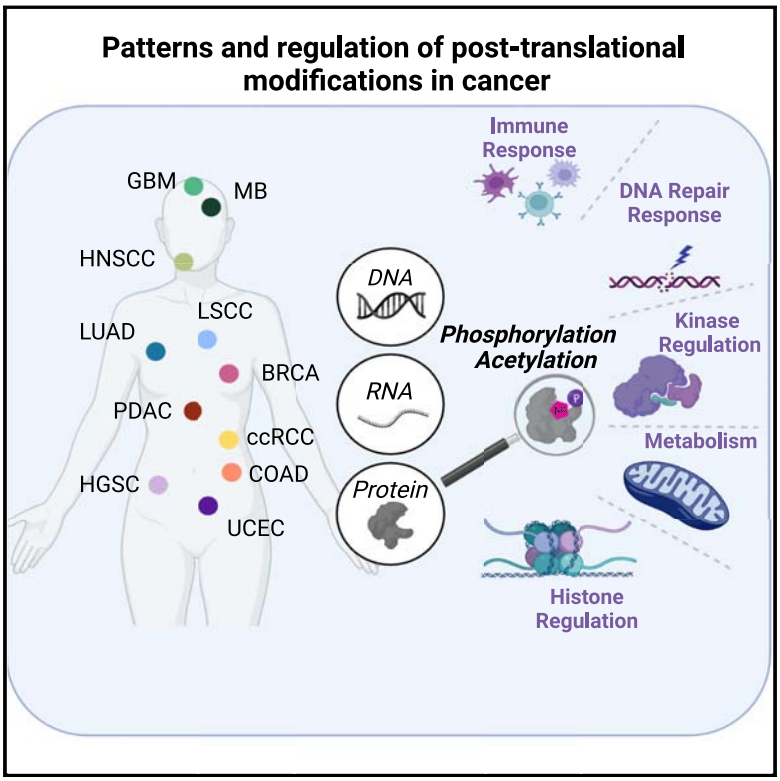
Cell

Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation

Graphical abstract



Highlights

- Unsupervised clustering reveals 33 pan-cancer multi-omic signatures
- PTM dysregulation is associated with distinct DNA damage repair mechanisms
- Changes in acetylation of metabolic proteins correlate with tumor immune state
- Phosphorylation of Thr/Ser kinases is affected by proximal acetylation

Authors

Yifat Geffen, Shankara Anand, Yo Akiyama, ..., Li Ding, Gad Getz, Clinical Proteomic Tumor Analysis Consortium

Correspondence

faguet@illumina.com (F.A.), lewis_cantley@dfci.harvard.edu (L.C.C.), lding@wustl.edu (L.D.), gadgetz@broadinstitute.org (G.G.)

In brief

An analytical resource of post-translational modifications from over 1,000 patients across 11 cancer types reveals pan-cancer changes involved in hallmark cancer processes and reveals potential new therapeutic avenues.

Geffen et al., 2023, Cell *186*, 1–23
August 31, 2023
https://doi.org/10.1016/j.cell.2023.07.013

CellPress

Cell

CellPress
OPEN ACCESS

Article

Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation

Yifat Geffen,[1,2,15] Shankara Anand,[1,15] Yo Akiyama,[1,15] Tomer M. Yaron,[3,15] Yizhe Song,[4,15] Jared L. Johnson,[3,16] Akshay Govindan,[4,16] Özgün Babur,[5,16] Yize Li,[4,16] Emily Huntsman,[3] Liang-Bo Wang,[4] Chet Birger,[1] David I. Heiman,[1] Qing Zhang,[1] Mendy Miller,[1] Yosef E. Maruvka,[6] Nicholas J. Haradhvala,[1] Anna Calinawan,[7] Saveliy Belkin,[1] Alexander Kerelsky,[3] Karl R. Clauser,[1] Karsten Krug,[1] Shankha Satpathy,[1] Samuel H. Payne,[8] D.R. Mani,[1] Michael A. Gillette,[1,13] Saravana M. Dhanasekaran,[9] Mathangi Thiagarajan,[10] Mehdi Mesri,[11] Henry Rodriguez,[11] Ana I. Robles,[11] Steven A. Carr,[1] Alexander J. Lazar,[12] François Aguet,[1,14,17,*] Lewis C. Cantley,[3,17,*] Li Ding,[4,17,*] Gad Getz,[1,2,13,17,18,*] and Clinical Proteomic Tumor Analysis Consortium

[1]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA 02142, USA
[2]Cancer Center and Department of Pathology, Massachusetts General Hospital, Boston, MA 02115, USA
[3]Weill Cornell Medical College, Meyer Cancer Center, New York, NY 10021, USA
[4]Washington University School of Medicine, St. Louis, MO 63110, USA
[5]Department of Computer Science, University of Massachusetts Boston, Boston, MA 02125, USA
[6]Biotechnology and Food Engineering, Lokey Center for Life Science and Engineering, Technion, Israel Institute of Technology, Haifa, Israel
[7]Department of Genetic and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[8]Department of Biology, Brigham Young University, Provo, UT 84602, USA
[9]Department of Pathology, University of Michigan, Ann Arbor, MI 48109, USA
[10]Leidos Biomedical Research Inc., Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA
[11]Office of Cancer Clinical Proteomics Research, National Cancer Institute, Rockville, MD 20850, USA
[12]Departments of Pathology & Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[13]Harvard Medical School, Boston, MA 02115, USA
[14]Present address: Illumina Artificial Intelligence Laboratory, Illumina, Inc., San Diego, CA 92121, USA
[15]These authors contributed equally
[16]These authors contributed equally
[17]These authors contributed equally
[18]Lead contact
*Correspondence: faguet@illumina.com (F.A.), lewis_cantley@dfci.harvard.edu (L.C.C.), lding@wustl.edu (L.D.), gadgetz@broadinstitute.org (G.G.)
https://doi.org/10.1016/j.cell.2023.07.013

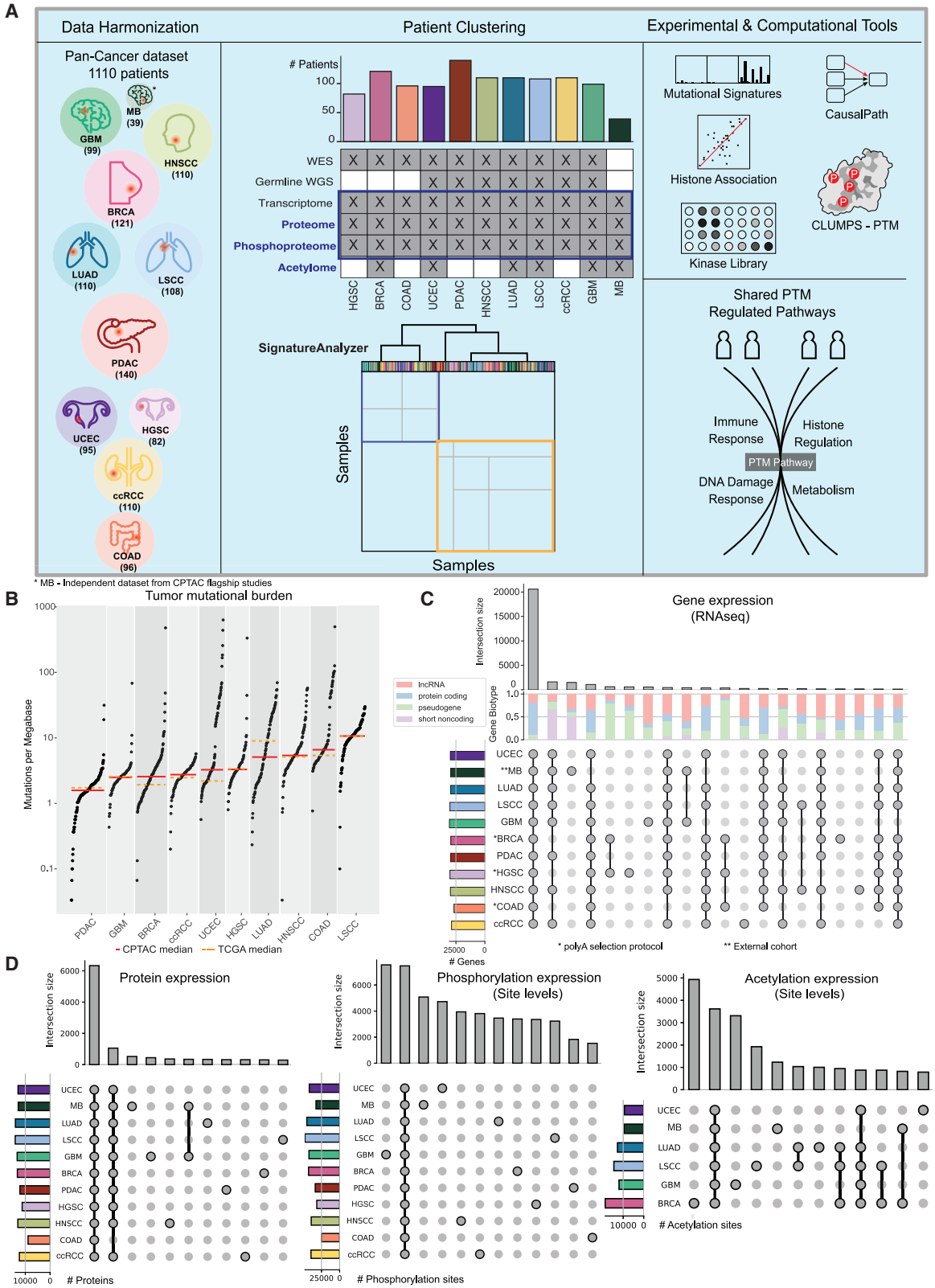
SUMMARY

Post-translational modifications (PTMs) play key roles in regulating cell signaling and physiology in both normal and cancer cells. Advances in mass spectrometry enable high-throughput, accurate, and sensitive measurement of PTM levels to better understand their role, prevalence, and crosstalk. Here, we analyze the largest collection of proteogenomics data from 1,110 patients with PTM profiles across 11 cancer types (10 from the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium [CPTAC]). Our study reveals pan-cancer patterns of changes in protein acetylation and phosphorylation involved in hallmark cancer processes. These patterns revealed subsets of tumors, from different cancer types, including those with dysregulated DNA repair driven by phosphorylation, altered metabolic regulation associated with immune response driven by acetylation, affected kinase specificity by crosstalk between acetylation and phosphorylation, and modified histone regulation. Overall, this resource highlights the rich biology governed by PTMs and exposes potential new therapeutic avenues.

INTRODUCTION

Systematic genomics-based studies of tumors have revolutionized our understanding of tumor biology[1] and significantly impacted patient care.[2] However, many cancers still lack effective treatments or remain poorly characterized, emphasizing their complex biology and their molecular and phenotypic heterogeneities.[3] Recent advances in sample processing and liquid chromatography-tandem mass spectrometry (LC-MS/MS) enable quantifying protein levels and post-translational modifications (PTMs) at a large scale.

Please cite this article in press as: Geffen et al., Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation, Cell (2023), https://doi.org/10.1016/j.cell.2023.07.013



(legend on next page)

Concerted efforts by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) have generated large proteogenomic datasets for individual cancer types.[4–13] These studies all included PTMs and started to bridge the gap between molecular features and phenotypic consequences, identifying new cancer subtypes with potential therapeutic vulnerabilities.[14] Despite these advances and the critical role PTMs play in regulation and fine-tuning of cellular signaling,[15] their shared patterns, crosstalk among PTMs (e.g., phosphorylation, acetylation, etc.), and how multiple PTMs form regulatory networks remain poorly understood, especially across cancer types.

Previous pan-cancer genomic studies have demonstrated that investigating recurrent gene and pathway alterations across different cancer types can promote our understanding of the fundamental molecular events that drive cancer.[16,17] Here, we set out to identify shared and divergent PTM patterns across cancer types to investigate common post-translational regulatory mechanisms that are altered in multiple cancers to both expand and complement genomic studies. To this end, we generated a harmonized pan-cancer cohort using data from 11 studies, encompassing samples from 1,110 treatment-naive patients, with complete genomic, transcriptomic, proteomic, and PTM (phosphorylation and acetylation) data (Figure 1A). This enabled us to search for patterns that could not be identified in a single cohort due to the limited sample size of individual studies (39–140 patients). To focus on shared and tissue-independent patterns across cancer, we regressed the tissue-specific effects in each data type as part of the harmonization process (STAR Methods).

We focused our analyses on (1) hallmark pathways known to be dysregulated in cancer[18] that are tightly controlled by PTMs,[19,20] including DNA damage and repair pathways, cell immuno-metabolism, and histone-level regulation of gene expression; and (2) potential crosstalk among different types of PTMs. PTMs have a range of potential regulatory effects—from quick, to ongoing, to long term.[19,20] In immune and metabolic responses, the transient and reversible nature of PTMs enables the quick response needed to adapt to changes in the microenvironment.[21–23] PTM effects on histone modifications, on the other hand, can affect the long-lasting regulation of cellular programs; indeed, in cancer, aberrant histone acetylation can inactivate tumor suppressors or activate oncogenes.[24,25] In DNA repair processes, phosphorylation plays a key role in regulating the activity of DNA repair proteins,[26] and PTM-focused analyses may better characterize the landscape of DNA repair, particularly in DNA repair-deficient cancers. Finally, serine/threonine phosphorylation and lysine acetylation are among the most widespread and conserved PTMs in eukaryotic organisms. Although most studies to date have focused on how a single PTM type can regulate cellular processes, the recognition that proteins harbor multiple PTM types suggests that they may act together to jointly manifest complex regulatory effects, many of which remain largely unexplored.

Overall, this is the first pan-cancer study that details the extensive regulation of acetylation and phosphorylation and their shared patterns across cancer types. Together, our results comprise a rich resource to explore and generate hypotheses regarding PTM-governed processes in cancer that, after further experimental validation, may identify new drug targets or suggest novel ways to affect cancer biology.

RESULTS

Pan-cancer dataset overview

Previous CPTAC proteogenomic studies revealed protein-based molecular tumor subtypes and identified cancer-specific pathways using PTMs. In this study, we integrate data across CPTAC cohorts to enable a pan-cancer analysis of gene, protein, and PTM patterns shared across cancer types. To accomplish this, CPTAC pan-cancer working groups harmonized data from all available cohorts using standardized pipelines for assessing somatic mutations, somatic copy-number alterations (SCNAs), mRNA expression, protein abundance, phosphorylation, acetylation, and clinical data (Figure S1).[27] The final combined dataset comprised 1,110 patients from 11 cohorts (Figure 1A). Ten tumor types were part of CPTAC, including glioblastoma (GBM),[4] head and neck squamous cell carcinoma (HNSCC),[5] lung adenocarcinoma (LUAD),[6] lung squamous cell carcinoma (LSCC),[7] breast cancer (BRCA),[8] pancreatic ductal adenocarcinoma (PDAC),[9] clear cell renal cell carcinoma (ccRCC),[10] high-grade serous ovarian cancer (HGSC),[11] uterine corpus endometrial carcinoma (UCEC),[12] and colorectal adenocarcinoma (COAD).[13] An external medulloblastoma (MB) dataset[28] was also included, generated following the same protocols as the CPTAC datasets for all available data types but lacking whole-exome DNA sequencing. For each patient, we identified both germline and somatic variants and quantified gene expression, protein abundance, and PTM levels (STAR Methods). We detected a median of ~25,000 exonic germline variants and ~320 exonic somatic mutations per patient, with median somatic mutation burdens that matched The Cancer Genome Atlas cohorts (Figure 1B). As expected, a subset of UCEC and COAD patients showed exceptionally high tumor mutational burden (TMB), reflective of microsatellite instability (MSI) and polymerase proofreading deficiencies (POLE and POLD1 exonuclease domain mutants). In addition, we found that an average of

Figure 1. Pan-cancer dataset overview

(A) Pan-cancer analysis workflow: (left) data harmonization of available cohorts; (middle top) available data types and discovery of multi-omic signatures based on RNA, proteins, and phosphosites, (middle bottom) clustering of samples based on signature activities; (right top) experimental and computational tools used to study clusters of tumors and pathways; (right bottom) highlighted cancer pathways with altered post-translational modifications.

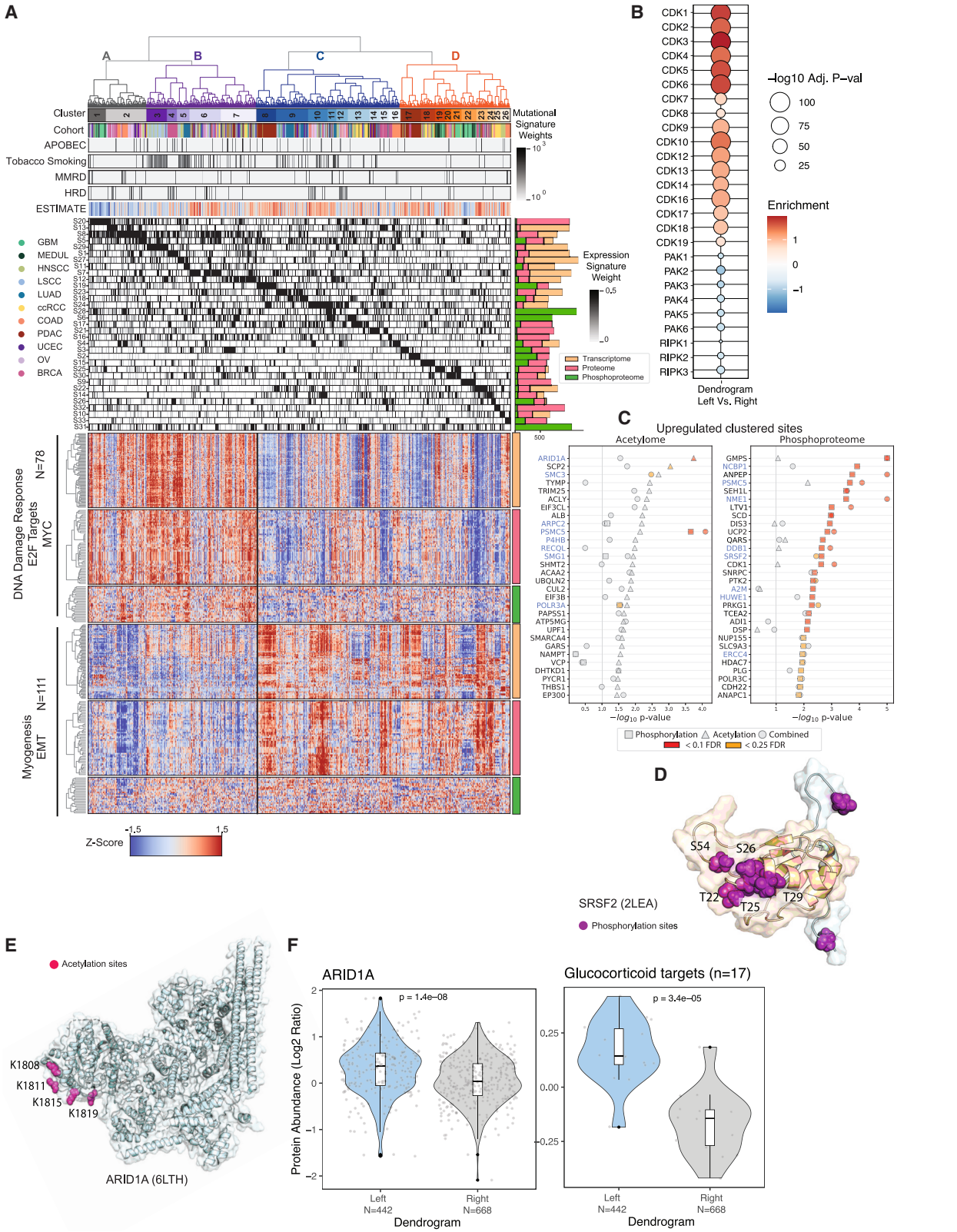
(B) Tumor mutational burden (TMB) across cohorts—CPTAC median, red; TCGA median, dotted orange.

(C) Upset plots showing the distribution of shared expressed genes (RNA) and the different RNA biotypes contribution.

(D) Upset plots showing the distribution of shared proteins (left) and site-level phosphorylation (middle) and acetylation (right) across the different cohorts (bars representing ~85% of the data for visibility).

See also Figure S1.

Please cite this article in press as: Geffen et al., Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation, Cell (2023), https://doi.org/10.1016/j.cell.2023.07.013



(legend on next page)

~24,000 genes, including coding and non-coding, were expressed in any cohort (transcripts per million [TPM] $\geq$ 0.1 and $\geq$6 reads each in at least 20% of patients). We detected an average of ~10,000 proteins, ~22,000 phosphosites, and ~6,000 acetylation sites per patient (available for 6 cohorts) (STAR Methods; Figures 1C and 1D).

Next, since we aimed to search for pan-cancer patterns, we analyzed the overlap among genes, proteins, and PTM sites. We found ~21,000 genes to be expressed across all cohorts (~14,500 protein coding and ~6,500 non-coding; Figure 1C); moreover, 6,333 proteins were detected across all cohorts and accounted for the majority of the data. Importantly, PTMs show a more discrete pattern in each tumor type, with relatively fewer shared across cohorts (Figure 1D, center and right panels; Table S1). This may reflect their role in fine-tuning responses at the cell- and tissue-type level beyond that of gene or protein expression alone.[29,30]

Pan-cancer PTM landscape

To explore shared PTM patterns across cancer, we first integrated the data types that were available across all 11 cohorts—specifically, gene expression, protein abundance, and phosphoprotein level data—while regressing out tissue-specific effects to remove obvious differences among tumor types (Figures S1A–S1D; STAR Methods). We applied SignatureAnalyzer, a Bayesian variant of non-negative matrix factorization (NMF),[31–33] across the 1,110 tumors represented by a combined set of 14,057 features (see SignatureAnalyzer section in STAR Methods) to obtain 33 pan-cancer multi-omic signatures. Notably, most signatures had contributions from all 3 features (Figure 2A; STAR Methods). In addition to defining the signatures, SignatureAnalyzer estimates the activity level of each signature for each tumor. By assigning each tumor to its most active signature, we found that most signatures span multiple tumor types (Figure S2A), suggesting that, in general, the signatures reflect pan-cancer biological processes.

To characterize tumor subsets with both shared and divergent biology, we performed hierarchical clustering of the samples based on their activities across the 33 signatures, which more robustly reflect the pan-cancer biological processes active in each sample (Table S2; STAR Methods). In addition, to define groups of samples that share their most prominent biology (used for certain downstream analyses), we traversed the dendrogram and defined clusters based on their most frequent dominant signature, identifying 26 non-overlapping terminal clusters (Figures S2B and S2C). To further explore the biology in each group, we performed pathway enrichment analyses at the RNA, protein, and PTM levels (Table S2) and applied multiple methods specifically tailored to identify PTM differences: (1) CLUMPS-PTM (Figures S2D–S2F; STAR Methods), (2) The Kinase Library,[34] (3) CausalPath,[35] (4) PTM signature enrichment analysis (PTM-SEA),[36] and (5) a method to predict differential activity of histone regulators (PTM dedicated tools in the STAR Methods). Aggregating the results allowed us to comprehensively characterize differences in tumor biology across our pan-cancer cohort.

We started our analysis by focusing on the top split in the dendrogram comprising a significant enrichment for DNA damage response (DDR) and proliferation pathways (MYC and E2F) on the left side of the split, and myogenesis and epithelial-mesenchymal transition (EMT) pathways on the right side (Figure 2A). We applied both a site-specific pathway enrichment analysis—PTM-SEA as well as The Kinase Library that predicts the PTM regulators based on their substrate specificity (PTM dedicated tools in the STAR Methods). Both tools showed significant enrichment of cyclin-dependent kinase (CDK) activity and downregulation of p21-activated kinases (PAKs) in the left side of the first split (Figure 2B; Table S2). These results are consistent with the pathway activation differences, since CDK-mediated phosphorylation is associated with rapid cell proliferation, whereas PAKs are associated with the actin cytoskeletal remodeling and increased migratory phenotype that accompanies EMT.[37]

Using CLUMPS-PTM to identify clusters of correlated PTMs in protein 3D structures (PTM dedicated tools in the STAR Methods), we found 22 proteins with significant clustering (false discovery rate [FDR] < 0.1) of phosphosites that were upregulated in the first split when comparing the left vs. right side of the dendrogram (Figure 2C). One of the top hits for significant phosphorylation clustering was SRSF2 (FDR = 0.044), a serine- and arginine-rich splicing factor. The cluster falls within the RRM-1 domain, and its phosphorylated form has been shown to interact with E2F1 to promote transcriptional control of cell cycle target genes such as cyclin E (Figures 2D and S2F), thereby promoting cell proliferation in lung carcinoma cell lines.[38]

For acetylation sites, we found only one protein, ARID1A, with a significant clustering (FDR = 0.085). ARID1A is a SWI/SNF chromatin remodeler that is commonly mutated in cancer[39] and plays a complex role in tumorigenesis.[40] The cluster is on the C-terminal tail of ARID1A, within the glucocorticoid receptor (GR)-binding

Figure 2. Pan-cancer PTM landscape

(A) Hierarchical clustering of sample similarity matrices across their signature activities (middle heatmap). Tracks: (top) cluster, and cohort annotations, (middle) whole-exome mutational signatures, and (below) ESTIMATE assignments. Lower panel heatmap shows RNA, proteins, and phosphosites in the top differentially expressed pathways between the left and right sides of the first split of the dendrogram.

(B) Bubble plot representation of The Kinase Library enrichment based on differentially expressed substrates of each kinase between the first split of the dendrogram. Enrichment (red), depletion (blue).

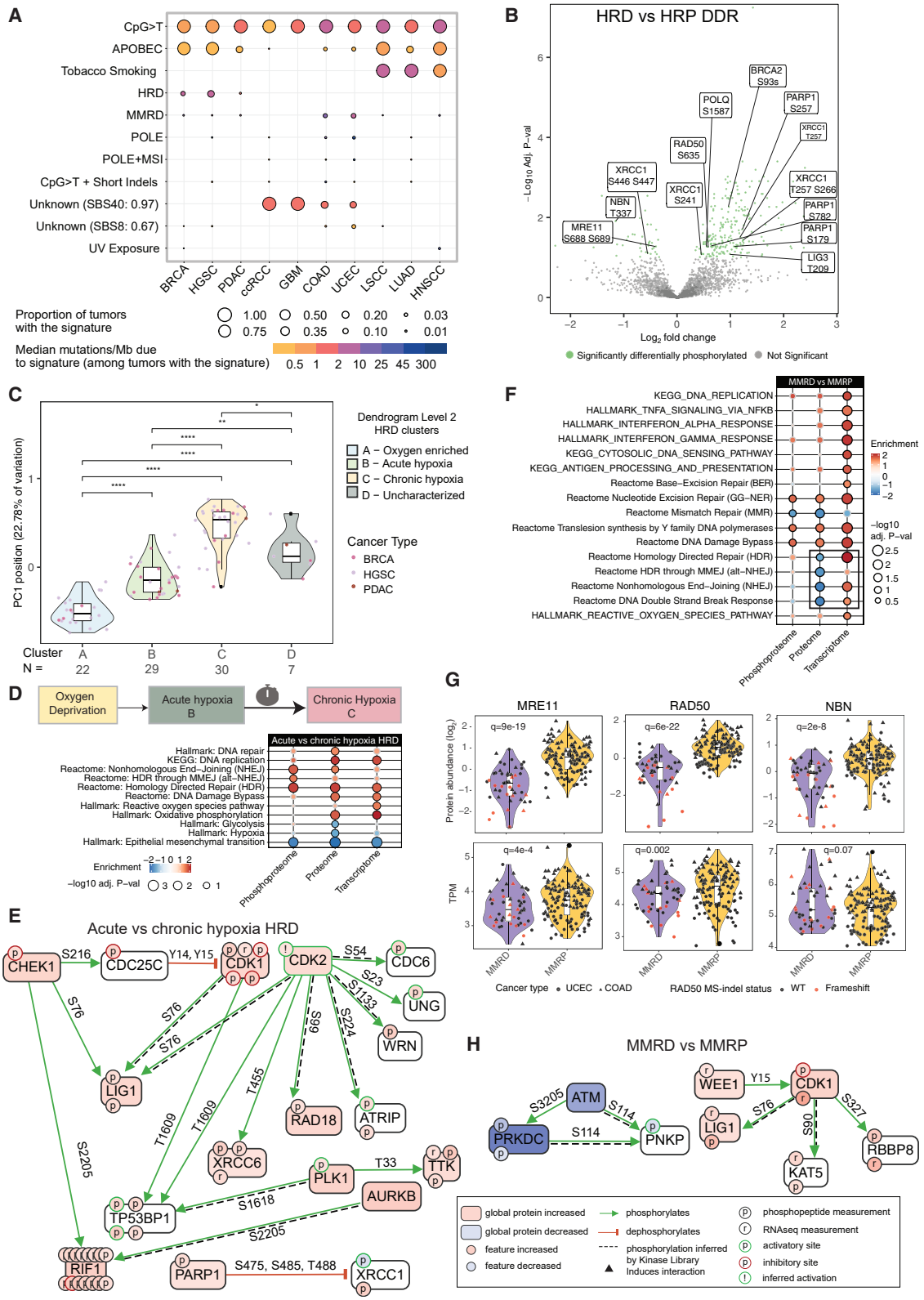
(C) CLUMPS-PTM results for the first split shows significant 3D spatial clustering of differentially acetylated (left, triangles) or phosphorylated (right, boxes) sites. Circles represent significance based on the union of both. DDR hallmark geneset (blue). Red, significant results, FDR < 0.1; yellow, near significance results, FDR < 0.25.

(D) SRSF2 phosphorylation cluster on 3D crystal structure (cyan, PDB: 2LEA), RRM-1 domain (amber, RNA recognition motif), phosphosites (purple).

(E) ARID1A acetylation cluster on 3D crystal structure (cyan, PDB: 6LTH), acetylsites (pink).

(F) Violin plots showing protein abundances of ARID1A (left), and glucocorticoid targets (right) between the first split of the dendrogram.

See also Figure S2.



