HHS Public Access

Author manuscript

Cancer Cell. Author manuscript; available in PMC 2026 March 25.

Published in final edited form as:

Cancer Cell. 2025 September 08; 43(9): 1731–1757.e17. doi:10.1016/j.ccell.2025.07.011.

Integrative Analysis of Lung Adenocarcinoma Across Diverse Ethnicities and Exposures

A full list of authors and affiliations appears at the end of the article.

Summary

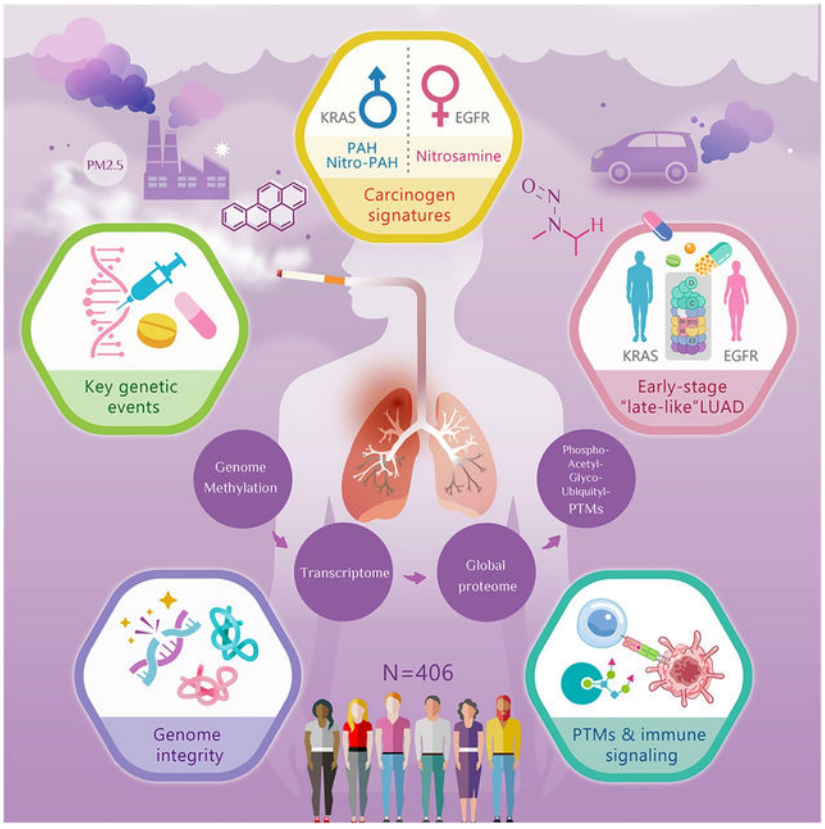
Lung adenocarcinomas (LUAD) are a pressing global health problem with enduring lethality and rapidly shifting epidemiology. Proteogenomic studies integrating proteomics and post-translational modifications with genomics can identify clinical strata and oncogenic mechanisms, but have been underpowered to examine effects of ethnicity, smoking and environmental exposures, or sex on this heterogeneous disease. This comprehensive proteogenomic analysis of LUAD tumors and matched normal adjacent tissues from 406 patients across diverse geographic and demographic backgrounds explores the impact of understudied driver mutations, prognostic role of chromosomal instability, patterns of immune signaling, differential and sex-specific effects of endogenous mutagens and environmental carcinogens, and pathobiology of early-stage tumors with “late-like” characteristics. Candidate protein biomarkers are proposed for unstable tumors with highly fragmented genomes and for carcinogen exposures, and a LUAD subtype-specific atlas of therapeutic vulnerabilities is presented. These observations and the associated data resource advance the objective of precision management strategies for this devastating disease.

Graphical Abstract

shankha@broadinstitute.org (S.S.), slyu@ntu.edu.tw (S.L.Y.), hychen@stat.sinica.edu.tw (H.C.), manidr@broadinstitute.org (D.R.M.), mcieslik@med.umich.edu (M.C.), yujuchen@gate.sinica.edu.tw (Y.J.C.), gillette@broadinstitute.org (M.A.G.).

Author contribution

Sample processing and QC: YiJC, DIH, HF, YA, ESC, HT, CY, YW, GH, MT, AIR; Data generation: SS, YiJC, JW, JA, SA, MHK, MH, EJB, DM, HF, HK, KS, MH, RM, QKL, SMD, AM, SG, JRT, ESC, HH, YH, ZL, HZ, AIR; Data freeze: NMC, YiJC, NH, YHC, YH, YG, DIH, HC, YA, AIN, DRM, MPC; Data interpretation: SS, NMC, YiJC, NH, YC, YH, JTL, FP, JC, YG, IP, HC, SMD, RM, QKL, GM, KL, EJJ, MH, KRC, KR, RG, GSO, PP, AM, AIR, AJL, PY, AIN, DRM, MPC, YuJC, MAG; Data analysis: SS, NMC, YiJC, NH, YHC, YH, JTL, FP, YG, IP, GXL, HC, SMD, RM, MH, GM, KL, WL, SA, GH, EJJ, BM, ED, AK, JL, YC, LD, AM, AIR, DRM, MPC, YuJC, MAG; Figures: SS, NMC, YC, NH, YHC, YH, JTL, FP, YG, IP, HC, SMD, RM, MH, GM, KL, SA, EJJ, BM, AM, DRM, MPC, YuJC, MAG; Manuscript writing: SS, NMC, YiJC, NH, YH, JTL, FP, YG, HC, KL, AM, AIR, MPC, YuJC, MAG; Supervision: SS, JC, DF, BZ, KR, PW, RG, PP, AM, GAG, AIR, GC, PY, SY, HC, SMD, AIN, DRM, MPC, YuJC, MAG; Funding acquisition: SS, JC, BZ, PW, AM, GAG, GC, SY, HC, SMD, AIN, SAC, DRM, YuJC, MAG.



eTOC Blurb

Satpathy et al. conduct a proteogenomic analysis of 406 lung adenocarcinoma tumors across diverse demographics revealing oncogenic mechanisms, the prognostic role of chromosomal instability, immune signaling patterns, sex-specific carcinogen effects, and early-stage tumors with “late-like” traits. These biomarkers and subtype-specific therapeutic atlas advance precision strategies for this lethal disease.

Keywords

Proteogenomics; LUAD; Genomics; Mass Spectrometry; PTMs; Biomarkers; Smokers; Non-smokers; CPTAC; ICPC

Introduction

Lung adenocarcinomas (LUAD), the most prevalent subtype of non-small cell lung cancers (NSCLC), represent a significant global health challenge with marked geographical and epidemiological diversity in incidence and disease progression[1]. Despite advances in targeted therapies and improved diagnostic techniques, LUAD continue to exhibit high mortality rates[2], underscoring the need for deeper understanding of their molecular underpinnings across diverse populations.

Building upon foundational knowledge of LUAD genetic and transcriptional alterations[3–9], recent proteogenomic studies have helped functionalize genomic aberrations by integrating expression of proteins and their post-translational modifications (PTMs) with genomic and transcriptomic data[10–15]. Collectively, these studies have probed the tumor immune microenvironment (TIME)[16], refined molecular subtypes, identified pathways associated with poor prognosis, helped illuminate biology of *ALK* fusions, examined the impact of specific carcinogens, and identified candidate drug targets.

Despite these advances, many important gaps in our understanding of LUAD biology remain. *EGFR* mutations are among the best-studied LUAD alterations and provided an early molecularly-targeted therapeutic success story[6,17], but the proteogenomic correlates of specific driver mutations (e.g., *L858R* and *E19del*) and co-mutations remain underexplored, and the comparative inefficacy of immunotherapies in that population[18] invites scrutiny. Incidence and outcome differences between males and females[19–24] or across races and ethnicities[25–27], even when controlled for smoking behavior or driver mutation, are poorly understood. Complex interactions among genetic ancestry, somatic mutations, and environmental exposures may underlie the observed clinical and molecular heterogeneity of LUAD[28,29]. Disparate outcomes amongst patients with early-stage disease call for improved stratification[11,30].

Challenges with the acquisition of high-quality specimens suitable for proteogenomic analysis, together with the cost and complexity of data generation, have limited the numbers of patients in comprehensive studies. Those focused on comparatively uniform cohorts are unsuitable for comparisons across important disease-modifying demographic variables[11–13], while those with greater diversity in ethnicity, smoking status, or sex can be statistically underpowered for subgroup analyses[10]. Here we present a proteogenomic analysis of LUAD involving tumors and normal adjacent tissues (NATs) from 406 treatment-naïve patients from North American, Eastern European (collectively “Western”/Caucasian), and Asian cohorts, broadly spanning smoking status, sex, and age. While we have hewed to established categorical frameworks for some our analyses (e.g. mutation-specific designations), a detailed and systematic examination of geographic origin, genetic ancestry, health behaviors, environment, and sex on oncogenesis and disease progression mandated greater flexibility in others. These cohorts enhance the generalizability of our findings and identify region-specific and exposure-related molecular signatures that could guide personalized diagnostic and therapeutic strategies.

Results

Proteogenomic characteristics of the ICPC-CPTAC LUAD cohort

The U.S. National Cancer Institute (NCI)’s Clinical Proteomic Tumor Analysis Consortium (CPTAC) and the Taiwan Cancer Moonshot team under the International Cancer Proteogenome Consortium (TW-ICPC) previously undertook separate analyses of distinct LUAD patient cohorts with diverse demographics[10,11]. The published CPTAC cohort (“CPTAC.A”) included Western and Asian smokers, former smokers, and never-smokers enriched for *KRAS* or *EGFR* mutations, respectively. In contrast, the TW-ICPC study focused on never-smoking Taiwanese LUAD cases, primarily females with tumors featuring

EGFR mutations (“ICPC.A”). Here, we present a joint analysis of these published cohorts plus additional cases from each team (“CPTAC.B” and “ICPC.B”). This large and diverse LUAD proteogenomic cohort includes 406 LUAD tumors and 388 matched NATs. The sample size allowed the exploration of features enriched across or within each of the cohorts at unprecedented granularity (Figure 1A, Table S1).

Proteogenomic data generated for all cohorts included copy number aberrations (CNA), somatic and germline mutations, transcriptomics, global proteomics, and phosphoproteomics (Figure 1A, Table S1–2). DNA methylation, acetylproteome, and ubiquitylproteome data were available exclusively for the CPTAC cohorts. N-linked glycosylation searches and potential site-specific carbohydrate composition were performed on the global proteomics data for the CPTAC.B cohort (Glycoproteome) (Figure 1A, Table S2). A schematic describing the strategy used to combine the common omics types between cohorts is shown in Figure S1A, with its effects examined in Figure S1B, C.

The oncoplot emphasized the expected mutual exclusivity of known oncogenic alterations involving *RAS/RTK* pathways and a wide range of co-mutations including tumor suppressor genes such as *TP53*, *STK11*, *RBM10*, and *CDKN2A* (Figure 1B). A cell of origin analysis based on cell phenotypes identified AT2 and Club cells as enriched in *EGFR*, *RAS* and *RAF* family mutated tumors, in contradistinction to a small subset of lung squamous cell carcinoma (LSCC) samples (n=14, otherwise excluded in this study), which showed enrichment of basal cells as expected (Figure 1C, Table S3). Intriguingly, RTK-altered cell lines for which a cell of origin could not be determined exhibited poorer responses to targeted therapies compared to lineage-specific or cell type-enriched counterparts (Figure S1D, E). This heterogeneity in therapeutic response underscores potential differences in cell transcriptional states.

Non-negative matrix factorization (NMF)-based multi-omics clustering defined four tumor clusters (Figure 1D) characterized by differential enrichment of clinical and biological annotations and molecular features, as summarized in Figures 1E–F. Samples in Cluster 1, “Unstable Proliferative (UP)”, exhibited evidence of genomic instability and a signature commensurate with high proliferation. Cluster 2, “Quiescent EGFR, (QE)”, displayed downregulation of several pivotal carcinogenic pathways and enrichment of *EGFR* mutants. Cluster 3, “Immune Active KRAS, (IAK)”, demonstrated upregulation of immune signaling pathways and enrichment of *KRAS* mutants. Lastly, Cluster 4, “Stable Early Stage (SES)”, exhibited notable enrichment of early-stage and genomically stable tumors. The multi-omics clusters derived from this combined cohort were compared to transcriptomic subtypes derived from TCGA LUAD studies[3,9] (LUAD Subtype and TCGA Subtype, respectively) as well as to multi-omics clusters previously derived by CPTAC[10,11] (CPTAC.A C1-C4) (Figures 1D, S1F). While the ICPC-CPTAC cohort is unique in the diversity of its demographics, some biological signatures were conserved across cohorts. For example, the IAK cluster was enriched for LUAD Subtype S3 and TCGA Subtype proximal_inflammatory and correlated well with CPTAC.A C1, all of which showed strong positive enrichment for immune signaling. In addition, the SES cluster was enriched for LUAD Subtype S5 and TCGA Subtype terminal_respiratory_unit and correlated well with CPTAC.A C4, all of which exhibited negative enrichment of cell cycle-related pathways.

NMF-based clustering was also performed in the space of individual -omes, including RNA, protein, phosphoprotein, and gene-level CNA. The proteome clustering (NMF Prot Consensus, Figure 1D) was noteworthy, stratifying patients into three clusters (C1-C3) associated with distinctive clinical and phenotypic features, including enrichment in proteome Cluster 2 of genomically unstable, immune “hot” tumors in patients with worse performance status and outcomes (Table S1). Sankey plots (Figure S1G) show the overall relationship between the multi-omic NMF clusters and individual omic-based clusters, with IAK only fully defined upon multiomic integration.

Biological impact of less frequent genetic events such as *RBM10* mutation and *ALK* fusions

The large size of the combined LUAD cohort allowed us to investigate the proteogenomic consequences of comparatively uncommon co-mutations and driver events including *RBM10* mutations (n=46) and *ALK* fusions (n=17). Unlike prior studies in smaller and less diverse cohorts, *RBM10* mutations were not mutually exclusive with mutations in *TP53*[31] or associated with male sex (Figure S2A,B), but truncating mutations were associated with proteomics-driven downregulation of spliceosome-related pathways (Figure 2A, S2C–E). Modulation of splicing kinases CLK1–3 and SPRK2,3 and phosphorylation-driven upregulation of EGFR, NOTCH and KIT receptor signaling were observed (Figure 2B). Downregulation of the spliceosome in *RBM10*-mutated samples stands in contradistinction to its upregulation in many cancers[32], and may confer dependency-associated sensitivity to spliceosome inhibition in the context of *EGFR* mutation[31]. Consistent with the observed metabolic alterations (Figure 2A), Hexokinase 2 (HK2), the enzyme catalyzing the first-step of glycolysis[33], was markedly upregulated in *RBM10/EGFR* co-mutated samples (Figure 2C, Table S3) while protein expression of SF3B1 was largely unchanged, suggesting that sensitivity to spliceosome inhibition might not be due to activity of the SF3B1 complex[31]. Finally, we observed significant upregulation of neutrophil degranulation (ND) signatures (p=0.0004, Wilcoxon test) in the *EGFR-L858R* subset with *RBM10* co-mutation (Figure S2F, G). ND, observed at both the RNA and protein levels, was validated in both an additional patient dataset[34] and an *Rbm10* knockout mouse model[31] (Figure S2H and I), providing further evidence for *de novo* ND upregulation in co-mutated *EGFR-L858R* and *RBM10* LUAD. RBM10 deficiency has been associated with resistance to EGFR inhibitor therapy[35]; our data suggest that the functional consequences of *RBM10* alterations extend to the immune microenvironment, with potential implications for cytotoxic and immune checkpoint inhibitor therapies.

Unlike *RBM10, EGFR*, and *KRAS* mutations, fusions associated with the receptor tyrosine kinase *ALK* are relatively uncommon but targetable LUAD driver events found largely in never-smokers. As constitutively activated ALK is both ectopic at the tissue level and mislocalized at the cellular level in *ALK* fusion-driven LUAD[36], phosphoproteomics affords a unique opportunity to expose oncogenic biology. Increased phosphorylation of specific tyrosine (Y) residues on proteins including SND1, EML4, HDLBP, LHDA, ARHGEF5, and PIK3R1, which we previously reported for *ALK* fusion-positive CPTAC.A LUAD samples[10], was confirmed in the combined LUAD cohort (Figure 2D, Figure S2J, Table S3). To assess the functional significance of these observations, we performed kinase inhibition

experiments in cell lines with *ALK* fusions or kinase driver mutations (*FGFR/2*, *KRAS*, *PIK3CA*) and examined ALK-associated pY sites by mass spectrometry. Treatment of *ALK* fusion-driven cell lines with ALK inhibitor, but not other kinase driver/inhibitor-matched treatments, resulted in significant downregulation of ALK-associated pY (Figure 2E), supporting their dependency on ALK. Peptide pull-down assays in lysates from *EML4-ALK* fusion-driven cell lines that used phosphorylated forms of the peptides as baits led to recovery of ALK as well as many proteins previously implicated as ALK interactors, such as PTPN11 and SND1 (Figure 2F). Altogether, these experiments confirm that *ALK* fusions promote a cascade of defined phosphorylation events that facilitate key protein interactions.

Functional status of post-translational modification sites may be supported by clustering on protein structure

PTMs, like mutations, can dramatically alter proteoform diversity and impact cancer biology[37–39]. Aggregated PTMs can potentially exert a stronger biological impact, or indicate sites of functional importance, especially when superimposed on protein structure. To identify clusters of correlated PTMs within protein 3D structures, we applied CLUMPS-PTM[39] (Table S3). In pairwise comparisons of tumors vs NATs, smokers vs never-smokers, and *EGFR* mutated vs WT tumors (Figure 2G), 32 proteins exhibited significant clustering of phosphosites, acetylsites, or ubiquitylsites (adj. $p<0.1$, Figure S2K). Within *EGFR* mutated and WT tumors, HSPB1, a small heat shock protein, emerged as a top hit (Figure 2H). Differential phosphorylation at HSPB1 S78 and S82 facilitates physical interactions with other proteins, including AKT1, leading to their phosphorylation and triggering an anti-inflammatory response by promoting neutrophil apoptosis[40]. Accordingly, increased HSPB1 phosphorylation was observed at sites S82/S83 in the Eosinophils/Endothelial (E/E) immune subtype, which was relatively depleted of neutrophils and enriched for *EGFR* mutations (Figure S2L). We also observed higher AKT1 protein in *EGFR* mutated samples (Figure 2H, right panel). Clusters of acetylation sites were observed in EP300, a well-defined tumor suppressor and histone acetyltransferase (Figure S2M). In tumor vs NAT comparisons, increased acetylation was observed at known EP300-activating sites K1546, K1549, K1554, and K1555. This increase correlated with elevated N-terminal acetylation of H2A and H2B, substrates regulated by CBP/EP300, which serve as markers for active enhancers—key elements whose aberrant activity in driving cancer progression is well-documented across multiple cancer types[41,42,43]. Protein acetylation can also have important implications for cancer metabolism, as reflected in the differential acetylation of MDH1, which supports glycolysis in tumors by producing metabolites such as NAD+ and malate[44] (Figure 2I). Lysine residues K328, K329, and K335 on MDH1 are known ubiquitination sites suggested to stabilize and enhance enzyme activity when acetylated[45]. We identified a cluster of acetylsites on SMC3, known to regulate double-strand break (DSB) repair and maintain genome integrity through an ATM-ESCO2-SMC3 axis[46]. Consequently, we examined the correlation between protein abundance and genome stability. Significant correlation was seen between the principal acetylation sites K105/K106 and genomic instability (Figure S2N, second panel). A similar acetylsite-specific association between instability and complex member SMC1A (Figure S2N) further underscores the functional insights gained exclusively from PTM analysis. Finally, ubiquitylsite clustering occurred directly within the proteasome domain of PSMA5, a 20S core proteasome subunit (Figure 2J). Ubiquitylation

at K187, K192, K196, and K203 suppresses proteasome activity[47]. PSMA5 promotes tumor progression and correlates with poor prognosis[48]; its expression was significantly increased in NMF protein-based cluster C2, which included the highest percentage of patients with late-stage tumors and worse survival (see “Proteomic Clustering Identifies Early-Stage Late-Like LUAD Subtype”, below).

Genome contiguity combined with Breakage Intensity Clustering is prognostic in LUAD

Chromosomal instability (CIN) is a hallmark of cancer onset and progression[49]. In concordance with previous studies[50], we found that common measures of CIN, such as the weighted genome instability index (wGII[51]) (Figure S3A), fail to stratify LUAD outcomes. These metrics primarily quantify genomic imbalance (aneuploidy) instead of degree of fragmentation (contiguity), the latter being reflected in genomic segment lengths. We examined genome contiguity through segment length distributions and identified the 75th percentile of this distribution (SegLen-Q3) as a proxy for the presence of long segments, using segment length to categorize samples into fully stable, partially stable, or unstable (Figure 3A, Table S3). SegLen-Q3-based classification was associated with overall survival (OS) ($p<0.02$) (Figure 3B). This association was validated in multiple independent cohorts across a range of cancers (Figure S3B–D). Unstable tumors showed focal amplifications of *TERT*, *NKX2-1* and *MYC* along with expected increase in structural variants (Figure 3C, Figure S3E,F), with TERT amplifications being also associated with OS (Figure S3G–I).

The distribution of segment lengths is determined not only by the number of genomic breakpoints in a sample, but also their genome-wide arrangement, i.e. whether they tend to cluster together resulting in a few long and many short segments or disperse, resulting in most segments being of average length. Accordingly, we independently estimated the intensity (int) and clustering (clust) of genome breakage by applying the theory of spatial point processes[52,53] to develop the Breakage Intensity Clustering (BIC) metric (STAR Methods). Classification of tumors along these two parameters (Figure 3D) identified patient strata with exceptionally good (BIC-contiguous), typical (BIC-fragmented), and poor (BIC-intense) survival (Figure 3E). The BIC classification uniquely identified tumors where significant breakage was localized to small genomic regions (Figure S3J).

We noted that focal *NKX2-1* amplifications were largely restricted to BIC-intense tumors (Figure 3C), with a significantly higher proportion of BIC-intense tumors (10%; Figure 3F) compared to other BIC groups. The amplifications were associated with a significant increase in *NKX2-1* RNA ($p<1.5e-5$) and protein expression (Figure S3K). We also observed complete transcriptional silencing of *NKX2-1* (null expression) in a subset of BIC-intense and BIC-fragmented tumors (Figure 3F), associated with increased methylation of the *NKX2-1* promoter in fragmented NKX2–1 null tumors. Unexpectedly, both *NKX2-1* amplification and silencing were associated with higher proliferation (Figure 3G). Tumors that lost *NKX2-1* also lost their alveolar cell identity as measured by both inference of transcriptional programs (Figure 3H) and expression of the marker gene Napsin A (NAPSA) (Figure S3L). Conversely, *NKX2-1* amplified tumors maintained their lineage and NAPSA expression (Figure S3M) while acquiring a high proliferative index (Figure 3G). Tumors without NKX2–1 expression had a significant upregulation of cancer testis antigens, such as

TESMIN, MAGEA, and MAGEA12, and markers associated both with LUAD progression and squamous differentiation, including KRT6A and KRT16 (Figure S3M).

IGF2BP3 is a robust biomarker of genome fragmentation and predicts immune checkpoint inhibitor response

Recent therapies targeting dependencies of chromosomally unstable tumors[54,55] provide strong rationale for the development of robust instability biomarkers. Genomic diagnostic methods based on WGS are challenging in the clinical setting due to cost, sample availability, and turnaround time. After an unsuccessful attempt to train deep-learning models to identify instability based on H&E images[56] (Figure S3N), we pursued proteomic biomarkers as a clinically accessible alternative. By comparing fragmented with other cancer genomes (Figure 3I, Table S3) we identified *bona fide* oncogene IGF2BP3[57] as a lead candidate biomarker associated with genome instability (FC: 1.50, adj. $p<0.009$). We optimized an immunohistochemistry (IHC) assay to detect IGF2BP3, which showed varying staining patterns across the range of genome instability levels, (Figure 3J) with undetectable expression in fully stable samples. IGF2BP3 IHC was validated on a tissue microarray featuring independently-collected samples from treatment-naive patients (78 LUAD, 92 LSCC, 69 metastatic LUAD and 60 matched NAT)[58]. IGF2BP3 expression was undetectable in normal lung, elevated in a subset of primary LUAD and LSCC, and highly elevated in almost all metastatic LUAD (Figure 3K), as well as unstable LUAD from the ICGC/PCAWG cohort (Figure 3L)[59]. Experimental induction of genomic instability resulted in increased IGF2BP3 protein levels in unstable clones (Figure S3O), implicating IGF2BP3 as a dynamic biomarker of chromosomal instability. The uniformity of cytosolic IGF2BP3 IHC staining stood in contrast to the patchy staining of cell-cycle dependent Ki67 protein (Figure S3P). Chromosomally unstable tumors evade anti-tumor immunity through a plethora of mechanisms[60]. We therefore examined whether IGF2BP3 could serve as a predictive biomarker for immune checkpoint inhibitor (ICI) therapy. Data from a recent randomized trial that compared ICI to docetaxel in NSCLC[61] showed that high IGF2BP3 expression was associated with worse outcomes for patients treated with ICI (Figure 3M, $p=0.013$), but not docetaxel (Figure S3Q), regardless of PD-L1 (CD274) expression (Figure 3N). High expression of IGF2BP3 associated with genome instability appears to be triggered by loss of promoter methylation (Figure 3O). The most aggressive BIC-intense tumors are characterized by increased frequency of *TP53* and *EGFR* mutations, whole-genome doubling (WGD), and *TERT* amplifications. BIC-intense tumors were only modestly associated with clinical stage or smoking status (Figure S3R) and had fewer *KRAS* mutations (Figure S3S).

Altogether, these data identify genomic fragmentation as a prognostic and putative predictive hallmark of LUAD, and highlight divergent, NKX2–1-dependent and -independent evolutionary trajectories towards acquisition of aggressive cellular phenotypes.

Differential immune signaling and inferred transcription factor and kinase activity in Hot and Cold LUAD tumors

Therapeutic successes with immune modulation have transformed oncology treatment, but many cancers, including *EGFR*-mutated LUAD, remain recalcitrant to immunotherapy.

To deepen our understanding of immune context and immunobiology, we performed deconvolution analysis by integrating proteomic and gene expression data for 406 tumors and 388 matched NATs[62]. To inform the deconvolution, we leveraged a comprehensive database of single-cell RNA data containing ~1 million cells from ~180 LUAD (STAR Methods, Table S4). Using a classifier based on pan-cancer RNA and proteomic data[62], we allocated LUAD samples into 5 immune subtypes: CD8+/IFNG+, Eosinophils/Endothelial (E/E), Fibroblast/TGFBeta, CD8−/IFNG+ and CD8−/IFNG− (Figure 4A, Table S4). As expected, samples in CD8+/IFNG+ were enriched for $CD8^+$ and $CD4^+$ T cells and showed activation of interferon gamma signaling, which was also observed in CD8−/IFNG+ (Figure 4B). The EGFR signaling pathway was upregulated in CD8+/IFNG+ (Figure 4B). Samples in E/E were enriched for fibroblasts, reflecting their stromal cell content, and associated with low smoke-exposure scores and low tumor mutational burden (TMB). This subtype was also characterized by the proteomic upregulation of the “MAPK1/ERK2 activation” signature and downstream “RSK activation” pathway, a promising therapeutic target in lung cancers[63] (Figure 4B). Both CD8−/IFNG− and CD8−/IFNG+ were associated with the upregulation of “MYC targets”, “E2F targets”, “DNA repair”, and “Cell Cycle Checkpoint”, suggesting activation of cell cycle-related pathways in these two immune subtypes.

Tumor cell percentages estimated from CNA and gene expression data were anti-correlated with immune cell fractions and decreased in highly immunogenic CD8+/IFNG+ tumors (Figure 4A). Immune subtypes aligned with the RNA-based LUAD classification[9], with CD8+/IFNG+ enriched in the “Immune Active KRAS (IAK)” cluster (S3) and E/E in RNA S5 tumors (Figure 4A). We observed an enrichment of CD8+/IFNG+ in the C2 proteomic cluster (Figure 4A), suggesting that some C2 tumors might benefit from immunotherapy.

Deconvolution analysis of NATs demonstrated a higher infiltration of myeloid cells such as macrophages and NK cells compared to tumor tissue (Figure S4A). Association analysis between cell type fractions and mutation profiles revealed that *TP53* mutation was associated with higher presence of $CD8^+$ T cells, neutrophils, and dendritic cells (DC) (Figure S4B), of potential therapeutic relevance given prior reports that $CD8^+$ T cells in *TP53*-mutated tumors conferred benefits in immunotherapy response[64,65]. In addition, we found that *EGFR exon19del* and *L858R* tumors were associated with reduced immune infiltration in NATs, not seen in the corresponding tumor samples (Figure S4C). This finding suggests that the poor response of *EGFR* mutant tumors to immunotherapy[66] could be influenced by the immune environment in NAT. Next, leveraging our phosphoproteomic data, we performed kinase-enrichment analysis (KEA3)[67] to derive kinase activation scores for each patient, and performed association analysis with immune subtypes (Table S4). We observed higher activation of tyrosine kinases such as HCK, CSK and BTK in CD8+/IFNG+ tumors (Figure 4C). Tyrosine kinases are regulators of signaling in the immune system, with Src family kinases such as HCK being more active in leukocytes. On the other hand, CD8−/IFNG+ and CD8−/IFNG− were associated with activation of cell cycle-related kinases such as CHEK2, CSNK2A2 and CDK9, reflecting colder tumor microenvironments. Consistent with the pathway analysis, we also observed activation of MAPK1 (ERK2) and RPS6KA1 (RSK) in the E/E subtype.

Since protein kinases serve as cell signaling switches to regulate intracellular programs such as transcription and translation, we aimed to identify more complete cell-signaling pathways downstream of the kinases that were differentially activated in Hot (CD8+/IFNG+) and Cold (CD8−/IFNG−) tumors. To achieve this, we evaluated the co-activities of inferred kinases and transcription factors (TFs). We analyzed the differentially expressed mRNAs and phosphoproteins from each patient with TF and kinase enrichment analyses, respectively. This enabled us to identify modules of kinases that regulate modules of TFs either positively or negatively in Hot vs Cold tumors (Figure S4F, 4D, Table S4). Two of the seven kinase modules contained immune-related tyrosine kinases including SYK, LYN, LCK, and FYN, and two of five TF modules were enriched for immune-related TFs such as RUNX3, FOXP3 and STAT4, which appeared downregulated in Cold tumors (Figure S4F). In contrast, a cell-cycle related kinase module containing CDK1/2 and AURKB was associated with upregulation of an immune TF module only in Hot tumors (Figure S4F). Another module, enriched for kinases related to focal adhesion, including MAPK8/12, MAP2K1, PTK6, TESK2, and IGF1R (Enrichr[68]), was downregulated in Hot tumors, in contrast with the aforementioned immune-related TF modules (Figure 4D). Inhibition of focal adhesion was associated with increased immune infiltration and Cold-Hot transition in a *KRAS/STK11* co-mutated mouse model of lung adenocarcinoma[69]. This suggests that targeting the kinases in this module may shift the tumor microenvironment toward a more permissive phenotype. This association was validated using the LINCS L1000 dataset, a resource that profiled gene expression changes in human cancer cell lines after the CRISPR knockouts of ~8,000 single genes and ~30,000 chemical perturbations with small molecules and drugs[70]. Specifically, we examined genes that were upregulated following the CRISPR knockout of kinases in the Focal Adhesion module and confirmed that this set of upregulated genes significantly overlapped with the Reactome Innate Immune pathway (Innate Immune System R-HSA-168249) and genes downstream of STAT4 obtained from published ChIP-seq studies and other sources (Figure 4D).

