Genomics Proteomics Bioinformatics 20 (2022) 928–938

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

METHOD

DGMP: Identifying Cancer Driver Genes by Jointing DGCN and MLP from Multi-omics Genomic Data

Shao-Wu Zhang *, **Jing-Yu Xu, Tong Zhang**

MOE Key Laboratory of Information Fusion Technology, School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

Received 2 November 2021; revised 21 October 2022; accepted 4 November 2022
Available online 1 December 2022

Handled by Feng Gao

KEYWORDS

Driver gene;
Directed graph convolutional network;
Multilayer perceptron;
Gene regulatory network;
Multi-omics data

Abstract Identification of cancer **driver genes** plays an important role in precision oncology research, which is helpful to understand cancer initiation and progression. However, most existing computational methods mainly used the protein–protein interaction (PPI) networks, or treated the directed **gene regulatory networks** (GRNs) as the undirected gene–gene association networks to identify the cancer driver genes, which will lose the unique structure regulatory information in the directed GRNs, and then affect the outcome of the cancer driver gene identification. Here, based on the multi-omics pan-cancer data (*i.e.*, gene expression, mutation, copy number variation, and DNA methylation), we propose a novel method (called DGMP) to identify cancer driver genes by jointing **directed graph convolutional network** (DGCN) and **multilayer perceptron** (MLP). DGMP learns the multi-omics features of genes as well as the topological structure features in GRN with the DGCN model and uses MLP to weigh more on gene features for mitigating the bias toward the graph topological features in the DGCN learning process. The results on three GRNs show that DGMP outperforms other existing state-of-the-art methods. The ablation experimental results on the DawnNet network indicate that introducing MLP into DGCN can offset the performance degradation of DGCN, and jointing MLP and DGCN can effectively improve the performance of identifying cancer driver genes. DGMP can identify not only the highly mutated cancer driver genes but also the driver genes harboring other kinds of alterations (*e.g.*, differential expression and aberrant DNA methylation) or genes involved in GRNs with other cancer genes. The source code of DGMP can be freely downloaded from https://github.com/NWPU-903PR/DGMP.

Introduction

Cancer is a heterogeneous disease that is driven by various kinds of genomic and epigenomic alterations, such as single

* Corresponding author.
E-mail: zhangsw@nwpu.edu.cn (Zhang SW).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2022.11.004

nucleotide variations, DNA methylation, and chromosomal aberrations [1]. Some of these alterations confer growth and positive selection advantages to the mutant cells, leading to intensive proliferation and tumors [2]. That is, the accumulation of diverse genetic mutations causes cancer progression, and these genetic mutations confer a selective growth advantage to the mutant cells [3]. Thus, identification and comprehensive understanding of cancer driver genes that play causal roles in cancer evolution are crucial for cancer diagnosis and therapy [4].

Several cancer sequencing projects [5–7] have generated a large volume of gene mutation data. Therefore, many computational methods have been proposed to identify the driver genes and disease genes from the cancer genomic data. Generally, these methods can be cataloged into three groups: 1) mutation frequency-based methods, 2) network-based methods, and 3) machine learning-based methods.

Mutation frequency-based methods identify the significantly hypermutated genes as the driver genes compared with a background mutation frequency distribution [8,9]. For example, MutSigCV [8] calculated the statistical significance of their mutation frequency among all the samples to identify the recurrently mutated genes as the driver genes. However, due to the tumor heterogeneity, it is difficult to build a reliable background mutation model. In addition, these methods cannot be used to detect the low-mutated frequency and non-mutated cancer driver genes, because part of driver genes are mutated at high frequencies ($>$ 20%), whereas most of the cancer genes are mutated at intermediate frequencies (2%–20%) or even lower frequencies [10], and even many genes involved in tumorigenesis are not altered on the DNA sequences, and these genes are dysregulated through various cellular mechanisms [3].

Network-based methods often adopted random walk with restart (RWR) [11,12], network diffusion [13,14], subnetwork enrichment analysis [15–17], matrix completion [18], and network structure control [19–22] to predict cancer driver genes and disease genes at the biological network level by incorporating the protein–protein interactions (PPIs), pathway knowledge, and so on. For example, pgWalk [11] constructed a disease–gene network by integrating the multiple genomic and phenomic data and then simulated the process of a random walker wandering on such a heterogeneous network to prioritize the candidate genes. MAXIF [15] constructed a phenome–interactome network by integrating the given phenotype similarity profile, PPI network, and associations between diseases and genes, and then maximized the information flow in this phenome–interactome network to uncover the candidate disease genes. Jiang et al. [16] constructed a gene semantic similarity network by the biological process domain of the gene ontology and then used the gene semantic similarity scores in the network to infer disease genes. Although these methods have been successfully used for detecting cancer driver genes and disease genes, they are still limited to the unreliable and incomplete interactions in biological network [23]. Developing an integrative framework by incorporating cancer multi-omics data (*e.g.*, somatic mutations, structural variations, gene expression, and methylation) and adopting the hybrid approaches would improve the prediction of cancer driver genes for the network-based methods.

Machine learning-based methods [24–28] usually train the classifier, *e.g.*, random forest and support vector machine (SVM), by extracting the diverse features from different types of cancer data to predict new cancer driver genes. For example, deepDriver [26] predicted cancer driver genes with a convolutional neural network (CNN) model that was trained with the gene mutation features and their neighbors in the similarity networks. Integrating the functional impact of mutations and the similarity of gene expression patterns with CNN model can improve the prediction accuracy of driver genes. NRFD [28] constructed a cancer gene interaction network by integrating various kinds of cancer-related information sources to obtain the feature vector of each gene and then used the random forest to predict the cancer driver genes.

Most existing machine learning-based methods just extract the network-based features by using network analysis. They cannot effectively combine the network topology features and the multi-omics features of genes. That is, very few methods can combine both multidimensional gene features with the graph representation features of gene–gene interaction networks. For example, the explainable multiomics graph integration (EMOGI) method [29] adopted the graph convolutional network (GCN) model to combine the multidimensional multi-omics gene features with the topological features of the PPI network to identify cancer genes. Although EMOGI [29] successfully identified not only highly mutated cancer genes but also other non-mutated cancer-dependency genes, it only used the association information between genes in an undirected PPI network and does not make full use of the regulation information between genes in gene regulatory network (GRN). In addition, the spectral-based GCN can only be applied to the undirected network, whereas GRN is a directed network that provides the specific causal links (*e.g.*, one gene activates or inhibits other genes) between genes, which helps to understand the molecular mechanism of gene regulation in cancers and the molecular basis of cancer subtypes [30,31]. Thus, we proposed a novel deep learning-based method (called DGMP) to identify the cancer driver genes by integrating the multidimensional multi-omics gene features as well as the topological structure features of the GRN through the directed graph convolutional network (DGCN) model. Compared to the GCN, DGCN [32] uses the first- and second-order proximity to extend the spectral-based graph convolutional to the directed graphs for retaining the connection properties in the directed graph, and also expanding the receptive field in convolutional operation without stacking more convolutional layers.

