNA-seq corresponding to the same individual were clustered together in triplicate groups.

After assuring appropriate concordance between DNA WGS and RNA-seq samples, we compared RNA-seq and proteomic datasets by expression correlation. We first identified linked RNA transcripts and proteins by gene symbol mapping and restricted analyses to only these matched pairs. For each dataset, we *Z* score transformed individual species' expression levels prior to comparisons. We then computed Spearman correlations between pairwise vectors of normalized sample expression levels for matched species. To confirm sample identity, we determined the best and expected sample-wise correlations among RNA-protein dataset comparisons. For MS total proteomics datasets, we observed a median dataset-wise Spearman correlation of 0.48 with matched RNA-seq datasets and, for the vast majority of cases (85 of 87), the best sample-to-sample correlations derived from expected sample pairs; in cases where this was not true, the expected correlation was highly numerically similar to the best correlation. Thus, this analysis confirmed appropriate sample-wise matching between RNA-seq and MS proteomic datasets.

Comparisons between MS total proteins and RPPA antibodies targeting total proteins (83 total species) revealed a median tumor-wise Spearman correlation of 0.42. When further restricting this analysis to individual proteins with strong MS to RPPA correlations (e.g. protein-wise $\rho > 0.4$), we observed a stronger median sample-wise correlation of 0.7. In the latter analysis, the majority of samples (81 of 87) were most strongly correlated with their expected pairs, while the remaining cases showed numerically similar correlations between their expected pairs and their observed bests, which may be related to the low number of targets on the RPPA platform. Protein-wise Spearman correlation across the 83 common proteins between the platforms was 0.36.

RNA-protein correlations

RNA:protein correlations were determined between $\log_2$(TPM +1) RNA expression data and total proteomic MS data. For species matching by gene symbol between the two datasets (n = 7,472), we row-standardized (i.e. z-scored) RNA and protein measurements

and calculated gene-wise Spearman correlations. We then corrected correlation p values for multiple hypotheses using the Benjamini-Hochberg FDR procedure and deemed RNA-protein pairs with FDR <0.05 and a positive Spearman correlation coefficient ($\rho > 0$) as significantly positively correlated. We identified enriched pathways and biological processes among uncorrelated species ($\rho < 0$ and FDR >0.05) and strongly positively correlated species ($\rho > 0.5$ and FDR <0.05) by running GO term enrichment on these foreground gene sets against all identified matched RNA-protein pairs.

For further RNA and protein comparative analysis, we restricted to genes with at least 2 mean TPM by RNAseq, yielding 7,322 genes, referred to as 'co-expressed genes'. Then, we calculated tumor-wise correlations as the Spearman correlation between a tumor's RNA expression and protein expression using co-detected expressed genes. In the Gillete et al. cohort gene-wise and protein-wise correlations were calculated similarly. RNAseq data was already restricted to expressed genes in Gillette et al.[13] Using species common to both Gilette et al. RNA and protein platforms (n = 10,069) gene-wise and tumor-wise Spearman correlations were calculated. Gene-wise correlations between APOLLO and Gilette cohorts were compared on common, expressed genes between the cohorts, n = 6,729 genes.

Somatic mutation recurrency analysis

We used MutEnricher (version 1.3.1)[74] to interrogate coding genes and non-coding regulatory elements for recurrent somatic mutations. For coding genes, we considered GENCODE (version 28) basic gene annotations,[55] skipping genes encoded on chromosomes Y or M and restricting to those with at least one coding domain sequence (CDS) and whose median RNA transcript expression across the full cohort was greater than one transcript per million (14,757 genes in total). For variant impact annotations, we ran ANNOVAR[60] on the filtered sample somatic VCF files against GENCODE v28 gene models. When running MutEnricher's coding module, we used the covariate clustering method to compute background mutation rates. As covariates we used gene full lengths, coding lengths, sequence GC and CpG contents, median RNA expression levels, and replication timing values from 23 different Repli-Seq datasets across a variety of cell lines obtained from ENCODE (https://www.encodeproject.org/; retaining the mean value per gene from each dataset). In addition, we used MutEnricher's default binomial testing strategy (i.e. stat-type = nsamples), which computes the significance of observing *n* samples out of *N* total with non-silent somatic mutations in a gene, and, for hotspots, restricted testing to regions with at least three non-silent somatic mutations from at least three samples. All other parameters were set to their default values. In cases where subsets of samples were run through MutEnricher (e.g. by expression subtypes), the same parameters and gene covariate clustering results were used. Genes with overall burden or combined burden plus hotspot FDRs <0.1 were deemed significant, except where otherwise noted.

