nature cancer

Resource

https://doi.org/10.1038/s43018-024-00817-x

A protein expression atlas on tissue samples and cell lines from cancer patients provides insights into tumor heterogeneity and dependencies

Received: 4 October 2023

Accepted: 5 August 2024

Jun Li[1], Wei Liu[1], Kamalika Mojumdar[1], Hong Kim[1], Zhicheng Zhou[1], Zhenlin Ju[1], Shwetha V. Kumar[1], Patrick Kwok-Shing Ng[2,8,9], Han Chen[1], Michael A. Davies[3], Yiling Lu[4], Rehan Akbani[1] ✉, Gordon B. Mills[5] ✉ & Han Liang[1,6,7] ✉

The Cancer Genome Atlas (TCGA) and the Cancer Cell Line Encyclopedia (CCLE) are foundational resources in cancer research, providing extensive molecular and phenotypic data. However, large-scale proteomic data across various cancer types for these cohorts remain limited. Here, we expand upon our previous work to generate high-quality protein expression data for approximately 8,000 TCGA patient samples and around 900 CCLE cell line samples, covering 447 clinically relevant proteins, using reverse-phase protein arrays. These protein expression profiles offer profound insights into intertumor heterogeneity and cancer dependency and serve as sensitive functional readouts for somatic alterations. We develop a systematic protein-centered strategy for identifying synthetic lethality pairs and experimentally validate an interaction between protein kinase A subunit α and epidermal growth factor receptor. We also identify metastasis-related protein markers with clinical relevance. This dataset represents a valuable resource for advancing our understanding of cancer mechanisms, discovering protein biomarkers and developing innovative therapeutic strategies.

Molecularly characterized patient tumors and cancer cell lines lay a critical foundation for modern cancer research, enabling the study of intertumor heterogeneity and oncogenic mechanisms, as well as the development of therapeutic strategies and biomarkers. Over the last decade, the cancer research community has made tremendous efforts to systematically characterize cancer samples through high-throughput molecular profiling data. Among these efforts, The Cancer Genome Atlas (TCGA) and the Cancer Cell Line Encyclopedia (CCLE) are the most widely used and deeply characterized resources, providing a comprehensive molecular portrait for ~11,000 patient samples across 33 cancer types[1,2] and ~1,000 cancer cell lines of diverse lineages[3–9]. Although DNA and RNA data from these resources have been extensively characterized, the parallel large-scale protein expression data remain limited. Because proteins are the basic functional units in various biological processes and are the direct drug targets, this limitation represents a notable knowledge gap in gaining a full understanding of the molecular bases of human cancers and developing more effective strategies for precision cancer medicine[10].

Reverse-phase protein arrays (RPPAs) are a quantitative antibody-based approach to assess the expression of selected proteins in a large

A full list of affiliations appears at the end of the paper. ✉e-mail: rakbani@mdanderson.org; millsg@ohsu.edu; hliang1@mdanderson.org

number of samples in a high-throughput, cost-effective and sensitive manner[11]. Using this platform, we previously generated high-quality protein expression data for TCGA patient tumors[12] and CCLE cell lines (including both baseline and perturbed protein expression profiles)[3,13,14]. These datasets provide deep insights into tumor subtype classification, cancer signaling rewiring in diverse tumor contexts, drug resistance mechanisms, novel biomarkers and therapeutic strategies[1,15–18]. Importantly, we developed user-friendly data portals to facilitate a broad community in analyzing and visualizing these data[19,20]. However, one major drawback of these earlier efforts (phase I) is the relatively small number of protein markers assessed (~200), leaving many cancer-related pathways uncovered or insufficiently characterized.

To increase the utility of TCGA and CCLE cohorts, we finished a phase II characterization of ~8,000 TCGA tumor samples and ~900 CCLE cell lines using an updated RPPA platform. This effort substantially increases the number of proteins profiled to ~500 (that is, RPPA500), which covers almost all known cancer hallmarks and a large portion of clinical markers and therapeutic targets. We then performed integrated analyses of this dataset with other pre-existing molecular and phenotypic data to demonstrate the unique value of RPPA-based protein expression profiling in translational research.

Results

Upgraded RPPA resource for TCGA and CCLE samples

We generated quantitative protein expression data from ~8,000 TCGA patient tumor samples and ~900 CCLE cell line samples using our updated RPPA platform (Fig. 1a), now covering ~500 high-quality protein markers (240 more protein markers than in phase I). We validated antibody candidates through a comprehensive workflow (Extended Data Fig. 1a). Following the established guidelines[11] and the phase I processing pipeline[12,21], we quantified and normalized the RPPA data through four steps, from raw images to batch-effect-corrected data (Extended Data Fig. 1b). After rigorous sample-wise and protein-wise correlation analyses to ensure data quality, the upgraded RPPA500 dataset is available on The Cancer Proteome Atlas (TCPA; https://tcpaportal.org)[20].

The final RPPA500 dataset (Supplementary Table 1) covers 7,828 TCGA patient samples across 32 cancer types, with the largest cohorts in breast invasive carcinoma (BRCA, $n = 881$), kidney renal clear cell (KIRC, $n = 478$) and uterine cancers (UCEC, $n = 440$) (Fig. 1b). In parallel, it contains 878 CCLE cell line samples of 24 lineages (Supplementary Table 2), primarily used in cancer research and characterized by various molecular profiling methods[3,4,22] and phenotypic assays[8,23–26] (Fig. 1c). The updated dataset contains 447 protein markers (357 total and 90 post-translationally modified (PTM) proteins; Supplementary Table 3), targeting crucial oncogenic pathways and therapeutic biomarkers. Compared to phase I (refs. 12,20,27), the phase II protein set covers 15 new hallmark sets (30%) with ≥5 protein markers; furthermore, for 60% of the hallmark sets covered by both phases, RPPA500 has, on average, an 80% increase in the protein numbers assessed per set (Fig. 1d, Extended Data Fig. 1c and Supplementary Table 4). This expansion, thus, greatly increases our ability to interpret cancer mechanisms and generate clinically relevant hypotheses. For example, by comparing different RPPA-based interferon-γ (IFNγ) response pathway scores to those from the literature[28], we found that the pathway score based on RPPA500 (phases I and II) showed significantly higher correlations than that based on phase I data, consistently across cancer types (Extended Data Fig. 1d).

To further evaluate the RPPA500 data quality, we compared RPPA protein levels with corresponding mRNA expression (mean Pearson's $R = 0.51$) and mass spectrometry (MS)-based protein expression (mean Pearson's $R = 0.52$) for total proteins[3]. Compared to correlations between MS-based protein and mRNA expression (mean Person's $R = 0.54$), our data showed overall good correlation and consistency (Fig. 1e). In addition, cross-platform analyses for both phase I and II showed no significant differences, suggesting that the data quality is consistent between the two phases (Extended Data Fig. 1e). Lineage-specific correlation analysis of RPPA total proteins with mRNA or MS data resulted in an average correlation of ~0.4 for both (Extended Data Fig. 1f), largely because of the reduced sample size in each lineage (Extended Data Fig. 1g). For PTM proteins, the average correlations were much lower: 0.23 and 0.24 for mRNA and MS-based total proteins[22], respectively, indicating that PTM protein levels could not be accurately imputed by expression levels of mRNA and total proteins (Fig. 1e). We also analyzed the expression of phosphorylated proteins in NCI60 cell lines[29]. The average correlation between RPPA and MS data was 0.44, consistent with our observation for total proteins (Extended Data Fig. 1h). For example, a phosphorylated heat-shock protein (HSP27 pS82), exhibited a Pearson's correlation of 0.7 between the two datasets ($P = 2.3 \times 10^{-6}$; Extended Data Fig. 1i). Overall, our thorough, multilayered quality control (QC) analyses indicate the high quality of RPPA data and the consistency across various platforms and phases.

RPPA protein expression patterns in TCGA samples

On the basis of TCGA RPPA500 data, we performed an unsupervised clustering analysis and identified ten clusters across 32 cancer types (Fig. 2a). Most clusters were dominated by individual cancer types or cancer types from the same tissue of origin. Furthermore, tumors from different tissue sites but similar organ systems tended to cluster together (for example, pan-GYN, encompassing breast cancer and gynecological malignancies, and pan-GI, encompassing malignancies occurring throughout the gastrointestinal tract; Fig. 2b). These patterns are generally consistent with the clustering patterns based on gene expression data, highlighting that lineage effects are dominant in protein expression[30].

For individual proteins, those from the same pathways tended to cluster together (Fig. 2a). For example, the phosphatidylinositol 3-OH kinase (PI3K) signaling pathway formed a cluster in which phosphorylated protein kinase B (AKT pS473 and pT308) and phosphorylated glycogen synthase kinase 3 (GSK3 pS9 and pS21S9) were strongly upregulated in K5 and phosphorylated ribosomal protein S6 (pS235S236 and pS240S244) was strongly upregulated in K1. Protein kinase C (PKC) formed a cluster and showed a strong signal in K5 (PKCα, PKCα pS657 and PKCABII pT638T641). Immune-related markers also formed a cluster, with granzyme B, cleaved caspase 7, cluster of differentiation 4 (CD4) and spleen tyrosine kinase being downregulated in K5 but upregulated in other clusters.

Interestingly, some cancer types were split into multiple clusters, showing distinct overall survival patterns. For example, KIRC samples in cluster K7 showed a worse prognosis than in K10 (log-rank test, $P = 6.6 \times 10^{-}4$; Fig. 2c), while cervical squamous cell carcinoma (CESC) samples in K7 showed a worse prognosis than in K1 (log-rank test, $P = 1.3 \times 10^{-2}$; Fig. 2d and Extended Data Fig. 2a,b). Pathway analysis showed a higher epithelial–mesenchymal transition (EMT) score in K7 than K10 in KIRC and K1 in CESC (Extended Data Fig. 2c) and a lower IFNα score in K7 than K10 in KIRC and K1 in CESC (Extended Data Fig. 2d). Downregulation of IFNα in K7 suggests a potential suppression of the immune response. These findings suggest EMT upregulation and IFNα downregulation as features of the K7 cluster, which may explain the worse prognosis associated with patients in the K7 cluster. We further analyzed molecular drivers associated with RPPA clusters (Extended Data Fig. 2e,f). Accounting for cancer type as a covariate, we identified several significant gene–cluster pairs. Interestingly, K1 and K8 were associated with multiple driver gene alterations (*TP53* and *PIK3CA* mutations; *PTEN* loss) but with reversed patterns. These results indicate that RPPA data provide a unique perspective to characterize intertumor heterogeneity and identify related clinical patterns and cancer drivers.

Protein profiling data represent a more direct readout of protein activity than mRNA expression; therefore, it is critical to assess

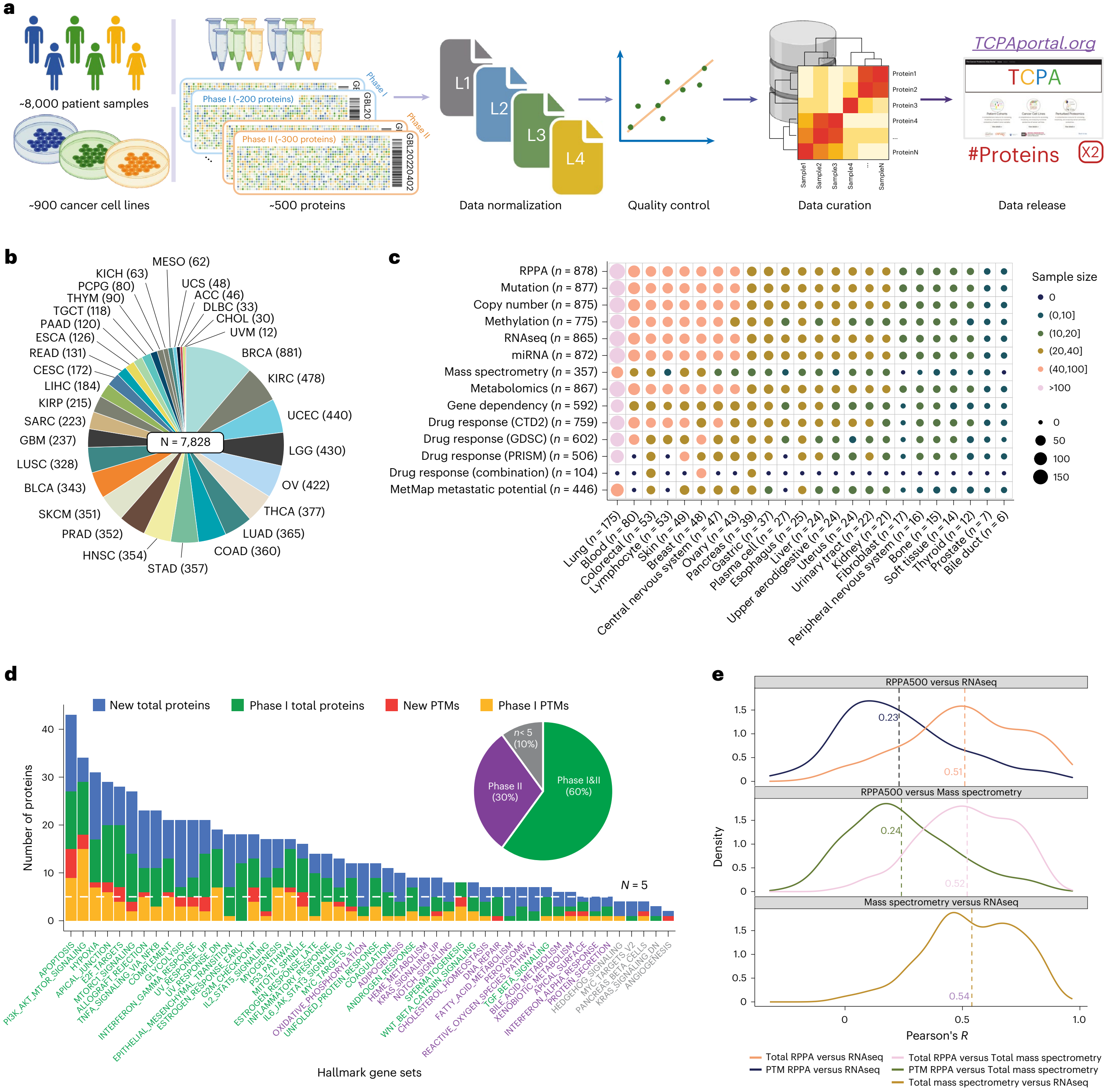
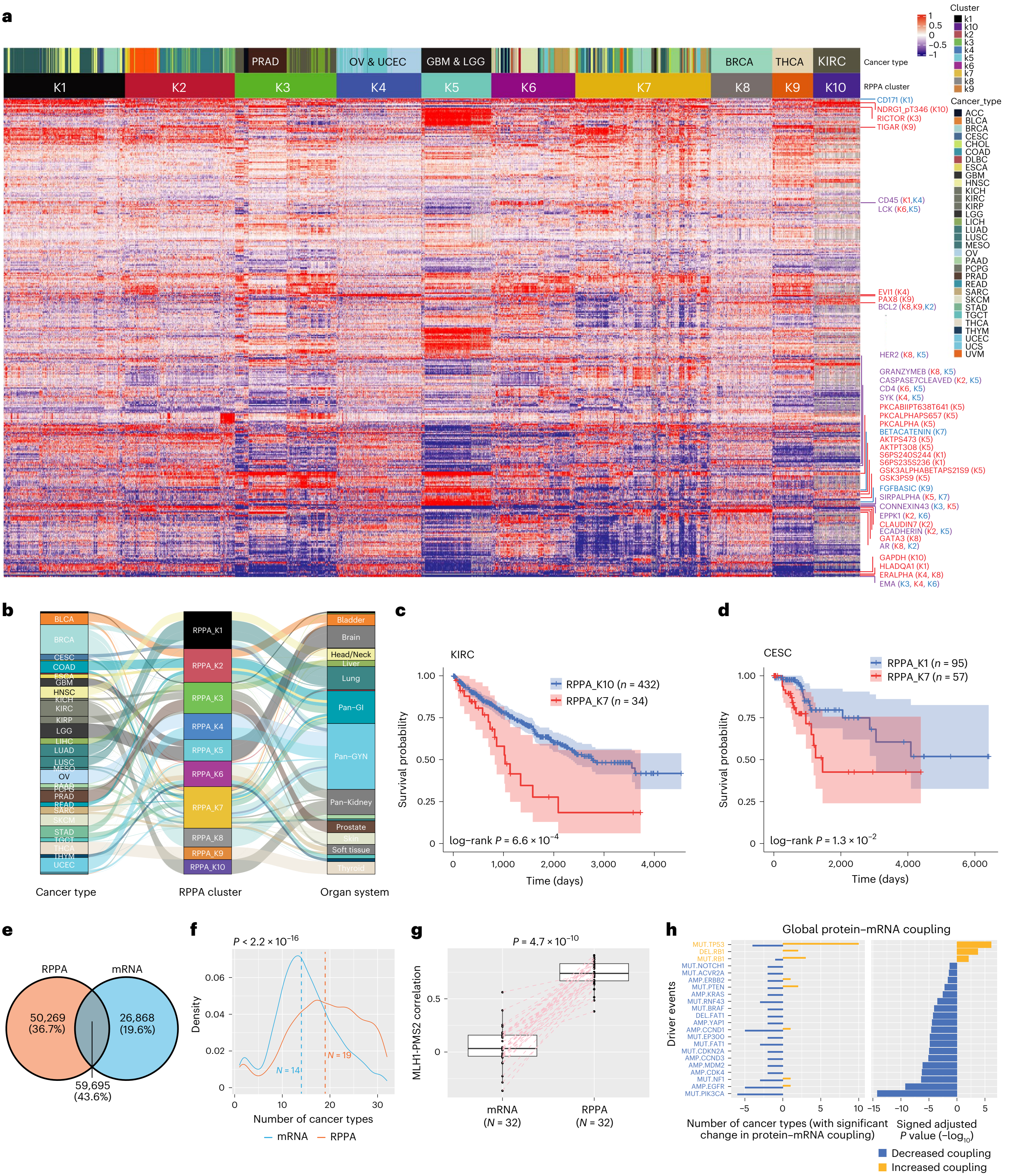


Fig. 1 | Overview of the upgraded RPPA resource of TCGA and CCLE samples. **a**, The pipeline of RPPA data generation and data sharing for ~8,000 patient samples and ~900 cancer cell lines across ~500 protein markers. The raw data were normalized through four levels (from L1 to L4; Methods). The full list of cancer classifications can be found on TCGA (https://portal.gdc.cancer.gov/). **b**, The cancer type distribution of TCGA patient samples (*n* = 7,828). Each number represents the sample size of each cancer type. **c**, The lineage distribution of CCLE samples (*n* = 878) and their parallel molecular and phenotypic data. The number for each column represents the sample size of each cell line lineage with available RPPA data and the number for each row represents the total sample size for each data type. **d**, The distribution of protein markers in different cancer hallmark gene sets, with newly covered pathways (with at least five protein markers) highlighted in purple. The new total proteins and PTM proteins are shown in blue and red, respectively, and the phase I total proteins and PTM proteins are shown in green and orange, respectively. The pie chart shows the distribution of hallmark sets covered in the different phases. **e**, The distributions of the expression correlations of RPPA-based proteins with RNA-seq-based mRNA (top) and MS-based total proteins (middle) and correlations between MS-based total proteins and RNA-seq-based mRNA (bottom). The distributions are shown for total proteins and PTM proteins. TCGA cancer type abbreviation codes are provided in Supplementary Table 1.

whether protein expression is more informative in inferring biological interactions. We performed a coexpression analysis and identified potential interaction pairs using RPPA and mRNA data (false discovery rate (FDR) < 0.1). These pairs were further mapped to the annotated physical protein–protein interactions (PPIs) in the STRING database[31]. In total, we identified 136,832 PPIs: 59,695 were identified by both RPPA and mRNA, whereas 50,269 were identified by RPPA only (that is, twofold higher than the 26,868 identified solely by mRNA; Fig. 2e),



indicating that RPPA data are more sensitive to detecting PPIs. We further found that RPPA-inferred PPIs occurred in more cancer types (median $n = 19$) than mRNA-inferred PPIs (median $n = 14$; Fig. 2f), indicating that RPPA-based PPIs are more robust. For example, the MLH1–PMS2 complex, a well-known PPI, forms a heterodimer that is essential in DNA mismatch repair, which was identified in all 32 cancer types by RPPA but only detected in five cancer types by mRNA ($P = 4.7 \times 10^{-10}$; Fig. 2g).

