doi: 10.1093/nar/gkaa1272

Network controllability-based algorithm to target personalized driver genes for discovering combinatorial drugs of individual patients

Wei-Feng Guo[1,2], Shao-Wu Zhang[1,*], Yue-Hua Feng[1], Jing Liang[2], Tao Zeng[3,4,*] and Luonan Chen[4,5,6,7,*]

[1]Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xian 710072, China, [2]School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China, [3]CAS Key Laboratory of Computational Biology, Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China, [4]Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy Science, Shanghai 200031, China, [5]School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China, [6]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China and [7]Key Laboratory of Systems Biology, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou 310024, China

Received October 07, 2020; Revised December 02, 2020; Editorial Decision December 18, 2020; Accepted December 22, 2020

ABSTRACT

Multiple driver genes in individual patient samples may cause resistance to individual drugs in precision medicine. However, current computational methods have not studied how to fill the gap between personalized driver gene identification and combinatorial drug discovery for individual patients. Here, we developed a novel structural network controllability-based personalized driver genes and combinatorial drug identification algorithm (CPGD), aiming to identify combinatorial drugs for an individual patient by targeting personalized driver genes from network controllability perspective. On two benchmark disease datasets (i.e. breast cancer and lung cancer datasets), performance of CPGD is superior to that of other state-of-the-art driver gene-focus methods in terms of discovery rate among prior-known clinical efficacious combinatorial drugs. Especially on breast cancer dataset, CPGD evaluated synergistic effect of pairwise drug combinations by measuring synergistic effect of their corresponding personalized driver gene modules, which are affected by a given targeting personalized driver gene set of drugs. The results showed that CPGD performs better than existing synergistic combinatorial strategies in identifying clinical efficacious paired combinatorial drugs. Furthermore, CPGD enhanced cancer subtyping by computationally providing personalized side effect signatures for individual patients. In addition, CPGD identified 90 drug combinations candidates from SARS-COV2 dataset as potential drug repurposing candidates for recently spreading COVID-19.

INTRODUCTION

The combination therapy has been widely used in the disease treatment, because it is difficult to achieve the desired clinical effect for monotherapy and the multiple drugs has demonstrated great advantages in overcoming drug resistance and improving clinical outcomes in disease therapy (1). As well known to us, many complex diseases such as cancer are heterogeneous diseases, and the tumor genes cooperate as well as adapt and evolve to the changing conditions for individual patients (2,3). Thus, it is essential to consider the individual heterogeneity during combination therapy in the disease treatment. However, the number of potential combinatorial drugs is astronomical, and these candidate compound combinations cannot all be validated in a rational and rigorous manner for individual patients. Therefore, it is quite challenging to predict the individual targeted combinatorial drugs in the era of precision medicine,

[*]To whom correspondence should be addressed. Tel: +86 29 88431308; Fax: +86 29 88431306; Email: zhangsw@nwpu.edu.cn
Correspondence may also be addressed to Tao Zeng. Tel: +86 21 54920102; Fax: +86 21 54920120; Email: zengtao@sibs.ac.cn
Correspondence may also be addressed to Luonan Chen. Tel: +86 21 54920100; Fax: +86 21 54920120; Email: lnchen@sibs.ac.cn

rather than conventional patient-cohort targeted combinatorial drugs.

Current methods of identifying combinatorial drugs have two main categories. (i) Machine learning-based methods extract the feature vector of the known synergistic drug combinations on a large variety of cancer cell lines, then utilize the machine learning to predict drug combinations (4–6). However, because of the limited number of the personalized samples information (e.g. the personalized omics data), it is difficult to select/find the effective individual patient-specific synergistic drug combinations. (ii) Along with the rapid development of high-throughput biological molecule screening, the emergence of systems biology or network biology has raised the possibility of exploring multi-targets intervention methods with drug synergistic effects for disease treatment (7–9). From a network perspective, many studies have demonstrated that targeting driver genes (i.e. candidate drug targets) can provide the critical information for drug discovery and drug repurposing (10–13). Consequently, some methods have been developed for driver gene identification with multi-dimensional genomic data, such as DriverML (14), DriverNet (15), MutSigCV (16), OncoDriveFM (17), SCS (18) and DawnRank (19). But those existing methods did not fill the gap between the anticancer combinatorial drugs discovery and the targeting personalized driver genes (PDGs) identification, and new algorithms are urgently required to recommend the personalized combinatorial drugs in the disease treatment.

In the past decade, some researches have studied the structural network controllability principles, such as maximum matching set (MMS) (20), minimum dominating set (MDS) (21) and feedback vertex set (FVS) (22). Furthermore, a wealth of sample-specific network construction methods in single samples has been proposed recently to support the personalized driver genes analysis on individual patient-specific biological data (23). These methodological advances have raised the possibility of exploring more precise mathematical models on high throughput personalized multi-omics data for the discovery of efficacious personalized drug combinations. However, the existing structural network controllability methods still face several limitations. The first is that a proper network structure is not available to characterize the gene regulatory mechanism of an individual patient, which is a rate-limiting step of structural network controllability methods. The second is that current structural network controllability methods focus on the selection of minimum number of driver nodes but overlook the weight information of network edges/relations, which may generate multiple configurations with same minimum number of driver nodes, resulting in a potential bottleneck for identifying the combinatorial drugs of individual patients. And the third is that gold standard evaluation metrics are not available when evaluating the performance of identifying the personalized combinatorial drugs with different structural network controllability methods.

To overcome above problems, we developed a novel structural network controllability-based algorithm (namely CPGD) to detect PDGs and further identified the personalized combinatorial drugs for an individual patient. We firstly used the paired single sample-network method (paired-SSN) (24) to construct the personalized gene interaction network (PGIN) whose interactions determine the state transition of an individual patient during disease development. Instead of directly using paired-SSN method on gene expression data, we introduced a measurement (i.e. network edge score) to score the edges of PGIN by integrating the co-mutation score across cancer type-specific data and the personalized co-expression score of each individual patient. Then, according to a FVS-based controllability perspective, we developed a novel nonlinear structural network controllability method (namely weight-NCUA) to identify the PDGs for determining the state transition of the individual biological system between disease state and normal state. In contrast with other structural network controllability approaches, weight-NCUA considers the edge weight information (i.e. network edge score) for the driver node optimization. Finally, the information of the drugs–PDGs interactions and PDGs interactions were used for: (i) prioritizing the personalized combinatorial drugs to evaluate the ability of predicting clinical efficacious combinatorial drugs; (ii) exploring the risk assessment of individual patients on the basis of paired combinatorial drugs and (iii) enhancing the cancer subtyping by side effect quantification on PDGs.

We have evaluated the effectiveness of CPGD on the breast and lung cancer datasets which were derived from The Cancer Genome Atlas (TCGA). On one hand, CPGD can effectively predict clinical efficacious combinatorial drugs, compared with other state-of-the-art driver gene methods. By simultaneously considering the disease related gene module information and multi-sources drug function information, CPGD can measure the synergistic effect of the corresponding personalized driver gene modules for evaluating the synergistic effect of pairwise drug combinations. On the other hand, CPGD has detected three novel pairwise drug combinations (i.e. CETUXIMAB and CARBOPLATIN, CARBOPLATIN and CYCLOPHOSP HAMIDE, CYCLOPHOSPHAMIDE and GEMCITAB INE), and they can significantly divide the breast cancer patients into the discriminative risk groups with the personalized co-targeting driver genes of paired combinatorial drugs. Those results were also supported by TCGA-BRCA cancer dataset and the independent GSE5327-BRCA cancer dataset. By quantifying the side effect of the personalized combinatorial drugs on the personalized driver genes for each individual patient, CPGD further identified two new subtypes on breast cancer with significant differences in survival. In addition, we have applied CPGD on severe acute respiratory syndrome coronavirus 2 (SARS-COV2) dataset, which consists of gene expression data of patients with SARS (25) and 332 SARS-COV2 related proteins for identifying drug combination candidates (26). Consequently, CPGD identified 90 drug combination candidates. Among these drug combination candidates, a pairwise drug combination (i.e. DEXAMETHASONE and TH ALIDOMIDE) is predicted as the potential promising drug combination candidates, both of which are currently being tested in clinical trials for coronavirus disease 2019 (COVID-19) as a recent report (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7280907/).

MATERIALS AND METHODS

Gene expression datasets

Considering the sufficiently available gold-standard clinical combinatorial drug information, we considered breast cancer and lung cancer datasets as two benchmark disease datasets and the detailed information of cancer samples used in this study was summarized in Supplementary note 4 of Supplementary file 1 and Supplementary file 2. Consequently, we collected two cancer gene expression datasets from breast and lung cancer patients. The breast cancer dataset is breast invasive carcinoma (BRCA), and the lung cancer dataset consists of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). The paired (or matched) samples for each individual patient (i.e. a normal sample and a tumor sample from the same patient) were filtered, then obtaining them from the TCGA data portal. In addition, to identify drug combination candidates for recently speading COVID-19, we also collected SARS-COV2 related dataset which consists of gene expression data of patients with SARS (25).

Somatic mutation data of cancer type-specific datasets

To score the edges in PGIN, we obtained the single nucleotide variations (SNVs) data of BRCA, LUAD and LUSC datasets from TCGA, which contains 90 490, 72 541 and 65 304 nonsynonymous somatic mutations, respectively.

Prior-known cancer subtype information

We collected subtype information of cancer patients from TCGA (https://xenabrowser.net/datapages/). For breast cancer patients, we obtained four subtypes of basal-like (Basal), HER2-enriched (HER2), Luminal A (Lum A) and Luminal B (Lum B). For lung cancer dataset, we obtained two subtypes of LUSC and LUAD.

Combinatorial drug–gene interaction network

The interactions between combinatorial drug and gene have been collected by Quan *et al.* (27) from Drug Combination Database (DCDB) (28), drug–gene interaction database (DGIdb) (29), DrugBank (30) and Therapeutic Target Database (TTD) (31). Such combinatorial drug-gene interaction network contains 342 combinatorial drugs and 5788 interaction edges (Supplementary file 2). Among these 342 combinatorial drugs, 122 of these combinatorial drugs are efficacious for treating cancer. Especially, 32 and 17 prior-known combinatorial drugs (Supplementary file 2) derived from DCDB are efficacious in treating breast cancer and lung cancer, respectively.

Collection of disease related genes

For breast cancer and SARS-COV2 disease dataset, we collected 2341 breast cancer related genes (27) and 332 SARS-COV2 related proteins (26) for identifying pairwise drug combination candidates respectively.

Human drug–target network with activation and inhibition interactions

To analyze the risk assessment of cancer patients, the human drug–target network with activation and inhibition interaction information was further added into the network-based investigation of drug–target interactions (32).

Construction of personalized gene interaction networks by using paired-SSN method

For paired-SSN method (24), the co-expression networks of the tumor sample and normal sample for an individual patient are separately built by using SSN method (33). Then, the personalized gene interaction networks are constructed by using the following criterion. When the *P*-value of edge between gene *i* and gene *j* is <0.05 in the tumor sample network but larger than 0.05 in the normal sample network (or the *P*-value of the edge is >0.05 in the tumor sample network but <0.05 in the normal sample network), this edge is retained to constitute the PGIN. More details of paired-SSN were shown in supplementary note 3 of Supplementary file 1.

Supplementaryly, we developed a weight measurement (i.e. network edge score) for scoring the edges of PGIN by integrating somatic mutation data across cancer type-specific data into PGIN based on following scores.

$$e_{ij}^{\text{Patient } k} = \text{Norm}(\text{co-mutation}(i, j)) \\ * \text{Norm}(\text{co-expression}(i, j, k)) \quad (1)$$

$$\text{co-mutation}(i, j) = \frac{|G(i) \cap G(j)|}{|G(i) \cup G(j)|} \quad (2)$$

$$\text{co-expression}(i, j, k) = \left| \log_2 \left| \frac{\Delta\text{PCC}_{ij,k}^{\text{Tumor}}}{\Delta\text{PCC}_{ij,k}^{\text{Normal}}} \right| \right| \quad (3)$$

$$\Delta\text{PCC}_{ij,k}^{\text{Tumor/Normal}} = \text{PCC}_{ij,k}^{n+1} - \text{PCC}_{ij}^{n} \quad (4)$$

where Norm denotes the min-max normalized function; *G*(*i*) and *G*(*j*) denotes the set of individual tumors for mutated gene *i* and gene *j* respectively by inspecting somatic mutations in a given cancer dataset; PCC_{ij}^{n} is the PCC between gene *i* and gene *j* in the reference network with *n* reference samples, and $\text{PCC}_{ij,k}^{n+1}$ is the PCC in the perturbed network with one additional sample (i.e. tumor sample or normal sample) for individual patient *k*.

The co-mutation score (i.e. Equation 2) is defined as the jaccard coefficient between two mutated genes/nodes of one edge among the population (i.e. the fraction of tumors in which both two genes are the mutated genes and tumors in which either of two genes is mutated gene), which indicates the co-mutation probability to promote tumorigenesis and anti-disease drug responses (34). The score of the personalized co-expression (i.e. Equation 3) represents the significant difference of two genes' expression association (i.e. Equation 4) between normal sample and tumor sample for individual patient *k*. Therefore, the measurement $e_{ij}^{\text{Patient } k}$(i.e. Equation 1) could more accurately represent the personalized state transition of an individual patient in cancer development, which integrates the gene somatic mutations, personalized gene expression and network topology information in the prior-known human genetic interaction network.

And based on the network edge scores related to each gene in PGIN, the node weight w_i was calculated with the following formula:

$$w_i = \sum_{j \in N(i)} e_{ij} \tag{5}$$

where $N(i)$ represents the neighboring node set of node i in PGIN. Therefore, the PGINs can reveal the significant gene interactions between the normal and tumor samples for each patient during disease development, with weight information on nodes and edges.

Weight-NCUA method

In network systems with adequate knowledge of the underlying wiring diagram, disregarding specific functional forms, the FVS-based controllability (FC) methods can identify driver nodes to drive the system state into any desired dynamical attractor. Under the FC framework, NCUA has been proposed to investigate the controllability of undirected structural networks, selecting a minimum set of driver nodes to realize a undirected structural network controllable (24). However, NCUA does not consider potential multiple sets of driver nodes, which would cause the underestimation of different drivers' importance. Thus, here we proposed the weight-NCUA method to find optimal driver node set by using the network edge weight information (i.e. network edge score).

