RESEARCH

Open Access

Exploring gene-patient association to identify personalized cancer driver genes by linear neighborhood propagation

Yiran Huang[1,2,3], Fuhao Chen[1], Hongtao Sun[1] and Cheng Zhong[1,2,3*]

*Correspondence:
chzhong@gxu.edu.cn

[1] School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China
[2] Key Laboratory of Parallel, Distributed and Intelligent Computing in Guangxi Universities and Colleges, Guangxi University, Nanning 530004, China
[3] Guangxi Key Laboratory of Multimedia Communications and Network Technology, Guangxi University, Nanning 530004, China

Abstract

Background: Driver genes play a vital role in the development of cancer. Identifying driver genes is critical for diagnosing and understanding cancer. However, challenges remain in identifying personalized driver genes due to tumor heterogeneity of cancer. Although many computational methods have been developed to solve this problem, few efforts have been undertaken to explore gene-patient associations to identify personalized driver genes.

Results: Here we propose a method called LPDriver to identify personalized cancer driver genes by employing linear neighborhood propagation model on individual genetic data. LPDriver builds personalized gene network based on the genetic data of individual patients, extracts the gene-patient associations from the bipartite graph of the personalized gene network and utilizes a linear neighborhood propagation model to mine gene-patient associations to detect personalized driver genes. The experimental results demonstrate that as compared to the existing methods, our method shows competitive performance and can predict cancer driver genes in a more accurate way. Furthermore, these results also show that besides revealing novel driver genes that have been reported to be related with cancer, LPDriver is also able to identify personalized cancer driver genes for individual patients by their network characteristics even if the mutation data of genes are hidden.

Conclusions: LPDriver can provide an effective approach to predict personalized cancer driver genes, which could promote the diagnosis and treatment of cancer. The source code and data are freely available at https://github.com/hyr0771/LPDriver.

Keywords: Personalized cancer driver gene, Gene-patient association, Linear neighborhood propagation, Gene interaction network

Background

Cancer is a heterogeneous disease driven by genetic alterations [1]. Identifying the cancer driver genes with alterations plays a crucial role in the treatment and diagnosis of cancer [2–5]. A number of computational methods have been proposed to identify cancer driver genes in recent years [2]. Most of these methods concentrate on identifying driver genes in specific types or subtypes of cancer [3]. Based on the rationale, these computational methods can mainly be grouped in mutation-based methods and the network-based methods. Mutation-based methods [6–10] employ the characteristics of gene mutations to identify cancer driver genes while the network-based methods [11–14] utilize gene interaction networks to assess the role of genes to predict cancer driver genes.

Several mutation-based methods have been proposed to identify cancer driver genes, each with its unique approach and hypothesis [15]. MutSigCV [7] evaluates the gene mutation frequencies that exceed what is expected to identify potential cancer driver genes. However, it may incorrectly identify the genes with frequent mutations that are non-contributory to cancer development as potential cancer driver genes. Unlike MutSigCV, OncodriveFM [8] hypothesizes that the genes with significantly functional impact are more likely to be candidate driver genes. OncodriveFM evaluates the bias of gene mutations with functional impacts rather than the sheer mutation count, enabling the detection of driver genes with low recurrence but significant roles in cancer development. OncodriveFML [6], similar to OncodriveFM, employs functional impact assessment but extends its scope to both coding and non-coding mutations. DriverML [9] adopts a different approach by considering how mutation types affect the functional impact of mutations. It uses a supervised machine learning approach with pan-cancer training data to optimize weight parameters for mutation types, scoring functional influences of gene alterations. ActiveDriver [10], on the other hand, identifies cancer drivers based on the structural consequences of gene mutations, particularly focusing on the enrichment in post-translational modification sites like phosphorylation, acetylation, and ubiquitination sites. Although these mutation-based methods offer valuable insights into the identification of cancer driver genes, they face the problems of the incomplete gene mutation databases caused by cancer heterogeneity. This may limit their ability to comprehensively identify driver genes.

On the other hand, the network-based methods identify cancer driver genes by incorporating the information of pathways, gene–gene or protein–protein interactions to measure the gene roles in the biological networks [15–17]. For instance, CBNA [11] combines the network controllability analysis with mutation data, allowing it to pinpoint both coding and miRNA driver genes within gene networks. Meanwhile CBNA can also be employed to uncover the drivers specific to particular types or subtypes of cancer. In contrast to CBNA, DriverNet [12] integrates various multi-omics data such as gene expression data and biological pathway information to construct bipartite gene network, and utilizes greedy optimization search to identifies driver genes with high outlying expression in the bipartite gene network. Similarly, Subdyquency [13] integrates mutated genes' variation frequency and its interactions with dysregulated genes in a certain compartment to build bipartite graph, then employs random walk method on the built graph to produce walking score for each mutated gene of patient to pinpoint candidate driver genes. Different from these methods, MEMo [14] (Mutual Exclusivity Modules) takes a module-centric approach, using mutual exclusivity techniques in biological networks to discover oncogenic network modules. MEMo suggests that genomic alterations in the same cancer type tend to occur within a limited number of pathways and are unlikely to coexist within the same patient. These abovementioned methods provide diverse strategies for uncovering cancer driver genes, contributing valuable insights into the molecular mechanisms of cancer.