Our RPPA data also provide a unique opportunity to assess how driver events affect protein–mRNA coupling. Given the copy-number alterations of driver genes, we confirmed that amplification increased its *cis* protein–mRNA coupling ($P = 3.9 \times 10^{-2}$; Extended Data Fig. 2g),

Fig. 2 | Global patterns of RPPA protein expression in different TCGA cancer types. a, Clustering heat map of TCGA samples based on the expression of 447 proteins. **b**, Sankey diagram showing the relationship among cancer types, RPPA clusters and organ systems. **c**,**d**, Kaplan–Meier plots showing differential survival patterns between corresponding clusters. The *P* values are based on a log-rank test. Shaded areas denote the 95% confidence intervals. The numbers represent the patient count in each group (KIRC: K10, $n = 432$; K7, $n = 34$; CESC: K1, $n = 95$; K7, $n = 57$). **e**, Venn diagram showing the number of PPIs identified by RPPA protein and/or mRNA coexpression patterns (RPPA only, $n = 50{,}269$; mRNA only, $n = 26{,}868$; both, $n = 59{,}695$). **f**, The distribution of the number of cancer types in which a specific PPI was identified by RPPA or mRNA data. The dotted lines indicate the median values (RPPA, $n = 19$; mRNA, $n = 14$). The *P* value is based on a Wilcoxon test. **g**, Box plot showing that the RPPA-based Mlh1–PMS2 correlation is much higher than the mRNA-based correlation across cancer types ($n = 32$). The *P* value is based on a paired Wilcoxon test. The middle lines in the boxes are the median, the upper and lower boundaries of the boxes are the first and third quartiles and the whiskers extend to 1.5× the interquartile range of the lower and the upper quartiles; *n* represents the number of cancer types. **h**, The effects of driver events on global protein–mRNA coupling, including mutation, copy-number amplification and deletion. The signed adjusted *P* value is based on an ANOVA with cancer type as a covariate. The sign was determined by the median of differences indicating the direction of change in protein–mRNA coupling. Increases and decreases in protein–mRNA coupling are shown in orange and blue, respectively.

while deletion decreased its *cis* coupling ($P = 1.8 \times 10^{-2}$; Extended Data Fig. 2h). We next focused on *trans* effects and identified 22 significant somatic driver events (Fig. 2h). Intriguingly, most driver events decreased protein–mRNA coupling in the altered group relative to the wild-type group. These results emphasize the importance of expression profiling of protein in addition to mRNA.

RPPA protein expression patterns in CCLE cell line samples

Using CCLE RPPA500 data, we performed an unsupervised clustering analysis and identified ten RPPA clusters (Fig. 3a). Similar to TCGA RPPA clusters, the CCLE RPPA clusters were also largely driven by cancer lineages (Extended Data Fig. 3a,b). To investigate the observed lineage effects, we further performed a supervised analysis and identified 39 lineage-specific proteins (Fig. 3b,c). Interestingly, differentiating powers of these proteins varied; some proteins showed high expression levels across a subset of lineages (for example, caveolin), some were specific to individual lineages (for example, MELANA; Extended Data Fig. 3c,d) and some could further distinguish lineage subtypes (for example, Bcl-2; Extended Data Fig. 3e). Proteins in the same pathway showed consistent signature patterns (for example, cell-cycle pathway; Extended Data Fig. 3f). In addition, compared to mRNA data, RPPA proteins showed a higher dynamic range (fold change) than the corresponding mRNA for each lineage-specific marker (Extended Data Fig. 3g)[3]. These findings highlight the superiority of RPPA500 protein markers in characterizing intrinsic tumor-specific contexts, suggesting their potential value in subtype classification and biomarker identification.

RPPA data reveal deep mechanistic insights in cancer dependency

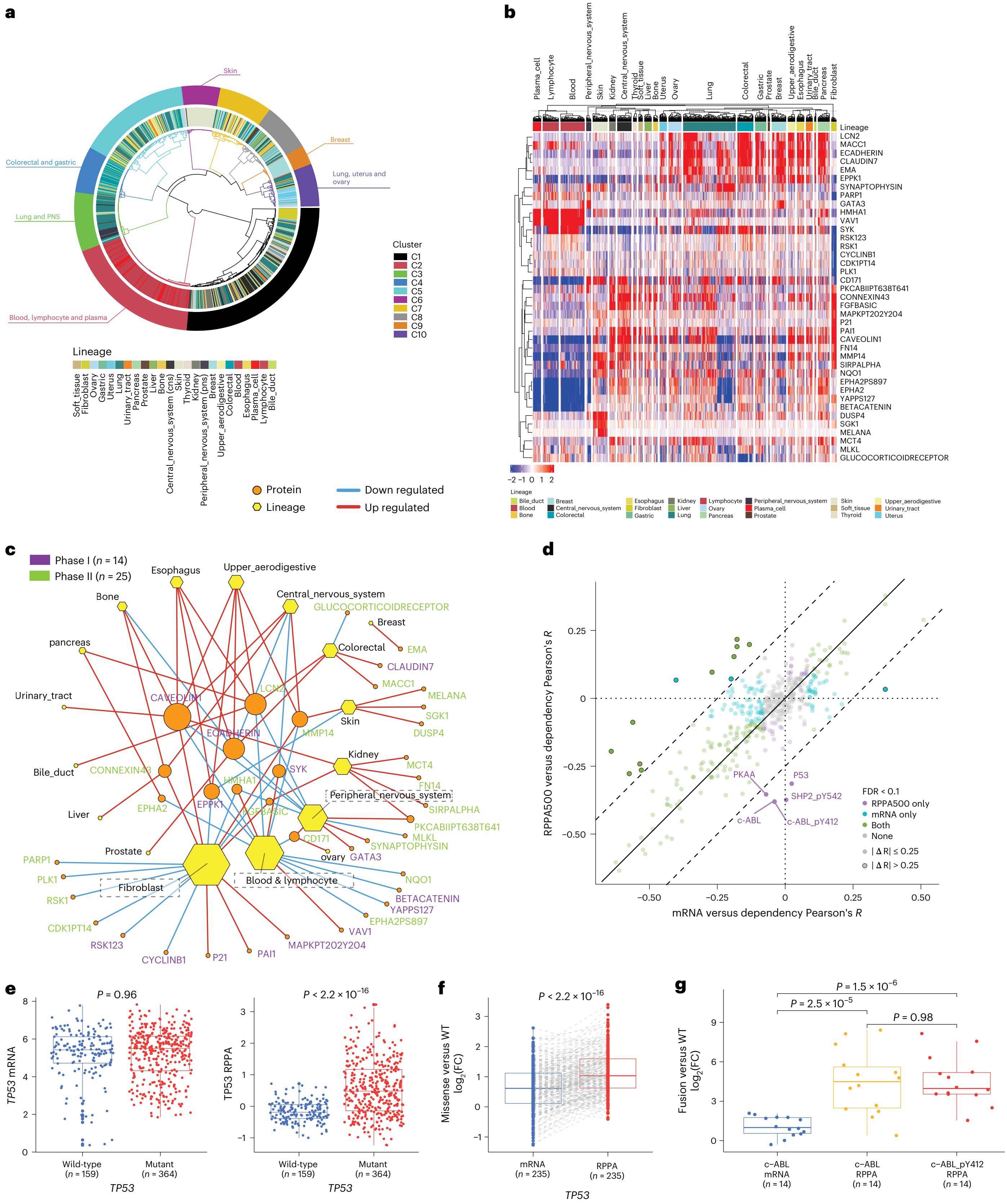
Cancer dependency is defined as the fitness effect on the growth or survival of cancer cells when an individual gene is knocked out or down[32], providing essential information for developing treatments and reducing toxicity. To assess the value of RPPA expression in elucidating gene dependency, we performed a comparative analysis using both RPPA and mRNA expressions with the gene dependency scores obtained from the DepMap project[7,33,34]. Overall, the correlations between gene dependency and its protein and mRNA levels were aligned tightly on the diagonal line, indicating good, consensus genotype–phenotype relationships (Fig. 3d). However, five proteins exhibited much higher absolute correlations than their mRNA counterparts ($|\Delta R| > 0.25$, FDR < 0.1). Among them, phosphatase SHP2 (pY542), rather than the mRNA, has been reported as a sensitive marker for receptor tyrosine kinase activation[3]; therefore, we focused on the remaining four proteins to dissect the difference in correlation between RPPA protein and mRNA levels.

The first case, *TP53*, is the most frequently mutated tumor suppressor gene in human cancers. We performed a differential expression analysis between *TP53* wild-type and mutant cell lines (Fig. 3e). Although *TP53* mRNA showed no difference (Wilcoxon test, $P = 0.96$), the P53 protein level was significantly higher in mutant cell lines (Wilcoxon test, $P < 2.2 \times 10^{-16}$), suggesting that protein level, rather than mRNA level, captures the mutational effects. This was also confirmed by TCGA data (paired Wilcoxon test, $P = 1.4 \times 10^{-5}$; Extended Data Fig. 3h). Focusing on different P53 domains, we found that only samples with a mutation in the DNA-binding domain (Wilcoxon test, $P = 3.9 \times 10^{-39}$) or the tetramerization domain (Wilcoxon test, $P = 3.6 \times 10^{-2}$) showed significantly higher RPPA levels than wild-type samples (Extended Data Fig. 3i). As the DNA-binding domain is the most important functional domain in *TP53*, we further investigated this pattern with mRNA data but observed no differences (Extended Data Fig. 3j). In addition, P53 protein expression showed a significantly higher $\log_2$ fold change than mRNA between cell lines harboring *TP53* missense mutations and wild-type samples (Fig. 3f). We also confirmed a significantly lower P53 RPPA expression for nonsense mutants compared to the wild type, suggesting that RPPA can effectively distinguish mutations in terms of their functional domains and variant classifications (Extended Data Fig. 3k). Lastly, we observed that the gene dependency of *TP53* mutants was closer to zero, suggesting a loss of function ($P < 2.2 \times 10^{-16}$; Extended Data Fig. 3l). These findings highlight the superior power of RPPA over mRNA for elucidating mutation functional effects.

In the second case, both total and phosphoproteins of tyrosine kinase c-ABL (c-ABL and c-ABL pY412) showed high correlations with gene dependency (c-ABL Pearson's $R = -0.38$, $P = 1.3 \times 10^{-21}$; c-ABL pY412 Pearson's $R = -0.38$, $P = 8.4 \times 10^{-22}$) while mRNA did not (Pearson's $R = -0.04$, $P = 0.34$). The *ABL1* gene frequently fuses with *BCR* in leukemia, forming an active kinase that promotes cellular proliferation and suppresses apoptosis[35] and is commonly used to diagnose blood cancers. We, thus, assessed whether RPPA expression more effectively captures this fusion event. Indeed, compared to c-ABL mRNA changes, both total and phosphoproteins exhibited significantly higher fold changes in cell lines with *BCR*–*ABL1* fusions relative to the wild type (Fig. 3g), suggesting that RPPA-based protein markers more sensitively detect the functional consequences of fusion events.

Fig. 3 | Global patterns of RPPA protein expression in different CCLE cancer lineages. a, Clustering patterns of CCLE samples based on the expression of 447 proteins. **b**, Lineage-specific protein expression in CCLE samples. **c**, Network view of differentially expressed proteins associated with each cancer lineage (phase I, $n = 14$; phase II, $n = 25$). The *P* values are based on a two-sided Wilcoxon test. **d**, A scatter plot of protein–dependency correlation versus mRNA–dependency correlation. Each dot represents a Pearson correlation colored on the basis of whether the correlation is significant in RPPA, mRNA or both. **e**, Box plots of *TP53* mRNA expression or protein expression between *TP53* wild-type ($n = 159$) and *TP53* mutant ($n = 364$) CCLE samples. The *P* values are based on a two-sided Wilcoxon test. **f**, Box plots showing differential fold changes of *TP53* mRNA expression and protein expression between *TP53* missense mutant ($n = 235$) and wild-type CCLE samples. The *P* value is based on a paired Wilcoxon test. **g**, Box plots showing differential fold changes of c-ABL mRNA and protein expression between CCLE *BCR*–*ABL1* fusion ($n = 14$) and wild-type blood CCLE samples. The *P* values are based on a Wilcoxon test. **e**–**g**, The middle line in the box is the median, the upper and lower boundaries of the box are the first and third quartiles and the whiskers extend to 1.5× the interquartile range of the lower and the upper quartiles; *n* represents the sample size in each group.

Lastly, protein kinase A subunit α (PKAA, encoded by *PRKAR1A*) showed a significant negative correlation with gene dependency at the protein level across cell lines, unlike with mRNA. We assessed whether such correlations persisted within each lineage. While no lineage showed significant mRNA correlation, seven lineages had significant negative correlations between PKAA protein levels and dependency (Extended Data Fig. 3m), highlighting the robustness of RPPA signals. In addition, lineage-specific analyses revealed many protein markers



with significant correlations across multiple lineages but not with mRNA (Extended Data Fig. 3n), suggesting that protein levels provide deeper phenotypic insights.

Phosphorylated mitogen-activated protein kinase (MAPK) kinase (MEK) is a sensitive functional readout of *BRAF* mutations

Somatic mutations are a major source of driver events in cancer development and clinical markers[36]. However, distinguishing driver mutations from passenger mutations remains challenging, even in the best-studied cancer genes[37,38]. To further demonstrate the power of RPPA-based protein signaling in interpreting the functional effects of somatic mutations, we performed a differential expression analysis of oncogene mutations between wild-type and mutant cell lines and identified the most prominent pair *BRAF* mutations and phosphorylated MEK protein level (Fig. 4a). *BRAF*, a key player in regulating the MAPK–ERK (extracellular signal-regulated kinase signaling) pathway[39], is frequently mutated in certain cancers such as melanoma (~50%). The most common *BRAF* mutation is V600. Clinically, treating persons with melanoma harboring non-V600 *BRAF* mutations with B-Raf inhibitors remains controversial[40]. MEK1/2 pS217S221 was the most significantly upregulated protein in *BRAF*-mutated samples because of the increased kinase activity caused by *BRAF* mutations[41]. Focusing on the Ras–Raf signaling pathway (Fig. 4b), we observed that both pMEK (MEK1/2 pS217S221) and pERK (MAPK pT202Y204) were significantly upregulated in *BRAF* mutant cell lines, but not at the levels of B-Raf, MEK and ERK total proteins, implying that *BRAF* mutations mainly alter the phosphorylation of its downstream MEK and ERK targets but not their total proteins. Furthermore, because ERK is a negative regulator of Ras through feedback inhibition[42], N-Ras showed a significantly lower protein expression in *BRAF* mutant cell lines (Wilcoxon test, $P = 3.4 \times 10^{-3}$). We next examined the correlation with gene dependency scores and found that MEK1/2 pS217S221 negatively correlated with *BRAF* dependency in mutant cell lines (Pearson's $R = -0.71$, $P < 2.2 \times 10^{-16}$; Fig. 4c) but not in the wild-type samples (Pearson's $R = -0.04$, $P = 0.4$; Fig. 4d), highlighting the oncogenic addiction effect of *BRAF* mutations.

