Network-based cancer gene relationship prediction method reveals perturbations in the cancer gene network

Jiajun Qiu[1, 2*#], Kui Chen[3#], Chunlong Zhong[3], Sihao Zhu[4], Xiao Ma[5*]

1 Department of Informatics, I12-Chair of Bioinformatics and Computational Biology, Technical University of Munich (TUM), Boltzmannstrasse 3, 85748, Garching, Munich, Germany
2 TUM Graduate School, Center of Doctoral Studies in Informatics and Its Applications (CeDoSIA), 85748, Garching, Munich, Germany
3 Department of Neurosurgery, Shanghai East Hospital, Tongji University School of Medicine, Shanghai 200120, P.R. China
4 Department of Biomedical Engineering, College of Engineering, Peking University, Peking 100080, P.R. China
5 Group of Signal Transduction, Department of Psychiatry and Psychotherapy, University Hospital, LMU Munich, Nussbaumstrasse 7, 80336 Munich, Germany

# Authors contributed equally to this work.

* Corresponding author:
Jiajun Qiu: jiajunqiu@hotmail.com
Xiao Ma: Xiao.Ma@med.uni-muenchen.de; xiaoma1989@hotmail.com

Abstract

The landscape of the gene relationship/network (such as activation, expression, phosphorylation, and binding) in cancer is found different from the general (non-disease) situation, and gene network perturbations are supposed to be the main cause of cancer. Thus, it makes no sense to use a regular gene relationship prediction method to map the cancer gene network. Here, we established a novel prediction method that we dubbed network-based cancer gene relationship (NECARE), which achieved a high performance with a Matthews correlation coefficient (MCC) = 0.71±0.01 and an F1 = 89±0.7%. Then, we investigated the cancer interactome atlas and revealed a large-scale perturbation in the gene network in cancer using NECARE. We found 2287 genes, which were named cancer hub genes, that were enriched with gene interaction perturbations, and over 56% of cancer treatment-related genes were hub genes. We further assessed the association of hub genes with the prognosis of 32 types of cancers and found that hub genes were significantly related to the cancer outcomes. Furthermore, the mutations occurring on residues that bind to macromolecules were overrepresented at cancer hub genes. By coimmunoprecipitation (co-IP), we confirmed that the NECARE prediction method was highly reliable and was 90% accurate. NECARE is available at: https://github.com/JiajunQiu/NECARE.

Key words: network-based prediction; cancer gene network perturbation; cancer hub genes; R-GCN; deep learning

Introduction

Cells are biological systems that employ a large number of genes and signaling pathways to coordinate multiple functions [1]. Therefore, genes do not act in isolation but instead interact with each other and work as part of complex networks [2]. Recent studies have already found, instead of dysregulation of single genes, network perturbation is the cause of cancers [2]. Gene network in cancer is perturbed by many factors, one of which could be mutations. Disease-causing mutations not only produce a mutated gene and thus a mutated protein but also disturb the interactions between the mutated protein and its normal molecular partners [3]. Additionally, distinct mutations will cause different molecular defects in proteins, and they may lead to distinct perturbations of gene networks, giving rise to distinct phenotypic outcomes [4]. Nonsense mutations that grossly destabilize a protein structure can be modeled as removing a protein node from the network (Fig. 1A). Alternatively, missense mutations may give rise to partially functional gene products with specific changes in distinct biophysical or biochemical interactions (Fig. 1B) [4]. Furthermore, studies have already found that missense mutations in cancer are more likely to occur on the interaction interface of proteins. Thus, network perturbation, instead of single gene dysregulation, has been found to be the reason for human diseases, especially cancers [5]. For example, in cancer, TP53, a well-known tumor suppressor gene (Fig. 1C), loses many interactions with other important genes, such as PTEN and MDM2 [6]. However, new genes, such as CDK4, have been discovered to interact with TP53. In the normal network, the cross-talk line from TP53 to CDKN2A is TP53-MDM2-CDKN2A, but in cancer, the cross-talk line is TP53-CDK4-STK11-CDKN2A [7]. Therefore, in cancer, mutations lead to reconstruction of the gene network rather than the simple destruction, making the gene network in cancer tissues very different from that in normal tissues.

>>> Fig. 1 <<<

There have been some studies about cancer network perturbations [2,8-11]. For example, James West et al. tried to identify genes with network perturbations by calculating the network entropy [8]. Maxim Grechkin et al. also identified perturbed genes through inferred gene regulators and their expression [2]. As these studies were based on only the coexpression of genes, their network was more likely to reflect the relationships (expression and repression) between transcriptional factors and their targets. However, these studies failed to consider physical relationships such as protein-protein interactions (PPIs), which are

also important types of gene relationships and are significantly different from coexpression networks based on topological comparisons [12].

To date, gene relationship prediction methods fall into three categories: 1) **Structure-based methods**, which are based on the 3D structure of proteins and limited to proteins with PDB structures [13-16]. Structure-based methods are better at predicting physical interactions. 2) **Sequence-based prediction methods,** which attempt to predict interactions by the sequences of two candidate proteins/genes [17-20]. 3) **Network-based methods** that predict interactions based on the known network. In addition to the two input genes, network-based methods also consider their known neighbors [21-27].

In our study, to investigate the whole cancer gene relationship map, we first established a prediction method, dubbed **ne**twork-based **ca**ncer gene **re**lationship, NECARE, by applying a relational graph convolutional network (R-GCN) with knowledge-based features. One crucial novelty of this work is that it is the first specifically established method for predicting the overall gene relationships (activation, inhibition, expression, phosphorylation, binding and dissociation) in cancer. Another novel aspect of this methodology is that, unlike previous network-based node relationship prediction algorithms, NECARE considers the type and direction of gene links at the input space, and it can also distinguish the direction of the output (predicted gene interaction). These advantages show that NECARE outperforms the other algorithms (both network- and sequence-based algorithms) in predicting cancer gene relationships. Thus, our tool can help other researchers to determine the possible upstream and downstream molecular partners of their target genes in cancer.

Furthermore, we mapped the cancer interactome and analyzed the gene network perturbations in cancer with NECARE. We found that the network perturbations were enriched in some specific genes that were defined as cancer hub genes in our study. These hub genes were significantly related to the outcome of 32 types of cancers. Many of these hub genes have already been well studied in previous cancer studies or served as drug targets. For the crosstalk of signaling pathways, some pathways were found to cross-talk with many other pathways, some of which are well-known pathways in cancer studies, such as the VEGF and Notch signaling pathways. These results indicated that our results can potentially be the targets of future cancer studies.

Methods

General gene relationship data

To build a knowledge graph for our modified R-GCN, which considers both the type and direction of the links [28], we extracted the general gene relationship data from the following

three databases: 1) STRING [29], a famous database for known protein-protein associations, from which we extracted data about the experimental annotated human protein-protein associations; 2) Kyoto Encyclopedia of Genes and Genomes (KEGG) [6], a well-known publicly accessible pathway database, from which we extracted human non-disease pathway and gene relationship data; and 3) HIPPIE [30], which contains experimentally detected PPIs from IntAct [31], MINT [32], BioGRID [33], HPRD [34], DIP [35], BIND [36] and MIPS [37]. Overall, our general gene relationship data contained 551850 pairs of relationships with 13 different types of links (Table S1). The whole dataset is available from (github.com/JiajunQiu/NECARE/dataset/NECARE.graph).

Cancer gene relationship data

Cancer gene relationship data, which contain diverse relationship types (i.e., activation, inhibition, expression, phosphorylation, binding and dissociation) served as the training data for the R-GCN. Unlike the above general gene relationship data, we did not consider the type of interactions in the training data. Therefore, for each pair of genes (gene A and gene B) in our training data, there were only two binary problems: 1) whether the A->B interaction exists and 2) whether the B->A interaction exists.

We obtained cancer-related gene relationship data from the KEGG and Reactome databases, which served as the positive training set [6,38]. We also included the OncoPPI database, which is a cancer-specific PPI database, in our positive training set [7].

The negative training set comprised three parts: 1) the pairs of relationships with “disassociation/missing interaction” or other negative annotations in the database; 2) the relationship of the reverse direction of the unidirected relationship in the positive set; and 3) the Negatome database, a collection of protein pairs that are not likely to be engaged in direct physical interactions [39]. Here, we included the data from Negatome to balance the data in the model training process (Fig. S1). This was because the total number of negative relationships from the first two parts of the negative training set were still below the total number of the positive relationships (positive: 4358 vs negative: 2773). Therefore, data from Negatome was selected as additional negative data in the training and cross-training sets (Fig. S1) [39]. Note that the Negatome data were not used in the testing set (performance evaluation).

Overall, in the training data, most of the relationships were unidirected except “binding” and “dissociation/missing interaction” (e.g., if gene A dissociated from gene B, then both link A->B and link B->A would be set as negative). The whole dataset is available from (github.com/JiajunQiu/NECARE/dataset/NECARE_TrainingData.txt).

The 5-fold cross-validation

We applied a 5-fold cross-validation approach for the training mode (Fig. S1). Technically, we divided the training set into five parts. To avoid overestimating the performance, we forced the relationships from the same pair of genes to exist only in the same parts. In each rotation, we used three of the five parts for training, one for cross-training (optimize hyperparameters, including number of hidden units in neural network, early stop, etc.), and one for testing. Overall, we optimized the parameters in the cross-training set and tested the final performance in the testing set. The testing set was never used in the hyperparameter optimization and feature selection.

Due to the lack of enough data for independent tests, we also applied this cross-validation approach to other methods that were included in the performance comparison.