Nevertheless, all these methods identify cancer driver genes at the population level. Due to tumor heterogeneity in cancer, different patients may have different genetic alterations and their tumors may be caused by different genes, and two patients who suffer from the same kind of cancer and receive the same remedy may have different prognosis [2, 3]. Thus, it is necessary to identify cancer drivers specific to an individual patient for personalized diagnosis and treatment [2, 3]. DawnRank [18] for the first time utilizes a ranking framework to assess the connectivity and the amount of differential expression genes in gene interaction network. By combining gene ranks with somatic alteration data, such as copy variation number, DawnRank effectively detects individual driver alterations. However, it relies on the same gene regulatory network for all patients, potentially missing patient-specific regulatory information. To address this limitation, SCS [19] constructs a personalized gene regulatory network for each patient using the gene expression data of patients and normal people. SCS identifies personalized cancer driver genes as the minimal set of the most differentially expressed genes in the constructed network. Further, PRODIGY [20] adopts Steiner tree model to evaluate the impact of the genes with mutations on the deregulated pathways to identify personalized cancer driver genes. Later, PersonaDrive [21] tries to construct a personalized bipartite graph that links mutated genes to differentially expressed genes for each patient, and calculates the edge weights of the graph based on the overlap between the mutated gene and the differentially expressed gene pair in biological pathways. Subsequently, it ranks the the potential driver genes based on their influence scores evaluated by the edge weights in the bipartite graph. Similarly, BetweenNET [22] combines patient genomic data with protein–protein interaction network to build customized gene interaction network and identifies personalized cancer driver genes in the customized network. Meanwhile, based on the structural controllability theory. Cheng et al. [23] proposed a weighted minimum dominating set network model WMDS.net to find the key regulators of gene co-expression networks to determine cancer driver genes. Distinguishing from these methods focusing on coding driver genes, Pham et al. [2] shifts the focus to the comprehensive exploration of coding and non-coding cancer drivers. They introduced a network-based approach named pDriver to identify personalized coding and miRNA cancer drivers. Recently, Guo et al. [24] proposed a structure-based network control method called PNC for identifying personalized cancer driver genes based on the network control method NCUA. In order to verify the effectiveness of NCUA, Guo et al. replaced NCUA with the state-of-the-art structure-based network control methods MMS [25] and MDS [25] in PNC respectively and compared their performance in identifying personalized cancer driver genes. The experimental comparison results showed that as compared to MMS and MDS, NCUA is more effective for PNC in identifying personalized cancer driver genes.

In this paper, we propose a novel method, called LPDriver, to identify personalized cancer driver genes. In comparison to existing methods for finding personalized cancer driver genes, LPDriver mainly includes the following advantages:

i. LPDriver attempts for the first time to explore the gene-patient associations extracted from personalized gene network to mine functionally similar genes among patients for identifying personalized driver genes. Nevertheless, LPDriver does not need to bind gene-patient associations with the known driver genes and the mutation genes to detect driver genes, which will promote and facilitate identification of cancer driver genes.
ii. Distinguishing from existing network-based methods, LPDriver finds the genes in the maximum matching set of the personalized gene network, which maximumly cover most but not all edges of the gene network, as potential driver genes. This could further extend exploration of driver genes.

To take advantages of the gene interaction network specific to a patient of a specific cancer, we first construct the personalized gene interaction network for each patient based on the tumor gene expression data of the patient. Next, we build the personalized gene network for the given patient upon the difference between the gene network built on all patients of the specific cancer dataset (e.g. the cancer dataset of The Cancer Genome Atlas) and the gene network built on all patients excluding the patient under consideration. Then we extract potential driver genes by finding the maximum matchings of the bipartite graph of the personalized gene network to build gene-patient associations. Finally, we utilize a linear neighborhood propagation model to mine the linear neighborhood similarity of genes among patients to infer the personalized driver genes from the built gene-patient associations.

