nature cancer

Analysis

https://doi.org/10.1038/s43018-024-00869-z

Mapping the functional network of human cancer through machine learning and pan-cancer proteogenomics

Received: 6 April 2023

Accepted: 25 October 2024

Published online: 11 December 2024

Check for updates

Zhiao Shi[1,2,3], Jonathan T. Lei[1,2,3], John M. Elizarraras[1,2] & Bing Zhang [1,2] ✉

Large-scale omics profiling has uncovered a vast array of somatic mutations and cancer-associated proteins, posing substantial challenges for their functional interpretation. Here we present a network-based approach centered on FunMap, a pan-cancer functional network constructed using supervised machine learning on extensive proteomics and RNA sequencing data from 1,194 individuals spanning 11 cancer types. Comprising 10,525 protein-coding genes, FunMap connects functionally associated genes with unprecedented precision, surpassing traditional protein–protein interaction maps. Network analysis identifies functional protein modules, reveals a hierarchical structure linked to cancer hallmarks and clinical phenotypes, provides deeper insights into established cancer drivers and predicts functions for understudied cancer-associated proteins. Additionally, applying graph-neural-network-based deep learning to FunMap uncovers drivers with low mutation frequency. This study establishes FunMap as a powerful and unbiased tool for interpreting somatic mutations and understudied proteins, with broad implications for advancing cancer biology and informing therapeutic strategies.

Advancements in next-generation sequencing and mass spectrometry (MS) have transformed cancer research. Large-scale initiatives such as The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) have greatly deepened our understanding of cancer, revealing a vast array of somatic mutations and cancer-associated proteins. These advancements present new challenges in the functional interpretation of identified mutations and proteins, especially for the numerous low-frequency mutations[1] and understudied proteins[2].

Protein–protein interaction networks have been instrumental in prioritizing somatic mutations and predicting the functions of uncharacterized proteins[3–5]. However, many of the known interactions were identified in noncancer contexts, limiting their relevance to cancer research. Recent efforts have started to address this gap by mapping interactions for selected proteins in specific cancer cell lines[6,7]. Despite these advances, unbiased, genome-scale identification of protein–protein interactions across diverse cancer types remains a daunting task. Moreover, in vitro cell line models have inherent limitations, such as the absence of the tumor microenvironment. mRNA coexpression has also been used to infer functional associations but with varied success[8,9]. Studies have shown that protein expression data are more closely aligned with gene function and that protein coexpression is a much stronger predictor of functional association than mRNA coexpression[10–14].

In this paper, we introduce FunMap, a functional network of 10,525 genes constructed using a supervised machine learning method that integrates proteomics and RNA sequencing (RNAseq) data from 11 cancer types, recently harmonized by the CPTAC pan-cancer working group[15]. FunMap connects functionally related genes with unprecedented precision, surpassing existing protein–protein interaction networks. Through network analysis, FunMap uncovers protein modules and a hierarchical modular organization linked to cancer hallmarks

[1]Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX, USA. [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. [3]These authors contributed equally: Zhiao Shi, Jonathan T. Lei. ✉e-mail: bing.zhang@bcm.edu

and clinical phenotypes, predicts the functions of understudied cancer proteins, offers deeper insights into established cancer drivers and identifies drivers with low mutation frequency. To facilitate broader use in cancer research, we provide an interactive web application (https://funmap.linkedomics.org/) and source code (https://github.com/bzhanglab/funmap).

Results

Protein coexpression strongly predicts cofunctionality

We used MS-based proteomics and RNAseq data from 11 tumor cohorts (Supplementary Table 1) to quantify gene coexpression at the protein and mRNA levels, respectively. Cancer types included breast invasive carcinoma (BRCA), clear cell renal cell carcinoma (CCRCC), colon adenocarcinoma (COAD), glioblastoma (GBM), hepatocellular carcinoma (HCC), head and neck squamous cell carcinoma (HNSCC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LSCC), ovarian serous cystadenocarcinoma (OV), pancreatic ductal adenocarcinoma (PDAC) and uterine corpus endometrial carcinoma (UCEC). Tumor samples ranged from 83 to 159 per cohort and five cancer types also had sufficient normal samples with proteomics and RNAseq data, leading to 16 proteomics and 16 RNAseq datasets (Fig. 1a). Each proteomics dataset included 7,961–11,815 genes (Fig. 1b), with a median of 10,441 and a union of 14,070 genes, among which 6,602 were identified across all 16 datasets and 10,024 were identified in 10 or more datasets (Fig. 1c). Each RNAseq dataset included 17,733–19,113 genes (Fig. 1b), with a median of 18,740 and a union of 19,855 genes, among which 15,603 were identified across all 16 datasets (Fig. 1c).

To assess the relationship between gene coexpression and cofunctionality, we used a previously published 'gold standard' derived from the Reactome pathway database[12]. This gold standard defines gene pairs coannotated in the same 'detailed' pathway (≤200 genes) as positive pairs and those without shared pathway annotations as negative pairs. It includes 205,284 positive and 11,327,528 negative gene pairs. This extensive dataset allowed us to quantify the functional relevance of any specific set of gene pairs by calculating the log likelihood ratio (LLR), with higher LLRs indicating stronger evidence of functional relevance (Methods).

For each proteomics and RNAseq dataset, we ranked gene pairs by their Pearson's correlation coefficients (PCCs) and computed LLRs for the top 10,000–300,000 pairs. LLRs showed a decreasing trend across all datasets (Fig. 1d). In most tumor datasets, proteomics data consistently yielded higher LLRs than RNAseq, indicating greater functional relevance. However, in normal datasets, proteomics LLRs were similar to or lower than RNAseq LLRs. This may be explained by the low intersample heterogeneity in normal protein datasets (Extended Data Fig. 1), hindering the detection of correlations between functionally related genes. The low intersample heterogeneity likely also contributed to the lower LLRs in normal protein datasets compared to tumor protein datasets. Interestingly, despite lower heterogeneity in tumor protein datasets compared to tumor RNA datasets (Extended Data Fig. 1), the higher LLRs in the protein data suggest that this level of heterogeneity is sufficient for detecting functionally relevant correlations.

To delve deeper into how mRNA and protein coexpression patterns relate to gene cofunctionality within the tumor datasets, we grouped gene pairs into 400 two-dimensional bins on the basis of their correlations in both proteomics and RNAseq data and then computed LLRs for each bin (Fig. 1e). Gene pairs with higher protein correlation consistently displayed elevated LLR scores, even when mRNA correlation was moderately positive or even negative. While gene pairs with higher mRNA correlation also tended to have higher LLR scores, these higher scores were more frequently observed in areas where there were strong correlations at both mRNA and protein levels. Together, these results demonstrate that, while both protein and mRNA correlations indicate gene cofunctionality, protein correlation is a much stronger predictor.

A machine-learned functional map

We used supervised machine learning to integrate the diverse predictive signals from all 32 proteomics and RNAseq datasets to construct a comprehensive functional network. Normal sample datasets were included because they were derived from tumor-adjacent normal tissues, which provide clinically relevant biological information[16]. Despite varying magnitudes, each dataset displayed functional relevance (LLR > 1; Fig. 1d). To account for differences in sample size and intersample heterogeneity across datasets, we computed PCC-based mutual rank (MR) scores for all gene pairs within each dataset (Methods), as MR is a robust metric for assessing gene coexpression across diverse datasets[17].

We used 50% of the gold-standard positive and negative gene pairs as training data to build an extreme gradient boosting (XGBoost) model, using MR scores from the 32 datasets as features to distinguish the positive and negative gene pairs (Methods). Feature importance analysis revealed that tumor protein features contributed the most (61.5%), followed by tumor RNA (20.7%), normal RNA (9.0%) and normal protein (8.8%) (Extended Data Fig. 2). Among individual datasets, the tumor protein data from LSCC contributed the most.

The trained model was applied to all 98,975,415 gene pairs, which were then sorted by predicted probabilities. LLRs were computed using the remaining 50% set-aside gold-standard gene pairs for the top-ranked gene pairs from the top 50,000 to 250,000 (Fig. 2a). Similarly, we trained two additional XGBoost models using only the 16 proteomics datasets or the 16 RNAseq datasets and plotted the LLR curves. For comparison, we included LLR curves from a baseline method based on average PCCs across the 32 datasets and the LSCC tumor protein data alone. Interestingly, the LSCC tumor proteomics dataset performed as well as or better than the combined RNAseq datasets, underscoring the pivotal role of protein-level regulation in coordinating gene function. The XGBoost model combining all datasets clearly outperformed the baseline method according to average PCCs, highlighting the advantage offered by machine learning. It also outperformed the model combining only the proteomics datasets, which in turn outperformed the model combining only the RNAseq datasets or the LSCC tumor proteomics data alone, demonstrating the value of data integration in gene cofunctionality prediction.

Applying an LLR cutoff of 3.912 (that is, a likelihood ratio (LR) of 50) to the results from the XGBoost model combining all 32 datasets yielded a functional association network with 10,525 genes and 196,800 edges, which was named FunMap (Supplementary Table 2). With an LR of 50, edges are 50 times more likely to connect functionally associated gene pairs than unrelated pairs. We compared FunMap's functional relevance and proteome coverage to other networks used in systems biology studies (Fig. 2b). FunMap and the ProHD[12], both based primarily on protein coexpression, showed similar LR scores (50 and 56, respectively), although ProHD covered only 2,680 genes. These scores were much higher than those of BioPlex[18] (LR = 28), HuRI[19] (LR = 10), HI-Union[19] (LR = 10) and BioGRID[20] (LR = 14), networks based on experimentally obtained protein–protein interaction data or curated protein and genetic interaction data. While FunMap showed higher proteome coverage than HuRI and HI-Union, BioPlex and BioGRID covered more genes (13,854 and 17,259, respectively). The STRING network[21] had the highest LR score (LR = 187) and deep coverage of 16,351 genes; however, unlike the other purely data-driven networks, it incorporated existing knowledge during network construction, including that used for our evaluation.

Genes in FunMap overlapped significantly with those in other networks (Fig. 2c) but its edges showed limited overlaps (Fig. 2d), indicating a substantial number of additional functional associations. While tumor versus normal differences were not used in FunMap's construction, analysis of the five cancer types with normal samples revealed that 60–74% of FunMap edges connected genes with consistent significant overexpression or underexpression in tumors (adjusted $P < 0.01$, Wilcoxon rank-sum test; Fig. 2e). These percentages significantly exceeded

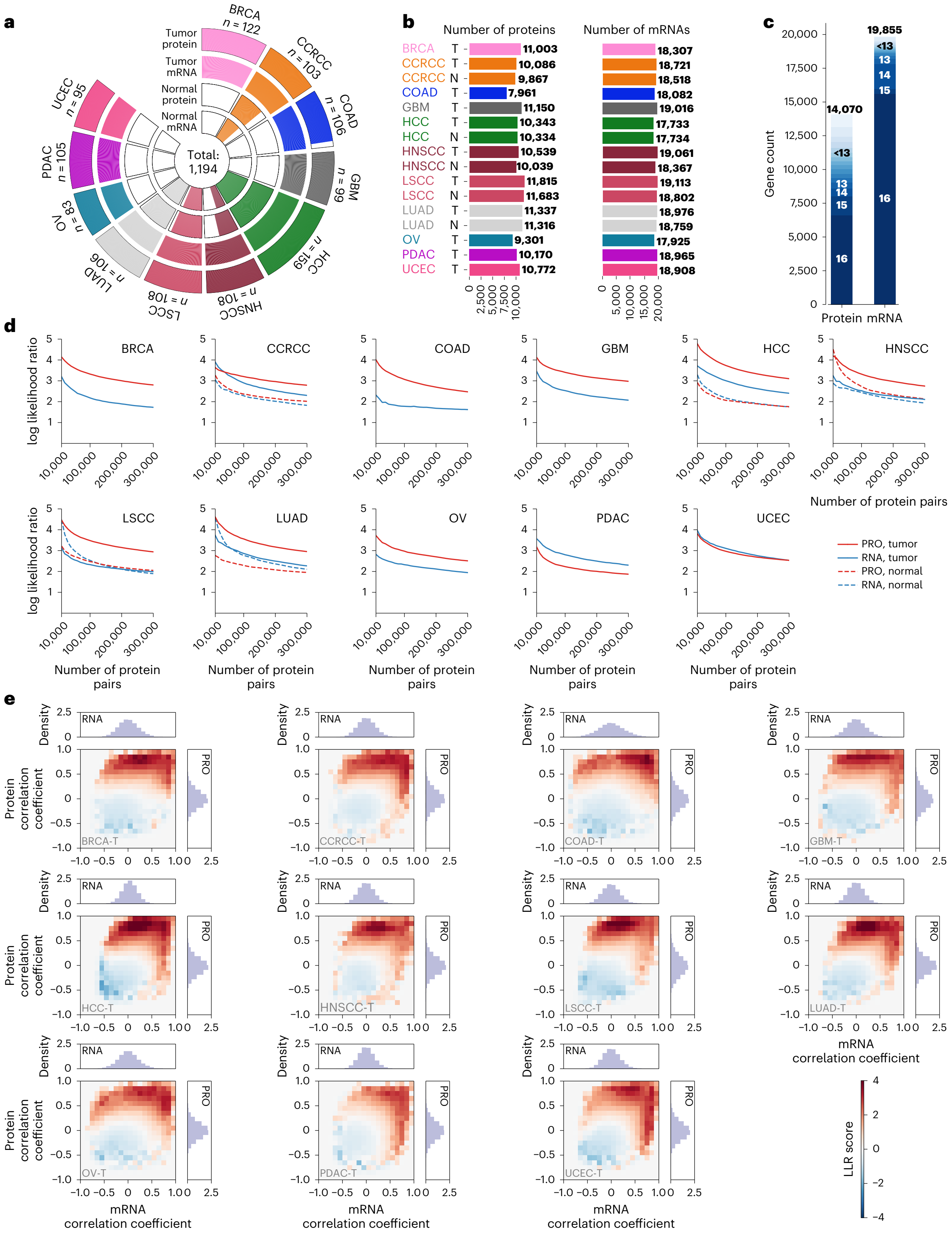


Fig. 1 | Protein coexpression is a strong predictor of gene cofunctionality. **a**, Proteomics and RNAseq data from tumor (T) and normal (N) samples across 11 cancer cohorts used in this study. The number of samples (*n*) is indicated in the plot. **b**, Numbers of quantified proteins and mRNAs in individual datasets. **c**, Numbers of proteins and mRNAs quantified across datasets. The numbers inside blue shaded boxes indicate the numbers of datasets with quantitative data. **d**, LLRs quantifying functional relevance of the top-ranking gene pairs based on the PCC from the top 10,000–300,000 in each dataset. **e**, Distributions of LLRs of the gene pairs with a given mRNA coexpression (*x* axis) and protein coexpression (*y* axis) pattern in the 11 tumor datasets. The density plots on the top and right visualize the mRNA and protein coexpression distributions, respectively.

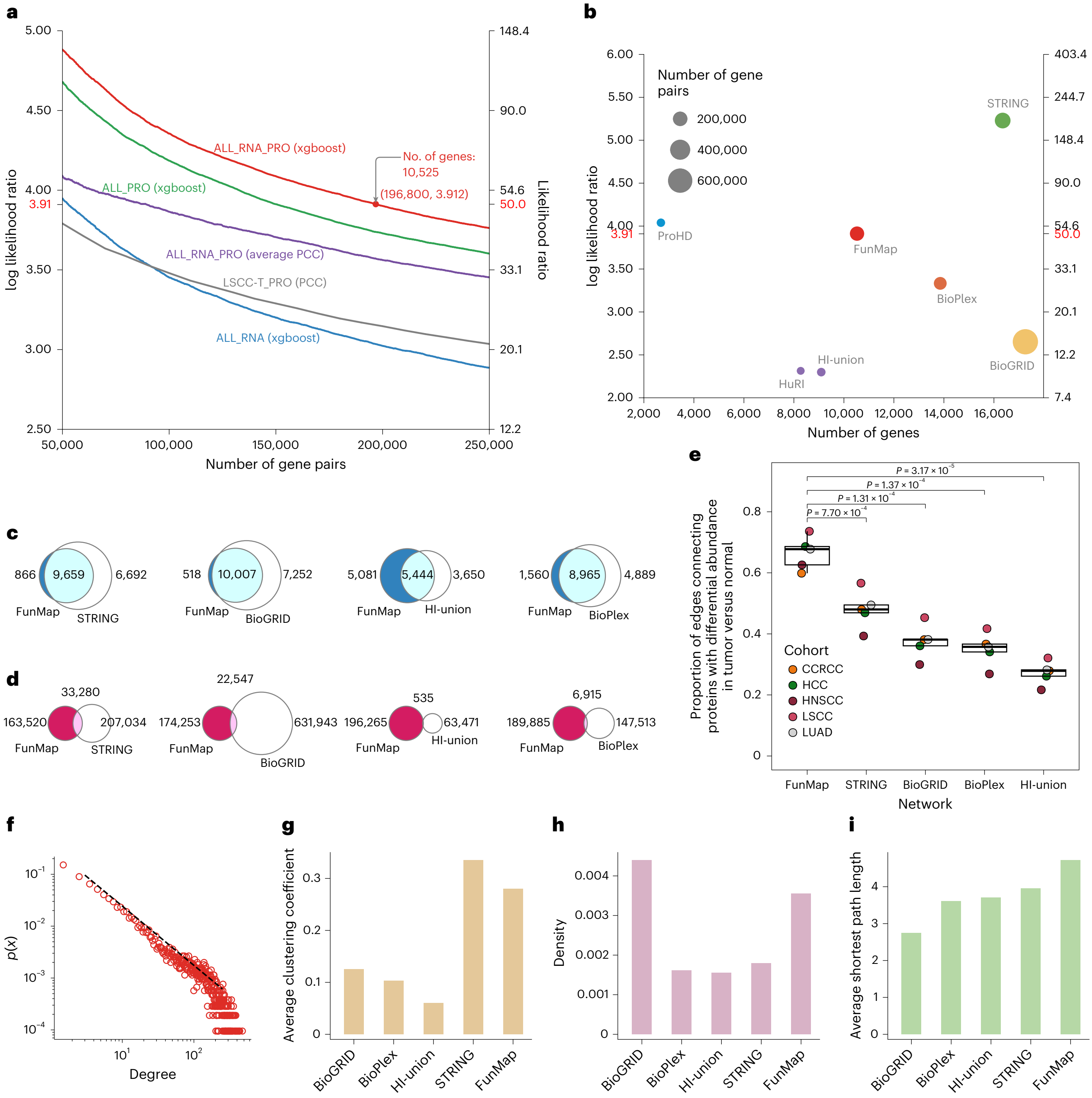


Fig. 2 | FunMap has high functional relevance, deep proteome coverage and a scale-free, modular and small-world network topology. a, A supervised machine learning model combining all 32 datasets (ALL_RNA_PRO (xgboost)) achieved higher LLRs across the whole range of top-scoring gene pair numbers from 50,000–250,000 compared with the models combining only proteomics datasets (ALL_PRO (xgboost)), only RNAseq datasets (ALL-RNA (xgboost)), the average PCC across the 32 datasets (ALL_RNA_PRO (average PCC)) or the PCCs from the LSCC tumor proteomics data alone (LSCC-T_Pro (PCC)). Applying an LLR cutoff of 3.912 (LR = 50) to results from the all-inclusive model produced a network with 10,525 genes and 196,800 edges, which was named FunMap. **b**, Scatter plot comparing functional relevance (*y* axis) and proteome coverage (*x* axis) of FunMap and other networks. The red horizontal lines in **a** and **b** indicate the LLR cutoff applied for FunMap, while the gray vertical line in **a** represents the number of gene pairs at the selected LLR cutoff. **c**, Gene overlap between FunMap and other networks. **d**, Edge overlap between FunMap and other networks. **e**, Box plots depicting proportion of edges connecting proteins with consistent significant overexpression or underexpression in tumors versus normal samples (*n* = 5 cohorts) for FunMap and other networks. For box plots, the center line indicates the median, box limits indicate the upper and lower quartiles and whiskers indicate 1.5× the interquartile range. *P* values were derived from a paired *t*-test followed by adjustment based on Holm's method. **f**, Degree distribution of FunMap. *p*(*x*) is the probability of nodes having a specific degree *x*. **g**–**i**, Plots comparing the average clustering coefficient (**g**), density (**h**) and average shortest path length (**i**) of FunMap and other networks.