Theoretically, weight-NCUA uses the following dynamic equation to represent the dynamic behavior of a PGIN.

$$dx_i^k/dt = f(x_i^k, x_{I_i}^k) \tag{6}$$

where x_i^k denotes the expression state of gene i for the patient k, I_i represents the neighborhood gene set of gene i,and $f\,f$ represents the dynamic behavior control law of PGIN for patient k, satisfying the continuous differentiability, dissipativity, and decay conditions (22). Equation (6) represents the dynamic behavior of the gene expression level in PGIN. We assume that each edge in PGIN is bidirectional, thus we can convert PGIN into a bipartite network in which the the upside nodes and bottom side nodes represent the nodes and the edges of the original network, respectively. If the node v_i in the up side is one of nodes for edge v_j in the bottom side, then v_i and v_j are linked in the bipartite network. Based on the FVS controllability theory, weight-NCUA selects the dominated nodes set M in the up side that cover the nodes in the bottom side as the driver nodes, which determine the state of PGIN.

It is known that different dominated nodes sets in the bipartite network may generate multiple sets of PDGs, resulting in a potential bottleneck for identifying the combinatorial drugs of the individual patients. Therefore, we introduced an index $W(M)$ to indicate the quality of the selected PDGs.

$$W(M) = \sum w_i r_i - \lambda \sum r_i \tag{7}$$

where w_i denotes the network edge score related to gene i; r_i is an indicative variable,when gene i is selected as the PDGs, $r_i = 1$, otherwise, $r_i = 0$; $\sum w_i r_i$ denotes the network edge scores of candidate gene sets; $\sum r_i$ denotes the number of candidate PDGs; and λ is the balance parameter to adjust the network edge scores and the number of candidate PDGs.

For weight-NCUA, we expect that the PDGs not only contain the minimum number, but also have the maximum network edge scores. It is required to further measure the quality of candidate set of PDGs. Thus, we select the PDGs by solving the following linear integer programming (LIP):

$$\begin{aligned} &\max\ W(M) = \sum w_i r_i - \lambda \sum r_i \\ &\text{s.t.} \sum_{i \in N(u)} r_i \geq 1\ (\forall u \in V_L), r_i = \{0, 1\} \end{aligned} \tag{8}$$

where V_L denotes the bottom side nodes in the bipartite graph and $N(u)\,N(u)$ denotes the neighborhood nodes in the bipartite graph. This LIP objective function is to obtain the PDGs with the minimum number and the maximum network edge scores. The restriction condition is to ensure that all the edges of PGIN in bipartite network can be covered. Under the dynamic behavior, the state of all genes in a PGIN can be regulated by the detected PDGs. Above optimization problem can be solved by using the LP-based classic branch and bound methods (35) or other objective optimization algorithms (36–38).

Synergistic effect evaluation

Because of the limited number of individual patient-specific cancer samples, it is possible to select the effective synergistic drug combinations with target network-based methods (10–13). The key point of target network-based methods is to construct a reliable drug-target network by using data from various sources, then develop a network-based learning method to predict the drug synergy. For example, DrugComboRanker was proposed to rank the paired drug combinations by targeting their corresponding signaling modules in cancer-specific networks. However, DrugComboRanker considers the combinatorial drug prediction of conventional patient-cohort, but ignores the discovery of personalized combinatorial drugs.

Considering the above facts, CPGD evaluates the synergistic effect of paired drug combinations by combining drug target related information with drug similarity from the Connectivity Map (CMAP) database (39), as well as the drug chemical structure similarity (40). The synergistic effect of CPGD consists of two parts, i.e. Drug Target (DT) score, Drug Function and Drug Chemical structure (DFDC) score. The synergistic effect of paired drugs on individual patient k is evaluated by using the following formulas:

$$\begin{aligned} Synergy\_score(\text{drug}A, \text{drug}B, \text{Patient}k) = {} & DT\_score(drug\,A, drug\,B, k) \\ & + DFDC\_score(drug\,A, drug\,B) \end{aligned} \tag{9}$$

On one hand, the DT score is defined as,

$$\begin{aligned} & DT\_score(\text{drug}A, \text{drug}B, k) = Disease\_score(\text{drug}A,\ M, k) \\ & * Disease\_score(\text{drug}B,\ M, k) * Jaccard\_score(\text{drug}A, \text{drug}B, k) \\ & * GO\_score(\text{drug}A, \text{drug}B, k) + DFDC\_score(\text{drug}A, \text{drug}B) \end{aligned} \tag{10}$$

$$Disease\_score(\text{drug}A,\ M, k) = -\log_{10}(P\text{−value}(DM_A^k,\ M)) \tag{11}$$

$$Jaccard\_score(\text{drug}A, \text{drug}B, k) = \text{Jaccard}(\text{drug}A, \text{drug}B, k) \tag{12}$$

where P-value is calculated for evaluating the significance level by hyper-geometric test.

$$P\text{-value}(DM_A^k,\ M) = \sum_{o_i=o_s}^{o} p(o, o_i) \tag{13}$$

$$P(o, o_i) = \frac{\binom{s}{o_i}\binom{N-s}{o-o_i}}{\binom{N}{o}} \tag{14}$$

where$o_s o_s$ denotes the number of intersected genes between driver gene module DM_A^k and disease related gene module M; N denotes the total number of genes in PGIN; o and s are the number of nodes in DM_A^k and M, respectively.

$$\text{Jaccard}(\text{drug}\,A, \text{drug}\,B, k) = \frac{\left|DM_A^k \cap DM_B^k\right|}{\left|DM_A^k \cup DM_B^k\right|} \tag{15}$$

where DM_A^k and DM_B^k denotes the driver gene module of drug A and drug B, respectively, i.e. $DM_A^k = \{t_{Ai}^k\}, DM_B^k = \{t_{Bj}^k\}, i \in [1,n], j \in [1,m]$.

$$GO\_score(\text{drug}\,A, \text{drug}\,B, k) = \frac{\sum_{i,j} sim(t_{Ai}^k, t_{Bj}^k)}{(m+n)(m+n-1)} \tag{16}$$

where $sim(t_{Ai}^k, t_{Bj}^k)$ is the semantic similarity of gene ontology (GO) annotations of t_{Ai}^k and t_{Bj}^k based on the Gene Ontology (GO) term profiles (41). $t_{Ai}^k t_{Bj}^k t_{Bj}^k$ denotes the i-th target of drug A and j-th target of drug B for individual patient k respectively. The semantic similarity is calculated with GOSemSim R package under the 'measure' parameter setting to 'wang' (42). The GO similarity was provided in the folder 'Drug_Targets_GO_similarity_Data' of https://github.com/NWPU-903PR/CPGD. On the other hand, the DFDC score are defined as following,

$$DFDC\_score(\text{drug}\,A, \text{drug}\,B) = C\_\text{map}(\text{drug}\,A, \text{drug}\,B) + C\text{hemical}(\text{drug}\,A, \text{drug}\,B) \tag{17}$$

where C_map(drug A, drug B) denotes the drug function similarity between drug A and drug B based on genomic profiling data of drugs, which are available in the CMAP (39). CMAP dataset consists of 6100 gene expression profiles of four cancer cell lines (i.e. MCF7, PC3, HL60 and SKMEL5) treated by 1309 drugs at different doses. Chemical(drug A, drug B) denotes the chemical structure similarity between drug A and drug B.

In details, the drug similarity metric proposed by Iorio *et al.* (43) is defined as the drug similarity in the CMAP dataset as follows. First, for each individual drug at each dose, genes were ranked based on their fold changes (i.e., drug treatment versus control). Then, gene rank lists at different doses were merged into one gene rank list by using a hierarchical majority voting scheme. Consequently, gene signatures for individual drugs were created by optimally selecting the top- and bottom-ranked 250 genes and the gene set enrichment analysis (44). The drug function similarity file in the CMAP dataset was provided in Supplementary file 3. The code for calculating above drug function similarity was provided in supplementary note 5 of Supplementary file 1. Finally, the dissimilarity $S_G(A, B)$ between drug A and drug B into the similarity score was converted as follows,

$$C\_\text{map}(\text{drug}\,A, \text{drug}\,B) = 1 - S_G(A, B) \tag{18}$$

To obtain the chemical structure similarity, we firstly collected the drug SMILES information from DrugBank (30), and then calculated the Extended Connectivity Fingerprints (ECFP) (45) with a radius of six (ECFP_6) by RDKit python package. Finally, the chemical structure similarity of each drug pairs was calculated by Tanimoto coefficient (40). The drug chemical structure similarity file was provided in Supplementary file 3.

Side effect evaluation

To calculate the *side effect score* of a given drug pair, we firstly collected the drug-target network with the information of activation and inhibition interaction. Then we gave the classification results of sharing targets of two drugs for characterizing the effect of a given drug pair. The configurations (+,+) and (–,–) of the sharing targets of two drugs are referred as coherent, where the action of one drug on the sharing targets is reinforced by the presence of a second drug. The configuration (+,–) of the sharing targets of two drugs is called incoherent, where the action of one drug on the sharing targets is mitigated by the presence of a second drug. Finally, the *side effect score* (32) can be calculated by using the following formula.

$$b_{ij} = \text{sign}(s_{ij}^{+-} - s_{ij}^{++} - s_{ij}^{--}) \tag{19}$$

where sign denotes the signum function; $s_{ij}^{++}, s_{ij}^{+-}, s_{ij}^{--}$ denote the number of sharing targets with configurations (+, +), (+, –) and (–, –) for drug pair (i, j), respectively. The computational procedure is shown in Supplementary Figure S2 (Supplementary file 1).

Enrichment analysis of the PDGs in prior-known cancer driver gene lists

To estimate the significance of overlap between the predicted PDGs and the gold-standard cancer driver gene lists such as Cancer Census Genes (CCG) (46) and Network of Cancer Genes (NCG) (47), we computed the P-value by the hyper geometric test (48) as follows.

$$p = P(x \geq k) = \sum_{x=k}^{\infty} \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \tag{20}$$

where N is the number of genes in PGIN, K is the number of a given cancer driver gene lists (e.g. lists of the personalized Differential Expression Genes (DEG), CCG and NCG), k is the number of the predicted PDGs overlapping with the given gene lists, and n is number of the predicted PDGs. The personalized DEGs are selected by calculating the fold-change between the normal sample and tumor sample ($|\log_2(\text{fold-change})| > 1$). If the enrichment P-value is <0.05, we regarded that the predicted PDGs are significantly enriched in certain gold-standard cancer driver gene lists.

RESULTS

Workflow overview and implementation of CPGD algorithm

From dynamical system viewpoint, gene expressions of an individual patient are the biological system variables, varying at different time points. It is the PGIN (i.e. system structure or network edges) that results in the change of gene expression value (i.e. system variable or network nodes) (49). Therefore, CPGD hold a key assumption that the PGIN determines the state transition between normal state and disease state of an individual patient. The input of CPGD is the gene expression data of paired samples (i.e., normal and tumor samples) for individual patients. Meanwhile, the main outputs of CPGD include: (i) prioritization of the personalized combinatorial drugs; (ii) risk assessment for individual patients based on personalized drug pairs; (iii) disease subtyping by side effect quantification of personalized drug pairs on personalized driver genes. As shown in Figure 1, CPGD mainly comprises two steps as follows:

Step 1. Identifying PDGs from genetic data of individual patients

i) **Constructing the PGINs**. CPGD firstly uses paired-SSN method (24) to construct the PGIN where the interactions determine the state transition of an individual patient during disease development. Then, instead of using paired-SSN method on gene expression data alone, we here developed a weight measurement for scoring the edges of PGIN by integrating somatic mutation data and individual gene expression data for an individual patient. This network edge score combines the co-mutation score, personalized co-expression score, and the prior-known human genetic interaction network information. The co-mutation score is defined as jaccard coefficient between two mutated genes/nodes among the population (34). The personalized co-expression score represents the personalized significant difference of two genes' expression association between normal sample and tumor sample of an individual patient. In this work, the PGINs are weighted graphs in which nodes represent genes, and edges denote the significant difference of gene interactions between the normal and tumor state at gene expression and mutation levels simultaneously.

ii) **Identifying PDGs with weight-NCUA method on PGINs**. In this work, one main focus is how to determine the state transition between disease state and normal state by targeting the PDGs at the gene expression level. Our recent work showed that traditional MMS-based controllability methods (20) ignore the fundamental nonlinear dynamics of system, which may lead to many false positive results. Thus, based on FVS control theory, a structural network controllability method (called NCUA), was developed in our previous work (24) which focus on how to choose proper subset nodes (i.e. driver nodes) for driving the network from initial state to desired stable state by proper input signals. However, NCUA overlooks the weight information of nodes/edges, which may result in a potential bottleneck for identifying the optimal PDGs. In this work, by considering the weight information of nodes/edges in PGIN, we introduced a novel structural network controllability method (namely weight-NCUA) to identify PDGs. Briefly, weight-NCUA tries to design a fitness index for representing the impact quality of the PDGs set, and to determine the PDGs by identifying optimal dominated node set with maximum impact quality to cover all the edges in PGIN. More details of weight-NCUA were shown in Materials and Methods.

Step 2 Screening the role of personalized combinatorial drugs by targeting the PDGs. The PDGs are thought to be the candidate drug targets, which can drive individual biological system from disease state to normal state (or approximate normal state) through drug activation signals. Holding this key assumption, we screened the role of personalized combinatorial drugs by using the following aspects:

i) **Prioritization of personalized combinatorial drugs.** We prioritized the potential personalized anti-disease combinatorial drugs by measuring the number of targeting PDGs for a given drug combination.

ii) **Evaluating the synergistic effect of pairwise drug combination.** Based on the ranking of candidate combinatorial drugs for each individual patient, we firstly selected the pairwise drug combinations among top 10 candidates for each patient, then evaluated the synergistic effect of pairwise drug combinations by integrating drug targets similarity, drug function similarity in the CMAP database and drug chemical structure information.

iii) **Exploring risk assessment of pairwise drug combination.** For a given pairwise drug combination, we firstly selected the union set of targeting PDGs of individual patients as the features of indicating risk assessment. Then, based on gene expression data with selected gene features, the unsupervised clustering method (called similarity network fusion, SNF) (50) was used for subtype identification. Finally, the survival outcomes for patients in these clusters were evaluated by Kaplan–Meier statistics.

iv) **Enhancing the cancer subtyping with side effect quantification on PDGs.** We calculated the *side effect score* of each drug pair by quantifying side effect on PDGs within known cancer driver genes (i.e. CCG and NCG) in the drug-target network with activation and inhibition interactions for each patient. Based on the *side effect score* of each drug pair, we obtained the number of drug pairs with an *aggravating effect* (i.e. *side effe*ct score >0), and the number of drug pairs with *enhancing effect* (i.e. s*ide effect* score < 0), which are two side effect signatures of individual patients.