(legend on next page)

domain (Figures 2E and S2F). GR regulates many genes whose products increase catabolism, reduce inflammation, and increase cell survival.[41] Increased acetylation in this cluster can potentially block a ubiquitination site at the C-terminal of the protein,[42] which would reduce ARID1A degradation and potentially increase GR signaling. We indeed observe a higher protein abundance of ARID1A and GR targets on the left side of the dendrogram (Figure 2F).

Mechanisms of PTM dysregulation in DNA repair-deficient tumors

Next, we leveraged our proteogenomic dataset to investigate the effects of DNA repair deficiencies that are undetectable at the genomic and transcriptomic levels. We first extracted mutational signatures across our cohort by applying SignatureAnalyzer[31–33,43,44] to five partitions of our dataset by distinct environment and cell-intrinsic mutational mechanisms: POLE/POLD1-exonuclease domain mutants, mismatch repair-deficient (MMRD), smoking-related, homologous recombination deficiency (HRD)-related, and not HRD-related (Figure S3A; STAR Methods). We extracted a total of 22 mutational signatures representing 11 distinct mutational processes, including MMRD and HRD (Figure 3A; Table S3; STAR Methods). Using these signatures, we identified 57 MMRD and 88 HRD tumors (Figure S3B; STAR Methods). Consistent with previous pan-cancer studies, most MMRD tumors were from the COAD and UCEC cohorts (21 and 28 tumors, respectively), whereas the HRD group encompassed 54 HGSC, 30 BRCA, and four PDAC tumors.[45,46]

HRD cancers rely on alternative repair pathways to mitigate double-strand break (DSB) damage.[47] To investigate the PTM-directed activities of repair proteins in HRD cancers, we performed differential expression analyses (across all feature types) between HRD and homologous recombination-proficient (HRP) tumors across DNA repair genes, followed by CausalPath analysis to identify causal relationships between PTMs and their mediators (PTM dedicated tools in the STAR Methods). These comparisons revealed significant differences in the phosphorylation of 268/1,596 sites residing in 112/310 measured DNA repair proteins. In particular, we found differences in 8/12 proteins representing the microhomology-mediated end-joining (MMEJ) pathway, which is the primary HRD compensatory pathway (FDR $\leq$ 0.1, Figure 3B; Table S3).[48,49] Notably, we found increased phosphorylation of three PARP1 phosphorylation sites, including PKA-mediated site S782 (FDR = 0.05) and ATR-mediated site S179 (FDR = 0.05), which are known to regulate PARP1 activity.[50–52] HRD tumors also exhibited significantly increased phosphorylation of POLQ on S1587 (FDR = 0.05). POLQ promotes MMEJ by inhibiting RAD51-mediated HR, and its loss has been shown to elicit synthetic lethality in HRD tumors, including in cell lines resistant to PARP inhibition[53,54] (Figure S3C); the functional effects of S1587 phosphorylation, however, have not been well-studied. We additionally found increased phosphorylation of EXO1 S714 (FDR = 0.04), an ATM-mediated site that has been proposed to attenuate EXO1 activity and hinder homologous recombination (HR) as a result.[55] Differential phosphorylation analysis thus revealed site-specific modifications that may regulate mechanisms that compensate for HR loss.

We observed that the 88 HRD tumors spread across the four main branches of the dendrogram (clusters A–D, Figure 2A). We therefore explored whether this partitioning reflected different DNA repair activities, which could potentially associate with different therapeutic vulnerabilities. First, we performed principal component analysis (PCA) of the multi-omic signature weights in HRD tumors to verify that their partitioning in the pan-cancer dendrogram was maintained when focusing only on these tumors. We found that even the first principal component (PC1) can separate these HRD clusters (Figure 3C). To characterize the biological processes associated with each cluster, we performed pairwise multi-omic differential expression analyses between A, B, and C (excluding D due to the small sample size of n = 7). Gene set enrichment analysis (GSEA) revealed that B exhibited significant upregulation of hypoxia-related proteins compared with A and significant downregulation compared with C (FDR = 0.08 and 0.03, respectively; Table S3). Previous cell line studies have described the relationship between hypoxia severity and DDR, showing that acute hypoxia with periodic re-oxygenation activated DNA repair pathways to mitigate reactive oxygen species (ROS)-related DNA damage, whereas chronic hypoxia stalled replication and suppressed DDR.[56–59] We hypothesized that B represented an acute hypoxia group, and C a chronic hypoxia group. Indeed, GSEA also highlighted an upregulation of the ROS pathway at the mRNA level (FDR = 0.04); DNA repair at the protein level (FDR = 0.02); and DNA replication at both mRNA and protein levels in B compared with C (FDR = 0.01, 0.02, respectively) (Figure 3D; Table S3). Similarly,

Figure 3. PTM analysis of DNA repair deficiencies

(A) Mutational signatures associated with each cohort. Circle size represents the proportion of tumors. Circle color indicates median mutations/Mb.

(B) Volcano plot illustrating the differential phosphorylation between HRD and HRP tumors. MMEJ genes are labeled.

(C) Violin plot of 1st principal component projections based on the multi-omic signature activities for HRD tumors. Points are colored by their cancer type and separated by HRD cluster. *p value < 0.05, **p value < 0.01, ***p value < 0.001, and ****p value < 0.0001.

(D) Schematic diagram of the acute and chronic hypoxia HRD clusters (top). Arrow length represents duration of hypoxia. Bubble plot showing GSEA results between the acute and chronic hypoxia HRD subgroups (bottom).

(E) CausalPath results of differentially expressed DDR genes between acute and chronic hypoxia HRD tumors. Acute hypoxia upregulation (red), downregulation (blue). Black dashed lines, 90th percentile scoring substrates based on The Kinase Library results.

(F) Bubble plot showing GSEA results between MMRD and MMRP tumors. MMRD pathways upregulated (red), downregulated (blue). Box indicates opposing effects at the RNA and protein levels in DSB pathways.

(G) Violin plots showing protein abundance (top) and RNA (bottom) levels of MRN complex proteins between MMRD and MMRP tumors (COAD [circle] and UCEC [triangle]). RAD50 microsatellite frameshift indel samples indicated in red.

(H) CausalPath results of differentially expressed DDR genes between MMRD and MMRP tumors as in (E).

See also Figure S3.

Please cite this article in press as: Geffen et al., Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation, Cell (2023), https://doi.org/10.1016/j.cell.2023.07.013

PTM-SEA detected in the acute hypoxia HRD cluster B increased activity of the DNA damage signaling kinases ATM, CHEK1, and CHEK2 (FDR = 0.09, 0.02, and 0.06, respectively) as well as an enrichment of CDK1/2/4/6 activities (all FDR = 0.018; Table S3), as expected.[56–58]

In order to identify specific differences in regulators of DDR proteins between acute and chronic hypoxia HRD tumors, we applied CausalPath on all differentially expressed features. CausalPath detected increased PARP1 and XRCC1 interaction in the acute hypoxia HRD group through increased PARP1 protein (FDR = 0.09) and decreased phosphorylation of XRCC1 sites S475, S485, and T488 (FDR = 0.102, Figure 3E).[60] This interaction facilitates XRCC1 recruitment to ROS-induced base lesions and single-strand break sites, suggesting increased PARP1 activity in the base-excision repair (BER) pathway,[61,62] which is necessary for the PARP trapping mechanism of PARP inhibitors.[63] Further supporting increased PARP activity, GSEA showed protein-level enrichment of oxidative phosphorylation and downregulation of glycolysis pathways (FDR = 0.0007 and 0.08, respectively), consistent with a known pro-survival metabolic shift from glycolysis to oxidative phosphorylation due to PARP consumption of NAD^+.[64] Furthermore, CausalPath highlighted CDK2 phosphorylation of WRN S1133, which is further supported by The Kinase Library (97.5th percentile of CDK2 substrates) and is a known response to collapsed replication forks[49,65] (Table S3). Overall, phosphorylation-focused analysis highlighted major differences in the PTM activity of DDR proteins between the acute and chronic hypoxic HRD tumors that are indistinguishable at the mutational signature level.

As opposed to the partitioning of the HRD tumors into four groups, the MMRD tumors showed only tissue-driven separation (Figure S3D), prompting us to analyze all the MMRD tumors together. Similar to the HRD analysis, we ran differential expression analysis between MMRD and mismatch repair-proficient (MMRP) tumors across all feature types associated with DNA repair genes (Table S3). As expected, due to common MLH1 promoter hypermethylation in MMRD tumors, depletion of MLH1 RNA expression and protein abundance (FDR = 1×10^{-33}, 5×10^{-30}, respectively) were among the most significant differences. To characterize pathway-level differences, we performed GSEA on all differentially expressed features (Figure 3F; Table S3). Intriguingly, MMRD tumors exhibited upregulation of DSB repair pathways at the mRNA level and downregulation at the protein level (FDR = 0.09 and 0.01, respectively; box in Figure 3F). This difference is likely due to the fact that different genes drive the mRNA vs. protein pathway activation levels (14 vs. 4 distinct leading edge genes, respectively). The 4 leading edge proteins were MRE11, RAD50, NBN (which together form the DSB sensing and signaling MRN complex[66]), and ATM (FDR = 2×10^{-17}, 2×10^{-18}, 2×10^{-7}, and 2×10^{-3}, respectively; Figure 3G). We investigated whether truncating microsatellite mutations, which were previously found to be enriched in RAD50, MRE11, and ATM,[67–69] could explain the decreased mRNA and protein levels of these genes. We found 16 patients with truncating alterations in *RAD50* (ten K722fs and four N934fs frameshift indels), all of which were in the MMRD group (16/49 vs. 0/142, FDR = 1×10^{-10}, Table S3; STAR Methods). The analysis of *MRE11* and *NBN* alterations was not sufficiently powered due to a few truncating events (three and one, respectively). Interestingly, we found a similar decrease in protein abundance of MRE11, RAD50, and NBN in MMRD vs. MMRP cell lines, but only a marginal decrease in *MRE11* mRNA, and no change (or even increase) in *RAD50* and *NBN*, respectively (Figure S3E), raising the possibility that reducing one protein in the complex may destabilize the complex and lead to degradation of the other complex proteins. Further studies are needed to explore the mechanisms of reduced expression of the MRN complex in MMRD tumors.

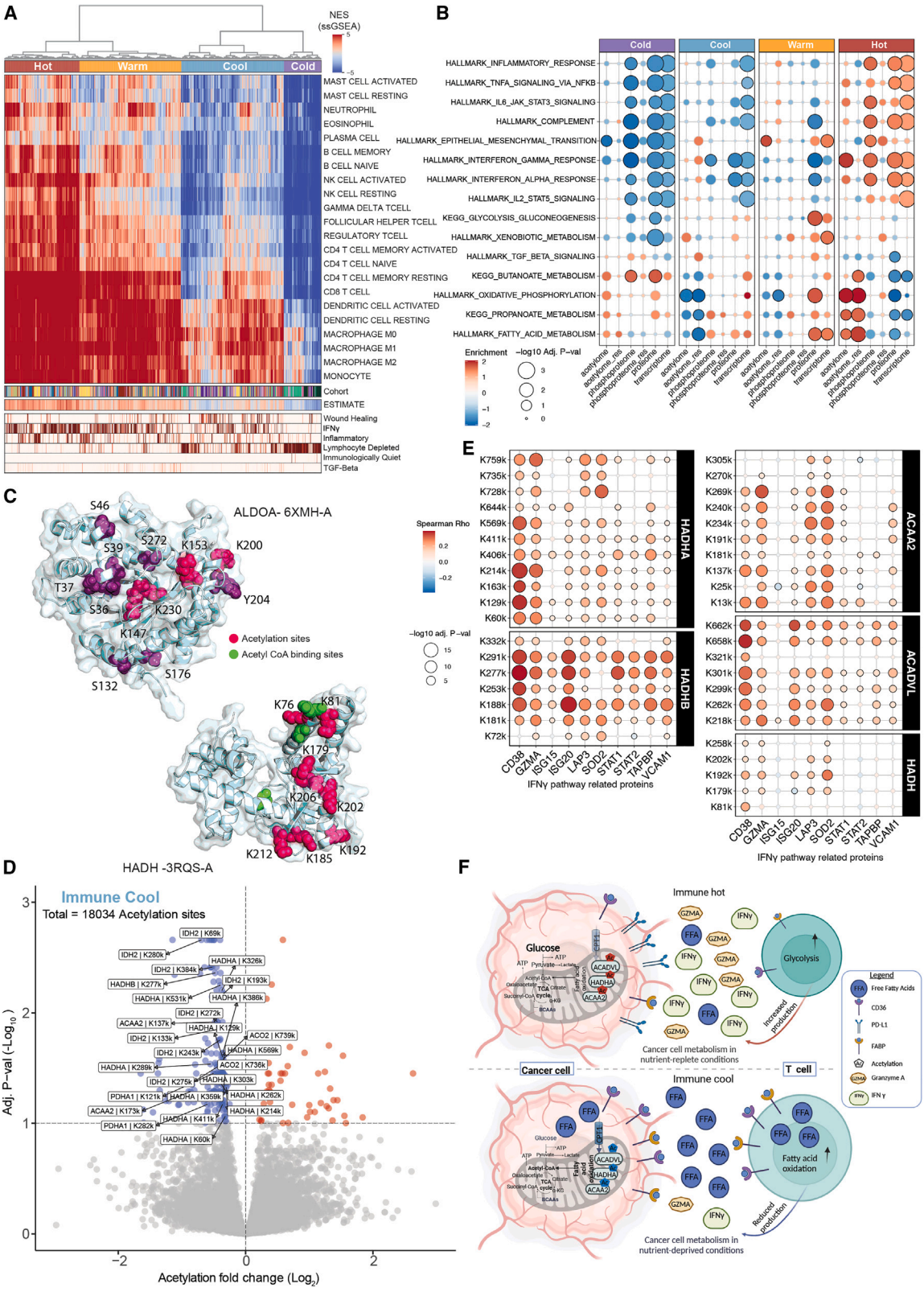
Consistent with previous reports, we also found significant enrichment of ATM microsatellite indels in MMRD tumors (16/49 vs. 3/142, FDR = 5×10^{-8}; Table S3). Applying CausalPath to the differential expression results revealed evidence of DSB sensing and signaling defects due to ATM loss (Figure 3H). MMRD tumors showed a decrease in ATM-mediated phosphorylation of PRKDC (DNA-PKcs) at S3205 (FDR = 0.1), which induces DSB repair signaling.[69] We also found a decrease in ATM and PRKDC-mediated phosphorylation of PNKP at S114 and T118 (FDR = 0.09, Table S3). Phosphorylation of these sites is critical for PNKP retention at DSB sites and subsequent processing of DSB ends prior to ligation in the NHEJ pathway.[70,71] These results highlight that proteogenomic analyses can expose effects of somatic deleterious alterations in MMRD tumors that cannot be observed at the mRNA level alone.

PTM regulation of metabolic pathways affects tumor-associated immune responses

The interplay between cell metabolism and the immune response was previously established[19,20]; here, we aimed to characterize the effects of PTM regulation on this interplay across tumor types. First, we applied multiple methods that infer immune infiltration and activity: (1) ESTIMATE,[72] which estimates the abundance of immune infiltration based on expression levels of curated genesets; (2) ImmuneSubtypeClassifier,[73] which provides immune phenotypes at a granular level using a classifier approach; and (3) unsupervised clustering based on enrichments of curated genesets from CIBERSORT.[74] Our unsupervised clustering approach revealed four broad immune subtypes across different cancer types: immune-cold, -cool, -warm, and -hot. These subtypes aligned with the results from ESTIMATE and ImmuneSubtypeClassifier (Figures 4A, S4A, and S4B; Table S4). We observed a mixed tumor distribution in the subtypes, except for the "immune-cold" subtype, which was predominantly composed of brain tumors (93 out of 130 samples, Figure S4C), consistent with brain tumors typically being immune-cold due in part to the blood-brain barrier.[75] Next, we performed differential expression and pathway analyses in these immune subtypes using estimated tumor-intrinsic expression by removing the contribution of the immune cells (Figure S4B; Table S4; STAR Methods). Similar to previous studies,[6,7] our immune-hot subtype showed an increase in immune-related pathways (Figure 4B) as well as a significant increase of immunosuppressive markers, including IDO1, CD163, ENTPD1, and PD-L1 (CD274) (immune-hot vs. immune-warm FDR < 0.1, Figure S4D). The median fold change was lower than was previously reported in LUAD, potentially due to the large heterogeneity across cancer types (median FC differences

Please cite this article in press as: Geffen et al., Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation, Cell (2023), https://doi.org/10.1016/j.cell.2023.07.013

CellPress
OPEN ACCESS



(legend on next page)

Please cite this article in press as: Geffen et al., Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation, Cell (2023), https://doi.org/10.1016/j.cell.2023.07.013

between immune-hot vs. immune-warm ranges between 0.32 and 1.5 vs. >1.9 in CPTAC LUAD[6]). We also found significant differences in acetylation levels across multiple metabolic pathways in the immune-cool subtype, including propanoate metabolism, oxidative phosphorylation, and fatty acid (FA) metabolism pathways (Figure 4B; FDR < 0.07). In these pathways, acetylation is known to play an important inhibitory role.[7,76,77] We observed high levels of acetylation in lipid metabolism pathways in the immune-hot group and low levels of acetylation in the immune-cool subtype (FDR < 0.01), even after correction for protein abundance (denoted as acetylome_res in Figure 4B and STAR Methods), suggesting that low acetylation levels potentially contribute to the high activation of these pathways in immune-cool tumors. Indeed, FA enzymes are known to control specific gene expression,[78] and this effect is mainly regulated at the protein and PTM levels.[79] Of note, immune-cold tumors show a similar metabolic pathway activation as immune-hot tumors, perhaps due to the fact that the brain mostly relies on glucose as its energy source since it requires less oxygen for ATP generation[80] (although this explanation does not explain why the few non-brain tumors clustered into the immune-cold subtype, which is yet to be determined) (Figure S4E).

Next, we employed CLUMPS-PTM on the differentially regulated sites among these immune subtypes to identify functional regions on the 3D protein structures. The glycolytic domain of the ALDOA enzyme, which is abundant in cancer,[81] was found to harbor a significant cluster of 4 increased acetylated sites in the immune-hot group (K147, K153, K200, and K230; FDR < 0.12, subset to glycolysis proteins), three of which are also known ubiquitination sites that can lead to protein degradation.[42] The same domain harbored a significant cluster of increased phosphorylated sites (FDR = 0.06, subset to glycolysis proteins) in the immune-warm group (Figure 4C top; Table S4). In contrast, in the immune-cool group, multiple FA metabolic-related proteins (e.g., HIBCH, FASN, and HADH) display clusters of sites with decreased acetylation (FDR < 0.12, subset to FA pathways). For instance, HADH has an essential role in FA β-oxidation, and the eight significantly reduced acetylation sites are clustered on the 3-hydroxyacyl-coenzyme A (CoA) dehydrogenase NAD-binding domain of the protein; this domain also contains the acetyl CoA-binding sites, which would allow the binding and subsequent oxidative activity of the enzyme[82,83] (Figure 4C, lower). Moreover, we detected a significant cluster of increased phosphorylation sites (S337, T338, and S339, FDR = 0.0026) on the dehydrogenase E1 domain of BCKDH known to catalyze the overall breakdown of alpha-keto acids to acetyl-CoA.[84] This phosphorylation was shown to be mediated by BCKDK and inhibit BCKDH activity, further limiting the levels of acetyl-CoA and increasing FA oxidation to support the cell's energy demand[85] (Figure S4F).

We then performed PTM-SEA to identify the main regulators of PTMs (e.g., kinases or phosphatases; Table S4). This analysis revealed high enrichment of (1) CDK activity in the immune-cool subtype, consistent with the high proliferation associated with this subtype (additionally supported by The Kinase Library enrichment results; Figure S4G) and (2) mTOR activity, a direct regulator of FA metabolism and oxidative phosphorylation as well as an indirect regulator of lipid homeostasis through SREBP1.[86] Moreover, the immune-cool subtype showed an increase of FA uptake, both by transporters within the cell, such as CPT1A (FDR < 7×10^{-7}), and by cell-surface transporters including (1) FABP4, (2) ABCA, and (3) CD36 (FDR < 1.3×10^{-4} immune-cool vs. immune-hot) (Figure S4H).

Recent studies have shown that lipid-enriched tumor microenvironments reduce the cytotoxicity of effector T cells[21,87] since they cannot metabolize long-chain FAs, leading to lipotoxicity and exhaustion.[23] Therefore, we tested the correlation of FA acetylation levels with protein levels of immune-related effectors (Figure 4D). We observed a significant positive correlation between downregulating FA acetylation sites and immune response marker proteins in the interferon gamma (IFNγ) and cell cytotoxic pathways (Figure 4E). Some of the most significant associations were between GZMA and HADHA K214 and K759 (rho −0.3, $p < 1 \times 10^{-13}$) and between CD38 and HADHB K277 and K253 (rho 0.35, $p < 1 \times 10^{-13}$). Moreover, a Spearman correlation of single-sample GSEA (ssGSEA) for CD8+ T cell abundance estimated from CIBERSORT showed FA as the fourth most significantly correlated association after interferon pathways (rho 0.23, FDR = 2×10^{-13}, Figure S4I). The observed associations between PTM-regulated biological processes in cancer cells and their neighboring immune cells are summarized in Figure 4F.

Alterations in histone regulation by PTMs in cancer-associated genes

Here, we leveraged the largest pan-cancer acetylation dataset to comprehensively study histone acetylation and phosphorylation patterns across the six cancer types with available acetylation

Figure 4. PTM regulation of immuno-metabolism across cancers

(A) ssGSEA hierarchical clustering for immune-related genesets (heatmap) showing four immune clusters: hot to cold. Tracks represent ESTIMATE and ImmuneSubtypeClassifier annotations. Normalized Enrichment Score (NES).

(B) Bubble plot representing MSigDB (Molecular Signature Database) hallmark and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways enrichment among the four immune subtypes.

(C) Significant clustering based on CLUMPS-PTM of both acetylation and phosphorylation sites on ALDOA (top; PDB: 6XMH-A) in the immune-hot group, and clustering of decreased acetylation sites on HADH in the immune-cool cluster (lower; PDB: 3RQS-A). Phosphosites (purple), acetylsites (pink), and acetyl CoA binding sites (green).

(D) Volcano plot showing differential acetylation between the immune-cool subtype and the other immune clusters. Acetylation sites on fatty acid metabolism proteins are highlighted.

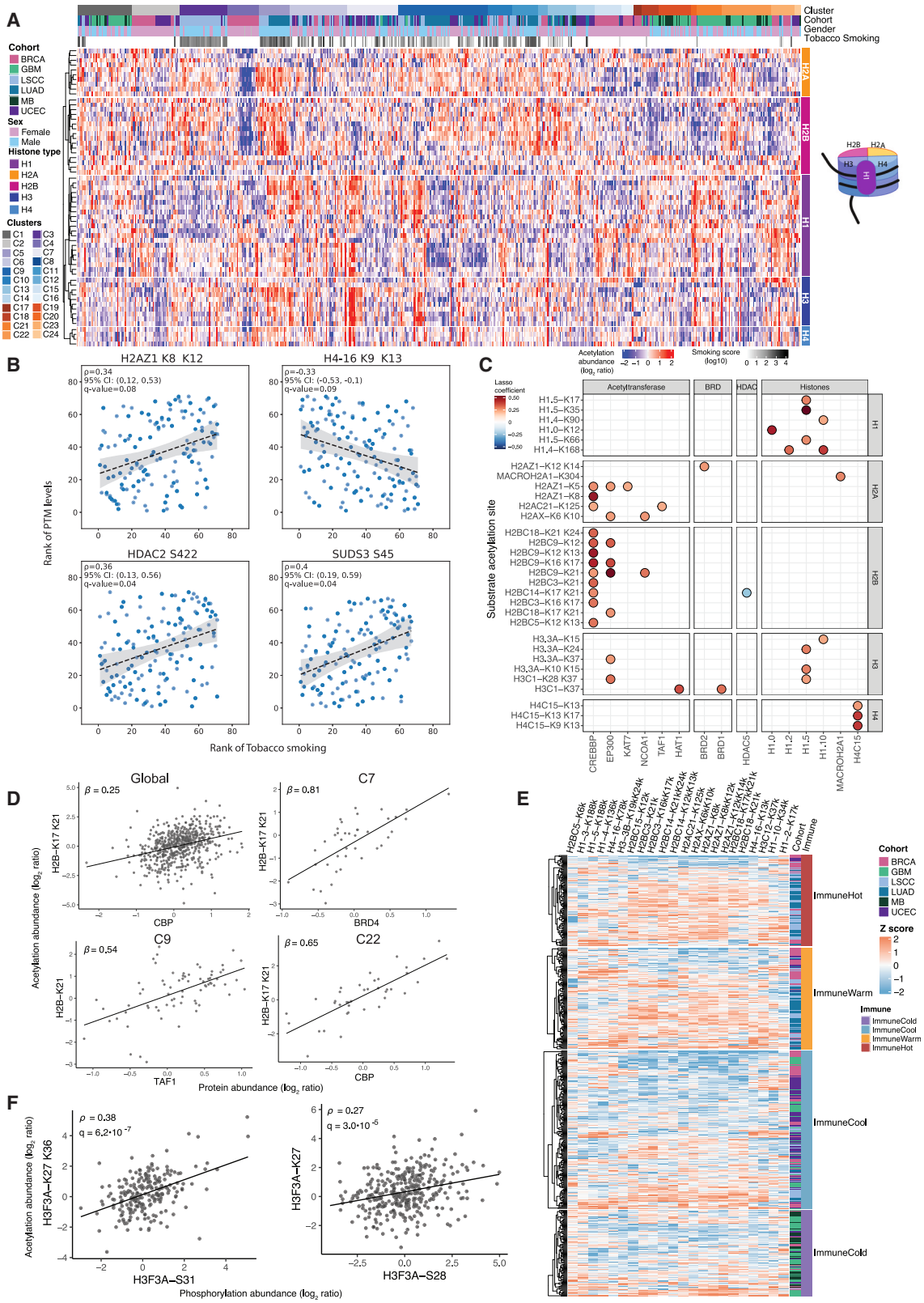
(E) Bubble plot representing significant correlations between fatty acid β-oxidation enzymes acetylation sites and protein levels from the IFNγ pathway.

(F) Schematic representation of PTM-based metabolic changes in immune-cool vs. immune-hot tumors showing key enzymes in the glycolysis and fatty acid β-oxidation pathways and their proposed effect on T cells.

See also Figure S4.

Please cite this article in press as: Geffen et al., Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation, Cell (2023), https://doi.org/10.1016/j.cell.2023.07.013

CellPress
OPEN ACCESS



(legend on next page)

data (STAR Methods). To identify specific histone acetylation patterns, we classified histone-related genes into the following five categories: (1) the linker histone H1, (2) the four core histones H2A (including MACROH2A1, MACROH2A2, H2A.X, and H2AZ1), (3) H2B, (4) H3, and (5) H4 (Figure 5A). We found that histone acetylation partitioned into two structural groups, group 1 (H3, H4, and H1) and group 2 (H2A and H2B), with significant correlations among the mean acetylation profiles of each group for pairs within the groups (all with rho > 0.4 and $p < 2.2 \times 10^{-16}$) and weaker to no correlation for pairs between the two groups ($-0.15 <$ rho < 0.17) (Figure S5A). This is consistent with acetylation state coordination leading to enhanced nucleosome opening and subsequent gene activation.[88,89]

Since tobacco smoking is known to impair histone deacetylase (HDAC) activity and affect histone acetylation,[90] we sought to better characterize these effects by evaluating the correlation between smoking mutational signature and histone acetylation in LUAD tumors, limiting to male patients to decouple effects due to strong association between gender and smoking ($p = 1 \times 10^{-4}$, Figure S5B). We found two positive and two negative significant correlations (FDR $\leq$ 0.1; Figure S5C).[91–93] Among these correlations was the previously described dose-dependent relationship between smoking and H4-16 K9 and K13 acetylation (rho = -0.33, FDR = 0.09). We also found positive correlations for H2AZ1 K8/K12 and H2AZ1 K12/K14 acetylation (rho = 0.34 and 0.39, respectively; FDR = 0.08 and 0.06, respectively) (Figure S5C; Table S6). Acetylation of these sites has been shown to localize H2AZ1 to promoter regions of several cancer genes (e.g., *ERBB3*, *CDK4*, and *RASEF*)[94,95] and facilitate their transcription[96] (Figure 5B). Next, to explore the effect of smoking on HDAC activity, we tested the correlation between smoking and phosphorylation of HDACs. We found six significant (FDR $\leq$ 0.1) positive correlations on four proteins, which are all components of SIN3/HDAC complexes (Figure S5D). Among these correlations was phosphorylation of HDAC2 S422 (rho = 0.36, FDR = 0.04), which was shown to reduce deacetylase activity.[97] Moreover, phosphorylation of this site was mediated by CSNK2A1 kinase upon exposure to cigarette smoke extract.[97]

Following this analysis, we investigated potential transcriptional consequences of smoking-related changes in histone acetylation by correlating the smoking signature and ssGSEA pathway scores (Figure S5E). Among the top significant associations (FDR $\leq$ 0.1 and rho $\geq$ 0.15) was the expected upregulation of G2/M checkpoint genes, consistent with previous studies that associated cigarette smoke and increased proliferation (FDR = 0.01, rho = 0.26, Figure S5F, left).[98,99] We also identified a significant positive correlation with mTOR signaling genes (FDR = 0.01, rho = 0.27, Figure S5F, right), which is consistent with previous studies and may play a role in lung tumorigenesis by altering cell proliferation and metabolism.[100,101]

Next, we investigated the association of key regulators (histone acetyltransferases [HATs], HDACs, and bromodomain proteins [BRDs]; STAR Methods) with histone acetylation levels. We identified multiple positive associations between the protein abundance of histone acetyltransferases CBP/p300 and various acetyl sites, including N-terminal H2B acetyl sites such as K11, K15, K16, and K20 ($0.2 < \beta < 0.52$; Table S5), consistent with the reported substrate specificity of CBP/p300 for these sites (Figure 5C).[102] The protein abundance of the histone acetyltransferase NCOA1, a known co-activator of CBP/p300, was also positively associated with H2BC9-K21 acetylation ($\beta = 0.29$).[102] Additionally, we identified novel associations such as the positive correlation of H3C1 at K36 acetylation with HAT1 ($\beta = 0.36$). HAT1 was shown to promote acetylation of H3 at K14, and its expression was associated with poor prognosis across cancer types.[103,104] We also observed negative associations: HDAC5 protein abundance negatively correlates with H2BC14 K16 and K20 acetylation ($\beta = -0.20$). HDAC5, a known therapeutic target, was shown to have a role in cell differentiation, stemness, and proliferation in several cancer types.[105]

We also tested these associations within our 26 clusters and indeed found cluster-specific correlations (Table S5). For example, CBP showed a low correlation with H2B using all samples ($\beta = 0.25$), but a much stronger correlation when evaluating tumors across cluster 22, which is enriched with brain tumors ($\beta = 0.65$, Figure 5D). Furthermore, we tested the association between histone regulation and cancer hallmark pathways. We found increased acetylation levels of EP300 on known activating sites K1558 and K1560 when comparing the left vs. right side of the rightmost side of the dendrogram's second level (C vs. D, Figure 2A) (FDR = 2.6×10^{-7}),[106] and a concordant increase in the acetylation of the N-terminal H2A and H2B acetylation sites regulated by CBP/p300 (H2AC21-K5K9, FDR = 0.052; H2BC18-K16K20, FDR = 1.3×10^{-4}).[102] Moreover, GSEA showed significant enrichment of E2F and MYC transcription targets in tumors from cluster C (both FDR = 0.08). Consistently, we found a significant positive correlation between H3 acetylation and E2F, MYC, and G2/M checkpoint pathways across the six cohorts, likely reflecting increased transcription associated with cell proliferation. We also observed a significant positive correlation between the acetylation of H3 at K27 and K36 and MTORC1 signaling, consistent with previous studies.[107]

Figure 5. Pan-cancer histone regulation

(A) Heatmap showing site-level acetylation of various histone protein substrates across 6 cohorts. Tracks above show the cohort, cluster assignment, gender, and smoking score.

