*Original Research*

# A network-based method for identifying cancer driver genes based on node control centrality

**Feng Li[1], Han Li[1], Junliang Shang[1], Jin-Xing Liu[1], Lingyun Dai[1], Xikui Liu[2] and Yan Li[2]**

[1]School of Computer Science, Qufu Normal University, Rizhao 276826, China; [2]Department of Electrical Engineering and Information Technology, Shandong University of Science and Technology, Jinan 250031, China
Corresponding author: Yan Li. Email: liyan@sdkd.net.cn

### Impact Statement

The complex mechanisms of cancer lead to difficulty in diagnosis and treatment. Coding and non-coding cancer drivers play crucial roles in the occurrence and progression of cancer. Therefore, we attempt to design a Network-based Method for identifying cancer Driver Genes based on node Control Centrality (NMDGCC). The results show that NMDGCC identifies a large number of coding and non-coding cancer drivers and has better performance than existing methods. Our research can help precision medicine and provide better treatment options for cancer patients.

### Abstract

Cancer is one of the major contributors to human mortality and has a serious influence on human survival and health. In biomedical research, the identification of cancer driver genes (cancer drivers for short) is an important task; cancer drivers can promote the progression and generation of cancer. To identify cancer drivers, many methods have been developed. These computational models only identify coding cancer drivers; however, non-coding drivers likewise play significant roles in the progression of cancer. Hence, we propose a Network-based Method for identifying cancer Driver Genes based on node Control Centrality (NMDGCC), which can identify coding and non-coding cancer driver genes. The process of NMDGCC for identifying driver genes mainly includes the following two steps. In the first step, we construct a gene interaction network by using mRNAs and miRNAs expression data in the cancer state. In the second step, the control centrality of the node is used to identify cancer drivers in the constructed network. We use the breast cancer dataset from The Cancer Genome Atlas (TCGA) to verify the effectiveness of NMDGCC. Compared with the existing methods of cancer driver genes identification, NMDGCC has a better performance. NMDGCC also identifies 295 miRNAs as non-coding cancer drivers, of which 158 are related to tumorigenesis of BRCA. We also apply NMDGCC to identify driver genes related to the different breast cancer subtypes. The result shows that NMDGCC detects many cancer drivers of specific cancer subtypes.

**Keywords:** Cancer, driver gene, node control centrality, interaction network

***Experimental Biology and Medicine*** **2022; 248: 232–241. DOI: 10.1177/15353702221139201**

## Introduction

Cancer is driven by different driver genes and is a complex disease. Some systemic cancer genomics projects have been used for cancer research, such as The Cancer Genome Atlas (TCGA),[1] and International Cancer Genome Consortium (ICGC)[2] have generated a lot of genomic data to provide a basis for cancer research. In biomedical research, identifying cancer drivers is one of the most important tasks. Identifying cancer drivers can help design better treatments for cancer patients. Research has shown that the occurrence and development of cancers are closely related to gene mutations. Gene mutations have many different types. Normal cells become tumor cells after undergoing gene mutations, thereby promoting the development of cancer. For example, BRCA1 and AKT1 can cause breast cancer when they are mutated.[3] Although gene mutations can cause cancer, not all gene mutations are related to cancer. Some mutations have nothing to do with cancer, which are called passenger mutations.[4] The mutations which can promote the occurrence and progression of cancer are driver mutations. The genes with driver mutation are considered as driver genes.[5] But some genes with no mutations can promote the development of cancer, which are also considered as cancer driver genes. Not only coding genes are related to cancer, but non-coding genes also account for most of the genome, and they are also closely associated with cancer.

Around this biological research task, some methods of identifying driver genes have been developed. We have divided these methods into two main categories based on their characteristics: (1) mutation-based methods and (2) network-based methods. Because the gene mutation rate in tumor cells is significantly higher than in normal cells,[6] the mutation-based methods identify cancer drivers by using

mutation characteristics and the network-based methods rely on the importance and influence of each gene to identify cancer drivers in different networks. In the mutation-based approaches, CoMEt[7] identifies cancer genes by testing the mutual exclusivity of genes. OncodriveFM[8] detects driver genes by using the functional effect of mutated genes. DriverSub[9] uses mutation data of genes through the subspace learning framework to identify cancer drivers. MutSigCV[10] predicts driver genes by evaluating the significance of mutations. DriverML[11] detects cancer drivers by using the functional effect of mutations and a machine learning approach. CHASM[12] predicts driver mutations by using a machine learning technique (random forest). WeSME[13] identifies driver genes by assessing the mutual exclusivity of mutations. HetRCNA[14] identifies cancer drivers by using a matrix decomposition framework. However, because the mutation data is incomplete, the mutation-based methods often have limitations in the identification of cancer drivers.

The second type of network-based methods rely on indicators such as the influence and importance of genes in the network to identify driver genes. DriverNet[15] predicts cancer drivers by combining different omics data to evaluate the influence of mutations on the transcriptional network. MEMo[16] uses mutation and network information to predict cancer drivers. TieDIE[17] predicts cancer drivers by using network diffusion. NetBox[18] uses biological networks to detect cancer drivers. Tri-NMF[19] uses an unsupervised learning model to predict cancer drivers. DawnRank[20] applies the PageRank algorithm to the gene network for detecting driver genes. CBNA[21] uses the controllability of complex networks to find critical nodes in the network, which are considered as cancer drivers. There is also a method that integrates expression data and mutation data of genes into a network to identify cancer drivers.[22] However, the network-based methods often use a general network, not a specific network for certain cancer types. In addition, most of these methods often only identify coding cancer drivers.

Recently, with the widespread application of control theory and the development of network science, many methods are developed to evaluate the importance of nodes in the network, such as control range[23] and control centrality.[24] The control range is the control ability exhibited by the node while maintaining the overall controllability of the network. However, control centrality is the maximum control ability that a node has in the network. Hence, control centrality better reflects the importance of nodes in the network than the control range. We design a Network-based Method for identifying cancer Driver Genes based on node Control Centrality[24] (NMDGCC), which can identify both coding driver genes and non-coding cancer drivers. Based on microRNAs (miRNAs), mRNAs, and transcription factors (TFs) expression data of breast invasive carcinoma (BRCA) from TCGA, a specific gene regulatory network for breast cancer is constructed. Then, we filter out those edges that are not in the protein–protein interaction (PPI) network,[25] TargetScan,[26] and TransmiR.[27] Based on the constructed network, we take advantage of the control centrality of complex network to identify the nodes with high control centrality values and consider them as candidate cancer drivers. Because some gene mutations are associated with the progression of cancer, we use the mutation frequency to rank candidate cancer drivers. The candidate cancer drivers with high mutation frequency are considered as driver genes. We use the BRCA dataset from TCGA and compare NMDGCC with the other four methods of identifying cancer drivers to test its effectiveness. To further assess the capabilities of NMDGCC, we also apply it to predict driver genes for different cancer subtypes.