those found in the other networks ($P < 0.001$, paired t-test; Fig. 2e), suggesting a stronger connection of FunMap to cancer.

FunMap showed a power-law degree distribution (Fig. 2f), indicating a scale-free topology with highly connected hubs. Compared to other networks, FunMap was characterized by a relatively higher average clustering coefficient (similar to STRING), relatively higher density (similar to BioGRID) and the highest average shortest path length (Fig. 2g–i). Together, these results suggest the high functional relevance, cancer relevance and proteome coverage of FunMap, as well as its scale-free, modular and small-world properties.

Cancer-associated dense modules

A high clustering coefficient of FunMap suggests that genes in the network tend to form clusters or modules. To assess FunMap's ability to connect genes encoding proteins in the same functional module, we used the CORUM database[22], which contains 5,204 manually annotated mammalian protein complexes involving 5,299 genes. Among the 196,800 edges in FunMap, 14,401 (7.3%) connected genes encoding proteins in the same CORUM complex (Fig. 3a). Strikingly, both the absolute count and the percentage of the edges overlapping with CORUM in FunMap were higher than those in the BioPlex network (6,747, 4.4%; Fig. 3a). As BioPlex was designed to experimentally identify protein complexes through affinity purification combined with MS, these results underscore FunMap's potential in unveiling tightly coregulated functional modules.

Some CORUM complexes associated with cancer-related processes displayed robust connectivity among their members in FunMap but not in BioPlex, such as complexes involved in cell cycle and DNA replication, gene expression and regulation, signal transduction, cell motility and innate immunity (Fig. 3b). Unlike BioPlex, which used data from only two in vitro cell lines, FunMap used data from over 1,000 human tumor samples, making it potentially more effective in uncovering functional modules relevant to in vivo cancer biology.

To extend our analysis beyond CORUM complexes, we applied the iterative clique enumeration (ICE) algorithm[23] to FunMap (Methods). This algorithm identifies relatively independent cliques, which are fully connected subnetworks (dense modules) with limited overlap to each other. Through this approach, we identified 281 dense modules, each with five or more genes (Supplementary Table 3). Of these, 130 (46%) overlapped significantly with CORUM complexes, an additional 37 (13%) overlapped with BioPlex complexes and another 49 (17%) overlapped with Gene Ontology (GO) annotations (false discovery rate (FDR) < 0.05, Fisher's exact test followed by Benjamini–Hochberg adjustment; Fig. 3c and Supplementary Table 3). These results emphasize the functional coherence of genes within these de novo identified dense modules.

To evaluate the cancer relevance of these dense modules, we compared the average standardized protein abundance between tumor and normal samples for each of the five cancer types (Supplementary Table 3). Of the 276 modules with sufficient data for statistical analysis, 273 showed significantly different abundance in tumors compared with normal samples in at least one cancer type (adjusted $P < 0.01$, Wilcoxon rank-sum test followed by Benjamini–Hochberg adjustment). Notably, 43 of the 273 (16%) had no significant overlap with CORUM, BioPlex or GO annotations (adjusted $P > 0.01$, hypergeometric test) and 203 (74%) had more than half of their edges unique to FunMap compared to other networks (Supplementary Table 3). These observations underscore the value of FunMap in uncovering previously unrecognized, cancer-relevant dense modules.

A total of 78 dense modules showed significant differential expression across all five cancer types, with 36 (46%) having less than 25% edge overlap with the other networks (Extended Data Fig. 3a). Many overexpressed modules were enriched in processes related to replication and proliferation. Moreover, three highly overexpressed modules (cliques 160, 96 and 54) were associated with extracellular matrix (ECM) organization (Fig. 3d,e and Extended Data Fig. 3b–e) and higher module levels were significantly associated with or trending toward worse overall survival (OS) in various cancer types (Fig. 3f, Extended Data Fig. 3f,g and Supplementary Table 3). Fewer modules were underexpressed and those related to cell adhesion (cliques 46 and 17; Fig. 3g,h and Extended Data Fig. 3a) may contribute to increased cell motility and tumor aggressiveness. This was supported by tumors with underexpression of clique 46 showing worse OS in HCC (Fig. 3i).

In summary, these results demonstrate the ability of FunMap to identify functionally and clinically relevant dense modules. Importantly, many of these modules were associated with cancers of diverse histological origin but had limited overlap with other networks, highlighting a unique connection of FunMap to cancer biology and disease progression.

Hierarchical modular organization linked to cancer hallmarks

The coexistence of scale-free topology (Fig. 2f) and a high clustering coefficient (Fig. 2g) in FunMap indicates a hierarchical modular organization, where genes form smaller modules that combine into larger ones across multiple scales[24]. Using the network seriation and modularization (NetSAM) algorithm[25], a specialized computational tool for uncovering the hierarchical organization in biological networks, we identified eight hierarchical levels and 255 modules with at least 20 genes in FunMap (Fig. 4 and Supplementary Table 4). Of these, 243 (95%) significantly overlapped with at least one GO annotation (FDR < 0.05, Fisher's exact test followed by Benjamini–Hochberg adjustment; Supplementary Table 4), indicating their functional coherence.

We focused on the enriched GO annotations that have been previously linked to cancer hallmarks[26,27] (Supplementary Table 4). The top ten largest branches were associated with various hallmarks (Fig. 4 and Methods), including tumor microenvironment-related hallmarks such as avoiding immune destruction and tumor-promoting inflammation, with the largest branch linked to tumor-promoting inflammation (1,118 genes). These findings underscore the strength of using tumor-derived data in network construction, which can capture complex, biologically important information that may be missed in cell-line-based protein–protein interaction networks.

To assess the clinical importance of these modules, we calculated meta P values for differential expression between tumors and normal samples across the five cancer cohorts (Supplemental Table 4). Tumor-overexpressed branches were linked to hallmarks such as enabling replicative immortality, genome instability and mutation, sustaining proliferative signaling, evading growth suppressors, avoiding immune destruction, resisting cell death and activating invasion and metastasis (Fig. 4). A detailed examination of these branches revealed their hierarchical functional organization. For example, the level 3 module L3_M55, associated with 'protein folding' and 'protein transport', was divided into two level 4 modules: L4_M58 (protein folding) and L4_M59 (protein transport) (Fig. 5a). The latter was further split into level 5 modules for 'protein targeting to the endoplasmic reticulum (ER)' (L5_M51) and 'ER to Golgi vesicle-mediated transport' (L5_M50). In tumor cells, ongoing replication, growth and genetic aberrations disrupt protein homeostasis[28], increasing the need for protein folding and related protein transport to resist cell death and avoid immune destruction, two hallmarks linked to this branch. Overexpression of the protein folding module (L4_M58) was associated with worse OS in CCRCC (Fig. 5b), with similar trends in HNSCC, LUSCC and LUAD (Supplementary Table 4), supporting its protumor role.

Tumor-underexpressed branches were linked to cancer hallmarks including deregulating cellular energetics, tumor-promoting inflammation, inducing angiogenesis and activating invasion and metastasis (Fig. 4). Although the association with tumor-promoting hallmarks

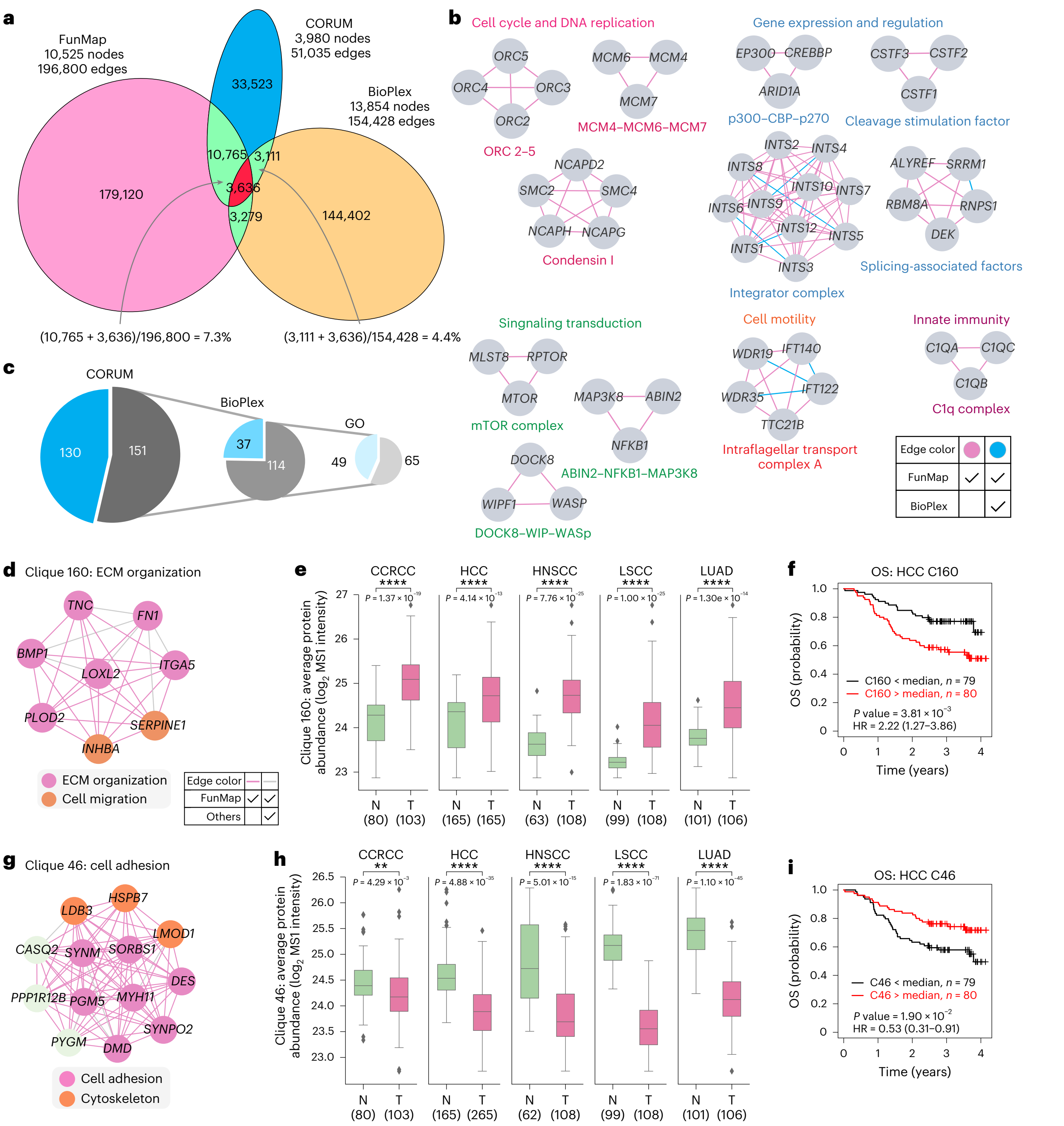


Fig. 3 | FunMap reveals known and previously unidentified dense modules associated with cancer biology and clinical phenotype. a, Overlap among gene pairs in FunMap, BioPlex and gene pairs encoding proteins in the same CORUM complex. **b**, Examples of CORUM complexes displaying robust connectivity among their complex members in FunMap but not in BioPlex. **c**, Numbers of de novo predicted FunMap dense modules with a significant overlap with CORUM complex, BioPlex complex or GO term ($P < 0.05$, Fisher's exact test, blue shaded sections). **d**, A tumor-overexpressed, ECM-associated dense module (clique 160). Edge color indicates a lack of overlap in BioGRID, BioPlex, HI-union, STRING and CORUM (pink) or overlap in any of these resources (gray). **e**, Box plots comparing the average protein abundance of clique 160 in tumor and normal samples demonstrating tumor overexpression in five cancer cohorts. The number of samples (n) is indicated in parentheses. P values were determined using a two-sided Wilcoxon rank-sum test. **f**, Kaplan–Meier plots depicting OS difference in persons with CCRCC, HCC and LUAD stratified by the median value of the average abundance of proteins in clique 160. The number of samples (n) is indicated on each plot. Log-rank P values and hazard ratios (HRs), shown with 95% confidence intervals, were derived from Cox proportional hazard models. **g**, A tumor-underexpressed, cell-adhesion-associated dense module (clique 46). The edge color is as described in **d**. **h**, Box plots comparing the average protein abundance of clique 46 in tumor and normal samples demonstrating tumor underexpression in five cancer cohorts. The number of samples (n) is indicated in parentheses. P values were determined using a two-sided Wilcoxon rank-sum test. **i**, Kaplan–Meier plots depicting OS difference in persons with HCC stratified by the median value of the average abundance of proteins in clique 46. The number of samples (n) is indicated in the plot. P values and HRs were obtained as described in **f**. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ and $^{****}P < 0.0001$. For box plots, the center line indicates the median, box limits indicate the upper and lower quartiles and whiskers indicate 1.5× the interquartile range; the number of samples per group is indicated in parentheses.