Modified relational graph convolutional networks

Graph convolutional networks (GCNs) can be understood as special cases of a simply differentiable message-passing framework. Information can be obtained from the neighbors of each node in the GCN. The R-GCN is an extension of the GCN [28]. It accounts for the edge type and direction and can calculate the forward-pass update of an entity or node denoted in relational (directed and labeled) multigraphs [28] (Fig. 1D).

$$h_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in N_i^r}\frac{1}{C_{i,r}}W_r^{(l)}h_j^{(l)} + W_O^{(l)}h_i^{(l)}\right) \qquad \text{Eqn. 1}$$

In Eqn. 1, if we define the directed and labeled multigraphs as G= $(\mathcal{V},\mathcal{E},\mathcal{R})$ with the nodes defined as $v_i \in \mathcal{V}$, labeled edges as $(v_i,r,v_j) \in \mathcal{E}$, and edge type as $r \in \mathcal{R}$, then $h_i^{(l)}$ is the hidden state of node v_i in the i-th layer of the neural network. N_i^r denotes the set of neighbor indices of node v_i under the relation $r \in \mathcal{R}$. $C_{i,r}$ is a normalization constant, which is defined as the degree of the target node of an edge. $W_r^{(l)}$ is a form of weight sharing among different relation types, and $W_O^{(l)}$ is a weight matrix for the linear message transformation. The incoming messages from neighbors are accumulated and then passed through an activation

function σ such as ReLU [28]. Therefore, in our study, instead of considering only the gene itself, information about each gene was obtained from other genes that contacted it.

In the original R-GCN, at the last step, the score was calculated by DistMult (the trilinear dot product) [28]. However, it cannot distinguish the direction in the final edge prediction. The A->B edge would have the same score as the B->A edge. Thus, here, we modified the final score calculation to:

$$Score = sigmoid(Embed_{source} W_{1,*} + W_{2,*} Embed_{target}) \quad \text{Eqn. 2}$$

W is a 2×length (Embed) weight matrix. With this formula, if the positions of the source and target node were reversed, the score would be totally different. Therefore, we can distinguish the direction of the link.

The feature we used was a combination of two parts. Part one was the OPA2Vec vector of each gene, which was a knowledge-based feature [40]. OPA2Vec is a tool that can be used to produce feature vectors for biological entities from ontology. OPA2Vec used mainly metadata from the ontology in the form of annotation properties as the main source of data. In this study, we used the OPA2Vec pretrained model based on PubMed data, and the annotation file was downloaded from http://purl.obolibrary.org/obo/go.owl. Part two was the cancer-specific feature based on The Cancer Genome Atlas (TCGA), including the expression profile of each gene in 32 different types of cancer and the mutation rate among patients for each type of cancer.

Performance evaluation

We evaluated the performance of the prediction via a variety of measures. For simplicity, we used the following standard annotations: true positives (TP) were the correctly predicted gene relationships in cancer, while false positives (FP) were the gene pairs that had no links in cancer and were incorrectly predicted to have interactions. True negatives (TN) were the

correctly predicted noninteractions, and false negatives (FN) were the gene pairs that had interactions but were not a correctly predicted.

$$\text{Precision} = TP/(TP+FP);\ \text{Sensitivity (Recall)} = TP/(TP+FN)$$

$$\text{Specificity} = TN/(TN+FN) \qquad \text{Eqn.}$$

$$F1(C) = 2*\text{Precision}*\text{Recall}/(\text{Precision}+\text{Recall})$$

We also calculated the Matthews correlation coefficient (MCC) and area under the curve (AUC):

$$MCC(C) = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad \text{Eqn. 4}$$

Error estimates

Error rates for the evaluation measures were estimated by bootstrapping (without replacement to render more conservative estimates), i.e., by resampling the set of samples used for the evaluation 1000 times and calculating the standard deviation of those 1000 different results. Each of these sample sets contained 50% of the original samples (picked randomly again, without replacement).

Pathway cross-talk

Pathway cross-talk was constructed following the steps [41]:

Remove pathways containing less than 5 genes or more than 200 genes. The reason is that pathways with too few genes may not have sufficient biological content, and pathways with too many genes might be too generic.

Fisher's exact test was applied to evaluate gene overlap between any given pair of pathways. Raw P-values were adjusted by the Benjamini-Hochberg (BH) procedure. Any pathway pair with an adjusted P-value<0.05 was removed due to significant overlap.

The number of interactions between any two pathways were counted. For each pathway pair, the interactions for pathway cross-talk were defined similarly to those in a previous

study: 1) two genes from two different pathways link directly to each other 2) or they link through a third gene.

The background interaction distribution of each pathway pair was calculated as follows. For all genes in a given pathway, we skipped the genes that did not have any interactions with the other pathways. Then, we counted the number of genes interacting with one gene under this criterion. Next, we randomized the pathway by replacing the original pathway gene with a randomly selected gene from our general gene relationship data with the same or similar interaction number. Once both pathways were randomized, a final step was performed to count the number of interactions between these pathways (we counted the number of interactions between them). This randomization step was repeated 1000 times.

One-sided Fisher's exact test on all pathway pairs was performed with a 2 × 2 contingency table that included the following numbers: n, N-n, r, R-r, where n was the number of interactions between the original pathways, N was the number of total interaction counts of all pathway pairs, r denoted the average number of interactions between the pair of corresponding randomized pathways after 1000 rounds of randomizations, and R denoted the average number of the total interactions of all randomized pathway pairs after 1000 rounds of randomizations.

Two overlapping pathways may interact with the same pathway. In this case, the two edges were considered redundant, and the edge that contained fewer genes was removed.

Cancer hub gene identification

Cancer hub genes were defined as those genes that lost (or gained) many links in the cancer network. We used the cancer gene links connecting with an equal likelihood at the genes in the network as a null model. We first changed the network to be undirected by the following two steps: 1) Every node (gene) in the network was assigned a different prime number, and a pair of links was represented by the product of two prime numbers of two nodes. 2) The links that had duplicated products were removed. Then, for each node in the network, we calculated how many links were connected to it.

We assumed that, for a particular gene (node) to be called a putative hub gene, more links (gained or lost) must connect to that gene than expected by chance if the links were randomly connected to the genes in the network. Randomly, the frequency of links connected to any particular residue followed a binomial distribution:

$$P(m = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{Eqn. 5}$$

where n is the 2x total number of links, k is the number of links connecting to a particular node, p is the probability of any individual link connecting at a particular node, and P (m=k) is precisely the probability of observed k links at a single node. Since our null model assumes an equal likelihood of links at any node, we used $p = 2/L$, where L is the overall number of unique nodes in the network.

Thus, to assign a probability to the observation of k links connecting at a particular node by chance (i.e., a P-value), we calculated the probability of at least k links connecting at a particular node from our null model:

$$P\ (m \geq k) = \sum_{i=k}^{n} \binom{n}{k} p^k (1-p)^{n-k} \quad \text{Eqn. 6}$$

To correct for and test multiple hypotheses, the p-values for all considered hub genes were adjusted using the Bonferroni correction method.

Eigenvector centrality was a measure of the influence of a node in a network. The regular eigenvector centrality of each gene in the network was the eigenvector of the adjacency matrix with the largest unique eigenvalue. Here, in our study, we applied a variant of eigenvector centrality [42]. The final centrality values followed the SoftMax probability: any node that you randomly picked up would reach a certain node in the network.

Clinically related cancer genes

Cancer genes related to clinical treatment were downloaded from the Tumor Alterations Relevant for GEnomics-driven Therapy (TARGET) database (https://software.broadinstitute.org/cancer/cga/target). TARGET (tumor alterations relevant for genomics-driven therapy) is a database of genes that, when somatically altered in cancer, are directly linked to a clinical action. TARGET genes is associated with response or resistance to a therapy, diagnosis, and/or prognosis.

Survival analysis of hub genes

To assess the association of hub genes with survival outcomes, we obtained the mutation and clinical outcome data of 32 different types of cancers from the TCGA (Table S2). For each cancer, we first calculated hazard ratios (HRs) and P-values (log-rank test) for each involved gene by Cox proportional hazards regression analysis using the coxph function of the R survival package (v. 2.37.2). Then, for each cancer, we integrated the hub genes with a significant P-value (cutoff: 0.05) into a combined mutation score (MS):

$$MS = \sum_{j=1}^{g} (w_j \times M_j) \qquad \text{Eqn. 7}$$

where M_j is whether gene j is mutated in the tumor sample of the patient (1 for mutated and 0 for nonmutated) and w_j is set to 1 or -1, depending on the HR of each gene (1 for HR≥1 and -1 for HR<1). The median value (50%) or the automatically selected best cutoff value of the MS was used to divide the corresponding patients into high- and low-MS groups for Kaplan–Meier analysis of their association with the 10-year survival.

Macromolecular binding mutation enrichment analysis

Cancer mutation data were downloaded from the Catalogue of Somatic Mutations in Cancer (COSMIC) database. Residues involved in protein-protein, DNA-protein and RNA-protein interactions were predicted by our new powerful high-throughput prediction method ProNA2020 [43]. Then, Fisher's exact test was performed with a 2 × 2 contingency table that included the following numbers: N, n, M, and m, where N is the number of binding mutations on hub genes, n denotes the number of nonbinding mutations on hub genes, M is the number of binding mutations on nonhub genes, and m denotes the number of nonbinding mutations on nonhub genes.

Results

Network-based cancer gene relationship (NECARE) prediction method

The gene network in cancer is different from that in normal (non-cancer) situations. To reveal gene network perturbation in cancer, we designed the first network-based cancer-specific gene relationship prediction method: NECARE. Basically, instead of looking at only the particular gene, NECARE also obtained the information of its neighboring genes.