To infer tumor cell-specific pathway activity associated with Hot/Cold tumors, we implemented a novel analysis pipeline via BayesDeBulk[62]. We found “MTOR Signaling”, “Signaling by WNT in Cancer”, and multiple metabolic pathways to be upregulated in tumor cells of Cold tumors based on both RNA and proteome data (Figure 4E, right panels; Table S4), while “Innate Immune System” and “Signaling by Notch” were upregulated in tumor cells of Hot tumors based on proteome data. WNT signaling has been associated not only with tumor development but with the capacity of tumoral cells to evade the host immune response[71]. To validate this pathway activation in tumor cells of Hot and Cold tumors, we leveraged a scRNA-seq database comprising 33 patients classified with low immune content (Cold) and 40 patients with high immune content (Hot). As shown in Figure 4E (left panel), several metabolic pathways and “MTOR signaling” were found upregulated in tumor cells of Cold tumors based on scRNA-seq data, confirming the findings of our deconvolution tool. On the other hand, only “Innate Immune Response” was confirmed upregulated in tumor cells of Hot tumors by scRNA-seq data.

Protein glycosylation, typically enriched in membrane and extracellular matrix proteins, was analyzed for associations with immune subtypes in LUAD (Table S4, Figure S4B). We found glycosylation on several proteins including FKBP10[72] and PLOD2[73], which have

been previously associated with LUAD progression and are members of the collagen 1 proteostasis network[74] (Figure S4D and S4E), to be downregulated in the E/E subtype (Figure 4F). This might be linked to lower neutrophil infiltration of E/E subtype and better survival (Figures 4G, 4H), as implicated in previous studies[75,76]. These findings align with prior studies on LUAD survival[77,78] and suggest the influence of glycosylation on subtype-specific outcomes, warranting further investigation.

Context-dependent role of STK11 mutation in LUAD

STK11 mutations in LUAD are associated with more aggressive disease and worse outcomes[79,80], but there are conflicting reports regarding their association with immunotherapy efficacy[81]. Pathway analysis done on significant positive outlier features in *STK11*-mutated samples converged on dysregulated metabolic pathways (Figure S4G). *STK11* mutation has been associated with a cold tumor microenvironment in LUAD[10] as well as in pan-cancer immune analysis[82]. Surprisingly, in this expanded, multinational LUAD cohort, only dendritic cell depletion was associated with *STK11* mutation (Figure 4I), while other metrics of an inflammatory TME, such as tumor cell percentage and immune score, were not. The distribution of *STK11*-mutant tumors across immune subtypes (Figure S4H) invited closer investigation of their heterogeneity, including examination of concurrent *STK11* and *KEAP1* mutations, previously shown to have unique properties in LUAD[83,84]. *STK11* mutants classified into proteomic C1/C3 clusters were strikingly different from those allocated to C2 (Figure S4I), with C1/C3 *STK11*-mutated tumors having lower wgII, lower TMB, and enrichment for the S5 RNA subtype rather than the major S4 subtypes as previously proposed[9]. Proteomic evidence of neutrophil degranulation, which we previously described as a characteristic of *STK11*-mutated LUAD[10], was seen in *STK11*-mutated tumors in the C2 but not C1/C3 clusters.

Several cancer-relevant molecular aberrations were differentially distributed between C1/C3 and C2 *STK11*-mutated tumors (Figure S4I and S4J, Table S4). C1/C3 samples did not show significantly more *STK11/KEAP1* co-mutations (Figure S4J), suggesting differential mechanisms other than co-mutations[83,84]. C1/C3 *STK11*-mutated samples had higher expression of CD47 (Figure 4J) and a more stable genome, manifested in downregulated MKI67, CDK1, TOP2A, and IGF2BP3 protein expression (Figure 4K). At the proteome pathway level, these tumors showed lower neutrophil degranulation and upregulation of xenobiotic metabolism, heme metabolism, and reactive oxygen species pathways (Figure S4K). Enhanced NRF2 pathway activity was observed at multiple molecular levels including PTMs (Figure 4K). *STK11*-mutated tumors with low neutrophil degranulation (C1/C3) had a ferroptosis resistance phenotype, similar to a recent report[81], reinforced by upregulation of NEDD4-like E3 ubiquitin protein ligase (NEDD4L) and downregulation of lactotransferrin (LTF), a novel mechanism limiting ferroptosis[85] not observable at the transcriptomics level (Figure 4K). The ferroptosis protective gene NQO1 has been implicated as a potential target for cancer therapeutics[86]. Some cluster-specific differences between *STK11*-mutated tumors were shared with *STK11*-WT tumors in the cluster; for instance, neutrophil degranulation and higher IGF2BP3 expression appeared to be common characteristics of C2 tumors. By contrast, CD47, NEDD4L and UGDH expression differences were specific for *STK11*-mutated tumors independent of NRF2 pathway activity (Table S4). C1/C3 *STK11*-mutated

tumors also showed gastric differentiation markers MUC5AC, GKN2, PGC, and CTSE expression, and downregulation of IGF2BP3 and EZH2 (Figure S4L, Table S4). This heterogeneity suggests that stratification within *STK11*-mutated tumors might be needed to exploit potential therapeutic vulnerabilities.

Distinct contribution of environmental carcinogens to LUAD development

To investigate the impact of key endogenous carcinogens and environmental exposures on lung cancers across diverse geographic regions and genetic backgrounds, we performed unsupervised clustering of tumors according to the relative contributions of mutational signatures derived from human cancers[87–89] and pluripotent stem cells exposed to environmental carcinogens[90] (STAR Methods), revealing ten clusters (Figure S5A) that further assembled into four major groups and a mixed mutation signature group (Figure 5A). Polycyclic aromatic hydrocarbons (PAHs) and nitro-PAHs, carcinogens from tobacco or air pollution[91], dominated tumors in smokers and high smoke-exposure score never-smokers irrespective of geography. These tumors were enriched for CD8−/IFNG− (in *EGFR*/*KRAS*-WT) or CD8+/IFNG+ (in *KRAS*-mutated) immune subtypes (p=0.0033, Figure 5B). In contrast, the nitrosamine-high signature, reflecting a potent carcinogen found in processed food, cosmetics, and cigarette smoke[92], occurred predominantly in Asian female never-smokers with early-stage *EGFR-L858R* or *-E19Del* mutated tumors (p<0.001, Figure 5B, Table S5). PAHs/nitro-PAHs were enriched in smokers and never-smokers with *KRAS*-mutated or *EGFR*-WT tumors, while nitrosamine was specific to never-smokers with *EGFR*-mutations (Figure 5C). Nitrosamine-low tumors included a small subset with a distinct APOBEC signature (yellow in the “mutation group” assignment) (Figure 5A–B, S5A, subclusters 5 and 9), aligning with our previous report of poor survival of our never-smoking cohort[11]. The four major groups showed significant differences in relapse-free survival (RFS) and OS (Figure 5D, Figure S5B). Patients with tumors in the nitro-PAHs/PAHs and APOBEC groups showed markedly worse RFS, with hazard ratios (HR, corrected for age, sex, cohort and stage) for relapse events of 2.07 (95% CI=1.06–4.04) and 1.87 (95% CI=1.07–3.27), respectively (Figure 5D–E). Furthermore, tumors with nitrosamine-high signatures, which were enriched for *EGFR-E19Del*, exhibited poorer RFS and OS (Figure 5F, Figure S5C), with an adjusted HR of 6.07 (95% CI=1.26–29.24) for RFS. These findings underscore the critical role of nitrosamines and nitro-PAHs in driving adverse clinical outcomes in lung cancers across both smokers and never-smokers.

Tobacco stream smoke and haze pollution are significant sources of carcinogens, including PAHs, heterocyclic compounds, and N-nitrosamines[91]. To investigate how their long-term exposures drive cancerization in normal tissues, we analyzed differential proteins and pathways in NATs, the paired tumors of which had nitrosamine (Figure 5G, left panel) and nitro-PAHs or PAHs signatures (Figure 5G, right panel) (Table S5). Notable overlap was seen between the carcinogen signature groups, most prominently in immune-related pathways, such as neutrophil degranulation, MHC-II antigen presentation, and FcγR-mediated phagocytosis (Table S5). NATs associated with high nitrosamine signature tumors exhibited metabolic and immune activation (AGER signaling, CXCR4 signaling), as well as GNRH and ERK-MAPK signaling, potentially promoting proliferation and invasion[93]. Nitro-PAHs/PAHs NATs showed IL-8 signaling. The results link carcinogen exposures to

inflammation, oncogenesis, and tumor progression in LUAD NATs, as well as carcinogen-specific alterations such as nitrosamine-induced metabolic dysregulation[94].

Smokers constitute the largest subset of patients enriched for exogeneous carcinogen signatures. We assessed field cancerization effects and distinguished smoking-related oncogenesis from other carcinogen exposures for never-smokers, applying the S-scoring method[95] to compare the transcriptome, proteome and phosphoproteome of NATs from individuals who had high or low smoke-exposure scores (HSS/LSS)[10] (Figure S5D). Among the 226 and 157 up- or down-regulated features, 197 were integrated into a protein-protein interaction network and organized into 19 interconnected modules (Figure S5E). Upregulation in immune receptors, complement system components, and chemokines indicated complex modulation of immune responses and immune cell recruitment within the HSS NAT microenvironment. Pathway analysis consistently revealed significant enrichment for immune-related processes as well as glucose, cholesterol, and lipid metabolism (Figure 5H). Interestingly, while both groups shared nodes of inflammatory and metabolic imbalance, nitro-PAHs exposure induced a higher number of upregulated phosphorylation events in extracellular matrix (ECM) degradation, inflammation and antigen presentation (Figure 5I), suggesting the more profound carcinogenic impacts of nitro-PAHs compared to PAHs.

Nitrosamines and nitro-PAHs/PAHs carcinogens promote different cancer-relevant pathways: Following the analysis of NATs, we evaluated the exogenous carcinogen contributions in tumor development by examining differentially enriched proteins, phosphosites, and activated pathways in tumors from high, low and absent carcinogen signature groups using Ingenuity Pathway Analysis. Among environmental carcinogens, nitrosamine CS was linked to predominantly female never-smokers (Figure 5B). Tumors with the nitrosamine signature shared enriched pathways with their paired NATs, but exhibited higher carcinogenic activities (Figure S6A, Table S6), including mitochondrial fatty acid oxidation, inflammation and immune response, aligning with the early toxic effects of nitrosamine exposure[96]. Tumors in nitro-PAHs/PAHs signature groups had evidence of smoke exposure-activated pathways, including xenobiotic metabolism AHR pathway, neutrophil degranulation, IL-8 signaling, mTOR signaling, and detoxification of reactive oxygen species pathways, while MAPK signaling and ERBB signaling were more activated in the nitrosamine signature group (Figure 6A, Table S6). These carcinogens influenced expression of proteins explicitly linked to activated metabolism and oncogenesis. In the nitro-PAHs/PAHs group, for instance, AHR downregulation suggested that carcinogen binding triggered its degradation, promoting immune activation and cancer progression (Figure 6B upper panel, Table S6).

The nitro-group addition to PAHs in nitro-PAHs has been reported to increase carcinogenic potential 10–1000 fold[97], a difference that was clearer in tumors than NATs (Figures 5I, S5I, 6B, lower panel, Figure S6B–C). The nitro-PAHs tumors showed stronger activation of aromatic compound metabolic processes, cell cycle, and metastasis, and exclusive enrichment of proteasome, ribosome, and ATP synthase subunits, proteins/phosphosites involved in organonitrogen or aromatic compound metabolism (e.g. CYP1B1, a key enzyme

for cigarette-smoke carcinogen metabolism)[98,99], immune response (e.g. CD274 (PD-L1), CD79A), and MMPs (Figure 6B, lower panel).

Of particular interest are the common and distinct characteristics of tumors from never-smokers and smokers that share the nitro-PAHs/PAHs signature. A subset of tumors from smokers, mostly *EGFR*-mutated from Asian individuals, had LSS and a stronger contribution of APOBEC and methylcytosine deamination signatures (Figure 5B, middle panel, Figure S5A, Table S5), with evidence of crosstalk between chemical carcinogenesis, immune response and trafficking of AMPA receptors. Swanton et al. recently reported that PM2.5 promotes *EGFR/KRAS* mutation-driven tumorigenesis, manifest as elevated IL-1β and CD274 expression in macrophages of a PM2.5-treated mice model[100]. We observed significantly elevated CD274 ($p<0.01$) expression and activation of cancer-relevant processes including RTK/MAPK/PI3K signaling, glycosylation metabolism, IL1 signaling and neutrophil degranulation pathways in never-smoker females without HSS, a cancer cohort that was also significantly younger (median age <60yr, $p=0.0083$) (Figure S6D, Figure S6E, Table S6). Our concordant molecular observations suggest that these never-smokers were likely more susceptible to air-pollution-induced LUAD. Neutrophil degranulation promotes lung cancer progression and metastasis[101], cytotoxicity that can be inhibited by CD274[102]; thus never-smokers with a nitro-PAHs signature may benefit from immunotherapy. To promote classification of tumors by carcinogen exposure signatures, a panel of biomarker candidates, many related to immune response, was established to identify the subset of smokers or never-smokers with highest susceptibility to nitro-PAHs/PAHs carcinogens (Figure 6C).

Our classification of carcinogenic signatures in human tissue samples suggests that endo- and exo-carcinogens are associated with distinct driver mutations and smoke exposure. We validated the effects of common carcinogens, nitro-PAH (1-nitropyrene, 1-NP) and PAH (benzo[a]pyrene, BaP), on lung cancer cell lines with different driver mutations (A549 [*KRAS* mutation] and PC9 [*EGFR* mutation]) below the effective concentration 10 (EC10) (Figure S6F). Proteomic and phosphoproteomic analyses revealed that 1-NP and BaP treatments upregulated pathways related to cell cycle, DNA repair, EGFR signaling, and invasion (Figure 6D, Table S6). Notably, the carcinogens induced different molecular signaling patterns for cancer progression in cells with different driver mutations in a dose-dependent fashion (Figure 6E). Furthermore, down-regulated AHR and up-regulated CYP1B1, SERPINE1, PTPN11, RB1, ATP5F1A, TNFRSF10B, which were observed in tumors with nitro-PAHs and PAHs signatures (Figure 6B, Figure S6B–C, Table S6), were also consistently regulated in 1-NP and BaP treated cells (Figure 6E). Functionally, 1-NP and BaP promoted A549 cell migration and invasion, consistent with poor outcomes in *KRAS*-mutant tumors enriched with nitro-PAHs/PAHs signatures, effects that were absent in PC9 cells (Figure 6F, 6G, Figure S6G–I). Conversely, 1-NP enhanced PC9 proliferation but did not significantly affect A549 cells (Figure 6H). Collectively, these findings demonstrate that the same carcinogens elicit distinct cancer phenotypes depending on underlying driver mutations, underscoring the complexity of environmental carcinogen impacts. Routine measurement of carcinogen exposures as a single risk panel could inform effective chemoprevention and monitoring strategies. Our nominated biomarker candidates (Figure 6C) may offer actionable insights to guide such strategies in both smokers and

never-smokers, and their evaluation in blood may provide the foundation for a minimally invasive assay of high-risk exposure to environmental carcinogens.

Proteomic clustering identifies early-stage late-like LUAD subtype

Global protein expression-based NMF clustering yielded three clusters (C1, C2, C3) (Figure S1D), with C2 showing significantly shorter median RFS (2.8 years) across stages I-III and C3 having better RFS (Figure 7A, Tables S1, S7). The C2 cluster included a large proportion of late-stage samples (58% of stage III and 50% of stage IV), had 1.7–2.1-fold higher recurrence compared to other clusters (60% of patients with recurrence), and showed the most cancer-related deaths (52%), more lymph node metastasis (58% of pN2), and higher ECOG scores (65% of score >2) (Table S7). C2 also included 36.4% of stage I patients, a subtype we designated “late-like” in our previous paper[11]. An extreme gradient boosted decision tree classifier trained only on annotated “late-like” tumors from our previous ICPC.A study was used to predict late-like status for all remaining tumors, and showed that, overall, “late-like” tumors were highly enriched in C2. Consistent with a late-like phenotype, stage I patients in C2 showed the poorest RFS among all stage I, independent of smoking status (Figure 7A, right panel, Figure S7A–B). In addition, self-reported smokers with high smoke-exposure score in the C2 cluster had significantly poorer survival than all other smokers, while RFS of self-reported never-smokers was not affected by the smoke-exposure score. (Figure S7C–D). Ingenuity Pathway Analysis (IPA) showed that the C2 subtype samples reflect many of the characteristics of pathologically late-stage (III, IV) samples, irrespective of their own staging (Figure 7B, Table S7). Relative to other clusters, C2 samples showed enrichment of immune- and metastasis-related pathways, such as T-cell receptor signaling, microautophagy, Notch signaling, the TNFR2 non-canonical NF-kB pathway, and cell cycle checkpoints (Figure 7B), with C2 never-smokers showing significant upregulation of immune-related pathways relative to non-C2 never-smokers (Figure 7C, Table S7). We denoted proteins that increased in abundance across stages I-IV and were annotated as druggable targets and secreted blood proteins as potential biomarkers for C2 patients (Figure 7D, Figure S7E, Table S7). Among the most significantly differential proteins, COL11A1 and THBS2 have been shown to promote tumor aggressiveness in LUAD via multiple mechanisms[103,104]. Furthermore, never-smokers in C2 markedly overexpressed KRT6A and KRT14 proteins, which have been associated with tumor progression in LUAD (Figure S7F)[105,106]. Taken together, these results identify a subset of early-stage patients with poor outcomes whose tumors are marked by molecular features seen in advanced-stage cancers, and who might benefit from more intensive monitoring and treatment.

Given their aggressive characteristics, we further explored whether C2 tumors had common or unique oncogenic factors relative to non-C2 samples. *TP53* mutations were particularly prevalent amongst C2 tumors (66.3%) compared to others (32.6% in C1 and 31.8% in C3). Carcinogen mutation signatures were heterogeneous, showing both cluster- and sex-specific contributions among the three proteome clusters (Figure 7E). As expected, the nitro-PAHs signature was significantly more frequent in males, especially in C2 (71.6%), than in females. APOBEC signatures are associated with sex, with APOBEC (C>T) significantly

overrepresented in males (67–80%) and APOBEC (C>G) more common in females (39–40%).

Notably, patients in the C2 proteomic subtype comprised two distinct clinical profiles, one enriched with smoking males, containing mostly *EGFR*-WT (70%) with or without *KRAS*-mutated tumors, and one with never-smoking females enriched with *EGFR*-mutated (68%) tumors. This led us to examine the contribution of mutational signatures in association with sex within the C2 subtype (Figure 7F). Nitrosamine and APOBEC (C>G) signatures predominated for C2 females with *EGFR* mutation (*L858R* and *E19Del*). In contrast, nitro-PAHs and APOBEC (C>T) predominantly occurred in male patients, with the PAHs signature being enriched for male patients with *KRAS* mutations and high smoke-exposure scores. Upregulation of APOBEC3A, APOBEC3B, and APOBEC3G proteins were also observed in the C2 compared to non-C2 subtypes (Figure S7G).

Though increasing evidence has shown that APOBEC-induced genetic heterogeneity contributes to carcinogenesis and represents a cancer hallmark, its differential association with sex is less reported. To further illuminate the molecular characteristics defining the tumors from distinct C2 subgroups of never-smoking females enriched for *EGFR*-mutated tumors and smoking males enriched for *EGFR*-WT/*KRAS*-mutated tumors, their associated pathway biology was investigated by IPA. Compared to C1/C3 subtypes, pathways commonly enriched in C2 samples included those related to cancer progression, cell death and survival, metabolism, immune response, and metastasis (Figure 7G, Figure S7H, Table S7), commensurate with worse outcomes even at early pathologic stages. In addition to higher immune scores (ESTIMATE) in C2 patients (Figure S7I), sex-specific differences were also observed, with EGFR and BCR signaling axes highly activated in females and KRAS pathway upregulation in males (Figure 7G), highlighting different routes of LUAD progression that converge into poor outcome. Many proteasome components were commonly activated in C2 patients (Table S7), with weighted correlation network analysis[107] revealing co-expression of proteasome complexes with APOBEC proteins (3A, 3B, 3G). Notably, however, immunoproteasome complex elements PSMB8, PSMB9 and PSMB10 were elevated in C2 females, while only the standard proteasome components were upregulated in C2 males (Figure 7G, Figure S7H). Expression of PSMB8, PSMB9 and PSMB10 have been reported to be associated with better response to immune checkpoint inhibition in cancers[108–110]. The active immune response in C2 females suggests that they may benefit from combined EGFR/immune checkpoint inhibition therapy, with immunoproteasome-selective inhibitors as potential combination options[111–114]. By contrast, treatment targeting proteasome inhibition and *KRAS* mutation might be a beneficial strategy for C2 late-like males[115,116].

Candidate therapies and biomarkers with potential translational utility

We applied a therapeutic target prioritization method that integrates characteristics of tumors with functional genomics data in cell lines[117]. A predicted drug target was identified if its protein, activating phosphorylation site, or other PTM site was significantly increased in tumors vs paired NATs for a particular LUAD subset (*EGFR, RBM10, STK11* mutation and *ALK* fusion), and if gene knockdown by shRNA induced loss of fitness in LUAD

cell lines that conformed with that subset (Figure S8A–F and Table S8). The top 15 overexpressed targetable dependencies found by this approach spanned a range of genomic aberrations and corroborated other findings reported in this study (Figure 8A). SF3B1, an established splicing factor, was found to be overexpressed across multiple aberrations (Figure S8A, B, E, and F), similar to our previous analyses in the pan-cancer setting[117,118]. MET was found to be an overexpressed dependency in tumors with *EGFR-L858R* (Figure S8G). MET inhibition has been successfully used in lung cancer patients harboring *EGFR* mutations and *MET* amplification[119,120,121], demonstrating the ability of this approach to identify relevant therapeutic opportunities. XPO1, a protein transporter that shuttles an array of tumor suppressors and growth regulatory proteins[122–124], was a dependency in multiple molecular contexts. The XPO1 inhibitor Selinexor is an approved drug for diffuse large B-cell lymphoma and is currently being investigated in several clinical trials[122] including NSCLC (NCT03095612). Proteome-centric landscapes of current and possible future therapeutic targets include those unique to particular LUAD subsets (Figure S8G–H) or shared amongst subsets (Figure 8A).

Targets were classified into five tiers according to their actionability (Table S8). Tier 1 included 11 genes like *EGFR* that are targeted by approved oncology drugs. Tier 2 consisted of six genes targeted by drugs approved for other indications, providing drug repurposing potential. Ten genes in Tier 3, such as heat shock protein HSP90AA1, are targeted by investigational or preclinical drugs, indicating potential utility as stratification biomarkers for experimental therapies. Tier 4 comprised eight genes from protein families frequently targeted by small molecule inhibitors, and Tier 5 included five membrane proteins, potential candidates for antibody or chimeric antigen receptor (CAR) T cell therapies. In addition to nominating therapeutic candidates, these results may provide insights into therapeutically relevant underlying biology. For example, multiple EGFR ubiquitination sites were found to be overexpressed relative to NATs in both *EGFR*-WT and *EGFR*-mutated tumors (Figure 8A), but ubiquitination at EGFR K716 was exclusively overexpressed in *EGFR*-mutated subsets despite EGFR protein being underexpressed (Figure 8B, left two panels). LUAD cell lines harboring these *EGFR* mutations also showed greater dependency to EGFR knockdown by shRNA compared to *EGFR*-WT cell lines (Figure 8B, right panel). Ubiquitination at K716 triggers internalization of activated surface EGFR[125]. Rapidly internalizing activated mutant forms of EGFR could be attractive targets for antibody-drug conjugates (ADCs), since internalization of ADC antigens is critical to delivery and activation of the targeted cytotoxic agent. Likewise, CSF1R is a surface receptor, internalization of which is regulated by ubiquitination[126]. The CSF1R K812 ubiquitination site was exclusively overexpressed in *EGFR-E19del* tumors independent of its global protein levels and demonstrated selective *EGFR* dependency in LUAD cell lines harboring *EGFR-E19del* (Figure 8C and Figure S8G). These results suggest targeting CSF1R as a therapeutic opportunity in *EGFR-E19del* tumors and prioritize CSF1R K812 as a candidate for follow-up mechanistic studies since its role in receptor internalization is not established. Other therapeutic hypotheses inviting further study include heat shock proteins HSP90B1 and HSP90AA1, overexpressed dependencies in multiple LUAD subsets (Figure 8A). In the Profiling Relative Inhibition Simultaneously in Mixtures (PRISM) drug screen[127], LUAD cell lines harboring *STK11* mutations were more sensitive to the HSP90AA1 inhibitor

tanepsimycin than *STK11* WT cell lines. Tanepsimycin has been tested in lung cancer patients in a phase I clinical trial (NCT00004065); stratification by *STK11* mutation status may be warranted (Figure 8D).

To identify candidate drug targets for samples assigned to proteomic NMF clusters, we first trained a global proteome-based XGBoost classifier to predict cluster membership (Figure S8I–J and STAR methods), achieving a mean AUROC of 0.97 using 100 proteins designated tumor epithelial-specific by deconvolution analysis, and then applied the classifier to LUAD cell line proteomic data in DepMap. Predicted LUAD cell line cluster assignments aligned reasonably with core NMF LUAD tumors based on protein Hallmark ssGSEA scores (Figure S8K). Numerous targetable dependencies for NMF proteomic clusters were identified (Figures 8E–G, Table S8). SF3B1 (C1/C2) reinforced our observations of splicing deregulation in subsets of tumors. Co-occurrence of TOP2A and AURKB highlighted a potential co-dependency for C2 unstable tumors, which were also found to have high IGF2BP3 relative to C1 and C3. KIF11, another dependency associated with poor-outcome C2 tumors, has been reported to be prognostic and a targetable LUAD vulnerability[128]. These and many other examples suggest the potential of our approach to nominate targets for further (pre)clinical evaluation.

In summary, our systematic analysis provides a druggable proteomic atlas for LUAD subsets and highlights the translational utility of leveraging multi-omic data to nominate promising drug targets that may be vulnerable to multiple therapeutic modalities (Figure 8H).

Discussion

The presented study, a product of the international and interdisciplinary CPTAC/ICPC program, provides a large and comprehensive molecular data resource for LUAD: a proteogenomic compendium representing the demographic, etiological, clinical, and molecular complexity of LUAD, addressing key gaps in our understanding of this diverse and deadly disease. These major findings are summarized in Figure 8H. The intrinsic molecular structure of LUAD was revealed in clusters enriched not only as expected for key driver mutations (*EGFR*, *KRAS*), but also for degrees of genomic instability, markers of proliferation, immune signaling, endogenous mutagen and environmental carcinogens, high-risk early stage, and outcome. Extensive PTM characterization provided insights into oncogenic mechanisms, for instance in *ALK*-fusion-driven disease; allowed the bridging of functional and structural proteomics by examining the impact and localization of PTMs on protein structures, for example in the acetyltransferase EP300; and supported the use of cell type-specific modification patterns to study cancer immunity and the tumor microenvironment through the lens of protein modifications. The scope and scale of the tumor cohort facilitated analyses of signature-determined carcinogen exposures ranging from nitrosamine to nitro-PAHs/PAHs, their complex interaction with sex determinants, and their relevance to early recurrence. Comprehensive molecular characterization of NATs revealed the contribution of environmental carcinogens to early events of persistent inflammation, ECM remodeling, immune modulation, and metabolic dysregulation in non-cancerous tissues, collectively contributing factors to LUAD tumorigenesis. These

data suggest a need to increase awareness of and develop a risk prediction model for environmental carcinogens in addition to tobacco smoke.

Several methodological and analytical innovations are presented herein. The heterogeneity of genomic instability led us to develop a synthetic metric designated “breakage intensity clustering” that provides a prognostic measure of chromosomal instability by characterizing the positional distribution of genomic breakpoints and identifies highly fragmented genomes associated with poor outcomes. We identified IGF2BP3 as a prognostic proteomic biomarker of chromosomally unstable, unfavorable lung cancer, and a panel of biomarkers that might facilitate the development of effective chemopreventive and monitoring strategies for environmental carcinogens. These biomarkers could be used to broaden lung cancer screening to other high-risk populations left out of current CT screening guidelines. We applied innovative statistical approaches to existing single-cell compendia to characterize gene expression in the LUAD tumor microenvironment with cell-type resolution. At the translational level, we built upon previous findings to improve the understanding of early stage, “late-like” LUAD, exposing underlying biology, generalizing this molecular subtype to female never-smokers and male smokers of diverse ethnicity, and identifying potential approaches to patient stratification beyond clinical staging. Further, by integrating tumor proteomics with cell-line drug screens we built a comprehensive atlas of therapeutic vulnerabilities for distinct LUAD genetic subtypes, including nomination of sex-specific combination therapies in late-like disease.

Limitations of the study

Focused on tissues collected from Caucasian and Asian patients, the cohort underrepresents other demographics, potentially affecting the generalizability of the findings. Though a large cohort, the statistical power is low for certain subgroup analyses, particularly those involving less frequent mutations or specific demographic groups. The cross-sectional nature of the study precludes longitudinal analysis, important for understanding the dynamics of oncogenesis and disease progression. The analysis relies on batch correction to support combination and integration of heterogeneous cohorts, which can dampen biological signals and introduce unintended biases in the data. All molecular data is derived from a single cryopulverized piece of bulk tissue, which is optimal for integration of the several data types but can obscure intratumoral heterogeneity and conflate signals from tumor epithelium and stroma. Our use of single-cell-based deconvolution mitigates the latter but does not afford the granularity of true single-cell analysis. Associations with environmental exposures are based on self-report, which may be subject to recall bias. This limitation is partially mitigated by using genomic data to assess carcinogen signatures as surrogates of exposure. Additionally, the high level of association between ancestry, sex, exposure, and mutation – for instance, the high frequency of *EGFR* mutation in Asian female never-smokers – makes it difficult to tease out the relative contributions of those variables to the underlying disease biology even when controlling for confounders.

