A novel network control model for identifying personalized driver genes in cancer

Wei-Feng Guo[1], Shao-Wu Zhang[1]*, Tao Zeng[2], Yan Li[1], Jianxi Gao[3]*, Luonan Chen[1,2,4]*

[1] Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xian, 710072, China

[2] Key Laboratory of Systems Biology, Shanghai Institutes for Biological Science, Chinese Academy Science, Shanghai 200031, China

[3] Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, 12180, USA

[4] Collaborative Research Center for Innovative Mathematical Modeling, Institute of Industrial Science, University of Tokyo, Tokyo 153-8505, Japan

[*] Corresponding authors: zhangsw@nwpu.edu.cn, gaoj8@rpi.edu, lnchen@sibs.ac.cn

Abstract

Although existing computational models have identified many common driver genes, it remains challenging to identify the personalized driver genes by using samples of an individual patient. Recently, the methods of exploiting the structure-based control principles of complex networks provide new clues for identifying minimum number of driver nodes to drive the state transition of large-scale complex networks from an initial state to the desired state. However, the network control methods cannot be directly applied to identify the personalized driver genes due to the unknown network dynamics of the personalized system. Here we proposed the personalized network control model (PNC) to identify the personalized driver genes by employing the network control principle on genetic data of individual patients. In PNC model, we firstly presented a paired single sample network construction method to construct the personalized state transition network for capturing the phenotype transitions between healthy and disease state. Then, we designed a novel structure network control method from the Feedback Vertex Sets-based control perspective to identify the personalized driver genes. The experimental results on 13 kinds of cancer datasets from The Cancer Genome Atlas show that our PNC model outperforms other methods in terms of F-measures for identifying the personalized driver genes enriched in the gold-standard cancer driver gene lists. Thus PNC can provide novel insights for understanding tumor heterogeneity in individual patients.

KEYWORDS: Driver genes; Structure network control theory; Tumor heterogeneity; Individual patients.

Author summary

Notably there may be unique personalized driver genes for an individual patient in cancer. Identifying personalized driver genes that lead to cancer initiation and progression of individual patient is one of the biggest challenges in precision medicine. However, most methods for cancer-driver identification have focused mainly on the cohort information rather than on individual information and fail to identify personalized driver genes. We here proposed personalized network control model (PNC) to identify personalized driver genes by applying the network control principle on personalized data of individual patients. By considering the progression from the healthy state to the disease state as the network control problem, our PNC aims to detect a small number of personalized driver genes that are altered in response to input signals for triggering the state transition in individual patients. The impetus behind PNC contains two main respects. One is to design a paired single sample network construction method for constructing personalized state transition networks to capture the phenotypic transitions between normal and disease attractors. The other one is to develop a structure network control method on personalized state transition networks for identifying personalized driver genes which can drive individual patient system state from healthy state to disease state through oncogene activations. The experimental results on multiple cancer datasets highlight that PNC is effective for identifying personalized driver genes in cancer.

Introduction

Cancer is a heterogeneous disease that is driven by oncogene activations such as genetic mutation, gene amplification, chromosomal rearrangement and transposable elements [1]. Researchers have recognized that personalized driver gene identification is required for discovering rare causal events in cancer and may provide vital information for selecting effective therapies in precision medicine[2, 3]. However, most computational models for cancer-driver identification look for genes that have

more mutations than expected from the average background mutation rate. They mainly focus on cohort-level driver gene identification, and fail to identify the personalized driver genes. Therefore, new computational models and methods are urgently needed to identify personalized driver genes for understanding tumor heterogeneity in cancer. Recently, with the development of network science, structural-based network control approaches that aim to find a minimum set of driver nodes for steering the states of large-scale networks to the desired states can help the personalized driver genes identification in cancer propagation [2]. So far, the methods of exploiting the structure-based control principles of complex networks can be mainly divided into two categories according to the styles of the networks [4-7]. One focuses on the directed (or non-symmetrical) networks [4-6], and the other focuses on the undirected (or symmetrical) networks [7]. For directed networks, some linear or local nonlinear structural control tools based on Maximum Matching Sets (MMS) are developed to identify the minimum number of driver nodes that need to be controlled by external signals for the system to achieve the desired control objectives [4-6]. Although those MMS based control tools have many applications in modeling bio-molecular systems [2,9-11], those tools may only give an incomplete view of the network control properties for a system with nonlinear dynamics. Recently, the method of Feedback Vertex Set (FVS)-based control (FC) [12, 13] is proposed to control the large-scale networks in a reliable and nonlinear manner, where the network structure is prior known and the functional form of the governing equations is not specified, but it must satisfy some properties. With such a scheme, a Directed FVS-based control method (DFVS) is applied to control the directed networks by identifying the source nodes and FVS nodes as the driver nodes [8]. The above approaches only focus on the mainly linear or nonlinear dynamics of the directed networks. There are few approaches to investigate the structural controllability of the undirected networks. For example, the method of Minimum Dominating Sets (MDS) [7] investigates the structural controllability of the undirected networks. Since it works with the strong assumption that the controllers can control its outgoing links independently, MDS requires higher costs in many kinds of networks which may

underestimate the structural control analysis of undirected networks. Therefore, efficient methods need to be developed for analyzing the structural controllability of the undirected networks.

Although these structure-based control methods of complex networks have offered powerful mathematical frameworks to understand diverse biological systems at a network level [2,9-11], they cannot be directly applied to the personalized driver genes identification of individual patients. This is primarily due to a gap between the structure network control methods and the applications in the personalized patient system. More efficient methods are needed to obtain personalized state transition networks that capture the phenotypic transitions between healthy and disease states, which is a rate-limiting step of structure network control methods. We addressed the limitation by introducing a novel personalized network control model (PNC), which sheds light on the cancer transition network between a disease state and a healthy state by employing the network control principle on the personalized genetic data of individual patients. PNC consists of the following two steps: (i) construct a personalized state transition network for characterizing the state transition of each individual during dynamical biological processes and (ii) design structure control method on the personalized state transition network. That is, we firstly developed a paired single sample network construction method (Paired-SSN) to construct the personalized state transition network for capturing the phenotype transitions between normal state and tumor state through the integration of the expression data and gene/protein interaction network. The personalized state transition network is a graph in which nodes represent genes, and edges denote the significant co-expression difference between normal state and tumor state in the gene interaction network. Then, on the personalized state transition networks, we designed a novel structure network control method (namely NCUA) according to the FC theory to drive a complex undirected network with nonlinear dynamics from initial attractor to desired attractor through perturbation to a feasible subset of driver nodes.The experimental results from 13 kinds of cancer datasets in The Cancer Genome Atlas (TCGA) demonstrate that:

(i) Considering the Cancer Census Genes (CCG)[14] and Network of Cancer Genes (NCG) [15] as the gold-standard of cancer driver genes, PNC performs better in terms of the F-measures for identifying the personalized driver genes in contrast to traditional personalized driver genes identification methods (i.e., Differential expression genes identification methods and Hub genes selection method) and other personalized driver genes identification methods such as single sample controller strategy (SCS) [2]; furthermore, our PNC shows the enrichment significance in CCG and NCG on multiple cancer datasets;

(ii) The F-measures of NCUA method for identifying the personalized driver genes enriched in CCG and NCG genes would be higher when Paired-SSN rather than LIONESS and SSN are used. When Paired-SSN are used for constructing the personalized state transition network, compared with other network control methods, NCUA can more effectively improve the performance in terms of the F-measures for identifying the personalized driver genes that are enriched in the CCG and NCG genes.