## Materials and methods

### Data collection

We obtain 747 tumor samples of the BRCA dataset from TCGA. The expression data of mRNAs, miRNAs, and TFs are obtained from tumor samples to construct the gene regulatory network. As the dataset we obtained has a large number of coding genes, we need to screen for these coding genes by using the PPI network. Finally, the expression data of 5168 mRNAs, 839 TFs, and 1719 miRNAs are selected to construct the gene regulatory network. The PPI network[25] is used to obtain the coding genes and refine the network. The TF list[28] is used to obtain TFs from coding genes. We refine the network by using several gene interaction databases, including TargetScan[26] version 7.0 to refine miRNA-mRNA/TF interaction, and TransmiR[27] version 2.0 to refine TF-miRNA interaction. We also use the Cancer Gene Census (CGC)[29] from the COSMIC database[30] as the gold standard for identified coding cancer drivers. The mutation data of breast cancer provided by TCGA[1] is used to calculate mutation frequency; among the mutation classification, we use functional mutation data.

### The overview of NMDGCC

The overview of NMDGCC is described in Figure 1. NMDGCC contains two stages: (1) Construct the gene regulatory network: (a) Construct mRNA-TF-miRNA regulatory network by using expression data and (b) Refine the gene interaction network by using several interaction databases and PPI network, and (2) Identify coding driver genes and miRNA drivers: (a) Calculate the control centrality values for each node and (b) Identify candidate driver genes for cancer patients.

### Construction of the gene interaction network

We utilize the expression data of mRNAs, miRNAs, and TFs to build a network. Each node in the network represents a gene, which can be miRNA, TF, or mRNA. We calculate the Pearson correlation coefficients (PCC) between all pairs of nodes. If the absolute value of PCC between two nodes is larger than or equal to the average of the absolute values of PCC between all nodes in the network, then there exists an edge between them.

### Refinement of gene regulatory network

In order to make the constructed network more reliable, we remove some false-positive edges in the constructed network. We utilize PPI network[25] and several gene interaction databases to refine the interaction network. The edges

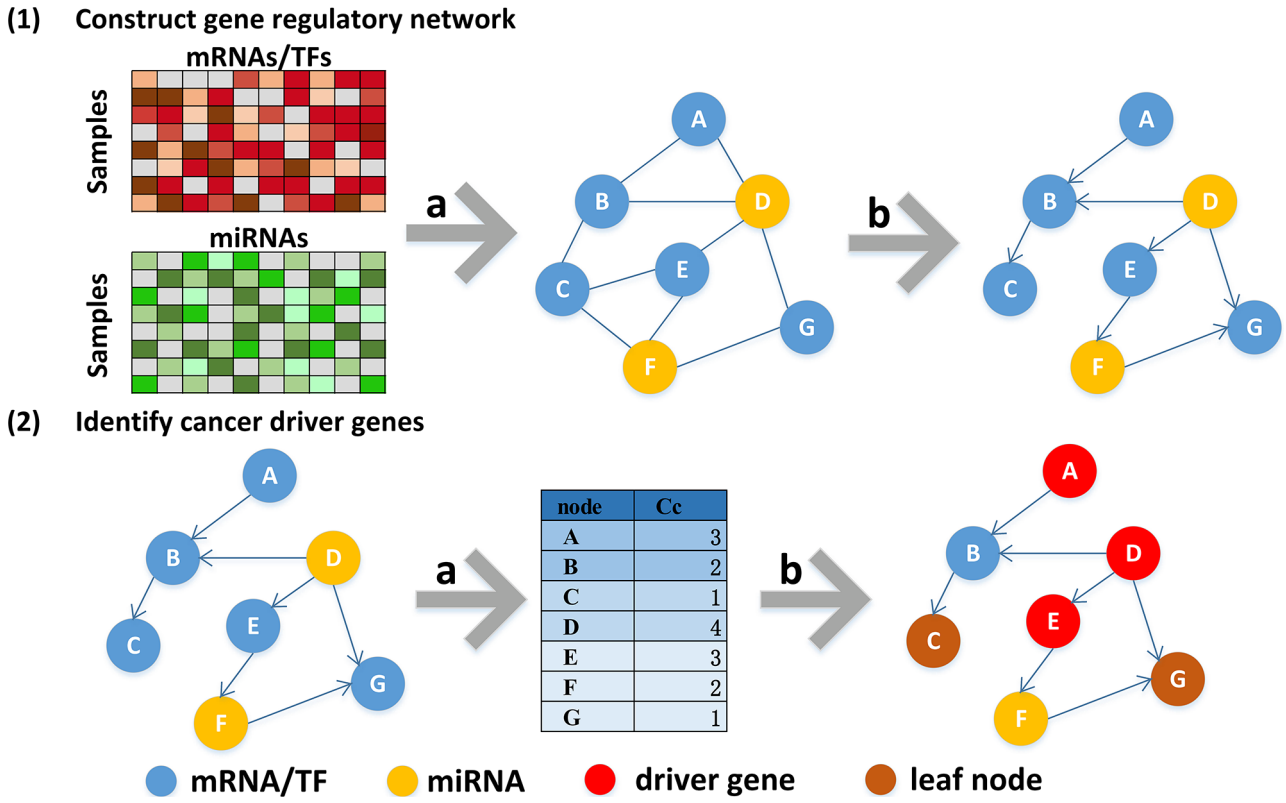


**Figure 1.** The overview of NMDGCC for identifying cancer drivers. (1) Construct gene regulatory network: (a) construct gene regulatory network by using expression data, (b) refine the gene regulatory network by using several interaction databases and PPI network, and (2) identify coding and non-coding driver genes: (a) calculate the control centrality values for each node and (b) identify candidate cancer drivers.

between mRNA-mRNA and TF-TF/mRNA in the network are removed if they are not in the PPI network. The edges between TF-miRNA in the network are removed if they are not in TransmiR.[27] The edges between miRNA-TF/mRNA in the network are removed if they are not in TragetScan.[26] The final interaction network contains 7726 nodes (including 5168 mRNAs, 839 TFs, and 1719 miRNAs) and 90,169 edges (including 814 TF-mRNA edges, 1228 TF-TF edges, 10,082 miRNA-TF edges, 11,249 mRNA-mRNA edges, 29,956 TF-miRNA edges, and 36,840 miRNA-mRNA edges).