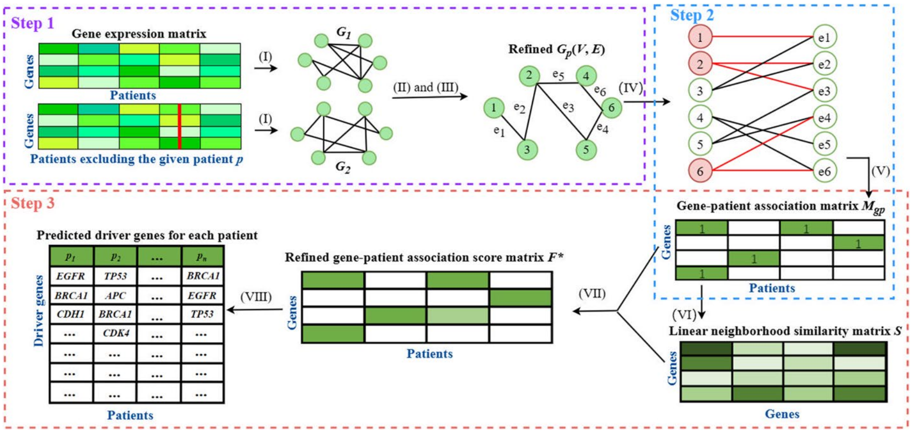
We applied LPDriver on multiple cancer datasets of The Cancer Genome Atlas (TCGA) [26] and validated the results by considering the cancer driver genes in the Network of Cancer Genes (NCG) [27] and Cancer Genes Census (CGC) [28] as the benchmark. The experimental comparison results demonstrate that LPDriver is more effective than the existing methods in detecting cancer driver genes. Moreover, the experimental results also show that LPDriver not only can reveal personalized cancer driver genes for individual patient, but also detect some potentially novel driver genes that have been documented to be related to cancer. Generally, LPDriver is an effective and applicable complement to the existing methods for identifying cancer driver genes.

Materials and methods

LPDriver identifies personalized cancer driver genes by three main steps: (1) Constructing personalized gene interaction network (PGIN) using the gene expression data of tumor samples of patients. (2) Identifying potential gene-patient associations by finding the maximum bipartite matchings from the bipartite graph of PGIN. (3) Predicting personalized driver genes from the gene-patient associations through linear neighborhood propagation. The overview of our method is summarized in Fig. 1.

Constructing personalized gene interaction network

Intrigued by LIONESS [29], for a given patient p and a group of patients as reference, such as the patients of a specific cancer data set in TCGA, we construct a personalized gene interaction network for the given patient p based on the statistical



p_1	p_2	...	p_n
EGFR	*TP53*	...	*BRCA1*
BRCA1	*APC*	...	*EGFR*
CDH1	*BRCA1*	...	*TP53*
...	*CDK4*	...	...
...	...	...	...
...	...	...	...
...	...	...	...

Fig. 1 Illustration of LPDriver. **I** Constructing the gene interaction network G_1 and G_2. **II** Removing the edges that exist in both G_1 and G_2. **III** Removing the edges(interactions) that are not supported by the known gene interaction network. **IV** Transform $G_p(V,E)$ into bipartite graph and identify driver genes of patient *p* in bipartite graph. **V** Identify the potential driver genes of all patients to construct the matrix M_{gp}. **VI** Compute the linear neighborhood similarity of genes among patients in M_{gp}. **VII** Adopt label propagation method based on matrix *S* to refine M_{gp}. **VIII** Identify highly rank genes as the personalized driver genes from *F**

difference between the gene network built on all patients and the gene network built on all patients except the given patient *p*.

Pearson Correlation Coefficient (PCC) has been widely adopted to assess the correlations between the patients' gene expression profiles to identify personalized driver genes. Following literature [19, 24], we first use all patients' tumor gene expression data to compute the Pearson Correlation Coefficient (PCC) between the patients' genes to evaluate the correlations between the patients' gene expression profiles. Briefly, we first identified a group of tumor samples for the studied cancer type. The PCC of each pair of genes was calculated according to the expression data of the patient *p* to construct the gene interaction network G_1. In the similar way, we use the tumor gene expression data of all patients excluding the patient *p* to compute the PCC between the patients' genes and construct another gene interaction network G_2. Then, all the edges with significantly differential correlations [23, 24] (i.e. p value < 0.05) were retained and used to construct the personalized gene interaction network for that cancer patient.

Finally, we remove the edges that exist in both G_1 and G_2, and remain the edges that only exist in G_1 or else G_2, and use the remaining edges of G_1 and G_2 to construct the personalized gene interaction network $G_p(V, E)$ for the patient *p*, where *V* and *E* are the node set and edge set of $G_p(V, E)$ respectively.

Note that, in the construction of $G_p(V, E)$ for the patient *p* of a cancer, remaining the edges, which only exist in G_1 or else G_2, between genes *i* and *j* is based on the observation: G_1 is produced with the patient *p* and G_2 is produced without the patient *p*, and the edge between genes *i* and *j* is only included in G_1 or else G_2, which implies the presence of this patient alters the association between genes *i* and *j* in $G_p(V, E)$, and therefore the interactions between genes *i* and *j* of this patient could have a relatively high correlation with this cancer. We thus remain these edges in $G_p(V, E)$.

On the other side, removing the edges, which exist in both of G_1 and G_2, between genes *i* and *j* is based on the observation: both of G_1 and G_2 include the edges between genes i and j, which implies the presence of this patient does not affect the association of genes i and j in $G_p(V, E)$, and therefore the interactions between genes i and j of this patient could have a relatively low correlation with this cancer. We thus remove these edges in $G_p(V, E)$.