(iii) For the cancer patient, the controllability and personalized driver genes evaluated by PNC vary in different cancer datasets; more heterogeneous the degree distribution of constructed personalized state transition network is, the easier it is to control the entire system. In conclusion, our PNC model provides a new powerful tool for identifying the personalized driver genes in individual patient system and gives novel insights for understanding tumor heterogeneity in cancer.

Results

An overview of PNC model

Cancer can be perceived as a dysfunction of molecular networks that regulate molecular communications and cellular processes [2-3]. To fully understand cancer progression, we need to understand the dynamics of these networks with respect to control theory [16]. In particular, we consider the progression from a healthy state to a disease state as a network control problem. By considering the gene expression profiles in normal and tumor samples as the respective state for a given patient and

applying adequate knowledge of the network structure, structure network control tools aim to detect a small number of genes that are altered in response to input signals and trigger a state transition in individual patients. The personalized driver genes are the altered nodes/genes in response to input signals which can trigger the state transition of the whole biological network. The input signals may be oncogene activation signals such as a genetic mutation, gene amplification, a chromosomal rearrangement, or transposable elements. The “controllers” in PNC model for identifying personalized driver genes are the genetic or environment factors which produce the oncogene activation signals.

Our PNC consists of two main parts. One is a paired single sample network construction method (Paired-SSN) rather than an original SSN [17], which we used to construct personalized state transition networks. In these networks, edges exist in both the gene interaction network and the differential co-expression network for each patient based on integration of the matched genetic information of normal and disease samples from the same patient (Figure 1). In this work, we used a gene interaction network from literature [18] by integrating multiple types of datasets, including Mutual Exclusivity Modules in cancer [19, 20], Reactome [21], NCI-Nature Curated PID [22] and Kyoto Encyclopedia of Genes and Genomes[23]. The gene interaction network consists of 11,648 genes and 211,794 interactions. The nodes in the personalized state transition networks represent the genes, and a pair of nodes denotes the significant co-expression difference of gene intaractions between normal sample and tumor sample. For the second part of PNC, we designed a novel nonlinear structural control method (namely, NCUA) [12, 13] for identifying personalized driver genes. This was done to ensure that the system state would asymptotically be changed in the personalized state transition network. These two primary methodologies are described in more detail in the section of Methods.

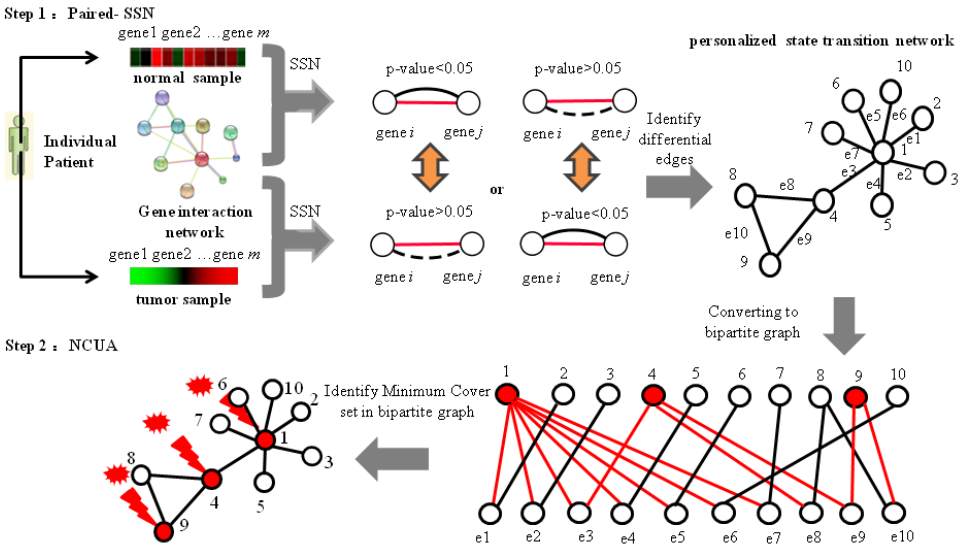

Figure 1 Overview of personalized network control model (PNC) for identifying the personalized driver genes. PNC model consists of two main parts. One is that a Paired-SSN method was developed to construct the personalized state transition networks by using the matched expression information of normal and tumor samples of an individual patient and gene interaction network. For an individual patient, we firstly respectively obtained co-expression p-value of the interaction edges for normal sample and tumor sample by using the SSN. Then the personalized state transition network was constructed. In the personalized state transition network, the edge between gene *i* and gene *j* will exist if co-expression p-value of the interaction edge is less than (greater than) 0.05 for the tumor sample but greater than (less than) 0.05 for the normal sample. The solid line denotes that co-expression p-value is less than 0.05 while the dotted line denotes that co-expression p-value is larger than 0.05. The red solid line denotes the interaction edge in gene interaction network. Therefore edges in the personalized state transition networks denote the significant co-expression difference of gene intaractions between normal sample and tumor sample. Another part is to design a novel nonlinear structural control method (NCUA) for identifying the personalized driver genes by considering personalized state transition network as an undirected network. For our NCUA method, we assumed that each edge in an undirected network is bi-directional (feedback loop) and constructed a bipartite graph from the undirected network. In bipartite graph, the nodes of top side are the nodes of original graph and the nodes of the bottom side are the edges of the original graph. In the end, we adopted an equivalent optimization procedure for obtaining the minimum dominating nodes in the up side to cover the bottom side nodes in the bipartite graph, which are considered as the minimum driver nodes of NCUA.