## Control centrality

Based on the constructed network, we utilize control centrality[24] to assess the importance of cancer drivers. In the network we constructed, the control centrality ($Cc$) of a node (gene) represents the capability of controlling the constructed interaction network. The larger the control centrality, the larger the control ability of node. Nodes with large control centrality values are more important than other nodes in the network. The N-node directed weighted network describes a complex system that can be modeled as

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{1}$$

where $A$ represents an $N \times N$ matrix that describes the network. The element $a_{ij}$ is the strength of the node $j$ that can affect the node $i$ in the matrix $A$. $x(t)$ represents the state of the node at time $t$. $B$ represents an $N \times M$ input matrix that detects the controlled nodes. $u(t)$ is the time-dependent input vector with $M$ independent signals. The $(A,B)$ denotes the system (1), and the $C = (B, AB, \ldots, A^{N-1}B)$ denotes the controllability matrix. In the system $(A,B)$, the $\mathrm{rank}(C)$ denotes the rank of the controllability matrix $C$, and it represents the dimension of the controllable subspace. When we control node $i$ only, the $\mathrm{rank}(C^{(i)})$ represents the ability of node $i$ to control the system. If node $i$ cannot control other nodes except itself, then $\mathrm{rank}(C^{(i)}) = 1$. We consider elements of $A$ and $B$ are independent free parameters or zeros. The $\mathrm{rank}_g(C)$ represents the generic rank of the $C$. For each node $i$, its control centrality can be defined as

$$C_C(i) \equiv \mathrm{rank}_g(C^{(i)}) \tag{2}$$

We can calculate the computation of $\mathrm{rank}_g(C)$ by solving a combinatorial optimization problem. Then, connecting $N$ state nodes and $M$ input nodes forms a directed graph $G(A,B)$. The $G_s$ represents stem-cycle disjoint subgraph of $G(A,B)$, and it contains cycles and stems only. The stem is a directed path, and there are no duplicate nodes in the stem. The $\mathrm{rank}_g(C)$ can be calculated by

$$\mathrm{rank}_g(C) = \max_{G_s \in G} |E(G_s)| \tag{3}$$

where $|E(G_s)|$ is the number of edges in the subgraph $G_s$. In the directed network, we can calculate $\max_{G_s \in G} |E(G)|$ by solving a linear programming problem.[31]

### Identification of coding drivers and non-coding drivers

Gene mutations are often associated with cancer. Therefore, we regard frequently mutated genes with high control centrality in the network as cancer drivers. The control centrality of the leaf node is 1, which means that it can only control itself in the network. This kind of node has little influence on the network, then we remove the nodes with control centrality of 1 (leaf nodes). Calculate the average of the control centrality values of all other nodes. Consider the nodes whose control centrality values are larger than the average value as candidate cancer drivers. Then, we use the mutation frequency to rank them. The higher the mutation frequency of the identified driver genes, the higher the ranking.

### Evaluation metrics

We use three measures to assess the performance of the five methods. These three metrics are *Precision*, *Recall*, and $F_1Score$, respectively. *Precision* represents the fraction of the validated driver genes in the identified cancer drivers. *Recall* represents the fraction of validated driver genes in the gold standard. $F_1Score$ calculates the harmonic mean of *Recall* and *Precision*. The mathematical formula for the three measures is as follows

$$Precision = \frac{tp}{tp + fp} \tag{4}$$

$$Recall = \frac{tp}{tp + fn} \tag{5}$$

$$F_1\text{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

where $tp$ represents the number of detected driver genes in the gold standard, $fp$ represents the number of detected driver genes that are not in the gold standard, and $fn$ represents the number of genes in the gold standard that are not in the detected driver genes.

## Results

We calculate the control centrality ($Cc$) values of each node in the constructed regulatory network by utilizing concept of control centrality. After calculating the $Cc$ values of all nodes, we remove the leaf nodes with $Cc$ value of 1. In the network, the leaf node only controls itself and does not control other nodes. Then, we calculate the average values of the $Cc$ values of other nodes. If the $Cc$ values of a node are larger than average values, we consider this node to be a node with high $Cc$ values. The higher the $Cc$ values of a node, the more nodes it can control in the network. Therefore, nodes with high $Cc$ values are very important in the network, so they are thought to be candidate cancer drivers. In the gene interaction network we constructed, a total of 703 nodes with high $Cc$ values are identified and considered as candidate cancer drivers.

### Comparison analysis with other methods on identifying coding driver genes

We apply NMDGCC to the BRCA dataset from TCGA and compare the performance with four existing methods. These methods include mutation-based method ActiveDriver[32] and network-based methods DriverNet,[15] DawnRank,[20] and CBNA.[21] Because most methods only identify coding driver genes, we only compare them with mutated coding driver genes identified by NMDGCC.

The cancer genes in CGC[29] are used as the gold standard to assess the performance of each method according to the predicted cancer drivers. The more the driver genes predicted by each method in the CGC, the better the performance of the method. Therefore, we compare NMDGCC with four identification methods on three aspects: (1) the number of validated cancer drivers in the predicted top (50, 100, 150, and 200) cancer drivers; (2) *Precision*, *Recall*, and $F_1score$ of five methods based on the predicted top (50, 100, 150, 200) cancer drivers; and (3) the total number of predicted cancer drivers and the fraction of the total predicted cancer drivers in CGC.

To facilitate the performance comparison of these five methods, we first compare the number of validated driver genes in the top (50, 100, 150, and 200) cancer drivers predicted by each method.

We validate the cancer drivers identified by five methods by CGC (Figure 2). The number of cancer drivers detected by NMDGCC and CBNA validated with the gold standard are similar for the top 50 identified drivers. Among the top (50, 100, 150, 200) cancer drivers detected by these five methods, our method has more cancer driver genes validated with CGC than the other four methods. In general, NMDGCC performs better than the other four methods. Furthermore, our method also identifies novel cancer driver genes, and these drivers can be prioritized as candidates for further experiments.

Then, to evaluate the effectiveness of the NMDGCC, we also compute *Precision*, *Recall*, and $F_1Score$ based on CGC and top (50, 100, 150, 200) identified cancer drivers by NMDGCC and other four methods. It can be seen from the results (Figure 3) that NMDGCC outperforms the other four methods in the three measures.

In order to further test the performance of the above methods in predicted cancer drivers, we calculate the total number of predicted cancer drivers and the fraction of the total identified cancer driver genes in the gold standard (Figure 4). The total number of cancer drivers detected by CBNA, DawnRank, and NMDGCC is all the genes they predicted. The total number of cancer drivers predicted for ActiveDriver and DriverNet is the genes with $P < 0.05$. The result shows that while the total number of cancer drivers identified by NMDGCC is less, the fraction of correct prediction validated by CGC is higher than other methods.