Determination of the reference gene interaction networks and parameters in CPGD

To assess the effect of different sources of prior-known gene interaction networks on the performance of CPGD, we have adopted six gene interaction networks from different literatures. The reference Network 1 was built by Hou *et al.* (19),

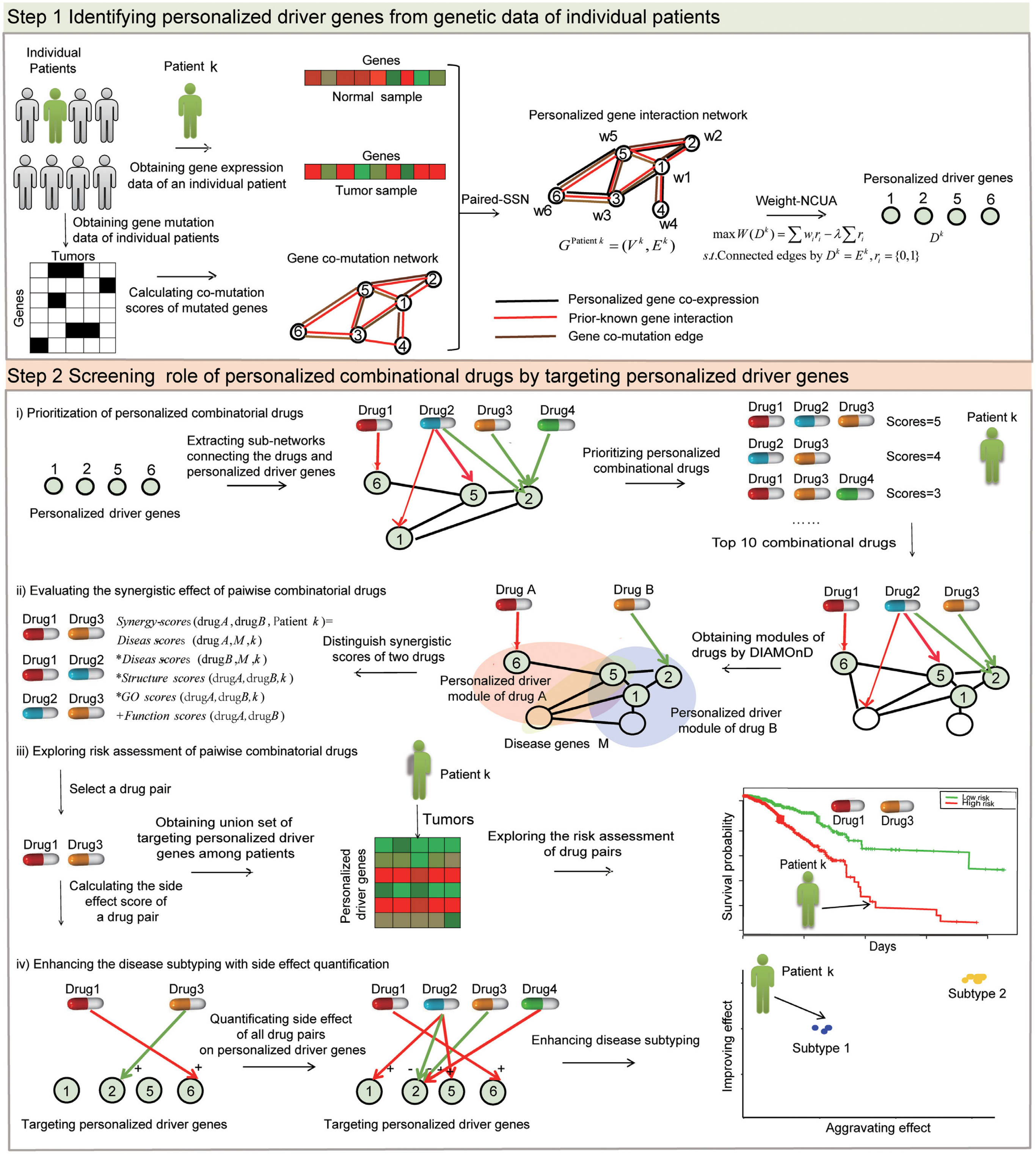
Step 1 Identifying personalized driver genes from genetic data of individual patients
Individual Patients
Patient k
Obtaining gene expression data of an individual patient
Genes
Normal sample
Genes
Tumor sample
Obtaining gene mutation data of individual patients
Tumors
Genes
Calculating co-mutation scores of mutated genes
Gene co-mutation network
Paired-SSN
Personalized gene interaction network
w5
w2
w1
w6
w3
w4
$G^{\text{Patient }k} = (V^k, E^k)$
Weight-NCUA
$\max W(D^k) = \sum w_i r_i - \lambda \sum r_i$
$s.t.$Connected edges by $D^k = E^k, r_i = \{0,1\}$
Personalized driver genes
1 2 5 6
D^k
Personalized gene co-expression
Prior-known gene interaction
Gene co-mutation edge
Step 2 Screening role of personalized combinational drugs by targeting personalized driver genes
i) Prioritization of personalized combinatorial drugs
Personalized driver genes
Extracting sub-networks connecting the drugs and personalized driver genes
Drug1 Drug2 Drug3 Drug4
Prioritizing personalized combinational drugs
Drug1 Drug2 Drug3 Scores=5
Drug2 Drug3 Scores=4
Drug1 Drug3 Drug4 Scores=3
……
Top 10 combinational drugs
Patient k
ii) Evaluating the synergistic effect of paiwise combinatorial drugs
Drug1 Drug3
Drug1 Drug2
Drug2 Drug3
$Synergy\text{-}scores$ (drug$A$, drug$B$, Patient k)=
$Diseas\ scores$ (drug A,M,k)
*$Diseas\ scores$ (drug$B$,M,k)
*$Structure\ scores$ (drug$A$, drug$B$,k)
*$GO\ scores$ (drug$A$, drug$B$,k)
+$Function\ scores$ (drug$A$, drug$B$)
Distinguish synergistic scores of two drugs
Drug A
Drug B
Personalized driver module of drug A
Personalized driver module of drug B
Disease genes M
Obtaining modules of drugs by DIAMOnD
Drug1 Drug2 Drug3
iii) Exploring risk assessment of paiwise combinatorial drugs
Select a drug pair
Drug1 Drug3
Calculating the side effect score of a drug pair
Obtaining union set of targeting personalized driver genes among patients
Patient k
Tumors
Personalized driver genes
Exploring the risk assessment of drug pairs
Survival probability
Days
Low risk
High risk
Drug1 Drug3
Patient k
iv) Enhancing the disease subtyping with side effect quantification
Drug1 Drug3
Targeting personalized driver genes
Quantificating side effect of all drug pairs on personalized driver genes
Drug1 Drug2 Drug3 Drug4
Targeting personalized driver genes
Enhancing disease subtyping
Patient k
Subtype 1
Subtype 2
Improving effect
Aggravating effect

Figure 1. CPGD overview. **Step 1:** Paired-SSN is used to construct the PGIN for capturing the phenotypic transitions between normal and disease states. Instead of using Paired-SSN method on gene expression data alone, we introduce network edge score for measuring the edges of PGIN by integrating cancer type-specific somatic mutation data into PGIN. The edges of PGIN integrate co-mutation scores, personalized gene co-expression scores, and prior-known interactions. Then, an improved structural network controllability method (called weight-NCUA) is developed to identify the PDGs, where the driver genes are considered as candidate drug targets towards the desired control objective by drug activation. **Step 2**: CPGD screens the role of personalized combinatorial drugs from several biomedical aspects. (i) At first, CPGD will prioritizes the personalized combinatorial drugs by measuring the number of targeting PDGs. (ii) Second, it can explore synergistic effect of drug pairs by a few sub-steps: to select drug pairs from top 10 candidate combinatorial drugs for each patient as candidate drug pairs; to evaluate the synergistic effect of these candidate pairwise drug combinations, which measures the synergistic

which consists of 11 648 genes and 211 794 edges by integrating a variety of data sources, such as MEMo (51), Reactome (52), NCI-Nature Curated PID (53) and KEGG (54). The reference Network 2 was built by Quan *et al*. (27) from the Synthetic Lethality genes interactions Database (SynLethDB), which consists of 6513 genes and 19 955 synthetic lethal gene pairs for humantumors. The reference Network 3 was constructed by Vinayagam *et al*. (55), which consists of 6339 proteins and 34 813 edges, where the edge denotes the hierarchy of signal flow between the interacting proteins.The reference Network 4 was constructed in reference (56). The Network 5 was collected from STRING dataset (https://string-db.org/) whose edge scores are higher than 900. The Network 6 consists of gene interactions by removing co-expression edges, the literature-derived interactions, and predicted interactions from Network 1 (24). Above six networks were provided in the folder 'All gene interaction networks used CPGD' of https://github.com/NWPU-903PR/CPGD.

By using each reference gene interaction network, CPGD outputs the PDGs for each individual patient and ranks the candidate combinatorial drugs according to the number of targeting PDGs. Based on the ranking of candidate combinatorial drugs for each individual patient, we assessed the effect of different reference gene interaction networks on CPGD in terms of discovery rate (DR) among prior-known clinical efficacious combinatorial drugs, $DR = \sum_{k=1}^{n} p_k/n$, where p_k denotes the fraction of the top k predicted combinatorial drugs within the Clinical Anti-disease Combinatorial drugs for treating disease, and n is the number of top ranked anti-disease drug (here, $n = 10$).

The performance of CPGD with different reference gene interaction networks on BRCA and LUNG datasets were shown in Figure 2A and B. We can find that: (i) compared with other networks, Network 1 with more complete gene interactions have more stable performance in BRCA and LUNG datasets, which may be a general suggestion as prior-known network; (ii) Network 1 has the highest *discovery rate* for LUNG, while Network 2 has the highest *discovery rate* in BRCA; (iii) the balance parameter (λ) has different effects to CPGD on BRCA and LUSC cancer datasets, and the reference networks. The discovery rate of CPGD with λ= 0.01 in BRCA and with λ= 10 in LUNG is the highest, which were selected for follow-up analysis.

The choice of proper prior-known network structure is an important factor for CPGD. According to the results in Figure 2A and B, several suggestions for choosing more proper prior-known networks were concluded as follows: (i) when the gold-standard of anticancer drug combinations is available, we can choose the prior-known network with highest performance; (ii) when the gold-standard of anticancer drug combinations is not available, Network 1 with more stable performance could be considered as prior-known network.

Furthermore, the performances of CPGD in different subtypes on BRCA (Network 2) and LUNG (Network 1) cancer datasets were shown in Figure 3A and B, from which we can see that the discovery rates of CPGD with λ= 0.01 and λ= 10 are the highest for BRCA and LUNG respectively, which are consistent in different cancer subtypes. The common driver genes predicted by CPGD in different subtypes from BRCA and LUNG cancer datasets were listed in Supplementary file 4. The comparisons of common driver genes for various different subtypes in breast and lung cancer datasets were shown in Supplementary Figure S3, from which we found that the subtype-specific driver genes are different in different cancer datasets.

In addition, we also found that CPGD can obtain novel predictions in the top-ranked drug combinations, besides those already in clinical trials (Supplementary Figures S11 and S12). CPGD can identify at least one novel drug combinations among top 10 predicted drug combinations for Lum A, Lum B, HER2, and Basal subtypes on BRCA cancer dataset, while CPGD can identify at least nine novel drug combinations among top 10 predicted drug combinations for LUSC and LUAD subtype-specific patients from LUNG cancer dataset. Based on the mean ranking of subtype-specific cancer patients, we also gave the full ranking list of drug combinations in different subtypes of BRCA and LUNG cancer datasets (Supplementary file 5). We found that there are some novel predictions in BRCA and LUNG cancer datasets. For example, DC000222 (i.e. TRASTUZUMAB, DOCETAXEL, CARBOPLATIN, LETROZOLE and LAPATINIB) and DC002977 (i.e. TRASTUZUMAB,DOCETAXEL, CARB OPLATIN,DOXORUBICIN and CYCLOPHOSPHAMI DE) are novel predicted drug combinations for breast and lung cancer patients, respectively.

Evaluation of detection robustness of PDGs on gene expression data

We used the gene expression data of subtype-specific patients to obtain the PDGs, and calculated the jaccard coefficient between the PDGs of subtype-specific gene expression data and those of all gene expression data (Supplementary Figure S4). We found that the jaccard coefficient is larger than 0.6 when the balance parameter λ varies from 0.6 to 1, demonstrating that the prediction results of subtype-specific gene expression data are similar or consistent with those of all gene expression data. When λ increases, the jac-

effect of their corresponding PDGMs by considering disease related gene module information and multi-sources drug function information. (iii) Third, CPGD could explore risk assessment of drug pairs. For a given pairwise drug combinations among above candidate drug pairs, CPGD selects the union set of targeting PDGs of individual patients as the gene features to explore the risk assessment. Based on gene expression data with selected gene features, SNF is used for subtype identification and the survival outcomes for patients in these clusters are evaluated by Kaplan–Meier statistics. iv) Finally, CPGD can enhance subtyping by side effect quantification on PDGs. When two drugs act simultaneously on the same target genes, there will be two kinds of actions of two combinations, e.g. *coherent* action as (+, +) and (−, −), and *incoherent* action as (+, −). By attaching signs to the mechanisms of different actions,the *side effect score* can be calculated in the drug-target network with activation and inhibition interactions for each patient. Based on the side effect score of each drug pair, we can obtain the number of drug pairs with aggravating effect (i.e. *side effect score*>0), and the number of drug pairs with enhancing effect (i.e. side effect score < 0), which are two side effect signatures of individual patients.