Given the high enrichment of *BRAF* mutations in melanoma, we confirmed the negative correlation between MEK1/2 pS217S221 and *BRAF* dependency (Extended Data Fig. 4a). This suggests that MEK1/2 pS217S221 not only distinguishes *BRAF* mutational effects from the wild type but also represents a sensitive marker for the functional impact of *BRAF* mutations on cell viability. To test this hypothesis, we performed cell viability assays to evaluate the effects of various *BRAF* mutations using two informer cell models[38] (Fig. 4e). By assessing the effect of overexpressing a mutant *BRAF* relative to that of overexpressing a wild-type *BRAF*, we evaluated 138 mutations and classified them into five functional types (that is, strong activating, moderate or weak activating, inactivating, inhibitory, no effect and inconclusive). Focusing on cell lines carrying different types of *BRAF* mutations, MEK1/2 pS217S221 exhibited a clear decreasing trend in the order activating, inactivating, wild-type, silent and inhibitory (Fig. 4f); moreover, pMEK could distinguish strong activating from moderate ot weak activating (Wilcoxon test, $P = 0.028$). In contrast, such a trend was not observed at the mRNA level (Extended Data Fig. 4b,c). Additionally, in TCGA melanoma samples, MEK1/2 pS217S221 RPPA levels were higher in tumors with activating *BRAF* mutations than in the wild type (Extended Data Fig. 4d). Particularly, *BRAF*^{V600E} mutations were enriched in the high and median MEK1/2 pS217S221 groups (Extended Data Fig. 4e). Interestingly, five *BRAF* mutations (V600D, N581Y, V600K, 487_492VTAPTP>A and T599_599>TT) in the high group were not included in our functional assays but all these mutations have been previously identified as oncogenic in the literature[43–46]. This independent validation strongly demonstrates the power of RPPA analysis, especially of PTMs, in interpreting the functional effects of somatic mutations.

To further explore the clinical utility of pMEK, we assessed the power of MEK1/2 pS217S221 and B-Raf in predicting the sensitivity to dabrafenib, a US Food and Drug Administration (FDA)-approved B-Raf inhibitor in melanoma. The MEK1/2 pS217S221 level, but not BRAF protein expression, showed a significant correlation with the drug's sensitivity in *BRAF* mutant cell lines (Fig. 4g,h). We further confirmed that this correlation was retained in melanoma cell lines (Pearson's $R = -0.63$, $P = 0.022$; Fig. 4i). In contrast, *BRAF*, *MEK1* and *MEK2* mRNA expressions did not exhibit predictive power (Pearson's $P = 0.38$, 0.39 and 0.57, respectively; Fig. 4i). These results suggest the potential of pMEK as a predictive marker of dabrafenib sensitivity to stratify patients with *BRAF* mutations.

Systematic identification of RPPA-based synthetic lethality

Synthetic lethality is an important cancer therapeutic strategy where two perturbations are not lethal when acting independently but become lethal when combined[47]. It has attracted wide attention following the success of poly(ADP-ribose) polymerase (PARP) inhibitors as synthetic lethal with *BRCA1/2* mutations[48]. Because proteins are the functional products of gene activities, we hypothesized that a protein-centered analysis could effectively identify synthetic lethality. We developed an RPPA-based approach to identify synthetic lethality pairs by incorporating both CCLE and TCGA samples (Fig. 5a). We first identified protein pairs where cell lines with a lower expression of protein A were more dependent on protein B (that is, with significantly lower dependency scores) and then confirmed that patients with low expression levels of both correlated with better prognosis. Through this framework, we identified 1,131 potential synthetic lethality pairs.

To validate our predictions, we assessed the efficacy of 2,025 clinically relevant two-drug combinations across 125 cancer cell lines[49]. By mapping our predictions onto this dataset, among 52 synthetic lethality pairs (Fig. 5b), 48 were synergistic and 4 were nonsynergistic, leading to a ratio of 12, which was fivefold higher than the background expectation (Fisher's exact test, $P = 8.6 \times 10^{-5}$; Fig. 5c). To further test whether our predictions were enriched in drug combinations showing synergy in more cell lines, we calculated the synergy rate, defined as the number of synergistic cell lines divided by the total number of tested cell lines, and found that our predictions had significantly higher synergy rates than other combination pairs (Wilcoxon test, $P = 9.9 \times 10^{-4}$) (Fig. 5d). These results suggest that our RPPA protein-centered strategy has substantial power in nominating synthetic lethality interactions.

We selected a synthetic lethality pair PKAA–EGFR (epidermal growth factor receptor) for experimental validation because (1) this pair had not been validated by the above drug screen data; (2) EGFR is a major therapeutic target with many FDA-approved inhibitors; and (3) PKAA RPPA, but not mRNA, showed significant associations with gene dependency (Fig. 3d). We first confirmed this association of PKAA protein level with *EGFR* gene dependency and their significant correlation with patient survival (Fig. 5e and Extended Data Fig. 5a–c). To evaluate the PKAA–drug association, we performed a correlation analysis of PKAA protein level with drug sensitivity data from four major datasets (that is, Cancer Target Discovery and Development, Genomics of Drug Sensitivity in Cancer (GDSC) and PRISM). Intriguingly, a high expression of PKAA level was dominantly associated with resistance to EGFR inhibitors (EGFRis) (Fig. 5f and Extended Data Fig. 5d; Fisher's exact test, $P = 2.7 \times 10^{-69}$). In total, we identified 33 EGFRis, including afatinib, lapatinib, gefitinib and erlotinib (eight FDA-approved EGFRis)[50].

For experimental validation, we selected two *EGFR*-resistant lung cancer cell lines, H226 and A549, and established corresponding PKAA-knockdown (KD) stable cell lines using short hairpin RNAs (shRNAs) (Extended Data Fig. 5e,f). We performed drug response assays using four EGFRi drugs at different concentrations (that is, DMSO and nine EGFRi concentrations). Indeed, PKAA-KD cell lines were much more sensitive to the EGFRi (Fig. 5g and Extended Data Fig. 5g–l). Across all four EGFRis, the relative cell viabilities of PKAA-KD cells were

https://doi.org/10.1038/s43018-024-00817-x

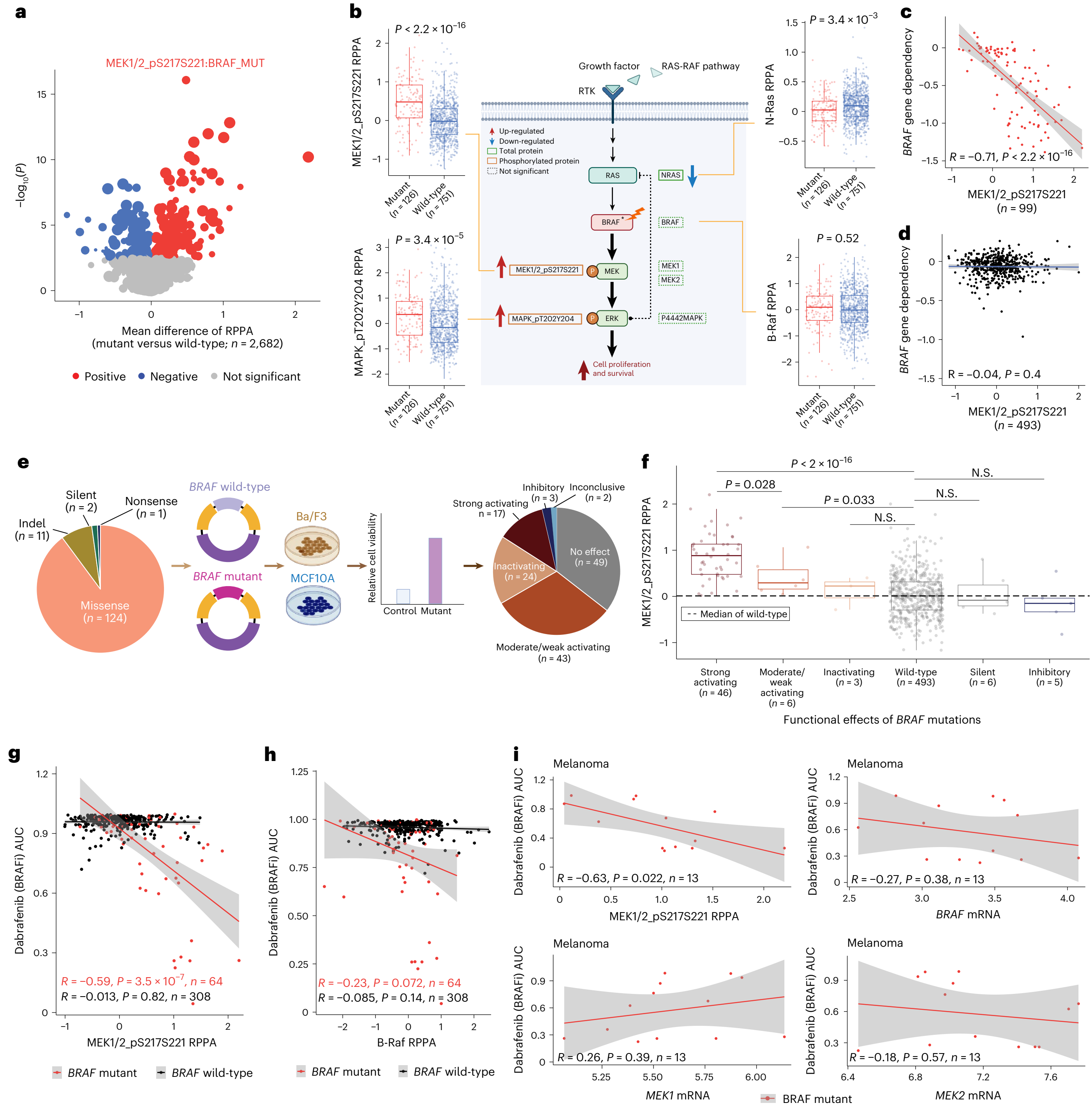


Fig. 4 | Effects of *BRAF* mutations on RPPA-based protein signaling. **a**, A volcano plot showing the differential expression of RPPA proteins between oncogene mutation-containing and wild-type cell lines; *n* = 2,682 represents the total number of comparisons. Significant proteins are shown in blue and red (FDR < 0.1), with the most significant one, MEK1/2 pS217S221:BRAF_MUT, highlighted. The *P* values are based on a Wilcoxon test. **b**, Differential expression of key proteins in the Ras–Raf pathway between *BRAF* mutant (*n* = 126) and wild-type (*n* = 751) samples. The *P* values are based on a Wilcoxon test. **c**,**d**, A scatter plot of MEK1/2 pS217S221 protein versus *BRAF* gene dependency in *BRAF* mutant (*n* = 99; **c**) and wild-type (*n* = 493; **d**) samples; *n* represents the sample size in each group. **e**, Cartoon summary showing a functional genomics approach to characterize the effects of *BRAF* mutations (*n* = 138) on cell viability, based on the readout of two 'informer' cell lines; *n* represents the number of mutations in each category. **f**, A box plot showing the differential expression of MEK1/2 pS217S221 based on the classified functional effects of *BRAF* mutations characterized by the cell viability assays (strong activating, *n* = 46; moderate or weak activating, *n* = 6; inactivating, *n* = 3; wild-type, *n* = 493; silent, *n* = 6; inhibitory, *n* = 5). The *P* values are based on a Wilcoxon test; *n* represents the number of samples in each group. **g**,**h**, The correlations of MEK1/2 pS217S221 (*n* represents the number of mutant (in red; *n* = 64) or wild-type (in black; *n* = 308) samples; **g**) and B-Raf protein (**h**) with dabrafenib (B-Raf inhibitor) sensitivity. The correlations are shown for *BRAF* mutant and wild-type cell lines. **i**, The correlations of MEK1/2 pS217S221 protein expression, *BRAF* mRNA, *MEK1* mRNA and *MEK2* mRNA with dabrafenib sensitivity in melanoma cell lines. *P* values are based on Pearson's correlations; *n* = 13 represents the number of samples in each scatter plot. **b**,**f**, The middle line in the box is the median, the lower and upper boundaries of the box are the first and third quartiles and the whiskers extend to 1.5× the interquartile range of the lower and the upper quartiles. **c**,**d**,**g**–**i**, Shaded areas denote the 95% confidence intervals.

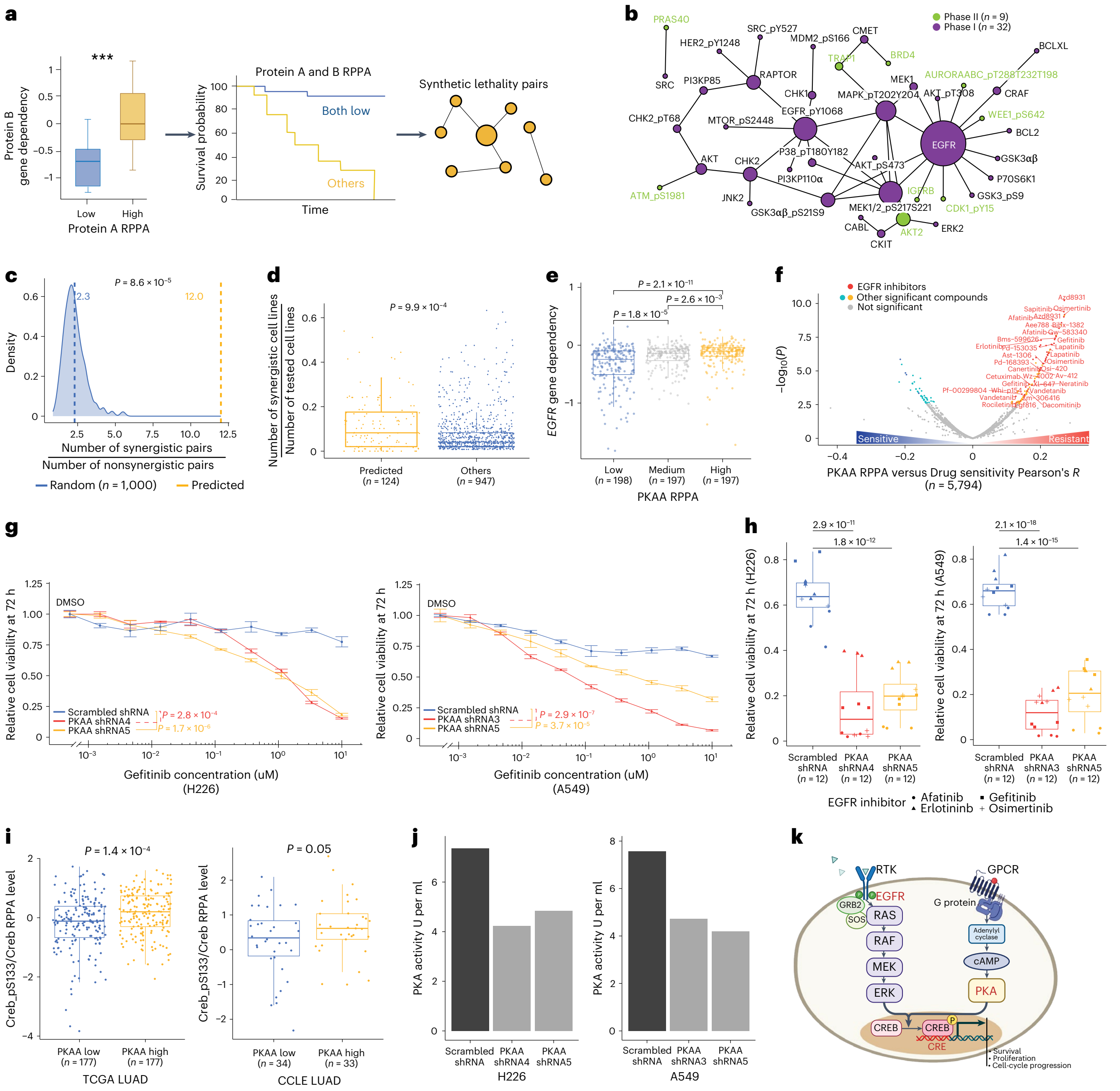
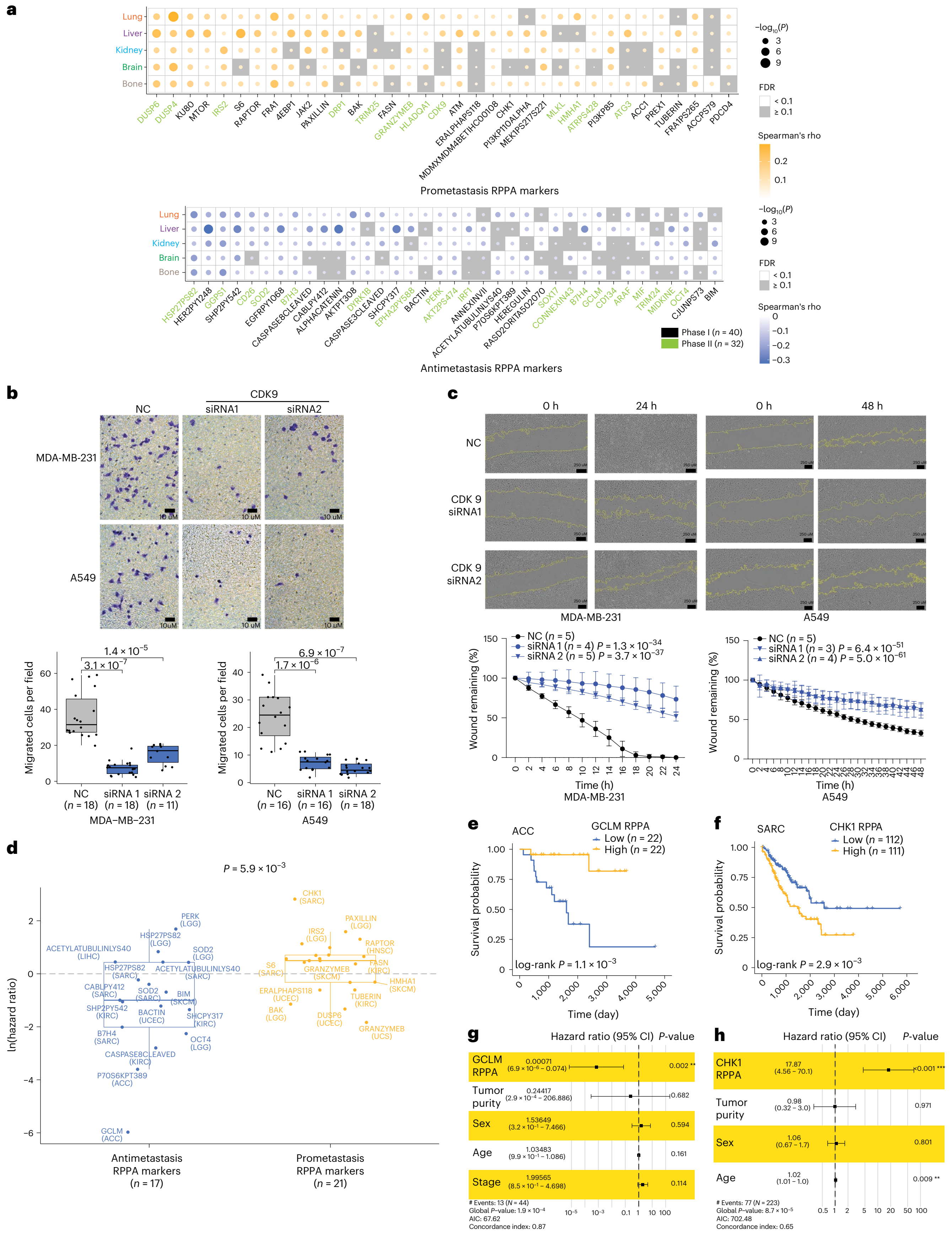