In order to obtain accurate and reliable regulatory mechanism of personalized gene interaction network for each patient, based on the known gene interaction network retrieved from the existing gene interaction databases, we refine the personalized gene interaction network $G_p(V, E)$ by removing the edges(interactions) that are not supported by the known gene interaction network [30, 31]. The constructed personalized networks for different types of cancer are specific significantly as the constructed network are built upon the individual patients of a specific cancer dataset in TCGA. Meanwhile, the known gene interaction network data only serves to refine the constructed personalized interaction network. Note that the specific known gene interaction network can be specified by the users and the known gene interaction network ConsensusPathDB [30, 31] used in this work can be found at Additional file 1.

Identifying potential gene-patient associations

After obtaining the personalized gene interaction network $G_p(V, E)$ for the patient p, we now try to determine potential gene-patient associations for the patient p based on $G_p(V, E)$. Note that, in $G_p(V, E)$, the interactions(edges) between genes i and j are possibly associated with current tumor for the patient p. A set of genes that are capable of interacting (connecting) with most genes of G_p could play central or driving role in controlling the gene interaction network and could most likely be potential driver genes for the patient p [2]. An intuitive way for discovering such potential genes is to find the genes that are able to maximumly cover the nodes of $G_p(V, E)$ [2, 24].

Based on this intuition, we adopt the following steps to identify potential driver genes for each patient in $G_p(V, E)$: (1) building a bipartite graph from the personalized gene interaction network $G_p(V, E)$, where the nodes of the left side are the nodes of $G_p(V, E)$ and the nodes of the right side are the edges of $G_p(V, E)$, and (2) determining the maximum matching set of the left side nodes to cover the right side nodes in the bipartite graph by using the well-known bipartite graph matching Hungarian algorithm [24, 32].

Specifically, for the personalized gene interaction network $G_p(V, E)$, we first transform $G_p(V, E)$ into a bipartite graph $G(L, R, E_B, W)$, where $L = V$, $R = E$, E_B is the edge set of the bipartite graph $G(L, R, E_B, W)$ and $W_{v,u} \in W$ is the weight of the edge $(v, u) \in E_B$, $v \in L$, $u \in R$. After building the bipartite graph $G(L, R, E_B, W)$, we try to find a maximum matching set M from L, which could maximumly cover the nodes of R, and choose the genes in the maximum matching set M as the potential driver genes for the patient p.

Next, we apply the well-known bipartite graph matching algorithm "Hungarian algorithm" [32] to find the maximum weighted bipartite matching in $G(L, R, E_B, W)$. In fact, the maximum weighted bipartite matching in $G(L, R, E_B, W)$ is an edge set $M \subseteq E_B$ such that the sum of the edge weights $\sum_{(v,u)\in M} W_{v,u}$ is maximum and the nodes of each edge in M are different [33]. Following Hungarian algorithm, we can obtain the maximum matching set M from L by solving the following linear programming relaxation [32]:

$$\begin{aligned} & max_{x_{v,u}} \sum_{(v,u)\in E_B} x_{v,u} W_{v,u} \\ & s.t \sum_{u\in R:(v,u)\in E_B} x_{v,u} = 1, \quad \forall v \in L \\ & \qquad\qquad x_{v,u} \geq 0 \end{aligned} \tag{1}$$

where the edge weight $W_{v,u}$ of the bipartite graph is set to 1, $x_{v,u}$ is an indicative variable, $(v, u)\in E_B$ is the edge between $v \in L$ and $u \in R$ in the bipartite graph $G(L, R, E_B, W)$. The solution of formula (1) is the maximum bipartite matching set M in $G(L, R, E_B, W)$.

Since the nodes(genes) in the maximum matching set M can maximumly cover the right side nodes of the bipartite graph $G(L, R, E_B, W)$ [33], these nodes(genes) could associate with most of the genes in $G_p(V, E)$ and play central or driving role in controlling the gene interaction network. Finally, we can choose the genes in the maximum matching set M as the potential driver genes for the patient p.

In the following, the potential driver genes of all patients in the tumor reference samples will be repeatedly produced in the similar way and the produced driver genes of all patients are used to construct the gene-patient association matrix M_{gp}, where the rows are genes and the columns are patients in M_{gp}. In matrix M_{gp}, if gene i is a potential driver gene for patient j then $M_{gp}(i,j) = 1$, otherwise $M_{gp}(i,j) = 0$.

Predicting driver genes

Previous studies indicated that the data point associations in matrix could be reconstructed and refined by using linear neighborhood similarity [34–36]. Intrigued by this, we utilize a linear neighborhood propagation model to refine the gene-patient associations in M_{gp} to detect personalized driver genes. In this model, we compute the linear neighborhood similarities of genes among patients in M_{gp}, use a label propagation method based on the linear neighborhood similarities of genes to infer the unobserved gene-patient associations to refine M_{gp}, and identify the personalized driver genes from the refined gene-patient associations.

