RESEARCH ARTICLE SUMMARY

SYSTEMS BIOLOGY

Interpretation of cancer mutations using a multiscale map of protein systems

Fan Zheng†, Marcus R. Kelly†, Dana J. Ramms, Marissa L. Heintschel, Kai Tao, Beril Tutuncuoglu, John J. Lee, Keiichiro Ono, Helene Foussard, Michael Chen, Kari A. Herrington, Erica Silva, Sophie N. Liu, Jing Chen, Christopher Churas, Nicholas Wilson, Anton Kratz, Rudolf T. Pillich, Devin N. Patel, Jisoo Park, Brent Kuenzi, Michael K. Yu, Katherine Licon, Dexter Pratt, Jason F. Kreisberg, Minkyu Kim, Danielle L. Swaney, Xiaolin Nan, Stephanie I. Fraley, J. Silvio Gutkind, Nevan J. Krogan*, Trey Ideker*

INTRODUCTION: Tumor genome sequencing has revealed that, beyond a few commonly mutated genes, most mutations that affect cancer genomes are rare. To interpret these rare events, a powerful approach has been to organize mutations by their effects on commonly dysregulated cellular systems. Understanding the cancer genome in this way requires surmounting two challenges: (i) How do we comprehensively map cancer cell systems? (ii) How do we identify which systems are under mutational selection?

RATIONALE: To address these questions, we used proteomic mass spectrometry and data integration to build a structured map of protein assemblies found in human cancer cells. We then developed a statistical model of mutation, pinpointing which assemblies are under strong mutational selection and in which cancer types. The goal was to interpret the many rare gene mutations that affect tumor genomes by their convergence on higher-order entities.

RESULTS: We amassed a large compendium of cancer protein interactions, combining the screens in breast cancer (Kim *et al.*, this issue) and head-and-neck cancer (Swaney *et al.*, this issue) with multi-omic evidence from 127 previous studies. Lines of evidence were integrated quantitatively to yield a continuous metric of association for each protein pair (integrated association stringency, or IAS). This network of protein associations exhibited clear multiscale and modular structure, revealing 2338 robust assemblies of interacting proteins (hereafter "protein systems") across different stringencies. Systems were organized hierarchically, with small high-stringency systems (e.g., specific complexes) combining in larger ones (e.g., processes and organelles) as stringency was relaxed.

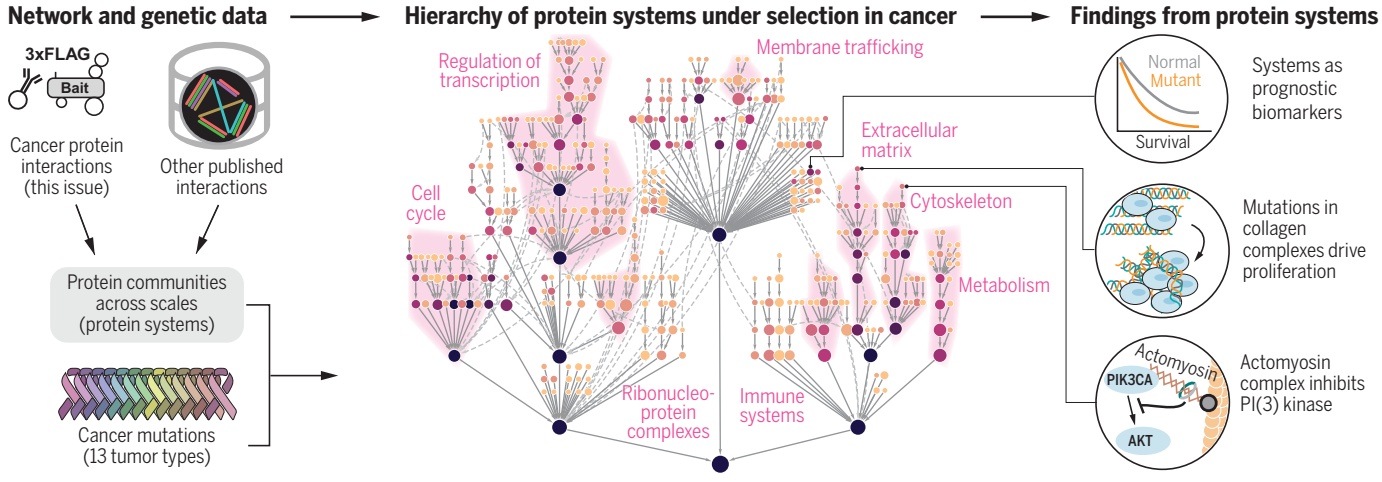
We next developed a statistical model, HiSig, to identify a parsimonious set of systems that best explains the gene mutation frequencies observed in tumors. HiSig analysis of 13 tumor types yielded a map of 395 mutated protein systems we call NeST (Nested Systems in Tumors, http://ccmi.org/nest/). NeST comprised numerous small complexes, most mutated within specific tumor types, organized within larger systems relevant to most cancers.

Although NeST recapitulated cancer hallmarks, the majority of systems had not been previously described or had not been associated with cancer mutation. Nonetheless, many were recurrently mutated in independent cohorts, supporting their significance. Notable systems included a PIK3CA-actomyosin complex that points to a new mode of phosphatidylinositol 3-kinase regulation, as well as recurrent mutations in collagen complexes that we found to disrupt the extracellular matrix, thereby promoting proliferation. Finally, we identified NeST systems that serve as biomarkers of cancer outcomes, leading to 548 genes for potential use in clinical sequencing panels.

CONCLUSION: In their classic description of the "Hallmarks of Cancer," Hanahan and Weinberg predicted that the "complexities of cancer ... will become understandable in terms of a small number of underlying principles." Around the same time, Alberts provided his seminal perspective of the cell as a collection of "protein assemblies [interacting] in an elaborate network." By organizing disparate tumor mutations into underlying principles captured by a multiscale map of protein assemblies, this work represents a synthesis of these visions. The strategies developed here may generalize to other diseases that are affected by rare genetic alterations. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: nevan.krogan@ucsf.edu (N.J.K.); tideker@health.ucsd.edu (T.I.)
†These authors contributed equally to this work.
Cite this article as F. Zheng *et al.*, *Science* **374**, eabf3067 (2021). DOI: 10.1126/science.abf3067

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abf3067



Mapping cancer protein systems. Protein interaction datasets were integrated to identify protein communities ("systems") at multiple scales of analysis (left). Each system was tested for cancer mutational selection as a system versus its substituent proteins, revealing a hierarchy of protein systems under selection in cancer (center). Discoveries from this hierarchy (right) were validated with clinical data and functional experiments.

RESEARCH ARTICLE

SYSTEMS BIOLOGY

Interpretation of cancer mutations using a multiscale map of protein systems

Fan Zheng[1,2]†, Marcus R. Kelly[1,2]†, Dana J. Ramms[2,3,4], Marissa L. Heintschel[5], Kai Tao[6,7], Beril Tutuncuoglu[2,8,9,10], John J. Lee[1], Keiichiro Ono[1], Helene Foussard[8,9,10], Michael Chen[1], Kari A. Herrington[11], Erica Silva[1], Sophie N. Liu[1], Jing Chen[1], Christopher Churas[1], Nicholas Wilson[1], Anton Kratz[1,2], Rudolf T. Pillich[1,2], Devin N. Patel[1,2], Jisoo Park[1,2], Brent Kuenzi[1,2], Michael K. Yu[1], Katherine Licon[1,2], Dexter Pratt[1], Jason F. Kreisberg[1,2], Minkyu Kim[2,8,9,10], Danielle L. Swaney[2,8,9,10], Xiaolin Nan[6,7,12], Stephanie I. Fraley[5], J. Silvio Gutkind[2,3,4], Nevan J. Krogan[2,8,9,10]*, Trey Ideker[1,2,3,5]*

A major goal of cancer research is to understand how mutations distributed across diverse genes affect common cellular systems, including multiprotein complexes and assemblies. Two challenges—how to comprehensively map such systems and how to identify which are under mutational selection—have hindered this understanding. Accordingly, we created a comprehensive map of cancer protein systems integrating both new and published multi-omic interaction data at multiple scales of analysis. We then developed a unified statistical model that pinpoints 395 specific systems under mutational selection across 13 cancer types. This map, called NeST (Nested Systems in Tumors), incorporates canonical processes and notable discoveries, including a PIK3CA-actomyosin complex that inhibits phosphatidylinositol 3-kinase signaling and recurrent mutations in collagen complexes that promote tumor proliferation. These systems can be used as clinical biomarkers and implicate a total of 548 genes in cancer evolution and progression. This work shows how disparate tumor mutations converge on protein assemblies at different scales.

Substantial progress in cataloging the molecular basis of cancer has come from genomic, transcriptomic, and proteomic profiling of thousands of patients by consortia such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). However, very different patterns of mutations are observed across different tumors and across different cells within the same tumor (*1*), hindering the interpretation of cancer genomes. Although some known cancer driver genes are mutated frequently in a statistically significant number of samples, the importance of the much greater number of low-frequency genetic events has remained largely unclear (*2*) (Fig. 1A).

A powerful way to interpret the many rare mutations is to organize them into networks of genes that participate in commonly dysregulated cellular components or processes (*3*–*5*). Alterations may be observed infrequently at the nucleotide or gene level but can be substantially more common when considering impacts in a larger biological process. For instance, the TCGA analysis of head and neck squamous cell carcinoma (*6*) reported genetic changes to a "cell differentiation pathway" in 64% of human papillomavirus–negative tumors, combining mutations across four genes that were each altered more rarely (*AJUBA*, *TP63*, *NOTCH1*, and *FAT1*, each individually in 7 to 32% of patients). Systematic aggregation of mutated genes in the form of gene sets (*7*–*10*) or molecular networks connecting pairs of functionally related genes (*11*–*20*) has been very useful in identifying higher-order systems of genes under mutational selection in cancer, many of which would otherwise be missed.

Although this strategy is promising, the realization of a complete systems-level description of cancer mutations will require overcoming two challenges in particular. The first is to assemble accurate and comprehensive knowledge maps of dysregulated cellular components and processes (henceforth called cellular "systems" for generality). Unlike whole-genome sequencing, which has provided complete information on tumor genomes, systematic efforts to map cancer cellular systems are just beginning (*21*–*28*). In this respect, proteomics efforts have used techniques such as affinity purification, proximity labeling, co-elution mass spectrometry, and yeast two-hybrid assays to create catalogs of human protein interactions, which have been useful for defining protein complexes and larger cellular components (*29*–*33*). Thus far, however, large-scale protein interaction surveys have not focused on cancer proteins in particular, and experiments have been typically conducted in model organisms or cell lines chosen for experimental tractability rather than cancer relevance. A second challenge is to identify mutational selection on biological systems larger than those encoded by single genes. Cellular components functioning in cancer can occupy a range of biophysical scales, from individual residues and domains in proteins (*34*–*37*) to impacts on multiprotein complexes (*38*–*41*), signaling networks (*42*, *43*), and classical and membraneless organelles (*44*, *45*). Analyzing the incidence of cancer mutations at only one of these scales misses components under mutational selection at all others. Accordingly, it remains largely unclear which multigene systems and scales represent the key focal points on which mutations converge.

To address these challenges, we integrated existing data resources with a collection of systematic protein interaction networks centered on cancer proteins in cancer-relevant conditions, with an emphasis on breast and head-and-neck cancers as described in (*46*, *47*). Using these data, we constructed a structured map of protein systems, not restricted to one scale but organized across a hierarchy of cellular components and processes. We then developed a unified statistical model to identify systems under mutational selection considering all scales simultaneously. Together, these analyses define a compendium of protein complexes, signaling pathways, and larger assemblies with evidence for recurrent mutation in cancer.

Interaction mapping and integrative analysis yield a hierarchy of protein systems

We amassed a large compendium of cancer protein-protein interactions (PPIs) based on affinity purification mass spectrometry (AP-MS) of 61 proteins with established roles in cancer, combining the separate screens in breast and head-and-neck tissues described in our two companion papers (*46*, *47*). In these companion studies, proteins were epitope-tagged (3×FLAG), expressed, then purified from a panel of cell lines representing malignant and nonmalignant breast tissues (tumor: MDA-MB-231, MCF-7; normal: MCF-10A; 40 tagged proteins) and/or head-and-neck tissues (tumor: CAL-33, SCC-25; normal: HET-1A; 30 tagged proteins, of which nine were also investigated in breast; table S1). Copurified proteins were then identified by

[1]Division of Genetics, Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA. [2]Cancer Cell Map Initiative (CCMI), La Jolla, CA, USA. [3]Moores Cancer Center, University of California San Diego, La Jolla, CA 92093, USA. [4]Department of Pharmacology, University of California San Diego, La Jolla, CA 92093, USA. [5]Department of Bioengineering, University of California San Diego, La Jolla, CA 92093, USA. [6]Department of Biomedical Engineering, Oregon Health and Science University, Portland, OR 97239, USA. [7]Center for Spatial Systems Biomedicine, Oregon Health and Science University, Portland, OR 97201, USA. [8]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, CA 94158, USA. [9]J. David Gladstone Institutes, San Francisco, CA 94158, USA. [10]Quantitative Biosciences Institute, University of California, San Francisco, CA 94158, USA. [11]Department of Biochemistry and Biophysics Center for Advanced Light Microscopy at UCSF, University of California, San Francisco, CA 94158, USA. [12]Knight Cancer Early Detection Advanced Research Center, Oregon Health and Science University, Portland, OR 97201, USA.
*Corresponding author. Email: nevan.krogan@ucsf.edu (N.J.K.); tideker@health.ucsd.edu (T.I.)
†These authors contributed equally to this work.

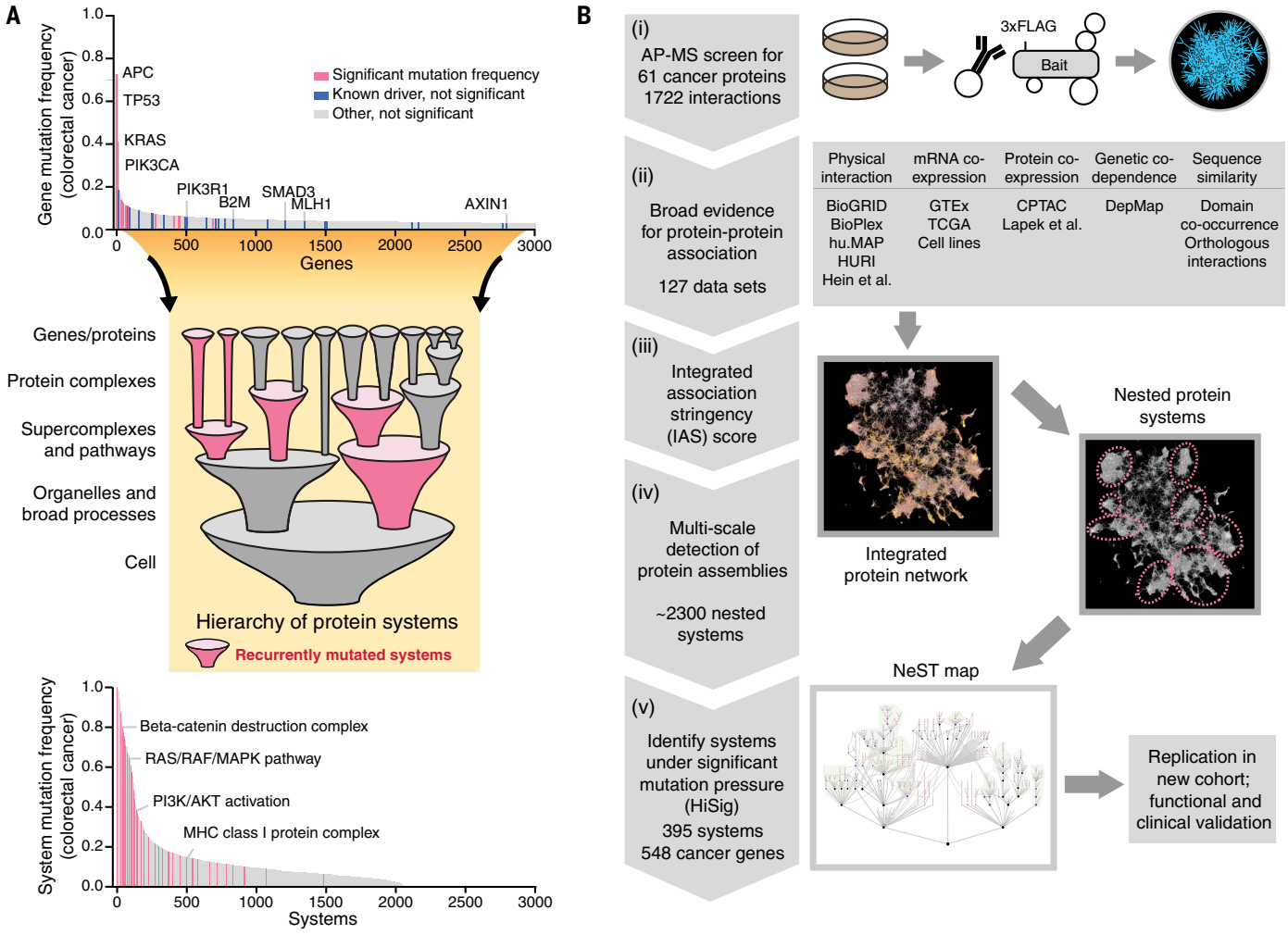


Fig. 1. Rationale for a multiscale map of cancer systems and overview of its assembly. (**A**) Cancer genes missed by single-gene mutation analysis can be recovered by identification of significantly mutated systems. In the distribution of gene mutation frequencies in colorectal adenocarcinoma (top), analysis of significantly mutated genes (pink, TCGA pan-cancer atlas) (*37*) misses a number of known colorectal cancer driver genes (blue, COSMIC Cancer Gene Census) (*72*). Representative genes from both categories are labeled. When evaluating mutation significance in a hierarchy of protein systems (middle), driver genes missed in the single-gene analysis can be recovered within significantly mutated systems (bottom, pink). (**B**) Pipeline for assembly of the cancer systems map. (i) Generate cancer protein interactions using affinity purification mass spectrometry (AP-MS). (ii) Collect previous protein association evidence of five major data types. (iii) Integrate all evidence to derive an integrated association stringency (IAS) score network for all pairs of 19,035 human proteins. (iv) Identify a hierarchy of protein systems by multiscale community detection. (v) Identify recurrently mutated systems in the hierarchy by HiSig, defining a cancer systems map, which is validated in independent cohorts and functional studies.

mass spectrometry using PPI confidence-scoring algorithms (see materials and methods). Here, we combined data across all proteins and cell lines, yielding a total of 1722 distinct interactions. Approximately 85% of PPIs had not been reported in previous AP-MS datasets or in curated databases (*30*, *48*–*50*), showing that these experiments had substantially enlarged the known interactomes of many cancer proteins (table S1).