Fig. 5 | Systematic identification of synthetic lethality based on RPPA500 data of CCLE and TCGA samples. a, A computational strategy to identify synthetic lethality pairs by integrating RPPA expression and gene dependency in CCLE samples and RPPA expression and patient survival patterns in TCGA samples. **b**, Network view of synthetic lethality pairs validated by independent drug screening data (phase I, $n = 32$; phase II, $n = 9$). The size of a node is proportional to the number of linked edges. **c**, Odds ratio of predicted synthetic lethality pairs and background expectation. The true synthetic lethality pairs were defined by Jin et al., who systematically assayed single drugs and their drug combinations. The P value is based on a Fisher's exact test. The density curve was generated on the basis of random samplings ($n = 1{,}000$) from all the tested drug combinations. **d**, A box plot showing differential proportions of synergistic cell lines between predicted synthetic lethality pairs and others. The P value is based on a Wilcoxon test; n represents the number of drug combinations assessed (predicted, $n = 124$; others, $n = 947$). **e**, Box plot showing differential patterns of *EGFR* dependency between PKAA low ($n = 198$), medium ($n = 197$) and high groups ($n = 197$). The P values are based on a Wilcoxon test; n represents the sample size in each group. **f**, A volcano plot showing the correlations between PKAA protein and drug sensitivity. Significantly positive and negative correlations (FDR < 0.1) are highlighted in red and blue, respectively. P values were obtained from a Pearson's correlation coefficient test; $n = 5{,}794$ represents the total number of tests performed. **g**, Drug response assays at 72 h for H226 or A549 PKAA-KD and control cells treated with gefitinib (DMSO and nine gefitinib drug concentrations; $n = 3$ independent replicates were examined for each treatment and perturbation). Data are shown as the mean ± s.e.m. The P values are based on an ANOVA. **h**, Box plots showing differential relative cell viabilities at 72 h for H226 or A549 PKAA-KD and control cells treated with 10 µM each of four EGFR inhibitors (that is, afatinib, erlotinib, gefitinib and osimertinib). The P values are based on an ANOVA; $n = 12$ represents the number of samples in each group. **i**, Box plots showing the differential pCREB expression in both PKAA low and high LUAD patient samples (PKAA low, $n = 177$; PKAA high, $n = 177$) or cell lines (PKAA low, $n = 34$; PKAA high, $n = 33$). The P values are based on one-tailed *t*-tests; n represents the number of samples in each group. **j**, Bar plots showing PKA activity in both control and PKAA-KD cells; $n = 2$ independent replicates were examined for each perturbation. **k**, A graphic summary showing that PKA and EGFR pathways provide complementary signals to activate CREB-mediated downstream signaling. **c**,**e**,**h**,**i**, The middle line in the box is the median, the lower and upper boundaries of the box are the first and third quartiles and the whiskers extend to 1.5× the interquartile range of the lower and the upper quartiles. Predicted synthetic lethality pairs are provided in Supplementary Table 5.


a
Lung
Liver
Kidney
Brain
Bone
DUSP6 DUSP4 KU80 MTOR IRS2 S6 RAPTOR FRA1 4EBP1 JAK2 PAXILLIN DRP1 BAK TRIM25 FASN GRANZYMEB HLADQA1 CDK9 ATM ERALPHAPS118 MDMXMDM4BETIHC00108 CHK1 PI3KP110ALPHA MEK1PS217S221 MLKL HMHA1 ATRPS428 PI3KP85 ATG3 ACC1 PREX1 TUBERIN FRA1PS265 ACCPS79 PDCD4
Prometastasis RPPA markers
$-\log_{10}(P)$
3
6
9
FDR
< 0.1
≥ 0.1
Spearman's rho
0.2
0.1
HSP27PS82 HER2PY1248 GGPS1 SHP2PY542 CD26 SOD2 EGFRPY1068 B7H3 CASPASE8CLEAVED CABLPY412 ALPHACATENIN AKTPT308 DYRK1B CASPASE3CLEAVED SHCPY317 EPHA2PY588 BACTIN PERK AKT2PS474 IRF1 ANNEXINVII ACETYLATUBULINLYS40 P70S6KPT389 HEREGULIN RASD2ORITAS02070 SOX17 CONNEXIN43 B7H4 GCLM CD134 ARAF MIF TRIM24 MIDKINE OCT4 CJUNPS73 BIM
Phase I (n = 40)
Phase II (n = 32)
Antimetastasis RPPA markers
0
−0.1
−0.2
−0.3
b
CDK9
NC
siRNA1
siRNA2
MDA-MB-231
A549
10 uM
Migrated cells per field
1.4×10^{-5}
3.1×10^{-7}
6.9×10^{-7}
1.7×10^{-6}
NC (n = 18) siRNA 1 (n = 18) siRNA 2 (n = 11)
MDA–MB–231
NC (n = 16) siRNA 1 (n = 16) siRNA 2 (n = 18)
A549
c
0 h
24 h
0 h
48 h
NC
CDK 9 siRNA1
CDK 9 siRNA2
250 uM
MDA-MB-231
A549
NC (n = 5)
siRNA 1 (n = 4) $P = 1.3 \times 10^{-34}$
siRNA 2 (n = 5) $P = 3.7 \times 10^{-37}$
NC (n = 5)
siRNA 1 (n = 3) $P = 6.4 \times 10^{-51}$
siRNA 2 (n = 4) $P = 5.0 \times 10^{-61}$
Wound remaining (%)
Time (h)
MDA-MB-231
A549
d
$P = 5.9 \times 10^{-3}$
ln(hazard ratio)
PERK (LGG)
HSP27PS82 (LGG)
ACETYLATUBULINLYS40 (LIHC)
SOD2 (LGG)
HSP27PS82 (SARC)
ACETYLATUBULINLYS40 (SARC)
CABLPY412 (SARC)
SOD2 (SARC)
BIM (SKCM)
SHP2PY542 (KIRC)
BACTIN (UCEC)
SHCPY317 (KIRC)
B7H4 (SARC)
OCT4 (LGG)
CASPASE8CLEAVED (KIRC)
P70S6KPT389 (ACC)
GCLM (ACC)
CHK1 (SARC)
PAXILLIN (LGG)
IRS2 (LGG)
RAPTOR (HNSC)
S6 (SARC)
GRANZYMEB (SKCM)
FASN (KIRC)
HMHA1 (SKCM)
ERALPHAPS118 (UCEC)
TUBERIN (KIRC)
BAK (LGG)
DUSP6 (UCEC)
GRANZYMEB (UCS)
Antimetastasis RPPA markers (n = 17)
Prometastasis RPPA markers (n = 21)
e
ACC
GCLM RPPA
Low (n = 22)
High (n = 22)
Survival probability
log-rank $P = 1.1 \times 10^{-3}$
Time (day)
f
SARC
CHK1 RPPA
Low (n = 112)
High (n = 111)
log-rank $P = 2.9 \times 10^{-3}$
Time (day)
g
Hazard ratio (95% CI) P-value
GCLM RPPA 0.00071 (6.9×10^{-6} – 0.074) 0.002 **
Tumor purity 0.24417 (2.9×10^{-4} – 206.886) 0.682
Sex 1.53649 (3.2×10^{-1} – 7.466) 0.594
Age 1.03483 (9.9×10^{-1} – 1.086) 0.161
Stage 1.99565 (8.5×10^{-1} – 4.698) 0.114
Events: 13 (N = 44)
Global P-value: 1.9×10^{-4}
AIC: 67.62
Concordance index: 0.87
h
Hazard ratio (95% CI) P-value
CHK1 RPPA 17.87 (4.56 – 70.1) <0.001 ***
Tumor purity 0.98 (0.32 – 3.0) 0.971
Sex 1.06 (0.67 – 1.7) 0.801
Age 1.02 (1.01 – 1.0) 0.009 **
Events: 77 (N = 223)
Global P-value: 8.7×10^{-5}
AIC: 702.48
Concordance index: 0.65

greatly reduced compared to the control cell lines for both H226 and A549, indicating that low PKAA indeed sensitized EGFR-resistant cells (Fig. 5h). After confirming the causal effects between PKAA loss and increased EGFRi sensitivity, we next sought to elucidate the underlying mechanism. A major phosphorylation substrate of PKAA protein is the cAMP response element-binding protein (CREB); EGFR signaling is also upstream of CREB and phosphorylated CREB is responsible for activating diverse cellular responses, including proliferation, survival and cell-cycle progression[51–54]. Thus, PKAA and EGFR may provide complementary effects on CREB signaling. To test this hypothesis, we assessed how PKAA protein levels were associated with pCREB in lung adenocarcinoma (LUAD) and observed that high PKAA was associated with high pCREB level (normalized by total CREB protein) in patient and cell line samples for LUAD (one-tailed *t*-test, TCGA, $P = 1.4 \times 10^{-4}$; CCLE, $P = 5 \times 10^{-2}$; Fig. 5i). We further confirmed that PKA activity decreased in PKAA-KD cells compared to the controls (Fig. 5j). Thus, we proposed a mechanistic model (Fig. 5k). Two distinct signaling pathways (EGFR and PKAA) can activate CREB; cancer cells with a higher level of PKAA protein expression are less dependent on the EGFR signaling, becoming more resistant to EGFRis.

Characterization of metastasis potential based on RPPA data

Metastasis is a hallmark of cancer and a major contributor to cancer-related deaths. A recent study generated a metastasis map of human cancer cell lines, MetMap, where the metastatic potential of cancer cell lines was determined in mouse xenografts at a large scale[26]. To gain a better understanding of metastasis from the perspective of RPPA-based protein expression, we examined the correlation of RPPA protein markers with MetMap metastasis potential scores. In total, we identified 35 positively and 37 negatively correlated proteins with overall metastasis potential scores (FDR < 0.1). We also evaluated correlations for each target organ, confirming that organ-specific correlations were consistent with overall correlations but with variations (Fig. 6a). Among prometastasis protein markers, mammalian target of rapamycin (MTOR) and its regulatory protein (RAPTOR) are the main protein components of mTORC1, PI3Ks (p110α and p85) are major players in PI3K–AKT signaling and pMEK (MEK1/2 pS217S221) is a key member in the Ras–Raf–MEK–ERK pathway. As reported by previous studies[55], high levels of mTORC1, PI3K–AKT and MEK are responsible for the reactivation of metastatic cancer cells. Among antimetastasis markers, we identified multiple proteins related to the apoptosis signaling pathway, including cleaved caspases 3 and 8. Caspases act as metastasis suppressors by inducing programmed cell death[56]. We further confirmed that prometastasis and antimetastasis markers exhibited significantly divergent expression between metastatic and primary cell lines (Wilcoxon test, $P = 2 \times 10^{-9}$; Extended Data Fig. 6a). Specifically, prometastasis RPPA markers demonstrated elevated expression in metastatic cell lines, while antimetastasis markers showed reduced expression in these cells when compared to primary cell lines. To further validate our prediction, we chose cyclin-dependent kinase 9 (CDK9), an inferred prometastasis RPPA marker, to assess the differences in cell migration and wound-healing capabilities between wild-type and *CDK9* small interfering RNA (siRNA) KD cells (Extended Data Fig. 6b,c). Indeed, both MDA-MB-231 and A549 cells with *CDK9* KD exhibited a significant reduction in their cell migration and wound-healing abilities (Fig. 6b,c), strongly suggesting a critical role of CDK9 in promoting these essential aspects of metastasis.

Because metastasis is the major cause of death in cancer patients, we next performed patient survival analysis on these identified protein markers using TCGA RPPA500 dataset. Among the significant hits (FDR < 0.1), we found that the hazard ratios of the prometastasis proteins were significantly higher than those of antimetastasis ones (Wilcoxon test, $P = 5.9 \times 10^{-3}$; Fig. 6d). The median log hazard ratio of prometastasis proteins was positive (0.5) and that of antimetastasis proteins was negative (−1.0), indicating that the patients would have an increased risk of death if their tumors exhibited higher levels of prometastasis protein markers or lower levels of antimetastasis protein markers. For example, two outliers in the prometastasis and antimetastasis groups were glutamate-cysteine ligase regulatory subunit (GCLM) in adrenocortical carcinoma (ACC) and checkpoint kinase 1 (CHK1) in sarcoma (SARC), respectively. Patients with worse survival showed a lower level of GCLM in ACC and a higher level of CHK1 in SARC (Fig. 6e,f). These patterns remained significant even after adjusting for other potential confounding factors (Cox $P < 0.05$; Fig. 6g,h). To further validate the findings, we extracted the metastasis status for ACC and SARC patients from their clinical annotations and found that lower protein levels of GCLM and a higher level of CHK1 indeed could predict the metastasis status of patients identified at their diagnosis (Extended Data Fig. 6d,e; GCLM, area under the curve (AUC) = 0.75, $P = 0.023$; CHK1, AUC = 0.70, $P = 3.4 \times 10^{-5}$).

Lastly, we examined the clinical development stage for targeting prometastasis markers using the Pharos database[57]. Notably, >50% of the markers were categorized as Tchem (38%) and Tclin (18%), including our validated marker CDK9 (Tchem), highlighting the potential clinical utility of our findings (Extended Data Fig. 6f). Collectively, these results provide a comprehensive view of the key proteins and pathways underlying the metastatic potential in human cancers, laying a foundation for predicting metastasis in patients and developing antimetastasis therapeutics.

Discussion

Here, we generated an expanded protein expression atlas based on TCGA and CCLE cohorts with several major advantages. First, we profiled ~9,000 samples across all major cancer types and lineages, featuring a large sample size and wide cancer diversity. RPPA500, thus, facilitates both conventional analyses and advanced data mining, such as deep learning, on a pan-cancer scale. Unlike other proteomics studies limited by cancer type and technical issues[58–60], RPPA500 enables direct comparisons across cancer types. Second, RPPA500 covers two complementary cohorts, enhancing translational research with

Fig. 6 | Characterization of tumor metastasis potential based on RPPA protein expression. **a**, Heat maps showing the correlations between metastasis potential and protein markers in cell lines. Positive and negative contributors are shown in orange and blue, respectively. The *P* values are based on Spearman's rank correlation coefficient test; *n* represents the number of protein markers (phase I, *n* = 40; phase II, *n* = 32). **b**, Cell migration assays for MDA-MB-231 or A549 *CDK9* siRNA KD and control cells. The *P* values are based on a Wilcoxon test; *n* represents the number of samples tested for each condition (MDA-MB-231, *n* = 18, 18 and 11 for negative control (NC), siRNA 1 and siRNA 2, respectively; A549, *n* = 16, 16 and 18 for NC, siRNA 1 and siRNA 2, respectively). **c**, Wound-healing assays for MDA-MB-231 (from 0 h to 24 h) or A549 (from 0 h to 48 h) *CDK9* siRNA KD and control cells. Data are shown as the mean ± s.d. The *P* values are based on an ANOVA; *n* represents the number of samples tested for each condition (MDA-MB-231, *n* = 5, 4 and 5 for NC, siRNA 1 and siRNA 2, respectively; A549, *n* = 5, 3 and 4 for NC, siRNA 1 and siRNA 2, respectively). **d**, Box plot showing the effects (log hazard ratios) of negative and positive contributors to patient survival based on TCGA patient cohorts. The *P* value is based on a Wilcoxon test; *n* represents the number of significant protein–cancer pairs in each group (antimetastasis RPPA markers, *n* = 17; prometastasis RPPA markers, *n* = 21). **e**,**f**, Kaplan–Meier plots showing differential survival patterns of ACC patients by GCLM protein expression (GCLM low, *n* = 22; GCLM high, *n* = 22) (**e**) and SARC patients by CHK1 protein expression (CHK1 low, *n* = 112; CHK1 high, *n* = 111) (**f**). *P* values are based on log-rank tests; *n* represents the number of patients in each group. **g**,**h**, Forest plots of hazard ratios for GCLM (**g**) or CHK1 (**h**) and clinical variables. The *P* values are based on a multivariate Cox proportional hazards model. The center point of each horizontal bar represents the estimated hazard ratio. **b**,**d**, The middle lines in the boxes are the medians, the lower and upper boundaries of the box are the first and third quartiles and the whiskers extend to 1.5× the interquartile range of the lower and the upper quartiles.

extensive molecular, clinical and phenotypic data. It bridges TCGA's patient-derived insights with CCLE's detailed cell line data, fostering the generation of actionable hypotheses and potential treatment strategies. Third, compared to phase I (refs. 12,20,27), our expanded RPPA500 substantially increased the protein set, which covers all hallmark gene sets, major biomarkers and therapeutic targets under clinical investigation. Taken together, the RPPA500 dataset provides a valuable resource for the cancer research community to translate big cancer omics data into the clinical practice of precision medicine.