The cancer driver genes identified through these methods are all important genes. Therefore, these methods identified cancer driver genes and may have overlap parts. Then, we compare the five methods of identifying driver genes to find the overlap parts (Figure 5). For this experiment, we use validated driver genes of each method. From the diagram, these methods identify some of the same driver genes, but

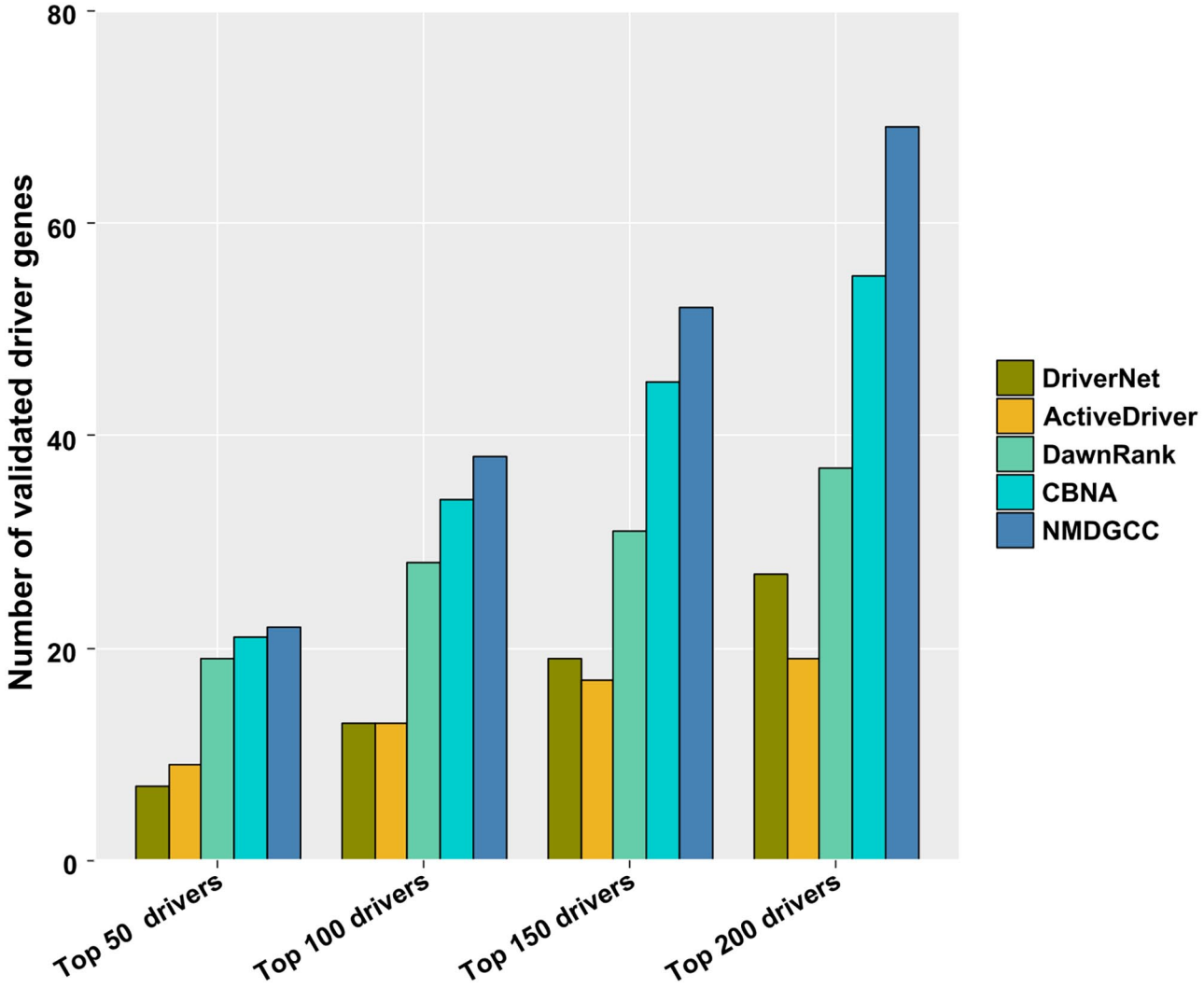


**Figure 2.** Comparison of the number of coding cancer drivers identified by five methods validated with CGC. In the figure, each bar represents the number of validated driver genes.

NMDGCC identifies many cancer drivers that cannot be identified by other methods. Therefore, NMDGCC can be combined with other methods to jointly advance the identification of cancer drivers and cancer research.

### Identification of miRNA cancer drivers

The occurrence and progression of cancer are not only associated with coding drivers but also with non-coding drivers. Therefore, NMDGCC also identifies miRNA cancer drivers. NMDGCC has identified a total of 295 miRNA cancer drivers as non-coding drivers, 158 of them have been confirmed involved in tumorigenesis of BRCA by OncomiR.[33] In the predicted top 10 miRNA cancer drivers, 6 miRNA drivers have been validated to be associated with tumorigenesis (Table 1).

### Identification of coding cancer drivers without mutation

The proposed method also explores coding cancer drivers without mutations. We apply gene ontology (GO) enrichment analysis to assess the ability of the NMDGCC in discovering coding driver genes without mutations through an online software of DAVID.[34] In the enrichment analysis results, we select the top 10 GO biological processes terms and top 10 GO molecular functions terms involved with the highest number of coding drivers and select the top 20 cancer drivers based on their occurrence in these GO terms. The results show that our method identified cancer drivers which are involved in numerous GO molecular functions and GO biological processes (Figure 6). The corresponding terms of the top 10 biological processes and top 10 molecular functions are provided in Tables 2 and 3. This promising result suggests that the driver genes predicted by NMDGCC are biologically meaningful.

### Rank of coding driver genes with mutation based on mutation density

In our method, final ranking of predicted driver genes is based on the mutation frequency, which may produce some false positives. The reason for this phenomenon is the length of the gene. If a gene is longer, its mutation frequency may be higher, and it will be ranked higher in the results. In order