(B) Scatter plots showing the rankings of site-specific histone PTM levels and tobacco smoking mutational signature activities. 95% confidence intervals of the Spearman's correlation coefficient determined by bootstrapping.

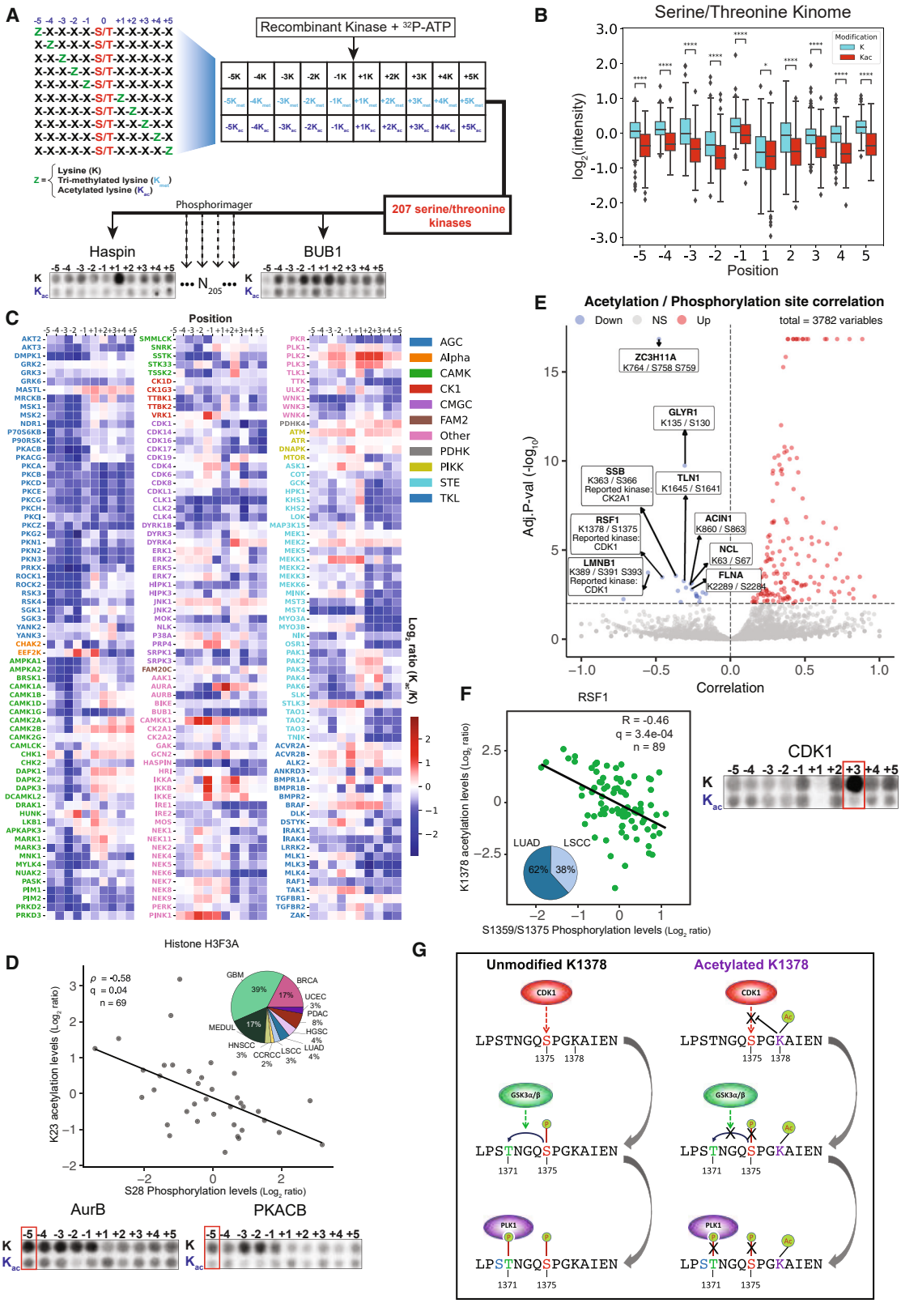
(C) Bubble plot showing pan-cancer associations between key regulators of histone acetylation and histone acetylation sites.

(D) Scatter plots showing the lasso regression associations between histone regulators and H2B acetylation levels across all tumors and in specific clusters in the dendrogram. β represents the lasso regression coefficient.

(E) Heatmap showing the differentially acetylated histone sites in the immune cold subtype compared to all other immune subtypes.

(F) Correlations between histone acetylation sites and close proximity phosphorylation sites.

See also Figure S5.



(legend on next page)

Please cite this article in press as: Geffen et al., Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation, Cell (2023), https://doi.org/10.1016/j.cell.2023.07.013

We then focused on how metabolic shifts across our pan-cancer immune subtypes affect the regulation of histones. Our findings above indicated an increase in FA metabolism in immune-cool tumors (relative to immune-hot), and we also observe decreased acetylation at 22/61 histone acetylation sites relative to immune-hot and at 31/61 sites relative to immune-warm (FDR < 0.1, Figure 5E; Table S5). As previously shown, these results reflect a possible association between histone acetylation and glycolytic flux as well as cellular acetyl-CoA abundance and availability in the different immune clusters.[108]

Finally, we analyzed the correlations between adjacent histone phosphorylation and acetylation sites to better understand their potential crosstalk (STAR Methods). Of the 81 histone acetylsite and phosphosite adjacent pairs (up to five amino acids apart) tested globally, we identified 12 as significantly correlated (FDR < 0.05; Table S5). For instance, H3F3A-S31 phosphorylation was strongly correlated with H3F3A-K27K36 acetylation across all samples (rho = 0.38, FDR = 6.2×10^{-7}), consistent with S31 phosphorylation stimulating H3-K2 acetylation through p300 activity[109] (Figure 5F). In addition, the phosphorylation levels of H3-S28 and acetylation of H3-K27 were positively correlated (rho = 0.27, FDR = 3×10^{-5}), consistent with S28 phosphorylation reducing K27 trimethylation and priming acetylation.[110]

Crosstalk between protein phosphorylation and acetylation in cancer

Motivated by the correlations between phosphorylation and acetylation of adjacent sites in histones, we aimed to systematically analyze this crosstalk across other proteins. Mechanistically, serine/threonine protein kinases are known for their substrate specificity based on the amino acid sequence surrounding their phosphorylation sites.[111–113] We therefore asked whether lysine acetylation adjacent to phosphoacceptor sites can impact their ability to be phosphorylated.

We experimentally characterized lysine-PTM selectivity across 207 recombinant Ser/Thr kinases using degenerate peptide substrates that compared modified and unmodified lysine at the five adjacent amino acid positions in both the N- and C-terminal directions from the phosphoacceptor site[111,113–116] (Figure 6A). Globally, we observed a general selection against substrates containing acetylated lysine across the kinome (Figure 6B). Nevertheless, there are some exceptions in which kinases favor acetylated lysine over unmodified lysine to carry out the phosphorylation of a nearby serine or threonine (Figure 6C). We observed a similar pattern of global selection against trimethylated lysine, but to a smaller extent than acetylation, possibly because trimethylation causes a smaller steric alteration and preserves the positive charge on lysine (Figures S6A and S6B). Together, our screen indicates that Ser/Thr kinases discriminate between the PTM states of the lysines surrounding the phosphoacceptor sites, and that lysine acetylation has the potential to regulate Ser/Thr kinase function.

Using these kinase specificity patterns, we could potentially identify the kinases involved in specific acetylation-phosphorylation crosstalks. Since different kinases may be active in various cell types and states, we first mapped the crosstalk in different dendrogram branches that may each display shared kinase activity due to their similar RNA, protein, and phosphorylation patterns. We searched for potential acetylation-phosphorylation crosstalk in adjacent pairs (up to five amino acids apart) on proteins globally as well as in the dendrogram clusters (STAR Methods). We identified a negative correlation between H3-3A acetylation at K23 and phosphorylation at S28 in a dendrogram branch that includes terminal clusters 22 and 23 (rho = −0.58, FDR = 0.04, Figure 6D, top), suggesting that inhibitory crosstalk is uniquely present in these tumors compared with all tumors (rho = 0.08, FDR = 0.35 globally). Phosphorylation of S28 has been reported to activate transcription,[110] whereas acetylation at K23 inhibits transcription.[118] S28 on H3-3A is a reported substrate for members of the aurora kinase (AURK) and PKA families of kinases.[119,120] Consistent with this finding, our substrate motifs for AurB and PKACB show selection against acetylated lysine in this context (Figure 6D, bottom), indicating that K23 acetylation inhibits the ability of S28 to be phosphorylated by these upstream kinases.

We then explored the acetylation-phosphorylation crosstalk across other proteins and identified 3,952 adjacent pairs among 579 patients (Figure 6E; STAR Methods). Among these, 74 pairs showed a significant negative correlation between their levels (FDR < 0.1). Among the most statistically significant examples (FDR = 3×10^{-4}) was S1375/K1378 on the centromeric protein RSF1 (Figures 6F left and S6D; Table S6). RSF1 is an essential mediator of mitosis known to be overexpressed in many types of cancers.[121] CDK1 has been reported to phosphorylate S1375 on RSF1 during G2/M, which facilitates recruitment of the kinase PLK1 that promotes subsequent mitotic events.[122,123] In our peptide substrate assays, CDK1 strongly

Figure 6. Pan-cancer acetylation and phosphorylation crosstalk

(A) The Kinase Library overview—biochemical assay of a combinatorial peptide library with unmodified, methylated, or acetylated lysine for testing kinases affinity to peptides with modified lysins (indicated as "K") at ±5 positions relative to the Ser/Thr phosphoacceptor residue (excluding serine, threonine, and cysteine); e.g., +3K means a lysine is present 3 amino acids towards the C terminus from the phosphoacceptor residue.

(B) Boxplot showing the average intensity for unmodified and acetylated lysine residues. *p value < 0.05 and ****p value < 0.0001.

(C) Heatmap showing ratio between mean intensities for acetylated and unmodified lysine residues. Kinases are colored according to their phylogenetic groups.[117]

(D) Scatter plot showing the correlation between K23 acetylation levels and S28 phosphorylation levels on histone H3-3A and their cohort distribution (top). Biochemical specificity assays showing AurB and PKACB phosphorylation between unmodified and acetylated peptides (bottom).

(E) Volcano plot showing correlations between pairs of phosphorylation and acetylation sites. Negative correlations are highlighted.

(F) Scatter plot showing the correlation between K1378 acetylation levels and S1375 phosphorylation levels on RSF1 and their cohort distribution (left). Biochemical specificity assays showing CDK1 phosphorylation between unmodified and acetylated peptides (right).

(G) Inhibitory crosstalk proposed mechanism on RSF1.

See also Figure S6.

favored serines/threonines that contained lysines three positions in the C-terminal direction (+3K), matching the known motif for RSF1; moreover, the phosphorylation was almost entirely abolished when the lysine was acetylated (Figure 6F, right), consistent with CDK1's reduced ability to phosphorylate S1375 when K1378 is acetylated, potentially explaining the observed negative correlation.

With our kinome-wide scoring system, we could also infer how phosphorylation of RSF1 facilitates the next step in this signaling cascade and the subsequent recruitment of PLK1 (Figure 6G). PLK1 binds to peptides containing phosphorylated serine or threonine that are directly preceded by an unmodified serine (S-pS/pT)[124] T1371 on RSF1 that matches this pattern. Therefore, when CDK1 phosphorylates S1375,[125,126] the phosphorylated peptide becomes a substrate for a second phosphorylation event by the ubiquitously expressed GSK3 kinases (GSK3 alpha/beta) on T1371, after which the phosphorylated peptide can be recognized by PLK1. Once lysine K1378 is acetylated, the entire cascade is inhibited. Altogether, this can explain how crosstalk between S1375/K1378 on RSF1 ultimately affects the recruitment of PLK1.

DISCUSSION

PTMs are core regulators of signal transduction, and they play major roles in protein-protein interactions, protein stability, and localization, among many other essential functions. In this study, we comprehensively investigated PTMs across 11 cancer types and highlighted the contribution of PTMs to known cancer hallmark processes: (1) DNA repair, (2) immune response, (3) metabolism, (4) histone regulation, and (5) kinase regulation. We noted commonalities of these processes and PTM patterns across cancer types as well as important distinctions. This rich resource will enable additional investigation of PTMs across cancer types beyond the PTM work we describe here.

DNA repair deficiencies, such as HRD and MMRD, generate patterns of somatic mutations throughout tumor development, providing evidence of a given repair deficiency. Importantly, these mutational signatures do not necessarily reflect the current activities of repair pathways, which may be pertinent to understanding the variation in response to therapies that target DNA repair genes (e.g., PARP and POLQ inhibitors, etc.). In-depth analyses of DNA repair-deficient cancers highlighted the ability of phosphorylation-focused analyses to reveal and characterize informative patterns that are undetectable at the genomic and transcriptomic levels. Through our analysis of HRD clusters, we found that significant differences in the phosphorylation of DNA repair proteins were strongly associated with hypoxia severity. We found decreased activity of several DNA repair proteins, including PARP1, in HRD tumors with chronic hypoxia, potentially affecting their response to PARP inhibitors. Our proteogenomic analysis of MMRD tumors linked recurrent RAD50 microsatellite indels with significant decreases in the abundance of the three proteins in the MRN complex, which is crucial for DSB sensing and signaling. Through site-specific phosphorylation analysis, we identified further evidence of DSB repair dysfunction, which may provide additional avenues for developing treatments for MMRD (i.e., MSI) cancers.

In general, immune responses are tightly regulated to tailor each response to a given threat encountered by the host, requiring rapid regulatory changes that can be achieved by PTMs. Similarly, cellular metabolism requires the same flexibility, and PTMs therefore play an essential role in regulating both immune and metabolic responses. As an example, these two processes can be linked with growing evidence that cancer cell regulation of lipid metabolism by PTMs on FA enzymes can have an effect on the immune response.[21–23] In this study, we identified four expression-based immune clusters with diverse metabolic phenotypes driven by acetylation. CLUMPS-PTM highlighted ALDOA, a glycolysis-related protein, which has both significant clusters of altered phosphorylation and acetylation sites in the immune-hot subtype that are associated with increased ALDOA activity. Inhibition of ALDOA in mice reduced lung metastases and prolonged survival.[127] The immune-cool subtype showed an increased activity of FA metabolism that was strongly correlated with reduced IFNγ expression, suggesting an important role of FA in immune suppression. Recent studies demonstrated that inhibition of FA oxidation in acute myeloid leukemia (AML) can restore sensitivity to venetoclax and azacitidine in cells that became resistant.[128,129] These results highlight that targeting lipid metabolism in cancer may not only reduce the ability of tumor cells to produce higher levels of energy but also promote a tumor microenvironment that is more conducive to immune cell infiltration and activation.[128,129] In addition, we identified an association between transcriptionally active metabolic pathways and reduced histone acetylation, potentially due to the reduced availability of acetyl CoA that is used by the cell for both acetylation and producing energy[107]; additional studies will be needed to further investigate this regulation.

Finally, we performed a comprehensive analysis of crosstalk between acetylation and phosphorylation using The Kinase Library.[34] This analysis revealed that most serine/threonine kinases disfavor acetylated lysines in close proximity to the phosphorylation site (shown by significantly negatively correlated pairs of neighboring acetyl-/phosphor-sites), allowing us to predict the likely responsible kinases for the crosstalk.

In summary, PTMs are an integral part of the tumor cell's adaptation and response to intracellular and environmental changes. A deeper understanding of PTM-governed processes leading to cancer initiation and progression has the potential to uncover novel therapeutic targets, identify biomarkers of response to existing therapies, and extend our knowledge of cancer biology.

Limitations of the study

Genomic-based pan-cancer studies have proved to be highly valuable resources for discovering new cancer driver genes and shared dysregulated pathways, as well as for identifying actionable therapeutic targets.[16–18] Our proteogenomic pan-cancer study was limited to 1,110 samples across 11 tumor types, and we anticipate that larger-scale studies, including both more cases and more cancer types, will increase our power to identify proteogenomic mechanisms underlying cancer. The analyses described in this study are all based on bulk tumor material. Similar to single-cell and spatial transcriptomic analyses,[130–133] technologies including single-cell proteomics, spatial proteomics, and laser capture microdissection[134–136] are likely to provide even more

valuable insights into tumor heterogeneity and the contribution of specific cell types to cancer. Although comprehensive PTM-focused research and analyses of PTM crosstalk are emerging fields, some shortcomings are worth noting: (1) current mass spectrometry analyses have relatively high false negative rates that limit our ability to perform crosstalk analyses among PTMs; (2) to fully establish crosstalk relationship between PTMs, a double MS search for simultaneous detection of 2 or more PTMs would be needed; and (3) although there is a growing body of phosphorylation databases and kinase prediction tools, parallel comprehensive acetylation databases and tools are currently lacking, and the functional effects of many of the acetylation sites reported in this study remain to be explored.

CONSORTIA

The members of the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium are François Aguet, Yo Akiyama, Eunkyung An, Shankara Anand, Meenakshi Anurag, Özgün Babur, Jasmin Bavarva, Chet Birger, Michael J. Birrer, Anna Calinawan, Lewis C. Cantley, Song Cao, Steven A. Carr, Michele Ceccarelli, Daniel W. Chan, Arul M. Chinnaiyan, Hanbyul Cho, Shrabanti Chowdhury, Marcin P. Cieslik, Karl R. Clauser, Antonio Colaprico, Daniel Cui Zhou, Felipe da Veiga Leprevost, Corbin Day, Saravana M. Dhanasekaran, Li Ding, Marcin J. Domagalski, Yongchao Dou, Brian J. Druker, Nathan Edwards, Matthew J. Ellis, Myvizhi Esai Selvan, David Fenyö, Steven M. Foltz, Alicia Francis, Yifat Geffen, Gad Getz, Michael A. Gillette, Tania J. Gonzalez Robles, Sara J.C. Gosline, Zeynep H. Gümüş, David I. Heiman, Tara Hiltke, Runyu Hong, Galen Hostetter, Yingwei Hu, Chen Huang, Emily Huntsman, Antonio Iavarone, Eric J. Jaehnig, Scott D. Jewell, Jiayi Ji, Wen Jiang, Jared L. Johnson, Lizabeth Katsnelson, Karen A. Ketchum, Iga Kolodziejczak, Karsten Krug, Chandan Kumar-Sinha, Alexander J. Lazar, Jonathan T. Lei, Yize Li, Wen-Wei Liang, Yuxing Liao, Caleb M. Lindgren, Tao Liu, Wenke Liu, Weiping Ma, D.R. Mani, Fernanda Martins Rodrigues, Wilson McKerrow, Mehdi Mesri, Alexey I. Nesvizhskii, Chelsea J. Newton, Robert Oldroyd, Gilbert S. Omenn, Amanda G. Paulovich, Samuel H. Payne, Francesca Petralia, Pietro Pugliese, Boris Reva, Ana I. Robles, Karin D. Rodland, Henry Rodriguez, Kelly V. Ruggles, Dmitry Rykunov, Shankha Satpathy, Sara R. Savage, Eric E. Schadt, Michael Schnaubelt, Tobias Schraink, Stephan Schürer, Zhiao Shi, Richard D. Smith, Xiaoyu Song, Yizhe Song, Vasileios Stathias, Erik P. Storrs, Jimin Tan, Nadezhda V. Terekhanova, Ratna R. Thangudu, Mathangi Thiagarajan, Nicole Tignor, Joshua M. Wang, Liang-Bo Wang, Pei Wang, Ying Wang, Bo Wen, Maciej Wiznerowicz, Yige Wu, Matthew A. Wyczalkowski, Lijun Yao, Tomer M. Yaron, Xinpei Yi, Bing Zhang, Hui Zhang, Qing Zhang, Xu Zhang, and Zhen Zhang.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human subjects and clinical data
- METHOD DETAILS
  - Genomics Data processing
  - RNAseq data processing and quantification
  - Proteomics data processing
  - Patient Signatures & Clustering
  - Immune clustering
  - Interpretive data analysis
  - Mutational signatures using SignatureAnalyzer
  - Differential Expression
  - fGSEA
  - Dedicated tools for PTM analysis
  - PTM-SEA
  - CLUMPS-PTM
  - The Kinase Library
  - CausalPath
  - Histone analysis
  - SignatureAnalyzer
  - PTM dedicated tools
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2023.07.013.

ACKNOWLEDGMENTS

This work was supported by the National Cancer Institute (NCI) Clinical Proteomic Tumor Analysis Consortium (CPTAC) grants U24CA210955, U24CA210985, U24CA210986, U24CA210954, U24CA210967, U24CA210972, U24CA210979, U24CA210993, U01CA214114, U01CA214116, U24 CA270823, U24CA270823, U01CA271402, U24CA271075, R35CA197588, and U01CA214125. In addition, this project has been funded in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract no. 75N91019D00024, Task Order 75N91020F00029, as well as P01CA206978. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. We thank the Pattern Team at the Broad Institute for figure design. Schematic figures and the graphical abstract were created using BioRender.com.

AUTHOR CONTRIBUTIONS

Study conception and design, Y.G., S.A., F.A., G.G., and L.D.; performed experiment or data collection, Y.G., S.A., Y.A., J.L.J., Y.L., C.B., D.I.H., Q.Z., A.C., F.A., K.R.C., D.R.M., S.B., and M.T.; computational, multi-omic, and statistical analyses, Y.G., S.A., Y.A., T.M.Y., Y.S., J.L.J., A.G., Ö.B., Y.L., E.H., L.-B.W., Y.E.M., N.J.H., A.K., K.K., and F.A.; data interpretation and biological analysis, Y.G., S.A., Y.A., T.M.Y., Y.S., J.L.J., A.G., Y.E.M., N.J.H., F.A., L.C.C., G.G., and L.D.; writing – original drafts, Y.G., S.A., Y.A., M. Miller, T.M.Y., Y.S., J.L.J., A.G., F.A., and G.G.; writing – review & editing, Y.G., S.A., Y.A., M. Miller, T.M.Y., Y.S., J.L.J., A.G., Ö.B., C.B., D.I.H., Y.E.M., N.J.H., K.R.C., S.S., S.H.P., M.A.G., S.M.D., A.I.R., S.A.C., A.J.L., F.A., L.C.C., L.D., and G.G.; supervision, Y.G., A.I.R., F.A., L.C.C., L.D., and G.G.; administration, Y.G., M. Mesri, H.R., A.I.R., L.D., and G.G.

DECLARATION OF INTERESTS

Y.G. is a consultant for Oriel Research Therapeutics. T.M.Y. is a co-founder, stockholder, and on the board of directors of DESTROKE, Inc., an early-stage start-up developing mobile technology for automated clinical stroke detection. J.L.J. has received consulting fees from Scorpion Therapeutics and Volastra Therapeutics. Y.E.M. is a consultant for ForseeGenomics and is also an inventor on patent applications filed by the Broad Institute related to MSMuTect, MSMutSig, and MSIDetect. N.J.H. is a consultant for MorphoSys. F.A. is an inventor on a patent application related to SignatureAnalyzer-GPU and has been an employee of Illumina, Inc., since 8 November 2021. L.C.C. is a founder and member of the board of directors of Agios Pharmaceuticals and is a founder of Petra Pharmaceuticals. L.C.C. is an inventor on patents (pending) for Combination Therapy for PI3K-associated Disease or Disorder, and The Identification of Therapeutic Interventions to Improve Response to PI3K Inhibitors for Cancer Treatment. L.C.C. is a co-founder and shareholder in Faeth Therapeutics. G.G. receives research funds from IBM, Pharmacyclics, and Ultima Genomics, and is also an inventor on patent applications filed by the Broad Institute related to MSMuTect, MSMutSig, POLYSOLVER, SignatureAnalyzer-GPU, MSIDetect, and MinimuMM-Seq. He is also a founder, consultant, and privately held equity in Scorpion Therapeutics.

Received: July 12, 2022
Revised: January 6, 2023
Accepted: July 10, 2023
Published: August 14, 2023

REFERENCES

1. Ding, L., Bailey, M.H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., Gibbs, D.L., Weerasinghe, A., Huang, K.-L., Tokheim, C., et al. (2018). Perspective on oncogenic processes at the end of the beginning of cancer genomics. Cell *173*, 305–320.e10. https://doi.org/10.1016/j.cell.2018.03.033.
2. Doroshow, D.B., and Doroshow, J.H. (2020). Genomics and the history of precision oncology. Surg. Oncol. Clin. N. Am. *29*, 35–49. https://doi.org/10.1016/j.soc.2019.08.003.
3. Rodriguez, H., Zenklusen, J.C., Staudt, L.M., Doroshow, J.H., and Lowy, D.R. (2021). The next horizon in precision oncology: proteogenomics to inform cancer diagnosis and treatment. Cell *184*, 1661–1670. https://doi.org/10.1016/j.cell.2021.02.055.
4. Wang, L.-B., Karpova, A., Gritsenko, M.A., Kyle, J.E., Cao, S., Li, Y., Rykunov, D., Colaprico, A., Rothstein, J.H., Hong, R., et al. (2021). Proteogenomic and metabolomic characterization of human glioblastoma. Cancer Cell *39*, 509–528.e20. https://doi.org/10.1016/j.ccell.2021.01.006.
5. Huang, C., Chen, L., Savage, S.R., Eguez, R.V., Dou, Y., Li, Y., da Veiga Leprevost, F., Jaehnig, E.J., Lei, J.T., Wen, B., et al. (2021). Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. Cancer Cell *39*, 361–379.e16. https://doi.org/10.1016/j.ccell.2020.12.007.
6. Gillette, M.A., Satpathy, S., Cao, S., Dhanasekaran, S.M., Vasaikar, S.V., Krug, K., Petralia, F., Li, Y., Liang, W.-W., Reva, B., et al. (2020). Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. Cell *182*, 200–225.e35. https://doi.org/10.1016/j.cell.2020.06.013.
7. Satpathy, S., Krug, K., Jean Beltran, P.M., Savage, S.R., Petralia, F., Kumar-Sinha, C., Dou, Y., Reva, B., Kane, M.H., Avanessian, S.C., et al. (2021). A proteogenomic portrait of lung squamous cell carcinoma. Cell *184*, 4348–4371.e40. https://doi.org/10.1016/j.cell.2021.07.016.
8. Krug, K., Jaehnig, E.J., Satpathy, S., Blumenberg, L., Karpova, A., Anurag, M., Miles, G., Mertins, P., Geffen, Y., Tang, L.C., et al. (2020). Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. Cell *183*, 1436–1456.e31. https://doi.org/10.1016/j.cell.2020.10.036.
9. Cao, L., Huang, C., Cui Zhou, D., Hu, Y., Lih, T.M., Savage, S.R., Krug, K., Clark, D.J., Schnaubelt, M., Chen, L., et al. (2021). Proteogenomic characterization of pancreatic ductal adenocarcinoma. Cell *184*, 5031–5052.e26. https://doi.org/10.1016/j.cell.2021.08.023.
10. Clark, D.J., Dhanasekaran, S.M., Petralia, F., Pan, J., Song, X., Hu, Y., da Veiga Leprevost, F., Reva, B., Lih, T.-S.M., Chang, H.-Y., et al. (2019). Integrated proteogenomic characterization of clear cell renal cell carcinoma. Cell *179*, 964–983.e31. https://doi.org/10.1016/j.cell.2019.10.007.
11. McDermott, J.E., Arshad, O.A., Petyuk, V.A., Fu, Y., Gritsenko, M.A., Clauss, T.R., Moore, R.J., Schepmoes, A.A., Zhao, R., Monroe, M.E., et al. (2020). Proteogenomic characterization of ovarian HGSC implicates mitotic kinases, replication stress in observed chromosomal instability. Cell Rep. Med. *1*, 100004. https://doi.org/10.1016/j.xcrm.2020.100004.
12. Dou, Y., Kawaler, E.A., Cui Zhou, D., Gritsenko, M.A., Huang, C., Blumenberg, L., Karpova, A., Petyuk, V.A., Savage, S.R., Satpathy, S., et al. (2020). Proteogenomic characterization of endometrial carcinoma. Cell *180*, 729–748.e26. https://doi.org/10.1016/j.cell.2020.01.026.
13. Vasaikar, S., Huang, C., Wang, X., Petyuk, V.A., Savage, S.R., Wen, B., Dou, Y., Zhang, Y., Shi, Z., Arshad, O.A., et al. (2019). Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. Cell *177*, 1035–1049.e19. https://doi.org/10.1016/j.cell.2019.03.030.
14. Mani, D.R., Krug, K., Zhang, B., Satpathy, S., Clauser, K.R., Ding, L., Ellis, M., Gillette, M.A., and Carr, S.A. (2022). Cancer proteogenomics: current impact and future prospects. Nat. Rev. Cancer *22*, 298–313. https://doi.org/10.1038/s41568-022-00446-5.
15. Deribe, Y.L., Pawson, T., and Dikic, I. (2010). Post-translational modifications in signal integration. Nat. Struct. Mol. Biol. *17*, 666–672. https://doi.org/10.1038/nsmb.1842.
16. Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W.K., Luna, A., La, K.C., Dimitriadoy, S., Liu, D.L., Kantheti, H.S., Saghafinia, S., et al. (2018). Oncogenic signaling pathways in The Cancer Genome Atlas. Cell *173*, 321–337.e10. https://doi.org/10.1016/j.cell.2018.03.035.
17. Bailey, M.H., Tokheim, C., Porta-Pardo, E., SenGupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. Cell *173*, 371–385.e18. https://doi.org/10.1016/j.cell.2018.02.060.
18. Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674. https://doi.org/10.1016/j.cell.2011.02.013.
19. Narita, T., Weinert, B.T., and Choudhary, C. (2019). Functions and mechanisms of non-histone protein acetylation. Nat. Rev. Mol. Cell Biol. *20*, 156–174. https://doi.org/10.1038/s41580-018-0081-3.
20. Biswas, S.K. (2015). Metabolic reprogramming of immune cells in cancer progression. Immunity *43*, 435–449. https://doi.org/10.1016/j.immuni.2015.09.001.
21. Yu, W., Lei, Q., Yang, L., Qin, G., Liu, S., Wang, D., Ping, Y., and Zhang, Y. (2021). Contradictory roles of lipid metabolism in immune response within the tumor microenvironment. J. Hematol. Oncol. *14*, 187. https://doi.org/10.1186/s13045-021-01200-4.
22. Rohatgi, N., Ghoshdastider, U., Baruah, P., Kulshrestha, T., and Skanderup, A.J. (2022). A pan-cancer metabolic atlas of the tumor microenvironment. Cell Rep. *39*, 110800. https://doi.org/10.1016/j.celrep.2022.110800.
23. Manzo, T., Prentice, B.M., Anderson, K.G., Raman, A., Schalck, A., Codreanu, G.S., Nava Lauson, C.B., Tiberti, S., Raimondi, A., Jones, M.A., et al. (2020). Accumulation of long-chain fatty acids in the tumor microenvironment drives dysfunction in intrapancreatic CD8+ T cells. J. Exp. Med. *217*, e20191920. https://doi.org/10.1084/jem.20191920.
24. Peleg, S., Feller, C., Ladurner, A.G., and Imhof, A. (2016). The metabolic impact on histone acetylation and transcription in ageing. Trends Biochem. Sci. *41*, 700–711. https://doi.org/10.1016/j.tibs.2016.05.008.