This study provides a comprehensive proteogenomic characterization of LUAD, addressing key questions about the disease's molecular underpinnings and highlighting potential clinical applications. The findings help pave the way for personalized treatment strategies and

offer promising directions for future research. Like other studies of its type, it intends to present actionable observations while also being hypothesis-generating and providing a resource to the wider community of basic science and clinical investigators. Validation studies have been incorporated to functionally assess key findings, such as the character of IGF2BP3 as a biomarker and the context-specific impact of select exogenous carcinogens, but additional studies will be needed to validate other biological findings, and further investigation into the functional roles of other identified biomarkers and therapeutic targets, such as specific post-translational modifications, will be crucial for translating these findings into clinical practice. Continued efforts to integrate diverse patient cohorts, longitudinal data, and detailed annotation regarding region-specific carcinogen exposure will be essential in advancing our understanding and treatment of this complex disease.

Resource availability

Lead contact

Further information and requests for resources should be directed to the lead contact, Michael A. Gillette (gillette@broadinstitute.org)

Material availability

This study did not generate new unique reagents.

Data and code availability

Genomic and transcriptomic data files for the ICPC cohort can be accessed at the database of Genotypes and Phenotypes (dbGaP) through Study accession: phs001954, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001954. Genomic and transcriptomic data files for the CPTAC cohort can be accessed at the Genomic Data Commons (GDC), https://portal.gdc.cancer.gov/, with dbGaP Study Accession: phs001287, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001287. Proteomic data (including PTMs) for both cohorts can be accessed at the Proteomics Data Commons, https://pdc.cancer.gov/pdc with Study accession numbers ICPC.A proteome: PDC000219, ICPC.A phosphoproteome: PDC000220, CPTAC.A proteome: PDC000153, CPTAC.A phosphoproteome: PDC000149, CPTAC.A acetylome: PDC000224. CPTAC.A ubliquitylome (PDC000627), ICPC.B proteome (PDC000563), phosphoproteome (PDC000564), CPTAC.B proteome (PDC000489), phosphoproteome (PDC000490), acetylome (PDC000491), and ubiquitylome (PDC000492) will be available upon publication. All data discussed in the paper, including those from previous publications, are available at: https://pdc.cancer.gov/pdc/cohort/cptac-icpc-luad

STAR★Methods

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Cohort description: Sample annotation (including demographics such as age, sex, race, etc.), processed and batch corrected and normalized data files are provided in Tables S1–S2.

CPTAC.A and CPTAC.B cohorts: The tumor, normal adjacent tissue (NAT) and whole blood samples used in this manuscript were prospectively collected for the CPTAC project and, while representing a substantial expansion of the previously published cohort, are otherwise identical to the sample acquisition described in our previous publication[10]. Institutional review boards at tissue source sites reviewed protocols and consent documentation adhering to the Clinical Proteomic Tumor Analysis Consortium (CPTAC) guidelines. Clinical data were obtained from tissue source sites and aggregated by an internal database called the CDR (Comprehensive Data Resource) that synchronizes with the CPTAC DCC (https://cptac-data-portal.georgetown.edu/). Clinical data can be accessed and downloaded from the DCC (Data Coordinating Center). Demographics, histopathologic information, and treatment details were collected. LUAD histopathology was confirmed for all cases by at least 2 expert pathologists based on high resolution images of H&E sections.

ICPC.A and ICPC.B cohorts: The paired tumor, NAT and whole blood samples used in this manuscript were prospectively collected between July 2016 and July 2020 for the Taiwan Cancer Moonshot project and, while representing a substantial expansion of the previously published cohort, are otherwise identical to sample acquisition described in our previous publication[11], and followed guidelines set by the research ethics committees of Academia Sinica, National Taiwan University Hospital (NTUH), Taichung Veterans General Hospital (TCVGH), and Chung Shan Medical University Hospital (CSMUH) which approved the study. All patients provided written informed consent.

Cell lines: Human lung cancer cell lines A549 and PC9 were cultured in RPMI medium supplemented with 10% fetal bovine serum. NCI-H1975 (ATCC, catalog number CRL-5908), and its derivative aneuploid clones were selected following treatment with reversine according to the protocol in[134].

METHOD DETAILS

Sequencing sample preparation and analysis

CPTAC.A and CPTAC.B cohorts: This study sampled a single site of the primary tumor from surgical resections, with an internal requirement to process a minimum of 125mg of tumor tissue and 50mg of NAT. DNA and RNA were extracted from tumor and NAT and downstream processing for whole exome sequencing (WES), whole genome sequencing (WGS), RNA and miRNA sequencing proceeded according to methods identical to those previously described in[10].

ICPC.A and ICPC.B cohorts: This study sampled the primary tumor from surgical resections from three medical centers, with an internal requirement to process a minimum of 40 mg of tumor tissue and NAT and 2 ml peripheral blood. DNA was extracted from tumor, NAT and Buffy coat, and RNA was extracted from tumor and NAT for downstream processing for whole exome sequencing (WES), whole genome sequencing (WGS) and RNA sequencing.

Array Based Methylation Analysis

CPTAC.A and CPTAC.B cohorts: Methylation EPIC arrays were used for DNA methylation data and are only available for the CPTAC.A and CPTAC.B cohorts. The methods employed are identical to those previously described in[10].

MS methods sample preparation and analysis

CPTAC.A and CPTAC.B cohorts: Proteome, phosphoproteome, and acetylome analyses were done in a serial fashion as described previously[10]. Ubiquitylome analysis was done in parallel as described in[131] and was performed in two batches. CPTAC.B samples, excluding the first TMT plex, were also enriched and profiled for N-glycosylation using flowthrough of acetylome experiments; these data have not been used in this study but will be available as a resource. The first batch of the CPTAC cohort (CPTAC.A) was processed using TMT10 reagents and spanned 25 TMT10 plexes. This cohort has been published previously[10] except for ubiquitylomes, which are new to this study. The CPTAC.B cohort (new cohort) was processed in two batches, with the first (CPTAC.B1) consisting of 11 TMT11 plexes and the second (CPTAC.B2) of 14 TMT11 plexes. For ubiquitylome analysis, the first batch consisted of 14 TMT10 plexes (CPTAC.A), and the second batch consisted of 15 TMT11 plexes (CPTAC.B). The protocols for protein extraction, tryptic digestion, TMT10/TMT11 labeling of peptides, peptide fractionation by basic reversed-phase liquid chromatography, phosphopeptide enrichment using immobilized metal affinity chromatography, and LC-MS/MS, were identical to[10] with the few exceptions listed below.

1. CPTAC.B cohort leveraged TMT11 instead of TMT10
2. CPTAC.B acetylomes were run on a 110-minute LC gradient instead of 240 minutes as was employed for CPTAC.A.
3. Ubiquitylomes for CPTAC.B cohorts were enriched using magnetic di-Gly remnant beads and a further-optimized method[132]
4. Proteomes for the CPTAC.B cohort were analyzed on a Thermo HFX instead of a Thermo Lumos instrument.

To facilitate quantitative comparison between all samples across experiments, a common reference (CR) sample was included in each TMT10/TMT11 plex as described in[10] .

ICPC.A and ICPC.B cohorts: The primary tumor and the paired NAT from surgical resections were processed following our previous publication[11] and were processed using TMT10plex reagent in two batches. The first batch of the ICPC cohort (ICPC.A) spanned 27 TMT10 plexes and has been published previously[11]. The ICPC.B cohort (new cohort) consists of 31 TMT10 plexes. The protocols for protein extraction, in-solution tryptic digestion, TMT10 labeling of peptides, peptide fractionation by basic pH reverse-phase liquid chromatography, phosphopeptide enrichment using immobilized metal affinity chromatography and LC-MS/MS were identical to the previous study[11] with the few exceptions listed below.

1. The reference samples were used for quality control in each TMT10 batch experiment. The third reference sample was prepared by pooling 15 tumor

tissues from early-stage patients and labeled for the TMT-131 channel in ICPC.B cohorts.

2. The phosphoproteome enrichment of ICPC.B cohorts was performed using the NTA resin (Qiagen#NTA30410).

3. All fractionated samples were analyzed in single injections on a 240-minute LC gradient for ICPC.B.

4. Phosphoproteomic analysis from batch 2 to 31 in ICPC.B was performed on a Thermo Orbitrap Eclipse instead of the Orbitrap Lumos used in ICPC.A.

Cell line experiments

Proteomics and peptide-pull down: The cell line phosphoproteome experiment was part of a larger set of experiments within the Novartis and Broad Institute Cancer Cell Line Encyclopedia (CCLE) consortium and will be published elsewhere. Data is shown only for the selected phosphopeptides and the detailed phosphoproteome analysis will be published separately. The 8 cell lines, their genomic alterations, the target, and the inhibitors used are listed as follows. None of the selected peptides were identified in KPL1 cells and therefore that line is excluded in Figure 2E.

1. Bladder cancer RT112 cell line with *FGFR-TACC3* fusion, treated with FGFR inhibitor Infigratinib at 500nM concentration

2. Endometrial cancer, AN3CA cell line with *FGFR2 N549K* mutation, treated with FGFR inhibitor, Infigratinib at 500nM concentration

3. Breast cancer, T47D cell line with *PIK3CA* mutation, treated with PI3KA inhibitor Alpelisib at 1uM concentration

4. Breast cancer, KPL1 cell line with *PIK3CA* mutation, treated with PI3KA inhibitor Alpelisib at 1uM concentration

5. Lung cancer, H3122 cell line with *ALK* fusion, treated with ALK inhibitor Ceritinib at 300nM concentration

6. Lung cancer, H2228 cell line with *ALK* fusion, treated with ALK inhibitor Ceritinib at 1uM concentration

7. Colon cancer, Ls513 cell line with *KRAS* mutation, treated with MEK inhibitor Trametinib at 30nM concentration

8. Colon cancer, H747 cell line with *KRAS* mutation, treated with MEK inhibitor Trametinib at 50nM concentration

RT112 and AN3CA cells were treated with DMSO for 6 hours and corresponding inhibitors for 6, 24, and 96 hours in duplicates. All the other cell lines were treated with DMSO in duplicates, or with inhibitors in triplicates for 6 and 24 hours. Each cell line was processed in a single TMT 10-plex experiment representing all the DMSO and drug treatment time points and replicates. Cells were lysed, digested, and TMT labeled as described above in the “CPTAC.A and CPTAC.B cohort” section. Mixed sample after labeling for each TMT plex was processed for phospho-tyrosine enrichments as previously described[133].

Resulting pY enriched samples were analyzed on a Fusion Lumos Mass Spectrometer coupled to an Easy-nLC 1200 using Picofrit 75uM ID columns (New Objective) packed in house with ReproSil-Pur C18-AQ 1.9uM beads to a length of 25cm. For each sample 2 × 4/9uL injections were performed and each sample was separated at 200nL/min flow rate with solvent A of 0.1% formic acid/3% acetonitrile, solvent B of 0.1% formic acid/90% acetonitrile and a gradient of 2 – 6% B in 2 min, 2 – 35% B in 12 min, 35 – 60% B in 8 min, 60 – 90% B in 3 min, followed by 2 × 10 min washes at 90% and 50% B. MS parameters included: MS1 resolution of 60,000, mass range of 350 to 1800 m/z, AGC target of 4e5, and max inject time of 50ms, isolation width of 0.7amu, normalized CE of 34%, MS2 resolution of 50,000, AGC target of 1e5, and max inject time of 250ms, cycle time of 2 sec and dynamic exclusion of 5 sec.

Data were searched and analyzed using Spectrum Mill MS Proteomics software (Broad Institute) using a Uniprot human database downloaded on 12/28/2017 as described previously[10,131]. Reporter ion ratio for each channel to a median of all channels was used for quantitation of samples in each plex.

For peptide pull down experiments, the LUAD H3122 cell line, which harbors an *EML4-ALK* fusion, was used. Cells were cultured in 10% fetal bovine serum containing Dulbecco's Modified Eagle Medium (DMEM). Cells at 70% confluence were washed twice with ice cold PBS followed by lysis in non-denaturing buffer (50mM Tris pH 7.5, 150mM NaCl, 1mM EDTA, 1% Triton X 100, Roche protease inhibitor, and phosphatase inhibitors NaF, PIC2 and 3). A total of 1mg cell lysate was incubated with 10ug of biotinylated peptides and 10 ul of magnetic streptavidin slurry and incubated at 4° C for 1 hour. The beads were washed three times with a lysis buffer followed by transfer to a fresh Eppendorf 1.5ml tube and two additional washes with ice-cold Tris wash buffer (59nN Tris HCL pH 7.5). After the beads were washed, the entirety of the wash buffer was removed and 80ul of Urea digestion buffer was added (1mM DTT, 5ug/mL Trypsin). Beads were incubated for 1 hour at room temperature on a Thermo-mixer (1000rpm). The ~80ul supernatant was transferred to a fresh 1.5ml tube. An additional 80ul of digestion buffer was added to the beads for another 30 min room-temperature incubation, followed by the transfer of the supernatant to the 1.5ml tube containing the previous 80ul, making the total volume ~ 160ul. The beads were washed twice with 60ug Urea-Tris buffer (2M Urea in 50mM Tris HCL pH 7.5) and pooled with the 160 ul supernatant, making the total volume ~ 280–300ul. The tubes were given a quick spin followed by another 1.5ml tube transfer step to avoid any residual beads. Proteins were reduced with 4mM DTT at room-temperature for 30 mins, followed by alkylation using 10mM IAA for another 45 minutes. 0.5ug of Trypsin was added and digestion was performed overnight. Samples were acidified to a final concentration of 1% formic acid, and stage-tipped as described above. Peptides were analyzed on a Thermo Orbitrap Exploris 480, coupled with a Thermo™ Vanquish Neo UHPLC system (ThermoFisher), equipped with a 25 cm house-packed New Objective PicoFrit™ column that was packed with 1.5 um C18 beads (Reprosil-Pur, Dr. Maisch), with a 68-min gradient. The separation was performed at 200 nL/min, starting at 1.6% B - 4.8% B in 1 min, 24% B in 50 min, 48% B in 5 min and 80% B in 2 min. The column was then washed at 72% B for 10 min. Data acquisition on the Thermo Orbitrap Exploris 480 was performed using data-independent acquisition (DIA). MS1 spectra were acquired in loop

control mode (All), and collected at a resolution of 120,000, with a scan range of m/z: 400–1200. Automatic gain control target was set to 3E6, with a maximum injection time of 25 ms. For DIA analysis, a variable isolation scheme consisting of 23 windows covered a mass range of m/z: 400–850. MS/MS scans were collected at a resolution of 30,000, with an AGC target of 1E6 and maximum injection time of 54 ms. DIA precursors were fragmented using higher-energy collisional dissociation (HCD) with a normalized collision energy (NCE) of 27% NCE with default charge state set to +3. The DIA method cycle time was ~1.5 seconds, equating to a median of 4–6 data points per peak (DPPP).

The peptide sequences are as follows, with the phospho residue indicated by Y(p):

HDLBP_Y437, unmodified: MVKDLINRMDY(p)VEINIDHKFH, MVKDLINRMDYVEINIDHKFH

EML4_Y226, unmodified: KDVIINQEGEY(p)IKMFMRGRPI, KDVIINQEGEYIKMFMRGRPI

SND1_Y109, unmodified: IENKTPQGREY(p)GMIYLGKDTN, IENKTPQGREYGMIYLGKDTN

LDHA_Y239, unmodified: EVHKQVVESAY(p)EVIKLKGYTS, EVHKQVVESAYEVIKLKGYTS

Cell line functional assays: Human lung cancer cell lines A549 and PC9 were cultured in RPMI medium supplemented with 10% fetal bovine serum. The cells were treated with 1-nitropyrene (1-NP) or Benzo[a]pyrene (BaP) at the indicated concentrations for 48 hours, with dimethyl sulfoxide (DMSO) as the solvent control (0μM). Subsequently, the cells were seeded for wound-healing and invasion assays. For the wound-healing assay, cells were seeded into a 2-well culture insert to create a defined cell-free gap. Removal of the insert was designated as the 0-hour time point, and the cell-free gap served as the baseline. Wound closure percentages were quantified at the indicated time points using the MRI wound-healing tool in ImageJ, analyzing three different fields for each sample. For the invasion assay, the 8 μm pores of Transwell 24-well inserts were coated with a thin layer of Matrigel matrix (2.08 mg/mL). A total of 1×10^4 A549 and PC9 cells were seeded onto the gel-coated inserts and incubated for 18 hours. Invasive cells were counted under a light microscope at 200× magnification. For the cell viability assay, cancer cells were seeded onto 96-well plates (3×10^3 cells/well) overnight and treated with carcinogens. After 48 hours incubation, cell viability was evaluated by thiazolyl blue tetrazolium bromide (MTT) assay. Briefly, 10 μl of the MTT solution (5mg/mL) was added to each well and incubated for 4 hours at 37° C. Then, 100 μl of DMSO was added and mixed vigorously to solubilize colored crystals. The absorbance at 570 nm (630 nm as the reference) was measured by using a multi-well scanning spectrophotometer. All data were statistically analyzed using GraphPad Prism (v9.0).

For experiments described in Figure S3O, establishment of aneuploid clones was performed according to the protocol described in[134]. Briefly, lung adenocarcinoma cell line NCI-H1975 purchased from ATCC (CRL-5908) was grown in ATCC-modified RPMI-1640 medium with 10% FBS as prescribed. A subconfluent plate of cells was treated with 0.5uM reversine

(ApexBio cat# A3760) and DMSO control for 48 hours, then split and allowed to recover without reversine for 5 to 7 days, followed by single cell cloning over 6–8 weeks in 96-well plates, gradually scaled up to 6-well plates and subsequently to T75 flasks. The clones were assessed for aneuploidy by microscopy and flow cytometry and select aneuploid clones were used to perform western blots for IGF2BP3 (Abcam cat# ab179807) and GAPDH (Cell Signaling cat# 3683) according to standard protocols.

Immunohistochemistry: The immunohistochemistry (IHC) assay was performed on the Ventana discovery automated system as per the previously described protocol[135]. Briefly, heat-mediated antigen retrieval under cell conditioning media 1 (CC1) was followed by application of the primary antibody and later development of chromogenic signal (in brown) utilizing OmniMap DAB kit.

Post-IHC staining, the signal was identified in the cytoplasmic and membranous compartments of the neoplastic cells. The percentage (%) and intensity of the staining (as 0= absent; 1= mild; 2= moderate; and 3= strong) were recorded. To provide a semiquantitative total product score (TS), a sum of percentage area multiplied by staining intensity was rendered out of 300 as previously described[136]. For every TMA core, a final product score was calculated and recorded in a tabulated fashion.

RNAseq data analysis

Data QC and filtering: To establish a unified RNAseq dataset, the GTEx/TOPMed pipeline used in the CPTAC Pan-Cancer projects[39] was applied to both the CPTAC and ICPC cohorts. During the initial QC phase, a high rate of mitochondrial RNA (mtRNA) and ribosomal RNA (rRNA) reads along with a lower-than-expected mapping rate (rRNA mapping rate >10% and overall mapping <90%) was encountered. These mapping rates were likely attributable to variations in library preparation protocols across different centers. To address this issue, a reference containing rRNA/mtRNA genes was developed to facilitate read mapping. Subsequently, any reads mapping to this reference were filtered out from the FASTQ files.

RNAseq data processing and quantification: Following the QC and filtering steps, the RNA samples were processed using the GTEx/TOPMed pipeline, as previously described[137]. In summary, the entire cohort was aligned to the human reference genome GRCh38 using STAR v2.7.5a. Subsequently, quality control checks and gene-level quantification, measured in Transcripts per Million (TPM) units, were conducted using RNA-SeQC 2.3.6[138]. Additionally, isoform expression was quantified using RSEM[139]. TPM values were calculated for the subset of protein-coding genes, $\log_2$-transformed, and median-centered prior to further analysis.

Genomics Data Analysis

Data processing: To standardize and harmonize the CPTAC and ICPC datasets, several QC and formatting procedures were conducted:

1. Adding read group information

2. Realigning the BAM files to the hg38 reference
3. Marking duplicates
4. Merging the BAM files
5. Sorting and indexing the BAM files
6. Running base recalibration using the latest GATK tools

After completing these steps, genomic data processing for this cohort was performed using the same pipelines as those used for the CPTAC Pan-Cancer analysis[39].

GISTIC: To identify significantly amplified or deleted focal as well as arm level copy number alterations (CNA), the Genomic Identification of Significant Targets in Cancer (GISTIC2.0) algorithm was applied[140]. Events with q value <0.25 were deemed significant. Run parameters were identical to the Pan-Cancer projects[39]. GISTIC was also used to map segment-level CNA to gene-level CNA. Normalized (i.e., divided by 2), $\log_2$-transformed gene-level CNA values were used for several subsequent analyses, including NMF clustering.

MutSig2CV: To mitigate potential sequencing artifacts and undetected germline variants, a dedicated panel of normals comprising samples from both ICPC and CPTAC cohorts was established and subsequently filtered the somatic reads. Following this, the MutSig2CV tool[141] was employed to pinpoint significantly mutated genes within the cohort, alongside their mutation frequencies. These findings were limited to genes listed in the Cancer Gene Census[142]. Genes with a q-value < 0.1 were considered statistically significant.

Proteomics Data Analysis: Global proteomic, phosphoproteomic, acetylproteomic, glycoproteomic, and ubiquitylproteomic datasets were analyzed using the MSFragger[143] search engine via the FragPipe computational platform[144].

Identification and quantification: Raw mass spectrometry files were converted into open mzML format using the msconvert utility of the Proteowizard software suite and analyzed using the FragPipe computational platform with a TMT-10/11 workflow where proteins and PTMs in a TMT plex are ratioed to a common reference (CR) sample included in each plex. The quantification was performed independently for each cohort (CPTAC.B1, CPTAC.B2, CPTAC.A, ICPC.A, ICPC.B). Each cohort had its own CR sample present in every TMT plex for that cohort. Where possible, replicates of CR samples from prior cohorts were included in newer cohorts to enable cohort combination via bridging (see below). For the glyco-enriched data, the TMT11-virtual workflow was employed, in which a virtual bridge was internally created to link CPTAC.B1 and CPTAC.B2 cohorts. MS/MS spectra were searched using the database search tool MSFragger v3.8 against a harmonized Homo sapiens GENCODE34 protein sequence database appended with an equal number of decoy sequences. Whole cell lysate MS/MS spectra were searched using a precursor-ion mass tolerance of 20 ppm and allowing C12/C13 isotope errors −1/0/1/2/3. Mass calibration and parameter optimization were enabled. Cysteine carbamidomethylation (+57.0215) and lysine TMT labeling (+229.1629) were specified as fixed modifications, and methionine oxidation

(+15.9949), N-terminal protein acetylation (+42.0106), and TMT labeling of peptide N terminus and serine residues were specified as variable modifications.

For PTMome datasets other than glycoproteome, specific variable modifications were additionally included: phosphorylation (+79.9663) of serine, threonine, and tyrosine residues for phospho-enriched peptide search; acetylation (+42.01056) of lysine for acetyl-enriched peptide search; and ubiquitination (+114.04293) of lysine for ubiquityl-enriched peptide search. The search was restricted to tryptic peptides, allowing up to two missed cleavage sites. Peptide to spectrum matches (PSMs) were further processed using Percolator to compute the posterior error probability, which was then converted to posterior probability of correct identification for each PSM. The resulting files from Percolator were converted to pep.xml format, and with the PTM-enriched data set, pep.xml files were additionally processed using PTMProphet to localize the expected modification sites. The resulting files were then processed together to assemble peptides into proteins (protein inference) using ProteinProphet run via the Philosopher[145] toolkit v5.0.0 to create a combined set of high confidence protein groups. The combined prot.xml file and the individual PSM lists for each TMT experiment were further processed using the Philosopher filter command as follows.

Each peptide was assigned either as a unique peptide to a particular protein group or assigned as a razor peptide to a single protein group that had the most peptide evidence. The protein groups assembled by Percolator were filtered to 1% protein-level False Discovery Rate (FDR) using the target-decoy strategy and the best peptide approach (allowing both unique and razor peptides). The PSM lists were filtered using a sequential FDR strategy, keeping only those PSMs that passed the 1% PSM-level FDR filter, and mapped to proteins that also passed the global 1% protein-level FDR filter. In addition, for all PSMs corresponding to a TMT-labeled peptide, reporter ion intensities were extracted from the MS/MS scans (with a 0.002 Da window) using Philosopher, and the precursor ion purity scores were calculated using the intensity of the sequenced precursor ion and that of other interfering ions observed in MS1 data (within a 0.7 Da isolation window). The PSM output files were further processed using TMT-Integrator v4.0.5[146] to generate summary reports at the gene level and modification site level. TMT-Integrator used the PSM tables generated by the Philosopher pipeline as input and created integrated reports with quantification across all samples. First, PSMs were filtered to remove all entries that did not pass at least one of the quality filters, such as PSMs with (a) no TMT label; (b) precursor-ion purity less than 50%; (c) summed reporter ion intensity (across all channels) in the lower fifth percentile of all PSMs in the corresponding PSM.tsv file (2.5% for phosphopeptide enriched data); (d) peptides without phosphorylation (for phosphopeptide enriched data). In the case of redundant PSMs (i.e., multiple PSMs in the same MS run sample corresponding to the same peptide ion), only the single PSM with the highest summed TMT intensity was retained for subsequent analysis. Both unique and razor peptides were used for quantification, while PSMs mapping to common external contaminant proteins (included in the searched protein sequence database) were excluded.

Next, for each PSM the intensity in each TMT channel was converted into a log2-based ratio to the reference channel. The PSMs were grouped to the gene level, and the gene ratios were computed as the median of the corresponding PSM ratios after outlier removal.

Ratios were then converted back to absolute intensity in each sample by using the reference gene intensity, estimated by the sum of all MS2 reporter ions from all corresponding PSMs. To generate peptide-level and site-level tables, additional post-processing was applied to generate all non-conflicting phosphosite configurations using a strategy similar to that described in Huang et al.[147]. In doing so, confidently localized sites were defined as sites with PTMProphet[146] localization probability of 0.75 or higher. The same peptide sequences but with different site configurations, i.e., different site localization configurations or peptides with unlocalized sites, were retained as separate entries in the site-level tables. In the peptide-level tables, different site-level configurations were combined into a single peptide-level index, grouping PSMs with all site configurations together if they corresponded to the same peptide sequence. The tutorial describing all steps of the analysis, including specific input parameter files, command-line option, and all software tools necessary to replicate the results, are available at https://github.com/Nesvilab.

Glycopeptide analysis: Raw files of the global proteome and glyco-enriched samples were additionally searched for N-linked glycopeptides via MSFragger (version 3.8) and Philosopher (version 5.0.0). Parameters were as described for whole proteome search, except as follows. C12/C13 isotope errors of 0/1/2 were allowed, and methionine oxidation (+15.99491) was the only specified variable modification for glyco-enriched samples. Phosphorylation of serine, threonine, and tyrosine (+79.96633) was specified for phospho-enriched samples. “Nglycan” search mode was used, restricting glycosylation sites to the consensus sequence N-X-S/T, where X is any residue other than proline. A customized human N-glycan database was used that contained 252 compositions[148,149]. Diagnostic ion filtering for glycopeptide spectra was enabled with a minimum intensity threshold of 10% and the following list of oxonium ions considered: 204.086646, 186.076086, 168.065526, 366.139466, 144.0656, 138.055, 512.197375, 292.1026925, 274.0921325, 657.2349, 243.026426, 405.079246, 485.045576, 308.09761. Glycan Y ions of 203.07937 and 406.15874 were included in the search, along with a remainder mass of 203.07937 on peptide b and y ions (“b~/y~” ions). PSMs were further processed with PeptideProphet, using the extended mass model with a mass width of 4000, as described in[150]. Protein inference, FDR filtering, and reporter ion intensity extraction were accomplished as in the whole proteome search.

Glycan assignment and glycan-specific FDR filtering were subsequently performed in PTM-Shepherd as previously described[151]. Briefly, possible glycan compositions given the observed delta mass recorded by MSFragger were scored using the glycan fragment ions observed in the spectrum and filtered to 1% FDR by comparison to spectrum-based decoy glycans. Default settings were used except for consideration of a single ammonium adduct on possible glycan compositions. PTM-Shepherd assigned glycan compositions and confidence scores were written back to the PSM tables. The PSM output files were then processed with TMT-Integrator[152] v4.0.5 to generate summary reports at the gene, protein, peptide, site, and “multi-mass” levels from glycopeptide spectra. Multi-mass refers to the combination of glycan and site, i.e., each distinct glycan identified at a given site generates a separate entry. The PSM filtering and summarization processes were the same as for whole proteome searches, with the exception of restricting the PSMs considered to those

of glycopeptides and using the MSFragger-reported localization of the glycosite within identified peptides rather than PTMProphet[146].

Bridging and batch correction: Harmonization of genomics data was accomplished by processing both CPTAC and ICPC genomic and transcriptomic data using the same characterization pipelines. Some residual differences in the RNAseq data were minimized (see “RNAseq data QC and filtering”), followed by batch correction to merge the CPTAC and ICPC cohorts.

Merging the global proteome, phosphoproteome, and glycoproteome from the global proteome data involved a series of bridging and batch correction operations to ensure that the final comprehensive dataset used for analysis was consistent and free of batch effects. The CPTAC.B1 and CPTAC.B2 cohorts used different common references (CR) and were bridged based on a few replicates of the CPTAC.B1 CR included in the CPTAC.B2 cohort. Bridging was followed by batch correction to create the complete CPTAC.B cohort. This cohort was then merged with the CPTAC.A cohort to create the combined CPTAC cohort using bridging (via CPTAC.A CRs included in the CPTAC.B TMT plexes) and batch correction. The ICPC.A and ICPC.B cohorts used the same CR and were combined to create the ICPC cohort using batch correction. The CPTAC and ICPC cohorts were finally merged using bridging (ICPC CRs were included in the CPTAC cohort) followed by batch correction. The CPTAC and the ICPC datasets were further batch corrected separately for tumor and NAT samples to remove residual cohort-specific batch effects (Figure S1A).

The acetylome was bridged and batch corrected in an identical manner except that this was restricted to only the CPTAC cohort. Ubiquitylome data for the CPTAC cohort were run in two batches with identical CRs and were combined simply by batch correction. ICPC cohorts did not include acetylome or ubiquitylome profiling.

In all the above, bridging was accomplished by scaling the protein or PTM-site ratio in the source dataset by the ratio of the source CR to target CR, thereby bridging the source cohort to the target cohort. Batch correction was accomplished using the removeBatchEffect function in the limma R package[153].