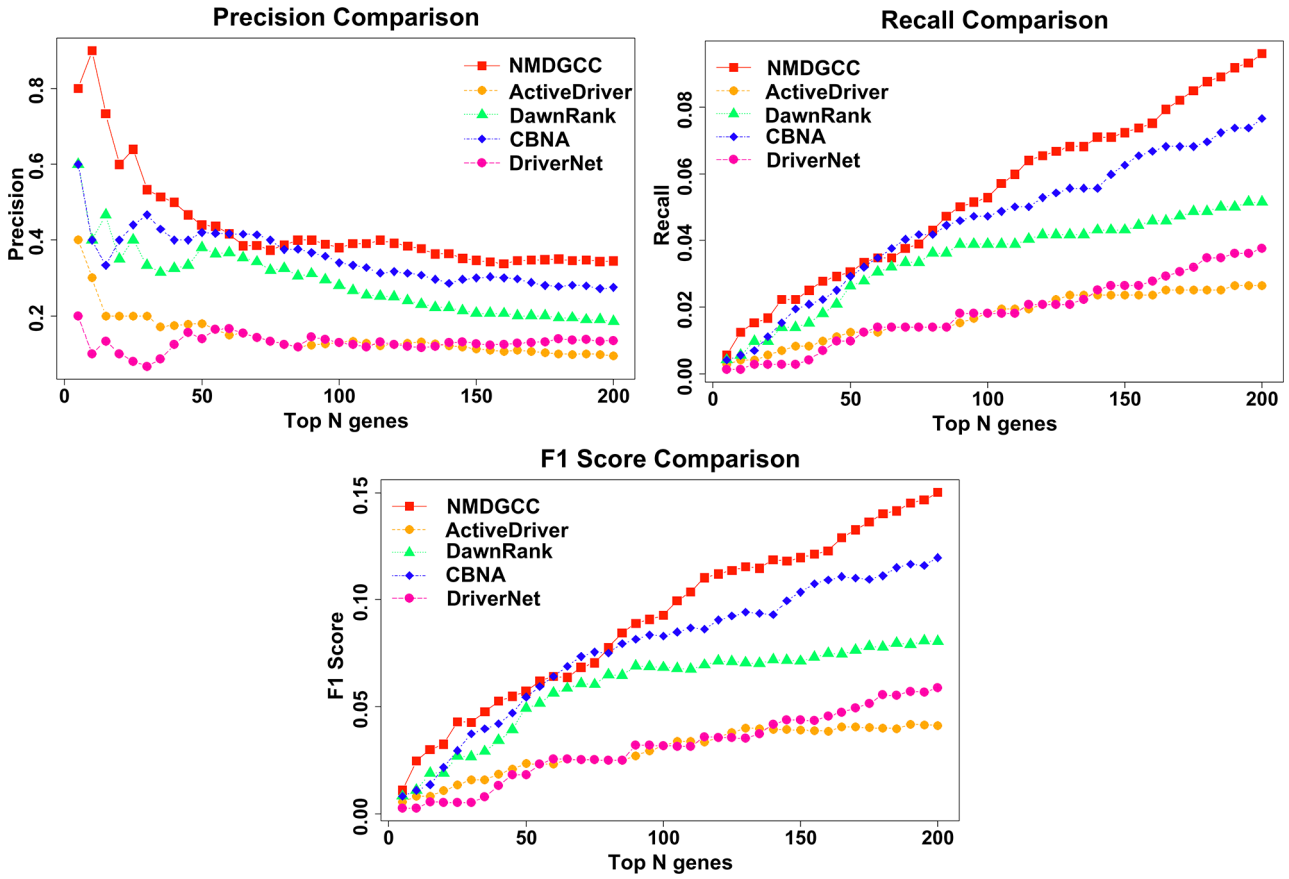


**Figure 3.** Comparison of *Precision*, *Recall*, and *F*-*score* of five methods identified top cancer drivers. The x-axis denotes the predicted top cancer drivers by each method. The y-axis denotes the value of three measures.

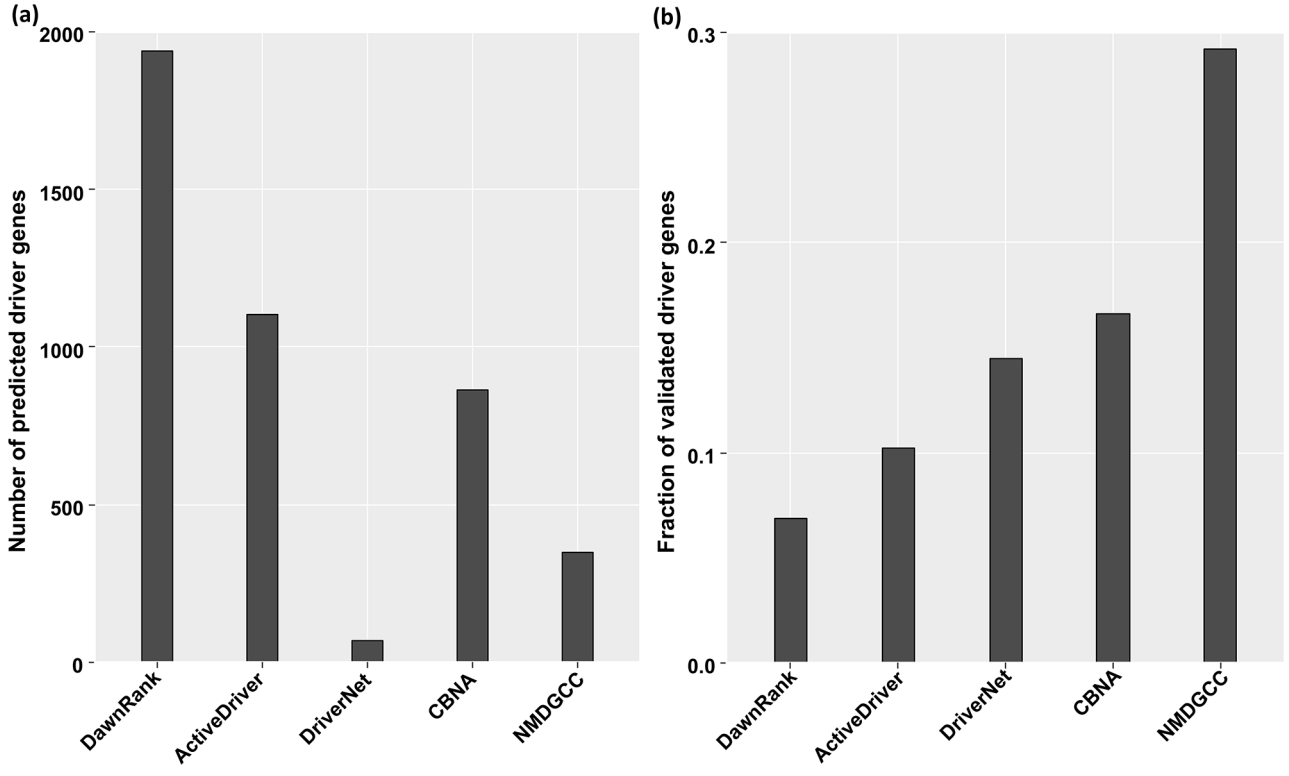


**Figure 4.** Comparison of the total number of cancer drivers identified by five methods: (a) the total number of predicted driver genes and (b) fraction of driver genes validated with gold standard.