NECARE is different from previous network-based algorithms: 1) at the input space, it accounts for the type and direction of edges (Fig. 1D); and 2) at the prediction output, it can also distinguish the direction of the predicted interaction.

In our study, we tested two kinds of features for the neural network: 1) ontology-based features (OPA2Vec) and 2) TCGA-based expression and mutation profiles. Their performance was compared in the cross-training set (Fig. S1). The combination of OPA2Vec and TCGA worked better than each of them alone, reaching an MCC = 0.75 (Fig. S2). Thus, the combination of OPA2Vec- and TCGA-based (expression and mutation) profiles was selected as the features for NECARE.

We also performed hyperparameter optimization in the cross-training set. Table S3 shows the hyperparameters we used in the final model of NECARE.

Finally, we evaluated the performance of NECARE in the testing set. Overall, NECARE achieved an MCC = 0.71±0.01 and an F1 = 89±0.7% (Table S4). In addition, we also determined the reliability index (RI) of NECARE (Fig. 2A). RI was correlated with its performance and can be used to measure its prediction performance. The RI ranged from -100 to 100 (-100 meant most reliable negative prediction and 100 meant most reliable positive prediction). For instance, the subset of predictions at RI ≥ 0 had a precision of >89% (Fig. 2A: red line at x=0). This level covered approximately 89% of all predictions (Fig. 2A: blue line at x=0). When increasing the RI to 80 (dashed line), the precision reached 96% (Fig. 2A: red line at x=80), but it can cover only 60% of all predictions (Fig. 2A: blue line at x=80). Therefore, basically, a higher RI represented a more reliable prediction. The RI was also calculated for the negative prediction (noninteracting prediction) (Fig. S3). At RI = 0, the precision for the negative prediction was 82%, and it increased to 92% at RI=-80 (Fig. S3).

NECARE outperformed other algorithms

As NECARE is a network-based method, we first compared it with other network-based node relationship prediction algorithms such as the state of art method L3 [24], and their methods were compared. We also compared NECARE with other state-of-the-art sequence-based deep learning PPI prediction methods such as PIPR [17] and DPPI [18] (Fig. 2B).

>>> Fig. 2 <<<

First, we drew the ROC (receiver operating characteristic) curves for all the methods (Fig. 2C) and calculated the AUC for them. Our method achieved the best performance with an AUC = 0.93 (Fig. 2C, Table S4), while most of the other methods had an AUC of approximately 0.5 (Table S4).

For the detailed metrics, NECARE reached the highest F1 (89±0.7%) and MCC (0.71±0.01) in the comparison (Fig. 2B, Table S4). The RCNN (recurrent convolutional neural network)-based method PIPR achieved the highest precision of 95±0.5% (precision for NECARE was 89±0.8%). However, PIPR had a low specificity of 30±5%, and the specificity of NECARE was 82±1% (Table S4).

Many network-based methods have a very high sensitivity (recall) but low precision. The probable reason is that most of those methods were designed for unidirected prediction to determine whether two proteins/genes are bound (or interact) only with each other. This meant that they could not distinguish the direction of the interaction, while some of the negative samples in our dataset were the reverse links of the positive interactions. To make a fairer comparison, we performed a further analysis by 1) removing all the interactions in our test set except “binding” and “missing interaction” and 2) making the interactions unidirected.

In the unidirected mode, NECARE still achieved the best performance with an AUC = 0.90 and MCC = 0.27±0.02 (Fig. 2C, Table S5). In the unidirected mode, the precision of most methods significantly increased to approximately 96%, which indicated our hypothesis that those methods cannot determine the direction of the links (Table S5). As the precision of NECARE also slightly improved to 99±0.3%, it was the method with the highest precision in the unidirected mode (Table S5).

Finally, we can conclude that NECARE is currently the best prediction method that can be used to identify gene relationships in cancer.

Cancer hub genes discovered by NECARE

By applying NECARE, unlike previous studies that were limited to the coexpression between genes [2], we were able to reveal the comprehensive and rigorous perturbation of the cancer gene network. We mapped the cancer gene interactome with its highly reliable predictions ($|RI| \geq 0.8$, Figs. 2A and S3). On average, each gene lost 36 edges with the other genes in the cancer gene network. However, they obtained approximately 659 new edges on average (Fig. S4, red dashed lines). This indicated our hypothesis that instead of simply being fractured, the gene network in cancer is reprogrammed.

Furthermore, we assumed that the perturbation was not evenly distributed among all the genes. Some genes may hold more perturbations than others. Genes enriched with network perturbations (gained/lost links) were defined as cancer hub genes. Finally, we identified 2287 genes enriched with network perturbations in cancer (Fig. 3A, Table S6). The KEGG pathway enrichment analysis (Fig. S5), as expected, showed that the 2287 cancer hub genes were enriched in some cancer-related pathways: the PI3K-Akt signaling pathway (P-

value =1.2×10^{-22}), Ras signaling pathway (P-value = 1.8×10^{-27}), Rap1 signaling pathway (P-value =1.3×10^{-22}), and ErbB signaling pathway (P-value =1.1×10^{-22}) (Fig. S5).

>>> Fig. 3 <<<

Then, we classified cancer hub genes into three types: Type 1, hub genes enriched with gained links; Type 2, hub genes enriched with lost links; and Type 3, hub genes enriched with both gained and lost links. Overall, we identified 1274 Type 1 hub genes, 1000 Type 2 hub genes and 13 Type 3 hub genes (Fig. 3A). Fig. 3B shows their distribution among chromosomes and the top 1000 links among all the hub genes. Even among hub genes, the top 1000 links were not distributed evenly, and some hub genes had more links than others. For example, MXD1 was engaged in 58 links among the top 1000 links. MXD1 encodes a member of the MYC/MAX/MAD network of basic helix-loop-helix leucine zipper transcription factors and mediates cellular proliferation, differentiation and apoptosis [44]. In contrast, 314 hub genes engaged in only one link from the top 1000 links.

Type 2 hub genes had many well-known cancer-related genes, such as BCL2, BCL3, BRCA1, CCNB1, CCNB2, CDK1, CDK4, CDK14, CDKN1B, E2F1, JUN, KRAS, MRAS MYC, PTEN, SAMD2 SAMD3, STAT3, TP53, YAP1, and YY1 (Table S6). According to the PubMed index, for Type 2 hub genes, there were 924 cancer-related publications per gene. Though Type 1 hub genes also contained a large number of cancer-related genes, such as genes from the RAB family, which are part of the Ras superfamily of small GTPases, RAB17, RAB1A, RAB27A, RAB29, RAB30, RAB31, RAB3D and RAB3IP, there were fewer studies for Type 1 hub genes than for Type 2 hub genes (225 publications per gene).

Regarding the 13 Type 3 hub genes that had both gained- and lost-link perturbations (Fig. 3A, Table S6), some of these genes, such as EGFR and NRAS, are already well-known in cancer research. However, some genes, such as MAPK12 from the mitogen-activated protein kinase (MAPK) family, have not yet been well studied. It is well known that the MAPK family consists of a large number of kinases that are altered in cancers, and many targeted therapies have already been developed against these kinases. Therefore, it is reasonable to speculate that MAPK12 also plays an important role in cancer development, and MAPK12 is worth studying further [45].

More interestingly, over 56% of genes that were found to be involved in cancer treatment were cancer hub genes. Among them, 13% were Type 1 hub genes, 42% were Type 2 hub genes, and 1% were Type 3 hub genes. In addition, the distribution of the gained edges had no difference between clinically related genes and the background (all genes) (Kolmogorov–Smirnov P-value = 0.59, Fig. S4A). However, there was a significant difference in the

distribution of the lost edges (Kolmogorov–Smirnov P-value < 2.2×10^{-16}; $Mean_{All\ genes}$ = 36 and $Mean_{Clinically\ related\ genes}$ = 122) (Fig. S4B). Furthermore, those hub genes were significantly associated with the 10-year survival outcomes of 32 distinct types of cancer (Fig. 4). Overall, patients with high MSs had a poor prognosis and low survival rate (red lines in Fig. 4).

>>> Fig. 4 <<<

Subsequently, we analyzed the centrality of those hub genes (Fig. 3C). Three types of hub genes and the nonhub genes could be clearly separated by the centrality. First, this suggested that our statistical analysis, which was applied to identify hub genes, was reliable because we did not consider centrality during the identification of genes. Then, we found that Type 1 hub genes tended to have a high centrality in the cancer network but a low centrality in the normal network. However, Type 2 hub genes showed the reverse trend (a high centrality in the normal network but a low centrality in the cancer network). Type 3 hub genes were balanced between Type 1 and Type 2 hub genes. These nonhub genes had a low centrality in both normal and cancer networks. The centrality change in Type 1 and Type 2 hub genes also reflected the perturbation of the cancer network.

We also found that cancer mutations in Type 1 and Type 3 hub genes were more likely to be protein-protein binding mutations (Fig. 5A, Table S7). However, mutations in Type 2 hub genes were enriched with nonbinding (protein-protein) mutations, which indicated that the links lost in Type 2 hub genes were less likely to have a PPI relationship (P-value < 2.2×10^{-16}) (Fig. 5A, Table S7). For protein-nucleotide (DNA/RNA) binding mutations, mutations in Type 1 and Type 2 hub genes were overrepresented as DNA-protein and RNA-protein binding mutations, which meant they were more involved in transcription and posttranscriptional regulation (P-value < 2.2×10^{-16}, Fig. 6B, Tables S8, S9). Though Type 3 hub genes were enriched with Protein-Protein and Protein-DNA binding mutations (Fig. 5A, 5B), the mutations in Type 3 hub genes were underrepresented as protein-RNA binding mutations (Fig. 5C, Table S9). This indicated that the mutations in Type 3 hub genes were more likely to affect PPIs and the processes of transcription.