We integrated these experiments with a broad collection of human PPI data from previous studies (*29*, *30*, *49*, *51*, *52*) alongside four additional types of evidence previously shown to inform protein physical associations (Fig. 1B). These additional types were correlated levels of protein abundance over many cell lines or tissues (*53*–*55*); correlated mRNA transcript levels over many cell lines or tissues (*56*); genetic codependencies, as revealed by correlated cell growth outcomes for CRISPR knockdowns performed over many cell lines (*57*); and sequence-based associations, including protein pairs with sequence domains that frequently interact and protein pairs with orthologs that interact in model species (*58*). For each evidence type, we performed a broad survey of studies relevant to tumor samples, tumor cell lines, and human tissues, resulting in a compendium of 127 datasets in total (Fig. 1B and table S2). We used an established method of biological network integration (*59*), based on supervised machine learning, to quantitatively weigh and combine all evidence to create a single integrated association stringency (IAS) score for each pair of human proteins (Fig. 2A and fig. S1A; 19,035 proteins and 1.8×10^8 scored protein pairs; see materials and methods). This integration system was trained for the ability to interconnect proteins in the same cellular component or biological process recorded in the Gene Ontology reference database (*60*). PPIs were the most informative evidence type for this task, followed by sequence similarity and protein coexpression (fig. S1, B and C). The resulting IAS network (Fig. 2B) has been made available for download, browsing, and query (http://ccmi.org/nest; data S1).

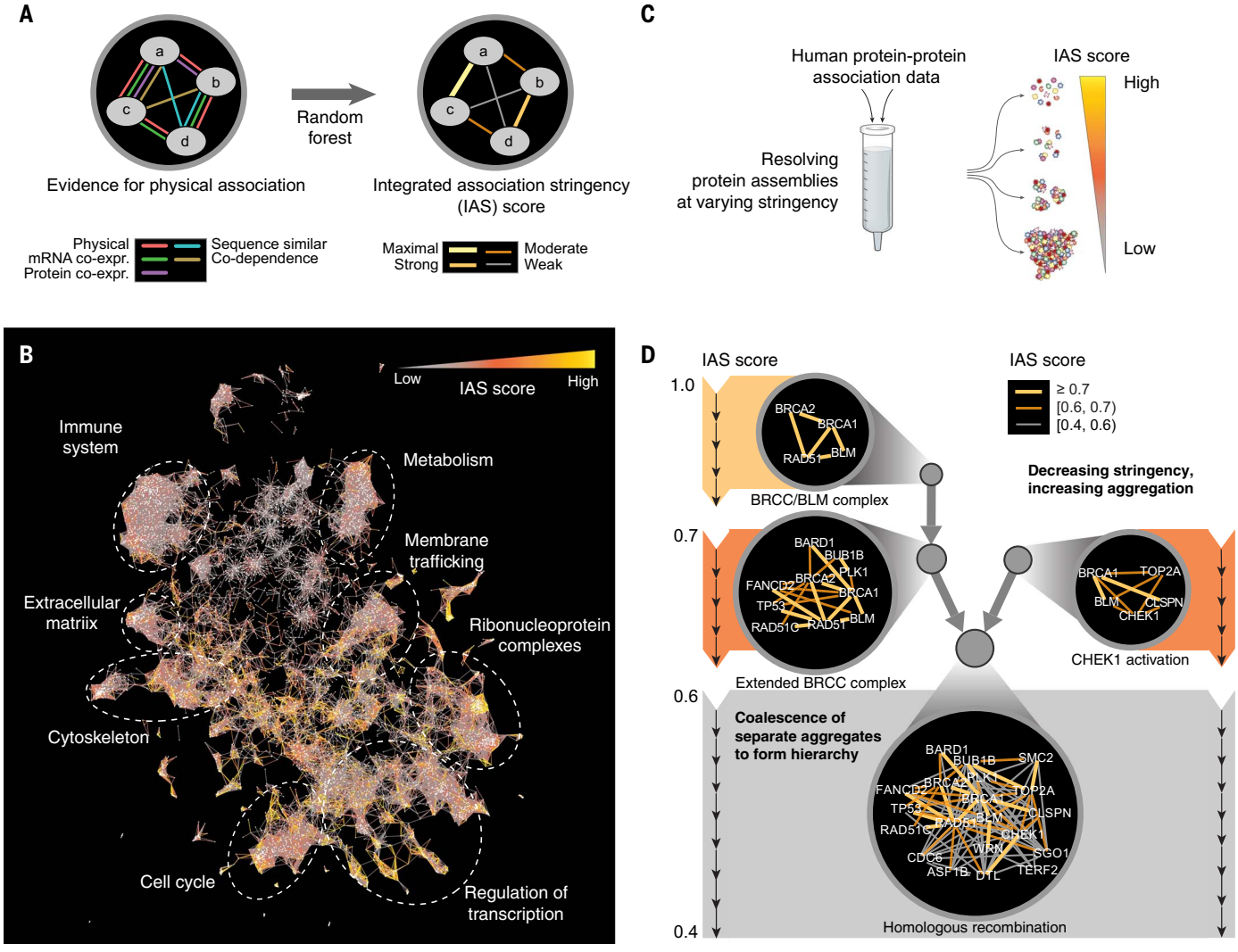


Fig. 2. Integration and multiscale organization of protein networks. (**A**) Types of network evidence are combined in a machine learning framework (random forest) to determine an integrated association stringency (IAS) score among protein pairs. (**B**) Visualization of overall IAS network structure highlighting major large-scale systems (text labels and dashed circles). (**C**) Process of multiscale community detection, whereby protein systems of increasing size are discovered as the threshold IAS score is progressively lowered. (**D**) As the IAS threshold decreases (top to bottom), strongly associated protein systems are detected and then expanded to include proteins with moderate-to-weak associations. Each circle indicates a protein system; edge colors (yellow, orange, gray) indicate decreasing stringencies of association.

Because the IAS score forms a continuous metric of physical association, robust assemblies of interacting proteins can be identified at specific IAS thresholds (Fig. 2C). To catalog these assemblies over the entire range, we performed a hierarchical community detection analysis (materials and methods). By this procedure, the IAS threshold was initialized to its maximum (most stringent) value and then iteratively relaxed to lower (more permissive) values; at each step, progressively larger and less stringently associated protein assemblies were detected. Small assemblies of strongly associating proteins were subsumed within larger assemblies as the IAS threshold decreased, creating a hierarchy. For example, four homologous recombination (HR) proteins—BRCA1, BRCA2, RAD51, and BLM—associated under a stringent IAS threshold, reconstituting a variant of the BRCC complex (*61*) (which we called the "BRCC/BLM complex"). As the IAS threshold decreased, this complex expanded to encompass a larger group incorporating BARD1, BUB1B, FANCD2, PLK1, RAD51C, and TP53 (the "extended BRCC complex"), which then consolidated with a distinct assembly containing BRCA1, BLM, TOP2A, CHEK1, and CLSPN ("CHEK1 activation") and other proteins to form a supercluster of 20 proteins broadly involved in HR (Fig. 2D). Rather than merely supply an inventory of HR factors, however, the hierarchical analysis reveals how larger protein assemblies are organized from smaller ones.

When applied to the entire IAS network, hierarchical community detection resulted in identification of a total of 2338 protein assemblies (data S2). Notably in this analysis, proteins were allowed to cluster in multiple distinct assemblies when such affiliation was supported by the interaction data. For example, β-catenin (CTNNB1) has well-established pleiotropic functions, with separate roles in the β-catenin destruction complex and adherens junction (*62*); accordingly, its interaction patterns placed it in distinct assemblies corresponding to each of these two aspects (fig. S1, D and E). Because small assemblies tended to correspond to protein complexes and signaling scaffolds while larger groups more closely represented broad cellular processes, we

adopted the general name "protein systems" to describe entities at any scale.

Identifying recurrently mutated systems at multiple scales

We next sought to identify which systems were under pressure for recurrent mutations in cancer. Given a list of systems, each consisting of a set of proteins, a straightforward analysis might be to test for statistical enrichment of mutations within each system individually (*63*). However, for overlapping or nested systems, such tests are confounded because mutations that affect one system create correlated enrichments in other systems with overlapping components. Our goal was to determine whether the gene mutation frequencies in a cancer cohort were best explained by separate selection pressures on individual genes, or more parsimoniously by pressure on a small set of systems. For this purpose, we developed a unified statistical model of mutation, called HiSig, to determine a set of systems that optimally explains the observed mutation pattern for all genes while seeking to minimize the number of systems required for such explanation (Fig. 3A and fig. S2). The HiSig model accounts for overall mutation burden, protein length, and other factors when analyzing mutation patterns to arrive at an expected mutation frequency but does not attempt to model phenotypic impacts of mutation (materials and methods). The significance of mutational enrichment was evaluated using permutation testing at a fixed false discovery rate (FDR) (materials and methods).

As an example of mutational analysis at the systems level, we again turned to HR and its subsystems as discussed above. HR defects can be caused by driving loss-of-function mutations in BRCA1, BRCA2, and related proteins in an effect that has been called "BRCAness" (*64*, *65*). Consistent with this expectation, HiSig identified significant mutational pressure on the "BRCC/BLM complex" and "extended BRCC complex" in breast and ovarian carcinomas, two tumor types for which BRCAness had been well studied (*64*, *65*) (Fig. 3B). Although mutations in individual genes encoding the BRCC/BLM complex were rare and failed to achieve strong statistical significance, with <3% for each gene in breast tumors (*66*), mutations in the four BRCC/BLM genes converged for an aggregate mutation frequency of 7%, exceeding random expectation [95% of confidence interval (CI): 0.4 to 3.6%; P = 0.0008, log-normal distribution]. Thus, the systems-level analysis was able to recognize evidence of a well-known systems-level effect, BRCAness, despite the lack of strong signals from individual genes.

HiSig also identified mutational selection for the same HR systems in bladder urothelial carcinoma (Fig. 3B, 18%, versus 95% of CI of random expectation: 1.2 to 11.1%; P = 0.0029, log-normal distribution), a tumor type for which significant mutation rates of individual BRCAness genes had not previously been identified (*67*), perhaps because this cancer has a higher background mutation burden. To corroborate this finding, we performed CRISPR-Cas9 disruptions to genes encoding the "extended BRCC complex" in bladder cancer cells

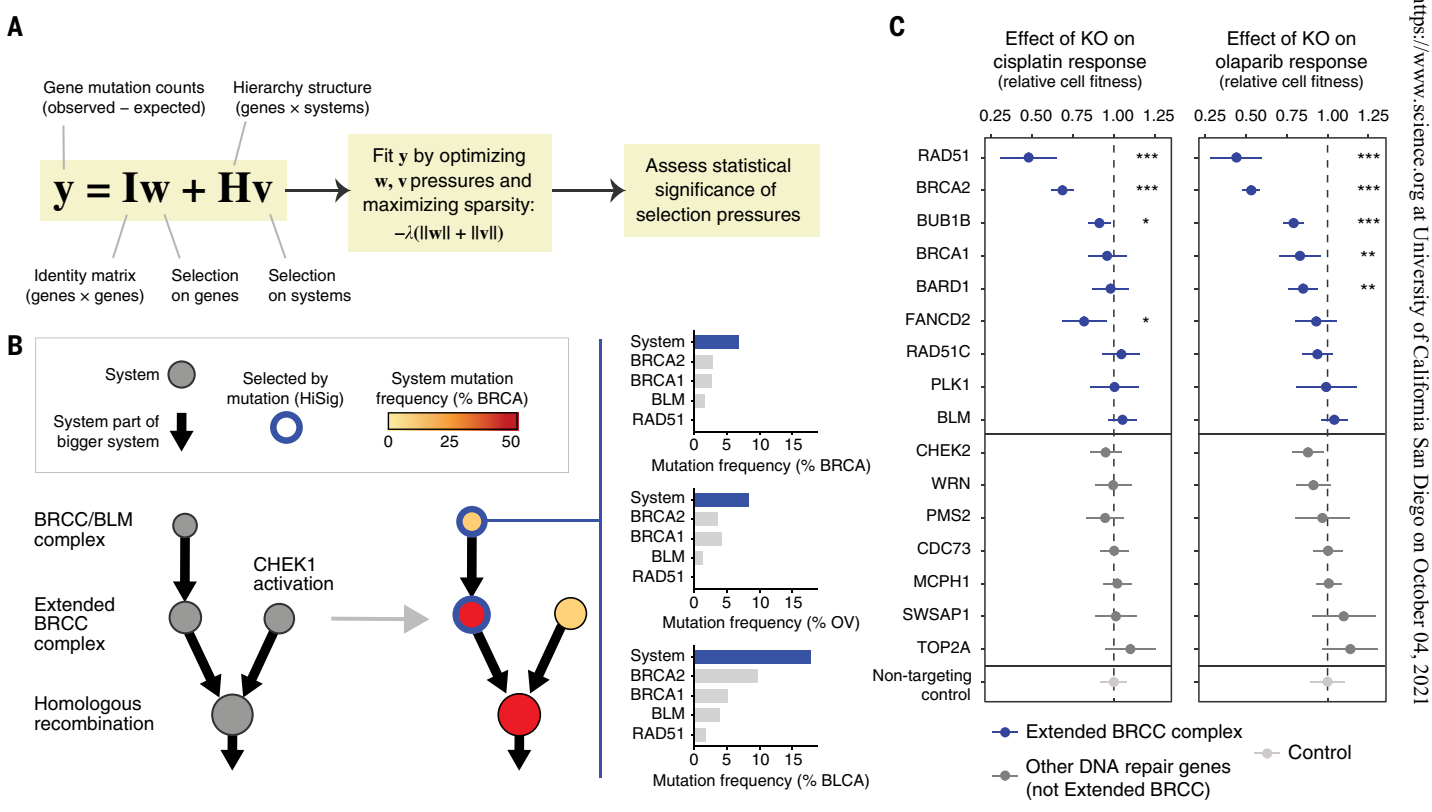


Fig. 3. Convergence of mutations within protein systems. (**A**) HiSig model for identification of significantly mutated systems. A regularized lasso regression model is fit to the observed mutation counts of genes in a cohort (**y**) by adjusting mutation pressures (**w**, **v**) on genes and systems, respectively. The λ term penalizes models that lack sparsity (i.e., placing pressures on many genes or systems). Systems harboring frequent mutations that are not well explained by pressures on any single gene will have positive pressures **v** after solving the penalized regression. (**B**) Hierarchy of systems related to homologous recombination, extracted from the IAS network and analyzed for mutation frequency in breast cancer (BRCA, red color gradient on nodes). Blue node borders indicate significantly mutated systems selected by HiSig. Bar charts at right show system- and gene-level mutation frequencies for the "BRCC/BLM complex" in breast, ovarian, and bladder cancers (BRCA, OV, BLCA). (**C**) Effects of CRISPR-Cas9 gene knockouts (rows) on the response to cisplatin (left) or olaparib (right) in the SW1710 bladder cancer cell line. Data are means ± 95% CI. The cell fitness resulting from each knockout is compared to that of nontargeting controls; *P < 0.05, **P < 0.01, ***P < 0.001 (t test with Benjamini-Hochberg multiple-testing correction).

and scored each for sensitivity to olaparib or cisplatin, which are phenotypes indicative of HR deficiency and thus BRCAness (*68*). Six gene disruptions to this system caused sensitization to one or both drugs, whereas none of a control set of gene disruptions produced significant sensitization (Fig. 3C).