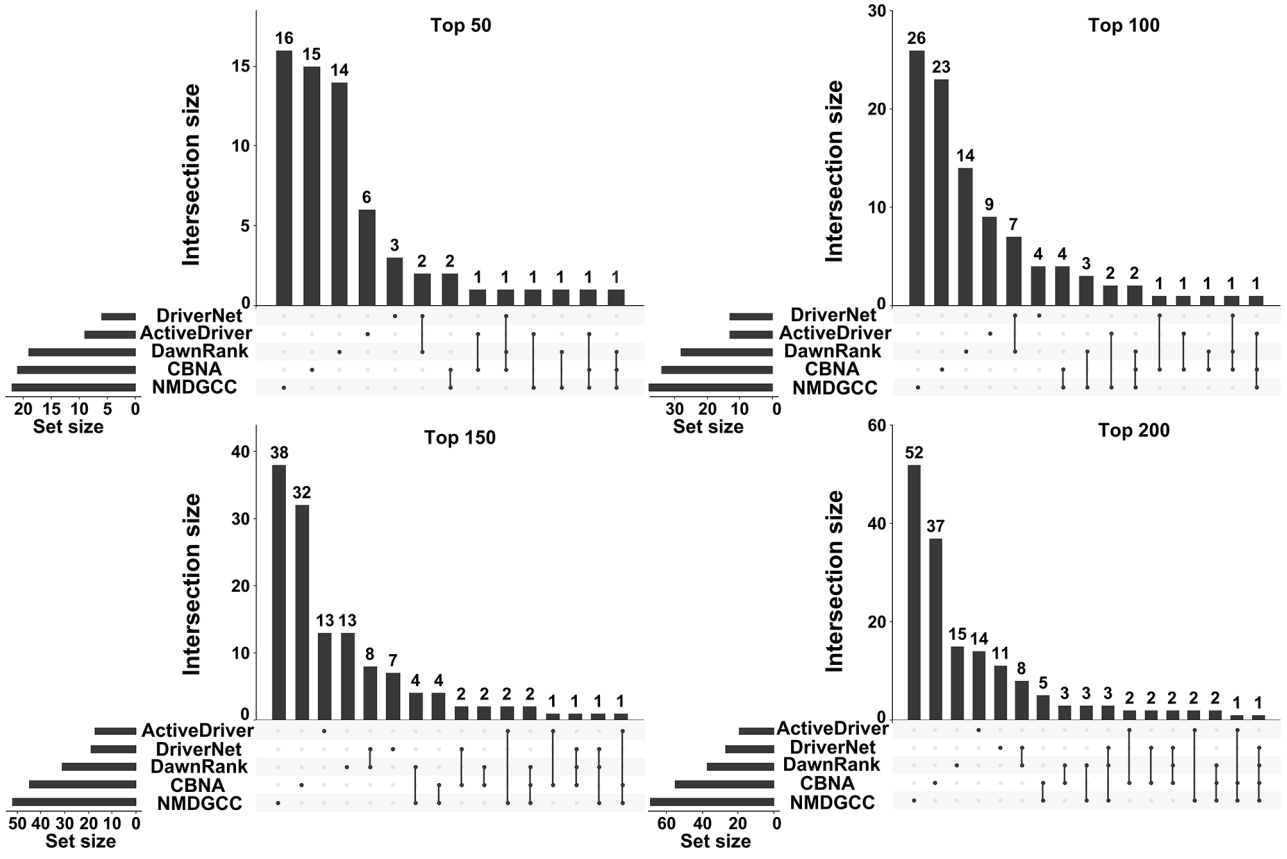


**Figure 5.** Overlap between predicted top-k driver genes by five methods. The horizontal bars at the bottom left represent the number of confirmed cancer drivers using CGC. The dotted lines and vertical bars indicate the number of confirmed cancer drivers which overlap with each other.

**Table 1.** miRNA drivers predicted by NMDGCC.

| No. | Identified miRNA drivers | Confirmed |
|---|---|---|
| 1 | hsa-miR-137 | √ |
| 2 | hsa-miR-520b | |
| 3 | hsa-miR-520d-5p | |
| 4 | hsa-miR-520e | |
| 5 | hsa-let-7a-5p | √ |
| 6 | hsa-let-7d-5p | √ |
| 7 | hsa-let-7e-5p | √ |
| 8 | hsa-let-7f-5p | |
| 9 | hsa-let-7g-5p | √ |
| 10 | hsa-let-7i-5p | √ |

miRNA: microRNA; NMDGCC: Network-based Method for identifying cancer Driver Genes based on node Control Centrality; √: miRNA was associated with the tumorigenesis of BRCA.

to have a more reliable ranking, we rank predicted driver genes by using the mutation density. The mutation density is the ratio of mutation count to the length of the gene. Among the top 10 driver genes ranked using mutation density, there are 4 genes in CGC which are TP53, FOXA1, GATA3, and HOXA11 (Table 4).

## Identification of driver genes specific to cancer condition

The network is constructed in our method by using the data of tumor samples, and the cancer driver genes are nodes with high control centrality[24] values in this network. These nodes are important in the constructed networks based on cancer data. There are some differences between the gene expression of normal people and cancer patients. But the nodes with high control centrality in the network constructed by the gene expression data of normal samples may also be nodes with high control centrality values in the cancer patient network. Hence, we identify the cancer drivers specific to the cancer condition. We construct a gene interaction network by utilizing gene expression data from normal samples and find nodes with high $Cc$ values in this network. Then compare with the node with high $Cc$ values in the cancer state and identify driver genes that are only specific to cancer conditions. As the number of coding genes without mutation and miRNA specific to cancer state is very small, we only analyze coding genes with mutation. We have identified a total of 33 cancer drivers specific to cancer condition, of which 7 are in CGC (Table 5).

## Prediction of driver genes for different cancer subtypes

There are many different subtypes of breast cancer, and patients with different subtypes of cancer will have different clinical manifestations and survival outcomes. Different driver genes may cause different cancer subtypes. To provide better treatments for patients with different cancer subtypes, we analyze the driver genes specific to the cancer subtype. First, the Pam50[35] method is used to divide breast cancer into

