

AWARD NUMBER: W81XWH-19-1-0703

TITLE: Novel Systems Biology Approach to Decoding Actionable Targets to Overcome Resistance in GI Cancer Monotherapies

PRINCIPAL INVESTIGATOR: Nidhi Sahni, PhD

CONTRACTING ORGANIZATION: The University of Texas MD Anderson Cancer Center
Houston, TX

REPORT DATE: December 2022

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE December 2022		2. REPORT TYPE Final		3. DATES COVERED 01Sep2019-31Aug2022	
4. TITLE AND SUBTITLE Novel Systems Biology Approach to Decoding Actionable Targets to Overcome Resistance in GI Cancer Monotherapies				5a. CONTRACT NUMBER W81XWH-19-1-0703	
				5b. GRANT NUMBER CA181455	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Nidhi Sahni, PhD E-Mail: nsahni@mdanderson.org				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) The University of Texas MD Anderson Cancer Center 1515 Holcombe Blvd. Houston, TX 77030				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Development Command Fort Detrick, Maryland 21702-5012				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT Although there are many existing tools to predict cancer-causing mutations, it remains challenging to effectively and systematically identify drug combination targets for better CRC therapeutics. The reasons are multi-fold: 1) Most tumors carry multiple driver alterations that trigger diverse oncogenic events that cannot be suppressed with mono-therapies; 2) diverse confounding factors such as intra-tumor and inter-tumor heterogeneity exist across patient populations; and 3) achieving analytical sensitivity often requires generating and analyzing a massive amount of data, which poses an unprecedented computational (big-data) challenge. Our objectives in this application are to devise a network-based framework to discover drug combinations to overcome resistance and conquer cancer progression. Here we propose an integrative approach to predict synergistic drug combinations and experimentally validate their functional effects on CRC using patient-derived tumor models. Together, this application is significant and innovative, because it will provide insights in prioritizing drug combination target pairs, and uncovering patient-specific signaling mechanisms, a critical step towards personalized precision medicine in CRC therapy. This work will also experimentally validate several drug combinations in CRC, which stand to not only find their potential cures, but also save more patient lives.					
15. SUBJECT TERMS Network model, drug combination, cell lines, patient-derived organoid, xenograft model, KRAS mutation, GI cancer					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRDC
Unclassified	Unclassified	Unclassified	Unclassified	58	19b. TELEPHONE NUMBER (include area code)

TABLE OF CONTENTS

	<u>Page</u>
1. Introduction	4
2. Keywords	4
3. Accomplishments	4
4. Impact	12
5. Changes/Problems	13
6. Products	13
7. Participants & Other Collaborating Organizations	14
8. Special Reporting Requirements	18
9. Appendices	18

1. INTRODUCTION:

Although there are many existing tools to predict cancer-causing mutations, it remains challenging to effectively and systematically identify drug combination targets for better CRC therapeutics. The reasons are multi-fold: 1) Most tumors carry multiple driver alterations that trigger diverse oncogenic events that cannot be suppressed with mono-therapies; 2) diverse confounding factors such as intra-tumor and inter-tumor heterogeneity exist across patient populations; and 3) achieving analytical sensitivity often requires generating and analyzing a massive amount of data, which poses an unprecedented computational (big-data) challenge. Our objectives in this application are to devise a network-based framework to discover drug combinations to overcome resistance and conquer cancer progression. Here we propose an integrative approach to predict synergistic drug combinations and experimentally validate their functional effects on CRC using patient-derived tumor models. Together, this application is significant and innovative because it will provide insights in prioritizing drug combination target pairs, and uncovering patient-specific signaling mechanisms, a critical step towards personalized precision medicine in CRC therapy. This work will also experimentally validate several drug combinations in CRC, which stand to not only find their potential cures, but also save more patient lives.

2. KEYWORDS:

Network model, drug combination, cell lines, patient-derived organoid, xenograft model, *KRAS* mutation, GI cancer

3. ACCOMPLISHMENTS:

What were the major goals of the project?

Specific Aim 1: Discover actionable drug targets to overcome resistance in GI cancer monotherapy via novel network-based approach.		
Major Task 1: Devise a network-based ‘functional variomics’ pipeline to identify synthetic lethality pairs for combination therapy in CRC	Months	% completion
Subtask 1: Using an integrative ‘functional variomics’ pipeline to identify driver variants in CRC that confer drug resistance (TCGA and COSMIC)	1-6	100
Subtask 2: Decipher specific signaling perturbations caused by drug-resistant mutations	7-9	100
Subtask 3: Drug combination design based on synthetic lethality and genetic interaction networks	10-12	100
Specific Aim 2: Functionally characterize the effects of actionable drug combinations in GI cancers <i>in vitro</i> and <i>in vivo</i>.		
Major Task 1: Assess the MEKi and PARPi/XPO1i combinations in CRC cell lines and patient-derived tumor organoid models	Months	% completion
Subtask 1: Submit documents for ACURO approvals	1	100

Subtask 2: Functional characterization of MEKi and PARPi/XPO1i combinations in CRC cell lines	4-24	100
Subtask 3: Functional characterization of MEKi and PARPi/XPO1i combinations in patient-derived tumor organoid (PDO) models	16-24	100
Major Task 2: Assess the MEKi and PARPi/XPO1i combinations in CRC patient-derived xenograft models	Months	% completion
Subtask 1: Submit documents for ACURO approvals	1	100
Subtask 2: Functional characterization of MEKi and PARPi/XPO1i combinations in patient-derived xenograft (PDX) models	14-36	100
Specific Aim 3: Determine molecular mechanisms of overcoming drug resistance mediated by somatic mutations.		
Major Task 1: Transcriptome-wide exploration of gene expression profiling in PDX models upon drug combination treatment using single cell RNA-seq	Months	% completion
Subtask 1: PDX sample preparation and single cell RNA-seq	18-30	100
Subtask 2: Data analysis and integration	20-36	100
Major Task 2: Protein expression profiling in PDX models upon drug combination treatment using Reverse Phase Protein Array (RPPA)	Months	% completion
Subtask 1: PDX sample preparation and RPPA experiment	18-30	100
Subtask 2: Data analysis and integration	20-36	100

What was accomplished under these goals?

Major activities, objectives and results:

We have successfully completed all the proposed work in this grant.

For **Specific Aim 1**, the objective was to identify key drug combination candidates for various CRC patients. We started out with designing a pan-cancer based predictive framework. We downloaded mutation and gene expression profile data from the TCGA project. We sought to identify candidate driver mutations by integrating functional omics datasets of cancer genomes at the systems level. It was increasingly appreciated that the interaction profiles (‘edgotype’) of a genomic mutation provides a fundamental link between genotype and phenotype. We therefore modeled mutational effect on protein-protein interaction networks, by developing a network-based computational method to identify candidate

‘edgetic’ mutations that perturb functional pathways. Our predictive framework could identify informative paths that could be used to distinguish disease risk factors from neutral elements and to stratify disease subtypes with clinical relevance. The predicted targets are enriched in cancer vulnerability genes, known drug targets but depleted for proteins associated with side effects, demonstrating the power of network-based strategies to investigate the functional impact and perturbation profiles of genomic mutations.

The expression of genes was measured by Fragments Per Kilobase Million (FPKM), and genes not expressed in more than 30% samples were excluded in our analysis. In addition, another set of gene expression profile data was downloaded from the International Cancer Genome Consortium (ICGC) project. The normalized read count for each gene was used in our analysis. A systematic map of ~16,000 high-quality human binary protein-protein interaction network, which was identified by yeast two-hybrid (Y2H) assay, was used in our analysis. The homodimer interactions were removed and there are 15,957 interactions among 4,743 genes for further analysis.

We have devised a network-based ‘functional variomics’ pipeline to identify synthetic lethality pairs for combination therapy in colorectal cancer (CRC). We used an integrative framework to discover driver variants in CRC that confer drug resistance, to decipher specific signaling perturbations caused by drug-resistant mutations, and to design drug combinations based on synthetic lethality and genetic interaction networks. To identify the mutations that might perturb signaling pathways, we hypothesized that gene-gene relationships would show perturbations in patients with specific driver mutations. We therefore developed e-MutPath as an open-source R package to identify candidate driver mutations that perturb functional pathways. Three types of omics datasets were integrated, including gene expression, somatic mutations and functional networks or pathways (**Fig. 1**). The output would provide prioritized mutations as well as the perturbed edges in signaling pathways or networks.

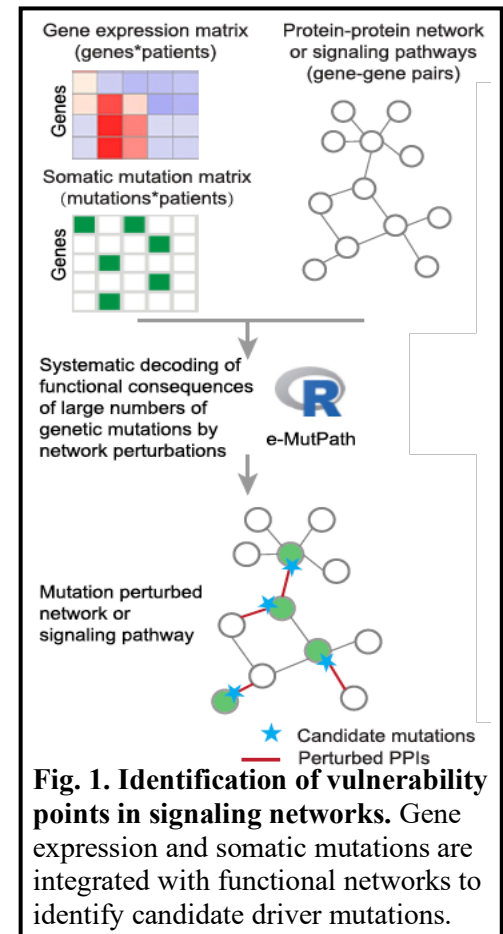


Fig. 1. Identification of vulnerability points in signaling networks. Gene expression and somatic mutations are integrated with functional networks to identify candidate driver mutations.

Specifically, three steps were performed in this computational method in the context of cancer. First, perturbed functional interactions were identified in each cancer patient based on a correlation perturbation analysis of RNA expression (**Fig. 1**, top panel). All the patients were mapped to a two-dimensional plane based on the expression levels of two interacting genes. If a patient showed a significant deviation from a normal gene-gene relationship distribution, the patient would be an outlier in the regression line modeled by all the patients. We used Grubb’s test for detecting the outliers. Second, sample-specific interaction perturbation profiles were constructed; if gene expression data in normal samples were present, we also

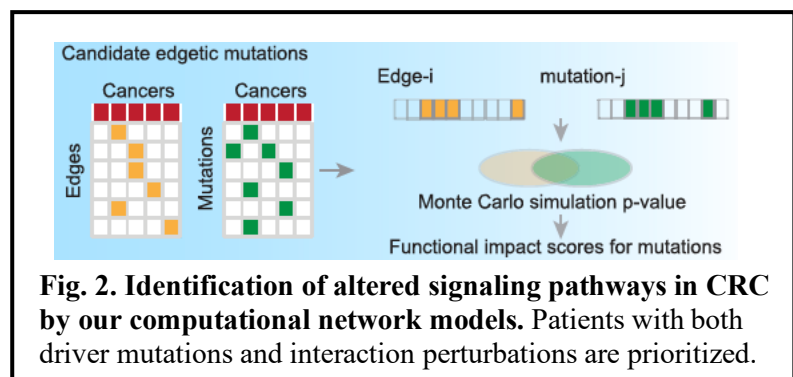


Fig. 2. Identification of altered signaling pathways in CRC by our computational network models. Patients with both driver mutations and interaction perturbations are prioritized.

identified the perturbed functional subnetworks or pathways that could distinguish cancer from normal samples (**Fig. 1**, middle panel). Finally, candidate driver mutations in each cancer sample that mediated interaction perturbations were identified by integration of interaction perturbation patterns with mutational profiles (**Fig. 1**, bottom panel). We used Monte Carlo simulation to evaluate whether the patients with specific mutations were significantly overlapped with those showing gene-gene relationship perturbations (**Fig. 2**). The mutations with p-value less than 0.05 were identified as candidate driver edgetic mutations.

We next trained three types of machine learning methods (Support Vector Machine, SVM; Logistic regression, logR; and Random forest, RF) based on the perturbation profiling of patients and normal controls. Based on ten-fold cross-validation, we found that these classifiers exhibited the AUC (Area under the receiver operating characteristic curve) greater than 0.9 (**Fig. 3**, AUC = 0.937-0.997). Moreover, these perturbed functional profiles were validated in another independent dataset obtained from International Cancer Genome Consortium (**Fig. 3**, AUC = 0.814-0.942). These results suggest that our computational model can identify the informative functional paths to distinguish disease from normal controls, which is helpful for novel biomarker and drug target discovery.

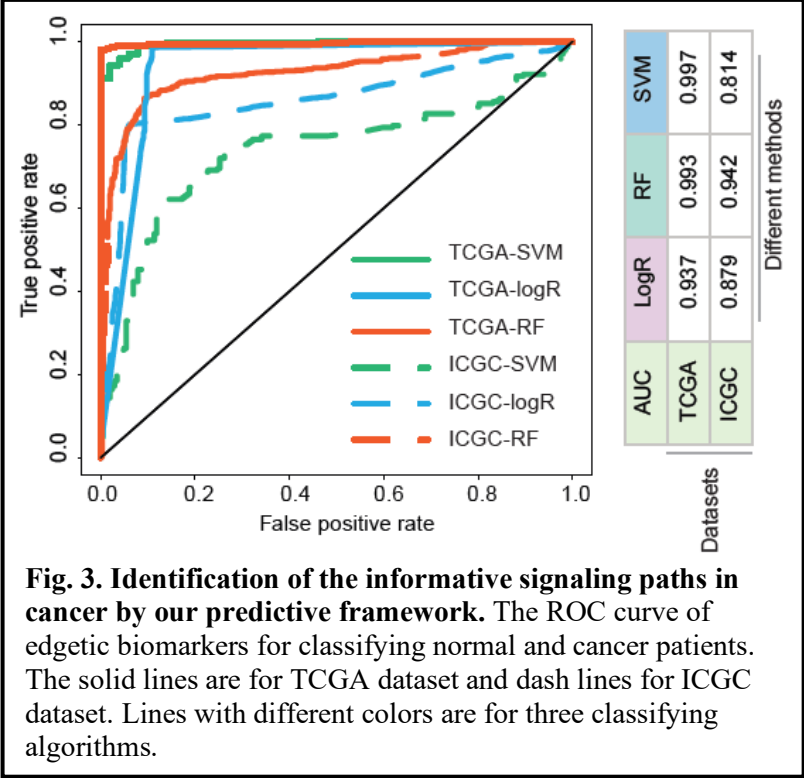


Fig. 3. Identification of the informative signaling paths in cancer by our predictive framework. The ROC curve of edgetic biomarkers for classifying normal and cancer patients. The solid lines are for TCGA dataset and dash lines for ICGC dataset. Lines with different colors are for three classifying algorithms.

Having shown e-MutPath could identify mutation-perturbed signaling pathways, we next evaluated its performance in uncovering cancer-related genes. We first considered the genes from Cancer Gene Census (CGC) and found that our top predictions included a high fraction of CGC genes. To illustrate the power of e-MutPath, we compared its performance with four widely used approaches-SIFT, Polyphen-2, MutationTaster and MutationAssessor. Our result showed that e-MutPath consistently prioritized a larger fraction of cancer genes than other methods, demonstrating the advantage of network integration. We

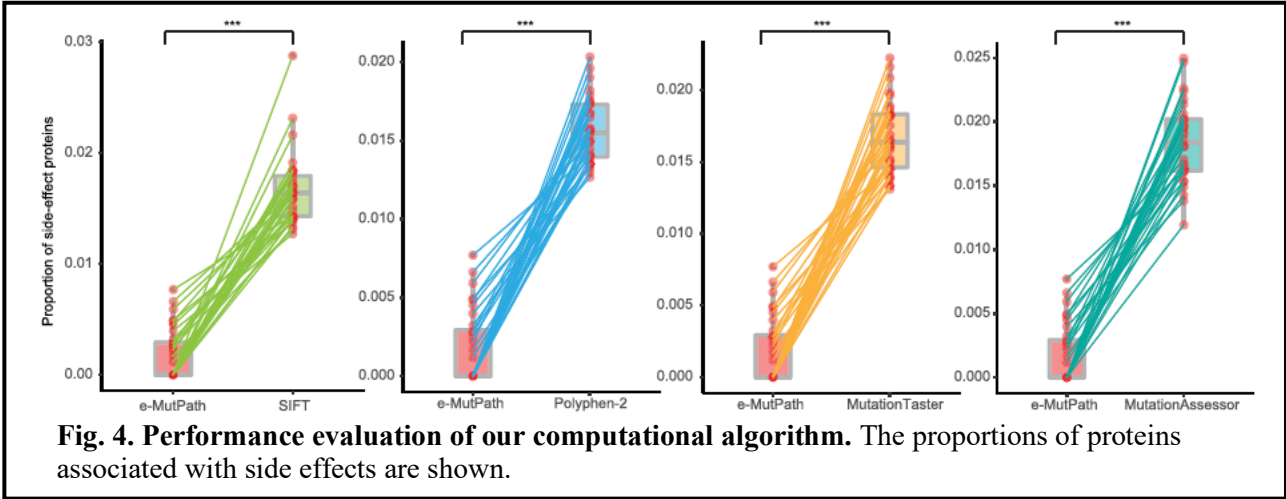


Fig. 4. Performance evaluation of our computational algorithm. The proportions of proteins associated with side effects are shown.

found that while e-MutPath predicted a smaller number of targets, they comprised larger fractions of ‘gold standard’ cancer genes. In addition, we obtained known cancer genes from CancerMine, which is a text-mined and routinely updated database of drivers, oncogenes and tumor suppressors. We found that the overall results were consistent for all genes, oncogenes, tumor suppressors and driver genes in CRC. Finally, to evaluate potential targetable and side effects of the predicted genes, we used a list of 151 clinically actionable genes and 237 proteins that are reported to be associated with side effects. We found that e-MutPath predictions exhibited similar fractions of actionable genes but were depleted of side effect-causing proteins across cancer types (Fig. 4). Taken together, these results demonstrate significant improvement of e-MutPath over previous state-of-the-art methods in identifying key cancer vulnerability targets.

In order to better understand mutation-perturbed signaling pathways in cancer, we applied our predictive framework to 33 cancer types. We identified a connected network perturbed by mutations (Fig. 5), including a number of genes (such as *MED4* and *TCF4*) known to be involved in cancer. Cancer genes often function as network hub proteins which are involved in many cellular processes. We next investigated the topological features of the prioritized genes with mutations and found that they showed significantly higher degrees, betweenness and closeness (Fig. 5, p -values<0.001). Moreover, we performed functional enrichment analysis and found that these genes were likely to be involved in cancer hallmark pathways. Collectively, these results suggest that these mutations play critical roles in cancer by

perturbing the interactions involved in cancer-related pathways. Together, our network model represents a robust computational tool to systematically assign functions to genetic mutations, especially in the context of their specific pathway perturbation effect. More importantly, we are able to predict novel drug candidate gene pairs for testing and validation in this project. Our work from Aim 1 resulted in a high-profile publication in 2021 in *Nucleic Acids Research* (PMID: 33211847).

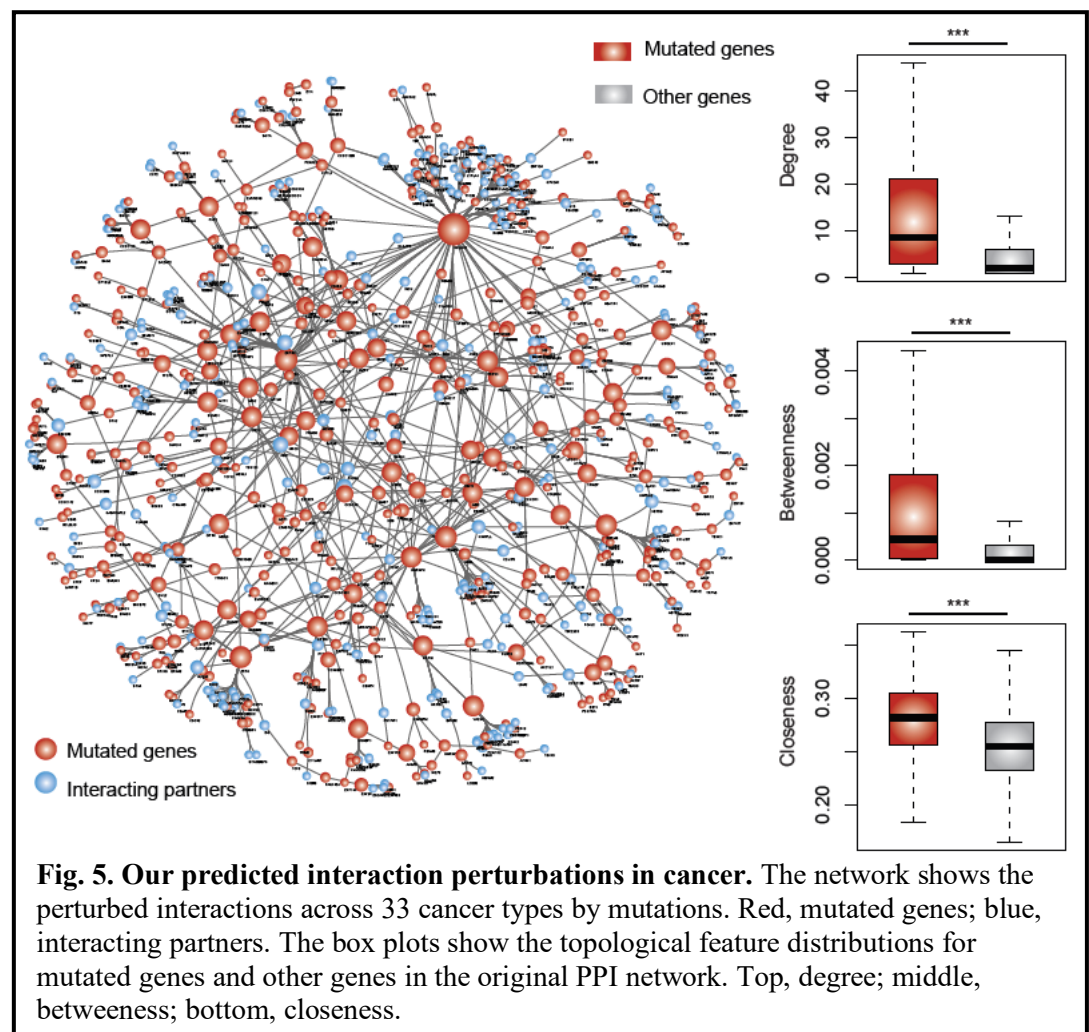
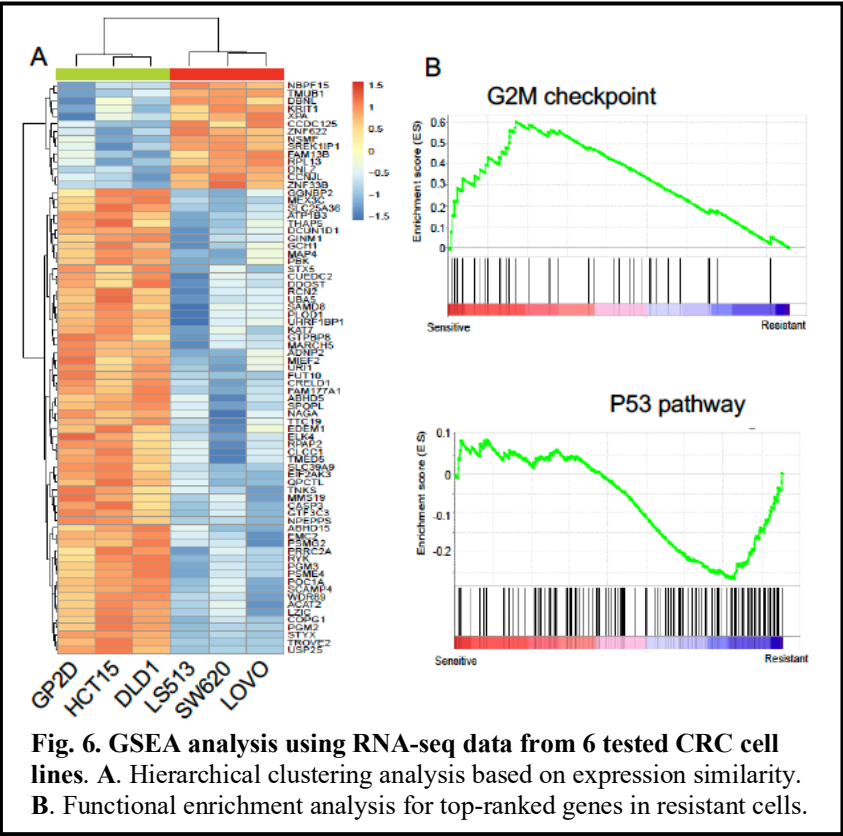
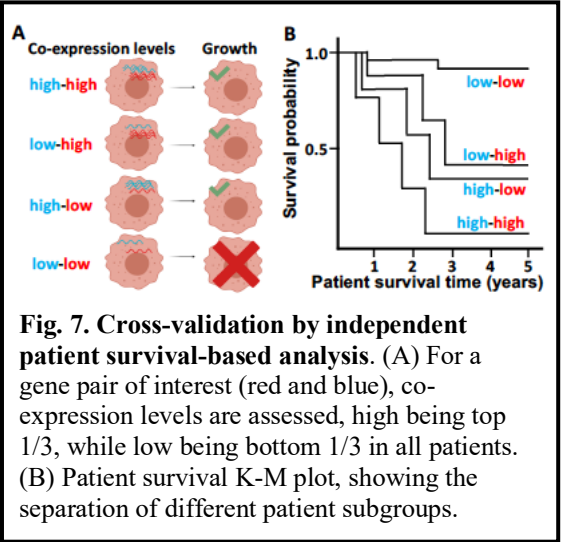


Fig. 5. Our predicted interaction perturbations in cancer. The network shows the perturbed interactions across 33 cancer types by mutations. Red, mutated genes; blue, interacting partners. The box plots show the topological feature distributions for mutated genes and other genes in the original PPI network. Top, degree; middle, betweenness; bottom, closeness.

For **Specific Aim 2**, we have assessed the MEKi-based combinations in CRC cell lines. We experimentally determined MEKi sensitive vs resistant CRC cell lines and performed RNA-seq gene expression to identify signaling pathways that are altered in different cell lines with distinct drug responses. Our data in 6 CRC cell lines identified 3 lines that are resistant to MEKi (GP2D, DLD1, HCT15), and 3 lines that are sensitive (SW620, CT26, LS513). We also determined their IC50 values. Initial RNA-seq was performed on these cell lines and GSEA analysis indicates that in drug resistant cells, G2M checkpoint (cell cycle) proteins are upregulated, while P53 pathway is downregulated (**Fig. 6**).

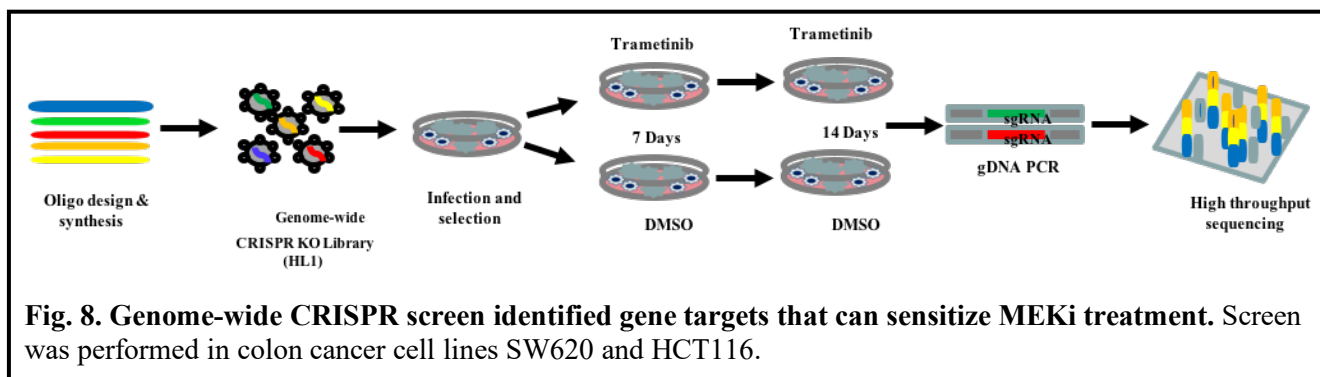


As a cross validation, we designed an independent computational method that leverages CRC patient survival (**Fig. 7**). Starting from patient gene expression data, we will identify candidate gene pairs based on (1) elevated expression of both genes in cancer cells, (2) actionability (drug availability); and (3) improved survival in patients with low expression in both genes. For synthetic lethal gene pairs, **we expected** that low expression in both genes offers disadvantage for cancer cell growth, and thus significantly improved survival in corresponding patients, as shown in the Kaplan-Meier plots. **Our results** indicated that CRC patients with reduced expression levels of MEK-gene pairs fare better than their high expression counterparts. We could successfully validate ~20 pairs of e-MutPath generated results so far. MEKi and PARPi/XPO1i combinations are among the top predicted and validated candidates.

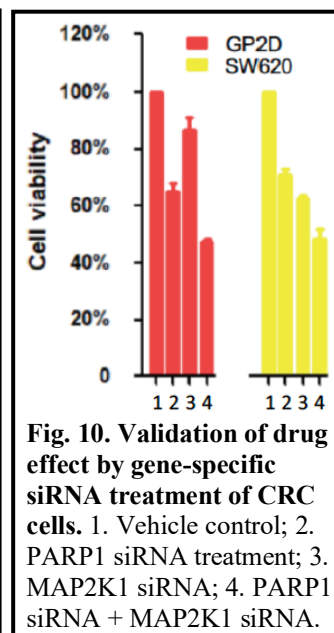
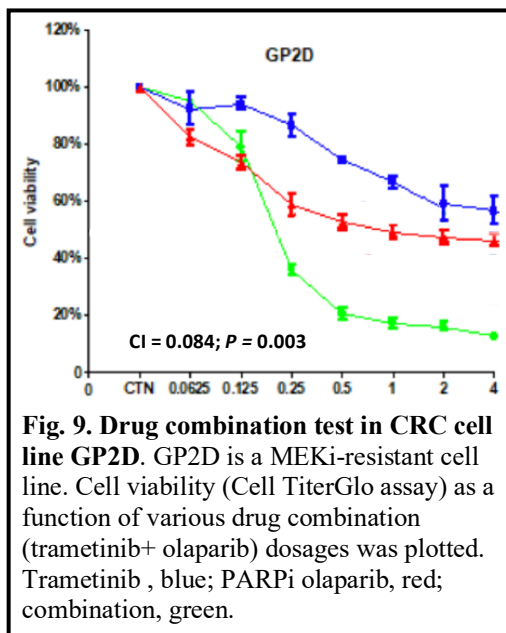


Next, we performed a genome-wide CRISPR screen in GI cancer cell lines, SW620 and HCT116 [ATCC®]. These were selected because their response to MEKi was moderate (i.e., not too sensitive). We sought to identify effective drug combinations that could work well in these cancer lines. The CRISPR screen data were processed and analyzed using MAGeCK algorithm. As a control, GI cancer-specific essential genes were those that were negatively selected in GI cancer cell lines. We further used MAGeCK to identify genes whose knockout leads to a stronger negative selection in trametinib (MEK inhibitor, MEKi) conditions compared with DMSO control in colon cancer cells (**Fig. 8**). We collected read counts of gRNAs in all conditions from Fastq files, and then normalized the counts using median normalization.

MAGeCK built a linear model to estimate the variance of read counts, evaluated the gRNA abundance changes between conditions, and assigned a P value. Finally, drug candidate genes were evaluated from the rankings of gRNAs (by their P values) using the α -robust rank aggregation (α -RRA) algorithm. Benjamini–Hochberg method was used for multiple comparison adjustments. To increase the statistical power, genes that had fewer than four gRNAs or fewer than two significant gRNAs were excluded from our analysis. MEKi and PARPi/XPO1i showed possible synergy from our CRISPR screen analysis, and were therefore subjected to further functional validation.

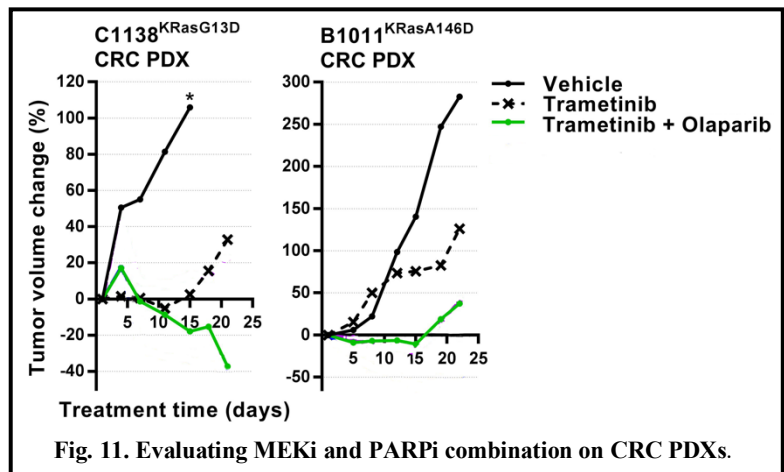


To assess the ability of MEKi and PARPi/XPO1i combination to inhibit cell growth *in vitro*, we tested a panel of 5 colorectal cancer cell lines (LS123, HCT15, DLD1, GP5D and T84) that were previously exposed to increasing concentrations of MEK inhibitors for 72h and exhibited intrinsically resistant (data not shown). These cell lines were tested in as monolayer growth conditions. Cell Titer Glo 2.0 cell viability assays were used to evaluate potential synergy between MEKi alone and in combination with the PARPi (Olaparib) /XPO1i (Eltanexor). Combination Indices were calculated using Calcsyn. We analyzed MAPK and other pathways using Western blot and qRT-PCR analysis. Our data using one resistant line (GP2D) (Fig. 9) and one sensitive line (SW620; data not shown) showed much improved synergistic (combination index $CI < 1$) effect over any single agent. These drug effects were further validated by PARP1-specific and MAP2K1-specific siRNAs (Fig. 10).