For non-coding elements, we interrogated LUAD-specific non-coding regulatory regions (promoters, 3′ UTRs, 5′ UTRs, and distal enhancers) identified by ATAC-seq.[31] For each element type, we removed regions overlapping any annotated CDSs and independently ran MutEnricher with the local background mutation rate method, which scans the local genomic neighborhood surrounding each non-coding element to compute a per-sample background rate. Again, we used MutEnricher's binomial testing strategy and tested candidate hotspots with a maximum distance between mutations of 50 basepairs and with at least three somatic mutations from at least two samples. All other parameters were set to their default values. Non-coding elements and/or hotspots were deemed significant if their FDR-corrected p values were less than 0.1.

Copy number alteration recurrency analysis

We used GISTIC2.0[68] to identify recurrent somatic copy number alterations among all tumor samples. As input, we used sample $\log_2$(CN) – 1 (logCN) values from copy number segments identified by Canvas (see above). As parameters we used: -genegistic 1 -armpeel 1 -brlen 0.98 -conf 0.99 -scent median -maxseg 5000 -ta 0.1 -td 0.1 -rx 0. All other parameters not explicitly indicated here were set to their defaults. We considered events with q-value < 0.25 to be significant (i.e. the default GISTIC threshold).

Somatic structural variant clustering

To identify recurrent somatic structural variants among tumor samples, we devised a clustering approach whereby SVs were grouped by the consistency of their breakend coordinates. For this cohort-wide analysis, intrachromosomal SVs (i.e. deletions, tandem duplications, insertions, and inversions) were handled separately from interchromosomal SVs (i.e. translocations) as we aimed to compare events targeting the same or similar genomic loci. We considered an intrachromosomal SV in the analysis if it passed Manta and Break Point Inspector filters (i.e. PASS), if its length was >30 bp, and the sum of the somatic paired read (PR) and split read (SR) evidences for the SV was $\geq$5. Deletion SVs less than 1000 bp in length were also excluded from the analysis and reciprocal inversion calls annotated as the same "event" by Manta were merged into singular calls, retaining the mean start and end coordinates of the source SVs. In addition, we filtered SVs whose breakend coordinates did not fully intersect with a mappable genomic locus defined by the Genome in a Bottle Consortium (version 3.2.2 with hg38 genome lift over; ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.2.2). Once the intrachromosomal SVs were loaded, we first identified overlapping SV intervals (with help from the interlap Python package: https://github.com/brentp/interlap) as well as closely neighboring intervals (e.g. within D_T bp). Next, for all intervals comprising more than one SV, we used single linkage hierarchical clustering with the following distance function (D_{ab}) to identify SV clusters:

$$D_{ab} = \min(d_1, d_2)$$

$$d_1 = d_{ss} * d_{ee} = (|a_s - b_s| / 3e9) * (|a_e - b_e| / 3e9)$$

$$d_2 = d_{se} * d_{es} = (|a_s - b_e| / 3e9) * (|a_e - b_s| / 3e9)$$

where *a* and *b* subscripts index two SVs and *s* and *e* subscripts indicate the start and end coordinates, respectively, of the two SVs ($s < e$ for all SVs). SV clusters were then called if the linkage distance between formed groups was less than a threshold C_T, where $C_T = (D_T/3e9)^2$. Any SVs not part of a cluster were reported as singletons. For identified SV clusters, we calculated the significance of observing *k* or more SVs as part of a cluster using a Poisson test, where the Poisson μ (mean) parameter was calculated as the average number of SVs within a cluster among non-singleton clusters.

For interchromosomal SVs (i.e. translocations), we again retained SVs passing both Manta and Break Point Inspector filters as well as those with PR + SR $\geq$ 5. To find clustered translocations, we created a breakend graph whereby an edge between two SVs was established if the two involved chromosomes matched and the distance D_{ab} between the two was less than C_T, as defined above. Translocation clusters were then identified from this breakend graph by identifying connected components within; any non-connected SVs were classified as singletons. Breakend graph building and manipulation were performed within the NetworkX package in Python (https://networkx.org/).[75] In addition, we computed Poisson p values for the translocation clusters as described above for intrachromosomal SVs.