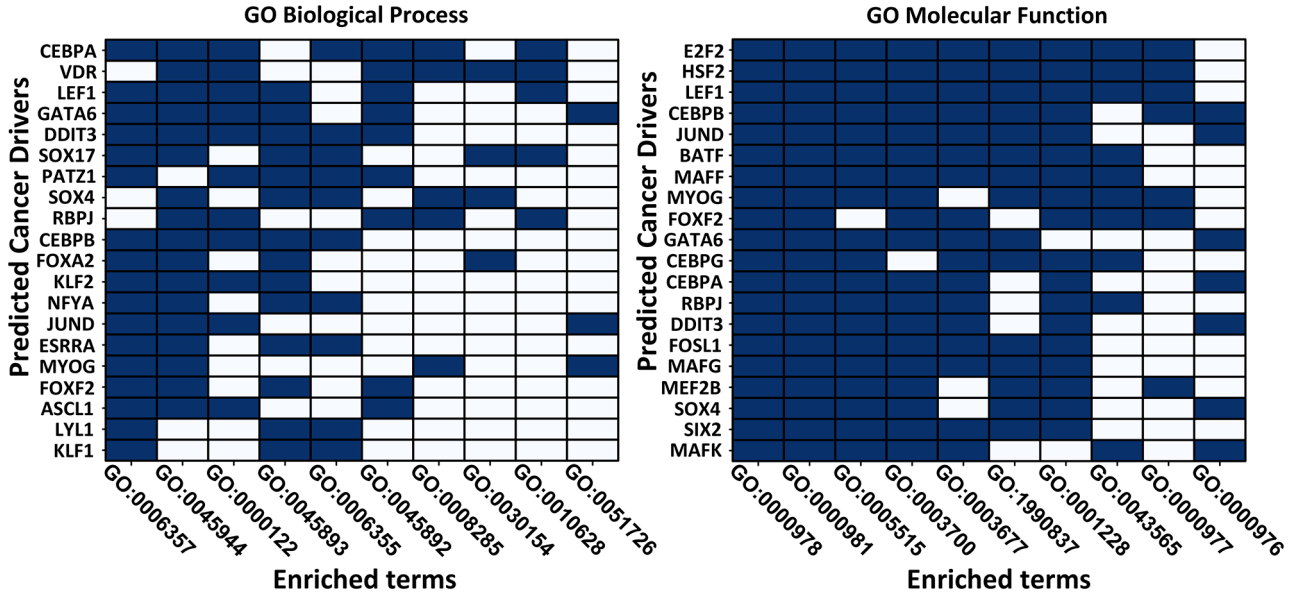


**Figure 6.** (a) and (b) The heatmap of the top 20 identified coding drivers across top 10 GO biological processes and GO molecular functions enriched terms, respectively. Blue cells in the matrix indicate that a driver gene is related to an enriched term.

**Table 2.** The top 10 enriched terms are shown in Figure 6(a) and corresponding biological processes.

| Enriched term | Biological process |
|---|---|
| GO:0006357 | Regulation of transcription from RNA polymerase II promoter |
| GO:0045944 | Positive regulation of transcription from RNA polymerase II promoter |
| GO:0000122 | Negative regulation of transcription from RNA polymerase II promoter |
| GO:0045893 | Positive regulation of transcription, DNA-templated |
| GO:0006355 | Regulation of transcription, DNA-templated |
| GO:0045892 | Negative regulation of transcription, DNA-templated |
| GO:0008285 | Negative regulation of cell proliferation |
| GO:0030154 | Cell differentiation |
| GO:0010628 | Positive regulation of gene expression |
| GO:0051726 | Regulation of cell cycle |

**Table 3.** The top 10 enriched terms are shown in Figure 6(b) and corresponding molecular functions.

| Enriched term | Molecular function |
|---|---|
| GO:0000978 | RNA polymerase II core promoter proximal region sequence-specific DNA binding |
| GO:0000981 | RNA polymerase II transcription factor activity, sequence-specific DNA binding |
| GO:0005515 | Protein binding |
| GO:0003700 | Transcription factor activity, sequence-specific DNA binding |
| GO:0003677 | DNA binding |
| GO:1990837 | Sequence-specific double-stranded DNA binding |
| GO:0001228 | Transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding |
| GO:0043565 | Sequence-specific DNA binding |
| GO:0000977 | RNA polymerase II regulatory region sequence-specific DNA binding |
| GO:0000976 | Transcription regulatory region sequence-specific DNA binding |

**Table 4.** The top 10 mutated coding drivers using mutation density ranked.

| No. | Identified cancer drivers | In CGC |
|---|---|---|
| 1 | TP53 | √ |
| 2 | FOXA1 | √ |
| 3 | GATA3 | √ |
| 4 | MAZ | |
| 5 | HOXA11 | √ |
| 6 | HOXA5 | |
| 7 | ATF4 | |
| 8 | IRF3 | |
| 9 | DLX5 | |
| 10 | FOXD3 | |

CGC: Cancer Gene Census; √: driver gene validated with CGC.

**Table 5.** List of validated coding drivers of specific cancer condition.

| No. | Identified driver gene |
|---|---|
| 1 | ZFHX3 |
| 2 | HOXA11 |
| 3 | ZNF521 |
| 4 | HOXA13 |
| 5 | MLLT1 |
| 6 | MLLT3 |
| 7 | SKI |

five subtypes, namely Her2, Basal, Normal-like, Luminal A, and Luminal B. After dividing the 747 samples of the BRCA dataset using the Pam50 method, we obtain 158 Basal subtype samples, 221 Luminal A subtype samples, 108 Her2 subtype samples, 165 Luminal B subtype samples, and 95 Normal-like subtype samples. Then, we use the NMDGCC to identify cancer drivers of different cancer subtypes. If the mutation frequency in one subtype is higher than the mutation frequency in any other subtype, it indicates that the

**Table 6.** Predicted driver genes are specific to each breast cancer subtype.

| Subtype | Predicted driver genes |
|---|---|
| LuminalA | NCOR1, RUNX1, ARID1A, CBFB, FOXA1, BRCA1, HIVEP1, NCOA6, MYB, TCF12 |
| LuminalB | GATA3, MGA, TAF1, CTCF, ZFPM2, GLI2, DNMT1, MAZ, DNMT3A, MBD1 |
| Basal | TP53, CREBBP, TCF20, ARID2, STAT6, BACH2, NFATC3, AFF1, IL1RAP, CD97 |
| Her2 | MLLT4, BPTF, SMARCA2, TCF4, NCOA3, STAT4, NCOR2, ZHX2, HNF1A, PBRM1 |
| Normal-like | SMARCC1, NPAS2, ETS1, NEUROD1, GATAD2A, NR2C2, TP73, SKI |

mutation of this gene is dominant in this subtype, and the mutation of this gene is also specific to this subtype. The results show that we have identified specific driver genes for each subtype (Table 6).

## Discussion

In this work, we have designed NMDGCC, a network-based method for identifying coding cancer drivers and non-coding cancer drivers based on node control centrality. In NMDGCC, we integrate mRNAs, TFs, and miRNAs expression data to construct a gene interaction network. We also use gene interaction databases to refine the constructed network. We remove some false-positive edges to make the edges in the constructed network closer to the real gene interaction. Then, the concept of the control centrality of complex network is used to calculate control centrality values of each node in the constructed network. The centrality values of node represent the size of its controllable subspace, and the node with higher centrality values has a larger control ability. If a node with higher $Cc$ values is mutated, it may change the cell from a normal state to a tumor state. Hence, we consider nodes with significantly larger control centrality values as cancer drivers.