Performance comparisons of PNC with other personalized driver genes identification methods

To identify personalized driver genes by using PNC, we collected paired samples (a control/normal sample and a tumor sample) from each individual. Here we used cancer datasets which contained enough normal-disease paired samples (>20 paired

samples) in TCGA for the case study. By searching TCGA, 13 cancer datasets met the requirements, i.e., the datasets for breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), kidney chromophobe (KICH) and kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), stomach adenocarcinoma (STAD), uterine corpus endometrial carcinoma(UCEC) , Head and Neck Squamous Cell Carcinoma (HNSC),Prostate Adenocarcinoma (PRAD) and Thyroid Papillary Carcinoma (THCA). More information is given in Table S1 of Additional File 1.

On the 13 cancer datasets, the identified personalized driver genes annotated in the Cancer Census Genes (CCG)[14] and Network of Cancer Genes (NCG) [15] (Additional File 2) were adopted to compute the F-measure scores (Methods) for assessing the performance of PNC. Figure 2(a) is the results of PNC and other traditional methods such as DEG-FoldChange, DEG-p-value, DEG-FDR and Hub genes selection methods (Methods). DEG-FoldChange selects the personalized driver genes by calculating the fold-change between normal sample and tumor sample ($|\log2(\text{fold-change})|>1$). DEG-p-value and DEG-FDR selects the personalized driver genes by respectively calculating the p-value and FDR (<0.05) between a cancer tumor sample and a group of control samples. Hub genes selection method regards the hub genes in the constructed network as cancer driver genes. From Figure 2(a), we can find that F-measures of PNC are higher than that of DEG methods (i.e., DEG-FoldChange, DEG-p-value and DEG-FDR) and Hub genes selection method. These results show that our network control method is more effective for discovering personalized driver genes than the DEG methods and the Hub genes selection method.

Recently a single-sample controller strategy (SCS) based on the MMS structural control theory [2] was proposed to identify personalized driver genes. SCS integrates mutation data (e.g., single nucleotide variations, SNV; copy number variations, CNV) and expression data (i.e., paired samples of each patient) into a reference molecular network for each patient to obtain the driver genes in a personalized-sample manner.

To further evaluate the advantage of PNC, we obtained the single nucleotide variations mutation data, the copy number variations mutation data and the gene expression data of the PRAD cancer dataset and compared the performance of PNC with that of SCS in terms of F-measures for identifying personalized driver genes enriched in CCG and NCG genes. Figure 2(b) shows the comparison results of PNC and SCS. We can see that PNC has higher F-measure than SCS, indicating that PNC can more effectively identify personalized driver genes which are enriched in the CCG and NCG genes.

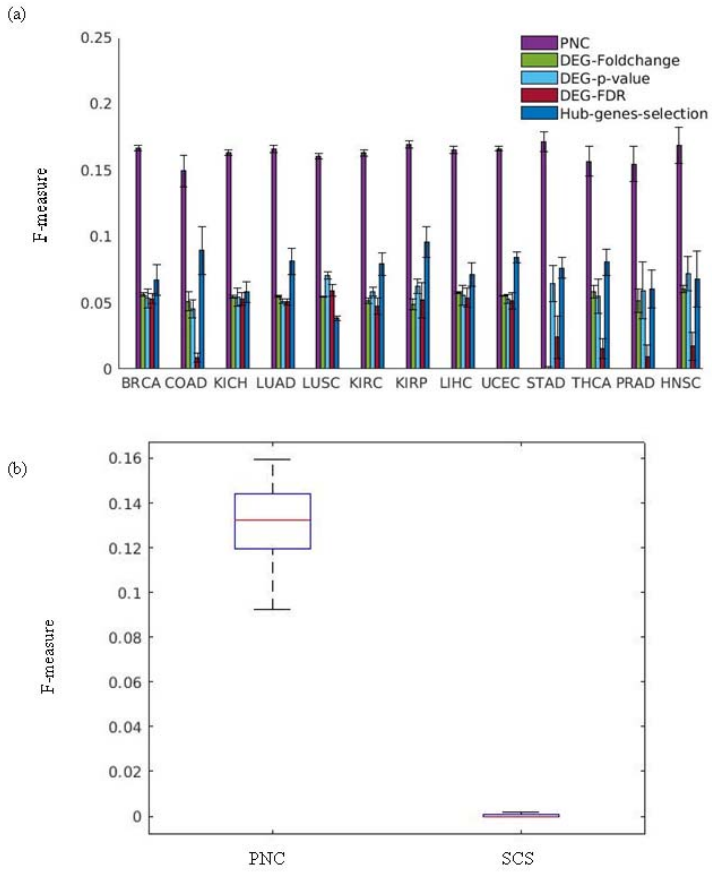

Figure 2 Results of PNC and other methods. (a) F-scores of PNC and other four traditional methods (i.e., DEG-FoldChange, DEG-p-value, DEG-FDR and Hub genes selection method) on 13 cancer datasets. (b) F-scores of PNC and single sample controller strategy (SCS) [2] on the PRAD cancer dataset.

The effect of the different single sample network construction methods

In addition, to evaluate the influence of potential different personalized state transition network construction methods, we used LIONESS and SSN for constructing personalized state transition networks on the collected 13 cancer datasets. Different from Paired-SSN, SSN considers the tumor sample network as the personalized state transition network. The tumor sample network is constructed by quantifying the differential network between the tumor sample and a group of normal samples. LIONESS[24] constructs the personalized state transition networks by calculating the edge statistical significance between all tumor samples and the tumor samples without a given single sample (Methods). As LIONESS can be applied to multiple aggregate network reconstruction approaches, we used the Pearson Correlation Coefficient (PCC) in LIONESS to guarantee a fair comparison with the SSN and Paired-SSN methods. To keep the edge direction and filter the noise of PCC correlation in the personalized state transition networks, the previously mentioned gene interaction network in Paired-SSN was used on the LIONESS and SSN.

Based on the personalized state transition networks constructed by LIONESS, SSN and Paired-SSN respectively, we applied the NCUA method for identifying the personalized driver genes (Figure 3 (a)). From Figure 3 (a), we found that considering the CCG and NCG genes as the gold-standard of cancer driver genes, the F-measures of network control methods would be higher when SSN and Paired-SSN networks rather than LIONESS were used. Furthermore, F-measures of NCUA on Paired-SSN are more robust than SSN method. For example, in the KIRP cancer dataset, NCUA on Paired-SSN had higher F-measure than SSN method [17] although NCUA on Paired-SSN and SSN exhibited a similar performance in other cancer datasets. Therefore compared with other methods, Paired-SSN can include more information of individual patients and be suitable for identifying the personalized driver genes of individual cancer patients in this work.