For these analyses we used a distance threshold (D_T) of 10 kb. For some analyses, we also included fusion genes called from sample RNA transcript data. For these, we used well-annotated fusion calls made by MapSplice. In an alternative analysis, we identified clusters of somatic SVs on a per-sample basis that allowed for "chains" of SVs. Here, somatic SVs were filtered using the same criteria described above and a breakend graph was built among all SV types (e.g. both interchromosomal and intrachromosomal SVs). An edge between two SVs was created in the graph if either of the two breakends of one SV intersected a breakend of another within a distance threshold D_T. SV clusters were identified from this graph by finding connected components and non-connected SVs were deemed singletons.

Somatic signatures analysis

We used MultiModalMuSig,[20] a multi-modal correlated topic modeling (MMCTM) approach implemented in the Julia programming environment (version 1.1.0), to jointly learn somatic mutation signatures from patient SNV, indel, and structural variant profiles. For each tumor sample, we extracted the counts of 1) somatic SNVs within their tri-nucleotide contexts (96 possible) using the deconstructSigs R package (version 1.8.0),[65] 2) somatic short indels classified by size, affected nucleotides, and repeat/microhomology contexts according to classes defined by COSMIC (https://cancer.sanger.ac.uk/cancergenome/assets/PCAWG7_indel_classification_2017_12_08.xlsx), and 3) somatic structural variants grouped by alteration type (deletions, tandem duplications, insertions, inversions, and interchromosomal translocations) and size (<1 kb, 1–10 kb, 10–100 kb, 100 kb–1 Mb, 1 Mb–10 Mb, and >10 Mb; applicable to intrachromosomal SVs only). To determine the optimal number of signatures (*K*) for each variant type (i.e. mode), we independently ran 100 iterations of MultiModalMuSig for each mode across a range of possible *K* (2–10; MultiModalMuSig parameters $\alpha = 0.1$ and tolerance = 1×10^{-6} for all modes) and assessed model log-likelihoods against each value of *K*. From this analysis we chose 3 SNV, 3 indel, and 4 SV signatures as the optimal per-mode *K* values. We then ran the algorithm 1000 times on the three modes jointly using their optimal *K* values ($\alpha = 0.1$ for all modes and tolerance = 1×10^{-7}), retaining the maximum scoring (i.e. highest log likelihood) result. Following identification, we mapped the SNV and indel signatures to known COSMIC (version 3) single base substitution (SBS) and small insertion and deletion (ID) signatures, respectively, using ridge regression and cosine similarity mapping. With the ridge regression procedure, all known COSMIC signatures were considered in a single model, whereas signatures were mapped individually with the cosine correlation method.

We then used hierarchical clustering (Euclidean distance and Ward linkage) to identify groups from the patient signature probabilities. This analysis revealed three distinct clusters. We additionally compared these signature cluster assignments to several clinical and molecular features, including smoking history and somatic mutation rates, using ANOVA for continuous variables and χ^2 statistics for categorical features.

Somatic quantitative trait loci (QTL) analysis

We performed quantitative trait loci (QTL) analyses to identify coding and non-coding somatic mutation impacts on RNA and/or protein expression. For all analyses, we used elastic net, a regularized regression procedure that combines both the L1 and L2 penalties on estimated coefficients, to evaluate mutation impacts on expression:

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}}\left(exp - \boldsymbol{X}\beta^2 + \lambda_1\beta_1 + \lambda_2\beta^2\right)$$