>>> Fig. 5 <<<

Pathway cross-talk in cancer

Usually, a series of genes or molecules perform functions in groups and lead to a certain product or a change in cells. The assembly of interactions among these genes is called the

biological/signaling pathway. To further understand the network perturbation, we also mapped out the interactome of pathways, and pathway cross-talk is the interaction between different signaling pathways. Among the 1403 cross-talks in Fig. 7, 645 of them were new cross-talks in cancer that did not exist in the normal gene network (Fig. 6, red edges). These new cross-talks were not distributed evenly. Some pathways had more new cross-talks, while some held fewer.

The glycosaminoglycan degradation pathway participated in 69 new cross-talks (Fig. 6, red edges). Glycosaminoglycans are natural heteropolysaccharides that are present in every mammalian tissue [46]. Glycosaminoglycans are engaged in the regulation of growth factor signaling, cellular behavior, inflammation, angiogenesis, and the proteolytic environment by forming proteoglycans, which are a type of posttranslational modification [46,47]. Dysregulated expression of glycosaminoglycans is present in cancer and is reported to correlate with clinical prognosis in several malignant neoplasms. Recent knowledge about the biological roles of these molecules in cancer biology, tumor angiogenesis, and metastasis has promoted the development of drugs targeting them [47].

The IL-17 signaling pathway was found to have cross-talks with 55 pathways, 38 of which did not exist in the normal gene network. Therapies inhibiting the IL-17 pathway to treat inflammatory diseases such as rheumatoid arthritis or psoriasis have been clearly established. Though IL-17 was already found to be related to some cancers, its detailed function in oncology is still not very clear [48].

>>> Fig. 6 <<<

In our study, the IL-17 signaling pathway was found to have cross-talks with many important cancer-related pathways such as the ErbB and Rap1 signaling pathways. A previous study reported that IL-17A could upregulate the phosphorylation level of ERK1/2, which is involved in those pathways [49]. The cross-talk between the IL-17 and Notch signaling pathways was also found in our study based on the cancer network. The Notch pathway is one of the most commonly activated signaling pathways in cancer and is involved in cell proliferation, differentiation and survival [50]. The Notch pathway was found to be involved in many pathway cross-talks (n=59) in the cancer network, and 23 of them were new cross-talks (Fig. 6, red edges). The Notch pathway had a cross-talk with the Wnt signaling pathway and apoptosis pathway in the cancer network. The Notch pathway also cross-talked with the cell cycle pathway in both the cancer and normal networks.

In addition, it was also interesting to find that the thiamine metabolism pathway (n=69) and pentose phosphate pathway (PPP) (n=46) are involved in many pathway cross-talks (Fig. 6). Interest in cancer metabolism has increased in recent years. Thiamine supplementation contributes to a high rate of tumor cell survival, proliferation and chemotherapy resistance [51]. Thus, we speculated that cancer cells may regulate the concentration of thiamine by integrating signals from other pathways (e.g., the Notch, Hedgehog and VEGF signaling pathways, which were found to cross-talk with the thiamine metabolism pathway). PPP is a metabolic pathway parallel to glycolysis and represents the first committed step of glucose metabolism [52]. The PPP plays a pivotal role in supporting cancer cell survival and growth by generating pentose phosphate for nucleic acid synthesis and providing nicotinamide-adenine dinucleotide phosphate (NADPH) [53]. Lin et al. found that 6-phosphogluconate dehydrogenase, the third enzyme in the PPP, linked oxidative PPP, lipogenesis and tumor growth by inhibiting LKB1-AMPK signaling [54]. In our cross-talk analysis, we also found cross-talk between PPP and the AMPK signaling pathway, which again showed that our results can provide important cues for future cancer research.

Experimental validation of NECARE

The Ras signaling pathway is one of the most important pathways in cancer. Considering that pathways with too many genes (> 200) might be too generic, the Ras signaling pathway was filtered out and not discussed in our cross-talk analysis. However, there were clearly cross-talks between the Ras signaling pathway and other pathways. For example, there is cross-talk between the Wnt and Ras signaling pathways. Fig. 8A shows 10 highly reliable (RI > 80, Fig. 2A) bidirected gene interactions predicted by NECARE between WNT3 (from the Wnt signaling pathway) and SHC2 (from the Ras signaling pathway) with the following genes: BCL10, CDK19, EBF3, FOXC2, IGFBP4, LHX1, NUP153, PRKAR1B, PTPN5 and RAB30. WNT3 is a member of the Wnt family and may play a key role in cancer through activation of the Wnt-beta-catenin-TCF signaling pathway [55]. SHC2 was located very upstream of the Ras signaling pathway and could be activated by many receptor tyrosine kinases (RTKs) in the Ras signaling pathway [6] (Fig. 7A).

>>> Fig. 7 <<<

As bidirected positive predictions in NECARE were most likely to be the “binding” relationships, we applied coimmunoprecipitation (co-IP) to validate the predictions (SOM: Section S1, coimmunoprecipitation). We co-transfected the expression vectors of these 10 genes together with WNT3 and SHC2 in glioblastoma cell line LN229 (Fig. 7B, C). Co-IP was

applied to confirm their binding interaction. Ninety percent (18 of 20) of NECARE predictions were confirmed (Fig. 7B, C). Only two pairs of interactions, PTPN5-WNT3 and IGFBP4-SHC2, obtained negative validation results in co-IP (Fig. 7B, C).

Discussion

Previous studies have already found several gene network perturbations in cancer gene networks. This result indicated that the gene network in cancer was significantly different from that in non-disease situations. In our study, we used R-GCN to establish a prediction method, NECARE, which is specific for cancer.

In the biological cell system, instead of isolation, genes act as a complex network. Genes may be regulated by others, control the expression of many other genes, or function together with other genes. Our model simulated this biological system by using a GCN. Information for each gene can be obtained from their neighbors in the network with weights for different relationship types and directions. Then, we compared our method with two different kinds of other algorithms: 1) sequence-based methods and 2) network-based methods. Our system outperformed all other algorithms in the task of predicting gene relationships in cancer. Sequence-based, state-of-the-art methods, such as PPI-Detect and PIPR [17,20], achieved good performance in PPI prediction but failed in our cancer-specific task. Since genes were acting as a network complex, the disorder information would be broadcasted among the network. And the interaction between two genes may also be affected by their neighbors in the network. Therefore, sequence-based methods which only considered the input genes themselves may not be very specific for cancer gene relationship prediction. This is also the reason why we used network-based algorithm combined with knowledge-based features such as OPA2Vec. Our system with R-GCN can distinguish the relationship types and directions, while other network-based algorithms are not able to do so. Thus, our method is currently the best solution for cancer gene network prediction.

With the help of NECARE, we identified 2287 cancer hub genes that were enriched with network perturbations in cancer. As gene network perturbation was already found to be the main reason for cancer, these cancer hub genes should be the focus of the pathological mechanisms and treatment targets. Indeed, we found that a high MS of hub genes was significantly related to a poor prognosis of 32 different types of cancers. Over half of the cancer treatment-related genes in the database TARGET were hub genes in our study. Thus, these hub genes we identified have a high potential to be the drug design targets for cancer treatment and other clinical research.

In addition, as mentioned before, we classified the hub genes into three types: Type 1 (gained links), Type 2 (lost links), and Type 3 (both gained and lost links). The majority of the well-known cancer-related genes belonged to Type 2 hub genes, and previous clinical

studies also focused more on these hub genes, which broke the network heavily. We speculated that this was because the dysfunction of the gene itself was considered the cause of cancer in the past, and strong dysfunction may make a gene completely removed from the network and lose all of its edges. Thus, most previous studies focused on Type 2 hub genes that are enriched with lost edges in the cancer network. However, instead of the dysfunction of the gene itself, network perturbations were found to be the reason for cancer [3]. With our cancer gene relationship prediction method NECARE, many genes enriched with gained edges in the cancer network were identified (Type 1 hub genes). Therefore, future studies focusing on Type 1 hub genes (also Type 3 hub genes that were enriched with both gained and lost edges) could become a new perspective of cancer research and lead to a better understanding of the pathological mechanism of cancer.

In our study, we also plotted the cross-talk map among different signaling pathways to decipher the underlying structure of cell-signaling networks in cancer (Fig. 6). As expected, some cancer-related pathways such as the Notch and VEGF signaling pathways, have cross-talks with many other signaling pathways. As we know, a cell has numerous receptors and is constantly and simultaneously receiving many signals. A protein can bind to signals by a specific receptor, leading to its activation and initiating signal-transduction pathways [56], which is part of any communication process that governs the basic activities of cells and coordinates multiple-cell actions [57]. Perturbations in signaling interactions and cellular information processing may cause diseases such as cancer [58]. Thus, our results may help researchers and clinicians better understand the pathogenesis of cancer to treat cancer more effectively.

Overall, in our study, we established the first cancer-specific gene relationship prediction method. With the help of our new method, we analyzed gene network perturbations in cancer and identified cancer hub genes. Our method provides a powerful tool for biology researchers and clinicians to find possible interacting partners of their input genes in cancer. They can also choose to focus their research on the cancer hub genes identified by our method to develop new targets for cancer treatment.

Conclusions

In conclusion, the gene relationship network in cancer was found to be different from that in non-disease situations in many previous studies. Here, we provide the first cancer-specific network-based gene relationship prediction method: NECARE. Using NECARE, we revealed cancer gene network perturbations. We identified cancer hub genes that were enriched in network perturbations. We analyzed pathway cross-talk in cancer. Finally, we showed an example of NECARE application and validated the predictions by biological experiments.