25. Audia, J.E., and Campbell, R.M. (2016). Histone modifications and cancer. Cold Spring Harb. Perspect. Biol. *8*, a019521. https://doi.org/10.1101/cshperspect.a019521.

26. Huen, M.S.Y., and Chen, J. (2008). The DNA damage response pathways: at the crossroad of protein modifications. Cell Res. *18*, 8–16. https://doi.org/10.1038/cr.2007.109.

27. Li, Y., Dou, Y., Da Veiga Leprevost, F., Geffen, Y., Calinawan, A.P., Aguet, F., Akiyama, Y., Anand, S., Birger, C., Cao, S., et al. (2023). Proteogenomic data and resources for pan-cancer analysis. Cancer Cell 41, ■.

28. Archer, T.C., Ehrenberger, T., Mundt, F., Gold, M.P., Krug, K., Mah, C.K., Mahoney, E.L., Daniel, C.J., LeNail, A., Ramamoorthy, D., et al. (2018). Proteomics, post-translational modifications, and integrative analyses reveal molecular heterogeneity within medulloblastoma subgroups. Cancer Cell *34*, 396–410.e8. https://doi.org/10.1016/j.ccell.2018.08.004.

29. Karabulut, N.P., and Frishman, D. (2016). Sequence- and structure-based analysis of tissue-specific phosphorylation sites. PLoS One *11*, e0157896. https://doi.org/10.1371/journal.pone.0157896.

30. Garcia, B.A., Thomas, C.E., Kelleher, N.L., and Mizzen, C.A. (2008). Tissue-specific expression and post-translational modification of histone H3 variants. J. Proteome Res. *7*, 4225–4236. https://doi.org/10.1021/pr800044q.

31. Kim, J., Mouw, K.W., Polak, P., Braunstein, L.Z., Kamburov, A., Kwiatkowski, D.J., Rosenberg, J.E., Van Allen, E.M., D'Andrea, A., and Getz, G. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nat. Genet. *48*, 600–606. https://doi.org/10.1038/ng.3557.

32. Kasar, S., Kim, J., Improgo, R., Tiao, G., Polak, P., Haradhvala, N., Lawrence, M.S., Kiezun, A., Fernandes, S.M., Bahl, S., et al. (2015). Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. Nat. Commun. *6*, 8866. https://doi.org/10.1038/ncomms9866.

33. Taylor-Weiner, A., Aguet, F., Haradhvala, N.J., Gosai, S., Anand, S., Kim, J., Ardlie, K., Van Allen, E.M., and Getz, G. (2019). Scaling computational genomics to millions of individuals with GPUs. Genome Biol. *20*, 228. https://doi.org/10.1186/s13059-019-1836-7.

34. Johnson, J.L., Yaron, T.M., Huntsman, E.M., Kerelsky, A., Song, J., Regev, A., Lin, T.Y., Liberatore, K., Cizin, D.M., Cohen, B.M., et al. (2023). An atlas of substrate specificities for the human serine/threonine kinome. Nature *613*, 759–766. https://doi.org/10.1038/s41586-022-05575-3.

35. Babur, Ö., Luna, A., Korkut, A., Durupinar, F., Siper, M.C., Dogrusoz, U., Vaca Jacome, A.S., Peckner, R., Christianson, K.E., Jaffe, J.D., et al. (2021). Causal interactions from proteomic profiles: molecular data meet pathway knowledge. Patterns (N Y) *2*, 100257. https://doi.org/10.1016/j.patter.2021.100257.

36. Krug, K., Mertins, P., Zhang, B., Hornbeck, P., Raju, R., Ahmad, R., Szucs, M., Mundt, F., Forestier, D., Jane-Valbuena, J., et al. (2019). A curated resource for phosphosite-specific signature analysis. Mol. Cell. Proteomics *18*, 576–593. https://doi.org/10.1074/mcp.TIR118.000943.

37. Dong, Y., Sun, Y., Huang, Y., Fang, X., Sun, P., Dwarakanath, B., Kong, L., and Lu, J.J. (2019). Depletion of MLKL inhibits invasion of radioresistant nasopharyngeal carcinoma cells by suppressing epithelial-mesenchymal transition. Ann. Transl. Med. *7*, 741. https://doi.org/10.21037/atm.2019.11.104.

38. Edmond, V., Merdzhanova, G., Gout, S., Brambilla, E., Gazzeri, S., and Eymin, B. (2013). A new function of the splicing factor SRSF2 in the control of E2F1-mediated cell cycle progression in neuroendocrine lung tumors. Cell Cycle *12*, 1267–1278. https://doi.org/10.4161/cc.24363.

39. Kadoch, C., Hargreaves, D.C., Hodges, C., Elias, L., Ho, L., Ranish, J., and Crabtree, G.R. (2013). Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. Nat. Genet. *45*, 592–601. https://doi.org/10.1038/ng.2628.

40. Sun, X., Wang, S.C., Wei, Y., Luo, X., Jia, Y., Li, L., Gopal, P., Zhu, M., Nassour, I., Chuang, J.-C., et al. (2017). Arid1a has context-dependent oncogenic and tumor suppressor functions in liver cancer. Cancer Cell *32*, 574–589.e6. https://doi.org/10.1016/j.ccell.2017.10.007.

41. Xu, S., and Tang, C. (2021). The role of ARID1A in tumors: tumor initiation or tumor suppression? Front. Oncol. *11*, 745187. https://doi.org/10.3389/fonc.2021.745187.

42. Akimov, V., Barrio-Hernandez, I., Hansen, S.V.F., Hallenborg, P., Pedersen, A.-K., Bekker-Jensen, D.B., Puglia, M., Christensen, S.D.K., Vanselow, J.T., Nielsen, M.M., et al. (2018). UbiSite approach for comprehensive mapping of lysine and N-terminal ubiquitination sites. Nat. Struct. Mol. Biol. *25*, 631–640. https://doi.org/10.1038/s41594-018-0084-y.

43. Haradhvala, N.J., Polak, P., Stojanov, P., Covington, K.R., Shinbrot, E., Hess, J.M., Rheinbay, E., Kim, J., Maruvka, Y.E., Braunstein, L.Z., et al. (2016). Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. Cell *164*, 538–549. https://doi.org/10.1016/j.cell.2015.12.050.

44. Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al. (2020). The repertoire of mutational signatures in human cancer. Nature *578*, 94–101. https://doi.org/10.1038/s41586-020-1943-3.

45. Polak, P., Kim, J., Braunstein, L.Z., Karlic, R., Haradhavala, N.J., Tiao, G., Rosebrock, D., Livitz, D., Kübler, K., Mouw, K.W., et al. (2017). A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. Nat. Genet. *49*, 1476–1486. https://doi.org/10.1038/ng.3934.

46. Degasperi, A., Zou, X., Amarante, T.D., Martinez-Martinez, A., Koh, G.C.C., Dias, J.M.L., Heskin, L., Chmelova, L., Rinaldi, G., Wang, V.Y.W., et al. (2022). Substitution mutational signatures in whole-genome-sequenced cancers in the UK population. Science *376*, abl9283. https://doi.org/10.1126/science.abl9283.

47. Polo, S.E., and Jackson, S.P. (2011). Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications. Genes Dev. *25*, 409–433. https://doi.org/10.1101/gad.2021311.

48. Wang, H., and Xu, X. (2017). Microhomology-mediated end joining: new players join the team. Cell Biosci. *7*, 6. https://doi.org/10.1186/s13578-017-0136-8.

49. Sfeir, A., and Symington, L.S. (2015). Microhomology-mediated end joining: A back-up survival mechanism or dedicated pathway? Trends Biochem. Sci. *40*, 701–714. https://doi.org/10.1016/j.tibs.2015.08.006.

50. Li, Z., Wang-Heaton, H., Cartwright, B.M., Makinwa, Y., Hilton, B.A., Musich, P.R., Shkriabai, N., Kvaratskhelia, M., Guan, S., Chen, Q., et al. (2021). ATR prevents Ca2+ overload-induced necrotic cell death through phosphorylation-mediated inactivation of PARP1 without DNA damage signaling. FASEB J. *35*, e21373. https://doi.org/10.1096/fj.202001636RRR.

51. Gupte, R., Liu, Z., and Kraus, W.L. (2017). PARPs and ADP-ribosylation: recent advances linking molecular functions to biological outcomes. Genes Dev. *31*, 101–126. https://doi.org/10.1101/gad.291518.116.

52. Brunyanszki, A., Olah, G., Coletta, C., Szczesny, B., and Szabo, C. (2014). Regulation of mitochondrial poly(ADP-ribose) polymerase activation by the β-adrenoceptor/cAMP/protein kinase A axis during oxidative stress. Mol. Pharmacol. *86*, 450–462. https://doi.org/10.1124/mol.114.094318.

53. Zatreanu, D., Robinson, H.M.R., Alkhatib, O., Boursier, M., Finch, H., Geo, L., Grande, D., Grinkevich, V., Heald, R.A., Langdon, S., et al. (2021). Polθ inhibitors elicit BRCA-gene synthetic lethality and target PARP inhibitor resistance. Nat. Commun. *12*, 3636. https://doi.org/10.1038/s41467-021-23463-8.

54. Ceccaldi, R., Liu, J.C., Amunugama, R., Hajdu, I., Primack, B., Petalcorin, M.I.R., O'Connor, K.W., Konstantinopoulos, P.A., Elledge, S.J., Boulton, S.J., et al. (2015). Homologous-recombination-deficient tumours are dependent on Polθ-mediated repair. Nature *518*, 258–262. https://doi.org/10.1038/nature14184.

55. Bolderson, E., Tomimatsu, N., Richard, D.J., Boucher, D., Kumar, R., Pandita, T.K., Burma, S., and Khanna, K.K. (2010). Phosphorylation of Exo1 modulates homologous recombination repair of DNA double-strand breaks. Nucleic Acids Res. *38*, 1821–1831. https://doi.org/10.1093/nar/gkp1164.

56. Bindra, R.S., Crosby, M.E., and Glazer, P.M. (2007). Regulation of DNA repair in hypoxic cancer cells. Cancer Metastasis Rev. *26*, 249–260. https://doi.org/10.1007/s10555-007-9061-3.

57. Ng, N., Purshouse, K., Foskolou, I.P., Olcina, M.M., and Hammond, E.M. (2018). Challenges to DNA replication in hypoxic conditions. FEBS Journal *285*, 1563–1571. https://doi.org/10.1111/febs.14377.

58. Pires, I.M., Bencokova, Z., Milani, M., Folkes, L.K., Li, J.-L., Stratford, M.R., Harris, A.L., and Hammond, E.M. (2010). Effects of acute versus chronic hypoxia on DNA damage responses and genomic instability. Cancer Res. *70*, 925–935. https://doi.org/10.1158/0008-5472.CAN-09-2715.

59. Chan, N., Koritzinsky, M., Zhao, H., Bindra, R., Glazer, P.M., Powell, S., Belmaaza, A., Wouters, B., and Bristow, R.G. (2008). Chronic hypoxia decreases synthesis of homologous recombination proteins to offset chemoresistance and radioresistance. Cancer Res. *68*, 605–614. https://doi.org/10.1158/0008-5472.CAN-07-5472.

60. Kim, I.-K., Stegeman, R.A., Brosey, C.A., and Ellenberger, T. (2015). A quantitative assay reveals ligand specificity of the DNA scaffold repair protein XRCC1 and efficient disassembly of complexes of XRCC1 and the poly(ADP-ribose) polymerase 1 by poly(ADP-ribose) glycohydrolase. J. Biol. Chem. *290*, 3775–3783. https://doi.org/10.1074/jbc.M114.624718.

61. Hegde, M.L., Izumi, T., and Mitra, S. (2012). Oxidized base damage and single-strand break repair in mammalian genomes: role of disordered regions and posttranslational modifications in early enzymes. Prog. Mol. Biol. Transl. Sci. *110*, 123–153. https://doi.org/10.1016/B978-0-12-387665-2.00006-7.

62. Zheng, F., Zhang, Y., Chen, S., Weng, X., Rao, Y., and Fang, H. (2020). Mechanism and current progress of poly ADP-ribose polymerase (PARP) inhibitors in the treatment of ovarian cancer. Biomed. Pharmacother. *123*, 109661. https://doi.org/10.1016/j.biopha.2019.109661.

63. Rose, M., Burgess, J.T., O'Byrne, K., Richard, D.J., and Bolderson, E. (2020). PARP inhibitors: clinical relevance, mechanisms of action and tumor resistance. Front. Cell Dev. Biol. *8*, 564601. https://doi.org/10.3389/fcell.2020.564601.

64. Murata, M.M., Kong, X., Moncada, E., Chen, Y., Imamura, H., Wang, P., Berns, M.W., Yokomori, K., and Digman, M.A. (2019). NAD^+ consumption by PARP1 in response to DNA damage triggers metabolic shift critical for damaged cell survival. Mol. Biol. Cell *30*, 2584–2597. https://doi.org/10.1091/mbc.E18-10-0650.

65. Palermo, V., Rinalducci, S., Sanchez, M., Grillini, F., Sommers, J.A., Brosh, R.M., Jr., Zolla, L., Franchitto, A., and Pichierri, P. (2016). CDK1 phosphorylates WRN at collapsed replication forks. Nat. Commun. *7*, 12880. https://doi.org/10.1038/ncomms12880.

66. Bian, L., Meng, Y., Zhang, M., and Li, D. (2019). MRE11-RAD50-NBS1 complex alterations and DNA damage response: implications for cancer treatment. Mol. Cancer *18*, 169. https://doi.org/10.1186/s12943-019-1100-5.

67. Ikenoue, T., Togo, G., Nagai, K., Ijichi, H., Kato, J., Yamaji, Y., Okamoto, M., Kato, N., Kawabe, T., Tanaka, A., et al. (2001). Frameshift mutations at mononucleotide repeats in RAD50 recombinational DNA repair gene in colorectal cancers with microsatellite instability. Jpn. J. Cancer Res. *92*, 587–591. https://doi.org/10.1111/j.1349-7006.2001.tb01134.x.

68. Alemayehu, A., and Fridrichova, I. (2007). The MRE11/RAD50/NBS1 complex destabilization in Lynch-syndrome patients. Eur. J. Hum. Genet. *15*, 922–929. https://doi.org/10.1038/sj.ejhg.5201858.

69. Neal, J.A., Dunger, K., Geith, K., and Meek, K. (2020). Deciphering the role of distinct DNA-PK phosphorylations at collapsed replication forks. DNA Repair *94*, 102925. https://doi.org/10.1016/j.dnarep.2020.102925.

70. Zolner, A.E., Abdou, I., Ye, R., Mani, R.S., Fanta, M., Yu, Y., Douglas, P., Tahbaz, N., Fang, S., Dobbs, T., et al. (2011). Phosphorylation of polynucleotide kinase/ phosphatase by DNA-dependent protein kinase and Ataxia-telangiectasia mutated regulates its association with sites of DNA damage. Nucleic Acids Res. *39*, 9224–9237. https://doi.org/10.1093/nar/gkr647.

71. Weinfeld, M., Mani, R.S., Abdou, I., Aceytuno, R.D., and Glover, J.N.M. (2011). Tidying up loose ends: the role of polynucleotide kinase/phosphatase in DNA strand break repair. Trends Biochem. Sci. *36*, 262–271. https://doi.org/10.1016/j.tibs.2011.01.006.

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

73. Gibbs, D.L. (2020). Robust classification of Immune Subtypes in Cancer. https://doi.org/10.1101/2020.01.17.910950.

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

75. Sevenich, L. (2019). Turning "cold" into "hot" tumors-opportunities and challenges for radio-immunotherapy against primary and metastatic brain cancers. Front. Oncol. *9*, 163. https://doi.org/10.3389/fonc.2019.00163.

76. DeBerardinis, R.J., and Chandel, N.S. (2016). Fundamentals of cancer metabolism. Sci. Adv. *2*, e1600200. https://doi.org/10.1126/sciadv.1600200.

77. Hitosugi, T., and Chen, J. (2014). Post-translational modifications and the Warburg effect. Oncogene *33*, 4279–4285. https://doi.org/10.1038/onc.2013.406.

78. Pégorier, J.-P., Le May, C., and Girard, J. (2004). Control of gene expression by fatty acids. J. Nutr. *134*, 2444S–2449S. https://doi.org/10.1093/jn/134.9.2444S.

79. Nusinow, D.P., Szpyt, J., Ghandi, M., Rose, C.M., McDonald, E.R., 3rd, Kalocsay, M., Jané-Valbuena, J., Gelfand, E., Schweppe, D.K., Jedrychowski, M., et al. (2020). Quantitative proteomics of the cancer cell line encyclopedia. Cell *180*, 387–402.e16. https://doi.org/10.1016/j.cell.2019.12.023.

80. Schönfeld, P., and Reiser, G. (2013). Why does brain metabolism not favor burning of fatty acids to provide energy? Reflections on disadvantages of the use of free fatty acids as fuel for brain. J. Cereb. Blood Flow Metab. *33*, 1493–1499. https://doi.org/10.1038/jcbfm.2013.128.

81. Gizak, A., Wiśniewski, J., Heron, P., Mamczur, P., Sygusch, J., and Rakus, D. (2019). Targeting a moonlighting function of aldolase induces apoptosis in cancer cells. Cell Death Dis. *10*, 712. https://doi.org/10.1038/s41419-019-1968-4.

82. Pan, A., Sun, X.-M., Huang, F.-Q., Liu, J.-F., Cai, Y.-Y., Wu, X., Alolga, R.N., Li, P., Liu, B.-L., Liu, Q., et al. (2022). The mitochondrial β-oxidation enzyme HADHA restrains hepatic glucagon response by promoting β-hydroxybutyrate production. Nat. Commun. *13*, 386. https://doi.org/10.1038/s41467-022-28044-x.

83. Yang, J.H., Kim, N.H., Yun, J.S., Cho, E.S., Cha, Y.H., Cho, S.B., Lee, S.-H., Cha, S.Y., Kim, S.-Y., Choi, J., et al. (2020). Snail augments fatty acid oxidation by suppression of mitochondrial ACC2 during cancer progression. Life Sci. Alliance *3*, 3. https://doi.org/10.26508/lsa.202000683.

84. Peng, H., Wang, Y., and Luo, W. (2020). Multifaceted role of branched-chain amino acid metabolism in cancer. Oncogene *39*, 6747–6756. https://doi.org/10.1038/s41388-020-01480-z.

85. White, P.J., McGarrah, R.W., Grimsrud, P.A., Tso, S.-C., Yang, W.-H., Haldeman, J.M., Grenier-Larouche, T., An, J., Lapworth, A.L., Astapova, I., et al. (2018). The BCKDH kinase and phosphatase integrate BCAA and

lipid metabolism via regulation of ATP-citrate lyase. Cell Metab. *27*, 1281–1293.e7. https://doi.org/10.1016/j.cmet.2018.04.015.

86. Koundouros, N., and Poulogiannis, G. (2020). Reprogramming of fatty acid metabolism in cancer. Br. J. Cancer *122*, 4–22. https://doi.org/10.1038/s41416-019-0650-z.

87. Gubser, P.M., Bantug, G.R., Razik, L., Fischer, M., Dimeloe, S., Hoenger, G., Durovic, B., Jauch, A., and Hess, C. (2013). Rapid effector function of memory CD8+ T cells requires an immediate-early glycolytic switch. Nat. Immunol. *14*, 1064–1072. https://doi.org/10.1038/ni.2687.

88. Furukawa, A., Wakamori, M., Arimura, Y., Ohtomo, H., Tsunaka, Y., Kurumizaka, H., Umehara, T., and Nishimura, Y. (2020). Acetylated histone H4 tail enhances histone H3 tail acetylation by altering their mutual dynamics in the nucleosome. Proc. Natl. Acad. Sci. USA *117*, 19661–19663. https://doi.org/10.1073/pnas.2010506117.

89. Hao, F., Murphy, K.J., Kujirai, T., Kamo, N., Kato, J., Koyama, M., Okamato, A., Hayashi, G., Kurumizaka, H., and Hayes, J.J. (2020). Acetylation-modulated communication between the H3 N-terminal tail domain and the intrinsically disordered H1 C-terminal domain. Nucleic Acids Res. *48*, 11510–11520. https://doi.org/10.1093/nar/gkaa949.

90. Chen, D., Fang, L., Li, H., Tang, M.-S., and Jin, C. (2013). Cigarette smoke component acrolein modulates chromatin assembly by inhibiting histone acetylation. J. Biol. Chem. *288*, 21678–21687. https://doi.org/10.1074/jbc.M113.476630.

91. Sundar, I.K., Nevid, M.Z., Friedman, A.E., and Rahman, I. (2014). Cigarette smoke induces distinct histone modifications in lung cells: implications for the pathogenesis of COPD and lung cancer. J. Proteome Res. *13*, 982–996. https://doi.org/10.1021/pr400998n.

92. Van Den Broeck, A., Brambilla, E., Moro-Sibilot, D., Lantuejoul, S., Brambilla, C., Eymin, B., and Gazzeri, S. (2008). Loss of histone H4K20 trimethylation occurs in preneoplasia and influences prognosis of non-small cell lung cancer. Clin. Cancer Res. *14*, 7237–7245. https://doi.org/10.1158/1078-0432.CCR-08-0869.

93. Liu, F., Killian, J.K., Yang, M., Walker, R.L., Hong, J.A., Zhang, M., Davis, S., Zhang, Y., Hussain, M., Xi, S., et al. (2010). Epigenomic alterations and gene expression profiles in respiratory epithelia exposed to cigarette smoke condensate. Oncogene *29*, 3650–3664. https://doi.org/10.1038/onc.2010.129.

94. Oshita, H., Nishino, R., Takano, A., Fujitomo, T., Aragaki, M., Kato, T., Akiyama, H., Tsuchiya, E., Kohno, N., Nakamura, Y., et al. (2013). RASEF is a novel diagnostic biomarker and a therapeutic target for lung cancer. Mol. Cancer Res. *11*, 937–951. https://doi.org/10.1158/1541-7786.MCR-12-0685-T.

95. Wu, A., Wu, B., Guo, J., Luo, W., Wu, D., Yang, H., Zhen, Y., Yu, X., Wang, H., Zhou, Y., et al. (2011). Elevated expression of CDK4 in lung cancer. J. Transl. Med. *9*, 38. https://doi.org/10.1186/1479-5876-9-38.

96. Valdés-Mora, F., Song, J.Z., Statham, A.L., Strbenac, D., Robinson, M.D., Nair, S.S., Patterson, K.I., Tremethick, D.J., Stirzaker, C., and Clark, S.J. (2012). Acetylation of H2A.Z is a key epigenetic modification associated with gene deregulation and epigenetic remodeling in cancer. Genome Res. *22*, 307–321. https://doi.org/10.1101/gr.118919.110.

97. Adenuga, D., and Rahman, I. (2010). Protein kinase CK2-mediated phosphorylation of HDAC2 regulates co-repressor formation, deacetylase activity and acetylation of HDAC2 by cigarette smoke and aldehydes. Arch. Biochem. Biophys. *498*, 62–73. https://doi.org/10.1016/j.abb.2010.04.002.

98. Ho, Y.-S., Chen, C.-H., Wang, Y.-J., Pestell, R.G., Albanese, C., Chen, R.-J., Chang, M.-C., Jeng, J.-H., Lin, S.-Y., Liang, Y.-C., et al. (2005). Tobacco-specific carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) induces cell proliferation in normal human bronchial epithelial cells through NFkappaB activation and cyclin D1 up-regulation. Toxicol. Appl. Pharmacol. *205*, 133–148. https://doi.org/10.1016/j.taap.2004.09.019.

99. Schaal, C., and Chellappan, S.P. (2014). Nicotine-mediated cell proliferation and tumor progression in smoking-related cancers. Mol. Cancer Res. *12*, 14–23. https://doi.org/10.1158/1541-7786.MCR-13-0541.

100. Memmott, R.M., and Dennis, P.A. (2010). The role of the Akt/mTOR pathway in tobacco carcinogen-induced lung tumorigenesis. Clin. Cancer Res. *16*, 4–10. https://doi.org/10.1158/1078-0432.CCR-09-0234.

101. Wang, Y., Liu, J., Zhou, J.-S., Huang, H.-Q., Li, Z.-Y., Xu, X.-C., Lai, T.-W., Hu, Y., Zhou, H.-B., Chen, H.-P., et al. (2018). MTOR suppresses cigarette smoke-induced epithelial cell death and airway inflammation in chronic obstructive pulmonary disease. J. Immunol. *200*, 2571–2580. https://doi.org/10.4049/jimmunol.1701681.

102. Weinert, B.T., Narita, T., Satpathy, S., Srinivasan, B., Hansen, B.K., Schölz, C., Hamilton, W.B., Zucconi, B.E., Wang, W.W., Liu, W.R., et al. (2018). Time-resolved analysis reveals rapid dynamics and broad scope of the CBP/p300 acetylome. Cell *174*, 231–244.e12. https://doi.org/10.1016/j.cell.2018.04.033.

103. Mishima, Y., Wang, C., Miyagi, S., Saraya, A., Hosokawa, H., Mochizuki-Kashio, M., Nakajima-Takagi, Y., Koide, S., Negishi, M., Sashida, G., et al. (2014). Histone acetylation mediated by Brd1 is crucial for Cd8 gene activation during early thymocyte development. Nat. Commun. *5*, 5872. https://doi.org/10.1038/ncomms6872.

104. Gruber, J.J., Geller, B., Lipchik, A.M., Chen, J., Salahudeen, A.A., Ram, A.N., Ford, J.M., Kuo, C.J., and Snyder, M.P. (2019). HAT1 coordinates histone production and acetylation via H4 promoter binding. Mol. Cell *75*, 711–724.e5. https://doi.org/10.1016/j.molcel.2019.05.034.

105. Yang, J., Gong, C., Ke, Q., Fang, Z., Chen, X., Ye, M., and Xu, X. (2021). Insights into the function and clinical application of HDAC5 in Cancer Management. Front. Oncol. *11*, 661620. https://doi.org/10.3389/fonc.2021.661620.

106. Thompson, P.R., Wang, D., Wang, L., Fulco, M., Pediconi, N., Zhang, D., An, W., Ge, Q., Roeder, R.G., Wong, J., et al. (2004). Regulation of the p300 HAT domain via a novel activation loop. Nat. Struct. Mol. Biol. *11*, 308–315. https://doi.org/10.1038/nsmb740.

107. Wan, W., You, Z., Xu, Y., Zhou, L., Guan, Z., Peng, C., Wong, C.C.L., Su, H., Zhou, T., Xia, H., et al. (2017). mTORC1 phosphorylates acetyltransferase p300 to regulate autophagy and lipogenesis. Mol. Cell *68*, 323–335.e6. https://doi.org/10.1016/j.molcel.2017.09.020.

108. Cluntun, A.A., Huang, H., Dai, L., Liu, X., Zhao, Y., and Locasale, J.W. (2015). The rate of glycolysis quantitatively mediates specific histone acetylation sites. Cancer Metab. *3*, 10. https://doi.org/10.1186/s40170-015-0135-3.

109. Martire, S., Gogate, A.A., Whitmill, A., Tafessu, A., Nguyen, J., Teng, Y.C., Tastemel, M., and Banaszynski, L.A. (2019). Phosphorylation of histone H3.3 at serine 31 promotes p300 activity and enhancer acetylation. Nat. Genet. *51*, 941–946. https://doi.org/10.1038/s41588-019-0428-5.