For coding analyses, we considered recurrently mutated genes (MutEnricher Fisher FDR <0.1) and grouped mutations within these genes by their variant impact types (i.e. nonsynonymous SNVs, stopgains, frameshift alterations, etc.). For each of these genes we built a regression model according to:

$$\exp_g = \beta_0 + \beta_{mut1}M_1 + \beta_{mut2}M_2 + \ldots + \beta_{mutn}M_n + \beta_{CNA}C + \beta_{anc}A + \beta_{sex}S + \beta_{hidden}H$$

where β values represent regression coefficients for mutation types (e.g. M_1, M_2, … M_n), gene copy number levels (C, sample $\log_2$(CN) – 1 values for gene g), background ancestry (A, either EUR or AFR for this cohort), gender (S, male or female), and additional hidden factors (H; see below). A similar approach was used for non-coding loci:

$$\exp_g = \beta_0 + \beta_{l1}L_1 + \beta_{l2}L_2 + \ldots + \beta_{ln}L_n + \beta_{CNA}C + \beta_{anc}A + \beta_{sex}S + \beta_{hidden}H$$

where L variables represent recurrently mutated non-coding loci linked to gene g. If a non-coding element contained one or more significant hotspot regions (FDR <0.05), mutations contained within these were considered part of a distinct element in the regression; then, if the full non-coding element was also significant (burden FDR <0.2), the remaining mutations not part of any significant hotspots were included as an additional locus. If no significant hotspots were identified but the overall region was deemed significant (burden FDR <0.2), the full element was included as a single term in the regression. For locus-gene links, we considered *cis* mappings, i.e. genes within +/− 1 Mb of the locus. Ultimately, we considered 426 regulatory elements mapped to 4,767 *cis* genes, creating 7,019 total locus-gene links. We note that this overall QTL approach to non-coding mutations is similar to methods described in Zhang et al.,[106] though with modifications.

For RNA transcript expression levels, we ran DESeq2's (version 1.16.0)[85] variance stabilizing transformation (VST) on the raw sample gene counts. We then ran probabilistic estimation of expression residuals (PEER; obtained from https://github.com/PMBio/peer for Python)[77,107] on the VST-normalized data to estimate the hidden factor variables in the QTL regression models (i.e. H factors). We estimated 10 potential hidden factors with PEER, including a mean term along with the known sample covariates gender and ancestry during estimation. After fitting, we examined a plot of the posterior variance of factor weights against the number of estimated hidden factors; this analysis indicated a natural choice of four optimal hidden factors to include in regressions involving RNA expression measurements. We also ran this procedure on the global proteomics data to estimate hidden factors for regressions involving total protein measurements. Here, we used the normalized and log-transformed total protein measurements as input to PEER, again considering up to 10 potential factors and running in the context of known covariates for gender and ancestry. Examination of the posterior factor weight variances additionally revealed four optimal hidden factors for use in protein QTL regressions. These estimated factors were included directly into subsequent regressions.

All regression were run in Python with the ElasticNet linear regression module implemented in the Scikit-learn package.[82] To estimate the optimal L1 and L2 penalty weights (i.e. λ_1 and λ_2), we ran regressions over a range of values for the elastic net parameters (alpha, l1_ratio); in this implementation, λ_1 = alpha *l1_ratio and λ_2 = 0.5 * alpha * (1 - L1_ratio). For all tested (alpha, l1_ratio) combinations, samples were randomly divided 50 times into 70% training and 30% testing sets and the mean training error (negative mean squared error) was recorded; the optimal set of parameters was then taken as the minimum error combination. We tested 120 logarithmically spaced values for alpha that ranged from 1×10^{-4} to 89.125 and 21 L1_ratio values ranging from 0 to 1 in 0.05 increments.

After fitting with the optimal parameter sets, we used F-statistics to derive an overall mutation effect p value for each tested gene/protein as well as individual mutation term significances (e.g. for multiple coding mutation types or multiple non-coding loci). For the overall effect, we compared the full model accuracy against that of a model fit with the same elastic net parameters but without any mutation terms. Similarly, for individual mutation terms, we compared the full model accuracy against a model excluding each mutation term individually. We then multiple hypothesis corrected (Benjamini-Hochberg) the overall mutation effect p values for each gene/protein separately from the individual mutation type to gene or locus to gene p values. For coding mutation analyses, we considered genes/proteins significant if their overall mutation effect FDR was less than 0.1 and if they possessed at least one mutation type with an FDR <0.1. Similarly, for non-coding mutations, we considered genes with overall FDR <0.1 and at least one locus-gene interaction effect with FDR <0.1. In coding mutation volcano plots, we computed an aggregate effect size estimate for each gene/protein as the significance-weighted average of the individual mutation effect terms (i.e. $w_i = log_{10}(p\text{-}value_i)/sum(log_{10}(p\ values))$; aggregate effect size = sum(β *w)). For non-coding plots, we reported individual locus-gene p values and effect sizes directly from the elastic net regressions.