Through our integrative analyses, we showcased a few exciting cases with potential translational impact. For example, RPPA effectively identifies the functional effects of driver somatic alterations (for example, *TP53* mutations and *BCR–ABL1* fusions). Specifically, pMEK expression can distinguish mutational effects from both the wild-type background and different types of mutations, as validated by our cell viability assays. This would help address an unmet clinical need on how to stratify melanoma patients with non-V600 *BRAF* mutations for more effective treatments. Through the protein-centered integrated analysis of TCGA and CCLE samples, we identified numerous synthetic lethality pairs, validated by independent drug combination data. The example of PKAA and EGFR suggests a strategy for treating LUAD patients with low PKAA levels using EGFRis. Because LUAD patients usually develop resistance to EGFR inhibitors in ~1–2 years, this strategy may represent an exciting opportunity to improve clinical outcomes for a sizable portion of patients. Further efforts are warranted to investigate the clinical importance of these findings using independent cohorts, animal models and clinical studies.

Of note, our results indicate that, in many cases, meaningful patterns can only be detected at the RPPA protein level rather than the mRNA level, highlighting a need to generate proteomics data for clinical sample characterization. While RPPA500's protein markers are fewer than those of MS-based methods, RPPA offers increased sensitivity (particularly for some key phosphoproteins) and cost-effectiveness, complementing MS's scopes and focuses. Moreover, we acknowledge the intricate challenges associated with the analysis of bulk tumor samples, primarily because of their heterogeneous cellular composition. To mitigate such effects, we meticulously integrated CCLE and TCGA data, aiming to identify consistent patterns across cohorts, and further reinforced them through rigorous experimental validation.

While our approach offers significant advantages, we advocate for continuous advancements in techniques that enable more precise analyses of specific cellular subpopulations within tumors. Such advancements are crucial for deepening our understanding of complex interactions among various cellular populations and their roles in cancer biology. It is of paramount importance to fully harness the capabilities of our RPPA platform alongside other protein profiling techniques and comprehensive tumor microenvironment data. This integrated approach stands poised to expedite the development and implementation of protein-centric diagnostics and therapeutic strategies, marking a substantial stride forward in the field of precision oncology.

Methods

Patient sample and cell line cohorts for RPPA profiling

The research complies with all relevant ethical regulations. We obtained patient samples from the National Institutes of Health (NIH) TCGA project and CCLE samples from the Broad Institute. The details of these samples were described in our previous studies[1–3]. Ethics oversight and written informed consent were obtained through TCGA project.

RPPA quantification and normalization. *Overview*. We collected and prepared cell line and patient tumor samples obtained from CCLE and TCGA. The antibodies were validated following the standard RPPA pipeline as documented in the previous publications[12,61,62]. All the RPPA protein expression data were generated by the RPPA core facility at the MD Anderson Cancer Center. CCLE samples were profiled on a single RPPA slide and TCGA samples were profiled on eight slides. RPPA slides were scanned and quantified using ArrayPro (Meda Cybernetics) to obtain the protein signal intensities as the raw level 1 (L1) RPPA data. We then performed curve-fitting analysis to generate L2 RPPA data by estimating protein expression levels using SuperCurve[63]. L2 data were then median-polished to obtain normalized L3 RPPA data. From the L3 data, the batch effects were removed using a replicate-based normalization method to generate L4 data, which were used in the downstream analyses[12,21].

Sample preparation. The frozen tumors or cell pellets were lysed and proteins were extracted using RPPA lysis buffer (containing 1% Triton X-100, 50 mM HEPES pH 7.4, 150 mM NaCl, 1.5 mM $MgCl_2$, 1 mM EGTA, 100 mM NaF, 10 mM Na pyrophosphate, 1 mM Na_3VO_4, 10% glycerol and freshly incorporated protease and phosphatase inhibitors from Roche Applied Science (cat. nos. 05056489001 and 04906837001, respectively))[61].

These lysates were adjusted for protein concentration and then serially diluted into five twofold dilutions using the same lysis buffer, followed by printing on nitrocellulose-coated slides using an Aushon Biosystems 2470 arrayer. The slides were then probed with ~500 prevalidated primary antibodies and detected with suitable biotinylated secondary antibodies: goat anti-rabbit IgG (Invitrogen, cat. no. 31822, lot no. XC3537505; dilution: 1:1,000), goat anti-mouse IgG (Vector Laboratories, cat. no. BA-9200, lot. no. ZD0801; dilution: 1:10,000) or rabbit anti-goat IgG (Fisher Scientific, cat. no. NC9376096, lot. no. ZB0923; dilution: 1:10,000).

Data quantification. The signal was amplified using streptavidin-conjugated horseradish peroxidase (HRP), which binds to the secondary antibody and catalyzes biotinylated tyramide to form insoluble biotinylated phenols. The signals were then made visible by a secondary streptavidin-conjugated HRP and DAB colorimetric reaction. Subsequently, the slides were scanned and analyzed and their quantification was performed using ArrayPro Analyzer software (MediaCybernetics) to generate spot intensity (L1 data)[63].

Data normalization. Our previously developed SuperCurve GUI tool was used to estimate relative protein levels (in $\log_2$ scale). A fitted curve or 'supercurve' was generated with signal intensities on the *y* axis and relative $\log_2$ quantities of each protein on the *x* axis using a non-parametric, increasing B-spline model[63]. Before model fitting, raw spot intensity data were adjusted to correct spatial bias using 'control spots' spread across slides[64]. Each slide generated a QC metric to determine its quality and only slides with a score greater than 0.8 on a 0–1 scale were considered for further processing[65]. For duplicate slides, the one with the highest QC score was chosen for analysis (L2 data). Protein measurements were then corrected for loading as detailed[63,66] using bidirectional median centering across samples and antibodies (L3 data). They were also debatched using technically replicated control samples for normalization across RPPA core sets (L4 data)[12].

RPPA QC. Following our standard data processing pipeline[12,21] (Extended Data Fig. 1b), we quantified and normalized the RPPA data of ~500 proteins for ~8,000 TCGA patient samples and ~900 CCLE cell line samples. We had QC metrics at each step to ensure a high quality of the data.

Slide-level QC. Using the 96 replicates of mixed lysates, we generated QC metrics to determine the quality of each RPPA slide following staining. Importantly, we printed additional slides for each sample run to replace slides that failed QC and to have slides to assess newly validated antibodies. The replicate slide can be used with the same antibody or, if necessary, with another antibody validated for the same protein or phosphoprotein. Thus, high QC metrics serve to remove 'bad' data, as well as indicate slides that need to be replaced with a newly stained replicate slide.

Sample-level QC. We assessed the total protein content in each sample by computing the median value across all stained proteins. That median served as an estimator for the total protein content in the sample. If the total protein content was too low, we considered it as an 'outlier' sample and discarded it from the report. From prior experience with TCGA data (~8,000 samples), a median value less than −3 ($\log_2$) signifies a low-quality outlier.

RPPA protein expression clustering analysis. We performed an unsupervised clustering analysis for TCGA and CCLE RPPA data using ConsensusClusterPlus[67]. The dendrogram was generated by the R package Circlize[68]. Protein markers showing significant lineage-specific differential expression were identified as lineage-specific RPPA markers. The supervised clustering heat map was generated using ComplexHeatmap[69]. The network view was generated using Cytoscape[70]. Each node represents an RPPA marker or cell line lineage. The node size was determined on the basis of the number of edges linking to it. For survival analysis, we used a log-rank test to assess the differential survival probability between RPPA clusters within each cancer under FDR < 0.1. For each cancer–cluster pair (for example, KIRC, RPPA_K10 versus RPPA_K7), we further performed differential expression analysis to identify significant protein markers on the basis of Wilcoxon test *P* values and mean fold changes. To assess the associations between RPPA clusters and gene-level somatic alterations, we performed enrichment analysis (chi-squared test) for each gene–cluster pair under FDR < 0.1. To further adjust for covariates, we added the cancer types as the strata and performed Cochran–Mantel–Haenszel chi-squared tests with FDR < 0.1 to identify significant gene–cluster pairs.

PPI and protein–mRNA coupling. To identify PPIs using both mRNA and RPPA data, we first calculated protein–protein correlations using RPPA500 protein and RNA sequencing (RNA-seq) mRNA data in each cancer type. The significant correlations (FDR < 0.1) were further mapped to the physical interactions obtained from the STRING database (https://string-db.org). The matched interactions were used to generate the Venn diagram and density curves shown in Fig. 2e,f. To assess the protein–mRNA coupling across cancer types, we first categorized samples in each cancer type into the altered and unaltered groups on the basis of alterations of each driver gene. Driver event status was obtained from a TCGA pan-cancer study on genetic alterations in oncogenic signaling pathways[71]. Within each group, we further performed analysis to assess the correlations between the expression of each protein and its corresponding mRNA expression. For each driver event, the changes in the global protein–mRNA coupling were evaluated by examining the difference in protein–mRNA correlations between the altered and wild-type groups. *P* values were determined using an analysis of variance (ANOVA) to consider both the effects of cancer type and the changes in protein–mRNA correlations.

Cancer dependency analysis. The correlation analysis between gene dependency and RPPA or mRNA was performed by Pearson's correlations, with FDR < 0.1. The CCLE mutation data were used to group cell lines into the wild type and *TP53* mutants. The residue position of each mutation was used to map each cell line onto the P53 protein domains. To test *TP53* differential expression in TCGA, we first calculated $\log_2$ fold changes of *TP53* mRNA and RPPA between *TP53* wild-type and mutant tumor samples in each cancer type, followed by differential analysis between mRNA and RPPA using the Wilcoxon test. To evaluate the differential expression pattern of c-ABL in fusion and wild-type cell lines, we first quantified $\log_2$ fold changes between *BCR–ABL1* fusion and wild-type blood cancer cell lines, after which we performed pairwise Wilcoxon tests among c-ABL mRNA, total protein and phosphorylated protein expressions.

Mutant construction and cell viability assays. The *BRAF* mutation candidates were obtained from CCLE and TCGA mutation data. We constructed the lentivirus vector of wild types, mutations and controls with pHAGE-EF1α-GFP or pHAGE-EF1α-PURO backbone using the high-throughput mutagenesis and molecular barcoding (HiTMMoB) technique, as described previously[72,73]. Cell viability assays were performed as previously described[37,38]. Briefly, we used two growth factor-dependent cell models, Ba/F3 and MCF10A. Both cell types die in the absence of the required factor(s). Mutations and their corresponding wild-type counterparts were assessed in parallel in the same experiment, the latter of which determined the basal activity of the genes in the cell models. For each experiment, pHAGE constructs of mutants and wild-type genes were freshly prepared from a single colony and used for generating lentivirus for Ba/F3 and MCF10A transductions. The cell viability of Ba/F3 and MCF10A cells was measured at four time points (at intervals of 3 or 4 days) during the 3-week assay period. The functional annotations of mutations were based on a comparison to the corresponding wild-type clones. The final annotation was made on the basis of consensus calls of the two cell lines. To further classify *BRAF* mutations into different 'activating' groups, the calls were based on comparison to *PIK3CA* mutations of known activity that were run in each experiment. Strong activating ones were equal or greater in activity compared to a strongly activating *PIK3CA* mutation (K1047R). Moderate activating ones were equal to or greater than a moderate activating *PIK3CA* mutation (M1043I). Weak activating ones were greater than the wild type but less than a moderate activating *PIK3CA* mutation.

Analysis of synthetic lethality pairs. To predict synthetic lethality pairs, we first performed a differential analysis on gene dependencies of a protein between two sample groups defined by the median RPPA expression of another protein. *P* values were calculated by the Wilcoxon test. We selected the protein pairs by considering both the direction of the difference Δ and the FDR, where Δ is the difference between the mean values of gene dependency scores in cell lines with high and low RPPA expressions. The selected protein pairs were further tested against patient survival and predicted as synthetic lethality pairs if they (1) showed significance for both log-rank and Cox proportional hazard tests and (2) the patient group with low RPPA of both proteins had better survival than others. To validate the predictions, we matched the synthetic lethality pairs to the targets of the drug combination set obtained from the GDSC drug combination data portal. A protein pair was marked as synergistic if it showed synergy in at least one combination and cell line pair. To generate the background set, we randomly sampled the same number of drug pairs from the combination set and counted the number of synergistic pairs. For each synergistic pair, we further defined the synergy rate as the number of cell lines showing synergy divided by the number of tested cell lines. The synergy rates of the predicted synthetic lethality pairs and all the other nonoverlapped synergy pairs in the drug combination set were used to evaluate differential synergy rates between predicted and other synergistic pairs by the Wilcoxon test. Validated synergistic pairs were then used to generate a synthetic lethality network using Cytoscape[70], in which each node represented a protein and each edge linked a synthetic lethality pair. To assess the association between *EGFR* gene dependency and PKAA protein expression, we performed Wilcoxon tests between three groups of cell lines (low, medium and high) defined by the terciles of PKAA RPPA levels. To identify PKAA-associated drugs, we performed a correlation analysis between drug sensitivity data and PKAA RPPA expression by Pearson's correlation test under FDR < 0.1. Fisher's exact test was used to evaluate the enrichment of EGFR inhibitors in drugs resistant to high PKAA RPPA levels. To assess whether high PKAA was associated with high pCREB in LUAD samples from TCGA and CCLE, we normalized pCREB by its total protein expression by calculating $\log_2$ fold changes and performed differential analysis using a one-tailed *t*-test.

In vitro validation of synthetic lethality. *Cell lines*. LentiX-293T cells (Clontech) used to generate lentivirus were cultured in complete DMEM (10% FBS and 50 U per ml penicillin–streptomycin), the human LUAD cell lines A549 and H226 cells were cultured in complete RPMI medium. All cell lines were propagated at 37 °C and 5% CO_2 in a humidified atmosphere. All cell lines were authenticated by short tandem repeat analysis at the MD Anderson Cancer Center and all cells were periodically monitored for *Mycoplasma* contamination using the universal mycoplasma detection kit (American Type Culture Collection).

shRNA KD. The pKLO.1 shRNA plasmids for PKAA KD were purchased from Sigma-Aldrich (shRNA 3: TRCN0000018369, shRNA 4: TRCN0000039939, shRNA 5: TRCN0000039942). Lentivirus production and transduction of A549 and H226 cells were described previously[38]. Briefly, LentiX-293T cells were transfected with the pKLO.1-PKAA shRNA and packaging plasmids (psPAX2 and pMD2.G) using Lipofectamine 3000 (Invitrogen). Lentivirus was collected 3 days after transfection and used to transduce the respective cells by spinoculation at 1,000*g* for 3 h in the presence of 8 μg ml^{-1} polybrene. The transduced cells were selected in a puromycin-supplemented medium (1–2 μg ml^{-1}) for 7 days, as described previously[74,75].

Drug response and PKA activity assays. To test the effect of PKAA expression on EGFRi sensitivity, control and PKAA-KD A549 and H226 cells were seeded in 96-well plates (2,500 cells per well) overnight and then treated with DMSO or four different inhibitors (afatinib, gefitinib, erlotinib and osimertinib) at nine threefold serial dilutions for 72 h. The cell viability was determined using the CellTiter-Glo 2.0 cell viability assay (Promega). All inhibitors were purchased from Selleck Chemicals. To assess the PKA activity level in PKAA-KD and control A549 and H226 cells, the PKA colorimetric activity kit (Invitrogen) was used according to the manufacturer's instructions.

Analysis of proteins associated with metastatic potential. The metastasis potential scores were obtained from MetMap. We assessed the Spearman's rank correlations between the total metastasis potential scores 'all5' with CCLE RPPA500 and identified significant protein markers with FDR < 0.1. The proteins with positive or negative correlations were defined as prometastasis or antimetastasis protein markers, respectively. The two heat maps show the detailed correlations between protein makers and the metastasis potential of each specific target organ (that is, lung, liver, kidney, brain and bone; Fig. 6a). To evaluate the correlations between the identified protein markers and patient survival data in TCGA cohorts, we used the Cox proportional hazard model to select the significant protein–cancer pairs with FDR < 0.1 and obtained the hazard ratio for each significant pair. A Wilcoxon test was then performed to evaluate the significance of the difference between the log hazard ratio of prometastasis and antimetastasis RPPA markers. The metastatic status of both ACC and SARC patients was extracted from TCGA clinical data to assess the predictive power of GCLM and CHK1 RPPA expression. For survival analysis, to control potential clinical confounding factors, we further performed a multivariate Cox regression analysis by including available clinical variables (that is, age, sex and stage) in the regression model and generated forest plots for visualization.

In vitro validation of a metastasis-related protein. *Cell lines and siRNA transfections*. A549 and MDA-MB-231 cells (3.5×10^5) were seeded on six-well plates overnight. Gene KD was achieved by transiently transfecting the cells with two *CDK9*-specific siRNAs (25 nM; Sigma, siRNA 1: SASI_Hs01_00112405, siRNA 2: SASI_Hs01_00112407) using Lipofectamine 3000 (Thermo Fisher Scientific). Universal negative control 1 (Sigma, SIC001) was used as control. Then, 48 h after transfection, KD efficiencies were confirmed by real-time qPCR and cells were seeded for further experiments. *CDK9* primer sequences were as follows: forward, GAAGCTGGCAGACTTTGGGC; reverse, ACCAGAGTGTCACCACACGG.

Migration assays. For migration assays, control and *CDK9* siRNA-transfected cells (5×10^4) were seeded onto transwell inserts (Corning; 8.0 μm, 24-well) in serum-free medium. Serum-supplemented medium in the bottom chamber was used as a chemoattractant. Cells were allowed to migrate in response to the chemoattractant for 24 h, at which time the cells on the lower side of the transwell were fixed and stained with crystal violet. A total of 10–15 cell-containing fields were imaged for each siRNA and ImageJ[76] software was used to count migrated cells.