110. Lau, P.N.I., and Cheung, P. (2011). Histone code pathway involving H3 S28 phosphorylation and K27 acetylation activates transcription and antagonizes polycomb silencing. Proc. Natl. Acad. Sci. USA *108*, 2801–2806. https://doi.org/10.1073/pnas.1012798108.

111. Hutti, J.E., Jarrell, E.T., Chang, J.D., Abbott, D.W., Storz, P., Toker, A., Cantley, L.C., and Turk, B.E. (2004). A rapid method for determining protein kinase phosphorylation specificity. Nat. Methods *1*, 27–29. https://doi.org/10.1038/nmeth708.

112. Songyang, Z., Carraway, K.L., 3rd, Eck, M.J., Harrison, S.C., Feldman, R.A., Mohammadi, M., Schlessinger, J., Hubbard, S.R., Smith, D.P., and Eng, C. (1995). Catalytic specificity of protein-tyrosine kinases is critical for selective signalling. Nature *373*, 536–539. https://doi.org/10.1038/373536a0.

113. Johnson, J.L., Yaron, T.M., Huntsman, E.M., Kerelsky, A., Song, J., Regev, A., Lin, T.-Y., Liberatore, K., Cizin, D.M., Cohen, B.M., et al. (2022). A global atlas of substrate specificities for the human serine/threonine kinome. https://doi.org/10.1101/2022.05.22.492882.

114. Songyang, Z., Blechner, S., Hoagland, N., Hoekstra, M.F., Piwnica-Worms, H., and Cantley, L.C. (1994). Use of an oriented peptide library to determine the optimal substrates of protein kinases. Curr. Biol. *4*, 973–982. https://doi.org/10.1016/s0960-9822(00)00221-9.

115. Yaffe, M.B., and Smerdon, S.J. (2004). The use of in vitro peptide-library screens in the analysis of phosphoserine/threonine-binding domain structure and function. Annu. Rev. Biophys. Biomol. Struct. *33*, 225–244. https://doi.org/10.1146/annurev.biophys.33.110502.133346.

116. Turk, B.E., Huang, L.L., Piro, E.T., and Cantley, L.C. (2001). Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. Nat. Biotechnol. *19*, 661–667. https://doi.org/10.1038/90273.

117. Manning, G., Whyte, D.B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. Science *298*, 1912–1934. https://doi.org/10.1126/science.1075762.

118. Agricola, E., Randall, R.A., Gaarenstroom, T., Dupont, S., and Hill, C.S. (2011). Recruitment of TIF1γ to chromatin via its PHD finger-bromodomain activates its ubiquitin ligase and transcriptional repressor activities. Mol. Cell *43*, 85–96. https://doi.org/10.1016/j.molcel.2011.05.020.

119. Yasui, Y., Urano, T., Kawajiri, A., Nagata, K.-I., Tatsuka, M., Saya, H., Furukawa, K., Takahashi, T., Izawa, I., and Inagaki, M. (2004). Autophosphorylation of a newly identified site of aurora-B is indispensable for cytokinesis. J. Biol. Chem. *279*, 12997–13003. https://doi.org/10.1074/jbc.M311128200.

120. Goto, H., Tomono, Y., Ajiro, K., Kosako, H., Fujita, M., Sakurai, M., Okawa, K., Iwamatsu, A., Okigaki, T., Takahashi, T., et al. (1999). Identification of a novel phosphorylation site on histone H3 coupled with mitotic chromosome condensation. J. Biol. Chem. *274*, 25543–25549. https://doi.org/10.1074/jbc.274.36.25543.

121. Cai, G., Yang, Q., and Sun, W. (2021). RSF1 in cancer: interactions and functions. Cancer Cell Int. *21*, 315. https://doi.org/10.1186/s12935-021-02012-9.

122. Lee, H.-S., Park, Y.-Y., Cho, M.-Y., Chae, S., Yoo, Y.-S., Kwon, M.-H., Lee, C.-W., and Cho, H. (2015). The chromatin remodeller RSF1 is essential for PLK1 deposition and function at mitotic kinetochores. Nat. Commun. *6*, 7904. https://doi.org/10.1038/ncomms8904.

123. Lee, H.-S., Lin, Z., Chae, S., Yoo, Y.-S., Kim, B.-G., Lee, Y., Johnson, J.L., Kim, Y.-S., Cantley, L.C., Lee, C.-W., et al. (2018). The chromatin remodeler RSF1 controls centromeric histone modifications to coordinate chromosome segregation. Nat. Commun. *9*, 3848. https://doi.org/10.1038/s41467-018-06377-w.

124. Elia, A.E.H., Cantley, L.C., and Yaffe, M.B. (2003). Proteomic screen finds pSer/pThr-binding domain localizing Plk1 to mitotic substrates. Science *299*, 1228–1231. https://doi.org/10.1126/science.1079079.

125. Yaron, T.M., Heaton, B.E., Levy, T.M., Johnson, J.L., Jordan, T.X., Cohen, B.M., Kerelsky, A., Lin, T.-Y., Liberatore, K.M., Bulaon, D.K., et al. (2020). The FDA-approved drug Alectinib compromises SARS-CoV-2 nucleocapsid phosphorylation and inhibits viral infection in vitro. https://doi.org/10.1101/2020.08.14.251207.

126. Zheng, Y., Ramsamooj, S., Li, Q., Johnson, J.L., Yaron, T.M., Sharra, K., and Cantley, L.C. (2019). Regulation of folate and methionine metabolism by multisite phosphorylation of human methylenetetrahydrofolate reductase. Sci. Rep. *9*, 4190. https://doi.org/10.1038/s41598-019-40950-7.

127. Chang, Y.-C., Chiou, J., Yang, Y.-F., Su, C.-Y., Lin, Y.-F., Yang, C.-N., Lu, P.-J., Huang, M.-S., Yang, C.-J., and Hsiao, M. (2019). Therapeutic targeting of aldolase A interactions inhibits lung cancer metastasis and prolongs survival. Cancer Res. *79*, 4754–4766. https://doi.org/10.1158/0008-5472.CAN-18-4080.

128. Luby, A., and Alves-Guerra, M.-C. (2021). Targeting metabolism to control immune responses in cancer and improve checkpoint blockade immunotherapy. Cancers *13*, 5912. https://doi.org/10.3390/cancers13235912.

129. Stevens, B.M., Jones, C.L., Pollyea, D.A., Culp-Hill, R., D'Alessandro, A., Winters, A., Krug, A., Abbott, D., Goosman, M., Pei, S., et al. (2020). Fatty acid metabolism underlies venetoclax resistance in acute myeloid leukemia stem cells. Nat. Cancer *1*, 1176–1187. https://doi.org/10.1038/s43018-020-00126-z.

130. van Galen, P., Hovestadt, V., Wadsworth, M.H., Hughes, T.K., Griffin, G.K., Battaglia, S., Verga, J.A., Stephansky, J., Pastika, T.J., Lombardi Story, J., et al. (2019). Single-cell RNA-Seq reveals AML hierarchies relevant to disease progression and immunity. Cell *176*, 1265–1281.e24. https://doi.org/10.1016/j.cell.2019.01.031.

131. Sade-Feldman, M., Yizhak, K., Bjorgaard, S.L., Ray, J.P., de Boer, C.G., Jenkins, R.W., Lieb, D.J., Chen, J.H., Frederick, D.T., Barzily-Rokni, M., et al. (2018). Defining T cell states associated with response to checkpoint immunotherapy in melanoma. Cell *175*, 998–1013.e20. https://doi.org/10.1016/j.cell.2018.10.038.

132. Pelka, K., Hofree, M., Chen, J.H., Sarkizova, S., Pirl, J.D., Jorgji, V., Bejnood, A., Dionne, D., Ge, W.H., Xu, K.H., et al. (2021). Spatially organized multicellular immune hubs in human colorectal cancer. Cell *184*, 4734–4752.e20. https://doi.org/10.1016/j.cell.2021.08.003.

133. Ji, A.L., Rubin, A.J., Thrane, K., Jiang, S., Reynolds, D.L., Meyers, R.M., Guo, M.G., George, B.M., Mollbrink, A., Bergenstråhle, J., et al. (2020). Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. Cell *182*, 497–514.e22. https://doi.org/10.1016/j.cell.2020.05.039.

134. Minakshi, P., Kumar, R., Ghosh, M., Saini, H.M., Ranjan, K., Brar, B., and Prasad, G. (2019). Chapter 14 - Single-cell proteomics: technology and applications. In Single-Cell Omics, D. Barh and V. Azevedo, eds. (Academic Press), pp. 283–318. https://doi.org/10.1016/B978-0-12-814919-5.00014-2.

135. Elyada, E., Bolisetty, M., Laise, P., Flynn, W.F., Courtois, E.T., Burkhart, R.A., Teinor, J.A., Belleau, P., Biffi, G., Lucito, M.S., et al. (2019). Cross-species single-cell analysis of pancreatic ductal adenocarcinoma reveals antigen-presenting cancer-associated fibroblasts. Cancer Discov. *9*, 1102–1123. https://doi.org/10.1158/2159-8290.CD-19-0094.

136. Le Large, T.Y., Mantini, G., Meijer, L.L., Pham, T.V., Funel, N., van Grieken, N.C., Kok, B., Knol, J., van Laarhoven, H.W., Piersma, S.R., et al. (2020). Microdissected pancreatic cancer proteomes reveal tumor heterogeneity and therapeutic targets. JCI Insight *5*, e138290. https://doi.org/10.1172/jci.insight.138290.

137. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222. https://doi.org/10.1093/bioinformatics/btv710.

138. Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., et al. (2018). Strelka2: fast and accurate calling of germline and somatic variants. Nat. Methods *15*, 591–594. https://doi.org/10.1038/s41592-018-0051-x.

139. Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219. https://doi.org/10.1038/nbt.2514.

140. Taylor-Weiner, A., Stewart, C., Giordano, T., Miller, M., Rosenberg, M., Macbeth, A., Lennon, N., Rheinbay, E., Landau, D.-A., Wu, C.J., et al. (2018). DeTiN: overcoming tumor-in-normal contamination. Nat. Methods *15*, 531–534. https://doi.org/10.1038/s41592-018-0036-9.

141. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. *22*, 568–576. https://doi.org/10.1101/gr.129684.111.

142. Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865–2871. https://doi.org/10.1093/bioinformatics/btp394.

143. Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human

cancers. Genome Biol. *12*, R41. https://doi.org/10.1186/gb-2011-12-4-r41.

144. Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501. https://doi.org/10.1038/nature12912.

145. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. https://doi.org/10.1126/science.aaz1776.

146. Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127. https://doi.org/10.1093/biostatistics/kxj037.

147. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.

148. Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics *14*, 7. https://doi.org/10.1186/1471-2105-14-7.

149. Aran, D., Hu, Z., and Butte, A.J. (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. *18*, 220. https://doi.org/10.1186/s13059-017-1349-1.

150. Wen, B., Wang, X., and Zhang, B. (2019). PepQuery enables fast, accurate, and convenient proteomic validation of novel genomic alterations. Genome Res. *29*, 485–493. https://doi.org/10.1101/gr.235028.118.

151. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47. https://doi.org/10.1093/nar/gkv007.

152. Hess, J.M., Bernards, A., Kim, J., Miller, M., Taylor-Weiner, A., Haradhvala, N.J., Lawrence, M.S., and Getz, G. (2019). Passenger hotspot mutations in cancer. Cancer Cell *36*, 288–301.e14. https://doi.org/10.1016/j.ccell.2019.08.002.

153. Costello, M., Pugh, T.J., Fennell, T.J., Stewart, C., Lichtenstein, L., Meldrim, J.C., Fostel, J.L., Friedrich, D.C., Perrin, D., Dionne, D., et al. (2013). Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. *41*, e67. https://doi.org/10.1093/nar/gks1443.

154. Li, Y., Porta-Pardo, E., Tokheim, C., Bailey, M.H., Yaron, T.M., Stathias, V., Geffen, Y., Imbach, K.J., Cao, S., Anand, S., et al. (2023). Pan-cancer proteogenomics connects oncogenic drivers to functional states. Cell *186*. Published Online: August 14, 2023.

155. Graubert, A., Aguet, F., Ravi, A., Ardlie, K.G., and Getz, G. (2021). RNA-SeQC 2: Efficient RNA-seq quality control and quantification for large cohorts. Bioinformatics *37*, 3048–3050. https://doi.org/10.1093/bioinformatics/btab135.

156. Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., and Dewey, C.N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics *26*, 493–500. https://doi.org/10.1093/bioinformatics/btp692.

157. Ouspenskaia, T., Law, T., Clauser, K.R., Klaeger, S., Sarkizova, S., Aguet, F., Li, B., Christian, E., Knisbacher, B.A., Le, P.M., et al. (2022). Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. Nat. Biotechnol. *40*, 209–217. https://doi.org/10.1038/s41587-021-01021-3.

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

159. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883. https://doi.org/10.1093/bioinformatics/bts034.

160. Tan, V.Y.F., and Févotte, C. (2013). Automatic Relevance Determination in Nonnegative Matrix Factorization with the /spl beta/-Divergence. IEEE Trans. Pattern Anal. Mach. Intell. *35*, 1592–1605. https://doi.org/10.1109/TPAMI.2012.240.

161. Kim, J., Kwiatkowski, D., McConkey, D.J., Meeks, J.J., Freeman, S.S., Bellmunt, J., Getz, G., and Lerner, S.P. (2019). The cancer genome atlas expression subtypes stratify response to checkpoint inhibition in advanced urothelial cancer and identify a subset of patients with high survival probability. Eur. Urol. *75*, 961–964. https://doi.org/10.1016/j.eururo.2019.02.017.

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

163. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer *18*, 696–705. https://doi.org/10.1038/s41568-018-0060-1.

164. Haradhvala, N.J., Kim, J., Maruvka, Y.E., Polak, P., Rosebrock, D., Livitz, D., Hess, J.M., Leshchiner, I., Kamburov, A., Mouw, K.W., et al. (2018). Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. Nat. Commun. *9*, 1746. https://doi.org/10.1038/s41467-018-04002-4.

165. Davies, H., Glodzik, D., Morganella, S., Yates, L.R., Staaf, J., Zou, X., Ramakrishna, M., Martin, S., Boyault, S., Sieuwerts, A.M., et al. (2017). HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat. Med. *23*, 517–525. https://doi.org/10.1038/nm.4292.

166. Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*, R29. https://doi.org/10.1186/gb-2014-15-2-r29.

167. Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25. https://doi.org/10.1186/gb-2010-11-3-r25.

168. Kamburov, A., Lawrence, M.S., Polak, P., Leshchiner, I., Lage, K., Golub, T.R., Lander, E.S., and Getz, G. (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc. Natl. Acad. Sci. USA *112*, E5486–E5495. https://doi.org/10.1073/pnas.1516373112.

169. Luna, A., Siper, M.C., Korkut, A., Durupinar, F., Dogrusoz, U., Aslan, J.E., Sander, C., Demir, E., and Babur, O. (2021). Analyzing causal relationships in proteomic profiles using CausalPath. Star Protoc. *2*, 100955. https://doi.org/10.1016/j.xpro.2021.100955.

170. Balci, H., Siper, M.C., Saleh, N., Safarli, I., Roy, L., Kilicarslan, M., Ozaydin, R., Mazein, A., Auffray, C., Babur, Ö., et al. (2021). Newt: a comprehensive web-based tool for viewing, constructing and analyzing biological maps. Bioinformatics *37*, 1475–1477. https://doi.org/10.1093/bioinformatics/btaa850.

171. Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401*, 788–791. https://doi.org/10.1038/44565.

172. Brunet, J.-P., Tamayo, P., Golub, T.R., and Mesirov, J.P. (2004). Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. USA *101*, 4164–4169. https://doi.org/10.1073/pnas.0308531101.

173. Knisbacher, B.A., Lin, Z., Hahn, C.K., Nadeu, F., Duran-Ferrer, M., Stevenson, K.E., Tausch, E., Delgado, J., Barbera-Mourelle, A., Taylor-

Weiner, A., et al. (2022). Molecular map of chronic lymphocytic leukemia and its impact on outcome. Nat. Genet. *54*, 1664–1674. https://doi.org/10.1038/s41588-022-01140-w.

174. Roh, W., Geffen, Y., Cha, H., Miller, M., Anand, S., Kim, J., Heiman, D.I., Gainor, J.F., Laird, P.W., Cherniack, A.D., et al. (2022). High-resolution profiling of lung adenocarcinoma identifies expression subtypes with specific biomarkers and clinically relevant vulnerabilities. Cancer Res. *82*, 3917–3931. https://doi.org/10.1158/0008-5472.CAN-22-0432.

175. Hunter, T. (2007). The age of crosstalk: phosphorylation, ubiquitination, and beyond. Mol. Cell *28*, 730–738. https://doi.org/10.1016/j.molcel.2007.11.019.

176. Woodsmith, J., Kamburov, A., and Stelzl, U. (2013). Dual coordination of post translational modifications in human protein networks. PLoS Comput. Biol. *9*, e1002933. https://doi.org/10.1371/journal.pcbi.1002933.

177. Beltrao, P., Bork, P., Krogan, N.J., and van Noort, V. (2013). Evolution and functional cross-talk of protein post-translational modifications. Mol. Syst. Biol. *9*, 714. https://doi.org/10.1002/msb.201304521.

178. Bludau, I., Willems, S., Zeng, W.-F., Strauss, M.T., Hansen, F.M., Tanzer, M.C., Karayel, O., Schulman, B.A., and Mann, M. (2022). The structural context of posttranslational modifications at a proteome-wide scale. PLoS Biol. *20*, e3001636. https://doi.org/10.1371/journal.pbio.3001636.

179. Babur, Ö., Melrose, A.R., Cunliffe, J.M., Klimek, J., Pang, J., Sepp, A.-L.I., Zilberman-Rudenko, J., Tassi Yunga, S., Zheng, T., Parra-Izquierdo, I., et al. (2020). Phosphoproteomic quantitation and causal analysis reveal pathways in GPVI/ITAM-mediated platelet activation programs. Blood *136*, 2346–2358. https://doi.org/10.1182/blood.2020005496.

180. Babur, Ö., Ngo, A.T.P., Rigg, R.A., Pang, J., Rub, Z.T., Buchanan, A.E., Mitrugno, A., David, L.L., McCarty, O.J.T., Demir, E., et al. (2018). Platelet procoagulant phenotype is modulated by a p38-MK2 axis that regulates RTN4/Nogo proximal to the endoplasmic reticulum: utility of pathway analysis. Am. J. Physiol. Cell Physiol. *314*, C603–C615. https://doi.org/10.1152/ajpcell.00177.2017.

181. Joshi, S.K., Nechiporuk, T., Bottomly, D., Piehowski, P.D., Reisz, J.A., Pittsenbarger, J., Kaempf, A., Gosline, S.J.C., Wang, Y.-T., Hansen, J.R., et al. (2021). The AML microenvironment catalyzes a stepwise evolution to gilteritinib resistance. Cancer Cell *39*, 999–1014.e8. https://doi.org/10.1016/j.ccell.2021.06.003.

182. Khadka, P., Reitman, Z.J., Lu, S., Buchan, G., Gionet, G., Dubois, F., Carvalho, D.M., Shih, J., Zhang, S., Greenwald, N.F., et al. (2022). PPM1D mutations are oncogenic drivers of de novo diffuse midline glioma formation. Nat. Commun. *13*, 604. https://doi.org/10.1038/s41467-022-28198-8.

183. Keshishian, H., McDonald, E.R., 3rd, Mundt, F., Melanson, R., Krug, K., Porter, D.A., Wallace, L., Forestier, D., Rabasha, B., Marlow, S.E., et al. (2021). A highly multiplexed quantitative phosphosite assay for biology and preclinical studies. Mol. Syst. Biol. *17*, e10156. https://doi.org/10.15252/msb.202010156.

184. Johnson, B.E., Creason, A.L., Stommel, J.M., Keck, J.M., Parmar, S., Betts, C.B., Blucher, A., Boniface, C., Bucher, E., Burlingame, E., et al. (2022). An omic and multidimensional spatial atlas from serial biopsies of an evolving metastatic breast cancer. Cell Rep. Med. *3*, 100525. https://doi.org/10.1016/j.xcrm.2022.100525.

185. Zhao, M., Scott, S., Evans, K.W., Yuca, E., Saridogan, T., Zheng, X., Wang, H., Korkut, A., Cruz Pico, C.X., Demirhan, M., et al. (2021). Combining neratinib with CDK4/6, mTOR, and MEK inhibitors in models of HER2-positive cancer. Clin. Cancer Res. *27*, 1681–1694. https://doi.org/10.1158/1078-0432.CCR-20-3017.

Please cite this article in press as: Geffen et al., Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation, Cell (2023), https://doi.org/10.1016/j.cell.2023.07.013

STAR★METHODS

KEY RESOURCES TABLE

RESOURCE or REAGENT	SOURCE	IDENTIFIER
Data Deposition		
CPTAC clinical and proteomic data	Li et al.[27]	https://pdc.cancer.gov/pdc/cptac-pancancer
Full proteomic data tables - raw, harmonized and imputed	This manuscript	https://pdc.cancer.gov/pdc/cptac-pancancer
CPTAC genomic and transcriptomic data	Li et al.[27]	https://pdc.cancer.gov/pdc/cptac-pancancer, and Cancer Data Service (CDS: https://dataservice.datacommons.cancer.gov/)
Software and Algorithms		
Somatic variant calling pipeline for CPTAC	Li Ding Lab	https://github.com/ding-lab/somaticwrapper
bam-readcount v0.8	McDonnell Genome Institute	https://github.com/genome/bam-readcount
GATK4's CalculateContamination	GATK	https://gatk.broadinstitute.org/hc/en-us/articles/360036888972-CalculateContamination
GATK4 Picard tools	GATK	https://github.com/broadinstitute/picard
GATK4 Funcotator	GATK	https://gatk.broadinstitute.org/hc/en-us/articles/360037224432-Funcotator
Panel-Of-Normal token tool	Getz Lab	https://github.com/getzlab/tokenizer_TOOL
SNVmerge	Getz Lab	https://github.com/getzlab/cptac_wxs_harmonize
Manta	Chen et al.[137]	https://github.com/Illumina/manta
Strelka v2	Kim et al.[138]	https://github.com/Illumina/strelka
MuTect	Cibulskis et al.[139]	https://software.broadinstitute.org/gatk/download/archive
DeTiN	Taylor-Weiner et al.[140]	https://github.com/getzlab/deTiN
VarScan2.3.8	Koboldt et al.[141]	http://varscan.sourceforge.net
Pindel0.2.5	Ye et al.[142]	http://gmt.genome.wustl.edu/packages/pindel/
SignatureAnalyzer	Kim et al.[31]	https://github.com/getzlab/getzlab-SignatureAnalyzer
GISTIC2.0	Mermel et al.[143]	https://github.com/broadinstitute/gistic2
MutSig2CV	Lawrence et al.[144]	https://github.com/getzlab/MutSig2CV
GTEx RNA-seq pipeline	The GTEx Consortium[145]	https://github.com/broadinstitute/gtex-pipeline
CLUMPS-PTM	Getz Lab	https://github.com/getzlab/CLUMPS-PTM
The Kinase Library	Johnson et al.[34]	Described in methods
CausalPath	Babur et al.[35]	https://github.com/PathwayAndDataAnalysis/causalpath
Spectrum Mill	Karl R. Clauser, Steven Carr Lab	https://proteomics.broadinstitute.org/
ComBat (v3.20.0)	Johnson et al.[146]	https://bioconductor.org/packages/release/bioc/html/sva.html
ESTIMATE	Yoshihara et al.[72]	https://bioinformatics.mdanderson.org/public-software/estimate/
fGSEA	Subramanian et al.[147]	http://bioconductor.org/packages/release/bioc/html/fgsea.html
GSVA	Hänzelmann et al.[148]	https://www.bioconductor.org/packages/release/bioc/html/GSVA.html
xCell	Aran et al.[149]	http://xcell.ucsf.edu/
PepQuery	Wen et al.[150]	http://pepquery.org
PTM-SEA	Krug et al.[36]	https://github.com/broadinstitute/ssGSEA2.0
Terra	Broad Institute's Data S ciences Platform	https://terra.bio/
LIMMA v3.36 (R Package)	Ritchie et al.[151]	https://bioconductor.org/packages/release/bioc/html/limma.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Gad Getz (gadgetz@broadinstitute.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Raw and processed proteomics as well as open access genomic data can be obtained via Proteomic Data Commons (PDC) at https://pdc.cancer.gov/pdc/cptac-pancancer. Raw genomic and transcriptomic data files can be accessed via the Genomic Data Commons (GDC) Data Portal at https://portal.gdc.cancer.gov with dbGaP Study Accession: phs001287.v16.p6. Complete CPTAC pan-cancer controlled and processed data can be accessed via the Cancer Data Service (CDS: https://dataservice.datacommons.cancer.gov/). The CPTAC pan-cancer data hosted in CDS is controlled data and can be accessed through the NCI DAC approved, dbGaP compiled whitelists. Users can access the data for analysis through the Seven Bridges Cancer Genomics Cloud (SB-CGC), which is one of the NCI-funded Cloud Resource/platforms for compute intensive analysis.

1. Create an account on CGC, Seven Bridges (https://cgc-accounts.sbgenomics.com/auth/register)
2. Get approval from dbGaP to access the controlled study (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001287.v16.p6)
3. Log into CGC to access Cancer Data Service (CDS) File Explore
4. Copy data into your own space and start analysis and exploration
5. Visit the CDS page on CGC to see what studies are available and instructions and guides to use the resources. (https://docs.cancergenomicscloud.org/page/cds-data)

Code for the analysis and figures in this paper can be found at https://github.com/getzlab/CPTAC_PanCan_PTM_2023.
Software and code used in this study are referenced in their corresponding STAR Methods sections and the key resources table.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human subjects and clinical data

In this study, a comprehensive dataset was assembled, comprising a total of 1110 patients from 11 different cohorts. The cohorts included 99 patients with glioblastoma (GBM),[4] 110 patients with head and neck squamous cell carcinoma (HNSCC),[5] 110 patients with lung adenocarcinoma (LUAD),[6] 108 patients with lung squamous cell carcinoma (LSCC),[7] 121 patients with breast cancer (BRCA),[8] 140 patients with pancreatic ductal adenocarcinoma (PDAC),[9] 110 patients with clear cell renal cell carcinoma (ccRCC),[10] 82 patients with high-grade serous ovarian cancer (HGSC),[11] 95 patients with uterine corpus endometrial carcinoma (UCEC),[12] 96 patients with colorectal adenocarcinoma (COAD)[13] and 39 patients with Medulloblastoma (MB).[28]

This combined dataset comprises 531 males and 579 females, with an age range of 2-90 years (median age 62). Detailed clinical data can be found in the companion Pan-Cancer resource manuscript.[27]

METHOD DETAILS

Genomics Data processing

Harmonized genome alignment

WGS, WES, RNA-Seq sequence data were harmonized by NCI Genomic Data Commons (GDC) https://gdc.cancer.gov/about-data/gdc-data-harmonization, which included alignment to GDC's hg38 human reference genome (GRCh38.d1.vd1) and additional quality checks. All the downstream genomic processing was based on the GDC-aligned BAMs to ensure reproducibility.

Somatic mutation detection

The Broad hg38 characterization pipeline. Patient whole exome sequencing (WES) data, i.e. WES sequences of patients' matched tumor and blood normal samples, were analyzed using the Getz Lab's production hg38 WES characterization pipeline. While somatic whole genome sequencing data were available for 7 out of 10 cancer types, they were sequenced to an average coverage of 15X, which would limit our discovery of subclonal mutations or clonal mutations in low tumor purity and/or high ploidy samples.

The hg38 characterization pipeline runs on the Terra cloud-based analysis platform (https://terra.bio/). This pipeline is the standard computational workflow used by the Getz Lab for characterizing a tumor sample's somatic variants through contrastive computational analysis of matched tumor-normal WES BAMs. The pipeline's analysis steps are organized into five modules: (1) DNA Sequence Data Quality Control; (2) Somatic Copy Number Analysis; (3) Somatic Variant Discovery, which includes the discovery of SNVs and indels; (4) Post-Discovery Filtering; and (5) merging of adjacent somatic SNPs into DNPs, TNPs and ONPs.

Please cite this article in press as: Geffen et al., Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation, Cell (2023), https://doi.org/10.1016/j.cell.2023.07.013

The DNA Sequence Quality Control module, at the head of the pipeline, employs (i) GATK4's CalculateContamination (Ver GATK 4.1.4.1) tool to calculate the fraction of reads coming from cross-sample contamination and (ii) GATK4 Picard tools (ver GATK 4.0.5.1) to validate the BAM files and collect multiple classes of metrics that can be used to evaluate sequencing data quality. The pipeline's Somatic Copy Number Analysis module runs the GATK4 Best Practices Workflow (ver GATK 4.1.4.1) for discovery of allele-specific copy-number alterations.

The Somatic Variant Discovery module employs MuTect[139] for detection of somatic single nucleotide variants and Manta+Strelka v2[137,138] for detecting small insertions and deletions (INDEL sizes up to 49 bases). Following this initial detection of somatic SNVs (SSNV) and INDELs DeTiN[140] (v1.8.9) was run to rescue SSNVs and INDELs called by MuTect and Strelka that may have been misclassified as germline variants due to contamination of normal tissue with tumor cells. The resulting SSNV and indel VCFs are each run through the GATK4 Funcotator (ver GATK 4.1.4.1) to analyze detected variants for their function and produce annotated MAFs, which are then merged into a single MAF containing candidate SSNVs and indels.

The Post-Discovery Filtering module runs a collection of filters in parallel on the merged annotated MAF to eliminate artifacts, germline variants and common sequencing artifacts that occur in normal panels. The filtered variant calls coming out of each filter are then aggregated to create an "intersection MAF" containing only variants that pass all filters. These variants are then run through a mutation validator that validates the calls with any available orthogonal sequencing data (e.g., from WGS sequencing, targeted sequencing, low pass sequencing, RNA sequencing).

The final stage of the pipeline employs a SNVmerger subworkflow which merges SNPs to DNP/TNP/ONPs and writes the resulting oligonucleotides to a VCF, which are then re-annotated and merged into the earlier filtered/validated variant MAF.