ACKNOWLEDGMENT

Financial support was obtained from the program of the China Scholarships Council (CSC201606230244, CSC201708080070). This work was also supported by a grant from the Outstanding Leaders Training Program of Pudong Health Bureau of Shanghai (PWR12018-07) and the Key Discipline Construction Project of Pudong Health Bureau of Shanghai (PWZxk2017-23, PWYgf2018-05). We also want to thank Prof. Jingbin Yan for helping revise the manuscript.

CONFLICT OF INTEREST

None declared.

References

1 del Sol, A., Balling, R., Hood, L. & Galas, D. Diseases as network perturbations. *Curr Opin Biotechnol* **21**, 566-571, doi:10.1016/j.copbio.2010.07.010 (2010).

2 Grechkin, M., Logsdon, B. A., Gentles, A. J. & Lee, S. I. Identifying Network Perturbation in Cancer. *PLoS Comput Biol* **12**, e1004888, doi:10.1371/journal.pcbi.1004888 (2016).

3 Park, J., Lee, D. S., Christakis, N. A. & Barabasi, A. L. The impact of cellular networks on disease comorbidity. *Mol Syst Biol* **5**, 262, doi:10.1038/msb.2009.16 (2009).

4 Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Mol Syst Biol* **5**, 321, doi:10.1038/msb.2009.80 (2009).

5 Engin, H. B., Kreisberg, J. F. & Carter, H. Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. *PLoS One* **11**, e0152929, doi:10.1371/journal.pone.0152929 (2016).

6 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-462, doi:10.1093/nar/gkv1070 (2016).

7 Li, Z. *et al.* The OncoPPi network of cancer-focused protein-protein interactions to inform biological insights and therapeutic strategies. *Nat Commun* **8**, 14356, doi:10.1038/ncomms14356 (2017).

8 West, J., Bianconi, G., Severini, S. & Teschendorff, A. E. Differential network entropy reveals cancer system hallmarks. *Sci Rep* **2**, 802, doi:10.1038/srep00802 (2012).

9 Guan, Y., Dunham, M. J., Troyanskaya, O. G. & Caudy, A. A. Comparative gene expression between two yeast species. *BMC Genomics* **14**, 33, doi:10.1186/1471-2164-14-33 (2013).

10 Wang, K. *et al.* Meta-analysis of inter-species liver co-expression networks elucidates traits associated with common human diseases. *PLoS Comput Biol* **5**, e1000616, doi:10.1371/journal.pcbi.1000616 (2009).

11 Bockmayr, M., Klauschen, F., Gyorffy, B., Denkert, C. & Budczies, J. New network topology approaches reveal differential correlation patterns in breast cancer. *BMC Syst Biol* **7**, 78, doi:10.1186/1752-0509-7-78 (2013).

12 Xulvi-Brunet, R. & Li, H. Co-expression networks: graph properties and topological comparisons. *Bioinformatics* **26**, 205-214, doi:10.1093/bioinformatics/btp632 (2010).

13 Hayashi, T., Matsuzaki, Y., Yanagisawa, K., Ohue, M. & Akiyama, Y. MEGADOCK-Web: an integrated database of high-throughput structure-based protein-protein interaction predictions. *BMC Bioinformatics* **19**, 62, doi:10.1186/s12859-018-2073-x (2018).

14 Fukuhara, N. & Kawabata, T. HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures. *Nucleic Acids Res* **36**, W185-189, doi:10.1093/nar/gkn218 (2008).

15 Chen, R., Li, L. & Weng, Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **52**, 80-87, doi:10.1002/prot.10389 (2003).

16 Zhang, C. & Lai, L. SDOCK: a global protein-protein docking program using stepwise force-field potentials. *J Comput Chem* **32**, 2598-2612, doi:10.1002/jcc.21839 (2011).

17 Chen, M. *et al.* Multifaceted protein-protein interaction prediction based on Siamese residual RCNN. *Bioinformatics* **35**, i305-i314, doi:10.1093/bioinformatics/btz328 (2019).

18 Hashemifar, S., Neyshabur, B., Khan, A. A. & Xu, J. Predicting protein-protein interactions through sequence-based deep learning. *Bioinformatics* **34**, i802-i810, doi:10.1093/bioinformatics/bty573 (2018).

19 Sun, T., Zhou, B., Lai, L. & Pei, J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics* **18**, 277, doi:10.1186/s12859-017-1700-2 (2017).

20 Romero-Molina, S., Ruiz-Blanco, Y. B., Harms, M., Munch, J. & Sanchez-Garcia, E. PPI-Detect: A support vector machine model for sequence-based prediction of protein-protein interactions. *J Comput Chem* **40**, 1233-1242, doi:10.1002/jcc.25780 (2019).

21 Adamic, L. A. & Adar, E. Friends and neighbors on the Web. *Social Networks* **25**, 211-230, doi:https://doi.org/10.1016/S0378-8733(03)00009-1 (2003).

22 Dice, L. R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **26**, 297-302, doi:10.2307/1932409 (1945).

23 Jaccard, P. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. *New Phytologist* **11**, 37-50, doi:10.1111/j.1469-8137.1912.tb05611.x (1912).

24 Kovacs, I. A. *et al.* Network-based prediction of protein interactions. *Nat Commun* **10**, 1240, doi:10.1038/s41467-019-09177-y (2019).

25 Newman, M. E. J. Clustering and preferential attachment in growing networks. *Physical Review E* **64**, 025102, doi:10.1103/PhysRevE.64.025102 (2001).

26 Zhou, T., Lü, L. & Zhang, Y.-C. Predicting missing links via local information. *The European Physical Journal B* **71**, 623-630, doi:10.1140/epjb/e2009-00335-8 (2009).

27 Cannistraci, C. V., Alanis-Lobato, G. & Ravasi, T. Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. *Bioinformatics* **29**, i199-209, doi:10.1093/bioinformatics/btt208 (2013).

28 Schlichtkrull, M. *et al.* in *The Semantic Web.* (eds Aldo Gangemi *et al.*) 593-607 (Springer International Publishing).

29 Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607-D613, doi:10.1093/nar/gky1131 (2019).

30 Alanis-Lobato, G., Andrade-Navarro, M. A. & Schaefer, M. H. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res* **45**, D408-D414, doi:10.1093/nar/gkw985 (2017).

31 Orchard, S. *et al.* The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**, D358-363, doi:10.1093/nar/gkt1115 (2014).

32 Ceol, A. *et al.* MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res* **38**, D532-539, doi:10.1093/nar/gkp983 (2010).

33 Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535-539, doi:10.1093/nar/gkj109 (2006).

34 Keshava Prasad, T. S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res* **37**, D767-772, doi:10.1093/nar/gkn892 (2009).

35 Xenarios, I. *et al.* DIP: The Database of Interacting Proteins: 2001 update. *Nucleic Acids Res* **29**, 239-241, doi:10.1093/nar/29.1.239 (2001).

36 Bader, G. D., Betel, D. & Hogue, C. W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31**, 248-250, doi:10.1093/nar/gkg056 (2003).

37 Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **30**, 31-34, doi:10.1093/nar/30.1.31 (2002).

38 Fabregat, A. *et al.* Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* **18**, 142, doi:10.1186/s12859-017-1559-2 (2017).

39 Blohm, P. *et al.* Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res* **42**, D396-400, doi:10.1093/nar/gkt1079 (2014).

40 Smaili, F. Z., Gao, X. & Hoehndorf, R. OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. *Bioinformatics* **35**, 2133-2140, doi:10.1093/bioinformatics/bty933 (2019).

41 Li, Y., Agarwal, P. & Rajagopalan, D. A global pathway crosstalk network. *Bioinformatics* **24**, 1442-1447, doi:10.1093/bioinformatics/btn200 (2008).

42 Page, L., Brin, S., Motwani, R. & Winograd, T. in *WWW 1999.*

43 Qiu, J. *et al.* ProNA2020 predicts protein-DNA, protein-RNA, and protein-protein binding proteins and residues from sequence. *J Mol Biol*, doi:10.1016/j.jmb.2020.02.026 (2020).

44 Grinberg, A. V., Hu, C. D. & Kerppola, T. K. Visualization of Myc/Max/Mad family dimers and the competition for dimerization in living cells. *Mol Cell Biol* **24**, 4294-4308, doi:10.1128/mcb.24.10.4294-4308.2004 (2004).

45 Braicu, C. *et al.* A Comprehensive Review on MAPK: A Promising Therapeutic Target in Cancer. *Cancers (Basel)* **11**, doi:10.3390/cancers11101618 (2019).

46 Afratis, N. *et al.* Glycosaminoglycans: key players in cancer cell biology and treatment. *FEBS J* **279**, 1177-1197, doi:10.1111/j.1742-4658.2012.08529.x (2012).

47 Yip, G. W., Smollich, M. & Gotte, M. Therapeutic value of glycosaminoglycans in cancer. *Mol Cancer Ther* **5**, 2139-2148, doi:10.1158/1535-7163.MCT-06-0082 (2006).

48 Fabre, J. *et al.* Targeting the Tumor Microenvironment: The Protumor Effects of IL-17 Related to Cancer Type. *Int J Mol Sci* **17**, doi:10.3390/ijms17091433 (2016).

49 Kim, G. *et al.* Interleukin-17 induces AP-1 activity and cellular transformation via upregulation of tumor progression locus 2 activity. *Carcinogenesis* **34**, 341-350, doi:10.1093/carcin/bgs342 (2013).