A cancer systems map integrating 13 tumor types

We used HiSig to analyze somatic mutation patterns from 6251 exomes, representing 13 tumor types with sufficient sample numbers and mutation burdens (*37*): bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), colorectal adenocarcinoma (COAD), glioblastoma (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), skin cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), and uterine corpus endometrial carcinoma (UCEC). HiSig analysis identified 319 systems with evidence for mutational selection in one or more cancer types. To unify these systems into a pan-cancer hierarchy, we added sufficient higher-order systems to link them, yielding a final map of 395 protein systems that we called NeST (Nested Systems in Tumors, http://ccmi.org/nest/; Fig. 4A and data S3). Systems were named by a team of in-house curators using a combination of expert knowledge, literature analysis, and gene set enrichment (materials and methods). In general, NeST organizes numerous small systems with tissue-specific mutation patterns within a few large systems relevant to greater numbers of tumor types (e.g., cell cycle progression, immune systems, and transcription; Fig. 4A), providing a systematic reconstruction of hallmark processes altered in cancer (*69*).

To systematically validate the collection of recurrently mutated NeST systems, we examined whole-exome sequencing data from independent patient cohorts representing nine tumor tissue types, for a total validation set of 4077 tumor exomes (materials and methods). Per tissue type, we observed that 37 to 66% of NeST systems were also recurrently mutated in the tissue-matched validation cohorts (Benjamini-Hochberg FDR < 0.1 by HiSig; Fig. 4B, fig. S3A, and table S3). This range of validation percentages was comparable to that of validating individual mutated genes and was significantly better than achieved if the tissue types used for discovery and validation were decoupled and randomized (e.g., pairing a BRCA validation cohort with the mutated systems discovered for LUSC; fig. S3A).

The number of recurrently mutated systems in each cancer type ranged from 11 (SKCM) to 84 (BLCA) and was anticorrelated with the genome-wide mutation burden of each type (Spearman $\rho = -0.59$, $P = 0.03$; fig. S3B), reflecting the difficulty of demonstrating significant mutation rates in very highly mutated cancers (*70*). Notably, the size distribution of recurrently mutated systems (as determined by HiSig) was similar to that of all systems (Fig. 4C), which suggests that selective pressures operate at all scales of cell biology rather than at one resolution only, such as protein complexes or narrowly defined pathways. Application of HiSig to alternative hierarchies of human protein systems curated from literature (Gene Ontology; fig. S3C) identified significantly fewer mutated systems, demonstrating the specific value and relevance of NeST as a resource for the study of cancer. We also repeated HiSig analysis on the IAS-derived hierarchy systems, considering transcription-altering copy number alterations (CNAs) rather than point mutations and indels (see supplementary text). This analysis found 20 additional systems enriched for CNAs (table S4), including a cyclin-containing system containing CDKN2A (p16/ARF) and a transcriptional regulator complex containing EP300 and CREBBP (fig. S4).

Comparing NeST to our recent comprehensive literature survey of published cancer pathways (*71*), we saw that 40% of significantly mutated systems (129/319) recapitulated an established cancer mechanism (Fig. 4D and table S5). Examples of these 129 established systems included "PIK3-cyclin signaling," which integrated mutations in subunits of the PI3K holoenzyme (PIK3CA, PIK3R1, PIK3R2) with mutations in the downstream oncoprotein cyclin D1 (CCND1), as well as "SMAD-TGFβ signaling," a system encompassing mutations to SMAD transcription factors, the TGFβ receptor TGFBR1, and functionally related proteins. Although both of these complexes contained individual cancer proteins with high mutation rates (PIK3CA, SMAD4), mutations in other proteins in these complexes (PIK3R1, SMAD3) had been too rare to meet the significance thresholds of previous analyses (*37*) despite having functionally validated roles in cancer (*72*).

Among the remaining 60% of systems (190/319), 75 recapitulated well-known cellular components that had not been previously associated with recurrent cancer mutations, whereas 115 were best described as novel protein assemblies (materials and methods). Below, we investigate several of the discoveries in greater detail with biophysical and functional assays, exemplifying the use of NeST to generate biological hypotheses and inform their investigation.

A PIK3CA-actomyosin assembly regulating the PI3K/AKT pathway

HiSig identified recurrent mutations in a novel variant of the actomyosin complex, a cytoskeletal component regulating cell shape, motility, and membrane organization (*73*, *74*). Multiple direct physical interactions from the AP-MS data newly linked actomyosin proteins with the p110α subunit (PIK3CA) of phosphatidylinositol 3-kinase (PI3K), forming a system we named the "PIK3CA-actomyosin complex" (Fig. 5A). This system was under significant mutational pressure in five cancer types (BLCA, 36%; BRCA, 38%; COAD, 42%; HNSC, 29%; STAD, 32%), integrating frequent mutations in PIK3CA with less common mutations in numerous actomyosin proteins including nonmuscle type II myosins (NM2 proteins MYH9 and MYH10; Fig. 5B). This mutation pressure was seen to validate in BRCA and COAD independent cohorts (fig. S5A and table S3; large secondary cohorts were unavailable for BLCA, HNSC, or STAD). Although cytoskeletal remodeling is a downstream effect of PI3K/AKT signaling (*75*, *76*), actomyosin proteins had not been previously shown to physically associate with PIK3CA, nor had their mutations in cancer been widely studied.

In support of the physical association, we found that actomyosins bind specifically to PIK3CA and generally not to other cancer protein baits assayed by our AP-MS experiments (Fig. 5C). We were also able to validate the association by proximity ligation assay, demonstrating that PIK3CA and MYH9 colocalize in CAL-33 cells (Fig. 5D). To determine more precisely where this colocalization occurs, we performed superresolution microscopy and found that both molecules associated in small, membrane-proximal puncta (Fig. 5E and fig. S5C).

We next investigated the functional consequences of PIK3CA-actomyosin physical interaction by assaying canonical readouts of PI3K signaling, phosphorylation of AKT (pAKT) and of ribosomal protein S6 (pS6), in response to NM2 inhibition by blebbistatin (*77*). Blebbistatin treatment increased the levels of pAKT and pS6 in CAL-33 cells, but this effect was suppressed by additional treatment with alpelisib, an FDA-approved PIK3CA inhibitor (Fig. 5F and materials and methods). Conversely, in SCC-25 cells, where PIK3CA-actoymosin interactions were not observed, blebbistatin treatment did not affect either readout (fig. S5D). PIK3CA inhibition by NM2 was further supported by reverse-phase protein array (RPPA) data for 899 cell lines from the Cancer Cell Line Encyclopedia (CCLE) (*78*), which showed that cells harboring MYH9 or MYH10 mutations have significantly elevated pAKT relative to cells lacking such mutations (Fig. 5G). Furthermore, genomic alterations in *MYH9* and *PIK3CA* were mutually exclusive, a sign of functional dependency ($P < 0.05$; Fig. 5H). Together, these results suggest that the actomyosin complex directly inhibits PIK3CA signaling.

Beyond the example of the PIK3CA-actomyosin complex, we noted that many of the systems in NeST were driven by new protein interactions from our AP-MS screens (112 systems; fig. S6A

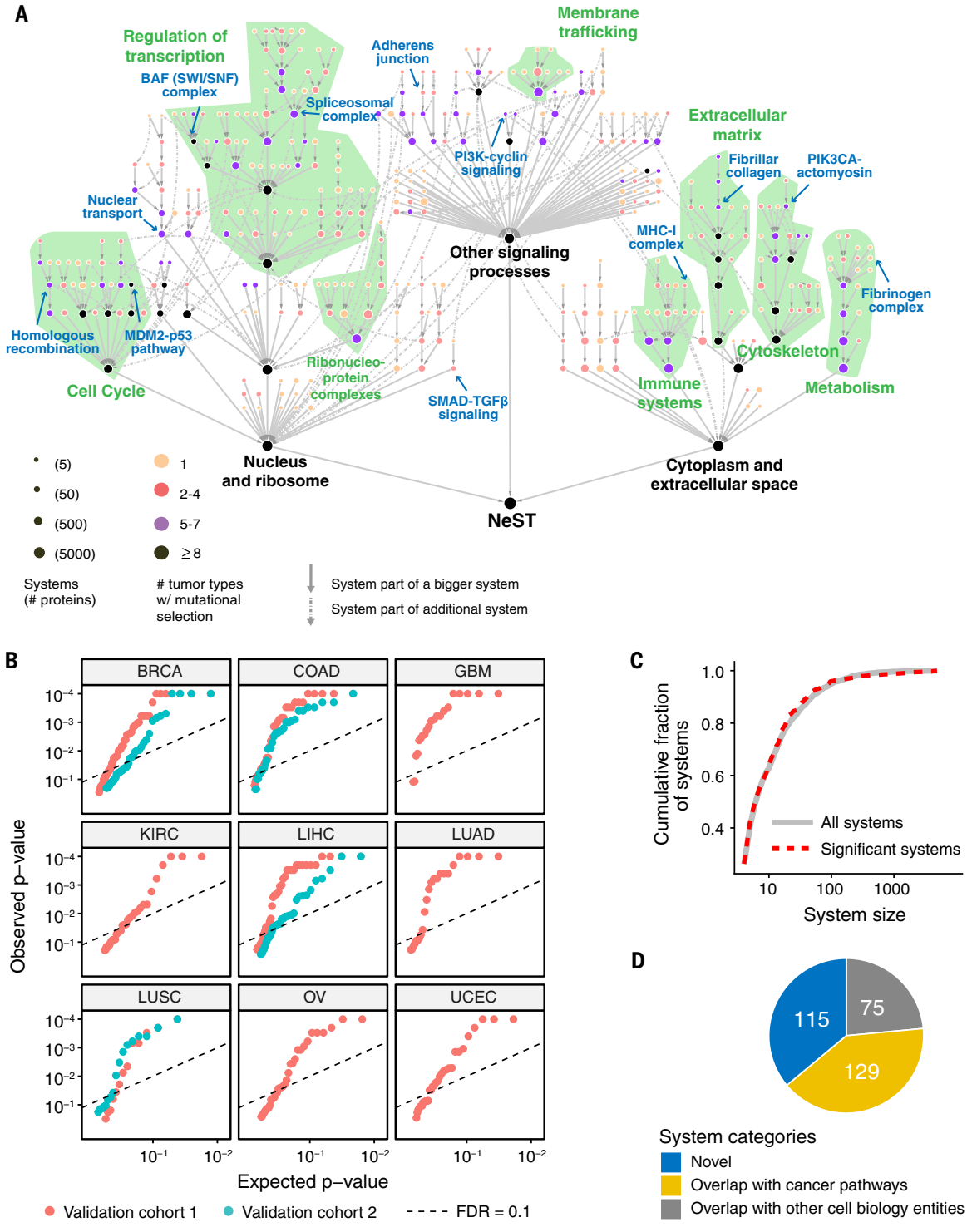


Fig. 4. Pan-cancer map of mutated systems (NeST). (**A**) NeST map assembled from the union of significantly mutated systems identified by HiSig in each of 13 tumor types. Nodes represent protein systems, with color indicating number of cancer types for which a system is mutated. Gray arrows between systems ($s \rightarrow t$) show hierarchical containment of the first system (s) in the second (t). For systems contained by two or more larger ones (representing pleiotropy), these additional containment relations are shown as dashed arrows. Green text and regions show correspondence to the large-scale components of the IAS network in Fig. 2B. (**B**) Recapitulation of significantly mutated systems in independent validation cohorts. For each system (points), the qq-plots show HiSig *P* values of recurrent mutation as determined in the validation cohort versus the values expected by chance. The 3 × 3 plot grid shows validation in nine tissue types. When multiple validation cohorts are available for a tissue type, values for the second validation cohort are shown in cyan. Dashed line shows a 10% FDR cutoff. (**C**) Numbers of proteins per system for significantly mutated systems versus all systems extracted from the IAS network. (**D**) Composition of NeST systems by relationships to prior knowledge of cell biology and cancer pathways.

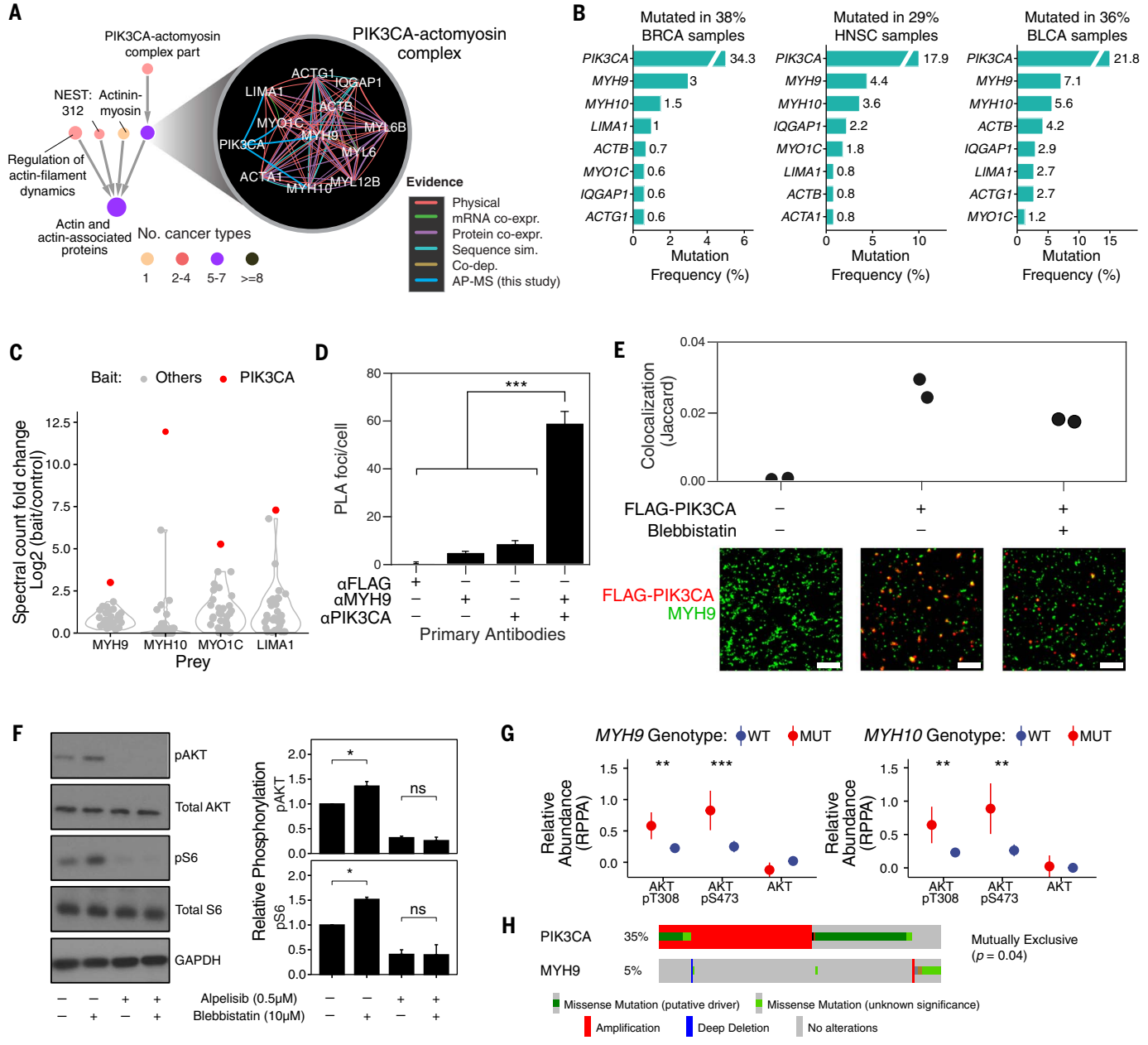


Fig. 5. A PIK3CA-actomyosin assembly regulating the PI3K/AKT pathway. (**A**) Interactions defining the PIK3CA-actomyosin complex and the context of this system in the NeST hierarchy. (**B**) Mutation frequencies of genes in this system (shown are the top eight frequently mutated genes in each cohort). (**C**) Signal (spectral count fold change versus control) of actomyosin proteins (MYH9, MYH10, MYO1C, LIMA1) in AP-MS experiments with cancer protein baits. Red dots represent signal as interactors of PIK3CA in the CAL-33 cell line; gray dots show corresponding data for other cancer proteins. (**D**) Per-cell counts of proximity ligation foci when probing for PIK3CA and MYH9 in CAL-33 cells; significantly more foci are observed when probing for both proteins. ***$P < 0.001$ (one-tailed Student's t test). (**E**) CAL-33 cell lines expressing FLAG-PIK3CA (as used in AP-MS experiments) were probed with antibodies to FLAG and MYH9 and imaged by DNA-PAINT. Substantial PIK3CA/MYH9 colocalization is observed in small, membrane-proximal puncta. Representative subfields are shown below; scale bar, 500 nm. (**F**) CAL-33 cells were treated with DMSO, blebbistatin (10 μM), and/or alpelisib (0.5 μM), harvested, and analyzed by immunoblotting with the indicated antibodies. Left: Immunoblot representative of three independent experiments. Right: Quantification of immunoblot signal. ($N = 3$; a representative image is displayed.) Signals were normalized relative to total protein and loading control. Data are means ± SE. *$P < 0.05$ (one-way ANOVA); ns, not significant; $N = 3$ for all measurements. (**G**) Comparison of the RPPA-measured abundance of indicated protein species in cancer cell lines with mutant (red, $N = 108$) versus wild-type (blue, $N = 791$) *MYH9* (left) and mutant (red, $N = 82$) versus wild-type (blue, $N = 817$) *MYH10* (right). **$P < 0.01$, ***$P < 0.001$ (Wilcoxon rank sum test). (**H**) Genomic alterations of *PIK3CA* and *MYH9* in the TCGA-HNSC cohort show a pattern of mutual exclusivity.

and table S6). For example, new AP-MS experiments showed mitogen-activated protein kinase (MAPK) proteins copurifying with HLA-A to form a new protein system (fig. S6B), which suggests that known crosstalk between the MAPK pathway and major histocompatibility complex internalization may be mediated by direct MAPK-HLA interaction (*79*). Enrichment for the new interactions was greatest in systems mutated in BRCA (31 systems) and substantial in HNSC tumors (25 systems), the two tumor types used for AP-MS data generation; such

enrichment was also observed in other tumor types (fig. S6C).