The effect of the different network control methods

Furthermore, to evaluate the efficiency of NCUA as the second step of PNC, we gave computational comparisons in terms of F-measures for identifying personalized driver genes enriched in CCG and NCG genes between our NCUA and other methods including MMS (full control) [4], MDS[7] and DFVS[8] (Figure 3 (b)). We should note that MMS and DFVS are directed-network-based control methods while MDS and NCUA are undirected-network-based control methods. Thus, for MMS and DFVS, the personalized state transition network is a directed graph in which the directed information of reference gene interaction is used. Meanwhile, for MDS and NCUA, the personalized state transition network is an undirected graph in which the directed information of reference gene interaction is not used. In Table S2 of Additional file 1, we gave the concept comparisons including the network types and targeted states and time complexity and network dynamics between our NCUA and other network control method including MMS, MDS and DFVS. From Figure 3 (b), we can find that F-measures of NCUA in 13 cancer datasets are higher than those of MMS, MDS and DFVS methods, so our NCUA could overall identify personalized driver genes more efficiently than other network control methods.

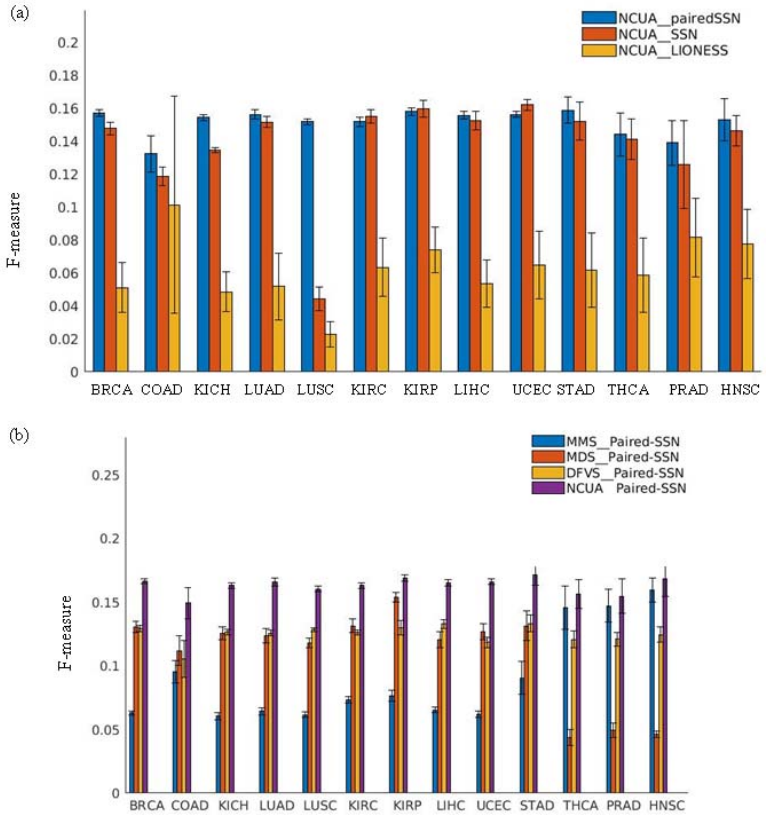

Figure 3 The effect of different single sample network control methods and network control methods on PNC. (a) The significant enrichment F-scores of NCUA on personalized state transition networks constructed by Paired-SSN (the first step of our PNC), SSN and LIONESS. (b)The enrichment F-scores of NCUA (the second step of our PNC) and other structure network control methods (MMS and MDS and DFVS) on the paired SSN networks (the first step of our PNC) in the list of CCG and NCG genes.

Statistic analysis of the personalized driver genes identified by PNC

In fact, PNC provides a set of personalized driver genes for each patient. On each cancer dataset, by calculating the frequency of genes that appear as the personalized driver genes among a group of patients, we defined high-frequency personalized driver genes (fh>0.6), medium-frequency personalized driver genes (0.3<fm≤ 0.6), and low-frequency personalized driver genes (fl≤ 0.3). In Figure 4, we gave the fractions of high-frequency personalized driver genes (yellow), medium-frequency personalized driver genes (red), and low-frequency personalized driver genes (blue)

on 13 cancer datasets. Figure 4 revealed that different cancer datasets had different distributions for personalized driver genes among groups of patients. For example, in the PRAD cancer dataset, the high-frequency personalized driver genes accounted for the majority (>60%); however in the UCEC cancer dataset, the high-frequency personalized driver genes accounted for a minority (<20%). These results demonstrated that the heterogeneity of personalized driver genes varied in different cancer datasets.

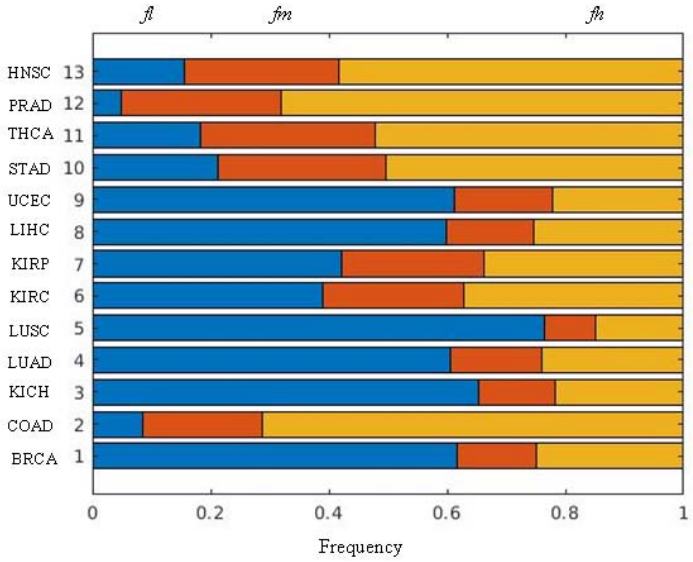

Figure 4 The fractions of high-frequency personalized driver genes (yellow, *fh*), medium-frequency personalized driver genes (red, *fm*), and low-frequency personalized driver genes (blue, *fl*) on 13 cancer datasets.

It is widely accepted that driver genes have low mutation frequency, which is called "long tail of rarely mutated genes" [25]. To demonstrate the ability of discovering the driver genes with rare mutation frequency, we first obtained the single nucleotide variations (SNVs) data of 13 kinds of cancer datasets from TCGA [26]. We then divided the personalized driver genes within CCG and NCG genes into driver genes with rare mutation frequency (0<frequency<0.05) and high mutation frequency (frequency>0.05) respectively. Finally we calculated the fraction of personalized driver genes with rare mutation frequency and high mutation frequency respectively on 13 kinds of cancer data as shown in Figure 5. Results in Figure 5 show that our

PNC tends to discover personalized driver genes with rare mutation frequency on most of cancer datasets. However the fraction of the driver genes with rare mutation frequency and high mutation frequency are similar on UCEC and STAD cancer datasets.