Specifically, assume that there are n genes and m patients in M_{gp}, we denote these n genes as feature vectors y_i, $i = 1, 2, \ldots, n$ and consider these genes as data objects. This optimization problem can be formulated as the objective function:

$$\begin{aligned} & min_{w_{ii_j}} \left\| y_i - \sum_{i_j : y_{i_j} \in N(y_i)} w_{ii_j} y_{i_j} \right\|^2 = min_{w_{ii_j}} \sum_{i_j, i_k : y_{i_j}, y_{i_k} \in N(y_i)} w_{ii_j} G^i_{i_j, i_k} w_{ii_k} = w_i^T G^i w_i \\ & \text{s.t. } \sum_{i_j} w_{ii_j} = 1, \; w_{ii_j} \geq 0. \end{aligned} \tag{2}$$

where $N(y_i)$ is a neighbor set of y_i with k ($k = 1, 2, \ldots, n - 1$) nearest neighbors, y_{ij} denotes the jth neighbor of y_i and w_{iij} represents the contribution of y_{ij} to reconstruct y_i, and w_{iij} can be regarded as the linear neighborhood similarity of y_{ij} and y_i, $w_i = (w_{ii1}, w_{ii2}, \ldots, w_{iik})^T$. $G^i_{ij,ik} = (y_i - y_{ij})^T (y_i - y_{ik})$ is the jth row and kth column of Gram matrix G^i [34–36].

In order to minimize the norm of reconstructive weight w_i, the Tikhonov regularization term was added to prevent overfitting [34–36]. Then, we can rewrite the objective function as:

$$\begin{aligned} & min_{w_{ii_j}} \left\| y_i - \sum_{i_j : y_{i_j} \in N(y_i)} w_{ii_j} y_{i_j} \right\|^2 = w_i^T G^i w_i + \gamma w_i^2 = w_i^T (G^i + \gamma I) w_i \\ & \text{s.t. } \sum_{i_j} w_{ii_j} = 1, \ w_{ii_j} \geq 0. \end{aligned} \tag{3}$$

where γ is the regularization coefficient and is set to 1 for simplicity and $I \in R^{n \times n}$ is identity matrix.

We first use standard quadratic programming [37] to solve (3), and the solutions $w_i = (w_{ii1}, w_{ii2}, \ldots, w_{iik})^T$ are the reconstructive weights of the data point y_i, $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, n - 1$. Then we use these solutions to construct a weight matrix $S \in R^{n \times n}$ and each entry of S can be regarded as the linear neighborhood similarity of genes. Based on the weight matrix S, we construct an undirected graph G_{dg} where the nodes are genes and the edge weights are the similarities of genes.

In the graph G_{dg}, we utilize a label propagation method, which iteratively propagates the label information of driver genes on G_{dg}, to discover the unobserved gene-patient associations to refine M_{gp}. The initial associations of n genes and the patient p_i in M_{gp} can be regarded as the initial labels of n genes for the patient p_i. In each propagation, every driver gene receives label information from the gene's neighbors with proportion α and reserves the initial label with proportion $1 - \alpha$. The iteration is defined as:

$$F_i^{t+1} = \alpha S F_i^t + (1 - \alpha) F_i^0 \tag{4}$$

where $F_i^0 = \left(f_{1i}^0, f_{2i}^0, \ldots, f_{ni}^0\right)^T$ represents the initial labels of n genes for the patient p_i and $F_i^t = \left(f_{1i}^t, f_{2i}^t, \ldots, f_{ni}^t\right)^T$ denotes the predicted labels of n genes for the patient p_i at iteration t [34]. Considering all m patients, let $F^t = (F_1^t, F_2^t, \ldots, F_m^t)$, and the iteration process can be formulated in matrix form as follows:

$$F^{t+1} = \alpha S F^t + (1 - \alpha) F^0 \tag{5}$$

We can use Eq. (5) to update the label matrix. Finally, Eq. (5) will be converged to the following:

$$F^* = (1 - \alpha)(I - \alpha S)^{-1} F^0 \tag{6}$$

F^* is the predicted gene-patient association score matrix. Then the gene-patient associations in M_{gp} are refined to F^* by inferring the unobserved gene-patient associations through propagating the label information of driver genes. We can obtain the prediction scores of the genes for each patient from F^* and identify the highly rank genes as the personalized driver genes of each patient for further analysis. More details on the convergence inference of the label propagation can be found in literature [35].

Results

Performance comparison

In this section, we validate the effectiveness of LPDriver by comparing it with other ten state-of-the-art methods including five personalized cancer driver identification methods including PNC [24], WMDS.netP [23], DawnRank [18], SCS [19] and SSN

[38], and five population level cancer driver identification methods including DriverNet [12], OncoDriveFM [8], MutSigCV [7], DriverML [9] and ActiveDriver [10].