Washington University characterization pipeline. In parallel to somatic mutation calling done by the Broad pipeline, somatic mutations and DNP calls were done by the Washington University characterization pipeline and are provided in detail in the companion Pan-Cancer Driver manuscript.[185]

Callset Harmonization. The per patient variant calls employed by the CPTAC PanCAN working group were derived from the harmonization of variant calls made independently by the Broad and Washington University.

ICE whole-exome capture technology was deployed by the Genomics Platform at the Broad Institute for all CPTAC projects. Therefore, as a first filtering step, we removed calls outside of the ICE capture interval list. A panel-of-normals built from an aggregation of normal blood samples from the CPTAC and TCGA cohorts, which is an integral part of the Broad's somatic mutation calling pipeline, was used to filter recurrent artifacts arising from calls made by the Washington University pipeline. In addition, indels were left-aligned to make sure their representations were comparable.

While we got better concordance between the two pipelines, we observed 2 key differences:

1. Majority of divergent calls are of low variant allele frequency (VAF < 0.05)
2. The Broad's pipeline calls long multiple nucleotide polymorphisms (MNPs)

To mitigate (1), we first removed all calls with VAF < 0.05 from both pipelines and rescued only high confident calls if either criteria is satisfied:

a. If a low VAF variant is called from both pipelines
b. If a low VAF is only called by either pipeline, but is a cancer hotspot defined in Hess 2019.[152]

To mitigate (2), long MNPs were collapsed to shorter MNPs by imposing a more stringent merging criteria that requires a 2bp gap length at max.

C>A artifact in CPTAC2 cohorts. Using Asymtools2[43] we were able to identify a sequencing artifact affecting CPTAC2 whole exome sequencing. Asymtools2 is a framework for visualizing mutational strand asymmetries. Asymtools2 illustrated a biased enrichment of cytosine to adenine SNVs over guanine to thymine SNVs on the genomic reference strand at low allelic fractions, suggesting a previously described process driven by oxidative damage of guanine to 8-oxoguanine after bait-DNA hybridization.[153] We further partitioned the C>A and G>T mutations with an allele fraction of less than 0.1 into their trinucleotide contexts, establishing the G>T contexts as the null model for correction. We then corrected for the sequencing artifact by ranking each C>A mutation by its allelic fraction and removed mutations until the number of C>A mutation counts were equal to those of the G>T mutations for each context.

Functional Impact. Finally, the functional impact of harmonized calls was annotated with GATK Funcotator.

Germline SNP and short indel discovery from WES (Washington University in St Louis)

Germline variant calling was performed using the Washington University pipeline and is provided in detail in the companion Pan-Cancer Driver manuscript.[154]

RNAseq data processing and quantification

We processed the RNA-seq data from all cohorts using the GTEx/TOPMed pipeline described at https://github.com/broadinstitute/gtex-pipeline/blob/master/TOPMed_RNAseq_pipeline.md.[145] The samples were aligned to the human reference genome GRCh38 with the GENCODE V34 gene annotation using STAR v2.7.5a; optical and PCR duplicates were identified with Picard 2.18.17 MarkDuplicates; quality control and gene-level quantification (in Transcripts per Million (TPM) units) were performed with

RNA-SeQC 2.3.6,[155] and isoform expression was quantified with RSEM.[156] We defined 'expressed' across the combined 11 cohorts using the following criteria: both (i) ≥0.1 TPM in ≥20% of samples and (ii) ≥6 reads in ≥20% of samples.

Proteomics data processing

Proteomics LC-MS/MS data interpretation

MS/MS spectra obtained from proteins/phosphosites and acetylation sites were interpreted by Spectrum Mill (SM) v 7.08 (proteomics.broadinstitute.org) to provide identification and relative quantitation at the protein, peptide, and post-translational modification (PTM) phospho- and acetyl-site levels.

Personalized sequence databases

For searching all available datasets with LC-MS/MS, we generated a cohort-level personalized protein sequence database for each tumor type starting with a base human reference proteome to which we appended non-redundant somatic and germline variants and indels for each of the ~100 participants/cohort. The base proteome consisted of the human reference proteome GENCODE 34 (ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_34/) with 47,429 non-redundant protein coding transcript biotypes mapped to the human reference genome GRCh38, 602 common laboratory contaminants, 2043 curated smORFs (lncRNA and uORFs), and 237,427 novel unannotated ORFs (nuORFs) supported by ribosomal profiling nuORF DB v1.0[157] for a total of 287,501 entries, which yield 16,645,198 distinct 9-mers. The nuORFs alone yield 8,612,372 distinct 9-mers and thus increase the peptide search space by only a factor of ~2. The personalized protein sequence entries were prepared by processing the individual participant's somatic and germline variant calls from whole exome sequencing data, described above, using QUILTS v3[158] with no further variant quality filtering using an Ensembl v100 reference proteome and reference genome for sequence identifiers consistent with the variant calling. GENCODE v34 is a contemporaneous subset of Ensembl v100 (March 2020). Using the SM Protein Database utilities, the base reference proteome and individual patient proteomes were combined and redundancy removed to produce a cohort-level protein sequence database and a variant summary table to enable subsequent mapping of sequence variants identified in TMT multiplexed LC-MS/MS datasets back to individual patients.

Spectrum quality filtering

For all datasets, similar MS/MS spectra with the same precursor m/z acquired in the same chromatographic peak were merged, the precursor MH+ inclusion range was 800–6000, and the spectral quality filter was a sequence tag length > 0 (i.e., minimum of two peaks separated by the in-chain mass of an amino acid).

MS/MS search conditions

Using the SM MS/MS search module for all datasets included the next parameters: ''trypsin allow P'' enzyme specificity with up to 4 missed cleavages; precursor and product mass tolerance of ± 20 ppm; 30% minimum matched peak intensityScoring parameters were ESI-QEXACTIVE- HCD-v2, for whole proteome datasets, and ESI-QEXACTIVE-HCD-v3, for phosphoproteome and acetylome. Allowed fixed modifications included carbamidomethylation of cysteine and selenocysteine. TMT labeling was required at lysine, but peptide N-termini were allowed to be either labeled or unlabeled. Allowed variable modifications for whole proteome datasets were acetylation of protein N-termini, oxidized methionine, deamidation of asparagine, hydroxylation of proline in PG motifs, pyro-glutamic acid at peptide N-terminal glutamine, and pyro-carbamidomethylation at peptide N-terminal cysteine with a precursor MH+ shift range of -18 to 97 Da. For all PTM-omes, variable modifications were revised to omit hydroxylation of proline and allow deamidation only in NG motifs. The phosphoproteome was revised to allow phosphorylation of serine, threonine, and tyrosine with a precursor MH+ shift range of -18 to 272 Da. The acetylome was revised to allow acetylation of lysine with a precursor MH+ shift range of -400 to 70 Da.

PTM site localization

Using the SM Autovalidation and Protein/Peptide Summary modules for the PTM-ome datasets results were filtered and reported at the phospho- and acetyl-site levels. When calculating scores at the variable modification (VM) site level and reporting the identified VM sites, redundancy was addressed in SM as follows: a VM-site table was assembled with columns for individual TMT-plex experiments and rows for individual VM-sites. PSMs were combined into a single row for all non-conflicting observations of a particular VM-site (e.g., different missed cleavage forms, different precursor charges, confident and ambiguous localizations, and different sample-handling modifications). For related peptides, neither observations with a different number of VM-sites nor different confident localizations were allowed to be combined. Selecting the representative peptide for a VM-site from the combined observations was done such that once confident VM-site localization was established, higher identification scores and longer peptide lengths were preferred. While an SM PSM identification score was based on the number of matching peaks, their ion type assignment, and the relative height of unmatched peaks, the VM site localization score was the difference in identification score between the top two localizations. The score threshold for confident localization, > 1.1, essentially corresponded to at least 1 b or y ion located between two candidate sites that has a peak height > 10% of the tallest fragment ion (neutral losses of phosphate from the precursor and related ions as well as immonium and TMT reporter ions were excluded from the relative height calculation). The ion type scores for b-H3PO4, y-H3PO4, b-H2O, and y-H2O ion types were all set to 0.5. This prevented inappropriate confident localization assignment when a spectrum lacked primary b or y ions between two possible sites but contained ions that could be assigned as either phosphate-loss ions for one localization or water loss ions for another localization.

Protein grouping of PSMs, peptides, and PTM sites

Using the SM Autovalidation and Protein/Peptide summary modules, results were filtered and reported at the protein level. Identified proteins were combined into the same protein group if they shared a peptide with sequence length greater than 8. A protein group

Please cite this article in press as: Geffen et al., Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation, Cell (2023), https://doi.org/10.1016/j.cell.2023.07.013

could be expanded into subgroups (isoforms or family members) when distinct peptides were present which uniquely represent a subset of the proteins in a group. For the proteome dataset the protein grouping method ''expand subgroups, top uses shared'' (SGT) was employed which allocates peptides shared by protein subgroups only to the highest scoring subgroup containing the peptide. For the PTM-ome datasets the protein grouping method ''unexpand subgroups'' was employed, which reports a VM-site only once per protein group allocated to the highest scoring subgroup containing the representative peptide. The SM protein score is the sum of the scores of distinct peptides. A distinct peptide is the single highest scoring instance of a peptide detected through an MS/MS spectrum. MS/MS spectra for a particular peptide may have been recorded multiple times (e.g., as different precursor charge states, in adjacent bRP fractions, modified by deamidation at Asn or oxidation of Met, or with different phosphosite localization), but are still counted as a single distinct peptide.

Peptide spectrum match (PSM) filtering and false discovery rates (FDR)

Using the SM Autovalidation module, peptide spectrum matches (PSMs) for individual spectra were confidently assigned by applying target-decoy based FDR estimation to achieve <1.0% FDR at the PSM, peptide, VM site and protein levels. For the whole proteome dataset, thresholding was done in 3 steps: at the PSM level, the protein level for each TMT-plex, and the protein level for the cohort of 2 TMT-plexes. For the PTM-omes, phosphoproteome and acetylome datasets thresholding was done in two steps: at the PSM level for each TMT-plex, and at the VM site level for the cohort of 2 TMT-plexes. In step 1 for all datasets, PSM-level autovalidation was done first and separately for each TMT-plex experiment using an auto-thresholds strategy with a minimum sequence length of 7; automatic variable range precursor mass filtering; with score and delta Rank1 - Rank2, score thresholds were optimized to yield a PSM level FDR estimate for precursor charges 2 through 4 of < 0.8% for each precursor charge state in each LC-MS/MS run. To achieve reasonable statistics for precursor charges 5-6, thresholds were optimized to yield a PSM-level FDR estimate of < 0.4% across all runs per TMT-plex experiment (instead of per each run), since many fewer spectra are generated for the higher charge states.

In step 2 for the PTM-omes: phosphoproteome and acetylome datasets VM site polishing autovalidation was applied across both TMT plexes to retain all VM site identifications with either a minimum id score of 8.0 or observation in n TMT plexes (n=4, 3, or 2 if > 20, 7, or 1 plexes/cohort, respectively). The intention of the VM site polishing step is to control FDR by eliminating unreliable VM site level identifications, particularly low scoring VM-sites that are only detected as low scoring peptides that are also infrequently detected across both TMT plexes in the study. Using the SM Protein/Peptide Summary module to make VM-site reports, the ubiquitylome and acetylome datasets are further filtered to remove peptides ending with the regular expression [∧K][K]k since trypsin and Lys-C cannot cleave at a acetylated lysine. The [K] means retain if unmodified Lys present in one of the last two positions to allow for a missed cleavage with ambiguous PTM-site localization. C-terminally acetylated lysines are present in the acetylome dataset, but have been shown to arise from artifactual modification during TMT-labeling after trypsin digestion.

In step 2 for the whole proteome dataset, protein polishing autovalidation was applied separately to each TMT-plex experiment to further filter the PSMs using a target protein level FDR threshold of zero. The primary goal of this step was to eliminate peptides identified with low scoring PSMs that represent proteins identified by a single peptide, so-called ''one-hit wonders.'' After assembling protein groups from the autovalidated PSMs, protein polishing determined the maximum protein level score of a protein group that consisted entirely of distinct peptides estimated to be false-positive identifications (PSMs with negative delta forward-reverse scores). PSMs were removed from the set obtained in the initial peptide level autovalidation step if they contributed to protein groups that had protein scores below the maximum false-positive protein score. Step 3 was then applied, consisting of protein polishing autovalidation across all TMT plexes in a cohort together using the protein grouping method ''expand subgroups, top uses shared'' to retain protein subgroups with either a minimum protein score of 25 or observation in TMT plexes (n=4, 3, or 2 if > 20, 7, or 1 plexes/cohort, respectively). The primary goal of this step was to eliminate low scoring proteins that were infrequently detected in a cohort. As a consequence of these two proteins- polishing steps, each identified protein reported in the study comprised multiple peptides, unless a single excellent scoring peptide was the sole match and that peptide was observed in multiple TMT-plexes.

Quantitation using TMT ratios

Using the SM Protein/Peptide Summary module, a protein comparison report was generated for the proteome dataset using the protein grouping method ''expand subgroups, top uses shared'' (SGT). For the PTM-omes: phosphoproteome and acetylome datasets Variable Modification site comparison reports limited to either phospho- or acetyl-sites, respectively, was generated using the protein grouping method ''unexpand subgroups.'' Relative abundances of proteins and VM-sites were determined in SM using TMT reporter ion $\log_2$ intensity ratios from each PSM. TMT reporter ion intensities were corrected for isotopic impurities in the SM Protein/Peptide Summary module using the afRICA correction method, which implements determinant calculations according to Cramer's Rule and correction factors obtained from the reagent manufacturer's certificate of analysis for each cohort. Each protein-level or PTM site-level TMT ratio was calculated as the median of all PSM-level ratios contributing to a protein subgroup or PTM site. PSMs were excluded from the calculation if they lacked a TMT label, had a precursor ion purity < 50% (MS/MS has significant precursor isolation contamination from co-eluting peptides), or had a negative delta forward-reverse identification score (half of all false-positive identifications). Using the SM Process Report module non-quantifiable proteins and PTM sites (e.g., unlabeled peptides containing an acetylated protein N-terminus and ending in arginine rather than lysine) were removed, and median/MAD normalization was performed on each TMT channel in each dataset to center and scale the aggregate distribution of protein-level or PTM site-level log-ratios around zero in order to nullify the effect of differential protein loading and/or systematic MS variation.

Normalization of phosphosites and acetylation sites

For downstream analyses with PTM data, we perform ordinary least squares fit using statsmodels.regression.linear_model.OLS for every matched value of protein and PTM site aligned by accession number (RefSeq) (Figure S1). The residuals from this are ''corrected'' phosphoproteome and acetylome values we term as ''phosphoproteome_res'' or ''acetylome_res.''

Patient Signatures & Clustering

Transcriptomics

For downstream analyses, TPM values were scaled using DESeq2 size factors and then $\log_2(x+1)$ transformed. Three cohorts (BRCA, HGSC, COAD, from CPTAC2) were sequenced with a polyA selection protocol, with the remainder sequenced with a total RNA protocol (rRNA-depleted using RiboZero, from CPTAC3). Medulloblastoma was also sequenced with a total RNA protocol. This protocol batch effect was regressed out from the log-transformed values using COMBAT.[159] Finally, the 5,000 genes with the highest coefficient of variation were selected as inputs to clustering.

Proteomics

We first selected phospho-sites that were fully localized using mass spectrometry. Then, within each cohort, we filtered out any proteins or phosphoproteins that were identified based on SM criteria in less than $\leq$25% of samples (Figure S1A). After this, we further subset the feature space to include only detected proteins & phospho-sites across all 11 cohorts to avoid imputation across entire cohorts. Within each cohort, we then used K-nearest neighbors to impute missing values with sklearn.impute.KNNImputer with K=5. Finally, to correct phosphoproteome levels for protein abundance, we performed ordinary least squares fit described in the proteomics data processing Section (above) with filtered and imputed data (Figure S1A).

Combining Multi-Omic Data

Across all 11 cohorts, the shared data types available for clustering were whole transcriptome RNA-seq, proteome, and phosphoproteome. To harmonize the RNA data with the normally distributed proteomic data, we applied an inverse normal transformation, median centered the data, and scaled the data using median absolute deviation. We selected samples that had matched RNA, protein, and phosphoprotein in the entire dataset and concatenated these 3 matrices. To reduce the transcriptome space to a comparable feature size to the proteome data, we selected the top 5,000 highly variable genes, ranked by coefficient of variation. The combined matrix contained gene expression for 5,000 highly variable mRNA protein coding genes, 5,716 proteins, and 3,341 phosphoproteins for 1,110 patients across 11 cancer types. Prior to any cohort-level correction, we found that cohort and tissue-specific effects were the dominant source of variation in multi-omic data (Figures S1C and S1D). We therefore performed cohort-level batch correction by regressing out the cohort effects (as dummy-coded covariates/ indicator variables) from the combined matrix.

Multi-Omic Signatures

We derived the expression signatures using SignatureAnalyzer (https://github.com/getzlab/SignatureAnalyzer), a Bayesian variant of non-negative matrix factorization (ARD-NMF).[33,160,161] To use this tool with zero-centered, normally distributed data, SignatureAnalyzer splits the input matrix into positive and negative matrices before running the decomposition (Table S2). L2 priors were imposed on both the W and H matrices, and a Gaussian objective function was used. SignatureAnalyzer was run 100 times with random initialization, and the solution with the best objective function with the mode number of signatures (k=33) was selected (robustness analysis and additional support is provided in Note S1). To annotate these 33 signatures, we performed ranked gene-set enrichment analysis using *fGSEA* with factor weight as the rank.

Sample-Sample Clustering

To further cluster samples using the 33 derived expression signatures, we created a similarity matrix based on cosine similarity of the H-matrix (samples x signatures) derived from the SignatureAnalyzer algorithm. We then performed hierarchical clustering using euclidean distance as the metric and Ward linkage using SciKit-learn. For downstream analysis, we compared each pair of sample clusters at each split of the dendrogram (see differential expression).

Immune clustering

To estimate the abundance of immune cell-types and signatures in each sample, we ran multiple bulk RNA deconvolution approaches. We ran CIBERSORT with the LM22 signature matrix[74] using the re-processed CPTAC transcriptomic TPM data for all 11 cohorts. We also ran ESTIMATE[72] as a separate approach for estimating tumor purity and immune infiltration using TPM data. Finally, we ran ImmuneSubtypeClassifier[73] to classify samples into different ''immune types''. The methods for the immune analysis are provided at https://github.com/getzlab/CPTAC_PanCan_2021/blob/master/analysis/Fig_immuno_metabolism/runImmuneDeconv.R. To identify broad immune clusters to probe differences in immune signals and changes in metabolic levels using matched protein and PTM data, we ran gene set enrichment analysis using the gene sets from CIBERSORT's LM22 matrix on the z-scored transcriptomic TPM data. Next, we performed hierarchical clustering using euclidean distance and Ward-linkage with SciKit-learn[162] and selected K=4 to identify 4 broad ''immune clusters'' for downstream use. To perform differential expression analysis for each immune subtype, we used a 'one vs. rest' approach. We additionally corrected for cancer type (i.e., cohort) by modeling it as a fixed covariate along with the estimated immune abundances from CIBERSORT for ''Macrophages_M0'', ''Macrophages_M1'', ''Macrophages_M2'', ''T_cells_CD8'', and ''T_cells_CD4_memory_resting'' to better separate tumor intrinsic expression from contributions of the immune microenvironment.

Please cite this article in press as: Geffen et al., Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation, Cell (2023), https://doi.org/10.1016/j.cell.2023.07.013

Interpretive data analysis

Variant call tools

GISTIC. The Genomic Identification of Significant Targets in Cancer (GISTIC2.0) algorithm[143] was used to identify significantly amplified or deleted focal-level and arm-level events, with q value <0.25 considered significant. The following parameters were used: Amplification Threshold = 0.1, Deletion Threshold = 0.1, Cap Values = 1.5, Broad Length Cutoff = 0.98, Remove X-Chromosome = 0, Confidence Level = 0.99, Join Segment Size = 4, Arm Level Peel Off = 1, Maximum Sample Segments = 2000, Gene GISTIC = 1.

Each gene of every sample is assigned a thresholded copy number level that reflects the magnitude of its deletion or amplification. These are integer values ranging from -2 to 2, where 0 means no amplification or deletion of magnitude greater than the threshold parameters described above. Amplifications are represented by positive numbers: 1 means amplification above the amplification threshold; 2 means amplification larger than the arm level amplifications observed in the sample. Deletions are represented by negative numbers: -1 means deletion beyond the threshold; -2 means deletions greater than the minimum arm-level copy number observed in the sample.

MutSig2CV. The somatic variants were filtered through a panel of normals to remove potential sequencing artifacts and undetected germline variants. MutSig2CV[144] was run on these filtered results to evaluate the significance of mutated genes and estimate mutation densities of samples. These results were constrained to genes in the Cancer Gene Census,[163] with false discovery rates (q values) recalculated. Genes of q value < 0.1 were declared significant.

Mutational signatures using SignatureAnalyzer

Mutational signatures were extracted for all cohorts, excluding MB since WES data was unavailable. This analysis was done using SignatureAnalyzer (https://github.com/getzlab/SignatureAnalyzer), a Bayesian Non-negative matrix factorization algorithm that infers an optimal collection of signatures from the data (additional details are provided in Note S1). We performed a multi-step signature extraction workflow to mitigate noise and bias introduced by hypermutated samples with high leverage, an issue particularly exacerbated when solely using exomes.

We first focus on the 13 patients with missense mutations in the exonuclease domain of POLE or POLD1 (referred to as POLE-exo* or POLD-exo*) since they have a unique mutational signature that requires a special analysis. For these tumors, we applied SignatureAnalyzer using the 96 tri-nucleotide sequence context for single-base substitutions (SBSs) and a length-based context for insertions and deletions.[164] Analyzing these tumors using this particular spectra allowed us to detect 7 distinct signatures, 6 of which had strong similarity (cosine similarity > 0.75) with the previously reported signatures in POLE/POLD-exo* tumors.[164] In order to control for a single DNA repair deficiency in our study, we do not include POLE/POLD-exo* samples in downstream MMRD (i.e., MSI) analyses.

We then applied SignatureAnalyzer on the remaining 1056 samples using the standard composite spectra that includes the 96 tri-nucleotide sequence context for single-base substitutions (SBSs), 78 double-base substitutions (DBSs), and 83 indel features (ID), as performed in Alexandrov et al.[44] Among the extracted signatures was one resembling a signature (SBS15) associated with MMRD in previous studies (cosine similarity 0.82). We inspected the distribution of this MMRD signature in COAD and UCEC samples, which are known to have higher incidences of MMRD (i.e., MSI). The minimum signature weight in these samples with evidence of mismatch repair deficiency, as evidenced by their non-trivial contributions from the signature, was 72 mutations, and 19% of these mutations were attributed to this MMRD signature. Thus, we set a threshold of 72 mutations and 19% of mutations from this SBS15-like signature to identify MMRD tumors across our dataset (Figure S3B). Finally, since several known signatures associated with MMRD were merged together in the composite analysis, we re-ran SignatureAnalyzer using only the 96 SBS features on the MMRD samples and indeed recapitulated (cosine similarity > 0.85) the known MMRD-associated signatures from the COSMIC v3 catalog of signatures.

We partitioned the remaining patients into three groups by their tissue type: tobacco smoking related included LSCC, LUAD, and HNSCC; homologous recombination deficiency (HRD) related included BRCA, OV, and PDAC; and non-homologous recombination deficiency related included CCRCC, COAD, GBM, and UCEC. We used the tri-nucleotide SBS context to extract signatures from the HRD related and non-HRD related tumors, and we characterized these signatures by computing cosine similarity with the COSMIC v3 signatures.

Tobacco smoking and UV exposure (in HNSCC cancers) serve as primary mechanisms of mutagenesis in the tobacco smoking related group, and these processes have been shown to cause a wide range of single and double base substitutions as well as indels.[44] Thus, we employed a composite spectra to isolate the effects of the various mutational processes in these patients, and characterized them by computing their cosine similarity to those found in the PCAWG study.[44]

Identifying homologous recombination deficiency using whole exome data remains challenging due to its flat single-base substitution landscape as well as a reduction in microhomology coverage in exomes. Further complicated by the composite reference's whole genome calibration, we opted to use the trinucleotide SBS spectra for this group of patients, as well as for the remaining COAD, UCEC, CCRCC, and GBM patients.

As previously described, whole exome sequencing provides less power for non-negative matrix factorization, such that signature bleeding can present particularly misleading attributions for low mutation burden patients. In order to more confidently classify HRD tumors, we leveraged germline exome data to identify tumors with known pathogenic variants from the ClinVar database or variants considered as "HIGH IMPACT" by ClinVar on *BRCA1*, *BRCA2*, and *PALB2*. We found that breast and ovarian tumors with these variants that also showed relatively high contributions from the HRD mutational signature had more than 45 mutations attributed to the

signature (Figure S3B). Thus, we set our threshold for HRD classification to any tumor with greater than or equal to 45 mutations attributed to the HRD mutational signature.

HRD vs HRP Analysis

We note that classifying HRD tumors using only exome somatic and germline mutation data remains challenging, particularly due to substantially low power to detect microhomology indels, a crucial component to HRD detection.[165] To prevent over-classification of HRP tumors for this analysis, we selected the bottom 20% of samples based on their HRD mutational signature contributions from both HGSC and BRCA cohorts. In order to limit repair deficiencies to HR, we excluded any MMRD or POLE exonuclease mutant tumors. In addition, due to the small number of PDAC classified as HRD (4 patients), we focused this analysis on breast and ovarian HRD tumors.

Differential Expression

RNA

Differential expression was performed using Limma-Voom.[166] The trimmed mean of M values (TMM) between-sample normalization[167] was applied to counts using calcNormFactors, the voom transformation was applied using limma::voom, and limma::lmFit was used for the moderated t-test, followed by empirical bayes shrinkage with limma::eBayes. The cancer type (i.e., cohort) was modeled as a fixed covariate in each analysis unless otherwise specified. FDR is computed using the Benjamini-Hochberg procedure. Versions used were edgeR_3.28.1, limma_4.32.2, and R version 3.6.1.

Protein, Phosphorylation, Acetylation

Differential expression was performed using Limma on median-MAD normalization matrices output from SpectrumMill v 7.08. limma::lmFit was used for the moderated t-test, followed by empirical bayes shrinkage with limma::eBayes. No imputation was performed prior to differential expression analyses. Proteins & PTM-sites were filtered out if they were present in <10 patients in either group being compared. The cancer type (i.e., cohort) is modeled as a fixed covariate in each analysis unless otherwise specified. Versions used: limma_4.32.2 and R version 3.6.1.

fGSEA

Gene-set enrichment analysis was performed using fGSEA (http://bioconductor.org/packages/release/bioc/html/fgsea.html). To evaluate RNA- and protein-level enrichment in the context of differential expression analyses, we ranked genes by the product of $\log_2$(Fold Change) and -$\log_{10}$(P-value). We analyzed the enrichment of phosphorylation and acetylation of gene sets in differential expression analyses by collapsing each gene to its maximum absolute value product of $\log_2$(Fold Change) and -$\log_{10}$(P-value) of all corresponding peptides. To apply fGSEA to the results of SignatureAnalyzer, we performed ranked gene-set enrichment analysis by ranking the factor weights (i.e., the weights within each signature were used to rank the genes to carry out ranked gene-set enrichment analysis for each signature).

Dedicated tools for PTM analysis

Additional details and use cases for the PTM dedicated tools can be found in Note S2.

PTM-SEA

We evaluated site-specific phosphorylation pathway enrichment using PTM-SEA (https://github.com/broadinstitute/ssGSEA2.0). Enrichment was run using the sites' flanking amino acids, using the ptm.sig.db.all.flanking.human.v1.9.0.gmt database. We employed the heuristic method introduced by Krug et al.[36] to deconvolute multiple phosphorylated peptides to separate data points (for differential expression analyses, ranks were based on -10 * $\log_{10}$ (p value) * sign($\log_2$(fold change)); for SignatureAnalyzer results, the factor weights of phosphorylation features were used at the rank).

CLUMPS-PTM

Mapping Sites

Every PTM site detected in this study was first mapped to UniprotKB to select for canonical isoforms and for ease of downstream alignment with PDB structures. *blastp+* was run on every fasta sequence from the CPTAC reference (RefSeq) used for mass-spec quantification for Spectrum Mill and queried to the entire UniprotKB Sequence database. To select the appropriate hit for each RefSeq ID blasted to UniprotKB, the top hit by identity overlaps from *blastp+* was selected after filtering for Uniprot IDs found in the SIFTS protein database. The SIFTS protein database was used for highly curated annotations between Uniprot protein IDs and their respective PDB structures. Next, each matched Uniprot was selected, and every matching PDB in the SIFTS protein database was checked for sequence overlap. The DBREF entry in each PDB header was used to identify the offset between the Uniprot sequence and PDB sequence. Finally, the PDB entry with the most overlap between sites found in CPTAC and longest sequence length was selected as the matching PDB entry.

Algorithm

The method is based on the CLUMPS method (CLUstering of Mutations in Protein Structures) for detecting significant clusters of mutations in 3D protein structures.[168] Here we search for significant clustering of differentially acetylated/phosphorylated sites. For sites mapped to an individual PDB structure, we computed an initial weighted average proximity (WAP) score based on matched PTM sites. Succinctly, WAP scores (see formula below) are a summation of all residue pairs (*q* and *r*), weighted by the product of the

strengths of each individual residues (n_q or n_r, respectively) and their distances (with a Gaussian decreasing weight using a scale parameter t) in the 3D protein structure. The original CLUMPS was designed to analyze mutations and the strength of each residue was the number of mutated patients at that residue. In CLUMPS-PTM, we use the differential phosphorylation/acetylation of each PTM-modified residue. The formula is:

$$WAP = \sum_{q,r} n_q n_r e^{-\frac{d_{q,r}^2}{2t^2}}$$

where q and r are protein residues; $d_{q,r}$ is the Euclidean distance (in Å) between the centroids of these residues; n_q and n_r are the weights of each modification (i.e $\log_2$(fold-change) * -$\log_{10}$(FDR) in CLUMPS-PTM); and t is distance scale parameter (in Å). Empirical p-values are calculated by permuting the altered residues in the protein. In CLUMPS-PTM, we only test permutations of PTMs to other PTM-possible sites (e.g., only lysines are sampled for acetyl-sites). In our analysis, we wanted to average across a range of scale parameters and therefore used t=3, 4.5, 6, 8, 10, and run 10,000 permutations for each, and generate a mean empirical p-value across these parameter values. Finally, we used sites that mapped to the PDB structure and were differentially expressed ($\log_2$(fold change) >0; see Differential Expression). Two steps were taken before FDR correction of the empirical p-values. First, if the empirical p-values were zero, representing an insufficient number of permutations, we set the p-values to 0.1 over the number of permutations. Next, we reduced the number of hypotheses tested by excluding proteins from the FDR procedure if the most significant p-value that the protein and associated sites could theoretically yield is greater than 0.1. The minimal theoretical p-value is given by $\frac{1}{\binom{N_{sites}}{C_{modifiable\ residues}}}$.