Generally, the role of the graph in GCN is to guide the weights training by averaging the features of a node with its graph neighbors [33]. When the relationships represented through the graph are consistent with the information of nodes (*i.e.*, the features of neighbor nodes in the graph are expected to be more similar than those of other nodes), GCN can improve the performance of node classification [33,34]. If feature similarities of neighborhood nodes in the graph are not congruent, this graph-based averaging is not beneficial in the training process. It will result in the performance of GCN being lower than that of these methods [*e.g.*, multilayer perceptron (MLP)] that rely exclusively on the node feature information [33,34]. Considering that the features of some genes that are graph neighbors in GRN may not be more similar than other genes, we introduced the MLP classifier into the DGMP model for further improving the performance of cancer driver gene identification. DGMP uses the DGCN model to learn the

multi-omics features of genes, as well as the topological structure features in GRN and adopts MLP to weigh more on gene features for mitigating the bias toward the graph topological features in the DGCN learning process. DGMP aims to not only identify the highly mutated cancer driver genes, but also predict the driver genes that harbor other kinds of alterations (*e.g.*, differential expression and aberrant DNA methylation), or identify the driver genes involved in GRN with other cancer genes. Overall, our main contributions include two points. 1) We use the DGCN model to perform the graph convolutional operation on the directed GRN for capturing the directed information (*i.e.*, regulation information) and expand the receptive field to the second-order neighbor of a gene for aggregating more its neighbor and the topological feature information. 2) We introduce MLP into the DGMP framework to weigh more on gene features for mitigating the bias toward the topological features in the DGCN learning process.

The results on three networks of DawnNet, Kyoto Encyclopedia of Genes and Genomes (KEGG), and RegNetwork show that our DGMP outperforms other existing state-of-the-art methods in terms of the area under the receiver operating characteristic (ROC) curve (AUROC) and the area under the precision–recall (PR) curve (AUPRC), demonstrating that our DGMP can effectively identify the cancer driver genes.

Method

DGMP model

DGMP is based on DGCN and MLP and trained in a semi-supervised manner to discriminate the cancer driver genes from the non-cancer driver genes. The inputs of DGMP are gene single nucleotide variants (SNVs), gene copy number aberrations (CNAs), gene expression information, DNA methylation in gene promoter regions, and the GRN in which some genes have labels, while most have no labels. The positive labels correspond to the annotated cancer driver genes, and the negative labels correspond to the non-caner driver genes in the partially labeled GRN. The output of DGMP is a fully labeled graph, in which each gene is assigned a probability to be a cancer-driver gene. As shown in **Figure 1**, DGMP mainly consists of three modules. The first module (Figure 1A) is that DGCN [32] is used to learn the embedding vectors of genes from GRN and genomic data (*i.e.*, SNVs, CNAs, DNA methylation, and gene expression) by utilizing the first-order proximity (A_F), the second-order in-degree proximity (A_{in}), and the second-order out-degree proximity (A_{out}). The first- and second-order proximity can expand the convolutional operation receptive field, and extract and leverage the directed graph information. The second module (Figure 1B) is the MLP which is used to obtain the embedding vectors of genes only from the genomic data of SNVs, CNAs, gene expression, and DNA methylation. The third module (Figure 1C) is a fully connected neural network that is built to predict the cancer driver genes by concatenating the output embedding vectors from DGCN and MLP.

In the process of training our DGMP model, the semi-supervised training manner is implemented in DGCN module. By inputting both the structural features of partially labeled GRN and the multi-omics features of genes, DGCN encodes GRN structure by directly using a neural network model to obtain the embedding feature vectors of all genes in GRN (in which some genes have labels, while most have no labels), and then trains on a supervised target for all genes with labels. That is, all the labeled and unlabeled genes participate in the generation of graph embedding vectors, and only the labeled genes are used to evaluate the loss of DGMP.

DGCN

In order to use the GCN model which can effectively learn the underlying pairwise relationship among vertices in a directed graph, Tong et al. [32] used the first- and second-order proximity to extend the spectral-based graph convolutional to the directed graphs and then developed a DGCN model to learn the embedding vectors of nodes in directed graphs.

For a directed gene interaction network G, it can be considered a multi-layer GCN by using first- and second-order proximity of the network G [32].

$$\begin{aligned}\hat{A}_F &= \tilde{D}_F^{-\frac{1}{2}}\tilde{A}_F\tilde{D}_F^{-\frac{1}{2}}\\ \hat{A}_{in} &= \tilde{D}_{in}^{-\frac{1}{2}}\tilde{A}_{in}\tilde{D}_{in}^{-\frac{1}{2}}\\ \hat{A}_{out} &= \tilde{D}_{out}^{-\frac{1}{2}}\tilde{A}_{out}\tilde{D}_{out}^{-\frac{1}{2}}\end{aligned} \tag{1}$$

and

$$\begin{aligned}Z_F &= \sigma\left(\hat{A}_F X\Theta\right)\\ Z_{in} &= \sigma\left(\hat{A}_{in} X\Theta\right)\\ Z_{out} &= \sigma\left(\hat{A}_{out} X\Theta\right)\end{aligned} \tag{2}$$

where σ is an activation function; $\tilde{A}_F$ is the first-order proximity matrix with self-loop derived from the gene interaction network matrix $A \in R^{N\times N}$ with partially labeled genes (*i.e.*, cancer driver genes, non-cancer driver genes, and unlabeled genes); $\tilde{A}_{in}$ is the second-order in-degree proximity with self-loop derived from the gene interaction network matrix A; $\tilde{A}_{out}$ is the second-order out-degree proximity with self-loop derived from the gene interaction network matrix A; X is the feature matrix of genes; Θ is a shared trainable weight matrix; Z_F is the first-order proximity convolutional output of DGCN; Z_{in} and Z_{out} are the second-order in-degree and out-degree proximity convolutional outputs of DGCN. Z_F, Z_{in}, and Z_{out} can not only obtain the first- and second-order neighbor feature information in network G, but also Z_{in} and Z_{out} retain the directed structure information in network G.

The first-order proximity entry $A_F(i,j)$, the second-order in-degree proximity entry $A_{in}(i,j)$, and the second-order out-degree proximity entry $A_{out}(i,j)$ between genes v_i and v_j in matrix A are defined as follows:

$$\begin{aligned}A_F(i,j) &= A^{sym}(i,j)\\ A_{in}(i,j) &= \sum_k A_{k,i}A_{k,j}\\ A_{out}(i,j) &= \sum_k A_{i,k}A_{j,k}\end{aligned} \tag{3}$$

where A^{sym} is the symmetric matrix of matrix A. If there is no link from v_i to v_j or v_j to v_i, then $A_F(i,j) = 0$. $A_{in}(i,j)$ is an entry in the in-degree matrix of A. $A_{out}(i,j)$ is an entry in the out-degree matrix of A.

