

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

Yongsheng Li<sup>1,2,†</sup>, Brandon Burgman<sup>1,3,†</sup>, Ishaani S. Khatri<sup>1,2</sup>, Sairahul R. Pentaparthi<sup>1</sup>, Zhe Su<sup>1,2</sup>, Daniel J. McGrail<sup>4</sup>, Yang Li<sup>5</sup>, Erxi Wu<sup>1,6,7,8</sup>, S. Gail Eckhardt<sup>1,3</sup>, Nidhi Sahni<sup>5,9,10,\*</sup> and S. Stephen Yi<sup>1,2,3,11,\*</sup>

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

<sup>†</sup>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

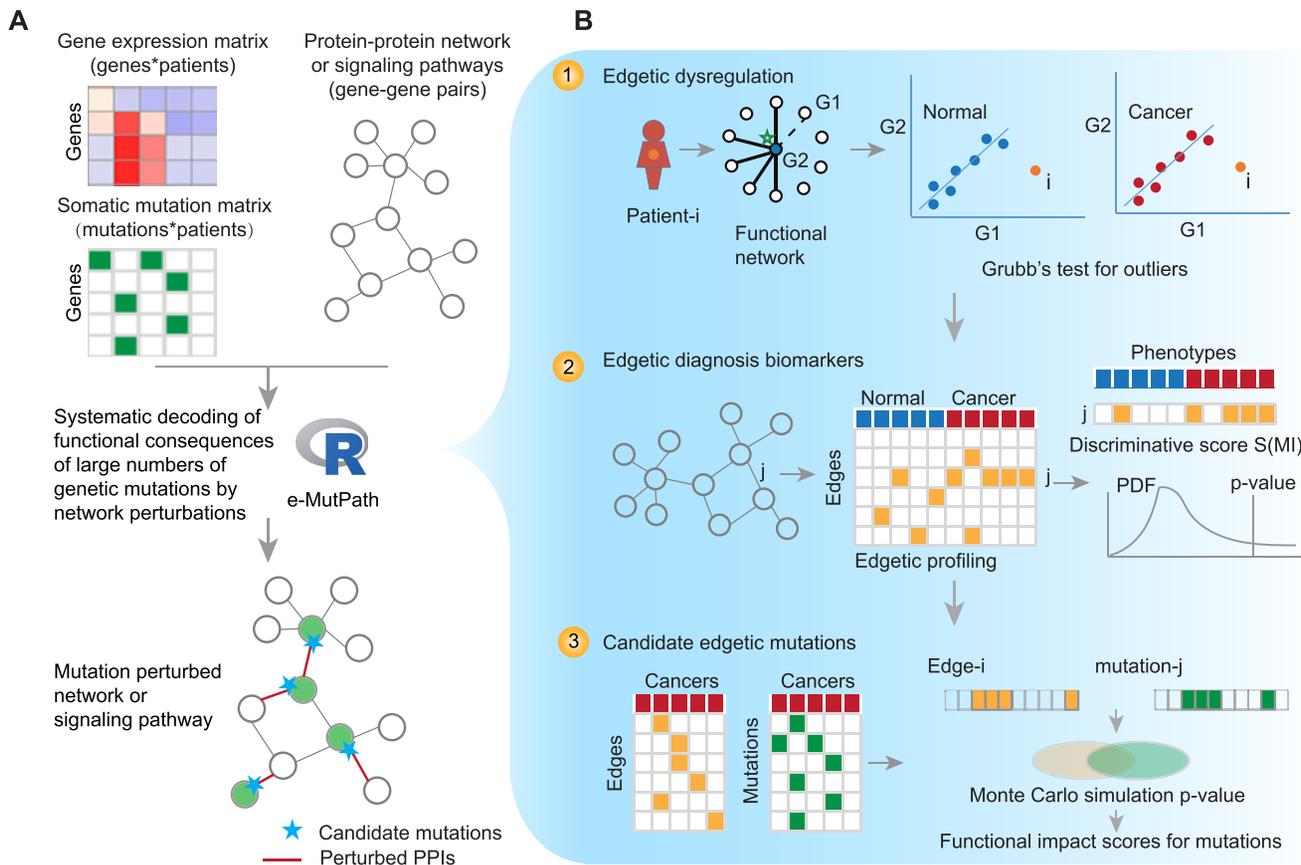
It has been demonstrated that perturbed molecular interactions can be identified by integration of protein or gene expression profile with molecular networks (25,26). However, there are limited protein expression profiles available to use in complex diseases (27,28). Here, we integrated genome-wide gene 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  $\mu$  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:

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

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  $N_{\text{rand}}$  is the number of patients with both the mutation and the specific perturbed interaction in random conditions,  $N_{\text{real}}$  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

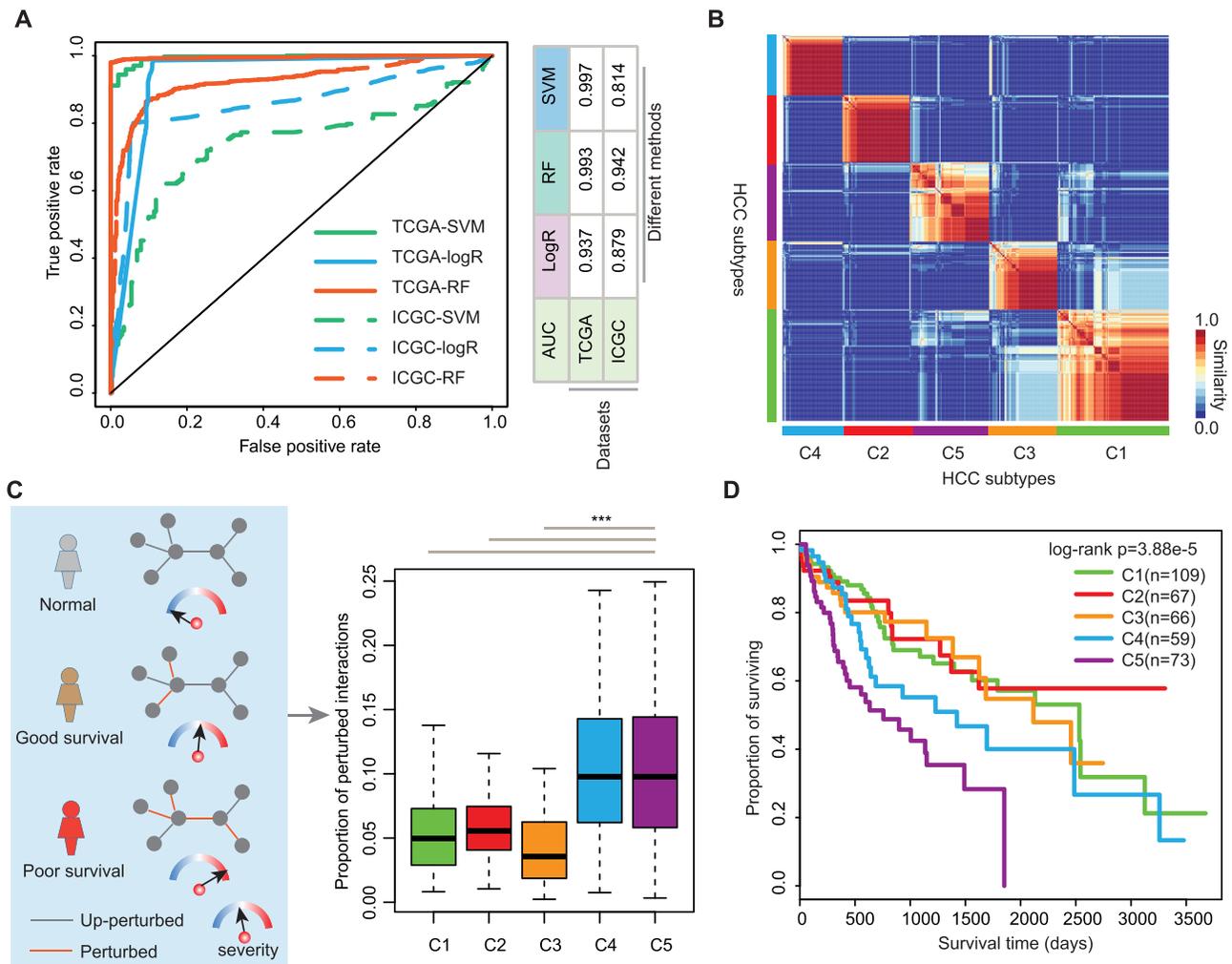
As a proof of principle, we applied the proposed method to HCC data obtained from TCGA project (18), which included 374 cancer and 50 normal samples. Based on the human PPI networks, we identified 477 interactions among 486 genes perturbed in HCC (Supplementary Figure S1A). We found that the majority 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) ( $P = 2.29E-5$ ) 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.