OmicsEV was utilized to investigate RNA and protein data quality before and after batch correction using default parameters except for the cpu parameter, which was set to 4 to conserve memory[154]. Median-normalized datasets for tumors from each cohort were concatenated into a single matrix of gene level log ratios for each sample, and unlog transformed data was used as input. In the sample list, both the batch and class were set to represent the individual cohorts (CPTAC.B, CPTAC.A, ICPC.A, and ICPC.B for RNA; for protein, CPTAC.B was separated into CPTAC.B1 and CPTAC.B2). However, to conserve memory given the size of the combined cohort, machine learning-based prediction of the class label was not performed. Relevant metrics from OmicsEV for assessing batch correction included principal component analysis to determine the extent of batch effect before and after correction as well as metrics to assess the impact of batch correction on biological signal (“Correlation between proteins: within vs between protein complexes" and “Co-expression network-based function prediction”). To further evaluate the quality of

biological signal in the final dataset, OmicsEV was also used to compare the batch corrected protein data to the batch corrected RNA data.

Unsupervised multi-omic clustering: We used non-negative matrix factorization (NMF)-based multi-omic clustering using protein, phosphosite, RNA transcript and gene copy number alterations (CNA) as previously described[10,131]. Data from RNA expression, CNA, proteome and phosphoproteome from 383 tumor samples were used as input for the integrative multi-omic clustering. The CNA data was further filtered to only retain genes significantly amplified or deleted (as determined by GISTIC2). All data tables were then concatenated and all rows with missing values were removed. To remove uninformative features from the dataset prior to NMF clustering we removed features with the lowest standard deviation (bottom 5th percentile) across all samples. Each column in the data matrix was further scaled and standardized such that all features from different data types were represented as z-scores. The resulting data matrix of z-scores was converted into a non-negative input matrix, as required by NMF, as follows:

1. Create one data matrix with all negative numbers zeroed.
2. Create another data matrix with all positive numbers zeroed and the signs of all negative numbers removed.
3. Concatenate both matrices resulting in a data matrix twice as large as the original but containing only positive values and zeros and hence appropriate for NMF.

The resulting matrix was then subjected to NMF analysis leveraging the NMF R-package[155] and using the factorization method described in[156]. To determine the optimal factorization rank k (number of clusters) for the multi-omic data matrix, we tested a range of clusters between $k = 2$ and 8. For each k, we factorized matrix V using 50 iterations with random initializations of W and H. To determine the optimal factorization rank we calculated two metrics for each k: 1) the cophenetic correlation coefficient measuring how well the intrinsic structure of the data was recapitulated after clustering and 2) the dispersion coefficient of the consensus matrix as defined in[157], measuring the reproducibility of the clustering across 50 iterations. The optimal k was defined as the maximum of the product of both metrics for cluster numbers between $k = 3$ and 8.

Having determined the optimal factorization rank k, in order to achieve robust factorization of the multi-omic data matrix V, we repeated the NMF analysis using 250 iterations with random initializations of W and H and performed the partitioning of samples into clusters as described above.

The clustering method outlined above was also run on each of the individual single-omes for gene-level CNA, proteome, phosphoproteome and RNA expression. Except for the proteome, all other omes resulted in optimal clustering with k=4 clusters. For the proteome, NMF clustering using 406 tumor samples identified three clusters. The cluster assignments for the multi-ome clustering were compared to each of the single-ome clusters using Sankey diagrams to underscore similarities and differences between the clusters.

For functional characterization of clustering results by single sample Gene Set Enrichment Analysis (ssGSEA), we calculated normalized enrichment scores (NES) of cancer-relevant gene sets by projecting the matrix of signed multi-omic feature weights (Wsigned) onto Hallmark pathway gene sets[130] using ssGSEA[158]. To derive a single weight for each gene measured across multiple omics data types (protein, RNA, phosphorylation site, CNA) we retained the weight with maximal absolute amplitude. We used the ssGSEA implementation available on https://github.com/broadinstitute/ssGSEA2.0 using the following parameters:

- gene.set.database = “h.all.v6.2.symbols.gmt”
- sample.norm.type = “rank”
- weight = 0.75
- statistic = ”area.under.RES”
- output.score.type = “NES”
- nperm = 1000
- global.fdr = FALSE
- min.overlap = 10
- correl.type = ”z.score”

To test the association of the resulting clusters to clinical variables we used Fisher’s exact test (R function fisher.test) to test for overrepresentation in the set of samples defining the cluster core as described above.

The entire workflow described above has been implemented as a module in PANOPLY[159], which runs on Broad’s Cloud platform Terra (https://app.terra.bio/). The docker containers encapsulating the source code and required R-packages for NMF clustering and ssGSEA have been submitted to Dockerhub (broadcptac/panoply_mo_nmf:1_3, broadcptac/panoply_ssgsea:1_3).

Mapping of TCGA-derived LUAD subtypes: TCGA-derived LUAD subtypes from[9] (LUAD Subtype) and[3] (TCGA Subtype) were mapped onto the CPTAC-ICPC cohort by building a classifier based on the TCGA RNA-Seq expression data. For LUAD Subtype, the TCGA RNA data were filtered for genes with the top 25% standard deviation as described in[9]. Missing values in the CPTAC-ICPC RNA data were imputed using the mice R package by averaging over 15 multiple imputations. Then, the shared genes between the filtered TCGA RNA data and the imputed CPTAC-ICPC RNA data (2909 genes) were used as input for the classifier. Values were z-scored prior to building the classifier. Two classifiers were trained on the TCGA RNA data using the randomForest and glmnet packages in R. The glmnet classifier had the highest accuracy and was therefore used to predict subtypes for the CPTAC-ICPC RNA. Samples with glmnet probability < 0.6 were considered “Unassigned.” For TCGA Subtype, the TCGA RNA was filtered in the same manner as for LUAD Subtype. Missing values in the CPTAC-ICPC RNA data were imputed using *k*-nearest neighbors (k-NN). As for LUAD Subtype, two classifiers were trained, and the glmnet classifier had the highest accuracy and was therefore used to predict subtypes.

***RBM10*-mutant and *ALK*-fusion analysis:** Differential expression analysis of *RBM10* and *ALK* fusion was performed using limma[153]. Due to co-mutation with *EGFR* mutation, differential expression analysis of *RBM10* was performed between *RBM10* MUT *EGFR* MUT and *RBM10* WT *EGFR* MUT samples. Only *RBM10* mutants resulting in truncated protein products were considered. Other forms of *RBM10* mutations were excluded from the analysis. Differential expression analysis results of proteome and phosphoproteome were subjected to GSEA[160] and PTM-SEA[161], respectively.

Two external validation analyses were performed to confirm our findings in human and mouse data, respectively. For the human data, RNAseq and sample mutation information of the East Asian were downloaded from OncoSG[34]. Differential expression analysis using Limma was performed, and the result was supplied to GSEA using the C2 gene set from the MSigDB[162]. For the sgRMB10 mouse model, GSEA was performed by reusing differential expression results provided by the original authors[31]. The mouse M2 gene set from MSigDB was adopted[163].

CLUMPS-PTM analysis: To investigate the role of PTMs in cancer patients,CLUMPS-PTM was employed to detect clusters of correlated PTMs within protein 3D structures, similar to the work done in the PTM-dedicated Pan-Cancer paper[39]. This method is adapted from the original CLUMPS approach for detecting significant clusters of mutations in 3D protein structures[164], but modified to search for significant clustering of differentially ubiquitinated, acetylated, or phosphorylated sites. The updates include: (1) Reconfiguring the tool to recognize output from multiple mass spectrometry pipelines, including Spectrum Mill and FragPipe, (2) Upgrading the Gencode pipeline to incorporate multi-PTM peptides, rather than solely single ones, (3) Integrating the latest available PDB structures (as of October 2023) and adding AlphaFold v4, (4) Adjusting the algorithm to include ubiquitination sites.

Cell phenotype enrichment analysis: A random forest model[165] was trained using single-cell data from normal lung epithelial cells[166], focusing on highly variable genes (HVGs). To reduce bias from unbalanced sample sizes, 200 cells were randomly selected from overrepresented clusters, such as the AT2 cluster. This model was then applied to bulk RNA-seq data from CPTAC and an independent study[75] to predict their closest normal cell type. Due to the distinct data structures of single-cell and bulk datasets, a rank-based inverse normal transformation[167] was applied to both datasets to standardize their distribution. Following this transformation, a random forest classifier[168] was used to analyze the normalized data and predict the normal cell types with the highest prediction score in the LUAD bulk RNA data.

Instability measure: To calculate the SegLen-Q3 measure of CIN, the lengths of CNA segments were determined for each tumor sample. The 75th percentile of the segment lengths was selected as the CIN measure for each tumor, as this percentile effectively captures the population of 'large segments' present within a tumor sample. Alternative percentiles, such as the median and 25th percentile, were also evaluated but did not demonstrate any significant biological relevance. In particular, the median segment length was deemed unsuitable due to the observation of binomial segment length distributions

in numerous LUAD tumors. Consequently, the median size fails to accurately reflect the presence of either large or short segments within these tumor samples. The 75th percentile, in contrast, provides a more robust and biologically meaningful measure of CIN by focusing on the larger segments that are likely to have a greater impact on tumor biology and hence, was selected for the CIN measure.

Breakage Intensity-Clustering (BIC): To further break down the instability measure based on segment length, BIC, a CIN measure based on properties of CNA segment breakpoints (called breakpoints hereinafter), was developed. This measure focuses on the influence of breakpoints on segment length distribution through their positional distribution (clustering or dispersion) and occurrence rate. If breakpoints occur at a high frequency, the 75th percentile length of the segments (SegLen-Q3) will decrease. Conversely, if breakpoints are clustered together, the distribution of segment lengths will be skewed, affecting the SegLen-Q3 measure differently than if breakpoints were uniformly distributed along the genome

The rate of breakpoints was measured by counting the number of detected breakpoints per tumor. To measure the breakpoints clustering, an adapted version of the K-ripley[169] score was used. The K-ripley score (Kscore) can be defined for any radius of t, where Kscore will reflect the expected degree of clustering at a radius t of a randomly chosen breakpoint on the genome. The one-dimensional Kscore is defined as:

$$K(t) = \lambda^{-1} \sum_{i} \sum_{i \neq j} W^{-1}(l_i, l_j) I(d_{ij} < t)$$

where λ is the rate of breakpoints, I is the identity function and d_{ij} is the distance between the ith and jth breakpoints. The W function is an 'edge-correction' function. When analyzing a specific radius t, it is crucial to consider cases where the distance of breakpoint i to the start of the genome is less than radius t. In such instances, the score may underestimate the number of breakpoints at the radius t. This underestimation occurs because, for the distance beyond the edge of the genome, no breakpoints are counted, although theoretically, breakpoints could have been present if the ith point was located at a distance greater than t from the edge of the genome. To address this issue, the W function was introduced, which is defined as 1 when t does not extend beyond the edge of the genome. In cases where t extends beyond the edge of the genome, W represents the proportion of radius t that falls within the genome length. By employing this approach, it is assumed that the length extending beyond the genome edge has the same expected number of points as the region within the genome.

In this study, K_{tumor} was calculated for each tumor at the radius of 5Mb. We selected 5Mb as the radius as it is large enough to capture the focal clustering of breakpoints without going over the edge of chromosomes. For each tumor, K_{null} was also calculated, which is the Kscore for a random distribution of points (from a homogeneous Gaussian process) with the same λ value as the tumor. The final measure of clustering is defined as $ClustScore = K_{tumor} - K_{null}$.

Finally, both ClustScore and Intensity score of breakpoints were used to group tumors into 3 main BIC groups: (I) 'BIC-contiguous' for tumors with high ClustScore, (II) 'BIC-fragmented' for any tumor with low ClustScore and low Intensity score and, (III) 'BIC-intense' for any tumor with low clustering but high Intensity of breakpoints.

Both single resolution (InceptionResnet V1 and V2) and multi resolution (Panoptes4)[170] deep neural network models were built to predict DNA instability features. All models were trained to perform binary classification tasks. For BIC status, only low-clust-high-int and high-clust samples were included as contrasting categories. For Q3, partial stable samples were left out from the training set. Model performance was evaluated by cross validation with training/validation/test ratio of 8:1:1. Mean and standard error of patient level AUROC on the test split were calculated for models trained over 5 random splits. Metrics of the best performing architecture were shown for each task.

Estimation of stromal and immune scores: ESTIMATE[171] was utilized to infer immune and stromal scores based on RNA-seq data. Tumor purity was estimated via TSNet[172] using immune and stromal signatures from[171].

Immune signaling and deconvolution: To estimate the fraction of different cell types in the tissue microenvironment, a multi-omic based deconvolution integrating global proteomic and RNAseq data was performed using BayesDeBulk[62]. To perform the deconvolution, BayesDeBulk requires a list of cell-type specific markers for each cell type. To derive cell-type markers, an extensive database of single-cell data including ~1M cells from ~180 confirmed LUAD cases was leveraged to identify 57 unique cell types based on a review of LUAD single cell literature. Markers between cell types were generated using the scanpy *sc.tl.rank_gene_groups* function. The *t-test-overestim_var* method was used for comparisons with Benjamini-Hochberg correction for multiple comparisons. For BayesDeBulk estimation, 10,000 Markov-Chain Monte Carlo iterations (MCMC) were considered. The estimated fractions were derived as the median between the MCMC iterations after discarding a burn-in of 1,000 iterations.

Differential analysis of Hot and Cold tumors: A new analysis pipeline was implemented to perform cell-type specific differential analysis to determine which proteins/RNA were activated in tumor cells. Specifically, BayesDeBulk was used to estimate the protein/RNA levels in tumor cells and immune/stromal cells of "Hot" versus "Cold" tumors. For this analysis, CD8+/IFNG+ was utilized as the "Hot" subtype, while "CD8−/IFNG−" was the "Cold" subtype. Considering this stratification, the expression level of the j-th gene was modeled as the following function:

$$y_{i,j} = \pi_i x_{j,H} 1(i \in H) + \pi_i x_{j,C} 1(i \in C) + (1 - \pi_i) x_{j,S/I}$$

with π_i being the estimated tumor purity for sample i; $1(i \in H)$ an indicator function equal to 1 if the i-th sample belong to the Hot cluster and 0 otherwise; $1(i \in C)$ being an indicator function equal to 1 if the i-th sample belongs to the Cold cluster and 0 otherwise; and $x_{j,H}$, $x_{j,C}$, $x_{j,S/I}$ being the expression of the j-th gene in tumor cells of the "Hot"

tumors, “Cold” tumors and immune stromal cells, respectively. Here, purity parameters were considered as fixed and estimated via TSNet. BayesDeBulk was utilized in order to estimate the parameters of this model, i.e., $x_{j,H}$, $x_{j,C}$, $x_{j,S/I}$, and for each gene j. For BayesDeBulk estimation, 10,000 Markov-Chain Monte Carlo iterations (MCMC) were considered. The estimated parameters were derived as the median between the MCMC iterations after discarding a burn-in of 1,000 iterations.

Pathway enrichment: For each MCMC iteration, a pathway score was computed as a mean of expression z-scores of genes belonging to that pathway, separately for RNAseq and protein data sets. The significantly enriched tumor-specific pathways in “hot” and “cold” clusters were those that were simultaneously elevated (at 95% CI within the 9,000 MCMC iterations) compared to the other cluster and compared to the immune/stromal component.

Single cell validation of pathway enrichment

Sample Selection: Samples were selected based on several criteria: lung biopsies with a confirmed LUAD diagnosis that were confirmed non-enriched (without CD45+ immune gating). For identification of tumor vs. adjacent normal biopsies, given that 65% of the samples did not explicitly report on sampling location (i.e tumor versus adjacent normal tissue) a qualitative analysis of the cell type fraction was used. This approach allowed for the exclusion of samples that had an excess of normal cell types, such as ATII cells and alveolar macrophages. Following this, cell type fractions across each patient were normalized to sum to 100%, and samples were selected that contained at least a 5% total fraction of malignant LUAD cells, resulting ih a cohort of 73 patients representing a total of 656,673 single cells.

Hot vs. Cold Segmentation Strategy: Upon analyzing these patients with respect to cell type fraction differences, there was a distinct separation between samples with higher versus lower immune-cell content, leading to the categorization of tumors as either “hot” or “cold” on this basis. This categorization resulted in 33 patients from 10 independent publications classified with low immune content (cold) and 40 patients from 13 independent publications with high immune content (hot), providing a balanced division for differential expression analysis. In terms of cell numbers, 57,006 LUAD single cells were classified as coming from “cold” samples, while 14,373 LUAD single cells were classified as coming from “hot” samples.

DE Approach: The analysis strategy involved a dual approach: examining malignant LUAD cells both at the single-cell level and through a meta cell analysis, where the expression of malignant cells was averaged by patient. For the differential expression analysis, a nonparametric Wilcoxon rank-sum test was employed across ~25,000 genes, applying the Benjamini-Hochberg method for adjusting p-values.

Identifying Immune Signaling Modules: For each patient, the differentially expressed mRNAs and phosphosites were identified using Z-scored log-fold change comparing tumor to NAT. The top up- and down-regulated 500 mRNAs for each patient were converted into gene sets and submitted for transcription factor (TF) enrichment analysis with ChEA3[173]. The top up- and down-regulated 500 differentially phosphosites were then converted to

protein sets and submitted for kinase enrichment analysis using KEA3[67]. The returned ranks from the enrichment analysis were adjusted based on a baseline distribution that is created from random protein sets of size 500 with 1,000 replicates. Ranks were then adjusted according to the following formula to create Z-scores to help mitigate the effect of set length and the number of substrates or targets associated with each kinase or TF in the enrichment analysis results:

$$-(observed\_rank - baseline\_mean)/baseline\_SD$$

The next step identifies the kinases and TFs that mostly co-occur. This was achieved by scoring co-occurrence in individual patients from the Hot and Cold patient groups, respectively. Pairs are only scored if both the kinase and TF rank in the top 30 for the patient according to the following formula:

$$(31 - [TF\ Rank])/30 + (31 - [Kinase\ Rank])/30$$

The most co-occurred kinases and TFs for both the Hot and Cold patient groups (top 1%) are then visualized as heatmaps using hierarchical clustering. There are four heatmaps that summarize the most common kinase-TF co-occurrences: kinases and TF that are enriched in the up-regulated phosphosites and up-regulated mRNAs; kinases and TFs that are enriched in the up-regulated phosphosites and down-regulated mRNAs; kinases and TFs that are enriched in the down-regulated phosphosites and up-regulated mRNAs; and kinases and TFs that are enriched in the down-regulated phosphosites and down-regulated mRNAs. We term these heatmaps as up-up, up-down, down-up, and down-down.

From these heatmaps, clusters of kinases and TFs that occur more than once were extracted. The regulatory relationships such as activation and inhibition are then inferred from these clusters and visualized as a network, enabling the creation of a bipartite graph that summarizes the relationships between kinase and TF modules in the Hot and Cold tumor groups.

Validation of Kinase-TF associations: *STAT4* targets were sourced from Enrichr[68] using the term search from three categories: Transcription, Pathways, and Diseases/Drugs. This resulted in 18 sets describing *STAT4* targets with a total of 2,248 genes in their union. Processed gene set signatures of *MAPK8*, *MAPK12*, *MAP2K1*, *PTK6*, *TESK2*, and *IGF1R* L1000[70] CRISPR/Cas knockouts were downloaded from the SigCom LINCS website[174]. The processed gene set signatures available from SigCom LINCS were computed using the Characteristic Direction method[175]. *STAT4* targets' overlap with each signature was assessed with the Fisher exact test retaining any with a significant overlap. The overlapping genes between significant signatures and the *STAT4* targets were submitted to Enrichr[68] for analysis against the Reactome_2022 library, and those both significantly enriched for 'Innate Immune System R-HSA-168249' and also appearing in the top three returned terms were retained. Additionally, overlapping genes appearing in at least six of the retained signatures were extracted and included in the diagram.

Kinase enrichment in immune subtypes: To define upregulation of phosphosites in tumor samples, we calculated normalized tumor sample intensity scores for all phosphosites by adjusting tumor samples with the average of normal samples and standard deviation of tumor samples within each cancer type in phosphoproteomic data. The top 500 upregulated phosphoproteins were selected from the normalized tumor sample scores for each sample. We then performed KEA3 rank analysis using the KEA3[67] consensus Appyter[176] on the selected protein sets. Once derived, the kinase activity score for samples in one subtype was compared to the activity scores for samples in the other subtypes via the Wilcoxon rank-sum test. *p*-values were adjusted for multiple hypothesis correction via the Benjamini-Hochberg method.

Outlier analysis: Since glycosylation data is known to be enriched for cell membrane and extracellular matrix proteins, associations between immune subtype and protein glycosylation were investigated by employing a previously developed outlier analysis method[177] to investigate glycosylation data (Figure S4B). Meanwhile, to study the difference between *STK11*-mutated tumors and *STK11* wild type tumors, TMT-based proteomic data, phosphoproteomic data, and RNA-Seq data were investigated. Differential extreme value analysis (DEVA) performed using the blacksheepr Bioconductor package facilitated the generation of SummarizedExperiment objects, followed by the execution of the differential extreme value analysis (DEVA) with a fraction samples cutoff of 0.1. This criterion ensured the inclusion of features with a minimum occurrence of 10% in the group of interest as outliers. Outliers were identified as features (genes) with intensity values surpassing the median plus 1.5 times the interquartile range (IQR) for each dataset. For glycosylation and phosphoproteomic data, the aggregateFeatures parameter was set to TRUE to aggregate data at the gene level. Furthermore, only genes exhibiting a higher frequency of outliers in the group of interest compared to the outgroup were considered. Comparison between the group of interest and the remaining samples was conducted using Fisher's exact test on outlier and non-outlier counts. Resulting p-values underwent correction for multiple comparisons via the Benjamini-Hochberg method and False Discovery Rate (FDR) correction. These procedures were adapted from outlier analysis methodologies outlined in a prior CPTAC LUAD paper[10]. Subsequently, a list of outliers was generated with a false discovery rate (FDR) cutoff of 0.05. Using this gene list, pathway analysis was performed employing the enrichKEGG function from the R package clusterprofiler for KEGG pathways and enricher for Reactome and Hallmark pathways.

Immune subtype and glycosylation: Outlier analysis (Figure S4B) was performed as described earlier. In the glycosylation outlier analysis, both positive (overexpression) and negative (low expression) outliers within the in-group (non-wild type immune subtypes) were examined. Protein-protein interaction analysis was performed via uploading genes that contributed to the observed downregulation of glycoprotein expression to the STRING database (Figure S4E).

Analysis of *STK11* mutant samples: To assess *STK11* mutation and neutrophil degranulation (ND) association previously described[10], both an ND score and an independent component analysis (ICA) score was calculated. For the ND score, the

GSVA[178] score was calculated using the WP_Neutrophil_Degranulation gene set from MSigDB[162]. The ICA score was calculated by using weights previously found to calculate a weighted sum[10]. Additionally, the CIN70 score[179] was calculated to assess chromosomal instability across samples using RNAseq data. Enrichment of the ferroptosis resistance gene set[180] was calculated using differential expression analysis results of proteomic data. Differential expression analysis was performed by limma[153] and further submitted to GSEA[160].

To compare *STK11* mutant tumors across C1/C3 and C2 proteomic clusters, differential expression analysis was performed across these two groups using limma. A covariate of NRF2 pathway activity estimated via GSVA, similar to the ND score described above, was further added when assessing *STK11* MUT effect between C1/C3 and C2 (Figure S4L). The NRF2 pathway gene set was fetched from the previous CPTAC LSCC study[131].

Network analysis based on phosphorylation: Phosphosites were clustered into co-regulated phosphosite modules based on PhosphoDisco[181]. Briefly, phosphorylation abundance was normalized for parent protein abundance using proteomic abundance. Then the resulting normalized phosphorylation dataset was clustered using the HDBSCAN method via the Python programming language hdbscan package. The two hyperparameters, minimum cluster size and minimum core sample size, were both set to a value of 3. Once clustering finished, there were a total of 62 unique modules containing two or more phosphosites deemed statistically to co-regulate. A module score was then calculated for each patient defined by the median abundance of the protein-normalized phosphosites contained within that module. Module scores that were significantly different between *STK11* WT and mutant were determined by an independent 2-sample T-test (Python programming language, scipy package, stats module, ttest_ind method). To determine significant signatures from PTMsigDB[161] present in either *STK11* WT or mutant samples, a Fisher's enrichment test was calculated.

Carcinogen mutation signature analysis: To detect mutational signatures from endogeneous and environmental carcinogens and their contributions in each tumor sample, we first categorized the samples based on smoking status and the presence of oncogenic *EGFR* and *KRAS* mutations. Then, non-negative matrix factorization (NMF) algorithms were utilized to decompose the matrix of sample frequencies of somatic substitutions into 96 mutation types[156,182]. These types consist of six base substitution categories and sequence immediately 5' and 3' to the mutated bases. To identify the mutational signatures associated with endogenous carcinogens and environmental exposures, the identified mutational signatures were matched to an in-house signature database that integrated 30 experimentally-derived human cancer signatures from the COSMIC database[183] and 54 carcinogen signatures derived from human-induced pluripotent stem cell (iPSC) models exposed to environmental carcinogens[90]. The potential sources to contribute to mutational signatures were identified by the highest cosine similarity. After decomposing the somatic mutation matrix of 96 mutation types into carcinogen signatures, the carcinogen contribution in each sample was estimated. To categorize the impact of carcinogens more precisely, samples were first clustered based on their carcinogen contributions into 10

subclusters using hierarchical clustering with Euclidean distance and the Ward linkage method. Subsequently, considering the carcinogen contribution clustering results, sample size, carcinogen categories, and their sources, we further classified samples into four major carcinogenic clusters (CS), including two endogenous CS with best matches to the COSMIC database (methylcytosine deamination CS, APOBEC CS) and two environmental CS with high similarity to the iPSC models (nitrosylated polyaromatic hydrocarbon/polyaromatic hydrocarbon (nitro-PAH/PAH) CS, nitrosamine CS). To investigate the impact of carcinogens on clinical outcomes, the Kaplan-Meier method was used to estimate the relapse-free survival probability and overall survival probability for different carcinogen groups. The Cox proportional hazard model was utilized to assess the effect of carcinogen groups.

Multiomics data integration using S-score: For integrative analysis, S-scores were computed for each gene, protein, and phosphorylation site using the methodology described by[184]. The S-score is a statistical measure that integrates multiple omics data types, providing a unified significance score at the gene level.

Normalization and weighting: The first step involved normalization and weighting of features within each dataset. For each feature (gene, protein, phosphosite) denoted by i in dataset j ($j = 1, 2, 3$), the log fold change ($logFC$) value was transformed into a z-score (z_i) using the following formula:

$$z_i(logFC_{ij} - \mu_j)/\sigma_j$$

where:

- μ_j is the mean $logFC$ value in dataset j
- σ_j is the standard deviation of $logFC$ values in dataset j

Z-scores center and standardize the data within each dataset, allowing for comparisons across datasets with different measurement units.

Next, a weight (w_j) was assigned to each feature based on the inverse square root of the number of observations (n_j) in its dataset:

$$w_j = 1/\sqrt{n_j}$$

This weighting scheme ensured a fairer comparison across features within datasets of varying sizes. The weighted z-score (z_{wij}) for each feature in each dataset was then calculated by multiplying z_i by its corresponding weight (w_j):

$$z_{wij} = z_i * w_j$$

Combining Z-scores: For each gene, a combined weighted z-score (z_{cwi}) was calculated by summing z_{wij} across all 3 datasets (transcriptome, proteome, phosphoproteome):

$$z_{cwi} = \sum_{j=1}^{3} z_{wij}$$

A combined weight w_c was then computed as the square root of the sum of squared w_j from all 3 datasets:

$$w_c = \sqrt{\sum_{j=1}^{3} w_j^2}$$

S-score Calculation: The S-score for each feature (S_i) was derived by normalizing z_{cwi} by w_c:

$$S_i = z_{cwi}/w_c$$

This normalization step ensured the S-score remains on a standard scale even when the number of datasets being integrated changes.

Significance Testing: p-values were computed using the standard normal cumulative distribution function, adjusted for two-tailed tests. This assessed the statistical significance of the S-score. Finally, the Benjamini-Hochberg procedure was used to adjust p-values for multiple testing, controlling the false discovery rate (FDR).

Network and Subnetwork Identification: The Prize-Collecting Steiner Forest (PCSF) algorithm was utilized for network analysis, as outlined by[185]. This algorithm is particularly suited for identifying relevant subnetworks within larger interaction networks, considering both the connectivity between important nodes (terminals) and the cost of including additional (Steiner) nodes.

In this implementation, the computed S_i were leveraged as node prizes ($S_i = p(v)$) within the PCSF framework. The network was represented as a graph $G = (V, E)$, where V is the set of nodes (genes) and E is the set of edges (interactions between nodes). Each edge $e \in E$ incurs a cost $c(e)$ based on the corresponding interaction's experimental confidence derived from the STRING database. Interactions with a combined score of ≥ 0.7 were used to ensure high-confidence interactions. The PCSF algorithm aims to find a subgraph $G' = (V', E')$ that maximizes the sum of the prizes collected from the included nodes ($\sum_{v \in V'} p(v)$) while minimizing the total edge cost ($\sum_{e \in E'} c(e)$). Here, λ is a tuning parameter that balances the trade-off between node importance (reflected by S-scores) and edge costs (representing interaction confidence). This optimization problem can be formulated as:

$$\max(\Sigma_{v \in V'} p(v) - \lambda \Sigma_{e \in E'} c(e)).$$

The resulting subnetwork, enriched with high-scoring nodes (terminals) and high-confidence interactions, is likely to capture the most significant biological pathways and interactions, providing valuable insights into the underlying biological mechanisms.

Activating Pathway Analysis by IPA: Differentially expressed proteins and phosphosites were significantly selected by the comparison between different subgroups, including clinical staging (stage I, III-IV, or all), smoking status (smoker vs never-smoker), sex, proteome subgroups (C1, C2, C3, or non-C2), or mutation-annotated carcinogen signatures (nitro-PAHs, PAHs, or nitrosamines) by Kruskal Wallis test, p-value < 0.05, and Wilcoxon rank sum test, B.H. p-value < 0.05, as well as a cut-off value of absolute $\log_2$ abundance >1.3 in each group. To investigate functional activity of cancer-related signaling pathways, we utilized Ingenuity Pathway Analysis (IPA, Qiagen) to examine the significantly enriched canonical signaling pathways. The Z-score was used to infer the activation state of implicated biological functions and signaling pathways, determined by the “up” or “down” regulation of the observed proteins and phosphoproteins and a literature-derived direction and effect (“activating” or “inhibiting”) of the proteins to the pathways[186,187]. Statistical significance of the pathways was considered as absolute |z-score| ≥ 1 and B.H. p-value < 0.05 (calculated by a Fisher’s exact test). An interaction network map and functional characterization of differentially expressed proteins or phosphoproteins were constructed by the network generation algorithm of IPA and STRING database.