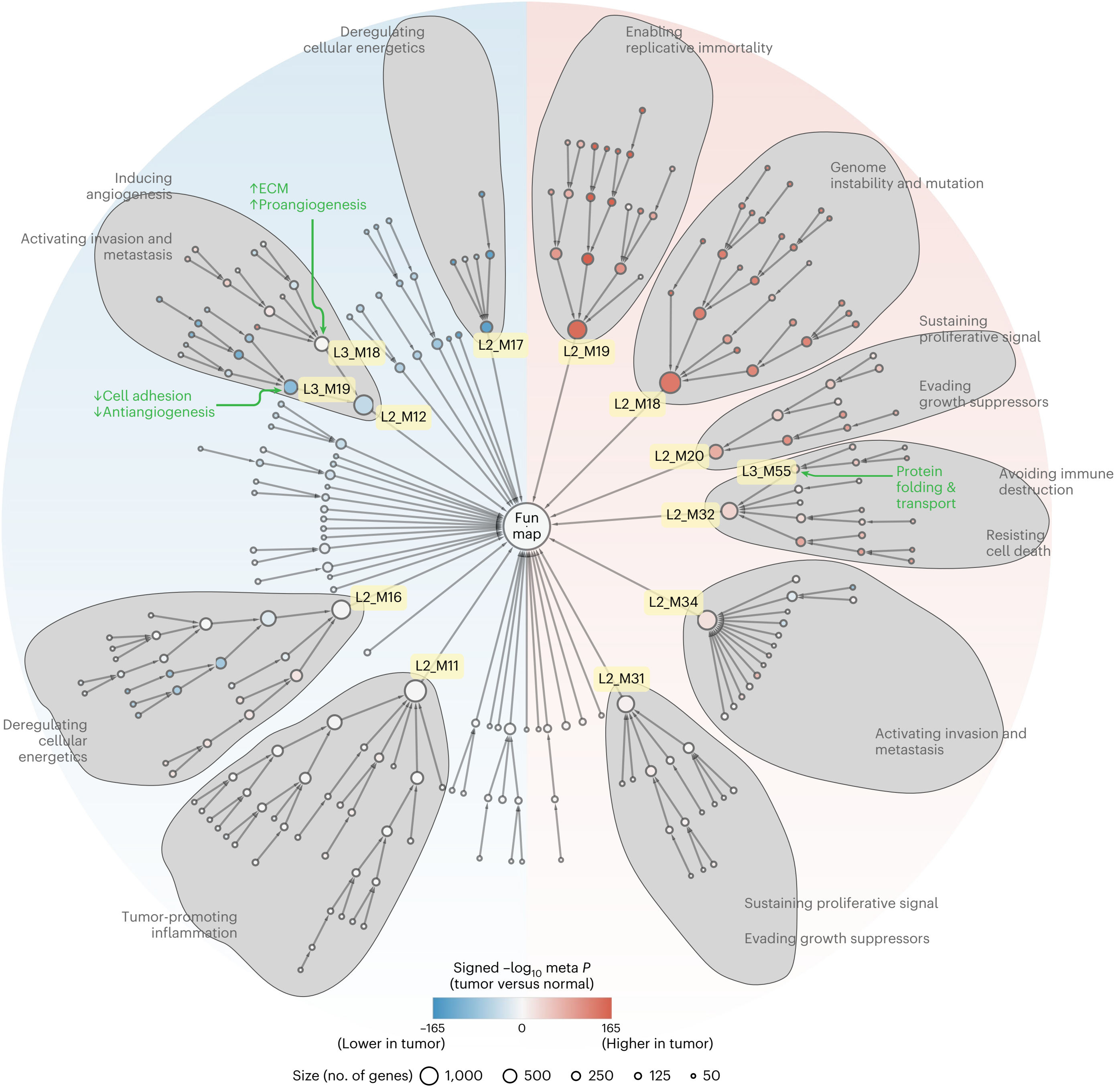

Deregulating cellular energetics
Enabling replicative immortality
Genome instability and mutation
Inducing angiogenesis
Activating invasion and metastasis
↑ECM
↑Proangiogenesis
↓Cell adhesion
↓Antiangiogenesis
Sustaining proliferative signal
Evading growth suppressors
Avoiding immune destruction
Resisting cell death
Protein folding & transport
L3_M18
L3_M19
L2_M12
L2_M17
L2_M19
L2_M18
L2_M20
L3_M55
L2_M32
L2_M34
L2_M16
L2_M11
L2_M31
Fun map
Deregulating cellular energetics
Tumor-promoting inflammation
Activating invasion and metastasis
Sustaining proliferative signal
Evading growth suppressors
Signed −log10 meta P (tumor versus normal)
−165
0
165
(Lower in tumor)
(Higher in tumor)
Size (no. of genes)
1,000
500
250
125
50

Fig. 4 | Hierarchical modular organization of FunMap statistically linked to cancer hallmarks. Hierarchical modular organization of FunMap. Nodes represent NetSAM-derived modules with node size proportional to module size. Nodes are colored on the basis of the significance of pan-cancer tumor versus normal protein abundance difference and ordered according to the significance levels of level 2 modules for each branch. The top-enriched cancer hallmark annotations for the ten largest level 2 branches are annotated in gray text. Green text indicates biological processes highlighted in Fig. 5.

initially seemed counterintuitive, further examination provided deeper insight. For example, the branch rooted in L2_M12, associated with inducing angiogenesis and activating invasion and metastasis, was enriched in functional categories including ECM structure, cell adhesion and angiogenesis, with modules deeper within the branch showing more specialized roles (Fig. 5c). While L2_M12 was overall underexpressed, it was divided into an underexpressed module (L3_M19) tied to antitumor functions such as cell adhesion and an overexpressed module (L3_M18) linked to protumor functions such as angiogenesis. Both overexpressed and underexpressed modules were enriched with ECM components but antiangiogenic ECM components were enriched in underexpressed modules, while proangiogenic ECM components were enriched in overexpressed modules (Supplementary Table 4). Interestingly, underexpressed dense modules related to cell adhesion (cliques 17 and 46) were entirely covered by L3_M19, whereas the overexpressed dense modules related to ECM (cliques 54, 96 and 160) were found entirely within L3_M18. Consistent with the good-prognosis association observed for clique 46 (Fig. 3i), higher expression of L3_M19 was correlated with a longer OS in HCC (Fig. 5d), with a similar trend observed for LUAD and CCRCC (Supplementary Table 4). In contrast, higher expression of the tumor-overexpressed module L4_M13, which was under L3_M18 and included most components from

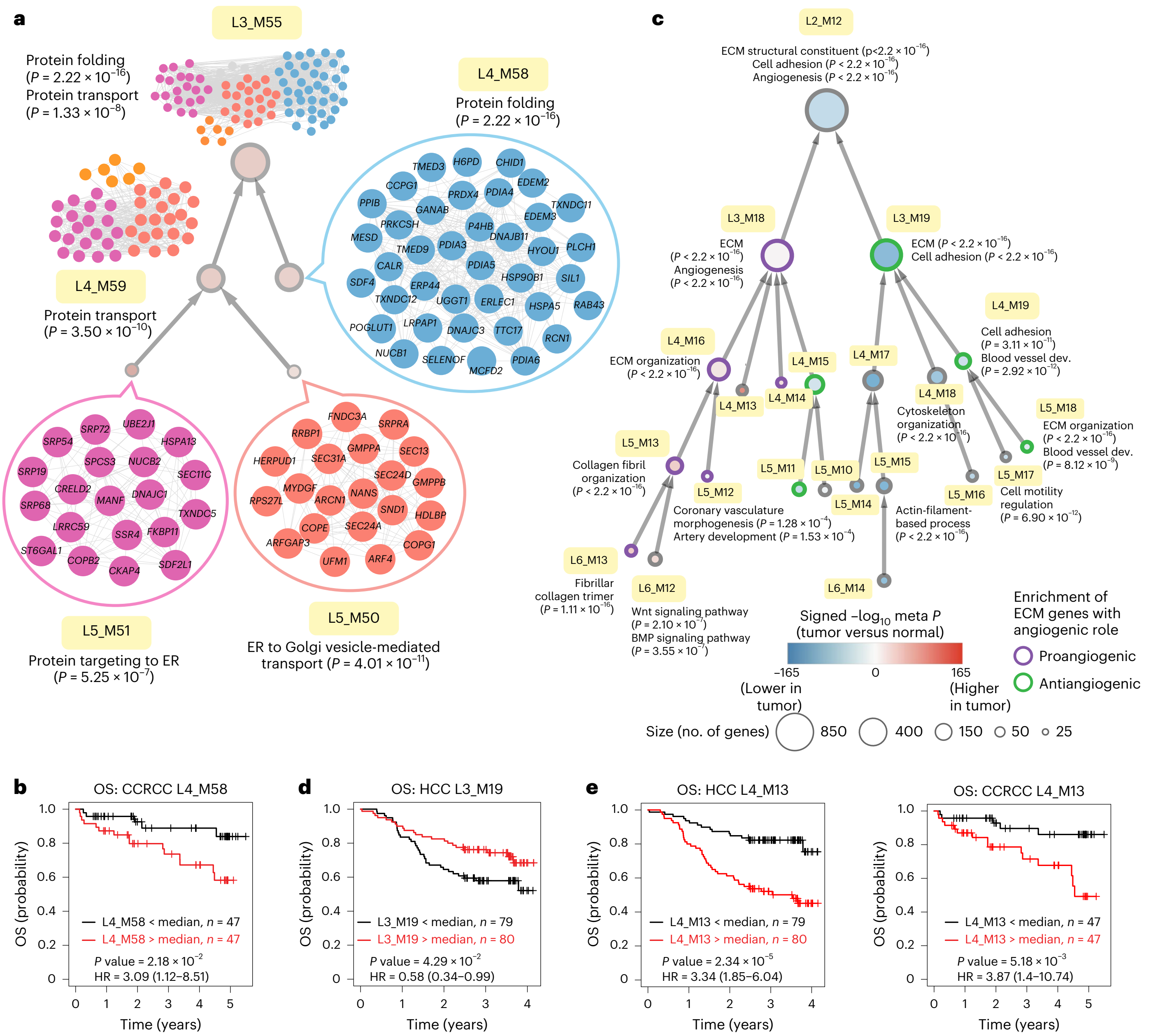


Fig. 5 | In-depth analysis of selected FunMap branches and their clinical associations. a, Hierarchical organization of five modules related to protein folding and protein transport. The node color and size of the modules are the same as in Fig. 4. *P* values were determined using a hypergeometric test. **b**, Kaplan–Meier plots depicting OS difference in persons with CCRCC stratified by the median value of the average abundance of proteins in module L4_M58. The number of samples (*n*) is indicated in the plot. Log-rank *P* values and HRs, shown with 95% confidence intervals, were derived from Cox proportional hazard models. **c**, Hierarchical organization of modules in an angiogenesis and metastasis associated branch. The node color and size are the same as in **a**. The node outline indicates the enrichment of ECM genes with proangiogenic versus antiangiogenic roles. *P* values were determined using a hypergeometric test. Blood vessel dev., blood vessel development. **d**,**e**, Kaplan–Meier plots depicting OS difference in persons with HCC stratified by the median value of the average abundance of proteins in module L3_M19 (**d**) or in persons with HCC and HNSCC stratified by L4_M13 abundance (**e**). The number of samples (*n*) is indicated in the plots. *P* values were derived as described in **b**.

the poor-prognosis cliques 54, 96 and 160 (Fig. 3f and Extended Data Fig. 3f,g), was correlated with a shorter OS in HCC (Fig. 5e) and other cancer types (Supplementary Table 4). Thus, the hierarchical module analysis not only reinforced the clique-based analysis results but also revealed the broader functional context and systematic organization of the dense modules.

In summary, network analysis revealed a hierarchical modular organization of FunMap, in which the major branches were statistically aligned to cancer hallmarks, supported by both functional analysis and the examination of clinical outcomes.

Connecting somatic mutations to protein modules

A major goal of cancer proteogenomics is to understand how somatic mutations impact the cancer proteome. Previous studies used univariate analysis to examine the *cis* and *trans* effects of individual mutations[29,30]. Here, we used a machine learning approach to simultaneously model the impact of all significant mutations on individual functional modules in FunMap to better capture the complexity of biological systems (Methods).

We identified 77 genes that were significantly mutated (q value < 0.1) in at least one of the ten CPTAC cancer types. For each of the 536 modules

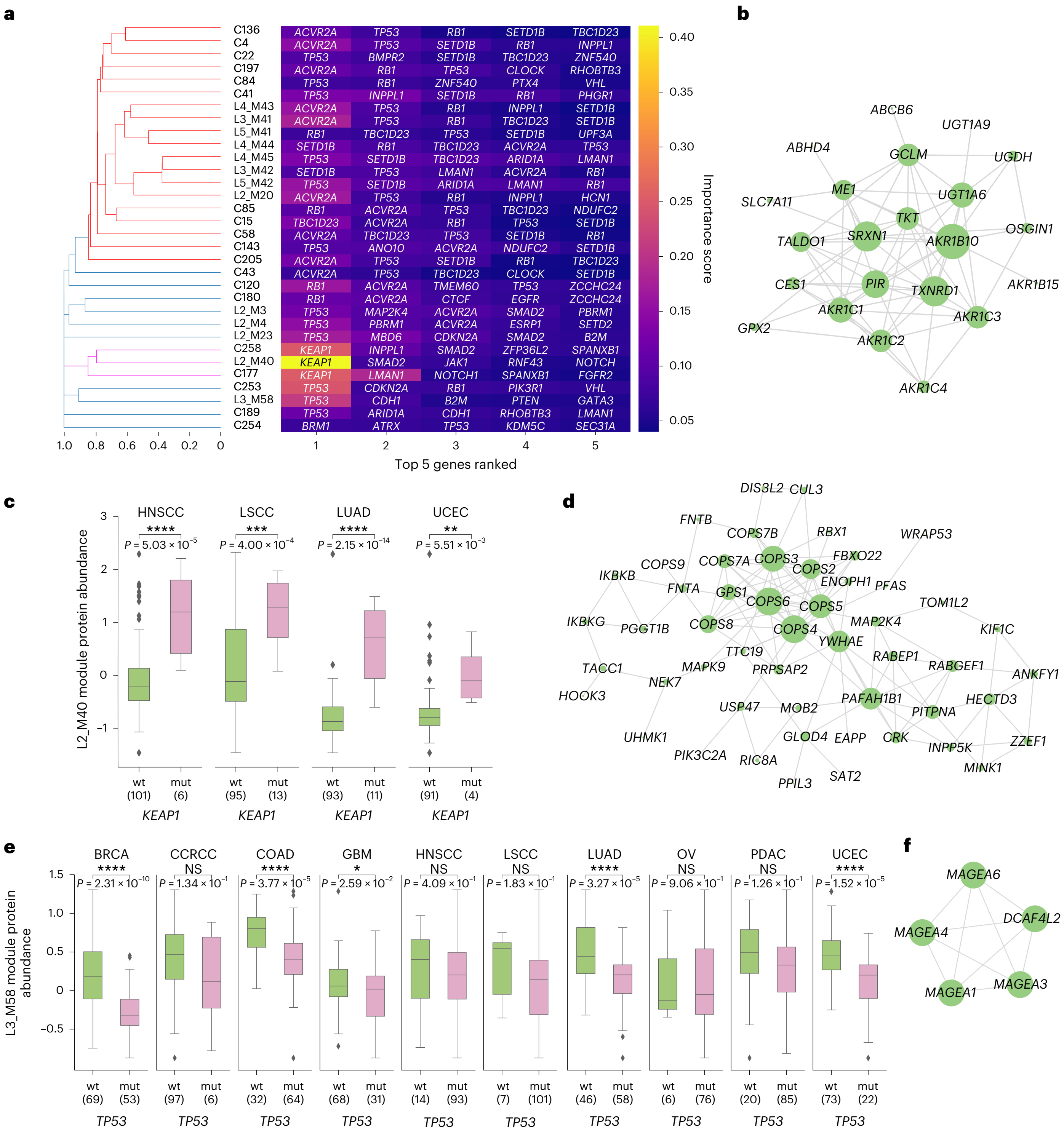


Fig. 6 | Connecting somatic mutations to functional protein modules. **a**, Heat map depicting the most important mutant genes in predicting the protein abundance of 32 modules. The modules were clustered on the basis of membership similarity. The heat map color corresponds to the relative importance in the XGBoost model. **b**, Associations defining module L2_M40. The node size corresponds to the node degree. **c**, Box plot comparing L2_M40 protein abundance in samples with and without *KEAP1* mutations in selected cancer cohorts. The number of samples (n) is indicated in parentheses. P values were derived from a two-sided Wilcoxon rank-sum test. **d**, Associations defining module L3_M58. The node size corresponds to the node degree. **e**, Box plot comparing L3_M58 protein abundance in samples with and without *TP53* mutations across cancer cohorts. The number of samples (n) is indicated in parentheses. P values were derived from a two-sided Wilcoxon rank-sum test. **f**, Clique 254, a CT antigen-associated dense module. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$ and ****$P < 0.0001$; NS, not significant. For box plots, the center line indicates the median, box limits indicate the upper and lower quartiles and whiskers indicate 1.5× the interquartile range; the number of samples per group is indicated in parentheses.

identified by NetSAM or ICE, we trained an XGBoost model to predict the average standardized protein abundance on the basis of the mutation status of the 77 genes. In a fivefold cross-validation based on data from 1,021 tumors across ten cancer types, 32 modules showed a nonrandom correlation (PCC > 0.2, $P < 0.00001$) between predicted and actual abundance, suggesting a significant connection between mutation status and protein abundance of these modules. Feature importance analysis highlighted *TP53* as a top predictor across all 32 modules, consistent with its role as a master regulator, while some other genes were specific to certain modules (Fig. 6a and Supplementary Table 5).

Hierarchical clustering of the 32 modules based on pairwise membership overlap revealed a predominant cluster with 19 modules (highlighted by red lines in the dendrogram in Fig. 6a). These modules comprised genes involved in the cell cycle or cellular division processes (Supplementary Tables 3 and 4). The most distinctive mutant genes defining this cluster included *RB1*, *ACVR2A*, *SETD1B* and *TBC1D23*. Mutations or deletions of *RB1* are common across various cancers and disrupt cell-cycle control, leading to uncontrolled cell proliferation[31]. While the roles of *ACVR2A*, *SETD1B* and *TBC1D23* are less extensively documented, mutations in these genes have been implicated in cell proliferation and tumorigenesis[32,33].

Another cluster of three modules were dominated by *KEAP1* mutations (highlighted by pink lines in the dendrogram in Fig. 6a), with L2_M40, a module comprising 22 genes (Fig. 6b), showing a particularly strong effect. L2_M40 exhibited increased protein abundance in tumors with *KEAP1* mutations across all cancer types that had a sufficient number of *KEAP1*-mutant samples for statistical comparison (Fig. 6c). Moreover, the expression of genes in this module showed the highest degree of coregulation at both mRNA and protein levels (average PCC > 0.5) in these four cancer types compared to the other cancer types (Extended Data Fig. 4a). Importantly, all genes in the module are known targets of nuclear factor erythroid 2-related factor 2 (NRF2)[34–41], which is activated by loss-of-function mutations in *KEAP1*, the gene encoding an inhibitor of NRF2. Therefore, this example serves as a strong positive control for our prediction.

Despite its broad importance, *TP53* mutations showed the strongest importance for modules C253 and L3_M58 (Fig. 6a). Module L3_M58, comprising 51 genes including highly interconnected constitutive photomorphogenesis 9 (COP9) signalosome subunits (Fig. 6d), showed decreased protein abundance in *TP53*-mutant tumors across nine of the ten cancer types, with a statistically significant decrease in five (Fig. 6e). Notably, gene expression in this module was more coregulated at the protein level than at the RNA level in most of the cancer types (Extended Data Fig. 4b). The COP9 signalosome is known to promote p53 degradation by targeting it for ubiquitination[42]. Our data suggest a negative feedback loop in which wild-type p53 activates the signalosome to suppress p53 levels and the process is disrupted by *TP53* mutations, leading to increased mutant p53 accumulation. This is consistent with the elevated p53 levels observed in *TP53*-mutant tumors (Extended Data Fig. 4c).