Destabilizing mutations in collagen systems promote tumor progression

HiSig detected selection pressure on a system consisting of 11 fibrillar collagen proteins in seven cancer types, with particularly high mutation frequencies for SKCM (71%), LUSC (49%), and LUAD (49%) (Fig. 6, A to C). Recurrent mutations in this system were validated in independent cohorts of lung cancers (fig. S7B and table S3). Fibrillar collagens are the most abundant proteins of the extracellular matrix (*80*, *81*), the structure and composition of which can promote a tumor microenvironment conducive to invasive growth (*82*, *83*). Although differential expression of collagen proteins between tumor and normal tissues had been reported (*84*), thus far somatic mutations in collagens had not been associated with cancer, save for very rare tumor types (*85*, *86*).

Collagen proteins contain helical repeat regions rich in glycine and proline residues (Fig. 6D), and these domains are critical for proper folding of collagen into trimeric bundles (*80*). In SKCM and LUAD, we found that nonsilent mutations in collagens significantly tend to modify these glycine or proline codons, even when correcting for the abundance of C or G transversions in these cancers (*87*) (Fig. 6E). Computational modeling (*88*) predicted that the specific mutations observed in tumor samples strongly destabilize collagen proteins (Fig. 6F).

We therefore hypothesized that mutations of collagens may promote tumor progression by disrupting protein folding and thus organization of the extracellular matrix. To test our hypothesis, we allowed fibroblasts with or without G/S mutations in collagen genes (Tu To or HFF-1 cell lines, respectively; see supplementary text) to deposit matrix in culture vessels. After decellularizing this matrix, we seeded it with A549 LUAD cells and observed expression of Ki67, a standard marker of cell proliferation (Fig. 6G and materials and methods). We found greatly increased cell proliferation specifically in the collagen-mutant matrix, which suggests that collagen mutations can perturb the cellular environment to favor tumor growth (Fig. 6, H and I). We also compared matrix deposited by HFF-1 fibroblasts that were engineered to overexpress COL1A1 with and without the G281S mutation, and we similarly observed increased proliferation of A549 cells on the mutant matrix (fig. S7, B and C). In yet further supporting analysis, we found that collagen mutations are associated with increased metastasis in mouse xenograft models (*89*) (Fig. 6J and fig. S7D). Together, these data support a role for collagen mutations in disrupting the tumor microenvironment to favor cancer progression.

Protein systems as clinical biomarkers

We examined the extent to which NeST systems could serve as prognostic biomarkers, nominating those systems in which several criteria were met: (i) HiSig identifies the system as under mutational selection in a given tumor type; (ii) significant differences in survival are observed between patients with mutations in the system and those without; and (iii) the survival association cannot be trivially explained by mutations in any single gene in the system. By these criteria, and with additional correction for mutational burden (materials and methods), we identified a total of 25 prognostic associations (FDR < 0.3; Fig. 7A and table S7). Among the prognostic systems, we observed poor prognoses in GBM tumors associated with mutations in a system integrating PI3K components (PIK3CA, PIK3R1, PIK3CB) with GRB2 and its binding partner GAB1 ["proximal receptor tyrosine kinase (RTK) signaling"; Fig. 7, B and C]. Although mutations in PIK3CA and PIK3R1 were each weakly associated with the progression-free survival time, integrating these mutations resulted in a stronger prognostic effect (Fig. 7C). Another prognostic system, representing histone functions in the DNA damage response, brought together 15 proteins that individually had low mutation frequencies (<3.5%) and no prognostic association. However, integrating all of these mutations led to an unexpectedly high rate of mutation (14.3%) with poor prognosis in COAD tumors (Fig. 7, D and E).

The preceding analyses demonstrate the advantages of interpreting cancer genomes through a map of protein systems. However, lists of cancer genes remain very useful for clinical applications, such as the design of diagnostic gene panels. In this respect, we considered that the recurrently mutated systems identified in NeST might include many genes not implicated in current cancer gene panels or databases (*37*, *72*, *90*, *91*). The designation of a NeST system requires mutations in multiple constituent genes that cannot be better explained by any other system (HiSig approach; materials and methods). Thus, we conservatively nominated two genes per system having the highest relative mutation frequencies, yielding a nonredundant catalog of 548 NeST cancer genes (table S8). This NeST systems gene catalog covered 101 of 179 significantly mutated genes (SMGs) reported in previous TCGA cancer-driver analyses for the 13 tissues studied (*37*); the remaining 78 SMGs were not components of a larger system under selection. Moreover, our analysis nominated 447 genes not previously associated with cancer by their somatic mutations (Fig. 7F). As an ensemble, mutations in NeST cancer genes were predicted as less deleterious than mutations in TCGA SMGs, while significantly more deleterious than mutations in background genes not in either category (fig. S8). Conversely, NeST cancer genes were differentially expressed between tumor and normal samples to a greater extent than previously reported SMGs (Fig. 7G) and, individually, their expression levels offered greater prognostic value (Fig. 7H). Additionally, orthologs of both NeST genes and SMGs were significantly enriched for tumor growth effects in transposon-based forward genetic screens in mice (Fig. 7I and materials and methods).

Discussion

In their original description of the "Hallmarks of Cancer," Hanahan and Weinberg (*92*) predicted that "complexities of cancer, described in the laboratory and clinic, will become understandable in terms of a small number of underlying principles." Independently, in his seminal vision regarding the future of molecular biology (*93*), Alberts described the cell as a "collection of protein machines" where "each of these protein assemblies interacts with other large complexes of proteins in an elaborate network." Our joint study, which includes the accompanying manuscripts (*46*, *47*), represents a systematic effort to combine these two visions: organizing the diverse mutation patterns of cancer into underlying principles represented by a multiscale map of complexes and larger protein machines. This map, NeST, derives from an end-to-end data generation and analysis pipeline consisting of protein network collection, multi-omics integration, structural and predictive modeling, and model visualization (Fig. 1B). It serves not only as a useful abstraction for understanding cancer cell biology, but as a tool to discover new protein systems and their association with cancer.

Defining a collection of biological systems requires recognizing physical and functional boundaries between these entities, a process with inherent ambiguities. Even when assigning mutations to genes, the gene boundary might or might not include introns, alternatively spliced exons, promoters, or enhancers. Likewise, the systems in NeST are defined on the basis of evidence that the proteins are densely connected by interactions of a certain stringency. One source of ambiguity concerns the parameters used for community detection. In constructing NeST, we optimized these parameters with respect to the ability of systems to explain mutation rates, although other means of defining systems boundaries could be explored. Nonetheless, we found that most systems could be validated by mutation patterns in second cohorts (Fig. 4B) and many of the associated genes had independent functional evidence (Fig. 7, G to I), supporting the relevance of the current pipeline.

Comprehensive interaction screens directed to a panel of 61 cancer proteins in multiple BRCA and HNSC cell lines (*46*, *47*) contributed greatly

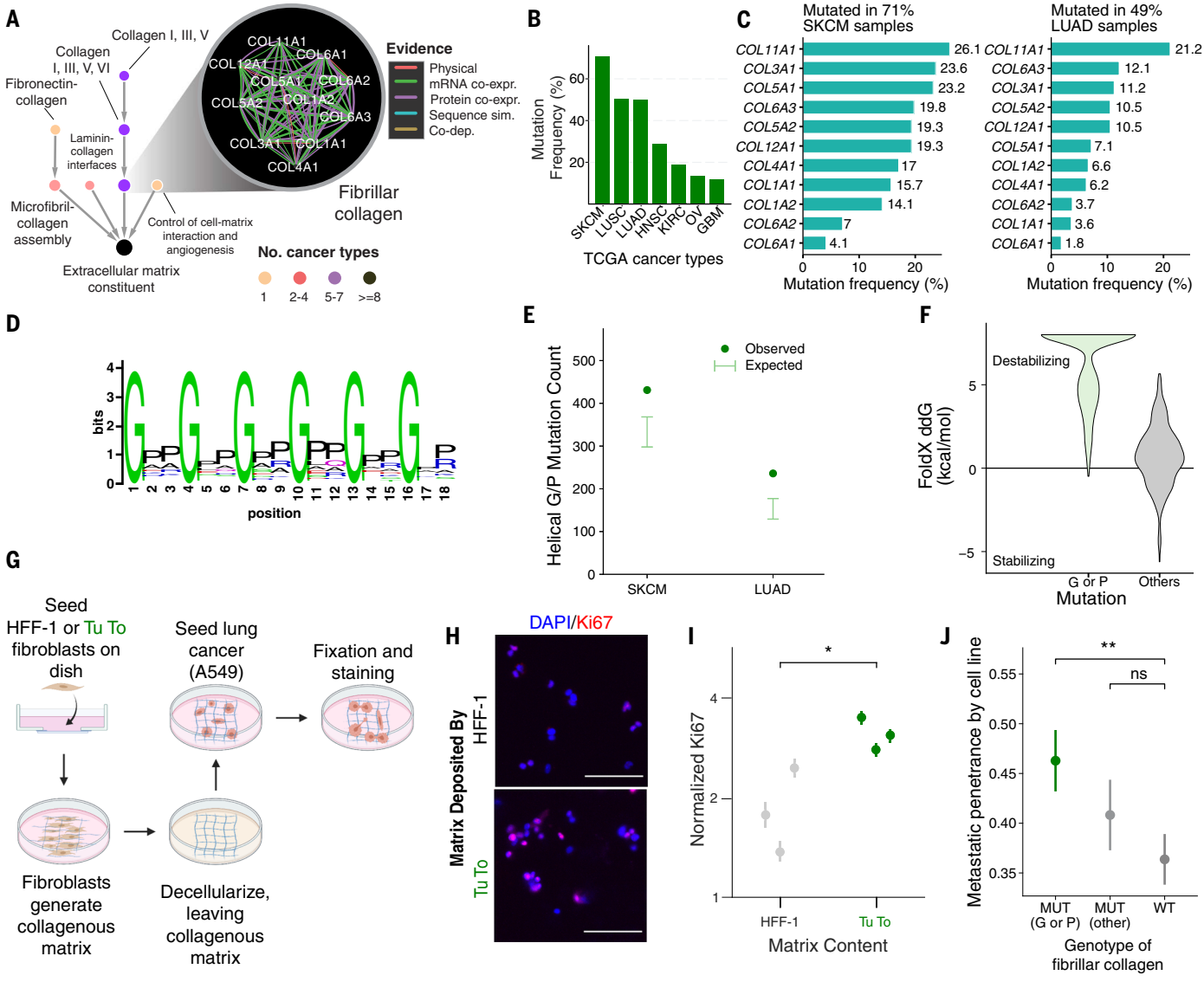


Fig. 6. Destabilizing mutations in collagen systems promote tumor progression. (**A**) Interactions defining the fibrillar collagen system and the context of this system in the NeST hierarchy. (**B**) Mutation frequencies of this system across TCGA tumor cohorts in which the system was identified by HiSig. (**C**) Mutation frequencies per collagen protein in TCGA melanoma (SKCM) and lung adenocarcinoma (LUAD) cohorts. (**D**) Amino acid sequence tendency of the collagen triple-helix repeats. Image was created on the basis of the PFAM domain PF01391. (**E**) Analysis of collagen mutations in SKCM and LUAD cohorts. Error bars denote 95% CI based on a binomial distribution according to the background mutation rates of collagen genes. (**F**) Protein stability change upon point mutations on the glycine/proline positions (G/P) versus other positions in the triple-helix repeats, predicted by FoldX 5.0 (materials and methods) (*88*). (**G**) Schematic of matrix deposition assay. (**H**) Immunofluorescence images of A549 cells on matrix deposited by the indicated cell lines. Scale bars, 100 μm. (**I**) Quantification of immunofluorescence across three biological replicates (points) for each condition. Error bars indicate 95% CI. For HFF-1, N = 445, 860, 774; for Tu To, N = 2316, 2114, 1810. *P < 0.05 (one-tailed Student's t test). (**J**) Associations between fibrillar collagen mutation status and the metastatic penetrance of cancer cell lines (*89*) (N = 146, 123, 219). **P < 0.01 (Wilcoxon rank sum test).

to the discovery of protein systems in this study. In general, physical protein interaction strongly informed IAS scores (fig. S1, A and C), and the specific baits chosen for AP-MS were known cancer proteins expressed in cancer cells. The new interaction data also contributed to protein systems mutated in other cancer types besides HNSC and BRCA (fig. S6). Similar observations have emerged from projects such as ENCODE and others, in which analysis of selected cell lines has defined transcriptional regulatory networks with broad relevance across tissues and diseases (*94*, *95*). These findings suggest that additional interaction screens in a modest number of cellular contexts might achieve reasonable coverage of protein complexes driving most cancer types.

Beyond recapitulating known cancer mechanisms, the NeST map identifies a number of recurrently mutated systems that do not overlap significantly with previously published cancer pathways (*71*) (Fig. 4D and table S5). Such systems might be newly identified for one or more of the following distinct aspects of the present study: (i) the expanded content and scope of protein-protein associations in the input network, which integrates new AP-MS experiments targeting cancer protein interactions with a compendium of diverse multi-omics data; (ii) the explicit identification of distinct protein assemblies in the network (systems) over a continuum of scales;

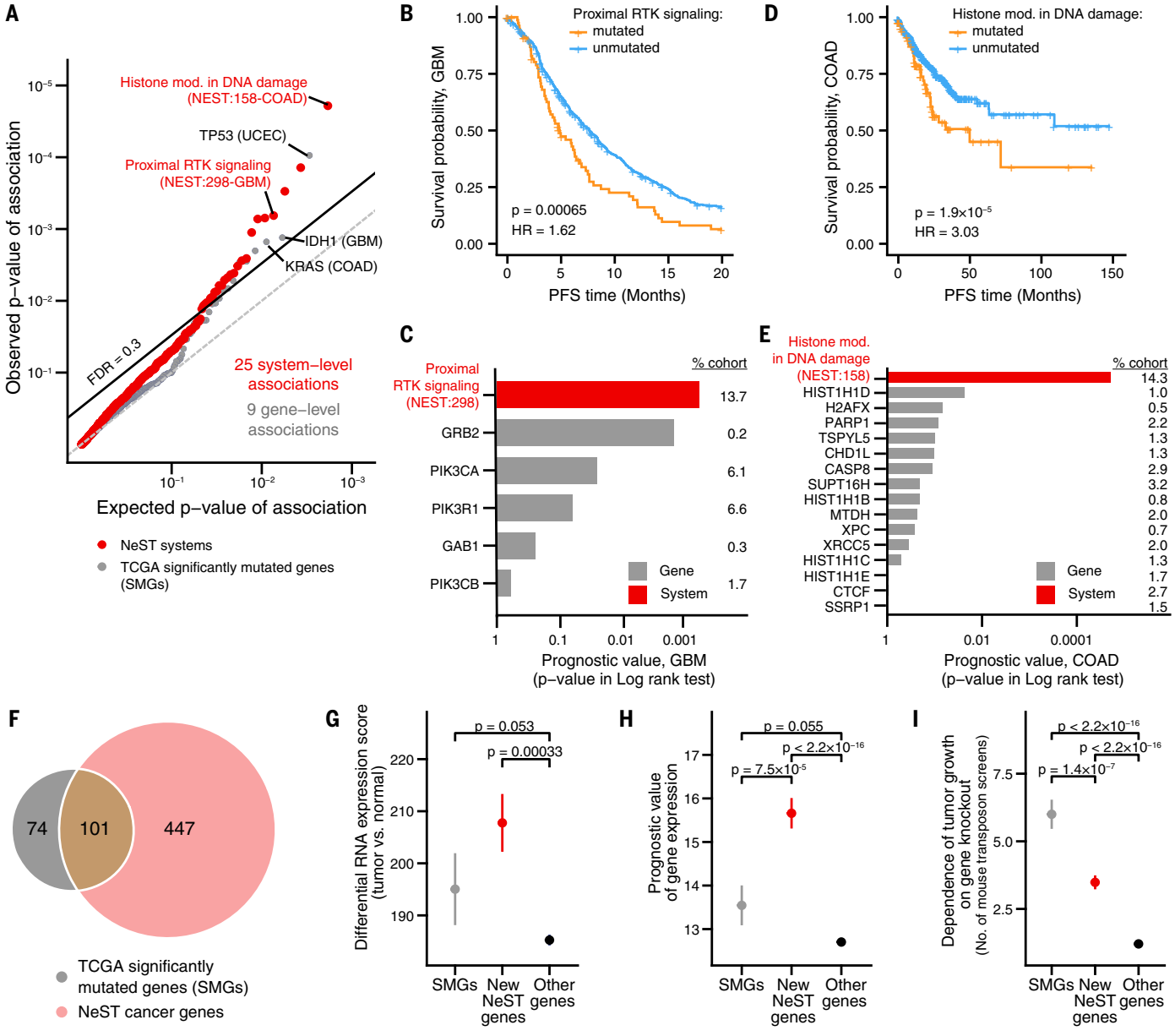


Fig. 7. Protein systems inform clinical markers and lists of cancer genes. (**A**) Significance of prognostic association between biomarker status (mutated/unmutated) and progression-free survival (PFS). Red points represent systems; gray points represent genes. Observed *P* values are plotted versus those expected when performing the same number of tests at random. The dashed line indicates a cutoff of FDR = 0.3. The relevant tissue type for each association is shown in parentheses. (**B** to **E**) Selected NeST systems associated with prognosis. [(B) and (C)] NeST:298, "Proximal RTK signaling." [(D) and (E)] NeST:158, "Histone modification in DNA damage response." Kaplan-Meier curves [(B) and (D)] stratify tumors according to the mutation status of each system. Bar charts [(C) and (E)] indicate significance of association (log-rank test) between mutation status and PFS for systems (black bars) versus individual genes in those systems (gray bars). Numbers on the right indicate the percentage of patients in that cohort with mutations in the indicated gene. (**F**) Venn diagram showing the common and distinct genes identified by significantly mutated systems in NeST (pink) versus TCGA PanCancer analysis (gray) (*37*). (**G** to **I**) Functional support for NeST cancer genes: (G) RNA-seq tumor-normal differential expression in TCGA (*128*). (H) Prognostic value of mRNA expression in TCGA (*129*). (I) Number of times a gene has been identified in independent cancer genetic screens in mice (*130*). Whisker plots show means ± SE. **P* < 0.05 (one-sided Wilcoxon rank sum test).

and (iii) the scoring of mutation significance at the level of systems rather than genes, in a manner that is distinct from approaches such as network propagation or gene set enrichment.