Tumor expression subtyping

Tumor RNA expression subtype assignments were calculated using the LUAD subtype predictor (https://github.com/mwilkers/lungCancerSubtypes/blob/main/lung_adenocarcinoma_subtypes/wilkerson.2012.LAD.predictor.centroids.csv).[8] Tumor RNA expression data were upper quartile read counts per gene, increased by one pseudocount, and log2 transformed. The maximum Pearson correlation of a tumor's gene median centered expression to the predictor centroids defined that tumor's subtype assignment, using common genes between the predictor centroids and the APOLLO RNA data (475 of 506 genes). Subtype assignments based on total protein expression were calculated in similar fashion. Proteomic expression were collapsed to gene symbols by taking the average of peptide mapping to the same gene symbol, and gene median centered. The maximum Pearson correlation of a tumor's protein expression to the predictor centroids defined its protein subtype assignment, using common genes between the predictor centroids and the APOLLO protein data (317 out of 506 genes).

Unsupervised RNA expression subtypes were detected with the same expression matrix but restricting to genes having a cohort mean TPM of at least 2 and annotated as coding with a complete open reading frame in GENCODE v28 (n = 14,374 genes). The top 3,000 genes by median absolute deviation were selected for unsupervised clustering and median centered. Consensus clustering was performed by ConsensusClusterPlus[62] with the following options: distance metric – Pearson, clustering algorithm – hierarchical, item resampling 80%, and 1,000 repetitions, supporting three clusters. Unsupervised protein expression subtypes were detected

similarly with gene-symbol collapsed protein expression data. The top 3,000 proteins by median absolute deviation were selected for ConsensusClusterPlus analysis using the same options, also supporting three clusters. Unsupervised phosphoproteomics expression subtypes were detected similarly using ConsensusClusterPlus using phosphoproteins that were detected in at least 50% of samples.

Joint and Individual Variation Explained (JIVE) decomposition was performed on tumor RNA and tumor proteomics expression data using r.jive,[79] identifying a joint rank of 3. RNA data were TPM, pseudocount incremented, log2 transformed. Protein data were gene symbol collapsed proteomics data. Data were reduced to genes detected by both RNA and protein. Clustering on the first 3 joint rank vectors by partitioning around medoids and a k = 3 defined joint unsupervised subtypes.

TP53 mutation signature

Tumors were scored according to a published pan-cancer *TP53* overexpression gene signature (*TP53* mutant vs wildtype, 20 genes).[108] Similar to prior work,[109] $\log_2$(TPM +1) expression values for these genes were standardized to z-scores, and the mean of a tumor's z-scores served as as the mutant *TP53* score per tumor.

Cell type estimation

We used the deconvolution methods CIBERSORT[61] and ESTIMATE[67] to estimate tumor immune and other cell type proportions from bulk RNA-seq datasets. In all cases, we used sample transcripts per million (TPM) values for protein coding genes as the input data to these algorithms. We ran CIBERSORT on the web (version 1.06; https://cibersort.stanford.edu/), using both relative and absolute modes together and disabling quantile normalization (as recommended for RNA-seq data), against the LM22 immune cell types signature set for 1000 permutations. ESTIMATE version 1.0.13 was also used to compute tumor immune and stromal scores. When comparing deconvolution features (e.g. absolute immune scores) between sample groups (e.g. by expression subtypes), we used either t-statistics or non-parametric statistics (Mann-Whitney U test) to compare mean differences and the Benjamini-Hochberg procedure to correct for multiple comparisons. We also used ESTIMATE with input of normalized gene-level proteomics data.

Gene ontology and other pathway enrichment analyses

Gene ontology (GO) and other pathway enrichments for differentially expressed genes and proteins were calculated with hypergeometric tests against a background of all expressed genes (mean TPM $\geq$1) or all measured proteins. For GO, we considered terms with less than or equal to 1000 total genes in the GO database downloaded on November 16, 2018. We also obtained gene sets from the Molecular Signatures Database (MSigDB; version 7.0; http://www.gsea-msigdb.org/gsea/downloads.jsp),[110] including hallmark,[111] C2 curated (canonical, KEGG, and Reactome pathways), and C3 regulatory target (transcription factor and miRNA targets) gene sets. Raw p values calculated for each term set were adjusted for multiple hypotheses using the Benjamini-Hochberg FDR procedure.