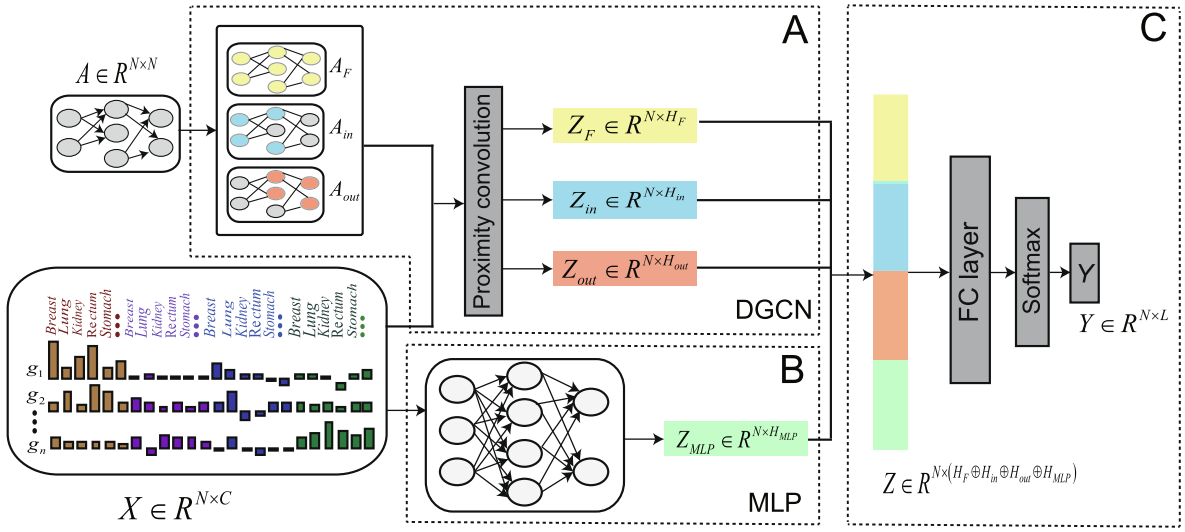

Figure 1 Schematic of DGMP model framework

A. DGCN module. GRN with partially labeled genes and the genomic information (*i.e.*, SNVs, CNAs, DNA methylation, and gene expression) are inputted into the DGCN module. According to the definition of first-order and second-order proximity, we can obtain three proximity networks (*i.e.*, three undirected graphs) of the first-order proximity network (A_F), second-order in-degree proximity network (A_{in}), and second-order out-degree proximity network (A_{out}) from GRN (*i.e.*, directed graph), and then implement the graph convolutional operation on these three proximity networks to achieve graph convolutional of directed graph for generating three embedding vectors (Z_F, Z_{in}, and Z_{out}) of genes. **B.** MLP module. The genomic data of SNVs, CNAs, gene expression, and DNA methylation are inputted into the MLP to obtain the embedding vectors (Z_{MLP}) of genes. **C.** Fully connected neural network module. The first-order proximity convolutional output Z_F, second-order in-degree proximity convolutional output Z_{in}, and second-order out-degree proximity convolutional output Z_{out} of DGCN, the output Z_{MLP} of MLP are concatenated in series to form an embedding matrix Z that is fed into a fully connected neural network to identify the cancer driver genes. g_n represents the n-th gene; N is the total number of genes; C represents the dimension of gene features; L represents the dimension of gene labels; H_F, H_{in}, H_{out}, and H_{MLP} represent the dimension of gene embedding feature vectors. GRN, gene regulatory network; SNV, single nucleotide variant; CNA, copy number aberration; MLP, multilayer perceptron; DGCN, directed graph convolutional network.

Since $A_{in}(i,j)$ is the sum of the input edges both of genes v_i and v_j, *i.e.*, $\sum_k A\{i \leftarrow k \rightarrow j\}$, which reflects the in-degree similarity between genes v_i and v_j. The greater the value of $A_{in}(i,j)$, the higher the similarity of the second-order in-degree. Similarly, $A_{out}(i,j)$ measures the second-order out-degree proximity by accumulating the links from both genes v_i and v_j, *i.e.*, $\sum_k A(i \rightarrow k \leftarrow j)$. If there are no shared genes linked from gene v_i to gene v_j, the second-order proximity is set to zero.

MLP

MLP is a forward-structured artificial neural network, which maps a set of input vectors to a set of output embedding vectors. The feature matrix X of genes is inputted into the MLP to get the embedding matrix Z_{MLP}.

$$Z_{MLP} = \sigma(XW) \tag{4}$$

where X is the feature matrix of genes, W is a shared trainable weight matrix, and $\sigma(\cdot)$ is an activation function.

Fully connected neural network

The first-order proximity convolutional output Z_F, the second-order in-degree proximity convolutional output Z_{in}, the second-order out-degree output Z_{out} from DGCN, and the output matrix Z_{MLP} from MLP are concatenated to form an embedding matrix Z, which is fed into a fully connected neural network to obtain the probability of a gene as a cancer driver gene.

$$Z = Concat(Z_F, Z_{in}, Z_{out}, Z_{MLP}) \tag{5}$$

In summary, the DGMP model can be written as follows:

$$Y = f(X,A) = \mathrm{softmax}\left(Concat\left(\mathrm{ReLU}\begin{pmatrix}\tilde{A}_F X\Theta^{(0)}\\ \tilde{A}_{in} X\Theta^{(0)}\\ \tilde{A}_{out} X\Theta^{(0)}\\ XW^{(0)}\end{pmatrix}\Theta^{(1)}\right)\right) \tag{6}$$

Three different proximity convolutions on GRN share the same filter weight matrix $\Theta^{(0)} \in R^{C\times H}$. It transforms the input dimension C to the embedding size H. The weight matrix $W^{(0)} \in R^{C\times H}$ in MLP also transforms the input dimension C to the embedding size H. The outputs of three different proximity convolutions on GRN and the output of MLP are concatenated to feed into a fully connected network layer, which is used to convert the feature dimension from $4H$ to F. $\Theta^{(1)} \in R^{4H\times F}$ is an embedding-to-output weight matrix. The activation function is defined as $softmax(x_i) = exp(x_i)/\sum_i exp(x_i)$. All labeled samples are used to calculate the cross-entropy error in this semi-supervised cancer driver gene identification task.

We utilized the ADAM optimizer for training 500 epochs and set the dropout rate to 0.6, learning rate to 0.001, and

weight decay to 0.01. In our DGMP model, the dimensions of all four embedding vectors (*i.e.*, Z_F, Z_{in}, Z_{out}, and Z_{MLP}) were set to 4.

Results and discussion

Datasets

The genomic data of gene expression, gene mutation, copy number, and DNA methylation used in EMOGI work are taken to assess the performance of our DGMP model. These genomics data are collected from 29,446 samples in The Cancer Genome Atlas (TCGA) database, covering 16 different cancer types. By performing the same data preprocessing pipeline as EMOGI, we can obtain the pan-cancer gene feature matrix X ($X \in R^{N\times 64}$) in which each gene is represented with a 16×4 dimensional vector, where N is the gene number; 16 is the number of cancer types; and 4 refers to the values of four omics types (*i.e.*, SNVs, CNAs, DNA methylation, and gene expression) that are computed for each cancer type.

The GRNs of DawnNet, KEGG, and RegNetwork are selected to implement our DGMP model and other competitive methods. For the GRN of DawnNet built-in DawnRank [35], we merge all the redundant genes into single genes, and combine their corresponding edges to generate a directed gene association network (namely DawnNet) that contains 9677 genes, 176,826 directed edges, and 10,150 undirected edges. For the KEGG network, we integrated the pathways in the KEGG database [36] with the pathview tool [37] to obtain the KEGG pathway network. The KEGG network contains 4798 genes and 61,520 directed edges. For the RegNetwork, we extracted the regulatory interactions of transcription factor (TF)–TF and TF–gene from the RegNetwork data repository [38] to generate a directed network, named as RegNetwork. RegNetwork contains 20,300 TFs/genes and 148,387 regulation interaction edges. RegNetwork data repository [38] was established by integrating the documented regulatory interactions among TFs, microRNAs (miRNAs), and target genes from 25 selected databases, including five-type transcriptional and posttranscriptional regulatory relationships (*i.e.*, TF–TF, TF–gene, TF–miRNA, miRNA–TF, and miRNA–gene) for human and mouse. The data of DawnNet, KEGG, and RegNetwork can be downloaded from Tables S1–S3.