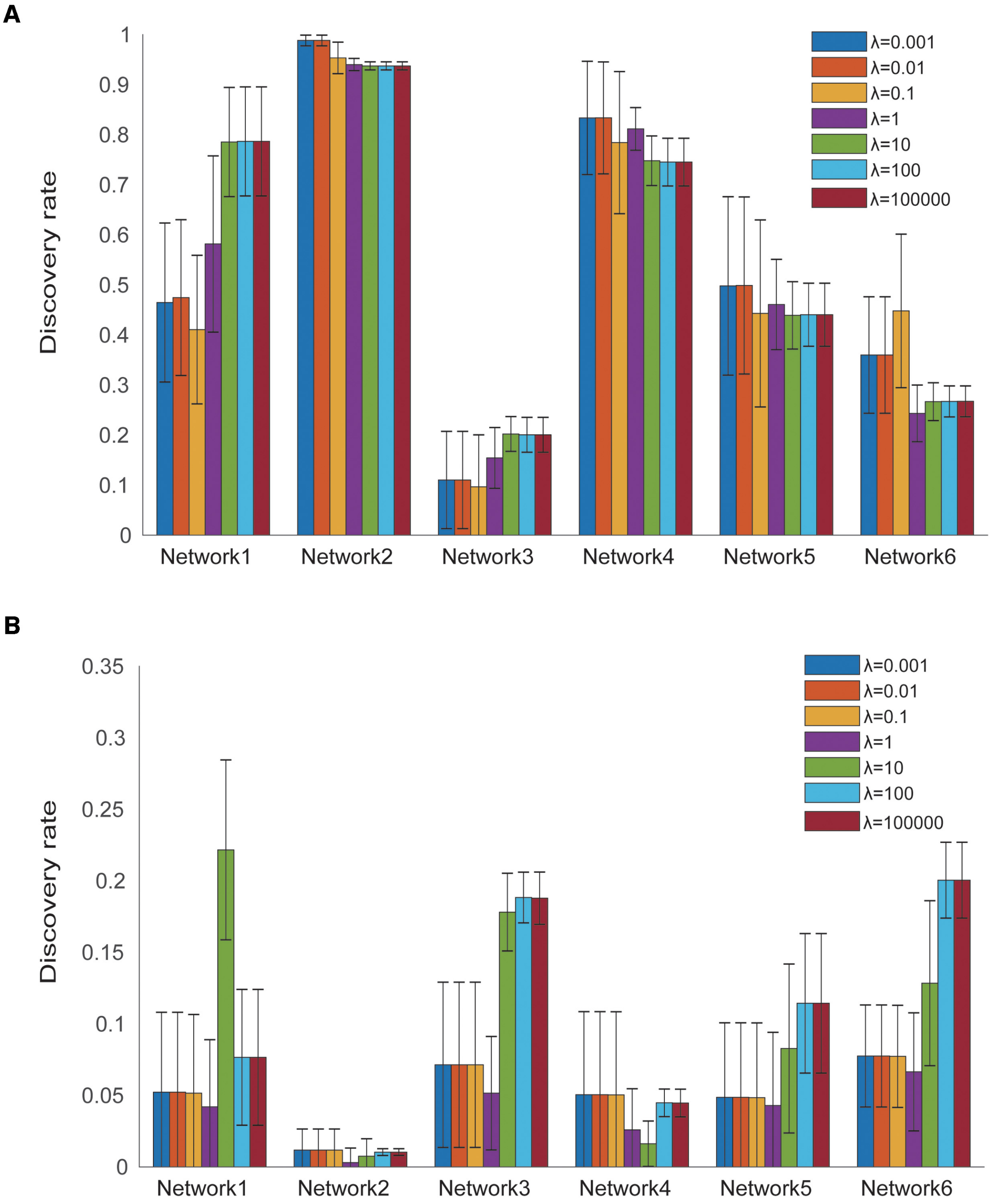

Figure 2. Effect of the reference gene interaction networks and parameters in CPGD. (**A**) Results on six prior-known networks with different balance parameters for BRCA cancer dataset. (**B**) Results on six prior-known networks with different balance parameters for LUNG cancer dataset.

card coefficient becomes larger. These results suggested that CPGD can robustly identify PDGs from high-throughput expression data.

Comparisons and evaluation of CPGD with existing driver gene-focus methods

One key contribution of CPGD is to identify driver genes for inferring combinatorial drugs. To evaluate the effectiveness of CPGD, we compared CPGD with other state-of-the-artd river gene identification methods, such as DriverML (14), MutSigCV (16), OncoDriveFM (17), SCS (18), DawnRank (19), PNC (24), pDriver (https://www.biorxiv.org/content/10.1101/2020.04.23.058727v1) and ActiveDriver (57) (Figure 4). We also compared CPGD with GeneRank (58,59), HotNet2 (60) and Hub genes-based methods (Figure 4 and Supplementary Figure S5 in Supplementary file 1). The number of PDGs for these methods

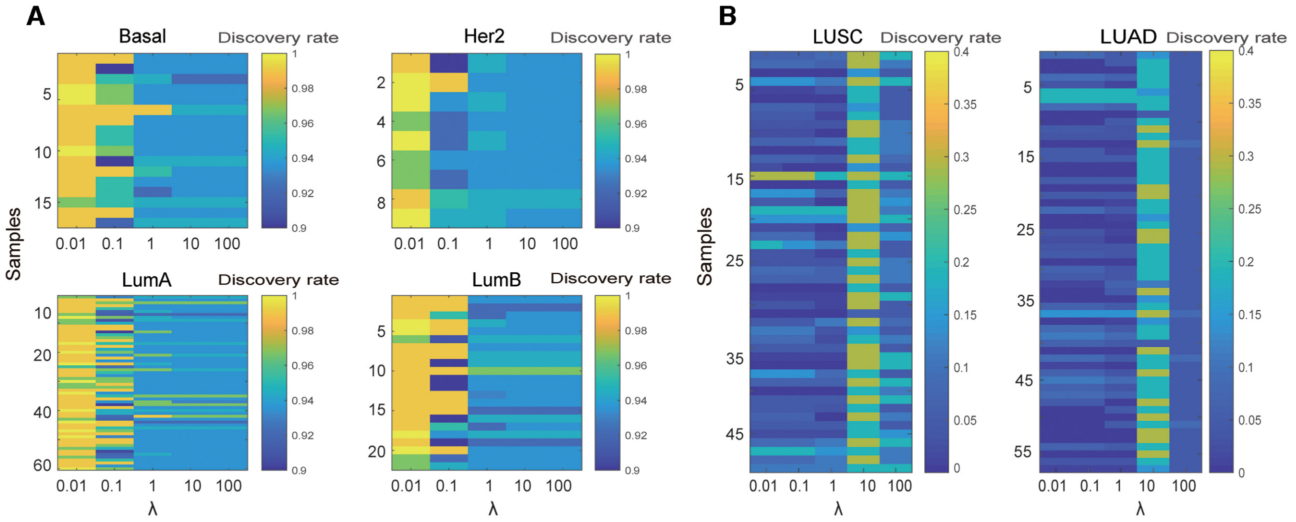

Figure 3. Drug combination prediction results in a subtype-specific manner. (**A**) Results on BRCA cancer dataset. (**B**) Results on LUNG cancer datasets.

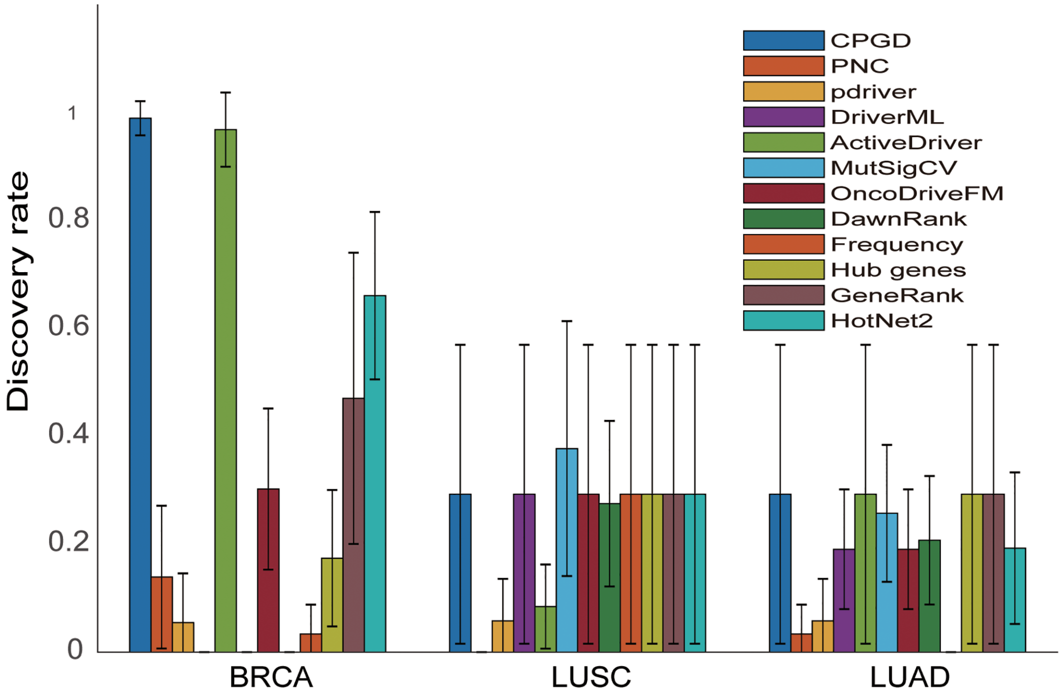

Figure 4. The comparison and evaluation of CPGD with other existing methods for predicting clinical efficacious combinatorial drugs on different cancer datasets. The y-axis denotes the mean *discovery rate* of the top 10 predicted combinatorial drugs. The error bar denotes the standard derivation of *discovery rate* among the top 10 predicted combinatorial drugs. The bar colors represent different algorithms.

is same as those of CPGD (Supplementary note 3 in Supplementary file 1). Based on the information of sub-networks between the drugs and PDGs, personalized combinatorial drugs including the drug–PDG interactions and PDG interactions were prioritized to evaluate the ability of predicting clinical efficacious combinatorial drugs. As shown in Figure 4, the *discovery rate* of CPGD has consistent higher performance than those of other 11 driver gene identification methods on two benchmark cancer datasets, indicating that the ability of CPGD is superior for predicting clinical efficacious combinatorial drugs.

Comparisons of CPGD with existing combinatorial drug prediction methods for predicting synergistic drug pairs

To demonstrate the effectiveness of CPGD on combinatorial drug prediction,by considering disease related gene module information and multi-sources drug function information, we evaluated the synergistic effect of drug pairs among top 10 candidate combinatorial drugs for each individual cancer patient on BRCA dataset. The main function of CPGD for measuring the synergistic effect of pairwise drug combinations includes the drug-gene interactions and disease related gene module collection, driver gene module construction and synergistic effect evaluation.

(i) For drug–gene interactions and disease related gene module collection, we collected the interactions between drugs and genes, as well as the disease related gene module to identify anti-disease drug pairs for risk assessment. On BRCA cancer dataset, the interactions between drugs and targeted genes were extracted from the combinatorial drugs and gene interaction network (Supplementary file 2). A list of 2341 breast cancer related genes collected by Quan *et al*. (27) from the Unified Medical Language System (UMLS) (61) was available in folder 'List of Breast_cancer_genes' of https://github.com/NWPU-903PR/CPGD.

(ii) For driver gene module construction, PDGMs were identified by DIseAse MOdule Detection (DIAMOnD) method (62). More computational details of DIAMOnD were shown in supplementary note 3 of Supplementary file 1. The targeting driver gene number, driver gene module number and driver gene module scores of candidate drugs in BRCA dataset were identified from top 10 combinatorial drugs for individual patients, which were shown in Supplementary Figures S6–S8, respectively. Different drugs have different distributions in terms of the size of PDGMs in individual patients. These results demonstrated that individual heterogeneity for combination therapy should be considered in the disease treatment. The results in Supplementary Figure S8 showed that the module scores of identified PDGMs could be convergent within 10 interaction times.

(iii) For synergistic effect evaluation, due to the rapid development of network biology, it raised the possibility of exploring individual samples based methods with synergistic effects for disease treatment (33,63–66). From the perspective of network-based methods, CPGD measures, the synergistic effect of their corresponding PDGMs by considering disease related gene module information and multi-sources drug information, such as the drug target similarity, as well as drug similarity from the CMAP database (39) and drug chemical structure similarity (40) into their corresponding PDGMs for evaluating the synergistic effect of pairwise drug combinations.

To validate if the predicted or top-ranked drug combinations generate the synergistic effects, some synthetic lethal gene pairs and their corresponding targeted drug pairs were firstly extracted. Among these drug pairs, pairwise drug combinations with clinically validated anticancer activity were considered as the benchmark for verifying whether the drug combination is a synergistic drug combination (Supplementary file 2) (27). Then, we ranked the drug pairs by calculating the mean synergistic scores among all BRCA cancer patients, obtaining the clinical *Discovery rate* by calculating the mean fraction of the top k ($k = 1, 2, \ldots, 10$) pairwise drug pairs within the Clinical Anti-cancer Combinatorial drugs for treating breast cancer patients. Finally, we evaluated synergistic effect of drug pairs among top 10 candidate combinatorial drugs for each individual patient. As comparison base, we used CombRanker on PDGMs of pairwise drug combinations to obtain the synergistic scores for each individual patient (Figure 5). Main conclusions could be derived from Figure 5 as follows:

(i) Figure 5A showed the Discovery rate results of CPGD and three synergistic combinatorial strategies in CombRanker (10). We can see that CPGD performs better than any of the existing synergistic combinatorial strategies.

(ii) In fact, synergistic combinatorial strategy of CPGD consists of two parts, i.e. DT score, Drug Function and DFDC score. To demonstrate the effeciency of these two parts, we calculated the discovery rate of CPGD, DT score alone and DFDC score alone, respectively. Figure 5B showed that DFDC score has more contributions on the overall synergistic effect than DT score.

(iii) To show the robustness of synergistic combinatorial strategy of CPGD, we obtained all possible drug pairs and the corresponding targets (i.e. 452 drug pairs) from Supplementary file 2 and obtained synergistic scores of these drug pairs by using CPGD. Based on synergistic scores of these 452 drug pairs, Figure 5C, D shows the Discovery rate results of CPGD and three synergistic combinatorial strategies in CombRanker (10), and CPGD performed better than any of the existing synergistic combinatorial strategies. Furthermore, DFDC score has more contributions on the overall synergistic effect. These conclusions are consistent with those from Figure 5A and B.

Influence of construction methods on single sample network during CPGD analysis

In order to investigate the effect of different network construction methods on CPGD, we adopted different single sample network construction methods, such as paired-SSN (24), SSN (33) CSN (63), SPCC (64,65) and LIONESS (66). For SPCC and LIONESS methods, after obtaining the SPCC and LIONESS co-expression distribution (S) of all gene pairs, we chose a threshold w to filter the edges with low co-expression value, $w = \mu(S) + 2\delta(S)$, where $\mu(S)$ and $\delta(S)$ are the mean value and standard variance for co-expression distribution (S) of all gene pairs. For all network construction methods, we chose Network 2 and Network 1 as the reference gene interaction network for BRCA and LUNG, respectively.

The results of CPGD on different subtypes of cancer patients with single sample network construction methods were shown in Figure 6, from which we can see that the discovery rate of paired-SSN and CSN is higher than that of SPCC and LIONESS methods, and the discovery rate of paired-SSN and CSN is almost equal. Thus, considering the number of samples, here we selected the paired-SSN method to construct the PGIN.