50 Yuan, X. *et al.* Notch signaling: an emerging therapeutic target for cancer treatment. *Cancer Lett* **369**, 20-27, doi:10.1016/j.canlet.2015.07.048 (2015).

51 Lu'o'ng, K. V. & Nguyen, L. T. The role of thiamine in cancer: possible genetic and cellular signaling mechanisms. *Cancer Genomics Proteomics* **10**, 169-185 (2013).

52 Ramos-Martinez, J. I. The regulation of the pentose phosphate pathway: Remember Krebs. *Arch Biochem Biophys* **614**, 50-52, doi:10.1016/j.abb.2016.12.012 (2017).

53 Jin, L. & Zhou, Y. Crucial role of the pentose phosphate pathway in malignant tumors. *Oncol Lett* **17**, 4213-4221, doi:10.3892/ol.2019.10112 (2019).

54 Lin, R. *et al.* 6-Phosphogluconate dehydrogenase links oxidative PPP, lipogenesis and tumour growth by inhibiting LKB1-AMPK signalling. *Nat Cell Biol* **17**, 1484-1496, doi:10.1038/ncb3255 (2015).

55 Chiurillo, M. A. Role of the Wnt/beta-catenin pathway in gastric cancer: An in-depth literature review. *World J Exp Med* **5**, 84-102, doi:10.5493/wjem.v5.i2.84 (2015).

56 Kumar, P. & Mina, U. *Life Sciences, Fundamentals and Practice, Part I.* (2016).

57 Vlahopoulos, S. A. *et al.* Dynamic aberrant NF-kappaB spurs tumorigenesis: a new model encompassing the microenvironment. *Cytokine Growth Factor Rev* **26**, 389-403, doi:10.1016/j.cytogfr.2015.06.001 (2015).

58 Wang, K., Grivennikov, S. I. & Karin, M. Implications of anti-cytokine therapy in colorectal cancer and autoimmune diseases. *Ann Rheum Dis* **72 Suppl 2**, ii100-103, doi:10.1136/annrheumdis-2012-202201 (2013).

Figures

Fig. 1 Illustration of the perturbation of the gene relationship network and NECARE algrithom.

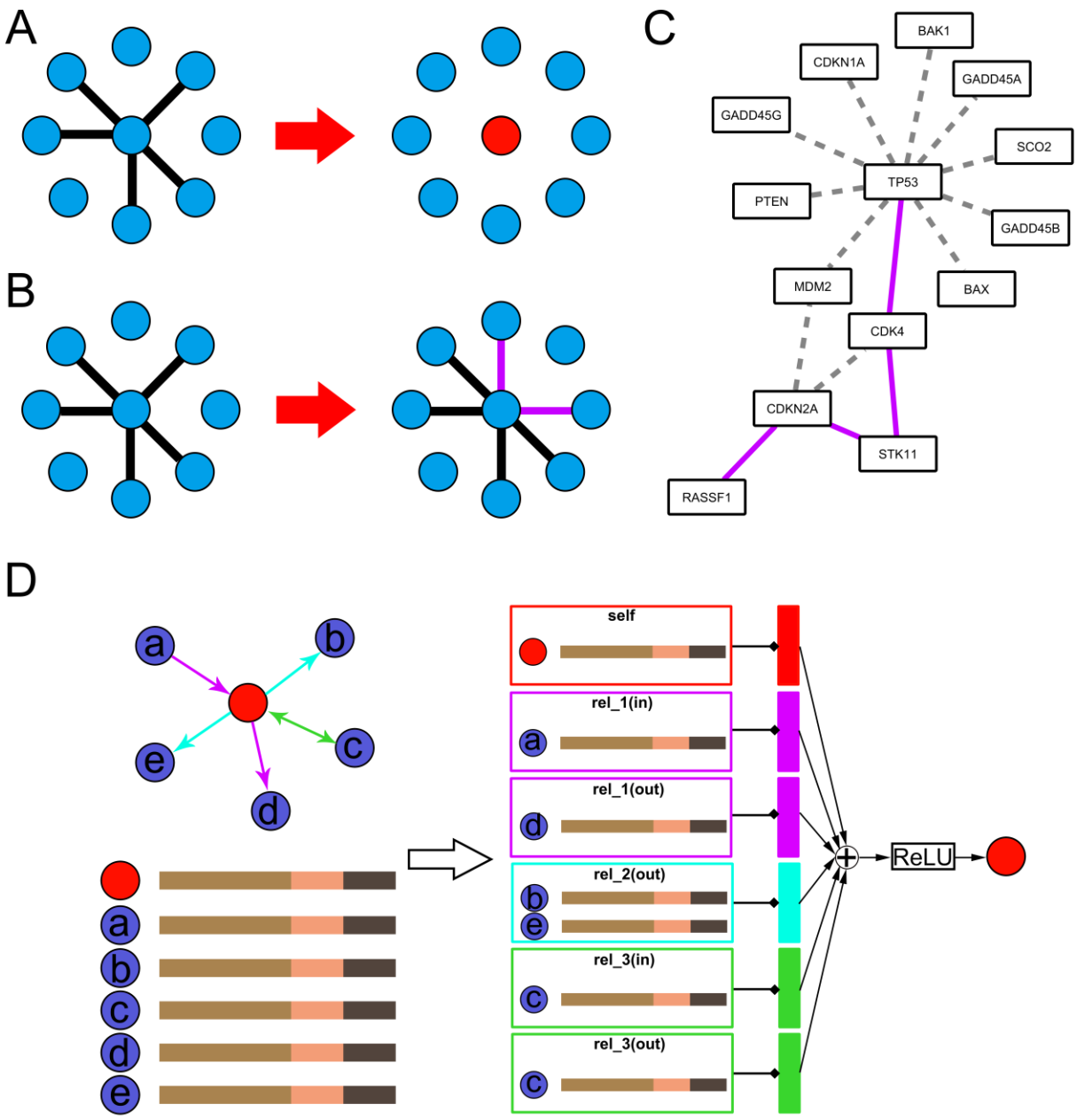

Fig. 1: Illustration of the perturbation of the gene relationship network and NECARE algorithm. **Panel A-C** introduce the concept of gene network perturbation. (A) Each node represents a gene. Mutations such as nonsense mutations could cause the node to be totally inactive or absent (red) and lose all the edges connected to this node. (B) Each node represents a gene. Mutations such as missense mutations could cause the gain or loss of specific edges (purple edges mean the new gained edges due to the mutations), while the center node is not totally inactive. (C) This is an example of the perturbation of the gene relationship network in cancer. The example

is based on the KEGG database [6]. Gray dashed edges are the interactions that are lost in cancer, and purple edges are the new interactions in which genes are involved in cancer. **Panel D** shows a simple example to show how we represent the gene (red node) by NECARE with R-GCN. Nodes a-e and the red node represent different genes, and the red node is set as the target gene. Nodes a-e are all in contact with the red node, and different color edges represent different types of interactions. First, each node is represented by a feature vector that contains three parts: (tan: OPA2Vec; salmon: TCGA-based expression feature; and taupe: TCGA-based mutation feature). Then, to represent the red node, the feature vectors are gathered and transformed for each relation type individually (for both in- and out-edges; also, a self-loop is included). The resulting representation (vertical rectangles with different colours for different relationship types) is summed up and passed to an activation function (ReLU).

Fig. 2 Network-based cancer gene relationship (NECARE) prediction.