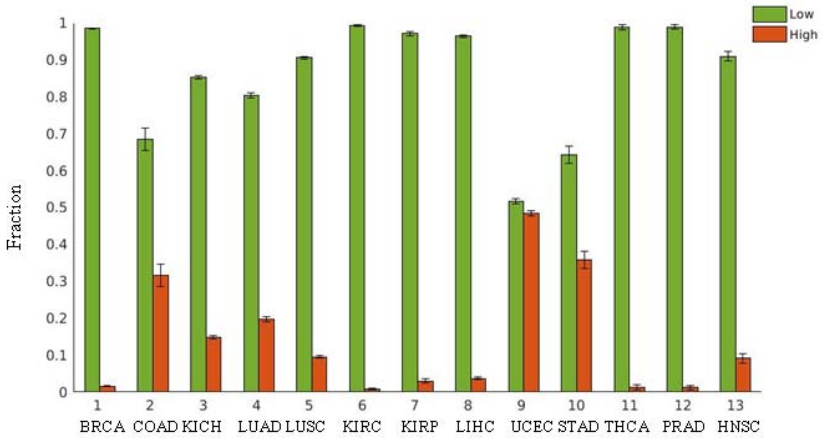

Figure 5 The fraction of personalized driver genes in CCG and NCG with low mutation frequency and high mutation frequency respectively on 13 kinds of cancer data.

The enrichment analysis of the personalized driver genes identified by PNC

To further support the efficiency of PNC by statistic significance, the enrichment p-values of the predicted driver genes in CCG and NCG lists were evaluated for PNC as shown in Figure 6. The enrichment p-values were calculated by using the hyper geometric test [27]. The computational details were shown in section of Methods. From the results of Figure 6, we can see that PNC is actually significant for predicting the personalized driver genes enriched in the CCG and NCG lists. Furthermore, from Figure 6, we can see the following facts, for the CCG and NCG lists, the enrichment p-values of PNC vary in the different cancer datasets. For example, for COAD ,PRAD, THCA and HNSC cancer dataset, not all patients have significant enrichment results (Figure 6). But for other cancer datasets such as BRCA , KICH , LUAD and UCEC cancer datasets, the personalized driver genes are significant enriched in CCG and NCG genes (p-value<0.05) for all patients by using the PNC method (Figure 6) .

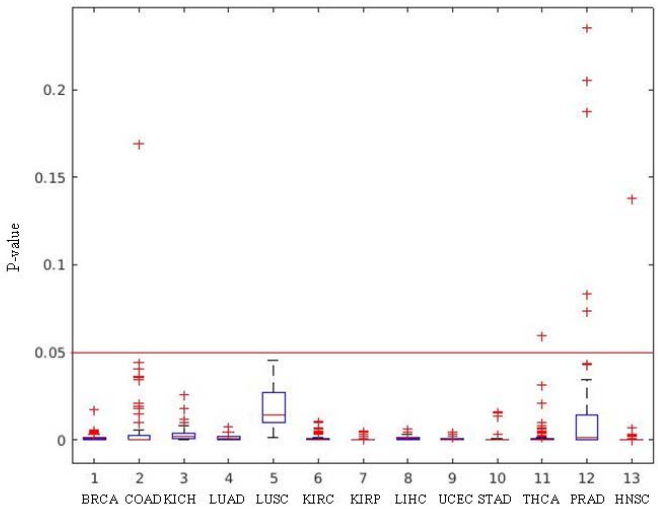

Figure 6 The enrichment p-value of personalized driver genes on 13 kinds of cancer data. The red line denotes the significant threshold value 0.05.

Controllability evaluation of individual patient system by using PNC

To discover the personalized controllability of individual patient system, we employed the PNC on 13 cancer datasets in TCGA. We defined controllability as follows,

$$Controllability = \|D\| / \|V\|, \tag{1}$$

where $\|D\|$ denotes the set of driver nodes to control the whole network and $\|V\|$ denotes the number of connected nodes in the network. In Eq. (1), the smaller the value of *controllability* is, the more easily the personalized state transition network will be controlled. The personalized controllability of PNC on 13 cancer datasets from TCGA are shown in Figure 7 (a). From Figure 7 (a), we can find that the personalized controllability of individual patients varies in different cancer datasets. For example, the personalized controllability in COAD cancer data is less than 0.4, while the personalized controllability in LUSC is higher than 0.5.

To demonstrate the relationship between personalized controllability and network structure, we firstly obtained the connected component (CC) of each personalized state transition network and then used the CC network to determine the

degree exponent γ by using the Kolmogorov-Smirnov goodness-of-fit test [28]. More computational details are shown in Supplementary Note 4 of Additional file 1. The network parameters in each CC on different cancer datasets including number of nodes, average degree and degree exponent are presented in Additional file 3. Finally in Figure 7 (b) we gave the controllability (defined in Eq. (1)) of personalized state transition networks which are significantly subject to the power-law distribution (p-value>0.05) as a function of the average degree and power law degree exponent. From Figure 7 (b), we found that the personalized state transition networks with smaller values of the average degree and power law degree exponent would be easier to control. That is, the more heterogeneous the degree distribution of a personalized state transition network is, the easier it is to control the entire system.

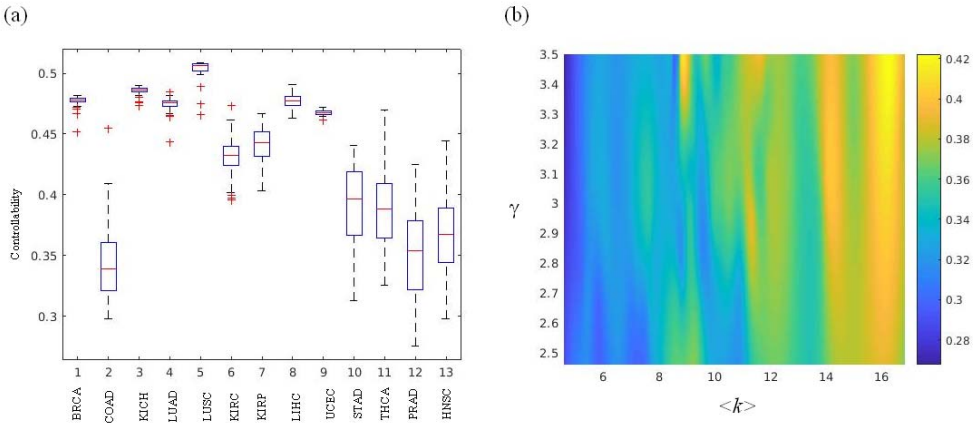

Figure 7 Controllability evaluation of individual patient system by using PNC. (a) Box plot with the distribution of the personalized controllability in 13 cancer datasets. **(b)** Heat map of the personalized controllability (different colors) which is as a function of the average degree and power law degree exponent.

