Oncoprotein-specific molecular interaction maps (SigMaps) for cancer network analyses

HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2021 February 12.

Published in final edited form as:

Nat Biotechnol. 2021 February ; 39(2): 215–224. doi:10.1038/s41587-020-0652-7.

Oncoprotein-Specific Molecular Interaction Maps (SIGMAPs) for cancer network analyses

Joshua Broyde[1,*], **David R. Simpson**[2,*], **Diana Murray**[1,*], **Evan O. Paull**[1], **Brennan W. Chu**[1], **Somnath Tagore**[1], **Sunny J. Jones**[1], **Aaron T. Griffin**[1], **Federico M. Giorgi**[3], **Alexander Lachmann**[4], **Peter K. Jackson**[5], **E. Alejandro Sweet-Cordero**[2], **Barry Honig**[1,6,7,8,9], **Andrea Califano**[1,6,7,10,11,12,13]

[1]Department of Systems Biology, Columbia University Medical Center, New York, NY 10032, USA

[2]Division of Pediatric Hematology/Oncology, Department of Pediatrics, UCSF Benioff Children's Hospital, San Francisco, CA 94158, USA

[3]Cancer Research UK Cambridge Institute, University of Cambridge, Robinson Way, Cambridge, CB2 1TN, UK

[4]Mount Sinai Center for Bioinformatics; Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place Box 1603, New York, NY 10029 USA

[5]Baxter Laboratory, Department of Microbiology & Immunology, Stanford University, Palo Alto, CA 94305, USA; Department of Pathology, Stanford University, Palo Alto, CA 94305, USA

[6]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA

[7]Department of Medicine, Columbia University, New York, NY 10032, USA

[8]Zuckerman Mind Brain and Behavior Institute, Columbia University, New York, NY 10027, USA

[9]Howard Hughes Medical Institute, Columbia University, New York, NY 10032, USA

[10]JP Sulzberger Columbia Genome Center, Columbia University Medical Center, New York, NY 10032, USA

[11]Department of Biomedical Informatics, Columbia University, New York, New York 10032, USA

[12]Institute for Cancer Genetics, Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, NY 10032, USA

Corresponding Authors: Sweet-Cordero, E. Alejandro (Alejandro.Sweet-Cordero@ucsf.edu), Honig, Barry (bh6@cumc.columbia.edu), Califano, Andrea (ac2248@cumc.columbia.edu).

*These authors contributed equally

Author Contributions

J.B., D.M., A.T.G., A.L. and F.M.G. performed computational analysis. J.B. performed machine learning. D.S. and P.K.J. performed, respectively, the knockdown and AP/MS experiments. E.O.P, B.W.C., and S.T created OncoSig.org. J.B., D.M., and S.J.J. compiled the codebook. J.B., D.S., E.A.S.C., A.C. and B.H. analyzed knockdown experiments. SD.M., B.H., A.C., and E.A.S.C. designed research. D.M. B.H and A.C designed computational and experimental work and analyzed data. D.M., D.S. J.B., A.C., and B.H assembled the data and wrote the paper.

Conflicts of Interest

A.C. is founder and equity holder of DarwinHealth Inc., a company that has licensed some of the algorithms used in this manuscript from Columbia University. Columbia University is also an equity holder in DarwinHealth Inc.

[13]Motor Neuron Center and Columbia Initiative in Stem Cells, Columbia University, New York, New York 10032, USA

Abstract

Tumor-specific elucidation of physical and functional oncoprotein interactions could improve tumorigenic mechanism characterization and therapeutic response prediction. Current interaction models and pathways, however, lack context-specificity and are not oncoprotein-specific. We introduce SigMaps as context-specific networks, comprising modulators, effectors and cognate binding-partners of a specific oncoprotein. SigMaps are reconstructed *de novo* by integrating diverse evidence sources—including protein structure, gene expression, and mutational profiles—via the OncoSig Machine Learning framework. We first generated a KRAS-specific SigMap for lung adenocarcinoma, which recapitulated published KRAS biology, identified novel synthetic lethal proteins that were experimentally validated in 3D spheroid models, and established uncharacterized crosstalk with Rab/Rho. To show that OncoSig is generalizable, we first inferred SigMaps for the ten most mutated human oncoproteins and then for the full repertoire of 715 proteins in the COSMIC Cancer Gene Census. Taken together, these SigMaps (interactive analyses at http://www.OncoSig.org) show that the cell's regulatory and signaling architecture is highly tissue specific.

Editors summary

Tumor-specific molecular-interaction signaling maps are produced for any oncoprotein using a machine learning algorithm.

INTRODUCTION

The explosion of 'omics data in cancer has fueled numerous efforts to elucidate pathways that underlie signaling and regulatory programs. However, these typically fail to provide sufficient detail to fully capture the complexity of the regulatory and signaling mechanisms responsible for mediating the effect of both genetic and pharmacological perturbations. Although networks derived from pairwise-interaction assays or computational inference may mitigate the excessive simplicity and linearity of cancer pathways, they generally do not account for nor discriminate between cellular contexts[1]. Recent approaches have started to address the challenge of "context-specific interactions" by incorporating cell line-, tumor-, or tissue-specific information[2-4]. However, comprehensive, proteome-wide depiction of human interactomes across different tissue contexts remains elusive.

To address these challenges, we developed an integrative Machine Learning (ML) framework (OncoSig) for the systematic, *de novo* reconstruction of tumor-specific molecular-interaction Signaling Maps (SigMaps), anchored on any oncoprotein of interest. Specifically, as shown in a conceptual diagram (Figure 1a), an oncoprotein-specific SigMap recapitulates the molecular architecture necessary to functionally modulate and mediate its activity within a specific cellular context, including its physical, cognate binding partners. To illustrate the unique information contained in SigMaps and to show their sensitivity and

specificity in elucidating novel biology, based on experimental validation, we first focus on KRAS and then show that SigMaps can be generalized to virtually any protein.

RESULTS

OncoSig Evidence Sources and Integration

As conceptually depicted (Figure 1b), OncoSig infers context-specific SigMaps by training a ML algorithm to integrate complementary evidence from transcriptional and post-translational interactions inferred from 3D-structural data, as well as from gene expression and mutational profiles in large-scale repositories, such as The Cancer Genome Atlas (TCGA)[5]. Nodes in Figure 1b are color-coded, as in Figure 1a, to highlight the functional role of OncoSig proteins.

Without loss of generality, we discuss the four complementary evidence sources integrated by OncoSig (Figure 1b). Additional evidence is easily incorporated in the framework. We first illustrate their use in generating a lung adenocarcinoma-specific (LUAD) SigMap for KRAS ($\mathrm{SigMap}^{KRAS}_{LUAD}$) and then extend to other proteins and contexts.

First, we integrate KRAS-specific structure-based protein-protein interactions (PPIs), as inferred by the PrePPI (Predicting Protein-Protein Interactions) algorithm[6,7], by combining structural homology to protein complexes in structural databases and non-structure-related data. PrePPI scores represent the likelihood ratio of any predicted PPI based on a random-interaction null model. The specific 3D complexes that support each inference provide key mechanistic insights on the underlying interaction. To avoid curation bias, we removed evidence from Gene Ontology (GO)[8] and from other PPI datasets, which were included in the original version of the algorithm.

Second, we include transcriptional interactions inferred by ARACNe (Algorithm for the Reconstruction of Gene Regulatory Networks)[9], a broadly-adopted reverse engineering algorithm that uses information theory to identify high-probability, direct transcriptional interactions (or least-indirect ones for signaling protein).

Third we use VIPER (Virtual Inference of Protein activity by Enriched Regulon analysis)[10] to associate the mutational state of a candidate upstream modulator protein with differential activity of the anchor protein or the mutational state of the latter with differential activity of its candidate downstream effectors. VIPER measures a protein's activity based on the expression of its transcriptional targets—akin to a tissue-specific, highly-multiplexed gene reporter assay.

Finally, we infer upstream modulators of the anchor-protein using CINDy (Conditional Inference of Network Dynamics)[11,12]—a refinement of the MINDy algorithm[11]. CINDy uses the conditional mutual information to assess changes in the mutual information between the anchor-protein and its transcriptional targets as a function of the candidate modulator expression or mutational state.

OncoSig further accounts for tumor-context specificity by leveraging evidence from the ARACNe, VIPER, and CINDy algorithms, whose predictions are based on the analysis of large-scale, tumor-specific molecular profile data, while PrePPI provides context-independent, structure-based evidence. For KRAS, for instance, analyses were based on sample-matched gene expression and mutational profiles from 488 LUAD samples in The Cancer Genome Atlas (TCGA)[13], including 326 $KRAS^{WT}$ and 134 $KRAS^{Mut}$ samples, as well as 28 samples lacking KRAS mutational information. VIPER assessment of protein activity covers ~6,200 proteins annotated as transcription factors (TFs), transcriptional cofactors (coTFs), and signaling proteins (SPs), based on Gene Ontology (GO) classification[8].

To integrate the evidence from these algorithms, we tested two established ML-algorithms: Naïve Bayes[14] (NB) and Random Forest[15] (RF). An advantage of the former is that the inference of specific protein-protein relationships can be easily traced back to their supporting evidence. In contrast, the latter is better suited to integrate non-statistically-independent evidence sources. However, tracing back predictions to the specific supporting evidence is challenging.

Input to the algorithm is schematically depicted as a matrix (Figure 1c), with ~20K rows—one for each protein in the human proteome. The Gold Standard Set (GSS) vector in the 2nd column describes proteins known to be functionally related (GSS^P) or unrelated (GSS^N) to the anchor protein (e.g., KRAS). To standardize the training process and facilitate extension to other proteins, the GSS^P_{KRAS}, comprising 250 proteins, was derived from the mSigDB–C2 curated gene set collection[16] and KEGG[17] (Supplementary File 1). The GSS^P_{KRAS} includes established KRAS-related proteins, such as RAF1 and PTPN11, whereas the GSS^N_{KRAS}, includes all other proteins (Figure 1c). Remaining columns represent ~36,000 features (independent variables) corresponding to PrePPI, ARACNe, CINDy, and VIPER confidence scores, respectively, supporting physical or functional interactions between row-specific and column-specific proteins (names shown in the second row). For each protein, statistically significant algorithms scores are reported in different columns, hence accounting for the greater number of columns than rows (Figure 1b), thus producing a very sparse matrix with most of its ~720M cells (20K x 36K) empty. For example, consider the LIMD1 row (LIM domain containing protein)—an adaptor protein in cytoskeleton organization and cell fate determination, not reported as functionally related to KRAS (light green highlight in Figure 1c). As shown, LIMD1 is predicted to physically interact with CTNNB1 by PrePPI, has ARACNe targets that significantly overlap those of SETD2, is predicted by CINDy to be post-translationally modulated by UBP1 and SETD2. Consistent with these results, its VIPER-inferred activity is significantly affected by SETD2 and CTNNB1 mutations.

The entire matrix, except for the Gold Standard column, is identical for any anchor protein of choice. Additionally, evidence sources are not restricted to those of this study and can be easily expanded in additional columns. The NB algorithm is trained using only scores representing interactions with the anchor protein, while the RF algorithm is trained on the entire matrix, thus accounting for the full network architecture.

When trained and cross-validated on the same GSS_{KRAS} (see methods), there was highly significant overlap between NB-based (Extended Data Figure 1) and RF-based (Figure 2) predictions (Supplementary File 2) ($p < 10^{-39}$ for Ingenuity-based GSS; $p < 10^{-103}$ for KEGG/MSigDB-based GSS, by GSEA analysis[18]) (Supplementary File 3, and see below). However, RF slightly outperformed NB (Extended Data Figure 2). As a result, we first focused on the RF-based version of the supervised learning algorithm ($OncoSig_{RF}$) to predict the $SigMap_{LUAD}^{KRAS}$. The algorithm was trained and tested using random, non-overlapping subsampling of the rows (Monte Carlo cross-validation, see Methods). For each random selection, the RF classifier was trained with 50 forests, each comprising 50 decision trees. A score $S_{RF} = 0.5$ means that 50% of the trees in all forests supported a protein's inclusion in the SigMap. We thus define novel predictions as proteins with $S_{RF} \geq 0.5$ that do not belong to the GSS_{KRAS}^{P} (i.e., not known as KRAS-related in the literature).

The KRAS SigMap

Receiver operating curve (ROC) analysis, by Monte Carlo cross-validation (Figure 2a), shows that out of 250 proteins in the GSS_{KRAS}^{P}, 61 (Recall = 24%) and 140 (Recall = 56%) were recovered at highly conservative False-Positive Rates (FPR = 1% and FPR = 5%), respectively (**red curve**). Figure 2b expands the portion of the ROC curve to show only the highest-confidence predictions (FPR ≤ 1%, $S_{RF} > 0.88$)—see Supplementary File 3 for the complete ranking of all human proteins. Predictions for established KRAS-related proteins (GSS_{KRAS}^{P}) are shown as gold circles ($N = 61$ of 262, True Positive Rate, TPR = 23%). These include a wide range of established KRAS-pathway proteins (labeled), such as other RAS superfamily members, growth factor receptors, P53, PI3K, and MAPK protein kinases (Figure 2b(i), gold circles in upper panel), emphasizing the SigMap concept's value in prioritizing established KRAS-pathway members and delineating crosstalk with other canonical pathways. While SOS1 and NF1—the canonical KRAS GEF and GAP—are not included in the FPR ≤ 1% set, they were both identified as highly significant, at FPR = 1.1% and 2.7%, respectively.

Since ROC statistics assess a classifier's ability to recover what is known, performance can be significantly underestimated, because correct, yet previously unknown predictions would be considered false positives (see experimental validation section). Moreover, since these predictions are LUAD-specific, whereas the gold standard ($PGSS_{KRAS}$) is non-context-specific, a 24% and 56% recall, at FPR = 1% and 5%, respectively, should be considered very high, especially when compared to co-expression-based methods (green curve, Figure 2a) or high-throughput experimental assays—such as Y2H or pull-down followed by tandem mass-spec. The latter typically have a 10% – 30% recall[19,20]. For instance, the KEGG/MSigDB and Ingenuity datasets used to assemble the $PGSS_{KRAS}$ share only 29 proteins, yielding a much lower 12% and 8% recall, when compared against each other, thus showcasing the imperfect nature of these resources.