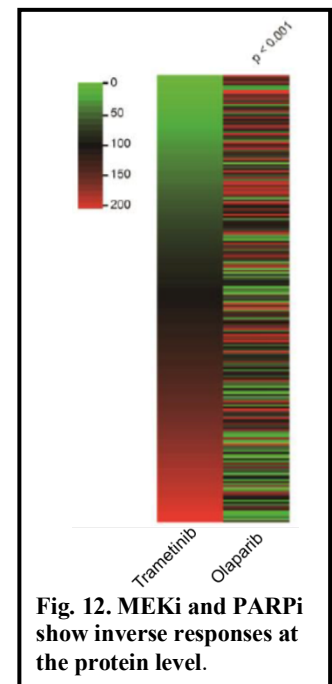


To assess organoid growth, organoids were mechanically disassociated into fragments and plated in 4ul of BME in 96-well plates. After three to four days of growth, organoids were treated with one of 6 groups: 1) control (vehicle); 2) MEKi (trametinib); 3) PARPi (olaparib); 4) MEKi+PARPi combination; 5) XPO1i (eltanexor); 6) MEKi+XPO1i combination. Cell viability was assayed using CellTiter-Glo (Promega). Treatment groups were statistically compared using a one-way ANOVA. Those reaching statistical significance were compared using the ad-hoc Student's T-test. PDO size and number were measured using

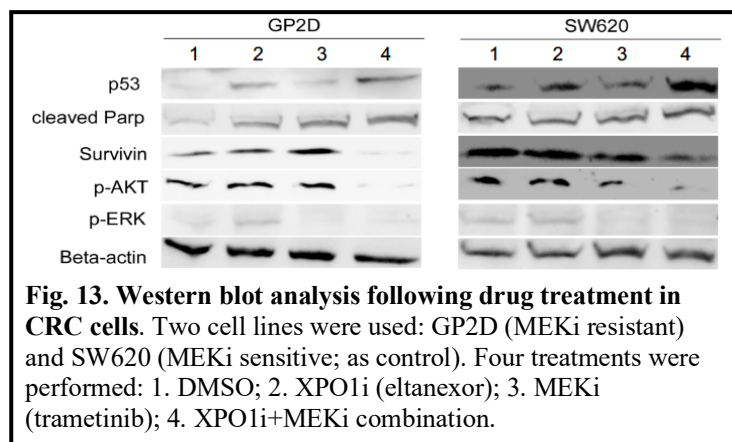
high throughput microscopy (using our Nikon HCA system). In both cases, the drug combination therapy appeared to be synergistic and showed great promise over monotherapies. For patient-derived tumor xenograft models, we completed the test of MEKi-PARPi combination. Tumor bearing mice (carrying *KRAS* mutations) were treated with vehicle (black line, mean N=3), trametinib as a single agent (0.25mpk, QD, PO, black dashed line, mean N=7) or in combination with olaparib (50mpk, BID, PO, mean N=3). The tumor volume was calculated according to the formula $V = ab^2/2$, in which a is the largest and b is the smallest diameter. Mice were euthanized when tumors reached $\sim 1500\text{mm}^3$. The response was determined by comparing the tumor volume change over time using the formula $\Delta\text{Vol} = 100\% \times ((V_t - V_b)/V_b)$. Our result (**Fig. 11**) suggests a strong synergistic effect of MEKi and PARPi in reducing CRC tumor growth.



For **Specific Aim 3**, to further dissect molecular mechanisms underlying the effective drug combinations, we have already performed one round of scRNA-seq and RPPA experiments. We profiled protein expression levels for HCT116 CRC cell lines upon treatment of MEKi (trametinib) or PARPi (olaparib), and found that the protein expression profiles are likely to be inverse (**Fig. 12**). In another study, we developed PARPi-resistant cells by culturing highly PARPi-sensitive cells in the continued presence of BMN673 (PARP inhibitor) for 3 to 4 months, at which time drug-resistant clones emerged. Our RPPA analysis demonstrates that RAS/MAPK pathway activity [increased pMEK, pBAD, and pFOXO3a (inactive form)] is upregulated in PARPi-resistant clones. Using this set of gene expression signature, we divided TCGA cancer samples into two subgroups, likely responders to PARPi, and likely resistant. We then identified common and frequent mutations that can account for an enhanced HR DNA repair. Our analysis revealed mutations in the gene *KRAS* as a marker for PARPi resistance.



Poly(ADP-ribose) polymerase 1 (PARP-1) and p53 are two key proteins in the DNA-damage response. PARP-1-mediated poly(ADP-ribosyl)ation blocks the interaction between p53 and the nuclear export receptor Crm1, resulting in nuclear accumulation of p53 and preventing genome mutation. Our analysis using western blot (**Fig. 13**) showed that XPO1 inhibition led to an increase in p53 and PARP levels. The surviving protein functions to inhibit caspase activation, thereby leading to negative regulation of apoptosis or programmed cell death, hence is depleted upon XPO1i-MEKi dual blockade (**Fig. 13**). It is also regulated by p53 and the cell cycle.



XPO1 inhibitor induces p-ERK and p-AKT (**Fig. 13**), leading to the accumulation of phosphorylated forms of MAPK proteins in the nucleus and cellular growth inhibition. Thus, inhibition of XPO1 induces cytostatic and pro-apoptotic effects in both MEKi-sensitive and MEKi-resistant cell lines at nanomolar concentrations.

To further extend the utility of our results, we built a predictive framework for drug combination predictions. Firstly, we assembled essential coding genes from the literature. Assembling the catalog of human cancer essentialome allowed us to identify genes, whose loss of function is lethal for all cell lines and those that are lethal only in a specific genetic background or environmental context. We found that while a number of core essential genes retrieved from different studies were context-dependent, the majority (960, 39.09%) of core essential genes were shared across studies. These results are consistent with the Daisy Model of gene essentiality, where the essential genes in each context are represented by a petal.

However, we still lack knowledge about the determinants of the context specificity of human essentialome. Thus, we hypothesized that the context-specific essentiality might emerge as a consequence of the genetic context of cancer. Thus, we further investigated to what extent the genetic context contributes to the context-specific essentiality. We firstly collected the mutated genes in five cell lines from GDSC. In addition, we mapped the synthetic lethal interactions from yeast to human. We found that the essential genes in each model were likely to be synthetic lethal with the mutated genes in a context-dependent manner (P -values <0.001). These results suggest a model to explain the context-specific essentiality and help understand the heterogeneity and diverse signaling in cancer. The genes that are synthetic lethal with the genetically mutated genes are more likely to be essential. These observations suggest that integrating the genome-wide mutation architecture and cell fitness landscape may help prioritizing synthetic lethal interactions.

Finally, we designed a network model SyE-NET (Synthetic Essentiality NETwork), and computationally predicted the synthetic lethal interactions in each cancer sample based on the hypothesis that context-specific essential genes and existing mutated genes in each sample form synthetic lethal pairs. In other words, if pairs of these genes are both impaired (either mutated or knocked out) in the same cell, that cell becomes lethal. As a quality control, we found that our predicted synthetic lethal interactions exhibited a significant overlap with previously published genetic interaction pairs using an orthogonal approach (P -values $<2.2E-16$). These results suggest that genetic context plays critical roles in the determination of gene essentiality as well as synthetic lethality.

What opportunities for training and professional development has the project provided?

Nothing to Report.

How were the results disseminated to communities of interest?

The PI was frequently invited to cancer research societies and universities to present her research findings.

What do you plan to do during the next reporting period to accomplish the goals?

Nothing to Report.

4. IMPACT:

What was the impact on the development of the principal discipline(s) of the project?

Nothing to Report.

What was the impact on other disciplines?

Nothing to Report.

What was the impact on technology transfer?

Nothing to Report.

What was the impact on society beyond science and technology?

Nothing to Report.

5. CHANGES/PROBLEMS:

Changes in approach and reasons for change

Nothing to Report

Actual or anticipated problems or delays and actions or plans to resolve them

Some delays occurred during the COVID pandemic which affected our research activities. However, we were able to catch up and complete the goals outlined in the project.

Changes that had a significant impact on expenditures

No significant changes.

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents:

Significant changes in use or care of human subjects

No significant changes.

Significant changes in use or care of vertebrate animals

No significant changes.

Significant changes in use of biohazards and/or select agents

No significant changes.

6. PRODUCTS:

• Publications, conference papers, and presentations

Journal publications.

Li Y, Burgman B, Khatri IS, Pentaparthi SR, Su Z, McGrail DJ, Li Y, Wu E, Eckhardt SG, **Sahni N[#]** and Yi S. e-MutPath: Computational modelling reveals the functional landscape of genetic mutations rewiring interactome networks. *Nucleic Acids Res.* 49(1):e2, 1/2021. PMCID: PMC7797045 (**#co-corresponding** author).

Chakravarty AK, McGrail DJ, Lozanoski TM, Dunn BS, Shih DJ, Cirillo KM, Cetinkaya S, Zheng WJ, Mills GB, Yi S, Jarosz DF and **Sahni N[#]**. (2022) Biomolecular condensation: a new phase in cancer research. *Cancer Discovery. OF1-OF13*. doi: 10.1158/2159-8290.CD-21-1605 (**#corresponding** author).

Hua X, Li Y, Pentaparthi SR, McGrail DJ, Zou R, Guo L, Shrawat A, Cirillo KM, Li Q, Bhat AV, Xu M, Qi D, Singh A, McGrath F, Andrews SB, Aung KL, Das J, Zhou Y, Lodi A, Mills GB, Eckhardt SG, Mendillo ML, Tiziani S, Wu E, Huang JH, **Sahni N[#]** and Yi SS. (2022) Landscape of microRNA regulatory network architecture and functional rerouting in cancer. *Cancer Research*. 2022 Oct 20;CAN-20-0371. doi: 10.1158/0008-5472.CAN-20-0371. PMID: 36265133 (**#co-corresponding** author).

Books or other non-periodical, one-time publications.

Nothing to Report.

Other publications, conference papers and presentations.

Nothing to Report.

• Website(s) or other Internet site(s)

Nothing to Report.

• Technologies or techniques

Nothing to report

• Inventions, patent applications, and/or licenses

Nothing to report

• Other Products

Nothing to report

7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

Name: Nidhi Sahni

Project Role: PI

Nearest person month worked: 1.62

Contribution to Project: Writing publications and technical assistance with all experiments

Name: Sharad Awasthi

Project Role: Postdoctoral Fellow

Nearest person month worked: 2.0

Contribution to Project: Technical assistance with all experiments

Name: Alexey Sorokin (no change)

Contribution to Project: Assistance with all PDX animal experiments

Name: Yang Li

Project Role: Postdoctoral Fellow

Nearest person month worked: 1.6

Contribution to Project: Technical assistance with experiments

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

	RECENTLY ENDED
Title:	Integrative bioinformatics and functional characterization of oncogenic driver aberrations in cancer.
Supporting Agency:	National Cancer Institute U01CA217842 (PI: Mills)
Grants Officer:	Jean C. Zenklusen zenklusj@mail.nih.gov
Performance Period:	09/06/2017 – 07/31/2022
Level of Funding:	
Project Goals	The major goal of this project is to understand which recurrent aberrations, and how the more abundant, low- frequency mutations, contribute to tumor progression.
Specific Aims	Aim 1: To implement an algorithmic framework for prediction of driver aberrations and their construction into expression-ready vectors. Aim 2: To functionalize oncogenic drivers and delineate driver relationships. Aim 3: To determine the effects of single and combination drug therapies on tumor heterogeneity and drug resistance. Aim 4: To elucidate underlying mechanisms and therapeutic liabilities engendered by driver aberrations.
Overlap:	None

	RECENTLY ENDED
Title:	Systems-guided rational design of novel epigenetics- based combination therapy in ovarian cancer
Supporting Agency:	The University of Texas MD Anderson Cancer Center Sprint for Life Research Award
Grants Officer:	Wesley Harrott, Associate VP, Research Administration 1515 Holcombe Blvd, Houston TX 77030
Performance Period:	06/01/2021 – 05/31/2022
Level of Funding:	
Project Goals	The goal of this proposal is to identify novel drug combinations in high grade serous ovarian cancer.
Specific Aims	Specific Aim 1: Exploit the efficacy of EP300-targeted combinations using in vitro organoid 3D model systems. We expect the combination will work well in our patient-derived organoid (PDO) model, providing informative data for this novel combination therapy. Specific Aim 2: Exploit the efficacy of our drug combinations using established patient-derived xenograft systems.
Overlap:	None

	RECENTLY ENDED
Title:	Functional analysis of cancer genomes
Supporting Agency:	Cancer Prevention & Research Institute of Texas Recruitment RR160021
Grants Officer:	James K.V. Willson, Chief Scientific Officer 1701 North Congress Avenue, Suite 6-127 Austin, TX 78701
Performance Period:	03/01/2016 – 02/28/2022 (NCE)
Level of Funding:	
Project Goals	The major goal of this proposal is to develop an integrative platform to functionalize missense mutations in cancers.
Specific Aims	Specific Aim 1: Establish an interdisciplinary research program integrating bioinformatics and experimental platform to study signaling networks and assess interaction perturbations in tumors with various mutations. Specific Aim 2: Interrogate interaction perturbation by mutation-induced post-translational modifications in cancer cell models. Specific Aim 3: Study non-coding variants by alteration in their binding affinity to transcription factors in cancer cell lines. Specific Aim 4: Rewire interaction networks using CRISPR- based genome editing to restore functions of cancer cells to the normal state.
Overlap:	None

	NEW		
Title:	High-throughput screening robot for drug perturbation and protein interaction alteration studies		
Effort:	Year (YYYY)	Person Months (##.##)	
	1. 2022	0.00 calendar	
Supporting Agency:	MD Anderson Cancer Center/Adverse Incident Bridge Funding, 2022-00060945-Y1		
Grants Officer:	Wesley Harrott, Associate VP, Research Administration 1515 Holcombe Blvd, Houston TX 77030		
Performance Period:	04/01/2022 – 03/31/2023		
Level of Funding:			
Project Goals	The high-throughput liquid handling robot (referred to as TECAN robot) in the Sahni laboratory has not been functioning properly since our relocation to the Houston campus, resulting in halting multiple projects. These funds are to repair this critical instrument and catch up/accelerate our research efforts.		
Specific Aims	N/A		
Overlap:	None		

	NEW		
Title:	i-SINFOP: Integrative Systems-scale Interrogation of Neomorphic Fusion OncoProteins		
Effort:	Year (YYYY)	Person Months (##.##)	
	1. 2022	1.80 calendar	
	2. 2023	1.80 calendar	
	3. 2024	1.80 calendar	
Supporting Agency:	CPRIT RP220292		
Grants Officer:	Michelle Le Beau 1701 North Congress Avenue, Suite 6-127 Austin, TX 78701		
Performance Period:	03/01/2022 – 02/28/2025		
Level of Funding:			
Project Goals	The goal of this proposal is to establish a robust systems platform to identify and characterize driver fusion oncoproteins.		
Specific Aims	Aim 1: To Identify and characterize candidate fusion oncoproteins by massively parallel functional variomics coupled with our established cellular phenotyping system. Aim 2: Goal 2: To systematically interrogate cancer fusion- mediated functional alterations in protein phase separation in live cells via our high-content imaging platform. Aim 3: Goal 3: To determine the molecular basis underlying fusion gene-driven cancer development.		
Overlap:	None		

	NEW		
Title:	Systems approach for rational design of novel arginine methylation-targeted combination therapy for TNBC		
Effort:	Year (YYYY)	Person Months (##.##)	
	1. 2022	0.00 calendar	
Supporting Agency:	MD Anderson Cancer Center/ Bridge Funding, 2022- 00060947-Y1		
Grants Officer:	Wesley Harrott, Associate VP, Research Administration 1515 Holcombe Blvd, Houston TX 77030		
Performance Period:	09/01/2022 – 08/31/2023		
Level of Funding:			

Project Goals	This funding will allow us to conduct new experiments in response to concerns from the reviewers of my DOD application. The new data obtained in the experiments, if successful, should strengthen the application and place it in strong consideration for funding upon resubmission.
Specific Aims	Aim 1: Elucidate how arginine methylation regulates ATR phosphorylation and DNA damage signaling. Aim 2: Identify optimal therapeutic efficacy and scheduling across TNBC subtypes. Aim 3: Assess the combined PRMT and PARP inhibition using our established in vivo animal models.
Overlap:	None

What other organizations were involved as partners?

Nothing to Report.

8. SPECIAL REPORTING REQUIREMENTS

None

9. APPENDICES:

e-MutPath: computational modeling reveals the functional landscape of genetic mutations rewiring interactome networks

Yongsheng Li^{1,2,†}, Brandon Burgman^{1,3,†}, Ishaani S. Khatri^{1,2}, Sairahul R. Pentaparthi¹, Zhe Su^{1,2}, Daniel J. McGrail⁴, Yang Li⁵, Erxi Wu^{1,6,7,8}, S. Gail Eckhardt^{1,3}, Nidhi Sahni^{5,9,10,*} and S. Stephen Yi^{1,2,3,11,*}

¹Department of Oncology, Livestrong Cancer Institutes, Dell Medical School, The University of Texas at Austin, Austin, TX 78712, USA, ²Oden Institute for Computational Engineering and Sciences (ICES), The University of Texas at Austin, Austin, TX 78712, USA, ³Interdisciplinary Life Sciences Graduate Programs (ILSGP), College of Natural Sciences, The University of Texas at Austin, Austin, TX 78712, USA, ⁴Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA, ⁵Department of Epigenetics and Molecular Carcinogenesis, The University of Texas MD Anderson Science Park, Smithville, TX 78957, USA, ⁶Neuroscience Institute and Department of Neurosurgery, Baylor Scott & White Health, Temple, TX 76502, USA, ⁷Department of Surgery, Texas A & M University Health Science Center, College of Medicine, Temple, TX 76508, USA, ⁸Department of Pharmaceutical Sciences, Texas A & M University Health Science Center, College of Pharmacy, College Station, TX 77843, USA, ⁹Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA, ¹⁰Program in Quantitative and Computational Biosciences (QCB), Baylor College of Medicine, Houston, TX 77030, USA and ¹¹Department of Biomedical Engineering, Cockrell School of Engineering, The University of Texas at Austin, Austin, TX 78712, USA

Received April 07, 2020; Revised October 07, 2020; Editorial Decision October 13, 2020; Accepted October 20, 2020

ABSTRACT

Understanding the functional impact of cancer somatic mutations represents a critical knowledge gap for implementing precision oncology. It has been increasingly appreciated that the interaction profile mediated by a genomic mutation provides a fundamental link between genotype and phenotype. However, specific effects on biological signaling networks for the majority of mutations are largely unknown by experimental approaches. To resolve this challenge, we developed e-MutPath (edgetic Mutation-mediated Pathway perturbations), a network-based computational method to identify candidate ‘edgetic’ mutations that perturb functional pathways. e-MutPath identifies informative paths that could be used to distinguish disease risk factors from neutral elements and to stratify disease subtypes with clinical relevance. The predicted targets are enriched in cancer vulnerability genes, known drug targets but depleted for proteins associated with side effects, demonstrating the power

of network-based strategies to investigate the functional impact and perturbation profiles of genomic mutations. Together, e-MutPath represents a robust computational tool to systematically assign functions to genetic mutations, especially in the context of their specific pathway perturbation effect.

INTRODUCTION

Genome sequencing and genome-wide association efforts have identified thousands of genetic variants across cancer types (1). Although mutations are traditionally thought to disrupt the entire gene function, it has become clear that many mutations (especially missense mutations) have a unique effect on signaling pathway perturbation (2,3), and they are therefore termed as ‘edgetic’ or ‘neomorphic’ (4–7). This leads to an exciting field of ‘functional variomics’, to study the functional effects of genomic mutations on specific molecular interactions, signaling pathways or cellular processes (2). These mutations have been observed to exert different functional effects, including oncogenic activation and tumor suppression (6). However, the functional pathways by which genetic variants lead to diverse phenotypic

*To whom correspondence should be addressed. Tel: +1 512 4955245; Fax: +1 512 4955480; Email: stephen.yi@austin.utexas.edu
Correspondence may also be addressed to Nidhi Sahni. Tel: +1 512 2379506; Fax: +1 512 4955480; Email: nsahni@mdanderson.org

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

consequences remain largely unresolved (8). Classical gene knockout or knockdown approaches have been performed to characterize the function of genes but not mutations (9–11). Indeed, it is difficult to create, express and characterize large numbers of specific mutants due to the enormous amount of time and cost involved.

Although experimental methods are critical to understand the function of individual genetic variants, it is increasingly appreciated that biological systems are formed by a large number of interacting genes or proteins (12). Interactome networks or signaling pathways exhibit distinct properties that cannot be understood by single gene-based analyses (3,13). Many mutations are believed to be edgetic, rewiring signal transduction pathways to alter cellular phenotypes (4–6). Identification of network or pathway perturbations and the consequences of such perturbations is crucial for understanding complex pleiotropic mutational effects and providing novel insights into genotype–phenotype relationships in disease (14).

While a number of computational methods have been developed to identify candidate driver mutations (15,16), the majority of these methods focus on the ‘nodes’ in the networks or pathways. The molecular interactions in the context of signal transduction networks are often neglected in these algorithms. To overcome this challenge, integration of available omics data from genome-scale projects is necessary, and could greatly facilitate the identification of candidate edgetic mutations and the perturbed pathways in disease (5,17). Thus, we proposed a novel computational method named e-MutPath (edgetic Mutation-mediated Pathway perturbations), to decode the function of mutations from the pathway perturbation perspective. We applied our network-based approach to 33 cancer types and uncovered both well-known and new cancer-associated genes and candidate driver mutations. Our results also demonstrate that e-MutPath is superior to previous methods in its unique network features, and more importantly, it systematically assigns functions to large numbers of genetic mutations, especially in the context of their specific pathway perturbation effect.

MATERIALS AND METHODS

Genome-wide genetic variants and gene expression profiles in cancer

Genome-wide somatic mutations and gene expression profile data of hepatocellular carcinoma (HCC) patients were downloaded from The Cancer Genome Atlas (TCGA) project (18). The expression of genes was measured by fragments per kilobase million (19), and genes not expressed in more than 30% patients were excluded in our analysis. In addition, another set of gene expression profile data for HCC was downloaded from the International Cancer Genome Consortium (ICGC) project (20). The normalized read count for each gene was used in our analysis. The gene expression profiles for another 32 types of cancer were also downloaded from the TCGA project (21) and processed in the same way as the HCC data.

Protein–protein interaction network and signaling pathways

A systematic map of ~16,000 high-quality human binary protein–protein interaction (PPI) network, which was identified by yeast two-hybrid (Y2H) assay, was used in our analysis (22,23). The homodimer interactions were removed and there were 15,957 interactions among 4,743 genes for further analysis. Human signaling pathways were obtained from (24).

e-MutPath algorithm

The network-based method (e-MutPath) integrates genome-wide somatic mutation profiles with gene expression and a protein interaction network to identify candidate driver mutations that perturb molecular interactions (Figure 1). Briefly, the method includes three steps: first, perturbed gene interactions were identified in cancer based on gene expression correlation analysis; second, patient-specific interaction perturbation profiles were constructed; and finally, candidate driver mutations in each cancer patient that mediated interaction perturbations were identified by integration of interaction perturbation profiles with mutation profiles.

Identifying perturbed interactions by edgetic mutations

It has been demonstrated that perturbed molecular interactions can be identified by integration of protein or gene expression profile with molecular networks (25,26). However, there are limited protein expression profiles available to use in complex diseases (27,28). Here, we integrated genome-wide gene expression profiles with interaction networks to identify patient-specific molecular interaction perturbations in cancer. The interaction networks could be defined as un-directed graph $G = (V, E)$, where $V = (v_1, v_2, v_3, \dots, v_i, \dots, v_n)$ is the set of genes represented as n nodes and $E = (e_1, e_2, e_3, \dots, e_j, \dots, e_m)$, the set of m interactions represented as edges. For each edge e_j formed by gene v_x and v_y in the network, we evaluated whether this edge was perturbed in patient k by calculating the distance d_k of this patient to other patients in the 2D space. The expression profiles of gene v_x and v_y in patient k were defined as $E_k = (e_{xk}, e_{yk})$. The distance was calculated based on gene expression profiles as follows:

$$d(k) = \sqrt{(E_k - \mu)^T S^{-1} (E_k - \mu)}$$

Where μ is the average expression and S is the covariance matrix of all patients. When we obtained the distance of each patient, the Grubbs’ test was used to determine whether there are outliers in the distribution of the distances. Specifically, the Grubbs’ test statistic is defined as:

$$G = \frac{\max(d_k - \bar{d})}{s}$$

Where $\bar{d}$ and s represent the mean and standard deviation of distance, respectively. We used the two-sided test, the hy-

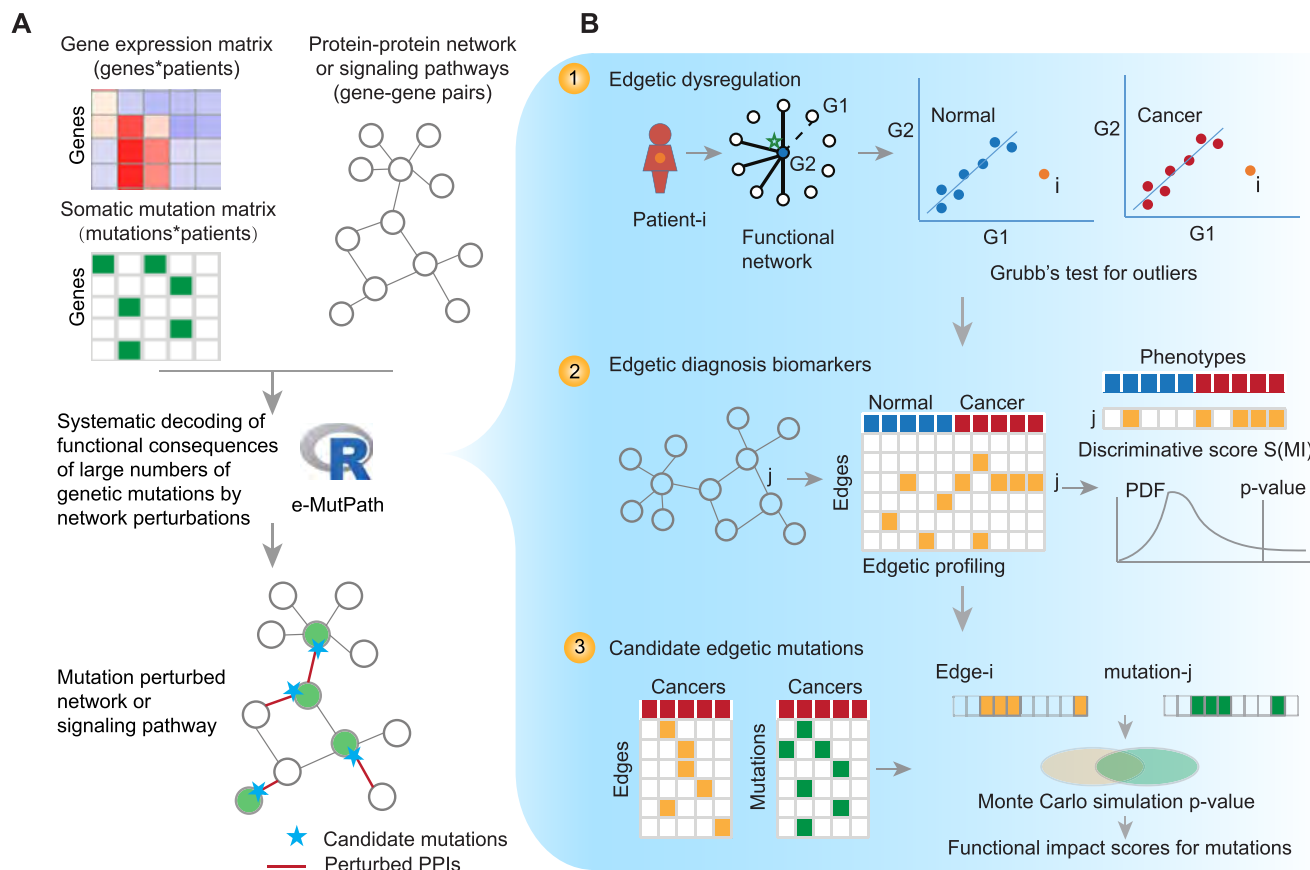


Figure 1. Overview of e-MutPath. (A) Gene expression and somatic mutations were integrated with functional networks to identify candidate driver mutations. (B) Identification of mutation-perturbed signaling pathways based on three steps. First, identification of patient-specific interaction perturbations based on gene expression correlation analysis. Second, identification of edgetic biomarkers based on discriminative scores. Third, identification of candidate edgetic driver mutations based on the number of overlapping patients with both mutations and interaction perturbations.

pothesis of no outliers was rejected if,

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{(t_{\alpha/(2N), N-2})^2}{N-2 + (t_{\alpha/(2N), N-2})^2}}$$

With $t_{\alpha/(2N), N-2}$ denoting the critical value of the t-distribution with $(N-2)$ degrees of freedom and a significance level of $\alpha/(2N)$, and N is the number of patients. We detected all the outliers in the distribution, and defined these samples as exhibiting interaction perturbation in cancer. After repeating this process for all edges in the interaction networks, we identified all the perturbed molecular interactions in specific patients.

Patient-specific molecular interaction perturbation profiles

Based on the above steps, we identified all the perturbed molecular interactions in specific patients. Next, we constructed the molecular interaction perturbation profiles. Each patient was represented as a profile of binary (1, 0) states on interaction edges, where rows represented the network edges and columns corresponded to patients. The ele-

ment M_{ij} in the matrix was defined as follow:

$$M_{ij} = \begin{cases} 1, & \text{if edge } i \text{ is perturbed in patient } j \\ 0, & \text{if edge } i \text{ is not perturbed in patient } j \end{cases}$$

In our analysis, if there were normal samples in the gene expression profiles, we required the interaction was perturbed both when comparing with normal versus cancer and among cancer patients.

Identifying driver mutation-mediated molecular interaction perturbations

To identify candidate driver mutations that mediated molecular interaction perturbations, we firstly constructed somatic mutation profiles of all patients. The somatic mutations for each patient are represented as a binary profile, in which a '1' indicates a specific mutation has occurred in the tumor. Based on the mutation profiles and molecular interaction perturbation profiles, we hypothesized that if a specific mutation and interaction perturbation co-occurred in more samples, the interaction perturbations were likely to be caused by the somatic mutation. Here, we required the mutations to occur in the genes of the patient who showed the perturbed interactions. Next, Monte Carlo simulation

was used to evaluate whether a specific mutation and a perturbed interaction significantly co-occurred in the same cancer patients. This process was repeated 1000 times and the P -value was defined as follows:

$$p = \frac{\#(N_{\text{rand}} > N_{\text{real}})}{n}$$

Where N_{rand} is the number of patients with both the mutation and the specific perturbed interaction in random conditions, N_{real} is the number of patients with both the mutation and the perturbed interaction observed, and n is the number of randomizations. Finally, we defined the functional ‘impact score’ for each mutation as ‘1 - P -value’.

Identification of informative molecular interactions for classification

To identify the discriminative edgetic biomarkers for classification of cancer and normal samples, we defined the discriminative score (S) for each edge in the interaction networks. Let $\mathbf{a}$ represent a vector of perturbation profiles for a specific edge over the tumor and normal samples, and let $\mathbf{c}$ represent the corresponding vector of class labels (normal or cancer). The discriminative score was defined as the mutual information MI between $\mathbf{a}$ and $\mathbf{c}$:

$$S = \text{MI}(\mathbf{a}, \mathbf{c}) = \sum_{x \in \mathbf{a}} \sum_{y \in \mathbf{c}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

Where $p(x, y)$ is the joint probability density function of $\mathbf{a}$ and $\mathbf{c}$, and $p(x)$ and $p(y)$ are the marginal probability density function of $\mathbf{a}$ and $\mathbf{c}$. x and y enumerate values of $\mathbf{a}$ and $\mathbf{c}$, respectively. Next, we tested whether the mutual information with the disease class is stronger than that obtained with random assignment of the labels to patients. This process was repeated 100 times, yielding a null distribution of MI scores. The real score of each edge was indexed on the null distribution to get the significant level. Network edges with Benjamini–Hochberg adjusted P -value < 0.01 were identified as candidate edgetic biomarkers.

Validation of informative molecular interaction perturbations

To validate the discriminative power of identified edgetic biomarkers for classification of normal and cancer samples, we trained three classifiers (logistic regression (logR), random forest (RF) and support vector machine (SVM)) based on edge-specific features. To measure unbiased classification performance, we used the 10-fold cross-validation method. All the samples were divided into ten subsets of equal size. Nine subsets were used as the training set to build these classifiers using the edge-specific features identified by MI score, and one subset was used as the validation set. The performance of each classifier was reported as the area under ROC curve (AUC). In addition, we constructed the classifiers based on liver cancer dataset from TCGA project and used another liver cancer dataset from ICGC as independent validation.

Cancer subtyping analysis

Based on the interaction network perturbation profiles of liver cancer samples, consensus non-negative matrix factor-

ization clustering was performed (29). To identify the subtypes, we clustered the patients by increasing $k = 2$ to $k = 7$ and the average silhouette width was calculated for selecting robust clusters. The clinical data for liver cancer patients were downloaded from the TCGA project and log-rank tests were performed to evaluate the survival difference among patients with different cancer subtypes. Moreover, we defined the severity score of each patient as the proportion of interactions perturbed by mutations.

Functional analysis of informative gene mutations

To analyze the functions of the informative gene mutations with interaction perturbations, we used a hypergeometric test to determine whether these genes were enriched in specific pathways. The P -value was calculated as:

$$P = 1 - F(N_{XY} - 1 | N, N_X, N_Y) \\ = 1 - \sum_{t=0}^{N_{XY}-1} \frac{\binom{N_X}{t} \binom{N - N_X}{N_Y - t}}{N_{XY}}$$

where N is the number of all genes (default background distribution), N_X and N_Y represent the total number of informative genes and genes in specific pathways, respectively, and N_{XY} is the number of overlapping genes that formed the informative edges in specific function, which had to be at least three. Kyoto Encyclopedia of Genes and Genomes pathways with adjusted P -values < 0.01 and including at least two genes were considered.

Evaluation of the performance

To explore whether e-MutPath could retrieve cancer-related genes, we firstly obtained 602 well-known cancer-associated genes in Cancer Gene Census (30). Moreover, we downloaded the cancer-related genes from the CancerMine resource (31), a text-mined and routinely updated database of drivers, oncogenes and tumor suppressors. The proportions of cancer-related genes, oncogenes, tumor suppressors and drivers were calculated separately for e-MutPath. In addition, we compared the proportions to four widely used computational methods: SIFT (32), Polyphen-2 (16), MutationTaster (15) and MutationAssessor (33). For each mutation, we calculated these four scores and the deleterious mutations predicted by these four methods were obtained. In SIFT, mutations were predicted as deleterious based on SIFT scores (< 0.05). For Polyphen-2, ‘probably damaging’ was called if the score was between 0.909 and 1. In MutationTaster, mutations predicted as ‘D’ (‘disease-causing’) were considered as driver mutations, and in MutationAssessor, mutations predicted as functional (H, M) were considered disease-relevant. All these mutations were mapped to genes and the proportion of cancer-related genes was calculated.

Moreover, we compared e-MutPath to a network-based method NetSig (34). First, we obtained single-gene MutSig suite Q values with the mutations as input (35). The output of MutSig and the PPI network were subjected to NetSig and the Q values were aggregated into P -values. Genes with

P -values < 0.05 were considered as cancer relevant genes. The proportions of cancer-related genes were compared.

Experimental validation of mutation-perturbed PPIs

We implemented a site-directed mutagenesis pipeline to generate specific mutations as in our previous studies (5,36). The mutation clones were sequenced by GENEWIZ and the sequences were confirmed based on the Basic Local Alignment Search Tool programs. We first checked whether the sequences were mapped to the exact genes of interest and then verified whether desired base mutations were successfully cloned into the correct position of each gene. We next used the Y2H assay to interrogate mutation-induced PPI alterations based on our previous pipeline (5,36). All protein interaction assays were performed twice independently.

RESULTS

Overview of e-MutPath

To identify the mutations that might perturb signaling pathways, we reasoned that gene–gene relationships would show perturbations in patients with specific driver mutations. We therefore developed e-MutPath as an open-source R package to identify candidate driver mutations that perturb functional pathways. Three types of omics datasets were integrated, including gene expression, somatic mutations and functional networks or pathways (Figure 1A). The output would provide prioritized mutations as well as the perturbed edges in signaling pathways or networks.