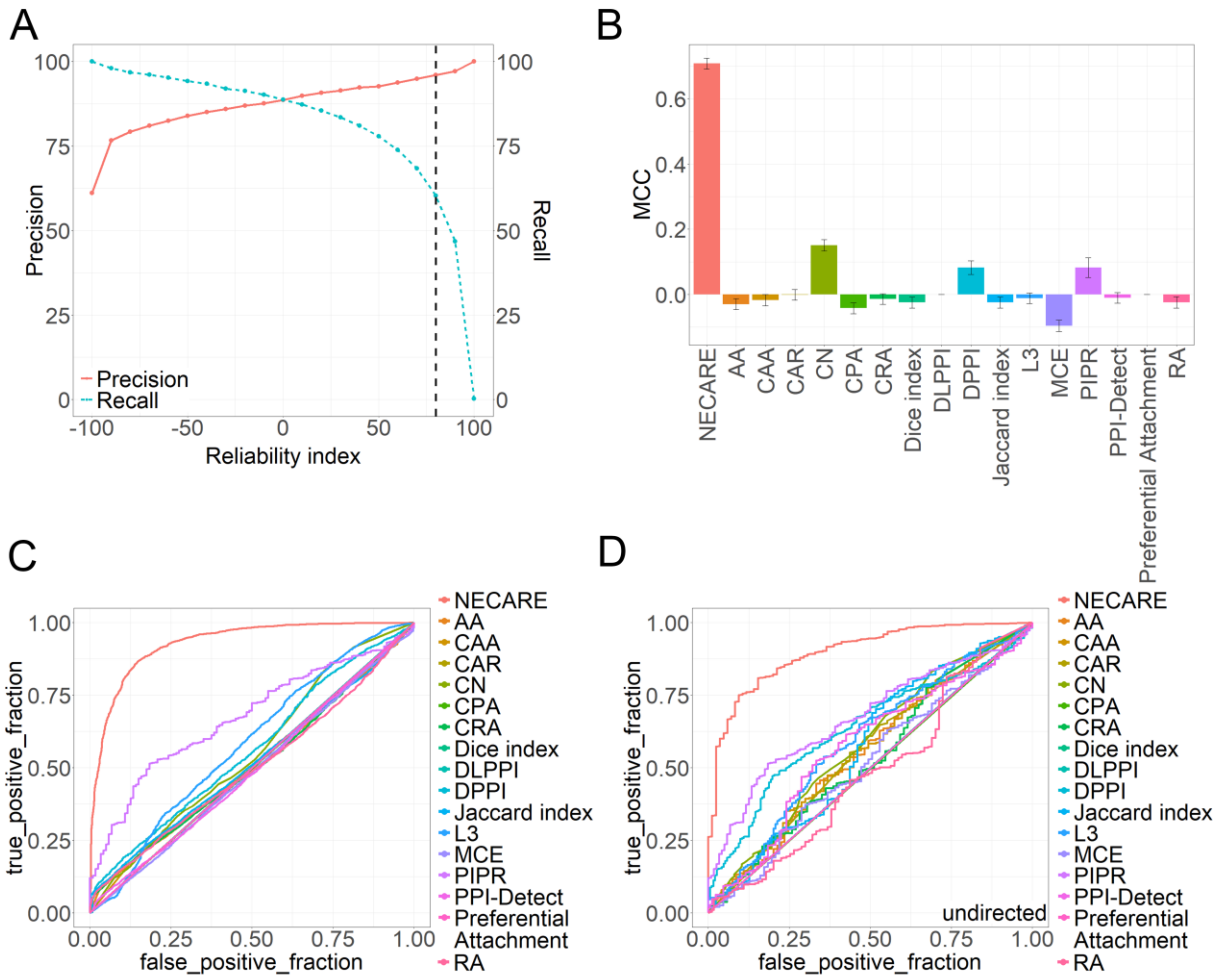

Fig. 2: Network-based cancer gene relationship (NECARE) prediction. (A) All machine learning solutions reflect the strength of a prediction even for binary classifications. This graph relates the prediction strength to the performance. The x-axes give the prediction strength as the RI (from -100: very reliable noninteraction to 100: very reliable interaction). The y-axes reflect the percentage precision (red line, Eqn. 3) and recall (blue line, Eqn. 3). The precision is proportional to the prediction strengths, i.e., predictions with a higher RI are, on average, better than predictions with a lower RI. For example, for all the gene relationship predictions with RI>80 (black dashed line), approximately 96% are correct predictions. (B) The MCC (Eqn. 4) was determined for a comparison among different methods on the test set, and our method NECARE obtains the highest MCC: 0.70. (C) ROC curve comparison for different methods based on the test set. NECARE has the largest AUC: 0.93. (D) ROC curve comparison for different methods (undirected mode). All links were removed except “binding” and “missing interaction”. In addition, links are further changed to be undirected. NECARE still has the largest AUC: 0.90.

Fig. 3 Cancer hub genes of the cancer gene relationship network.

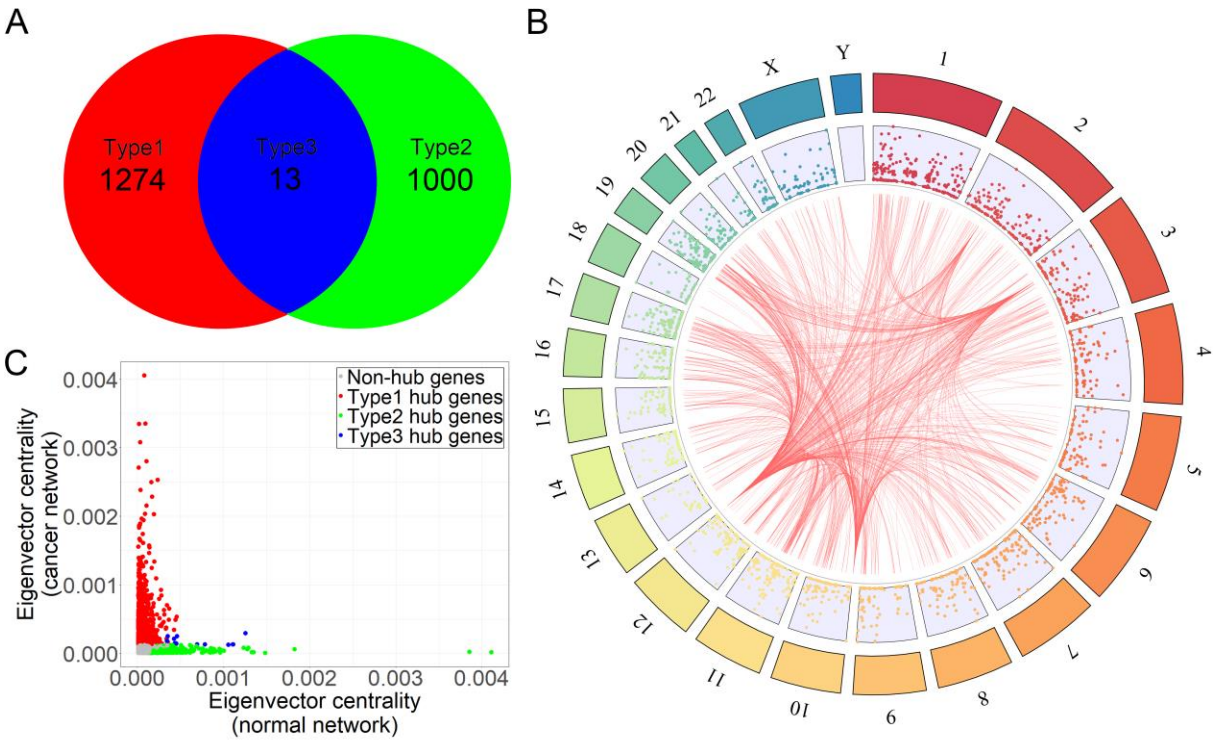

Fig. 3: Cancer hub genes of the cancer gene relationship network. Type 1: hub genes enriched for only gained links; Type 2: hub genes enriched for only lost links; Type 3: hub genes enriched for both gained and lost links. (A) The number of three different types of cancer hub genes. (B) The distribution of cancer hub genes among chromosomes. The links inside the circle are the top 1000 links between cancer hub genes based on the NECARE output scores. (C) The centrality eigenvector of cancer hub genes. The x-axis is the centrality in the normal network, and the y-axis is the centrality in the cancer network.

Fig. 4: The prognostic landscape of hub genes.

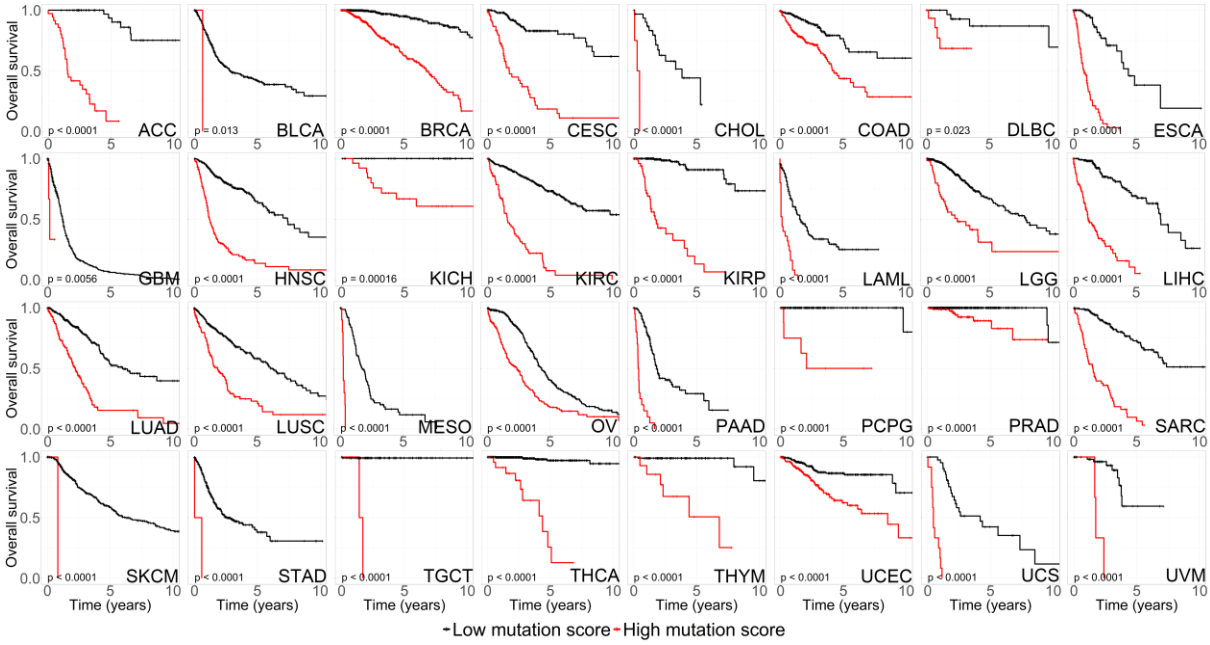

Fig. 4: The prognostic landscape of hub genes. Kaplan–Meier plots for the patients from 32 different types of cancers from TCGA divided into high- and low-MS groups (Methods). The P-value was calculated by the log-rank test.

Fig. 5: Macromolecular binding mutations of cancer hub genes.