Multi-omics gene set analysis

We used the multi-omics gene set analysis (MOGSA) software package (version 1.22.1) in R to perform multivariate single sample gene-set analysis.[73] Within this framework, we calculated integrated single sample MSigDB hallmark gene-set pathway scores (GSS) from sample transcriptomic, global proteomic, and phosphoproteomic data. To identify pathways enriched in specific sample groups (e.g. by expression subtypes), we first selected pathways in which individual sample GSS FDRs were smaller than 0.01 in 50% of all samples. From these pathways, we used generalized linear models (GLMs) to estimate the difference in sample GSS values between tumors in a group of interest against all others, selecting pathways with FDR <0.05. For visualization, we selected representative pathways ranked by GLM T values.

Survival analysis

Patient overall survival (OS) was defined as the interval from surgery to death or last follow up. This interval was censored at 5 years and patients deceased within 30 days were removed from OS analyses. Metastasis-free (MFS) survival time was defined as the time interval from surgery until the first appearance of metastasis to a distant organ site (brain, bone, adrenal, liver, colon, contralateral lung, pancreas) or death. Patients with MFS <30 days were removed from analyses involving this feature and MFS was censored at 5 years.

Categorical features (e.g. tumor subtype assignments) were tested for associations with OS and MFS by log-rank tests. Continuous data (e.g. expression of individual RNAs, proteins, phosphoproteins) were tested against OS and MFS by Cox proportional hazards model and Wald tests for the given molecular marker. False discovery rates were calculated by the Benjamini-Hochberg method. MS phosphoproteins were tested if detected in at least 50% of samples.

Using gene sets significantly associated with MFS (FDR <0.25), survival expression signatures were defined for these gene sets and expression measures from RNA or protein by the following function, similar in form to prior studies[112]:

$$Score = \sum_{i=1}^{k} \beta_i * expression_i$$

in which k is the size of the RNA/protein set, β is the log hazard ratio for RNA/protein i, and *expression* is a given tumor's RNA or protein expression for that gene. Survival expression signatures were defined using the APOLLO cohort to establish training performance and subsequently tested against the CPTAC cohort. The majority of survival signature genes/proteins from the APOLLO cohort were also available in the CPTAC cohort.

Phosphoprotein and kinase enrichment analysis

For phosphoproteomic analyses, we combined phosphosite data quantified by phospho MS and RPPA, requiring phosphosite measurements in 50% of samples, imputing missing values with k-nearest neighbors (K = 10), and excluding MS peptides/RPPA antibodies mapping to multiple parent proteins, to make an integrated phosphoproteomic dataset. We derived kinase-substrate (i.e. phosphosite) links from PhosphoSitePlus,[58] retaining kinases with at least three measured substrates in our phosphorylation data. We used three methods to infer active kinases from target phosphorylation site data:

1. *Kinase-substrate regression*: We built a Ridge regression model for each patient that inferred kinase activities from phosphosite abundance measurements according to:

$$\underset{\beta_i}{\operatorname{argmin}}\, p_i - \boldsymbol{X}\beta_{i2}^2 + \lambda\beta_{i2}^2$$

where p_i are median centered phosphosite abundances for patient i (i.e. $p_i = p_o - \text{median}(\mathbf{p})$), $\boldsymbol{X}$ is a binary matrix of kinase-phosphosite links ($x_i = 1$ if a link exists, 0 otherwise), and β_i are kinase regression coefficients. For all regressions, we set the L2 regularization parameter λ to a fixed value of 0.1. Following fitting, we computed a score for each kinase against each patient as the sign of the kinase regression coefficient multiplied by the negative base 10 logarithm of the coefficient p value (i.e. score = -sign(b_{ik}) x $\log_{10}$(p value)). To infer differential kinase activities between groups of samples (e.g. by expression subtypes), we compared kinase scores between a target group of interest against scores from remaining samples not in this group using t-statistics.