The known cancer driver genes (KCGs) are obtained from the expert-curated list in NCG [6] and IntOGen [39] to form a positive set S^+. The non-cancer driver genes can be selected from those most likely to be unassociated with cancers. We use the following criteria to recursively remove the genes from the set of all genes to get the non-cancer driver genes, thus forming a negative set S^-. 1) Removing the genes that are part of the KCGs in NCG; 2) removing the genes that present in the Online Mendelian Inheritance in Man (OMIM) disease database [40]; 3) removing the genes that associate with cancer pathways in the KEGG database [36]; and 4) removing the genes whose expression is correlated to the expression of cancer driver genes [41]. Generally, the number of non-cancer driver genes is far more than that of the KCGs. In order to avoid the bias of the prediction model toward the negative samples in the training process, we randomly sample the non-cancer driver genes from the negative set S^-, whose numbers are same as those of the KCGs. In addition, only the positive and negative samples that are included in the directed networks of DawnNet, KEGG, and RegNetwork are used for training. That is, we used 693 positive samples and 693 negative samples to train the prediction models for DawnNet network, and 406 positive samples and 406 negative samples for KEGG network, and 826 positive samples and 826 negative samples for RegNetwork network.

Performance comparison of DGMP with other methods

In this work, we take the AUROC and AUPRC metrics to evaluate the prediction power of different methods. AUROC value is defined as the area under the ROC curve, which plots the false positive rate (FPR) against the true positive rate (TPR) at different thresholds. AUPRC value is defined as the area under the PR curve, which plots the ratio of true positives among all positive predictions for each given recall rate. AUPRC is a more significant quality metric than AUROC for identifying the cancer driver genes, because it punishes much more the existence of false-positive cancer driver genes among the best-ranked prediction scores [42].

To evaluate the performance of our DGMP for identifying the pan-cancer driver genes, we first compared our DGMP with the machine learning-based methods of NRFD [28], EMOGI [29], and DeepWalk [43] + SVM on the DawnNet network in 5-fold cross-validation (5CV) test, and then compared it with the network-based methods of PageRank [44] and HotNet2 [13], and the mutation frequency-based method of MutSigCV [8]. For the 5CV test [45], all the labeled genes (*i.e.*, KCGs and the non-cancer driver genes selected according to some criteria) are randomly partitioned into 5 non-overlapping subsets of roughly equal size. One of these subsets is singled out in turn as the test set and the other four subsets are used as the training sets. This process is repeated for 5 iterations until all the labeled genes are tested in turn. In addition, we also performed these methods on KEGG and RegNetwork networks. The results of our DGMP and other six methods on three directed networks of DawnNet, KEGG, and RegNetwork are shown in **Table 1**. The ROC curves and PR curves of these seven methods on DawnNet, KEGG, and RegNetwork networks are shown in Figures S1–S6, and the statistical results of significance between DGMP and other six methods are shown in Figure S7, respectively.

From Table 1 and Figures S1–S7, we can see that the performance of our DGMP outperforms all of the other six state-of-the-art methods of EMOGI, NRFD, DeepWalk + SVM, HotNet2, PageRank, and MutSigCV for identifying the cancer diver genes on three directed networks of DawnNet, KEGG, and RegNetwork. The AUPRC and AUROC values of DGMP on the DawnNet network are 0.875 and 0.889, which are 0.054–0.355 and 0.053–0.284 higher than those of the other six methods, respectively; AUPRC and AUROC values of DGMP on KEGG pathway network are 0.854 and 0.876, which are 0.035–0.272 and 0.043–0.307 higher than those of other six methods, respectively; AUPRC and AUROC values of DGMP on RegNetwork network are 0.915 and 0.904, which are 0.097–0.279 and 0.086–0.228 higher than those of other six methods, respectively. These results demonstrate that our DGMP method has superior performance in identifying the cancer driver genes.

Table 1 AUPRCs and AUROCs of DGMP and other six methods on three GRNs in 5CV test

Method	DawnNet		KEGG		RegNetwork	
	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
DGMP	0.875	0.885	0.854	0.876	0.915	0.904
EMOGI	0.821	0.832	0.819	0.833	0.818	0.818
NRFD	0.788	0.792	0.758	0.795	0.790	0.748
DeepWalk + SVM	0.770	0.788	0.755	0.801	0.816	0.789
PageRank	0.755	0.722	0.746	0.698	0.817	0.801
HotNet2	0.778	0.724	0.801	0.759	0.784	0.738
MutSigCv	0.520	0.601	0.582	0.569	0.636	0.676

Note: For DeepWalk + SVM, we first transform the directed GRN network to the undirected network, and then use node2vec to learn the embedding vector for every gene. The learned embedding vectors of genes are fed into a radial basis function SVM classifier for identifying the cancer driver genes. For PageRank, we fix the known cancer driver genes as the seed genes and set their probability as 1, and then implement RWR on GRN network to prioritize the genes for predicting the cancer driver genes. KEGG, Kyoto Encyclopedia of Genes and Genomes; AUPRC, the area under the precision–recall curve; AUROC, the area under the receiver operating characteristic curve; SVM, support vector machine; GRN, gene regulatory network; 5CV, 5-fold cross-validation.

To further assess the performance of DGMP, we designed other two scenarios using the unbalanced positive and negative training samples to train the prediction models on DawnNet and STRING PPI [46] networks. That is, we used 693 cancer driver genes and 1763 non-cancer driver genes to train the prediction models on the DawnNet network and utilized 734 cancer driver genes and 1152 non-cancer driver genes to train the prediction models on the STRING PPI network that contains 12,412 genes (Table S4). The experimental results of two scenarios in the 5CV test are shown in Table S5, from which we can see that DGMP achieves the best performance on two networks in terms of AUROC and AUPRC values compared with all other methods, demonstrating the effectiveness of DGMP for identifying the cancer driver genes.

In order to evaluate the generalization performance of our DGMP, we built an independent test set of the cancer driver genes from the CancerMine database [47]. CancerMine is a literature-mined resource of cancer-related genes, which collects the genes of drivers, oncogenes, and tumor suppressors in different types of cancer. The genes in the independent test set (Table S6) do not overlap with those in the training set. To calculate the AUPRC and AUROC values, we counted hits in the independent test set as true positive samples, and all other predicted genes not contained in the independent test set as false positive samples. The results of our DGMP and other comparison methods in the independent test are shown in **Table 2**, from which we can see that DGMP achieves the best performance on DawnNet and RegNetwork networks compared with all other methods. Although AUPRC and AUROC values of DGMP on the KEGG network are slightly lower than those of the DeepWalk + SVM method, the TPR of DGMP is higher than that of the deepwalk + SVM method. For the independent test that only knows the driver gene labels, TPR is more objective than AUPRC and AUROC to measure the generalization performance of prediction models, because we counted all other unlabeled driver genes as the non-cancer driver genes in the process of calculating AUPRC and AUROC, while all these unlabeled (or unannotated) genes are not really non-cancer driver genes. In addition, AUPRC and AUROC values of DeepWalk + SVM on the KEGG network are slightly higher than those of our DGMP, the reason may be that DeepWalk + SVM obtains better performance on a smaller network, while our DGMP achieves superior performance on larger networks. These results in independent tests further demonstrate the power of DGMP to identify the cancer driver genes.