Wound-healing assays. For wound-healing assays, cells were seeded onto 96-well plates (7×10^4) and grown as a confluent monolayer. Cell monolayers were manually scratched using sterile pipette tips to generate wounds. Cell layers were washed with PBS to remove debris, supplemented with 2% FBS-containing medium and imaged every 2 h to monitor wound closure. ImageJ[76] was used for analysis.

Statistical analysis and data visualization. The sample collection for RPPA profiling was obtained from TCGA and CCLE projects and the data were not randomized. For in vitro assays, cell lines were randomly assigned to different treatment or control groups to ensure an unbiased assessment of drug responses and gene KD effects. No statistical methods were used to predetermine sample sizes but our sample sizes are similar to those reported in previous publications. Data collection and analysis were not performed blind to the conditions of the experiments. Samples that did not meet the QC metrics were excluded from the analysis. All differential analyses were tested by the Wilcoxon test or *t*-test. Multiple tests were corrected by FDR. Pearson's or Spearman's rank correlation analyses were performed to test the associations between continuous variables. Survival analyses were performed using the survival package in R; the significance was assessed by either the Cox proportional hazards regression model or the log-rank test, as determined by the type of the tested variables. The multivariate Cox proportional hazards model was used to assess the potential bias introduced by confounding factors. Data distribution was assumed to be normal but was not formally tested; nonparametric tests were used wherever possible. Detailed descriptions of the statistical tests are provided in the Methods and the respective figure legends.

Reporting summary

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

Data availability

The RPPA dataset generated in this study is accessible through TCPA data portal (https://tcpaportal.org). This portal includes one subplatform for TCGA patient tumor samples and another for CCLE cell lines. The 'dataset summary' module provides detailed information about the number of samples for each type of cancer or cell line lineage (related to Fig. 1). In TCGA patient subplatform, several analysis modules are available, including protein–protein correlation analysis, differential analysis and survival analysis (related to Fig. 2). Data related to the CCLE cell lines are hosted at the MD Anderson Cell Lines Project, a subplatform under TCPA. The analyses include protein–protein correlation analysis, protein–drug correlation analysis, protein–mutation correlation analysis and protein–dependency correlation analysis (related to Figs. 3–5). Furthermore, the comprehensive annotation for each antibody is available in the 'my protein' module on both subplatforms. Each entry in this module corresponds to a protein marker, showing relevant gene information, as well as the validation status of the antibody and its origin, source, catalog number and RRID.

We obtained CCLE-related data from DepMap (https://depmap.org/portal/), including the genomic (mutations, copy number and

DNA methylation), transcriptomic (RNA-seq and microRNA), MS, drug sensitivity, gene dependency and metabolomics data. Additional drug sensitivity data were downloaded from GDSC (https://www.cancerrxgene.org), PRISM (https://depmap.org/repurposing/) and GDSC drug combinations (https://gdsc-combinations.depmap.sanger.ac.uk). The metastatic potential data were downloaded from MetMap (https://depmap.org/metmap/). For TCGA samples, we downloaded molecular, tumor purity and clinical data from TCGA PanCanAtlas (https://gdc.cancer.gov/about- data/publications/pancanatlas). The annotations of hallmark gene sets were downloaded from Gene Set Enrichment Analysis (http://www.gsea-msigdb.org).

All other data supporting the findings of this study are available from the corresponding author on reasonable request. Source data are provided with this paper.

Code availability

All the software tools used for analysis in this study are accessible in public repositories. We used R to process the data and perform the computational analysis. SuperCurve can be found at https://bioinformatics.mdanderson.org/public-software/supercurve/. Cytoscape is available at https://cytoscape.org. ComplexHeatmap[69] and ConsensusClusterPlus[67] are R packages available on Bioconductor. We used BioRender (https://www.biorender.com) to generate the schematic diagrams and ggplot2 (ref. 77) to generate the data analysis plots. No custom code was generated in the course of this analysis.

References

1. Ding, L. et al. Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell* **173**, 305–320 (2018).
2. Hutter, C. & Zenklusen, J. C. The Cancer Genome Atlas: creating lasting value beyond its data. *Cell* **173**, 283–285 (2018).
3. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
4. Li, H. et al. The landscape of cancer cell line metabolism. *Nat. Med.* **25**, 850–860 (2019).
5. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
6. Behan, F. M. et al. Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature* **568**, 511–516 (2019).
7. Dempster, J. M. et al. Agreement between two large pan-cancer CRISPR–Cas9 gene dependency data sets. *Nat. Commun.* **10**, 5817 (2019).
8. Garnett, M. J. et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
9. Meyers, R. M. et al. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).
10. Rodriguez, H., Zenklusen, J. C., Staudt, L. M., Doroshow, J. H. & Lowy, D. R. The next horizon in precision oncology: proteogenomics to inform cancer diagnosis and treatment. *Cell* **184**, 1661–1670 (2021).
11. Akbani, R. et al. Realizing the promise of reverse phase protein arrays for clinical, translational, and basic research: a workshop report: the RPPA (Reverse Phase Protein Array) society. *Mol. Cell. Proteomics* **13**, 1625–1643 (2014).
12. Akbani, R. et al. A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat. Commun.* **5**, 3887 (2014).
13. Li, J. et al. Characterization of human cancer cell lines by reverse-phase protein arrays. *Cancer Cell* **31**, 225–239 (2017).
14. Zhao, W. et al. Large-scale characterization of drug responses of clinically relevant proteins in cancer cell lines. *Cancer Cell* **38**, 829–843 (2020).
15. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385 (2018).
16. Berger, A. C. et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* **33**, 690–705 (2018).
17. Fang, Y. et al. Sequential therapy with PARP and WEE1 inhibitors minimizes toxicity while maintaining efficacy. *Cancer Cell* **35**, 851–867 (2019).
18. Zhang, Y. et al. A pan-cancer proteogenomic atlas of PI3K/AKT/mTOR pathway alterations. *Cancer Cell* **31**, 820–832 (2017).
19. Li, J. et al. Explore, visualize, and analyze functional cancer proteomic data using The Cancer Proteome Atlas. *Cancer Res.* **77**, e51–e54 (2017).
20. Li, J. et al. TCPA: a resource for cancer functional proteomics data. *Nat. Methods* **10**, 1046–1047 (2013).
21. Siwak, D. R., Li, J., Akbani, R., Liang, H. & Lu, Y. Analytical platforms 3: processing samples via the RPPA pipeline to generate large-scale data for clinical studies. *Adv. Exp. Med. Biol.* **1188**, 113–147 (2019).
22. Nusinow, D. P. et al. Quantitative proteomics of the Cancer Cell Line Encyclopedia. *Cell* **180**, 387–402 (2020).
23. Basu, A. et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* **154**, 1151–1161 (2013).
24. Corsello, S. M. et al. Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nat. Cancer* **1**, 235–248 (2020).
25. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
26. Jin, X. et al. A metastasis map of human cancer cell lines. *Nature* **588**, 331–336 (2020).
27. Chen, M. M. et al. TCPA v3.0: an integrative platform to explore the pan-cancer analysis of functional proteomic data. *Mol. Cell. Proteomics* **18**, S15–S25 (2019).
28. Thorsson, V. et al. The immune landscape of cancer. *Immunity* **48**, 812–830 (2018).
29. Frejno, M. et al. Proteome activity landscapes of tumor cell lines determine drug responses. *Nat. Commun.* **11**, 3639 (2020).
30. Hoadley, K. A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304 (2018).
31. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
32. Tsherniak, A. et al. Defining a cancer dependency map. *Cell* **170**, 564–576 (2017).
33. Dempster, J. M. et al. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. *Genome Biol.* **22**, 343 (2021).
34. Pacini, C. et al. Integrated cross-study datasets of genetic dependencies in cancer. *Nat. Commun.* **12**, 1661 (2021).
35. Quintas-Cardama, A. & Cortes, J. Molecular biology of *BCR–ABL1*-positive chronic myeloid leukemia. *Blood* **113**, 1619–1630 (2009).
36. Pleasance, E. D. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
37. Chen, H. et al. Comprehensive assessment of computational algorithms in predicting cancer driver mutations. *Genome Biol* **21**, 43 (2020).
38. Ng, P. K. et al. Systematic functional annotation of somatic mutations in cancer. *Cancer Cell* **33**, 450–462 (2018).
39. Davies, H. et al. Mutations of the *BRAF* gene in human cancer. *Nature* **417**, 949–954 (2002).
40. Menzer, C. et al. Targeted therapy in advanced melanoma with rare *BRAF* mutations. *J. Clin. Oncol.* **37**, 3142–3151 (2019).

41. Yaeger, R. & Corcoran, R. B. Targeting alterations in the Raf–MEK pathway. *Cancer Discov.* **9**, 329–341 (2019).
42. Lavoie, H., Gagnon, J. & Therrien, M. ERK signalling: a master regulator of cell behaviour, life and fate. *Nat. Rev. Mol. Cell Biol.* **21**, 607–632 (2020).
43. Yao, Z. et al. BRAF mutants evade ERK-dependent feedback by different mechanisms that determine their sensitivity to pharmacologic inhibition. *Cancer Cell* **28**, 370–383 (2015).
44. Negrao, M. V. et al. Molecular landscape of *BRAF*-mutant NSCLC reveals an association between clonality and driver mutations and identifies targetable non-V600 driver mutations. *J. Thorac. Oncol.* **15**, 1611–1623 (2020).
45. Chen, S. H. et al. Oncogenic *BRAF* deletions that function as homodimers and are sensitive to inhibition by Raf dimer inhibitor LY3009120. *Cancer Discov.* **6**, 300–315 (2016).
46. Eisenhardt, A. E. et al. Functional characterization of a *BRAF* insertion mutant associated with pilocytic astrocytoma. *Int. J. Cancer* **129**, 2297–2303 (2011).
47. O'Neil, N. J., Bailey, M. L. & Hieter, P. Synthetic lethality and cancer. *Nat. Rev. Genet.* **18**, 613–623 (2017).
48. Lee, J. S. et al. Synthetic lethality-mediated precision oncology via the tumor transcriptome. *Cell* **184**, 2487–2502 (2021).
49. Jaaks, P. et al. Effective drug combinations in breast, colon and pancreatic cancer cells. *Nature* **603**, 166–173 (2022).
50. Abourehab, M. A. S., Alqahtani, A. M., Youssif, B. G. M. & Gouda, A. M.Globally approved EGFR inhibitors: insights into their syntheses, target kinases, biological activities, receptor interactions, and metabolism. *Molecules* **26**, 6677 (2021).
51. Sakamoto, K. M. & Frank, D. A. CREB in the pathophysiology of cancer: implications for targeting transcription factors for cancer therapy. *Clin. Cancer Res.* **15**, 2583–2587 (2009).
52. Riccio, A., Ahn, S., Davenport, C. M., Blendy, J. A. & Ginty, D. D. Mediation by a CREB family transcription factor of NGF-dependent survival of sympathetic neurons. *Science* **286**, 2358–2361 (1999).
53. Srinivasan, S. et al. Tobacco carcinogen-induced production of GM-CSF activates CREB to promote pancreatic cancer. *Cancer Res.* **78**, 6146–6158 (2018).
54. Qin, Y. et al. Interfering MSN–NONO complex-activated CREB signaling serves as a therapeutic strategy for triple-negative breast cancer. *Sci. Adv.* **6**, eaaw9960 (2020).
55. Fares, J., Fares, M. Y., Khachfe, H. H., Salhab, H. A. & Fares, Y. Molecular principles of metastasis: a hallmark of cancer revisited. *Signal Transduct. Target. Ther.* **5**, 28 (2020).
56. Khan, I. & Steeg, P. S. Metastasis suppressors: functional pathways. *Lab. Invest.* **98**, 198–210 (2017).
57. Nguyen, D. T. et al. Pharos: collating protein information to shed light on the druggable genome. *Nucleic Acids Res.* **45**, D995–D1002 (2017).
58. Dou, Y. et al. Proteogenomic characterization of endometrial carcinoma. *Cell* **180**, 729–748 (2020).
59. Gillette, M. A. et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell* **182**, 200–225 (2020).
60. Vasaikar, S. et al. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. *Cell* **177**, 1035–1049 (2019).
61. Tibes, R. et al. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Mol. Cancer Ther.* **5**, 2512–2521 (2006).
62. Hennessy, B. T. et al. A technical assessment of the utility of reverse phase protein arrays for the study of the functional proteome in non-microdissected human breast cancers. *Clin. Proteomics* **6**, 129–151 (2010).
63. Hu, J. et al. Non-parametric quantification of protein lysate arrays. *Bioinformatics* **23**, 1986–1994 (2007).
64. Neeley, E. S., Baggerly, K. A. & Kornblau, S. M. Surface adjustment of reverse phase protein arrays using positive control spots. *Cancer Inform.* **11**, 77–86 (2012).
65. Ju, Z. et al. Development of a robust classifier for quality control of reverse-phase protein arrays. *Bioinformatics* **31**, 912–918 (2015).
66. Gonzalez-Angulo, A. M. et al. Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer. *Clin. Proteomics* **8**, 11 (2011).
67. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).
68. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
69. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
70. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
71. Sanchez-Vega, F. et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337 (2018).
72. Dogruluk, T. et al. Identification of variant-specific functions of *PIK3CA* by rapid phenotyping of rare mutations. *Cancer Res.* **75**, 5341–5354 (2015).
73. Tsang, Y. H. et al. Functional annotation of rare gene aberration drivers of pancreatic cancer. *Nat. Commun.* **7**, 10500 (2016).
74. Cheung, L. W. et al. Naturally occurring neomorphic *PIK3R1* mutations activate the MAPK pathway, dictating therapeutic response to MAPK pathway inhibitors. *Cancer Cell* **26**, 479–494 (2014).
75. Liang, H. et al. Whole-exome sequencing combined with functional genomics reveals novel candidate driver cancer genes in endometrial cancer. *Genome Res.* **22**, 2120–2129 (2012).
76. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
77. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* 2nd edn (Springer, 2016).

Acknowledgements

This study was supported by the NIH (U01CA217842, U01CA281902 and U01CA253472 to H.L. and G.B.M., U24CA264128 to H.L., G.B.M. and R.A., R01CA251150 and P50CA281701 to H.L., P50CA221703 to H.L. and M.A.D., U24CA210950, U24CA210949 and U24CA264006 to R.A. and G.B.M., R50CA221675 to Y.L., P50CA270907 to J.L. and the Cancer Center Support Grant P30CA016672), a kind gift from the Sheldon and Miriam Adelson Medical Research Foundation, Susan G. Komen (SAC110052), the Ovarian Cancer Research Foundation (545152) and the Breast Cancer Research Foundation (BCRF-18-110) to G.B.M., a Department of Defense Congressionally Directed Medical Research Program award (W81XWH-16-1-0237) to R.A., grants from the Cancer Prevention and Research Institute of Texas (RP210042 and RP160015) to S.V.K. and the University Cancer Foundation through the Institutional Research Grant program at the University of Texas MD Anderson Cancer Center to J.L.

Author contributions

R.A., G.B.M. and H.L. conceptualized the project. J.L., W.L., Z.J., S.V.K., P.K.-S.N., H.C., M.A.D., Y.L., R.A., G.B.M. and H.L. contributed to the data analysis and result discussion. K.M., H.K., Z.Z., P.K.-S.N. and Y.L. contributed to the experiments. J.L. and H.L. wrote the manuscript with input from other authors. H.L. supervised the project.

Competing interests

M.A.D. has been a consultant to Roche/Genentech, Array, Pfizer, Novartis, BMS, GSK, Sanofi-Aventis, Vaccinex, Apexigen, Eisai, Iovance and ABM Therapeutics and he has been the PI of research grants to MD Anderson from Roche/Genentech, GSK, Sanofi-Aventis, Merck, Myriad, Oncothyreon and ABM Therapeutics. R.A. is a bioinformatics consultant for the University of Houston. G.B.M. is a scientific advisory board member or consultant for AstraZeneca, Chrysalis Biotechnology, GSK, ImmunoMET, Ionis, Lilly, PDX Pharmaceuticals, Signalchem Lifesciences, Symphogen, Tarveda, Turbine and Zentalis Pharmaceuticals, has stocks, options or financial considerations with Catena Pharmaceuticals, ImmunoMet, SignalChem and Tarveda and has licensed the HRD assay to Myriad Genetics and DSP patents to Nanostring. H.L. is a shareholder and advisor for Precision Scientific. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43018-024-00817-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43018-024-00817-x.