To more clearly explain how network parameters affect the controllability, we applied NCUA on multiple synthetic scale-free (SF) networks [29], as shown in Figure S1 of Additional file 1. Figure S1 show that the more heterogeneous an SF network degree distribution is, the easier it is to control the entire system. Those simulation results match well with the results of experimental results in cancer. Furthermore when the degree exponent γ is above 3, the fraction of minimum driver nodes is larger than that of the Erdös-Rényi random networks (Figure S1). These

results gave insight into heterogeneous networks will be easier to control with the minimum number of driver nodes than the homogeneous networks. More details are shown in supplementary note 3 and Supplementary Note 5 of Additional file 1.

Discussion

Cancer is known as a disease mainly caused by gene alterations [30]. Identification of driver genes is a key point of focus in cancer genomics. Computational models and methods are required to prioritize biologically efficient driver genes dependent on cancer high-throughput sequencing data. However, few methods can efficiently distinguish the personalized driver genes that change the state of genes in each patient. Recently, structure-based network control approaches have enabled us to investigate how to control complex networks by using a minimum set of driver nodes which could enable us to understand mechanisms of cancer progression. For structure-based network control methods, a minimum number of driver nodes need to be identified to drive the network state from the initial state to the desired state based on adequate knowledge of the network structure. The application of network control methods for personalized driver genes identification requires two key steps, i.e., (i) to construct personalized state transition networks which are involved in the state transition during disease development for each patient and (ii) to design the optimal network control methods based on the structure of the personalized state transition network. Here, we propose PNC based structure network control theory to personalized driver genes identification. PNC considers how to construct the personalized state transition network capturing the phenotype transitions between healthy and disease state through the utilize of Paired-SSN method and find the personalized driver genes on phenotype transitions by applying NCUA, which can control the individual from the normal attractor to the disease attractor.

To further demonstrate the advantage of our PNC, it has been employed to investigate personalized driver genes of cancer samples from TCGA by screening known driver genes in the CCG and NCG genes as controls of their phenotype transitions. We find that in contrast to other models on the personalized driver genes

identification such as Differential expression genes identification methods, Hub genes selection method and single sample controller strategy (SCS), our PNC shows a higher performance in identifying the personalized driver genes in the CCG and NCG genes. Particularly we have validated that Paired-SSN has better performance for constructing personalized state transition networks compared with other methods, such as SSN and LIONESS. Furthermore NCUA is a better choice for the identification of personalized driver genes on the constructed personalzied state transition networks compared with other network control methods. Although NCUA already performs better than other network control methods, efforts to achieve promising improvements for the utilization of the NCUA approach are still ongoing. For example, NCUA does not consider multiple driver-node sets to control the network. Therefore in the future it is worth studying how to find multiple driver nodes configurations in large-scale networks, which will benefit the application of network control methods in the fields of biology and biomedicine. Furthermore NCUA ignores the weight and direction information of network edges, causing a potential bottleneck in characterizing the edge control rather than only node control. Therefore, another important research direction will be the extension of our NCUA method to edge dynamics[31] Taken together, our analysis results support the practical applications of our PNC, which is effective for identifying Personalized driver genes by applying the network control theory to the personalized patient system. Therefore, the PNC reveals new biological insights for clinical applications of personalized therapy and in prognosis analysis.

Methods

Paired-SSN

For the Paired-SSN method, the co-expression network of the tumor sample network or normal sample network for each patient is first constructed based on statistical perturbation analysis of one sample against a group of given reference samples (e.g., choosing the expression data of all normal samples of all patients in a given cancer data set as the reference data here) with the SSN method [17]. If the differential PCC

(ΔPCC) of an edge is statistically significantly large based on the evaluation of the SSN method, this edge would be kept for the single sample (normal sample or tumor sample) from each patient. The p-value for an edge can be obtained from the statistical Z-value measuring ΔPCC. Of note, the ΔPCC of an edge between genes *i* and *j* and its significant Z-score are calculated using the following formula:

$$\Delta PCC_{ij} = \left| PCC_{ij}^{n+1} - PCC_{ij}^{n} \right|$$
$$Z_{ij} = \frac{\Delta PCC_{ij}}{(1-(PCC_{ij}^{n})^2)/(n-1)}, \quad (2)$$

where PCC_{ij}^{n} is the PCC of an edge between genes *i* and *j* in the reference network with *n* samples; and PCC_{ij}^{n+1} is the PCC of the edge between genes *i* and *j* in the perturbed network with one additional sample, given that this single sample (e.g. normal sample or tumor sample) for each patient is added to the reference sample group. All of the edges with significantly differential correlations (e.g., p-value < 0.05) are used to constitute the SSN for one normal sample or tumor sample.

Then, the personalized differential co-expression network between the normal sample network and tumor sample network can be constructed in which the edge between gene *i* and gene *j* will exist if the *p*-value of the edge is less than (greater than) 0.05 in the tumor network but greater than (less than) 0.05 in the normal network for their corresponding patient. To more precisely demonstrate the regulatory mechanism of personalized state transition networks, we use the gene interaction network to filter the noise of PCC between gene pairs. Finally the personalized state transition networks are obtained in which edges exist in both the gene interaction network and the differential co-expression network for each patient.

LIONESS

LIONESS reconstructs the individual specific network in a population of tumor samples as the personalized state transition network for each tumor sample[24]. LIONESS constructs the state transition network by calculating the edge statistical significance between all of the tumor samples and the tumor samples without a given

single sample. In order to guarantee a fair comparison with SSN and Paired-SSN, we use PCC in LIONESS to model the sample-specific state transition network. The network specific to one sample in terms of the aggregate networks is then:

$$e_{ij}^{q} = N_s(e_{ij}^{N_s} - e_{ij}^{N_s-q}) + e_{ij}^{N_s-q}, \tag{3}$$

where $e_{ij}^{N_s}$ and $e_{ij}^{N_s-q}$ are the PCC values of gene pairs *i* and *j* in sample *n* and all but sample *q*, respectively; N_s is the number of the samples. To construct the sample-specific network, after we obtain the PCC absolute value distribution *S* of all gene pairs we choose a threshold to determine the differential expression edges in the sample-specific network as follows:

$$w = \mu(S) + 2\delta(S), \tag{4}$$

where $\mu(S)$ and $\delta(S)$ are the mean and standard variance of the PCC absolute value distribution *S* of all gene pairs.