Ablation experiments of diverse architecture components in DGMP

To evaluate the contributions of diverse architecture components in our DGMP, we conducted ablation experiments on the DawnNet network in the 5CV test. The ablation experimental results of DGMP are shown in **Table 3**. In Table 3, $DGMP_{-direction}$ denotes that we neglect the directionality of regulation edges in DawnNet directed network, and adopt DGMP to identify the cancer driver genes. MLP denotes that we remove the DGCN module from the DGMP model

Table 2 Results of DGMP and other six methods on three GRNs in independent test

Method	DawnNet			KEGG			RegNetwork		
	AUPRC	AUROC	TPR	AUPRC	AUROC	TPR	AUPRC	AUROC	TPR
DGMP	0.136	0.706	0.769	0.107	0.635	0.801	0.115	0.757	0.864
EMOGI	0.107	0.659	0.747	0.100	0.602	0.767	0.103	0.684	0.847
NRFD	0.122	0.659	0.706	0.092	0.594	0.781	0.092	0.698	0.843
DeepWalk + SVM	0.124	0.696	0.671	0.119	0.659	0.790	0.098	0.697	0.662
PageRank	0.101	0.589	–	0.089	0.545	–	0.069	0.648	–
HotNet2	0.079	0.512	–	0.101	0.538	–	0.055	0.555	–
MutSigCv	0.053	0.491	–	0.073	0.483	–	0.031	0.447	–

Note: Since the outputs of PageRank, HotNet2, and MutSigCV methods are the rank orders of genes, not gene labels, we cannot calculate their TPRs. TPR = TP/(TP + FN). TPR, true positive rate; TP, true positive; FN, false negative; –, division operation.

Table 3 The ablation experimental results of DGMP on DawnNet network in 5CV test

	AUPRC	AUROC
DGMP	0.875 ± 0.020	0.885 ± 0.019
$DGMP_{-direction}$	0.871 ± 0.0017	0.881 ± 0.018
MLP	0.801 ± 0.017	0.839 ± 0.018
DGCN	0.828 ± 0.018	0.856 ± 0.022
$DGCN_{-direction}$	0.825 ± 0.019	0.838 ± 0.018
$DGCN_{-X}$	0.758 ± 0.013	0.743 ± 0.016
$DGCN_{-X}MLP$	0.841 ± 0.021	0.857 ± 0.016

Note: MLP, multilayer perceptron.

architecture to identify the cancer driver genes; DGCN denotes that we remove the MLP module from DGMP model architecture to identify the cancer driver genes; $DGCN_{-direction}$ denotes that we neglect the directionality of regulation edges in DawnNet directed network, and adopt DGCN to identify the cancer driver genes. $DGCN_{-X}$ denotes that we just use the topological information of GRN in the DGCN module without considering the genomic information (*i.e.*, SNVs, CNAs, DNA methylation, and gene expression) of genes, and also remove the MLP module from DGMP model architecture to identify the cancer driver genes; $DGCN_{-X}MLP$ denotes that we combine $DGCN_{-X}$ with MLP and the full connected neural network to identify the cancer driver genes.

As shown in Table 3, we can see that the AUPRC of DGMP that consists of DGCN and MLP is 0.875, which is 0.047 and 0.074 higher than that of DGCN and MLP, respectively, indicating that jointing DGCN and MLP can improve the performance of identifying the cancer driver genes. AUPRC of $DGCN_{-X}MLP$ is 0.083 higher than that of $DGCN_{-X}$, indicating that MLP does mitigate the bias toward the graph topological features in the DGCN learning process, further enhancing the prediction performance of DGMP. AUPRC of DGCN is 0.07 higher than that of $DGCN_{-X}$, indicating that feeding the multi-omics information of genes into DGCN can improve the prediction performance of DGMP. AUPRC of $DGMP_{-direction}$ is 0.004 lower than that of DGMP, and $DGCN_{-direction}$ is 0.003 and 0.05 lower than that of DGCN and DGMP, respectively, indicating that there is a trend to improve the performance of GCN by considering the regulation information (*i.e.*, the directionality of regulation edges). The results in Table 3 show that all the proposed components for building DGMP are valid and contribute to the final performance of DGMP, and jointing MLP and DGCN can effectively improve the performance of identifying cancer driver genes.

Previous studies [33,34] show that if the features of neighbors of a center node are not similar, further graph convolutional operation will result in lower performance for GCN. Considering that the neighbors' features of some center genes in GRN may not be more similar than other genes, performing graph convolutional operation on these neighbors will reduce the performance of DGCN, we introduce MLP to offset the performance degradation of DGCN. In order to further show the effectiveness of MLP mitigating the bias toward the graph topological features in the DGCN learning process, the neighborhood discrete entropy $Score_{etp}(u)$ [34] (its definition and formulation are given in File S1) was used to measure the diversity of neighborhoods of a gene. We picked the top 50 predicted cancer driver genes and the top 50 predicted non-cancer driver genes with the highest discrete entropy and then employed t-distributed stochastic neighbor embedding (t-SNE) to visualize the distribution (**Figure 2**) of these genes by extracting their embedding vectors (*i.e.*, Z_F, Z_{in}, Z_{out}, and Z_{MLP}). From Figure 2, we can see that after concatenating these embedding vectors (*i.e.*, Z_F, Z_{in}, and Z_{out}) generated by DGCN with the embedding vectors (*i.e.*, Z_{MLP}) generated by MLP, the within-class variance of cancer/non-cancer driver genes is smaller than that of DGCN, and the between-class distances of cancer/non-cancer driver genes are larger than those of DGCN in embedding space. These results demonstrate that graph convolutional operation is not so good for distinguishing these genes when the features of their neighbors are dissimilar, while MLP can offset the performance degradation of DGCN, that is, MLP can mitigate the bias toward the graph topological features in DGCN learning process.

De novo cancer driver gene analysis

To analyze the *de novo* cancer driver genes, we selected the genes that are not annotated as driver genes in the NCG database [6] from top 100 cancer driver genes (Table S7) predicted by DGMP on DawnNet, and considered these genes as the newly predicted cancer driver genes (namely NPCGs). As a result, we obtained 52 NPCGs (Table S8), of which 41 NPCGs are labeled as the cancer driver genes, or oncogene/tumor suppressor genes in the CancerMine database [47]. For example, transcriptional activator Sp3 as a driver gene competes with the tumor suppressor AP-2 for binding the VEGF promoter in prostate cancer, thereby repressing AP-2 expression [48]; TFF2 expression inhibits the gastric cancer cell growth and invasion *in vitro* via interactions with the transcription factor Sp3, and Sp3 knockdown in gastric cancer cells antagonizes TFF2 antitumor activity [49]. NFKB1 is a tumor suppressor in cervical cancer by inhibiting cell proliferation, colony formation, and migration, and its mutation will affect the radiotherapy sensitivity in cervical cancer [50]. GBP1 is downregulated and acts as a tumor suppressor in colorectal cancer cells [51]. These lines of evidence show that our DGMP can effectively predict the new and candidate cancer driver genes (CCGs). We also designed the following experiments to demonstrate the effectiveness of our DGMP in predicting *de novo* cancer driver genes.