Some modules, such as C254, lacked a dominant predictor (Fig. 6a). This module, comprising four melanoma antigen gene family cancer/testis (CT) antigens and a testis-specific protein DCAF4L2 (ref. 43) (Fig. 6f), showed no significant associations with any top-ranked mutant genes in univariate analysis. However, several top predictors, such as *PBRM1*, *ATRX*, *TP53* and *KDM5C*, have been linked to immunosuppression and immunotherapy response[44–48], aligning with the role of C/T antigens in triggering immune responses.

In summary, our machine learning approach effectively connected somatic mutations with protein abundance across various functional modules. While some modules had clear dominant predictors and others did not, our models consistently identified key mutant genes whose functions aligned with the overarching function of the modules, demonstrating a clear functional basis for our predictions.

Illuminating understudied cancer proteins

Despite the massive disparity in our knowledge of individual genes (ranging from 9,282 publications in the Gene Reference Into Function (GeneRIF) database for *TP53* to zero publications for 700 'dark' genes), protein degrees in FunMap (that is, the number of edges) were comparable across the entire spectrum of knowledge depth (Fig. 7a), offering a great opportunity to illuminate understudied genes. Notably, while known cancer driver genes were concentrated among well-studied genes, proteins differentially expressed between tumor and normal samples, according to a meta-analysis of five cancer types, were evenly distributed across the proteome, including the 700 dark genes with no publications (Fig. 7a and Supplementary Table 6). Specifically, 125 of these dark genes were highly significantly overexpressed in tumors, whereas 92 were highly significantly underexpressed (meta P value $< 1.0 \times 10^{-16}$; Fig. 7b).

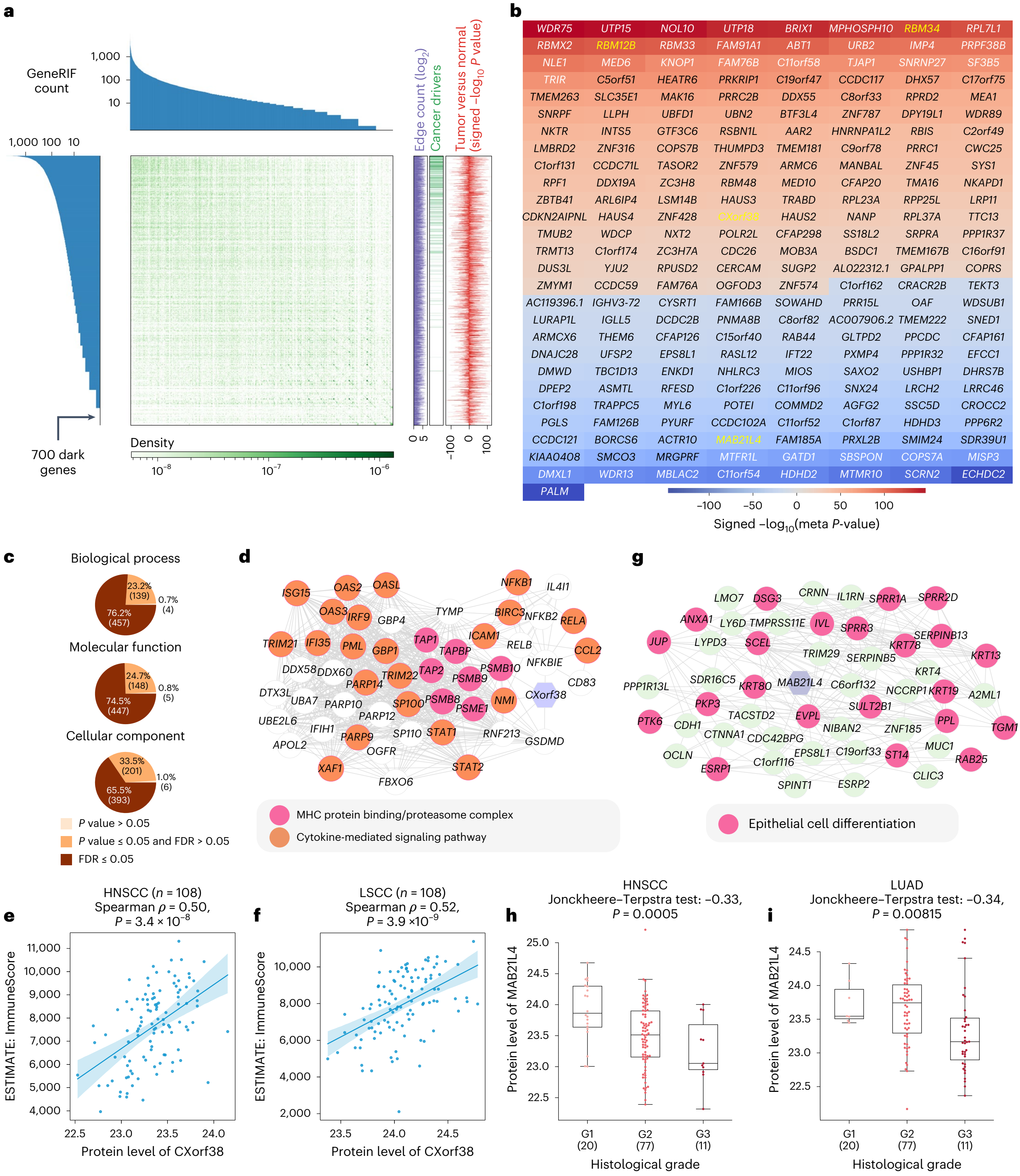
To gain functional insights into the 700 dark genes, we used the network topology analysis algorithm in WebGestalt[49] to establish a neighborhood of 50 genes for each dark gene and performed GO enrichment analysis (Methods). We found significant enrichment in biological processes for 76.2% of the genes, in molecular functions for 74.5% of the genes and in cellular components for 65.5% of the genes (FDR < 0.05, Fisher's exact test followed by Benjamini–Hochberg adjustment; Fig. 7c). This analysis connected 496 of the 700 dark genes, including the 200 shown in Fig. 7b, to at least one GO annotation. Although these genes lack publication records in GeneRIF, 315 have existing GO annotations. Of these, 183 (58%) had their top ten predicted GO terms overlap with one or more existing annotations. This high overlap, compared to just 0.63 from random gene sets, represents a 290-fold increase, underscoring the effectiveness of our approach in predicting gene function.

The dark genes *RBM34* and *RBM12B* were among the most significantly overexpressed genes in tumors (meta $P < 1.0 \times 10^{-100}$; Fig. 7b and Extended Data Fig. 5a), consistent with their frequent somatic amplification across various cancers (Extended Data Fig. 5b). Both genes encode RNA-binding motif (RBM) proteins, although their functions

Fig. 7 | FunMap predicts functions of understudied proteins. a, Heat map of the adjacency matrix of FunMap with genes sorted on the basis of GeneRIF counts. Genes with a GeneRIF count of 0 are defined as dark genes. The edge count depicts the $\log_2$ count of the number of edges per gene. The cancer driver annotation indicates whether a gene is annotated as a cancer gene in the CGC database. Tumor versus normal annotation plots of the signed $-\log_{10}$ meta P value comparing protein abundance in tumor versus normal across cancer cohorts. A positive sign indicates higher abundance in tumor and a negative sign indicates lower abundance in tumor compared to normal. **b**, Heat map depicting the signed $-\log_{10}$ meta P values ($P < 1.0 \times 10^{-16}$) computed as described in **a**. The yellow text indicates the genes analyzed in subsequent panels. **c**, Proportions of the dark genes with significantly enriched GO terms in enrichment analysis of the network neighborhood. P values were derived from a hypergeometric test and FDR-adjusted P values were derived using the Benjiamini–Hochberg method. **d**, Network neighborhood of *CXorf38* with genes associated with the enriched GO terms highlighted. **e**,**f**, Relationship between protein abundance of *CXorf38* and RNAseq-inferred ESTIMATE ImmunoScore in HNSCC (**e**) and LSCC (**f**) tumors. The number of samples (n) is indicated in the plots. P values were derived from two-sided Spearman's rank correlation. The shaded area depicts the 95% confidence interval. **g**, Network neighborhood of *MAB21L4* with genes associated with the enriched GO term highlighted. **h**,**i**, Protein abundance ($\log_2$ MS1 intensity) of *MAB21L4* by histological tumor grade in HNSCC (**h**) and LUAD (**i**) tumors. The number of samples (n) is indicated in parentheses. P values were derived from a two-sided Jonckheere–Terpstra test. For box plots, the center line indicates the median, box limits indicate the upper and lower quartiles and whiskers indicate 1.5× the interquartile range; the number of samples per group is indicated in parentheses.

have not been experimentally characterized. The network neighborhood of *RBM34* was enriched for genes involved in ribosomal RNA processing (Extended Data Fig. 5c), whereas that of RBM12B was enriched for genes associated with RNA splicing (Extended Data Fig. 5d). This analysis connected their amplification and overexpression to distinct functional roles, supported by computational inference from the GO consortium on the basis of an orthogonal phylogenetic approach[50].

The dark gene *CXorf38* was significantly overexpressed in tumors compared to normal samples in four of the five cancer types (meta $P = 8.6 \times 10^{-31}$; Extended Data Fig. 6a). Its network neighborhood was enriched for genes associated with the cytokine-mediated signaling pathway, major histocompatibility complex protein binding and proteasome complex (Fig. 7d), suggesting an immune function. As supporting evidence, CXorf38 protein levels correlated significantly



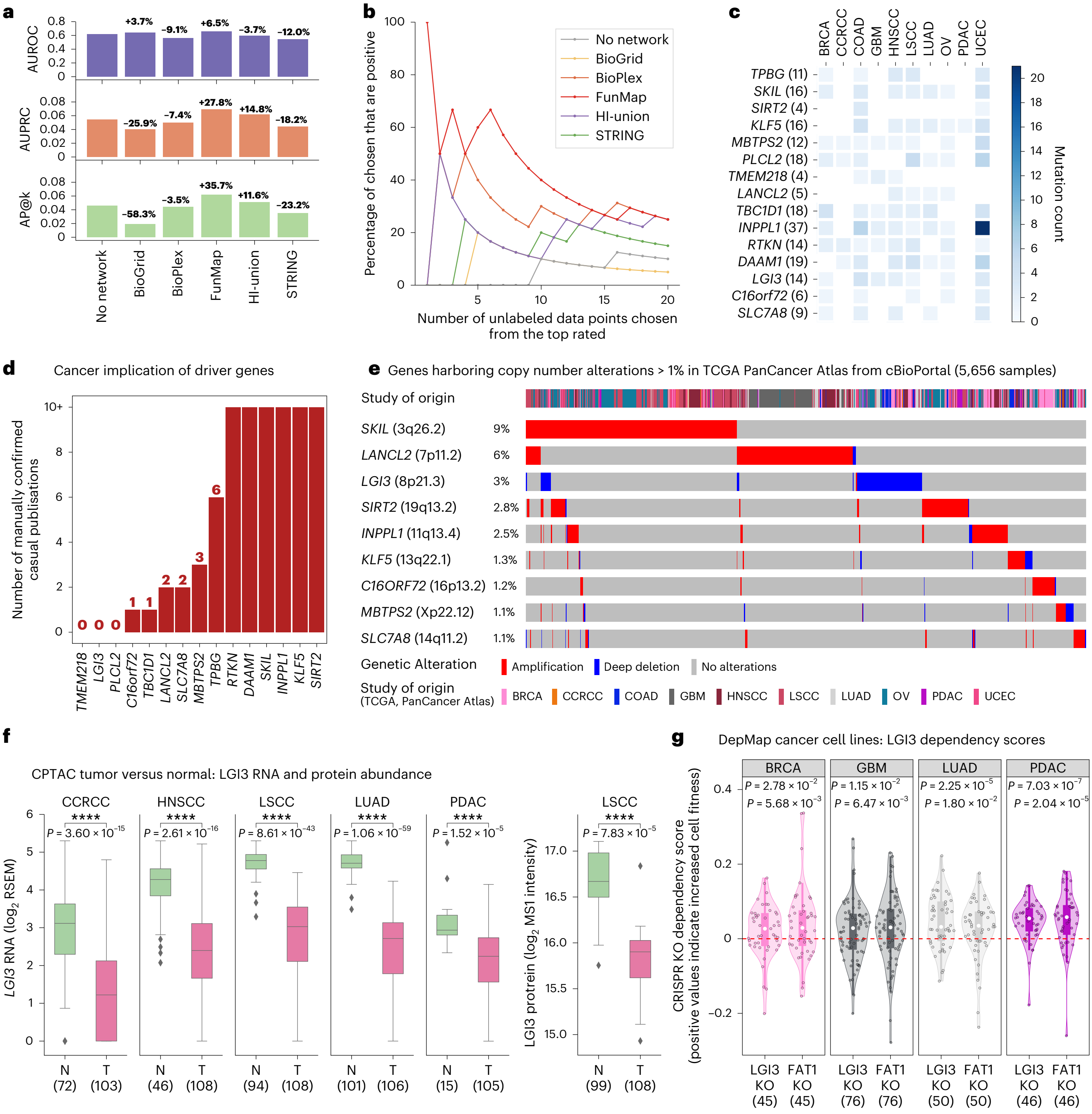


Fig. 8 | Discovery of cancer drivers with low mutation frequency using FunMap. a, Performance comparison between models trained with various networks and without network information, using AUROC, AUPRC or AP@k as evaluation metrics. **b**, Percentages of hidden positive genes among the top 20 predictions generated by various models. **c**, Mutation frequencies across various cancer types for the top 15 newly predicted cancer drivers by the FunMap-based model. **d**, Number of manually confirmed publications with direct experimental evidence implicating a causal role for a given predicted cancer driver. **e**, Oncoplot depicting copy number alterations in the newly predicted cancer drivers with >1% alteration frequencies in TCGA Pan-Cancer Atlas in cBioPortal. **f**, Box plots comparing *LGI3* RNA expression and LGI3 protein abundance in tumor versus normal samples demonstrating tumor underexpression in the cohorts shown. The number of samples (n) is indicated in parentheses. P values were derived from a two-sided Wilcoxon rank-sum test. **g**, Violin plots depicting dependency scores after *LGI3* or *FAT1* CRISPR KO in cell lines from annotated lineages downloaded from the DepMap resource. The number of samples (n) is indicated in parentheses. P values were derived from a one-sample, one-tailed t-test. For each cancer type, the first and second P values correspond to the significance of *LIG3* KO and *FAT1* KO, respectively. For box plots, the center line indicates the median, box limits indicate the upper and lower quartiles and whiskers indicate 1.5× the interquartile range; the number of samples per group is indicated in parentheses.

with the immune infiltration scores computed on the basis of RNAseq data in most CPTAC cancer types (Fig. 7e,f and Extended Data Fig. 6b). Moreover, single-cell data from the Human Protein Atlas show that CXorf38 is highly expressed in immune cells (Extended Data Fig. 6c), reinforcing its inferred immune role.

The dark gene *MAB21L4* was significantly underexpressed in tumors in three cancer types (meta $P = 9.9 \times 10^{-56}$) (Extended Data Fig. 6d). Its network neighborhood was enriched for genes associated with epithelial cell differentiation (Fig. 7g), the suppression of which has a critical role in tumorigenesis. Remarkably, MAB21L4 protein abundance was lower in poorly differentiated tumors (G3) compared to well differentiated (G1) and moderately differentiated (G2) tumors in both HNSCC and LUAD (Fig. 7h,i). These findings, consistent with a recent study showing that loss of MAB21L4 blocks differentiation to drive the development of squamous cell carcinoma[51], provide strong evidence to support a tumor suppressor role of MAB21L4.

Together, our systematic evaluation using existing GO annotation and the specific examples illustrate the utility of FunMap as a systematic framework to illuminate understudied genes, including many understudied cancer-associated proteins.

Discovery of drivers with low mutation frequency

Leveraging advancements in graph neural network (GNN)-based deep learning, we developed a positive–unlabeled (PU) learning algorithm that integrates the FunMap network, gene mutation significance scores from ten CPTAC cohorts and known cancer genes to train a graph attention network (GAT) model for classifying unlabeled genes as cancer or noncancer genes (Extended Data Fig. 7 and Methods).

For performance evaluation, we used 274 cancer genes from the original Cancer Gene Census (CGC)[52] as the positive set and 449 genes added later as hidden positives (Supplementary Table 7). The FunMap GAT model outperformed a random forest classifier trained without using network data, with a 6.5% improvement in area under the receiver operating characteristic (AUROC), 27.8% improvement in area under the precision–recall curve (AUPRC) and 35.7% improvement in the average precision at k (AP@k) (Methods). We also trained alternative GAT models using other networks including BioGrid[20], BioPlex[18], HI-union[19] and STRING[21]. The FunMap GAT model outperformed all alternative models for all three evaluation metrics (Fig. 8a).

Among the top FunMap GAT predictions, 60.0% of the top 5, 40% of the top 10 and 25% of the top 20 were hidden positives, far exceeding the expected 4.3% by random chance ($P < 0.01$, Fisher's exact test). In this analysis, models incorporating network data clearly outperformed those that did not (Fig. 8b), and there was minimal overlap among the top 20 predictions when different networks were used or when network data were not used (Supplementary Table 7). These results underscore the notable impact of network information on prediction outcomes.

Despite low mutation frequencies (Fig. 8c), 12 of the top 15 (80%) putative driver genes predicted by FunMap and not covered by CGC had at least one publication that supports a causal role in cancer through genetic and/or pharmacologic perturbation in model systems (Fig. 8d, Supplementary Table 7 and Methods). Moreover, nine genes showed frequent copy number alterations in TCGA data (Fig. 8e), providing independent support for our predictions because copy number data were not used in the FunMap GAT model. Notably, *LGI3*, although lacking causal evidence in the literature (Fig. 8d), was recurrently deleted in 3% of the 5,656 TCGA samples and significantly downregulated at both RNA and protein levels in tumors from CPTAC cancer cohorts where LGI3 was quantified in both tumor and normal samples (Fig. 8f). Furthermore, an analysis of clustered regularly interspaced short palindromic repeats (CRISPR) knockout (KO) dependency scores for cancer cell lines available through DepMap revealed a significant increase in cell fitness across various lineages following *LGI3* KO ($P < 0.05$, one-sample t-test) and the effect was on par with that observed for well-known tumor suppressor genes listed in the CGC such as *FAT1* (ref. 53) (Fig. 8g). These results collectively suggest *LGI3* as a putative tumor suppressor gene.