NeST serves as a resource for tumor genome interpretation and prognosis. The search for a well-defined and interpretable set of causative mutations and prognostic markers has been hindered by the "long tail" of cancer mutations (*1*) (Fig. 1A). This distribution of mutations across genes simultaneously makes candidate mutations numerous and statistical tests to assert their significance underpowered. By rationally consolidating mutations into fewer classes that maintain functional association, fewer statistical tests are required and each class contains more samples. Indeed, we are able to identify more prognostic systems than prognostic gene-level mutations (Fig. 7A). As interactome-mapping and patient-sequencing efforts continue over the next years, we expect

to improve our identification of protein systems and their associations with cancers.

Finally, although this study has focused on the analysis of the cancer genome, cancer is by no means the only disease in which diverse mutations converge on a narrower set of common processes and phenotypes. The concept of a hierarchy of protein systems, likewise, is not specific to cancer cells or their cell types of origin. The analysis presented here thus serves as a model to understand the large array of other diseases influenced by complex genetic alterations to somatic cells or to the germline.

Materials and methods

Targeted interrogation of cancer protein networks by AP-MS

Experimental acquisition of AP-MS data for 61 cancer proteins (baits) is described in (*46*, *47*). Briefly, protein spectral counts were determined by MaxQuant (*96*) and used for PPI confidence scoring by two complementary algorithms, SAINTexpress (*97*) and CompPASS (*98*, *99*). Outputs of these two algorithms were combined into a unified metric called the PPI score. We retained interactions above a PPI score cutoff (0.8 for breast cell lines, 0.9 for head and neck cell lines), resulting in 1722 nonredundant interactions discovered in one or more of six cell lines (table S1).

Acquisition of network data in major categories of experimental features

In addition to the new AP-MS data, we considered major types of protein pairwise associations recorded in public databases: physical protein-protein interaction, mRNA coexpression, protein coexpression, codependence, and sequence-based relationships. Data obtained under each of these feature categories resulted in a total of 127 individual features on human protein pairs (table S2), described below. As the namespace for human proteins, we used the symbols of protein-coding genes in the HUGO Gene Nomenclature Committee database (HGNC) (*100*), yielding 19,035 distinct proteins. We converted gene and protein names from other namespaces (Ensembl, Entrez, Uniprot) to HUGO nomenclature during the preprocessing of networks.

Physical protein-protein interaction (PPI) networks. PPI networks were drawn from five recent data-driven studies: (i) BioPlex 2.0, based on immunoprecipitation of tagged proteins followed by detection of interactors by mass spectrometry (MS) (*30*, *98*); (ii) a network based on MS that considers interaction stoichiometry and cellular abundances of interactors (*52*); (iii) hu.Map, based on protein cofractionation followed by MS (*29*); (iv) Human Reference Interactome (the "HI-II-14" dataset), a comprehensive yeast two-hybrid network (*51*); and (v) BioGRID, a literature-curated PPI database (*49*), restricted to high-confidence PPIs (the "multi-validated" category) and excluding entries based on high-throughput techniques to reduce redundancy with the other four networks. Each PPI network was sparse, with direct interaction relationships for a minority of protein pairs. However, two proteins that fall in the same network neighborhood may be associated with one another through network proximity, even if they are not directly connected by physical interaction. Thus, each of the five networks mentioned above was processed using a network embedding algorithm, *node2vec* (*101*), which represents the network neighborhood of each node (i.e., protein) as a high-dimensional vector and calculates the Euclidean distance between these vectors for each node pair. We specified 32 node2vec dimensions with all other parameters set to default. Each PPI network thus contributed one sparse feature (the original network) and one dense feature (the embedded pairwise distances), for 10 features in this category in total.

mRNA coexpression networks. Because the amount of available mRNA coexpression data is currently higher and more diverse than for other types, we split these datasets into three categories: coexpression in cell line collections (36 features), coexpression in human tumor samples (28 features), and coexpression in healthy human tissues (28 features). RNA expression levels in cell lines were obtained from the CCLE (Cancer Cell Line Encyclopedia) and GDSC (Genomics of Drug Sensitivity in Cancer) projects (*102*, *103*). RNA expression levels for human tumor samples were obtained from TCGA studies documented in the cBioPortal repository (*104*). RNA expression levels for healthy human tissue samples were obtained from the GTEx (Genotype-Tissue Expression) data portal (*105*). Genes with very low expression levels (median TPM ≤ 1) in these data sets were excluded. Pearson correlation coefficients of expression were calculated for all remaining gene pairs across all samples in a collection, and also across subsets of samples from the same tissue of origin if the subset size was larger than 40.

Protein coexpression networks. In parallel to RNA expression, protein expression levels were obtained from CPTAC studies of breast and ovarian tumor samples (*54*, *55*) and a study of breast cancer cell lines (*53*). Pearson correlation coefficients were calculated for all pairs of characterized proteins.

Codependency networks. For each human gene, we accessed its gene dependency profile, the vector of fitness scores resulting from CRISPR knockdown of that gene across a panel of 485 cell lines. These data were obtained from DepMap project release 18Q3 (https://depmap.org/portal/) (*106*). We used these data to compute a codependency score for each gene pair as the Pearson correlation of the two corresponding gene dependency profiles. A global correlation profile over all cell-line tissue types as well as correlation profiles per tissue of origin were generated. We focused on seven tissues with more than 20 cell lines, resulting in a total of eight features.

Sequence-based relationships. For features based on sequence relationships, such as protein domain co-occurrence and interactions of orthologous genes in other organisms, we used the log-likelihood score for each evidence code documented in HumanNet 1.0 (*58*) including CE-CX, CE-GT, CE-LC, CE-YH, DM-PI, HS-DC, HS-GN, HS-PG, SC-CX, SC-GT, SC-LC, SC-MS, SC-TS, and SC-YH (defined in www.functionalnet.org/humannet/HumanNet.v1.evidence_code.txt).

Using network features to formulate the integrated association stringency (IAS) score

AP-MS interaction data and the 127 additional network features were used as inputs to a two-stage random forest regression model, trained to best recover the proximity of protein pairs in the Gene Ontology (GO). To compute this proximity we used the *GossTo* tool (*107*) to calculate the Resnik semantic similarity score (*108*) for all protein pairs, based on the GO Biological Process (BP) and Cellular Component (CC) branches as of May 2017. Only evidence codes related to experimental support were used (evidence codes IDA, IPI, IMP, IGI, IEP, TAS, IC) considering "is_a" and "part_of" relationships. This procedure resulted in approximately 9.5×10^7 protein pairs with semantic similarity values, which were used as training data. In the first model stage, multiple random forest predictors were trained, each representing one category of network features (see above), each with 640 decision trees. In the second model stage, the predictions of the regression models were used as input features and trained again against semantic similarity values to produce a final model output. Outputs of both stages used out-of-bag (OOB) predictions, a standard approach for random forest models (*109*). We also explored standard fivefold cross-validation and did not observe a significant difference in prediction performance versus the OOB procedure. New AP-MS interactions identified for protein pairs (see above) were incorporated by setting the physical interaction PPI feature of that pair to the maximum in the input of the second-stage random forest.

The output of this model was taken as the unified IAS score for each protein pair (1.8×10^8 pairs, which also included protein pairs not used in the training). Random forests were implemented using the Scikit-learn Python package (*110*) with "max_tree_depth" set to 20. This setting was determined to be near-optimal for most feature categories when sweeping

over max_tree_depths of 5, 10, 15, ..., 35, 40. Comparison between the OOB accuracy and training accuracy showed that the OOB procedure did not cause overfitting (fig. S1A). The parameter "max_features" was set to 1 for categories with fewer than 10 features (protein coexpression and codependence) and to 0.5 otherwise. Other parameters were set to default. Network features were set to 8-bit signed integer arrays (NumPy data type "int8") during the training to reduce running time. The two-stage random forest model performed similarly to an alternative procedure of training on all 127 features simultaneously in terms of correlation to the GO proximity score (fig. S1B). However, the two-stage approach enabled an analysis of the contribution of features by their broad categories (fig. S1C).

Multiscale protein network community detection

Hierarchical protein community detection was based on the CliXO clique detection algorithm, which was revised for bug fixes, redesigned parameters, and code optimization (*111*). CliXO inputs a weighted network and outputs a hierarchy of communities detected in that network as the threshold of edge weight (here, the IAS score) is lowered (Fig. 2, C and D). The new version of CliXO (v1.0) had better performance than the previous version (v0.3) in a benchmark task, in which algorithms were used to recover a ground-truth GO hierarchy from an input weighted network reflecting the Resnik semantic similarity scores defined by the same hierarchy (fig. S2A; see also supplementary text).

The algorithm consists of three parameters: α, β, and m. The parameter α reflects the step size by which network stringency (threshold IAS score) is lowered for progressive cycles of clique detection; a smaller α tends to generate a deeper hierarchy, in which the differences between parent and child communities tend to be smaller. β reflects the stringency of merging overlapping cliques; a higher β tends to merge cliques less frequently, resulting in a broader hierarchy with more sibling systems with larger overlaps. Finally, each community is assigned a score adapted from the Newman-Girvan modularity (*112*), and those with a modularity score less than m (i.e., communities that are more likely to emerge by chance) are rejected from the hierarchy (fig. S2B).

This CliXO algorithm was applied to identify protein communities at moderate-to-high interaction stringency thresholds (IAS ≥ 0.3), which captured the vast majority of protein associations driven by physical interaction (fig. S1C). Below this threshold, the integrated network had a much higher edge density, leading to impractical run-time requirements. Instead, we used HiDeF, a scalable community detection method we recently developed (*113*), to efficiently extract large-scale protein systems at low stringency (IAS < 0.3; see supplementary text).

HiSig: Identification of recurrently mutated cancer systems

Given a set of partially overlapping and/or nested systems, each consisting of a set of proteins, we developed a unified statistical model of mutation to precisely pinpoint the systems with strong evidence for mutational selection. This model, HiSig, was inspired by the technique of overlapping group lasso regression (*114*), albeit with a different mathematical formulation. HiSig code is available online (*115*).

A comprehensive list of exome-wide somatic mutations identified in the 13 TCGA tumor cohorts considered in this study (BLCA, BRCA, COAD, GBM, HNSC, KIRC, LIHC, LUAD, LUSC, OV, STAD, SKCM, UCEC) was obtained from the NCI Genomic Data Commons as a MAF (Mutation Annotation Format) file (*116*) (https://gdc.cancer.gov/about-data/publications/mc3-2017). We considered the following types of somatic mutation events: "Missense_Mutation," "Nonsense_Mutation," "Frame_Shift_Del," "Frame_Shift_Ins," "Splice_Site," "Splice_Region," and "Nonstop_Mutation"; we removed others, such as silent mutations. We did not opt to incorporate any model of phenotypic impact from mutations into HiSig, given the model's complexity and the general lack of consensus about which method of phenotypic prediction is optimal. For each cohort and gene, we recorded the number of tumors in that cohort in which that gene was observed to have at least one somatic mutation event ($N_{g,obs}$). This observed number was compared to the expected number of mutations for that gene and cohort ($N_{g,exp}$), computed using MutSigCV 1.4 with default settings (https://software.broadinstitute.org/cancer/cga/mutsig_download) (*36*). The expected mutation value accounts for covariates of mutation tendency, including gene length, mRNA expression level, replication timing, and trinucleotide context of the mutation, which are integral parts of the MutSigCV statistical model (*36*). Note that $N_{g,exp}$ is an internal variable not included in the MutSigCV output, requiring a trivial code modification to access its value. We then defined the corrected mutation count of each gene g as

$$M_g = \log\left[\max\left(N_{g,obs} - N_{g,exp}, 0\right) + \varepsilon\right]$$

with $\varepsilon = 5$ as a pseudocount to avoid taking the logarithm of zero. The vector of corrected mutation counts for all m genes, denoted $\mathbf{y}$, was fit using the following model of mutation pressure, illustrated graphically in Fig. 3A:

$$\mathbf{y} = \mathbf{Iw} + \mathbf{Hv}$$

In this model, $\mathbf{H}$ is an $m \times n$ matrix representing the assignment of m genes to n systems, in which $\mathbf{H}_{ij} = 1$ if gene i is a member of system j and 0 otherwise. $\mathbf{I}$ is an $m \times m$ identity matrix. The vectors $\mathbf{w}$ and $\mathbf{v}$ model the positive selection pressures on genes and systems, respectively, that have given rise to the mutation counts in $\mathbf{y}$. Values for these vectors were solved by linear regression with L1 lasso regularization using the R package *glmnet* (*117*) with parameters lambda.min = 10^{-4}, nlambda = 500, standardize = False, and lower.limit = 0. These settings produce a family of optimal solutions to $\mathbf{w}$ and $\mathbf{v}$ under different strengths of regularization λ. Large λ tends to select a few large systems (zero $\mathbf{w}$, sparse $\mathbf{v}$), whereas small λ tends to select every gene as a model of its own mutation counts (dense $\mathbf{w}$, zero $\mathbf{v}$). Each system t, among the set of systems T, was assigned a selective pressure S over all regularization penalties:

$$S(t) = \max_{\lambda} \frac{v_t(\lambda)}{\sum_{t' \in T} v_{t'}(\lambda) + \sum_{g \in G} w_g(\lambda)}$$

For each λ, this equation calculates the fraction of the weight of a system t among all weights in the linear equation; the maximum fraction attained is returned as $S(t)$. An empirical P value was calculated by comparing $S(t)$ of the actual hierarchy against 10,000 random hierarchies in which the hierarchy structure $\mathbf{H}$ is permuted with respect to gene labels (i.e., permuting the rows in $\mathbf{H}$). The FDR is calculated using the Benjamini-Hochberg procedure with the Python package *statsmodels* (v0.9). In this study, we defined systems with FDR < 0.25 as recurrently mutated.

To optimize the hierarchical structure for HiSig analysis, we scanned the CliXO parameters (α, β, m; see above), yielding an ensemble of systems hierarchies of varying size and complexity (fig. S2, F and G). The HiSig procedure was applied to each of these hierarchies to determine recurrently mutated systems in each tumor type. Notably, a higher overall number of systems in a hierarchy did not necessarily imply that more of these systems would be scored as under mutation pressure, because larger hierarchies also must test more systems, adversely affecting the FDR. Among this ensemble, we chose the parameters $\alpha = 0.07$, $\beta = 0.5$, $m = 0.005$, as the hierarchy generated under these parameters was an optimal tradeoff between the model parsimony and the power of detecting more significantly mutated systems (fig. S2, F and G). We used this parameter set for all subsequent analyses.

Naming of NeST systems

NeST systems were named by a team of in-house curators, based on expert knowledge, literature analysis, and GSEA analysis. Where possible, names were chosen to agree with existing literature about the functional relationship between systems' substituent genes. The naming process had no influence on the composition of these systems or their inclusion in NeST, which remains a truly data-driven construct.