Firstly, we calculated the interaction percentages of KCGs from the NCG database [6] with NPCGs, CCGs [6], and KCGs, as well as the interaction percentage of KCGs with other genes that neither belong to NPCGs nor KCGs and CCGs. The statistical results are shown in **Figure 3**A. As shown in Figure 3A, we can find that genes in NPCGs generally have more interactions with KCGs than other genes, indicating that the NPCGs predicted by DGMP are closely related to the initiation and progression of cancers. We also found that SP1 and SHC1 have the largest number of interactions with KCGs. Among them, SP1 is strongly associated with Ser345-phosphorylated PR-B receptors to regulate growth-promoting (EGFR) target genes and PR cell cycle (p21) for breast cancer cell proliferation [52]; SHC1 may be an important route of DEPDC1B regulating the development of bladder cancer. In DEPDC1B-overexpressed cancer cells, the knockdown of SHC1 could abolish the promotion effects

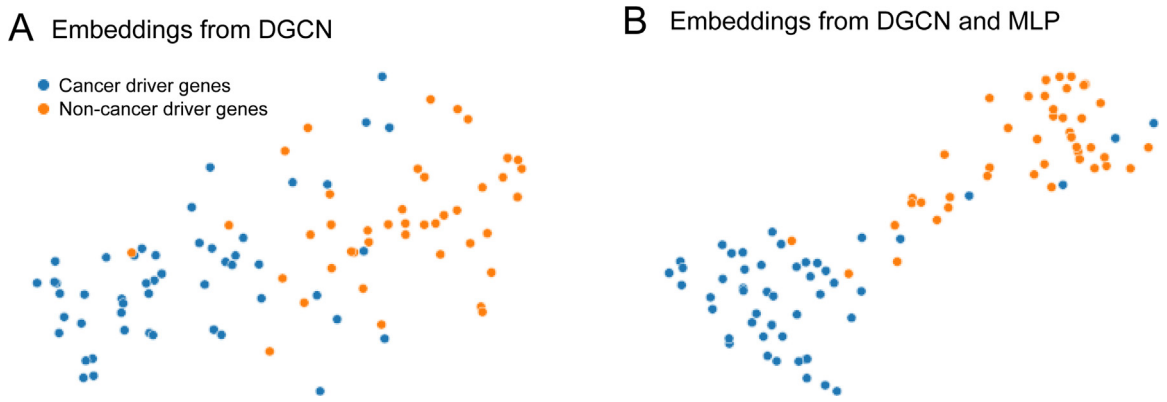

Figure 2 t-SNE visualization of top 50 predicted cancer/non-cancer driver genes with high neighbor discrete entropy
A. Visualization of 50 cancer/non-cancer driver genes by concatenating Z_F, Z_{in}, and Z_{out} from DGCN. **B.** Visualization of 50 cancer/non-cancer driver genes by concatenating Z_F, Z_{in}, and Z_{out} from DGCN and Z_{MLP} from MLP. t-SNE, t-distributed stochastic neighbor embedding.

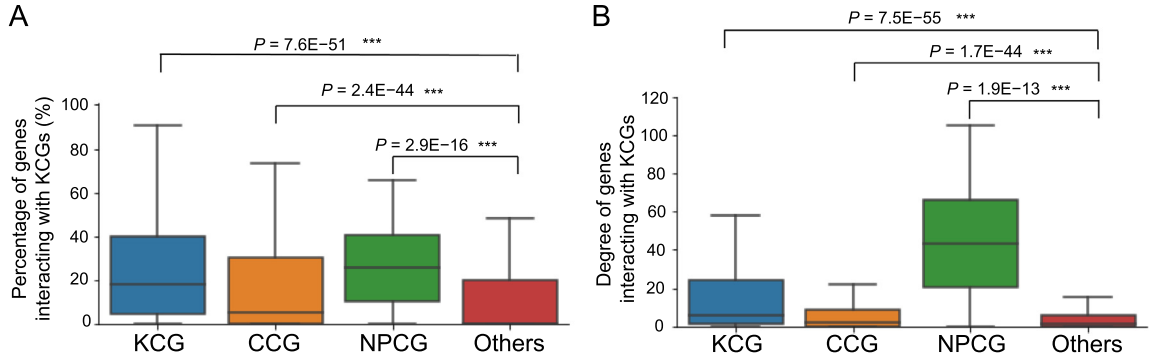

Figure 3 Statistical results of KCGs, CCGs, NPCGs, and other genes interacting with KCGs
A. Percentage of KCGs, CCGs, NPCGs, and other genes interacting with KCGs. **B.** Degree of KCGs, CCGs, NPCGs, and other genes interacting with KCGs. Mann-Whitney U test was carried out to test the difference significance of each comparison. The *P* values between the two methods are marked on their connecting lines. KCG, known cancer driver gene; CCG, candidate cancer driver gene; NPCG, newly predicted cancer driver gene.

caused by DEPDC1B [53]. In addition, we also compared the degrees of genes in NPCGs, KCGs, and CCGs with the degrees of other genes that neither belong to NPCGs nor KCGs and CCGs. As shown in Figure 3B, we can see that the degrees of genes in NPCGs, KCGs, and CCGs are significantly larger than those in other gene sets, indicating that the greater the network degree of a gene, the more likely it is to be as a cancer driver gene.

Secondly, we analyzed the multi-omics gene features of NPCGs, KCGs, CCGs, and other genes by calculating the average frequency of SNVs, average CNA rate, average DNA methylation change, and average gene differential expression across 16 cancer types. The mutation types contain both truncating and gain-of-function mutations, and all SNVs have the potential to affect cell growth.

As shown in **Figure 4**, we found that there are significant differences in SNVs (P = 2.1E–04), CNAs (P = 0.04), gene differential expression (P = 4.3E–03) between NPCGs and other genes, whereas there is no significant difference (P = 0.2) between NPCGs and other genes for DNA methylation. These results indicate that the omics gene features of SNVs, CNAs, and gene differential expression are important factors in distinguishing NPCGs from other genes. The NPCGs identified by DGMP are more frequently mutated across samples than other genes, indicating that DGMP can effectively identify the cancer driver genes from highly mutated genes. For example, *EGF* and *ROCK1* are the high mutation rate genes across different cancer types in NPCGs, which are demonstrated to be correlated with cancers [54,55]. EGF enhances the phosphorylation and acetylation of histone H3 to promote the DKK1 transcription in hepatocellular carcinoma [54]; ROCK1 is overexpressed in human hepatocellular carcinoma (HCC) cell lines and tissues, and knockdown of ROCK1 or ROCK2 can inhibit the HCC cell growth [55]. In addition, NPCGs also contain highly DNA methylation genes and highly mRNA differential expressed genes, which rarely mutate across cancers. For example, the *ITGB3* gene has high DNA methylation but with relatively low SNV frequency, whose high expression is correlated with overall survival and worse progression-free survival of multiple myeloma patients [56]; the *TFAP2A* gene is highly overexpressed compared to normal tissue in multiple cancer types, which modulates ferroptosis in gallbladder carcinoma cells through the Nrf2 signaling axis [57]. These results show that our DGMP can not