Taken together, our data highlight the effectiveness of FunMap in uncovering genes with a low mutation frequency as putative cancer genes, presenting them as promising candidates for further experimental validation.

Discussion

Large-scale omics profiling has massively expanded the landscape of somatic mutations and cancer-associated proteins but the difficulty in functional interpretation hinders their prioritization and translation into clinical practice. By using machine learning techniques on pan-cancer proteogenomics data, FunMap provides a systematic framework to tackle this challenge.

With 196,800 associations among 10,525 proteins and an LR of 50, FunMap provides both a comprehensive and unbiased proteomic coverage and a high level of functional relevance. The key differences between our approach and previous studies on gene coexpression networks include the use of protein profiling data obtained from over 1,000 human tumor samples spanning 11 cancer types and a supervised machine learning approach for functional network construction. Consistent with previous reports, protein coexpression is a much more reliable predictor of gene cofunctionality than mRNA coexpression[10,12]; however, combining both protein and mRNA coexpression provides the highest level of predictive power. One unexpected observation is that our coexpression-based functional network outperforms protein–protein interaction networks in discriminating between functionally relevant and irrelevant gene pairs. Thus, functional networks constructed from proteomic and proteogenomic data offer a complementary approach to protein–protein interaction networks, thereby expanding systems biology frameworks for functional genomics research. Indeed, analyses from our study clearly demonstrate the utilities of FunMap in providing a functional annotation of understudied cancer proteins, obtaining functional insights into somatic mutations and shedding global insights into cancer proteome organization and function.

A limitation of this study is that data from only 11 cancer types were included in the pan-cancer FunMap construction. We expect that proteomic and proteogenomic profiling will be applied to more cancer types in the future and a more comprehensive analysis can be performed as more cancer types are included in future studies. Moreover, the CPTAC cohorts used in the study have limited follow-up duration, with the incidence of death events varying substantially among different cancer types. Therefore, the statistical power to detect associations with survival is generally low and varies considerably across cohorts, which constrains the scope of our prognostic analysis. To mitigate this limitation, it would be beneficial to seek out cancer cohorts that have been followed for a longer period. For some cancer types such as breast cancer and lung cancer, there are already multiple independent proteomic and proteogenomic studies. In this scenario, our approach can also be used to integrate independent datasets from a single cancer type to build cancer-type-specific FunMaps. Additionally, this study focused on assessing the value of proteogenomic profiling data in mapping the functional network of human cancer but the approach can be easily expanded to integrate expression data with other types of data, such as protein–protein interaction data, to generate a more comprehensive functional network. Although FunMap GAT outperformed other models to some extent in distinguishing between driver and passenger mutations, the accuracy was far from satisfactory for all models, highlighting the difficulty of this persistent challenge. Further improvements may be made in both FunMap construction and network-based driver gene prediction. Lastly, the associations identified in our analysis represent pairs of genes that work in coordination within the complex tumor tissue system, which includes not only cancer cells but also the surrounding microenvironment. Because the data we used originated from bulk tissues, it is impossible to determine associations within specific cell types. The emerging single-cell proteomics technology would be ideal for addressing this limitation[54].

In conclusion, this study highlights the great potential of integrating machine learning and proteogenomic profiling to gain a deeper understanding of complex cancer systems. By generating a comprehensive functional network, this approach provides a robust framework for cancer functional genomics research, offering valuable insights into somatic mutations and cancer-associated proteins. These findings can greatly aid in prioritizing targets for clinical translation, ultimately contributing to the development of more effective cancer therapies.

Methods

Data acquisition

CPTAC data for ten cancer cohorts were harmonized by the CPTAC pan-cancer working group as previously described[15]. HCC data were downloaded from the original publication[55]. In total, we collected mRNA and proteomics data for 11 cancer cohorts, where five cohorts also included data for matched normal samples for both mRNA and protein. For each of the 32 mRNA or proteomics datasets, we required that each gene or protein had at least 20 valid data points to be included in the analysis. The union set of all valid genes was denoted as g_{valid}.

Network construction

A machine learning model using XGBoost[56] was trained to predict the probability of cofunctionality for a gene pair. For each gene pair (A, B), the PCC PCC_{AB} was computed between their mRNA expression vectors or protein expression vectors in each of the 32 datasets. We further calculated the MR of a gene pair in each dataset using a modified version of a previously published definition[17], $\text{MR}(A, B) = \frac{1}{n-1}\sqrt{r_{AB}r_{BA}}$, where r_{AB} is the rank of PCC_{AB} among all PCCs between gene A and its partners. The rank starts at 0 and a larger PCC results in higher ranks. The total number of genes is denoted as n. The MR values are in the range of [0,1]. In the case of PCC_{AB} missing in a cohort, we treat r_{AB} as a missing value. The 32 MRs for a gene pair were used as input features for training the XGBoost model.

To prepare the data for training and validating the XGBoost model, we downloaded a gold-standard set that was previously constructed using the Reactome pathway database[12]. In brief, functionally associated protein pairs (labeled as positive) are defined as pairs that are found in the same detailed pathway. Here, each protein is annotated to a subset of the lowest-level pathways. Only pathways that contain ≤200 proteins were included to make sure that only closely related protein pairs were positively labeled. Protein pairs that are not included in the same pathway at any annotation level are labeled negative. We included only those pairs where both proteins are in g_{valid} as the final dataset D for training the classification model. We partitioned the data into training (D_{train}) and test (D_{test}) sets, with a 50–50 split. The ratio of positive and negative labels was kept the same in the training and test sets using a stratified splitting technique. However, it is worth noting that the original dataset exhibited a substantial class imbalance issue, with a considerably larger number of instances in the negative class compared to the positive class. To tackle this challenge, we applied undersampling specifically to the negative class within the training dataset. This step involved reducing the number of negative class instances, aligning them with the number of positive class instances. The goal was to create a balanced training dataset that allowed the machine learning model to learn from both classes more equitably. We then performed hyperparameter tuning by applying grid search with fivefold cross-validation. The parameter grid was defined as follows: {'n_estimators': [50, 150, 250], 'max_features': [0.2, 0.4, 0.6, 0.8], 'min_samples_split': [2, 4, 6]}. We used AUROC as the performance metric for hyperparameter tuning. After the model was trained, we predicted the labels for all possible pairs of proteins in g_{valid}. We required that the MR of a pair must be larger than 0.95 (that is, top 5% among all gene pairs) in at least one data cohort. The final prediction performance was measured with LLR using the gold-standard subset D_{test}. Here, LLR is defined as

$$\text{LLR} = \ln\left(\frac{C(PP\&P)/C(PP\&N)}{C(P)/C(N)}\right)$$

where PP is the set of predicted positive protein pairs, while P and N are sets of positive and negative pairs in D_{test}, respectively. Set intersection is denoted as $\&$, while function $C(\cdot)$ returns the size of a set. To determine the number of pairs to be included in the final network, we first sorted the pairs in descending order of being positive (according to the predicted probability). We then computed the LLRs while designating more top pairs with a step size of 100 as PP. The LLR drops with the inclusion of less confidently predicted pairs. We stopped the process as soon as LLR dropped below 3.912 (LR = 50). All protein pairs selected with this procedure were included as edges in a functional association network named FunMap.

Detection of network modules

We used two complementary algorithms to identify modules from FunMap. First, we applied the ICE algorithm[23] to identify relatively independent maximal cliques in the network as functional modules. Overlap between the modules is allowed but restrained because of the inherent design of the algorithm. The stringent requirement imposed by the module definition in the algorithm ensures high-level of cofunctionality among all proteins in a module. The input to the software (http://ice.zhang-lab.org) is the network edge list file and the only required parameter is the minimal module size C. In this study, we set C to 5.

In contrast to the bottom-up approach taken in ICE, the top-down hierarchical modular organization of FunMap was uncovered using the NetSAM algorithm[25] implemented in R (https://bioconductor.org/packages/release/bioc/html/NetSAM.html). The main function of the package takes as input an network edge list file and outputs an 'nsm' file that describes all detected modules organized in a hierarchical fashion. The most important parameters to the function include 'minModule' and 'modularityThr'. The parameter 'minModule' specifies the ratio between the size of the smallest module and the total number of nodes in the network. If the size of a module identified by the function is less than the minimum size, the module is not further partitioned into submodules. We set 'minModule' such that the minimum size of a module was 20. To test whether a network under consideration had a nonrandom internal modular organization, we set the parameter 'modularityThr' to 0.2 such that the network would be considered to have internal organization and would be further partitioned when its modularity[57] was above this threshold value. This parameter reflects the stringency of splitting a module into submodules. A higher threshold value tends to split the modules less frequently.

Connecting hierarchical modules to cancer hallmarks

Overlap between FunMap's hierarchically organized modules and cancer hallmarks was evaluated according to 146 literature-curated GO terms[26,27,58–60]. These terms are categorized into ten themes that map to ten cancer hallmarks[61]. For each FunMap module, we performed overrepresentation analysis (ORA) and obtained the top ten enriched terms for that module. To annotate each branch of the tree structure rooted on a second-level module with the most relevant hallmark, we designed a voting scheme that works as follows: for each branch, we first designated the most overlapped hallmark as that with the largest sum of associated negative logarithm of P values for that hallmark over all modules in that branch. In essence, each module can vote for a representative hallmark for its residing branch using its level of overlap with that hallmark. The designated hallmark for each branch represents the consensus annotation for the whole branch. The top associated consensus annotation for the ten largest branches are shown in Fig. 4. For selected branches, a second consensus hallmark annotation was also shown that was both closely related to the top annotation and had a sufficiently significant *P* value.

Connecting hierarchical modules to ECM and angiogenesis

ECM genes encoding proteins with documented exclusive proangiogenic or antiangiogenic activity[62] along with collagen type VI were first used to calculate the proportion of proangiogenic and antiangiogenic genes within nodes downstream of the FunMap branch rooted in hierarchical module L2_M12. A final enrichment ratio for angiogenic impact was then computed by taking the previous proangiogenic ratio over the antiangiogenic ratio. Values > 1 indicate a higher proportion of proangiogenic ECM genes in a module while values < 1 indicate a higher proportion of antiangiogenic ECM genes. Some modules did not contain any antiangiogenic genes and were annotated as proangiogenic exclusive (Supplementary Table 4).

Connecting network modules to somatic mutations

We trained an XGBoost model to evaluate the importance of gene mutation in predicting network module abundance. A total of 536 modules were considered, including those revealed by NetSAM (255) and ICE (281). To compute module abundance, we first transformed the raw protein expression data in each cohort into *z* scores by performing feature-wise standardization. The module abundance of a sample is defined as the average *z* score of all genes in the module for that sample.

We used mutation data from ten CPTAC tumor cohorts in this part of the study because of the lack of mutation data from the HCC study. First, we selected genes that were significantly mutated in at least one cohort (*q* value < 0.1). We then retrieved the actual binary mutation data of the selected genes from each cohort and merged them into a final feature dataset. The resulting mutation dataset was composed of 1,021 samples and 77 genes.

For each module, we trained a regressor with XGBoost to predict module abundance based on the 77 significantly mutated genes. We applied fivefold cross-validation for hyperparameter tuning using the grid search technique. The parameter grid was defined as {'learning_rate': [0.1, 0.2, 0.3, 0.4, 0.5], 'n_estimators': [20, 50], 'max_depth': [2, 3, 4]}. We used the PCC between the predicted and actual abundance scores as the scoring metric for model assessment. Best parameters were used to fit a final model with the whole training data. We only included those modules that could be predicted with PCC > 0.25 in downstream analyses. This resulted in a total of 17 modules. The built-in feature importance scores of the trained model were used to estimate the contribution of each mutated gene in predicting the module abundance. Specifically, we used the 'gain' type importance, which implies the relative contribution of the corresponding feature to the model, calculated by taking each feature's contribution for each tree in the model. A higher value of this metric when compared to another feature implies that it is more important for generating a prediction. This allows features to be ranked and compared with each other.

Function prediction of understudied genes

On the basis of the assumption that genes with similar functions are located in proximity to each other in the functional association network, we made function prediction of the dark genes in FunMap. We used the network topology analysis algorithm in WebGestalt[49] to establish a neighborhood of 50 genes for each dark gene and then performed GO enrichment analysis. Specifically, the algorithm lets the random walker start from each dark gene. It repeatedly moves to its neighboring nodes with an equal likelihood. At each step, it also has some probability ($P = 0.5$) of returning to the starting point. The restart probability controls how far the random walker moves away from the dark gene. The final score of a gene is defined as the steady-state probability that the walker will stay at the gene in the long run. For each dark gene, we chose the top 50 genes with the highest scores as its network neighbors and then performed ORA against GO terms for these network neighbors.

Cancer driver gene prediction

To predict cancer driver genes, we trained GAT-based[63] neural network models on FunMap and compared the performance with models trained with other publicly available networks, including BioPlex[18], HI-union[19], BioGrid[20] and STRING[21]. For the STRING network, we only kept interactions with a combined score higher than 700. As a baseline, we also trained a random forest classifier without using network data.

We used mutation data from the ten CPTAC tumor cohorts in this part of the study. First, we selected genes that were significantly mutated in at least one cohort (*q* value < 0.1). We then performed $-\log_{10}$ transformation to the raw *P* values. Each gene was characterized by a ten-dimensional vector as its features, representing mutation significance in ten cancer cohorts.

Given the uncertainty regarding the role of an unlabeled gene as a driver or nondriver gene, the standard supervised machine learning approach is not well suited for our task. This is because of the fact that typical supervised learning algorithms necessitate the presence of both positive and negative examples for training purposes. Therefore, we formulated our prediction task as a PU learning problem[64] where genes in the network are divided into positive genes (known drivers) and unlabeled genes, which can contain both hidden driver genes (to be predicted positives) and nondriver genes (negatives). The goal is to train a model that uses known drivers to identify hidden drivers in the network. For known drivers, we downloaded a list of cancer drivers from the original CGC publication[52], which included 274 genes. To test our trained model, we downloaded the 449 driver genes that were included in the CGC database after the original publication (Supplementary Table 7). Only known and hidden driver genes presented in the respective networks were used in training and performance evaluation.

We used the bagging based PU learning approach[65] to tackle the driver gene prediction task. The approach can be broken down into four steps: (1) create a training set by combining all positive data points with a random bootstrapped sample set B of the same size from the unlabeled samples; (2) train a classifier with the newly assembled sample set, treating positive and unlabeled data points as positives and negatives, respectively; (3) apply the classifier to those unlabeled samples that were not included in B, the out-of-bag (OOB) sample set, and record their predicted scores; and (4) repeat the previous three steps T times ($T = 10$ in this study) and assign to each sample the average of the OOB scores it has received.

To train a node classifier in step 2 using GNN, we used the GAT architecture. The learning of a GAT attention layer involves four key steps. First, to obtain sufficient expressive power, a linear transformation is applied to the feature vectors of the nodes. Second, attention coefficients determining the relative importance of neighboring features to each other are computed. To obtain the attention score between two neighbors, it first concatenates the embeddings z of the two nodes obtained from the previous step, and then takes a dot product of it with a learnable weight vector a and finally applies a leaky rectified linear unit (LeakyReLU). This step can be formulated as

$$e_{ij} = \text{LeakyReLU}(a^T(z_i \| z_j))$$

where || denotes concatenation. Third, to make the scores easily comparable, the attention coefficients are normalized across all neighborhoods using the softmax function. The fourth and final step works similarly to a graph convolutional network. The embeddings from neighbors are aggregated together, weighted by the attention coefficients and then transformed by a nonlinear activation function. Similar to multiple channels in a convolutional neural network, GAT uses multihead attention to enhance the model capacity and to stabilize the learning process. Specifically, K independent attention mechanisms apply the transformations of steps 1–3. During the last step, embeddings from different heads are averaged before applying the nonlinear transformation. In this study, we trained a model consisting of two GAT layers each with eight attention heads.

For performance evaluation, in addition to the standard metrics such as AUROC and AUPRC that treat all unlabeled samples as negative, we also included the more appropriate AP@k metric, which is widely used in the areas of information retrieval and recommendation systems. Essentially we treated our task as a ranking problem where we aimed to assign the test positive samples with higher scores (likelihood of being a driver gene) such that they ranked higher in the list of sorted prediction scores[66]. After the samples were sorted by their predicted scores, AP@k was computed as $\mathrm{AP@k} = \frac{1}{\min(m,k)} \sum_{i=1}^{k} \frac{\mathrm{TP}(i)}{i}$, where m is the total number of positive samples in the test dataset. $\mathrm{TP}(i)$ is set to 0 if the ith sample is not a positive test sample. Otherwise, it is set to the number of positive test samples seen up to the ith position in the ranked list. AP@k is a measure that combines recall and precision for ranked results. It is considered a reasonable evaluation metric for emphasizing the return of more highly likely positive samples at the top of the ranked list[67].

We trained our GAT models using the Pytorch Geometric framework[68]. The inputs to the model included a feature matrix $X \in R^{N \times p}$ and network edge list (Extended Data Fig. 7). In this study, p was set to 10, representing the significance of gene mutation in ten cancer cohorts. Cross-entropy loss was computed as $L = -(y\log(h) + (1-y)\log(1-h))$, where h is the output of the network after sigmoidal activation and y is the node label (0 or 1). The ADAM optimizer[69] was used for training with an exponentially decaying learning rate ($\gamma = 0.99$) starting at 0.001. We applied early stopping to prevent overfitting. For the baseline random forest model, only the feature matrix was needed. Default parameters provided in the scikit-learn package[70] were used.