In addition to the 61 True Positive predictions, the $SigMap_{LUAD}^{KRAS}$ includes 201 novel predictions at FDR = 1% (red circles) (Figure 2b(ii)), including 30 proteins predicted as KRAS physical interactors (11%, blue and black circles) in BioGRID[21], 134 druggable

proteins (51%, red and black circles) in the Drug Repurposing Hub[22], and 33 proteins meeting both criteria (13%, black circles). Furthermore, 190 of 478 (40%) and 136 of 401 (34%) KRAS interactors in STRING[23] and HumanNet[24] were recapitulated by the SigMap, respectively, while ~1,700 novel predictions at FDR = 10% significantly expand KRAS biology knowledge. Taken together, these data strongly suggest that $\text{OncoSig}_{\text{RF}}$-inferred SigMaps effectively recapitulate known biology and can help prioritize novel functional or physical interactors, including many druggable ones, for validation.

Figure 2c shows the $\text{SigMap}_{LUAD}^{KRAS}$ as conceptually described in Figure 1a, (KRAS is the red node). To avoid overcrowding, we only include the top 68 predicted proteins (FPR ≤ 1%) supported by either VIPER ($p \leq 0.05$) or PrePPI evidence. This graph is provided strictly for illustrative purposes as it would be impossible to include all 262 proteins without compromising legibility. The architecture presentation follows Figure 1b (Supplementary File 3) and provides information that would be missing in typical pairwise-interaction networks, allowing formulation of mechanistic hypotheses for more effective design of experimental assays.

KRAS SigMap Validation in Primary Tumor Organoids

We selected the top 20 novel predictions by $\text{OncoSig}_{\text{NB}}$, trained on Ingenuity, i.e., not included in the $\text{GSS}_{\text{KRAS}}^{\text{P}}$ (see methods) and estimated the False Positive Rate (FPR) based on $\text{OncoSig}_{\text{RF}}$ trained on MSigDB/KEGG. This provides a reproducible methodology to estimate realistic experimental FPRs. The validation set comprised APPL1, ARHGAP26, ARHGDIA, ARHGEF1, IPO7, MAP3K6, MINK1, RAB13, RAB14, RAB1A, RAB25, RAB27A, RAB3D, RAB8A, RHOG, RHOT1, RPS6KA5, TBC1D4, VAV3, YWHAH (see Extended Data Figure 1b, Supplementary File 2, and Methods for details).

We performed a pooled, shRNA-based loss-of-function screen in 3D organoid cultures derived from a LUAD $\text{KRAS}^{\text{G12D/+}}/\text{p53}^{-/-}$ mouse model (Figure 3a).

To identify proteins that are essential for primary KRAS^{Mut} tumors to grow as organoids, primary tumor cells were isolated and separated from non-tumor stroma by lineage-depletion, prior to lentiviral infection with a lentiviral-mediated shRNA pool targeting each novel prediction (Extended Data Figure 1b), with 3 to 5 hairpins/gene. Positive controls included TBK1[25] and NUP205[26] (bright green), both established KRAS^{Mut} synthetic lethal partners, and established KRAS effectors MAPK1, AKT1, RALGDS, and RASA1 (purple). As negative controls, to estimate the background rate of essential genes, we used 25 independent shRNAs pools, targeting 515 different genes not expected to participate in KRAS signaling (BPS, black). Negative controls were screened in multiple independent pools such that all pooled libraries—whether representing OncoSig-inferred genes (red), positive controls (green and purple) or negative controls (black)—were of similar size (~100 individual shRNAs per pool). All shRNA sequences are provided in Supplementary File 4.

Comparison of the ranked $\log_2$(FC) (Fold Change) between 6 days and 12 days, for each individual gene, is shown in Figure 3b (ranked by average $\log_2$(FC) across all targeting hairpins) and in Extended Data Figure 3 (ranked by each hairpin). Consistent with previous

studies[25], growth was significantly inhibited in organoids incorporating shRNAs targeting known synthetic lethal and known KRAS signaling genes (green and purple dots). Strikingly, however, a majority of shRNAs targeting predicted KRAS SigMap proteins (red dots) also inhibited organoid growth, confirming statistically significant enrichment in $KRAS^{Mut}$ dependencies. The average $\log_2$(FC) distribution for novel predictions (red) is highly skewed toward lower values with >1/3rd of novel predictions showing a four-fold or greater decrease (Figure 3c, red curve; mean $\mu = -0.852$, $\sigma = 1.952$). In contrast, only 3 of the 515 negative controls (black dots) significantly affected organoid viability, thus producing an average $\log_2$(FC) distribution centered around zero, with most shRNAs showing ≤ 2-fold effect (Figure 3c, bold black curve; $\mu = -0.067$, $\sigma = 0.539$). The difference between the two distributions in Figure 3c is highly statistically significant ($p \leq 2.2\times10^{-16}$, Kolmogrov-Smirnoff test), suggesting that predicted KRAS SigMap partners are indeed highly enriched in $KRAS^{Mut}$ dependencies, compared to genes in the BPS negative control set.

Of the 20 novel KRAS SigMap genes, 16 (80%) produced statistically significant average fold-change reduction in organoid growth (FDR < 0.05). Figure 3d shows the fold-change (FC) viability reduction ($\log_2$(FC)), averaged after removing up to 2 outlier hairpins, for these statistically significant genes, as a function of their significance ($\log_{10}$(FDR-adjusted p-values) (Supplementary File 2). Many of the novel KRAS SigMap predictions (e.g. RAB1A, TBC1D4, RAB25, ARHGDIA; red) inhibited organoid growth as well as, or better than, established members of KRAS signaling pathways (green) and proteins representing dependencies of $KRAS^{Mut}$ cells purple).

We estimated the probability that 16 of the 20 predicted genes would be validated based on their individual p-values with an empirical multinomial model and found 95% of the simulations produced 2 or fewer false positives. Considering that not all KRAS interactors are $KRAS^{Mut}$ synthetic lethal, this is roughly consistent with a 16/20 validation rate (i.e. 4 false positives).

Moreover, the difference between OncoSig-predicted (80%) and negative-control genes (0.58%) inducing statistically significant viability reduction was highly significant ($p = 2.7 \times 10^{-24}$; Fisher's Exact Test).

As described in recent reviews, there is a wealth of studies to assess oncogenic $KRAS^{Mut}$ dependencies[25,27,28]. To further benchmark our predictions, we thus assessed the extent to which $OncoSig_{RF}$ may recapitulate these results. Since some of the genes in these studies appear in the GSS^{P}_{KRAS}, we retrained $OncoSig_{RF}$ on a modified, non-overlapping training set for each analysis. The modified GSS^{P}_{KRAS} used to retrain $OncoSig_{RF}$ and the resulting KRAS SigMaps are provided in Supplementary Files 1 and 3 respectively. Our results show that, (a) top-scoring $OncoSig_{RF}$ predictions were highly enriched in $KRAS^{Mut}$ synthetic-lethal partners identified by Barbie et al.[25] ($p = 2.4 \times 10^{-14}$, Extended Data Figure 4a); (b) the 24 genes showed by Hayes et al.[28] to contribute to ERK-inhibitor resistance in $KRAS^{Mut}$ cells were also highly enriched in SigMap proteins ($p = 3.4 \times 10^{-7}$, Extended Data Figure 4b); and (c) genes inducing oxidative stress-mediated lethality in $KRAS^{Mut}$ cells[8,29] were

similarly enriched ($p = 2.0 \times 10^{-11}$, Extended Data Figure 4c). These results, along with the high rate of experimental validation for KRAS SigMap predictions (Figure 3b), strongly support the ability of OncoSig$_{RF}$ to identify effector and modulator proteins representing *bona fide* KRAS^{Mut} dependencies.

Crosstalk in KRAS and RAB/RHO signaling

KRAS regulation of RAB and RHO GTPase signaling remains poorly understood[30,31]. As shown in Figure 2, OncoSig predicts novel physical and functional interactions between KRAS and RAB/RHO family members and regulators. Indeed, among the experimentally validated proteins, OncoSig identified RAB-family members/regulators RAB1A, RAB8A, RAB14, RAB25, RAB27A, TBC1D4, and APPL1, as well as RHO-family members/regulators RHOG, RHOT1, ARHGDIA, *VAV3* and *ARHGAP26* as downstream effectors and physical binding partners of KRAS and predicted additional RAB/RHO-family members, including RAB1B, RAB8B, and RAB32, and CDC42. Among these, RAB1A has emerged as a novel putative oncogene, stimulating tumorigenic growth independent of HRAS signal transduction[32], TBC1D4 is a putative RAB GTPase activating protein)[33], and APPL1 is an adapter protein binding to the GTP-bound, active form of RAB5[30]. Similarly, *ARHGDIA* is a RHO GDP-dissociation inhibitor (GDI) and negative regulator of RHO/RAC signaling[34]; *VAV3* and *ARHGEF16* are guanine nucleotide exchange factors (GEFs) for RHO proteins[35,36]; finally, *ARHGAP26* is a CDC42 activator.

Taken together, these results suggest far more extensive crosstalk between KRAS and RAB/RHO signaling than previously appreciated[37,38]. They further suggest that KRAS-mediated post-translational regulation of other small-GTPases may be dysregulated in LUAD.

Validated and predicted OncoSig interactions between KRAS and RAB family members also suggest a compelling role of KRAS in the regulation of intracellular trafficking[31], which is still poorly characterized[39]. Four RAB-family members validated in the organoid assays are predicted as KRAS cognate binding partners by PrePPI (bold lines, Figure 2). RAB1A and RAB25 mediate ER-Golgi trafficking and transport through apical recycling endosomes, respectively[40], while *TBC1D4* and *APPL1* promote endosomal vesicular trafficking[30,41]. MAP4Ks, such as *MINK1* (MAP4K6, downstream) and MAP4K1 (upstream), are also implicated in vesicular trafficking through their association with Striatin family complexes, whose dysregulation leads to cancer[42].

KRAS SigMap context specificity

To assess whether OncoSig$_{RF}$ can effectively discriminate KRAS SigMap context-specificity, we compared LUAD-specific predictions with predictions based on 482 TCGA lung squamous cell carcinoma (LUSC) samples, 434 TCGA colon adenocarcinoma (COAD) samples, and 176 TCGA pancreatic adenocarcinoma samples (PAAD). These data were used to create LUSC, PAAD and COAD-specific ARACNe, CINDy, and VIPER-based molecular interaction predictions, which were then used to train the RF classifier (Figure 1b).

Extended Data Figure 5a shows the ROC curves for $SigMap^{KRAS}_{LUAD}$ (red), $SigMap^{KRAS}_{LUSC}$ (gray) $SigMap^{KRAS}_{COAD}$(brown), and $SigMap^{KRAS}_{PAAD}$ (orange) (Supplementary File 5). These show that $OncoSig_{RF}$'s performance was equivalent in LUAD, LUSC and PAAD but weaker in COAD: At FPR ≤ 5%, only 12% of the GSS^{P}_{KRAS} was recovered in COAD versus 16% for LUAD, likely due to a much larger number of KRAS-related studies in LUAD vs. COAD. However, the $SigMap^{KRAS}_{COAD}$ and $SigMap^{KRAS}_{LUAD}$ were equally enriched in COAD and LUAD-specific $KRAS^{Mut}$ synthetic-lethal partners[27], respectively ($p = 3.3 \times 10^{-11}$, Extended Data Figure 5b).

Extended Data Figure 5 shows scatterplots of $OncoSig_{RF}$ scores for KRAS SigMap proteins in LUSC-vs-LUAD (Figure 4a), COAD-vs-LUAD (Figure 4b), and PAAD-vs-LUAD (Extended Data Figure 5c). Gold and gray points represent GSS^{P}_{KRAS} proteins and novel predictions, respectively. Darker colored points have high scores ($S_{RF} \geq 0.5$) in at least one context and lighter colored points score poorly in both contexts ($S_{RF} \leq 0.5$). The significant overlap between these maps ($R^2_{LUSC/LUAD} = 0.35$, $p < 10^{-267}$; and $R^2_{COAD/LUAD} = 0.10$, $p < 10^{-165}$) suggests the existence of a core of context-independent physical/functional KRAS interactors (see Supplementary File 5). However, there were also many off-diagonal points, representing both established and novel KRAS interactors that are specific to only one of the two contexts. Even in two relatively related lung cancer contexts (LUAD and LUSC)—as shown by a 100-fold improvement in overlap *p*-value, compared to COAD—critical differences are predicted (Figure 4b). For instance, CSF1 (macrophage colony-stimulating factor 1, black) is a LUSC-specific, significant survival marker[43], while downregulation of RASAL2 (a Ras GTPase-activating protein, green) promotes metastatic progression in LUAD[44]. Similar meaningful context-specific differences are predicted in COAD versus LUAD (Figure 4c). For instance, IFITM1 (interferon-induced transmembrane protein 1, black) promotes COAD-specific metastatic progression[45], while IL22RA1 (interleukin-22 receptor 1, green) is a LUAD-specific marker of poor survival[46,47]. For visualization purposes, we only show the top 33 proteins (FPR ≤ 1%, VIPER p ≤ 0.05, PrePPI-predicted physical interactors, or both) that overlap in the KRAS COAD and LUAD SigMaps (Extended Data Figure 5d) according to the conceptual SigMap architecture (Figure 1a). As shown, the underlying KRAS signaling architecture is quite different in the two contexts, consistent with the different activity of MEK inhibitors in COAD and LUAD[48-51].

PAAD to LUAD comparison did not identify substantial context-specificity. While this may simply reflect stronger conservation of KRAS biology in these two tumors, PAAD analysis is challenging because most samples (>90%) harbor KRAS mutations and virtually all present significant KRAS pathway activity. Thus, in this context, using $KRAS^{WT}$ samples as negative controls is not as effective as in cohorts where mutations occur only in a relatively small subset of the samples. This may also account for the failure to identify context-specific effectors and modulators.