LUAD Cell Lines from DepMap NMF clustering: An XGBoost method was applied to global proteome data filtered to retain proteins designated as tumor epithelium-specific in deconvolution analysis to train a protein-based classifier. The Minimum Redundancy Maximum Relevance (MRMR) algorithm was used for feature selection and evaluated by the Monte-Carlo Cross Validation (MCCV) process 20 times. Each time, the dataset was randomly split 70% for classifier training (T) and 30% for validation (V). During classifier training, hyperparameters were tuned with 10-fold cross-validation within the training set, T. The trained classifiers were then evaluated with the validation set, V (different for each of the 20 repeats). The number of markers giving the highest mean MCCV AUROC without meaningful numerical increases by adding more markers, which in our case was 100 features, was selected to repeat the feature selection and classifier building processes using all the data to produce a final fully trained, fixed model. This model was then applied to TMT proteomic data for LUAD cell lines from the DepMap resource (protein_quant_current_normalized.csv). Predictions were evaluated visually by UMAP projection of tumors and cell lines using Hallmark protein ssGSEA scores.

Predicting Effective Drug Targets: Details of the method for prioritizing over-expressed cancer dependencies as effective drug targets have been previously reported[117]. Briefly, drug target tiers are as follows: Tier 1 are the primary inhibited targets of drugs approved for any cancer type by any regulatory agency, tier 2 includes primary inhibited targets of drugs approved for any other indications, tier 3 includes targets inhibited by investigational or experimental agents, tier 4 are targets of undeveloped drugs but include protein families frequently targeted by small molecules, and tier 5 includes surface membrane proteins. Membrane proteins were also included as follows:

The COMPARTMENTS database generates subcellular localization predictions for human genes[188]. A filter was applied for genes with “Plasma membrane” and “Cellsurface” localization and a confidence score greater than or equal to 3. The Human Protein Atlas (HPA) maintains a list of membrane proteins where a filter was applied for “Evidence at protein level” and removal of genes with “Low tissue specificity.” From the DepMap Public 22Q2 files (https://depmap.org/portal), cancer cell line annotations (sample_info.csv) and mutation data (CCLE_mutations.csv) were downloaded. LUAD cell lines were identified from the cell line annotation file by applying the following filters: primary_disease = “Lung Cancer”, lineage = “lung”, lineage_sub_subtype = “NSCLC_adenocarcinoma”. Fusion annotations were also retrieved from DepMap along with shRNA knockdown gene effect scores from combined screens published by Broad’s Achilles, Novartis’ Project DRIVE, and the Marcotte et al. breast cell line datasets (DEMETER2 Data v6, D2_combined_gene_dep_scores.csv) and drug response data from Broad’s Profiling Relative Simultaneously in Mixtures (PRISM) (PRISM Repurposing 19Q4). A one-sample, one-tailed t-test was used to identify RNA/protein/activating phosphosite/acetylsite/ubiquitylsite/glycosite candidates associated with significantly reduced cell growth following gene knockdown. For those significantly increased omics candidates in tumors vs corresponding NATs (adjusted p-value ≤ 0.05, Wilcoxon rank sum test followed by Benjamini-Hochberg adjustment) for specified LUAD subsets, a p-value of gene effect score below zero lower or equal to 0.05 were defined as therapeutic target candidates.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical parameter reporting: All relevant statistical parameters for the analyses performed in this study such as sample size (n), center and spread (e.g. mean/median, standard deviation/error), statistical tests (e.g. two-sample t-test) and significance cutoffs (e.g. adjusted p-value < 0.05) are reported in the main text, figure legends, STAR Methods, and/or supplementary information. Where appropriate, methods used to determine the validity of certain statistical assumptions are discussed within this section.

ADDITIONAL RESOURCES

The CPTAC program website, which includes details about program initiatives, investigators, and datasets, can be accessed at: https://proteomics.cancer.gov/programs/cptac. All PDC accessions associated with this study can be accessed at: https://pdc.cancer.gov/pdc/cohort/cptac-icpc-luad.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Shankha Satpathy[1,24,25,26,27], Natalie M. Clark[1,25], Yi-Ju Chen[2,25], Noshad Hosseini[3,25], Ya-Hsuan Chang[4,25], Yi Hsiao[3,25], Jonathan T. Lei[5,25], Francesca Petralia[6,25], Jin-Shing Chen[7,25], Yifat Geffen[1,21], David I. Heiman[1], Indranil Paul[1], Hanbyul Cho[3], Michelle Hollenberg[8], Giacomo B. Marino[6], Kuen-Tyng Lin[2], Rahul

Mannan[3], C. Jackson White[1], Joe Allen[1], Shayan C. Avanessian[1], M. Harry Kane[1], Ashley Wolfe[3], Miloni Kinarivala[3], Wenke Liu[8], Shankara Anand[1], Mong-Wei Lin[7], Moe Haines[1], Erik J. Bergstrom[1], Grant Hussey[8], Ginny Xiaohe Li[3], Deepak C. Mani[1], Hao Fang[2], Eric J. Jaehnig[5], Hasmik Keshishian[1], Brecca Miller[8], Kang-Yi Su[9], Yi-Jing Hsiao[2], Hsao-Hsun Hsu[7], Min-Shu Hsieh[7], Kuo-Hsuan Hsu[10], Alexi Monovoukas[1], Simone Gohsman[1], John R. Thorup[1], Yamei Deng[3], Yo Akiyama[1], Eden Deng[6], Eric Sheng-Wen Chen[2], Azra Krek[6], Rodrigo Espinoza[2], Weiping Ma[6], Daniel Charytonowicz[6], Robert Sebra[6], Jyun-Hong Lin[2], Yan-Si Chen[2], Yin-Chen Hsu[9], Ze-Shiang Lin[2], Kun-Chieh Chen[11], Chang-Wei Yeh[12], Yu-Tai Wang[12], Alexander J. Lazar[13], Mehdi Mesri[14], Eunkyung An[14], Xu Zhang[14], Karl R. Clauser[1], David Fenyö[8], Arul M. Chinnaiyan[3], Bing Zhang[5], Li Ding[15], Kelly Ruggles[8], Chelsea Newton[16], Hui Zhang[17], Pei Wang[6], Galen Hostetter[16], Gilbert S. Omenn[3], Chandan Kumar-Sinha[3], Mathangi Thiagarajan[18], Ramaswamy Govindan[15], Paul Paik[19,22], Abhijit Parolia[3], Qing K. Li[20], Avi Ma'ayan[6], Gad A. Getz[1,21], Saravana M. Dhanasekaran[3], Ana I. Robles[14], Gee-Chen Chang[11,26], Pan-Chyr Yang[9,26], Sung-Liang Yu[9,26,27], Hsuan-Yu Chen[2,23,26,27], Alexey I. Nesvizhskii[3,26], Steven A. Carr[1,26], D. R. Mani[1,26,27], Marcin P. Cieslik[3,26,27], Yu-Ju Chen[2,9,26,27], Michael A. Gillette[1,21,26,27,28], Taiwan Cancer Moonshot Program, Clinical Proteomic Tumor Analysis Consortium

Affiliations

[1]Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA

[2]Academia Sinica, Taipei, TW

[3]University of Michigan, Ann Arbor, MI 48109, USA

[4]National Health Research Institutes, Miaoli County, TW

[5]Baylor College of Medicine, Houston, TX, 77030, USA

[6]Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA

[7]National Taiwan University Hospital and National Taiwan University College of Medicine, Taipei, TW

[8]NYU Grossman School of Medicine, New York, NY 10016, USA

[9]National Taiwan University, Taipei, TW

[10]Taichung Veterans General Hospital, Taichung, TW

[11]Chung Shan Medical University Hospital, Taichung, TW

[12]National Center for High-performance Computing, National Applied Research Laboratories, Hsinchu, TW

[13]The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

[14]Office of Cancer Clinical Proteomics Research, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Rockville, MD 20850, USA

[15]Washington University in St. Louis, St. Louis, MO 63110, USA

[16]Van Andel Research Institute, Grand Rapids, MI 49503, USA

[17]Johns Hopkins University, Baltimore, MD 21231, USA

[18]Leidos Biomedical Research Inc., Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA

[19]Memorial Sloan Kettering Cancer Center, New York, NY

[20]Johns Hopkins Medical Institutions, Baltimore, MD 21231, USA

[21]Massachusetts General Hospital, Boston, MA 02114, USA

[22]Weill Cornell Medical College, New York, NY, USA

[23]National Chung Hsing University, Taichung, TW

[24]Current address: AstraZeneca, Boston, MA, 02421, USA

[25]These authors contributed equally

[26]Senior authors

[27]Corresponding authors

[28]Lead contact

Acknowledgements

This work was supported by the National Cancer Institute (NCI) Clinical Proteomic Tumor Analysis Consortium (CPTAC) grants U24CA210986, U24CA210979, U24CA210972, U24CA270823, U24CA271012, U24CA271079, U24CA271037, U24CA271075, U24CA271076, U24CA271114, U01CA271402, U01CA271407, U01CA271410 and U01CA271412; the NCI Cancer Center Support Grant (P30CA196521), awarded to the Tisch Cancer Institute of the Icahn School of Medicine at Mount Sinai (MSSM); and the Next-Generation Pathway of Taiwan Cancer Precision Medicine Program (AS-KPQ-107-TCPMP) and Key and Novel Therapeutics Development Program for Major Diseases (AS-KPQ-111-KNT) at Academia Sinica in Taiwan. The ICPC team thanks the MS Facilities at Academia Sinica (AS-NBRPCF-111-201, AS-CFII-111-209) and National Center for High-performance Computing of Taiwan for providing computational resources and storage resources. We also acknowledge funding from Dr. Miriam and Sheldon G. Adelson Medical Research Foundation to S.A.C.

Declaration of Interests

Grants or contracts (institutional research) for P.P. from EMD Serono, Bicara, Novartis and Advisory Board/ Consulting for Novartis, Mirati, Janssen, EMD Serono, Takeda, Bicara, AstraZeneca. S.S. is currently employed by AstraZeneca. This work was conducted while S.S. was employed at the Broad Institute and AstraZeneca has no role in this study. The rest of the authors have no competing interests.

Consortia

Abdul Samad Hashimi, Abhijit Parolia, Adrij Mohan, Akhilesh Pandey, Alex Webster, Alexander J. Lazar, Alexander Pilozzi, Alexey I. Nesvizhskii, Alexi Monovoukas, Amanda G. Paulovich, Ana I. Robles, Aniket Dagar, Andrew K. Godwin, Arul M. Chinnaiyan, Ashley Wolfe, Avi Ma'ayan, Azra Krek, Barbara Pruetz, Bart O.. Williams, Bing Zhang, Brecca Miller, Brian J. Druker, Chandan Kumar-Sinha, Chang-Wei Yeh, Chao-Wen Lu, Chelsea Newton, Cheng-Hsiang Chu, Chi-Ya Shen, Chia-Li Han, Chien-Chia Lin, Chien-Yu Lin, Ching-Wen Chen, Chung-Hsien Lin, D. R. Mani, Danail Petrov, Daniel C. Rohrer, Daniel Charytonowicz, Daniel W. Chan, David Chesla, David Fenyö, David Heiman, Deepak C. Mani, Diwakar Davaar, Eden Deng, Eden Deng, Elizabeth Duffy, Eric J. Jaehnig,

Erik J. Bergstrom, Eunkyung An, Francesca Petralia, Gad A. Getz, Galen Hostetter, Gee-Chen Chang, George D Wilson, Giacomo B. Marino, Gilbert S. Omenn, Ginny Xiaohe Li, Grace Zhao, Grant Hussey, Hannah Hanbyul Cho, Hao Fang, Hasmik Keshishian, Henry Rodriguez, Hsao-Hsun Hsu, Hsiang-En Hsu, Hsing-Jui Tsai, Hsuan-Yu Chen, Hui Zhang, Iga Kołodziejczak, Indranil Paul, Jackson White, Jan Lubinski, Jasmin Bavarva, Jasmine Huang, Jason Hafron, Jeffrey W. Tyner, Jia-Jun Wu, Jin-Shing Chen, Jing-Wei Lin, John Koomen, John Thorup, Jonathan T. Lei, Joseph Allen, Juani Waniwan, Jyun-Hong Lin, Kakhaber Zaalishvili, Kang-Yi Su, Karen A. Ketchum, Karl R. Clauser, Kelly Ruggles, Ki-Hok Liao, Kuen-Tyng Lin, Kun-Chieh Chen, Kuo-Hsuan Hsu, Li Ding, M. Harry Kane, Maciej Wiznerowicz, Marcin J. Domagalski, Marcin P. Cieslik, Mathangi Thiagarajan, Meenakshi Anurag, Mehdi Mesri, Melissa Borucki, Michael A. Gillette, Michelle Hollenberg, Miloni Kinarivala, Min-Shu Hsieh, Moe Haines, Mong-Wei Lin, Natalie M. Clark, Nathan J. Edwards, Negin Vatanian, Noshad Hosseini, Pamela Grady, Pan-Chyr Yang, Paul D. Piehowski, Paul Paik, Pei Wang, Pei-Hsing Chen, Pei-Rong Huang, Popovici Bogdan, Qin Li, Qing K. Li, Rafael Fonseca, Rahul Mannan, Ramaswamy Govindan, Rashna Madan, Ratna R. Thangudu, Reese Crispen, Robert Sebra, Rodrigo Espinoza, Ronald Matteotti, Ross Bremner, Sandra Cerda, Sandra L. Cottingham, Saravana M. Dhanasekaran, Shankara Anand, Shankha Satpathy, Shayan Avanessian, Shirley Tsang, Shuang Cai, Simone Gohsman, Sin-Ming Huang, Steven A. Carr, Sung-Liang Yu, Tai-Ching Lin, Tao Liu, Tara Hiltke, Thomas Bauer, Wan-Chun Lai, Wei-Tzu Chiu, Weiping Ma, Wenke Liu, William W. Maggio, Xiaojun Jing, Xu Zhang, Xu-Heng Chiang, Ya-Hsuan Chang, Ya-Ling Chang, Yamei Deng, Yan-Ming Chen, Yan-Si Chen, Yi Hsiao, Yi-Jing Hsiao, Yi-Ju Chen, Yi-Ling Chen, Yi-Wei Lin, Yi-Wen Wang, Yifat Geffen, Yin-Chen Hsu, Yo Akiyama, Yuping Zhang, Yu-Cheng Chang, Yu-Ju Chen, Yu-Tai Wang, Yu-Ting Huang, Yuju Lien, Yvonne Shutack, Ze-Shiang Lin, Zhe-Rong Zheng, Zoran Andric

Secondary authors

Clinical Proteomic Tumor Analysis Consortium: Abdul Samad Hashimi, Adrij Mohan, Akhilesh Pandey, Alexander Pilozzi, Alex Webster, Amanda G. Paulovich, Aniket Dagar, Andrew K. Godwin, Barbara Pruetz, Bart O. Williams, Brian J. Druker, Daniel C. Rohrer, Daniel W. Chan, Danail Petrov, David Chesla, Diwakar Davaar, Elizabeth Duffy, George D Wilson, Grace Zhao, Iga Kołodziejczak, Jan Lubinski, Jasmine Huang, Jason Hafron, Jeffrey W. Tyner, John Koomen, Kakhaber Zaalishvili, Karen A. Ketchum, Maciej Wiznerowicz, Marcin J. Domagalski, Meenakshi Anurag, Melissa Borucki, Nathan J. Edwards, Negin Vatanian, Pamela Grady, Paul D. Piehowski, Popovici Bogdan, Qin Li, Rafael Fonseca, Rashna Madan, Ratna R. Thangudu, Reese Crispen, Ronald Matteotti, Ross Bremner, Sandra Cerda, Sandra L. Cottingham, Shirley Tsang, Shuang Cai, Tao Liu, Thomas Bauer, William W. Maggio, Xiaojun Jing, Yuping Zhang, Yvonne Shutack, Zoran Andric.

Taiwan Cancer Moonshot Program: Chao-Wen Lu, Cheng-Hsiang Chu, Chi-Ya Shen, Chia-Li Han, Chien-Chia Lin, Chien-Yu Lin, Ching-Wen Chen, Chung-Hsien Lin, Hsiang-En Hsu, Hsing-Jui Tsai, Jia-Jun Wu, Jing-Wei Lin, Juani Waniwan, Ki-Hok Liao, Pei-Hsing Chen, Pei-Rong Huang, Sin-Ming Huang, Tai-Ching Lin, Wan-Chun Lai, Wei-Tzu Chiu, Xu-Heng Chiang, Ya-Ling Chang, Yan-Ming Chen, Yi-Ling Chen, Yi-Wen Wang, Yi-Wei Lin, Yu-Cheng Chang, Yu-Ting Huang, Yuju Lien, Zhe-Rong Zheng.

References

1. Leiter A, Veluswamy RR, and Wisnivesky JP (2023). The global burden of lung cancer: current status and future trends. Nat. Rev. Clin. Oncol 20, 624–639. [PubMed: 37479810]
2. LoPiccolo J, Gusev A, Christiani DC, and Jänne PA (2024). Lung cancer in patients who have never smoked - an emerging disease. Nat. Rev. Clin. Oncol 21, 121–146. [PubMed: 38195910]
3. Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature 511, 543–550. [PubMed: 25079552]
4. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, 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. [PubMed: 22980975]
5. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. Nature 455, 1069–1075. [PubMed: 18948947]
6. Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, et al. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 304, 1497–1500. [PubMed: 15118125]
7. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, et al. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. U. S. A 98, 13790–13795. [PubMed: 11707567]
8. Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, Lin WM, Province MA, Kraja A, Johnson LA, et al. (2007). Characterizing the cancer genome in lung adenocarcinoma. Nature 450, 893–898. [PubMed: 17982442]
9. Roh W, Geffen Y, Cha H, Miller M, Anand S, Kim J, Heiman DI, Gainor JF, Laird PW, Cherniack AD, et al. (2022). High-Resolution Profiling of Lung Adenocarcinoma Identifies Expression Subtypes with Specific Biomarkers and Clinically Relevant Vulnerabilities. Cancer Res. 82, 3917–3931. [PubMed: 36040373]
10. Gillette MA, Satpathy S, Cao S, Dhanasekaran SM, Vasaikar SV, 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. [PubMed: 32649874]
11. Chen Y-J, Roumeliotis TI, 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. [PubMed: 32649875]
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. [PubMed: 32649877]
13. Soltis AR, Bateman NW, Liu J, Nguyen T, Franks TJ, Zhang X, Dalgard CL, Viollet C, Somiari S, Yan C, et al. (2022). Proteogenomic analysis of lung adenocarcinoma reveals tumor heterogeneity, survival determinants, and therapeutically relevant pathways. Cell Rep Med 3, 100819. [PubMed: 36384096]
14. Lehtiö J, Arslan T, Siavelis I, Pan Y, Socciarelli F, Berkovska O, Umer HM, Mermelekas G, Pirmoradian M, Jönsson M, et al. (2021). Proteogenomics of non-small cell lung cancer reveals molecular subtypes associated with specific therapeutic targets and immune evasion mechanisms. Nat Cancer 2, 1224–1242. [PubMed: 34870237]
15. Park S-J, Ju S, Goh S-H, Yoon B-H, Park J-L, Kim J-H, Lee S, Lee S-J, Kwon Y, Lee W, et al. (2024). Proteogenomic Characterization Reveals Estrogen Signaling as a Target for Never-Smoker Lung Adenocarcinoma Patients without EGFR or ALK Alterations. Cancer Res. 84, 1491–1503. [PubMed: 38607364]
16. Binnewies M, Roberts EW, Kersten K, Chan V, Fearon DF, Merad M, Coussens LM, Gabrilovich DI, Ostrand-Rosenberg S, Hedrick CC, et al. (2018). Understanding the tumor immune microenvironment (TIME) for effective therapy. Nat. Med 24, 541–550. [PubMed: 29686425]

17. Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, et al. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N. Engl. J. Med 350, 2129–2139. [PubMed: 15118073]

18. Gainor JF, Shaw AT, Sequist LV, Fu X, Azzoli CG, Piotrowska Z, Huynh TG, Zhao L, Fulton L, Schultz KR, et al. (2016). EGFR Mutations and ALK Rearrangements Are Associated with Low Response Rates to PD-1 Pathway Blockade in Non-Small Cell Lung Cancer: A Retrospective Analysis. Clin. Cancer Res 22, 4585–4593. [PubMed: 27225694]

19. Barta JA, Powell CA, and Wisnivesky JP (2019). Global Epidemiology of Lung Cancer. Ann Glob Health 85. 10.5334/aogh.2419.

20. Jemal A, Schafer EJ, Sung H, Bandi P, Kratzer T, Islami F, and Siegel RL (2023). The Burden of Lung Cancer in Women Compared With Men in the US. JAMA Oncol 9, 1727–1728. [PubMed: 37824139]

21. Siegel RL, Miller KD, Wagle NS, and Jemal A (2023). Cancer statistics, 2023. CA Cancer J. Clin 73, 17–48. [PubMed: 36633525]

22. Fidler-Benaoudia MM, Torre LA, Bray F, Ferlay J, and Jemal A (2020). Lung cancer incidence in young women vs. young men: A systematic analysis in 40 countries. Int. J. Cancer 147, 811–819. [PubMed: 32020598]

23. Wang Z, Mo M, Zhou C, Feng X, Shen J, Ye T, Zhang Y, Hu H, Chen H, and Zheng Y (2021). Time-varying effect of sex on prognosis of lung adenocarcinoma surgical patients in China. Thorac Cancer 12, 1699–1707. [PubMed: 33939309]

24. Nakamura H, Ando K, Shinmyo T, Morita K, Mochizuki A, Kurimoto N, and Tatsunami S (2011). Female gender is an independent prognostic factor in non-small-cell lung cancer: a meta-analysis. Ann. Thorac. Cardiovasc. Surg 17, 469–480. [PubMed: 21881356]

25. Woolston A, Sintupisut N, Lu T-P, Lai L-C, Tsai M-H, Chuang EY, and Yeang C-H (2015). Putative effectors for prognosis in lung adenocarcinoma are ethnic and gender specific. Oncotarget 6, 19483–19499. [PubMed: 26160836]

26. Schabath MB, Cress D, and Munoz-Antonia T (2016). Racial and Ethnic Differences in the Epidemiology and Genomics of Lung Cancer. Cancer Control 23, 338–346. [PubMed: 27842323]

27. Haiman CA, Stram DO, Wilkens LR, Pike MC, Kolonel LN, Henderson BE, and Le Marchand L (2006). Ethnic and racial differences in the smoking-related risk of lung cancer. N. Engl. J. Med 354, 333–342. [PubMed: 16436765]

28. Dogan S, Shen R, Ang DC, Johnson ML, D'Angelo SP, Paik PK, Brzostowski EB, Riely GJ, Kris MG, Zakowski MF, et al. (2012). Molecular epidemiology of EGFR and KRAS mutations in 3,026 lung adenocarcinomas: higher susceptibility of women to smoking-related KRAS-mutant cancers. Clin. Cancer Res 18, 6169–6177. [PubMed: 23014527]

29. Carrot-Zhang J, Soca-Chafre G, Patterson N, Thorner AR, Nag A, Watson J, Genovese G, Rodriguez J, Gelbard MK, Corrales-Rodriguez L, et al. (2021). Genetic ancestry contributes to somatic mutations in lung cancers from admixed Latin American populations. Cancer Discov. 11, 591–598. [PubMed: 33268447]

30. Woodard GA, Jones KD, and Jablons DM (2016). Lung Cancer Staging and Prognosis. Cancer Treat. Res 170, 47–75. [PubMed: 27535389]

31. Bao Y, Zhang S, Zhang X, Pan Y, Yan Y, Wang N, Ren Y, Zuo J, Zong W-X, Wang Z, et al. (2023). RBM10 Loss Promotes EGFR-Driven Lung Cancer and Confers Sensitivity to Spliceosome Inhibition. Cancer Res. 83, 1490–1502. [PubMed: 36853175]

32. Li Y, Porta-Pardo E, Tokheim C, Bailey MH, Yaron TM, Stathias V, Geffen Y, Imbach KJ, Cao S, Anand S, et al. (2023). Pan-cancer proteogenomics connects oncogenic drivers to functional states. Cell 186, 3921–3944.e25. [PubMed: 37582357]

33. Stine ZE, Schug ZT, Salvino JM, and Dang CV (2022). Targeting cancer metabolism in the era of precision oncology. Nat. Rev. Drug Discov 21, 141–162. [PubMed: 34862480]

34. Chen J, Yang H, Teo ASM, Amer LB, Sherbaf FG, Tan CQ, Alvarez JJS, Lu B, Lim JQ, Takano A, et al. (2020). Genomic landscape of lung adenocarcinoma in East Asians. Nat. Genet 52, 177–186. [PubMed: 32015526]

35. Nanjo S, Wu W, Karachaliou N, Blakely CM, Suzuki J, Chou Y-T, Ali SM, Kerr DL, Olivas VR, Shue J, et al. (2022). Deficiency of the splicing factor RBM10 limits EGFR inhibitor response in EGFR-mutant lung cancer. J. Clin. Invest 132. 10.1172/JCI145099.
36. Ducray SP, Natarajan K, Garland GD, Turner SD, and Egger G (2019). The Transcriptional Roles of ALK Fusion Proteins in Tumorigenesis. Cancers (Basel) 11. 10.3390/cancers11081074.
37. Olsen JV, and Mann M (2013). Status of large-scale analysis of post-translational modifications by mass spectrometry. Mol Cell Proteomics 12, 3444–3452. [PubMed: 24187339]
38. Dutta H, and Jain N (2023). Post-translational modifications and their implications in cancer. Front Oncol 13, 1240115. [PubMed: 37795435]
39. Geffen Y, Anand S, Akiyama Y, Yaron TM, Song Y, Johnson JL, Govindan A, Babur Ö, Li Y, Huntsman E, et al. (2023). Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation. Cell 186, 3945–3967.e26. [PubMed: 37582358]
40. Vidyasagar A, Wilson NA, and Djamali A (2012). Heat shock protein 27 (HSP27): biomarker of disease and therapeutic target. Fibrogenesis Tissue Repair 5, 7. [PubMed: 22564335]
41. Weinert BT, Narita T, Satpathy S, Srinivasan B, Hansen BK, Schölz C, Hamilton WB, Zucconi BE, Wang WW, Liu WR, et al. (2018). Time-Resolved Analysis Reveals Rapid Dynamics and Broad Scope of the CBP/p300 Acetylome. Cell 174, 231–244.e12. [PubMed: 29804834]
42. Narita T, Higashijima Y, Kilic S, Liebner T, Walter J, and Choudhary C (2023). Acetylation of histone H2B marks active enhancers and predicts CBP/p300 target genes. Nat. Genet 55, 679–692. [PubMed: 37024579]
43. Sur I, and Taipale J (2016). The role of enhancers in cancer. Nat. Rev. Cancer 16, 483–493. [PubMed: 27364481]
44. Hanse EA, Ruan C, Kachman M, Wang D, Lowman XH, and Kelekar A (2017). Cytosolic malate dehydrogenase activity helps support glycolysis in actively proliferating cells and cancer. Oncogene 36, 3915–3924. [PubMed: 28263970]
45. Akimov V, Barrio-Hernandez I, Hansen SVF, Hallenborg P, Pedersen A-K, Bekker-Jensen DB, Puglia M, Christensen SDK, Vanselow JT, Nielsen MM, et al. (2018). UbiSite approach for comprehensive mapping of lysine and N-terminal ubiquitination sites. Nat. Struct. Mol. Biol 25, 631–640. [PubMed: 29967540]
46. Fu J, Zhou S, Xu H, Liao L, Shen H, Du P, and Zheng X (2023). ATM-ESCO2-SMC3 axis promotes 53BP1 recruitment in response to DNA damage and safeguards genome integrity by stabilizing cohesin complex. Nucleic Acids Res. 51, 7376–7391. [PubMed: 37377435]
47. Tsimokha AS, Artamonova TO, Diakonov EE, Khodorkovskii MA, and Tomilin AN (2020). Post-Translational Modifications of Extracellular Proteasome. Molecules 25. 10.3390/molecules25153504.
48. Lu F, Zhou J, Chen Q, Zhu J, Zheng X, Fang N, and Qiao L (2022). PSMA5 contributes to progression of lung adenocarcinoma in association with the JAK/STAT pathway. Carcinogenesis 43, 624–634. [PubMed: 35605971]
49. Bakhoum SF, and Landau DA (2017). Chromosomal Instability as a Driver of Tumor Heterogeneity and Evolution. Cold Spring Harb. Perspect. Med 7. 10.1101/cshperspect.a029611.
50. Zhang T, Joubert P, Ansari-Pour N, Zhao W, Hoang PH, Lokanga R, Moye AL, Rosenbaum J, Gonzalez-Perez A, Martínez-Jiménez F, et al. (2021). Genomic and evolutionary classification of lung cancer in never smokers. Nat. Genet 53, 1348–1359. [PubMed: 34493867]
51. Burrell RA, McClelland SE, Endesfelder D, Groth P, Weller M-C, Shaikh N, Domingo E, Kanu N, Dewhurst SM, Gronroos E, et al. (2013). Replication stress links structural and numerical cancer chromosomal instability. Nature 494, 492–496. [PubMed: 23446422]
52. Ripley BD (1976). The second-order analysis of stationary point processes. Preprint, 10.1017/s0021900200094328 10.1017/s0021900200094328.
53. Dixon P (2006). Ripley's K Function. 3, 1796.
54. Bakhoum SF (2023). Targeting the undruggable. Science 380, 47.
55. Payton M, Belmontes B, Hanestad K, Moriguchi J, Chen K, McCarter JD, Chung G, Ninniri MS, Sun J, Manoukian R, et al. (2024). Small-molecule inhibition of kinesin KIF18A reveals a mitotic vulnerability enriched in chromosomally unstable cancers. Nat Cancer 5, 66–84. [PubMed: 38151625]