In addition, to demonstrate the effect of mutation data on the results of CPGD, we calculated the *Discovery rate* of CPGD with the integrated (i.e. both mutation and expression) data and the expression (i.e. expression only) data on BRCA and LUNG datasets respectively. As shown in Figure 7, among the top 10 predicted combinatorial drugs on

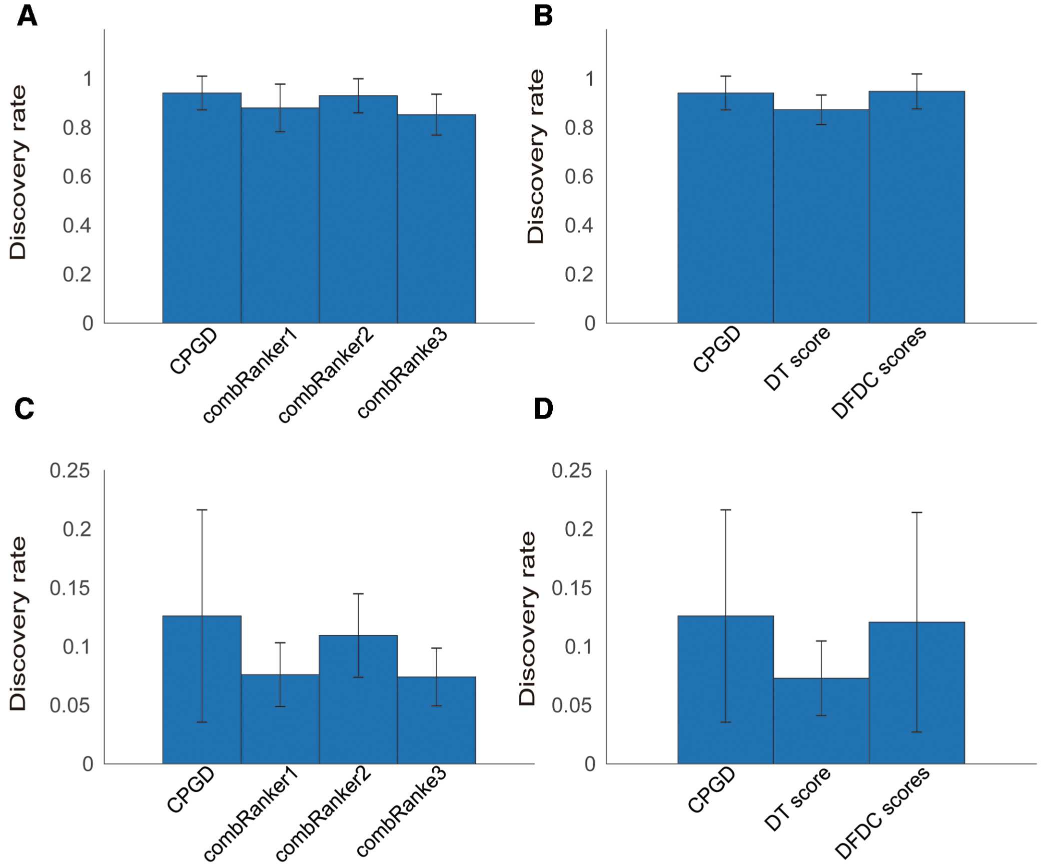

Figure 5. Comparison of CPGD and other synergistic combinatorial strategies for evaluating the synergistic effect of drug pairs. The error bar denotes the standard derivation of *discovery rate* among the predicted drug pairs. (**A**) The discovery rate of CPGD and three existing synergistic combinatorial strategies among top 10 candidate combinatorial drugs for each individual patient. (**B**) The discovery rate of CPGD, CPGD with DT score alone and CPGD with DFDC score alone among top 10 candidate combinatorial drugs for each individual patient. (**C**) The discovery rate of CPGD and three existing synergistic combinatorial strategies among all drug pairs. (**D**) The discovery rate of CPGD, CPGD with DT score alone and CPGD with DFDC score alone among all drug pairs.

BRCA and LUNG datasets, the *Discovery rate* of CPGD with the integrated data are higher than that with the expression data, demonstrating that the integrated multi-omics data can help improve the accuracy of driver gene identification.

We also compared with the mean *discovery rate* of random selected genes for individual patients. From the driver genes determined by CPGD, we randomly generated a gene set in which the number is same as the number of driver genes of CPGD. This random simulation was repeated 100 times to generate the distribution of mean discovery rate among all patients. From Supplementary Figure S9, we can see that the discovery rate of PDGs predicted with CPGD is significantly higher than mean discovery rate of those genes chose with random selection, which is consistent on two benchmark cancer datasets. These results further supported that CPGD can effectively discover anticancer combinatorial drugs.

Influence of different network controllability methods to CPGD

To evaluate the influence of different network controllability methods for CPGD, we compared weight-NCUA with other structural network controllability methods, such as MMS (20), MDS (21), DFVS (22) and NCUA (24) on above constructed PGIN (Figure 8). We also compared weight-NCUA method with another MMS-based critical nodes selection method (called MMS_critical), which identified the critical nodes such that removing such a node will require more nodes to control the network and is also a candidate method for predicting PDGs in PGIN (67,68). As shown in Figure 8, among the top 10 predicted combinatorial drugs on BRCA and LUNG, weight-NCUA has higher performance and robustness than most of other structural network controllability methods on different subtypes of cancer patients. The main reason is that weight-NCUA considers the network edge scores for the optimization of driver nodes, which are usually disregarded by other structural network controllability methods.

We note that Hu *et al*. recently introduced a network controllability-based method, called OptiCon, to discover synergistic driver genes as candidate targets for combination therapy (56). Although OptiCon is related to combination therapy research from network controllability perspective, OptiCon is to discover synergistic paired driver genes on a population of patients and does not focus on evaluating the synergistic effect of paired drugs on individual patients.

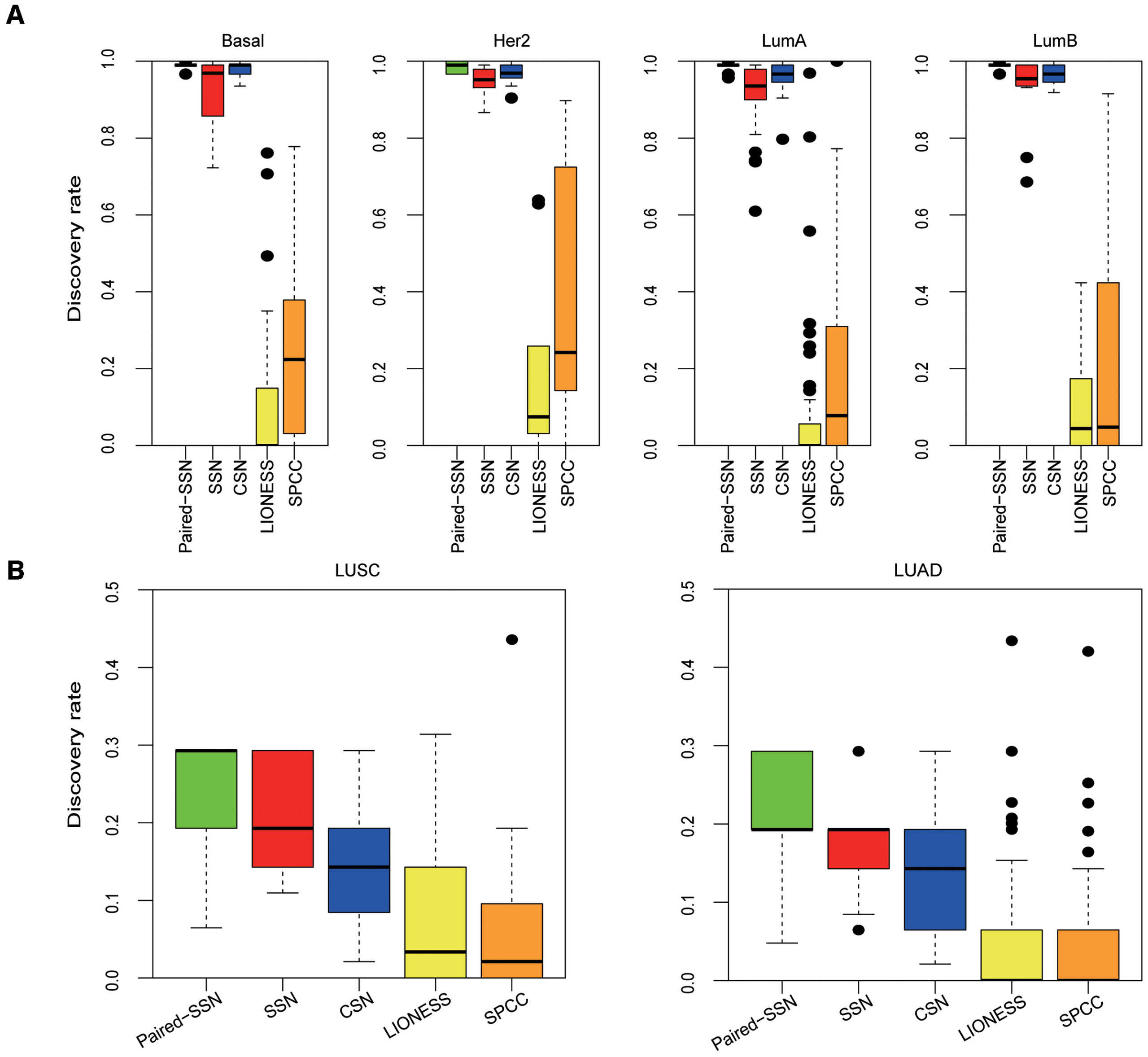

Figure 6. Influence of different single sample network construction methods on CPGD. To evaluate the usage efficiency of single sample network construction methods (i.e., Paired-SSN, CSN, SPCC and LIONESS) for personalized drug discovery, the combinatorial drugs annotated in the Clinical Anti-disease Combinatorial drugs are applied to obtain the *discovery rate* of the top-ranked/predicted anti-disease combinatorial drugs for different subtypes of BRCA (**A**) and LUNG (**B**).

Structural and functional property of personalized driver genes

After determining the suitable reference network, single sample network construction method and network controllability method, CPGD can be effectively applied to identify PDGs. For analyzing the functional and structural properties of detected PDGs, the enrichment *P*-values of PDGs in DEG, CCG and NCG lists were evaluated for CPGD (More computational details were shown in Materials and Methods). As shown in Figure 9, we obtained some new insight on tumor heterogeneity. For the DEG list, not all of patients have significant enrichment results. By contrast, the *P*-values of most patients are <0.05 (ESg = $-\log_{10}(P\text{-value}) < 1.3$) for CCG and NCG lists. These results showed that PDGs can be more completely characterized by CCG and NCG lists than those by DEG lists, indicating the importance and relevance of PDGs on biological and disease functions.

To further demonstrate the functional properties of personalized driver genes, we performed the enrichment pathway analysis for the personalized driver genes to determine if the personalized driver genes are enriched in Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. To identify the significantly enriched pathways of PDGs, we computed the P-value of PDG enriched pathways using the hyper geometric test (48) as described in formula (20), where *N* is the number of genes in the gene interaction network, *n* is the number of PDGs, *k* is the number of

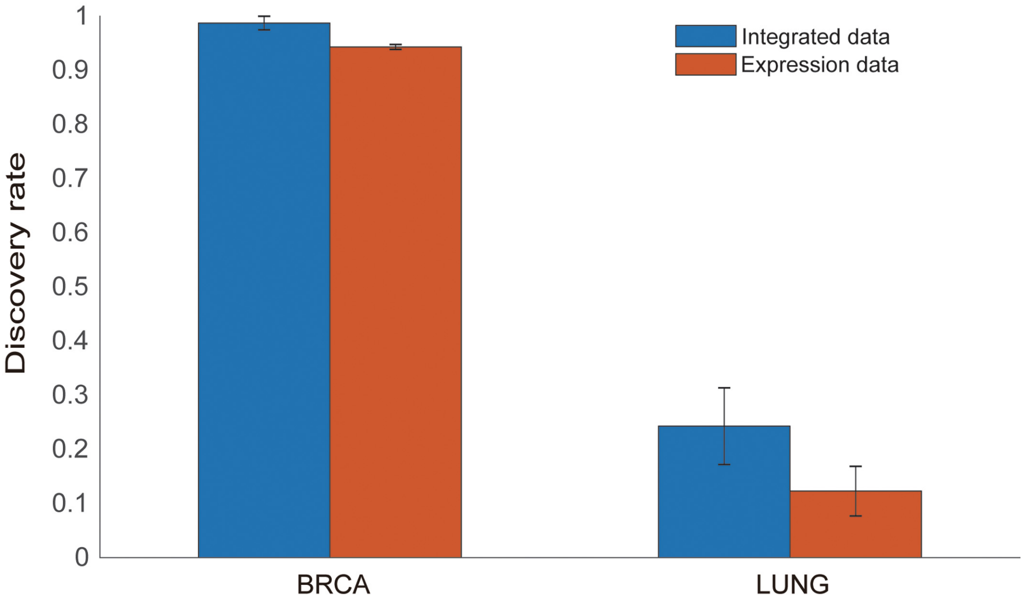

Figure 7. Results of CPGD method with the integrated multi-omics (both mutation and expression) data and the expression (expression only) data for top-ranked/predicted anti-cancer combinatorial drugs on BRCA and LUNG, respectively. The error bar denotes the standard derivation of discovery rate among all patients.

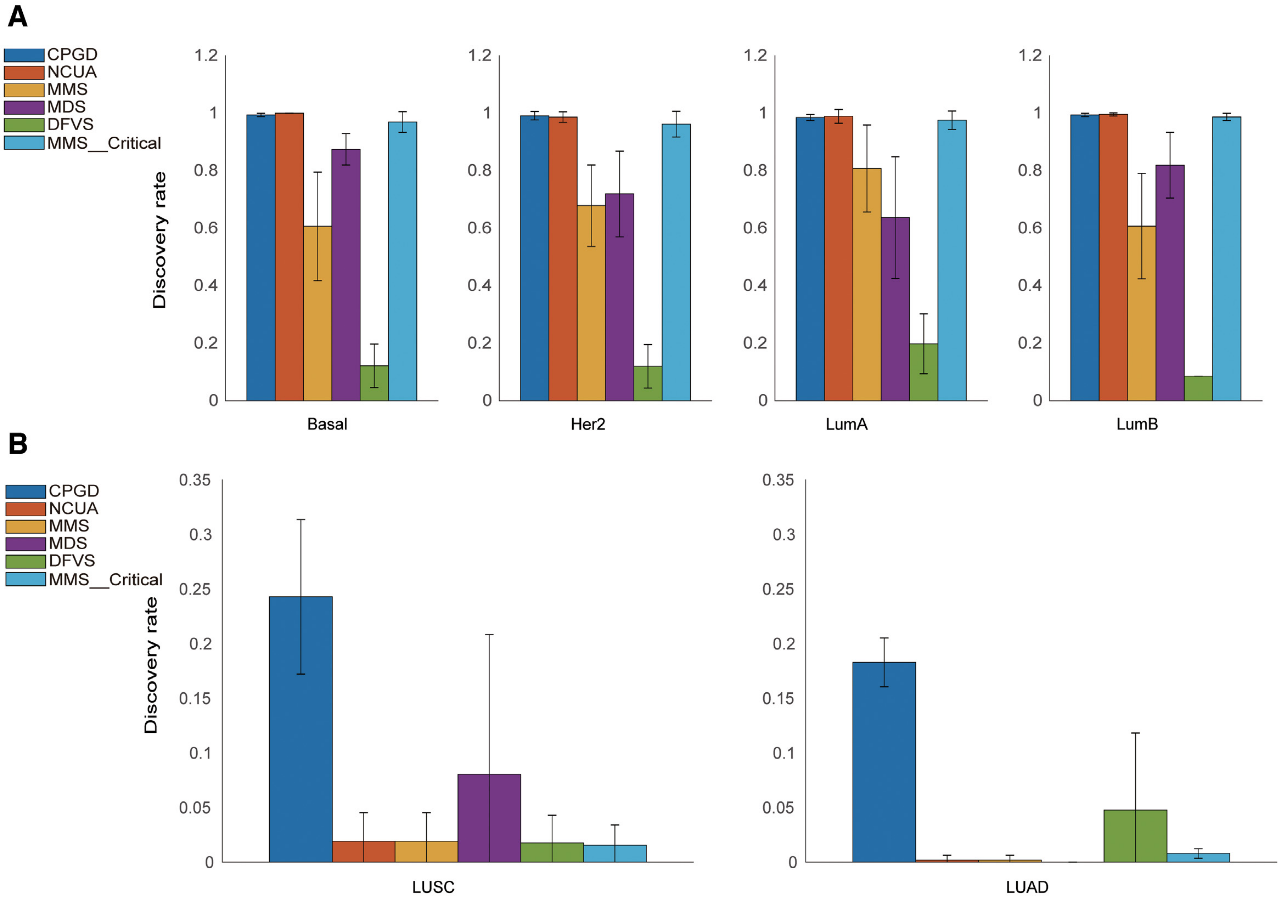

Figure 8. Comparison of weight-NCUA, MDS, MMS, DFVS and NCUA as well as the MMS_critical methods for predicted anti-disease combinatorial drugs. (**A**) Results on BRCA. (**B**) Results on LUNG.