Based on the data availability of the compared methods, we used twelve TCGA [26] cancer datasets as the test datasets: Breast invasive carcinoma (BRCA), Colon adenocarcinoma (COAD), Head and neck squamous cell carcinoma (HNSC), Kidney chromophobe (KICH), Kidney renal clear cell carcinoma (KIRC), Kidney renal papillary cell carcinoma (KIRP), Liver hepatocellular carcinoma (LIHC), Lung adenocarcinoma (LUAD), Lung squamous cellcarcinoma (LUSC), Prostate adenocarcinoma (PRAD), Papillary thyroid carcinoma (THCA) and Uterine corpus endometrial carcinoma (UCEC).

The known driver genes of Cancer Gene Census (CGC v.84) [28] and the Network of Cancer Genes (NCG 6.0) [27] database are used as the ground truth for assessing predicted driver genes. In cancer research, CGC and NCG are commonly used cancer gene datasets for validating driver genes predicted by computational methods. In total, 711 known cancer genes and 616 cancer census genes are downloaded from CGC and NCG gene lists (see Additional file 2).

The predicted driver genes annotated in the NCG and CGC were utilized to compute the F-measure to evaluate the performance of different methods [2, 24]. F-measure is computed by the following equation:

$$F - measure = \frac{2 * precision * recall}{precision + recall} \tag{7}$$

where the precision denotes the ratio of correctly identified driver genes to all identified driver genes and the recall denotes the ratio of correctly identified driver genes to the driver genes of NCG and CGC [2, 24]. In the performance comparisons, LPDriver finds potential driver genes from 12 cancer datasets with the proportion parameter $\alpha = 0.5$ and choose the identified genes appearing among over 80% patients in each dataset as the resulting driver genes. For these ten comparative methods, we obtained their identified driver genes for twelve TCGA cancer datasets from the WMDS.netP paper [23]. These identified driver genes of the comparison methods were obtained by using the same TCGA cancer type datasets based on the default parameter values provided in their papers.

Figure 2 shows the F-measures of the predicted cancer driver genes from different methods. As can be seen in Fig. 2, we can find that LPDriver outperforms other comparative methods in terms of the average values of F-measure. This result indicates that LPDriver is an effective method for predicting cancer driver genes.

Moreover, to verify whether LPDriver detects similar driver genes as other top 5 performing methods such as PNC, WMDS.netP, DawnRank, ActiveDriver and DriverML, we also compare the overlap between the prediction results of these methods. The discovered cancer drivers of these comparison methods are validated with both NCG and CGC and intersected to find the overlaps. Figure 3 shows the overlap among different methods for BRCA. The prediction overlap of the remaining eleven datasets can be found in the Additional file 3. As we can see in Fig. 3, although different methods have identified similar driver genes, the prediction overlap between LPDriver and other methods demonstrate that LPDriver is able to identify cancer driver genes that have not

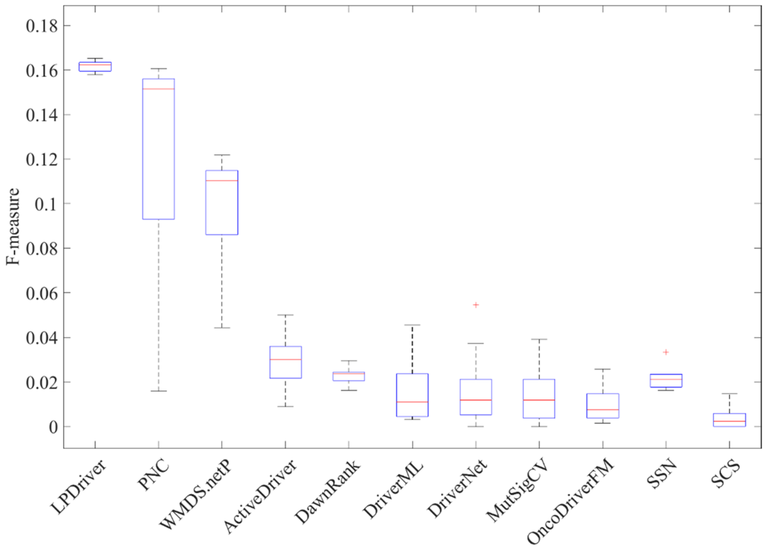


Fig. 2 Significant enrichment F-measures of the results from 11 methods. For each method, the F-measure values are the average results of twelve TCGA cancer datasets including BRCA, COAD, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, THCA and UCEC