NCUA

In many biological systems, there is adequate knowledge of the underlying wiring diagram, but not of the specific functional forms [32, 33]. Deep analysis of such complicated systems requires structure-based control approaches, which only should know whether there are edges or not. In past decades, the structure based network control studies have shifted from linear dynamics to nonlinear dynamics [34-36]. One of these methods, the FC control method[12, 13] can be reliably applied to large scale complex networks in which the structure is well known and the functional form of the governing equations is not specified. To drive the state of a network to any one of its naturally occurring end states (i.e., dynamical attractors), FC needs to manipulate a set of nodes (i.e., the FVS) that intersects every feedback loop in the network. Under the FC framework, DFVS is proposed to study the structural control problem of directed networks, selecting a minimum set of driver nodes rendering a directed network structural controllable, which has been recently discussed [8]. DFVS assumes that the

state of the nodes in a directed network, characterized by source nodes $s_j(t)$ and internal nodes $x_i(t)$, obeys the following equations:

$$\begin{aligned} dx_i / dt &= F_i(x_i, I_i, t), \\ ds_j / dt &= G_j(t) \end{aligned}, \tag{5}$$

where I_i is a set of predecessor nodes of node i, F_i captures the nonlinear response of node i to its predecessor nodes I_i,and G_j governs the nonlinear response of source node s_i.

If the undirected networks are considered as bi-directional networks, there are no source nodes in the undirected network nodes. Therefore Eq. (3) of an undirected network $G(V, E)$, is reduced to,

$$dx_i / dt = F_i(x_i, I_i), \tag{6}$$

where x_i denotes the state variable of the i-th node at time t in undirected networks, I_i is a set of neighborhood nodes of node i, and F_i captures the nonlinear response of node i to its neighborhood nodes in undirected networks.

In fact, the selection of driver nodes may be dependent on the network style (directed or undirected). Motivated by this fact, under the FC framework we formalize the Nonlinear structural Control problem of the Undirected networks (NCU), i.e., how to choose the minimum set of driver nodes for changing the network state from one attractor to the desired attractor in large-scale networks with symmetric edges information. More details can be found in Supplementary Note 1 of Additional File 1. Here we develop a graphic-theoretic algorithm called Nonlinear Control of Undirected network Algorithm (NCUA) under the nonlinear FC framework for determining the minimum driver nodes in undirected networks. Since NCUA and DFVS study the structural network control of undirected and directed biological networks respectively based on the framework of FC, therefore NCUA and DFVS methods completely depict the structure based network control of the large scale system with nonlinear dynamics from the respective of FC.

By assuming that each bi-directional edge is considered as a feedback loop, our

NCUA searches the sets of minimum nodes whose removal leaves the graph without edges in undirected networks (Figure 1). The implementation of NCUA consists of two main steps: (i) constructing a bipartite graph from the original undirected network, in which the nodes of the top side are the nodes of the original graph and the nodes of the bottom side are the edges of the original graph, and (ii) determining the MDS of the top side nodes to cover the bottom side nodes in the bipartite graph by using Integer Linear Programming (ILP). The computational details are as follows,

Constructing a bipartite graph from the original undirected network

For a given undirected network G (V, E), we assume that each edge is bi-direct, converting G (V, E) into a bipartite graph G (V_T,$V_\perp$,E_1), where $V_T \equiv V$ and $V_\perp \equiv E$. If $v_i \in V_T$ was one of the nodes for $v_j \in V_\perp$, we added one edge connecting $v_i \in V_T$ and $v_j \in V_\perp$ into the set E_1.

Obtaining the cover set with minimum cost by using Integer Linear Programming

After obtaining the bipartite graph G (V_T,$V_\perp$,E_1), we select a minimum dominating set S from V_T , which must cover all the nodes in $V_\perp$. This minimum dominating set is used to determine the driver nodes for controlling the whole network. In NCUA, we select the driver nodes by solving the following ILP:

$$\begin{aligned} &\min f = \sum_{v \in V_T} y_v \\ &s.t. \sum_{\{v,u\} \in E} y_v \geq 1 (\text{every } u \in V_\perp),\ y_v \in \{0,1\} \end{aligned}, \qquad (7)$$

where $\sum_{v \in V_T} y_v$ denotes the number of candidate nodes and y_v is an indicative variable; when node v in the up side node set in the bipartite network is within the driver nodes**,** $y_v = 1$, and otherwise, $y_v = 0$; Although it is an NP-hard problem, the optional solution can be obtained for the moderate size graphs with up to a few tens of thousands of nodes by utilizing the LP-based classic branch and bound method [37].

Hub genes selection method

The Hub genes selection method regards the hub genes in the constructed network as driver genes. After obtaining the degree distribution of all genes U in the personalized state transition network, we use a threshold as introduced in following formula to obtain the hub genes,

$$w = \mu(U) + 2\delta(U), \tag{8}$$

where $\mu(U)$ and $\delta(U)$ are the mean and standard variance of the degree distribution U of all gene, respectively.

F-measure

To verify the effectiveness of the PNC analysis mainly based on structural control methods, the F-measure is considered to assess the enrichment ability of identified personalized driver genes in a given gold standard list (CCG and NCG genes), which considers both the precision and the recall using the formula:

$$F_i = \frac{2P_iR_i}{P_i + R_i}, \tag{9}$$

where P_i denotes the fraction of correctly identified genes among all of the identified genes (precision) and R_i denotes the fraction of correctly identified personalized driver genes among the given dataset (recall).

Enrichment analysis of the personalized driver genes by using hyper geometric test

To estimate the significance of overlap between the predicted personalized driver genes for PNC method and a gold-standard cancer driver gene lists such as Cancer Census Genes and Network of Cancer Genes, we compute the *p*-value by the hyper geometric test [27] as follow,

$$p = P(X \geq k) = \sum_{k=1}^{\infty} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \quad , \tag{10}$$

where N is the number of genes in background gene interaction network, K is the number of gold-standard cancer driver gene lists, k is the number of the predicted personalized driver genes overlapping with the genes in the gold-standard cancer driver gene lists, and n is number of the predicted personalized driver genes. If the enrichment p-value is less than 0.05, then we regard that this predicted personalized driver genes for PNC method is significantly enriched in the gold-standard cancer driver gene lists.

Availability of code and datasets

The PNC code and the data resources used in this work can be freely downloaded from https://github.com/WilfongGuo/Personalized_Network_Control.

Appendices

Additional file 1: *Additional_file1_Supplemntary_manuscript.docx*. Supplementary manuscript of Structure based control analysis of undirected complex networks with applications to personalized driver genes identification in cancer.

Additional file 2: *Additional_file2_CGC_NCG_genes.xlsx*. The list of Cancer Census Genes (CCG) and Network of Cancer Genes (NCG) to assess the F-measure of the identified personalized driver genes of different methods.

Additional file 3: *Additional_file3_KS _Personalized_network.xlsx*. The results of personalized network using the Kolmogorov-Smirnov goodness-of-fit test on 13 kinds of cancer datasets from TCGA.