Generating SigMaps for hyper-mutated oncoproteins

We extended $OncoSig_{RF}$ analysis by generating LUAD-specific SigMaps for the nine additional recurrently mutated oncoproteins (CDKN2A, EGFR, MAPK, NTRK3, PI3K,

TP53, STK11, YAP1, and CTNNB1). PGSS datasets were derived from mSigDB–C2 sand KEGG, for each oncoprotein (Supplementary File 1). We evaluated $\text{OncoSig}_{\text{RF}}$'s ability to recapitulate the GSS^{P} of each oncoprotein by ROC analysis (Figure 5a, KRAS ROC highlighted in red, Supplementary File 6). This confirmed significant improvement over random classification and gene-expression-based correlation ($p < 10^{-10}$ in all cases), see black and green curves in Figure 2a. Classifier performance is, however, variable, depending on both GSS_{KRAS} quality and LUAD specificity: For instance, at FPR ≤ 1%, 35%-37% of established NTRK3-, TP53-, and CDKN2A-pathway members were recovered (dark blue, dark gray, and light gray curves), vs. 12%-20% of established STK11-, YAP1-, and PI3K-pathway members (orange, light blue, and brown curves). Figure 5b shows that experimentally validated novel KRAS SigMap proteins (Figure 3), which have high score in the KRAS map, are also recapitulated in the other maps, albeit with lower scores, suggesting significant cross-talk. Indeed, while there is little overlap among the PGSSs, the overlap among SigMaps (at FPR ≤ 1%) is quite substantial (Extended Data Figure 6a). Predictions for each of the ten SigMaps are provided in Supplementary File 6.

As a measure of retrospective validation, we assessed the algorithm's ability to recapitulate proteins in a previously published, 600-protein EGFR-specific network[52], using a PGSS that excluded proteins in that network (Supplementary Files 1 and 3). SigMap predictions were highly enriched in EGFR-centric network proteins ($p = 2.3 \times 10^{-43}$, Extended Data Figure 6b). Furthermore, predictions were also highly enriched in 58 genes whose knockdown was shown to sensitize cells to EGFR-targeted inhibitors ($p = 1.4 \times 10^{-9}$). Finally, the EGFR SigMap discriminates between genes that sensitize cells to EGFR-targeting drugs versus those that do not ($p = 2.0 \times 10^{-4}$, by Welch's two sample t-test; Extended Data Figure 6c).

$\text{OncoSig}_{\text{RF}}$ Generalization

Finally, we developed an unsupervised version of the algorithm ($\text{OncoSig}_{\text{UN}}$) to extend the analysis to arbitrary proteins of interest, without protein-specific training sets. For this purpose, the feature matrix (Figure 1c) was reduced to contain only interactions with the specific protein of interest, leaving only four of the 36K columns, one for each of the algorithms (Figure 1b). Proteins were then scored based on aggregate voting across the ten $\text{OncoSig}_{\text{RF}}$ classifiers described above. The rationale is that once a sufficient number of diverse training sets is available, they can be used to assess the generic contribution of each evidence source (i.e., its weight) toward classification of a *bona fide* interaction. For comparison purposes, we tested the unsupervised version of algorithm on KRAS. As shown by ROC analysis (Figure 6a), the performance of $\text{OncoSig}_{\text{UN}}$ (blue curve) was similar—and even slightly better in the highest precision range—compared to $\text{OncoSig}_{\text{RF}}$ (red curve). Note that GSS_{KRAS} was used only to evaluate $\text{OncoSig}_{\text{UN}}$'s performance, but not to train the unsupervised algorithm. The KRAS $\text{OncoSig}_{\text{RF}}$ classifier was excluded from aggregate voting (see Methods).

We then applied the procedure to the Cancer Gene Census proteins in COSMIC (715 proteins, as of December 2018). SigMaps are provided in Supplementary File 7. We also provide examples, of the conceptual SigMap architecture (Figure 1a) for SMARCA4 (Figure 6b), MET (Figure 6c), and BIRC6 (Figure 6d), which are among the most LUAD-specific

mutated genes[53,54]. Their SigMaps are rich in their experimentally observed interactors from BioGRID[21] (blue node labels), drug targets from the Drug Repurposing Hub[22] (red node border), and functional partners as described by Gene Ontology[8]. The maps make a number of interesting predictions, further highlighting the utility of the generalized approach.

SMARCA4 is an ATP-dependent chromatin remodeling enzyme and transcriptional activator. As shown in Figure 6b, chromatin remodeling complex proteins are predicted both upstream (SMARCC2, CHD1) and downstream (ATAD2, TAF1) of SMARCA4[55-57]. Most of the proteins in the map (85%) are involved in chromatin organization (gray dashed box, GO:0016568) and fall into two Gene Ontology biological process subgroups: Histone acetylation (GO:0016573, light green shading) and chromatin remodeling (GO:0006338, light blue shading)[58]. $OncoSig_{UN}$ predicts mechanistic connections between chromatin remodeling/histone acetylation and 1) DNA damage signaling through the acetylation reader protein, ZMYND8[59], and 2) RNA splicing through the RNA helicase, DDX23, and the nuclear cap-binding protein, NCBP2[60,61].

The SigMap for the receptor tyrosine kinase (RTK) MET (Figure 6c) contains both established and novel interactions. RTKs (EPHA2 and INSR) are correctly predicted as upstream regulators, and tyrosine kinases (LCK and PTK2B) are correctly predicted as downstream effectors. Co-activation among RTKs and RTK activation of intracellular kinases well-established mechanisms[62,63]. Although Src-family kinases (like LCK) are effector proteins of MET as well as regulators of STATs[63,64], STAT1 is predicted to regulate LCK (green dotted arrow). However, there is ample evidence that STAT1 functions as both a tumor suppressor and a tumor promoter[65,66], and, recently, LCK was shown to be a critical gene for cell proliferation in KRAS-dependent lung cancers[67]. Thus, the predicted regulatory interaction may provide a mechanism for increased activation of LCK in LUAD.

DISCUSSION

The term “pathway,” though widely used, is a quite loosely defined biological concept. Here, we propose a fundamentally different representation (*SigMap*) of the signaling and regulatory machinery necessary to modulate and effect the function of a specific protein of interest in a specific tissue context, i.e. equivalent to a protein’s “Mechanism of Action” (MOA).

Our data suggest that SigMaps provide a more unbiased, compact, and realistic representation of a protein’s MoA, compared to available network representations and algorithms. Moreover, based on a wealth of experimental assays—both novel and previously published—SigMaps effectively recapitulated the complex biology of signal transduction, with a high validation rate for novel prediction (16/20, 80%) (Figure 3). Finally, we have shown that equally-informative SigMaps can be constructed for arbitrary proteins, whether cancer-related or not, thus extending the value of the algorithm well beyond cancer.

SigMaps may also provide critical hypotheses, often missing in related resources, regarding autoregulatory interactions (loops) (gray lines, Figure 1a), which are critical to ensure the

stability of cellular phenotypes and may be responsible for complex adaptive behavior, such as in response to pharmacological perturbations.

Targeting mutation-specific dependencies is an approach for discovering novel $KRAS^{Mut}$ specific therapeutics[68]. The KRAS SigMap is highly enriched in $KRAS^{Mut}$ dependencies, many of which are druggable (Figures 2 and 3). A number of previous studies[25,26,69-71] used high-throughput screens to discover $KRAS^{Mut}$ dependencies, although the overlap in their predictions is poor[25,68]. This is due to the context specificity of KRAS-mediated dependencies and synthetic lethality and on reliance on traditional monolayer cell line cultures. However, the enrichment of $OncoSig_{RF}$ predictions in $KRAS^{Mut}$ dependencies identified by other studies and the high validation rate achieved here (Figure 3) suggest that many additional *bona fide* and more reproducible modulators and effectors of KRAS function may be identified, even among predictions with relatively lower scores in experimental assays (Supplementary File 3) and can be defined by their tissue-specific context, thus further increasing the repertoire of druggable KRAS signaling partners. Thus, SigMaps may provide additional, pharmacologically accessible candidate targets for many mutated oncoproteins, including KRAS, thus providing a valuable resource for guiding hypothesis-based studies to validate their disease-related relevance.

In summary, OncoSig generates a single integrated score representing the probability that a protein belongs to a specific SigMap. Use of PrePPI is instrumental for identifying physical protein-protein interactions, whereas ARACNe, VIPER and CINDy provide critical tissue-specificity and additional evidence supporting both physical and functional interactions. Taken together, these individual evidence sources can effectively assign place proteins within the conceptual molecular-interaction architecture schematically depicted in Figure 1a.

The code used in these analyses is available from GitHub and a graphical web application (http://www.OncoSig.org) allows interactive query and visualization of all SigMaps generated by this study.

ONLINE METHODS

Lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), colon adenocarcinoma (COAD), and pancreatic adenocarcinoma gene expression datasets (N = 488, 482, 434, and 176 samples respectively) were retrieved from The Cancer Genome Atlas (TCGA) and normalized as previously described[10]. We collated 1,813 transcription factors and transcriptional regulators (TFs), 969 transcriptional cofactors (coTFs), and 3,370 signaling proteins (SPs) as described[10].

Naive Bayes-based OncoSig ($OncoSig_{NB}$)

To ascertain upstream regulators of KRAS, we inferred the activity of KRAS in LUAD samples using VIPER[10] and computed two-tailed Normalized Enrichment Score of the KRAS activity. aREA[10] was used to assess the statistical significance of the co-segregation between nonsynonymous (missense) Single Nucleotide Polymorphisms in other genes and KRAS activity. To identify downstream effectors of aberrant KRAS signaling, we used VIPER to infer the differential activity of TFs, CoTFs, and SPs in $KRAS^{Mut}$ samples and

closest (based on Spearman correlation) matched $KRAS^{WT}$ samples. A differential gene expression signature ΔE_i was computed for each matched $KRAS^{Mut}/KRAS^{WT}$ pair and the activity change for each TF/CoTF/SP was calculated. Bonferroni-corrected *p*-values were integrated, using Stouffer's method producing a *p*-value for the co-segregation of $KRAS^{Mut}$ and the activity of other proteins.

A variant of the DeMAND algorithm[73] was used to discover proteins with dysregulated interactions in $KRAS^{WT}$ versus $KRAS^{Mut}$ LUAD samples using a previously developed, context-specific LUAD molecular interaction network[10]. Specifically, instead of considering drug-perturbed vs. control samples, as in the original algorithm, we analyzed molecular interactions that were dysregulated in $KRAS^{Mut}$ versus $KRAS^{WT}$ samples and then, for each network protein, we integrated the *p*-value of its dysregulated edges using Brown's method as discussed in the original manuscript. The latter accounts for potential statistical dependencies between dysregulated edges.

MINDy was used to predict post-translational modifications of TFs by SPs, as previously described[11,12]. A Fisher Exact Test was performed between TFs predicted to be regulated by KRAS and TFs predicted to be regulated by SPs. Each SP was thus assigned a *p*-value representing the statistical significance of the overlap between the TFs KRAS is predicted to regulate and the TFs other signaling molecules are predicted to regulate.

Predictions of KRAS protein-protein interactions were retrieved from the PrePPI database[6]. Each prediction has an associated Likelihood Ratio (LR) representing the odds above random of the protein-protein interaction occurring.

75 samples with KRAS knockdowns (KDs) in A549 cell lines were retrieved from The Library of Network-Based Cellular Signatures (LINCS) project[74] (http://www.lincsproject.org/). Averaging over all 75 samples, a single gene expression profile was obtained for each gene.

Affinity-purification/mass-spec assays (AP-MS) were used to characterize candidate protein-protein interactions for four established KRAS effectors–TBK1, RALGDS, RALA, and RALB–in the $KRAS^{Mut}$ LUAD cell line A549. AP-MS scores correspond to the protein peptide count for each effector (Supplementary File 1).

$OncoSig_{NB}$ was trained on a set of 350 proteins annotated as participating in KRAS signaling pathways by Ingenuity Pathway Analysis[75] (Supplementary File 1). Each clue was split into bins, which were populated by the raw evidence values such that an equal number of members of the positive gold standard set (PGSS) was distributed across bins as possible. Training was performed using two-fold cross validation with holdout, which creates an independent training and testing set and produces a final LR for each protein that is parameterized on the set to which it does not belong. $OncoSig_{NB}$ results are provided in Supplementary File 2.

Random Forest-based OncoSig ($OncoSig_{RF}$)

The ARACNe, VIPER and CINDY algorithms were applied to LUAD, LUSC, COAD, and PAAD gene expression profiles with the same parameters and *p*-value thresholds as

previously described[10]. Protein-protein interactions from the PrePPI database were retrieved[7], and the LR_{GO} and LR_{Exp} components were removed from LR_{PrePPI}, and interactions with modified LR_{PrePPI} scores ≥ 600 were used.

The PGSSs for the 10 pathways used with $OncoSig_{RF}$ were compiled as the union of the KEGG, Biocarta, and Reactome databases from the MSigDB–C2 category[16] and complemented with associated pathway members from the KEGG website[17] (Supplementary File 1). KRAS synthetic lethality and drug-dependency data were compiled from Barbie et al.[25], Corcoran et al.[27], Hayes et al.[28], Astsaturov et al.[52] and GO:0000302: "Response to Reactive Oxygen Species"[8,29]. CD-HIT[76] was used to generate a non-redundant $PGSS_{KRAS}$. $PGSS_{KRAS}$ was clustered at an 80% sequence identity threshold and, for each cluster, the representative with the longest sequence was selected as per CD-HIT protocol.

The features derived from the networks were as follows: Mutual information for ARACNe, number of statistically significant triplets for CINDy, negative log *p*-value for VIPER, and LR for PrePPI. We coded each feature symmetrically, so that interactions between protein A and protein B were input into the matrix twice, once in the feature vector for A and once in the feature vector for B; all other elements in the in the Random Forest feature matrix were set to zero. For each of the 10 oncogene-centric interactomes, proteins that are part of the PGSS were assigned a "1" within the PGSS vector, while all other proteins were assigned a "0" to represent membership in the NGSS. $OncoSig_{RF}$ was trained and tested with each pathway's PGSS and NGSS using Monte Carlo cross-validation[77], creating 50 forests each with 50 trees. We performed 100 OncoSig runs each with the LUAD, LUSC, COAD and PPAD KRAS PGSS networks, and distributions of Pearson correlation coefficients were estimated by calculating all pairwise Pearson correlation coefficients. LUAD-specific SigMaps for KRAS and the other nine hyper-mutated oncoproteins are provided in Supplementary File 3 and Supplementary File 6, respectively. Tissue-specific KRAS SigMaps are provided in Supplementary File 5.