Specifically, three steps were performed in this computational method in the context of cancer. First, perturbed functional interactions were identified in each cancer patient based on a correlation perturbation analysis of RNA expression (Figure 1B, top panel). We hypothesized that cells might adapt to change in the expression pattern of the interacting proteins to respond to cellular perturbations, such as genetic mutations. All the patients were mapped to a 2D plane based on the expression levels of two interacting genes. Thus, an outlier (distant from the population) on this 2D plane might be an indicator of interaction perturbation. If a patient showed a significant deviation from a normal gene-gene relationship distribution, the patient would be an outlier in the regression line modeled by all the patients. We used Grubb's test for detecting the outliers (see 'Materials and Methods' section). Second, sample-specific interaction perturbation profiles were constructed; if gene expression data in normal samples were present, we also identified the perturbed functional subnetworks or pathways that could distinguish cancer from normal samples (Figure 1B, middle panel). Finally, candidate driver mutations in each cancer sample that mediated interaction perturbations were identified by integration of interaction perturbation patterns with mutational profiles (Figure 1B, bottom panel). We used Monte Carlo simulation to evaluate whether the patients with specific mutations were significantly overlapped with those showing gene-gene relationship perturbations (see 'Materials and Methods' section). The mutations with P -value < 0.05 were identified as candidate driver edgetic mutations.

e-MutPath identifies informative paths to distinguish disease

As a proof of principle, we applied the proposed method to HCC data obtained from TCGA project (18), which included 374 cancer and 50 normal samples. Based on the human PPI networks, we identified 477 interactions among 486 genes perturbed in HCC (Supplementary Figure S1A). We found that the majority of these interactions involved genes that had been demonstrated to play critical roles in cancer, such as TRIP13, MDFI, SPRY2 and MEOX2. Particularly, TRIP13 was found to promote cell proliferation, invasion and migration in cancer (37). In addition, MEOX2 was identified as a target gene of the TGF-beta/Smad pathway and was known to regulate cell proliferation (38). To systematically understand the function of the perturbed PPIs, we performed functional enrichment analysis based on the genes involved. Functional enrichment analysis revealed that the perturbed functional subnetworks were enriched in HCC-related functions, such as insulin signaling pathway (39) ($P = 2.29\text{E-}5$) and FoxO signaling pathway (40) ($P = 0.001$, Supplementary Figure S1B).

Next, we investigated whether the perturbed PPI profiling could provide information for cancer subtype classification. We found that based on sample-specific functional perturbation profiles, cancer and normal samples were effectively distinguished from each other (Supplementary Figure S1C). We next trained three types of machine learning methods (SVM, logR and RF) based on the perturbation profiling of patients and normal controls. Based on 10-fold cross-validation, we found that these classifiers exhibited the AUC (area under the receiver operating characteristic curve) > 0.9 (Figure 2A, AUC = 0.937–0.997). Moreover, these perturbed functional profiles were validated in another independent HCC dataset obtained from ICGC (Figure 2A, AUC = 0.814–0.942). These results suggest that e-MutPath can identify the informative functional paths to distinguish disease from normal controls, which is helpful for cancer early diagnosis.

e-MutPath stratifies cancer subtypes with clinical relevance

After identifying the perturbed PPIs in cancer, we found that some patients were clustered together (Supplementary Figure S1C), suggesting that there were several subtypes in HCC. We thus used the interaction perturbing profiles for in-depth characterization of HCC subtypes. We stratified the HCC patients based on the similarity of their perturbation profiles by consensus clustering. Based on the k -values from 2 to 7, we identified an optimum number of five HCC subtypes (Figure 2B and Supplementary Figure S2A), each consisting of 109, 67, 66, 59 and 73 patients. These subtypes were henceforth termed C1–C5. We found that there were no stage differences among these subtypes (Supplementary Figure S2B, $P = 0.06$).

Furthermore, we calculated the proportion of perturbed PPIs for each patient and defined this as a severity score (Figure 2C). We hypothesized that if the patients had higher severity scores, they would exhibit a higher risk of cancer. We found that the patients in C4 and C5 showed significantly higher severity scores (Figure 2C, $P < 0.001$). To evaluate if a high proportion of perturbed PPIs could reflect disease severity in patients harboring specific mutations,

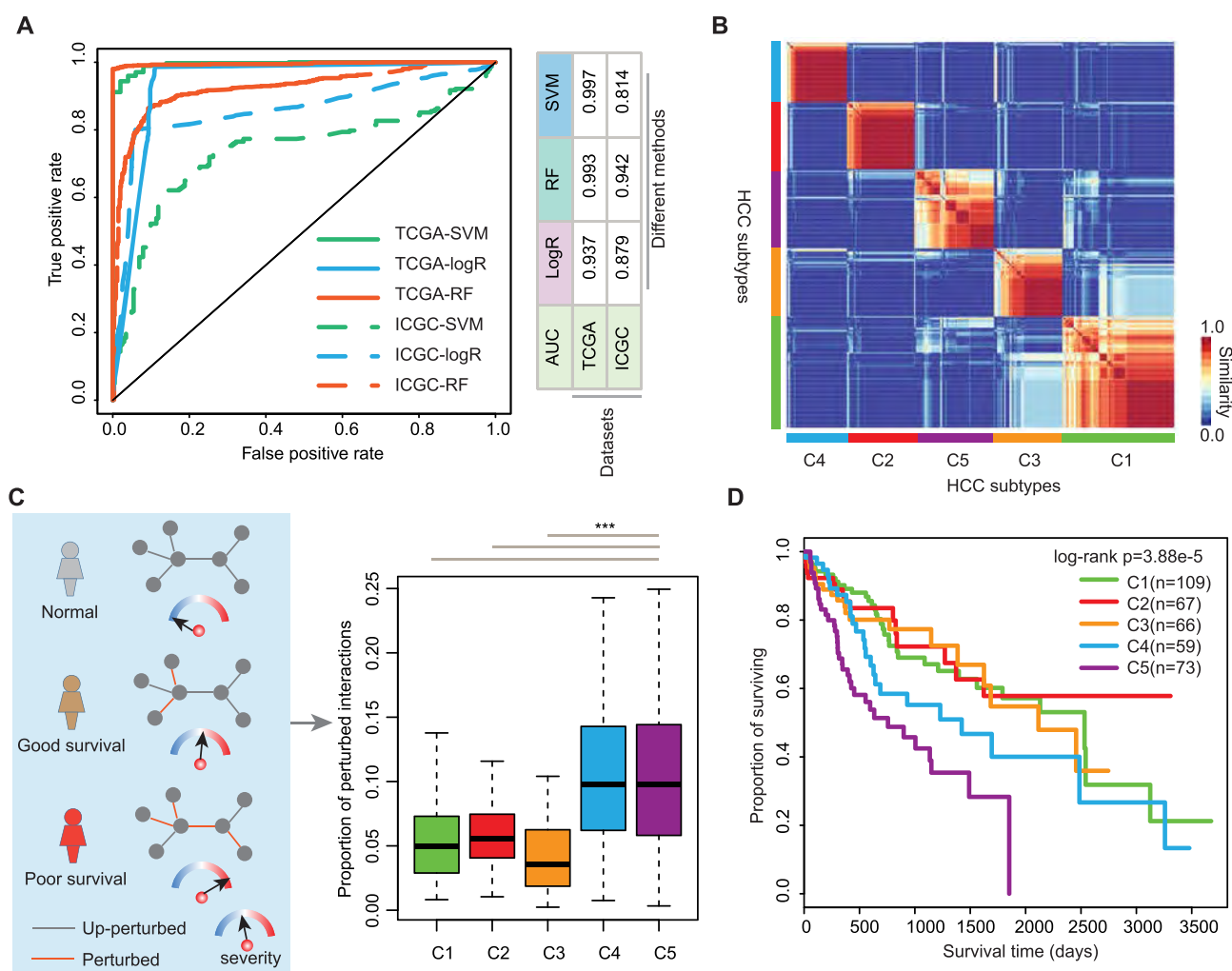


Figure 2. Identification of informative signaling perturbation paths based on e-MutPath in cancer. (A) ROC curves of different e-MutPath predictions for classifying normal and cancer patients. The solid lines are for TCGA dataset and dash lines for ICGC dataset. Lines with different colors represent three machine learning algorithms. (B) Heat map shows the similarity of HCC patients. Five subtypes are identified based on consensus clustering. (C) The proportion of perturbed PPIs in five subtypes. (D) Kaplan–Meier curve for overall survival in subtypes of HCC.

we next examined the survival rates of patients. We found indeed it was the case. Patients with mutations that perturbed a higher proportion of interactions tended to correlate with a significantly shorter survival time, indicating the deleterious nature of these mutations (Figure 2D, log-rank $P = 3.88\text{E-}5$). In addition, we analyzed cancer risks in patients by assessing their response to immunotherapy, which represents an alternative treatment approach that has been successful in many different cancer types (41). We found that the patients in C3 had higher immune scores, MHC scores and cytolytic activity than other subtypes (Supplementary Figure S2C). These results suggest that these patients are likely to respond to immunotherapy. Together, our approach is able to stratify cancer patients with distinct clinical outcomes.

e-MutPath outperforms other methods in prioritizing cancer genes

Having shown in previous sections e-MutPath could identify mutation-perturbed signaling pathways, we next evaluated its performance in uncovering cancer-related genes in 33 cancer types. We first considered the genes from Can-

cer Gene Census (CGC) (42) and found that our top predictions included a high fraction of CGC genes (Figure 3A). To illustrate the power of e-MutPath, we compared its performance with four widely used approaches-SIFT (32), Polyphen-2 (16), MutationTaster (15) and MutationAssessor (33). For the majority of cancer types, e-MutPath consistently prioritized a larger fraction of CGC genes than other methods (Figure 3A and Supplementary Table S1), demonstrating the advantage of network integration. We found that while e-MutPath predicted a smaller number of targets, they comprised larger fractions of ‘gold standard’ cancer genes by CGC (Figure 3B).

In addition, we obtained known cancer genes from CancerMine (31), which is a text-mined and routinely updated database of drivers, oncogenes and tumor suppressors. We found that the overall results were consistent for all genes (Figure 3C and Supplementary Figure S3A), oncogenes, tumor suppressors and driver genes (Figure 3D and Supplementary Figure S3B–D). Recently, a number of large-scale screens for cancer vulnerability genes using CRISPR-Cas9 and RNAi systems have been conducted (43,44). Using these datasets, we found that the predictions by e-MutPath

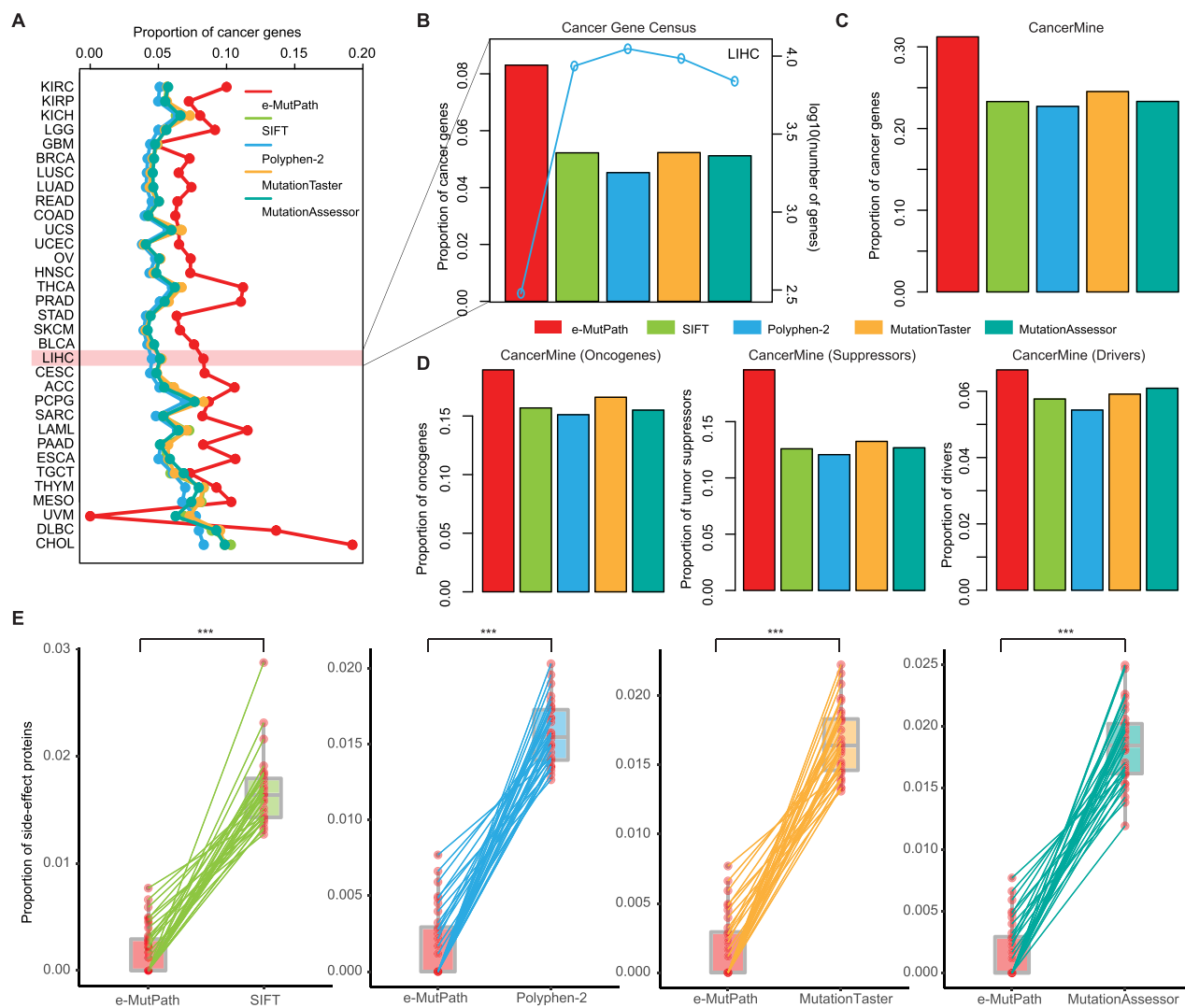


Figure 3. Performance evaluation of e-MutPath. (A) The proportion of cancer genes identified by different methods across 33 cancer types. (B) The left y-axis shows the proportion of cancer genes identified by five methods and the right y-axis shows the number of genes identified by each method. (C) The proportion of cancer genes in CancerMine identified in HCC by different methods. (D) The proportion of oncogene, tumor suppressors and driver genes in CancerMine identified in HCC by different methods. (E) The proportion of proteins associated with side-effects identified by different methods.

were more enriched for essential genes than other methods in the majority of cancer types (Supplementary Figure S4A). Particularly, the predicted genes by e-MutPath in LUSC, BLCA and LIHC showed significantly lower essentiality scores than predictions by other methods (Supplementary Figure S4B–G).

Finally, to evaluate potential targetable and side effects of the predicted genes, we used a list of 151 clinically actionable genes (45) and 237 proteins that are reported to be associated with side effects (46). We found that e-MutPath predictions exhibited similar fractions of actionable genes (Supplementary Figure S5) but were depleted of side effect-causing proteins across cancer types (Figure 3E). As methods that integrate network information could complement gene-based methods to identify new cancer genes, we then compared e-MutPath to NetSig that integrates protein interaction networks (34). e-MutPath exhib-

ited a similar power in identifying cancer-related genes as NetSig in the majority of cancer types (Supplementary Figure S6). However, it seems much more robust across cancers than NetSig (Supplementary Figure S6), suggesting that e-MutPath represents a novel network-based method to expand drug target discovery from cancer genomes. Taken together, these results demonstrate significant improvement of e-MutPath over previous state-of-the-art methods in identifying cancer-related genes.

Application of e-MutPath in pan-cancer

In order to better understand mutation-perturbed signaling pathways in cancer, we applied e-MutPath to 33 cancer types. We identified a connected network perturbed by mutations (Figure 4A and Supplementary Table S2), including a number of genes (such as MED4, TCF4 and EWSR1)

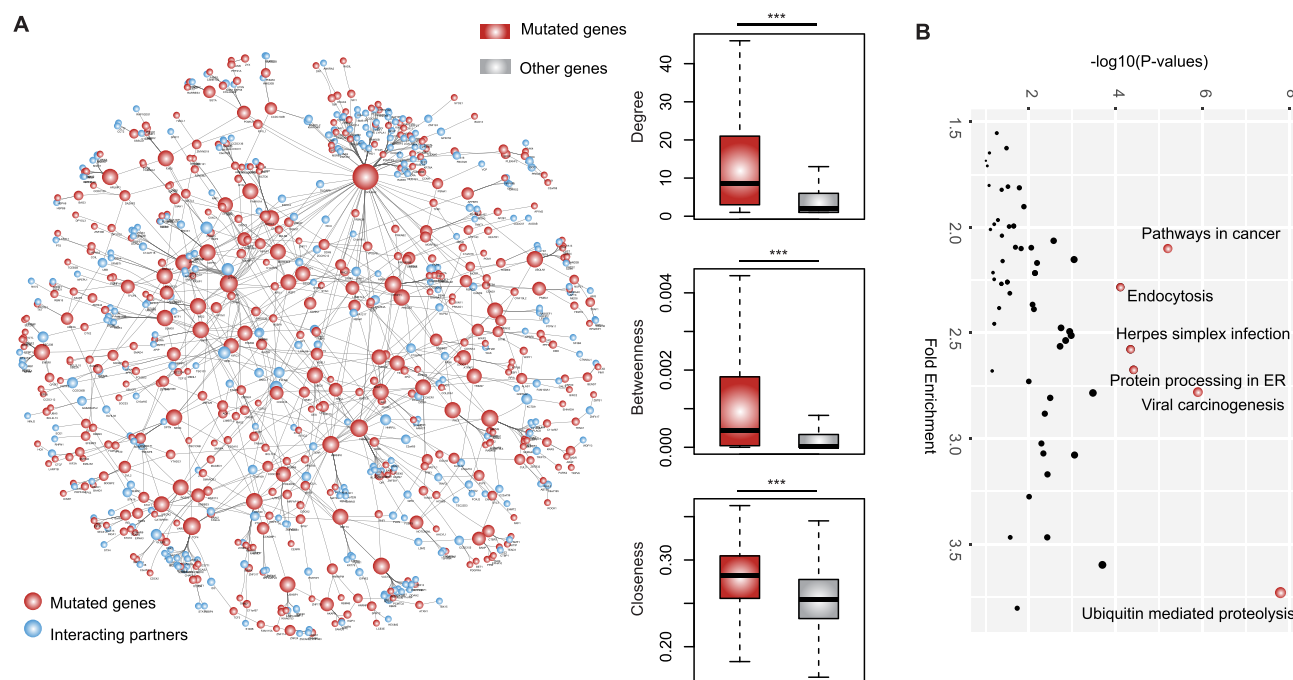


Figure 4. Interaction perturbations across cancer types. (A) The network shows perturbed interactions across 33 cancer types by mutations. Red, mutated genes; blue, interacting partners. The box plots show the topological features distribution for mutated genes and other genes in the original PPI network. Top, degree; middle, betweenness; bottom, closeness. (B) The enriched pathways by the mutated genes.

known to be involved in cancer. Cancer genes often function as network hub proteins which are involved in many cellular processes (47). We next investigated the topological features of the prioritized genes with mutations and found that they showed significantly higher degrees, betweenness and closeness (Figure 4A, P -values < 0.001). Moreover, we performed functional enrichment analysis and found that these genes were likely to be involved in cancer hallmark pathways (Figure 4B).

Particularly, we found that a candidate gene KEAP1 and the interacting partner DPP3 were highly expressed in HCC compared to normal tissues (Figure 5A). Moreover, we investigated the correlation of gene expression with patients' survival time. We found that high expression levels of the interacting partner gene were slightly associated with poor prognosis in HCC (Figure 5B). However, we found that the combined KEAP1-DPP3 signature could significantly distinguish the patients with different survival (Figure 5B, $P = 0.02$). e-MutPath identified one missense mutation in KEAP1 (G379V) that perturbed the interaction between KEAP1 and DPP3 in HCC. This perturbation prediction between KEAP1 and DPP3 was further validated by Y2H experiment. Lines of evidence have implied that the p62-Keap1-Nrf2 axis plays an important role in tumorigenesis (48,49). Collectively, our results suggest that these mutations likely play roles in cancer by perturbing the interactions involved in cancer-related pathways.

Signaling pathway perturbations in cancer

To further assess if the perturbed pathways by edgetic mutations were functionally relevant in specific disease context, we next integrated mutational profiles, gene expression patterns and signaling pathways to identify the perturbed pathways in HCC. Based on e-MutPath, we identified the

top ranked HCC mutations which were mostly enriched in centrosomal genes (CEP70, CEP76 and CEP135), JAK1, EGFR and HNF1A (Figure 6A). JAK1 encodes a cytoplasmic tyrosine kinase that is associated with a variety of cytokine receptors and plays a critical role in cell proliferation, survival and differentiation (50). The mutated forms of JAK1 might alter the activation of JAK/STAT pathways, and thus contribute to cancer development and progression. Moreover, we identified the hotspot in-frame mutation EGFR A767_V769dup, which perturbed the EGFR receptor tyrosine kinase signaling pathway in HCC (Figure 6B). The EGFR signaling pathway had been shown to play a key role in chronic liver damage, as well as in cirrhosis and HCC (51). Our results highlighted the importance of EGFR mutations in the development of liver diseases. Moreover, we identified five mutations in HNF1A ranked in the top 50 hits, which perturbed the protein interactions among HNF1A, MAPK8, PPARD and IGF1BP1 (Figure 6B). The dysregulation of the HNF1A and PPARD signaling had been demonstrated to play fundamental roles in HCC (52). In summary, e-MutPath not only identifies critical mutations in cancer, but also provides a systems-level mechanistic view to illustrate the functions of large numbers of candidate driver mutations.

DISCUSSION

In this study, we introduced a network-based method, e-MutPath, for decoding the functional consequences of mutations by network or pathway perturbations. This method incorporates individual mutational profiles with interaction networks or signaling pathways, which is a powerful approach for uncovering cancer-related genes. Our method is based on the 'edgotype' concept proposed recently and therefore complements existing frequency-based or node

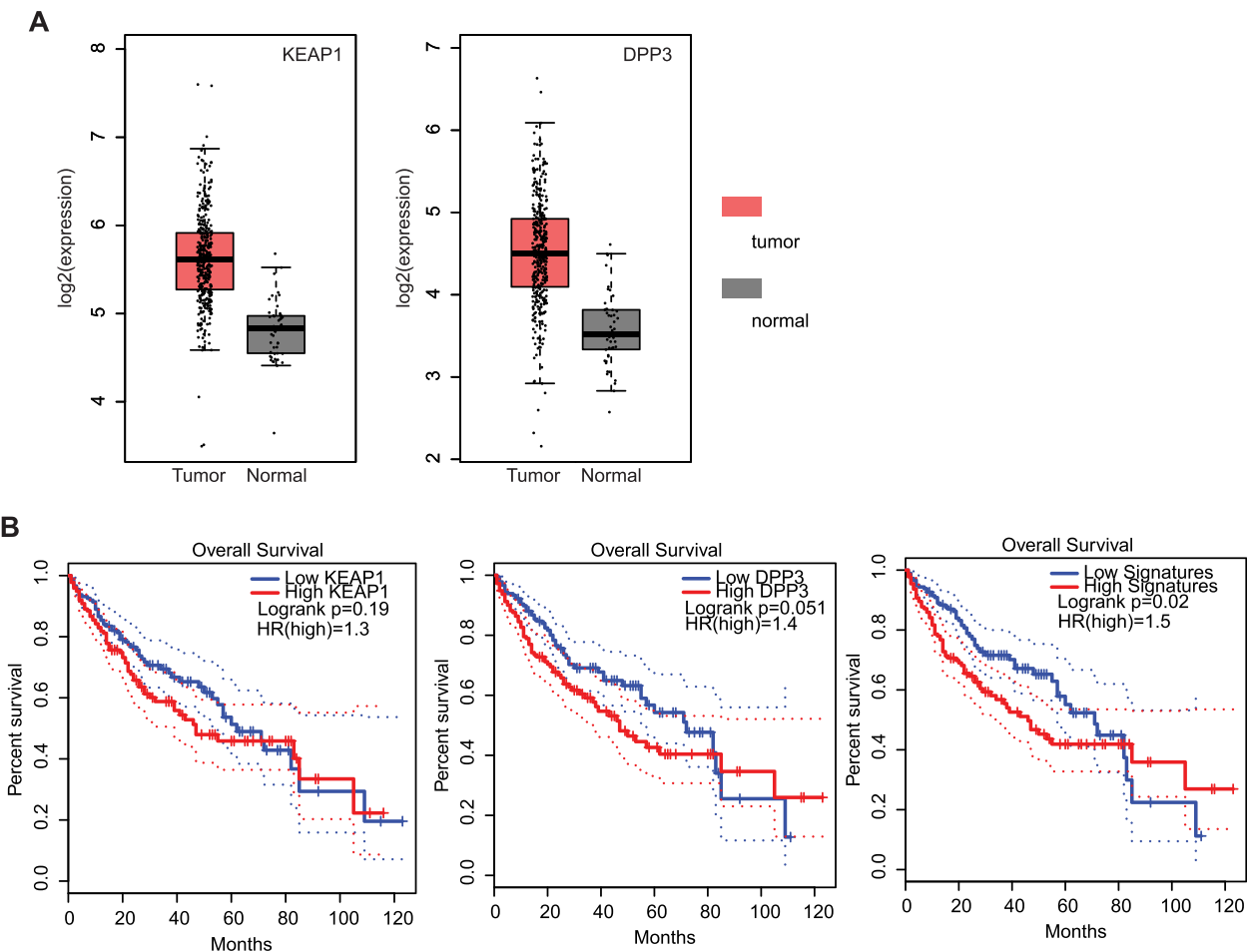


Figure 5. Perturbation of the KEAP1-DPP3 interaction by mutations in HCC. (A) The distribution of KEAP1 and DPP3 expression in cancer and normal samples. Red, tumor; gray, normal. (B) Kaplan–Meier curve for overall survival for patients based on their KEAP1, DPP3 expression or combination.

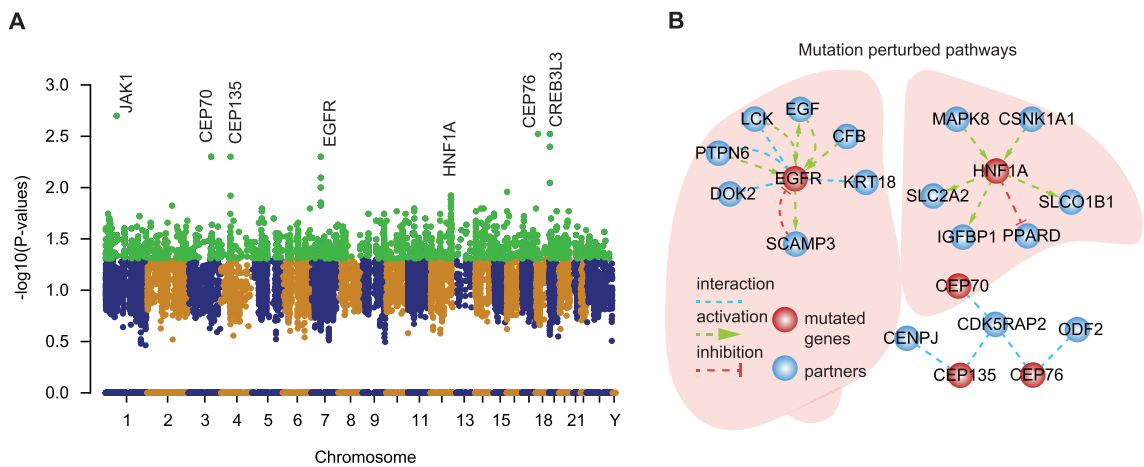


Figure 6. Perturbed signaling pathways by mutations identified by e-MutPath. (A) Manhattan plot of the mutations in HCC. (B) Top ranked mutation-perturbed signaling pathways in HCC.

(protein)-based methods (53,54). We have demonstrated by the wide application of e-MutPath to 33 cancer types that, functional mutations disrupt interactions involving genes previously implicated in the development and progression of cancer, providing complementary evidence for their functional impact. The identified genes are also depleted in proteins that are associated with side-effect. Even more interestingly, we found that e-MutPath could identify the perturbed pathways to stratify patients into different cancer subtypes. Network-based stratification (NBS) is another method to integrate somatic tumor genomes with gene networks (55). However, we found that there were no survival differences among the patients in different subtypes identified by NBS (Supplementary Figure S7). These results suggest that e-MutPath is a versatile tool that provides an effective framework for functional characterization of mutations in cancer and with immediate clinical relevance.

It is important for identifying edgetic perturbations by analyzing deviations in the gene expression correlation profiles of cancer patients. We thus evaluated the robustness of e-MutPath by using different proportions of samples in the gene expression profiles. We found that there were high consistencies among perturbed interactions when using different proportions of patients (Supplementary Figure S8). These results suggest that e-MutPath is robust to noise in the gene expression dataset. Moreover, we applied e-MutPath to pan-cancer and identified a connected network perturbed by mutations. Although increasing numbers of methods have been developed to identify the mutated subnetworks, such as HotNet2 (56), e-MutPath is conceptually distinguished from these methods. In brief, HotNet2 uses a new heat diffusion kernel analogous to random walk with restart that better captures the local topology of interaction networks. However, we are still lack of knowledge whether the interactions among genes were perturbed by mutations. In the contrast, e-MutPath addresses this issue by integrating mutation data and gene expression profiles.

In the future, e-MutPath can be extended in a number of ways. For example, while it currently analyzes only mutations within genes, other alterations are also observed in cancer. Alternative splicing and gene-fusions have been found to perturb signaling pathways (57,58), considering both types of alterations will increase the power of our approach. Second, e-MutPath may also benefit from incorporating the development of sample-specific network. While we currently consider the interaction networks or signaling pathways, considering the context specificity of these networks or pathways may be derived from recently published methods (59). Moreover, we have demonstrated e-MutPath to identify the perturbed PPIs and signaling pathways. It can also extend to other regulatory networks, such as RNA binding protein regulatory networks, miRNA-gene regulatory networks (13,60). We have applied e-MutPath across 33 cancer types and have shown that it is broadly effective in identifying cancer-related genes. However, cancers of the same tissue can often be grouped into distinct subtypes based on molecular features (18). With the development of high-throughput sequencing technology or single cell sequencing (61), e-MutPath could be used to study how different cancer subtypes yield different network or path-

way perturbations and decode the function of mutations in the subtype context. As it becomes feasible to obtain the DNA-seq and RNA-seq data from single cell sequencing, e-MutPath can be potentially extended to these datasets.

Collectively, e-MutPath could systematically uncover and stratify the functional consequences of myriad genetic mutations identified in distinct patients stricken by a variety of human diseases. Our new method identifies specific candidate driver mutations as well as their perturbed functional networks or pathways by integrating functional omics datasets at the systems level. We anticipate the proposed computational method will facilitate improved biological understanding of the function of disease variants toward personalized precision medicine.

DATA AVAILABILITY

Methods developed in this study and the codes for e-MutPath are freely available at the GitHub (<https://github.com/lyshaerbin/eMutPath>).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We were also grateful to contributions from TCGA Research Network Analysis Working Group. We also acknowledge the High-Performance Computing Facility at MD Anderson, Biomedical Research Computing Facility at UT Austin and Texas Advanced Computing Center (TACC) for computing assistance.

FUNDING

National Institutes of Health [R35 GM133658 to S.Y.]; Susan G. Komen [CCR19609287 to S.Y.]; American Association for the Study of Liver Diseases (AASLD) (Pinnacle Research Award in Liver Disease to N.S.); U.S. Department of Defense [W81XWH-18-PRCRP-CDA CA181455 to N.S.]; Cancer Prevention and Research Institute of Texas [RR160093 to S.G.E.]; NCI [K99CA240689 to D.J.M.]. Funding for open access charge: Susan G. Komen [CCR19609287 to S.Y.].

Conflict of interest statement. None declared.

REFERENCES

1. Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C. and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
2. Yi, S., Lin, S., Li, Y., Zhao, W., Mills, G.B. and Sahni, N. (2017) Functional variomics and network perturbation: connecting genotype to phenotype in cancer. *Nat. Rev. Genet.*, **18**, 395–410.
3. Wang, Y., Sahni, N. and Vidal, M. (2015) Global edgetic rewiring in cancer networks. *Cell Syst.*, **1**, 251–253.
4. Fuxman Bass, J.I., Sahni, N., Shrestha, S., Garcia-Gonzalez, A., Mori, A., Bhat, N., Yi, S., Hill, D.E., Vidal, M. and Walhout, A.J.M. (2015) Human gene-centered transcription factor networks for enhancers and disease variants. *Cell*, **161**, 661–673.

5. Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J.I., Coulombe-Huntington, J., Yang, F., Peng, J., Karras, G.I., Wang, Y. *et al.* (2015) Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, **161**, 647–660.
6. Ng, P.K., Li, J., Jeong, K.J., Shao, S., Chen, H., Tsang, Y.H., Sengupta, S., Wang, Z., Bhavana, V.H., Tran, R. *et al.* (2018) Systematic functional annotation of somatic mutations in cancer. *Cancer Cell*, **33**, 450–462.
7. Zhong, Q., Simonis, N., Li, Q.R., Charleatoux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., Venkatesan, K., Mou, D. *et al.* (2009) Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.*, **5**, 321.
8. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A. Jr and Kinzler, K.W. (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.
9. Bouhaddou, M., Eckhardt, M., Chi Naing, Z.Z., Kim, M., Iderker, T. and Krogan, N.J. (2019) Mapping the protein-protein and genetic interactions of cancer to guide precision medicine. *Curr. Opin. Genet. Dev.*, **54**, 110–117.
10. Shendure, J. and Akey, J.M. (2015) The origins, determinants, and consequences of human mutations. *Science*, **349**, 1478–1483.
11. Stein, A., Fowler, D.M., Hartmann-Petersen, R. and Lindorff-Larsen, K. (2019) Biophysical and mechanistic models for disease-causing protein variants. *Trends Biochem. Sci.*, **44**, 575–588.
12. Vidal, M., Cusick, M.E. and Barabasi, A.L. (2011) Interactome networks and human disease. *Cell*, **144**, 986–998.
13. Li, Y., McGrail, D.J., Xu, J., Mills, G.B., Sahni, N. and Yi, S. (2018) Gene regulatory network perturbation by genetic and epigenetic variation. *Trends Biochem. Sci.*, **43**, 576–592.
14. Li, Y., McGrail, D.J., Xu, J., Li, J., Liu, N.N., Sun, M., Lin, R., Pancsa, R., Zhang, J., Lee, J.S. *et al.* (2019) MERIT: systematic analysis and characterization of mutational effect on rna interactome topology. *Hepatology*, **70**, 532–546.
15. Schwarz, J.M., Cooper, D.N., Schuelke, M. and Seelow, D. (2014) MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods*, **11**, 361–362.
16. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
17. Woodsmith, J., Apelt, L., Casado-Medrano, V., Ozkan, Z., Timmermann, B. and Stelzl, U. (2017) Protein interaction perturbation profiling at amino-acid resolution. *Nat. Methods*, **14**, 1213–1221.
18. Cancer Genome Atlas Research Network. Electronic address, w.b.e. and Cancer Genome Atlas Research, N. (2017) Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell*, **169**, 1327–1341.
19. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
20. Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B. *et al.* (2011) International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database*, **2011**, bar026.
21. Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V. *et al.* (2018) Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, **173**, 291–304.
22. Rolland, T., Tasan, M., Charleatoux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R. *et al.* (2014) A proteome-scale map of the human interactome network. *Cell*, **159**, 1212–1226.
23. Li, Y., Sahni, N. and Yi, S. (2016) Comparative analysis of protein interactome networks prioritizes candidate genes with cancer signatures. *Oncotarget*, **7**, 78841–78849.
24. Cui, Q., Ma, Y., Jaramillo, M., Bari, H., Awan, A., Yang, S., Zhang, S., Liu, L., Lu, M., O'Connor-McCourt, M. *et al.* (2007) A map of human cancer signaling. *Mol. Syst. Biol.*, **3**, 152.
25. Xin, J., Ren, X., Chen, L. and Wang, Y. (2015) Identifying network biomarkers based on protein-protein interactions and expression data. *BMC Med. Genet.*, **8**(Suppl. 2), S11.
26. Liu, K.Q., Liu, Z.P., Hao, J.K., Chen, L. and Zhao, X.M. (2012) Identifying dysregulated pathways in cancers from pathway interaction networks. *BMC Bioinformatics*, **13**, 126.
27. Li, J., Lu, Y., Akbani, R., Ju, Z., Roebuck, P.L., Liu, W., Yang, J.Y., Broom, B.M., Verhaak, R.G., Kane, D.W. *et al.* (2013) TCPA: a resource for cancer functional proteomics data. *Nat. Methods*, **10**, 1046–1047.
28. Li, J., Zhao, W., Akbani, R., Liu, W., Ju, Z., Ling, S., Vellano, C.P., Roebuck, P., Yu, Q., Eterovic, A.K. *et al.* (2017) Characterization of human cancer cell lines by reverse-phase protein arrays. *Cancer Cell*, **31**, 225–239.
29. Brunet, J.P., Tamayo, P., Golub, T.R. and Mesirov, J.P. (2004) Metagenes and molecular pattern discovery using matrix factorization. *PNAS*, **101**, 4164–4169.
30. Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L. *et al.* (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.*, **45**, D777–D783.
31. Lever, J., Zhao, E.Y., Grewal, J., Jones, M.R. and Jones, S.J.M. (2019) CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat. Methods*, **16**, 505–507.
32. Ng, P.C. and Henikoff, S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
33. Reva, B., Antipin, Y. and Sander, C. (2007) Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.*, **8**, R232.
34. Horn, H., Lawrence, M.S., Chouinard, C.R., Shrestha, Y., Hu, J.X., Worstell, E., Shea, E., Ilic, N., Kim, E., Kamburov, A. *et al.* (2018) NetSig: network-based discovery from cancer genomes. *Nat. Methods*, **15**, 61–66.
35. Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S. and Getz, G. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, **505**, 495–501.
36. Yi, S., Liu, N.N., Hu, L., Wang, H. and Sahni, N. (2017) Base-resolution stratification of cancer mutations using functional variomics. *Nat. Protoc.*, **12**, 2323–2341.
37. Sheng, N., Yan, L., Wu, K., You, W., Gong, J., Hu, L., Tan, G., Chen, H. and Wang, Z. (2018) TRIP13 promotes tumor growth and is associated with poor prognosis in colorectal cancer. *Cell Death. Dis.*, **9**, 402.
38. Valcourt, U., Thuault, S., Pardali, K., Heldin, C.H. and Moustakas, A. (2007) Functional role of Meox2 during the epithelial cytostatic response to TGF-beta. *Mol. Oncol.*, **1**, 55–71.
39. Bechmann, L.P., Hannivoort, R.A., Gerken, G., Hotamisligil, G.S., Trauner, M. and Canbay, A. (2012) The interaction of hepatic lipid and glucose metabolism in liver diseases. *J. Hepatol.*, **56**, 952–964.
40. Tikhonovich, I., Cox, J. and Weinman, S.A. (2013) Forkhead box class O transcription factors in liver function and disease. *J. Gastroenterol. Hepatol.*, **28**(Suppl. 1), 125–131.
41. Hu, Z., Ott, P.A. and Wu, C.J. (2018) Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nat. Rev. Immunol.*, **18**, 168–182.
42. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I. and Forbes, S.A. (2018) The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer*, **18**, 696–705.
43. Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M. *et al.* (2017) Defining a cancer dependency map. *Cell*, **170**, 564–576.
44. Aguirre, A.J., Meyers, R.M., Weir, B.A., Vazquez, F., Zhang, C.Z., Ben-David, U., Cook, A., Ha, G., Harrington, W.F., Doshi, M.B. *et al.* (2016) Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. *Cancer Discov.*, **6**, 914–929.
45. Li, J., Han, L., Roebuck, P., Diao, L., Liu, L., Yuan, Y., Weinstein, J.N. and Liang, H. (2015) TANRIC: an interactive open platform to explore the function of lncRNAs in cancer. *Cancer Res.*, **75**, 3728–3737.
46. Kuhn, M., Al Banchaabouchi, M., Campillos, M., Jensen, L.J., Gross, C., Gavin, A.C. and Bork, P. (2013) Systematic identification of proteins that elicit drug side effects. *Mol. Syst. Biol.*, **9**, 663.
47. Wang, E., Lenferink, A. and O'Connor-McCourt, M. (2007) Cancer systems biology: exploring cancer-associated genes on cellular networks. *Cell. Mol. Life Sci.*, **64**, 1752–1762.

48. Kageyama, S., Saito, T., Obata, M., Koide, R.H., Ichimura, Y. and Komatsu, M. (2018) Negative regulation of the Keap1-Nrf2 pathway by a p62/Sqstm1 splicing variant. *Mol. Cell. Biol.*, **38**, e00642-17.
49. Katsuragi, Y., Ichimura, Y. and Komatsu, M. (2015) p62/SQSTM1 functions as a signaling hub and an autophagy adaptor. *FEBS J.*, **282**, 4672–4678.
50. Xie, H.J., Bae, H.J., Noh, J.H., Eun, J.W., Kim, J.K., Jung, K.H., Ryu, J.C., Ahn, Y.M., Kim, S.Y., Lee, S.H. *et al.* (2009) Mutational analysis of JAK1 gene in human hepatocellular carcinoma. *Neoplasia*, **56**, 136–140.
51. Komposch, K. and Sibilio, M. (2015) EGFR signaling in liver diseases. *Int. J. Mol. Sci.*, **17**, 30.
52. Patitucci, C., Couchy, G., Bagattin, A., Caneque, T., de Reynies, A., Scoazec, J.Y., Rodriguez, R., Pontoglio, M., Zucman-Rossi, J., Pende, M. *et al.* (2017) Hepatocyte nuclear factor 1alpha suppresses steatosis-associated liver cancer by inhibiting PPARgamma transcription. *J. Clin. Invest.*, **127**, 1873–1888.
53. Jia, P. and Zhao, Z. (2014) VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data. *PLoS Comput. Biol.*, **10**, e1003460.
54. Cho, A., Shim, J.E., Kim, E., Supek, F., Lehner, B. and Lee, I. (2016) MUFFINN: cancer gene discovery via network analysis of somatic mutation data. *Genome Biol.*, **17**, 129.
55. Hofree, M., Shen, J.P., Carter, H., Gross, A. and Ideker, T. (2013) Network-based stratification of tumor mutations. *Nat. Methods*, **10**, 1108–1115.
56. Leiserson, M.D., Vandin, F., Wu, H.T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M. *et al.* (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.*, **47**, 106–114.
57. Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G.M., Hao, T., Richardson, A., Sun, S., Yang, F., Shen, Y.A., Murray, R.R. *et al.* (2016) Widespread expansion of protein interaction capabilities by alternative splicing. *Cell*, **164**, 805–817.
58. Latysheva, N.S., Oates, M.E., Maddox, L., Flock, T., Gough, J., Buljan, M., Weatheritt, R.J. and Babu, M.M. (2016) Molecular principles of gene fusion mediated rewiring of protein interaction networks in cancer. *Mol. Cell*, **63**, 579–592.
59. Magger, O., Waldman, Y.Y., Rupp, E. and Sharan, R. (2012) Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Comput. Biol.*, **8**, e1002690.
60. Li, Y., Zhang, Y., Li, X., Yi, S. and Xu, J. (2019) Gain-of-function mutations: an emerging advantage for cancer biology. *Trends Biochem. Sci.*, **44**, 659–674.
61. Baslan, T. and Hicks, J. (2017) Unravelling biology and shifting paradigms in cancer with single-cell sequencing. *Nat. Rev. Cancer*, **17**, 557–569.

Biomolecular Condensation: A New Phase in Cancer Research



Anupam K. Chakravarty¹, Daniel J. McGrail², Thomas M. Lozanoski³, Brandon S. Dunn⁴, David J.H. Shih⁵, Kara M. Cirillo⁴, Sueda H. Cetinkaya⁴, Wenjin Jim Zheng⁵, Gordon B. Mills⁶, S. Stephen Yi^{7,8,9}, Daniel F. Jarosz^{10,11}, and Nidhi Sahni^{4,12,13}

ABSTRACT

Multicellularity was a watershed development in evolution. However, it also meant that individual cells could escape regulatory mechanisms that restrict proliferation at a severe cost to the organism: cancer. From the standpoint of cellular organization, evolutionary complexity scales to organize different molecules within the intracellular milieu. The recent realization that many biomolecules can “phase-separate” into membraneless organelles, reorganizing cellular biochemistry in space and time, has led to an explosion of research activity in this area. In this review, we explore mechanistic connections between phase separation and cancer-associated processes and emerging examples of how these become deranged in malignancy.

Significance: One of the fundamental functions of phase separation is to rapidly and dynamically respond to environmental perturbations. Importantly, these changes often lead to alterations in cancer-relevant pathways and processes. This review covers recent advances in the field, including emerging principles and mechanisms of phase separation in cancer.

INTRODUCTION

Malignant transformation is exemplified by the breakdown of regulatory processes that keep the proliferation of postmitotic cells under control. Key among these are maintenance of genomic integrity, regulation of gene expression, and cell signaling. Maintenance of proper subcellular localization is critical for these processes, and the roles of organelles such as the nucleus, mitochondria, and the Golgi apparatus in these functions have been investigated for decades. Yet it has recently been appreciated that subcellular organization includes myriad organelle-like structures that are not membrane bound. Many biomolecules can undergo a process known as liquid–liquid phase separation (LLPS), which can

sequester specific proteins and nucleic acids into membraneless organelles (MLO), reorganizing cellular biochemistry in space and time (Fig. 1A). The past decade has featured an explosion of research activity in this area, and many excellent reviews have summarized its mechanistic underpinnings and roles in development and neurologic diseases (1–6). By contrast, the exploration of these concepts in cancer has begun more recently (1, 7–13). In this review, we present evidence that known cancer drivers are enriched in genes that encode proteins that undergo phase separation (Fig. 1B). Furthermore, we explore mechanistic connections between phase separation and cancer-associated processes and how these can become deranged in disease. Finally, we discuss how LLPS may provide new therapeutic opportunities in oncology.

¹Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, Michigan. ²Center for Immunotherapy and Precision Immuno-Oncology, Cleveland Clinic, Cleveland, Ohio. ³Department of Bioengineering, Stanford University, Stanford, California. ⁴Department of Epigenetics and Molecular Carcinogenesis, The University of Texas MD Anderson Cancer Center, Houston, Texas. ⁵School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, Texas. ⁶Department of Cell, Developmental and Cancer Biology, Knight Cancer Institute, Oregon Health and Sciences University, Portland, Oregon. ⁷Department of Oncology, Livestrong Cancer Institutes, The University of Texas at Austin, Austin, Texas. ⁸Department of Biomedical Engineering, The University of Texas at Austin, Austin, Texas. ⁹Interdisciplinary Life Sciences Graduate Programs (ILSGP) and Oden Institute for Computational Engineering and Sciences (ICES), The University of Texas at Austin, Austin, Texas. ¹⁰Department of Chemical and Systems Biology, Stanford University School of Medicine, Stanford, California. ¹¹Department of Developmental Biology, Stanford University School of Medicine, Stanford, California. ¹²Program in Quantitative and Computational Biosciences (QCB), Baylor College of Medicine, Houston,

Texas. ¹³Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas.

Note: A.K. Chakravarty and D.J. McGrail contributed equally as co-first authors of this article.

Corresponding Authors: Nidhi Sahni, The University of Texas MD Anderson Cancer Center, 1901 East Road, Houston, TX 77230. Phone: 713-563-7742; E-mail: nsahni@mdanderson.org; Daniel F. Jarosz, 318 Campus Drive, Stanford, CA 94305. Phone: 650-725-6859; E-mail: jarosz@stanford.edu; and S. Stephen Yi, The University of Texas at Austin, 1601 Trinity Street, Austin, TX 78712. Phone: 512-495-5245; E-mail: stephen.yi@austin.utexas.edu

Cancer Discov 2022;12:2031–43

doi: 10.1158/2159-8290.CD-21-1605

This open access article is distributed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.

©2022 The Authors; Published by the American Association for Cancer Research

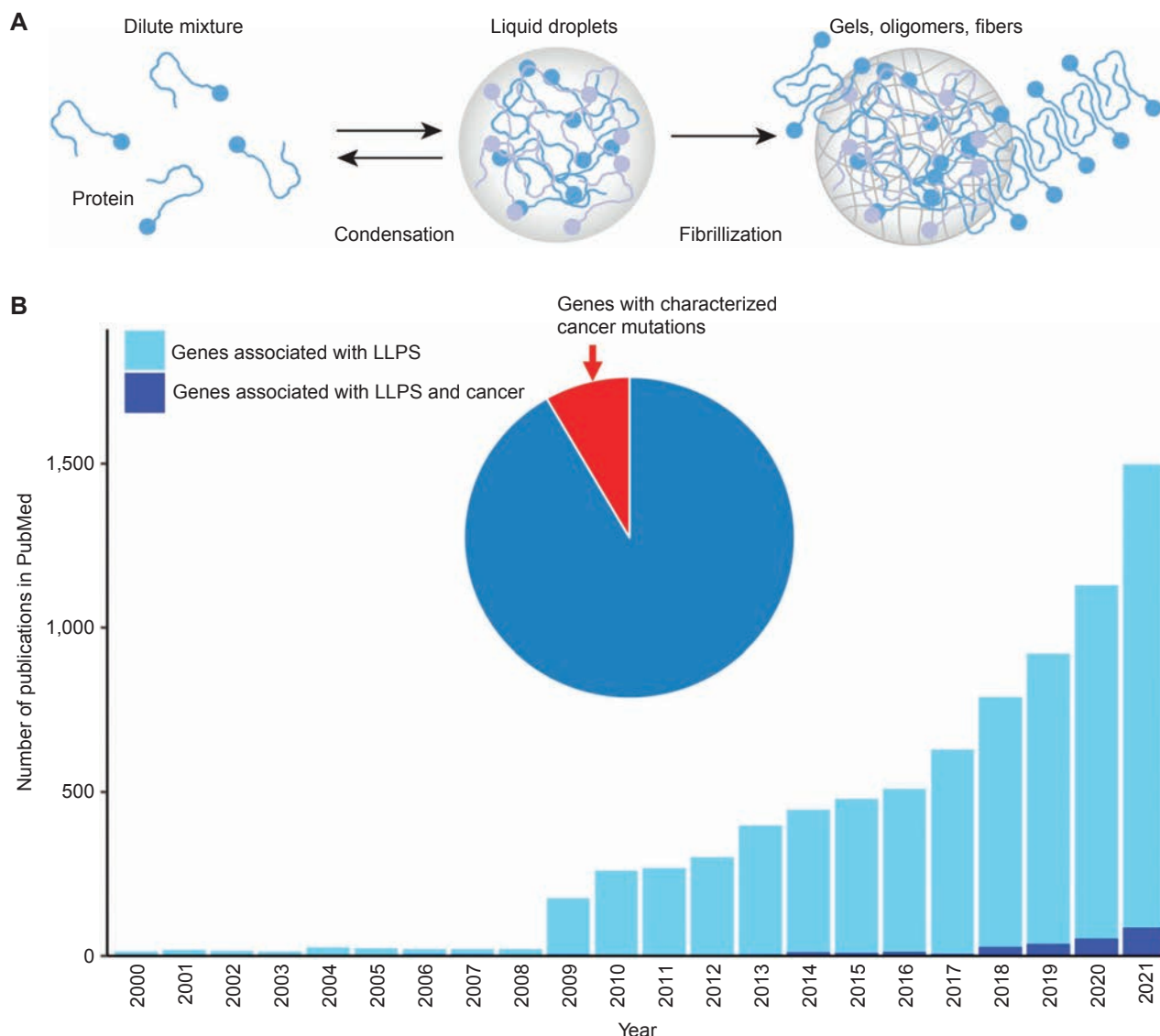


Figure 1. An emerging field in LLPS in cancer. **A**, Biomolecular polymers, such as proteins with intrinsically disordered regions, can undergo phase transitions resulting in liquid droplets with higher local density of the protein and physical properties distinct from the surroundings. Over time, the liquid droplets can mature into other phase-separated species such as gels, oligomers, or fibers. **B**, Increasing appreciation in studies over the last two decades exploring the connections between LLPS and cancer. Cyan bars represent the number of publications found in PubMed related to LLPS, which has steadily but continuously increased since 2000. Blue bars show a tremendous increase in the number of publications related to cancer and LLPS over the past 4 years. The pie chart reveals that only about 10% of LLPS genes annotated as cancer genes by the Cancer Gene Census have been studied to understand the molecular mechanisms behind cancer mutations.

MECHANISMS OF PHASE SEPARATION

Phase separation refers to a phenomenon in which biological macromolecules demix into separate dilute and dense phases that are bounded by a surface and exhibit distinct physical properties (2). This phenomenon, driven by noncovalent, transient interactions between biopolymers, organizes proteins, and nucleic acids into MLOs. MLOs can manifest as liquids, gels, glasses, crystalline solids, and even multiphase bodies (e.g., a solid core enveloped by a liquid) and can be classified by the arrangement (e.g., crystalline or amorphous), density, and mobility of their constituent

molecules (14, 15). MLOs form in a concentration-dependent manner. Once the concentration of a phase-separating protein reaches a critical threshold, protein-protein interactions become more thermodynamically favorable than protein-solvent interactions, and the system undergoes a phase transition into dense and dilute phases. Remarkably, in single-component phase separation, further increases in protein concentration cause the MLOs to grow in volume but retain the same concentration. Changing concentrations in multicomponent phase separation by contrast vary partition coefficients of individual components in a manner that minimizes the free energy of the condensate (16). Phase

separation is often reversible and can be controlled by the posttranslational modification of constituent proteins or changes to the local physicochemical environment such as salt concentration, pH, temperature, and pressure (17). It is also possible for the nucleating protein species that initiated the MLO formation to recruit other molecules (e.g., members of multiprotein complexes, ligands, chaperones, other MLO-forming proteins), thereby creating a mixture of species in the MLO. MLOs are dynamic structures, and over time condensates can even harden into solids or evolve into multiphase bodies (refs. 15, 18; Fig. 1A).

Deciphering the biological importance of phase separation has sparked intense research interest. Multiple well-studied cellular structures are now appreciated to exist as MLOs, such as the nucleolus and stress granules (19, 20). MLOs allow cells to spatiotemporally organize metabolic pathways and signaling cascades, store biomolecules for later use, or adapt to stressful conditions (21, 22). However, a number of neurodegenerative diseases have been associated with the formation of aberrant aggregates, such as amyloid fibers, that may arise from phase-separated MLOs (4). The potential for phase-separated, membraneless organelles to participate in normal cellular processes—and the derangement of such processes typical of disease states—depends on the functions of constituent molecules, the biophysical properties of the phase-separated body, and their interaction with the proteostasis network.

A key feature of proteins that phase separate is the capacity to form multiple weak, transient, noncovalent interactions with other molecules. One way to achieve this property is via repeated binding domains within a biomolecule (typically referred to as multivalency). Many proteins with extended intrinsically disordered regions (IDR) are also capable of forming an MLO. IDRs do not have a single, stable conformation in solution but rather explore a range of conformations, which allows them to form weak, transient interactions with many binding partners. These regions are often depleted of hydrophobic residues and typically contain more polar and charged residues, but the mechanisms through which IDRs drive phase transition are still being uncovered. For example, some mechanisms involve the patterning of charged residues throughout the IDR, whereas others rely on infrequent aromatic residues for π - π stacking interactions. Proteins with IDRs are very common in the human proteome (~35% of all amino acids are predicted to be disordered) and have weak sequence level conservation over evolutionary time (23). The presence of IDRs, despite their lack of primary sequence conservation, and specific physicochemical features of these regions are often maintained across large evolutionary distances in regulators of transcription, translation, and signaling (24, 25).

SYSTEMS-WIDE ENRICHMENT OF PHASE SEPARATION IN CANCER-ASSOCIATED PROTEINS

Given the enrichment for IDRs in signaling and gene regulatory functions—each biochemical processes with strong links to malignancy—we explored the 287 genes with annotated cancer hallmarks from the Catalogue of Somatic Mutations

in Cancer (COSMIC) for their ability to encode proteins that can undergo phase separation (26). There are multiple annotated public databases (Table 1) available for studying LLPS that cover phase-separating proteins, MLOs, phase-separating RNA, IDRs, posttranslational modifications, and short linear motifs. Furthermore, there are many computational tools available for analyzing molecular features related to LLPS within protein, RNA, and DNA sequences (Table 1). Fifty-three COSMIC proteins are annotated to undergo LLPS in at least two databases (27–30), a strong enrichment relative to all genes with annotated biological processes from the Gene Ontology Consortium (GO; ref. 31; $P < 1.41 \times 10^{-6}$ by permutation test; Fig. 2A). Remarkably, comparisons with other disease processes indicate that phase-separating proteins are more enriched in cancer driver genes than any other evaluated disease process, including neurodegeneration (ref. 32; Fig. 2B). This enrichment spans the majority of cancer hallmarks, including maintenance of genomic stability (e.g., *BLM*, *ERCC2*, *ATM*, and *ATR*), aberrant proliferation signaling (e.g., *MYC*, *ESR1*, and *ABL1*), and regulation of gene expression (e.g., *BRD4*, *CTCF*, and *SF3B1*; Fig. 3A and B).

DERANGEMENT OF PHASE SEPARATION IN CANCER

Spatiotemporal Reorganization of Cellular Material

Despite enrichment for phase separation among cancer genes, and the expanding appreciation for this phenomenon as an organizing principle in cell biology, only relatively recently have investigators begun to examine connections between phase separation and malignancy. The maintenance of a proper proteostatic equilibrium is essential for healthy cells (33). Oncogenic transformation places many additional demands on protein quality control systems, including molecular chaperones, the ubiquitin proteasome system, and autophagy (34, 35). Hubs of many of these protein quality control systems, including the autophagosome and proteasome, have now been proposed to be organized via the principles of phase separation (36, 37). Cancer-specific mutations in SPOP, a tumor suppressor and adaptor for the BTB-CUL3-RBX1 E3 ubiquitin ligase responsible for proteasomal targeting, provide a striking example. Mutations in SPOP are associated with many types of cancer (e.g., prostate, breast, endometrial, gastric, kidney, and ovarian). It has long been known that these mutations can impair substrate recruitment by the ligase. But deeper analysis has recently revealed that substrate interactions with the SPOP adapter drive phase separation that is critical for E3 ligase activity and ensuing proteolytic turnover (38). Inhibition of these interactions, as occurs for SPOP mutations associated with diverse cancers, thus leads to the accumulation of proto-oncogenic proteins (38).

In addition to its role in substrate-directed phase separation, SPOP regulates the assembly of stress granules, which form via phase separation (8). Stress granules play key roles in cellular adaptation to stress and are often upregulated in cancers. Analysis of SPOP-interacting proteins provided a clue to a possible mechanism. The SPOP interactor Caprin1 promotes nucleation of stress granules and is upregulated

Table 1. Computational resources, tools, and methods for studying LLPS

Annotated public databases for studying LLPS				
Database	Entries	Type	Description	Availability
DrLLPS	9,285	Protein	Manually curated to compile known and predicted LLPS proteins.	http://llps.biocuckoo.cn/
PhaSePro	121	Protein	Manually curated to list proteins that drive LLPS.	https://phasepro.elte.hu/
LLPSDB	273	Protein	Manually curated to list LLPS-related proteins.	http://www.bio-comp.org.cn/llpsdb/home.html
PhaSepDB	961	Protein	Manually curated to list LLPS- and MLO-related proteins.	http://db.phasep.pro/
AmyPro	162	Protein	Manually curated to list validated amyloid precursor proteins and prions.	https://amypro.net/#/
DisProt	932	Protein	Manually curated to identify IDPs.	https://disprot.org/
ZipperDB	20,000+	Protein	Collection of protein regions that have the tendency to form fibrils.	https://services.mbi.ucla.edu/zipperdb/
MloDisDB	719	MLO	Manually curated to showcase the relation between MLOs and diseases.	http://mloDis.phasep.pro/
RPS	21,613	RNA	Manually curated to list validated RNA involved in LLPS.	http://rps.renlab.org/#/Home
RNAPhaSep	1,113	RNA	Manually curated to compile known LLPS-related RNAs.	http://www.rnaphasep.cn/#/Home
Phospho-SitePlus	484,496	PTM	Curated to list commonly studied PTMs that were observed experimentally.	https://www.phosphosite.org/homeAction.action
dbPTM	2,235,664	PTM	Collection of experimentally verified protein PTMs.	https://awi.cuhk.edu.cn/dbPTM/
MobiDB	219,740,215	IDR	Collection of IDRs and similar features on proteins.	https://mobidb.bio.unipd.it/
LMPID	1,762	Motif	Manually curated to identify linear motifs that mediate protein interactions.	http://biresources.jcbose.ac.in/ssaha4/lmpid/
ELM DB	3,934	Motif	Manually curated to identify short linear motif instances.	http://elm.eu.org/
Computational tools for analyzing molecular features related to LLPS				
Tool	Focus and methodology			Availability
PLAAC	Identifies prion-like domains in proteins using a hidden Markov model.			http://plaac.wi.mit.edu/
IUPred3	Identifies disordered protein regions based on an energy estimation method.			https://iupred3.elte.hu/
PSPer	First <i>in silico</i> tool to identify prion-like RNA-binding phase separation proteins.			https://www.bio2byte.be/b2btools/psp/
PSPredictor	Predicts phase separation proteins using machine learning with LLPSDB.			http://www.pkumdl.cn:8000/PSPredictor/
PScore	Predicts phase separation proteins based on pi-interaction frequencies.			https://elifesciences.org/articles/31486
catGRANULE	Predicts granule propensity based on multiple features like RNA binding.			http://service.tartagliolab.com/update_submission/475471/abfa5da35f
dSCOPE	Predicts key regions of proteins for LLPS using a machine learning algorithm, random forest.			http://dscope.omicsbio.info/
BIAPSS	Provides interactive statistical analyses of LLPS proteins listed in PhaSePro and LLPSDB for inferring phase-separating affinities and determining biophysical statistics of LLPS proteins.			https://biapss.chem.iastate.edu/
PrionW	Predicts prion-like domains and amyloid nucleating cores from protein sequences.			http://bioinf.uab.cat/prionw/
MusiteDeep	Provides a deep-learning framework for predicting and visualizing PTMs in proteins.			https://www.musite.net/

Table 1. Computational resources, tools, and methods for studying LLPS (Continued)

Computational tools for analyzing molecular features related to LLPS		
Tool	Focus and methodology	Availability
SLiMAN	Provides a webserver for analyzing and predicting linear motifs and interactions in proteins.	http://sliman.cbs.cnrs.fr/LIMIP/limip_index.py
ESpritz	Predicts protein disorder using bidirectional recursive neural networks.	http://old.protein.bio.unipd.it/espritz/
PLATINUM	Analyzes hydrophobic and hydrophilic characteristics of proteins and other molecules.	https://model.nmr.ru/platinum/
PSAP	Provides a machine learning classifier that analyzes proteins and predicts phase separation.	https://github.com/Guido497/phase-separation
CIDER	Predicts and annotates protein sequences for IDRs/IDPs for their conformational properties.	http://pappulab.wustl.edu/CIDER/
DisMeta	Provides a meta-server that has predictors for disordered and low-complexity regions in proteins.	https://montelionelab.chem.rpi.edu/dismeta/
DILIMOT	Predicts linear motifs within protein sequences using interaction maps.	http://dilimot.russelllab.org/
ELM Prediction	Predicts functional sites on proteins using linear motif analyses and extensive filters.	http://elm.eu.org/search.html
PAPA	Predicts prion-forming propensity for proteins by focusing on amino acid compositions.	https://combi.cs.colostate.edu/supplements/papa/
DisoRDPbind	Predicts residues in IDRs that bind to protein, RNA, or DNA using a multilayered model.	http://biomine.cs.cvu.edu/servers/DisoRDPbind/
LARKSdb	Predicts LARKS formation and regions involved with reversible amyloid formation and LLPS.	https://srv.mbi.ucla.edu/LARKSdb/
MobiDB-lite	Predicts IDRs after executing 9 different predictors and filtering results for accuracy.	https://github.com/BioComputingUP/MobiDB-lite
PrionScan	Predicts prion domains on proteins using a model based on amino acid propensities.	http://webapps.bifi.es/prionscan
ProteinLIPS	Predicts locally unstable regions of proteins by identifying light interfaces of high polarity.	http://webapps.bifi.es/lips
ANCHOR	Predicts protein-binding regions within disordered proteins using pairwise energy estimation.	http://anchor.elte.hu/

Abbreviations: IDR, intrinsically disordered protein; PTM, posttranslational modification.

in prostate cancer (where SPOP is most frequently mutated; ref. 39). This arises because mutant SPOP can no longer recognize Caprin1 as a substrate, leading to its accumulation and subsequent increases in stress granule assembly. As a collateral consequence, cancer cells gain the ability to adapt and even thrive under stressful environmental conditions.

Cancer Fusions

In addition to aberrations in phase separation induced by mutations to single genes, cancer fusions can exert a strong impact on this emergent property, leading to either gain or loss of condensates (Fig. 4). Cancer-related fusion proteins often rewire protein interaction networks (40–43). Well-known examples that alter phase separation include members of the FET protein family (EWS and FUS) and the RTK family (ALK and RET). Fusions of EWS and FUS to transcription factors (TF) are known drivers of both Ewing sarcoma and myxoid liposarcomas (44, 45). Aberrant phase separation driven by the prion-like domain within the disordered

sequence of EWS–FLI1 promotes altered BAF-dependent chromatin remodeling and the establishment of oncogenic gene expression programs (44). The formation of phase-separated foci of EML4–ALK preserves the kinase activity of ALK but excludes repressive factors, mediating oncogenic signaling involving MAPK, PLC γ , and PI3K pathways (46). These oncogenic fusion condensates locally concentrate the RAS-activating complex and activate the MAPK signaling cascade via membraneless protein granules (47). In addition, NUP98 fusion oncoproteins such as NUP98–HOXA9, which is associated with pediatric leukemias, can form nuclear puncta through combined homotypic and heterotypic interactions, fueling aberrant transcriptional rewiring and oncogenic transformation (41–43). Finally, the oncogenic fusion DnAJB1–PKA, which is frequently detected in liver cancer, abolishes PKA-mediated phase separation that is critical for cAMP compartmentation and cancer signaling (48).

Cancer-related fusion proteins are often enriched in IDRs. Dysregulation of IDRs could cause errors in the location and

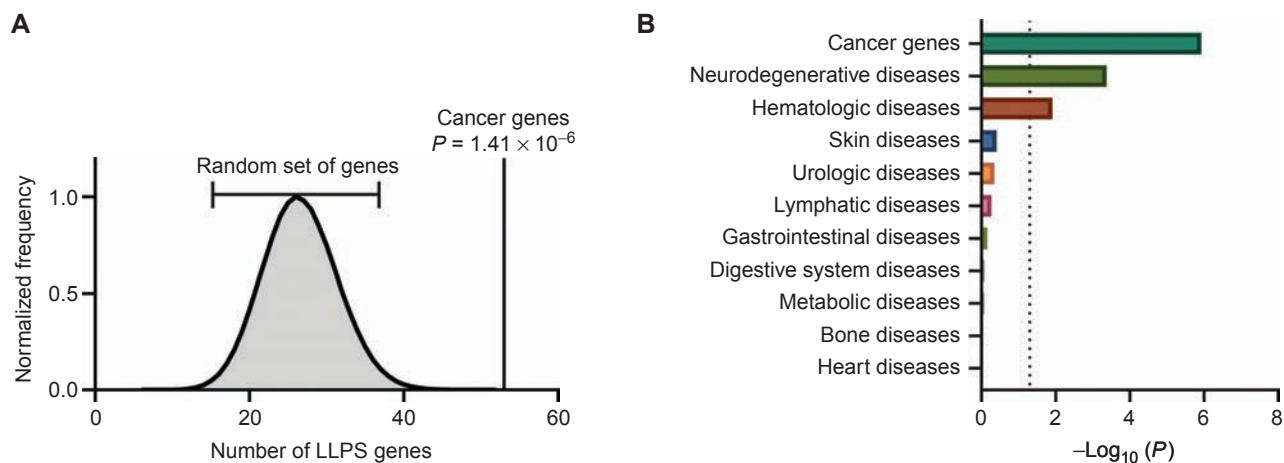


Figure 2. LLPS is enriched in cancer. **A**, Of the 287 genes with annotated cancer hallmarks in COSMIC, 53 undergo LLPS (vertical line), significantly more than observed when selecting 287 random genes with GO annotated biological processes (histogram). **B**, Comparison of enrichment for cancer-associated genes from **A** with those linked to other human diseases as well as genes that are associated with multiple diseases.

trafficking of molecules in the cell (40, 49). There has been a rapidly increasing appreciation of the role of IDRs—long considered the dark matter of the proteome—in the proteome (40). IDRs evolve more rapidly than ordered protein regions (23). This has primarily been attributed to reduced constraints on preserving globular structures, but others have argued that IDRs can frequently drive the adaptive evolution of cells that harbor them (50). Fusion of IDRs can occur via chromosomal translocations that often drive cancers; the relatively high rates at which such fusion events occur and their potential to rewire intracellular organization, signaling, and gene expression may provide a crucible for rapid cancer clonal evolution (51).

PHASE SEPARATION AND AGGREGATION IN CANCER-RELATED PROCESSES

DNA Damage Response

Ionizing radiation and radiomimetic substances such as DNA cross-linkers and alkylating agents cause double-strand breaks (DSB) in DNA. If left unrepaired, DSBs can drive the genomic instability that is a hallmark of most tumors. The rapid, site-specific, and transient assembly of DNA repair factors at these sites of damage is essential to repair DSBs and prevent genomic instability. A multivalent, polyanionic, nucleic acid-like molecule called poly-ADP ribose (PAR) is synthesized at sites of DNA damage and plays a crucial role in the DNA damage signaling network (52). Because multivalency in biomolecules often fuels phase separation, the capacity of PAR to influence such spatiotemporal reorganization was investigated. Live cell imaging demonstrated that PAR nucleates the phase separation of several intrinsically disordered proteins at sites of DNA damage (53). Indeed, this is one of the earliest steps in DNA damage signaling. Notable among these proteins are FET family members such as FUS, EWS, and TAF15. These proteins harbor repetitive RGG motifs that are cationic at cellular pH, favoring interaction with negatively charged PAR, which drives phase separation.

Importantly, these liquid droplets concentrate a selective cohort of proteins and do not readily mix with foci formed at DNA damage sites by the genomic caretaker protein 53BP1 (p53-binding protein 1; refs. 53, 54).