2. *Kinase-substrate enrichment analysis (KSEA):* For this method, we performed differential phosphosite expression analysis between groups of samples (e.g. by expression subtypes) using t-statistics and ranked these results in descending order according to: score = -sign(b_{ik}) x $\log_{10}$(p value). We then ran 100,000 permutations of pre-ranked GSEA (version 4.0.3)[106] with these values against PhosphoSitePlus kinase-phosphosite links. We retained p values and associated normalized enrichment scores for kinases up-regulated and down-regulated in the sample groups according to the ranked phosphosite data.
3. *Hypergeometric test (HGT) kinase enrichment*: With this method, we computed differential phosphosite expression (as in (2) above) and performed hypergeometric tests of up-regulated (regression coefficient >0 and FDR <0.25) and down-regulated (regression coefficient <0 and FDR <0.25) kinase targets against backgrounds of all tested phosphosites. From this analysis we retained HGT enrichment p values and foreground enrichment levels, the latter values signed as positive if the p value for the up-regulated phosphosite set was less than the p value for the down-regulated set and negative otherwise.

We then combined enrichment p values from these three methods using Fisher's method and called significant kinases as those with Benjamini-Hochberg FDR-corrected Fisher p values < 0.01 and whose effect size estimates among the three inference methods all possessed the same sign.

Transcription factor enrichment analysis

We inferred transcriptional regulator activities from *cis* transcription factor (TF) motif matches near differentially regulated RNA transcripts. To better isolate transcription factor-specific effects on gene expression, we regressed out the influences of somatic copy number alterations and recurrent somatic mutations from patient expression data (i.e. using the residuals from somatic QTL analyses). We obtained TF motif position frequency matrices (PFMs) from JASPAR[56] (2020 database, non-redundant vertebrate set, http://jaspar.genereg.net/downloads/) and ENCODE[53] (http://compbio.mit.edu/encode-motifs/) and used TAMO[84] to process and store this data.

We used TCGA LUAD ATAC-seq data to define active regulatory regions[31] within −50/+10 kb of expressed gene transcription start sites and identified motif matches within these. For each motif, we scanned each region and computed a normalized log likelihood ratio (LLR) score as $LLR_{norm} = (LLR - LLR_{min})/(LLR_{max} - LLR_{min})$ for every k-base-pair sub-sequence, where k is the length of the motif, and a motif match was called if LLR_{norm} was greater than or equal to the TFM-Pvalue[113] computed score threshold corresponding to a match p value < 1×10^{-6} for that motif against the human genome. We retained the best scoring motif match for each motif from each region.

To identify TF motif to gene links, we computed a distance-weighted affinity score for each motif against each gene as:

$$MA_{m,g} = \sum_{i=1}^{n} LLR_{norm-m,g,i} * e^{-d_{m,g,i}/D_o}$$

where $LLR_{norm\text{-}m,g,i}$ is the motif match score for motif *m* against gene *g* from *cis* regulatory region *i*, $d_{m,g,i}$ is the absolute distance in basepairs from the midpoint of the motif match location to the TSS of gene *g*, and D_o is the exponential distance constant (set to 10,000 bases for all motifs). We then created a binary motif-gene match matrix ($\boldsymbol{X_{TF}}$) by setting *MA* scores greater than or equal to 0.5 to 1 for that motif-gene link (0 otherwise). Motifs with links to fewer than 10 genes were removed from the final matrix.

We implemented two approaches for finding significant TF enrichments based on discovered motif-gene links and the alteration-adjusted expression data:

1. *Motif regression:* We used a linear regression framework to infer TF activities for each TF motif in each tumor sample (similar in principle to[114]). As TF motifs can show high levels of similarity (i.e. collinearity), we created univariate ordinary least squares (OLS) regression models for each TF motif separately. Patient-specific gene expression values were taken as robust *Z* score values (median centered and median absolute deviation scaled) against the full sample set. Patient TF motif scores were then taken as the sign of the OLS regression coefficient multiplied by the negative log10 p value.
2. *TFSEA*: For this method, we performed differential ANOVA stabilized and alteration-adjusted (i.e. somatic mutations and copy alterations) gene expression data between groups of samples using t-statistics and ranked these results in descending order according to: score = -sign(b_{im}) x $\log_{10}$(p value). We then ran 10,000 permutations of pre-ranked GSEA (version 4.0.3)[57] with these values against the motif-gene links stored in $\boldsymbol{X_{TF}}$ and retained p values and associated normalized enrichment scores for TF motifs up-regulated and down-regulated in the sample groups according to the ranked expression data.