General procedure for SigMaps, OncoSig

The feature matrix for a protein of interest consists only of those features that correspond to interactions with the protein of interest and, thus, has only four columns, one for each of the four interactomes: PrePPI, ARACNe, CINDy, and VIPER. SigMap membership of each protein in the human proteome is determined by aggregate voting after querying the ten $OncoSig_{RF}$ classifiers described above (KRAS, PI3KCA, TP53, EGFR, BRAF, STK11, CDKN2A, NTRK3, YAP1, and CTNNB1). LUAD-specific SigMaps for Cancer Gene Census proteins are provided in Supplementary File 7.

Tandem affinity purification

5 mL packed cell volume of RPE-hTERT cells expressing LAP-tagged proteins were resuspended with 20 mL of LAP-resuspension buffer, lysed, and then incubated on ice for 10 min. The lysate was first centrifuged at 14,000 rpm (27,000 g) at 4°C for 10 min, and the resulting supernatant was centrifuged at 43,000 rpm (100,000 g) for 1 hr at 4°C to further clarify the lysate. High speed supernatant was mixed with 500 μL of GFP-coupled beads[78]

and rotated for 1 hr at 4°C to capture GFP-tagged proteins, and washed five times with 1 mL LAP200N. After re-suspending the beads with 1 mL LAP200N buffer lacking DTT and protease inhibitors, the GFP-tag was cleaved by adding 5 μg of TEV protease and rotating tubes at 4°C overnight. TEV-eluted supernatant was added to 100 μL of S-protein agarose to capture S-tagged protein. After washing three times with LAP200N buffer lacking DTT and twice with LAP100 buffer, purified protein complexes were eluted with 50 μL of 2X LDS buffer and boiled at 95°C for 3 min. Samples were then run on Bolt® Bis-Tris Plus Gels in Bolt® MES SDS Running Buffer. Gels were fixed in 100 mL of fixing solution at room temperature, and stained with Colloidal Blue Staining Kit. After the buffer was replaced with Optima™ water, the bands were cut into eight pieces, followed by washing twice with 500 μL of 50% acetonitrile in Optima™ water. The gel slices were then reduced and alkylated followed by destaining and in-gel digestion using 125 ng Trypsin/LysC as previously described[79]. Tryptic peptides were extracted from the gel bands and dried in a speed vac. Prior to LC-MS, each sample was reconstituted in 0.1% formic acid, 2% acetonitrile, and water. NanoAcquity (Waters) LC instrument was set at a flow rate of either 300 nL/min or 450 nL/min where mobile phase A was 0.2% formic acid in water and mobile phase B was 0.2% formic acid in acetonitrile. The analytical column was in-house pulled and packed using C18 Reprosil Pur 2.4 uM where the I.D. was 100 uM and the column length was 20-25 cm. Peptide pools were directly injected onto the analytical column in which linear gradients (4-40% B) were of either 80 or 120 min eluting peptides into the mass spectrometer. MS/MS was acquired using CID with a collisional energy of 32-35. In a typical analysis, RAW files were processed using Byonic (Protein Metrics) using 12 ppm mass accuracy limits for precursors and 0.4 Da mass accuracy limits for MS/MS spectra. MS/MS data was compared to an NCBI GenBank FASTA database containing all human proteomic isoforms with the exception of the tandem affinity bait construct sequence and common contaminant proteins. Spectral counts were assumed to have undergone fully specific proteolysis and allowing up to two missed cleavages per peptide. Data is included in Supplementary File 1.

Primary tumor propagating cell culture and screening methodology

Primary lung tumor cells from $KRAS^{G12D/+}$; $p53^{fl/fl}$ mice were cultured in Matrigel as described previously[80]. Prior to seeding, primary cells were infected with a pool of 100-150 lentiviral pLKO shRNAs composed of 3-5 shRNAs per gene at a Multiplicity of Infection <0.5 to ensure single shRNA integration and selected with 1ug/ml puromycin 24 hours after seeding (Supplementary File 4). We screened the top 22 predicted genes from the NB classifier in two pools. Pools also included other candidate vulnerabilities identified by literature review and other methods. 25 pools consisting of 2,286 shRNAs targeting 515 genes not anticipated to be involved in KRAS-regulated signaling were used as a background comparison. After 7 days of spheroid growth, spheroids were dissociated with trypsin into single cells, and half of the 3D culture was re-seeded. The remaining half of each sample was retained for gDNA isolation (T0) until secondary spheroids fully formed 7 days later (T1). The integrated pLKO shRNA was PCR amplified using ExTaq (Clontech), barcoded, multiplexed, and sequenced on an Illumina GAIIx (primer sequences available on request). Sequencing reads were processed into count files in R (v. 3.1.1) using the edgeR package (v. 3.6.8) and analyzed using generalized linear models with edgeR using a time-

course design to compare the initial (T0) and final (T1) timepoints and perform a likelihood ratio test[81].

To calculate the statistical significance of the fold change in growth induced by each individual shRNA, we fit a density plot of all the background screens. For each shRNA, we integrated from the minimum $\log_2$FC of the entire BPS to the $\log_2$FC observed for that shRNA, producing a one-tailed p-value for the observed $\log_2$FC. We used Fisher's method to integrate the p-values of all shRNAs that mapped to the same protein.

Multiple hypothesis correction

All p-values reported for all analyses (except where noted otherwise) were corrected using the Benjamini & Hochberg False Discovery Rate (FDR).

Enrichment analysis

Enrichment analysis was performed with the aREA (analytic Rank-based Enrichment Analysis) algorithm[10].

Empirical Multinomial Model

The goal of this analysis was to estimate the probability that N out of the 20 genes prioritized by $\text{OncoSig}_{\text{NB}}$ for validation would emerge as statistically significant KRAS^{Mut} synthetic lethals following RNAi-mediated silencing. We thus used the individual unadjusted p-values assessed by $\text{OncoSig}_{\text{RF}}$ analysis as input probabilities for a 10,000-iteration Monte Carlo simulation. The total number of false positives in each Monte Carlo simulation was calculated and used to compute an empirical probability distribution for the number of false positives in the set of 20 genes. From this empirical distribution, we calculated the expected number of false positive genes to be 0.8645 and that 95% of the probability was accounted for by ≤ 2 false positives. These predictions were roughly consistent with the results of the RNAi screen, where 16/20 genes predicted as significant by $\text{OncoSig}_{\text{RF}}$ (at FDR = 5%) were statistically significantly depleted in the pooled screen. This number was roughly double of what would have been expected (4 vs. 2) by the empirical multinomial analysis.

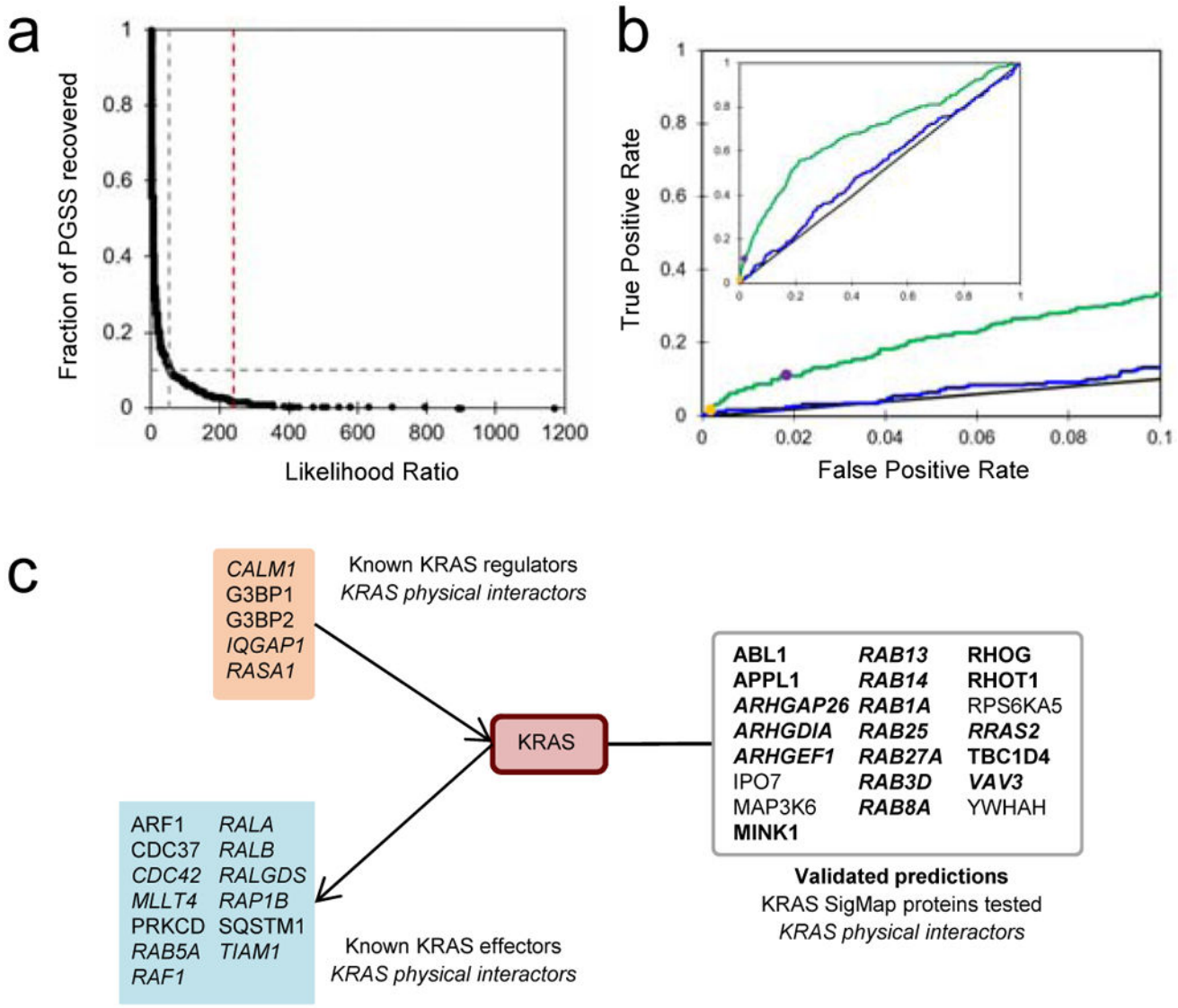
Data Availability Statement

All data are provided in Supplementary Files.

Code Availability Statement

R scripts and code for Naïve Bayes and Random Forest classifiers and input data files to reproduce the results described are freely available at https://github.com/califano-lab/OncoSig.

Extended Data



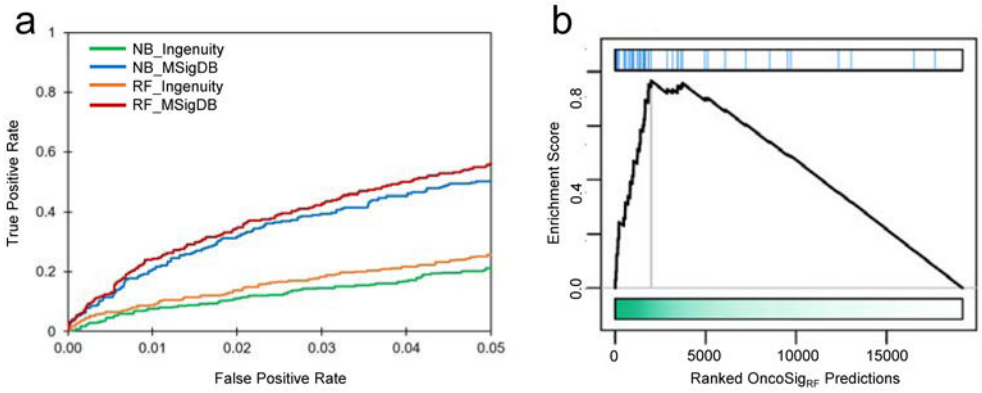
Extended Data Figure 1. The top 40 predictions from the OncoSig$_{NB}$ algorithm for the KRAS LUAD SigMap were chosen for validation.

(a): Performance of OncoSig$_{NB}$ at recovering the Ingenuity-derived PGSS as a function of LR$_{Post}$. At an LR$_{Post}$ = 53 (probability = 0.50) (vertical gray line), the OncoSig$_{NB}$ LUAD-specific KRAS SigMap contains 10% of the PGSS (horizontal gray line). The vertical red line corresponds to LR$_{Post}$ = 240, the cutoff used to obtain candidates for experimental validation.

(b): ROC curve analysis, evaluated as the recovery of the Ingenuity-derived PGSS (FPR ≤ 0.05), for 1) OncoSig$_{NB}$ (green curve, N = 1028)), 2) Pearson's correlation between mRNA expression of KRAS and mRNA expression of other proteins in LUAD (blue curve), and 3) random performance (black curve). Recovery using 2-fold cross-validation (green) is essentially indistinguishable from recovery using 100-fold Monte-Carlo Cross-validation (not shown). 393 OncoSig$_{NB}$ LUAD-specific KRAS SigMap predictions are made for

$LR_{Post} \geq 53$, which corresponds to probability ≥ 0.50 and FPR ≤ 0.019 (purple dot). 40 $OncoSig_{NB}$ LUAD-specific KRAS SigMap predictions are made for $LR_{Post} \geq 240$, which corresponds to probability ≥ 0.82 and FPR ≤ 0.0018 (yellow dot). The top 40 predictions are listed by gene name in **(c)**.