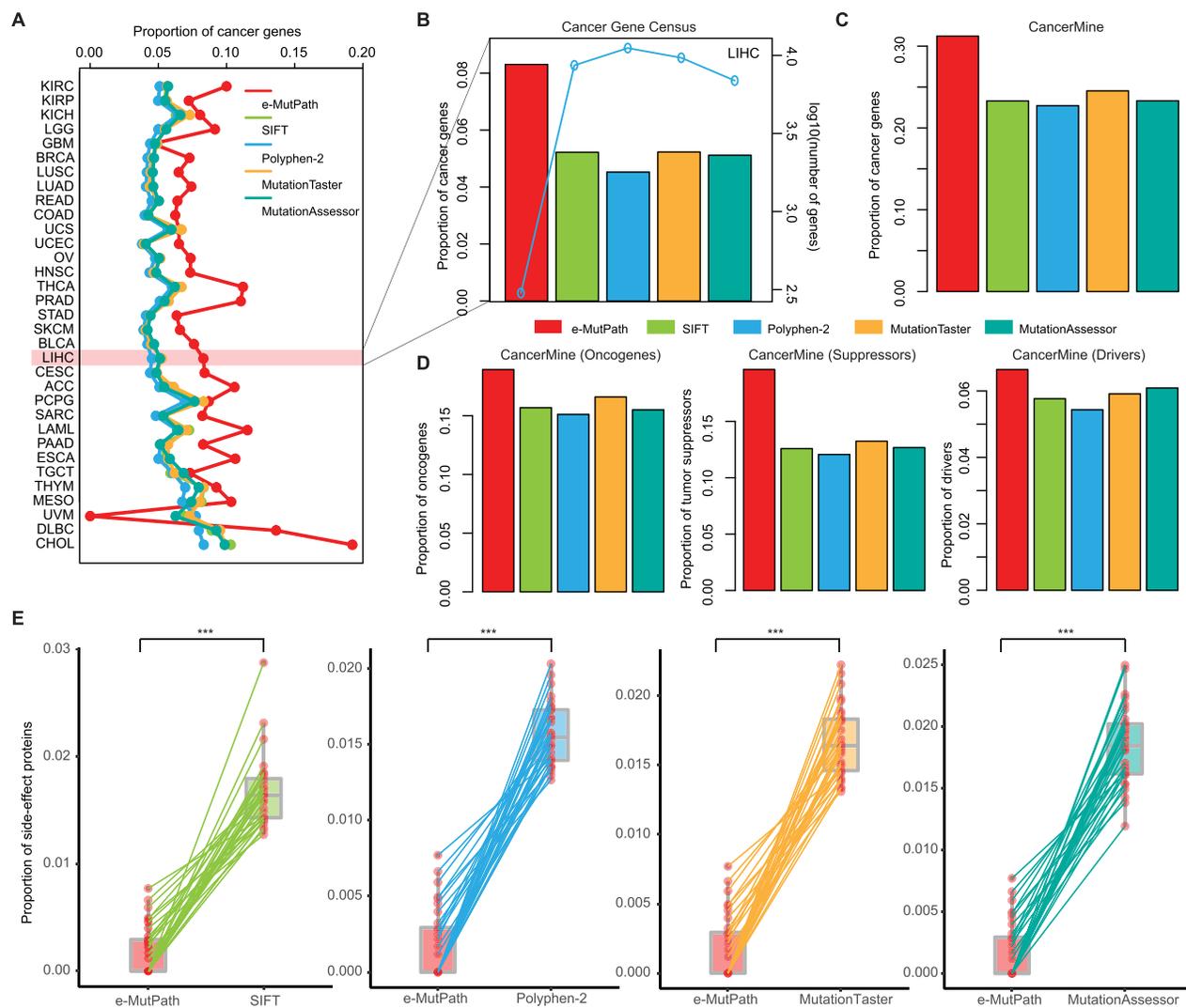
we next examined the survival rates of patients. We found indeed it was the case. Patients with mutations that perturbed a higher proportion of interactions tended to correlate with a significantly shorter survival time, indicating the deleterious nature of these mutations (Figure 2D, log-rank  $P = 3.88E-5$ ). 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.