Published causal evidence supporting predicted cancer drivers

Each of the predicted cancer drivers described above was used to search PubMed with the following terms on December 20, 2023: 'gene (CRISPR OR KO OR shRNA OR siRNA knockdown OR silencing OR overexpression OR over-expression) cancer', where 'gene' was replaced by the predicted cancer driver. Search results were sorted in descending order with respect to published date. Abstracts or manuscript texts were then manually vetted for causal evidence that genetic and/or pharmacologic perturbation of the predicted cancer driver functionally impacted cancer phenotypes (proliferation, migration, invasion, etc.) or augmented drug responses in model systems. This continued for each gene until all search records were verified or until ten publications by recent publication date were found with causal evidence impacting cancer phenotypes and/or drug response (Supplementary Table 7).

Genetic dependency in cancer cell lines

Cancer cell line annotations (sample_info.csv) and gene effect dependency scores derived from the integration of CRISPR KO screens published by Broad's Achilles and Sanger's SCORE projects were retrieved from DepMap Public 22Q2 (CRISPR_gene_effect_.csv)[71,72]. Cancer cell lines were matched to tumor cancer types by using the following filters: BRCA: primary_disease = 'breast cancer' and lineage = 'breast'; GBM: primary_disease = 'brain cancer' and lineage = 'central_nervous_system'; LUAD: primary_disease = 'lung cancer', lineage = 'lung' and lineage_sub_subtype = 'NSCLC_adenocarcinoma'; PDAC: primary_disease = 'pancreatic cancer' and lineage = 'pancreas'. For each cancer cell lineage, a one-sample, one-tailed t-test was used to identify *LGI3* and *FAT1* associated with significantly higher cell growth following gene KO.

Statistics and reproducibility

All data used for machine learning and gene dependency analysis are from publicly available resources[15,71,72] with detailed methodologies for data collection, blinding, randomization and protection. Sample sizes were from the original publications and they were sufficient for all statistical tests performed. Nonparametric statistical tests were used whenever possible. For parametric tests, normality of data distributions was assumed, although this was not formally tested. No data were excluded from analyses. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

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

Data availability

Proteomics and RNAseq data for the ten CPTAC cancer types were derived from the CPTAC pan-cancer study[15] (https://proteomic.datacommons.cancer.gov/pdc/cptac-pancancer). Proteomics and RNAseq data for HCC were downloaded from the original publication[55]. The data tables derived from these resources and used as input for FunMap construction are available from Zenodo (https://doi.org/10.5281/zenodo.7948943)[73]. Derived feature data for XGBoost model training are available from Zenodo (https://doi.org/10.5281/zenodo.7949374)[74]. XGBoost prediction scores for all gene pairs are available from Zenodo (https://doi.org/10.5281/zenodo.10080763)[75]. The FunMap edge list, dense modules and hierarchical modules are available online (https://funmap.linkedomics.org/). The same website also provides visualization tools to explore the gene neighborhoods, dense modules and hierarchical organization of FunMap. Additionally, the FunMap network and modules were integrated into WebGestalt[76] for enrichment analysis of user-provided gene lists. Cell line annotations and CRISPR KO dependency scores can be retrieved from the DepMap website (https://www.depmap.org). Other datasets used in the study included the gene cofunctionality gold standard derived from the Reactome pathway database[12], ProHD[12], BioPlex[18], HuRI[19], HI-Union[19] and BioGRID[20]. Source data are provided with this paper.

Code availability

The FunMap Python package is fully open source and available for download from the Python Package Index (https://pypi.org/project/funmap). The source code is hosted on GitHub (https://github.com/bzhanglab/funmap). Other supporting software is available as follows: scikit-learn 1.3.2 (https://scikit-learn.org/stable/index.html), ICE 1.0.2 (http://ice.zhang-lab.org), NetSAM 1.44.0 (https://www.bioconductor.org/packages/release/bioc/html/NetSAM.html), WebGestaltR 0.4.6 (https://cran.r-project.org/web/packages/WebGestaltR/index.html) and pytorch_geometric 1.7.2 (https://github.com/pyg-team/pytorch_geometric).

References

1. Ostroverkhova, D., Przytycka, T. M. & Panchenko, A. R. Cancer driver mutations: predictions and reality. *Trends Mol. Med.* **29**, 554–566 (2023).
2. Kustatscher, G. et al. Understudied proteins: opportunities and challenges for functional proteomics. *Nat. Methods* **19**, 774–779 (2022).
3. Dinstag, G. & Shamir, R. PRODIGY: personalized prioritization of driver genes. *Bioinformatics* **36**, 1831–1839 (2020).
4. Leiserson, M. D. M. et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).
5. Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Mol. Syst. Biol.* **3**, 88 (2007).
6. Kim, M. et al. A protein interaction landscape of breast cancer. *Science* **374**, eabf3066 (2021).

7. Swaney, D. L. et al. A protein network map of head and neck cancer reveals *PIK3CA* mutant drug sensitivity. *Science* **374**, eabf2911 (2021).
8. Quackenbush, J. Microarrays—guilt by association. *Science* **302**, 240–241 (2003).
9. Yanai, I. et al. Similar gene expression profiles do not imply similar tissue functions. *Trends Genet.* **22**, 132–138 (2006).
10. Wang, J. et al. Proteome profiling outperforms transcriptome profiling for coexpression based gene function prediction. *Mol. Cell. Proteomics* **16**, 121–134 (2017).
11. Ribeiro, D. M., Ziyani, C. & Delaneau, O. Shared regulation and functional relevance of local gene co-expression revealed by single cell analysis. *Commun. Biol.* **5**, 876 (2022).
12. Kustatscher, G. et al. Co-regulation map of the human proteome enables identification of protein functions. *Nat. Biotechnol.* **37**, 1361–1371 (2019).
13. Wu, L. et al. Variation and genetic control of protein abundance in humans. *Nature* **499**, 79–82 (2013).
14. Lapek, J. D. Jr et al. Detection of dysregulated protein-association networks by high-throughput proteomics predicts cancer vulnerabilities. *Nat. Biotechnol.* **35**, 983–989 (2017).
15. Li, Y. et al. Proteogenomic data and resources for pan-cancer analysis. *Cancer Cell* **41**, 1397–1406 (2023).
16. Zhu, H. et al. Proteomics of adjacent-to-tumor samples uncovers clinically relevant biological events in hepatocellular carcinoma. *Natl Sci. Rev.* **10**, nwad167 (2023).
17. Obayashi, T. & Kinoshita, K. Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA Res.* **16**, 249–260 (2009).
18. Huttlin, E. L. et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell* **184**, 3022–3040 (2021).
19. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).
20. Oughtred, R. et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **30**, 187–200 (2021).
21. Szklarczyk, D. et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).
22. Tsitsiridis, G. et al. CORUM: the comprehensive resource of mammalian protein complexes—2022. *Nucleic Acids Res.* **51**, D539–D545 (2023).
23. Shi, Z., Derow, C. K. & Zhang, B. Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. *BMC Syst. Biol.* **4**, 74 (2010).
24. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
25. Shi, Z., Wang, J. & Zhang, B. NetGestalt: integrating multidimensional omics data over biological networks. *Nat. Methods* **10**, 597–598 (2013).
26. Knijnenburg, T. A., Bismeijer, T., Wessels, L. F. A. & Shmulevich, I. A multilevel pan-cancer map links gene mutations to cancer hallmarks. *Chin. J. Cancer* **34**, 439–449 (2015).
27. Chen, Y., Verbeek, F. J. & Wolstencroft, K. Establishing a consensus for the hallmarks of cancer based on gene ontology and pathway annotations. *BMC Bioinformatics* **22**, 178 (2021).
28. Chen, X. & Cubillos-Ruiz, J. R. Endoplasmic reticulum stress signals in the tumour and its microenvironment. *Nat. Rev. Cancer* **21**, 71–88 (2021).
29. Vasaikar, S. et al. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. *Cell* **177**, 1035–1049 (2019).
30. Zhang, B. et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382–387 (2014).
31. Giacinti, C. & Giordano, A. RB and cell cycle progression. *Oncogene* **25**, 5220–5227 (2006).
32. Deacu, E. et al. Activin type II receptor restoration in *ACVR2*-deficient colon cancer cells induces transforming growth factor-β response pathway genes. *Cancer Res.* **64**, 7690–7696 (2004).
33. Yang, P. et al. SET domain containing 1B gene is mutated in primary hepatic neuroendocrine tumors. *Int. J. Cancer* **145**, 2986–2995 (2019).
34. Chorley, B. N. et al. Identification of novel NRF2-regulated genes by ChIP-Seq: influence on retinoid X receptor alpha. *Nucleic Acids Res.* **40**, 7416–7429 (2012).
35. Penning, T. M. Aldo-keto reductase regulation by the NRF2 system: implications for stress response, chemotherapy drug resistance, and carcinogenesis. *Chem. Res. Toxicol.* **30**, 162–176 (2017).
36. Chen, Y.-T., Shi, D., Yang, D. & Yan, B. Antioxidant sulforaphane and sensitizer trinitrobenzene sulfonate induce carboxylesterase-1 through a novel element transactivated by nuclear factor-E2 related factor-2. *Biochem. Pharmacol.* **84**, 864–871 (2012).
37. Thimmulappa, R. K. et al. Identification of NRF2-regulated genes induced by the chemopreventive agent sulforaphane by oligonucleotide microarray. *Cancer Res.* **62**, 5196–5203 (2002).
38. Rojo de la Vega, M., Chapman, E. & Zhang, D. D. NRF2 and the hallmarks of cancer. *Cancer Cell* **34**, 21–43 (2018).
39. Xu, I. M.-J. et al. Transketolase counteracts oxidative stress to drive cancer development. *Proc. Natl Acad. Sci. USA* **113**, E725–E734 (2016).
40. Loignon, M. et al. *Cul3* overexpression depletes NRF2 in breast cancer and is associated with sensitivity to carcinogens, to oxidative stress, and to chemotherapy. *Mol. Cancer Ther.* **8**, 2432–2440 (2009).
41. Kalthoff, S., Ehmer, U., Freiberg, N., Manns, M. P. & Strassburg, C. P. Interaction between oxidative stress sensor NRF2 and xenobiotic-activated aryl hydrocarbon receptor in the regulation of the human phase II detoxifying UDP-glucuronosyltransferase 1A10. *J. Biol. Chem.* **285**, 5993–6002 (2010).
42. Bech-Otschir, D. et al. COP9 signalosome-specific phosphorylation targets p53 to degradation by the ubiquitin system. *EMBO J.* **20**, 1630–1639 (2001).
43. Pineau, C. et al. Cell type-specific expression of testis elevated genes based on transcriptomics and antibody-based proteomics. *J. Proteome Res.* **18**, 4215–4230 (2019).
44. Pan, D. et al. A major chromatin regulator determines resistance of tumor cells to T cell-mediated killing. *Science* **359**, 770–775 (2018).
45. Miao, D. et al. Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science* **359**, 801–806 (2018).
46. Wu, L. et al. KDM5 histone demethylases repress immune response via suppression of STING. *PLoS Biol.* **16**, e2006134 (2018).
47. Liu, S., Liu, T., Jiang, J., Guo, H. & Yang, R. p53 mutation and deletion contribute to tumor immune evasion. *Front. Genet.* **14**, 1088455 (2023).
48. Hu, C. et al. ATRX loss promotes immunosuppressive mechanisms in *IDH1* mutant glioma. *Neuro. Oncol.* **24**, 888–900 (2022).
49. Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res.* **45**, W130–W137 (2017).

50. Gaudet, P., Livstone, M. S., Lewis, S. E. & Thomas, P. D. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief. Bioinform.* **12**, 449–462 (2011).
51. Srivastava, A. et al. *MAB21L4* deficiency drives squamous cell carcinoma via activation of RET. *Cancer Res.* **82**, 3143–3157 (2022).
52. Futreal, P. A. et al. A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
53. Pastushenko, I. et al. *Fat1* deletion promotes hybrid EMT state, tumour stemness and metastasis. *Nature* **589**, 448–455 (2021).
54. Bennett, H. M., Stephenson, W., Rose, C. M. & Darmanis, S. Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. *Nat. Methods* **20**, 363–374 (2023).
55. Gao, Q. et al. Integrated proteogenomic characterization of HBV-related hepatocellular carcinoma. *Cell* **179**, 561–577 (2019).
56. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ed. Krishnapuram, B.) (Association for Computing Machinery, 2016).
57. Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl Acad. Sci. USA* **103**, 8577–8582 (2006).
58. Plaisier, C. L., Pan, M. & Baliga, N. S. A miRNA-regulatory network explains how dysregulated miRNAs perturb oncogenic processes across diverse cancers. *Genome Res.* **22**, 2302–2314 (2012).
59. Hirsch, T. et al. Regeneration of the entire human epidermis using transgenic stem cells. *Nature* **551**, 327–332 (2017).
60. Kiefer, J. et al. Abstract 3589: a systematic approach toward gene annotation of the hallmarks of cancer. *Cancer Res.* **77**, 3589 (2017).
61. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
62. Mongiat, M., Andreuzzi, E., Tarticchio, G. & Paulitti, A. Extracellular matrix, a hard player in angiogenesis. *Int. J. Mol. Sci.* **17**, 1822 (2016).
63. Veličković, P. et al. Graph attention networks. Preprint at https://arxiv.org/abs/1710.10903 (2017).
64. Bekker, J. & Davis, J. Learning from positive and unlabeled data: a survey. *Mach. Learn.* **109**, 719–760 (2020).
65. Mordelet, F. & Vert, J.-P. A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognit. Lett.* **37**, 201–209 (2014).
66. Liu, T.-Y. *Learning to Rank for Information Retrieval* 1st edn (Springer, 2009).
67. Zhang, E. & Zhang, Y. Average precision. In *Encyclopedia of Database Systems* (eds Liu, L. & Özsu, M. T.) 192–193 (Springer, 2009).
68. Fey, M. & Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. Preprint at https://arxiv.org/abs/1903.02428 (2019).
69. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
70. Pedregosa, F. et al. Scikit-learn: machine learning in python. Preprint at https://arxiv.org/abs/1201.0490 (2012).
71. 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).
72. Pacini, C. et al. Integrated cross-study datasets of genetic dependencies in cancer. *Nat. Commun.* **12**, 1661 (2021).
73. Shi, Z. FunMap input expression data matrices. *Zenodo* https://doi.org/10.5281/zenodo.7948943 (2023).
74. Shi, Z. FunMap feature data file used for model training. *Zenodo* https://doi.org/10.5281/zenodo.7949374 (2023).
75. Shi, Z. FunMap prediction scores for all gene pairs. *Zenodo* https://doi.org/10.5281/zenodo.10080763 (2023).
76. Elizarraras, J. M. et al. WebGestalt 2024: faster gene set analysis and new support for metabolomics and multi-omics. *Nucleic Acids Res.* **52**, W415–W421 (2024).

Acknowledgements

We gratefully acknowledge contributions from the CPTAC and its Pan-Cancer Analysis working group. This work was supported by National Institutes of Health grants from the National Cancer Institute (U24 CA210954, U24 CA271076, R01 CA245903 and U01 CA271247 to B.Z.), by the Cancer Prevention and Research Institute of Texas (CPRIT; award RR160027 to B.Z.) and by the McNair Medical Institute at The Robert and Janice McNair Foundation (to B.Z.). B.Z. is a CPRIT scholar in cancer research and a McNair scholar.

Author contributions

Conceptualization, B.Z.; methodology, Z.S. and B.Z.; formal analysis, Z.S. and J.T.L.; investigation, Z.S., J.T.L. and B.Z.; resources, Z.S. and J.M.E.; data curation, Z.S. and J.T.L.; writing—original draft, Z.S., J.T.L. and B.Z.; visualization, Z.S., J.T.L. and J.M.E.; supervision, B.Z.; funding acquisition, B.Z.

Competing interests

B.Z. received research funding from AstraZeneca and consulting fees from Inotiv. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43018-024-00869-z.

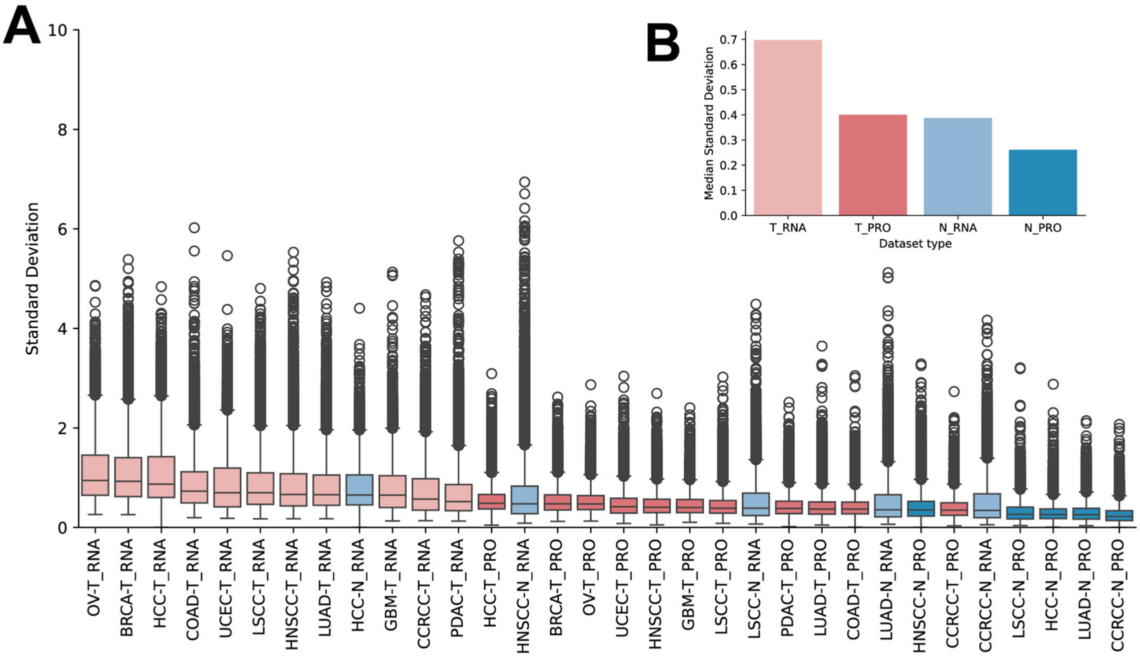
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43018-024-00869-z.