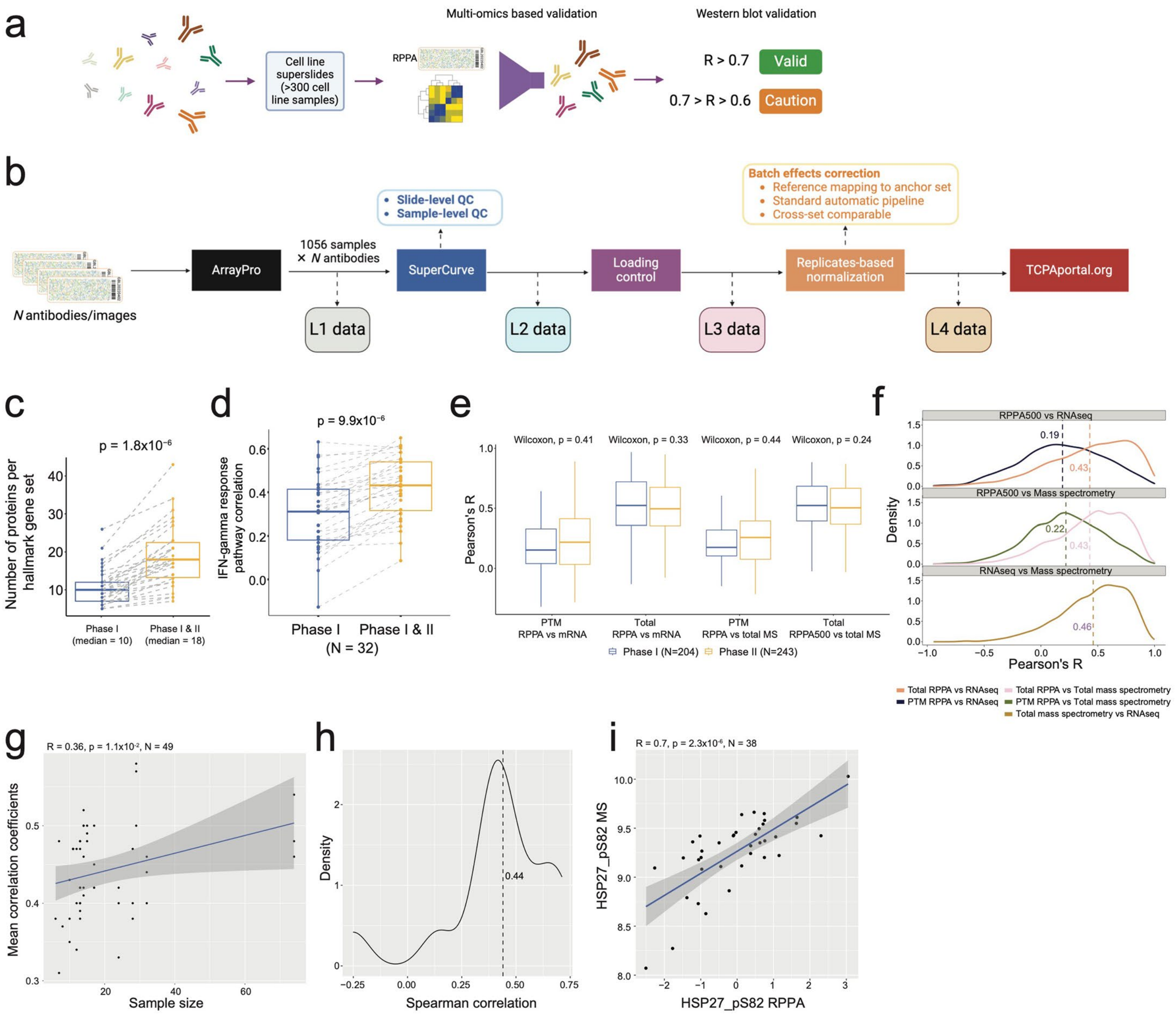
Correspondence and requests for materials should be addressed to Rehan Akbani, Gordon B. Mills or Han Liang.

Peer review information *Nature Cancer* thanks Bing Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

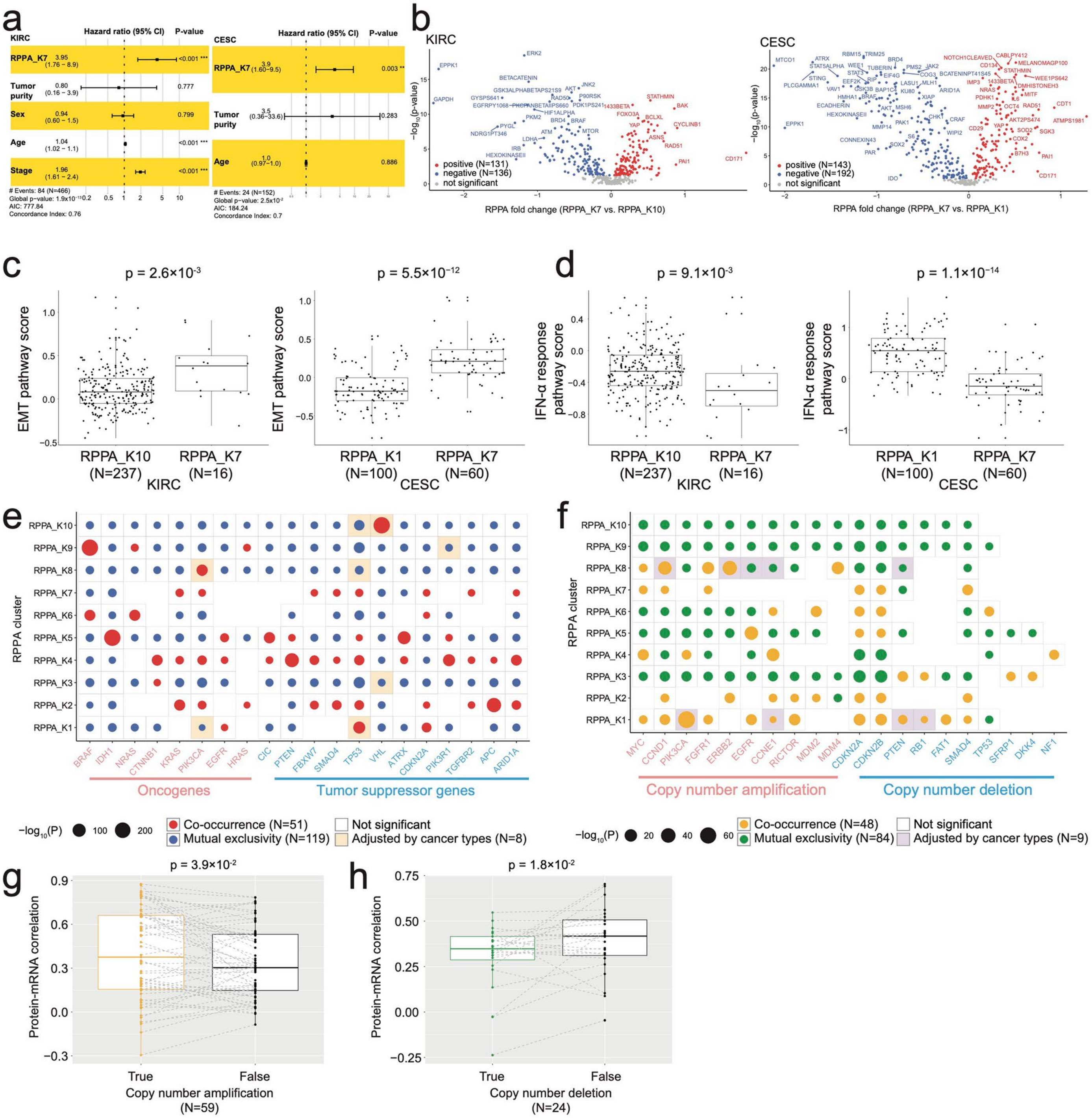
Reprints and permissions information is available at www.nature.com/reprints.

---

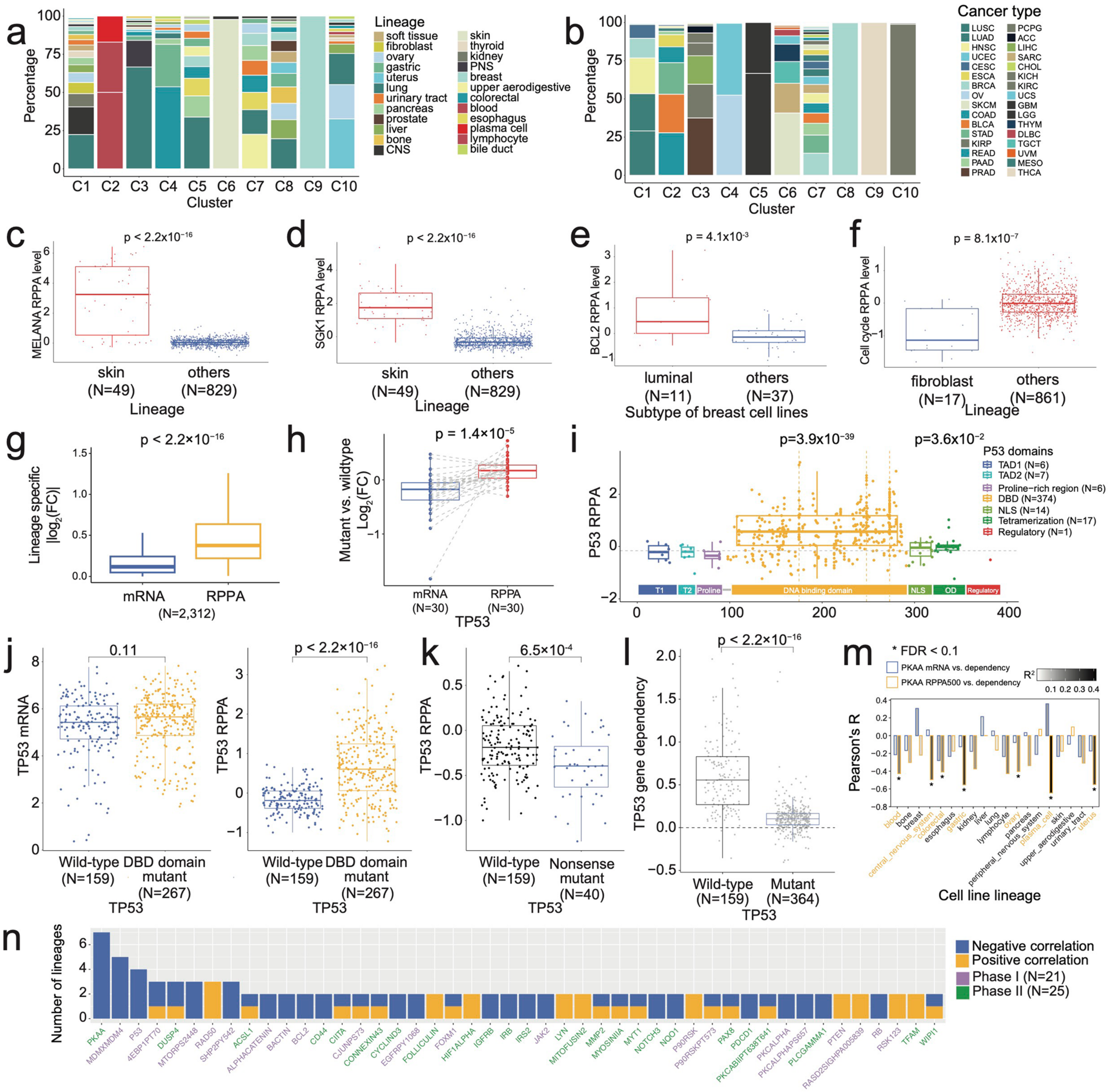
[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [2]Institute for Personalized Cancer Therapy, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [3]Department of Melanoma Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [4]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [5]Knight Cancer Institute and Cell, Developmental and Cancer Biology, Oregon Health and Science University, Portland, OR, USA. [6]Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [7]Institute for Data Science in Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [8]Present address: The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. [9]Present address: Department of Pediatrics, University of Connecticut Health Center, Farmington, CT, USA. ✉e-mail: rakbani@mdanderson.org; millsg@ohsu.edu; hliang1@mdanderson.org



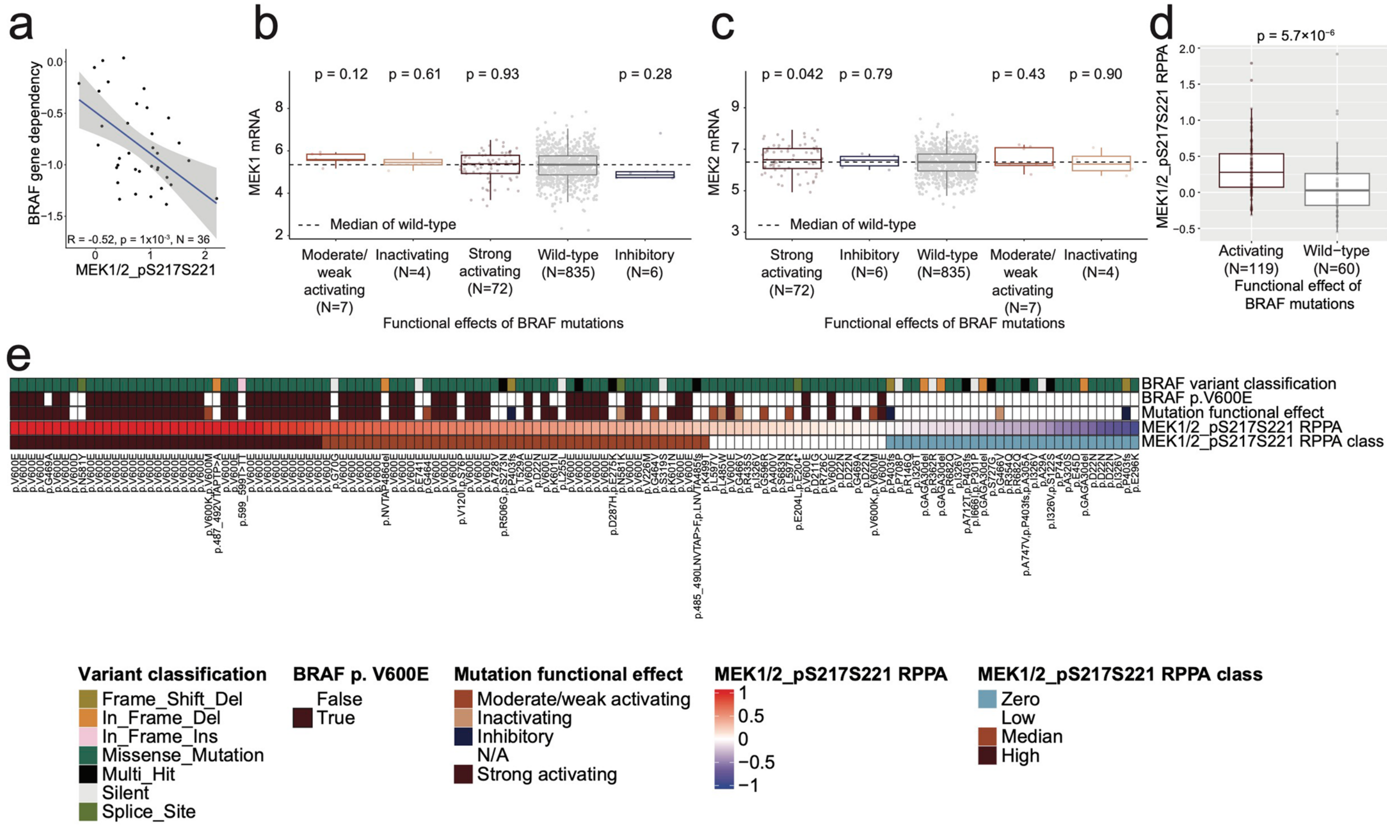
Extended Data Fig. 1 | Quality control of RPPA data. (**a**) Summary of antibody selection and validation process. (**b**) Overview of RPPA data normalization, with rigorous QC parameters or controls at each step. (**c**) Comparison of protein markers in Phase I and Phase I & II. (**d**) Boxplots showing differential correlations of RPPA-based IFN-gamma response pathway score and that reported by literature in Phase I and Phase I & II. N = 32 represents the number of cancer types. (**e**) Comparison of correlations between RPPA and MS/RNA-seq in Phase I (N = 204) and Phase II (N = 243). (**c**-**e**) P-values are based on paired Wilcoxon tests. N represents the number of protein markers. (**f**) The distributions of the lineage-specific expression correlations of RPPA-based proteins with RNA-seq-based mRNA (top panel), MS-based total proteins (middle panel), and correlations between MS-based total proteins and RNA-seq-based mRNA (bottom panel). The distributions are shown for total proteins and PTM proteins, respectively. (**g**) A scatter plot showing the correlation between the sample sizes and the mean correlations across different lineages. N = 49 represents the number of mean correlations. (**h**) The distribution of expression correlations between RPPA-based and MS-based phosphorylated proteins in NCI60 cell lines. (**i**) A scatter plot showing a representative example of the phosphoprotein, HSP27_pS82, between the RPPA and the MS data. N = 38 represents the number of cell line samples. (**c**-e) The middle line in the box is the median, the bottom and top of the box are the first and third quartiles, and the whiskers extend to the 1.5× interquartile range of the lower and the upper quartiles, respectively. (**g**, **i**) Shaded areas denote the 95% confidence intervals. The p-values are based on Spearman's correlation coefficient test.



Extended Data Fig. 2 | Pan-cancer analysis of TCGA samples. (**a**) Forest plots of hazard ratios for cluster RPPA_K7 and clinical variables. The p-values are based on a multivariate Cox proportional hazards model. The center point of each horizontal bar represents the estimated hazard ratio. The p-values are based on a multivariate Cox proportional hazards model. (**b**) Volcano plots showing differentially expressed protein markers between the corresponding cluster pairs identified from the patient survival analysis. The p-values are based on the Wilcoxon test. N represents the number of significant positive or negative protein markers (KIRC: N = 131, and 136 for significantly positive and negative protein markers respectively; CESC: N = 143, and 192 for significantly positive and negative protein markers respectively). (**c**, **d**) Significantly up and downregulated pathways, EMT (**c**) and IFN-α (**d**), identified by pathway analysis for K7 (N = 16) vs. K10 (N = 237) in KIRC and K7 (N = 60) vs. K1 (N = 100) in CESC. The p-values are based on the Wilcoxon test. N represents the number of patient samples in each group. (**e**, **f**) Associations of RPPA clusters with driver mutations (**e**) and copy number alterations (**f**). Co-occurrence and mutual exclusivity are shown in different colors. The p-values are based on the Chi-squared test. Significant hits with adjustments for cancer types are shaded. N represents the number of significant hits. (**g**, **h**) Boxplots showing differential patterns of protein-mRNA coupling across TCGA patient cohorts by copy number alteration of *cis*-cancer driver genes (copy number amplification: N = 59; and copy number deletion: N = 24). The p-values are based on the one-sided paired Wilcoxon test. N represents the number of protein-alteration pairs. (**c**, **d**, **g**, **h**) The middle line in the box is the median, the bottom and top of the box are the first and third quartiles, and the whiskers extend to the 1.5× interquartile range of the lower and the upper quartiles, respectively.