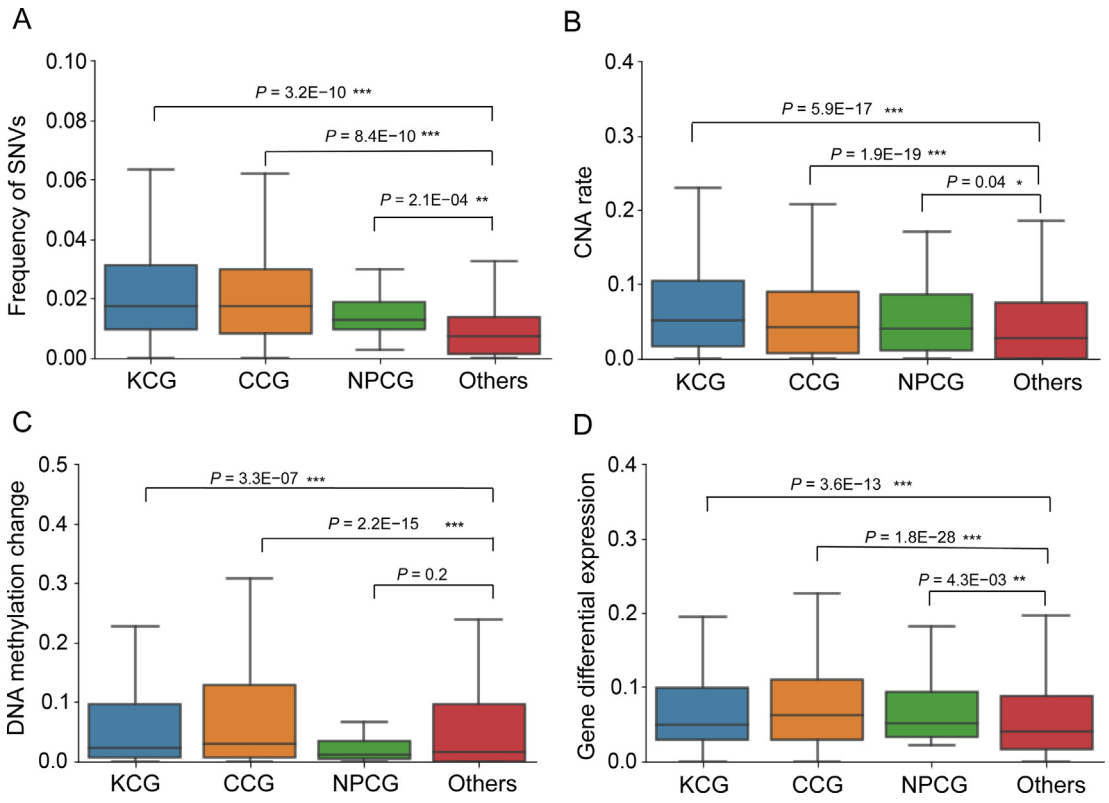

Figure 4 Averaged SNVs, CNAs, DNA methylation change, gene differential expression across 16 cancer types for NPCGs, KCGs, CCGs, and other genes

A. The average frequency of SNVs. **B.** The average CNA rate. **C.** The average DNA methylation change. **D.** The average gene differential expression. Mann–Whitney U test was carried out to test the difference significance of each comparison, and the corresponding *P* values between the two methods were marked on their connecting lines. The average DNA methylation change refers to the average of differences of methylation signals between cancer and normal samples across all samples of a cancer type, and the average gene differential expression refers to the $\log_2$ fold change between a gene's expression values in normal and cancer samples and then averaged across all samples of a cancer type.

only identify the driver genes involved in GRN with other known cancer genes, but also the highly mutated cancer driver genes or driver genes harboring other kinds of alterations (*e.g.*, differential expression and aberrant DNA methylation).

Conclusion

In this work, we presented a novel method of DGMP to identify cancer driver genes by integrating multi-omics genomic data and jointing the DGCN and MLP. DGMP uses DGCN to learn the multi-omics features of genes as well as the topological structure features of GRN and employs MLP to learn the gene features from multi-omics genomic data. Three embedding feature vectors from DGCN and one embedding feature vector from MLP are concatenated to feed into a fully connected neural network to output the probability that one gene is a cancer-driver gene. The results on three networks of DawnNet, KEGG, and RegNetwork show that DGMP is superior to other existing state-of-the-art methods in identifying the cancer driver genes. The ablation experimental results indicate that introducing MLP into DGCN can offset the performance degradation of DGCN, and considering the directionality of regulation edges has a trend to improve the performance of GCN. Jointing MLP and DGCN can effectively improve the performance of identifying cancer driver genes. The analysis results of the top 100 predicted cancer driver genes demonstrate that DGMP not only identifies more KCGs, but also can effectively predict the highly mutated cancer driver genes, and the driver genes harboring other kinds of alterations (*e.g.*, differential expression and aberrant DNA methylation), or genes involved in GRN with other cancer genes. The t-SNE visualization distribution of 50 cancer/non-cancer driver genes with high neighbor discrete entropy shows that concatenating the embedding feature vectors of DGCN and MLP can improve the aggregation of cancer/non-cancer driver genes, that is, MLP indeed mitigates the bias toward the graph topological features in DGCN learning process. In addition, DGMP can also be used to successfully identify driver genes of specific cancers, such as breast cancer and thyroid cancer (Figure S8).

There are two potential reasons which are responsible for the remarkable performance of DGMP. The first aspect is that DGMP makes full use of the regulation information among genes by adopting the DGCN model. The second aspect is that DGMP introduces MLP to weight more on gene features for mitigating the bias toward the graph topological structure features in DGCN learning process, which offsets the

performance degradation of DGCN caused by the convolutional operation of GCN on the neighbor genes with dissimilar features.

Although DGMP has achieved good performance in pan-cancer driver gene prediction, it can be improved from the following two aspects. First, DGMP averages the SNVs, CNAs, DNA methylation, and gene expression of all samples to obtain 4-omics features to represent the genes for each cancer, which will ignore the specific characteristics of individual cancer patients. Second, DGMP just simply concatenated three embedding feature vectors from DGCN and one embedding feature vector from MLP, whereas the contributions of these four feature vectors may be different. Thus, we can consider using the weighted fusion way to improve the performance of DGMP.

Code availability

The source code and datasets used in this work can be downloaded from https://github.com/NWPU-903PR/DGMP.

CRediT author statement

Shao-Wu Zhang: Conceptualization, Methodology, Investigation, Formal analysis, Writing - review & editing. **Jing-Yu Xu:** Conceptualization, Methodology, Software, Investigation, Writing - original draft. **Tong Zhang:** Investigation, Formal analysis, Writing - original draft. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

We would like to thank Drs. Yan Li and Wei-Feng Guo for their work on the manuscript revision. This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62173271 and 61873202 to SWZ).

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2022.11.004.