56. Wang JM, Hong R, Demicco EG, Tan J, Lazcano R, Moreira AL, Li Y, Calinawan A, Razavian N, Schraink T, et al. (2023). Deep learning integrates histopathology and proteogenomics at a pan-cancer level. Cell Rep. Med 4, 101173. [PubMed: 37582371]
57. Panebianco F, Kelly LM, Liu P, Zhong S, Dacic S, Wang X, Singhi AD, Dhir R, Chiosea SI, Kuan S-F, et al. (2017). THADA fusion is a mechanism of IGF2BP3 activation and IGF1R signaling in thyroid cancer. Proc. Natl. Acad. Sci. U. S. A 114, 2307–2312. [PubMed: 28193878]
58. Rodriguez E, Chen L, Ao M-H, Geddes S, Gabrielson E, Askin F, Zhang H, and Li QK (2014). Expression of transcript factors SALL4 and OCT4 in a subset of non-small cell lung carcinomas (NSCLC). Transl. Respir. Med 2, 10. [PubMed: 25346886]
59. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature 578, 82–93. [PubMed: 32025007]
60. Tijhuis AE, Johnson SC, and McClelland SE (2019). The emerging links between chromosomal instability (CIN), metastasis, inflammation and tumour immunity. Mol. Cytogenet 12, 17. [PubMed: 31114634]
61. Rittmeyer A, Barlesi F, Waterkamp D, Park K, Ciardiello F, von Pawel J, Gadgeel SM, Hida T, Kowalski DM, Dols MC, et al. (2017). Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. Lancet 389, 255–265. [PubMed: 27979383]
62. Petralia F, Krek A, Calinawan AP, Charytonowicz D, Sebra R, Feng S, Gosline S, Pugliese P, Paulovich AG, Kennedy JJ, et al. (2023). BayesDeBulk: A Flexible Bayesian Algorithm for the Deconvolution of Bulk Tumor Data. bioRxiv, 2021.06.25.449763.
63. Poomakkoth N, Issa A, Abdulrahman N, Abdelaziz SG, and Mraiche F (2016). p90 ribosomal S6 kinase: a potential therapeutic target in lung cancer. J. Transl. Med 14, 14. [PubMed: 26791782]
64. Fu J, Li Y, Li C, Tong Y, Li M, and Cang S (2021). A special prognostic indicator: tumor mutation burden combined with immune infiltrates in lung adenocarcinoma with TP53 mutation. Transl. Cancer Res 10, 3963–3978. [PubMed: 35116695]
65. Li H, Yang L, Wang Y, Wang L, Chen G, Zhang L, and Wang D (2023). Integrative analysis of TP53 mutations in lung adenocarcinoma for immunotherapies and prognosis. BMC Bioinformatics 24, 155. [PubMed: 37072703]
66. To KKW, Fong W, and Cho WCS (2021). Immunotherapy in Treating EGFR-Mutant Lung Cancer: Current Challenges and New Strategies. Front Oncol 11, 635007. [PubMed: 34113560]
67. Kuleshov MV, Xie Z, London ABK, Yang J, Evangelista JE, Lachmann A, Shu I, Torre D, and Ma'ayan A (2021). KEA3: improved kinase enrichment analysis via data integration. Nucleic Acids Res. 49, W304–W316. [PubMed: 34019655]
68. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, and Ma'ayan A (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 14, 128. [PubMed: 23586463]
69. Qiao M, Zhou F, Liu X, Jiang T, Wang H, Li X, Zhao C, Cheng L, Chen X, Ren S, et al. (2024). Targeting focal adhesion kinase boosts immune response in KRAS/LKB1 co-mutated lung adenocarcinoma via remodeling the tumor microenvironment. Exp. Hematol. Oncol 13, 11. [PubMed: 38291516]
70. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, et al. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell 171, 1437–1452.e17. [PubMed: 29195078]
71. Martin-Orozco E, Sanchez-Fernandez A, Ortiz-Parra I, and Ayala-San Nicolas M (2019). WNT Signaling in Tumors: The Way to Evade Drugs and Immunity. Front. Immunol 10, 2854. [PubMed: 31921125]
72. Ramadori G, Ioris RM, Villanyi Z, Firnkes R, Panasenko OO, Allen G, Konstantinidou G, Aras E, Brenachot X, Biscotti T, et al. (2020). FKBP10 Regulates Protein Translation to Sustain Lung Cancer Growth. Cell Rep. 30, 3851–3863.e6. [PubMed: 32187554]
73. Du H, Chen Y, Hou X, Huang Y, Wei X, Yu X, Feng S, Wu Y, Zhan M, Shi X, et al. (2017). PLOD2 regulated by transcription factor FOXA1 promotes metastasis in NSCLC. Cell Death Dis. 8, e3143. [PubMed: 29072684]

74. DiChiara AS, Taylor RJ, Wong MY, Doan N-D, Rosario AMD, and Shoulders MD (2016). Mapping and Exploring the Collagen-I Proteostasis Network. ACS Chem. Biol 11, 1408–1421. [PubMed: 26848503]
75. Zhang Y, Fu F, Zhang Q, Li L, Liu H, Deng C, Xue Q, Zhao Y, Sun W, Han H, et al. (2024). Evolutionary proteogenomic landscape from pre-invasive to invasive lung adenocarcinoma. Cell Rep Med 5, 101358. [PubMed: 38183982]
76. Gardelli C, Russo L, Cipolla L, Moro M, Andriani F, Rondinone O, Nicotra F, Sozzi G, Bertolini G, and Roz L (2021). Differential glycosylation of collagen modulates lung cancer stem cell subsets through β1 integrin-mediated interactions. Cancer Sci. 112, 217–230. [PubMed: 33068069]
77. Peng DH, Rodriguez BL, Diao L, Chen L, Wang J, Byers LA, Wei Y, Chapman HA, Yamauchi M, Behrens C, et al. (2020). Collagen promotes anti-PD-1/PD-L1 resistance in cancer through LAIR1-dependent CD8+ T cell exhaustion. Nat. Commun 11, 4520. [PubMed: 32908154]
78. Fang S, Dai Y, Mei Y, Yang M, Hu L, Yang H, Guan X, and Li J (2019). Clinical significance and biological role of cancer-derived Type I collagen in lung and esophageal cancers. Thorac Cancer 10, 277–288. [PubMed: 30604926]
79. Xiao J, Zou Y, Chen X, Gao Y, Xie M, Lu X, Li W, He B, He S, You S, et al. (2016). The Prognostic Value of Decreased LKB1 in Solid Tumors: A Meta-Analysis. PLoS One 11, e0152674. [PubMed: 27035914]
80. Ji H, Ramsey MR, Hayes DN, Fan C, McNamara K, Kozlowski P, Torrice C, Wu MC, Shimamura T, Perera SA, et al. (2007). LKB1 modulates lung cancer differentiation and metastasis. Nature 448, 807–810. [PubMed: 17676035]
81. Li A, Wang Y, Yu Z, Tan Z, He L, Fu S, Shi M, Du W, Luo L, Li Z, et al. (2023). STK11/LKB1-Deficient Phenotype Rather Than Mutation Diminishes Immunotherapy Efficacy and Represents STING/Type I Interferon/CD8+ T-Cell Dysfunction in NSCLC. J. Thorac. Oncol 18, 1714–1730. [PubMed: 37495171]
82. Petralia F, Ma W, Yaron TM, Caruso FP, Tignor N, Wang JM, Charytonowicz D, Johnson JL, Huntsman EM, Marino GB, et al. (2024). Pan-cancer proteogenomics characterization of tumor immunity. Cell 187, 1255–1277.e27. [PubMed: 38359819]
83. Lee HM, Muhammad N, Lieu EL, Cai F, Mu J, Ha Y-S, Cao G, Suchors C, Joves K, Chronis C, et al. (2024). Concurrent loss of LKB1 and KEAP1 enhances SHMT-mediated antioxidant defence in KRAS-mutant lung cancer. Nat Metab. 10.1038/s42255-024-01066-z.
84. Wohlhieter CA, Richards AL, Uddin F, Hulton CH, Quintanal-Villalonga À, Martin A, de Stanchina E, Bhanot U, Asher M, Shah NS, et al. (2020). Concurrent Mutations in STK11 and KEAP1 Promote Ferroptosis Protection and SCD1 Dependence in Lung Cancer. Cell Rep 33, 108444. [PubMed: 33264619]
85. Wang Y, Liu Y, Liu J, Kang R, and Tang D (2020). NEDD4L-mediated LTF protein degradation limits ferroptosis. Biochem. Biophys. Res. Commun 531, 581–587. [PubMed: 32811647]
86. Oh E-T, and Park HJ (2015). Implications of NQO1 in cancer therapy. BMB Rep. 48, 609–617. [PubMed: 26424559]
87. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, et al. (2013). Signatures of mutational processes in human cancer. Nature 500, 415–421. [PubMed: 23945592]
88. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, and Stratton MR (2013). Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 3, 246–259. [PubMed: 23318258]
89. Mayakonda A, Lin D-C, Assenov Y, Plass C, and Koeffler HP (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 28, 1747–1756. [PubMed: 30341162]
90. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, Gomez C, Degasperi A, Harris R, Jackson SP, et al. (2019). A Compendium of Mutational Signatures of Environmental Agents. Cell 177, 821–836.e16. [PubMed: 30982602]
91. Zhou G (2019). Tobacco, air pollution, environmental carcinogenesis, and thoughts on conquering strategies of lung cancer. Cancer Biol Med 16, 700–713. [PubMed: 31908889]

92. Park J-E, Seo J-E, Lee J-Y, and Kwon H (2015). Distribution of Seven N-Nitrosamines in Food. Toxicol Res 31, 279–288. [PubMed: 26483887]

93. Wang Q, Zhu W, Xiao G, Ding M, Chang J, and Liao H (2020). Effect of AGER on the biological behavior of non-small cell lung cancer H1299 cells. Mol. Med. Rep 22, 810–818. [PubMed: 32468030]

94. Snodin DJ, Trejo-Martin A, Ponting DJ, Smith GF, Czich A, Cross K, Custer L, Elloway J, Greene N, Kalgutkar AS, et al. (2024). Mechanisms of Nitrosamine Mutagenicity and Their Relationship to Rodent Carcinogenic Potency. Chem. Res. Toxicol 37, 181–198. [PubMed: 38316048]

95. de Souza JES, Fonseca AF, Valieris R, Carraro DM, Wang JYJ, Kolodner RD, and de Souza SJ (2014). S-score: a scoring system for the identification and prioritization of predicted cancer genes. PLoS One 9, e94147. [PubMed: 24710071]

96. Zhang H, Lu L, Zhao C, Liu Q, Zhou Q, Zhang Y, Pu Y, Wang S, Liu R, and Yin L (2022). Lipid metabolism disorders contribute to hepatotoxicity of ICR mice induced by nitrosamines exposure. Environ. Int 167, 107423. [PubMed: 35908391]

97. Zhang Y, Hu Q, Fu J, Li X, Mao H, and Wang T (2023). Influence of Exposure Pathways on Tissue Distribution and Health Impact of Polycyclic Aromatic Hydrocarbon Derivatives. Environ. Health 1, 150–167.

98. Church TR, Haznadar M, Geisser MS, Anderson KE, Caporaso NE, Le C, Abdullah SB, Hecht SS, Oken MM, and Van Ness B (2010). Interaction of CYP1B1, cigarette-smoke carcinogen metabolism, and lung cancer risk. Int. J. Mol. Epidemiol. Genet 1, 295–309. [PubMed: 21532841]

99. Øvrevik J, Refsnes M, Holme JA, Schwarze PE, and Låg M (2013). Mechanisms of chemokine responses by polycyclic aromatic hydrocarbons in bronchial epithelial cells: sensitization through toll-like receptor-3 priming. Toxicol. Lett 219, 125–132. [PubMed: 23458896]

100. Hill W, Lim EL, Weeden CE, Lee C, Augustine M, Chen K, Kuan F-C, Marongiu F, Evans EJ Jr, Moore DA, et al. (2023). Lung adenocarcinoma promotion by air pollutants. Nature 616, 159–167. [PubMed: 37020004]

101. El Rayes T, Catena R, Lee S, Stawowczyk M, Joshi N, Fischbach C, Powell CA, Dannenberg AJ, Altorki NK, Gao D, et al. (2015). Lung inflammation promotes metastasis through neutrophil protease-mediated degradation of Tsp-1. Proc. Natl. Acad. Sci. U. S. A 112, 16000–16005. [PubMed: 26668367]

102. Zhang S, Sun L, Zuo J, and Feng D (2024). Tumor associated neutrophils governs tumor progression through an IL-10/STAT3/PD-L1 feedback signaling loop in lung cancer. Transl. Oncol 40, 101866. [PubMed: 38128466]

103. Liao X, Wang W, Yu B, and Tan S (2022). Thrombospondin-2 acts as a bridge between tumor extracellular matrix and immune infiltration in pancreatic and stomach adenocarcinomas: an integrative pan-cancer analysis. Cancer Cell Int. 22, 213. [PubMed: 35701829]

104. Nallanthighal S, Heiserman JP, and Cheon D-J (2021). Collagen Type XI Alpha 1 (COL11A1): A Novel Biomarker and a Key Player in Cancer. Cancers 13. 10.3390/cancers13050935.

105. Relli V, Trerotola M, Guerra E, and Alberti S (2019). Abandoning the Notion of Non-Small Cell Lung Cancer. Trends Mol. Med 25, 585–594. [PubMed: 31155338]

106. Yao S, Huang H-Y, Han X, Ye Y, Qin Z, Zhao G, Li F, Hu G, Hu L, and Ji H (2019). Keratin 14-high subpopulation mediates lung cancer metastasis potentially through Gkn1 upregulation. Oncogene 38, 6354–6369. [PubMed: 31320708]

107. Langfelder P, and Horvath S (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559. [PubMed: 19114008]

108. Newell F, Pires da Silva I, Johansson PA, Menzies AM, Wilmott JS, Addala V, Carlino MS, Rizos H, Nones K, Edwards JJ, et al. (2022). Multiomic profiling of checkpoint inhibitor-treated melanoma: Identifying predictors of response and resistance, and markers of biological discordance. Cancer Cell 40, 88–102.e7. [PubMed: 34951955]

109. Hwang S, Kwon A-Y, Jeong J-Y, Kim S, Kang H, Park J, Kim J-H, Han OJ, Lim SM, and An HJ (2020). Immune gene signatures for predicting durable clinical benefit of anti-PD-1 immunotherapy in patients with non-small cell lung cancer. Sci. Rep 10, 643. [PubMed: 31959763]

110. Wang Y, Yan K, Guo Y, Lu Y, Su H, and Li H (2022). IP-score correlated to endogenous tumour antigen peptide processing: A candidate clinical response score algorithm of immune checkpoint inhibitors therapy in multiple cohorts. Front. Immunol 13, 1085491. [PubMed: 36700205]
111. Kirk CJ, Muchamuel T, Wang J, and Fan RA (2021). Discovery and Early Clinical Development of Selective Immunoproteasome Inhibitors. Cells 11. 10.3390/cells11010009.
112. Jenkins TW, Downey-Kopyscinski SL, Fields JL, Rahme GJ, Colley WC, Israel MA, Maksimenko AV, Fiering SN, and Kisselev AF (2021). Activity of immunoproteasome inhibitor ONX-0914 in acute lymphoblastic leukemia expressing MLL-AF4 fusion protein. Sci. Rep 11, 10883. [PubMed: 34035431]
113. Xi J, Zhuang R, Kong L, He R, Zhu H, and Zhang J (2019). Immunoproteasome-selective inhibitors: An overview of recent developments as potential drugs for hematologic malignancies and autoimmune diseases. Eur. J. Med. Chem 182, 111646. [PubMed: 31521028]
114. Wehenkel M, Ban J-O, Ho Y-K, Carmony KC, Hong JT, and Kim KB (2012). A selective inhibitor of the immunoproteasome subunit LMP2 induces apoptosis in PC-3 cells and suppresses tumour growth in nude mice. Br. J. Cancer 107, 53–62. [PubMed: 22677907]
115. Kim YJ, Lee Y, Shin H, Hwang S, Park J, and Song EJ (2023). Ubiquitin-proteasome system as a target for anticancer treatment-an update. Arch. Pharm. Res 46, 573–597. [PubMed: 37541992]
116. Mohanty A, Nam A, Srivastava S, Jones J, Lomenick B, Singhal SS, Guo L, Cho H, Li A, Behal A, et al. (2023). Acquired resistance to KRAS G12C small-molecule inhibitors via genetic/nongenetic mechanisms in lung cancer. Sci Adv 9, eade3816. [PubMed: 37831779]
117. Savage SR, Yi X, Lei JT, Wen B, Zhao H, Liao Y, Jaehnig EJ, Somes LK, Shafer PW, Lee TD, et al. (2024). Pan-cancer proteogenomics expands the landscape of therapeutic targets. Cell 0. 10.1016/j.cell.2024.05.039.
118. Li Y, Dou Y, Da Veiga Leprevost F, Geffen Y, Calinawan AP, Aguet F, Akiyama Y, Anand S, Birger C, Cao S, et al. (2023). Proteogenomic data and resources for pan-cancer analysis. Cancer Cell 41, 1397–1406. [PubMed: 37582339]
119. Le X, Paz-Ares LG, Van Meerbeeck J, Viteri S, Galvez CC, Smit EF, Garassino M, Veillon R, Baz DV, Pradera JF, et al. (2023). Tepotinib in patients with non-small cell lung cancer with high-level MET amplification detected by liquid biopsy: VISION Cohort B. Cell Rep Med 4, 101280. [PubMed: 37944528]
120. Sequist LV, Han J-Y, Ahn M-J, Cho BC, Yu H, Kim S-W, Yang JC-H, Lee JS, Su W-C, Kowalski D, et al. (2020). Osimertinib plus savolitinib in patients with EGFR mutation-positive, MET-amplified, non-small-cell lung cancer after progression on EGFR tyrosine kinase inhibitors: interim results from a multicentre, open-label, phase 1b study. Lancet Oncol 21, 373–386. [PubMed: 32027846]
121. Ma P, Huang R, Gu Y, Fang Y, Wu X, Chen D-S, Zhang H-W, Gao W, and Shu Y (2022). Savolitinib monotherapy exerted significant benefit in a non-small cell lung cancer patient with osimertinib resistance harboring primary EGFR L858R mutation and MET amplification: a case report. Anticancer Drugs 33, 1186–1190. [PubMed: 35946569]
122. Wang AY, and Liu H (2019). The past, present, and future of CRM1/XPO1 inhibitors. Stem Cell Investig 6, 6.
123. Kim J, McMillan E, Kim HS, Venkateswaran N, Makkar G, Rodriguez-Canales J, Villalobos P, Neggers JE, Mendiratta S, Wei S, et al. (2016). XPO1-dependent nuclear export is a druggable vulnerability in KRAS-mutant lung cancer. Nature 538, 114–117. [PubMed: 27680702]
124. Quintanal-Villalonga A, Durani V, Sabet A, Redin E, Kawasaki K, Shafer M, Karthaus WR, Zaidi S, Zhan YA, Manoj P, et al. (2023). Exportin 1 inhibition prevents neuroendocrine transformation through SOX2 down-regulation in lung and prostate cancers. Sci. Transl. Med 15, eadf7006. [PubMed: 37531417]
125. Akimov V, Fehling-Kaschek M, Barrio-Hernandez I, Puglia M, Bunkenborg J, Nielsen MM, Timmer J, Dengjel J, and Blagoev B (2021). Magnitude of Ubiquitination Determines the Fate of Epidermal Growth Factor Receptor Upon Ligand Stimulation. J. Mol. Biol 433, 167240. [PubMed: 34508725]

126. Rohde CM, Schrum J, and Lee AW-M (2004). A juxtamembrane tyrosine in the colony stimulating factor-1 receptor regulates ligand-induced Src association, receptor kinase function, and down-regulation. J. Biol. Chem 279, 43448–43461. [PubMed: 15297464]

127. Corsello SM, Nagari RT, Spangler RD, Rossen J, Kocak M, Bryan JG, Humeidi R, Peck D, Wu X, Tang AA, et al. (2020). Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. Nat. Cancer 1, 235–248. [PubMed: 32613204]

128. Li Z, Yu B, Qi F, and Li F (2021). KIF11 Serves as an Independent Prognostic Factor and Therapeutic Target for Patients With Lung Adenocarcinoma. Front Oncol 11, 670218. [PubMed: 33968780]

129. Oughtred R, Rust J, Chang C, Breitkreutz B-J, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, et al. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 30, 187–200. [PubMed: 33070389]

130. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, and Tamayo P (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 1, 417–425. [PubMed: 26771021]

131. Satpathy S, Krug K, Jean Beltran PM, Savage SR, Petralia F, Kumar-Sinha C, Dou Y, Reva B, Kane MH, Avanessian SC, et al. (2021). A proteogenomic portrait of lung squamous cell carcinoma. Cell 184, 4348–4371.e40. [PubMed: 34358469]

132. Rivera KD, Olive ME, Bergstrom EJ, Nelson AJ, Lee KA, Satpathy S, Carr SA, and Udeshi ND (2021). Automating UbiFast for High-throughput and Multiplexed Ubiquitin Enrichment. Mol. Cell. Proteomics 20, 100154. [PubMed: 34592423]

133. Keshishian H, McDonald ER 3rd, Mundt F, Melanson R, Krug K, Porter DA, Wallace L, Forestier D, Rabasha B, Marlow SE, et al. (2021). A highly multiplexed quantitative phosphosite assay for biology and preclinical studies. Mol. Syst. Biol 17, e10156. [PubMed: 34569154]

134. Watson EV, Lee JJ-K, Gulhan DC, Melloni GEM, Venev SV, Magesh RY, Frederick A, Chiba K, Wooten EC, Naxerova K, et al. (2024). Chromosome evolution screens recapitulate tissue-specific tumor aneuploidy patterns. Nat. Genet 56, 900–912. [PubMed: 38388848]

135. Mannan R, Wang X, Bawa PS, Chugh S, Chinnaiyan AK, Rangaswamy R, Zhang Y, Cao X, Smith SC, Trpkov K, et al. (2023). Characterization of protein S-(2-succino)-cysteine (2SC) succination as a biomarker for fumarate hydratase-deficient renal cell carcinoma. Hum. Pathol 134, 102–113. [PubMed: 36581128]

136. Mannan R, Wang X, Bawa PS, Zhang Y, Skala SL, Chinnaiyan AK, Dagar A, Wang L, Zelenka-Wang SB, McMurry LM, et al. (2023). Characterization of Intercalated Cell Markers KIT and LINC01187 in Chromophobe Renal Cell Carcinoma and Other Renal Neoplasms. Int. J. Surg. Pathol 31, 1027–1040. [PubMed: 36250542]

137. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318–1330. [PubMed: 32913098]

138. Graubert A, Aguet F, Ravi A, Ardlie KG, and Getz G (2021). RNA-SeQC 2: efficient RNA-seq quality control and quantification for large cohorts. Bioinformatics 37, 3048–3050. [PubMed: 33677499]

139. Li B, Ruotti V, Stewart RM, Thomson JA, and Dewey CN (2010). RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics 26, 493–500. [PubMed: 20022975]

140. Mermel CH, Schumacher SE, Hill B, Meyerson ML, 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. [PubMed: 21527027]

141. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, and Getz G (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505, 495–501. [PubMed: 24390350]

142. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, and Forbes SA (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer 18, 696–705. [PubMed: 30293088]

143. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, and Nesvizhskii AI (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat. Methods 14, 513–520. [PubMed: 28394336]

144. Yu F, Teo GC, Kong AT, Fröhlich K, Li GX, Demichev V, and Nesvizhskii AI (2023). Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. Nat. Commun 14, 4154. [PubMed: 37438352]

145. da Veiga Leprevost F, Haynes SE, Avtonomov DM, Chang H-Y, Shanmugam AK, Mellacheruvu D, Kong AT, and Nesvizhskii AI (2020). Philosopher: a versatile toolkit for shotgun proteomics data analysis. Nat. Methods 17, 869–870. [PubMed: 32669682]

146. Shteynberg DD, Deutsch EW, Campbell DS, Hoopmann MR, Kusebauch U, Lee D, Mendoza L, Midha MK, Sun Z, Whetton AD, et al. (2019). PTMProphet: Fast and Accurate Mass Modification Localization for the Trans-Proteomic Pipeline. J. Proteome Res 18, 4262–4272. [PubMed: 31290668]

147. Huang K-L, Li S, Mertins P, Cao S, Gunawardena HP, Ruggles KV, Mani DR, Clauser KR, Tanioka M, Usary J, et al. (2017). Proteogenomic integration reveals therapeutic targets in breast cancer xenografts. Nat. Commun 8, 14864. [PubMed: 28348404]

148. Cao L, Huang C, Cui Zhou D, Hu Y, Lih TM, Savage SR, Krug K, Clark DJ, Schnaubelt M, Chen L, et al. (2021). Proteogenomic characterization of pancreatic ductal adenocarcinoma. Cell 184, 5031–5052.e26. [PubMed: 34534465]

149. Li GX, Chen L, Hsiao Y, Mannan R, Zhang Y, Luo J, Petralia F, Cho H, Hosseini N, Leprevost F. da V., et al. (2024). Comprehensive proteogenomic characterization of rare kidney tumors. Cell Rep Med 5, 101547. [PubMed: 38703764]

150. Polasky DA, Yu F, Teo GC, and Nesvizhskii AI (2020). Fast and comprehensive N- and O-glycoproteomics analysis with MSFragger-Glyco. Nat. Methods 17, 1125–1132. [PubMed: 33020657]

151. Polasky DA, Geiszler DJ, Yu F, and Nesvizhskii AI (2022). Multiattribute Glycan Identification and FDR Control for Glycoproteomics. Mol. Cell. Proteomics 21, 100205. [PubMed: 35091091]

152. Djomehri SI, Gonzalez ME, da Veiga Leprevost F, Tekula SR, Chang H-Y, White MJ, Cimino-Mathews A, Burman B, Basrur V, Argani P, et al. (2020). Quantitative proteomic landscape of metaplastic breast carcinoma pathological subtypes and their relationship to triple-negative tumors. Nat. Commun 11, 1723. [PubMed: 32265444]

153. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, and Smyth GK (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47. [PubMed: 25605792]

154. Wen B, Jaehnig EJ, and Zhang B (2022). OmicsEV: a tool for comprehensive quality evaluation of omics data tables. Bioinformatics 38, 5463–5465. [PubMed: 36271853]

155. Gaujoux R, and Seoighe C (2010). A flexible R package for nonnegative matrix factorization. BMC Bioinformatics 11, 367. [PubMed: 20598126]

156. Brunet J-P, Tamayo P, Golub TR, and Mesirov JP (2004). Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. U. S. A 101, 4164–4169. [PubMed: 15016911]

157. Kim H, and Park H (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics 23, 1495–1502. [PubMed: 17483501]

158. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, Schinzel AC, Sandy P, Meylan E, Scholl C, et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 462, 108–112. [PubMed: 19847166]

159. Mani DR, Maynard M, Kothadia R, Krug K, Christianson KE, Heiman D, Clauser KR, Birger C, Getz G, and Carr SA (2021). PANOPLY: a cloud-based platform for automated and reproducible proteogenomic data analysis. Nat. Methods 10.1038/s41592-021-01176-6.

160. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A 102, 15545–15550. [PubMed: 16199517]

161. Krug K, Mertins P, Zhang B, Hornbeck P, Raju R, Ahmad R, Szucs M, Mundt F, Forestier D, Jane-Valbuena J, et al. (2018). A curated resource for phosphosite-specific signature analysis. Mol. Cell. Proteomics 10.1074/mcp.TIR118.000943.

162. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, and Mesirov JP (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740. [PubMed: 21546393]

163. Castanza AS, Recla JM, Eby D, Thorvaldsdóttir H, Bult CJ, and Mesirov JP (2023). Extending support for mouse data in the Molecular Signatures Database (MSigDB). Nat. Methods 20, 1619–1620. [PubMed: 37704782]

164. Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, Lander ES, and Getz G (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc. Natl. Acad. Sci. U. S. A 112, E5486–E5495. [PubMed: 26392535]

165. Zhang Y, Narayanan SP, Mannan R, Raskind G, Wang X, Vats P, Su F, Hosseini N, Cao X, Kumar-Sinha C, et al. (2021). Single-cell analyses of renal cell cancers reveal insights into tumor microenvironment, cell of origin, and therapy response. Proc. Natl. Acad. Sci. U. S. A 118. 10.1073/pnas.2103240118.

166. Madissoon E, Oliver AJ, Kleshchevnikov V, Wilbrey-Clark A, Polanski K, Richoz N, Ribeiro Orsi A, Mamanova L, Bolt L, Elmentaite R, et al. (2023). A spatially resolved atlas of the human lung characterizes a gland-associated immune niche. Nat. Genet 55, 66–77. [PubMed: 36543915]

167. Aulchenko YS, Ripke S, Isaacs A, and van Duijn CM (2007). GenABEL: an R library for genome-wide association analysis. Bioinformatics 23, 1294–1296. [PubMed: 17384015]

168. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, and Feuston BP (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci 43, 1947–1958. [PubMed: 14632445]

169. Ripley BD (1976). The Second-Order Analysis of Stationary Point Processes. J. Appl. Probab 13, 255–266.

170. Hong R, Liu W, DeLair D, Razavian N, and Fenyö D (2021). Predicting endometrial cancer subtypes and molecular features from histopathology images using multi-resolution deep learning models. Cell Rep Med 2, 100400. [PubMed: 34622237]

171. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, Treviño V, Shen H, Laird PW, Levine DA, et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun 4, 2612. [PubMed: 24113773]

172. Petralia F, Wang L, Peng J, Yan A, Zhu J, and Wang P (2018). A new method for constructing tumor specific gene co-expression networks based on samples with tumor purity heterogeneity. Bioinformatics 34, i528–i536. [PubMed: 29949994]

173. Keenan AB, Torre D, Lachmann A, Leong AK, Wojciechowicz ML, Utti V, Jagodnik KM, Kropiwnicki E, Wang Z, and Ma'ayan A (2019). ChEA3: transcription factor enrichment analysis by orthogonal omics integration. Nucleic Acids Res. 47, W212–W224. [PubMed: 31114921]

174. Evangelista JE, Clarke DJB, Xie Z, Lachmann A, Jeon M, Chen K, Jagodnik KM, Jenkins SL, Kuleshov MV, Wojciechowicz ML, et al. (2022). SigCom LINCS: data and metadata search engine for a million gene expression signatures. Nucleic Acids Res. 50, W697–W709. [PubMed: 35524556]

175. Clark NR, Hu KS, Feldmann AS, Kou Y, Chen EY, Duan Q, and Ma'ayan A (2014). The characteristic direction: a geometrical approach to identify differentially expressed genes. BMC Bioinformatics 15, 79. [PubMed: 24650281]

176. Clarke DJB, Jeon M, Stein DJ, Moiseyev N, Kropiwnicki E, Dai C, Xie Z, Wojciechowicz ML, Litz S, Hom J, et al. (2021). Appyters: Turning Jupyter Notebooks into data-driven web apps. Patterns (N Y) 2, 100213. [PubMed: 33748796]

177. Blumenberg L, Kawaler EA, Cornwell M, Smith S, Ruggles KV, and Fenyö D (2021). BlackSheep: A Bioconductor and Bioconda Package for Differential Extreme Value Analysis. J. Proteome Res 20, 3767–3773. [PubMed: 34165986]

178. Hänzelmann S, Castelo R, and Guinney J (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics 14, 7. [PubMed: 23323831]

179. Carter SL, Eklund AC, Kohane IS, Harris LN, and Szallasi Z (2006). A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nat. Genet 38, 1043–1048. [PubMed: 16921376]

180. Dixon SJ, Patel DN, Welsch M, Skouta R, Lee ED, Hayano M, Thomas AG, Gleason CE, Tatonetti NP, Slusher BS, et al. (2014). Pharmacological inhibition of cystine-glutamate exchange induces endoplasmic reticulum stress and ferroptosis. Elife 3, e02523. [PubMed: 24844246]

181. Schraink T, Blumenberg L, Hussey G, George S, Miller B, Mathew N, González-Robles TJ, Sviderskiy V, Papagiannakopoulos T, Possemato R, et al. (2023). PhosphoDisco: A Toolkit for Co-regulated Phosphorylation Module Discovery in Phosphoproteomic Data. Mol. Cell. Proteomics 22, 100596. [PubMed: 37394063]

182. Lee DD, and Seung HS (1999). Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791. [PubMed: 10548103]

183. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res 47, D941–D947. [PubMed: 30371878]

184. Balbin OA, Prensner JR, Sahu A, Yocum A, Shankar S, Malik R, Fermin D, Dhanasekaran SM, Chandler B, Thomas D, et al. (2013). Reconstructing targetable pathways in lung cancer by integrating diverse omics data. Nat Commun 4, 2617. [PubMed: 24135919]

185. Akhmedov M, Kedaigle A, Chong RE, Montemanni R, Bertoni F, Fraenkel E, and Kwee I (2017). PCSF: An R-package for network-based interpretation of high-throughput data. PLoS Comput. Biol 13, e1005694. [PubMed: 28759592]

186. So J, Tai AK, Lichtenstein AH, Wu D, and Lamon-Fava S (2021). Sexual dimorphism of monocyte transcriptome in individuals with chronic low-grade inflammation. Biol. Sex Differ 12, 43. [PubMed: 34321081]

187. Iwata T, Sedukhina AS, Kubota M, Oonuma S, Maeda I, Yoshiike M, Usuba W, Minagawa K, Hames E, Meguro R, et al. (2021). A new bioinformatics approach identifies overexpression of GRB2 as a poor prognostic biomarker for prostate cancer. Sci. Rep 11, 5696. [PubMed: 33707553]

188. Binder JX, Pletscher-Frankild S, Tsafou K, Stolte C, O'Donoghue SI, Schneider R, and Jensen LJ (2014). COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database 2014, bau012. [PubMed: 24573882]

Highlights:

- Genomic fragmentation is a prognostic hallmark of LUAD.
- Carcinogens influence signaling in tumors and adjacent tissues.
- Proteomic clustering finds early-stage LUAD with advanced features.
- Integrating proteogenomics and screens nominates potential drugs for LUAD subtypes.