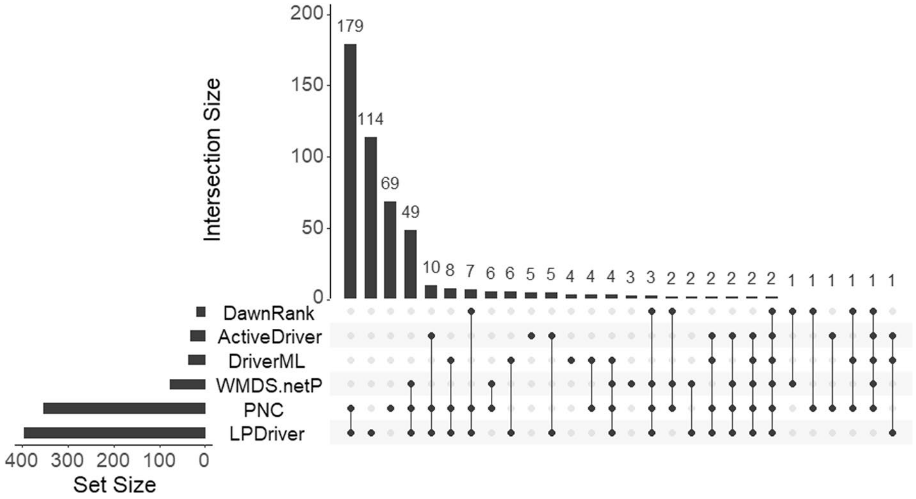


Fig. 3 Overlap among different methods for BRCA. The chart depicts the overlap between the driver genes detected by the six methods (LPDriver, PNC, WMDS.netP, DawnRank, ActiveDriver and DriverML) for BRCA. The horizontal bars situated at the bottom left signify the number of the identified driver genes that have been validated in both NCG and CGC. Meanwhile, the vertical bars, in conjunction with the dotted lines, collectively depict the overlaps among the validated driver genes of different methods

yet been identified by other methods. The complementarity of these methods can be utilized to maximize the prediction accuracy of cancer driver genes.

Effect of proportion parameter α

LPDriver relies on propagating label information of genes in the graph G_{dg} to mine gene-patient associations to detect personalized driver genes. The parameter α is used to adjust the gene label propagation proportion for each gene's initial label and the label information from neighborhood genes, which is critical in propagating gene labels.

In order to learn the impact of α on the performance of LPDriver, we calculate the F-measures of LPDriver with different α on twelve cancer datasets and the results are shown in Table 1. As we can see in Table 1, for 7 out of 12 datasets (i.e. COAD, HNSC, KIRC, LUAD, LUSC, PRAD and THCA), LPDriver receives the best F-measure in the case of $\alpha = 0.5$. Meanwhile, LPDriver yields the best F-measure for BRCA and LIHC when $\alpha = 0.4$, and gives the best F-measure for KICH and UCEC when $\alpha = 0.6$. This result indicates that the gene's initial label and the label information from neighborhood genes contribute almost equally to the identification of cancer driver genes. We thus identically set α as 0.5.

Influence of different reference networks

In order to comprehensively learn the influence of the reference gene interaction network on the network-based methods, we also evaluated the performance of LPDriver and other top 2 performing network-based methods PNC [24] and WMDS.netP [23] using different reference networks. These reference gene interaction networks are ConsensusPathDB [30, 31], HumanNet [39], StringNet [40] and the best performing reference network used in PNC, which is marked as network 6 in literature [24] and is called PNCNet in this work. Table 2 summarizes the number of genes and interactions from each of these four networks. (See Additional file 1 for more details of these networks).

To obtain fair and convincing comparison results, we obtained the source codes of PNC and WMDS.netP from their literatures and ran these two comparative algorithms according to the default values suggested by their papers. Specifically, PNC sets to its default p value threshold 0.05 and WMDS.netP sets to its default hyperparameter γ 0.01. In the comparison, all compared algorithms were run on a computer with an Intel Xeon 6130 and 208GB RAM. The running operating system is Linux. We then compared the performance of LPDriver, PNC and WMDS.netP using four reference gene interaction networks on 12 cancer datasets and the comparison results are shown in Fig. 4.

As shown in Fig. 4, for using different reference gene interaction networks, LPDriver achieves better performance than PNC and WMDS.netP. This indicates that LPDriver is an effective network-based method for predicting cancer driver genes using different reference networks. Moreover, as we can see in Fig. 4, the F-measures of these three methods vary for different reference networks. This demonstrates that reference network has direct impact on these four network-based methods in deed. Interestingly, in Fig. 4, we can observe that, the variances of the F-measures for PNC and LPDriver using different reference networks are larger than those of WMDS.netP. These results reveal that PNC and LPDriver are more sensitive to reference network than WMDS.netP, and the use of proper reference network may enable LPDriver and PNC to obtain better performance.