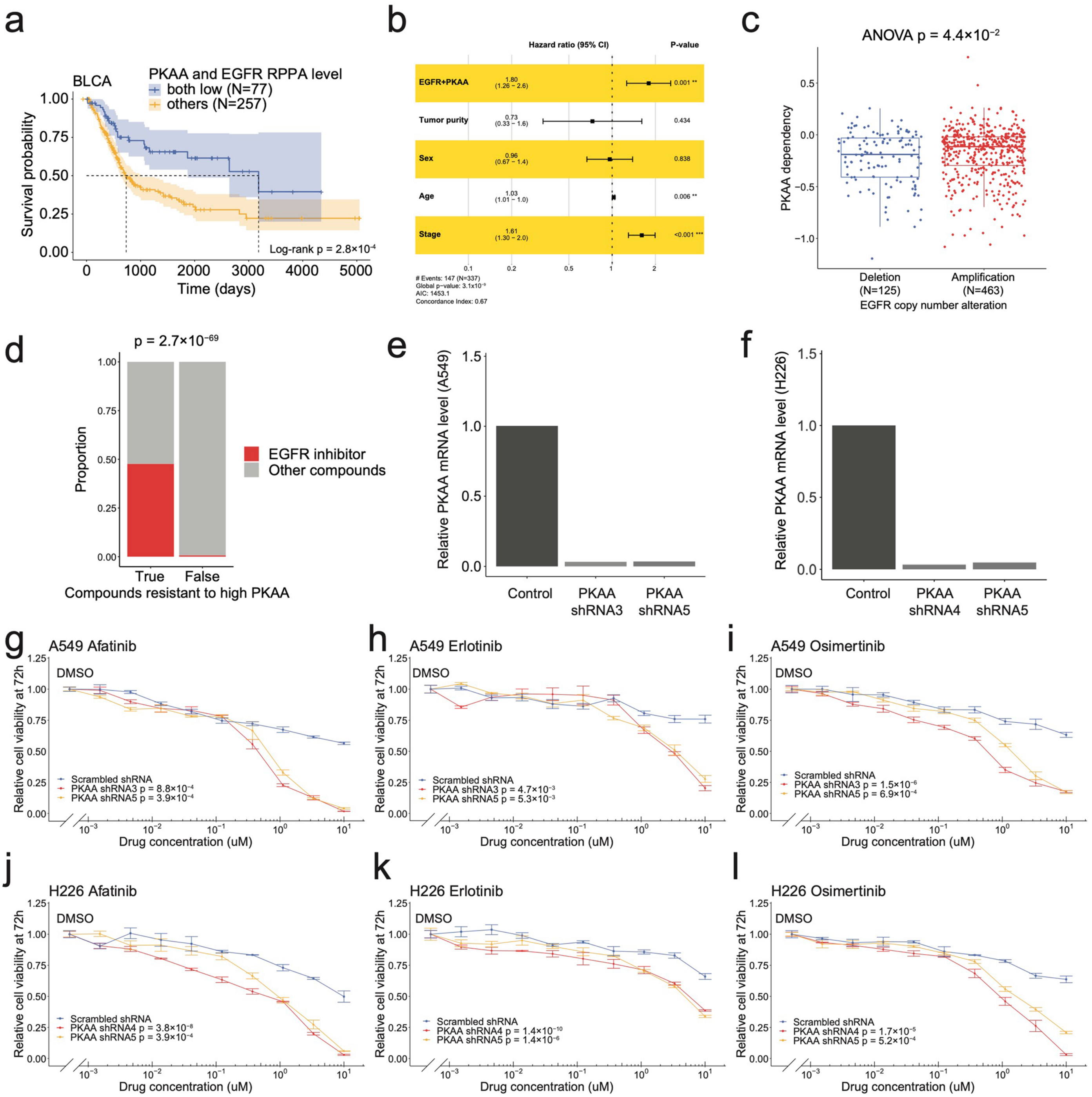


Extended Data Fig. 3 | Pan-cancer analysis of CCLE samples. (**a**, **b**) Sample distributions of different cell lineages (**a**) and cancer types (**b**) in RPPA-based clusters. (**c**, **d**) Box plots of MELANA (**c**) and SGK1 (**d**) in skin cancer (N = 49) and other cell lines (N = 829). (**e**) Box plot of BCL2 expression in luminal (N = 11) and other subtypes (N = 37) of breast cancer. (**f**) Box plot of expression of the proteins involved in the cell cycle in fibroblasts (N = 17) and other cell lines (N = 861). (**c**-**f**) N represents the number of cell line samples. (**g**) Box plots showing the comparison of mRNA-based $|\log_2(FC)|$ vs. protein-based $|\log_2(FC)|$ between lineages for the lineage-specific markers. N = 2,312 represents the number of protein-lineage pairs. (**c**-**g**) The p-values are based on the two-sided Wilcoxon test. (**h**) Box plots showing differential fold changes of TP53 mRNA and protein expression between *TP53* mutant and wild-type TCGA tumor samples. The p-value is based on the paired Wilcoxon test. N = 30 represents the number of cancer types. (**i**) Effect of mutations in different P53 domains on P53 protein expression. The p-values are based on Wilcoxon test. (**j**) Box plots showing differential *TP53* mRNA (left) and P53 protein (right) expression in *TP53* wild-type samples (N = 159) and those harboring mutations in P53 DNA binding domain (DBD domain; N = 267). (**k**) Box plots showing differential RPPA expression of P53 between *TP53* nonsense mutant (N = 40) and wild-type samples (N = 159). (**l**) Box plots showing differential gene dependency of *TP53* between *TP53* mutant (N = 364) and wild-type samples (N = 159). (**i**-**l**) N represents the number of cell line samples in each group. (**m**) A bar plot showing the correlation of gene dependency with PKAA protein and mRNA expression in different cell line lineages. Significant correlations (FDR < 0.1) are marked by an asterisk. (**n**) A bar plot summarizing the number of lineages in which each protein marker is significantly associated with their gene dependency but not with mRNA (FDR < 0.1). N represents the number of protein markers (phase I: N = 21; phase II: N = 25). (**c**-**l**) The middle line in the box is the median, the bottom and top of the box are the first and third quartiles, and the whiskers extend to the 1.5× interquartile range of the lower and the upper quartiles, respectively.

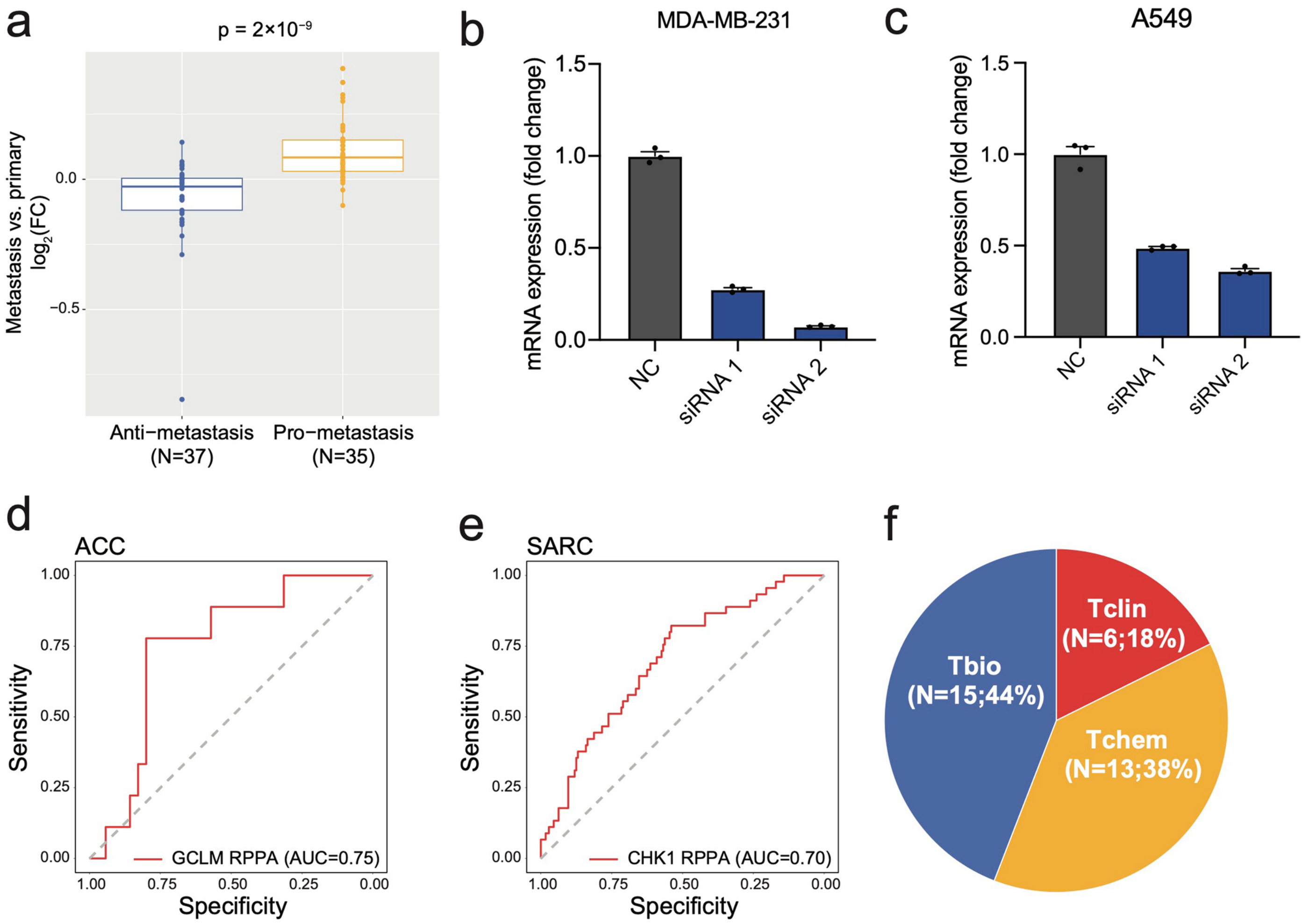


Extended Data Fig. 4 | Differential effects of BRAF mutations on MEK1/2 mRNA and protein expression. (**a**) A scatter plot showing the correlation between MEK1/2_pS217S221 RPPA and *BRAF* gene dependency in melanoma cell lines (N = 36). Shaded areas denote the 95% confidence intervals. (**b**-**c**) Box plots showing differential expression of MEK1 mRNA (**b**), MEK2 mRNA (**c**) based on the functional effects of *BRAF* mutations characterized by the cell viability assays. (**a**-**c**) N represents the number of cell line samples. (**d**) Box plots showing the differential expression of MEK1/2_pS217S221 RPPA between the patient samples with an activating *BRAF* mutation (N = 119) and those with wild-type *BRAF* (N = 60). (**b**-**d**) The p-values are based on Wilcoxon test. N represents the number of patient samples. (**e**) A heatmap showing detailed information on the *BRAF* mutations, their functional effects, and the corresponding MEK1/2_pS217S221 RPPA expression pattern. (**b**-**d**) The middle line in the box is the median, the bottom and top of the box are the first and third quartiles, and the whiskers extend to the 1.5× interquartile range of the lower and the upper quartiles, respectively.

https://doi.org/10.1038/s43018-024-00817-x



Extended Data Fig. 5 | In vitro validation of PKAA-EGFR synthetic lethal interaction. (**a**) A Kaplan-Meier plot showing distinct survival probabilities of bladder cancer patients with both low PKAA and EGFR protein levels. The p-value is based on a log-rank test. Shaded areas denote the 95% confidence intervals. N represents the number of patients in each group (both low: N = 77; and others: N = 257). (**b**) A forest plot of hazard ratios for PKAA protein and clinical variables. The p-values are based on a multivariate Cox proportional hazards model. The center point of each horizontal bar represents the estimated hazard ratio. N represents the number of patients. (**c**) Box plots showing the differential gene dependency of PKAA between samples with EGFR deletion (N = 125) and amplification (N = 463). The middle line in the box is the median, the bottom and top of the box are the first and third quartiles, and the whiskers extend to the 1.5× interquartile range of the lower and the upper quartiles, respectively. N represents the number of cell line samples. (**d**) Enrichment of EGFR inhibitors in drugs resistant to high PKAA levels. The p-value is based on Fisher's exact test. (**e**, **f**) Relative mRNA level of PKAA in PKAA-KD A549 or H226 cells. N = 2 independent replicates were examined for each condition. (**g**-**l**) Drug response assays at 72 h for A549 (**g**-**i**) or H226 (**j**-**l**) PKAA-KD and control cells treated with three EGFR inhibitors, Afatinib, Erlotinib, and Osimertinib (DMSO and 9 drug concentrations). N = 3 independent replicates were examined for each treatment and perturbation. Data are shown as mean ± SEM. The p-values are based on ANOVA.



Extended Data Fig. 6 | Evaluation of metastasis markers in cancer cell lines and patient samples. (**a**) Box plots showing the $\log_2$(FC) between metastatic and primary cell lines of anti- (N = 37) and pro-metastasis (N = 35) marker RPPA expression. The p-value is based on the Wilcoxon test. The middle line in the box is the median, the bottom and top of the box are the first and third quartiles, and the whiskers extend to the 1.5× interquartile range of the lower and the upper quartiles, respectively. N represents the number of protein markers in each group. (**b**, **c**) Relative mRNA level of *CDK9* in *CDK9*-KD and control MD-MB-231 or A549 cells. N = 3 independent replicates were examined for each perturbation. Data are shown as mean ± SEM. (**d**, **e**) ROC curves showing predictive powers of GCLM expression in ACC patients (N = 44) (**d**) and CHK1 in SARC patients (N = 223) (**e**) between metastatic and non-metastatic primary tumor samples. N represents the number of patients. (**f**) A pie chart showing the distribution of drug development levels for all the identified pro-metastasis protein markers. The annotation data was obtained from the Pharos database. N represents the number of protein markers in each group (Tclin: N = 6; Tchem: N = 13; and Tbio: N = 15).

nature portfolio

Corresponding author(s): Han Liang

Last updated by author(s): Jun 7, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	The analyses were performed using R (v4.3.1). RPPA data curve fitting analysis was based on SuperCurve (v1.5). Network visualization was performed using Cytoscape (v3.9.1). Heatmaps were built and visualized using ComplexHeatmap (v2.18.0) and ConsensusClusterPlus (v1.66.0). BioRender was used to generate schematic diagrams. The ggplot2 (v3.4.4) R package was used to generate data analysis plots. ImageJ (v1.53) was used for migration and wound healing analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The RPPA dataset generated in this study is accessible through the TCPA data portal at: https://tcpaportal.org. CCLE-related data were downloaded from DepMap (https://depmap.org/portal/), including genomic (mutations, copy number, DNA methylation), transcriptomic (RNA-seq and miRNA), mass spectrometry, drug sensitivity, gene dependency, and metabolomics data. Additional drug sensitivity data were downloaded from GDSC (https://www.cancerrxgene.org), PRISM (https://depmap.org/repurposing/), and GDSC drug combinations (https://gdsc-combinations.depmap.sanger.ac.uk). Metastatic potential data were downloaded from MetMap (https://depmap.org/metmap/). For TCGA samples, molecular, tumor purity, and clinical data were downloaded from the TCGA PanCanAtlas (https://gdc.cancer.gov/about-data/publications/pancanatlas). The annotations of hallmark gene sets were downloaded from GSEA (http://www.gsea-msigdb.org).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Yes. Our study is based on pre-existing TCGA samples, which has been widely used in the community.
Reporting on race, ethnicity, or other socially relevant groupings	This was not considered in the study.
Population characteristics	The summary information on TCGA samples profiled in this study is included in the supplementary table. All the clinical information is available at the TCGA GDC portal, as documented in the Methods section.
Recruitment	Patients were recruited through the TCGA project.
Ethics oversight	Ethics oversight and written informed consent were through the TCGA project.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample sizes for both cell lines and patient tumors were predefined in the RPPA phase I study. The details were included in the original marker papers from CCLE and TCGA projects (PMIDs: 24037243, 24871328, 31201206, 31068700).
Data exclusions	In each analysis, samples were filtered by considering both the quality and availability across different molecular and phenotype datasets.
Replication	Replicate samples were used to merge different RPPA slides to minimize the batch effects. All attempts at replication were successful.
Randomization	Sample randomization were performed when arranging the samples on each RPPA array to control the confounding factors. The detailed information was documented in previous publications.
Blinding	The investigators were not blinded to the sample information during the study. However, this lack of blinding did not impact the data collection and generation phases, as sample information was not utilized in these initial steps. Instead, specific details such as cell line lineage and patient clinical data were only applied in the downstream analyses, where blinding is less critical due to the objective nature of the analytical methods used.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Biotinylated Goat anti-Rabbit IgG H+L (Invitrogen, Cat#: 31822, Lot#: XC3537505, dilution: 1:1000) Biotinylated Goat anti-Mouse IgG H+L (Vector Laboratories, Cat#: BA-9200, Lot#: ZD0801, dilution: 1:10000) Biotinylated Rabbit anti-Goat IgG H+L (Fisher Scientific, Cat#: NC9376096, Lot#: ZB0923, dilution: 1:10000)
Validation	Each antibody was assessed based on the protein expression obtained from cultured cell lines or tumor tissues. Antibodies with a good correlation Western blot were considered as valid.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	The MDA-MB-231 and NCI-H226 cell lines were obtained from the MD Anderson Characterized Cell Line Core Facility. The A549 cell line was purchased from American Type Culture Collection. Lenti-X293T cells were obtained from Dr. Gordon B. Mills' laboratory.
Authentication	All cell lines were authenticated using short tandem repeat analysis at the MD Anderson Characterized Cell Line Core Facility.
Mycoplasma contamination	Cell lines were routinely tested and were negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No misidentified cell lines were used in this study.

Plants

Seed stocks	n/a
Novel plant genotypes	n/a
Authentication	n/a