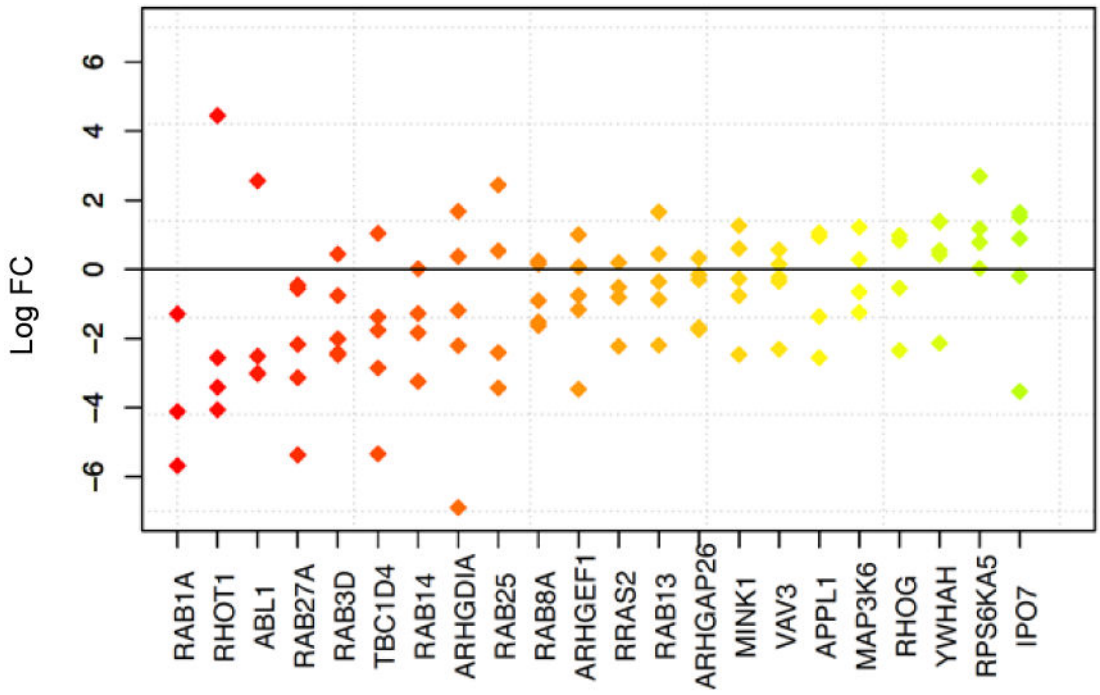
(c): Orange and blue boxes contain, respectively, known upstream regulators and downstream effectors that are successfully recovered by $OncoSig_{NB}$. Italicized text indicates proteins known to interact with KRAS via a physical protein-protein interaction. The box titled "validated predictions" shows the novel $OncoSig_{NB}$ predictions tested with the RNAi negative screen; those that were experimentally found to affect cell growth in a KRAS-dependent context are highlighted in bold text.



Extended Data Figure 2. The $OncoSig_{RF}$ and $OncoSig_{NB}$ algorithms produce highly similar KRAS LUAD SigMaps.

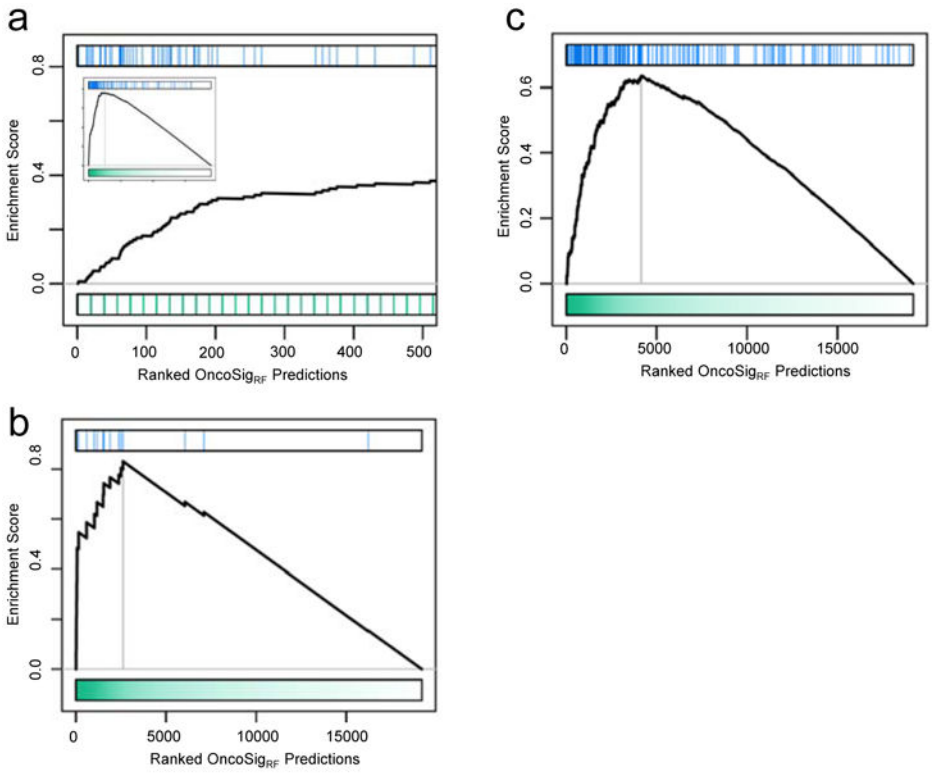
(a) Comparison of ROC curves (FPR ≤ 0.05) for LUAD-specific KRAS SigMaps predicted by $OncoSig_{NB}$ (green and blue curves) and $OncoSig_{RF}$ (orange and red curve) trained on the Ingenuity PGSS and the MSigDB PGSS, respectively.

(b) Gene Set Enrichment Analysis (GSEA) of the top 100 $OncoSig_{NB}$ LUAD-specific KRAS SigMap predictions at the top of the $OncoSig_{RF}$ LUAD-specific KRAS SigMap predictions. Ranking is based on $OncoSig_{RF}$ score (S_{RF}). Both the $OncoSig_{NB}$ predictions tested in the knockdown experiments (red lines) and the remaining top 100 $OncoSig_{NB}$ predictions (blue lines) are highly enriched at the top of the $OncoSig_{RF}$ predictions ($p = 5.6 \times 10^{-8}$ and $p = 1.7 \times 10^{-19}$, respectively).



Extended Data Figure 3. The log_2FC of shRNA abundance is plotted against the novel proteins tested in the KRAS negative selection screen.

The 3-5 points plotted for a given protein are shRNAs that target the mRNA for that protein (N = 100). The X-axis is sorted by mean log_2FC for all shRNAs targeting each gene. Colors change from red to green with mean log_2FC.

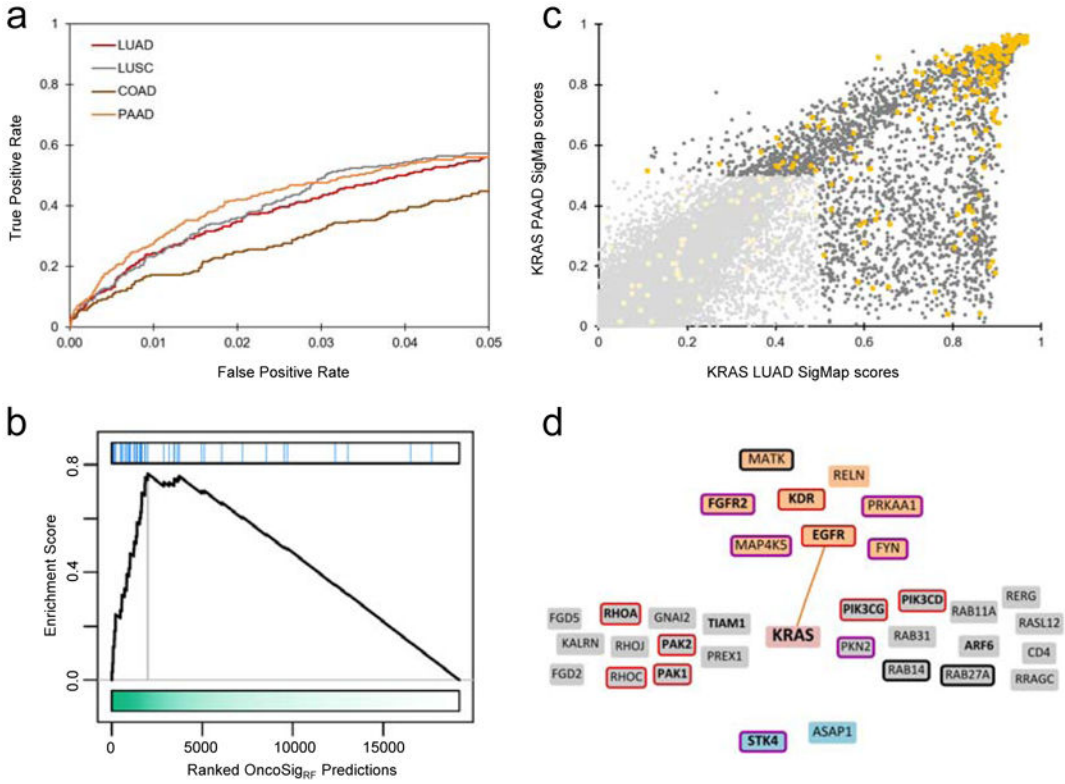


Extended Data Figure 4. OncoSig$_{RF}$ predictions are highly enriched in oncogenic KRAS^{Mut} dependencies.

(a): GSEA of KRAS^{Mut} synthetic lethal partners (25) (blue lines, N =216) and the top 500 OncoSig$_{RF}$ LUAD-specific KRAS SigMap predictions obtained by training on a modified PGSS for which the intersection with the synthetic lethal set was removed. Inset is the GSEA using all OncoSig$_{RF}$ predictions obtained in this way, where the ranking is OncoSig$_{RF}$ score. Enrichment analysis was performed with the aREA (analytic Rank-based Enrichment Analysis) algorithm (10).

(b): Enrichment of the protein resistance-signature to ERK inhibitor SCH772984 (28) (blue lines, N = 24) within OncoSig$_{RF}$ LUAD-specific KRAS SigMap predictions.

(c): Enrichment of proteins involved in response to Reactive Oxygen Species (GO:0000302) (8, 29) (blue lines, N = 276) within OncoSig$_{RF}$ LUAD-specific KRAS SigMap predictions.



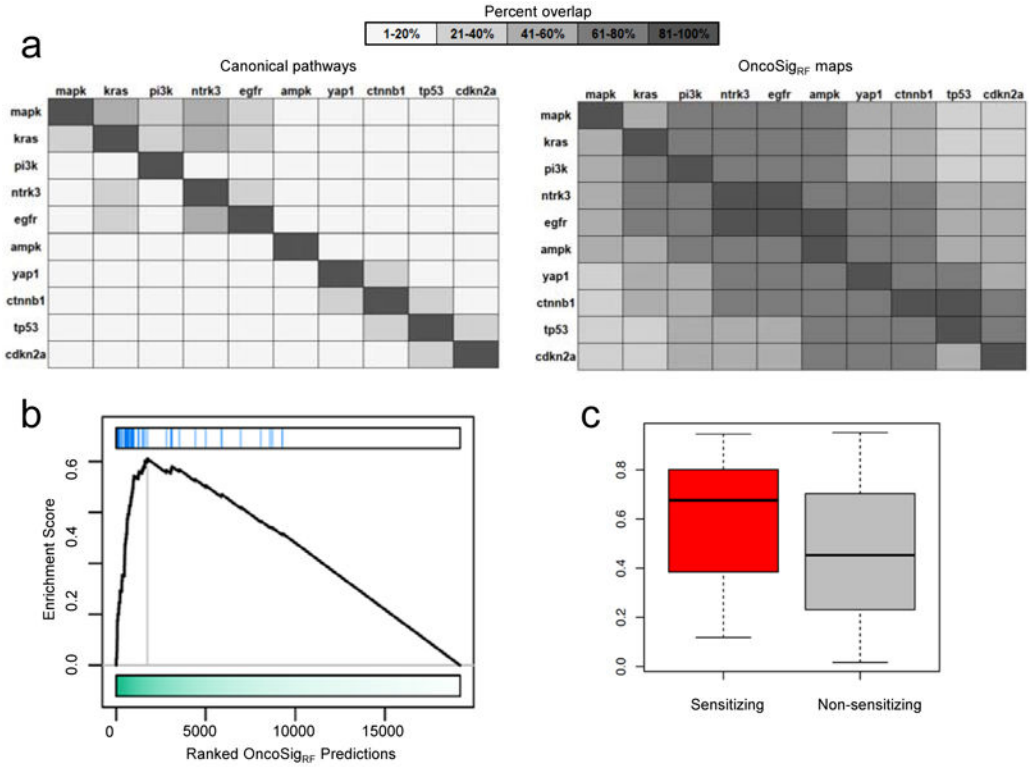
Extended Data Figure 5. OncoSig$_{RF}$ KRAS SigMaps exhibit tissue context specificity.

(a): ROC curves for the OncoSig$_{RF}$ KRAS SigMaps in LUAD (red), LUSC (gray) COAD (brown), and PAAD (orange) for FPR ≤ 0.05. Performance is evaluated as the recovery of established KRAS pathway proteins.

(b): Gene set enrichment analysis (GSEA) [32] of KRAS^{Mut} synthetic lethal partners, as determined by Corcoran et al. (27) (N = 48, blue lines). To avoid training and testing on the same proteins, OncoSig$_{RF}$ predictions for COAD-specific KRAS SigMap proteins were obtained by training on a modified PGSS from which any established KRAS^{Mut} synthetic lethal protein had been previously removed. Enrichment analysis was performed with the aREA (analytic Rank-based Enrichment Analysis) algorithm[28].

(c): Scatterplot of OncoSig$_{RF}$ scores for KRAS SigMap proteins in PAAD-vs-LUAD (N = 19,789). Each dot represents the scores for one protein. Darker colored points have high scores ($S_{RF} \geq 0.5$) in at least one context, and lighter colored points score poorly in both contexts ($S_{RF} \leq 0.5$). $R^2_{PAAD/LUAD} = 0.037$.

(d): OncoSig$_{RF}$ COAD-specific KRAS SigMap in the form depicted conceptually in Figure 1a. To prevent visual cluttering, only the top 33 OncoSig$_{RF}$ predictions (FPR ≤ 0.01) that are also VIPER-inferred KRAS interactors ($p \leq 0.01$), PrePPI-predicted KRAS physical interactors, or both, are depicted. Bold and regular text node labels represent established and novel predictions, respectively; orange and blue node colors represent upstream regulators and downstream effectors, respectively; red, blue, and black node borders represent predictions that are druggable (Drug Repurposing Hub (22)), KRAS^{Mut} synthetic lethal from the literature and validated here (see text), and both, respectively; orange and blue solid lines and gray nodes represent PrePPI-predicted physical interactors of KRAS.