Finally, the Benjamini-Hochberg procedure was used to correct empirical p-values for multiple hypothesis testing. The code to perform all analyses is available in a python package at https://github.com/getzlab/CLUMPS-PTM.

The Kinase Library

Serine/threonine kinase substrate specificity assays: Assays, matrix processing, and scoring process were previously described at Johnson et al.[34]

The Kinase Library enrichment analysis

The phosphorylation sites detected in this study were scored by all the characterized kinases (303 Ser/Thr kinases), and their ranks in the known phosphoproteome score distribution were determined as described above (percentile score). For every non-duplicate, singly phosphorylated site, kinases that ranked within the top-15 kinases for the Ser/ Thr kinases were considered as biochemically predicted kinases for that phosphorylation site. Towards assessing a kinase motif enrichment, we compared the percentage of phosphorylation sites for which each kinase was predicted among the significantly downregulated/upregulated phosphorylation sites (i.e., sites with FDR $\leq$ 0.1), versus the percentage of biochemically favored phosphorylation sites for that kinase within the set of unregulated (non-significant) sites in this study (sites with FDR > 0.1). Contingency tables were corrected using Haldane correction (adding 0.5 to the cases with zero in one of the counts). Statistical significance was determined using one-sided Fisher's Exact test, and the corresponding p-values were adjusted using the Benjamini-Hochberg procedure. Kinases that were significant (FDR $\leq$ 0.1) for both upregulated and downregulated analysis were excluded from downstream analysis. Then, for every kinase, the most significant enrichment side (upregulated or downregulated) was selected based on the adjusted p-value and presented in the bubble plot. Bubble plots were generated with size and color strength representing the adjusted p-values and frequency factors respectively, only displaying significant kinases.

Modified lysine peptide library assays

Reagents used for the peptide library experiments include: Kinase substrate library (Anaspec); Streptavidin-conjugated membranes (Promega). A list of kinase information can be found in a previous study.[34] To determine the substrate motifs, we performed *in vitro* phosphorylation assays with recombinant kinases on an oriented peptide array library of design:

Y-A-X_{-5}-X_{-4}-X_{-3}-X_{-2}-X_{-1}-S_0/T_0-X_1-X_2-X_3-X_4-X_5-G-K-K-biotin in the presence of ATP[γ-^{32}P]. Unmodified, tri-methylated, and acetylated lysines were fixed in every position across the peptide. Reactions were carried out in their designated buffers plus 20 μM ATP and 0.4 μCi of (33 nM) [γ-^{32}P]ATP) at 30°C for 90 min. The peptides were spotted onto streptavidin-coated filter sheets (Promega SAM^2 biotin capture membrane) and visualized by phosphorimaging on Typhoon FLA 7000. Detailed information on the protocol is provided elsewhere.[111,116]

CausalPath

For each differential expression table that we generated in this study, we applied the CausalPath method[35,169] to identify possible cause-effect relationships between the detected differential values, and to understand the molecular signaling behind those changes. CausalPath uses literature-curated human mechanistic pathways to identify which protein activities have control over the observable features in the omics studies. The method uses this information to perform logical reasoning over the observed differences, and generates a network of causal relations. Generated networks are tested by data label randomization to identify enrichments that indicate activation/inhibition of certain regulators.

In these CausalPath analyses, we integrated global protein, phosphoprotein, acetylprotein and mRNAseq datasets, and used a 0.1 FDR threshold. To include biologically relevant edge cases, we included RAD18 S99, XRCC1 S475/S485/T488, and CDK1 T14 (all

FDR=0.102) for the acute versus chronic hypoxia HRD analysis, TOP2A S1247 (FDR=0.102) for the global HRD vs. HRP analysis, and PRKDC S3205 (FDR=0.1009) for the MMRD vs. MMRP analysis. To manage the complexity of the result networks and to focus on specific pathways, we generated sub-networks of results by taking graph neighborhoods of pre-defined gene sets encompassing DNA damage response (DDR). The DDR gene set (Table S3) includes the union of the Reactome DNA Repair gene set and the CPTAC DNA-Damage Response (DDR) Working Group gene set (unpublished). We also included CDK1/2/4/6, AURKA, AURKB, PLK1/2/3/4/5, and WEE1 to account for cell cycle checkpoint kinases.

We also used CausalPath to investigate the downstream effects of the kinase activity predictions made by the Kinase Library method. We inserted the predictions of the Kinase Library as custom hypotheses to CausalPath to find out which differential changes are compatible/explainable by those predicted activities.

We used the Newt software[170] to visualize the CausalPath results and to generate Figures 3E, 3H, and S3C.

Histone analysis

For Histone analysis we used 6 cancer types with available acetylation data (5 from the CPTAC cohorts): breast, uterine, glioblastoma, lung squamous, and lung adenocarcinoma in addition to one external dataset, medulloblastoma, that was generated and harmonized in the same manner as the CPTAC cohorts. For the global analysis, we included only acetylation sites that were detected in all cohorts. These summed up to a total of 61 histone acetylation sites: 21 sites on H1, 10 sites on H2A, 16 sites on H2B, 10 sites on H3, and 4 sites on H4.

Histone-level correlations

The imputed acetylation data was subset for histone genes by mapping RefSeq IDs to HGNC IDs, and within each group of histones (H1, H2A, H2B, H3, H4), the relative abundance of all sites within this group was averaged. Spearman correlations and p values were calculated for each of the 10 pairs.

Histone acetylation and smoking analysis

We used Spearman's rank correlation to evaluate associations between the tobacco smoking mutational signature and histone acetylation and deacetylase phosphorylation, and we used Benjamini-Hochberg to control the FDR. We additionally calculate the 95% confidence intervals for each correlation shown in Figure 5B through bootstrapping. We sampled from the data with replacement 10,000 times and computing the Spearman's correlation coefficient each time. Using this distribution of correlation coefficients, we determined the 95% confidence interval for the correlations between the PTM sites and tobacco smoking mutagenesis. We limited this analysis to LUAD tumors, since this cancer type presents a broad spectrum of tobacco smoking mutagenesis. Due to the strong correlation between gender and tobacco smoking mutational signature weights as well as smoking history ($p=1x10^{-4}$, Figure S5B), we further limit this analysis to the 71 cancers from male patients (which are the majority of the smokers) in order to decouple these variables. Additionally, we computed the Spearman's rank correlation between smoking signature weights and normalized enrichment scores from mRNA-based ssGSEA using the MSigDB HALLMARK gene sets. We used the 71 male LUAD and 86 male LSCC for this analysis in order to prevent tissue-driven biases at low tobacco smoking mutation levels from other cancer types that typically have relatively low levels of tobacco smoking mutagenesis.

Lasso regression analysis

Acetylation sites with detection rates of at least 80% across the entire cohort were collected, and the abundances of these sites were imputed using the "impute.knn" function from the R package "impute" using the default parameters (k = 10, rowmax = 0.5, colmax = 0.8, maxp = 1500, rng.seed = 362436069). For the protein abundance of histones and histone regulators, the abundance of proteins mapping to multiple RefSeq IDs were averaged. Within each cluster, we tested the association between the abundance of histone acetylation sites and the abundance of histone acetyltransferases (HATs), histone deacetylases (HDACs) and bromodomain proteins (BRDs) using lasso regression. In particular, each histone acetylation site was treated as the outcome variable with the HATs/HDACs/BRDs protein treated as covariates. We split the data so that 80% (training set) are randomly designated for fitting the lasso regression model and the other 20% (testing set) are used to test its performance. Lasso uses a regularization parameter, λ, to control the weight of the L1 norm in the Lasso cost function. Within the training set, we identified the site-specific optimal λ value with 10-fold cross-validation that minimizes the mean squared error (MSE) for each acetylation site in each bootstrapping. We also calculated the lasso coefficient for each outcome-covariate pairs, where a non-zero lasso coefficient indicated that the covariate would have a significant effect on the outcome. We predicted the abundance of each acetylation site in the testing set using the lasso coefficient generated from the training set and calculated the test MSE between predicted value and actual value. We bootstrapped these processes for 100 replications in order to minimize any bias introduced by the random sampling. We reported average lasso coefficients (Table S5) from 100 bootstrapping of lasso regression to improve precision in variable selections. To demonstrate the reproducibility and stability of the variable selections, we computed the mean and standard error of the test MSE across all histone acetylation sites separately from 25, 50, 75 and 100 replications. The test MSE was generated from the best model selected using optimal λ in each bootstrapping of lasso regression. The mean and standard error of the MSE were averaged across all histone acetylation sites, resulting in values of 1.14 ± 0.78, 1.14 ± 0.79, 1.15 ± 0.79, 1.15 ± 0.79, across 25, 50, 75, and 100 replications, respectively. We also reported site-specific mean MSE along with their standard errors in Table S5. This demonstrates that the mean MSE is very stable across different numbers of replications.

Differential acetylation of histone sites across dendro groups

The results from the global differential expression analysis (see differential expression) were used to subset histone genes by mapping RefSeq IDs to HGNC IDs. Sites with $q < 0.1$ were considered significant.

Transcriptional signature analysis

Across all samples, the correlation between the ssGSEA scores of HALLMARK pathways and histone acetylation sites was assessed using the imputed PTM data. For each histone acetylsite-pathway pair, we computed the Spearman correlation and p value. P values were then adjusted to FDR values across the 50 HALLMARK pathways for each histone acetylation site tested.

Acetylation-phosphorylation crosstalk analysis

For each acetylation site detected, phosphorylation sites with the same RefSeq protein ID, and within 5 amino acid residues of the acetylation site, were collected. For each cluster and across all samples, Spearman correlations were then computed for each adjacent acetylation sitephosphorylation site pair and p-values were adjusted to FDR values across all site pairs tested.

Histone acetylation across immune groups

The imputed acetylation data was filtered for histone acetylation sites by mapping RefSeq IDs to HGNC IDs. The abundance of these sites was then compared between the Immune-cool samples versus all other samples using a two-sided Wilcoxon test. P values were corrected to FDR values, and sites with FDR < 0.1 were considered significant.

SignatureAnalyzer

Efficient algorithms have been developed by the machine learning community for matrix factorization with a strict non-negativity requisite. One of the benefits of using a non-negative factorization method is that it typically yields sparse solutions (i.e., fewer 'signatures' are used to represent the data of one tumor), which typically have clear interpretations. Non-negative Matrix Factorization (NMF) is an algorithm that computes a low-rank factorization of an input matrix V into two matrices W and H, such that $V \sim W * H$ and all elements of V, W, and H are strictly non-negative. The W matrix is a N-by-k matrix, where N is the number of features, and k is the number of signatures. This matrix provides the weights for each feature in each of the signatures, and thus the columns of W are called the signatures. The H matrix is a k-by-S matrix, where k, as before, is the number of signatures, and S is the number of samples. This matrix provides the contribution of each signature in each sample. This matrix can be normalized for downstream analyses such that the sum of contributions for a given sample sums to 1.

NMF was originally developed for analysis of image data[171] but has a long history of use in biology to deconvolute expression data[172] and identify 'meta genes,' which we now call 'signatures'. NMF was later used to deconvolute mutation profiles and identify mutational signatures that were then matched with different mutational processes.[44] The use of non-negative methods in the mutation space is intuitive since both mutation counts, and mutation rates, are positive numbers, and the results have a clear interpretation (i.e., the number of mutations a particular process contributes to each tumor). In these cases, the input V is a matrix of transcriptomic or genomic features by samples, respectively. The derived W matrix represents signatures and their feature weights, while the H matrix represents signature contributions to each sample's observed phenotype.

A common challenge with NMF is the tendency to overfit the data given an increasing input number of signatures, K. SignatureAnalyzer uses a Bayesian version of NMF, called automatic relevance determination-NMF (ARD-NMF),[160] which assumes a sparse prior on both the W and H matrices and searches for a maximum posterior solution. It initializes the analysis with a large number of signatures ($K_0 = 50$ for multi-omic NMF and $K_0 = \#$ features for mutational signatures) and by virtue of the priors, signatures unnecessary to explain the data have their weights driven to zero in the respective columns/rows of the W/H matrix. We run SignatureAnalyzer multiple times (n=100) and obtain a distribution on the number of signatures (as it converges to different local maxima of the posterior distribution). We select k as the mode of the distribution and the W and H matrices that correspond to the highest posterior solution with that value of k. This allows highly interpretable and sparse representations for both signature profiles and attributions that strike a balance between data fitting and model complexity.[160] This method, among others, has been widely applied to analyze mutational signatures and expression signatures for molecular data (mRNA, protein, etc).[43,44,164,172–174]

In this study, we apply SignatureAnalyzer to somatic mutation data to extract mutational signatures, and to the multi-omic phenotypes to identify patterns of regulation across cancer. Mutation counts are by nature non-negative, and we can adjust the multi-omic phenotype matrix by splitting each gene, protein, and phosphorylation feature into a positive and negative row in the matrix. We negate the negative row such that all values are non-negative. This creates a less dense matrix since every feature is replaced with two values, one of which is zero. The sparse matrix can be easily handled by NMF. At the end, the resulting signatures also contain rows corresponding to positive and negative values, and therefore the possibility of upregulation and downregulation is maintained. Since mutations can be represented as count data generated by independent, low frequency events (with the exception of clustered mutations such as APOBEC generated kataegis), we can approximate these mutation counts with a Poisson distribution. For multi-omic data, the Poisson assumptions do not hold, and we model the likelihood of the zero-centered data using a truncated (half) Gaussian distribution (Figure S1C).

In order to reward sparsity of the resulting W and H matrices, we apply an exponential prior to all entries in the W and H matrices for mutational signatures, and a half-normal prior to all entries for the multi-omic signatures. Applying an exponential prior to the weights

in the W and H matrices is equivalent to adding an L1 regularization term to the cost function being optimized. To illustrate this point, consider the pdf of the exponential distribution:

$$f(x;\lambda) = \lambda e^{-\lambda x}$$

In our cases, x represents an entry in the W and H matrices of the factorization. As SignatureAnalyzer follows the procedure in Tan et al.[160] to minimize a cost function given by the negative log posterior, this term is incorporated as follows:

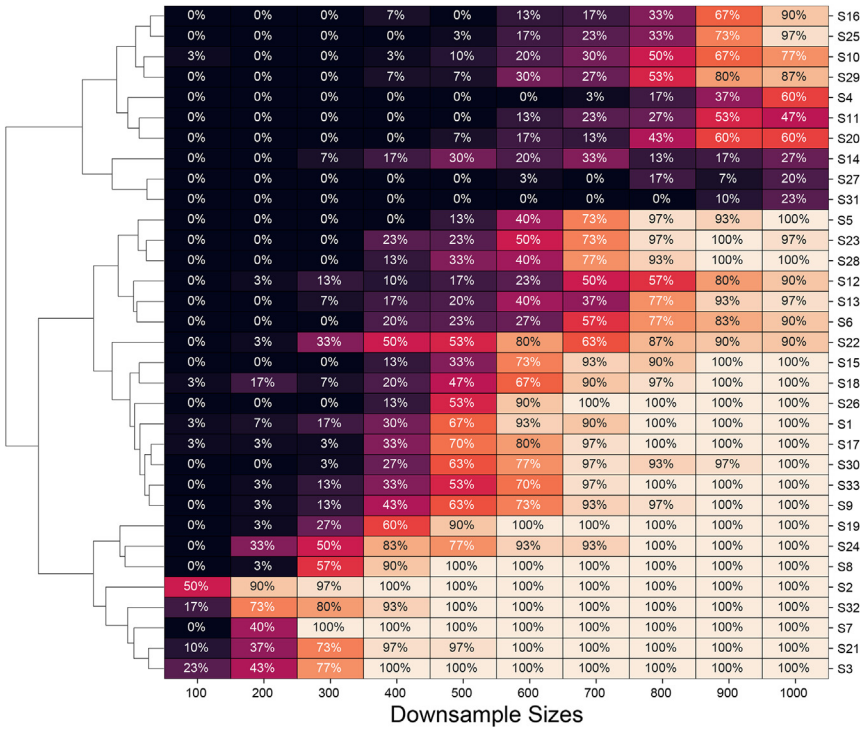
$$-\log f(x) = -\log \lambda + \lambda x$$

Since -logλ is constant with respect to the entries of the W and H matrix, the choice of an exponential prior effectively puts an L1 penalty on the value of x with regularization strength determined by λ. Note that L1 penalty uses the absolute value of x, ie. |x|, but since all values in the matrices are positive, we do not need the absolute value. We can work out similar math for the half normal case to see that this is indeed equivalent to L2 regularization.

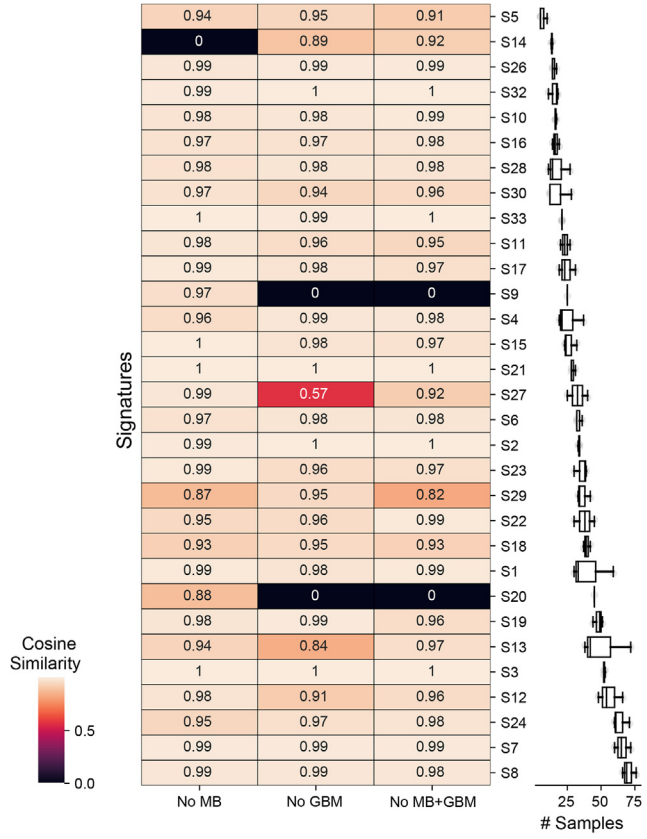
$$f(x;\sigma) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

$$-\log f(x;\sigma) = \frac{1}{2\sigma^2}x^2 + const$$

Mutational signature discovery requires deconvolution of somatic mutation counts, stratified by mutation sequence context, into a set of characteristic patterns (signatures) and inferring their activity across each of the samples in the study. The common stratification of somatic single nucleotide variants (SNVs) is based on the six base substitutions (C>A, C>G, C>T, T>A, T>C, and T>G) within the trinucleotide sequence context, which includes the bases located at positions immediately adjacent to the 5' and 3' directions relative to the mutated base. Thus, each of the six base substitutions have 16 possible combinations of neighboring bases, yielding 96 possible mutation types (or sequence contexts). The input matrix for mutational signature analysis is a 96 x S matrix, where S is the number of samples, and each element x_{ij} represents the number of observed mutations of type i in sample j. Using this matrix, we identified the mutational signatures and their contributions to each sample. This type of analysis has been used in numerous publications starting with Lawrence et al.[144] and most recently applied across cancer in the TCGA/ICGC Pan-Cancer Analysis of Whole Genomes (PCAWG) project.[44] We compared the signatures that we obtained to the ones reported in the COSMIC database and PCAWG paper in order to connect them to their proposed etiologies. We then used these signatures to classify tumors with mismatch repair deficiency and homologous recombination deficiency based on the activity of mutational signatures associated with these repair defects across all of our samples.



To provide additional internal validation for the robustness of the subtypes and test the effect of having fewer samples, we ran SignatureAnalyzer on downsampled cohorts of sizes N=100, 200,...1000 from our full Pan-Cancer cohort (n=1110). We performed 30 random downsamplings without replacement per value of N ("runs"), and within each run, we performed 20 independent Bayesian NMF iterations (each initialized with a unique random seed). We then matched the signatures derived in each downsampling run with the signatures obtained when analyzing the full cohort using cosine similarity with a threshold of 0.8. The percent of downsamples out of 30 runs in which a particular signature was identified is plotted in the heatmap below. As shown in the heatmap, even when downsampling to 800 samples (~72% of the dataset), 24 signatures are seen in >1/2 of the runs. Moreover, signatures S24 and S7 (which are the dominant signature in a large number of samples; 71 and 73 samples, respectively) are detected even in very small subsampled cohorts, making them more robust signatures. On the other hand, signatures S27 and S31 are less robust (i.e., "weaker") and are only found in a minority of downsampling runs, even with large cohort sizes (800-1000).



Further exploring the effect of cohort composition on our signatures, and specifically our external MB cohort, we investigated how the inclusion of MB impacts our signatures. To this end, we reran SignatureAnalyzer excluding (i) the MB cohort, (ii) the GBM cohort, and (iii) both MB and GBM cohorts. We provide a heatmap showing the cosine similarities between the signatures derived from each of these cohort exclusion runs, and the signatures derived in the full dataset, colored by cosine similarity. The boxplot to the right of the heatmap indicates the number of samples assigned to each signature per decomposition (assignment based on maximally contributing signature). First, we find that Signature 31 is not derived in each of these decompositions, consistent with downsampling analyses indicating its lower robustness ('weakness'). Next, we see that upon removal of the MB cohort, all signatures with the exception of S14 are derived with >0.85 cosine similarity to the full cohort signatures, suggesting MB itself does not yield a significant effect on the signatures. S14 is not derived when MB is excluded, but this signature is the second "weakest" signature derived from downsampling and is mixed across multiple cohorts (maximal contributions to S14 are with 47% from UCEC, 29% from CCRCC, and then 12% from MB) defined by strong transcriptomic activation of EMT (Table S2). Interestingly, we find Signatures 9 and 20 are not derived when excluding GBM and both brain cohorts (MB and GBM). S20 is a weaker signature with a 56% dominance of GBM. S9, however, is a stronger signature split between MB (34.29%) and GBM (65.71%), defined by downregulation of MYC with strong positive transcriptomic and negative phosphorylation effects. This suggests that S9 is a tissue-specific signature of biologic interest rather than a specific batch effect. S7 (immune upregulation) and S22 (myogenesis/adipogenesis in phosphoproteome) are examples of signatures without any MB samples. Notably, for these two signatures, 1 GBM patient (out of 73 in S7) maximally maps to S7 and 1 GBM patient (out of 39 in S22) maximally maps to S22 (Table S2). We believe more samples would be required to definitively suggest if these signatures are truly devoid of MB specific samples, or lack biological pathways associated with brain tissue.

PTM dedicated tools

To explore and interpret site specific PTM data, we developed and applied PTM-tailored tools. We provide here a short description of each:

CLUMPS-PTM

PTMs play key roles in numerous cell signaling processes in the context of cancer. Previous studies have shown that close proximity of PTM sites, either in linear sequence or on the 3D protein structure, increases the chance of crosstalk among such sites and may have functional outcomes.[175–177] Despite advances in LC-MS increasing the throughput of PTM quantification, identifying these functionally relevant PTMs and protein domains remains a challenge. We developed *CLUMPS-PTM*, an algorithm for spatial clustering of PTMs, to identify clusters of correlated PTMs in protein structures that likely reflect stronger, mutually dependent signaling effects (Figure S2D). *CLUMPS-PTM* is built on a previous tool we developed, *CLUMPS*,[168] initially designed to identify novel driver mutations based on clustering of missense mutations in 3D protein structures. *CLUMPS-PTM* was developed for phosphorylation and acetylation events, but is flexibly designed to incorporate any PTM modifications (such as ubiquitination, etc.). We developed pipelines (compatible with RefSeq and GENCODE isoforms) to map over 14K phospho-sites and 13K acetyl-sites detected in the CPTAC dataset to protein data bank (PDB) files (Figure S2E). We expand this to the predicted AlphaFold proteome and recover an additional 12K phospho-sites and 10K acetyl-sites exceeding model confidence (pLDDT) of 70%. Consistent with previous research,[178] phosphosites are found on residues with lower AlphaFold prediction confidence than acetylsites due to their abundance in unstructured domains ($p<1e{-}4$). *CLUMPS-PTM* may be run on either a database of PTM sites or differential expression results. This tool has been applied to multiple CPTAC projects[7] in addition to the Pan-Cancer context to propose functional domains for 1) tumor vs. normal analyses, 2) immune infiltration / subtyping, 3) phospho/acetyl co-clustering, and 4) tumor-specific subtyping. Overall, *CLUMPS-PTM* is an open-source tool that allows near proteome-wide spatial analysis with the growing availability of PTM data. We anticipate that it will be useful to the broader proteomic community for the discovery of novel targets and generation of insights into functional mechanisms (Figure S2F; STAR Methods).

The Kinase Library

The Kinase Library experimentally characterizes the substrate sequence specificity of over 300 protein Ser/Thr kinases.[34] This dataset is then used to computationally identify the most likely protein kinases for differentially phosphorylated sites and predict kinases that are up- or downregulated (STAR Methods). Moreover, this library also tests the change of sequence specificity when a close proximity acetylation is added and enables the study of crosstalk and its effect on kinase regulation. In addition, this method was also applied in our companion Driver paper[154] to characterize the activity states of kinases under different oncogenic driver events.

CausalPath

CausalPath[35] aims to identify potential cause–effect relationships between the observed omic changes (proteomic and transcriptomic) using a variety of prior information, such as (i) the regulation of protein modification (phosphorylation, acetylation, and methylation), (ii) the effect of the modified protein site on the protein activity, and (iii) the regulation of gene expression. In general, the method has two essential steps. In step 1, CausalPath works on the Pathway Commons database to identify which protein activities control which measurable omic features, and in what direction. In step 2, CausalPath uses a logical equation to detect the subset of the extracted priors that can causally explain the observed coordinated changes. The logical function links omic changes to protein activities and checks if the observed downstream changes are in the expected direction. On top of these steps, CausalPath runs enrichment analysis to detect the proteins on the result network with a significant number of affected downstream targets. CausalPath was previously used in a myriad of analyses and diseases such as: platelet activation,[179,180] lung squamous cell carcinoma,[7] glioblastoma,[4] acute myeloid leukemia,[181] pediatric high-grade glioma,[182] cell line drug treatments,[183] and breast cancer.[184,185]

PTM-SEA

PTM-SEA, similar to GSEA, is a pathway enrichment approach that identifies activated or deactivated regulators like kinases or phosphatases, using an enrichment of experimentally validated substrates.[36] Most proteomic tools are performed on a gene-centric level and collapse multiple sites to their average or dominant site; this process causes loss of important information encoded at the site-specific level. The tool is based on the PTMsigDB curated database with a large advantage of site-specific annotation and directionality of the PTM regulation.

QUANTIFICATION AND STATISTICAL ANALYSIS

The process of RNAseq data processing and quantification, as well as proteome quantification, has been outlined in the sections titled "RNAseq data processing and quantification" and "Proteomics data processing," respectively. The statistical analysis methodology and its corresponding details can be found both within the main text and in the relevant sections of the STAR Methods.

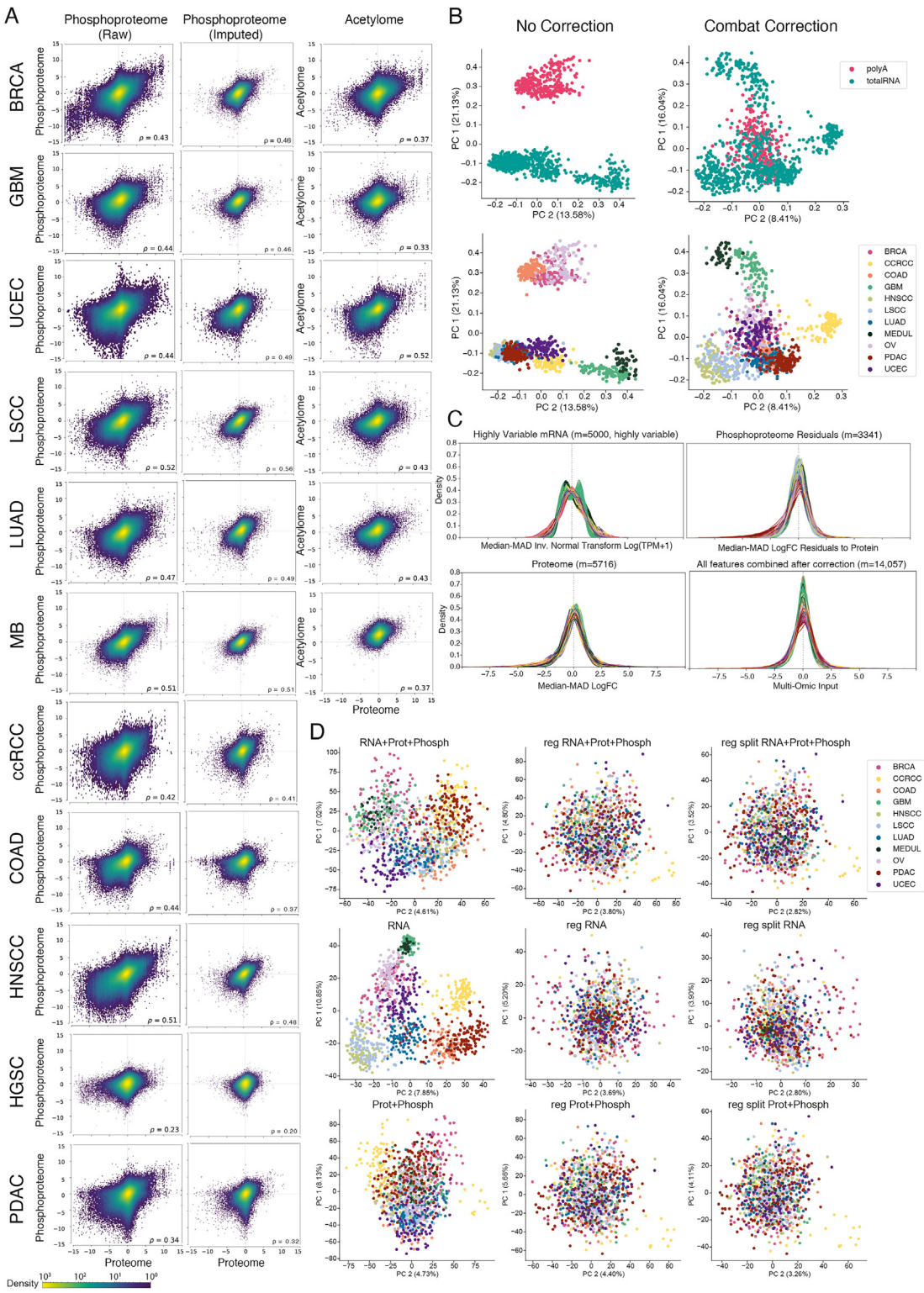
ADDITIONAL RESOURCES

Comprehensive information about the CPTAC program, including program initiatives, investigators, and datasets, are available at the CPTAC program website: https://proteomics.cancer.gov/programs/cptac.