Authors contributions

WFG and TZ developed the methodology. WFG and YL executed the experiments, WFG wrote this paper. WFG, SWZ, TZ, JG and LC revised the manuscript. SWZ, JG and LC supervised the work, made critical revisions of the paper, and approved the submission of the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgements

This paper was supported by the National Natural Science Foundation of China (61873202, 61473232, 91430111, 91439103, 91529303, 31771476, 81471047 and 11871456) and National Key R&D Program (2017YFA0505500), Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB13040700), National Key R&D Program (Special Project on Precision Medicine) (2016YFC0903400), and Natural Science Foundation of Shanghai (17ZR1446100). We thank Professor Tatsuya Akutsu in Kyoto University for giving us valuable comments during the preparation of our manuscript.

References

[1] Porta-Pardo E, Kamburov A, Tamborero D, et al. Comparison of algorithms for the detection of cancer drivers at subgene resolution[J]. Nature methods, 2017, 14: 782.

[2] Guo W-F, Zhang S-W, Liu L-L et al. Discovering personalized driver mutation profiles of single samples in cancer by network control strategy[J]. Bioinformatics 2018,34:1893-1903.

[3] Cheng F, Zhao J, Zhao Z. Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes[J]. Briefings in Bioinformatics, 2015, 17: 642-656.

[4] Liu Y-Y, Slotine J-J, Barabási A-L. Controllability of complex networks[J]. Nature, 2011, 473: 167.

[5] Gao J, Liu Y-Y, D'souza R M, et al. Target control of complex networks[J]. Nature communications, 2014, 5: 5415.

[6] Wu F-X, Wu L, Wang J, et al. Transittability of complex networks and its applications to regulatory biomolecular networks[J]. Scientific Reports, 2014, 4: 4819.

[7] Nacher J C, Akutsu T. Dominating scale-free networks with variable scaling exponent: heterogeneous networks are not difficult to control[J]. New Journal of Physics, 2012, 14: 073005.

[8] Zañudo J G T, Yang G, Albert R. Structure-based control of complex networks with nonlinear dynamics[J]. Proceedings of the National Academy of Sciences, 2017, 114: 7234-7239.

[9] Hu Y, Chen C-H, Ding Y-Y, et al. Optimal control nodes in disease-perturbed

networks as targets for combination therapy[J]. Nature communications, 2019, 10: 2180.

[10] Nacher J C, Akutsu T. Minimum dominating set-based methods for analyzing biological networks[J]. Methods, 2016, 102: 57-63.

[11] Vinayagam A, Gibson T E, Lee H-J, et al. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets[J]. Proceedings of the National Academy of Sciences, 2016, 113: 4976-4981.

[12] Futreal P A, Coin L, Marshall M, et al. A census of human cancer genes[J]. Nature reviews cancer, 2004, 4: 177.

[13] Repana D, Nulsen J, Dressler L, et al. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens[J]. Genome biology, 2019, 20: 1.

[14] Fiedler B, Mochizuki A, Kurosawa G, et al. Dynamics and control at feedback vertex sets. I: Informative and determining nodes in regulatory networks[J]. Journal of Dynamics and Differential Equations, 2013, 25: 563-604.

[15] Mochizuki A, Fiedler B, Kurosawa G, et al. Dynamics and control at feedback vertex sets. II: A faithful monitor to determine the diversity of molecular activities in regulatory networks[J]. Journal of theoretical biology, 2013, 335: 130-146.

[16] Kolch W, Halasz M, Granovskaya M, et al. The dynamic control of signal transduction networks in cancer cells[J]. Nature Reviews Cancer, 2015, 15: 515.

[17] Liu X, Wang Y, Ji H, et al. Personalized characterization of diseases using sample-specific networks[J]. Nucleic Acids Research, 2016, 44: e164.

[18] Hou J P, Ma J. DawnRank: discovering personalized driver genes in cancer[J]. Genome medicine, 2014, 6: 56.

[19] Ciriello G, Cerami E, Sander C, et al. Mutual exclusivity analysis identifies oncogenic network modules[J]. Genome research, 2012, 22: 398-406.

[20] Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis[J]. Genome biology, 2010, 11: R53.

[21] Croft D, O'kelly G, Wu G, et al. Reactome: a database of reactions, pathways and biological processes[J]. Nucleic Acids Research, 2010, 39: D691-D697.

[22] Schaefer C F, Anthony K, Krupa S, et al. PID: the pathway interaction database[J]. Nucleic Acids Research, 2008, 37: D674-D679.

[23] Kanehisa M, Goto S, Sato Y, et al. KEGG for integration and interpretation of large-scale molecular datasets[J]. Nucleic Acids Research, 2011, 40: D109-D114.

[24] Kuijjer M L, Tung M G, Yuan G, et al. Estimating sample-specific regulatory networks[J]. iScience, 2019, 14: 226-240.

[25] Leiserson M D M, Vandin F, Wu H T, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes[J]. Nature genetics, 2015, 47: 106.

[26] Samur M K. RTCGAToolbox: a new tool for exporting TCGA Firehose data[J].

PloS one, 2014, 9(9): e106397.
[27] Rivals I, Personnaz L, Taing L, et al. Enrichment or depletion of a GO category within a class of genes: which test?[J]. Bioinformatics, 2006, 23: 401-407.
[28] Clauset A, Shalizi CR, Newman MEJ Power-Law Distributions in Empirical Data. Siam Review 2009, 51: 661-703.
[29] Barabási A-L, Bonabeau E. Scale-free networks[J]. Scientific american, 2003, 288: 60-69.
[30] Zhang J, Zhang S. Discovery of cancer common and specific driver gene sets[J]. Nucleic acids research, 2017, 45: e86.
[31] Nepusz T, Vicsek T. Controlling edge dynamics in complex networks[J]. Nature Physics, 2011, 8: 568-573.
[32] Newman M E. The structure and function of complex networks[J]. SIAM review, 2003, 45: 167-256.
[33] Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease[J]. Nature reviews genetics, 2011, 12: 56.
[34] Cornelius S P, Kath W L, Motter A E. Realistic control of network dynamics[J]. Nature communications, 2013, 4: 1942.
[35] Wang L-Z, Su R-Q, Huang Z-G, et al. A geometrical approach to control and controllability of nonlinear dynamical networks[J]. Nature communications, 2016, 7: 11323.
[36] Karl S, Dandekar T. Convergence behaviour and control in non-linear biological networks[J]. Scientific Reports, 2015, 5: 9746.
[37] Lenstra Jr H W. Integer programming with a fixed number of variables[J]. Mathematics of operations research, 1983, 8: 538-548.