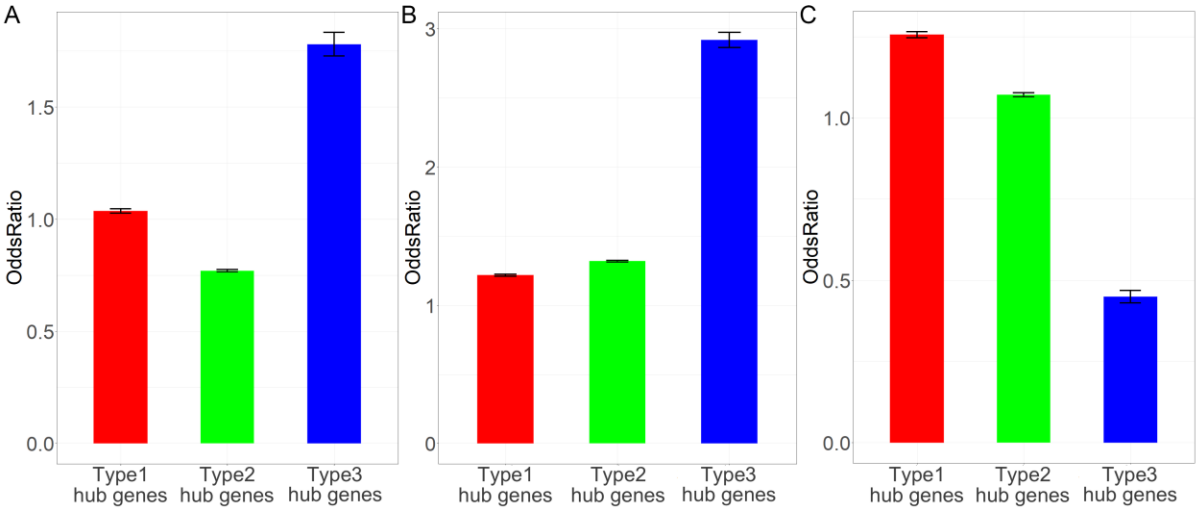

Fig. 5: Macromolecular binding mutations of cancer hub genes. (A) Fisher's exact test for protein-protein binding mutations in three different kinds of cancer hub genes. Shown are the odds ratios and 95% confidence intervals within each set of cancer hub genes. An odds ratio greater than 1 represents the overrepresentation of binding mutations, while an odds ratio less than 1 means the underrepresentation of binding mutations. Binding prediction was performed by ProNA2020 [43]. (B) Fisher's exact test for protein-DNA binding mutations of three different kinds of cancer hub genes. (C) Fisher's exact test for protein-RNA binding mutations of three different kinds of cancer hub genes.

Fig. 6: Pathway cross-talk in the cancer gene network.

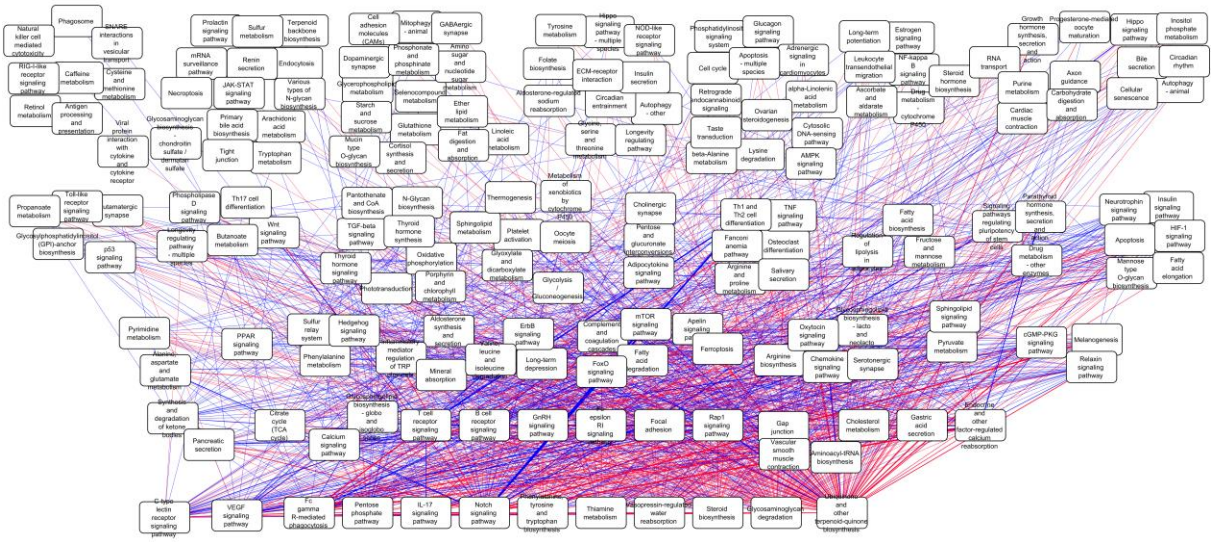

Fig. 6: Pathway cross-talk in the cancer gene network. The plot shows the cross-talk between different pathways. The blue edges indicate the cross-talk existing in the general (nondisease) gene relationship network, and red edges are the new cross-talk in the cancer gene network. The thickness of the edges corresponds to the P-value.

Fig. 7: Experimental validation of the NECARE predictions.

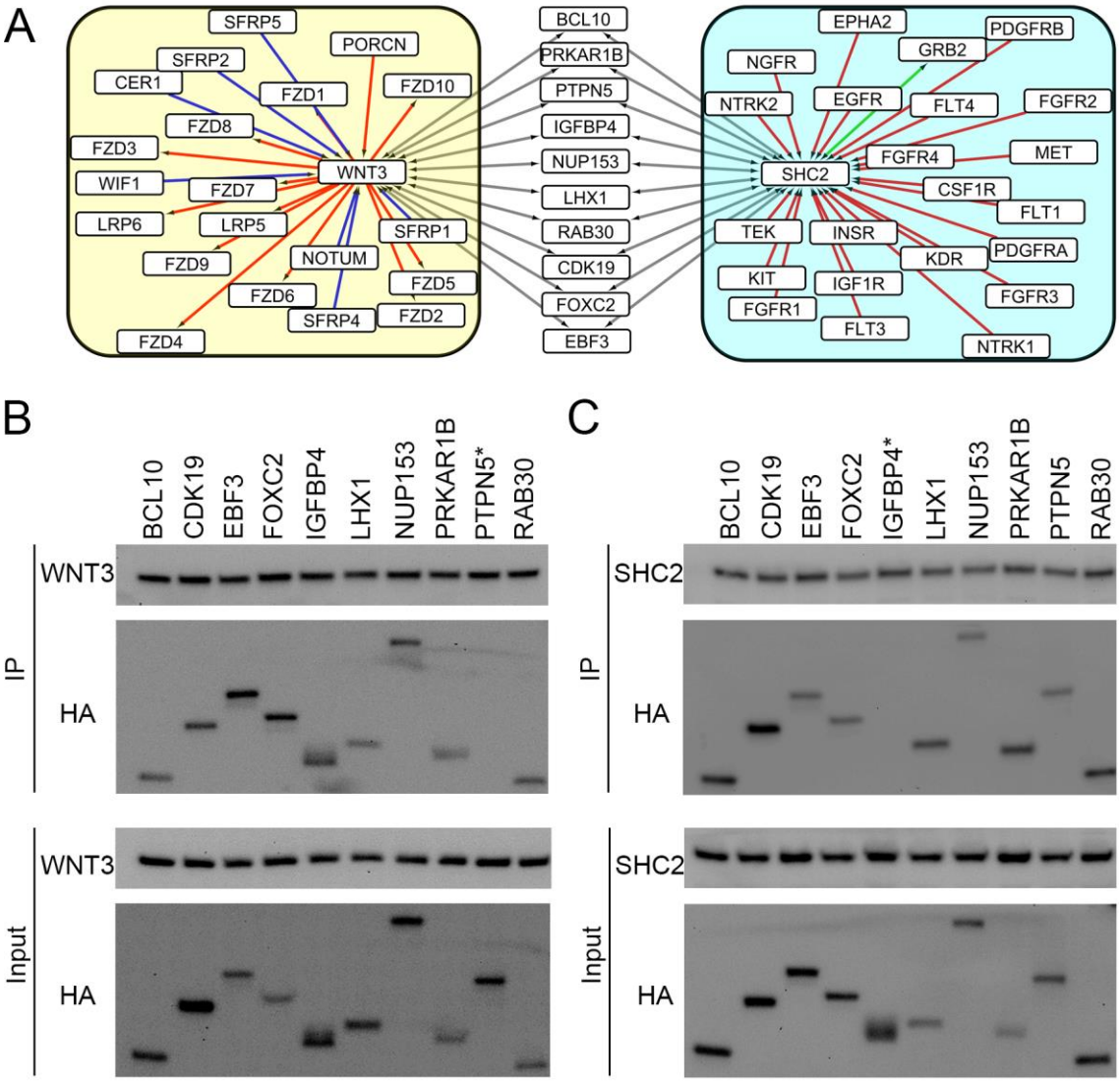

Fig. 7: Experimental validation of the NECARE predictions. Panel A shows the genes that cross-talk with WNT3 and SHC2 in each pathway. Different color edges represent different types of interactions. The red edge indicates activation; the blue edge indicates inhibition; the green edge is the KEGG annotated binding; the gray edge is NECARE predicted binding. The left yellow group shows the genes interacting with WNT3 in the Wnt signaling pathway. The right cyan group shows the genes in contact with SHC2 in the Ras signaling pathway. Those 10 genes in the middle with gray edges are NECARE predicted genes binding to WNT3 and SHC2 with a high RI (> 80). **Panels B** and **C** are co-IPs that validated the interactions of 10 predicted genes with WNT3 and SHC2 in LN229 cells. The interactions were determined by immunoblotting. The labeled "*" indicates a negative result of the co-IP validation experiment. **Panel B:** LN229 cells were cotransfected with the indicated HA-tagged constructs of 10 predicted genes and FLAG-tagged WNT3. **Panel C:** LN229 cells were cotransfected with the indicated HA-tagged constructs of 10 predicted genes and FLAG-tagged SHC2.