We apply NMDGCC to the breast cancer data and identify coding and non-coding drivers associated with the development of breast cancer. Comparing the validated coding drivers with four methods, the results show that NMDGCC has better performance. NMDGCC also identifies miRNA cancer drivers. More than 50% of the miRNA cancer drivers we identified have been confirmed to be associated with tumorigenesis of BRCA through a database. NMDGCC also succeeded in identifying the driver genes of different cancer subtypes in breast cancer. In summary, NMDGCC is an efficient method for the identification of cancer coding drivers and non-coding drivers. It can complement existing methods and jointly promote the identification and prediction of cancer drivers, and it can help in the treatment of cancer patients and provide better treatment options.

In the feature, we will apply NMDGCC to predict driver genes of other cancer types and also use more different types of data to improve NMDGCC.

### AUTHORS' CONTRIBUTIONS

FL and HL conceived and designed the study. J-XL, JS, and LD collected data. XL and YL performed the data analysis. HL and FL performed the experiments, implemented the method, and drafted the manuscript. All authors reviewed the manuscript and approved the final manuscript.

### DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### FUNDING

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work has been supported by the National Natural Science Foundation of China (61902216, 61972236, 61972226, and 61902215) and Natural Science Foundation of Shan-dong Province (No. ZR2018MF013).

### ORCID ID

Feng Li https://orcid.org/0000-0002-5556-3789

### REFERENCES

1. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The cancer genome atlas pan-cancer analysis project. *Nat Genet* 2013;**45**:1113–20
2. International Cancer Genome Consortium. International network of cancer genome projects. *Nature* 2010;**464**:993–8
3. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, Nik-Zainal S, Martin S, Varela I, Bignell GR. The landscape of cancer genes and mutational processes in breast cancer. *Nature* 2012;**486**:400–4
4. Vandin F. Computational methods for characterizing cancer mutational heterogeneity. *Front Genet* 2017;**8**:83
5. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci USA* 2016;**113**:14330–5
6. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. *Science* 2015;**349**:1483–9
7. Leiserson MD, Wu H-T, Vandin F, Raphael BJ. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol* 2015;**16**:1–20
8. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res* 2012;**40**:e169
9. Xi J, Yuan X, Wang M, Li A, Li X, Huang Q. Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication. *Bioinformatics* 2020;**36**:1855–63
10. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013;**499**:214–8
11. Han Y, Yang J, Qian X, Cheng W-C, Liu S-H, Hua X, Zhou L, Yang Y, Wu Q, Liu P. DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies. *Nucleic Acids Res* 2019;**47**:e45
12. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* 2009;**69**:6660–7
13. Kim Y-A, Madan S, Przytycka TM. WeSME: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics* 2017;**33**:814–21
14. Xi J, Li A, Wang M. HetRCNA: a novel method to identify recurrent copy number alternations from heterogeneous tumor samples based on matrix decomposition framework. *IEEE/ACM Trans Comput Biol Bioinf* 2018;**17**:422–34
15. Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, Huntsman DG, Caldas C, Aparicio SA, Shah SP. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol* 2012;**13**:1–14

16. Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res* 2012;**22**:398–406
17. Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D, Stuart JM. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* 2013;**29**:2757–64
18. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. *PLoS ONE* 2010;**5**:e8918
19. Xi J, Li A, Wang M. A novel unsupervised learning model for detecting driver genes from pan-cancer data through matrix tri-factorization framework with pairwise similarities constraints. *Neurocomputing* 2018; **296**:64–73
20. Hou JP, Ma J. DawnRank: discovering personalized driver genes in cancer. *Genome Med* 2014;**6**:56
21. Pham VVH, Liu L, Bracken CP, Goodall GJ, Long Q, Li J, Le TD. CBNA: a control theory based method for identifying coding and non-coding cancer drivers. *PLoS Comput Biol* 2019;**15**:e1007538
22. Shi K, Gao L, Wang B. Discovering potential cancer driver genes by an integrated network-based approach. *Mol Biosyst* 2016;**12**:2921–31
23. Wang B, Gao L, Gao Y. Control range: a controllability-based index for node significance in directed networks. *J Stat Mech: Theory Exp* 2012; **4**:P04011
24. Liu Y-Y, Slotine J-J, Barabási A-L. Control centrality and hierarchical structure in complex networks. *PLoS ONE* 2012;**7**:e44459
25. Vinayagam A, Stelzl U, Foulle R, Plassmann S, Zenkner M, Timm J, Assmus HE, Andrade-Navarro MA, Wanker EE. A directed protein interaction network for investigating intracellular signal transduction. *Sci Signaling* 2011;**4**:rs8
26. Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* 2015;**4**:e05005
27. Wang J, Lu M, Qiu C, Cui Q. TransmiR: a transcription factor–microRNA regulation database. *Nucleic Acids Res* 2010;**38**:D119–22
28. Lizio M, Harshbarger J, Abugessaisa I, Noguchi S, Kondo A, Severin J, Mungall C, Arenillas D, Mathelier A, Medvedeva YA. Update of the FANTOM web resource: high resolution transcriptome of diverse cell types in mammals. *Nucleic Acids Res* 2017;**45**:D737–43
29. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. *Nat Rev Cancer* 2004;**4**:177–83
30. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, Kok CY, Jia M, De T, Teague JW, Stratton MR, McDermott U, Campbell PJ. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 2015;**43**:D805–11
31. Poljak S. On the generic dimension of controllable subspaces. *IEEE Trans Autom Control* 1990;**35**:367–9
32. Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol Syst Biol* 2013;**9**:637
33. Wong NW, Chen Y, Chen S, Wang X. OncomiR: an online resource for exploring pan-cancer microRNA dysregulation. *Bioinformatics* 2018; **34**:713–5
34. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* 2003;**4**:P3
35. Liu MC, Pitcher BN, Mardis ER, Davies SR, Friedman PN, Snider JE, Vickery TL, Reed JP, DeSchryver K, Singh B, Gradishar WJ, Perez EA, Martino S, Citron ML, Norton L, Winer EP, Hudis CA, Carey LA, Bernard PS, Nielsen TO, Perou CM, Ellis MJ, Barry WT. PAM50 gene signatures and breast cancer prognosis with adjuvant anthracycline- and taxane-based chemotherapy: correlative analysis of C9741 (Alliance). *NPJ Breast Cancer* 2016;**2**:15023

*(Received August 2, 2022, Accepted October 12, 2022)*