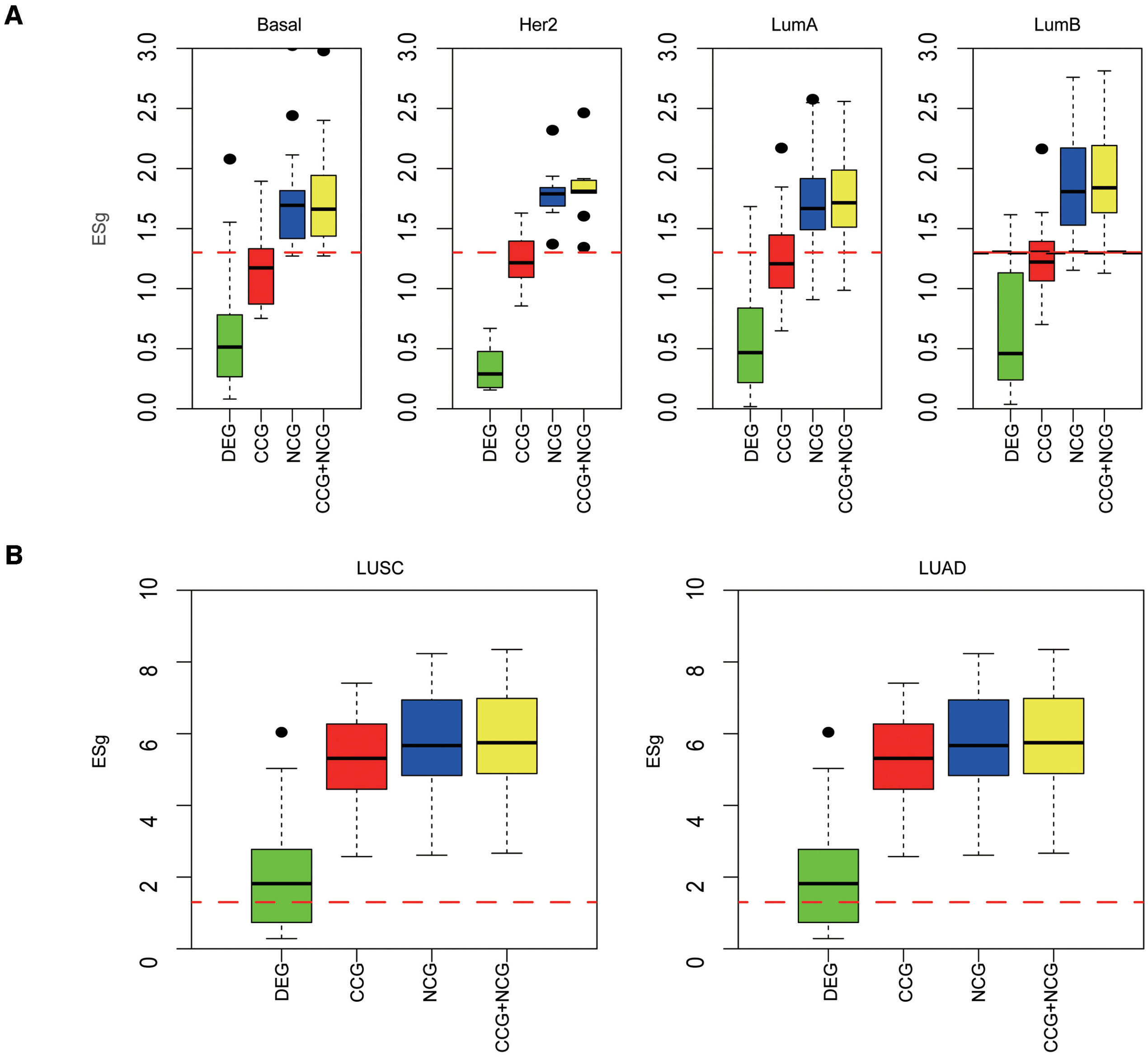

Figure 9. Functional and structural property of personalized driver genes. (**A**) Enrichment results of PDGs identified by CPGD in DEG, CCG and NCG lists on different subtypes of BRCA cancer patients. (**B**) Enrichment results of PDGs identified by CPGD in DEG, CCG and NCG lists on different subtypes of LUNG cancer patients. The enrichment score of driver gene is defined as ESg = –$\log_{10}$(P-value). The red line denotes ESg = –$\log_{10}$(0.05).

PDGs within a given pathway, and K is the gene number in a given pathway. The enrichment results (including pathway name, patient-occurred frequency and related combinatorial drugs) were shown in Supplementary file 6, from which we can find that 63.64% of these identified biological pathways are related with breast cancer, 51.61% related with lung cancer. These results indicated that CPGD can effectively identify the cancer-related pathways which are potentially targeted by drugs. We also found that 81.25% of the reported breast cancer-related pathways in previous studies are enriched in many patients' data with high frequency (>0.6), while 50% of the reported lung cancer-related pathways in previous studies are enriched in many patients' data with high frequency (>0.6), indicating patient heterogeneity varies in different cancer datasets.

Risk assessment of drug pairs with co-targeting of personalized driver genes on breast cancer

To explore risk assessment of pairwise drug combinations, the main outcome of CPGD includes the gene signature selection, subtype identification, and survival evaluation. For gene signature selection, we selected top 10 candidate combinatorial drugs for all patients (#112 samples) on TCGA-BRCA cancer dataset, and further selected the union set of targeted PDGs of individual patients as the signatures to explore the risk assessment of a given pairwise drug combination on patients. For subtype identification, we re-collected the gene expression data of all the tumors on TCGA-BRCA cancer dataset (#1006 samples). Based on targeted PDGs of drug pairs, the SNF (50) was applied on the gene expression data to select the gene signatures for identifying cancer subtypes/clusters. For survival evaluation, the survival outcomes of patients in the identified clusters were evaluated by Kaplan–Meier statistics. We chose the efficacious drug pairs with significant survival analysis results (P-value < 0.05) for risk assessment.

The results of risk assessment of pairwise drug combinations on TCGA-BRCA cancer dataset were shown in Figure 10 and Supplementary file 7. We found that the P-value of 6 pairwise drug combinations is less than that of each single drug, indicating that the combined therapeutic effect of these drug pairs are better than monotherapy effect of single drug.

To further evaluate 6 anti-disease drug pairs for risk assessment of BRCA cancer on independent data, we carried out the risk assessment of these drug pairs using the *SurvExpress* tool (69) on TCGA BRCA data (Supplementary file 7), and the PROGgeneV2 tool (70) on GSE5327-BRCA cancer dataset (Figure 11 and Supplementary Figure S10). We found that (i) six drug pairs can actually divide all patients into discriminative two clusters (P-value $<$ 0.05) on TCGA BRCA data; (ii) among these six drug pairs, three drug pairs (i.e. CETUXIMAB and CARBOPLATIN, CARBOPLATIN and CYCLOPHOSPHAMIDE, CYCL OPHOSPHAMIDE and GEMCITABINE) can actually divide all patients into discriminative two groups (P-value < 0.05) on GSE5327-BRCA cancer dataset; (iii) a few targeting driver genes (also the drug targets) of six paired drugs combination (#<30 targeting driver genes) are able to well (P-value <0.05) partition the cancer patients into subtypes with different survival time; (iv) the CETUXIMAB and CARBOPLATIN was predicted as a novel pairwise drug combination, which can significantly partition the breast cancer patients into two clusters with different survival risk on TCGA-BRCA cancer dataset (Supplementary file 7, P-value = 0.01548) and GSE5327-BRCA cancer dataset (Figure 11, P-value = 0.01624).

Disease subtyping by quantifying side effect signatures for breast cancer

To quantify the side effect of drug pairs on corresponding disease subtypes, we first collected the PDGs within two reliable prior-known cancer driver genes sets (i.e. the CCG and NCG lists), then calculated the *side effect score* of a given drug pair by quantifying their side effect on PDGs within CCG and NCG for each patient, as shown in Supplementary Figure S2 (Supplementary file 1). Finally, we got the number of drug pairs with an *aggravating effect (side effe*ct score > 0), and the number of drug pairs with *enhancing effect* (side effect score < 0), which are used as two side effect signatures of individual patients. By considering these side effect signatures of individual patients, we found that the patients in the breast cancer dataset can be significantly classified into two distinct subtypes (Figure 12A). Furthermore, by exploring the survival analysis of these two subtypes, we found that the greater the value of aggravating effect, the less the survival time on a specific patient subtype (Figure 12B). We also gave the list of common drug pairs with an *aggravating effect* and *enhancing effect* among these two subtypes (Supplementary file 8).

To further demonstrate the efficiency of CPGD on patient subtype recognition, SNF (50) was applied on the gene expression data for separating cancer patients into two subtypes as comparisons. We explored the differences between patient subtype results of using our CPGD and those of using all the gene expression data with SNF directly (Figure 12A–D). For subtypes identified by CPGD, the survival time of patients in high risk subtype are significantly shorter than that of patients in low risk subtype (P-value = 0.0108, Figure 12A and B). For subtypes identified by SNF directly on all gene expression data, there are no significant differences in survival curves among patients in high and low risk subtypes (P-value = 0.2476, Figure 12C). Therefore, compared with SNF of using all gene expression data directly, CPGD can more significantly partition cancer patients into subtypes with different survival time. These results indicated that CPGD can simultaneously distinguish high and low risk subtypes with different survival time well.We also calculated the jaccard coefficient between patient subtypes of CPGD and those of SNF using all the gene expression data directly (Figure 12D) to explore the difference between these two subtyping results, finding that the patients in high risk subtype are similar with those of using SNF (jaccard coefficient = 0.7238). These results indicated that CPGD can obtain some similar subtyping results compared with SNF using all the gene expression data.

In addition, we divided the patients of four prior-known clinical subtypes (Basal, Her2, Lum A and Lum B) into high and low risk groups according to our predicted subtypes (Figure 12E). The results showed that there are significant differences of survival curves among two patient groups in some clinical subtypes (e.g. Her2 subtype, P-value = 0.0432, Figure 12E), indicating that our predicted subtypes can help further classify the prior-known clinical subtypes for more detailed patient risk assessments. In addition, the patients in high risk subtype of CPGD are similar with that of conventional Lum A (Figure 12F, jaccard coefficient = 0.5143).

Identification of the potential drug-repurpusing candidates for COVID-19

To identify the potential drug-repurpusing candidates for recently spreading COVID-19 caused by the SARS-COV2 virus, we also used CPGD to identify the drug combination candidates on SARS-COV2 dataset. The computational details of CPGD to identify the potential drug-repurpusing candidates for COVID-19 were shown as follow,

(i) **Collection of SARS-COV2 related datasets.** The gene expression profiling of patients with SARS was collected (25) which consists of 60 SARS disease samples and 10 normal samples. Furthermore we collected 332 SARS-COV2 related proteins (26). We should note that we here chose Network 1 which have a more complete gene interactions as the prior-known gene gene interaction network.

(ii) **Modified version of CPGD.** Since we lacked the gene mutation data of SARS-COV2, we did not consider the construction of gene co-mutation network. Furthermore, due to the lack of the gene expression data

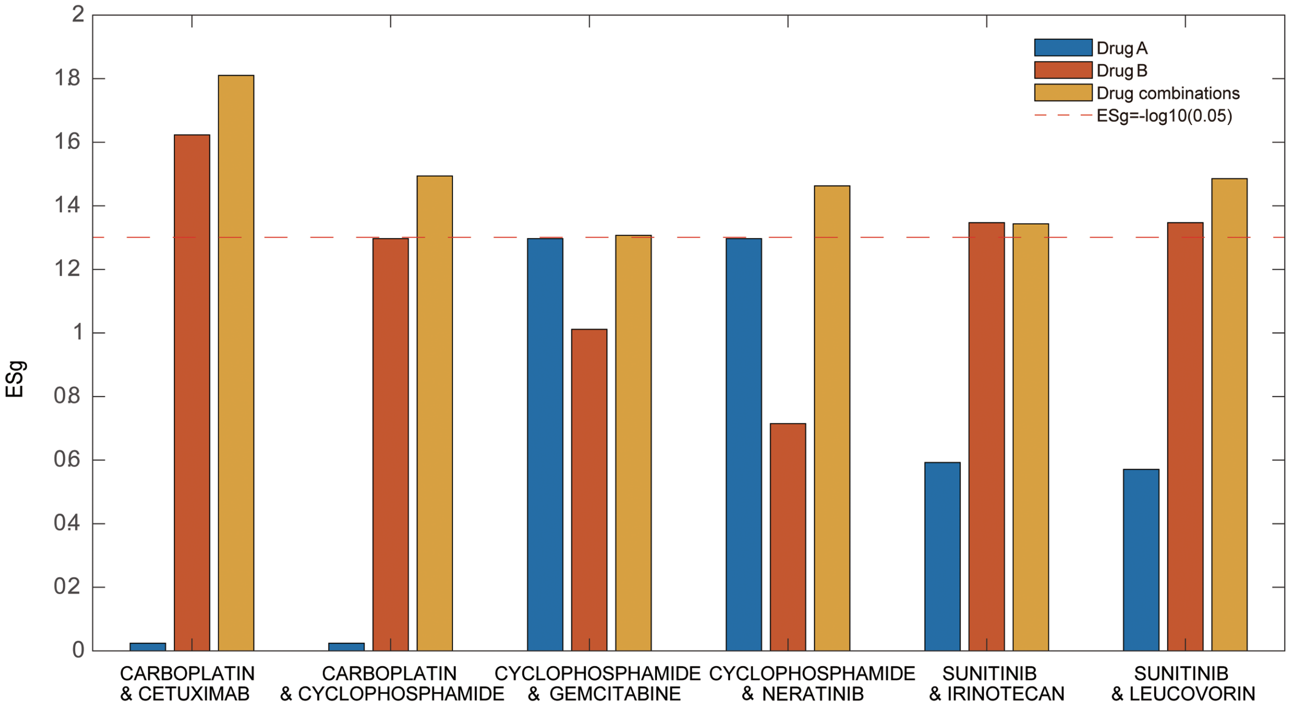

Figure 10. The *P*-value of combined drugs therapy and single drug therapy on BRCA cancer dataset. ESg = $-\log_{10}$(*P*-value).