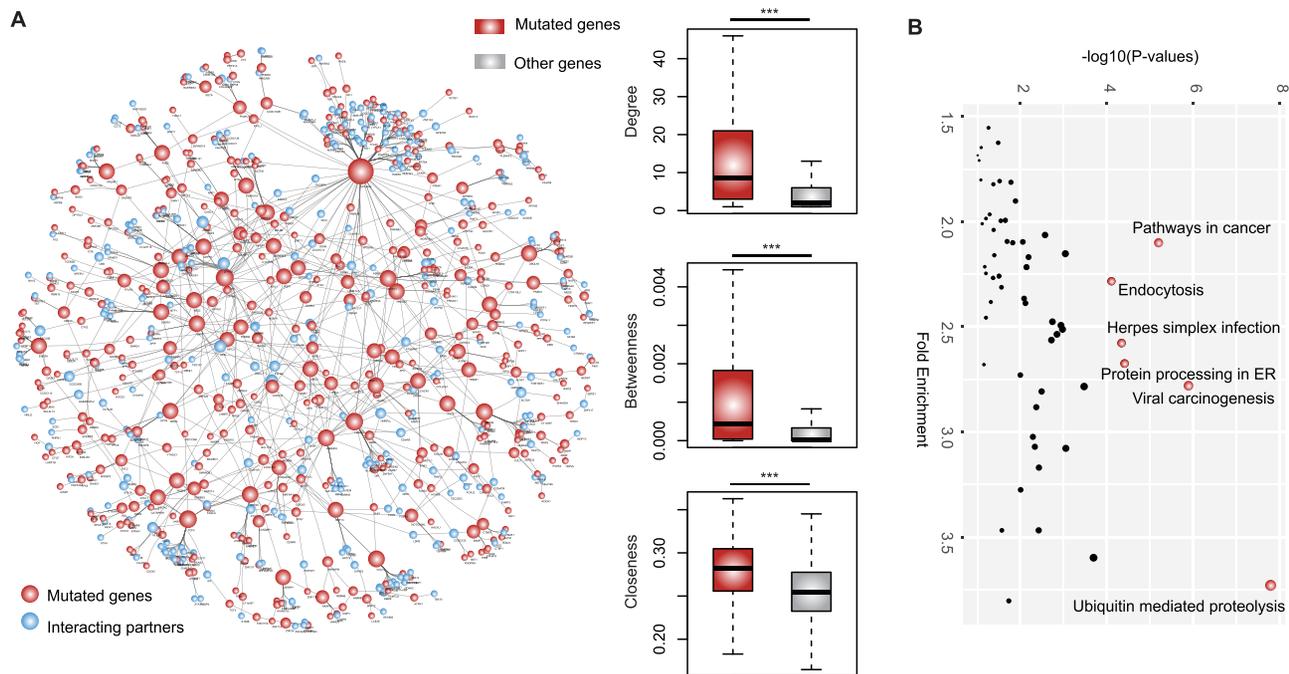
were more enriched for essential genes than other methods in the majority of cancer types (Supplementary Figure S4A). Particularly, the predicted genes by e-MutPath in LUSC, BLCA and LIHC showed significantly lower essentiality scores than predictions by other methods (Supplementary Figure S4B–G).

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.