ORCID

ORCID 0000-0003-1305-7447 (Shao-Wu Zhang)
ORCID 0000-0001-6556-9503 (Jing-Yu Xu)
ORCID 0000-0002-2909-5854 (Tong Zhang)

References

[1] Dinstag G, Shamir R. PRODIGY: personalized prioritization of driver genes. Bioinformatics 2020;36:1831–9.
[2] Shrestha R, Hodzic E, Sauerwald T, Dao P, Wang K, Yeung J, et al. HIT'nDRIVE: patient-specific multidriver gene prioritization for precision oncology. Genome Res 2017;27:1573–88.
[3] Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr LA, Kinzler KW. Cancer genome landscapes. Science 2013;339:1546–58.
[4] Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive characterization of cancer driver genes and mutations. Cell 2018;174:1034–5.
[5] ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature 2020;578:82–93.
[6] Repana D, Nulsen J, Dressler L, Bortolomeazzi M, Venkata SK, Tourna A, et al. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol 2019;20:1.
[7] Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. Nat Rev Cancer 2018;18:696–705.
[8] Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–8.
[9] Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics 2013;29:2238–44.
[10] Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 2014;505:495–501.
[11] Jiang R. Walking on multiple disease-gene networks to prioritize candidate genes. J Mol Cell Biol 2015;7:214–30.
[12] Zhou Y, Wang S, Yan H, Pang B, Zhang X, Pang L, et al. Identifying key somatic copy number alterations driving dysregulation of cancer hallmarks in lower-grade glioma. Front Genet 2021;12:654736.
[13] Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet 2015;47:106–14.
[14] Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nat Rev Genet 2017;18:551–62.
[15] Chen Y, Jiang T, Jiang R. Uncover disease genes by maximizing information flow in the phenome–interactome network. Bioinformatics 2011;27:i167–76.
[16] Jiang R, Gan M, He P. Constructing a gene semantic similarity network for the inference of disease genes. BMC Syst Biol 2011;5:S2.
[17] Zhang D, Bin YN. DriverSubNet: a novel algorithm for identifying cancer driver genes by subnetwork enrichment analysis. Front Genet 2021;11:10.
[18] Zhang T, Zhang SW, Li Y. Identifying driver genes for individual patients through inductive matrix completion. Bioinformatics 2021;37:4477–84.
[19] Guo WF, Zhang SW, Liu LL, Liu F, Shi QQ, Zhang L, et al. Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. Bioinformatics 2018;34:1893–903.
[20] Guo WF, Zhang SW, Zeng T, Li Y, Gao JX, Chen LN. A novel network control model for identifying personalized driver genes in cancer. PLoS Comput Biol 2019;15:27.
[21] Guo WF, Zhang SW, Zeng T, Akutsu T, Chen LN. Network control principles for identifying personalized driver genes in cancer. Brief Bioinform 2020;21:1641–62.
[22] Guo WF, Zhang SW, Feng YH, Liang J, Zeng T, Chen LN. Network controllability-based algorithm to target personalized

driver genes for discovering combinatorial drugs of individual patients. Nucleic Acids Res 2021;49:e37.
[23] Cheng FX, Zhao JF, Zhao ZM. Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. Brief Bioinform 2016;17:642–56.
[24] Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics 2011;27:2147–8.
[25] Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the evaluation of cancer driver genes. Proc Natl Acad Sci U S A 2016;113:14330–5.
[26] Luo P, Ding YL, Lei XJ, Wu FX. deepDriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. Front Genet 2019;10:13.
[27] Rogers MF, Gaunt TR, Campbell C. Prediction of driver variants in the cancer genome via machine learning methodologies. Brief Bioinform 2021;22:bbaa250.
[28] Liu C, Dai Y, Yu K, Zhang ZK. Enhancing cancer driver gene prediction by protein-protein interaction network. IEEE/ACM Trans Comput Biol Bioinform 2022;19:2231–40.
[29] Schulte-Sasse R, Budach S, Hnisz D, Marsico A. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. Nat Mach Intell 2021;3:513–26.
[30] Qin S, Ma F, Chen LM. Gene regulatory networks by transcription factors and microRNAs in breast cancer. Bioinformatics 2015;31:76–83.
[31] Coghlin C, Murray GI. The role of gene regulatory networks in promoting cancer progression and metastasis. Future Oncol 2014;10:735–48.
[32] Tong Z, Liang Y, Sun C, Rosenblum D, Lim A. Directed graph convolutional network. arXiv 2020;2004.13970.
[33] Qian Y, Expert P, Rieu T, Panzarasa P, Barahona M. Quantifying the alignment of graph and features in deep learning. IEEE Trans Neural Netw Learn Syst 2022;33:1663–72.
[34] Xie Y, Li S, Yang C, Wong R, Han J. When do GNNs work: understanding and improving neighborhood aggregation. Proceedings of the 29th International Joint Conference on Artificial Intelligence 2020:1303–9.
[35] Hou JP, Ma J. DawnRank: discovering personalized driver genes in cancer. Genome Med 2014;6:56.
[36] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 1999;27:29–34.
[37] Luo WJ, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics 2013;29:1830–1.
[38] Liu ZP, Wu C, Miao H, Wu H. RegNetwork: an integrated database of transcriptional and posttranscriptional regulatory networks in human and mouse. Database 2015;2015:bav095.
[39] Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Alba Jene-Sanz A, et al. IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods 2013;10:1081–2.
[40] McKusick VA. Mendelian inheritance in man and its online version. OMIM Am J Hum Genet 2007;80:588–604.
[41] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.
[42] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning 2006:233–40.
[43] Perozzi B, Al-Rfou R, Skiena S. DeepWalk: online learning of social representations. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2014:701–10.
[44] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems 1998;30:107–17.
[45] Zhang TH, Zhang SW. Advances in the prediction of protein subcellular locations with machine learning. Curr Bioinform 2019;14:406–21.
[46] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019;47:D607–13.
[47] Lever J, Zhao EY, Grewal J, Jones MR, Jones SJM. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. Nat Methods 2019;16:505–7.
[48] Egener T, Roulet E, Zehnder M, Bucher P, Mermod N. Proof of concept for microarray-based detection of DNA-binding oncogenes in cell extracts. Nucleic Acids Res 2005;33:e79.
[49] Cai Y, Yi M, Chen D, Liu J, Guleng B, Ren J, et al. Trefoil factor family 2 expression inhibits gastric cancer cell growth and invasion *in vitro* via interactions with the transcription factor Sp3. Int J Mol Med 2016;38:1474–80.
[50] Yang D, Zhang W, Liang J, Ma K, Chen P, Lu D, et al. Single cell whole genome sequencing reveals that *NFKB1* mutation affects radiotherapy sensitivity in cervical cancer. Oncotarget 2018;9:7332–40.
[51] Lan Q, Wang A, Cheng Y, Mukasa A, Ma J, Hong L, et al. Guanylate binding protein-1 mediates EGFRvIII and promotes glioblastoma growth *in vivo* but not *in vitro*. Oncotarget 2016;7:9680–91.
[52] Faivre EJ, Daniel AR, Hillard CJ, Lange CA. Progesterone receptor rapid signaling mediates serine 345 phosphorylation and tethering to specificity protein 1 transcription factors. Mol Endocrinol 2008;22:823–37.
[53] Lai CH, Xu KX, Zhou JH, Wang MR, Zhang WY, Liu XH, et al. DEPDC1B is a tumor promotor in development of bladder cancer through targeting SHC1. Cell Death Dis 2020;11:986.
[54] Niu J, Li W, Liang C, Wang X, Yao X, Yang RH, et al. EGF promotes DKK1 transcription in hepatocellular carcinoma by enhancing the phosphorylation and acetylation of histone H3. Sci Signal 2020;13:eabb5727.
[55] Wu H, Chen YY, Li B, Li C, Guo J, You J, et al. Targeting ROCK1/2 blocks cell division and induces mitotic catastrophe in hepatocellular carcinoma. Biochem Pharmacol 2021;184:114353.
[56] Li YQ, Zhang LL, Gong JC. Relation among EGFL7, ITGB3, and KLF2 and their clinical implication in multiple myeloma patients: a prospective study. Ir J Med Sci 2022;191:1995–2001.
[57] Huang HX, Yang G, Yang Y, Yan J, Tang XY, Pan Q. TFAP2A is a novel regulator that modulates ferroptosis in gallbladder carcinoma cells via the Nrf2 signalling axis. Eur Rev Med Pharmacol Sci 2020;24:4745–55.