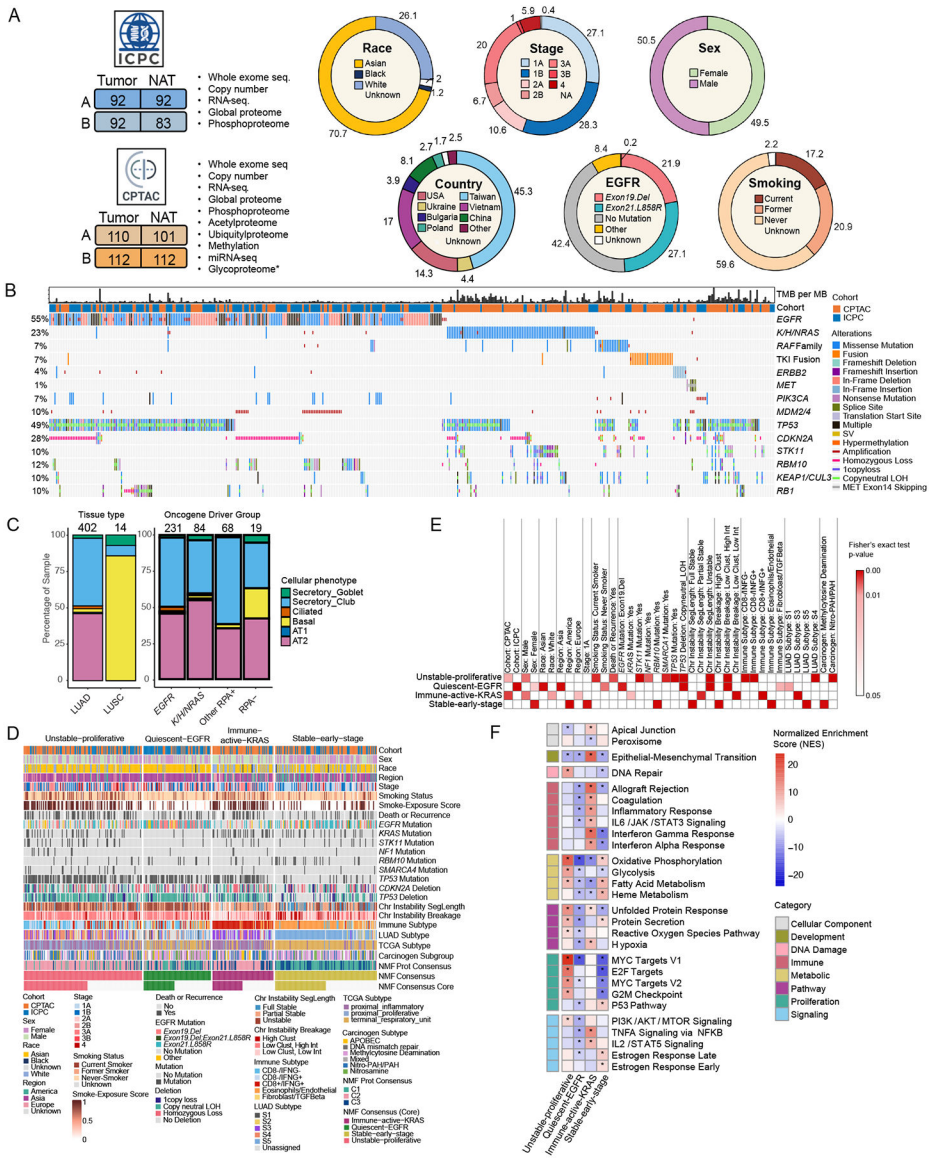


Figure 1. Proteogenomic landscape of the CPTAC and ICPC cohorts

A. Left: Summary of included samples and -omes. Sample numbers are stratified into tumor and normal adjacent tissue (NAT). The -omes available are listed to the right of each cohort (ICPC, top; CPTAC, bottom). * denotes that glycoproteome data is available only for the CPTAC.B cohort. Right: Distribution of selected annotations across 406 patients profiled in this study. All annotations are self-reported except for *EGFR* mutation, which was determined using whole exome mutation calls. “Unknown” denotes that the data were not reported, or in the case of *EGFR* mutation, that mutation calls were not available for those samples.

B. Oncoplot of mutations of interest across the cohort of 406 patients. TMB per MB: Tumor Mutation Burden per Megabase. Selected mutation calls are grouped together in the following categories: RAS Family (*HRAS*, *KRAS*, *NRAS*); RAF Family (*ARAF*, *BRAF*, *RAF1*); TKI Fusion (*ALK*, *RET*, *ROS1*).

C. Cell-of-origin analysis across the 402 tumor samples for which RNA-seq passed QC. The 402 samples are grouped based on mutation status. An additional 14 squamous samples collected and analyzed under the auspices of this study were included exclusively in this analysis.

D. Multi-omic NMF clustering across 383 tumor samples with available RNA-seq, copy number variation, proteomic, and phosphoproteomic data.

E. Clinical annotations significantly enriched in each multi-omic cluster. Color represents nominal p-value from Fisher's exact test on the NMF consensus core samples. Annotations shown have nominal p-value < 0.01 in at least one of the clusters.

F. Gene Set Enrichment Analysis (GSEA) on the four multi-omic clusters. Heatmap shows the Normalized Enrichment Score (NES), which is derived from the feature weights for each cluster. * indicates significant enrichment at adj. p<0.01.

See also Figure S1 and Tables S1 and S2.

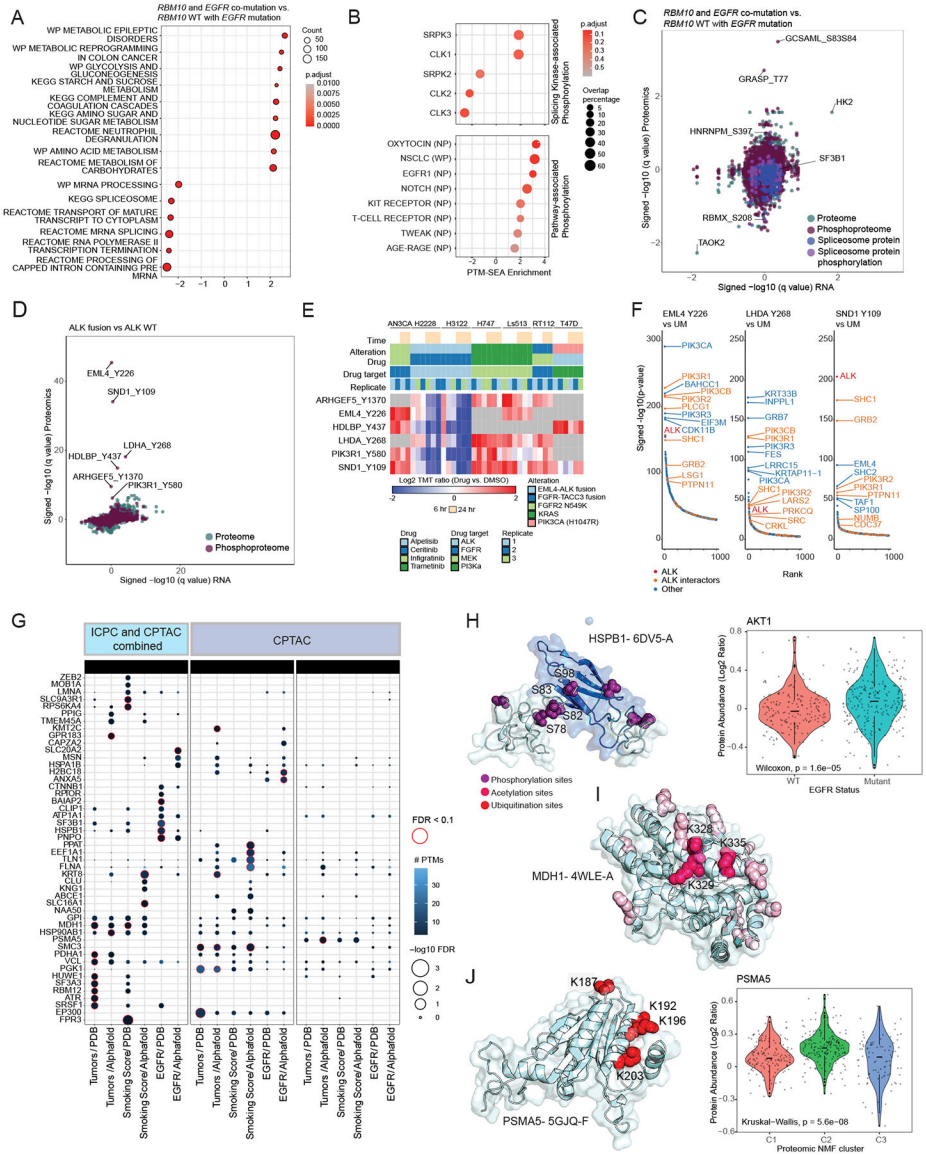


Figure 2. Impact of mutation and PTMs on downstream signaling and structure

A. GSEA result of protein-level differential expression analysis comparing *RBM10* truncated mutation tumors versus *RBM10* WT in the context of *EGFR* mutation.

B. Post-translational modification-Signature Enrichment Analysis (PTM-SEA) showed splicing kinase- and pathway-associated phosphorylation enrichment.

C. Differentially expressed features between *RBM10* mutant and WT tumors in the context of *EGFR* mutation at the levels of proteome and phosphoproteome.

D. Differential features between *ALK* fusion and WT tumors at the level of proteome and phosphoproteome.

E. Candidate pY sites and their abundance in a range of cell lines treated with driver-matched kinase inhibitors including *ALK*-fusion LUAD cell lines treated with the *ALK* inhibitor ceritinib. pY sites were downregulated upon kinase inhibition specifically in the *ALK* lines, and in a time dependent fashion (6 and 24 hours).

F. Rank plots showing protein abundance (indicated by signed −log10 p-value) in candidate pY vs. unmodified peptide pull-downs. The red dots show known ALK interactors in the BioGRID database[129]. Key interactors are indicated in the panel.

G. Bubble plot representing CLUMPS-PTM results. Phosphoproteome shows clustering results for both ICPC and CPTAC cohorts; acetylproteome and ubiquitylproteome results are shown for the CPTAC cohort. Red circle: Significant results: permutation test, adj. $p<0.1$.

H. HSPB1 phosphorylation cluster on 3D crystal structure (cyan, PDB ID: 6DV5-A, left). Violin plots showing relative protein abundances of AKT1 between *EGFR* mutant and wild type (right). In boxplots inset in violin plots, the centerline represents median; box bounds represent 25th and 75th percentiles; whiskers represent minimum and maximum; and filled points represent individual data points.

I. MDH1 acetylation cluster on 3D crystal structure (cyan, PDB ID: 4WLE-A). Highly differentially expressed acetylation sites are denoted by bright pink.

J. PSMA5 ubiquitylation cluster on 3D crystal structure (cyan, PDB ID: 5GJQ-F, left). Violin plots showing protein abundances of PSMA5 across the proteomic NMF clusters (right). In boxplots inset in violin plots, the centerline represents median; box bounds represent 25th and 75th percentiles; whiskers represent minimum and maximum; and filled points represent individual data points.

See also Figure S2 and Table S3.

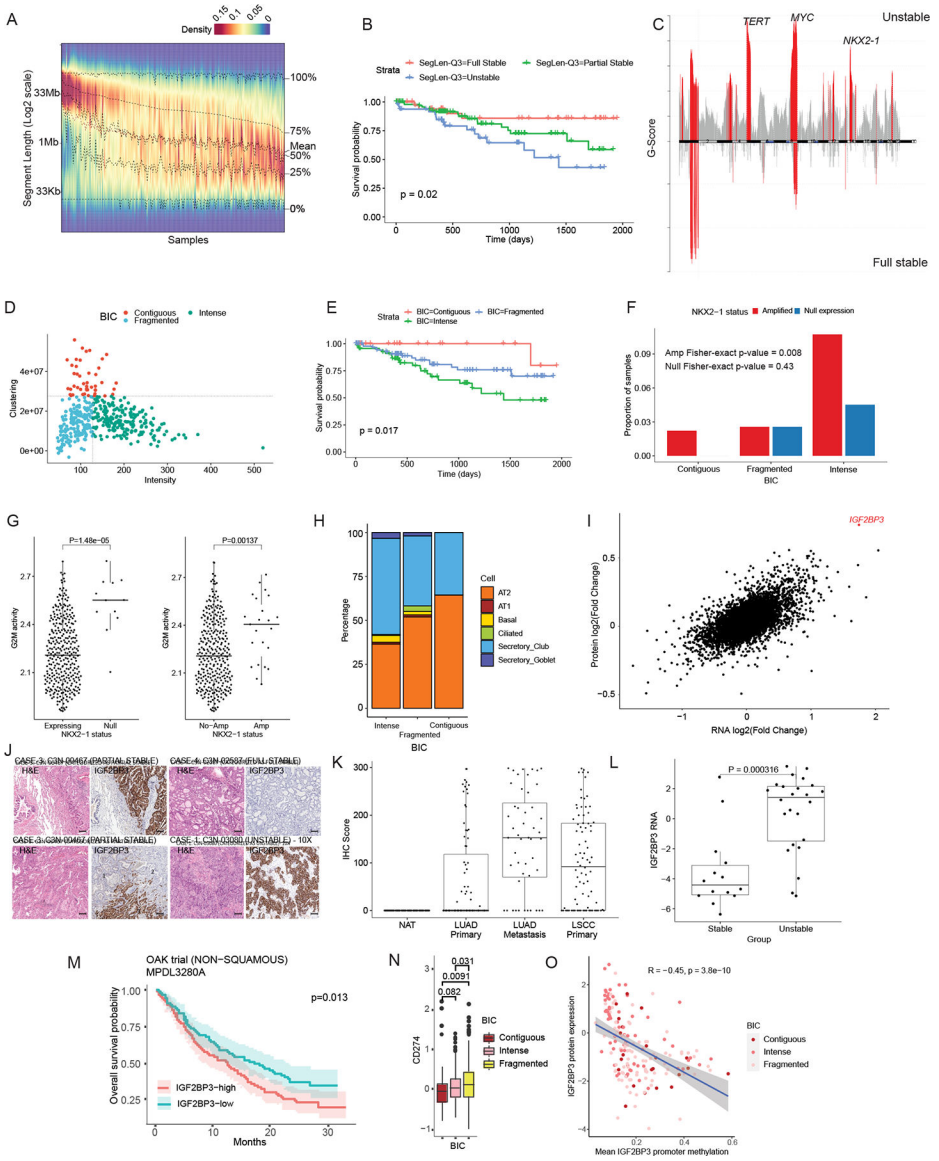


Figure 3. Genomic instability in LUAD and its impact

A. Distribution of CNV segment lengths for each sample. Each column on the x-axis represents one sample and the colors represent the density of the segment length for that sample. Samples are sorted based on 75th percentile of segment lengths. Samples on the left with a high density of large segments represent “Full-stable” and samples on the right with a high density of short segments represent “Unstable”.

B. Kaplan-Meier (KM) survival curves showing the overall survival of patients, stratified by SegLen-Q3. P-value obtained using log-rank test.

C. Comparison of recurrently amplified regions in unstable (top) and fully stable (bottom) tumors.

D. Classification of tumors according to their breakpoint intensity (x-axis) and clustering score (y-axis) to derive BIC groups. Tumors are categorized into three groups: tumors with highly clustered breakpoints (Contiguous), tumors with low clustering and intensity of

breakpoints (Fragmented), and tumors with low clustering and high intensity of breakpoints (Intense).

E. KM survival curves showing the overall survival of patients, stratified by BIC. Tumors with high clustering show exceptionally good survival independent of their breakpoint intensity. In groups with low clustering, high intensity of breakpoints results in worse outcomes. P-value obtained using log-rank test.

F. The proportion of tumors with *NKX2-1* amplification (red) and null expression (blue). Tumors with amplified *NKX2-1* show significant enrichment in the BIC-intense group.

G. Proliferative activity of tumors stratified by *NKX2-1* amplification (left) and expression (right) status. Proliferative activity is measured by ssGSEA analysis of RNA expression for the HALLMARK_G2M_CHECKPOINT[130] pathway. P values are based on the Wilcoxon Rank Sum test. In boxplots, the centerline represents median; box bounds represent 25th and 75th percentiles; whiskers represent minimum and maximum; and filled points represent individual data points.

H. Transcriptional state of tumors stratified by BIC groups. The BIC-intense group shows a reduced contribution of the AT2 phenotype.

I. Log2 fold changes of genes from differential expression analysis comparing BIC-intense to other tumors at both RNA (x-axis) and protein (y-axis) levels.

J. IGF2BP3 IHC and H&E staining shows heterogeneous protein expression of IGF2BP3 in case-3 (partial stable LUAD), uniformly strong expression in case-1 (unstable), and complete abrogation of IGF2BP3 expression in case-4 (full stable). (IHC; H&E). Scale bar = 50μm.

K. Independent validation in 3 patient TMAs reveals higher IGF2BP3 expression in metastatic LUAD compared to primary LUAD and primary LSCC. Also noteworthy is the absence of IGF2BP3 in benign lung parenchyma cores. In boxplots, the centerline represents median; box bounds represent 25th and 75th percentiles; whiskers represent minimum and maximum; and filled points represent individual data points.

L. Expression of *IGF2BP3* (RNA) in LUAD patients from ICGC/PCAWG, stratified by instability. In boxplots, the centerline represents median; box bounds represent 25th and 75th percentiles; whiskers represent minimum and maximum; and filled points represent individual data points.

M. KM survival curves showing the overall survival of non-squamous patients treated with the anti-PDL1 antibody MPDL3280A, stratified by *IGF2BP3* RNA expression in the OAK trial. P-value obtained using log-rank test.

N. PD-L1 (CD274) protein expression in the CPTAC cohort stratified by BIC status. In boxplots, the centerline represents median; box bounds represent 25th and 75th percentiles; whiskers represent minimum and maximum; and filled points represent individual data points.

O. Correlation of IGF2BP3 protein expression and average IGF2BP3 promoter methylation. See also Figure S3 and Table S3.

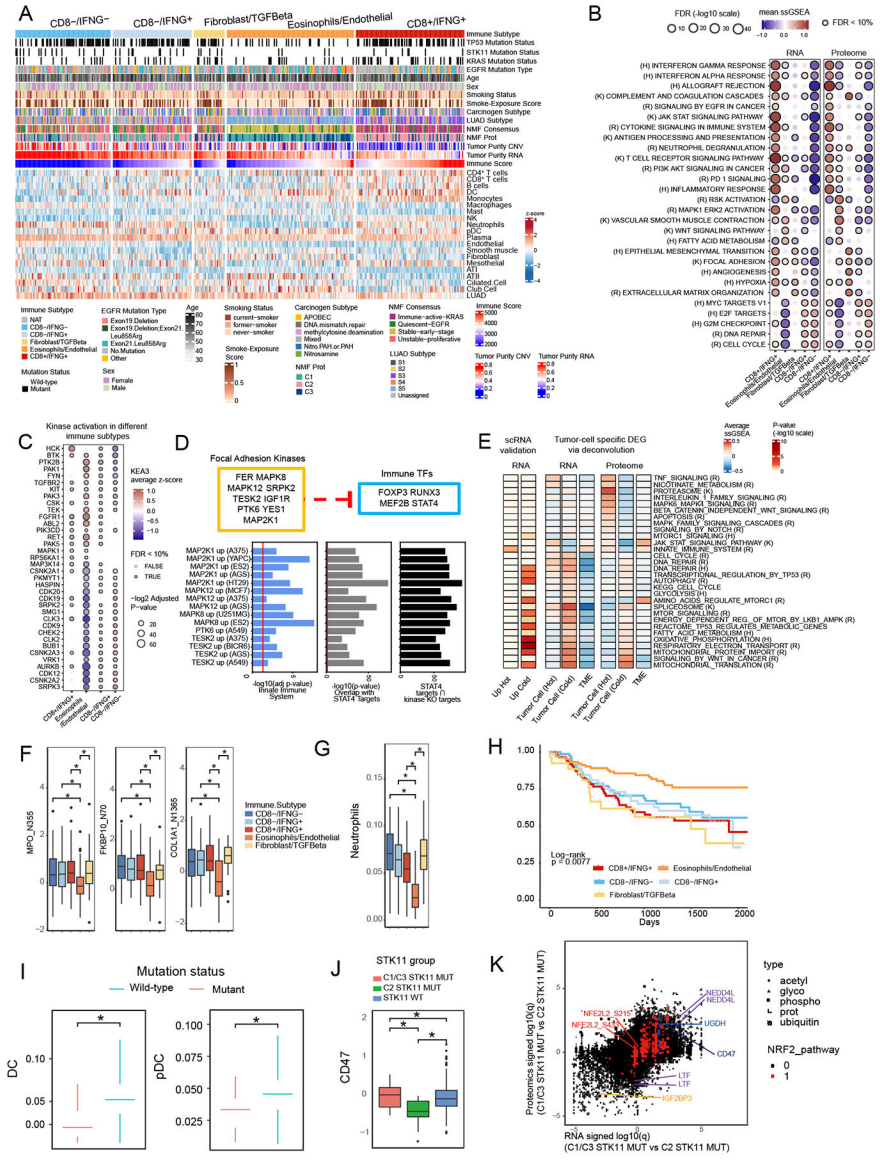


Figure 4. Immune landscape of LUAD

A. Cell type fractions estimated via multi-omics-based deconvolution analysis (top), protein (middle), and RNA expression (bottom) of cell type-specific markers. The annotation tracks show immune subtypes derived from proteogenomic data, immune score derived via ESTIMATE from RNA-seq data, tumor purity derived via TSNet from RNA-seq data, tumor purity estimated from CNV data, and various clinical and mutation parameters.

B. Bubble plot showing summary statistics of association analyses between immune subtypes and biological pathways (from Hallmark, KEGG and Reactome). Bubble size corresponds to adjusted p-value (Wilcoxon test, −log10 scale), while bubble color corresponds to the difference between the averaged pathway score for samples in a particular immune subtype and that of samples allocated to other immune subtypes.

C. Bubble-plot showing the association analysis between immune subtypes and kinase activity score for tumor samples. The color of the bubble corresponds to the average kinase

activity score in a particular immune subtype, while the size of the bubble corresponds to the adjusted p-value (Wilcoxon test, −log2 transformed). Associations significant at 10% FDR are denoted with an outer black circle.

D. Focal adhesion-related kinase cluster associated with immune-related TF cluster in Hot tumors. The bar plots show experimental validation via the LINCS L1000 database. The leftmost blue bar chart shows the enrichment of Innate Immune System Reactome pathway in the set of genes upregulated after kinase CRISPR-Cas knockouts from the LINCS L1000 dataset (p-values from Fisher's exact test). The centered bar chart shows the enrichment of STAT4 targets in the set of up-regulated genes after L1000 kinase knockouts. The rightmost chart displays the number of overlapping genes between STAT4 targets and the set of up-regulated genes after L1000 kinase knockouts. The red vertical line in the bar plots denotes significance (p = 0.05). Cell lines are in parentheses and their primary disease associations are: A549: Lung Cancer, AGS: Gastric Cancer, YAPC: Pancreatic Cancer, BICR6: Head and Neck Cancer, A375: Skin Cancer, ES2: Ovarian Cancer, HT29: Colon/Colorectal Cancer, MCF7: Breast Cancer, U251MG: Brain Cancer.

E. Cell-type specific pathway analysis to characterize pathway activation in tumor cells of Hot and Cold tumors based on RNA (left) and proteome data (right). The heatmap shows the average pathway score for each cell type (i.e., tumor cells of Hot tumors, tumor cells of Cold tumors and immune/stromal cells.

F. Selected glycosites stratified by immune subtype. P-values from the Wilcoxon test are reported. Asterisks represent p-values<0.001. In boxplots, the centerline represents median; box bounds represent 25th and 75th percentiles; whiskers represent minimum and maximum; and filled points represent outliers.

G. Neutrophil infiltration stratified by immune subtype. P-values from the Wilcoxon test are reported. Asterisks represent p-values<0.001. In boxplots, the centerline represents median; box bounds represent 25th and 75th percentiles; and whiskers represent minimum and maximum.

H. KM curves of progression-free survival stratified by immune subtypes. P-value from a log-rank test is reported.

I. Association between *STK11* mutation status and fraction of different cell types in tumor samples. P-values from the Wilcoxon-rank-sum test are reported. In boxplots, the centerline represents median; box bounds represent 25th and 75th percentiles; and whiskers represent minimum and maximum.

J. CD47 stratified by *STK11* mutation status and NMF proteomic subtype. P-values from the Wilcoxon test are reported. Asterisks represent p-values<0.001. In boxplots, the centerline represents median; box bounds represent 25th and 75th percentiles; whiskers represent minimum and maximum; and filled points represent outliers.

K. Scatter plot showing molecular aberrations between C1/C3 *STK11*-mutant vs C2 *STK11*-mutant samples.

See also Figure S4 and Table S4.

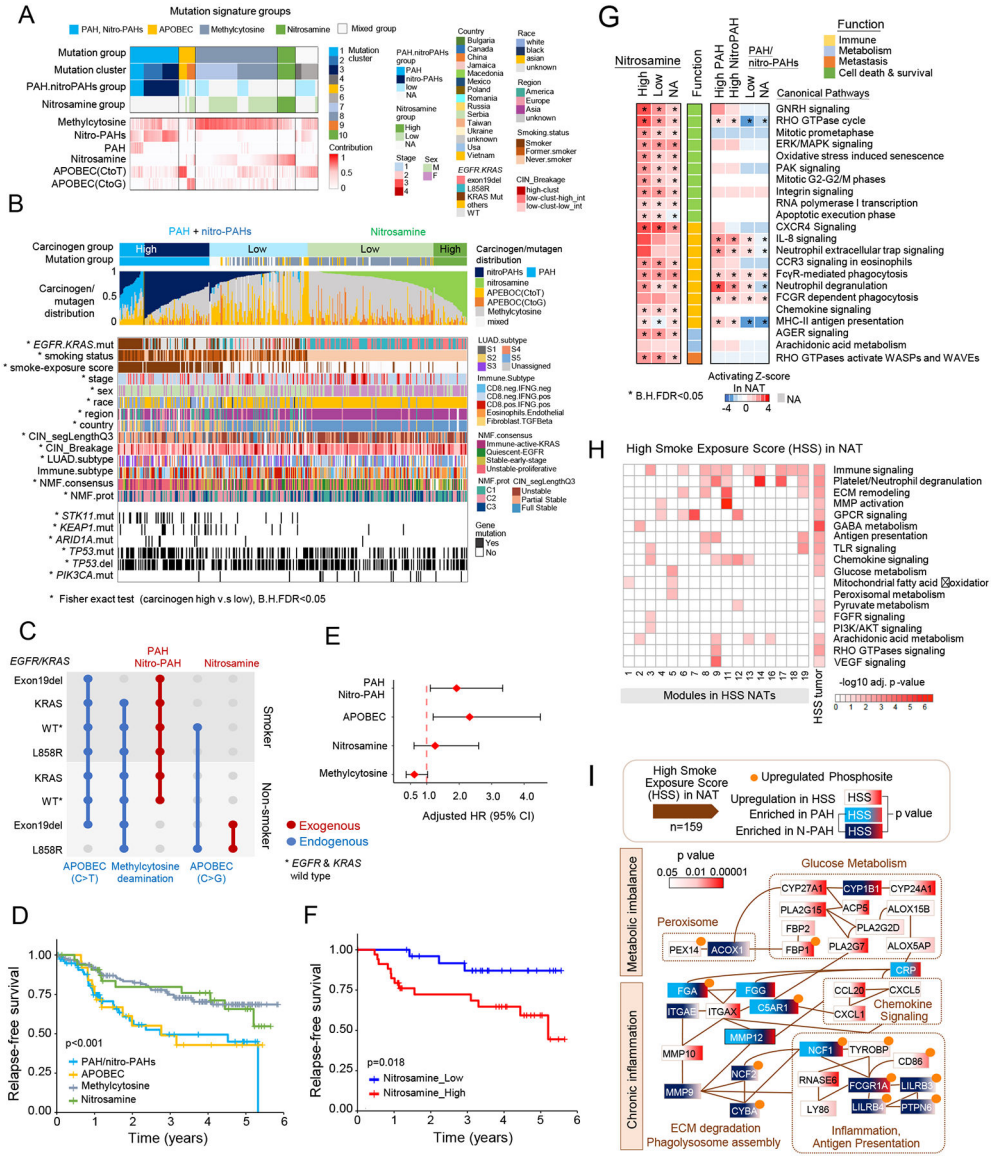


Figure 5. Environmental Carcinogens Synergistically Induced LUAD Development.