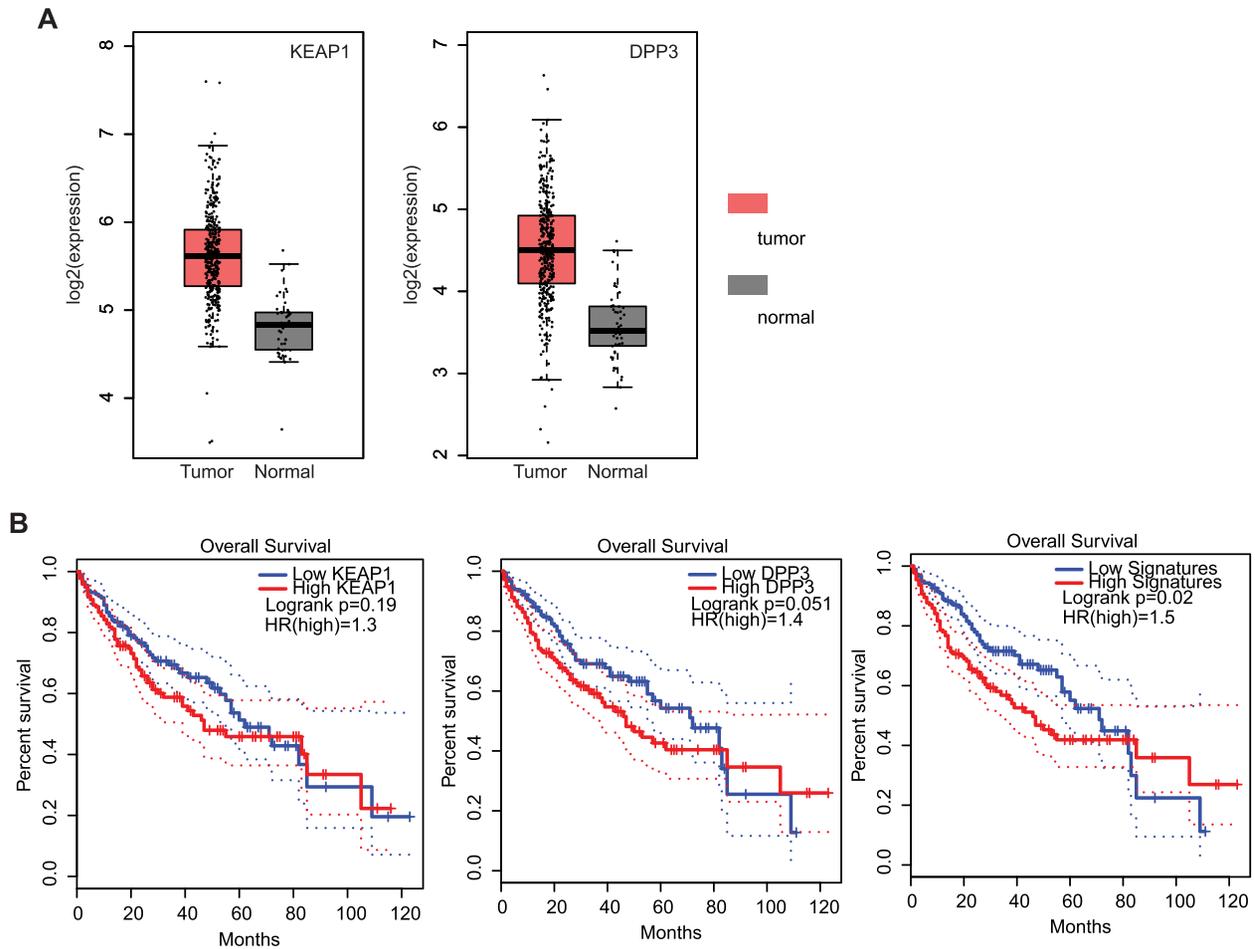
### Signaling pathway perturbations in cancer

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

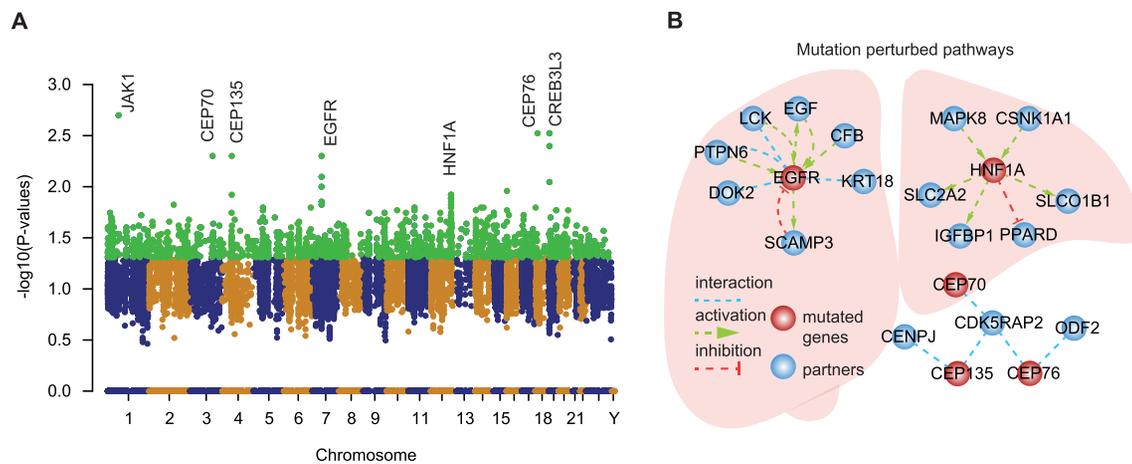
top ranked HCC mutations which were mostly enriched in centrosomal genes (CEP70, CEP76 and CEP135), JAK1, EGFR and HNF1A (Figure 6A). JAK1 encodes a cytoplasmic tyrosine kinase that is associated with a variety of cytokine receptors and plays a critical role in cell proliferation, survival and differentiation (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, PPAR and IGFBP1 (Figure 6B). The dysregulation of the HNF1A and PPAR 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., Weile, 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., Ideker, 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 Sibilica,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., Scaoazec,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., Ruppin,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.