Extended Data Figure 6. OncoSig$_{RF}$ SigMaps for hypermutated oncoproteins are retrospectively validated.

(a) Pairwise overlap of established pathway proteins (left) and the OncoSig$_{RF}$ LUAD-specific SigMaps (FPR ≤ 0.01, right) for the ten hyper-mutated oncoproteins (names of columns and rows). Percent overlap is color-coded according to the scale at top.

(b) SigMap predictions are highly enriched in 600 EGFR-centric network proteins (52) (p = 2.3 x 10^{-43}). Enrichment analysis was performed with the aREA (analytic Rank-based Enrichment Analysis) algorithm[28].

(c) Box plots of the OncoSig$_{RF}$ LUAD-specific EGFR SigMap scores for two subsets of the curated EGFR pathway proteins from Astsaturov et al. (52): those identified as EGFR synthetic lethal partners (red, N = 58) and those not identified as synthetic lethal (grey, N = 542). The p-value (2 x 10^{-4}) was calculated using Welch's two sample t-test.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the NCI Outstanding Investigator Award R35CA197745 to AC; the NCI Cancer Target Discovery and Development Program U01CA168426 to A.C.; the NCI Research Centers for Cancer Systems Biology Consortium U54CA209997 to A.C. and B.H.; NIGMS grant R01GM30518 to B.H.; NCI grant R01CA129562 to E.A.S.C.; Innovative Research Grant from Stand up to Cancer to E.A.S.C.; NIH High-end Instrumentation Program grant S10OD012351 to A.C.; the NIH Shared Instrumentation Program grant S10OD021764 to A.C. J.B. was supported in part by the Ruth L. Kirschstein National Research Service Award Institutional Research Training Grant T32GM082797. D.R.S. was supported by the Ruth L. Kirschstein National Research Service Award Institutional Research Training Grant T32CA09302.

REFERENCES

1. Prahallad A et al. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. Nature 483, 100–103, doi:10.1038/nature10868 (2012). [PubMed: 22281684]
2. Bild AH et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439, 353–357, doi:10.1038/nature04296 (2006). [PubMed: 16273092]
3. Krogan NJ, Lippman S, Agard DA, Ashworth A & Ideker T The cancer cell map initiative: defining the hallmark networks of cancer. Mol Cell 58, 690–698, doi:10.1016/j.molcel.2015.05.008 (2015). [PubMed: 26000852]
4. Greene CS et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet 47, 569–576, doi:10.1038/ng.3259 (2015). [PubMed: 25915600]
5. Cancer Genome Atlas Research, N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45, 1113–1120, doi:10.1038/ng.2764 (2013). [PubMed: 24071849]
6. Zhang QC et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature 490, 556–560, doi:10.1038/nature11503 (2012). [PubMed: 23023127]
7. Garzon JI et al. A computational interactome and functional annotation for the human proteome. Elife 5, doi:10.7554/eLife.18715 (2016).
8. Ashburner M et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25, 25–29, doi:10.1038/75556 (2000). [PubMed: 10802651]
9. Margolin AA et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 Suppl 1, S7, doi:10.1186/1471-2105-7-S1-S7 (2006).
10. Alvarez MJ et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat Genet 48, 838–847, doi:10.1038/ng.3593 (2016). [PubMed: 27322546]
11. Wang K et al. Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. Nat Biotechnol 27, 829–839, doi:10.1038/nbt.1563 (2009). [PubMed: 19741643]
12. Giorgi FM et al. Inferring protein modulation from gene expression data using conditional mutual information. PLoS One 9, e109569, doi:10.1371/journal.pone.0109569 (2014). [PubMed: 25314274]
13. Network, C. G. A. R. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511, 543–550, doi:10.1038/nature13385 (2014). [PubMed: 25079552]
14. Jansen R et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302, 449–453, doi:10.1126/science.1087361 (2003). [PubMed: 14564010]
15. Breiman L Random Forests. Machine Learning 45, 5–32 (2001).
16. Liberzon A et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 1, 417–425, doi:10.1016/j.cels.2015.12.004 (2015). [PubMed: 26771021]
17. Kanehisa M, Sato Y, Kawashima M, Furumichi M & Tanabe M KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44, D457–462, doi:10.1093/nar/gkv1070 (2016). [PubMed: 26476454]
18. Subramanian A 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, doi:10.1073/pnas.0506580102 (2005). [PubMed: 16199517]
19. Huttlin EL et al. The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell 162, 425–440, doi:10.1016/j.cell.2015.06.043 (2015). [PubMed: 26186194]
20. Rolland T et al. A proteome-scale map of the human interactome network. Cell 159, 1212–1226, doi:10.1016/j.cell.2014.10.050 (2014). [PubMed: 25416956]
21. Chatr-Aryamontri A et al. The BioGRID interaction database: 2017 update. Nucleic Acids Res 45, D369–D379, doi:10.1093/nar/gkw1102 (2017). [PubMed: 27980099]
22. Corsello SM et al. The Drug Repurposing Hub: a next-generation drug library and information resource. Nat Med 23, 405–408, doi:10.1038/nm.4306 (2017). [PubMed: 28388612]

23. Franceschini A et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41, D808–815, doi:10.1093/nar/gks1094 (2013). [PubMed: 23203871]
24. Lee I, Blom UM, Wang PI, Shim JE & Marcotte EM Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res 21, 1109–1121, doi:10.1101/gr.118992.110 (2011). [PubMed: 21536720]
25. Barbie DA et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 462, 108–112, doi:10.1038/nature08460 (2009). [PubMed: 19847166]
26. Kim J et al. XPO1-dependent nuclear export is a druggable vulnerability in KRAS-mutant lung cancer. Nature 538, 114–117, doi:10.1038/nature19771 (2016). [PubMed: 27680702]
27. Corcoran RB et al. Synthetic lethal interaction of combined BCL-XL and MEK inhibition promotes tumor regressions in KRAS mutant cancer models. Cancer Cell 23, 121–128, doi:10.1016/j.ccr.2012.11.007 (2013). [PubMed: 23245996]
28. Hayes TK et al. Long-Term ERK Inhibition in KRAS-Mutant Pancreatic Cancer Is Associated with MYC Degradation and Senescence-like Growth Suppression. Cancer Cell 29, 75–89, doi:10.1016/j.ccell.2015.11.011 (2016). [PubMed: 26725216]
29. Shaw AT et al. Selective killing of K-ras mutant cancer cells by small molecule inducers of oxidative stress. Proc Natl Acad Sci U S A 108, 8773–8778, doi:10.1073/pnas.1105941108 (2011). [PubMed: 21555567]
30. Liu Z, Xiao T, Peng X, Li G & Hu F APPLs: More than just adiponectin receptor binding proteins. Cell Signal 32, 76–84, doi:10.1016/j.cellsig.2017.01.018 (2017). [PubMed: 28108259]
31. Tzeng HT & Wang YC Rab-mediated vesicle trafficking in cancer. J Biomed Sci 23, 70, doi:10.1186/s12929-016-0287-7 (2016). [PubMed: 27716280]
32. Thomas JD et al. Rab1A is an mTORC1 activator and a colorectal oncogene. Cancer Cell 26, 754–769, doi:10.1016/j.ccell.2014.09.008 (2014). [PubMed: 25446900]
33. Gabernet-Castello C, O'Reilly AJ, Dacks JB & Field MC Evolution of Tre-2/Bub2/Cdc16 (TBC) Rab GTPase-activating proteins. Mol Biol Cell 24, 1574–1583, doi:10.1091/mbc.E12-07-0557 (2013). [PubMed: 23485563]
34. Lu W et al. Downregulation of ARHGDIA contributes to human glioma progression through activation of Rho GTPase signaling pathway. Tumour Biol, doi:10.1007/s13277-016-5374-6 (2016).
35. Hornstein I, Alcover A & Katzav S Vav proteins, masters of the world of cytoskeleton organization. Cell Signal 16, 1–11, doi:10.1016/s0898-6568(03)00110-4 (2004). [PubMed: 14607270]
36. Oliver AW et al. The HPV16 E6 binding protein Tip-1 interacts with ARHGEF16, which activates Cdc42. Br J Cancer 104, 324–331, doi:10.1038/sj.bjc.6606026 (2011). [PubMed: 21139582]
37. Boulter E, Estrach S, Garcia-Mata R & Feral CC Off the beaten paths: alternative and crosstalk regulation of Rho GTPases. FASEB J 26, 469–479, doi:10.1096/fj.11-192252 (2012). [PubMed: 22038046]
38. Cox AD & Der CJ Ras history: The saga continues. Small GTPases 1, 2–27, doi:10.4161/sgtp.1.1.12178 (2010). [PubMed: 21686117]
39. Prior IA & Hancock JF Ras trafficking, localization and compartmentalized signalling. Semin Cell Dev Biol 23, 145–153, doi:10.1016/j.semcdb.2011.09.002 (2012). [PubMed: 21924373]
40. Bhuin T & Roy JK Rab proteins: the key regulators of intracellular vesicle transport. Exp Cell Res 328, 1–19, doi:10.1016/j.yexcr.2014.07.027 (2014). [PubMed: 25088255]
41. Fukuda M TBC proteins: GAPs for mammalian small GTPase Rab? Biosci Rep 31, 159–168, doi:10.1042/BSR20100112 (2011). [PubMed: 21250943]
42. Hwang J & Pallas DC STRIPAK complexes: structure, biological function, and involvement in human diseases. Int J Biochem Cell Biol 47, 118–148, doi:10.1016/j.biocel.2013.11.021 (2014). [PubMed: 24333164]
43. Skrzypski M et al. Three-gene expression signature predicts survival in early-stage squamous cell carcinoma of the lung. Clin Cancer Res 14, 4794–4799, doi:10.1158/1078-0432.CCR-08-0576 (2008). [PubMed: 18676750]

44. Li N & Li S RASAL2 promotes lung cancer metastasis through epithelial-mesenchymal transition. Biochem Biophys Res Commun 455, 358–362, doi:10.1016/j.bbrc.2014.11.020 (2014). [PubMed: 25446096]
45. Yu F et al. IFITM1 promotes the metastasis of human colorectal cancer via CAV-1. Cancer Lett 368, 135–143, doi:10.1016/j.canlet.2015.07.034 (2015). [PubMed: 26259513]
46. Weinberg FD & Ramnath N Targeting IL22: a potential therapeutic approach for Kras mutant lung cancer? Transl Lung Cancer Res 7, S243–S247, doi:10.21037/tlcr.2018.09.04 (2018). [PubMed: 30393613]
47. Guillon A et al. Interleukin-22 receptor is overexpressed in nonsmall cell lung cancer and portends a poor prognosis. Eur Respir J 47, 1277–1280, doi:10.1183/13993003.01580-2015 (2016). [PubMed: 26846835]
48. Janne PA et al. Selumetinib plus docetaxel for KRAS-mutant advanced non-small-cell lung cancer: a randomised, multicentre, placebo-controlled, phase 2 study. Lancet Oncol 14, 38–47, doi:10.1016/S1470-2045(12)70489-8 (2013). [PubMed: 23200175]
49. Migliardi G et al. Inhibition of MEK and PI3K/mTOR suppresses tumor growth but does not cause tumor regression in patient-derived xenografts of RAS-mutant colorectal carcinomas. Clin Cancer Res 18, 2515–2525, doi:10.1158/1078-0432.CCR-11-2683 (2012). [PubMed: 22392911]
50. Adjei AA et al. Phase I pharmacokinetic and pharmacodynamic study of the oral, small-molecule mitogen-activated protein kinase kinase 1/2 inhibitor AZD6244 (ARRY-142886) in patients with advanced cancers. J Clin Oncol 26, 2139–2146, doi:10.1200/JCO.2007.14.4956 (2008). [PubMed: 18390968]
51. Sustic T, Bosdriesz E, van Wageningen S, Wessels LFA & Bernards R RUNX2/CBFB modulates the response to MEK inhibitors through activation of receptor tyrosine kinases in KRAS-mutant colorectal cancer. Transl Oncol 13, 201–211, doi:10.1016/j.tranon.2019.10.006 (2019). [PubMed: 31865182]
52. Astsaturov I et al. Synthetic lethal screen of an EGFR-centered network to improve targeted therapies. Sci Signal 3, ra67, doi:10.1126/scisignal.2001083 (2010). [PubMed: 20858866]
53. Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511, 543–550, doi:10.1038/nature13385 (2014). [PubMed: 25079552]
54. Tate JG et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res 47, D941–D947, doi:10.1093/nar/gky1015 (2019). [PubMed: 30371878]
55. Narlikar GJ, Sundaramoorthy R & Owen-Hughes T Mechanisms and functions of ATP-dependent chromatin-remodeling enzymes. Cell 154, 490–503, doi:10.1016/j.cell.2013.07.011 (2013). [PubMed: 23911317]
56. Morozumi Y et al. Atad2 is a generalist facilitator of chromatin dynamics in embryonic stem cells. Journal of molecular cell biology 8, 349–362, doi:10.1093/jmcb/mjv060 (2016). [PubMed: 26459632]
57. Sharma VM, Li B & Reese JC SWI/SNF-dependent chromatin remodeling of RNR3 requires TAF(II)s and the general transcription machinery. Genes Dev 17, 502–515, doi:10.1101/gad.1039503 (2003). [PubMed: 12600943]
58. Mi H, Muruganujan A, Ebert D, Huang X & Thomas PD PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res 47, D419–D426, doi:10.1093/nar/gky1038 (2019). [PubMed: 30407594]
59. Gong F & Miller KM Double duty: ZMYND8 in the DNA damage response and cancer. Cell Cycle 17, 414–420, doi:10.1080/15384101.2017.1376150 (2018). [PubMed: 29393731]
60. Sridhara SC et al. Transcription Dynamics Prevent RNA-Mediated Genomic Instability through SRPK2-Dependent DDX23 Phosphorylation. Cell Rep 18, 334–343, doi:10.1016/j.celrep.2016.12.050 (2017). [PubMed: 28076779]
61. Allemand E et al. A Broad Set of Chromatin Factors Influences Splicing. PLoS Genet 12, e1006318, doi:10.1371/journal.pgen.1006318 (2016). [PubMed: 27662573]
62. Blume-Jensen P & Hunter T Oncogenic kinase signalling. Nature 411, 355–365, doi:10.1038/35077225 (2001). [PubMed: 11357143]
63. Lemmon MA & Schlessinger J Cell signaling by receptor tyrosine kinases. Cell 141, 1117–1134, doi:10.1016/j.cell.2010.06.011 (2010). [PubMed: 20602996]

64. Organ SL & Tsao MS An overview of the c-MET signaling pathway. Ther Adv Med Oncol 3, S7–S19, doi:10.1177/1758834011422556 (2011). [PubMed: 22128289]

65. Meissl K, Macho-Maschler S, Muller M & Strobl B The good and the bad faces of STAT1 in solid tumours. Cytokine 89, 12–20, doi:10.1016/j.cyto.2015.11.011 (2017). [PubMed: 26631912]

66. Zhang Y & Liu Z STAT1 in cancer: friend or foe? Discov Med 24, 19–29 (2017). [PubMed: 28950072]

67. Balbin OA et al. Reconstructing targetable pathways in lung cancer by integrating diverse omics data. Nature communications 4, 2617, doi:10.1038/ncomms3617 (2013).