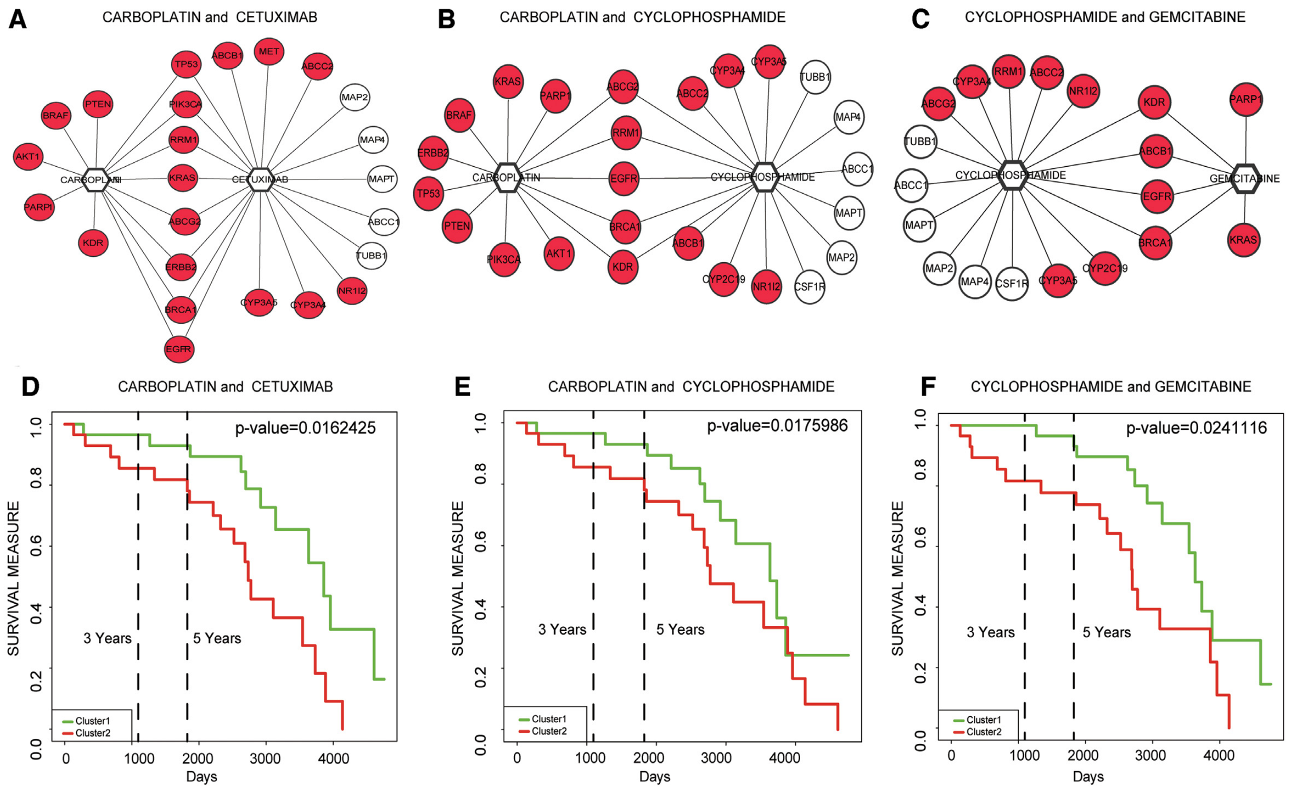

Figure 11. The risk assessment of three predicted anti-disease drug pairs on independent GSE5327-BRCA cancer dataset. (**A**–**C**) The interaction network between drug pairs and targeting driver genes for three predicted anti-disease drug pairs. The nodes with red color denote the breast-related genes and hexagon nodes denote the drugs. (**D**–**F**) Results of survival analysis of three predicted anti-disease drug pairs on independent dataset.

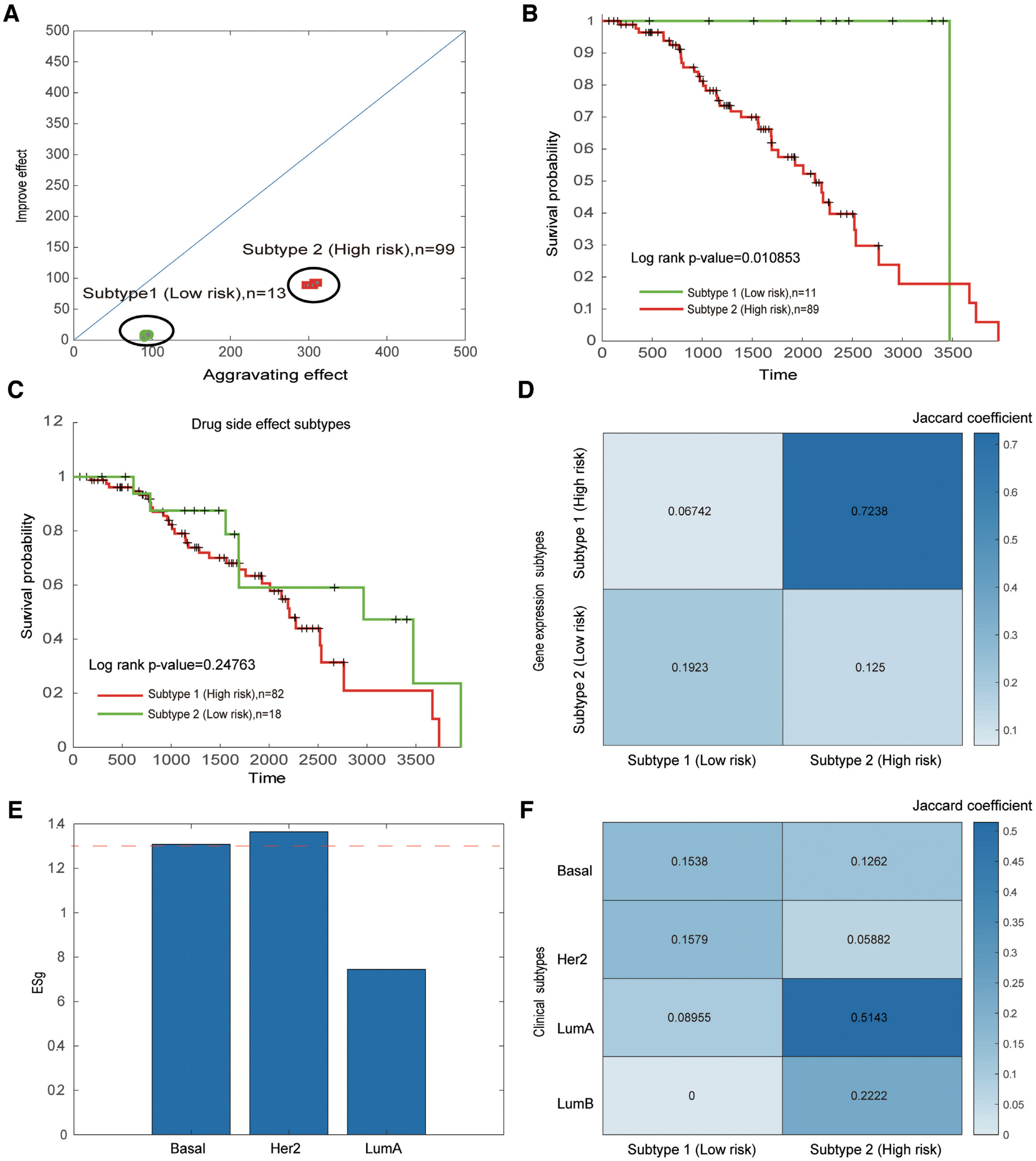

Figure 12. Cancer subtypes generated by quantifying side effect of drug pairs on the personalized driver genes. (**A**) The subtype classification of patients in the breast cancer dataset based on the number of drug pairs with an aggravating effect and an enhancing effect. (**B**) The survival analysis of two subtypes identified by CPGD. (**C**) The survival analysis of two subtypes identified by SNF using all the gene expression data. (**D**) The jaccard coefficient between subtypes identified by CPGD and SNF. (**E**) The *P-value* of survival curves among two groups of patients in four prior-known subtypes (Basal, Her2, Lum A and Lum B). (**F**) The jaccard coefficient between molecular subtypes of CPGD and four clinical subtypes (Basal, Her2, Lum A and Lum B).

of paired samples, paired-SSN was reduced to the original SSN. The scores of edges in PGIN was reduced as,

$$e_{ij}^{\text{Patient } k} = \left| \log_2 \left| \frac{\Delta \text{PCC}_{ij,k}^{\text{Disease}}}{\Delta \text{PCC}_{ij,k}^{\text{Normal}}} \right| \right|.$$

(iii) Identification of drug pairs on SARS-COV2 dataset. We used CPGD for obtaining the synergistic scores of pairwise drug combinations for each individual patient. We ranked drug pairs according to the mean synergistic scores of pairwise drug combinations among all patients.Since there exist multiple rankings for each drug combination by using different balance parameters of CPGD, we used condorce algorithm (71) to combine these multiple rankings into a single rank for drug combination candidates, which in turn determined drug combinations' prioity.

Supplementary file 9 listed 90 identified ranking drug combination candidates on SARS-COV2 dataset. We found that there are 15 drug pairs among 90 predicted drug combination candidates (16.67%) for which either of two drugs is reported as drug repurposing candidate in a recent report (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7280907/). Furthermore, among these 90 predicted drug combination candidates, both of two drugs for a pairwise drug combination, i.e. DEXAMETHASONE and THALIDOMIDE are reported as drug repurposing candidates and further currently being tested in clinical trials for COVID-19 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7280907/).

DISCUSSION

The genomic profiles of individual patients in complex disease (e.g. cancer patients) are diverse and heterogeneous, which is believed to be responsible for heterogeneous drug response. In the past years, many computational tools for the personalized driver gene identification have presented promising clues in determining personalized drug targets for drugs discovery of individual patients. However, it has not been studied how to fill gap between personalized driver gene identification and combinatorial drug discovery of individual patients. To this end, in this work a novel structural network controllability-based algorithm (CPGD) was developed to investigate the driver genes of individual cancer patients and discover drug combinations to target on multiple driver genes as potential combinatory therapies for personalized medicine. By exploring more precise mathematical models on high-throughput personalized multi-omics data, CPGD contains three advances in methodology. The first is that CPGD introduces a measure for scoring the edges of PGIN by integrating co-mutation scores on the somatic mutation data across cancer type-specific data with personalized co-expression scores on gene expression data of individual patients. The score of co-mutation edge indices the co-mutation probability of two mutated genes in individual tumors to promote tumorigenesis and anticancer drug responses. The score of the personalized co-expression edge represents the significant personalized co-expression difference in the human genetic interaction network between the normal sample and tumor sample of an individual patient. Therefore, the measure could more accurately represent the personalized state transition of an individual patient in caner development by combining the gene somatic mutations, personalized gene expression and network topology information in the prior-known human genetic interaction network. The second is that CPGD develops a novel network controllability-based algorithm of weight-NCUA for the driver node optimization by considering the edge weight information (i.e., network edge scores) of PGIN. Compared with the existing network controllability-based algorithms, weight-NCUA considers the edge weight information of the personalized gene regulatory network for the driver node optimization to avoid the existence of multiple minimum driver nodes configurations. The third is that CPGD designs the proper evaluation metrics from diverse biomedical aspects, such as (i) the prioritization of personalized combinatorial drugs for evaluating the ability of predicting clinical efficacious combinatorial drugs; (ii) the evaluation of synergistic effect of pairwise drug combinations by measuring the synergistic effect of their corresponding PDGMs; (iii) the exploration of risk assessment of paired combinatorial drugs and (iv) the enhancement of disease subtyping by side effect quantification on PDGs.

Using breast and lung cancer data as two gold-standard datasets, CPGD is better than other existing methods in terms of *discovery rate*. The drug pairs identified by CPGD can partition the patients into discriminative groups for effective risk assessment. Especially, CPGD can effectively recgonize the cancer subtypes of breast cancer by quantifying the side effect of combinatorial drugs onco-targeting PDGs for individual patients. We also implemented our CPGD on the SARS-COV2 dataset, finding that both of two drugs for a predicted pairwise drug combination candidates, i.e. DEXAMETHASONE and THALIDOMIDE are currently being tested in clinical trials for COVID-19 as reported in a recent research (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7280907/).

All the results suggest that the discovery of personalized combinatorial drugs in complex diseases could benefit from CPGD.However, our CPGD doesn't consider the dynamic change of drug concentration, which may lead to some false positive for the identification of efficacious combinatorial drugs. Some biological experimental validation and prospective clinical trials would be further conducted to verify the discovery of CPGD. And a more complete, systematic gene interaction network and drug-target network may further improve the performance of CPGD, supporting advanced combinatorial drug screen of individual patients.

DATA AVAILABILITY

CPGD was created using the MATLAB software. The implementation of our CPGD in MATLAB can be freely downloaded from https://github.com/NWPU-903PR/CPGD, where the gene expression data, combinatorial drug-gene interaction data and drug-gene interaction data with activation and inhibition interactions can also be freely downloaded.

The semantic similarity for generating GO similarity matrix was available in the folder 'Drug_Targets_GO_similarity_Data' of https://github.com/NWPU-903PR/CPGD. The drug function similarity in CMAP and drug chemical structure similarity was available in Supplementary file 3.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We appreciate Dr. Meng-Yuan Liu from Huazhong Agricultural University, China, for providing us the combinatorial drugs and gene interaction data, synthetic lethality genes interactions data. We thank Dr. Da-Chang Dang and Dr. Qing-Qing Zhang from Northwestern Polytechnical University, China, for giving us useful suggestions. We also thank all the reviewers for giving us valuable comments, which benefit for significant improvement of our manuscript.