We then combined enrichment p values from these two methods using Fisher's method. As the TF motifs can be highly similar (i.e. non-independent), we considered significant TF motif enrichments as those possessing a raw Fisher p value $< 1 \times 10^{-4}$ and an absolute motif regression coefficient greater than 0.75.

We report JASPAR motif enrichments for the main results; the ENCODE motif enrichments generally revealed similar overall motif type/group results, confirming the findings based on JASPAR motifs. For downstream analyses, we mapped enriched motifs to TF proteins. To do this we first took direct protein mappings reported by JASPAR. Next, we examined JASPAR motif clustering results to identify highly similar motifs and additionally used these protein mappings. Finally, we interrogated the human transcription factors database[115] (http://humantfs.ccbr.utoronto.ca/) to identify inferred TF mappings of these motifs based on amino acid similarity >75% to human and mouse proteins.

Integrative network modeling

We built network models integrating the wide array of omic data collected in this study, including enriched protein kinases, transcription factors, mutated/altered genes, phosphorylation sites, global proteins, and pathways. We focused these efforts on modeling data associated with samples assigned to expression subtypes (i.e. proximal inflammatory, proximal proliferative, and terminal respiratory unit subtypes).

Interactome building: We created a combined interactome from which network models were generated, integrating kinase-phosphosite links from PhosphoSitePlus,[58] MSigDB Hallmark protein-pathway associations (version 7.0)[102], and protein-protein interactions from the Reactome functional interactions (ReactomeFI) database[59] (version 071,718 from https://reactome.org/download-data). All reported interactions from PhosphoSitePlus and MSigDB Hallmark sets were used. In addition, we manually included a kinase-substrate link from STK11 to T172 of PRKAA2 based on established evidence.[116,117] For ReactomeFI interactions, we updated gene symbols to valid HGNC identifiers (as of May 24, 2019) and excluded interactions annotated as "predicted," interactions annotated as expression regulation without further annotation as activation, catalysis, or complex, and all interactions with ubiquitin (UBC gene symbol).

We used the NetworkX Python package to create and store this interactome as a directed graph object.[75] Kinase-phosphosite interactions were included as directed edges from kinases to substrates, protein-pathway links were encoded as directed edges from proteins to pathways, and protein-protein interactions were included according to ReactomeFI annotations, e.g. "-" edges were encoded as undirected edges between species, "→" edges were included as directed activation edges from protein A to protein B, "-|" edges were included as directed inhibitory edges from protein A to protein B, etc.

Network model creation: We extracted omic data species associated with subtype groups (e.g. enriched kinases, differential phosphosites, etc.) and labeled these data nodes as "terminals." We built network models around these terminal sets by extracting short paths between specific data types, linking upstream enriched kinase and significantly altered protein nodes to downstream active transcription factors and enriched pathways (similar in principle to the KiPNA method described in Brubaker and Paulo et al.,[118] but extended here).

Beginning with enriched kinases, we identified links between these and differential phosphosites; when found, we created an automatic kinase → phosphosite → parent protein path. From the parent protein, we extracted simple short paths targeting enriched transcription factors using the NetworkX all_simple_paths function with cutoff = 2. These short paths were included in the output network model if all nodes along the path were part of the terminal set, with the exception of parent proteins of enriched phosphosites. To include nodes associated with enriched somatic alterations (if not already included), we identified simple paths between these proteins and enriched kinases and/or transcription factors (cutoff = 3 in NetworkX function). Again, we required all nodes identified along such candidate paths to be part of the initial terminal set for inclusion in the final network. Lastly, we interrogated the

terminal set nodes for links to enriched pathways and included these when found. We then scanned the current network model proteins for additional enriched phosphosites in the terminal set mapping to these proteins and included these in the final network model.

Network clustering and visualization: We visualized networks in Cytoscape.[64] To simplify interpretation, we applied the GLay Girvan-Newman community clustering procedure[119] to the networks, retaining inter-cluster edges (as implemented in the clusterMaker Cytoscape plugin - http://www.cgl.ucsf.edu/cytoscape/cluster/clusterMaker.shtml). From these clustered models, we selected subnetworks around specific enriched kinases, transcription factors, and pathways for final visualization and presentation.

All p values are from two-sided tests calculated with R, except where indicated.