68. Downward J RAS Synthetic Lethal Screens Revisited: Still Seeking the Elusive Prize? Clin Cancer Res 21, 1802–1809, doi:10.1158/1078-0432.CCR-14-2180 (2015). [PubMed: 25878361]

69. Luo J et al. A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. Cell 137, 835–848, doi:10.1016/j.cell.2009.05.006 (2009). [PubMed: 19490893]

70. Wang T et al. Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. Cell 168, 890–903 e815, doi:10.1016/j.cell.2017.01.013 (2017). [PubMed: 28162770]

71. McDonald ER 3rd et al. Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. Cell 170, 577–592 e510, doi:10.1016/j.cell.2017.07.005 (2017). [PubMed: 28753431]

72. Aguirre AJ & Hahn WC Synthetic Lethal Vulnerabilities in KRAS-Mutant Cancers. Cold Spring Harb Perspect Med 8, doi:10.1101/cshperspect.a031518 (2018).

Methods-only References

73. Woo JH et al. Elucidating Compound Mechanism of Action by Network Perturbation Analysis. Cell 162, 441–451, doi:10.1016/j.cell.2015.05.056 (2015). [PubMed: 26186195]

74. Duan Q et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Nucleic Acids Res 42, W449–460, doi:10.1093/nar/gku476 (2014). [PubMed: 24906883]

75. Kramer A, Green J, Pollard J Jr. & Tugendreich S Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics 30, 523–530, doi:10.1093/bioinformatics/btt703 (2014). [PubMed: 24336805]

76. Li W & Godzik A Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659, doi:10.1093/bioinformatics/btl158 (2006). [PubMed: 16731699]

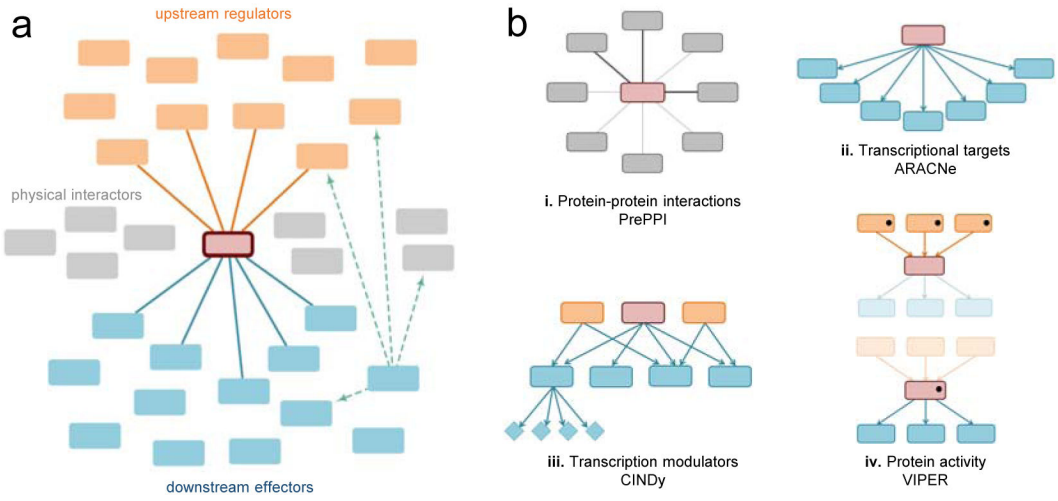
77. Arlot S & Celisse A A survey of cross-validation procedures for model selection. Statist. Surv 4, 40–79, doi:doi:10.1214/09-SS054 (2010).

78. Torres JZ, Miller JJ & Jackson PK High-throughput generation of tagged stable cell lines for proteomic analysis. Proteomics 9, 2888–2891, doi:10.1002/pmic.200800873 (2009). [PubMed: 19405035]

79. Shevchenko A, Tomas H, Havlis J, Olsen JV & Mann M In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat Protoc 1, 2856–2860, doi:10.1038/nprot.2006.468 (2006). [PubMed: 17406544]

80. Zheng Y et al. A rare population of CD24(+)ITGB4(+)Notch(hi) cells drives tumor propagation in NSCLC and requires Notch3 for self-renewal. Cancer Cell 24, 59–74, doi:10.1016/j.ccr.2013.05.021 (2013). [PubMed: 23845442]

81. Dai Z et al. edgeR: a versatile tool for the analysis of shRNA-seq and CRISPR-Cas9 genetic screens. F1000Res 3, 95, doi:10.12688/f1000research.3928.2 (2014). [PubMed: 24860646]



c

Protein	PGSS (0, 1)	Protein-protein interactions PrePPI (likelihood ratio)				Gene regulation ARACNe (mutual information)				Post-translational modification CINDy (SP-coF/TF-gene triplets)				Mutation activity VIPER -log (p-value)				OncoSig score S_{RF}
		...	UBP1	SETD2	CTNNB1	...	UBP1	SETD2	CTNNB1	...	UBP1	SETD2	CTNNB1	...	UBP1	SETD2	CTNNB1	...	
RAF1	1		0	634	11689		0.27	0.30	0		55	80	0		4.20	4.73	0		0.94
NRG1	0		0	0	0		0	0	0		0	0	0		0	0	0		0.50
ARF1	0		0	0	671		0	0	0.18		55	84	86		0	0	0		0.70
KLF10	0		0	0	0		0	0	0		0	0	0		0	0	9.54		0.25
PTPN11	1		3155	1158	8964		0	0	0		48	123	0		0	0	0		0.90
TRIB3	0		0	0	0		0	0	0		0	0	0		0	0	0		0.96
LIMD1	0		0	0	695		0	0.27	0		78	59	0		0	3.69	12.0		0.85
...																			

Figure 1: Protein-specific molecular interaction Signaling Map (SigMap) and the OncoSig$_{RF}$ algorithm

(a) Graphical representation of a SigMap for an anchor oncoprotein (red node). The SigMap comprises: (*i*) upstream activity modulators (orange nodes), (*ii*) downstream effectors, responsible for mediating its pathophysiologic function (blue nodes), (*iii*) structural cognate binding partners (gray nodes), which may be either modulators (solid orange lines) or effectors (solid blue lines), and (*iv*) auto-regulatory loops connecting downstream effectors to upstream modulators (dashed green lines). To avoid unnecessary clutter, implicit arrows connecting upstream modulators to the red node and the latter to its downstream effectors are omitted. Thus, the only interactions explicitly denoted by an edge are physical protein-protein interactions and autoregulatory interactions between modulator and effector proteins.
(b) Networks used to train the OncoSig$_{RF}$ algorithm. (*i*) PrePPI[6] predicts interactions between a protein (red), and its physical and/or functional interactors (gray). (*ii*) The ARACNe algorithm[9] predicts transcription factors or signaling molecules (red) that transcriptionally regulate target genes (blue). (*iii*) CINDy[12] predicts signaling molecules (orange/red) that post-translationally modify transcription factors (blue boxes), which in turn leads to differential expression of a transcription factor's targets (blue diamonds). (*iv*) The VIPER algorithm[10] infers downstream effectors (blue) and upstream regulators (orange) for a given protein (red). VIPER associates 1) the protein (red) with a missense mutation (black dot) with the activity change of transcription factors (blue) and 2) signaling molecules (orange) with missense mutations (black dots) with activity of the protein (red).
(c) Feature matrix for OncoSig$_{RF}$ algorithm. Networks in (b) are encoded as a feature matrix (dark green box), where rows correspond to proteins in the human proteome, columns

correspond to proteins for which clues exist in PrePPI[7], ARACNe[9], CINDy[11], and VIPER[10], respectively, and each entry is a scalar proportional to the confidence in the corresponding interaction as described in the literature. The latter include likelihood ratios for PrePPI, mutual information for ARACNe, number of SP-coF/TF-gene triplets for CINDy, and $-\log_{10}$ p-value for VIPER. The gold column corresponds to whether a protein is an established member of a particular pathway (PGSS, value of 1) or not (value of 0). Only a few components of a small subset of proteins are shown. The feature vector for LIMD1 (highlighted in light green) is described in the text. The last column provides the OncoSig_{RF} score of the subset proteins for the LUAD-specific KRAS SigMap (see Figure 2). SP = signaling protein, coF = co-factor, and TF = transcription factor.

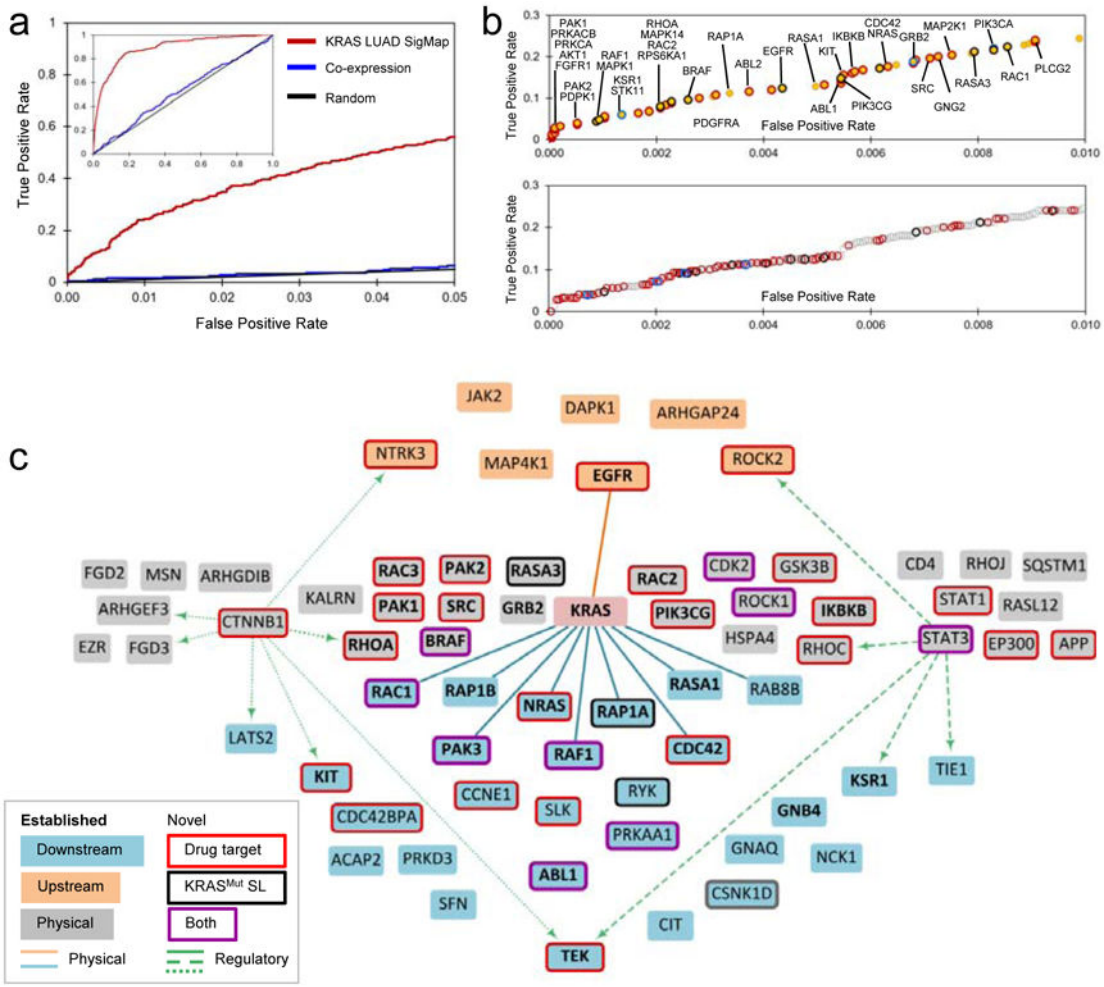


Figure 2: The OncoSig$_{RF}$ LUAD-specific KRAS SigMap

(a): ROC curves are displayed for the performance at recovering established KRAS-pathway proteins (FPR ≤ 0.05 (5%)) of OncoSig$_{RF}$ (red curve, N = 1,114), Pearson's correlation between mRNA expression of KRAS and mRNA expression of other proteins in LUAD (green curve, N = 957), and random prediction (black curve). The inset shows the full ROC curves (red curve, N = 19,548; green curve, N = 18,891).

(b): The ROC curve in **(a)** for FPR ≤ 0.01 (1%, N = 263) is separated into two according to whether predictions correspond to **(i)** established KRAS-pathway proteins (top panel, yellow circles), with best-known KRAS-pathway proteins individually labeled or **(ii)** novel KRAS SigMap proteins (bottom panel, white circles). Circles annotate predictions as either druggable (Drug Repurposing Hub[22]) (red), experimentally-validated KRAS interactors (BioGRID[21]) (blue), or both (black). .