53BP1 is a master regulator of DSB signaling and repair by nonhomologous end joining (NHEJ). During the G_1 phase of the cell cycle, in the absence of a sister chromatid, NHEJ pathways are the primary mode of DNA repair responsible for placing the broken DNA ends in proximity and ligating them. Spatiotemporally distinct from PAR-driven phase separation events, 53BP1 also undergoes DNA damage-induced LLPS. This features a hallmark of liquid droplets: frequent fusions and occasional fission of these foci. Condensation is primarily driven by the C-terminal oligomerization domain of 53BP1. Disruption of 53BP1 condensates destabilizes interactions with p53, reducing its ability to induce expression of downstream effectors such as p21 (55). 53BP1 condensates have therefore been implicated in coordinating local DNA damage recognition with global changes in downstream gene expression programs (55). A critical role of a noncoding damage-inducible long noncoding RNA transcribed at DNA damage sites in controlling the liquid-like properties of 53BP1 condensates has also been investigated (56). These few examples already suggest that phase separation plays a critical role in the DNA damage response. Indeed, proteins involved in DNA repair often form dynamic foci associated with activity (57, 58). Further investigations will no doubt continue to shed light on the role of posttranslational modifications in regulating such phenomena and how this process is altered in, or perhaps even acts as a cause of, oncogenic transformation.

Classic Proliferation Marker Ki-67

Histologic markers of cancer were the gold standard in the clinic prior to the advent of genetic biomarkers and remain widely used. For example, the Ki-67 labeling index formulated on the eponymous protein is a classic marker for proliferation (59, 60). It has been routinely used in the

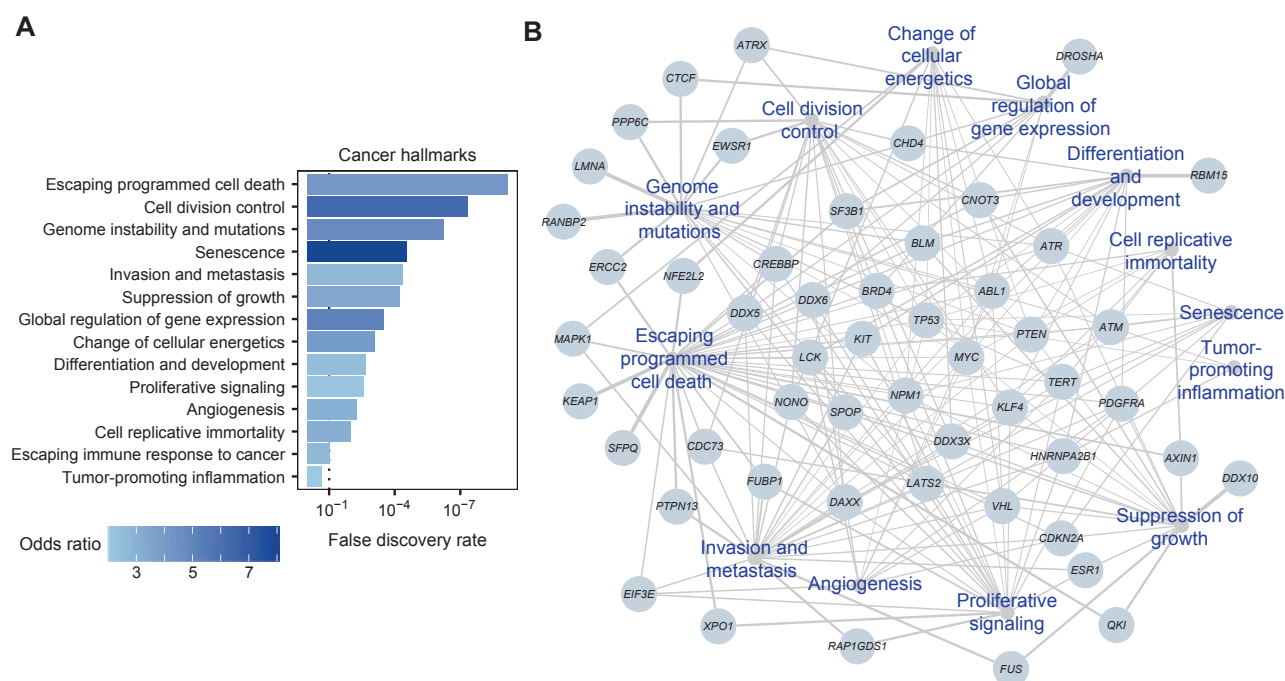


Figure 3. Cancer hallmarks represented by phase separation genes. **A**, Enrichment of specific cancer hallmarks from LLPS genes. **B**, Bipartite network representing LLPS genes and corresponding pathways from **A**.

clinic to identify and stage a wide variety of cancers ranging from breast cancers to neuroendocrine tumors (61, 62). Despite its enormous importance, little was known about the cellular mechanism of the Ki-67 protein. Recently, in a targeted siRNA screen, Ki-67 was found to play a key role in maintaining chromosome separation (63), consistent with its localization to the chromosome periphery (64). Further investigation revealed that Ki-67 was required to maintain spatial separation after nuclear envelope breakdown. The Ki-67 protein consists of an N-terminal region comprising a phosphopeptide-binding Forkhead-associated domain and a protein phosphatase 1-binding site, a central region comprising 16 tandem repeats, and a C-terminal region enriched in leucine-arginine pairs (65). The C-terminal region interacts with chromatin and is essential for the individualization of mitotic chromosomes. However, it is not sufficient on its own. Mutational analyses revealed that coupling the chromatin-targeting region with either the N-terminal region or the central region with tandem repeats restored this function. Strikingly, even a scrambled N-terminal region coupled to the essential chromatin-targeting region is sufficient for restoring its function. This leads to the model in which the general physicochemical properties of these regions and not their sequence are of primary importance in its function. The existing model suggests that Ki-67 behaves as an amphiphilic surfactant, coating the chromosome and forming a phase boundary that allows appropriate spatial separation of chromosomes (63). This example highlights the unexpected ways in which the emergent properties of phase separation lie at the heart of many hallmarks of cancer.

p53

The ability of MLOs to compartmentalize cellular space occurs in the cytoplasm as well as the nucleoplasm. Cajal bodies and promyelocytic leukemia (PML) nuclear bodies are examples of LLPS phenomena in the nucleoplasm. The PML protein provides the primary scaffold for these micron-sized, subnuclear organelles, and additional multivalent interactions further stabilize them (66). Several lines of evidence implicate PML nuclear bodies as key mediators of the tumor suppressor functions of p53. Transcriptional activation of classic p53 target genes such as *Bax* and *p21* is severely impaired when p53 does not localize to PML bodies (67). Furthermore, PML^{-/-} cells have a reduced propensity to senesce or undergo apoptosis, even upon p53 activation (68). In the MCF-7 breast carcinoma cell line, the p53 signaling cascade is impaired by PML knockdown (69). The PML protein also protects p53 from proteasomal degradation. Under conditions of DNA damage stress, PML sequesters Mdm2 in the nucleolus, thus preventing p53 ubiquitination and degradation (70). *WRAP53*, a gene implicated in cancer development, encodes both an antisense transcript to p53 and a WD40 repeat containing protein that promotes the formation of Cajal bodies (71). Although there are obvious implications of this antisense transcript regulating p53 activity, the connection to phase-separated Cajal bodies in this process remains to be investigated in depth.

In addition to functionally associating with phase-separated biomolecular condensates, wild-type and certain p53 mutants have been demonstrated to undergo liquid-like phase separation on their own, in a protein-autonomous manner (72, 73). *In vitro* experiments suggest that the

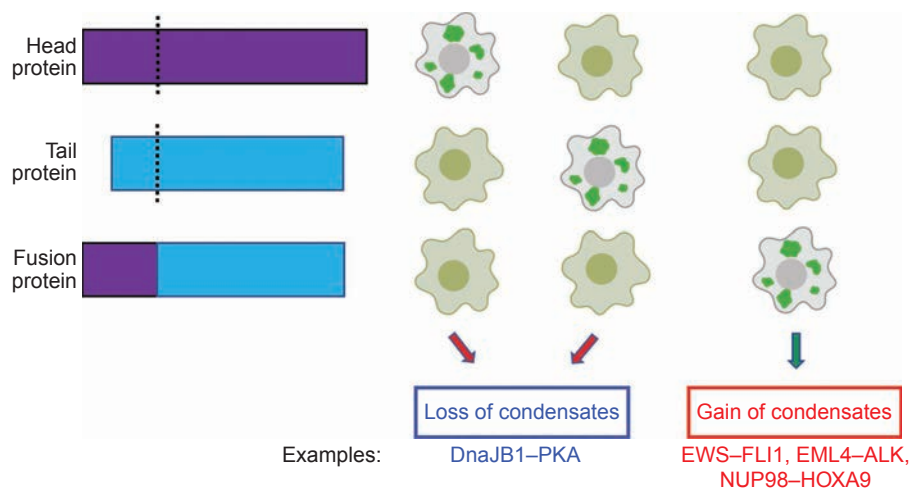


Figure 4. Alteration of phase separation behavior by fusion proteins. Fusion proteins (bottom) formed by aberrant joining of head protein (top) and tail protein (middle) at their breakpoints (dash lines). The corresponding phase separation phenotype (loss or gain of condensates) for each of these proteins is shown on the right. Each column of cell images represents one group for comparative purpose (cells harboring fusion proteins or their corresponding head or tail parent proteins). Specific examples for the gain or loss of condensates from the literature are also shown.

liquid-like properties of p53 can be modulated by interaction with ATP and nucleic acids, and a phosphomimetic mutant of the protein p53^{S392E} abrogates droplet formation (72). A few other mutations that commonly occur in patients have been found to render the protein amyloidogenic (74). The amyloid fibers formed by the mutant protein can sequester wild-type p53, leading to a dominant-negative effect and loss of p53 function (75). Peptide inhibitors of amyloid assembly block proliferation and restore p53 tumor suppression in ovarian carcinoma cells harboring these mutations (76).

EMERGENT ONCOGENIC PROPERTIES OF PROTEIN-NUCLEIC ACID CONDENSATES

Regulation of Gene Expression

Dysregulation of gene expression is a pervasive theme across all cancers. A string of recent articles have begun to delineate a potential role for biomolecular condensates in transcriptional gene control. Comparative analyses between healthy and cancerous cells have established that gene deserts around oncogenes such as *MYC* acquire many clusters of enhancer elements known as superenhancers (77). Recent findings have suggested that superenhancers and many of the integral components of the transcription machinery that interact with them including Mediator complex proteins, Brd4, as well as RNA Pol II can undergo phase separation and form MLOs (78–80), although there has been vigorous debate in the literature about whether some of these structures should instead be defined as “hubs” (81).

Phase separation has been speculated to explain many features of transcriptional control including local concentration of many of the factors necessary for gene activation. IDRs in these proteins are thought to be the driving force for the formation of these condensates. In this context, DNA translocations that are common in cancer are often associated with the acquisition of superenhancers at driver oncogenes (77). Furthermore, the presence of such IDRs on many TFs across

evolution and their frequent shuffling in cancer fusions collectively suggest an important role for IDR-driven phase separation in cancer (82).

Another series of investigations in the model eukaryote *Saccharomyces cerevisiae* have revealed how some intrinsically disordered TFs and RNA-binding proteins (RBP) form gel-like assemblies that can propagate in a prion-like fashion, acting as protein-based epigenetic elements that can transform gene expression over medium to long biological timescales. Even transient overexpression of these TFs and RBPs, as can occur for their human orthologs in cancer, can trigger the formation of self-templating protein assemblies that rewire the gene expression landscape for many cell divisions (24, 83, 84). Studies with the RBP Vts1/Smaug, whose human ortholog SAMD4A is overexpressed in chronic lymphocytic leukemia, revealed that its transient overexpression drove the formation of condensates that hyperactivate the protein's ability to degrade target RNAs (24). Consequently, cells alter how they couple metabolic information to decisions about proliferation and differentiation (85). Notably, SAMD4A retains the capacity to form condensates with heritable prion-like properties (24). Likewise, the Snt1 subunit of the Set3 histone deacetylase complex (ortholog of the NCOR–SMRT deacetylase complex that is dysregulated in many endocrine-related cancers; ref. 86) can self-assemble into self-templating gels, driving epigenetic inheritance of active chromatin and robust drug resistance (83). These findings have provoked efforts from us and others to understand how condensation of gene regulators can affect phenotype over medium to long biological timescales.

Alteration of Environmental Sensing and Response

Biochemical selectivity, a foundational requirement for the precise regulation of signaling pathways, can also be afforded by phase-separated condensates (87). This may have important implications for cancer, as the rewiring of metabolic equilibria and biomolecular interactions play a key role in

driving the initiation and progression of the disease. In line with this, phase separation of T-cell receptor signaling protein components fuels the formation of biochemically selective compartments (i.e., including kinases while excluding phosphatases) that play key roles across many cancers (88). Thus, alterations in phase separation can drive changes in biochemical selectivity relevant to cancer signaling.

Many signaling pathways rely on secondary messengers to relay and amplify responses across the cell. As one example of how phase separation can affect this behavior, DNA-induced condensation can promote the production of secondary messengers in cGAMP synthase (cGAS) signaling (89). In addition to secondary messenger production, phase separation plays important roles in target gene activation. For example, activation of the HIPPO pathway, which is critical for cell growth and proliferation, is mediated by phase separation of pathway-specific transcription coactivators (90, 91). Because cancer is often driven by changes in gene expression, it is tempting to speculate that altered phase separation of key TFs and their regulators might provide a common mechanism through which this can be achieved.

Although the aberrant activation of signaling pathways is often observed in cancer, it is common that cancer cells lose the capacity to turn off signaling pathways via negative feedback systems. Here, too, phase separation can play an important regulatory role. For example, the WNT/beta-catenin pathway is negatively regulated through the beta-catenin destruction complex. Recent work has shown that the destruction complex, which promotes the proteolysis of beta-catenin to keep its levels low in the absence of WNT, forms via phase separation (92). Furthermore, cytoplasmic DACT1 condensates that are induced by TGF β repress WNT signaling. These DACT1 condensates are maintained *in vivo* and are associated with the bone metastasis of breast and prostate cancers (93). Future work will be needed to fully comprehend how phase separation functions in the negative regulation of signaling pathways in both normal biology and disease states.

Alteration of Cellular Metabolism

One of the classic examples of alteration of metabolic flux in cancer cells was described by Otto Warburg and colleagues. The harnessing of glycolytic pathways over oxidative phosphorylation by cancer cells has been extensively described (94). Recent experiments in model organisms such as yeast, *Caenorhabditis elegans*, and mammalian cells have demonstrated that many of the enzymes involved in glycolysis assemble into membraneless cytoplasmic organelles, called glycolytic bodies (G-bodies), under hypoxic stress (95, 96). In yeast, G-bodies form when oxidative phosphorylation is inhibited, and this improves their fitness under hypoxia (97). G-bodies arise under hypoxic conditions in HepG2 cells, a hepatocellular carcinoma cell line (96). These subcellular organelles accelerate glucose consumption and enhance glycolytic flux by increasing the local concentration of the glycolysis enzymes. The enzyme phosphofructokinase, Pfk2 in yeast and PFKL in humans, drives the formation of G-bodies (96). The N-terminal disordered region of this protein is important for their formation (97). Recently, unbiased searches for RNA-binding activity have demonstrated that many metabolic enzymes moonlight as RBPs. In fact, many

RNAs are strongly associated with G-bodies as well and contribute to their structural integrity; targeting RNase to G-bodies disrupts them (97). From a structural standpoint, G-bodies do not resemble liquids. Rather, they are a gel-like species emphasizing the diverse palette of material properties that can contribute to function in the complex cellular environment. This example may be just the tip of the iceberg. Phase separation of proteins (98, 99) associated with oxidative pathways can sense cellular redox states, and some have been demonstrated to alter the flux of key metabolic regulators such as the TORC1 complex (100, 101).

Alternative Splicing

The modular addition of IDRs to protein sequences via alternative splicing offers an intriguing path to diversify protein function. Given the ability of IDRs to endow proteins with the ability to phase separate, this leads to a natural diversification of the material properties of condensates formed by the protein. This is perhaps best exemplified by heterogeneous nuclear ribonucleoprotein (hnRNP) A and D family members (102). These proteins play diverse roles in mammalian RNA metabolism, and hnRNPD in particular is upregulated in a variety of cancers including prostate, CML, hepatocellular carcinoma, and colon. hnRNPD has three isoforms: DL1, DL2, and DL3. These isoforms differ in the way the IDRs are arranged with consequent changes in their cellular behavior: DL3 does not harbor any IDRs and is completely soluble, DL2 harbors the C-terminal IDR, and DL1 harbors IDRs at both termini (103). The C-terminal IDR is tyrosine rich and confers amyloidogenic properties to the DL2 isoform. Strikingly, the presence of the arginine-rich N-terminal IDR in the DL1 isoform abrogates fibril formation and enhances the formation of large amorphous condensates. It has been suggested that the structural differences among the isoforms arise from the differences between π - π interactions among aromatic residues in the dominant DL2 isoform and π -cation interactions in DL1. The structural differences also correlate with differential localization within the cell. Certain point mutations in the C-terminal IDR increase the aggregation propensity of both DL1 and DL2 isoforms and are associated with an autosomal dominant form of muscular dystrophy. Interestingly, SRSF, which is a key upstream regulator of alternative splicing of hnRNPD, is also dysregulated in multiple cancers (103, 104).

TARGETING PHASE SEPARATION AS A THERAPEUTIC STRATEGY

Phase separation provides several new avenues for potential therapeutic intervention. Despite the widespread occurrence of higher order protein aggregates in many diseases of aging, therapeutic targeting of the associative properties of biomolecules has been notoriously difficult, with the exception of groundbreaking approaches to develop chemical chaperones/stabilizers for specific misfolded proteins (105, 106). Therapeutic targeting of phase separation could significantly expand the palette of drug development strategies for TFs and other proteins involved in gene regulation that have been notoriously difficult to drug. Yet doing so will require a deepened consideration of physicochemical

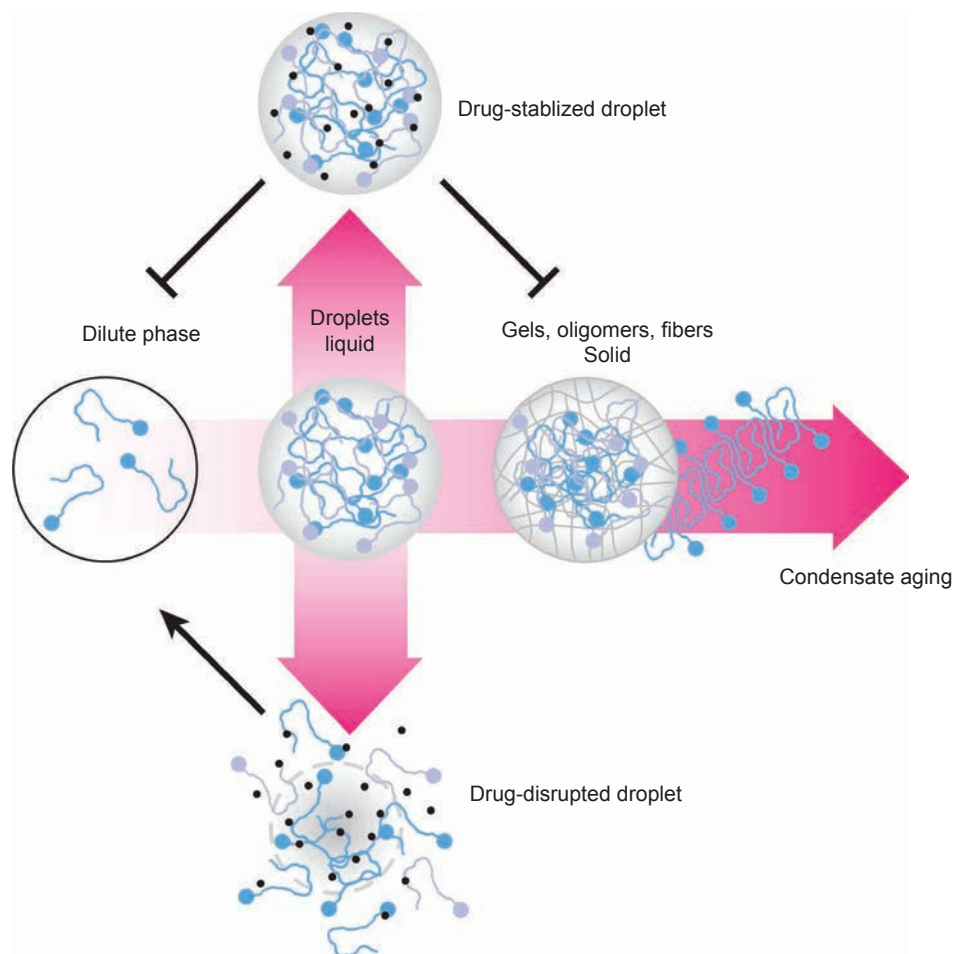


Figure 5. Therapies targeting condensates. The prevalence of cancer genes that form condensates presents an opportunity to develop new classes of therapeutics that modulate the phase separation of their targets. New drugs might be made to disrupt or prevent specific molecules from forming liquid droplets (bottom arrow), while others might stabilize the liquid droplet, blocking its dissolution or maturation into solid-phase MLOs.

principles that go beyond the lock and key approaches that are often deployed to target ordered proteins. One strategy for direct targeting of phase-separated entities relies on the differential partitioning of a small molecule or a biologic between the dilute and dense phases (Fig. 5). The disruption of many LLPS entities by 1,6-hexanediol first shed light on this possibility. From a therapeutic standpoint, recent work has demonstrated the feasibility of several compounds to disrupt the RNA-dependent recruitment of several RBPs—namely, TDP-43, FUS, and hnRNP A2B1—into stress granules (107). Furthermore, drugs such as cisplatin

can be enriched within condensates (108). A recent chemical screen has also identified small molecules that can modulate the phase separation properties of p53 mutants (109), foreshadowing a new genre of therapeutics for future cancer treatment. However, many challenges remain to overcome, including the development of molecules that can selectively target different phase-separated condensates in the same cellular environment. Furthermore, the differential sequestration of drugs within condensates raises the possibility that resistance may occur through mutations that alter phase separation.

BOX: FUTURE OUTLOOK

- The functional role of mutations in altering phase separation needs to be systematically evaluated, especially in cancer.
- Cancer gene fusions could have a profound impact on phase separation and indeed use this property to drive cancer initiation and progression. Systematic efforts are needed to understand this mechanism.
- Cancer mutations could drive either gain or loss of condensates, contributing to disease.
- Many missense mutations could affect phase separation, through rewiring relevant signaling pathways, and

known disease mutations may function by altering the presence or composition of condensates.

- The extent to which protein folding and stability contribute to phase separation activities is mostly unknown.
- Most importantly, many disease alleles, especially cancer mutants, could be dominant, sequestering wild-type counterparts to form condensates. This remarkable behavior needs to be systematically examined, which could reveal fundamental principles of genetic inheritance.

Authors' Disclosures

W.J. Zheng reports grants from the NIH and the Cancer Prevention & Research Institute of Texas during the conduct of the study. G.B. Mills is on scientific advisory boards for Amphista, AstraZeneca, Chrysalis Biotechnology, GSK, ImmunoMET, Ionis, Lilly, PDX Pharmaceuticals, Signalchem Lifesciences, Symphogen, Tarveda, Turbine, and Zentalis Pharmaceuticals, and reports other support from Catena, Genentech, NanoString Center of Excellence, and Ionis (provision of tool compounds) during the conduct of the study, as well as a patent for an HRD assay to Myriad Genetics and a patent for DSP to NanoString. D.F. Jarosz reports grants from the NIH, the David and Lucile Packard Foundation, and the Vallee Foundation during the conduct of the study, as well as personal fees from Transition Bio outside the submitted work. No disclosures were reported by the other authors.

Acknowledgments

This work was supported by Komen Foundation grant CCR19609287 (to S.S. Yi); National Institute of General Medical Sciences grants R35GM133658 (to S.S. Yi) and R35GM137836 (to N. Sahni); NIH grant DP2-GM119140 (to D.F. Jarosz); Department of Defense fund W81XWH-18-PRCRP-CDA CA181455 (to N. Sahni); and NCI grants P50CA217685 and U01CA217842 (to G.B. Mills). N. Sahni is a CPRIT Scholar in Cancer Research with funding from the Cancer Prevention & Research Institute of Texas (CPRIT) New Investigator Grant RR160021 and Research Grant RP220292. S.S. Yi was also supported by the Scialog program sponsored jointly by the Research Corporation for Science Advancement and the Gordon and Betty Moore Foundation (award number 28418) as well as a pilot project program sponsored jointly by the Oden Institute, MD Anderson, and Texas Advanced Computing Center. G.B. Mills also received a kind gift from the Miriam and Sheldon Adelson Medical Research Foundation and the Breast Cancer Research Foundation. D.J. McGrail was supported by NCI grant K99CA240689. A.K. Chakravarty was supported as a Howard Hughes Medical Institute fellow of the Damon Runyon Cancer Research Foundation (DRG2221-15) and by an NIH Pathway to Independence Award (1K99GM128180-01). T.M. Lozano was supported by a National Science Foundation (NSF) graduate research fellowship. D.F. Jarosz was also supported as a Searle Scholar, as a Kimmel Scholar, as a Vallee Scholar, by a Science and Engineering Fellowship from the David and Lucile Packard Foundation, and by an NSF CAREER Award (1453762) from the NSF. W.J. Zheng and D.J.H. Shih were partly supported by the NIH through grant 1UL1TR003167 and CPRIT through grant RP170668. We also acknowledge support from BioRender in figure generation.

Received December 7, 2021; revised April 6, 2022; accepted June 8, 2022; published first July 19, 2022.

REFERENCES

- Alberti S, Dormann D. Liquid-liquid phase separation in disease. *Annu Rev Genet* 2019;53:171-94.
- Shin Y, Brangwynne CP. Liquid phase condensation in cell physiology and disease. *Science* 2017;357:eaaf4382.
- Banani SF, Lee HO, Hyman AA, Rosen MK. Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol* 2017;18:285-98.
- Zbinden A, Perez-Berlanga M, De Rossi P, Polymenidou M. Phase separation and neurodegenerative diseases: a disturbance in the force. *Dev Cell* 2020;55:45-68.
- Hyman AA, Weber CA, Julicher F. Liquid-liquid phase separation in biology. *Annu Rev Cell Dev Biol* 2014;30:39-58.
- Agarwal A, Mukhopadhyay S. Prion protein biology through the lens of liquid-liquid phase separation. *J Mol Biol* 2022;434:167368.
- Bojja A, Klein IA, Young RA. Biomolecular condensates and cancer. *Cancer Cell* 2021;39:174-92.
- Jiang S, Fagman JB, Chen C, Alberti S, Liu B. Protein phase separation and its role in tumorigenesis. *Elife* 2020;9:e60264.
- Tsang B, Pritsanac I, Scherer SW, Moses AM, Forman-Kay JD. Phase separation as a missing mechanism for interpretation of disease mutations. *Cell* 2020;183:1742-56.
- Cai D, Liu Z, Lippincott-Schwartz J. Biomolecular condensates and their links to cancer progression. *Trends Biochem Sci* 2021;46:535-49.
- Peng PH, Hsu KW, Wu KJ. Liquid-liquid phase separation (LLPS) in cellular physiology and tumor biology. *Am J Cancer Res* 2021;11:3766-76.
- Suzuki HI, Onimaru K. Biomolecular condensates in cancer biology. *Cancer Sci* 2021;113:382-91.
- Taniue K, Akimitsu N. Aberrant phase separation and cancer. *FEBS J* 2022;289:17-39.
- Brangwynne CP, Eckmann CR, Courson DS, Rybarska A, Hoeg C, Gharakhani J, et al. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* 2009;324:1729-32.
- Molliex A, Temirov J, Lee J, Coughlin M, Kanagaraj AP, Kim HJ, et al. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* 2015;163:123-33.
- Riback JA, Zhu L, Ferrolino MC, Tolbert M, Mitrea DM, Sanders DW, et al. Composition-dependent thermodynamics of intracellular phase separation. *Nature* 2020;581:209-14.
- Riback JA, Katanski CD, Kear-Scott JL, Pilipenko EV, Rojek AE, Sosnick TR, et al. Stress-triggered phase separation is an adaptive, evolutionarily tuned response. *Cell* 2017;168:1028-40.
- Patel A, Lee HO, Jawerth L, Maharana S, Jahnel M, Hein MY, et al. A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell* 2015;162:1066-77.
- Brangwynne CP, Mitchison TJ, Hyman AA. Active liquid-like behavior of nucleoli determines their size and shape in *Xenopus laevis* oocytes. *Proc Natl Acad Sci U S A* 2011;108:4334-9.
- Wippich F, Bodenmiller B, Trajkovska MG, Wanka S, Aebersold R, Pelkmans L. Dual specificity kinase DYRK3 couples stress granule condensation/dissolution to mTORC1 signaling. *Cell* 2013;152:791-805.
- Saad S, Cereghetti G, Feng Y, Picotti P, Peter M, Dechant R. Reversible protein aggregation is a protective mechanism to ensure cell cycle restart after stress. *Nat Cell Biol* 2017;19:1202-13.
- Wallace EW, Kear-Scott JL, Pilipenko EV, Schwartz MH, Laskowski PR, Rojek AE, et al. Reversible, specific, active aggregates of endogenous proteins assemble upon heat stress. *Cell* 2015;162:1286-98.
- van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. Classification of intrinsically disordered regions and proteins. *Chem Rev* 2014;114:6589-631.
- Chakravarty AK, Smejkal T, Itakura AK, Garcia DM, Jarosz DF. A non-amyloid prion particle that activates a heritable gene expression program. *Mol Cell* 2020;77:251-65.
- Uversky VN. A decade and a half of protein intrinsic disorder: biology still waits for physics. *Protein Sci* 2013;22:693-724.
- Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res* 2019;47:D941-D7.
- Li Q, Peng X, Li Y, Tang W, Zhu J, Huang J, et al. LLPDB: a database of proteins undergoing liquid-liquid phase separation in vitro. *Nucleic Acids Res* 2020;48:D320-D7.
- You K, Huang Q, Yu C, Shen B, Sevilla C, Shi M, et al. PhaSepDB: a database of liquid-liquid phase separation related proteins. *Nucleic Acids Res* 2020;48:D354-D9.
- Meszaros B, Erdos G, Szabo B, Schad E, Tantos A, Abukhairan R, et al. PhaSePro: the database of proteins driving liquid-liquid phase separation. *Nucleic Acids Res* 2020;48:D360-D7.
- Ning W, Guo Y, Lin S, Mei B, Wu Y, Jiang P, et al. DrLLPS: a data resource of liquid-liquid phase separation in eukaryotes. *Nucleic Acids Res* 2020;48:D288-D95.