For the Pan-Cancer proteogenomics collection papers, along with links to the data and supplementary materials associated with these publications, please visit the Proteomic Data Commons (PDC) at: https://pdc.cancer.gov/pdc/cptac-pancancer.

CellPress
OPEN ACCESS

Supplemental figures



(legend on next page)

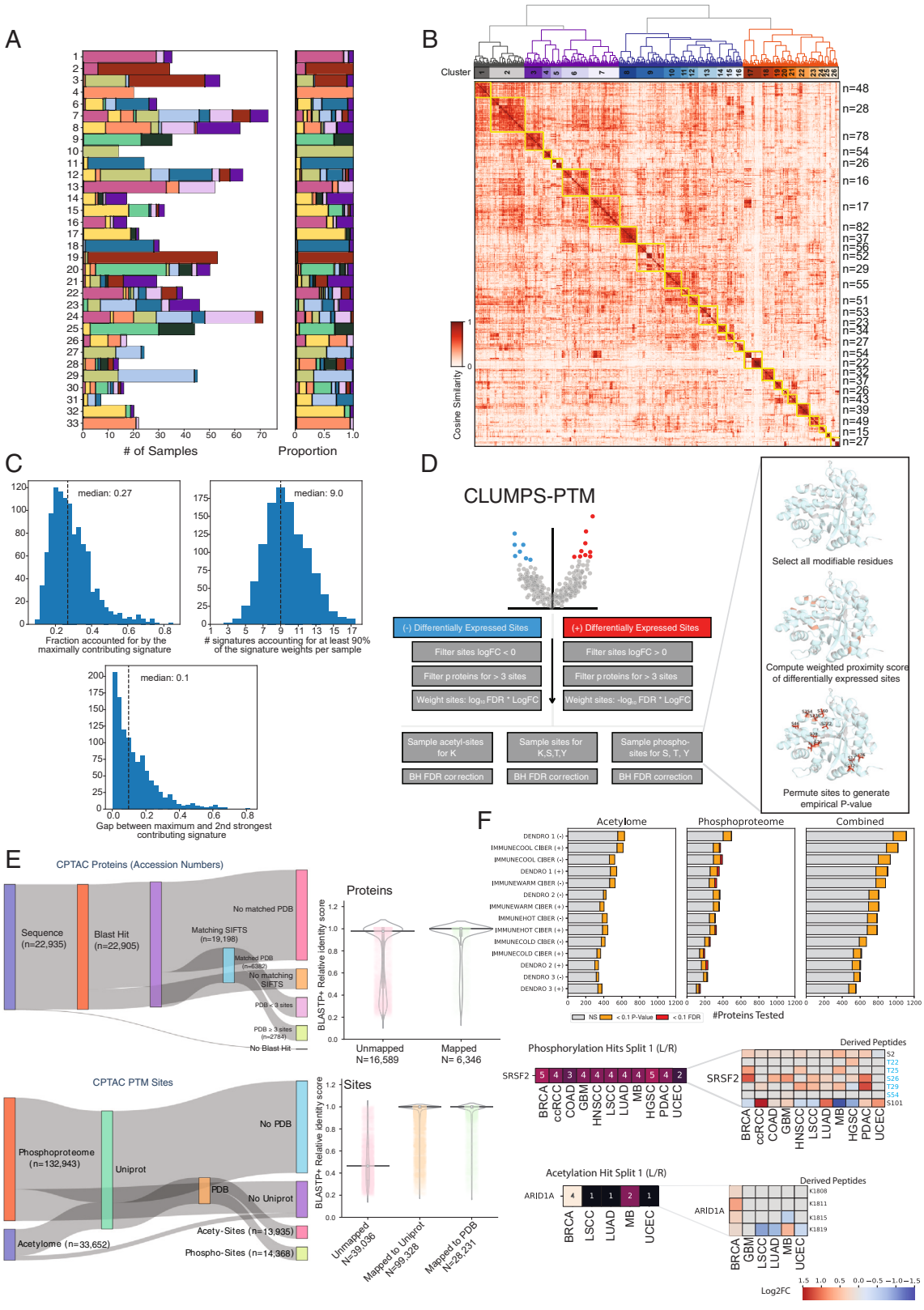
Figure S1. Pan-cancer data harmonization, related to Figure 1

(A) Density plots showing PTM levels and their matched protein levels. Ordinary least squares (OLSs) were fitted for each PTM-protein pair, and the residuals from the fit represented PTM levels corrected for protein abundance. Raw phosphoproteome abundance (left), imputed phosphoproteome abundance (used for clustering, middle), and raw acetylome abundance (right).

(B) Principal component analysis (PCA) projection of transcriptome TPM across the 10 CPTAC cohorts and the medulloblastoma cohort before and after applying batch correction (using COMBAT). Top PCA plots colored by sequencing protocol (poly-adenylated mRNA enrichment [red] and total RNA [blue]), and bottom PCA plots colored by cohort.

(C) Feature distribution showing the final normalized RNA data, the proteomic data, phosphorylation data corrected for protein abundance and the final multi-omic feature space following cohort correction. Distributions are colored by cohort.

(D) Projection of multi-omic features input into NMF on a PCA of all 1,110 samples colored by cohort; (top row) all features combined—mRNA, protein, and phosphorylation, (middle row) mRNA, (bottom row) protein and phosphorylation. (left + middle column). Features before and after regressing out cohort-specific effects, (right column) features after regression and split into positive and negative values for NMF.



(legend on next page)

Figure S2. Patient clustering and CLUMPS-PTM method, related to Figure 2

(A) Stacked barplot of patient-wise assigned maximally weighted signature colored by cohort (left). Stacked bar plot shows the breakdown, normalized by the number of patients maximally assigned to a given signature (right).

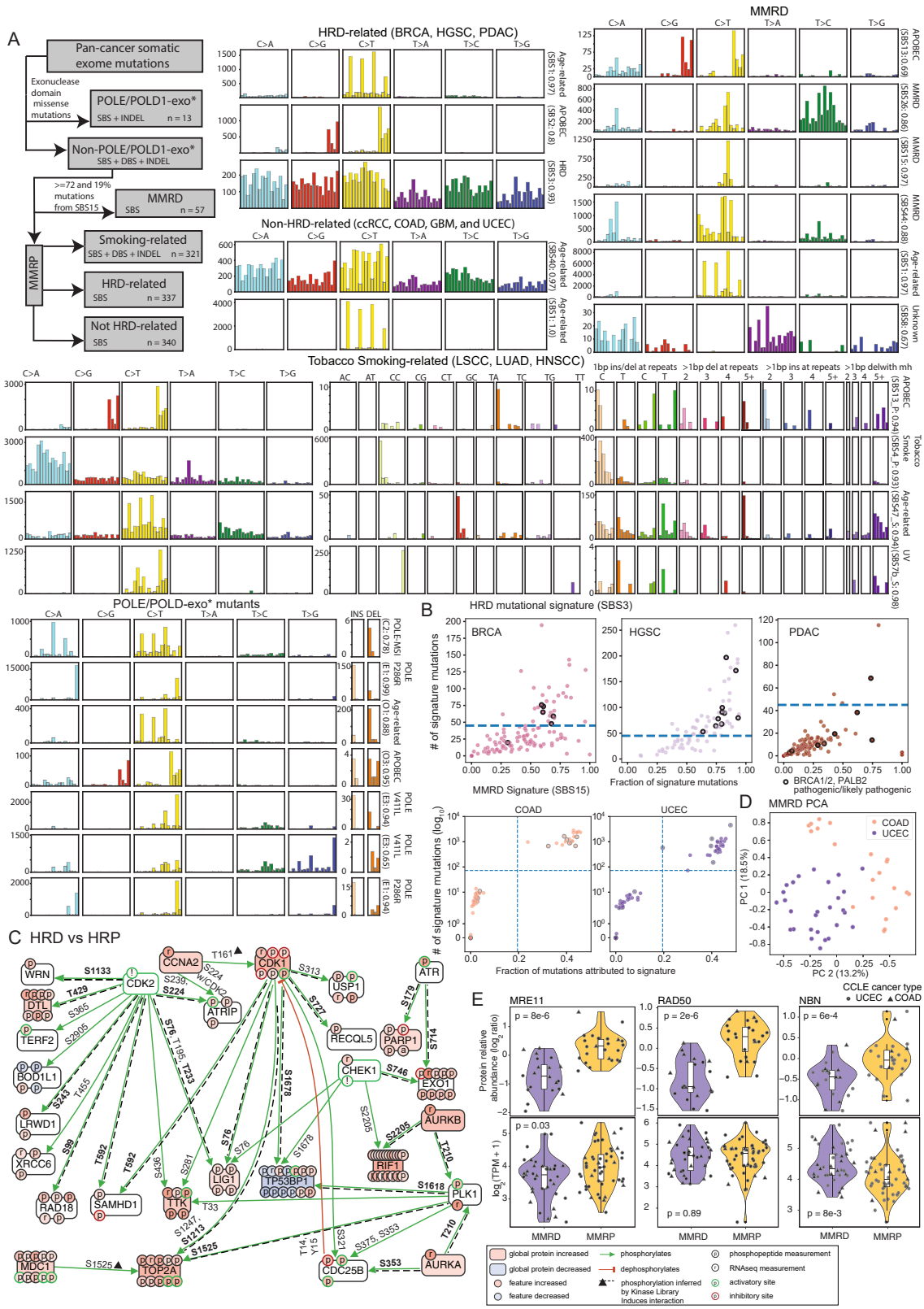
(B) Multi-omic hierarchical clustering of tumors based on multi-omic NMF signature weights. Heatmap represents the pairwise cosine similarity between tumors. Non-overlapping clusters, yellow boxes. Top track shows the cluster assignments. Dendrogram is colored based on the second level of the tree.

(C) Distribution of the fraction of a sample's NMF signature weights accounted for by the maximally contributing signature (top, left); distribution of the number of signatures required to explain at least 90% of a sample's signature weights (top, right); distribution of the difference between the fraction of a sample's NMF signature weights accounted for by the maximally contributing signature and the second strongest signature (bottom).

(D) Schematic workflow of the newly developed CLUMPS-PTM algorithm.

(E) Sankey plot showing the mapping between protein sequences to PDB structures. Protein sequences were aligned to the UniProtKB database (using BLASTP+ hits), then filtered to those with matching well-annotated proteins identified using the structure integration with function, taxonomy and sequence (SIFTS) method. These were then mapped to available crystal structures with at least 3 modified PTM sites across the entire CPTAC dataset (upper left). Sankey plot of all PTM sites in the CPTAC dataset, splitting to those with accession-numbers mapped to available Uniprot input proteins, and then to proteins with available PDB structures verified in the SIFTS database (lower left). Violin plots of the BLASTP+ relative identity score to the Uniprot database for proteins and PTM sites that map or do not map to PDB structures (right).

(F) Stacked bar plots summarizing CLUMPS-PTM results showing the total number of proteins tested based on group-specific differential expression test and direction (top). Stacks show the number of structures that are nominally significant ($p < 0.1$; yellow) and pass multiple hypothesis testing (FDR < 0.1; red) for acetylome, phosphoproteome, and combined samplers. CLUMPS-PTM phosphorylation and acetylation hits (q value < 0.1) for the left (L) vs. right (R) top split of the pan-cancer dendrogram (bottom). The first heatmap (on the left) indicates the number of distinct peptides found in each cohort that contributed to the differential PTM cluster. The x axis represents the cohort. The second heatmap (on the right) expands on each CLUMPS-PTM hit and shows the log-fold change between the left vs. right branches of the top split in the dendrogram for each represented peptide. Blue indicates phosphorylated clustered sites.



(legend on next page)

Figure S3. Mutational signature extraction and analysis of DNA repair deficiency, related to Figure 3

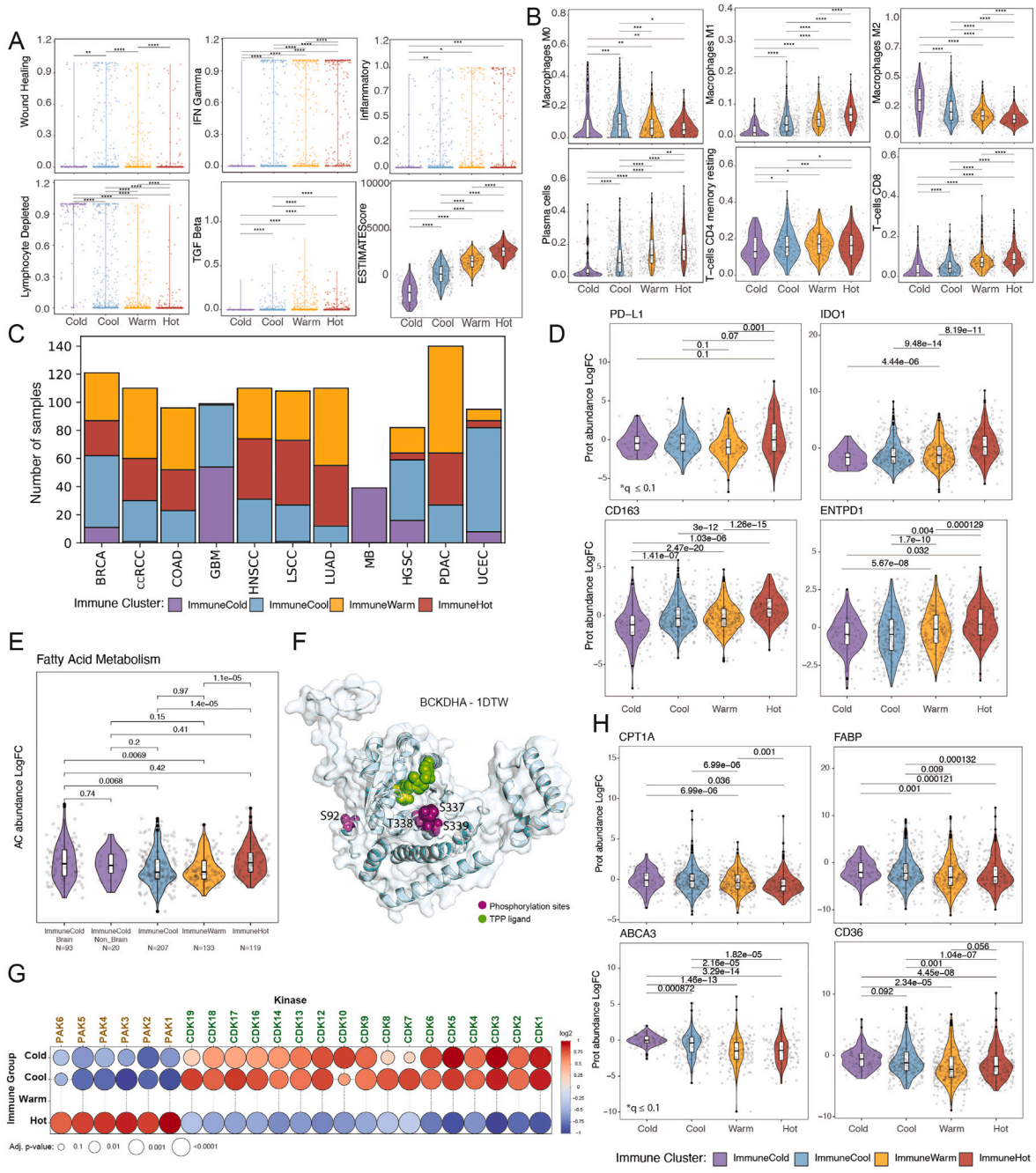
(A) Schematic of the mutational signature extraction workflow and barplot representations of the resulting mutational signatures.

(B) Scatter plots illustrating the distributions for the number of mutations and the fraction of mutations attributed to the HRD-associated mutational signature (top). HRD classification threshold at 45 mutations indicated by a dotted line. Tumors with known pathogenic or high impact germline variants in *BRCA1*, *BRCA2*, or *PALB2* genes are highlighted with a black outline; scatter plots showing the distributions for the number of mutations and the fraction of mutations attributed to the mutational signature associated with MMRD in COAD and UCEC tumors (bottom). MMRD classification thresholds at 72 mutations and at 19% of mutations, respectively indicated by horizontal and vertical dotted lines. Tumors with known pathogenic or high impact variants in MSH6, MSH2, MLH1, and PMS2 are highlighted with a black outline.

(C) Pathway of differentially expressed DDR genes (at the RNA [r circle], protein [box], and phosphorylation site [p circles] levels) between the HRD and homologous recombination proficient (HRP) tumors and their inferred effect on other pathway members. Upregulation in the HRD tumors in red and downregulation in blue. Dotted edges between the kinases and their substrates indicates that the interaction is within the 90th percentile of all top scoring substrates for the specific kinase (The Kinase Library).

(D) PCA of multi-omic signature weights across UCEC and COAD MMRD tumors. Points are colored by tumor type (COAD, orange; UCEC, purple).

(E) Distributions of MRN complex proteins (MRE11, RAD50, and NBN) relative protein abundance (top) and RNA expression in CCLE COAD and UCEC cancer cell lines from the DepMap dataset. Groups are separated and colored by mismatch repair proficiency (MMRD, purple; MMRD, yellow). Shapes indicate tumor type (UCEC, circle; COAD, triangle).



(legend on next page)

Figure S4. Immune clustering, activity, and associated metabolic markers, related to Figure 4

(A) Violin plots showing the levels of different immune signatures based on ImmuneSubtypeClassifier and purity estimates from ESTIMATE for the 4 identified immune clusters across all tumors.

(B) Violin plots showing the top 6 immune cell abundances from CIBERSORT that were used as covariates in differential expression to estimate tumor-intrinsic gene expression in the 4 immune clusters.

(C) Stacked barplot showing the distribution of 4 immune clusters within each cohort.

(D) Violin plots showing significant differences between the 4 identified immune clusters for the immunosuppressive markers—PD-L1, IDO1, CD163, ENTPD1 (pairwise comparisons with q value < 0.1 are shown).

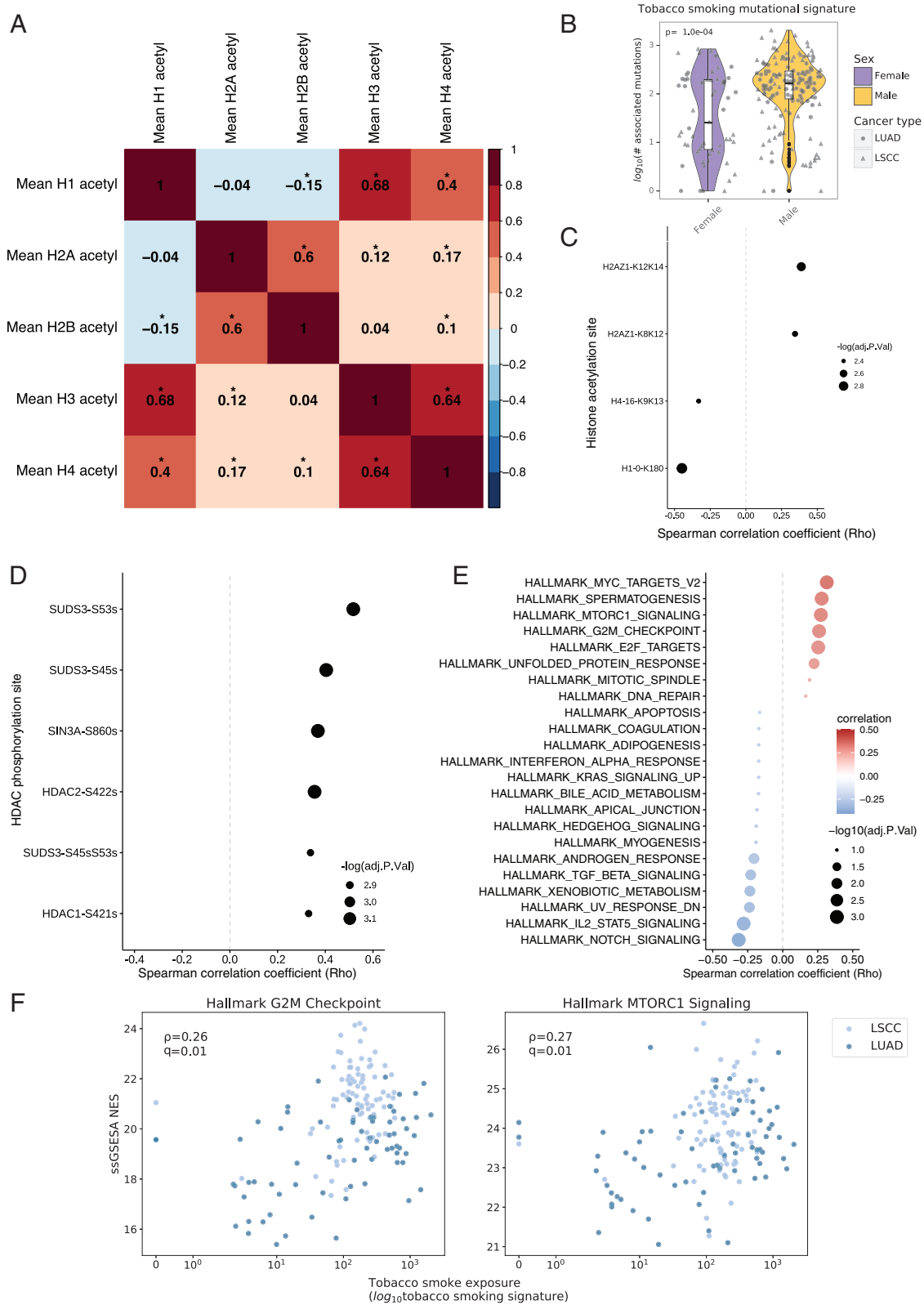
(E) Violin plots showing acetylation levels of fatty acid related proteins among different immune subtypes when immune cold is split to brain and non-brain tumors.

(F) BCKDK (PDB: 1DTW) showing significant spatial clustering of phosphorylation sites by CLUMPS-PTM of differentially regulated sites for the immune-cool cluster. Protein 3D crystal structure space-filling model (cyan) and highlighted clumped phosphosites (purple) and thymidine-5′-triphosphate (TTP) sites are marked in green.

(G) Bubble plot representing The Kinase Library enrichment score and significance, differentially expressed substrates of each kinase for the 4 immune clusters. Colors represent enrichment (red) and depletion (blue). Kinases are colored by to their phylogenetic groups.[117]

(H) Violin plots showing significant differences between the 4 identified immune clusters for fatty acid transporters—CPT1A, FABP, ABCA3, CD36 (only pairwise comparisons with q value < 0.1 are shown).

(I) Spearman correlation between the ssGSEA values for each sample of the MSigDB hallmark genesets to the CD8+ T cell abundance estimated from CIBERSORT.



(legend on next page)

Figure S5. Histone analysis, related to Figure 5

(A) Correlation matrix showing the correlations between mean acetylation of H1, H2A, H2B, H3, and H4. Asterisks denote correlations with $p < 0.05$.

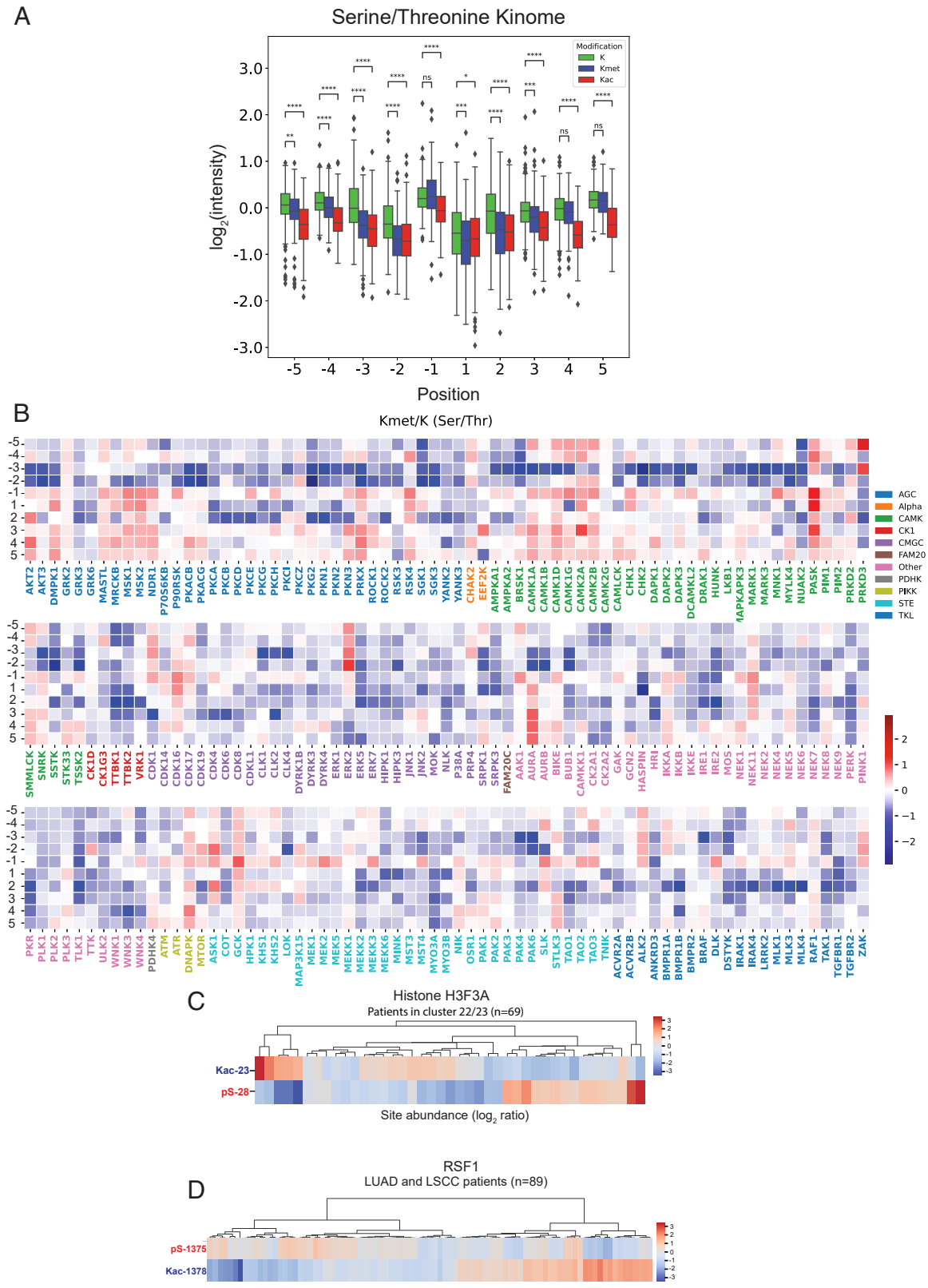
(B) Violin plot showing the distribution of tobacco smoking mutational signature contributions (log_{10}-scale) for males and females in LSCC and LUAD. Shapes indicate the cancer type—LUAD (circle) and LSCC (triangle). p value (1×10^{-4}) from Wilcoxon rank-sum test.

(C) Dot plot showing significant Spearman correlations ($q \leq 0.1$) across male LUAD patients (n = 71) between histone acetylation levels and tobacco smoking mutational signature contributions.

(D) Dot plot showing significant Spearman correlations ($q \leq 0.1$) across male LUAD patients (n = 71) between deacetylases phosphorylation levels and tobacco smoking mutational signature contributions.

(E) Dot plot illustrating significant Spearman correlations ($q \leq 0.1$, $|rho| > 0.15$) across male LSCC and male LUAD patients (n = 157) between tobacco smoking mutational signature contributions and the mRNA ssGSEA scores for the MSigDB hallmark genesets.

(F) Scatter plot showing the positive correlation (rho = 0.332) between tobacco smoking mutational signature contributions and the ssGSEA score for the G2/M checkpoint (left) and MTORC1 (right) hallmark geneset in male LUAD and male LSCC tumors (n = 157).



(legend on next page)

Figure S6. The Kinase Library, related to Figure 6

(A) Boxplots showing The Kinase Library normalized intensity ratios of 207 Ser/Thr kinases for unmodified, trimethylated, and acetylated lysines, across 10 positions (−5 to +5) around the central phosphoacceptor site.

(B) Heatmap showing The Kinase Library normalized intensity ratios between trimethylated lysine and unmodified lysine across the 207 Ser/Thr kinases. Kinases are colored according to their phylogenetic groups.[117]

(C) Heatmap showing the expression levels of acetylated K23 and phosphorylated S28 on H3F3A across 41 patients in cluster #8. Samples are sorted based on hierarchical clustering of their correlations.

(D) Heatmap showing the expression levels of acetylated K1378 and phosphorylated S1375 on RSF1 across the 89 LUAD and LUSC patients with acetylation and phosphorylation data. Samples are sorted based on hierarchical clustering of their correlations.