(c): OncoSig$_{RF}$ LUAD-specific KRAS SigMap in the form depicted conceptually in Figure 1a. To prevent visual cluttering, only the top 68 OncoSig$_{RF}$ predictions that are also VIPER-inferred KRAS interactors, PrePPI-predicted physical interactors, or both, are depicted. Bold and regular text node labels represent established and novel predictions, respectively; orange and blue node colors represent upstream regulators and downstream effectors, respectively; red, black, and purple node borders represent predictions that are druggable (Drug

Repurposing Hub[22]), $KRAS^{Mut}$ synthetic lethal partners from the literature[72] or validated in this study, and both, respectively; solid orange and blue lines and gray nodes represent PrePPI-predicted physical KRAS interactors; green dashed lines represent auto-regulatory and feed-forward loop interactions.

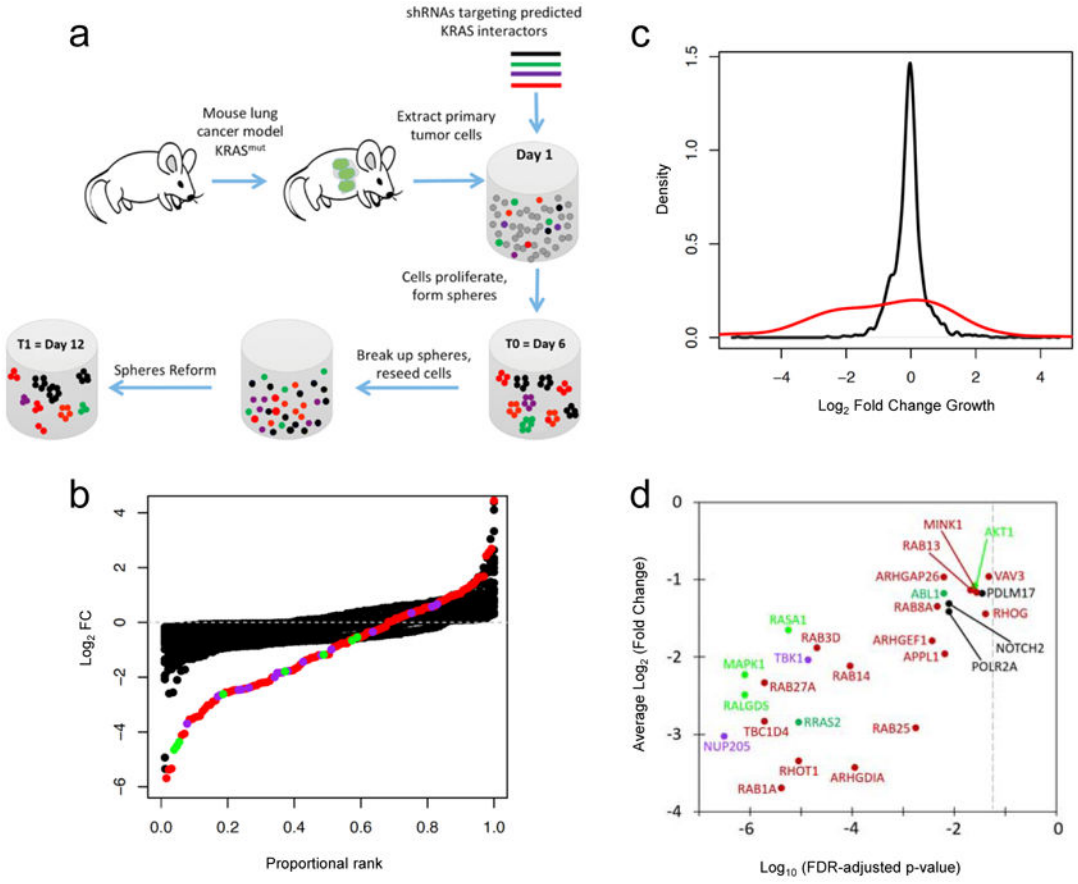


Figure 3: Experimental validation of the OncoSig$_{RF}$ LUAD-specific KRAS SigMap

(a): Schematic of the pooled shRNA negative screen experiments performed. An average of four shRNAs target each gene in the protocol implemented. KRAS$^{G12D/+}$/p53$^{fl/fl}$ primary tumor cells (green patches) are isolated from the mouse and placed in a semi-solid 3D matrix (cylinder). A pooled shRNA knockdown is performed (Day 1), and each cell stochastically integrates one shRNA into its DNA. Cells that integrate different shRNAs are shown as, red (representing shRNAs for novel predictions), green and purple (for positive controls), and black (for the background pool). Some cells and their daughter cells form spheroids (Day 6). The spheroids are dissociated, reseeded in a new matrix, and reform (Day 12). Fold Change (FC) of shRNA abundance is measured by deep sequencing the shRNAs at days 6 and 12.

(b): Plot of $\log_2$FC of shRNAs targeting predicted KRAS functional partners (red, N = 100), known members of the KRAS signaling pathways (RALGDS, MAPK1, RASA1 and AKT1) (purple, N = 17) and two synthetic lethal positive controls (NUP205 and TBK1) (green, N = 8). The black dots show $\log_2$FC of shRNAs targeting 515 genes within the Background Pooled Screens (BPS, N = 2286) not expected to be involved in KRAS regulated signaling. The X-axis is the normalized rank, calculated by ranking $\log_2$FC of each set of shRNAs and dividing by the number of shRNAs in that set. Each gene is represented by several dots, which correspond to different shRNAs. See Extended Data Figure 3 for more details.

(c): Density plots of $\log_2$FC for predicted KRAS functional partners (red), all individual BPS (grey), and the average of all BPS (black).

(d): Fold change versus significance for genes that significantly reduce organoid growth in the pooled shRNA negative screen experiments (FDR < 0.05, gray dotted line). Log_2FC of shRNAs, averaged after removing one or two outlier hairpins, is plotted against log_{10}-transformed, FDR-adjusted p-values. shRNAs are colored as described in **(b)**.

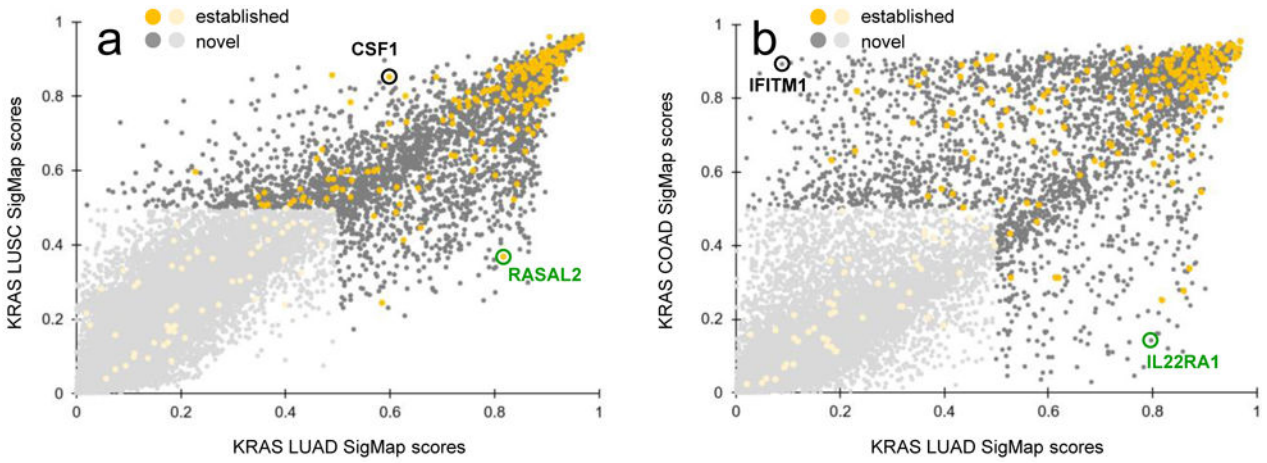


Figure 4: KRAS SigMap tumor context specificity

Scatterplots of OncoSig$_{RF}$ scores for KRAS SigMap proteins in **(a)** the LUSC-vs-LUAD (N = 19,790) and **(b)** the COAD-vs-LUAD (N = 19,438) tumor contexts **(b)**. Each dot represents one protein. Gold and gray points represent established and novel predictions, respectively. Darker colored points have high scores ($S_{RF} \geq 0.5$) in at least one context, and lighter colored points have low scores ($S_{RF} < 0.5$) in both contexts and should thus not be compared. Correlation coefficients are $R^2_{LUSC/LUAD} = 0.35$ ($p < 10^{-267}$) and $R^2_{COAD/LUAD} = 0.10$ ($p < 10^{-165}$), respectively (Welch's Two Sample t-test). Specific points highlighted in black and green are discussed in the text.

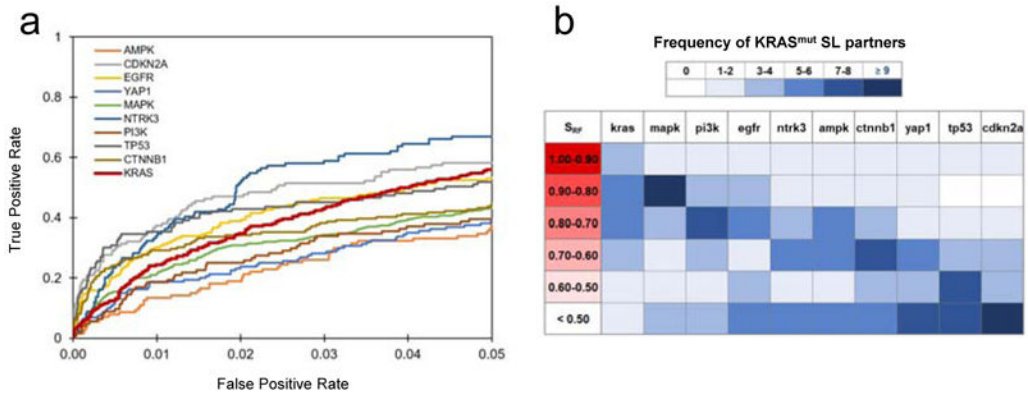


Figure 5: OncoSig$_{RF}$ LUAD-specific SigMap analysis of hyper-mutated oncoproteins

(a): ROC curves showing OncoSig$_{RF}$'s performance in terms of identifying established pathway proteins in LUAD-specific SigMaps for the 10 oncoproteins listed in the legend (FPR ≤ 0.05, N ~ 1000). The thick red line represents performance of OncoSig$_{RF}$ for the LUAD-specific KRAS SigMap from Figure 2a as a reference.

(b): Number of literature-derived KRAS^{Mut} synthetic lethal partners[72] (predicted by each of the ten SigMaps) as a function of OncoSig$_{RF}$ score. OncoSig$_{RF}$ scores are binned, and the bins are colored from dark red (highest scores) to white (S_{RF} < 0.50), as depicted in the leftmost column. The number of predictions per score bin is color-coded according to the legend at the top.

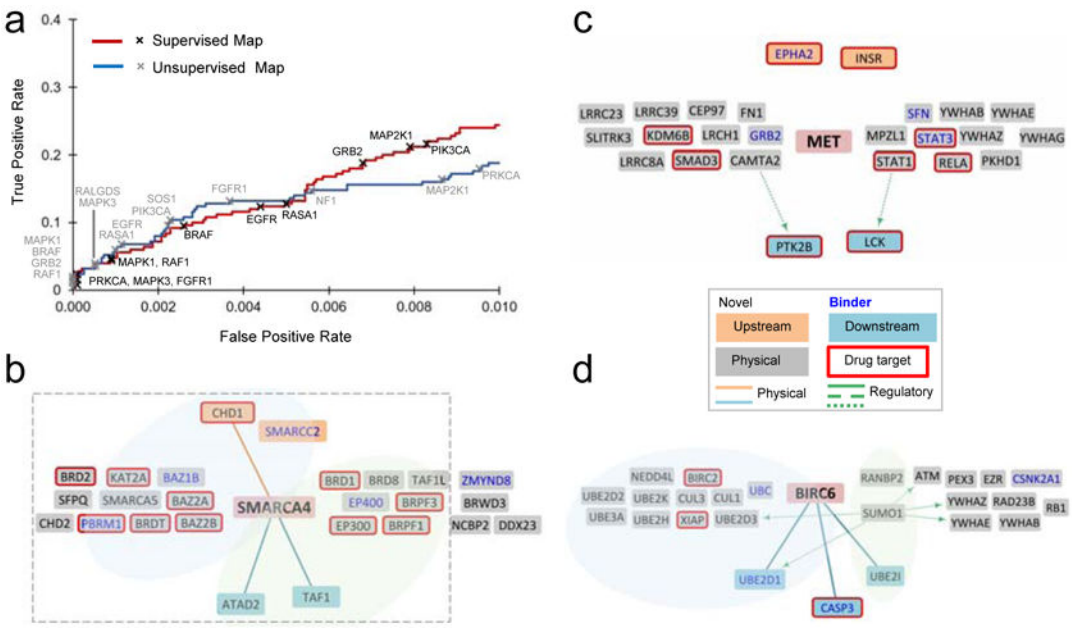


Figure 6: OncoSig: Generalization of OncoSig$_{RF}$ to Cancer Gene Census proteins

(a): ROC curve analysis of LUAD-specific KRAS SigMaps generated by OncoSig (blue curve) versus OncoSig$_{RF}$ (red curve), FPR ≤ 0.10. Established KRAS signaling proteins recovered by OncoSig and OncoSig$_{RF}$ are labeled in gray and black, respectively.

(b-d): LUAD-specific OncoSig SigMaps for three Cancer Gene Census proteins in the form depicted conceptually in Figure 1a: SMARCA4 (**b**), MET (**c**), and BIRC6 (**d**). Orange and blue node colors represent upstream regulators and downstream effectors, respectively; solid orange and blue lines and gray nodes represent PrePPI-predicted physical interactors; red node borders represent predictions that are druggable (Drug Repurposing Hub[22]); blue node text represents experimentally observed interactions (BioGRID[21]); and dashed lines represent regulatory interactions. In **(b)**, the gray dashed box includes map members involved in chromatin organization (GO:0016568); green shading includes map members involved in histone acetylation (GO:0016573); and blue shading includes map members involved in chromatin remodeling (GO:0006338)[58]. In **(d)**, blue shading includes the map members involved in protein ubiquitination (GO:0016567), and green shading includes map members involved in protein sumoylation (GO:0016925)[58].