Correspondence and requests for materials should be addressed to Bing Zhang.

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

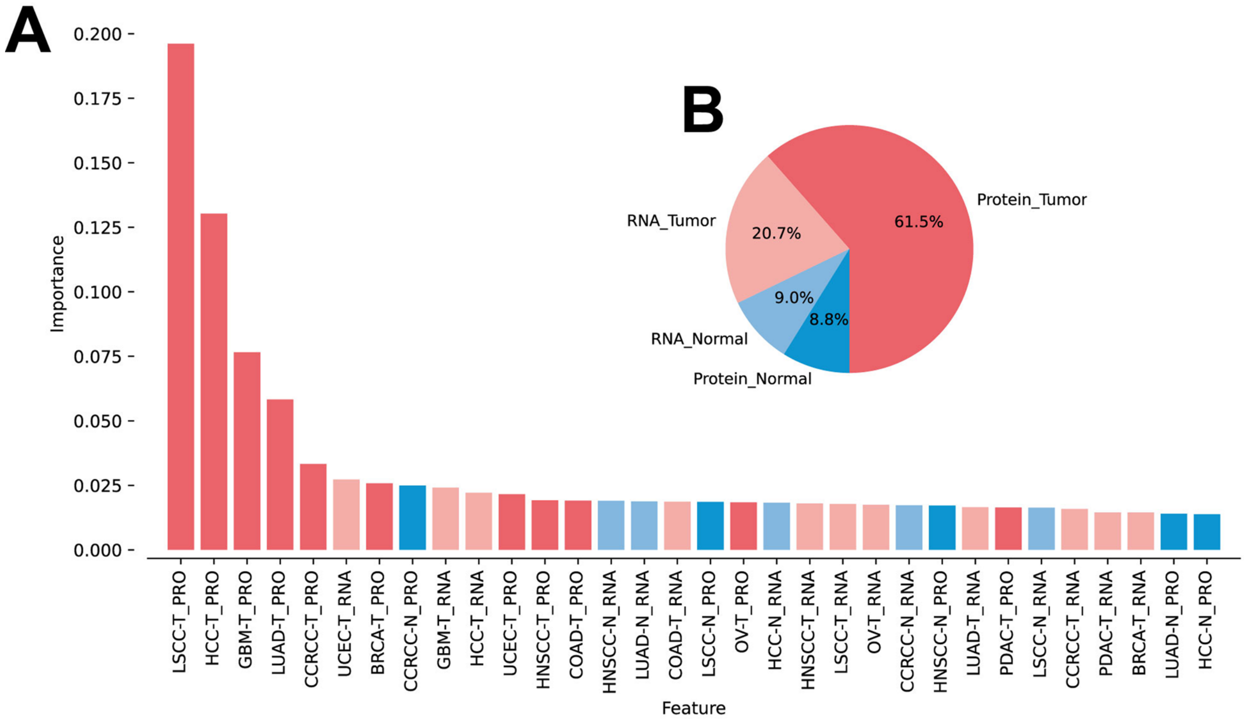
Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s43018-024-00869-z



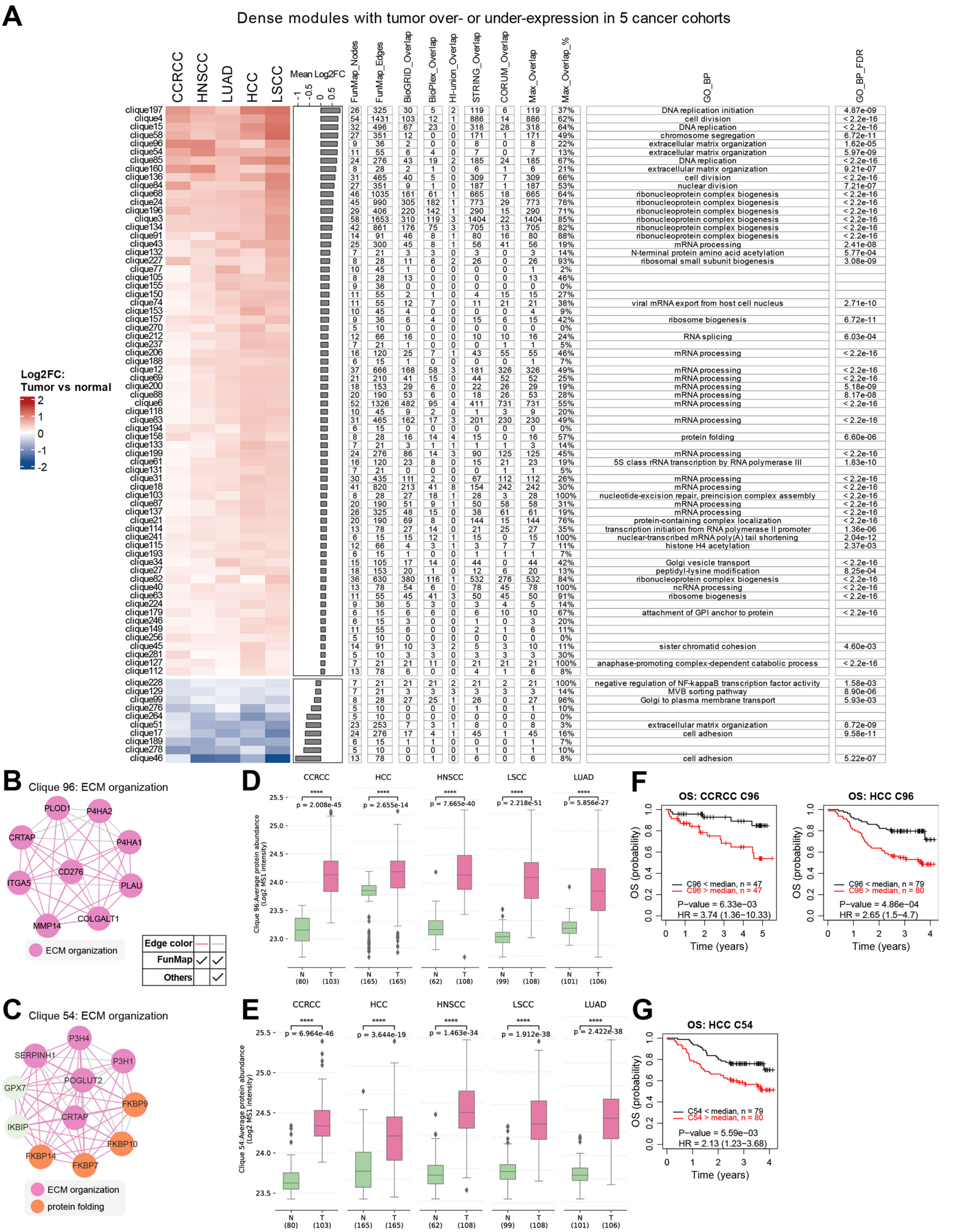
Extended Data Fig. 1 | Quantification of inter-sample heterogeneity through gene-wise standard deviation. A) Distributions of gene-wise standard deviations across individual datasets (n = 17,733 to 19,113 mRNAs and n = 7,961 to 11,815 proteins). For boxplots, centerline indicates the median, box limits indicate upper and lower quartiles, whiskers indicate the 1.5 interquartile range. **B**) Median values of the median standard deviations across various dataset groups. T: Tumor; N: Normal.

https://doi.org/10.1038/s43018-024-00869-z



Extended Data Fig. 2 | Breakdown of feature importance in the XGBoost model. A) Barplot showing importance of individual features. **B**) Pie chart depicting aggregated importance by data and sample type pairs.

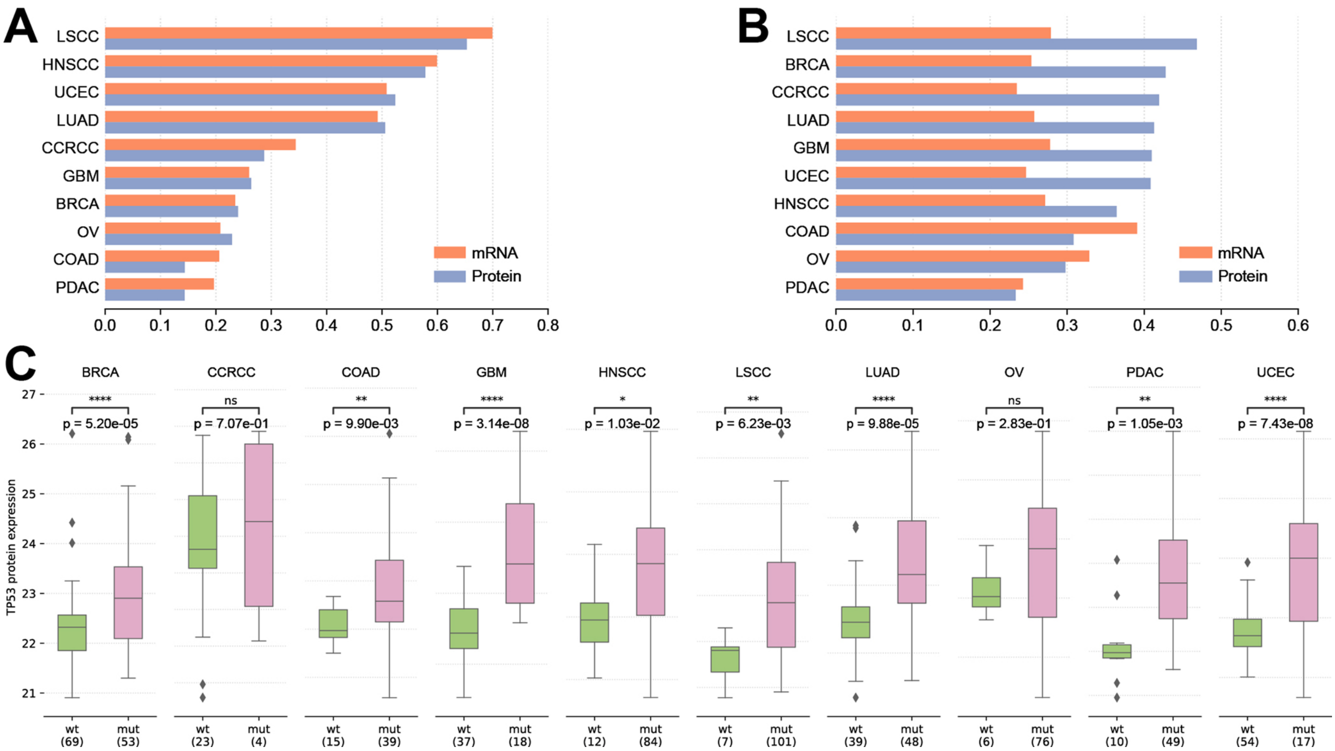
https://doi.org/10.1038/s43018-024-00869-z



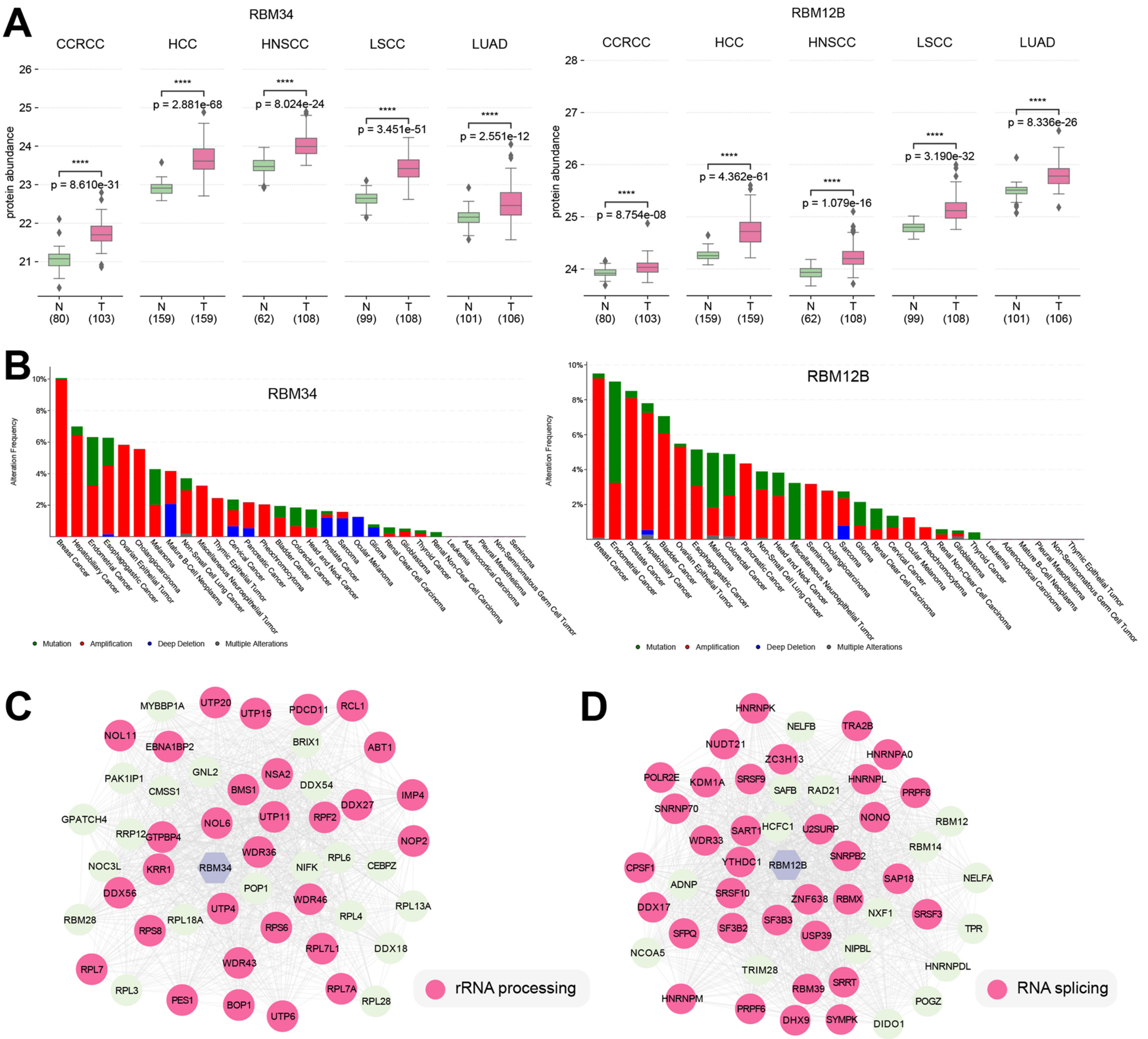
	FunMap_Nodes	FunMap_Edges	BioGRID_Overlap	BioPlex_Overlap	HI-union_Overlap	STRING_Overlap	CORUM_Overlap	Max_Overlap	Max_Overlap_%	GO_BP	GO_BP_FDR
clique197	26	325	30	5	2	119	6	119	37%	DNA replication initiation	4.87e-09
clique4	54	1431	103	12	1	886	14	886	62%	cell division	< 2.2e-16
clique15	32	496	67	23	0	318	28	318	64%	DNA replication	< 2.2e-16
clique58	27	351	12	0	0	171	1	171	49%	chromosome segregation	6.72e-11
clique96	9	36	2	0	0	8	0	8	22%	extracellular matrix organization	1.62e-05
clique54	11	55	6	4	0	7	0	7	13%	extracellular matrix organization	5.97e-09
clique85	24	276	43	19	2	185	24	185	67%	DNA replication	< 2.2e-16
clique160	8	28	2	1	0	6	1	6	21%	extracellular matrix organization	9.21e-07
clique136	31	465	40	5	0	309	7	309	66%	cell division	< 2.2e-16
clique84	27	351	9	1	0	187	1	187	53%	nuclear division	7.21e-07
clique68	46	1035	161	61	1	665	18	665	64%	ribonucleoprotein complex biogenesis	< 2.2e-16
clique24	45	990	305	182	1	773	29	773	78%	ribonucleoprotein complex biogenesis	< 2.2e-16
clique196	29	406	220	142	1	290	15	290	71%	ribonucleoprotein complex biogenesis	< 2.2e-16
clique3	58	1653	310	119	3	1404	22	1404	85%	ribonucleoprotein complex biogenesis	< 2.2e-16
clique134	42	861	176	75	3	705	13	705	82%	ribonucleoprotein complex biogenesis	< 2.2e-16
clique91	14	91	46	8	1	80	16	80	88%	ribonucleoprotein complex biogenesis	< 2.2e-16
clique43	25	300	45	8	1	56	41	56	19%	mRNA processing	2.41e-08
clique132	7	21	3	3	0	3	0	3	14%	N-terminal protein amino acid acetylation	5.77e-04
clique227	8	28	11	6	2	26	0	26	93%	ribosomal small subunit biogenesis	3.08e-09
clique77	10	45	1	0	0	0	0	1	2%		
clique105	8	28	13	0	0	0	0	13	46%		
clique155	9	36	0	0	0	0	0	0	0%		
clique150	11	55	2	1	0	4	15	15	27%		
clique74	11	55	12	7	0	11	21	21	38%	viral mRNA export from host cell nucleus	2.71e-10
clique153	10	45	4	0	0	0	0	4	9%		
clique157	9	36	6	4	0	15	6	15	42%	ribosome biogenesis	6.72e-11
clique270	5	10	0	0	0	0	0	0	0%		
clique212	12	66	16	0	0	10	10	16	24%	RNA splicing	6.03e-04
clique237	7	21	1	0	0	0	1	1	5%		
clique206	16	120	25	7	1	43	55	55	46%	mRNA processing	< 2.2e-16
clique188	6	15	1	0	0	0	0	1	7%		
clique12	37	666	168	58	3	181	326	326	49%	mRNA processing	< 2.2e-16
clique69	21	210	41	15	0	44	52	52	25%	mRNA processing	< 2.2e-16
clique200	18	153	29	6	0	22	26	29	19%	mRNA processing	5.18e-09
clique88	20	190	53	6	0	18	26	53	28%	mRNA processing	8.17e-08
clique6	52	1326	482	95	4	411	731	731	55%	mRNA processing	< 2.2e-16
clique118	10	45	9	2	0	1	3	9	20%		
clique83	31	465	162	17	3	201	230	230	49%	mRNA processing	< 2.2e-16
clique194	6	15	0	0	0	0	0	0	0%		
clique158	8	28	16	14	4	15	0	16	57%	protein folding	6.60e-06
clique133	7	21	3	1	1	1	1	3	14%		
clique199	24	276	86	14	3	90	125	125	45%	mRNA processing	< 2.2e-16
clique61	16	120	23	8	0	15	21	23	19%	5S class rRNA transcription by RNA polymerase III	1.83e-10
clique131	7	21	0	0	0	0	1	1	5%		
clique31	30	435	111	2	0	67	112	112	26%	mRNA processing	< 2.2e-16
clique18	41	820	213	41	8	154	242	242	30%	mRNA processing	< 2.2e-16
clique103	8	28	27	18	1	28	3	28	100%	nucleotide-excision repair, preincision complex assembly	< 2.2e-16
clique87	20	190	51	9	1	50	58	58	31%	mRNA processing	< 2.2e-16
clique137	26	325	48	15	0	38	61	61	19%	mRNA processing	< 2.2e-16
clique21	20	190	69	8	0	144	15	144	76%	protein-containing complex localization	< 2.2e-16
clique114	13	78	27	14	0	21	25	27	35%	transcription initiation from RNA polymerase II promoter	1.36e-06
clique241	6	15	15	12	1	15	0	15	100%	nuclear-transcribed mRNA poly(A) tail shortening	2.04e-12
clique115	12	66	4	3	1	3	7	7	11%	histone H4 acetylation	2.37e-03
clique193	6	15	1	0	0	1	1	1	7%		
clique34	15	105	17	14	0	44	0	44	42%	Golgi vesicle transport	< 2.2e-16
clique27	18	153	20	1	0	12	6	20	13%	peptidyl-lysine modification	8.25e-04
clique82	36	630	380	116	1	532	276	532	84%	ribonucleoprotein complex biogenesis	< 2.2e-16
clique40	13	78	54	6	0	78	45	78	100%	ncRNA processing	< 2.2e-16
clique63	11	55	45	41	3	50	45	50	91%	ribosome biogenesis	< 2.2e-16
clique224	9	36	5	3	0	3	4	5	14%		
clique179	6	15	6	6	0	6	10	10	67%	attachment of GPI anchor to protein	< 2.2e-16
clique246	6	15	3	0	0	1	0	3	20%		
clique149	11	55	6	0	0	2	1	6	11%		
clique256	5	10	0	0	0	0	0	0	0%		
clique45	14	91	10	3	2	5	3	10	11%	sister chromatid cohesion	4.60e-03
clique281	5	10	3	3	0	3	3	3	30%		
clique127	7	21	21	11	0	21	21	21	100%	anaphase-promoting complex-dependent catabolic process	< 2.2e-16
clique112	13	78	6	0	0	4	1	6	8%		
clique228	7	21	21	21	2	21	2	21	100%	negative regulation of NF-kappaB transcription factor activity	1.58e-03
clique129	7	21	3	3	3	3	3	3	14%	MVB sorting pathway	8.90e-06
clique99	8	28	27	25	1	26	0	27	96%	Golgi to plasma membrane transport	5.93e-03
clique276	5	10	0	0	0	1	0	1	10%		
clique264	5	10	0	0	0	0	0	0	0%		
clique51	23	253	7	3	1	8	0	8	3%	extracellular matrix organization	8.72e-09
clique17	24	276	17	4	1	45	1	45	16%	cell adhesion	9.58e-11
clique189	6	15	1	1	0	0	0	1	7%		
clique278	5	10	0	0	0	1	0	1	10%		
clique46	13	78	0	1	0	6	0	6	8%	cell adhesion	5.22e-07

Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Characterization of dense modules. A) Heatmap depicting log2 fold change (log2FC) of average protein abundance of dense modules (cliques) in tumor vs normal for each of the five cancer cohorts shown. All 78 cliques have concordant tumor over- or under-expression in all five cohorts (FDR < 0.01 in each cohort). Table shows the number and maximum number of overlapping edges with other networks as indicated. Gene ontology biological processes (GO_BP) indicates the top enriched term of a given clique (GO_BP_FDR). **B**-**C**) Tumor overexpressed, ECM-associated dense modules, Clique 96 (**B**) and Clique 54 (**C**). Edge color indicates lack of overlap in BioGRID, BioPlex, HI-union, STRING, and CORUM (pink) or overlap in any of these resources (gray). **D**-**E**) Boxplots comparing average protein abundance of Clique 96 (**D**) and Clique 54 (**E**) in tumor and normal samples demonstrating tumor overexpression in five cancer cohorts. Number of samples, n, are indicated in parenthesis. P-values determined by two-sided Wilcoxon rank-sum test. **F**-**G**) Kaplan-Meier plots depicting overall survival (OS) difference in patients from indicated cohorts stratified by median value of the average abundance of proteins in Clique 96 (**F**) and Clique 54 (**G**). Logrank p-values and hazard ratio (HR) shown with 95% confidence intervals derived from Cox-proportional hazard models. Significance is indicated as ****p < 0.0001. For boxplots, centerline indicates the median, box limits indicate upper and lower quartiles, whiskers indicate the 1.5 interquartile range, and number of samples per group indicated in parentheses.

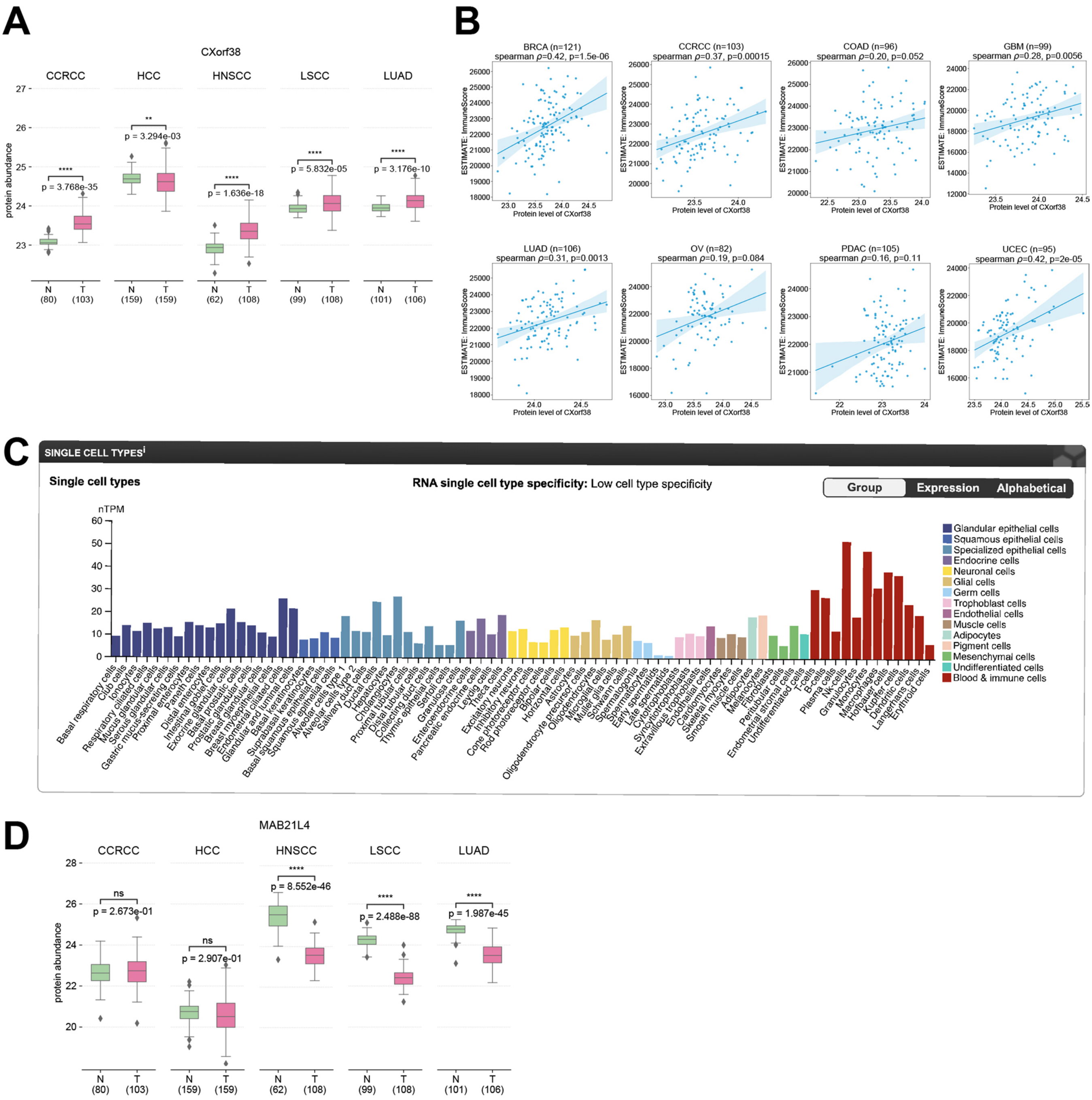


Extended Data Fig. 4 | Connecting somatic mutations to functional protein modules. A) Average pairwise Pearson's correlation coefficient for genes in L2_M40 based on mRNA or protein data in different cancer types. **B**) Average pairwise Pearson's correlation coefficient for genes in L3_M58 based on mRNA or protein data in different cancer types. **C**) Comparison of TP53 protein abundance (log2 MS1 intensity) in TP53 wildtype (wt) and mutant (mut) samples across 10 cancer types. Number of samples, n, are indicated in parenthesis. P-values were derived from two-sided Wilcoxon rank-sum test. Significance is indicated as *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, ns: not significant. For boxplots, centerline indicates the median, box limits indicate upper and lower quartiles, whiskers indicate the 1.5 interquartile range, and number of samples per group indicated in parentheses.

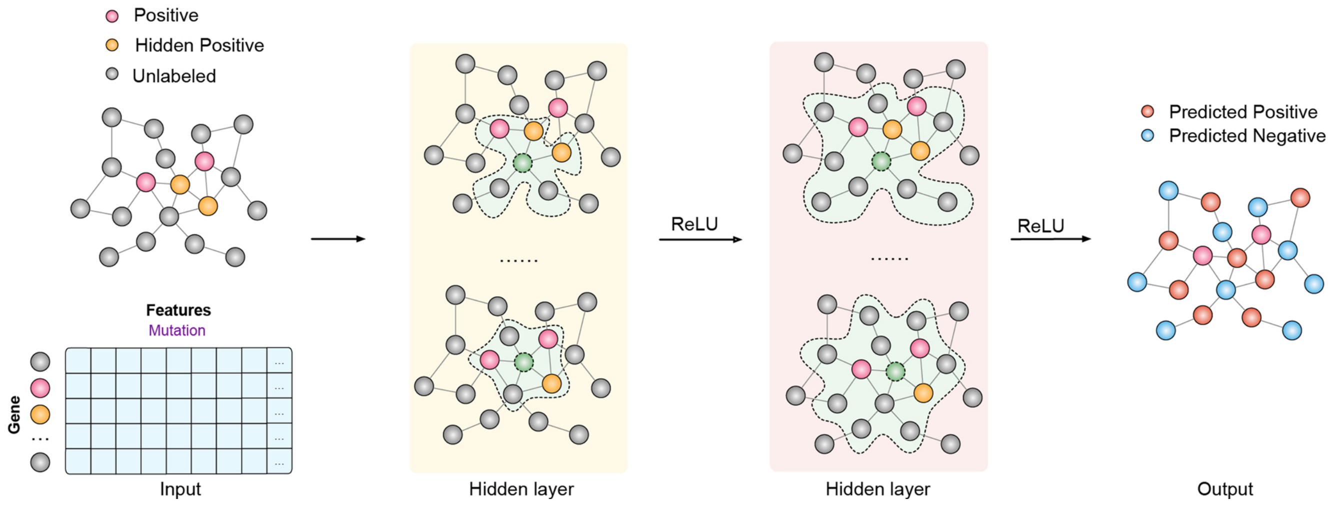


Extended Data Fig. 5 | Illuminating understudied cancer proteins RBM34 and RBM12B. A) Boxplots comparing protein abundance of RBM34 and RBM12B in tumor and normal samples demonstrating tumor over-expression in five cancer cohorts. Number of samples, n, are indicated in parenthesis. P-values determined by two-sided Wilcoxon rank-sum test. For boxplots, centerline indicates the median, box limits indicate upper and lower quartiles, whiskers indicate the 1.5 interquartile range, and number of samples per group indicated in parentheses. **B**) Barplots depicting frequency of somatic copy number and mutations in RBM34 and RBM12B from TCGA PanCancer Atlas Studies in cBioPortal. **C**-**D**) Network neighborhood of RBM34 (**C**) or RBM12B (**D**) with genes associated with the enriched GO terms highlighted.

https://doi.org/10.1038/s43018-024-00869-z



Extended Data Fig. 6 | Illuminating understudied cancer proteins CXorf38 and MAB21L4. A) Boxplots comparing protein abundance of CXorf38 in tumor and normal samples demonstrating tumor over-expression in five cancer cohorts. Number of samples, n, are indicated in parenthesis. P-values determined by two-sided Wilcoxon rank-sum test. **B**) Relationship between protein abundance of CXorf38 and RNA-seq inferred ESTIMATE ImmunoScore in eight cancer types. P-values were derived from two-sided Spearman's rank correlation. Shaded area depicts the 95% confidence interval. **C**) Single cell data from the Human Protein Atlas showing that CXorf38 is expressed across all cell types, but the highest expression occurs in immune cells. **D**) Boxplots comparing protein abundance of MAB21L4 in tumor and normal samples in five cancer cohorts. Number of samples, n, are indicated in parenthesis. P-values determined by two-sided Wilcoxon rank-sum test. Significance is indicated as *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, ns: not significant. For boxplots, centerline indicates the median, box limits indicate upper and lower quartiles, whiskers indicate the 1.5 interquartile range, and number of samples per group indicated in parentheses.



Extended Data Fig. 7 | Graph neural network architecture for predicting cancer driver genes based on network topology and mutation data. The model takes as input mutation data for genes represented in a feature matrix. Nodes in the graph correspond to genes, where pink nodes are known positive driver genes, orange nodes are hidden positive genes, and gray nodes are unlabeled genes. Both the node features and network topology are processed through hidden layers with ReLU activations. The output layer predicts gene classifications, with red nodes indicating predicted positive driver genes and blue nodes indicating predicted negative genes.

nature portfolio

Corresponding author(s): Bing Zhang

Last updated by author(s): Aug 27, 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 FunMap Python package is fully open source and available for download from the Python Package Index (PyPI) at https://pypi.org/project/funmap. The source code is hosted on GitHub at: https://github.com/bzhanglab/funmap. Other supporting software is available as follows: scikit-learn 1.3.2 (https://scikit-learn.org/stable/index.html), ICE 1.0.2 (http://ice.zhang-lab.org), NetSAM 1.44.0 (https://www.bioconductor.org/packages/release/bioc/html/NetSAM.html), WebGestaltR 0.4.6 (https://cran.r-project.org/web/packages/WebGestaltR/index.html). pytorch_geometric 1.7.2 (https://github.com/pyg-team/pytorch_geometric).

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

Proteomics and RNASeq data for the 10 CPTAC cancer types were derived from the CPTAC pan-cancer study15: https://proteomic.datacommons.cancer.gov/pdc/cptac-pancancer. Proteomics and RNASeq data for HCC were downloaded from the original publication55. The data tables derived from these resources and used as input for FunMap construction are available at https://zenodo.org/record/7948944. Derived feature data for XGBoost model training are available at https://zenodo.org/records/7949375. XGBoost prediction scores for all gene pairs are available at https://zenodo.org/records/10080764. FunMap edge list, dense modules, and hierarchical modules can be downloaded at: https://funmap.linkedomics.org/. FunMap edge list, dense modules, and hierarchical modules can be downloaded at: https://funmap.linkedomics.org/. The same web site also provides visualization tools to explore gene neighborhoods, dense modules, and hierarchical organization of FunMap. Additionally, FunMap network and modules have been integrated into WebGestalt73 for enrichment analysis of user provided gene lists. Cell line annotations and CRISPR KO dependency scores can be retrieved from the DepMap website: https://www.depmap.org. Other datasets used in the study included gene co-functionality "gold standard" derived from the Reactome pathway database12, ProHD12, BioPlex18, HuRI19, HI-Union19, and BioGRID20.

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	Publicly available data were used, and sex and gender were not considered in the analysis
Reporting on race, ethnicity, or other socially relevant groupings	Publicly available data were used, and race, ethnicity, or other socially relevant groupings were not considered in the analysis
Population characteristics	The datasets were not selected and analyzed based on specific population characteristics beyond availability of public datasets.
Recruitment	The datasets were not selected and analyzed based on specific population characteristics beyond availability of public datasets.
Ethics oversight	N/A

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	No new data generation. Sample sizes were from the original publications, and they were sufficient for all statistical testes performed.
Data exclusions	None.
Replication	No new data generation, this study reanalyzes previously published data.
Randomization	Randomization is not applicable because there was no new experiments.
Blinding	Blinding is not applicable because there was no new experiments.

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

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A