Independent cancer cohorts for validation

For validation of NeST systems in independent tumor cohorts, we selected 13 tumor cohorts with sufficient whole-exome sequencing data (>100 samples per cohort) that were independent of the TCGA cohorts used to define NeST. The validation cohorts include a total of 4077 tumor samples from three ICGC (International Cancer Genome Consortium) datasets (breast, liver, lung; samples from TCGA studies were removed), seven CPTAC (Clinical Proteomic Tumor Analysis Consortium) datasets (brain, breast, colon, lung, kidney, ovary, uterus), and two datasets from focused studies (colon, liver). MAF files recording the somatic mutations of these studies were obtained from ICGC data portal release 27 (https://dcc.icgc.org/releases/release_27), CPTAC data portal at the Genomic Data Commons (https://portal.gdc.cancer.gov/), and the cBioPortal (*104*), respectively. The MAF files of ICGC samples were processed by the *icgcSimpleMutationToMAF* function in the *maftools* package (*118*). MAF files were then used as input to MutSigCV 1.4 and followed by HiSig analysis, similar to the procedure described above for analysis of TCGA data.

Each validation cohort was paired with the TCGA discovery cohort of matching tissue type (Fig. 4B). For each of these pairwise comparisons, systems identified by HiSig as recurrently mutated in the discovery cohort (see above) were examined for significance in the validation cohort; if HiSig FDR < 0.1, the system was marked as "validated." FDR was computed from the HiSig *P* value using the Benjamini-Hochberg procedure, with the number of multiple tests equal to the number of recurrently mutated systems in the discovery cohort. To provide a reference for this analysis, we performed similar analyses for individual genes identified by MutSigCV, and for systems across cohorts of mismatched cancer types (fig. S3A).

Evaluation of the novelty of NeST systems

Overlap between NeST and known cancer pathways was determined according to 241 literature-curated pathways collected in our previous survey (*71*). NeST systems with significant overlap with any of these pathways (at least two overlapping genes and $P < 0.05$ by hypergeometric test adjusted by Bonferroni correction) were marked as having "overlap with cancer pathways." Remaining systems were tested for significant overlap with 16,064 gene sets in Gene Ontology (including terms in Cellular Component, Biological Process, and Molecular Function), and those with significant overlaps (same statistical definition as above) were marked as "overlap with other cell biology entities" (Fig. 4D and table S5).

CRISPR-Cas9 screen for determination of BRCAness

The protocol for arrayed CRISPR-Cas9 screening was adapted from previous work (*119*). To obtain reliable cell counts for cell proliferation, SW1710 bladder cancer cells were RFP-labeled using the Incucyte NucLight Red Lentivirus Reagent (cat. no. 4476), and transfected cells were selected using puromycin (3 μg/ml). Three different crRNAs were obtained for each gene from Dharmacon. crRNA:tracrRNA duplexes were formed by initially incubating 4 μl of crRNA (160 μM) with 4 μl of tracrRNA (160 μM) for 45 min at 37°C. Duplexes were incubated with 8 μl of Cas9-NLS protein (40 μM) at 37°C for 15 min. These crRNPs were then aliquoted (4 μl) into 96-well V-bottom plates in arrayed format in a random order to mitigate potential positional effects. Nucleofection of crRNPs into cells was conducted using the SE cell line 4D-nucleofector kit (Lonza, cat. no. V4SC-1960); 200,000 SW1710 RFP-labeled cells per well were resuspended in 20 μl of supplemented SE buffer and mixed with crRNPs. Cells were nucleofected on the Amaxa 4D-Nucleofector System, using program CM-137.

After nucleofection, 80 μl of prewarmed media was added and the cells were recovered at 37°C. Nucleofected cells from 96-well plates were then transferred to a 384-well plate such that each well has four technical replicates of ~500 cells. After 24 hours of recovery, the initial cell counts were measured using an ArrayScan (Thermo Fisher Scientific). Immediately after obtaining the initial cell counts, cells were treated with olaparib (10 μM) or cisplatin (600 nM). The selected doses were those resulting in 80% cell death compared to no treatment in wild-type cells based on dose-response curves determined in prescreens. Final cell counts were measured after ~6 days, when the wells nucleofected with nontargeting control were confluent.

As a control, BRCA1 (cat. no. CM-003461-02, Dharmacon) was tested to have a gene-editing efficiency of 87.6% under these conditions through TIDE analysis (*120*). Phenotypic analysis was conducted by normalizing the cell proliferation in drug-treated conditions to the cell proliferation in the no-treatment condition. This approach eliminated potential relative cell growth defects in the no-treatment condition as a result of individual gene knockouts. Differential drug sensitivity was determined by comparison of the mean cell proliferation of 24 total replicates per gene to the mean value obtained from nontargeting control. The statistical significance of the phenotypic effect (Fig. 3C) was determined by a paired-sample *t* test.

Analysis of the PIK3CA-actomyosin system

Proximity ligation assay. CAL-33 cell lines were fixed in 4% paraformaldehyde (in PBS) for 15 min at room temperature, and then washed in PBS. Cells were permeabilized with 0.25% Triton X-100 (in PBS) for 10 min at room temperature. Unspecific binding sites were blocked by incubating cells in blocking solution included in the Duolink In Situ Red Starter Kit Mouse/Rabbit (Sigma-Aldrich DUO92101). Primary antibody incubation was performed at 4°C overnight. The different combinations of primary antibodies used were: anti-PIK3CA (Invitrogen MA5-17149) diluted at 1/400, anti-FLAG (Cell Signaling Technologies 14793S) diluted at 1/800, and anti-MYH9 (ThermoFisher PA-29673) diluted at 1/500. Primary antibodies were diluted in Duolink Antibody Diluent. Detection was performed according to the manufacturer's protocol. Briefly, after probe incubation for an hour at 37°C, a ligation-ligase solution was added to each sample for 30 min at 37°C, and then washed twice for 2 min with 1× Duolink In Situ Wash Buffer A. An Amplification-Polymerase solution was added for 100 min at 37°C, and then washed twice for 10 min with 1× Duolink In Situ Wash Buffer B. The slides were mounted using Duolink In Situ Mounting Medium with DAPI. Protein-protein interactions appear as red spots.

Samples were imaged on a Nikon Ti2-E (Nikon) microscope equipped with a CREST X-Light spinning disk confocal (Crest Optics), Celeste Light Engine (Lumencor), Piezo stage (Mad City Labs), and a Prime 95B 25mm CMOS camera (Photometrics) using a Plan Apo VC 100×/1.4 Oil (Nikon). The red PLA dye was measured by exciting with 561-nm laser and capturing with 607/36-m filter. Nuclei/DAPI were excited with a 405-nm laser and captured with 450/50-m filter. Z stacks were set to capture the height of all cells in the field of view and images were taken to capture >150 cells per condition. PLA spots in cells were segmented in 3D and counted using GA3 analysis in NIS Elements (v. 5.30.01 build 1541, Nikon).

Cell culture and immunostaining. CAL-33 cells were maintained in Gibco DMEM (ThermoFisher, 11995073) supplemented with 10% fetal bovine serum (Fisher Scientific, 26-140-079) and 1% penicillin-streptomycin (Corning, Cat# 30-002-CI), and incubated at 37°C with 5% CO_2. For superresolution imaging experiments, cells were cultured in a poly-L-lysine–coated (Sigma-Aldrich, P9155) Lab-Tek II 8-well chambered coverglass (ThermoFisher, 155360). Doxycycline (Clontech, 631311) was added to each well at 1 μg/ml at the time of cell plating to induce 3×FLAG-PIK3CA expression. After 36 to 48 hours, the cells reached 60 to 70% confluency and were fixed and stained using the same protocols as described (*121*). Briefly, fixation was done with 3.7% paraformaldehyde (PFA) in 1× PHEM buffer at room temperature for 20 min. After three quick PBS (ThermoFisher, 14190144) washes, the sample was quenched with fresh 0.1% sodium borohydride in PBS for 7 min, washed with PBS (3×), and permeabilized with 0.3% saponin (Sigma-Aldrich, 47036) in PBS for 30 min. The sample was then rinsed

with PBS and blocked with Image-iT FX signal enhancer (ThermoFisher, I36933) at room temperature for 30 min. After three quick rinses with PBS, the sample underwent a second blocking in PBS with 3% goat serum (Abcam, ab7481) and 5% salmon sperm DNA (ThermoFisher, AM9680) for 30 min with gentle rocking. The sample was then incubated with a mouse monoclonal anti-FLAG M2 antibody (Sigma-Aldrich, F1804, dilution ratio 1:1000) and a rabbit anti-Myosin IIA antibody (Sigma-Aldrich, M8064, dilution ratio 1:150) in the same BSA and Salmon DNA blocking buffer at room temperature on a rocker for 90 min. The sample was then rinsed with PBS (3×, 5 min each) before incubating with DNA-conjugated donkey anti-rabbit IgG (H+L) (Jackson Immuno Research, 711-005-152, dilution ratio 1:100) and donkey anti-mouse IgG (H+L) (Jackson Immuno Research, 715-005-150, dilution ratio 1:200) in blocking buffer. Here, the anti-mouse and anti-rabbit secondary antibodies were pre-conjugated with different DNA oligos, namely DS1 (sequence: 5′-TATACATCTAAATACATCTAAT) and DS2 (sequence: 5′-TTATCTACATATATCTACATAT), respectively, using a procedure reported in (*121*). The incubation also took place on a rocker at room temperature for 1 hour, followed by thorough PBS rinses. The sample was then post-fixed with 3.7% PFA and 0.1% glutaraldehyde in a PHEM buffer at room temperature for 10 min. Of note, this post-fixation step is critical to subsequent DNA-PAINT imaging.

DNA-PAINT imaging. All superresolution fluorescence data were taken on a custom-built single-molecule imaging system used in our previous work (*121*). For multicolor imaging, the targets were imaged sequentially with exchange-PAINT using our modified protocol (*121*, *122*). Gold nanorods (50 nm, BBI Solutions, EM. GC50/4) added to the samples were used as fiduciary markers for both stage drift correction and image registration. All imaging was performed using Atto643 (Atto-tec, AD643) labeled imager strands IS1 (sequence: 5′-TTAG-ATGTAT) or IS2 (sequence: 5′ ATATGTAGAT) in imaging buffer C (1× PBS with 500 mM NaCl) containing 10 to 15% ethylene carbonate (EC) (Sigma-Aldrich, 676802). For each target, typically 40,000 frames of raw DNA-PAINT image data were acquired at 50-ms exposure time using micromanager (*123*). Image analysis, including single-molecule localization and subsequent coordinate filtering, sorting, and rendering, was performed using in-house Matlab scripts as described (*124*). For localization sorting, events that appeared within a defined number of frames (typically 5) and distance (typically 85 nm or 0.5 pixel on our setup) were combined into a single event with averaged coordinates. The assorted localizations were then used for final image rendering, and the rendered images were exported as TIF files for further analysis and annotations in Fiji (*125*).

Signaling assay. Serum-starved cells from head-and-neck cancer cell lines (CAL-33, SCC-25) were treated with a combination of 0.5 μM alpelisib (MedChemExpress HY-15244) and/or 10 μM (−)-blebbistatin (Selleckchem S7099) with the appropriate vehicle control for 30 min. Cells were lysed and subsequently cleared by centrifugation. After normalization of protein concentrations, samples were boiled in Laemmli sample buffer (BioRad) and subjected to immunoblot analysis. Antibodies (Cell Signaling Technologies) were used at the indicated dilutions: p-AKT Ser473 (cat. no. 4060, 1:5000), AKT (cat. no. 9272, 1:1000), p-S6 Ser235/236 (cat. no. 2211, 1:5000), S6 (cat. no. 2317, 1:1000), GAPDH (cat. no. 2118, 1:10,000). Results were quantified using ImageJ using the ratio of phosphorylated/total signal normalized to the vehicle control of each replicate ($N = 3$, Fig. 5F and fig. S5D).

Association of mutation status with RPPA data. Reverse-phase protein array (RPPA) quantifications of phosphorylated/total proteins in each of 899 cancer cell lines, (CCLE_RPPA_20181003.csv), along with the mutation profiles of these cell lines, (CCLE_ DepMap_18q3_maf_20180718.txt), were downloaded from the CCLE data portal (https://portals.broadinstitute.org/ccle/data) (*78*). Synonymous mutations and mutations outside of protein-coding regions were excluded from this analysis.

Analysis of collagen mutations

dN/dS analysis. To analyze mutational selection on collagen complexes, we considered the most prominent collagen mutation types for SKCM and LUAD, i.e., C → T or G → A transitions for SKCM, and transitions and transversions on C/G nucleotides for LUAD. Background mutation rates were calculated by examining the silent positions of the collagen genes in the analyzed systems. Triple-helical regions of collagen proteins were defined using annotations in the Uniprot database. Statistical significance and 95% CIs (Fig. 6E) were calculated based on a one-tailed binomial test.

Structure-based stability analysis. All point mutations in the trimeric helix region were considered. A crystal structure of collagen I trimeric helix region (PDB ID: 1BKV) was used as a template, which contains three identical polypeptide chains, each with 10 trimeric repeats. For each point mutation, nine residues surrounding the mutated position were threaded onto the middle of the template (fourth to sixth repeats). Two structure models representing the wild-type and the mutant were created (differing by one residue at the mutated position) and scored by the FoldX 5.0 suite (*88*) with the "BuildModel" command. The change of protein stability upon point mutation ($\Delta\Delta G$) was defined as the "total energy" of the mutant model subtracted by that of the wild-type model. Because the magnitude of destabilization is often nonphysiological, and because of the fixed backbone structures in this type of analysis, we set the maximum of $\Delta\Delta G$ to 8.0 kcal/mol when displaying results (Fig. 6F).

Association of mutation status with metastatic phenotypes. Metastatic potential and penetrance of 488 cancer cell lines were obtained from the MetMap 500 dataset (https://depmap.org/metmap/data) (*89*). Mutation profiles of these cell lines (CCLE_DepMap_18q3_maf_20180718.txt) were downloaded from the CCLE data portal (https://portals.broadinstitute.org/ccle/data) (*78*). Synonymous mutations and mutations outside of protein-coding regions were excluded from this analysis.

Cell lines and plasmids. HFF-1 (ATCC SCRC-1041), Tu To (ATCC CRL-1298), and A549 cells (ATCC CCL-185) were obtained from the American Type Culture Collection (ATCC). pEGFP-N2-COL1A1 was a gift from D Stephens (Addgene plasmid #66602; http://n2t.net/addgene:66602; RRID:Addgene_66602). pMD2.G was a gift from D. Trono (Addgene plasmid #12259; http://n2t.net/addgene:12259; RRID:Addgene_12259). pCMV-dR8.2 dvpr was a gift from B. Weinberg (Addgene plasmid #8455; http://n2t.net/addgene:8455; RRID:Addgene_8455). The COL1A1 ORF was introduced into pLenti CMV Blast DEST (706-1), which was a gift from E. Campeau and P. Kaufman (Addgene plasmid #17451; http://n2t.net/addgene:17451; RRID:Addgene_17451). With pCMV-dR8.2 dvpr and pMD2.G, IT was used to stably introduce CMV-COL1A1 into HFF-1 cells via standard lentiviral packaging and infection protocols as described (*28*). Site-directed mutagenesis was used to introduce the G281S mutation into COL1A1 as described (*28*).

Whole-exome sequencing of Tu To. All cells were cultured in Dulbecco's modified Eagle's medium supplemented with 10% (v/v) fetal bovine serum (FBS; Corning) and 0.1% gentamicin (Gibco Thermofisher). These cells were harvested to isolate their genomic DNA using a Wizard Genomic DNA Purification Kit (Promega), following the manufacturer's instructions. The isolated genomic DNA was sequenced utilizing an Illumina NovaSeq 6000. We used Terra (https://terra.bio/), a cloud computing platform for genomics, to perform germline mutation calling of Tu To. Paired-end FASTQ files of Tu To whole-exome sequencing were used as the input, and we used the "Sequence-Format-Conversion/Paired-FASTQ-to-Unmapped-BAM" and "Exome-Analysis-Pipeline/ExomeGermlineSingleSample" workflows to identify the germline variants.

Fibroblast-derived matrix experiments. Wild-type and mutant fibroblast-derived matrices were generated following the basic protocol outlined in (*126*) with adaptations. Media was supplemented with dextran sulfate sodium

salt from *Leuconostoc* spp., MW ~ 500,000 g/mol (D8906, Sigma-Aldrich) as described in (*127*) with a concentration of 100 μg/ml instead of the addition of ascorbic acid. Thickness of each matrix was measured using a Nikon TiE microscope with a 10× objective as the z-distance of the cell layer focal plane from the dish bottom focal plane. Once the fibroblast-derived matrix thickness reached 20 to 30 μm, the matrices were decellularized following the protocol in (*126*) including the addition of DNase I (Thermo Fisher Scientific Baltics, Vilnius, Lithuania). After decellularization, A549 cells were seeded on top of the matrices at a density of 30,000 cells/ml and incubated for 25 hours. The cells were then fixed using 4% paraformaldehyde (15710-s, Electron Microscopy Sciences, Hatfield, PA) for 30 min at room temperature. The fixed samples were stained with primary antibody against Ki67 (1:400 dilution, 8D5, Cell Signaling Technology) and secondary antibody Alexa Fluor 546 goat anti-mouse (1:2000 dilution; A11003, Life Technologies). DAPI (1:1000 dilution; 4083, Cell Signaling) was used to counterstain the nuclei. Fluorescent images were obtained using a Nikon TiE inverted microscope (Nikon Instruments Inc., Melville, NY) with a 10× objective lens. The images were observed and taken in NIS Elements Software (Nikon).