A. Ten mutation-based subclusters (Figure S5A) are condensed into four major mutagen/carcinogen groups based primarily on the relative contribution of nitro-PAHs (dark blue), PAHs (light blue), nitrosamines (green), APOBEC (yellow) and methylcytosine (gray) signatures. A mixed cluster is indicated in white.

B. Summary of association of environmental carcinogens with clinical features and gene mutations. Tumor samples are categorized into four groups based on the relative contribution of exogenous carcinogen signatures: PAHs/nitro-PAHs high, PAHs/nitro-PAHs low, Nitrosamines high, and Nitrosamines low. Clinical features with significantly different enrichment between these groups are annotated and marked with an asterisk. (Fisher's exact test, adj. $p<0.05$).

C. Summary plot showing different endogenous mutagen and exogenous carcinogen signatures (x-axis) in tumors and their enriched *EGFR* and *KRAS* mutations segregated by smoking status (y-axis).

D. KM plot showing significant differences in relapse-free survival (log-rank test, $p<0.001$) between the four carcinogen groups defined in Figure 5A.

E. Forest plot of the impact of carcinogen groups on relapse-free survival using the Hazard ratio (HR) with 95% confidence interval (CI).

F. KM plot showing significantly worse relapse-free survival (log-rank test, $p=0.018$) in patients in the high compared to low nitrosamine signature group.

G. Canonical pathway activation (activating z-score >1, adj. $p<0.05$) based on Ingenuity Pathway Analysis of proteins differentially expressed (Fisher's exact test, adj. $p<0.05$) in NATs from patients in the PAHs/nitro-PAHs and nitrosamine carcinogen signature groups.

H. Signaling pathway enrichments in high smoke-exposure score (HSS) NATs based on a protein-protein interaction network derived from proteins differentially expressed between HSS and low smoke-exposure score (LSS) NATs and partitioned into 19 strongly intra-connected modules (hypergeometric test, adj. $p <0.05$, STAR methods).

I. Upregulated proteins and phosphosites associated with metabolic imbalance and chronic inflammation from NATs with HSS, mostly contributed by samples with PAHs/nitro-PAHs signatures (adj. $p<0.05$). The protein interaction networks and functional categories are represented as connecting lines according to the STRING database.

See also Figure S5 and Table S5.

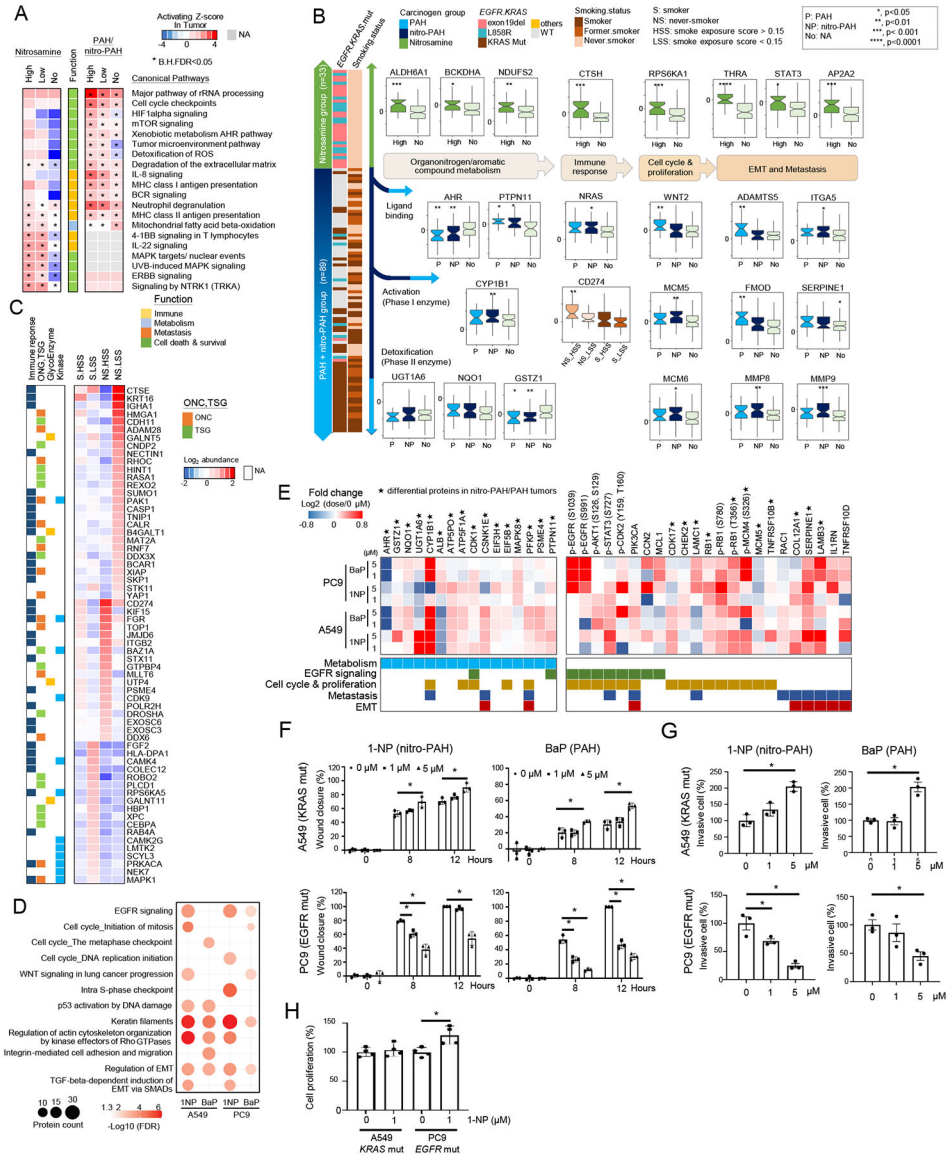


Figure 6. Nitrosamines and Nitro-PAHs/PAHs carcinogens promote different cancer-relevant pathways

A. Canonical pathway activation (activating z-score >1, adj. p<0.05) based on Ingenuity Pathway Analysis of proteins differentially expressed (Fisher's exact test, adj. p<0.05) in tumors from patients in the PAHs/nitro-PAHs and nitrosamine carcinogen signature groups. B. Differential protein expression profiles demonstrate the impact of specific carcinogens on LUAD tumors. Comparison between tumors manifesting high or absence of nitrosamine signatures (upper panel) or between high, low, or absence of PAHs/nitro-PAHs signatures (lower panel; Table S6) reveal activation of organonitrogen/aromatic compound metabolic processes, immune response, tumor progression and metastasis (activating z-score >1; Fisher's exact test, adj. p<0.05). Specific enriched proteins and phosphosites (Fisher's exact test, adj. p<0.05) underpinning shared carcinogenic processes are color-coded for the associated carcinogen signature (nitrosamine, green; PAHs, light blue; nitro-PAHs, dark blue). Vertical tracks annotate smoking status and key driver mutations (*EGFR* and *KRAS*)

for the carcinogen groups. Asterisks represent p-values (Wilcoxon rank sum test, * $p< 0.05$, ** $p<0.01$, ***$p<0.001$ and ****$p<0.0001$).

C. Protein biomarker candidates differentially upregulated in tumors with HSS or LSS from self-reported smokers or never-smokers (Kruskal Wallis tests, adj. $p<0.05$) further filtered for the ability to distinguish between any two groups (Wilcoxon rank sum test, adj. $p<0.05$). The vertical tracks annotate proteins as kinases, glycosylation-regulated enzymes, oncogenes, tumor suppressors, or immune response markers (ONC, oncogene; TSG, tumor suppressor gene.)

D. Metacore pathway analysis of differential proteomic and phosphoproteomic profiles of 1-NP- (5 uM) or BaP- (5 uM) treated A549 and PC9 cells. The differential protein numbers in the enriched pathways (Fisher's exact test, adj. $p<0.05$) are presented by the bubble plot.

E. Selected proteins or phosphosites from differential expression profiles of 1-NP- and BaP-treated A549 and PC9 cells, with tracks showing enriched functions in metabolism, EGFR signaling, cell cycle and proliferation, metastasis or EMT from Metacore pathway analysis. Asterisks indicate proteins consistently regulated in nitro-PAHs and PAHs tumor signatures.

F. Migration ability of lung cancer cells treated with carcinogens 1-NP and BaP. For panels F, G, and H, at least two independent experiments were performed, data are presented as mean ± SD, and p-values were determined by one-way ANOVA followed by Dunnett's multiple comparisons test. Asterisks indicate $p< 0.05$.

G. Invasion ability of lung cancer cells treated with carcinogens 1-NP and BaP. Asterisks indicate $p< 0.05$.

H. Cell proliferation assay for lung cancer cells treated with carcinogens 1-NP and BaP. Asterisks indicate $p< 0.05$.

See also Figure S6 and Table S6.

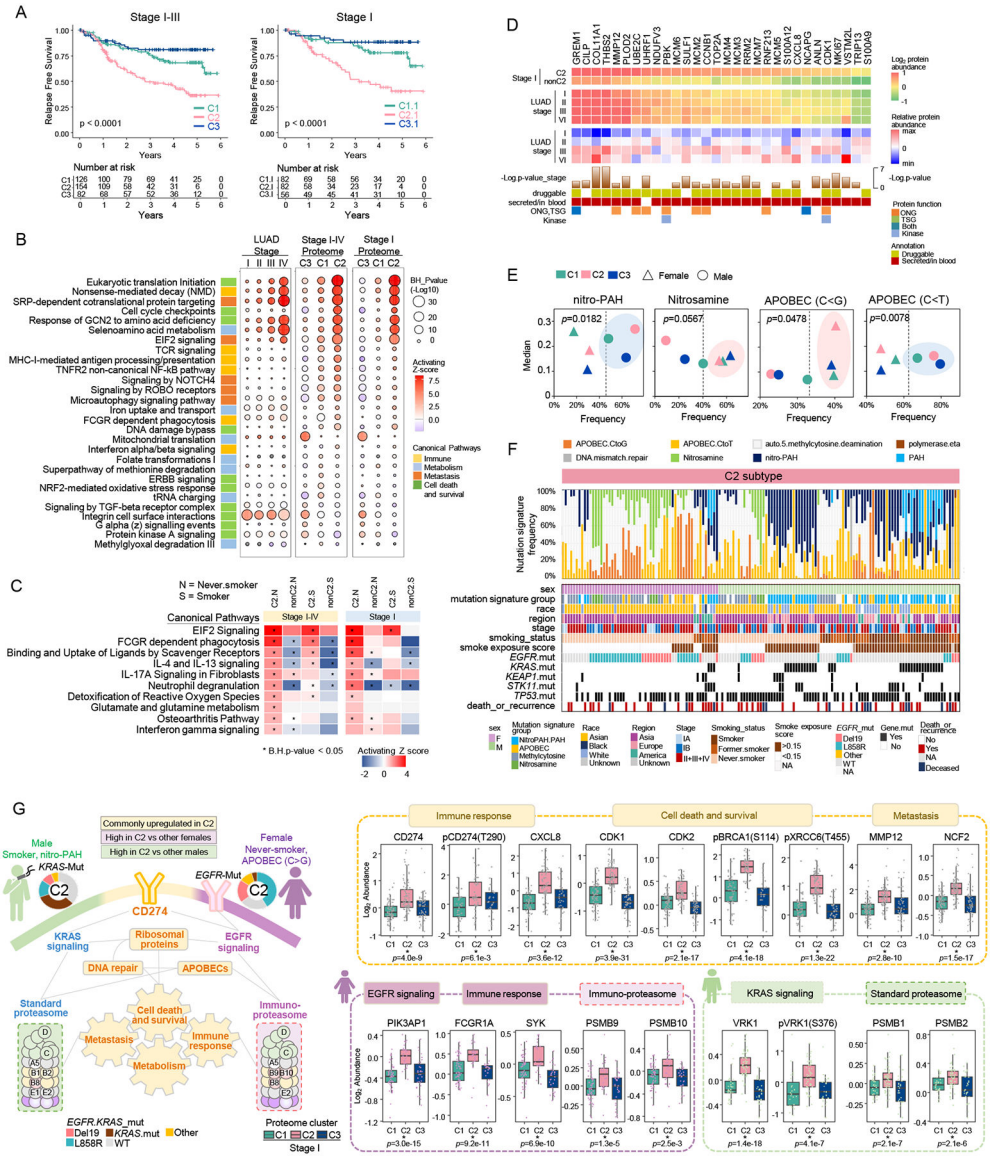


Figure 7. Proteomics Clustering Identifies Advanced Molecular Features Driving Late-like Early-stage LUAD

A. KM plots for relapse-free survival in tumors of stages I–III (left) and stage I alone (right) (log-rank test, p<0.001).

B. Summary of activated canonical pathways (activating z-score >1; Fisher's exact test, adj. p<0.05) based on Ingenuity Pathway Analysis of proteins differentially expressed in tumors of all stages (I–IV, left), C1-C3 clusters including tumors of all stages (center), and C1-C3 clusters subsetted for stage I tumors only (right).

C. Summary of activated canonical pathways (activating z-score >1; Fisher's exact test; asterisks indicate Benjamini-Hochberg adj. p<0.05) based on Ingenuity Pathway Analysis of proteins differentially expressed in C2 vs non-C2 tumors of all stages (I–IV, left) and stage I alone (right), separated by smoking status.

D. Top 30 candidate protein biomarkers significantly enriched in “late-like” early-stage tumors in the proteomic C2 subtype (Fisher's exact test, adj. p<0.01). The $\log_2$ and relative

abundances in stage I-IV or stage I in the C2/non-C2 subtype, −log p-value, protein function and annotation of these candidate biomarkers are presented.

E. Sex-specific differences in the incidence frequency (x-axis) and median values of mutagen/carcinogen exposure signatures (y-axis) in proteomic clusters C1–3 (p-values from Fisher's exact test).

F. Sex-dependent mutagen/carcinogen signature distributions within the proteomic C2 subtype. Lower tracks show clinical and mutational annotations.

G. Summary of shared and sex-specific enrichment of activated pathways in the C2 proteomic subtype (left). EGFR and BCR signaling-related proteins as well as the immunoproteasome are highly activated in C2 females, while C2 male patients have upregulated KRAS pathways and standard proteasome proteins. Shared activating proteins in the C2 subtype are involved in immune response, cell death and survival, metabolism, and metastasis (right).

See also Figure S7 and Tables S1 and S7.

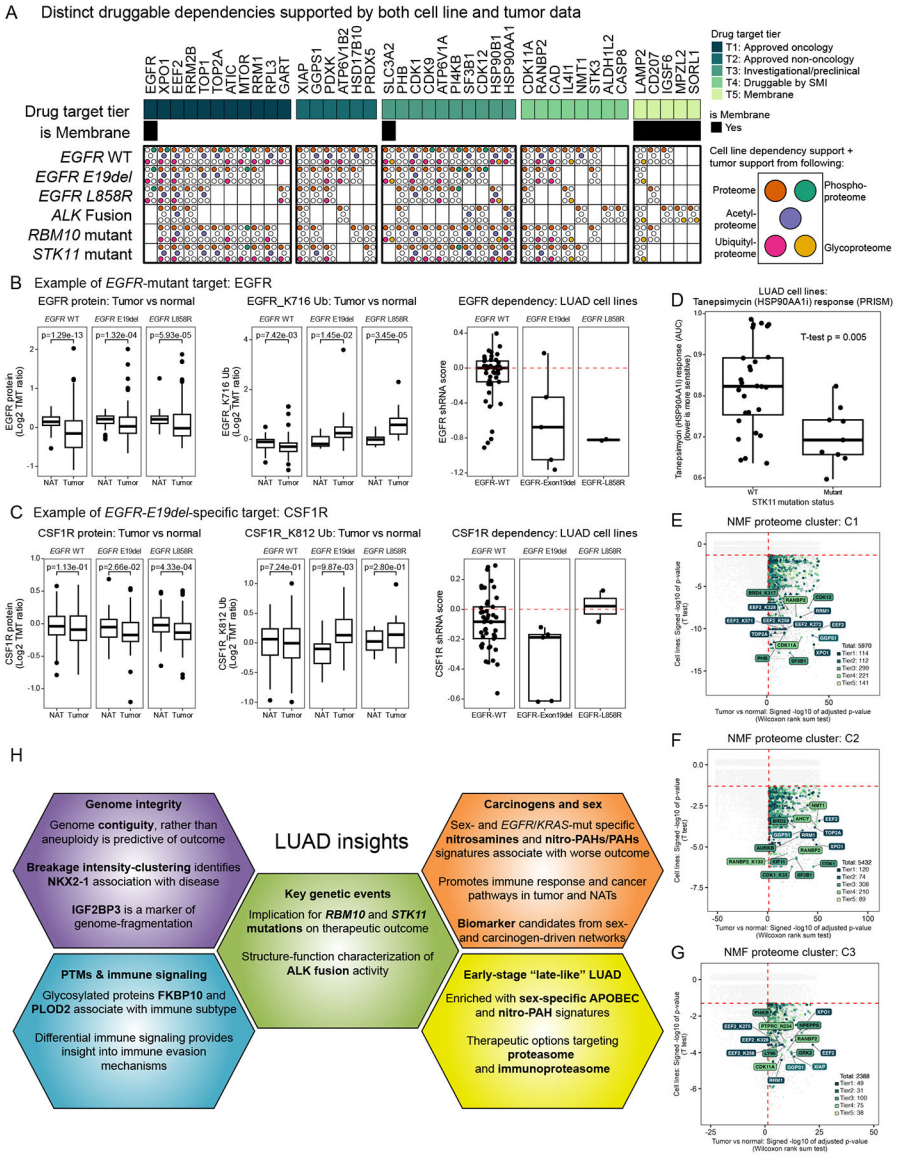


Figure 8. Candidate drug targets and biomarkers with potential translational utility

A. Plots depicting the top 15 potentially druggable targets identified in LUAD subtypes.

B. Relative abundance of EGFR protein (top panels) and EGFR_K716 ubiquitylation sites (middle panels) in tumors and corresponding NATs in *EGFR*-WT and *EGFR*-mutant tumors (p-values derived from Wilcoxon rank-sum test) and EGFR shRNA dependency scores from DepMap (lower panels; lower scores indicate greater dependency). In boxplots, the centerline represents median; box bounds represent 25th and 75th percentiles; whiskers represent minimum and maximum; and filled points represent outliers.

C. Same as in (B) except depicting CSF1R protein and CSF1R_K812 ubiquitination sites. Statistical analyses and boxplots are as described in (B).

D. Boxplots comparing response to a HSP90AA1 inhibitor, tanepsimycin (PRISM) in *STK11*-mutated or -WT LUAD cell lines. P-value derived from t-test. In boxplots, the

centerline represents median; box bounds represent 25th and 75th percentiles; whiskers represent minimum and maximum; and filled points represent outliers.

E. Potentially druggable targets for NMF proteomic cluster C1 tumors as defined by candidate upregulation in tumors vs paired NATs using multiple proteomic and PTM data types and a CRISPR effect score below zero in cluster-matched LUAD cell lines. The top 15 overexpressed, targetable dependencies are labeled. Numbers in the bottom right corner indicate the total number of overexpressed dependencies and targets from each drug target tier.

F. Same as in (E) except for NMF proteomic cluster C2 tumors.

G. Same as in (E) except for NMF proteomic cluster C3 tumors.

H. Summary of major translational findings from this study.

See also Figure S8 and Table S8.

Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
PTMScan® Ubiquitin Remnant Motif (K-ε-GG) Kit	Cell Signaling Technology	Catalog #59322, 5562
PTMScan Acetyl-lysine Kit	Cell Signaling Technology	Catalog: 13416
Rabbit Monoclonal IgG Human IGF2BP3	Abcam	Catalog No. ab179807, RRID: AB_2889895
GAPDH	Cell Signaling Technology	Catalog: 3683
Biological samples		
Primary tumor samples	This paper	Table S1
Tissue microarrays (TMAs)	Rodriguez et al. 58	N/A
Chemicals, peptides, and recombinant proteins		
HPLC-grade water	J.T. Baker	Catalog: 4218–03
Urea	Sigma	Catalog: U0631
Sodium chloride	Sigma	Catalog: 71376
1M Tris, pH 8.0	Invitrogen	Catalog: AM9855G
Ethylenediaminetetraacetic acid	Sigma	Catalog: E7889
Aprotinin	Sigma	Catalog: A6103
Leupeptin	Roche	Catalog: 11017101001
Phenylmethylsulfonyl fluoride	Sigma	Catalog: 78830
Sodium fluoride	Sigma	Catalog: S7920
Phosphatase inhibitor cocktail 2	Sigma	Catalog: P5726
Phosphatase inhibitor cocktail 3	Sigma	Catalog: P0044
Dithiothretiol, No-Weigh Format	ThermoScientific	Catalog: 20291
Iodoacetamide	Sigma	Catalog: A3221
Lysyl endopeptidase	Wako Chemicals	Catalog: 129–02541
Sequencing-grade modified trypsin	Promega	Catalog: V511X
Formic acid	Sigma, Fluka	Catalog: F0507, 56302
Acetonitrile, LC-MS grade	Honeywell	Catalog: 34967
Acetonitrile, anhydrous	Sigma	Catalog: 271004
Trifluoroacetic acid	Sigma	Catalog: 302031
Tandem Mass Tag reagent kit – 11plex	ThermoFisher	Catalog: A34808
0.5M HEPES, pH 8.5	Alfa Aesar	Catalog: J63218
Hydroxylamine solution, 50% (vol/vol) in H_2O	Aldrich	Catalog: 467804
Methanol	Honeywell	Catalog: 34966
Ammonium hydroxide solution, 28% (wt/vol) in H_2O	Sigma	Catalog: 338818
Ni-NTA agarose beads	Qiagen	Catalog: 30410
Iron (III) chloride	Sigma	Catalog: 451649
Potassium phosphate, monobasic	Sigma	Catalog: P0662

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Potassium phosphate, dibasic	Sigma	Catalog: P3786
MOPS	Sigma	Catalog: M5162
Sodium hydroxide	VWR	Catalog: BDH7225
Sodium phosphate, dibasic	Sigma	Catalog: S9763
Phosphate-buffered saline	Fisher Scientific	Catalog: 10010023
Discovery CC1	Roche-Ventana Medical System	Catalog No. 760–107
OmniMap Universal DAB Detection Kit	Roche-Ventana Medical System	Catalog No. 760–4310
Critical Commercial Assays		
TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero Gold	Illumina	Catalog: RS-122–2301
Infinium MethylationEPIC Kit	Illumina	Catalog: WG-317–1003
Nextera DNA Exosome Kit	Illumina	Catalog: 20020617
KAPA Hyper Prep Kit, PCR-free	Roche	Catalog: 07962371001
BCA Protein Assay Kit	ThermoFisher	Catalog: 23225
Deposited Data		
ICPC genomics and transcriptomics	Database of Genotypes and Phenotypes (dbGaP)	Accession: phs001954
CPTAC genomics and transcriptomics	dbGaP	Accession: phs001287
ICPC.A proteome	Proteomic Data Commons (PDC)	Accession: PDC000219
ICPC.A phosphoproteome	PDC	Accession: PDC000220
CPTAC.A proteome	PDC	Accession: PDC000153
CPTAC.A phosphoproteome	PDC	Accession: PDC000149
CPTAC.A acetylome	PDC	Accession: PDC000224
CPTAC.A ubiquitylome	PDC	Accession: PDC000627
ICPC.B proteome	PDC	Accession: PDC000563
ICPC.B phosphoproteome	PDC	Accession: PDC000564
CPTAC.B proteome	PDC	Accession: PDC000489
CPTAC.B phosphoproteome	PDC	Accession: PDC000490
CPTAC.B acetylome	PDC	Accession: PDC000491
CPTAC.B ubiquitylome	PDC	Accession: PDC000492
Link to all PDC accessions in this study	PDC	https://pdc.cancer.gov/pdc/cohort/cptac-icpc-luad
Targets and their assigned tiers/membrane	Savage et al.117	Table S2
DepMap: Mutation	DepMap Public 22Q2	https://depmap.org/portal/download
DepMap: Fusion	DepMap Public 23Q4	https://depmap.org/portal/download
DepMap: shRNA screen	DEMETER2 Data v6	https://depmap.org/portal/download
DepMap: TMT proteomics	Proteomics Files	https://depmap.org/portal/download
DepMap: PRISM drug screen	PRISM Repurposing 19Q4 Secondary Screen	https://depmap.org/portal/download
Digital Histopathology Images (CPTAC samples only)	The Cancer Imaging Archive (TCIA)	https://doi.org/10.7937/K9/TCIA.2018.PAT12TBS

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Digital Histopathology Images (CPTAC samples only)	Imaging Data Commons (IDC)	https://portal.imaging.datacommons.cancer.gov/explore/filters/?collection_id=CPTAC&collection_id=cptac_luad
Experimental Models: Cell Lines		
RT112	Leibniz Institute DSMZ	Catalog: ACC-418
AN3CA	American Type Culture Collection (ATCC)	Catalog: HTB-111
T47D	ATCC	Catalog: HTB-111
KPL1	DSMZ	Catalog: ACC-317
H3122	ATCC	Catalog: CRL-2868
H2228	ATCC	Catalog: CRL-5935
LS513	ATCC	Catalog: CRL-2134
H747	ATCC	Catalog: CCL-252
NCI-H1975	ATCC	Catalog: CRL-5908
A549	ATCC	Catalog: CCL-185
PC9	AcceGen	Catalog: ABC-TC0907
Software and Algorithms		
Methylation Array Harmonization Workflow	GDC	https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Methylation_Pipeline/
GTEx/TOPMed pipeline	GTEx Consortium.137	https://github.com/broadinstitute/gtex-pipeline/blob/master/TOPMed_RNAseq_pipeline.md
GISTIC2.0	Mermel et al.140	ftp://ftp.broadinstitute.org/pub/GISTIC2.0/GISTIC_2_0_23.tar.gz
MutSig2CV	Lawrence et al.141	https://github.com/getzlab/MutSig2CV
MSFragger	Kong et al.143	https://www.nesvilab.org/software.html
Fragpipe	Yu et al.144	fragpipe.nesvilab.org
TMT Integrator	Shteynberg et al.146	https://github.com/Nesvilab/TMT-Integrator
limma (R Package)	Ritchie et al.153	https://bioconductor.org/packages/release/bioc/html/limma.html
OmicsEV	Wen et al.154	https://github.com/bzhanglab/OmicsEV
NMF (R Package)	Gaujoux et al.156	https://cran.r-project.org/web/packages/NMF/index.html
ssGSEA 2.0	Barbie et al.158	https://github.com/broadinstitute/ssGSEA2.0
PTM-SEA	Krug et al.161	https://github.com/broadinstitute/ssGSEA2.0
PANOPLY	Mani et al.159	https://github.com/broadinstitute/PANOPLY
Terra	Broad Institute Data Science Platform	https://terra.bio/
CLUMPS-PTM	Geffen et al.39	https://github.com/getzlab/CLUMPS-PTM
ESTIMATE	Yoshihara et al.171	https://bioinformatics.mdanderson.org/estimate/index.html
TSNet	Petralia et al.172	https://github.com/WangLab-MSSM/TSNet
BayesDeBulk	Petralia et al.62	https://github.com/WangLab-MSSM/BayesDeBulk
ChEA3	Keenan et al.173	https://maayanlab.cloud/chea3/
KEA3	Kuleshov et al.67	https://maayanlab.cloud/kea3/
Enrichr	Chen et al.68	https://maayanlab.cloud/Enrichr/
KEA3 Consensus Appyter	Clarke et al.176	https://appyters.maayanlab.cloud/KEA3_Consensus_Kinases/

REAGENT or RESOURCE	SOURCE	IDENTIFIER
PhosphoDisco	Schraink et al.181	https://github.com/ruggleslab/phosphodisco
GATK4's CalculateContamination	GATK	https://gatk.broadinstitute.org/hc/en-us/articles/360036888972-CalculateContamination
GATK4 Picard tools	GATK	https://github.com/broadinstitute/picard
GATK4 Funcotator	GATK	https://gatk.broadinstitute.org/hc/en-us/articles/360037224432-Funcotator
Other		
Reversed-phase tC18 SepPak, 3cc 200mg	Waters	Catalog: WAT054925
Solid-phase C18 disk, for Stage-tips	Empore	Catalog: 66883-U
Stage-tip needle	Cadence	Catalog: 7928
Stage-tip puncher, PEEK tubing	Idex Health & Science	Catalog: 1581
PicoFrit LC-MS column	New Objective	Catalog: PF360-75-10-N-5
ReproSil-Pur, 120 Å, C18-AQ, 1.9-µm resin	Dr. Maisch	Catalog: r119.aq
Nanospray column heater	Phoenix S&T	Catalog: PST-CH-20U
Column heater controller	Phoenix S&T	Catalog: PST-CHC
300 µL LC-MS autosampler vial and cap	Waters	Catalog: 186002639
Offline HPLC column, 3.5-µm particle size, 4.6 um × 250 mm	Agilent	Catalog: Custom order
Offline 96-well fractionation plate	Whatman	Catalog: 77015200
700 µL bRP fractionation autosampler vial	ThermoFisher	Catalog: C4010–14
700 µL bRP fractionation autosampler cap	ThermoFisher	Catalog: C4010–55A
96-well microplate for BCA	Greiner	Catalog: 655101
Microplate foil cover	Corning	Catalog: PCR-AS-200
Vacuum centrifuge	ThermoFisher	Catalog: SPD121P-115
Centrifuge	Eppendorf	Catalog: 5427 R
Benchtop mini centrifuge	Corning	Catalog: 6765
Benchtop vortex	Scientific Industries	Catalog: SI-0236
Incubating shaker	VWR	Catalog: 12620–942
15 mL centrifuge tube	Corning	Catalog: 352097
50 mL centrifuge tube	Corning	Catalog: 352070
1.5 mL microtube w/o cap	Sarstedt	Catalog: 72.607
2.0 mL microtube w/o cap	Sarstedt	Catalog: 72.608
Microtube caps	Sarstedt	Catalog: 72.692
1.5 mL snapcap tube	ThermoFisher	Catalog: AM12450
2.0 mL snapcap tube	ThermoFisher	Catalog: AM12475
Microplate Reader	Molecular Devices	Catalog: M2
Offline HPLC System for bRP fractionation	Agilent 1260	Catalog: G1380–90000
EASY-nLC 1200 System	ThermoFisher	Catalog: LC140
UltiMate™ 3000 RSLCnano System	ThermoFisher	Catalog: 5200.0355
Vanquish Neo UHPLC	ThermoFisher	Catalog: VN-S10-A-01

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Q Exactive Plus Mass Spectrometer	ThermoFisher	Catalog: IQLAAEGAAPFALGMBDK
Q Exactive HF-X Mass Spectrometer	ThermoFisher	Catalog: 0726042
Orbitrap Fusion Lumos Tribrid Mass Spectrometer	ThermoFisher	Catalog: IQLAAEGAAPFADBMBHQ
Thermo Orbitrap Exploris 480	ThermoFisher	Catalog: BRE725539