31. The Gene Ontology C. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res* 2019;47:D330–D8.
32. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Disease networks: uncovering disease-disease relationships through the incomplete interactome. *Science* 2015;347:1257601.
33. Balch WE, Morimoto RI, Dillin A, Kelly JW. Adapting proteostasis for disease intervention. *Science* 2008;319:916–9.
34. Oromendia AB, Amon A. Aneuploidy: implications for protein homeostasis and disease. *Dis Model Mech* 2014;7:15–20.
35. Joshi S, Wang T, Araujo TLS, Sharma S, Brodsky JL, Chiosis G. Adapting to stress: chaperome networks in cancer. *Nat Rev Cancer* 2018;18:562–75.
36. Fujioka Y, Alam JM, Noshiro D, Mouri K, Ando T, Okada Y, et al. Phase separation organizes the site of autophagosome formation. *Nature* 2020;578:301–5.
37. Yasuda S, Tsuchiya H, Kaiho A, Guo Q, Ikeuchi K, Endo A, et al. Stress- and ubiquitylation-dependent phase separation of the proteasome. *Nature* 2020;578:296–300.
38. Bouchard JJ, Otero JH, Scott DC, Szulc E, Martin EW, Sabri N, et al. Cancer mutations of the tumor suppressor SPOP disrupt the formation of active, phase-separated compartments. *Mol Cell* 2018;72:19–36.
39. Shi Q, Zhu Y, Ma J, Chang K, Ding D, Bai Y, et al. Prostate cancer-associated SPOP mutations enhance cancer cell survival and docetaxel resistance by upregulating Caprin1-dependent stress granule assembly. *Mol Cancer* 2019;18:170.
40. Latysheva NS, Oates ME, Maddox L, Flock T, Gough J, Buljan M, et al. Molecular principles of gene fusion mediated rewiring of protein interaction networks in cancer. *Mol Cell* 2016;63:579–92.
41. Terlecki-Zaniewicz S, Humer T, Eder T, Schmoeller J, Heyes E, Manhart G, et al. Biomolecular condensation of NUP98 fusion proteins drives leukemogenic gene expression. *Nat Struct Mol Biol* 2021;28:190–201.
42. Ahn JH, Davis ES, Daugird TA, Zhao S, Quiroga IY, Uryu H, et al. Phase separation drives aberrant chromatin looping and cancer development. *Nature* 2021;595:591–5.
43. Chandra B, Michmerhuizen NL, Shirnekhi HK, Tripathi S, Pioso BJ, Baggett DW, et al. Phase separation mediates NUP98 fusion oncoprotein leukemic transformation. *Cancer Discov* 2021;12:1152–69.
44. Boulay G, Sandoval GJ, Riggi N, Iyer S, Buisson R, Naigles B, et al. Cancer-specific retargeting of BAF complexes by a prion-like domain. *Cell* 2017;171:163–78.
45. Chen M, Xu ES, Leisenring NH, Cardona DM, Luo L, Ma Y, et al. The fusion oncogene FUS-CHOP drives sarcomagenesis of high-grade spindle cell sarcomas in mice. *Sarcoma* 2019;2019:1340261.
46. Sampson J, Richards MW, Choi J, Fry AM, Bayliss R. Phase-separated foci of EML4-ALK facilitate signalling and depend upon an active kinase conformation. *EMBO Rep* 2021;22:e53693.
47. Tulpule A, Guan J, Neel DS, Allegakoen HR, Lin YP, Brown D, et al. Kinase-mediated RAS signaling via membraneless cytoplasmic protein granules. *Cell* 2021;184:2649–64.
48. Zhang JZ, Lu TW, Stoleran LM, Tenner B, Yang JR, Zhang JF, et al. Phase separation of a PKA regulatory subunit controls cAMP compartmentation and oncogenic signaling. *Cell* 2020;182:1531–44.
49. Meyer K, Kirchner M, Uyar B, Cheng JY, Russo G, Hernandez-Miranda LR, et al. Mutations in disordered regions can cause disease by creating dileucine motifs. *Cell* 2018;175:239–53.
50. Nilsson J, Grahn M, Wright AP. Proteome-wide evidence for enhanced positive Darwinian selection within intrinsically disordered regions in proteins. *Genome Biol* 2011;12:R65.
51. Hegyi H, Buday L, Tompa P. Intrinsic structural disorder confers cellular viability on oncogenic fusion proteins. *PLoS Comput Biol* 2009;5:e1000552.
52. Malanga M, Althaus FR. The role of poly(ADP-ribose) in the DNA damage signaling network. *Biochem Cell Biol* 2005;83:354–64.
53. Altmeyer M, Neelsen KJ, Teloni F, Pozdnyakova I, Pellegrino S, Grofte M, et al. Liquid demixing of intrinsically disordered proteins is seeded by poly(ADP-ribose). *Nat Commun* 2015;6:8088.
54. Teloni F, Altmeyer M. Readers of poly(ADP-ribose): designed to be fit for purpose. *Nucleic Acids Res* 2016;44:993–1006.
55. Kilic S, Lezaja A, Gatti M, Bianco E, Michelena J, Imhof R, et al. Phase separation of 53BP1 determines liquid-like behavior of DNA repair compartments. *EMBO J* 2019;38:e101379.
56. Pessina F, Giavazzi F, Yin Y, Gioia U, Vitelli V, Galbiati A, et al. Functional transcription promoters at DNA double-strand breaks mediate RNA-driven phase separation of damage-response factors. *Nat Cell Biol* 2019;21:1286–99.
57. Fijen C, Rothenberg E. The evolving complexity of DNA damage foci: RNA, condensates and chromatin in DNA double-strand break repair. *DNA Repair* 2021;105:103170.
58. Nozawa RS, Yamamoto T, Takahashi M, Tachiwana H, Maruyama R, Hirota T, et al. Nuclear microenvironment in cancer: control through liquid-liquid phase separation. *Cancer Sci* 2020;111:3155–63.
59. Sun X, Kaufman PD. Ki-67: more than a proliferation marker. *Chromosoma* 2018;127:175–86.
60. Whitfield ML, George LK, Grant GD, Perou CM. Common markers of proliferation. *Nat Rev Cancer* 2006;6:99–106.
61. Dowsett M, Nielsen TO, A'Hern R, Bartlett J, Coombes RC, Cuzick J, et al. Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *J Natl Cancer Inst* 2011;103:1656–64.
62. Richards-Taylor S, Ewings SM, Jaynes E, Tilley C, Ellis SG, Armstrong T, et al. The assessment of Ki-67 as a prognostic marker in neuroendocrine tumours: a systematic review and meta-analysis. *J Clin Pathol* 2016;69:612–8.
63. Cuylen S, Blaukopf C, Politi AZ, Muller-Reichert T, Neumann B, Poser I, et al. Ki-67 acts as a biological surfactant to disperse mitotic chromosomes. *Nature* 2016;535:308–12.
64. Verheijen R, Kuijpers HJ, van Driel R, Beck JL, van Dierendonck JH, Brakenhoff GJ, et al. Ki-67 detects a nuclear matrix-associated proliferation-related antigen. II. Localization in mitotic cells and association with chromosomes. *J Cell Sci* 1989;92:531–40.
65. Schluter C, Duchrow M, Wohlenberg C, Becker MH, Key G, Flad HD, et al. The cell proliferation-associated antigen of antibody Ki-67: a very large, ubiquitous nuclear protein with numerous repeated elements, representing a new kind of cell cycle-maintaining proteins. *J Cell Biol* 1993;123:513–22.
66. Banani SF, Rice AM, Peeples WB, Lin Y, Jain S, Parker R, et al. Compositional control of phase-separated cellular bodies. *Cell* 2016;166:651–63.
67. Guo A, Salomoni P, Luo J, Shih A, Zhong S, Gu W, et al. The function of PML in p53-dependent apoptosis. *Nat Cell Biol* 2000;2:730–6.
68. de Stanchina E, Querido E, Narita M, Davuluri RV, Pandolfi PP, Ferbeyre G, et al. PML is a direct p53 target that modulates p53 effector functions. *Mol Cell* 2004;13:523–35.
69. Bao-Lei T, Zhu-Zhong M, Yi S, Jun-Jie Q, Yan D, Hua L, et al. Knocking down PML impairs p53 signaling transduction pathway and suppresses irradiation induced apoptosis in breast carcinoma cell MCF-7. *J Cell Biochem* 2006;97:561–71.
70. Bernardi R, Scaglioni PP, Bergmann S, Horn HF, Vousden KH, Pandolfi PP. PML regulates p53 stability by sequestering Mdm2 to the nucleolus. *Nat Cell Biol* 2004;6:665–72.
71. Henriksson S, Farnebo M. On the road with WRAP53beta: guardian of Cajal bodies and genome integrity. *Front Genet* 2015;6:91.
72. Kamagata K, Kanbayashi S, Honda M, Itoh Y, Takahashi H, Kameda T, et al. Liquid-like droplet formation by tumor suppressor p53 induced by multivalent electrostatic interactions between two disordered domains. *Sci Rep* 2020;10:580.
73. Safari MS, Wang Z, Taylor K, Kolomeisky AB, Conrad JC, Vekilov PG. Anomalous dense liquid condensates host the nucleation of tumor suppressor p53 fibrils. *iScience* 2019;12:342–55.
74. Costa DC, de Oliveira GA, Cino EA, Soares IN, Rangel LP, Silva JL. Aggregation and prion-like properties of misfolded tumor suppressors: is cancer a prion disease? *Cold Spring Harb Perspect Biol* 2016;8:a023614.
75. Silva JL, De Moura Gallo CV, Costa DC, Rangel LP. Prion-like aggregation of mutant p53 in cancer. *Trends Biochem Sci* 2014;39:260–7.
76. Soragni A, Janzen DM, Johnson LM, Lindgren AG, Thai-Quynh Nguyen A, Tiourin E, et al. A designed inhibitor of p53 aggregation

- rescues p53 tumor suppression in ovarian carcinomas. *Cancer Cell* 2016;29:90–103.
77. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, et al. Super-enhancers in the control of cell identity and disease. *Cell* 2013;155:934–47.
 78. Cho WK, Spille JH, Hecht M, Lee C, Li C, Grube V, et al. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* 2018;361:412–5.
 79. Sabari BR, Dall'Agnese A, Boija A, Klein IA, Coffey EL, Shrinivas K, et al. Coactivator condensation at super-enhancers links phase separation and gene control. *Science* 2018;361:eaar3958.
 80. Shrinivas K, Sabari BR, Coffey EL, Klein IA, Boija A, Zamudio AV, et al. Enhancer features that drive formation of transcriptional condensates. *Mol Cell* 2019;75:549–61.
 81. Chong S, Dugast-Darzacq C, Liu Z, Dong P, Dailey GM, Cattoglio C, et al. Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science* 2018;361:eaar2555.
 82. Song K, Minami JK, Huang A, Dehkordi SR, Lomeli SH, Luebeck J, et al. Plasticity of extrachromosomal and intrachromosomal BRAF amplifications in overcoming targeted therapy dosage challenges. *Cancer Discov* 2021;12:1046–69.
 83. Harvey ZH, Chakravarty AK, Futia RA, Jarosz DF. A prion epigenetic switch establishes an active chromatin state. *Cell* 2020;180:928–40.
 84. Kechavarzi B, Janga SC. Dissecting the expression landscape of RNA-binding proteins in human cancers. *Genome Biol* 2014;15:R14.
 85. Itakura AK, Chakravarty AK, Jakobson CM, Jarosz DF. Widespread prion-based control of growth and differentiation strategies in *Saccharomyces cerevisiae*. *Mol Cell* 2020;77:266–78.
 86. Wong MM, Guo C, Zhang J. Nuclear receptor corepressor complexes in cancer: mechanism, function and regulation. *Am J Clin Exp Urol* 2014;2:169–87.
 87. Laflamme G, Mekhail K. Biomolecular condensates as arbiters of biochemical reactions inside the nucleus. *Commun Biol* 2020;3:773.
 88. Su X, Ditlev JA, Hui E, Xing W, Banjade S, Okrut J, et al. Phase separation of signaling molecules promotes T cell receptor signal transduction. *Science* 2016;352:595–9.
 89. Du M, Chen ZJ. DNA-induced liquid phase condensation of cGAS activates innate immune signaling. *Science* 2018;361:704–9.
 90. Lu Y, Wu T, Gutman O, Lu H, Zhou Q, Henis YI, et al. Phase separation of TAZ compartmentalizes the transcription machinery to promote gene expression. *Nat Cell Biol* 2020;22:453–64.
 91. Cai D, Feliciano D, Dong P, Flores E, Grubele M, Porat-Shliom N, et al. Phase separation of YAP reorganizes genome topology for long-term YAP target gene expression. *Nat Cell Biol* 2019;21:1578–89.
 92. Schaefer KN, Peifer M. Wnt/beta-catenin signaling regulation and a role for biomolecular condensates. *Dev Cell* 2019;48:429–44.
 93. Esposito M, Fang C, Cook KC, Park N, Wei Y, Spadazzi C, et al. TGF-beta-induced DACT1 biomolecular condensates repress Wnt signaling to promote bone metastasis. *Nat Cell Biol* 2021;23:257–67.
 94. Vander Heiden MG, DeBerardinis RJ. Understanding the intersections between metabolism and cancer biology. *Cell* 2017;168:657–69.
 95. Jang S, Nelson JC, Bend EG, Rodriguez-Laureano L, Tueros FG, Cartagenaova L, et al. Glycolytic enzymes localize to synapses under energy stress to support synaptic function. *Neuron* 2016;90:278–91.
 96. Jin M, Fuller GG, Han T, Yao Y, Alessi AF, Freeberg MA, et al. Glycolytic enzymes coalesce in G bodies under hypoxic stress. *Cell Rep* 2017;20:895–908.
 97. Fuller GG, Han T, Freeberg MA, Moresco JJ, Ghanbari Niaki A, Roach NP, et al. RNA promotes phase separation of glycolysis enzymes into yeast G bodies in hypoxia. *Elife* 2020;9:e48480.
 98. Miura N, Shinohara M, Tatsukami Y, Sato Y, Morisaka H, Kuroda K, et al. Spatial reorganization of *Saccharomyces cerevisiae* enolase to alter carbon metabolism under hypoxia. *Eukaryot Cell* 2013;12:1106–19.
 99. Prouteau M, Loewith R. Regulation of cellular metabolism through phase separation of enzymes. *Biomolecules* 2018;8:160.
 100. Kato M, Yang YS, Sutter BM, Wang Y, McKnight SL, Tu BP. Redox state controls phase separation of the yeast ataxin-2 protein via reversible oxidation of its methionine-rich low-complexity domain. *Cell* 2019;177:711–21.
 101. Yang YS, Kato M, Wu X, Litsios A, Sutter BM, Wang Y, et al. Yeast ataxin-2 forms an intracellular condensate required for the inhibition of TORC1 signaling during respiratory growth. *Cell* 2019;177:697–710.
 102. Gueroussov S, Weatheritt RJ, O'Hanlon D, Lin ZY, Narula A, Gingras AC, et al. Regulatory expansion in mammals of multivalent hnRNP assemblies that globally control alternative splicing. *Cell* 2017;170:324–39.
 103. Batlle C, Yang P, Coughlin M, Messing J, Pesarrodonna M, Szulc E, et al. hnRNPDL phase separation is regulated by alternative splicing and disease-causing mutations accelerate its aggregation. *Cell Rep* 2020;30:1117–28.
 104. Feng H, Bao S, Rahman MA, Weyn-Vanhenhenryck SM, Khan A, Wong J, et al. Modeling RNA-binding protein specificity in vivo by precisely registering protein-RNA crosslink sites. *Mol Cell* 2019;74:1189–204.
 105. Maurer MS, Schwartz JH, Gundapaneni B, Elliott PM, Merlini G, Waddington-Cruz M, et al. Tafamidis treatment for patients with transthyretin amyloid cardiomyopathy. *N Engl J Med* 2018;379:1007–16.
 106. Carlile GW, Yang Q, Matthes E, Liao J, Radinovic S, Miyamoto C, et al. A novel triple combination of pharmacological chaperones improves F508del-CFTR correction. *Sci Rep* 2018;8:11404.
 107. Fang MY, Markmiller S, Vu AQ, Javaherian A, Dowdle WE, Jolivet P, et al. Small-molecule modulation of TDP-43 recruitment to stress granules prevents persistent TDP-43 accumulation in ALS/FTD. *Neuron* 2019;103:802–19.
 108. Klein IA, Boija A, Afeyan LK, Hawken SW, Fan M, Dall'Agnese A, et al. Partitioning of cancer therapeutics in nuclear condensates. *Science* 2020;368:1386–92.
 109. Lemos C, Schulze L, Weiske J, Meyer H, Brauer N, Barak N, et al. Identification of small molecules that modulate mutant p53 condensation. *iScience* 2020;23:101517.



Landscape of MicroRNA Regulatory Network Architecture and Functional Rerouting in Cancer

Xu Hua¹, Yongsheng Li², Sairahul R. Pentaparthi², Daniel J. McGrail¹, Raymond Zou³, Li Guo¹, Aditya Shrawat⁴, Kara M. Cirillo³, Qing Li³, Akshay Bhat², Min Xu⁵, Dan Qi⁵, Ashok Singh⁶, Francis McGrath⁶, Steven Andrews⁶, Kyaw Lwin Aung², Jishnu Das⁷, Yunyun Zhou⁸, Alessia Lodi^{9,10}, Gordon B. Mills^{11,12}, S. Gail Eckhardt^{2,13}, Marc L. Mendillo¹⁴, Stefano Tiziani^{2,9,10,13}, Erxi Wu^{2,5,15,16}, Jason H. Huang^{5,15}, Nidhi Sahni^{3,17,18}, and S. Stephen Yi^{2,13,19,20}

ABSTRACT

Somatic mutations are a major source of cancer development, and many driver mutations have been identified in protein coding regions. However, the function of mutations located in miRNA and their target binding sites throughout the human genome remains largely unknown. Here, we built detailed cancer-specific miRNA regulatory networks across 30 cancer types to systematically analyze the effect of mutations in miRNAs and their target sites in 3' untranslated region (3' UTR), coding sequence (CDS), and 5' UTR regions. A total of 3,518,261 mutations from 9,819 samples were mapped to miRNA–gene interactions (mGI). Mutations in miRNAs showed a mutually exclusive pattern with mutations in their target genes in almost all cancer types. A linear regression method identified 148 candidate driver mutations that can significantly perturb miRNA regulatory networks. Driver mutations in 3'UTRs played their roles by altering RNA binding

energy and the expression of target genes. Finally, mutated driver gene targets in 3' UTRs were significantly downregulated in cancer and functioned as tumor suppressors during cancer progression, suggesting potential miRNA candidates with significant clinical implications. A user-friendly, open-access web portal (mGI-map) was developed to facilitate further use of this data resource. Together, these results will facilitate novel non-coding biomarker identification and therapeutic drug design targeting the miRNA regulatory networks.

Significance: A detailed miRNA–gene interaction map reveals extensive miRNA-mediated gene regulatory networks with mutation-induced perturbations across multiple cancers, serving as a resource for noncoding biomarker discovery and drug development.

Introduction

Genetic changes are a primary source of oncogenesis. Somatic mutations accumulate in cells during one's lifetime. Among them, some may contribute to a cell's selective advantage by altering activity in pathways associated with tumor formation and therapeutic resistance, and are therefore often referred to as “driver mutations.” Other variants may not have either phenotypic consequences or biological

effects that are selectively advantageous to cancer cells, and are thus defined as “passenger mutations.” Many studies have paid attention to missense mutations in protein coding regions that can drive cancer development. A number of driver mutations have been identified in previous research efforts, such as BRAF (V600E), IDH1 (R132H), PIK3CA (E545K), EGFR (L858R), and KRAS (G12D; ref. 1). However, in addition to missense mutations that affect the protein coding components of human genome, miRNAs and their target binding

¹Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas. ²Livestrong Cancer Institutes, Department of Oncology, Dell Medical School, The University of Texas at Austin, Austin, Texas. ³Department of Epigenetics and Molecular Carcinogenesis, The University of Texas MD Anderson Cancer Center, Houston, Texas. ⁴College of Natural Sciences, The University of Texas at Austin, Austin, Texas. ⁵Neuroscience Institute and Department of Neurosurgery, Baylor Scott & White Health, Temple, Texas. ⁶Dell Medical School, The University of Texas at Austin, Austin, Texas. ⁷Center for Systems Immunology, Department of Immunology, and Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania. ⁸Children's Hospital of Philadelphia, Philadelphia, Pennsylvania. ⁹Department of Nutritional Sciences, College of Natural Sciences, The University of Texas at Austin, Austin, Texas. ¹⁰Department of Pediatrics, Dell Medical School, The University of Texas at Austin, Austin, Texas. ¹¹Department of Cell, Developmental and Cancer Biology, School of Medicine, Oregon Health & Science University, Portland, Oregon. ¹²Precision Oncology, Knight Cancer Institute, Portland, Oregon. ¹³Interdisciplinary Life Sciences Graduate Programs (ILSGP), The University of Texas at Austin, Austin, Texas. ¹⁴Department of Biochemistry and Molecular Genetics, and Robert H. Lurie Comprehensive Cancer Center, Northwestern University Feinberg School of Medicine, Chicago, Illinois. ¹⁵Department of Surgery, Texas A & M University Health Science Center, College of Medicine, Temple, Texas. ¹⁶Department of Pharmaceutical Sciences, Texas A & M University Health Science Center, College of Pharmacy, College Station, Texas. ¹⁷Department of Bioinformatics and

Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas. ¹⁸Quantitative and Computational Biosciences Program, Baylor College of Medicine, Houston, Texas. ¹⁹Oden Institute for Computational Engineering and Sciences (ICES), The University of Texas at Austin, Austin, Texas. ²⁰Department of Biomedical Engineering, Cockrell School of Engineering, The University of Texas at Austin, Austin, Texas.

X. Hua, Y. Li, S.R. Pentaparthi, D.J. McGrail, R. Zou, and L. Guo contributed equally to this article.

Corresponding Authors: S. Stephen Yi, The University of Texas at Austin, 1601B Trinity Street, Stop Z1100, Austin, TX, 78712. E-mail: stephen.yi@austin.utexas.edu; Nidhi Sahni, The University of Texas MD Anderson Cancer Center, 1881 East Road, Houston, TX 77054. E-mail: nsahni@mdanderson.org; Jason H. Huang, Baylor Research Institute, Temple, TX 76502. E-mail: Jason.Huang@BSWHealth.org; and Erxi Wu, Baylor Research Institute, Temple, TX 76502. E-mail: erxi.wu@BSWHEALTH.org

Cancer Res 2023;83:1–15

doi: 10.1158/0008-5472.CAN-20-0371

This open access article is distributed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.

©2022 The Authors; Published by the American Association for Cancer Research

sites occupy a significant proportion of the genome and can harbor somatic mutations, which play driver roles through miRNA-related pathways.

miRNAs are endogenous regulatory noncoding RNAs that are ~22nt in length and act by targeting messenger RNAs (mRNA) for cleavage or translational repression. The diversity and abundance of miRNA targets contribute to the complexity of gene regulatory networks. Increasing lines of evidence have demonstrated that miRNAs play critical functions in various developmental, physiological, and pathological processes including cancer. Deregulation in the expression of miRNAs and their targets have been observed in various types of human cancer. In medulloblastoma, miR-217 was found to promote an oncogenic phenotype by blocking target genes such as histone deacetylase SIRT1 and SMAD family member SMAD7, inhibiting apoptosis (2). Furthermore, IGF2BP1 was found to interfere with miRNA-mediated mRNA suppression, enabling increased expression of mRNAs that contribute to tumorigenesis in ovarian cancer (3).

Mutations in miRNAs or their target genes may exert important effects on their deregulated expression. A number of studies have suggested that somatic mutations could impact miRNA–gene interactions (mGI) and are related to cancer development (4). Nucleotide variants in mRNAs within or near miRNA target regions may alter miRNA–mRNA binding by modifying mRNA structure nearby. For instance, a mutation within the E2F1 mRNA's miRNA target site was found to deregulate miRNA binding, contributing to oncogenesis in colorectal cancer (5). Beyond cancer, a gain-of-function mutation in miRNA-140 drastically altered its mRNA target profile, contributing to pathogenesis in human skeletal dysplasia (6). Taken together, mutations may potentially deregulate miRNA–mRNA interactions by several different mechanisms.

Several initial studies explored the mutational effect on miRNAs and their targets. SomamiR (7) mapped mutations to miRNAs and their targets, and developed a tool to calculate enrichment of mutated targets on KEGG pathways. PolymiRTS (8) used a TargetScan score to evaluate the effects changed by mutations on miRNA targets. However, mutations in specific cancer types and in individual patients were not considered in these studies. Many miRNA family members can form opposing associations with mRNAs, with some miRNAs exhibiting a positive correlation in expression levels whereas other members in the same family exhibit a negative correlation (9). Moreover, mRNAs found to bind many miRNA partners can change partners strongly based on the cellular environment (10). Such findings point to a need to assess miRNA–mRNA interactions independently in cancer-specific and patient-specific contexts.

miRNA targeting occurs not only in 3' untranslated regions (3' UTR), but also in coding regions and 5' UTRs (11–16). Silent and nonsilent mutations in these regions have been shown to be relevant for cancer and other diseases (5, 11, 17). However, the mechanism of how mutations influence miRNA target genes was not clear in such studies. Several possible mechanisms have been elucidated, including the creation of novel splice variants, dosage control of oncogenes, and modifications in mRNA secondary structure (18–20). A study by Stegeman and colleagues used miRNA mimics and reporter gene assays and showed that mutations could change the expression of target genes in prostate cancer (21). Previously, miRNA-driven deregulation of cell-cycle genes and oncogenes has been documented to drive instances of human cytomegalovirus and pancreatic cancer (22, 23). In addition, compound mutations affecting both miRNA and mRNAs involved in an interaction are associated with a deregulated disease phenotype (24, 25). But it remains elusive the extent to which mutations could impair mGIs, and a global view of driver

mutation-mediated gene regulatory network perturbations on a pan-cancer scale is needed.

Recently, with the development of high throughput sequencing projects, such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC), numerous somatic mutations have been identified in various cancer types. Large-scale computational analysis has proven to be useful in characterizing mutations of unclear functional significance in cancer, suggesting that such a study with a focus on miRNA–mRNA interactions could lend novel insights (26–28). These mutations might create new miRNA binding sites or lose several binding sites, which further perturb miRNA–gene regulation. The quantity of binding sites is crucial to mGIs, as multiple binding sites for a single miRNA show stronger regulatory significance regardless of the sites' location in the mRNA (29). To address these goals, we interrogated to what extent and how somatic mutations perturb miRNA–gene regulatory networks in cancer. In this study, we derived elaborate miRNA–target interaction networks in each cancer type by integrating empirically validated binding sites by CLIP-seq, and expression correlation between miRNAs and target genes (30). We demonstrate that somatic mutations are likely to occur selectively in miRNAs or target genes to perturb cancer hallmark-related functions. A mutually exclusive pattern is found for mutations in miRNAs and their targets. Driver mutations that significantly perturb miRNA regulatory networks in cancer are further identified. We show that driver mutations could exert their functions by impairing RNA binding affinity, resulting in alteration of target gene expression profiles. Intriguingly, driver gene targets that are significantly down-regulated in cancer often function as tumor suppressors during cancer progression. Our study provides a valuable resource for systematic investigation of the functional impact of somatic mutations on miRNA regulation in cancer.

Materials and Methods

Identification of global mGIs in 3' UTRs

Argonaute (AGO) proteins are RNA-binding proteins (RBP) and essential components of the RNA-induced silencing complex (RISC), which is the molecular machinery for miRNA-induced silencing. High-throughput sequencing of immunoprecipitated RNAs after cross-link (CLIP-seq) to AGO proteins provides powerful ways to trace the footprints of miRNA binding sites (31). We downloaded peaks of 36 AGO CLIP-seq datasets from starBase v2.0 (32) and considered these binding sites as evidence of physical interactions between miRNAs and target genes (Fig. 1A). All the peaks of AGO CLIP-seq data were merged together based on their genomic locations to obtain a global pool of miRNA–gene physical interactions. Peaks that were overlapped with at least 1bp were merged together to provide evidence of permissive regions for miRNA binding. A total of 434,701 merged peaks were finally obtained after converting to hg38 genome assembly and ready for feeding to software prediction of precise sites.

Although the AGO CLIP-seq peaks could capture the footprints of miRNA binding sites, which miRNAs could bind to specific target sites still remains unclear. Agarwal and colleagues showed that many noncanonical sites detected by cross-linking method do not mediate repression despite binding the miRNAs whereas the vast majority of functional sites are canonical and can be identified by TargetScan (v7.0; ref. 30). To identify mGIs and functional binding sites on the mRNA, 16,347,639 interactions between all the miRNAs in miRBase release 21 and target genes in whole human genome were downloaded from TargetScan (v7.0; ref. 30). Binding sites of each miRNA–gene pair were further intersected with AGO-merged peaks and we finally

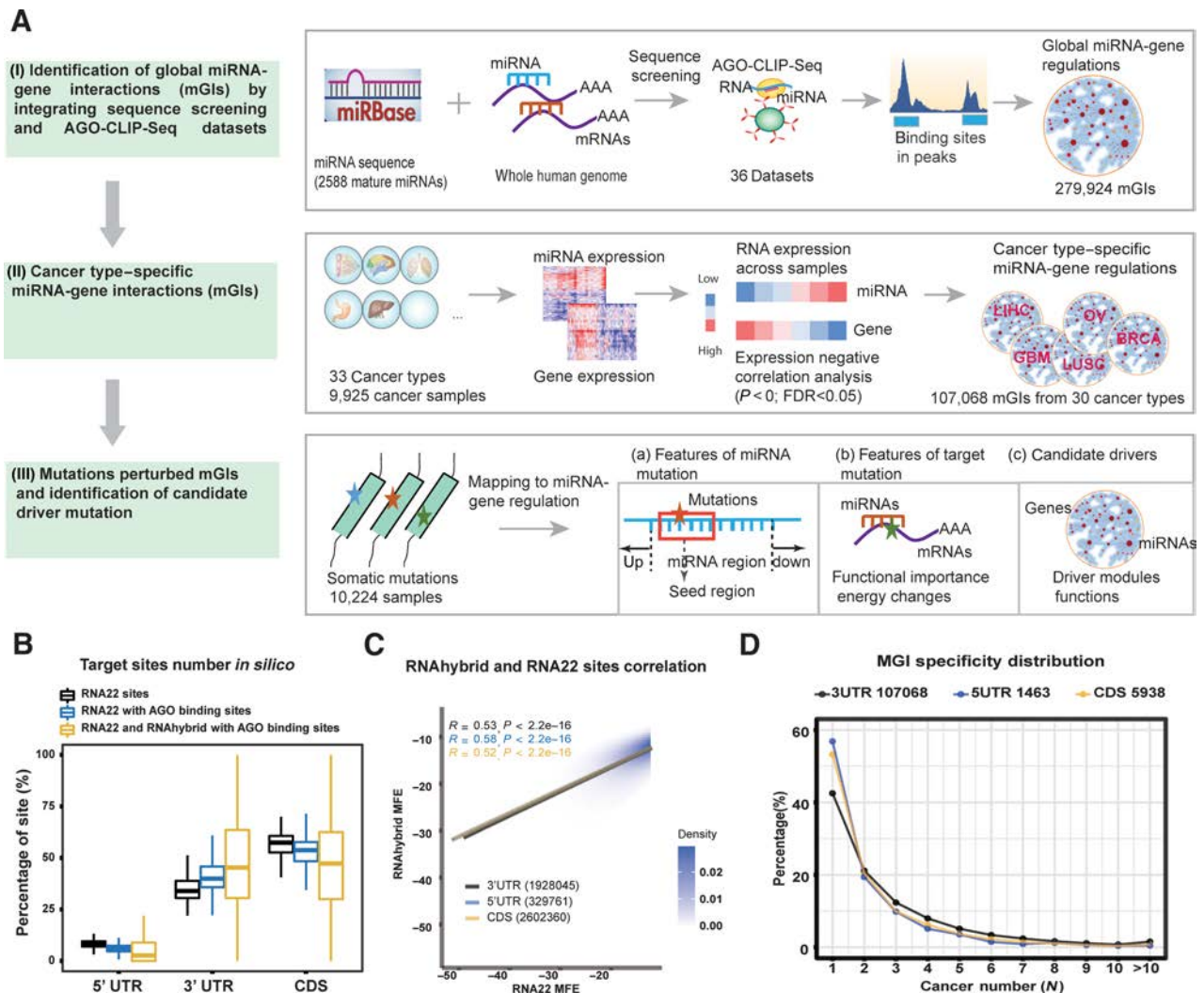


Figure 1. The landscape of miRNA-gene regulatory networks across cancer types. **A**, The workflow to identify genome-wide miRNA-gene regulation in cancer. Firstly, global mGIs were identified by integrating sequence alignment and AGO-CLIP-seq datasets. Then miRNA and mRNA expression profiles were integrated to identify cancer-specific interactions. Somatic mutations were mapped to mGIs to identify candidate driver mutations. **B**, Target sites distribution in 3' UTR, CDS, and 5' UTR regions from RNA22 raw data; RNA22 and AGO binding overlapped sites; RNA22 significant ($P < 0.05$), RNAhybrid significant ($P < 0.05$), and AGO binding overlapped sites. **C**, MFE correlation between RNA22 and RNAhybrid in CDS, 5' UTR, and 3' UTR regions at all AGO binding sites. **D**, The proportion of active mGIs from 3' UTR, 5' UTR, and CDS regions in different numbers of cancer types. (**A**, Created with BioRender.com.)

obtained 279,924 mGIs among 2,586 mature miRNAs and 4,198 target genes as global reference in 3' UTRs.

TargetScan (v7.0) predicts miRNA targets by searching for conserved sites six to eight nucleotides long that match to the seed region of 3' UTR only (30), whereas RNA22 (v2) and RNAhybrid (v2.1.2) are seedless prediction tools available for detection of target sites in CDS, 5' UTR, and 3' UTR.

Identification of global mGIs in coding sequence and 5' UTR

An assumption made by using TargetScan (7.0; ref. 30) to identify mGIs is that targets are only detected in 3' UTRs, so miRNA targets in coding sequences (CDS) and 5' UTRs are not detected by TargetScan (v7.0; ref. 30). To overcome this limitation, we downloaded ~250M (million) predicted miRNA target sites from RNA22 (v2; ref. 33) for the 2,588 mature human miRNAs from miRBase (release 21; ref. 34),

which targeted 140M sites in CDSs, 84M sites in 3' UTRs, and 23M sites in 5' UTRs from ENSEMBL78 (**Fig. 1B**; **Table 1**). To obtain more reliable sites, we overlapped these sites with AGO CLIP-seq sites collected from starBase (v2.0; ref. 32), which resulted in 2.7M sites in CDSs, 1.9M sites in 3' UTRs, and 0.3M sites in 5' UTRs (4.9M total sites; **Fig. 1B**; **Table 1**). These 4.9M target sequences and all 2,588 miRNA sequences were fed into RNAhybrid (v2.1.2; ref. 35) using default parameters with no helix constraint to ensure seedless predictions. Only the 0.15M sites participating in significant predictions ($P < 0.05$) by RNAhybrid (v2.1.2; ref. 35) and RNA22 (v2; ref. 33), and overlapping with AGO sites were kept.

Because RNA22 (v2) and RNAhybrid (v2.1.2) differ in that RNA22 (v2) is a pattern-based prediction tool that identifies miRNA targets through associating miRNAs with "target islands" (33), whereas RNAhybrid (v2.1.2) identifies miRNA targets through identifying

Table 1. miRNA targets by different algorithms.

Algorithms	5' UTR	3' UTR	CDS	Total
RNA22 raw	22558002	83939429	145391270	251888701
RNA22 raw and overlapped AGO sites	329761	1928045	2673408	4931214
RNA22 and RNAhybrid significant and overlapped AGO sites	12089	64839	68381	145309

energetically favorable hybridizations between miRNA and targets (35), we compared minimum free energies (MFE) of the 4.9M reliable sites overlapped by two algorithms in all types of sites (3' UTR, CDS, and 5' UTR) and calculated the Pearson correlation coefficients. We observed highly significant correlations ($P < 2.2 \times 10^{-16}$) between RNA22 and RNAhybrid in all three types of sites, where the correlation is 0.53 for 3' UTR sites, 0.58 for 5' UTR sites and 0.52 for CDS sites (Fig. 1C). To further evaluate the consistency of predicted sites in CDS and 5' UTR by two algorithms, we overlapped the significant sites detected by RNA22 (v2) and RNAhybrid (v2.1.2).

Identification of cancer type-specific mGIs

To obtain the functional mGIs in specific cancer type, we considered the expression correlation among samples for each mGI of global reference in the cancer type (Supplementary Table S1). The mGIs obtained from CDS and 5' UTRs were then combined with mGIs from 3' UTRs as the global reference for further analysis (Supplementary Table S2). Gene expression quantification by RNA-seq data and isoform expression quantification of mature miRNAs by miRNA-seq data of 33 cancer types were downloaded from the TCGA. These data were collected with written informed consent from the patients. This study was conducted in accordance with recognized ethical, and approved by an institutional review board. For each cancer type, datasets with more than 50 common samples (Supplementary Table S1) between RNA-seq and miRNA-seq data were extracted and Spearman's rank correlation of expression for each pair of miRNA and target gene was calculated.

$$\rho_i = \frac{\sum_k (rm_{ik} - \bar{rm}_i)(rg_{ik} - \bar{rg}_i)}{\sqrt{\sum_k (rm_{ik} - \bar{rm}_i)^2 \sum_k (rg_{ik} - \bar{rg}_i)^2}},$$

where ρ_i is the rank correlation coefficient between the expression of miRNA_{*i*} and gene_{*i*} in the *i*_{th} mGI. The ranks of miRNA_{*i*} and gene_{*i*} expression in *k*_{th} sample are represented as *rm_{ik}* and *rg_{ik}*. *N* is the total number of samples in a cancer type. To control the FDR, *P* values of correlations for mGIs in each cancer type were further adjusted by Benjamini and Hochberg method. Finally, mGIs with $\rho < 0$ and FDR < 0.05 were considered as functional and reliable interactions in each cancer type (Supplementary Tables S3–S5).

Validation of identified mGIs in 3' UTR

To validate the mGIs identified from cancer samples, we built a benchmark dataset by collecting various sources of experimentally validated mGIs. Both validated positive and negative mGIs were collected from mirRecords (36), miRTarBase (37), TargetMiner (38), and TarBase (39). After mapping gene IDs to approved HGNC gene symbols and miRBase IDs (34) and integrating redundant mGIs from different sources, we finally retrieved 5,577 validated mGIs including 5,139 positive mGIs, which are validated as functional and 438 negative mGIs, which are validated as non-functional. Functional

impacts of mutations were not considered during validation, as they would not be consistently common enough across cancer samples to skew the Spearman coefficient and affect mGI identification. Considering the incomplete nature for both benchmark and our predicted datasets, 166 common mGIs involved in our predictions were extracted for evaluation. Contingence table of prediction result in 3' UTR was generated, and Fisher exact tests were calculated on the basis of the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Moreover, evaluations of performance were also calculated as below:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$F_1 = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$$

$$\text{Accuracy(ACC)} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FN} + \text{FP} + \text{TN})$$

Cancer mutations

The cancer mutation dataset of 33 cancer types was achieved from “Multi-Center Mutation Calling in Multiple Cancers” (MC3) and was produced using six different algorithms on data from TCGA. A total of 3,518,261 mutations in 9,819 cancer samples were included in our analysis.

Mutation probability

We mapped cancer mutations to the specified region by bed-tools (40). To calculate the mutation probability in a genomic region, we first defined mutation rate of *mut_i* as: $R(\text{mut}_i) = N(\text{mut}_i) / N$, where $N(\text{mut}_i)$ is the number of samples with mutation_{*i*} in a cancer and *N* is the number of all samples in a cancer. Then mutation probability of a region *r* was calculated as:

$$P(r) = \sum_r^{\text{mut}_i} R(\text{mut}_i) / (\text{length}(r)).$$

In brief, *P(r)* was calculated by summing up the mutation rate of all mutations in the region *r* and normalized by the length of region. In this way, we could compare the level of mutation probability for regions with different lengths and different number of mutations. To evaluate the level of mutation probability in mature miRNAs, the mature miRNA regions, seed regions (2–8 bp of each miRNA), upstream and downstream flanking 50bp regions of mature miRNAs were calculated for comparison.