FUNDING

National Natural Science Foundation of China [61873202, 62002329, 12026608, 31930022, 31771476, 81471047, 11871456]; National Key R&D Program [2016YFC0903400, 2017YFA0505500, 2017YFC1201200]; Strategic Priority Research Program of the Chinese Academy of Sciences [XDB38040400]; Shanghai Municipal Science and Technology Major Project [2017SHZDZX01]. Funding for open access charge: National Natural Science Foundation of China [61873202, 62002329, 12026608, 31930022, 31771476, 81471047, 11871456]; National Key R&D Program [2016YFC0903400, 2017YFA0505500, 2017YFC1201200]; Strategic Priority Research Program of the Chinese Academy of Sciences [XDB38040400]; Shanghai Municipal Science and Technology Major Project [2017SHZDZX01].

Conflict of interest statement. None declared.

REFERENCES

1. Zheng,W., Sun,W. and Simeonov,A. (2018) Drug repurposing screens and synergistic drug-combinations for infectious diseases. *Br. J. Pharmacol.*, **175**, 181–191.
2. Schork,N.J. (2015) Personalized medicine: time for one-person trials. *Nature*, **520**, 609–611.
3. van der Wijst,M.G.P., de Vries,D.H., Brugge,H., Westra,H.J. and Franke,L. (2018) An integrative approach for building personalized gene regulatory networks for precision medicine. *Genome medicine*, **10**, 96.
4. Preuer,K., Lewis,R.P.I., Hochreiter,S., Bender,A., Bulusu,K.C. and Klambauer,G. (2018) DeepSynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics*, **34**, 1538–1546.
5. Karimi,M., Hasanzadeh,A. and Shen,Y. (2020) Network-principled deep generative models for designing drug combinations as graph sets. *Bioinformatics*, **36**, i445–i454.
6. Deng,Y., Xu,X., Qiu,Y., Xia,J., Zhang,W. and Liu,S. (2020) A multimodal deep learning framework for predicting drug-drug interaction events. *Bioinformatics*, **36**, 4316–4322.
7. Talevi,A. (2015) Multi-target pharmacology: possibilities and limitations of the "skeleton key approach" from a medicinal chemist perspective. *Front. Pharmacol.*, **6**, 205.
8. Wang,L., Yu,X., Zhang,C. and Zeng,T. (2018) Detecting personalized determinants during drug treatment from omics big data. *Curr. Pharm. Des.*, **24**, 3727–3738.
9. Zeng,T., Zhang,W., Yu,X., Liu,X., Li,M. and Chen,L. (2016) Big-data-based edge biomarkers: study on dynamical drug sensitivity and resistance in individuals. *Brief. Bioinform.*, **17**, 576–592.
10. Huang,L., Li,F., Sheng,J., Xia,X., Ma,J., Zhan,M. and Wong,S.T. (2014) DrugComboRanker: drug combination discovery based on target network analysis. *Bioinformatics*, **30**, i228–236.
11. Huang,L., Brunell,D., Stephan,C., Mancuso,J., Yu,X., He,B., Thompson,T.C., Zinner,R., Kim,J., Davies,P. *et al.* (2019) Driver network as a biomarker: systematic integration and network modeling of multi-omics data to derive driver signaling pathways for drug combination prediction. *Bioinformatics*, **35**, 3709–3717.
12. Zhou,Y., Hou,Y., Shen,J., Huang,Y., Martin,W. and Cheng,F. (2020) Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov.*, **6**, 14.
13. Cheng,F., Kovács,I.A. and Barabási,A.L. (2019) Network-based prediction of drug combinations. *Nat. Commun.*, **10**, 1197.
14. Han,Y., Yang,J., Qian,X., Cheng,W.C., Liu,S.H., Hua,X., Zhou,L., Yang,Y., Wu,Q., Liu,P. *et al.* (2019) DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies. *Nucleic Acids Res.*, **47**, e45.
15. Bashashati,A., Haffari,G., Ding,J., Ha,G., Lui,K., Rosner,J., Huntsman,D.G., Caldas,C., Aparicio,S.A. and Shah,S.P. (2012) DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.*, **13**, R124.
16. Lawrence,M.S., Stojanov,P., Polak,P., Kryukov,G.V., Cibulskis,K., Sivachenko,A., Carter,S.L., Stewart,C., Mermel,C.H., Roberts,S.A. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.
17. Gonzalez-Perez,A. and Lopez-Bigas,N. (2012) Functional impact bias reveals cancer drivers. *Nucleic Acids Res.*, **40**, e169.
18. Guo,W.F., Zhang,S.W., Liu,L.L., Liu,F., Shi,Q.Q., Zhang,L., Tang,Y., Zeng,T. and Chen,L. (2018) Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. *Bioinformatics*, **34**, 1893–1903.
19. Hou,J.P. and Ma,J. (2014) DawnRank: discovering personalized driver genes in cancer. *Genome Med.*, **6**, 56.
20. Liu,Y.-Y., Slotine,J.-J. and Barabási,A.-L. (2011) Controllability of complex networks. *Nature*, **473**, 167–173.
21. Nacher,J.C. and Akutsu,T. (2012) Dominating scale-free networks with variable scaling exponent: heterogeneous networks are not difficult to control. *New J. Phys.*, **14**, 073005.
22. Mochizuki,A., Fiedler,B., Kurosawa,G. and Saito,D. (2013) Dynamics and control at feedback vertex sets. II: a faithful monitor to determine the diversity of molecular activities in regulatory networks. *J. Theor. Biol.*, **335**, 130–146.
23. Guo,W.F., Zhang,S.W., Zeng,T., Akutsu,T. and Chen,L. (2020) Network control principles for identifying personalized driver genes in cancer. *Brief. Bioinform.*, **21**, 1641–1662.
24. Guo,W.F., Zhang,S.W., Zeng,T., Li,Y., Gao,J. and Chen,L. (2019) A novel network control model for identifying personalized driver genes in cancer. *PLoS Comput. Biol.*, **15**, e1007520.
25. Cameron,M.J., Ran,L., Xu,L., Danesh,A., Bermejo-Martin,J.F., Cameron,C.M., Muller,M.P., Gold,W.L., Richardson,S.E., Poutanen,S.M. *et al.* (2007) Interferon-mediated immunopathological events are associated with atypical innate and adaptive immune responses in patients with severe acute respiratory syndrome. *J. Virol.*, **81**, 8692–8706.
26. Gordon,D.E., Jang,G.M., Bouhaddou,M., Xu,J., Obernier,K., White,K.M., O'Meara,M.J., Rezelj,V.V., Guo,J.Z., Swaney,D.L. *et al.* (2020) A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, **583**, 459–468.
27. Quan,Y., Liu,M.Y., Liu,Y.M., Zhu,L.D., Wu,Y.S., Luo,Z.H., Zhang,X.Z., Xu,S.Z., Yang,Q.Y. and Zhang,H.Y. (2018) Facilitating Anti-Cancer combinatorial drug discovery by targeting epistatic disease genes. *Molecules*, **23**, 736.
28. Liu,Y., Wei,Q., Yu,G., Gai,W., Li,Y. and Chen,X. (2014) DCDB 2.0: a major update of the drug combination database. *Database*, **2014**, bau124.
29. Wagner,A.H., Coffman,A.C., Ainscough,B.J., Spies,N.C., Skidmore,Z.L., Campbell,K.M., Krysiak,K., Pan,D., McMichael,J.F., Eldred,J.M. *et al.* (2016) DGIdb 2.0: mining clinically relevant drug-gene interactions. *Nucleic Acids Res.*, **44**, D1036–D1044.
30. Wishart,D.S., Feunang,Y.D., Guo,A.C., Lo,E.J., Marcu,A., Grant,J.R., Sajed,T., Johnson,D., Li,C., Sayeeda,Z. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.

31. Yang,H., Qin,C., Li,Y.H., Tao,L., Zhou,J., Yu,C.Y., Xu,F., Chen,Z., Zhu,F. and Chen,Y.Z. (2016) Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res.*, **44**, D1069–D1074.
32. Torres,N.B. and Altafini,C. (2016) Drug combinatorics and side effect estimation on the signed human drug-target network. *BMC Syst. Biol.*, **10**, 74.
33. Liu,X., Wang,Y., Ji,H., Aihara,K. and Chen,L. (2016) Personalized characterization of diseases using sample-specific networks. *Nucleic Acids Res.*, **44**, e164.
34. Liu,C., Zhao,J., Lu,W., Dai,Y., Hockings,J., Zhou,Y., Nussinov,R., Eng,C. and Cheng,F. (2020) Individualized genetic network analysis reveals new therapeutic vulnerabilities in 6,700 cancer genomes. *PLoS Comput. Biol.*, **16**, e1007701.
35. Bazin,J.C., Li,H., Kweon,I.S., Demonceaux,C., Vasseur,P. and Ikeuchi,K. (2013) A branch-and-bound approach to correspondence and grouping problems. *IEEE Trans. Pattern Anal. Mach. Intell.*, **35**, 1565–1576.
36. Yu,K., Liang,J.J., Qu,B.Y., Cheng,Z. and Wang,H. (2018) Multiple learning backtracking search algorithm for estimating parameters of photovoltaic models. *Appl. Energy*, **226**, 408–422.
37. Yu,K., Qu,B., Yue,C., Ge,S., Chen,X. and Liang,J. (2019) A performance-guided JAYA algorithm for parameters identification of photovoltaic cell and module. *Appl. Energy*, **237**, 241–257.
38. Yu,K., Liang,J.J., Qu,B.Y., Chen,X. and Wang,H. (2017) Parameters identification of photovoltaic models using an improved JAYA optimization algorithm. *Energy Convers. Manage.*, **150**, 742–753.
39. Lamb,J., Crawford,E.D., Peck,D., Modell,J.W., Blat,I.C., Wrobel,M.J., Lerner,J., Brunet,J.P., Subramanian,A., Ross,K.N. *et al.* (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.
40. Martin,Y.C., Kofron,J.L. and Traphagen,L.M. (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, **45**, 4350–4358.
41. The Gene Ontology Consortium. (2017) Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.*, **45**, D331–D338.
42. Peng,J., Uygun,S., Kim,T., Wang,Y., Rhee,S.Y. and Chen,J. (2015) Measuring semantic similarities by combining gene ontology annotations and gene co-function networks. *BMC Bioinformatics*, **16**, 44.
43. Iorio,F., Bosotti,R., Scacheri,E., Belcastro,V., Mithbaokar,P., Ferriero,R., Murino,L., Tagliaferri,R., Brunetti-Pierri,N., Isacchi,A. *et al.* (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 14621–14626.
44. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
45. Rogers,D. and Hahn,M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, **50**, 742–754.
46. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
47. Repana,D., Nulsen,J., Dressler,L., Bortolomeazzi,M., Venkata,S.K., Tourna,A., Yakovleva,A., Palmieri,T. and Ciccarelli,F.D. (2019) The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol.*, **20**, 1.
48. Rivals,I., Personnaz,L., Taing,L. and Potier,M.C. (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.
49. Gao,J., Barzel,B. and Barabási,A.L. (2016) Universal resilience patterns in complex networks. *Nature*, **536**, 238.
50. Wang,B., Mezlini,A.M., Demir,F., Fiume,M., Tu,Z., Brudno,M., Haibe-Kains,B. and Goldenberg,A. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.
51. Ciriello,G., Cerami,E., Sander,C. and Schultz,N. (2012) Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.*, **22**, 398–406.
52. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
53. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–D679.
54. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
55. Vinayagam,A., Stelzl,U., Foulle,R., Plassmann,S., Zenkner,M., Timm,J., Assmus,H.E., Andrade-Navarro,M.A. and Wanker,E.E. (2011) A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal*, **4**, rs8.
56. Hu,Y., Chen,C.H., Ding,Y.Y., Wen,X., Wang,B., Gao,L. and Tan,K. (2019) Optimal control nodes in disease-perturbed networks as targets for combination therapy. *Nat. Commun.*, **10**, 2180.
57. Reimand,J. and Bader,G.D. (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.*, **9**, 637.
58. Morrison,J.L., Breitling,R., Higham,D.J. and Gilbert,D.R. (2005) GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics*, **6**, 233.
59. Zhou,X. and Liu,J. (2014) Inferring gene dependency network specific to phenotypic alteration based on gene expression data and clinical information of breast cancer. *PLoS One*, **9**, e92023.
60. Leiserson,M.D., Vandin,F., Wu,H.T., Dobson,J.R., Eldridge,J.V., Thomas,J.L., Papoutsaki,A., Kim,Y., Niu,B., McLellan,M. *et al.* (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.*, **47**, 106–114.
61. Aronson,A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.*, 17–21.
62. Ghiassian,S.D., Menche,J. and Barabási,A.L. (2015) A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.*, **11**, e1004120.
63. Dai,H., Li,L., Zeng,T. and Chen,L. (2019) Cell-specific network constructed by single-cell RNA sequencing data. *Nucleic Acids Res.*, **47**, e62.
64. Yu,X., Zeng,T., Wang,X., Li,G. and Chen,L. (2015) Unravelling personalized dysfunctional gene network of complex diseases based on differential network model. *J. Transl. Med.*, **13**, 189.
65. Zhang,W., Zeng,T., Liu,X. and Chen,L. (2015) Diagnosing phenotypes of single-sample individuals by edge biomarkers. *J. Mol. Cell Biol.*, **7**, 231–241.
66. Kuijjer,M.L., Tung,M.G., Yuan,G., Quackenbush,J. and Glass,K. (2019) Estimating sample-specific regulatory networks. *iScience*, **14**, 226–240.
67. Pham,V.V.H., Liu,L., Bracken,C.P., Goodall,G.J., Long,Q., Li,J. and Le,T.D. (2019) CBNA: a control theory based method for identifying coding and non-coding cancer drivers. *PLoS Comput. Biol.*, **15**, e1007538.
68. Vinayagam,A., Gibson,T.E., Lee,H.J., Yilmazel,B., Roesel,C., Hu,Y., Kwon,Y., Sharma,A., Liu,Y.Y., Perrimon,N. *et al.* (2016) Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 4976–4981.
69. Aguirre-Gamboa,R., Gomez-Rueda,H., Martínez-Ledesma,E., Martínez-Torteya,A., Chacolla-Huaringa,R., Rodriguez-Barrientos,A., Tamez-Peña,J.G. and Treviño,V. (2013) SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS One*, **8**, e74250.
70. Goswami,C.P. and Nakshatri,H. (2014) PROGgeneV2: enhancements on the existing database. *BMC Cancer*, **14**, 970.
71. McLean,I. (1990) The borda and condorcet principles: three medieval applications. *Social Choice Welfare*, **7**, 99–108.