Clinical analysis of protein systems

Each NeST system was tested for association between the mutation status of the system and the progression-free patient survival (PFS) time. Associations were tested only in the tumor type(s) for which the system was recurrently mutated (HiSig). Clinical data from TCGA patients (Pancancer Atlas), including PFS, were downloaded from the cBioPortal (*104*). Statistical significance was evaluated by log rank test, with P values adjusted by the Benjamini-Hochberg procedure. We noted that PFS time is significantly associated with the overall mutation burden in several tumor types (BLCA, OV, STAD, UCEC); thus, we used a multivariable Cox proportional-hazards model for all analysis, using the mutation status and the total number of mutations in the sample ($\log_{10}$-transformed) as independent variables. The P values assigned to mutation status by this model considered the statistical association after removing the confounding effects of mutation burden. This process identified a total of 38 associations (FDR < 0.3) between a system and PFS. Next, we tested the association of PFS with the mutation status of individual genes defined as cancer drivers by the TCGA Pancancer Atlas (*37*) for each of the 13 analyzed tumor types. From all combinations of single genes and tumor types, nine significant prognostic associations (i.e., gene/PFS) were detected (FDR < 0.3; table S7). We then excluded any PFS-associated system containing a gene which was (i) determined to be prognostic in that same tumor type and (ii) accounted for more than 80% of tumor cases with mutations in that system. After subtracting these systems, a total of 25 (system, tumor type) associations were found remaining, i.e., associations which could not be attributed to mutations within only a single gene.

Survey of cancer genes based on systems-level analysis

To create a catalog of cancer genes based on systems in the NeST map (table S8), we conservatively selected the top two genes with the highest mutation rates per system, adjusted for pleiotropic genes included in multiple systems. Raw gene mutation frequency is not a suitable metric to rank the importance of a gene to a system, because a pleiotropic gene with high mutation frequency (e.g., TP53) would be ranked first for every system in which it participates, despite contributing to many different systems. Rather, we distributed the total mutation count of a gene across the systems in which it participates, proportioned according to the weights assigned to each system by HiSig (**v** vector; see HiSig method above). In particular, we translated M_g, the total mutation counts for gene g, to M_{gs}, the mutation counts for gene g ascribed to system s, according to the following function:

$$\log M_{gs} = \frac{v_s}{w_g + \sum_{s' \in S(g)} v_{s'}} \log M_g$$

where w_g and v_s are the weights given to gene g and system s in the regression model by HiSig. Thus, for each system s, the genes were ranked by their values of M_{gs}, which penalized genes with high pleiotropy. The resulting NeST gene catalog was analyzed against three types of validation as follows:

Differential expression in tumors. For each gene, we examined differential expression between tumor and normal samples in nine TCGA cohorts, which each had >40 normal samples (BRCA, HNSC, KIRC, LIHC, LUAD, LUSC, PRAD, THCA, COAD). Differential expression was evaluated using a Kolmogorov-Smirnov (KS) test, by comparing the percentile rankings of gene RSEM values downloaded from the UCSC Xena platform (*128*) (https://xenabrowser.net/datapages/). A summary statistic (Fig. 7G) was derived by summing up the negative log of the P values across these cohorts.

Prognostic value of gene expression. For each gene and tissue type, we examined log-rank P values scoring the significance of association between mRNA expression level and patient survival, using Kaplan-Meier analysis. These P values were obtained from the Human Protein Atlas (HPA) (*129*) for each of the 13 TCGA cohorts analyzed in our study (www.proteinatlas.org/about/download). A summary statistic (Fig. 7H) was derived by summing up the negative log of the P values across these cohorts.

Mouse genetic screens. We used the Candidate Cancer Gene Database (CCGD) (*130*) (http://ccgd-starrlab.oit.umn.edu, downloaded on 4 June 2019), which contains literature-curated evidence from transposon-based forward genetic screens in mice. The number of studies in which a gene was disrupted by transposon insertional mutagenesis in mice tumors was used as the summary statistic (Fig. 7I).

REFERENCES AND NOTES

1. L. A. Garraway, E. S. Lander, Lessons from the cancer genome. *Cell* **153**, 17–37 (2013). doi: 10.1016/j.cell.2013.03.002; pmid: 23540688
2. R. A. Burrell, N. McGranahan, J. Bartek, C. Swanton, The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013). doi: 10.1038/nature12625; pmid: 24048066
3. Mutation Consequences and Pathway Analysis Working Group of the International Cancer Genome Consortium, Pathway and network analysis of cancer genomes. *Nat. Methods* **12**, 615–621 (2015). doi: 10.1038/nmeth.3440; pmid: 26125594
4. Y.-A. Kim, D.-Y. Cho, T. M. Przytycka, Understanding Genotype-Phenotype Effects in Cancer via Network Approaches. *PLOS Comput. Biol.* **12**, e1004747 (2016). doi: 10.1371/journal.pcbi.1004747; pmid: 26963104
5. A. Laddach, J. C.-F. Ng, S. S. Chung, F. Fraternali, Genetic variants and protein-protein interactions: A multidimensional network-centric view. *Curr. Opin. Struct. Biol.* **50**, 82–90 (2018). doi: 10.1016/j.sbi.2017.12.006; pmid: 29306755
6. Cancer Genome Atlas Network, Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015). doi: 10.1038/nature14129; pmid: 25631445
7. A. Liberzon *et al.*, The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015). doi: 10.1016/j.cels.2015.12.004; pmid: 26771021
8. M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016). doi: 10.1093/nar/gkv1070; pmid: 26476454
9. F. Sanchez-Vega *et al.*, Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337.e10 (2018). doi: 10.1016/j.cell.2018.03.035; pmid: 29625050
10. M. Paczkowska *et al.*, Integrative pathway enrichment analysis of multivariate omics data. *Nat. Commun.* **11**, 735 (2020). doi: 10.1038/s41467-019-13983-9; pmid: 32024846
11. H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, T. Ideker, Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3**, 140 (2007). doi: 10.1038/msb4100180; pmid: 17940530
12. F. Vandin, E. Upfal, B. J. Raphael, Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* **18**, 507–522 (2011). doi: 10.1089/cmb.2010.0265; pmid: 21385051
13. C. J. Vaske *et al.*, Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237–i245 (2010). doi: 10.1093/bioinformatics/btq182; pmid: 20529912
14. M. Hofree, J. P. Shen, H. Carter, A. Gross, T. Ideker, Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108–1115 (2013). doi: 10.1038/nmeth.2651; pmid: 24037242
15. M. D. M. Leiserson *et al.*, Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015). doi: 10.1038/ng.3168; pmid: 25501392
16. A. Cho *et al.*, MUFFINN: Cancer gene discovery via network analysis of somatic mutation data. *Genome Biol.* **17**, 129 (2016). doi: 10.1186/s13059-016-0989-x; pmid: 27333808

17. H. Horn *et al*., NetSig: Network-based discovery from cancer genomes. *Nat. Methods* **15**, 61–66 (2018). doi: 10.1038/nmeth.4514; pmid: 29200198
18. C. Dimitrakopoulos *et al*., Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* **34**, 2441–2448 (2018). doi: 10.1093/bioinformatics/bty148; pmid: 29547932
19. M. A. Reyna, M. D. M. Leiserson, B. J. Raphael, Hierarchical HotNet: Identifying hierarchies of altered subnetworks. *Bioinformatics* **34**, i972–i980 (2018). doi: 10.1093/bioinformatics/bty613; pmid: 30423088
20. M. A. Reyna *et al*., Pathway and network analysis of more than 2500 whole cancer genomes. *Nat. Commun.* **11**, 729 (2020). doi: 10.1038/s41467-020-14367-0; pmid: 32024854
21. M. A. Pujana *et al*., Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat. Genet.* **39**, 1338–1349 (2007). doi: 10.1038/ng.2007.2; pmid: 17922014
22. O. Rozenblatt-Rosen *et al*., Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. *Nature* **487**, 491–495 (2012). doi: 10.1038/nature11288; pmid: 22810586
23. Y. Yang *et al*., Essential role of the linear ubiquitin chain assembly complex in lymphoma revealed by rare germline polymorphisms. *Cancer Discov.* **4**, 480–493 (2014). doi: 10.1158/2159-8290.CD-13-0915; pmid: 24491438
24. N. J. Krogan, S. Lippman, D. A. Agard, A. Ashworth, T. Ideker, The cancer cell map initiative: Defining the hallmark networks of cancer. *Mol. Cell* **58**, 690–698 (2015). doi: 10.1016/j.molcel.2015.05.008; pmid: 26000852
25. E. H. Bowler, Z. Wang, R. M. Ewing, How do oncoprotein mutations rewire protein-protein interaction networks? *Expert Rev. Proteomics* **12**, 449–455 (2015). doi: 10.1586/14789450.2015.1084875; pmid: 26325016
26. M. Eckhardt *et al*., Multiple Routes to Oncogenesis Are Promoted by the Human Papillomavirus-Host Protein Network. *Cancer Discov.* **8**, 1474–1489 (2018). doi: 10.1158/2159-8290.CD-17-1018; pmid: 30209081
27. J. R. Kovalski *et al*., The Functional Proximal Proteome of Oncogenic Ras Includes mTORC2. *Mol. Cell* **73**, 830–844.e12 (2019). doi: 10.1016/j.molcel.2018.12.001; pmid: 30639242
28. M. R. Kelly *et al*., Combined Proteomic and Genetic Interaction Mapping Reveals New RAS Effector Pathways and Susceptibilities. *Cancer Discov.* **10**, 1950–1967 (2020). doi: 10.1158/2159-8290.CD-19-1274; pmid: 32727735
29. K. Drew *et al*., Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Mol. Syst. Biol.* **13**, 932 (2017). doi: 10.15252/msb.20167490; pmid: 28596423
30. E. L. Huttlin *et al*., Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505–509 (2017). doi: 10.1038/nature22366; pmid: 28514442
31. K. J. Roux, D. I. Kim, B. Burke, D. G. May, BioID: A Screen for Protein-Protein Interactions. *Curr. Protoc. Protein Sci.* **91**, 19.23.11–19.23.15 (2018).
32. K. Luck *et al*., A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020). doi: 10.1038/s41586-020-2188-x; pmid: 32296183
33. C. D. Go *et al*., A proximity biotinylation map of a human cell. *Nature* **595**, 120–124 (2021). doi: 10.1101/796391
34. Y. Samuels *et al*., High frequency of mutations of the PIK3CA gene in human cancers. *Science* **304**, 554 (2004). doi: 10.1126/science.1096502; pmid: 15016963
35. P. A. J. Muller, K. H. Vousden, p53 mutations in cancer. *Nat. Cell Biol.* **15**, 2–8 (2013). doi: 10.1038/ncb2641; pmid: 23263379
36. M. S. Lawrence *et al*., Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013). doi: 10.1038/nature12213; pmid: 23770567
37. M. H. Bailey *et al*., Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371–385.e18 (2018). doi: 10.1016/j.cell.2018.02.060; pmid: 29625053
38. B. Niu *et al*., Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat. Genet.* **48**, 827–837 (2016). doi: 10.1038/ng.3586; pmid: 27294619
39. C. Tokheim *et al*., Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. *Cancer Res.* **76**, 3719–3731 (2016). doi: 10.1158/0008-5472.CAN-15-3190; pmid: 27197156
40. J. Gao *et al*., 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Med.* **9**, 4 (2017). doi: 10.1186/s13073-016-0393-x; pmid: 28115009
41. F. Cheng *et al*., Comprehensive characterization of protein-protein interactions perturbed by disease mutations. *Nat. Genet.* **53**, 342–353 (2021). pmid: 33558758
42. P. Creixell *et al*., Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell* **163**, 202–217 (2015). doi: 10.1016/j.cell.2015.08.056; pmid: 26388441
43. J. Reimand, G. D. Bader, Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* **9**, 637 (2013). doi: 10.1038/msb.2012.68; pmid: 23340843
44. J. J. Bouchard *et al*., Cancer Mutations of the Tumor Suppressor SPOP Disrupt the Formation of Active, Phase-Separated Compartments. *Mol. Cell* **72**, 19–36.e8 (2018). doi: 10.1016/j.molcel.2018.08.027; pmid: 30244836
45. S. Vyas, E. Zaganjor, M. C. Haigis, Mitochondria and Cancer. *Cell* **166**, 555–566 (2016). doi: 10.1016/j.cell.2016.07.002; pmid: 27471965
46. D. L. Swaney *et al*., A protein network map of head and neck cancer reveals PIK3CA mutant drug sensitivity. *Science* **374**, eabf2911 (2021). doi: 10.1126/science.eabf2911
47. M. Kim *et al*., A protein interaction landscape of breast cancer. *Science* **374**, eabf3066 (2021). doi: 10.1126/science.eabf3066
48. M. Giurgiu *et al*., CORUM: The comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.* **47**, D559–D563 (2019). doi: 10.1093/nar/gky973; pmid: 30357367
49. A. Chatr-aryamontri *et al*., The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **45**, D369–D379 (2017). doi: 10.1093/nar/gkw1102; pmid: 27980099
50. S. Orchard *et al*., Protein interaction data curation: The International Molecular Exchange (IMEx) consortium. *Nat. Methods* **9**, 345–350 (2012). doi: 10.1038/nmeth.1931; pmid: 22453911
51. T. Rolland *et al*., A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014). doi: 10.1016/j.cell.2014.10.050; pmid: 25416956
52. M. Y. Hein *et al*., A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712–723 (2015). doi: 10.1016/j.cell.2015.09.053; pmid: 26496610
53. J. D. Lapek Jr. *et al*., Detection of dysregulated protein-association networks by high-throughput proteomics predicts cancer vulnerabilities. *Nat. Biotechnol.* **35**, 983–989 (2017). doi: 10.1038/nbt.3955; pmid: 28892078
54. P. Mertins *et al*., Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–62 (2016). doi: 10.1038/nature18003; pmid: 27251275
55. H. Zhang *et al*., Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166**, 755–765 (2016). doi: 10.1016/j.cell.2016.05.069; pmid: 27372738
56. A. K. Ramani *et al*., A map of human protein interactions derived from co-expression of human mRNAs and their orthologs. *Mol. Syst. Biol.* **4**, 180 (2008). doi: 10.1038/msb.2008.19; pmid: 18414481
57. E. A. Boyle, J. K. Pritchard, W. J. Greenleaf, High-resolution mapping of cancer cell networks using co-functional interactions. *Mol. Syst. Biol.* **14**, e8594 (2018). doi: 10.15252/msb.20188594; pmid: 30573688
58. I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, E. M. Marcotte, Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **21**, 1109–1121 (2011). doi: 10.1101/gr.118992.110; pmid: 21536720
59. M. H. Kramer *et al*., Active Interaction Mapping Reveals the Hierarchical Organization of Autophagy. *Mol. Cell* **65**, 761–774.e5 (2017). doi: 10.1016/j.molcel.2016.12.024; pmid: 28132844
60. M. Ashburner *et al*., Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000). doi: 10.1038/75556; pmid: 10802651
61. Y. Dong *et al*., Regulation of BRCC, a holoenzyme complex containing BRCA1 and BRCA2, by a signalosome-like subunit and its role in DNA repair. *Mol. Cell* **12**, 1087–1099 (2003). doi: 10.1016/S1097-2765(03)00424-6; pmid: 14636569
62. T. Valenta, G. Hausmann, K. Basler, The many faces and functions of β-catenin. *EMBO J.* **31**, 2714–2736 (2012). doi: 10.1038/emboj.2012.150; pmid: 22617422
63. A. Subramanian *et al*., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005). doi: 10.1073/pnas.0506580102; pmid: 16199517
64. C. J. Lord, A. Ashworth, BRCAness revisited. *Nat. Rev. Cancer* **16**, 110–120 (2016). doi: 10.1038/nrc.2015.21; pmid: 26775620
65. N. Turner, A. Tutt, A. Ashworth, Hallmarks of 'BRCAness' in sporadic cancers. *Nat. Rev. Cancer* **4**, 814–819 (2004). doi: 10.1038/nrc1457; pmid: 15510162
66. G. Ciriello *et al*., Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* **163**, 506–519 (2015). doi: 10.1016/j.cell.2015.09.033; pmid: 26451490
67. A. G. Robertson *et al*., Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell* **171**, 540–556.e25 (2017). doi: 10.1016/j.cell.2017.09.007; pmid: 28988769
68. C. J. Lord, A. Ashworth, PARP inhibitors: Synthetic lethality in the clinic. *Science* **355**, 1152–1158 (2017). doi: 10.1126/science.aam7344; pmid: 28302823
69. D. Hanahan, R. A. Weinberg, Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011). doi: 10.1016/j.cell.2011.02.013; pmid: 21376230
70. M. S. Lawrence *et al*., Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014). doi: 10.1038/nature12912; pmid: 24390350
71. B. M. Kuenzi, T. Ideker, A census of pathway maps in cancer systems biology. *Nat. Rev. Cancer* **20**, 233–246 (2020). doi: 10.1038/s41568-020-0240-7; pmid: 32066900
72. P. A. Futreal *et al*., A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004). doi: 10.1038/nrc1299; pmid: 14993899
73. A. Hall, The cytoskeleton and cancer. *Cancer Metastasis Rev.* **28**, 5–14 (2009). doi: 10.1007/s10555-008-9166-3; pmid: 19153674
74. D. V. Köster *et al*., Actomyosin dynamics drive local membrane component organization in an in vitro active composite layer. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E1645–E1654 (2016). doi: 10.1073/pnas.1514030113; pmid: 26929326
75. G. Xue, B. A. Hemmings, PKB/Akt-dependent regulation of cell motility. *J. Natl. Cancer Inst.* **105**, 393–404 (2013). doi: 10.1093/jnci/djs648; pmid: 23355761
76. X. Wu *et al*., Activation of diverse signalling pathways by oncogenic PIK3CA mutations. *Nat. Commun.* **5**, 4961 (2014). doi: 10.1038/ncomms5961; pmid: 25247763
77. A. F. Straight *et al*., Dissecting temporal and spatial control of cytokinesis with a myosin II inhibitor. *Science* **299**, 1743–1747 (2003). doi: 10.1126/science.1081412; pmid: 12637748
78. M. Ghandi *et al*., Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019). doi: 10.1038/s41586-019-1186-3; pmid: 31068700
79. S. D. Bradley *et al*., BRAFV600E Co-opts a Conserved MHC Class I Internalization Pathway to Diminish Antigen Presentation and $CD8^+$ T-cell Recognition of Melanoma. *Cancer Immunol. Res.* **3**, 602–609 (2015). doi: 10.1158/2326-6066.CIR-15-0030; pmid: 25795007
80. M. D. Shoulders, R. T. Raines, Collagen structure and stability. *Annu. Rev. Biochem.* **78**, 929–958 (2009). doi: 10.1146/annurev.biochem.77.032207.120833; pmid: 19344236
81. M. Fang, J. Yuan, C. Peng, Y. Li, Collagen as a double-edged sword in tumor progression. *Tumour Biol.* **35**, 2871–2882 (2014). doi: 10.1007/s13277-013-1511-7; pmid: 24338768
82. D. O. Velez *et al*., 3D collagen architecture induces a conserved migratory and transcriptional response linked to vasculogenic mimicry. *Nat. Commun.* **8**, 1651 (2017). doi: 10.1038/s41467-017-01556-7; pmid: 29162797
83. J. Riegler *et al*., Tumor Elastography and Its Association with Collagen and the Tumor Microenvironment. *Clin. Cancer Res.* **24**, 4455–4467 (2018). doi: 10.1158/1078-0432.CCR-17-3262; pmid: 29798909
84. S. E. Bell *et al*., Differential gene expression during capillary morphogenesis in 3D collagen matrices: Regulated expression of genes involved in basement membrane matrix assembly, cell cycle progression, cellular differentiation and G-protein signaling. *J. Cell Sci.* **114**, 2755–2773 (2001). doi: 10.1242/jcs.114.15.2755; pmid: 11683410
85. M.-P. Simon *et al*., Deregulation of the platelet-derived growth factor B-chain gene via fusion with collagen gene COL1A1 in dermatofibrosarcoma protuberans and giant-cell fibroblastoma. *Nat. Genet.* **15**, 95–98 (1997). doi: 10.1038/ng0197-95; pmid: 8988177
86. P. S. Tarpey *et al*., Frequent mutation of the major cartilage collagen gene COL2A1 in chondrosarcoma. *Nat. Genet.* **45**, 923–926 (2013). doi: 10.1038/ng.2668; pmid: 23770606
87. L. B. Alexandrov *et al*., The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020). doi: 10.1038/s41586-020-1943-3; pmid: 32025018