Significance of mutation exclusivity on mGI

For each mutated mGI, mutations could act on either the miRNA or target gene, or both of them. In this case, we classified mutated mGIs as miRNA-mutation mGIs, target-mutation mGIs, and dual-mutation mGIs. To find out whether the mutations on mGIs work in a synergetic way or in an exclusive way, we evaluated the mutation exclusivity of mGIs based on the occurrence of dual-mutation mGIs. In this way,

mutation exclusivity on miRNA–gene pair was tested by hypergeometric distribution based on the number of all mutated mGIs, miRNA–mutation mGIs, target–mutation mGIs, and dual–mutation mGIs in each cancer type.

$$p(X \leq d) = F(X|N, K, M) = \sum_{X=0}^d \frac{\binom{K}{X} \binom{N-K}{M-X}}{\binom{N}{M}},$$

where N is the number of mutated mGIs, K is the number of mGIs with miRNA mutation, which equals $m+d$, M is the number of mGIs with target mutation, which equals $t+d$, and X is the number of dual mutation mGIs. In the case, m is the number of miRNA exclusively mutated mGIs, t is the number of target exclusively mutated mGIs, and d is the number of dual–mutation mGIs.

Identification of driver mutations on mGI

For each mGI_{*j*} identified in a type of cancer_{*i*}, first all the mutations within mature miRNA_{*j*} and binding sites of target gene_{*j*} were considered as candidate driver mutations. For each mutation_{*k*} on mGI_{*j*}, we searched samples with mutation_{*k*} in cancer_{*i*} and integrated them with gene expression of miRNA_{*j*} and gene_{*j*} in mutated samples. To identify the driver mutations that could affect miRNA–mRNA binding, we assumed the driver mutation on either miRNA or target site could alter the inhibitory role of miRNA_{*j*} and cause abnormal expression of target gene_{*j*}, which is isolated from the distribution of nonmutated control samples. Here two sets of control samples were considered, thus nonmutated cancer samples in 25 cancer types and nonmutated normal samples in five cancer types. For each mutated mGI_{*j*} in cancer_{*i*}, we used the linear regression model L_j to fit the expression distribution of control samples and calculated the prediction interval with a probability of 0.95.

$$L_j : g_j \sim \beta_0 + \beta_1 * m_j,$$

where g_j and m_j are the expression of gene_{*j*} and miRNA_{*j*} among control samples. β_0 and β_1 are estimated parameters trained by expression of control samples, where only the models with significant P value ($P < 0.05$) are considered as successful. In this way, the control samples should fall into the predictive confident interval whereas samples with driver mutations on mGI_{*j*} should fall out of it (Supplementary Table S6).

Minimum-free energy change by driver mutations on mGI

We calculated the MFE for wild-type mGIs and driver-mutated mGIs (Supplementary Table S6) using RNAhybrid (41). The energy change of mGI altered by driver mutation is defined as:

$$\Delta E = \text{MFE}_{\text{mut}} - \text{MFE}_{\text{wild}}$$

The lower the ΔE is the more stable, as it is for the binding of miRNA–mRNA. Thus, it can be more easily for the miRNA to carry out its inhibitory function and downregulate the expression of target gene more effectively.

When calculating the proportions of mGI with changed or unchanged MFE, we defined mGIs with $|\Delta E|$ in top 30% as energy changed and all the other 70% as unchanged. Similarly for the gene

expression, mGIs with $|\Delta \text{Expression}|$ in top 30% was defined as expression changed and the others as unchanged.

The differential impact of mutations in miRNAs compared with mRNA targets was further assessed using general linear mixed modeling. We defined our model L_k as:

$$L_k : \Delta E_k \sim (\beta_0 + b_{C, oc}) + \beta_1 M_k + e_{mk},$$

where ΔE_k , M_k , and e_{mk} represent the change in free energy, the site of the mutation (miRNA vs. mRNA), and error for mGI_{*k*} retained after RNAhybrid analysis, respectively. β_0 and β_1 are estimated parameters trained on the mGIs, and $b_{C, oc}$ represents a control term for energy changes in the different cancer types. β_1 was used to assess which type of mutation was more destabilizing, as $\beta_1 > 0$ would suggest that mRNA mutations had a greater impact whereas $\beta_1 < 0$ suggested that mRNA mutations would have a reduced impact. Significance was assessed using the t value associated with β_1 .

Sex bias across driver mutations in mGIs

Certain cancer type exhibit differential mutation patterns between different sexes (42). To investigate the potential for such a bias in mGIs, we counted patients with mGI driver mutations for each gene in each cancer. To assess sex bias, we selected cancers that featured at least 1 male and 1 female patient. Within each cancer we selected genes that were part of an mGI driver mutation in at least 2 patients. With these criteria, 90 unique combinations of cancer types and genes were studied across 15 TCGA cancer cohorts.

mGI network analysis

Network visualization was conducted by the software Cytoscape. The layout of driver network was grouped by the classification of driver gene nodes, which are “mutated miRNAs,” “nonmutated miRNAs,” “mutated genes,” and “nonmutated genes.” Node degree is the number of interactions for each node in the global network.

Cancer hallmark and enrichment

To investigate the functional importance of the driver genes and mutations, functional enrichment analysis of the driver genes was carried out to investigate whether they were enriched in cancer hallmarks. The gene list of each cancer hallmark was obtained from one of the previous studies. We used hypergeometric test for exploring the statistical significance and the P value was computed as follow:

$$p(X \geq t) = 1 - F(X|N, K, M) = 1 - \sum_{X=0}^t \frac{\binom{K}{X} \binom{N-K}{M-X}}{\binom{N}{M}},$$

where N is the number of genes in the whole genome, of which, K genes were involved in the functional category under investigation (such as cancer hallmarks), and the number of candidate driver genes is M , of which, t genes were involved in the functional category.

Differential expression analysis

Differential expression analysis was carried out for all genes in cancer types having both tumor and normal samples. Count of raw reads in each gene was downloaded in “HTSeq–Counts” format from TCGA. R package “DESeq2” was used to carry out differential analysis. Genes with reads count smaller than 10 in total samples were filtered out. Finally, genes with $|\text{fold change}| > 2$ and adjusted P value < 0.05 were considered as significantly differentially expressed.

Random permutation test for differentially expressed genes

For a given set of N genes, we calculated the proportion of significantly downregulated, upregulated, and nondifferentially expressed genes, which are denoted as D , U , and M . As the definition, $D+U+M=1$. To test whether the observed values of D , U , and M are significant, we randomly picked N genes from all genes in all cancer types and calculated the proportions of them as d , u , and m in each random case. By repeating 10,000 times, we got the random distribution of variable d , u , and m . By comparing the D , U , and M with the random distributions of d , u , and m , we got the significance for each classification of differentially expressed genes.

Data availability

Mutation information (MC3), gene expression RNA-seq data, miRNA-seq data as well as patient clinical features (survival and gender) in this study were obtained from The Cancer Genome Atlas Project (TCGA) project available at the Genomic Data Commons portal at <https://portal.gdc.cancer.gov/>. AGO CLIP-seq datasets are available from starBase v2.0 (32). miRNA details are available in miRBase (release 21; ref. 34), and miRNA target gene sequences are available in TargetScan (v7.0; ref. 30), RNA22 (v2; ref. 33), and RNAhybrid (v2.1.2; ref. 35). Validated positive and negative mGIs were compiled from miRecords (36), miRTarBase (37), TargetMiner (38), and TarBase (39) to build a benchmark dataset. The data generated in this study are available in a user-friendly website resource, “mGI-map” (<https://mgimap.dellmed.utexas.edu>).

Results

Integrative method for construction of pan-cancer miRNA regulatory networks

To analyze miRNA–gene regulation across cancer types, we proposed a computational method for identifying active miRNA–gene regulation in specific cancer types (Fig. 1A). In brief, high-throughput sequencing of immunoprecipitated RNAs after cross-link (CLIP-seq) to AGO proteins was used to identify endogenous genome-wide interaction maps for miRNAs (31) and bioinformatics tools were developed to infer specific target sites each miRNA could bind to (30, 33, 35). In a previous study, Agarwal and colleagues showed that many noncanonical sites detected by cross-linking methods did not mediate gene repression despite their ability to bind miRNAs whereas the vast majority of functional sites were canonical and could be identified by TargetScan (v7.0; ref. 30). In this study, AGO CLIP-seq data from starBase v2.0 (32) were integrated with TargetScan 7.0 (30) to get mGIs in 3′ UTR, and were integrated with RNA22 (v2; ref. 33) and RNAhybrid (v2.1.2; ref. 35) to get mGIs in CDS and 5′ UTR regions (Supplementary Table S7). We found predictions for either CDS or 5′ UTRs by RNA22 (v2) and RNAhybrid (v2.1.2) were highly similar (Supplementary Figs. S1A and S1B). To evaluate the consistency of predicted sites in 3′ UTR by different algorithms, we overlapped the significant sites in 3′ UTR detected by RNA22 (v2), RNAhybrid (v2.1.2), and TargetScan (v7.0). We found TargetScan (v7.0) sites covered almost all of the sites detected by RNA22 (v2) and RNAhybrid (v2.1.2; Supplementary Fig. S1C).

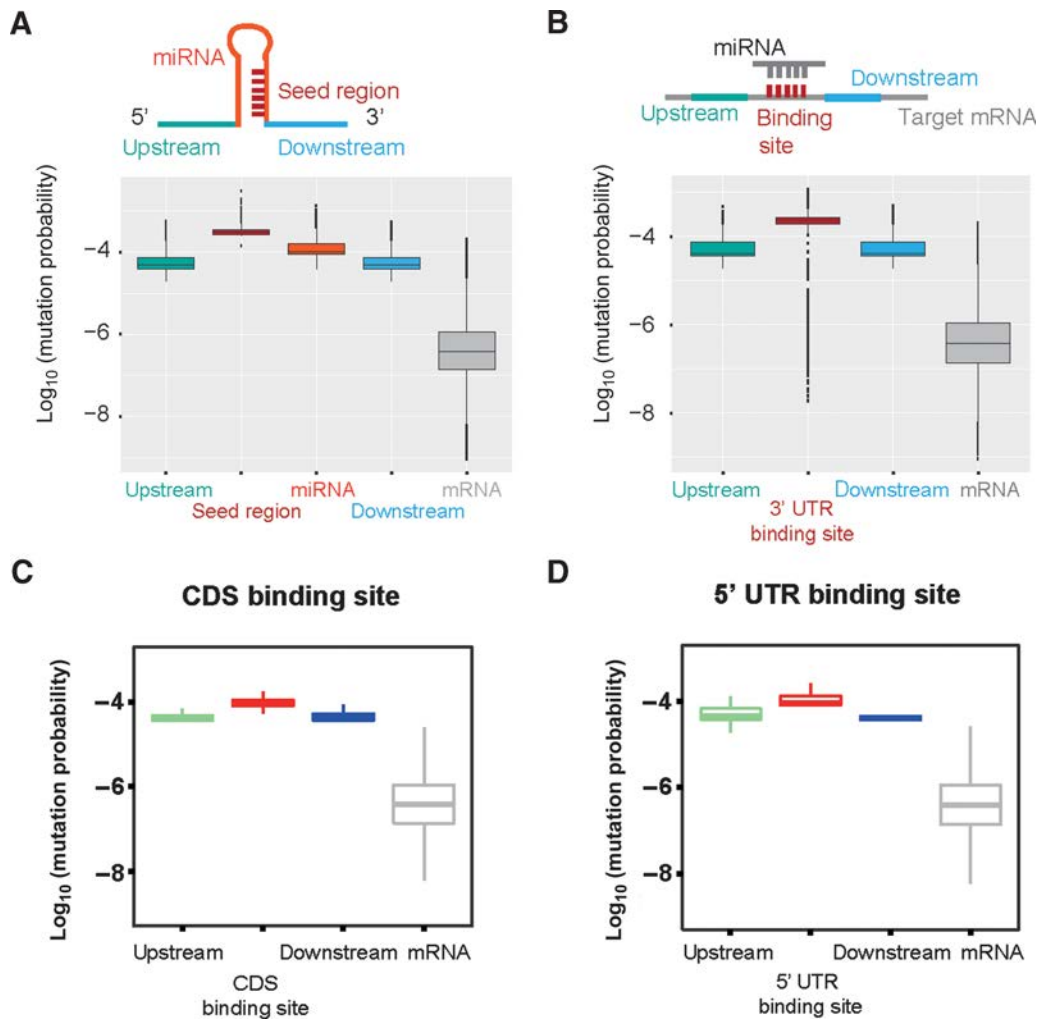
We derived 314,121 mGIs among 2,588 mature miRNAs and 5,207 target genes as a global reference network, including 279,924 mGIs in 3′ UTR, 27,016 mGIs in CDS, and 7,181 mGIs in 5′ UTR regions (Supplementary Table S2). After the inference of the global mGI network, we integrated the miRNA and gene expression profiles in specific cancer types to identify active miRNA–gene regulation. Only negatively correlated miRNA–gene pairs ($FDR < 0.05$) were kept for

further analysis. In total, 114,469 unique mGIs were obtained from 30 of 33 cancer types, including 107,068 in 3′ UTR, 5,938 in CDS, and 1,463 in 5′ UTR regions (Fig. 1D). The Spearman correlation coefficients ranged from -0.07 to -0.83 with a mean of -0.23 , as the correlation associated with the FDR cutoff changed on the basis of the cancer cohort. Next, we validated 3′ UTR results using a benchmark dataset of experimental results collected from various sources. The performance of our predictions was: recall rate = 93.5%, precision = 94.1%, F1 score = 93.8%, accuracy = 88.6%. In addition, our predictions exhibited a significant overlap with the benchmark dataset (Supplementary Fig. S1D: OR, 6.36; $P = 0.015$). These results indicate that integration of multi-omics datasets enables us to identify cancer relevant miRNA–gene regulation at the systems level.

Comparing miRNA–gene regulatory networks across cancer types, we identified marked rewiring in the miRNA regulatory programs among different cancer types, with a unique “on/off” switch depending upon the cancer type. We found that $\sim 42.47\%$ of 3′ UTR mGIs, 53.22% of CDS mGIs, and 56.87% of 5′ UTR mGIs occurred only in one cancer type, but only 1.55% of 3′ UTR mGIs, 0.72% of CDS mGIs, and 0.48% of 5′ UTR mGIs were observed in greater than 10 cancer types (Fig. 1D; Supplementary Figs. S2A and S2B). Notably, most of the active mGIs reside in 3′ UTR sites if compared with the number in CDS and 5′ UTR sites. We also observed that the distribution of cancer general (common) mGIs ($N \geq 10$ cancer types) in 3′ UTR sites was more homogeneous than that of total mGIs (Supplementary Figs. S3A and S3B). The low conservation of miRNA–gene regulation could be explained in part by the cancer-specific expression of miRNAs or target genes. This indicates although the proportion of general mGIs is small, they are relatively stable across cancer types. In contrast, the number of cancer specific mGIs increased as the total number of mGIs grew (Supplementary Fig. S3C). Furthermore, hierarchical clustering analysis was conducted based on the occurrence of cancer frequent mGIs in 3′ UTR. Several cancer types with similar origins were clustered together (Supplementary Fig. S3D). OV (ovary) and CESC (cervix) were clustered into gynecologic system, showing low occurrences of cancer frequent mGIs. KIRP and KIRC (kidney) were clustered into urologic system, showing medium occurrences of frequent mGIs. COAD, READ (colorectal), and LUAD, LUSC (lung) were separately clustered into gastrointestinal and thoracic systems, respectively, showing relatively high occurrences of frequent mGIs. All these results indicate the intricate functional roles of miRNA regulation in cancer. Moreover, we performed functional enrichment analysis using the target genes from 3′ UTR that were observed in cancer frequent mGIs. We found that these genes were involved in general cancer-related functions, including FoxO signaling pathway, TGF β signaling pathway and cell adhesion-related functions (Supplementary Fig. S3E). Taken together, our integrative analysis reveals cancer-specific miRNA–gene regulation, providing a valuable resource for mechanistic investigation of the function of miRNAs in cancer.

miRNA mutations and target gene mutations in cancer

miRNAs play important roles in the development and progression of cancer. Several miRNA-associated mutations have been identified before, but systematic analysis of mutational landscape on miRNA regulation is still lacking. In this study, cancer somatic mutations obtained from the TCGA project were mapped to miRNAs and their target genes, or other regions in the human genome. We found that the mutation probability in miRNAs was greater than that in mRNAs (Fig. 2A, $P < 2.2 \times 10^{-16}$, Wilcoxon rank-sum test). Moreover, we found that the genomic regions of miRNAs also had a higher mutation probability than upstream or downstream flanking regions (Fig. 2A,

**Figure 2.**

miRNA and target gene mutation profiles. **A**, The distribution of mutation probability for miRNA-related regions, including upstream, downstream of miRNA regions, miRNAs, seed regions, and mRNA regions. **B**, The distribution of mutation probability for 3' UTR binding sites, including upstream, downstream, miRNA binding sites in targets, and all target mRNA regions. **C**, The distribution of mutation probability for CDS binding sites, including their upstream, downstream, target binding sites, and all target mRNA regions. **D**, The distribution of mutation probability for 5' UTR binding sites, including their upstream, downstream, target binding sites, and all target mRNA regions.

$P < 2.2 \times 10^{-16}$). Specifically, the mutation probability of the seed regions within miRNAs was greater than other parts within miRNAs or the flanking regions (Supplementary Fig. S4A, $P < 2.2 \times 10^{-16}$), suggesting the seed regions of miRNAs are required for miRNA regulation. In addition, we identified thousands of mutations located in target coding genes of miRNAs. Compared with the flanking regions, the miRNA binding sites in 3' UTR, CDS, and 5' UTR of target genes all exhibited a higher mutation probability (Fig. 2B–D; Supplementary Fig. S4B). These results indicate that mutations are likely to occur in the genomic regions that are critical for miRNA binding.

Next, we explored the signaling pathways that were likely perturbed by these mutations. As the hallmarks of cancer comprise biological capabilities acquired during the multistep development of human tumors, we evaluated whether these mutations could possibly perturb cancer hallmark related functions. As shown in Supplementary Fig. S4C, we found that the mutated miRNA target genes in 3' UTRs were significantly enriched in various types of cancer hallmarks, such

as regulation of cell proliferation and DNA repair. To further explore if these mutations could play a functional role in perturbing miRNA–gene regulation, we estimated the functional impact of somatic mutations identified in 3' UTRs. Several methods have been proposed to assess the effects of mutations on protein function, and these methods are complementary to each other. We herein used Combined Annotation–Dependent Depletion (CADD; ref. 43), which is a framework that integrates multiple annotations into one metric, to explore the functional impact of mutations in target genes of miRNAs. For each cancer type, we randomly selected the same number of mutations as background controls. We then compared the CADD scores of mutations in target genes with those of randomly selected mutations. In the majority (23/31) of cancer types, mutations in target genes had significantly higher scores (Wilcoxon rank-sum test $P < 0.05$) than random controls (Supplementary Fig. S4D), suggesting that the identified miRNA target mutations could have deleterious effects in cancer. Evolutionary conservation of mutated residues has also been

demonstrated to reflect the functional importance of the mutational event (44). We therefore explored the conservation feature of the mutations in target genes. We found that positions harboring target gene mutations were more likely to be conserved than positions harboring randomly selected mutations in most (29/31) cancer types (Supplementary Fig. S4D). All these results indicate that there are widespread mutations enriched within miRNA sites and their target genes, which could functionally perturb the binding of miRNAs, thus playing critical roles in cancer.

Mutually exclusive mutations in miRNA–gene regulation

It has been observed that sets of genes that are co-involved in the same cancer pathways tend not to harbor mutations together in the same patient, which is called mutual exclusivity (45, 46). However, it remains elusive if the mutual exclusivity hypothesis could be extended in the context of miRNA regulation. We therefore assessed whether the same patients with cancer could simultaneously carry miRNA mutations and target gene mutations. To do so, we calculated the number of mGIs in 3' UTR and CDS regions for three mutation perturbation models. The first one was that mutations located only in miRNAs to perturb the interactions; the second one was that mutations located only in target genes to perturb the interactions; and the third one was that mutations were located in both miRNAs and their target genes (Fig. 3A). We found that a vast majority of mGIs harbored miRNA mutations for 3' UTR mGIs (Fig. 3B, left) or target mutations for CDS mGIs (Fig. 3B, right), and only a small fraction of mGIs had both miRNA and target mutations (Fig. 3B). 5' UTR mGIs were too few to make statistical calculations. In addition, we found that patients in different cancer types exhibited different regulatory patterns (Fig. 3C and D; Supplementary Figs. S5A and S5B). For example, in 3' UTR mGIs the majority of patients with ACC, LAML, PAAD, and READ carried miRNA mutations. In contrast, TGCT, THYM, and PCPG patients were likely to have target binding site mutations (Fig. 3C). In CDS mGIs, most of the cancer types carried target mutations except for ESCA, PAAD, READ, and THYM (Fig. 3D). In summary, we found that the mutual exclusivity of miRNA versus target mutations was more significant across cancer types than expected by chance (Fig. 3C and D, $P < 0.05$), which occurred in 23 of 25 cancer types for 3' UTR mGIs and 14 of 23 cancer types for CDS mGIs.

Specifically, we identified mutually exclusive 3' UTR mGIs that had been reported in various types of cancer. For example, the let-7 family is a conserved family of miRNAs and the deregulation of this family was observed in many cancer types, such as breast cancer, lung cancer, and melanoma. Here, several mutations were observed in let-7f-1–3p, let-7a-5p, and let-7b-3p. In addition, several mutations located within target genes were also observed in patients, such as LIFR, SPIRE1, and E2F3. However, no patient was found with both miRNA mutations and target gene mutations. These observations suggest that the mutations selectively target miRNAs or target genes to perturb miRNA–gene regulation in cancer, which is consistent with the pathway redundancy model in cancer.

Identification of driver mutations on mGIs

Distinguishing driver mutations from passenger mutations in individual genomes is a big challenge in cancer genomic studies. In this study, we integrated the mGI regulatory networks with target gene transcriptomic networks to identify candidate driver mutations in each cancer type. Our hypothesis was that candidate driver mutations were likely to perturb mGIs, which could be reflected by a significant change in the expression levels of target genes (Fig. 4A). We thus used a linear regression model to fit the expression distribution among non-

mutated control samples and identified driver mutations that elicited significant deviations from the normal correlation between miRNAs and their target genes (see details in Materials and Methods). In total, we identified 89 driver mutations for 3' UTR mGIs, 51 driver mutations for CDS mGIs, and 8 driver mutations for 5' UTR mGIs across 20 cancer types (Fig. 4B; Supplementary Figs. S6A and S6B; Supplementary Table S6). These driver mutations perturbed 154 mGIs in 3' UTRs, 57 mGIs in CDS regions, and 8 mGIs in 5' UTRs (Fig. 4B). Among the 89 driver mutations in 3' UTRs, 75 of them occurred in target genes and 14 of them occurred in miRNAs. In addition, 9 of 75 driver mutations in target genes had indeed been characterized as cancer mutations in the Catalogue of Somatic Mutations in Cancer (COSMIC) database and in this study we inferred their possible functions through miRNA regulatory networks.

Mutations at mGI interfaces could impair the binding of miRNAs to target genes by changing the MFE of binding sites, thus further tuning the regulation of target gene expression. We calculated the changes of MFE using RNAhybrid (41) and changes of target gene expression between wild-type and mutated 3' UTR and CDS mGIs. 5' UTR mGIs were again too few to make statistical calculations. By comparative analysis of different types of mGIs, we found that the energy changes were significantly correlated with the changes in target gene expression levels by Fisher exact test for 3' UTR mGIs ($P = 0.002$; OR, 3.3) but not for CDS mGIs ($P = 0.793$; OR, 1; Fig. 4C; Supplementary Table S6). We next speculated that if driver mutations disrupt miRNA–target binding, one would expect the repression of target gene expression by miRNAs to be reversed (i.e., increased). Consistently, we found that target genes in 3' UTR mGIs with increased expression (e.g., top 5%) showed a significantly higher alteration of MFE than those with decreased expression (e.g., bottom 5%) by Wilcoxon rank-sum test ($P = 0.038$, Fig. 4D, left). However, we did not observe significantly different alteration of MFE for target genes in CDS mGIs ($P = 0.575$, Fig. 4D, right). This difference could be due to the repression efficiency of target sites in 3' UTR and CDS regions. Although both 3' UTR and CDS sites are functionally conserved across mammalian species, 3' UTR sites are more conserved and appear to have more repressive effects than CDS sites (47). This may be a result of the complementary functions of 3' UTR and CDS sites: 3' UTR sites more efficiently trigger mRNA degradation whereas CDS sites more efficiently inhibit translation (48, 49). Fang and colleagues (50) have also investigated an additional synergistic function of CDS sites such that the presence of CDS sites appear to enhance the regulation of mRNAs at their 3' UTR sites. Anyway, it is interesting that driver mutations in 3' UTR showing greater influences through disruption of binding sites than those in CDS regions.

Furthermore, we investigated whether miRNA mutations and target mutations impact binding affinity in the same manner, we found that target mutations tended to have a greater impact on destabilizing free energy than miRNA mutations in 3' UTR mGIs ($P < 0.01$, Supplementary Fig. S7). Together, these results suggest that driver mutations are likely to perturb mGIs in cancer through their functional influence on miRNA binding affinity to target genes.

Finally, recent studies yielded gender-specific regulatory networks in many human tissues (42). Our analysis of sexual dimorphism in 3' UTR mGI mutation patterns yielded further insights. Across 15 TCGA cancer cohorts examined, we found that only EGFR exhibited a significant bias ($P < 0.05$) based on biological sex in LGG (Supplementary Fig. S8A; Supplementary Table S8). In this cohort, only females (9/285; 3.2%) possessed an EGFR mutation in mGI interaction sites. Although EGFR was only found to have a statistically significant bias in LGG, bias in EGFR mutation rates in SKCM and HNSC also

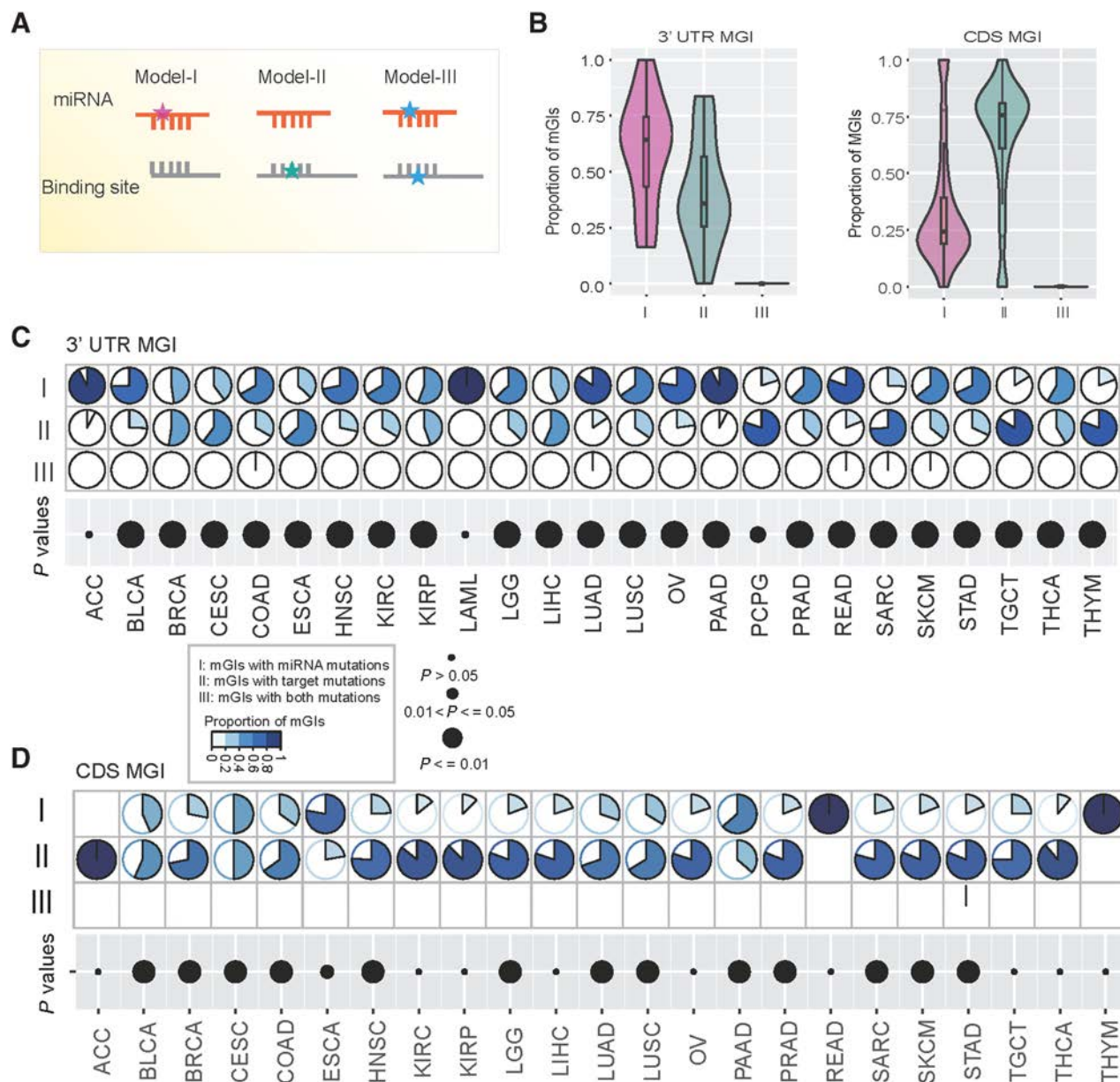


Figure 3.

The mutual exclusivity of mutations in mGIs. **A**, Three models to illustrate the mutual exclusivity of mutations in mGIs. **B**, The proportion of 3' UTR and CDS mGIs with different types of mutations. Purple, mGIs with miRNA mutations; green, mGIs with target mutations; blue, mGIs with both miRNA and target mutations. **C** and **D**, Pie plots show the proportion of samples with different types of mutated 3' UTR mGIs (**C**) and CDS mGIs (**D**). I, the proportion of samples with miRNA mutations; II, the proportion of samples with target mutations; III, the proportion of samples with both miRNA and target mutations. Dot plots at the bottom indicate the significance level for mutual exclusivity test.

ranked quite highly. Interestingly, the bias in EGFR-associated mGI mutations changes with the cancer type, as males were found to be at higher risk than females in SKCM (OR, 2.23) and lower risk than females in HNSC (OR, 0.27; Supplementary Fig. S8A). In addition, 8 of 9 mGI driver mutations in EGFR in LGG were missense mutations. Of these, five were clustered in the tyrosine kinase domain, and all 9 mutations were contained within or directly upstream or downstream of this domain (Supplementary Fig. S8B). Altogether, these findings indicate that sex bias associated with mGI mutations warrants further exploration.

mGI network perturbations by driver mutations

To analyze the structure and properties of perturbed mGI networks by driver mutations, we extracted all mGIs from 3' UTRs, CDS, and 5' UTRs involving both miRNAs and target genes that harbored driver mutations. As shown in Fig. 5A, the bipartite mGI networks had four types of node components: mutated miRNAs, nonmutated target genes, nonmutated miRNAs, and mutated target genes. mGIs could be perturbed by mutations in the miRNAs (mostly interactions between mutated miRNAs and nonmutated targets), or perturbed by mutations in the target genes (mostly interactions between

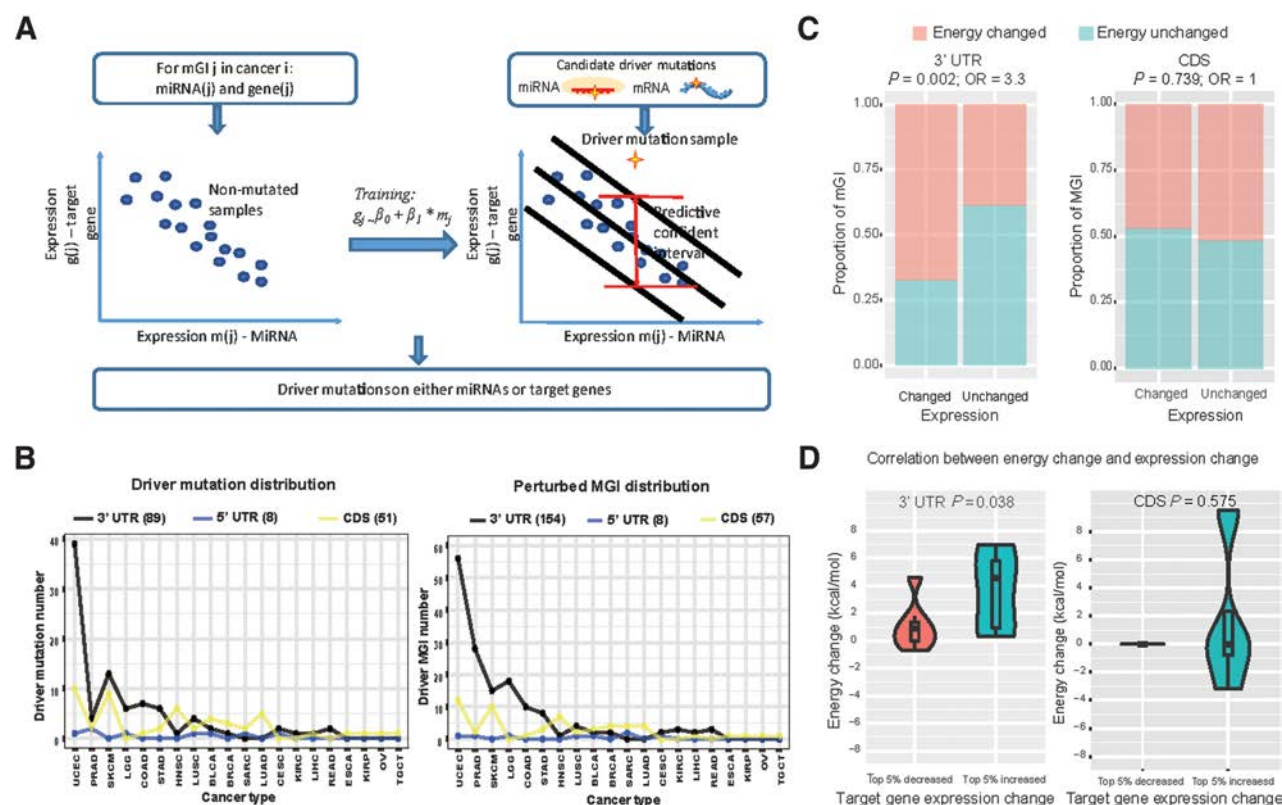


Figure 4.

The candidate driver mutated mGIs across cancer types. **A**, The linear regression model is used to identify the candidate driver mutated mGIs in cancer. **B**, The statistics of driver mutations (left) and perturbed mGIs (right) in 3' UTR, CDS, and 5' UTR regions across cancer types. **C**, The proportion of mGIs in 3' UTR (left) and CDS (right) regions with energy changed or not for expression changed/unchanged mGIs. Orange, energy changed; green, energy unchanged. **D**, The correlation between energy change and expression change for perturbed mGIs from 3' UTR (left) and CDS (right) regions.

nonmutated miRNAs and mutated targets). We then examined the functional pathways impacted by the driver mutation-induced mGI network perturbations. Through analysis of cancer hallmark signatures, we found that perturbed target genes were enriched in “insensitivity to antiproliferative signals” and “limitless replicative potential” (Fig. 5B). These two functions could help explain the increased growth and escaping from antiproliferative signals in the cancer cells carrying these driver mutations. Furthermore, we investigated the expression profiles of all the target genes in the perturbed mGI networks. As shown in Fig. 5C and D, 60 mutated target genes in perturbed networks were differentially expressed in at least one cancer type, and 48 nonmutated target genes (but targeted by mutated miRNAs) were differentially expressed in at least one cancer type. These results indicate that the target genes perturbed by driver mutations are associated with cancer development.