88. J. Delgado, L. G. Radusky, D. Cianferoni, L. Serrano, FoldX 5.0: Working with RNA, small molecules and a new graphical interface. *Bioinformatics* **35**, 4168–4169 (2019). doi: 10.1093/bioinformatics/btz184; pmid: 30874800
89. X. Jin *et al*., A metastasis map of human cancer cell lines. *Nature* **588**, 331–336 (2020). doi: 10.1038/s41586-020-2969-2; pmid: 33299191
90. N. Beaubier *et al*., Clinical validation of the Tempus xO assay. *Oncotarget* **9**, 25826–25832 (2018). doi: 10.18632/oncotarget.25381; pmid: 29899824
91. R. Sharaf *et al*., FoundationOne CDx testing accurately determines whole arm 1p19q codeletion status in gliomas. *Neurooncol. Adv.* **3**, vdab017 (2021). doi: 10.1093/noajnl/vdab017
92. D. Hanahan, R. A. Weinberg, The hallmarks of cancer. *Cell* **100**, 57–70 (2000). doi: 10.1016/S0092-8674(00)81683-9; pmid: 10647931
93. B. Alberts, The cell as a collection of protein machines: Preparing the next generation of molecular biologists. *Cell* **92**, 291–294 (1998). doi: 10.1016/s0092-8674(00)80922-8
94. Y. Cheng *et al*., Principles of regulatory information conservation between mouse and human. *Nature* **515**, 371–375 (2014). doi: 10.1038/nature13985; pmid: 25409826
95. H. Ding *et al*., Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nat. Commun.* **9**, 1471 (2018). doi: 10.1038/s41467-018-03843-3; pmid: 29662057
96. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008). doi: 10.1038/nbt.1511; pmid: 19029910
97. G. Teo *et al*., SAINTexpress: Improvements and additional features in Significance Analysis of INTeractome software. *J. Proteomics* **100**, 37–43 (2014). doi: 10.1016/j.jprot.2013.10.023; pmid: 24513533
98. E. L. Huttlin *et al*., The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425–440 (2015). doi: 10.1016/j.cell.2015.06.043; pmid: 26186194
99. M. E. Sowa, E. J. Bennett, S. P. Gygi, J. W. Harper, Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138**, 389–403 (2009). doi: 10.1016/j.cell.2009.04.042; pmid: 19615732
100. B. Yates *et al*., Genenames.org: The HGNC and VGNC resources in 2017. *Nucleic Acids Res.* **45**, D619–D625 (2017). doi: 10.1093/nar/gkw1033; pmid: 27799471
101. A. Grover, J. Leskovec, node2vec: Scalable Feature Learning for Networks. *KDD* **2016**, 855–864 (2016). doi: 10.1145/2939672.2939754; pmid: 27853626
102. J. Barretina *et al*., The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012). doi: 10.1038/nature11003; pmid: 22460905
103. F. Iorio *et al*., A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–754 (2016). doi: 10.1016/j.cell.2016.06.017; pmid: 27397505
104. J. Gao *et al*., Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013). doi: 10.1126/scisignal.2004088; pmid: 23550210
105. GTEx Consortium, The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015). doi: 10.1126/science.1262110; pmid: 25954001
106. R. M. Meyers *et al*., Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017). doi: 10.1038/ng.3984; pmid: 29083409
107. H. Caniza *et al*., GOssTo: A stand-alone application and a web tool for calculating semantic similarities on the Gene Ontology. *Bioinformatics* **30**, 2235–2236 (2014). doi: 10.1093/bioinformatics/btu144; pmid: 24659104
108. P. Resnik, Others, Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.* **11**, 95–130 (1999). doi: 10.1613/jair.514
109. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001). doi: 10.1023/A:1010933404324
110. F. Pedregosa *et al*., Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
111. F. Zheng, CliXO-1.0: v1.0 beta (2021); https://doi.org/10.5281/zenodo.4717237.
112. M. E. J. Newman, Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 8577–8582 (2006). doi: 10.1073/pnas.0601602103; pmid: 16723398
113. F. Zheng *et al*., HiDeF: Identifying persistent structures in multiscale 'omics data. *Genome Biol.* **22**, 21 (2021). doi: 10.1186/s13059-020-02228-4; pmid: 33413539
114. L. Jacob, G. Obozinski, J.-P. Vert, "Group Lasso with Overlap and Graph Lasso," paper presented at the Proceedings of the 26th Annual International Conference on Machine Learning (2009). doi: 10.1145/1553374.1553431
115. F. Zheng, HiSig (2021); https://doi.org/10.5281/zenodo.4722465. doi: 10.5281/zenodo.4722465
116. K. Ellrott *et al*., Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* **6**, 271–281.e7 (2018). doi: 10.1016/j.cels.2018.03.002; pmid: 29596782
117. J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010). doi: 10.18637/jss.v033.i01; pmid: 20808728
118. A. Mayakonda, D.-C. Lin, Y. Assenov, C. Plass, H. P. Koeffler, Maftools: Efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756 (2018). doi: 10.1101/gr.239244.118; pmid: 30341162
119. J. F. Hultquist *et al*., A Cas9 Ribonucleoprotein Platform for Functional Genetic Studies of HIV-Host Interactions in Primary Human T Cells. *Cell Rep.* **17**, 1438–1452 (2016). doi: 10.1016/j.celrep.2016.09.080; pmid: 27783955
120. E. K. Brinkman, T. Chen, M. Amendola, B. van Steensel, Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* **42**, e168 (2014). doi: 10.1093/nar/gku936; pmid: 25300484
121. F. Civitci *et al*., Fast and multiplexed superresolution imaging with DNA-PAINT-ERS. *Nat. Commun.* **11**, 4339 (2020). doi: 10.1038/s41467-020-18181-6; pmid: 32859909
122. R. Jungmann *et al*., Multiplexed 3D cellular super-resolution imaging with DNA-PAINT and Exchange-PAINT. *Nat. Methods* **11**, 313–318 (2014). doi: 10.1038/nmeth.2835; pmid: 24487583
123. A. D. Edelstein *et al*., Advanced methods of microscope control using μManager software. *J. Biol. Methods* **1**, e10 (2014). doi: 10.14440/jbm.2014.36; pmid: 25606571
124. X. Nan *et al*., Single-molecule superresolution imaging allows quantitative analysis of RAF multimer formation and signaling. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 18519–18524 (2013). doi: 10.1073/pnas.1318188110; pmid: 24158481
125. J. Schindelin *et al*., Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012). doi: 10.1038/nmeth.2019; pmid: 22743772
126. D. A. Beacham, M. D. Amatangelo, E. Cukierman, Preparation of extracellular matrices produced by cultured and primary fibroblasts. *Curr. Protoc. Cell Biol.* **33**, 10.19.11–10.19.21 (2006).
127. R. R. Lareu *et al*., Collagen matrix deposition is dramatically enhanced in vitro when crowded with charged macromolecules: The biological relevance of the excluded volume effect. *FEBS Lett.* **581**, 2709–2714 (2007). doi: 10.1016/j.febslet.2007.05.020; pmid: 17531987
128. M. J. Goldman *et al*., Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* **38**, 675–678 (2020). doi: 10.1038/s41587-020-0546-8; pmid: 32444850
129. M. Uhlen *et al*., A pathology atlas of the human cancer transcriptome. *Science* **357**, eaan2507 (2017). doi: 10.1126/science.aan2507; pmid: 28818916
130. K. L. Abbott *et al*., The Candidate Cancer Gene Database: A database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Res.* **43**, D844–D848 (2015). doi: 10.1093/nar/gku770; pmid: 25190456
131. F. Zheng, S. Zhang, C. Churas, HiDeF v1.0.0 (2020); http://doi.org/10.5281/zenodo.4059074.

ACKNOWLEDGMENTS

We thank C. Ng and J. Ma for very helpful comments and suggestions. **Funding:** Supported by the Cancer Cell Map Initiative (NCI U54 CA209891), the National Resource for Network Biology (P41 GM103504), and the Cytoscape Project (R01 HG009979). Also supported by NIH grant F32 CA239336 (B.T.); NCI grant 5F30CA236404-02 (E.S.); NIH grant R50 CA243885 (J.F.K.); NSF CAREER award MCB-1651855 and American Cancer Society Research Scholar Grant RSG-21-033-01-CSM (S.I.F. and M.L.H.); the Cancer Systems Biology Consortium (NCI grant U54 CA209988) and NIGMS grant R01 GM132322 (K.T. and X.N.); NIH K00 grant CA212456 (B.K.); the Martha and Bruce Atwater Breast Cancer Research Program, UCSF Prostate Cancer Program, and Benioff Initiative for Prostate Cancer Research (M.K.); and NIH award R01DE026870 (J.S.G.). This publication includes data generated at the UC San Diego IGM Genomics Center utilizing an Illumina NovaSeq 6000 that was purchased with funding from NIH SIG grant S10 OD026929. **Author contributions:** Conceptualization: T.I., N.J.K., F.Z., M.R.K., J.S.G., M.K., D.L.S., X.N., S.I.F. Data curation: F.Z., M.R.K., M.K., D.L.S., K.O., E.S., R.T.P., D.N.P., N.W., J.P., B.K. Formal analysis: F.Z., M.R.K., D.J.R., K.T. Funding acquisition: T.I., N.J.K., J.S.G., S.I.F., X.N., D.L.S., M.K., J.F.K., D.P. Investigation: F.Z., M.R.K., D.J.R., K.T., M.H., B.T., J.L., M.K., D.L.S., K.O., H.F., M.C., K.A.H., J.F.K., K.L. Methodology: F.Z., M.R.K., D.J.R., K.T., M.H., B.T., J.L., K.O., H.F., M.C., M.K.Y., K.A.H., A.K., J.P. Project administration: T.I., N.J.K., J.S.G., S.I.F., X.N., D.L.S., M.K., J.F.K., D.P., K.L. Resources: T.I., N.J.K., J.S.G., S.I.F., X.N., D.L.S., M.K., J.F.K., D.P., K.L., B.K., E.S., K.A.H. Software: F.Z., K.O., S.N.L., J.C., C.C., A.K., M.K.Y., D.P.; Supervision: T.I., N.J.K., M.R.K., F.Z., K.L., D.P., J.F.K., M.K., D.L.S., X.N., S.I.F., J.S.G. Validation: F.Z., M.R.K., T.I., D.J.R., K.T., M.H., B.T., J.L., H.F., M.C., K.A.H., N.W. Visualization: F.Z., M.R.K., T.I., D.J.R., K.T., M.H., K.O., S.N.L., J.F.K., M.K., D.L.S., X.N., S.I.F. Writing–original draft: F.Z., T.I., J.F.K. Writing–review and editing: F.Z., M.R.K., T.I., D.J.R., K.T., M.H., J.F.K., M.K., D.L.S., X.N., S.I.F., J.S.G., N.J.K. **Competing interests:** T.I. is co-founder of Data4Cure Inc., is on its scientific advisory board, and has an equity interest. T.I. is on the scientific advisory board of Ideaya BioSciences Inc., has an equity interest, and receives sponsored research funding. The terms of these arrangements have been reviewed and approved by the University of California San Diego in accordance with its conflict-of-interest policies. J.S.G. is a member of the board of scientific advisors for Vividion, Oncoceutics Pharmaceuticals, and Domain Pharmaceuticals (not directly relevant to this study). The laboratory of N.J.K. has received research support from Vir Biotechnology and F. Hoffmann-La Roche. N.J.K. is a shareholder of Tenaya Therapeutics, has received stock from Maze Therapeutics and Interline Therapeutics, and has consulting agreements with the Icahn School of Medicine at Mount Sinai, Maze Therapeutics, and Interline Therapeutics. **Data and materials availability:** All data are available in the main text and supplementary materials. The software used in this work is archived on Zenodo (*111*, *115*, *131*).

SUPPLEMENTARY MATERIALS

https://science.org/doi/10.1126/science.abf3067
Supplementary Text
Figs. S1 to S8
Tables S1 to S8
Data S1 to S3
References (*132*–*143*)
MDAR Reproducibility Checklist

17 October 2020; accepted 28 July 2021
10.1126/science.abf3067

Science

Interpretation of cancer mutations using a multiscale map of protein systems

Fan ZhengMarcus R. KellyDana J. RammsMarissa L. HeintschelKai TaoBeril TutuncuogluJohn J. LeeKeiichiro OnoHelene FoussardMichael ChenKari A. HerringtonErica SilvaSophie N. LiuJing ChenChristopher ChurasNicholas WilsonAnton KratzRudolf T. PillichDevin N. PatelJisoo ParkBrent KuenziMichael K. YuKatherine LiconDexter PrattJason F. KreisbergMinkyu KimDanielle L. SwaneyXiaolin NanStephanie I. FraleyJ. Silvio GutkindNevan J. KroganTrey Ideker

Science, 374 (6563), eabf3067. • DOI: 10.1126/science.abf3067

Mapping protein interactions driving cancer

Cancer is a genetic disease, and much cancer research is focused on identifying carcinogenic mutations and determining how they relate to disease progression. Three papers demonstrate how mutations are processed through networks of protein interactions to promote cancer (see the Perspective by Cheng and Jackson). Swaney *et al*. focus on head and neck cancer and identify cancer-enriched interactions, demonstrating how point mutant–dependent interactions of PIK3CA, a kinase frequently mutated in human cancers, are predictive of drug response. Kim *et al*. focus on breast cancer and identify two proteins functionally connected to the tumor-suppressor gene BRCA1 and two proteins that regulate PIK3CA. Zheng *et al*. developed a statistical model that identifies protein networks that are under mutation pressure across different cancer types, including a complex bringing together PIK3CA with actomyosin proteins. These papers provide a resource that will be helpful in interpreting cancer genomic data. —VV

Science (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.