Intriguingly, for the miRNAs carrying driver mutations (shown in top left of Fig. 5A), many of them were from the let-7 family including hsa-let-7a-2-3p, hsa-let-7a-3p, hsa-let-7a-5p, hsa-let-7b-3p, hsa-let-7c-3p, hsa-let-7c-5p, hsa-let-7f-1-3p, hsa-let-7f-2-3p, hsa-let-7f-5p, and hsa-let-7g-3p. As a well-known miRNA family with functions of tumor suppression, aberrant regulation of the let-7 family had been reported in many cancer types (51). These included increased cellular proliferation in lung cancer (52), increased proliferation and migration in liver cancer (53), increased invasion and metastasis in gastric adenocarcinoma, lymph node metastasis in breast cancer (54), and so on. Our study further recaptures the functional importance of the

let-7 family in cancer by pointing out that the let-7 family could rewire molecular signaling through mutations within miRNAs themselves to dysregulate their targets.

Driver mutations in 3' UTRs primarily target tumor suppressor genes

To take a close investigation on the mutated 3' UTR, we made a volcano plot for them in their matched cancer types and found 26.3% of the mutated 3' UTR targets were downregulated (Fig. 6A). To explore the significance of these proportions, we randomly picked the same number of genes as the mutated target gene set in all cancer types 10,000 times and generated the random distribution as a background control. We found that the observed proportion of downregulated 3' UTR target genes in reality was significantly higher than expected by chance (Fig. 6B, $P < 0.001$). Similarly, we found that 12.9% of the mutated CDS targets were downregulated (Fig. 6C), but this observed proportion of downregulated CDS target genes did not distinguish from random controls (Fig. 6D, $P = 0.0925$). In addition, the proportions of upregulated and nondifferentially expressed genes in 3' UTRs were not significant (Supplementary Figs. S9A–S9C), which supports the conclusion that mutated 3' UTR target genes are significantly downregulated. This result also matches with the different influences of driver mutations in 3' UTR and CDS regions on target gene expression through disruption of binding sites. It indicates that 3' UTR driver mutations are more likely to be involved in miRNA regulated pathways than CDS driver mutations.

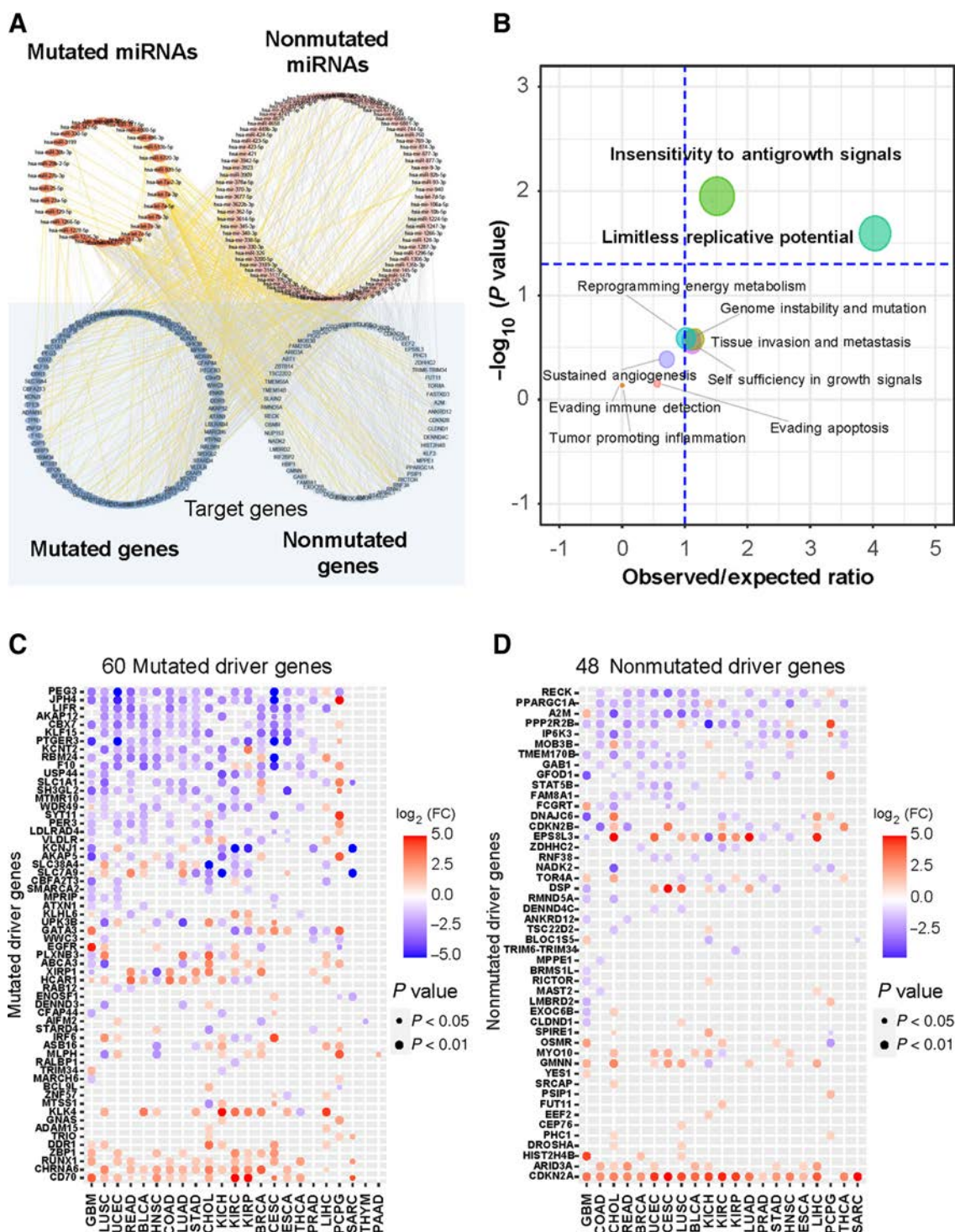


Figure 5.

Network perturbation by driver mutations. **A**, Gene regulatory networks perturbed by driver mutations in 3' UTR, CDS, and 5' UTR mGIs. Yellow edges, perturbed mGIs; gray edges, unperturbed mGIs. Red nodes, mutated miRNAs; orange nodes, mutated target genes; cyan nodes, nonmutated genes. **B**, Cancer hallmark enrichment of driver target genes in 3' UTR, CDS, and 5' UTR mGIs. Hallmarks with OR (x -axis) > 1 and P value (y -axis) < 0.05 are significantly enriched. **C** and **D**, Differential expression of mutated driver target genes (**C**) and nonmutated driver genes (**D**) from 3' UTR, CDS, and 5' UTR mGIs. Each row represents a driver gene that is differentially expressed in at least one cancer type. Each column represents a cancer type. The occurrence of each point indicates the gene is significantly differentially expressed in that cancer, which means the fold change of expression is larger than 2 or smaller than 1/2 and the P value calculated by DESeq2 is smaller than 0.05. Red indicates the gene is upregulated in cancer samples versus normal samples, whereas blue indicates downregulation.

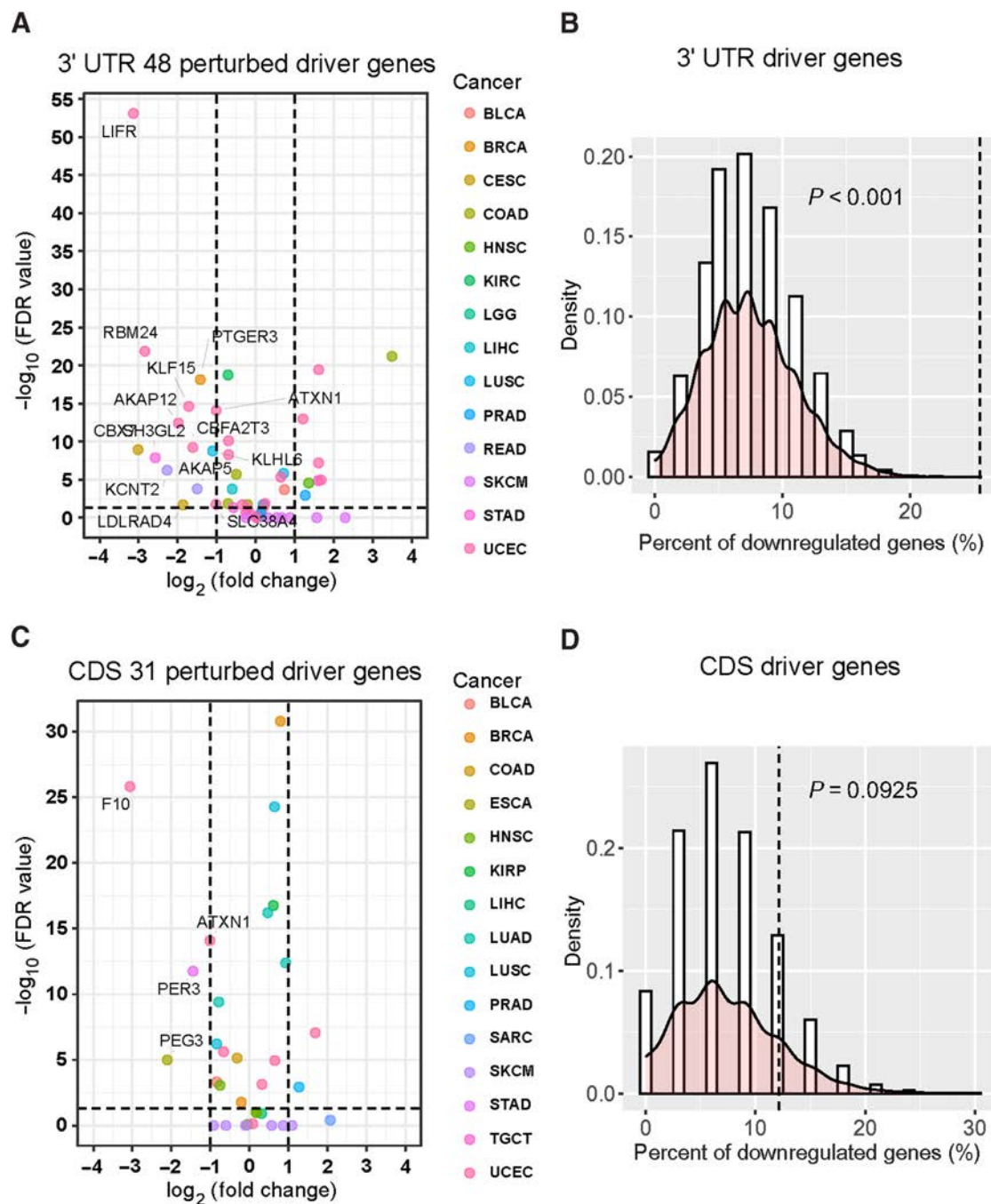


Figure 6.

Tumor suppressors are often targeted in perturbed miRNA regulatory networks. **A** and **C**, Volcano plots of 48 mutated 3' UTR target driver genes (**A**) and 31 mutated CDS target driver genes (**C**) in the cancer type where the driver mutation occurs. X-axis shows the $\log_2$ value of expression fold change and y-axis shows the value of $-\log_{10}$ (FDR). Each point indicates the gene in the cancer type where the driver mutation is found. The color of each point indicates the cancer type. **B**, Mutated driver genes in 3' UTR mGIs are significantly downregulated. Bar plot shows the random distribution of the proportion of downregulated genes in mutated driver genes. The smoothed curve is the probability density of random distribution. The dashed vertical line indicates the observed proportion of downregulated genes. **D**, Mutated driver genes in CDS mGIs are not significantly downregulated ($P > 0.05$). Legend is the same as **B**.

By curation from literature, we found that most of the down-regulated mutated 3' UTR target genes were tumor suppressor genes (Supplementary Fig. S9D). This finding is corroborated by prior analyses that found a similar pattern (9). The top-mutated target

genes were all cancer related and mostly annotated as inhibiting cancer progression. We further extracted the perturbed networks of these tumor suppressors (Supplementary Fig. S9E). Upstream miRNAs (red nodes) and their interactions with these tumor suppressors were

perturbed by driver mutations. Although the other target genes of these upstream miRNAs were unperturbed, no significant functional pathways were enriched for unperturbed targets.

In the perturbed tumor suppressor networks, all the target genes involved and most of the miRNAs had previous evidence to be associated with cancer. For example, *LIFR* could confer a dormancy phenotype in breast cancer cells and also negatively regulate the metastasis of pancreatic cancer cells (55, 56). *RBM24* was demonstrated to suppress cancer progression in nasopharyngeal carcinoma and also reported as a novel player in the p53 pathway (57, 58). *KLF15* could inhibit the cell proliferation in gastric cancer (59). *AKAP12* was reported as a tumor suppressor in glioblastoma and prostate cancer, and regulated by miRNA pathways (60, 61). *CBX7* was reported as a tumor suppressor in colon, thyroid carcinomas, glioblastoma, and pancreatic cancer (60, 62, 63). It could play its role by modulating the *KRAS* gene and could be regulated by miRNAs (60, 62). For the upstream miRNAs of these tumor suppressors, hsa-miR-6881-3p was known to play its role through the p53-mediated ceRNAs network in hepatocellular carcinoma (64). hsa-miR-15b-5p was used to distinguish human ovarian cancer tissues from normal tissues (65). hsa-miR-193b-3p was found differentially expressed in hepatocellular carcinoma (66). hsa-miR-15a-5p was shown to decrease melanoma cell viability and could cause cell-cycle arrest at the G₀-G₁ phase (67). hsa-miR-769-3p was differentially expressed in patients with non-small cell lung cancer (NSCLC) harboring EGFR mutations (68). In conclusion, the mutated driver target genes perturbed through miRNA pathways play their roles probably as tumor suppressors and are downregulated in many cancer types thus accelerating the growth of cancer cells.

mGI-map: a web-based resource for mutation-mediated mGI network perturbations in cancer

To help researchers apply the principles described in this work to any mGIs or mutations of interest, we developed a comprehensive and interactive web resource, mGI-map (<https://mgimap.dellmed.utexas.edu>). The features provided in the resource should serve as a guide for biologists interested in identifying miRNA-gene regulation in specific cancer types and understanding the consequences of mutations that perturb specific mGIs in cancer. All the data could also be downloaded for further analysis of the functional effect of mutations and miRNA-gene regulation (Supplementary Fig. S10).

Discussion

In this study, we derived *de novo* cancer-specific mGIs with high quality by integration of *in silico* predictive models, AGO CLIP-seq data, and transcriptome networks. We built out these mGI networks of various cancer types as a framework for pan-cancer analysis of the functional consequences of somatic mutations on mGIs. Further analysis showed that the mutations on mGIs exhibited a mutually exclusive pattern. Mutual exclusivity was previously observed for mutations in gene members of the same pathways. Yeang and colleagues found that when a member of a pathway was altered, the selection pressure on the other members was diminished or even nullified (69). As a result, significantly less overlap in mutations of the other genes was expected, creating a mutually exclusive pattern between their alterations. Supporting this expectation, it was previously shown that some functionally related genes were mutated in a mutually exclusive manner in thyroid tumors (70) and in leukemia (71). In our study, we observed this pattern on mGIs in most of the cancer types.

To identify driver mutations that can perturb the mGIs, we compared the samples with mutations on mGIs with the fitted expression distribution in nonmutated samples and evaluated the deviations of each candidate mutations. After identification of driver mutations, we analyzed the alteration of MFE and target gene expression between driver mutant and wild-type mGIs. We found that the target gene expression was associated with the alteration of MFE, indicating that the driver mutations could perturb mGIs by influencing mGI binding sites.

The data and design of our study produce several limitations. Because AGO-CLIP data are derived from cell lines rather than live tumor samples, some mGIs may have been lost given the high context dependency of the interactions (72). In addition, our analysis grouped together the many miRNA isomers (isomiRs) present in cancer as their source gene. This is significant due to prior evidence pointing to the differential involvement of isomiRs in cancer (73). Although this study offered increased statistical confidence in the identified mGIs, some resolution in the specificity of different isomiRs involved in the mGIs was lost. Telonis and colleagues found an abundant of isomiRs were also expressed in tissues and they were nearly 10 times higher than the reference sequences reported in miRBase from the analysis on TCGA dataset (73). In addition, around half of the most abundant isomiRs in various cancer types are not the reference miRNAs in the miRBase. The isomiR world makes the mGI regulation much more complicated and elaborated. Although we analyzed all the reference miRNAs in miRBase, we may still have missed ~9/10 miRNA targets due to the fact of isomiRs. Furthermore, the elaborated forms of isomiRs raise more challenges in their target sites identification. The traditional CLIP protocols include an RNase digestion step and different RNase strengths would trim the endpoints of miRNA-target pair differently (74). Therefore the digestion step could affect the miRNA member of the pair at both 5' and 3' ends, which prevents the accurate identification of isomiR target sites. We do believe there was still an opening gap in the miRNA regulation world when considering the isomiRs, and we realize that has been exceeding our capacity in this research. These areas will offer key directions for further study from our presented mGIs on mGI-map.

The genetic changes that contribute to cancer tend to affect three main types of genes—proto-oncogenes, tumor suppressor genes, and DNA repair genes, which are called “drivers” of cancer. In our analysis, we have shown that miRNA-related driver mutations tend to play their roles on tumor suppressor genes by modulating the mGIs. Tumor suppressor genes are involved in controlling cell growth and division. Cells with certain alterations in tumor suppressor genes may divide in an uncontrolled manner. Concordantly, our analyses show that the identified target genes in the perturbed miRNA regulatory networks are enriched in the cancer hallmarks of “insensitivity to antigrowth signals” and “self-sufficiency in growth signals,” which are consistent with the function of tumor suppressors. In this study, although several known tumor suppressors have been revealed and perturbed by miRNA related driver mutations, such as *LIFR*, *RBM24*, *PTGER3*, several identified novel mutated driver genes such as *LDLRAD4*, *AKAP5*, *KLHL6*, and so on, could become potential biomarkers and therapeutic targets for tumor suppression.

Authors' Disclosures

J. Das reports grants from NIAID, NHGRI, and DoD during the conduct of the study and personal fees from Seromyx outside the submitted work. G.B. Mills reports grants, personal fees, nonfinancial support, and other support from Amphista, AstraZeneca, Chrysalis Biotechnology, GSK, ImmunoMET, Ionis, Lilly, PDX Pharmaceuticals, Signalchem Lifesciences, Symphogen, Tarveda, Turbine, Zentalis

Pharmaceuticals, Catena, HRD assay to Myriad Genetics DSP to Nanostring, Adelson Medical Research Foundation, Breast Cancer Research Foundation, Komen Research Foundation, Ovarian Cancer Research Foundation, Prospect Creek Foundation, Nanostring Center of Excellence, Ionis (Provision of tool compounds), and Genentech during the conduct of the study. No disclosures were reported by the other authors.

Authors' Contributions

X. Hua: Resources, data curation, formal analysis, validation, investigation, visualization, methodology, writing—original draft. **Y. Li:** Data curation, formal analysis, validation, visualization, methodology, writing—original draft. **S.R. Pentaparthi:** Resources, data curation, formal analysis, validation, investigation, writing—original draft. **D.J. McGrail:** Data curation, validation, writing—review and editing. **R. Zou:** Data curation, validation, investigation, visualization, writing—review and editing. **L. Guo:** Data curation, validation, investigation, writing—review and editing. **A. Shrawat:** Data curation, validation. **K.M. Cirillo:** Data curation, validation. **Q. Li:** Data curation, validation. **A.V. Bhat:** Resources, data curation. **M. Xu:** Resources, writing—review and editing. **D. Qi:** Resources, data curation, writing—review and editing. **A. Singh:** Resources, data curation, validation, investigation, visualization, writing—review and editing. **F. McGrath:** Resources. **S. Andrews:** Resources. **K.L. Aung:** Writing—review and editing. **J. Das:** Conceptualization, resources, data curation, formal analysis, supervision, funding acquisition, validation, investigation, methodology, writing—original draft, project administration, writing—review and editing. **Y. Zhou:** Data curation, writing—review and editing. **A. Lodi:** Writing—review and editing. **G.B. Mills:** Writing—review and editing. **S.G. Eckhardt:** Writing—review and editing. **M.L. Mendillo:** Writing—review and editing. **S. Tiziani:** Writing—review and editing. **E. Wu:** Conceptualization, resources, data curation, supervision, funding acquisition, validation, investigation, methodology, writing—original draft, project administration, writing—review and editing. **J.H. Huang:** Resources, data curation, writing—review and editing. **N. Sahni:** Conceptualization, resources, supervision, funding acquisition, investigation, writing—original draft, project administration, writing—review and editing. **S.S. Yi:** Conceptualization, resources, data curation, formal analysis, supervision, funding acquisition, validation, investigation, methodology, writing—original draft, project administration.

References

- McGranahan N, Favero F, de Bruin EC, Birkbak NJ, Szallasi Z, Swanton C. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci Transl Med* 2015;7:283ra54.
- Kumar V, Chaudhary AK, Coulter DW, McGuire T, Mahato RI. Impact of miRNA-mRNA profiling and their correlation on medulloblastoma tumorigenesis. *Molecular Therapy-Nucleic Acids* 2018;12:490–503.
- Muller S, Bley N, Glass M, Busch B, Rousseau V, Misiak D, et al. IGF2BP1 enhances an aggressive tumor cell phenotype by impairing miRNA-directed downregulation of oncogenic factors. *Nucleic Acids Res* 2018;46:6285–303.
- Liu CC, Rennie WA, Carmack CS, Kanoria S, Cheng JJ, Lu J, et al. Effects of genetic variations on microRNA: target interactions. *Nucleic Acids Res* 2014;42:9543–52.
- Lopes-Ramos CM, Barros BP, Koyama FC, Carpinetti PA, Pezuk J, Doimo NTS, et al. E2F1 somatic mutation within miRNA target site impairs gene regulation in colorectal cancer. *PLoS One* 2017;12:e0181153.
- Grigelioniene G, Suzuki HI, Taylan F, Mirzamohammadi F, Borochowitz ZU, Ayturk UM, et al. Gain-of-function mutation of microRNA-140 in human skeletal dysplasia. *Nat Med* 2019;25:583–90.
- Bhattacharya A, Cui Y. SomamiR 2.0: a database of cancer somatic mutations altering microRNA-ceRNA interactions. *Nucleic Acids Res* 2016;44:D1005–10.
- Bhattacharya A, Ziebarth JD, Cui Y. PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. *Nucleic Acids Res* 2014;42:D86–91.
- Dhawan A, Scott JG, Harris AL, Buffa FM. Pan-cancer characterisation of microRNA across cancer hallmarks reveals microRNA-mediated downregulation of tumour suppressors. *Nat Commun* 2018;9:5228.
- Plotnikova O, Baranova A, Skoblov M. Comprehensive analysis of human microRNA-mRNA interactome. *Front Genet* 2019;10:933.
- Lal A, Kim HH, Abdelmohsen K, Kuwano Y, Pullmann R Jr., Srikantan S, et al. p16(INK4a) translation suppressed by miR-24. *PLoS One* 2008;3:e1864.
- Duursma AM, Kedde M, Schrier M, le Sage C, Agami R. miR-148 targets human DNMT3b protein coding region. *RNA* 2008;14:872–7.
- Tay Y, Zhang J, Thomson AM, Lim B, Rigoutsos I. MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature* 2008;455:1124–8.
- Shen WF, Hu YL, Uttarwar L, Passegue E, Largman C. MicroRNA-126 regulates HOXA9 by binding to the homeobox. *Mol Cell Biol* 2008;28:4609–19.
- Tsai NP, Lin YL, Wei LN. MicroRNA mir-346 targets the 5'-untranslated region of receptor-interacting protein 140 (RIP140) mRNA and up-regulates its protein expression. *Biochem J* 2009;424:411–8.
- Lee I, Ajay SS, Yook JI, Kim HS, Hong SH, Kim NH, et al. New class of microRNA targets containing simultaneous 5'-UTR and 3'-UTR interaction sites. *Genome Res* 2009;19:1175–83.
- Maqbool R, Ismail R, Hussain M. Mutations in MicroRNA genes and their binding sites are infrequently associated with human colorectal cancer in the kashmiri population. *Microna* 2014;2:219–24.
- Supek F, Minana B, Valcarcel J, Gabaldon T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 2014;156:1324–35.
- Shabalina SA, Spiridonov NA, Kashina A. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic Acids Res* 2013;41:2073–94.
- Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet* 2011;12:683–91.
- Stegeman S, Amankwah E, Klein K, O'Mara TA, Kim D, Lin HY, et al. A large-scale analysis of genetic variants within putative miRNA binding sites in prostate cancer. *Cancer Discov* 2015;5:368–79.
- Grey F, Tirabassi R, Meyers H, Wu GM, McWeeney S, Hook L, et al. A viral microRNA down-regulates multiple cell cycle genes through mRNA 5' UTRs. *PLoS Pathog* 2010;6:e1000967.
- Zhou HL, Rigoutsos I. MiR-103a-3p targets the 5' UTR of GPRC5A in pancreatic cells. *RNA* 2014;20:1431–9.
- Brest P, Lapaquette P, Souidi M, Lebrigand K, Cesaro A, Vouret-Craviari V, et al. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nat Genet* 2011;43:242–U24.

Acknowledgments

The authors are grateful to contributions from TCGA Research Network Analysis Working Group and also acknowledge the High-Performance Computing Cluster Facility at MD Anderson Cancer Center, the Biomedical Research Computing Facility at UT Austin, and Texas Advanced Computing Center (TACC) for assistance. This work was supported by the NIH grant R35GM133658 (to S. Yi), Komen Foundation grant CCR19609287 (to S. Yi), Early Career Investigator Grant by Ovarian Cancer Research Alliance Grant# 649968 (to N. Sahni), Department of Defense W81XWH-18-PRCRP-CDA CA181455 (to N. Sahni), NIGMS grant GM137836 (to N. Sahni), Cancer Prevention and Research Institute of Texas grant RR160093 (to S.G. Eckhardt), and NCI grants P50CA217685 and U01CA217842 (to G.B. Mills). S. Yi was also supported by 2021–2022 and 2022–2023 program in Oncological Data and Computational Sciences sponsored by The Joint Center for Computational Oncology between the Oden Institute, MD Anderson, and Texas Advanced Computing Center (TACC). G.B. Mills also received a kind gift from the Miriam and Sheldon Adelson Medical Research Foundation and Breast Cancer Research Foundation. N. Sahni is a CPRIT Scholar in Cancer Research with funding from the Cancer Prevention and Research Institute of Texas (CPRIT) New Investigator Grant RR160021. D.J. McGrail was supported by NCI grant K99CA240689. The authors would also like to acknowledge the support from BioRender in figure generation.

The publication costs of this article were defrayed in part by the payment of publication fees. Therefore, and solely to indicate this fact, this article is hereby marked “advertisement” in accordance with 18 USC section 1734.

Note

Supplementary data for this article are available at Cancer Research Online (<http://cancerres.aacrjournals.org/>).

Received February 3, 2020; revised December 15, 2020; accepted October 14, 2022; published first October 20, 2022.

25. Shah SP, Morin RD, Khattri J, Prentice L, Pugh T, Burleigh A, et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* 2009;461:809–U67.
26. Gartner JJ, Parker SCJ, Prickett TD, Dutton-Regester K, Stitzel ML, Lin JC, et al. Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. *Proc Nat Acad Sci U S A* 2013;110:13481–6.
27. Seiler M, Peng SY, Agrawal AA, Palacino J, Teng T, Zhu P, et al. Somatic mutational landscape of splicing factor genes and their functional consequences across 33 cancer types. *Cell Rep* 2018;23:282–96.
28. Biankin AV, Waddell N, Kassahn KS, Gingras MC, Muthuswamy LB, Johns AL, et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* 2012;491:399–405.
29. Moretti F, Thermann R, Hentze MW. Mechanism of translational regulation by miR-2 from sites in the 5' untranslated region or the open reading frame. *RNA* 2010;16:2493–502.
30. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 2015;4:e05005.
31. König J, Zarnack K, Luscombe NM, Ule J. Protein-RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet* 2012;13:77–83.
32. Li JH, Liu S, Zhou H, Qu LH, Yang JH. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* 2014;42:D92–7.
33. Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, et al. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* 2006;126:1203–17.
34. Kozomara A, G-JS miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 2014;42:D68–73.
35. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. *RNA* 2004;10:1507–17.
36. Xiao FF, Zuo ZX, Cai GS, Kang SL, Gao XL, Li TB. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res* 2009;37: D105–D110.
37. Chou CH, Chang NW, Shrestha S, Hsu SD, Lin YL, Lee WH, et al. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res* 2016;44:D239–D47.
38. Bandyopadhyay S, Mitra R. TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics* 2009;25:2625–31.
39. Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, Hatzigeorgiou AG. The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res* 2009;37:D155–D8.
40. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–2.
41. Kruger J, Rehmsmeier M. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res* 2006;34:W451–W4.
42. Lopes-Ramos CM, Chen CY, Kuijjer ML, Paulson JN, Sonawane AR, Fagny M, et al. Sex differences in gene expression and regulatory networks across 29 human tissues. *Cell Rep* 2020;31:107795.
43. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;46:310–5.
44. Watson IR, Takahashi K, Futreal PA, Chin L. Emerging patterns of somatic mutations in cancer. *Nat Rev Genet* 2013;14:703–18.
45. Babur O, Gonen M, Aksoy BA, Schultz N, Ciriello G, Sander C, et al. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biol* 2015;16:45.
46. Thomas RK, Baker AC, Debiassi RM, Winckler W, Laframboise T, Lin WM, et al. High-throughput oncogene mutation profiling in human cancer. *Nat Genet* 2007;39:347–51.
47. Liu G, Zhang R, Xu J, Wu CI, Lu X. Functional conservation of both CDS- and 3'-UTR-located microRNA binding sites between species. *Mol Biol Evol* 2015;32: 623–8.
48. Hausser J, Syed AP, Bilen B, Zavolan M. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res* 2013;23: 604–15.
49. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 2010;141:129–41.
50. Fang Z, Rajewsky N. The impact of miRNA target sites in coding sequences and in 3'UTRs. *PLoS One* 2011;6:e18067.
51. Wang T, Wang G, Hao D, Liu X, Wang D, Ning N, et al. Aberrant regulation of the LIN28A/LIN28B and let-7 loop in human malignant tumors and its effects on the hallmarks of cancer. *Mol Cancer* 2015;14:125.
52. Takamizawa J, Konishi H, Yanagisawa K, Tomida S, Osada H, Endoh H, et al. Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res* 2004;64:3753–6.
53. Ji J, Zhao L, Budhu A, Forgues M, Jia HL, Qin LX, et al. Let-7g targets collagen type I alpha2 and inhibits cell migration in hepatocellular carcinoma. *J Hepatol* 2010;52:690–7.
54. Hu X, Guo J, Zheng L, Li C, Zheng TM, Tanyi JL, et al. The heterochronic microRNA let-7 inhibits cell motility by regulating the genes in the actin cytoskeleton pathway in breast cancer. *Mol Cancer Res* 2013;11:240–50.
55. Johnson RW, Finger EC, Olcina MM, Vilalta M, Aguilera T, Miao Y, et al. Induction of LIFR confers a dormancy phenotype in breast cancer cells disseminated to the bone marrow. *Nat Cell Biol* 2016;18:1078–89.
56. Ma D, Jing X, Shen B, Liu X, Cheng X, Wang B, et al. Leukemia inhibitory factor receptor negatively regulates the metastasis of pancreatic cancer cells in vitro and in vivo. *Oncol Rep* 2016;36:827–36.
57. Hua WF, Zhong Q, Xia TL, Chen Q, Zhang MY, Zhou AJ, et al. RBM24 suppresses cancer progression by upregulating miR-25 to target MALAT1 in nasopharyngeal carcinoma. *Cell Death Dis* 2016;7:e2352.
58. Jiang Y, Zhang M, Qian Y, Xu E, Zhang J, Chen X. Rbm24, an RNA-binding protein and a target of p53, regulates p21 expression via mRNA stability. *J Biol Chem* 2014;289:3164–75.
59. Sun C, Ma P, Wang Y, Liu W, Chen Q, Pan Y, et al. KLF15 inhibits cell proliferation in gastric cancer cells via up-regulating CDKN1A/p21 and CDKN1C/p57 expression. *Dig Dis Sci* 2017;62:1518–26.
60. Ko HK, Akakura S, Peresie J, Goodrich DW, Foster BA, Gelman IH. A transgenic mouse model for early prostate metastasis to lymph nodes. *Cancer Res* 2014;74: 945–53.
61. Peixoto P, Blomme A, Costanza B, Ronca R, Rezzola S, Palacios AP, et al. HDAC7 inhibition resets STAT3 tumorigenic activity in human glioblastoma independently of EGFR and PTEN: new opportunities for selected targeted therapies. *Oncogene* 2016;35:4481–94.
62. Wu W, Zhou X, Yu T, Bao Z, Zhi T, Jiang K, et al. The malignancy of miR-18a in human glioblastoma via directly targeting CBX7. *Am J Cancer Res* 2017;7:64–76.
63. Ni S, Wang H, Zhu X, Wan C, Xu J, Lu C, et al. CBX7 suppresses cell proliferation, migration, and invasion through the inhibition of PTEN/Akt signaling in pancreatic cancer. *Oncotarget* 2017;8:8010–21.
64. Zhang Y, Kang R, Liu W, Yang Y, Ding R, Huang Q, et al. Identification and analysis of P53-mediated competing endogenous RNA network in human hepatocellular carcinoma. *Int J Biol Sci* 2017;13:1213–21.
65. Wang L, Zhu MJ, Ren AM, Wu HF, Han WM, Tan RY, et al. A ten-microRNA signature identified from a genome-wide microRNA expression profiling in human epithelial ovarian cancer. *PLoS One* 2014;9:e96472.
66. Jiang Y, Gao H, Liu M, Mao Q. Sorting and biological characteristics analysis for side population cells in human primary hepatocellular carcinoma. *Am J Cancer Res* 2016;6:1890–905.
67. Alderman C, Yang Y. The anti-melanoma activity and oncogenic targets of hsa-miR-15a-5p. *RNA Dis* 2016;3:e1450.
68. Ma Y, Pan X, Xu P, Mi Y, Wang W, Wu X, et al. Plasma microRNA alterations between EGFR-activating mutational NSCLC patients with and without primary resistance to TKI. *Oncotarget* 2017;8:88529–36.
69. Yeang CH, McCormick F, Levine A. Combinatorial patterns of somatic gene mutations in cancer. *FASEB J* 2008;22:2605–22.
70. Xing M, Cohen Y, Mambo E, Tallini G, Udelsman R, Ladenson PW, et al. Early occurrence of RASSF1A hypermethylation and its mutual exclusion with BRAF mutation in thyroid tumorigenesis. *Cancer Res* 2004;64:1664–8.
71. Carubaccia N, Trouplin V, Gelsi-Boyer V, Murati A, Rocquain J, Adelaide J, et al. Mutual exclusion of ASXL1 and NPM1 mutations in a series of acute myeloid leukemias. *Leukemia* 2010;24:469–73.
72. Ro S, Park C, Young D, Sanders KM, Yan W. Tissue-dependent paired expression of miRNAs. *Nucleic Acids Res* 2007;35:5944–53.
73. Telonis AG, Magee R, Lohrer P, Chervoneva I, Londin E, Rigoutsos I. Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. *Nucleic Acids Res* 2017;45:2973–85.
74. Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods* 2011;8:559–64.