Ablation study

In this section, we will evaluate the effectiveness of different parts of LPDriver by ablation study. In the second step, LPDriver identifies the potential driver genes by finding the genes in the maximum matching set of gene-interaction bipartite graph and we call this step as BGGM for simplicity. In order to evaluate the effectiveness of the step BGGM, we replace the step BGGM of LPDriver with three classical network control methods including NCUA [24], MMS [24, 25] and MDS [24, 25] respectively to find

Table 1 Performance comparison with different α on twelve cancer datasets

α	BRCA	COAD	HNSC	KICH	KIRC	KIRP	LIHC	LUAD	LUSC	PRAD	THCA	UCEC
0.1	0.1501	0.1537	0.1521	0.1426	0.1465	0.1446	0.1564	0.1589	0.1564	0.1489	0.1496	0.1464
0.2	0.1572	0.1553	0.1613	0.1413	0.1489	0.1475	0.1588	0.1565	0.1585	0.1523	0.1534	0.1423
0.3	0.1630	0.1573	0.1565	0.1535	0.1534	0.1496	0.1613	0.1603	0.1605	0.1589	0.1610	0.1520
0.4	**0.1665**	0.1616	0.1593	0.1598	0.1580	0.1553	**0.1623**	0.1591	0.1626	0.1623	0.1586	0.1480
0.5	0.1619	**0.1631**	**0.1624**	0.1578	**0.1651**	0.1582	0.1605	**0.1615**	**0.1626**	**0.1648**	**0.1635**	0.1584
0.6	0.1617	0.1605	0.1612	**0.1612**	0.1621	0.1610	0.1605	0.1603	0.1579	0.1617	0.1586	**0.1612**
0.7	0.1577	0.1586	0.1583	0.1576	0.1634	**0.1634**	0.1574	0.1584	0.1586	0.1638	0.1565	0.1576
0.8	0.1536	0.1592	0.1546	0.1562	0.1602	0.1621	0.1532	0.1565	0.1543	0.1623	0.1555	0.1563
0.9	0.1488	0.1571	0.1525	0.1553	0.1580	0.1582	0.1517	0.1543	0.1486	0.1589	0.1584	0.1534

*Significant of bold values demonstrate the best F-measure of LPDriver with different α on each dataset

Table 2 The gene number (Nodes) and interaction number (Edges) in four networks ConsensusPathDB, PNCNet, HumanNet and StringNet

	ConsensusPathDB	PNCNet	HumanNet	StringNet
Nodes	4912	9160	15,351	11,302
Edges	96,256	104,153	158,499	273,210

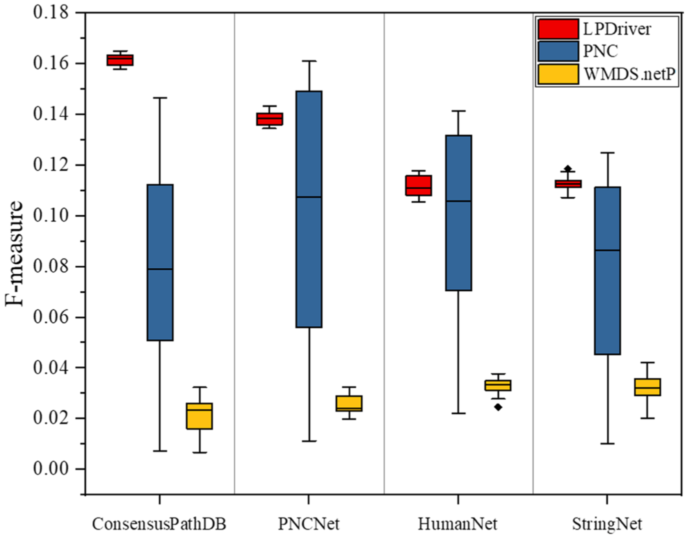


Fig. 4 Average F-measures of LPDriver, PNC and WMDS.netP on twelve cancer data sets using different reference gene interaction networks (i.e. ConsensusPathDB, PNCNet, HumanNet and StringNet)

driver genes in the gene-interaction bipartite graph. We constructed three modes for LPDriver, namely LPDriver_NCUA, LPDriver_MMS and LPDriver_MDS. Specifically, LPDriver_NCUA denotes the step BGGM of LPDriver is replaced by NCUA. LPDriver_MMS denotes the step BGGM of LPDriver is replaced by MMS. LPDriver_MDS denotes the step BGGM of LPDriver is replaced by MDS.

Figure 5 shows the F-measures of the predicted cancer driver genes of four different LPDriver's modes on twelve cancer datasets. Furthermore, in Fig. 5, to estimate the effect of predicting driver genes using linear neighborhood propagation in LPDriver, we performed these four LPDriver's modes with and without using linear neighborhood propagation respectively.

From Fig. 5, we can see that, for the mode performed with using linear neighborhood propagation, the F-measure of LPDriver is better than that of LPDriver_NCUA, LPDriver_MMS and LPDriver_MDS. Similarly, in Fig. 5, for the mode performed without using linear neighborhood propagation, the F-measure of LPDriver is higher than that of LPDriver_NCUA, LPDriver_MMS and LPDriver_MDS as well. These results demonstrate that finding genes in the maximum matching set of the gene-interaction bipartite graph could be an effective way of identifying personalized driver genes for LPDriver.

Moreover, as can be seen in Fig. 5, the F-measures of the LPDriver's modes performed with using linear neighborhood propagation are much better than those of the

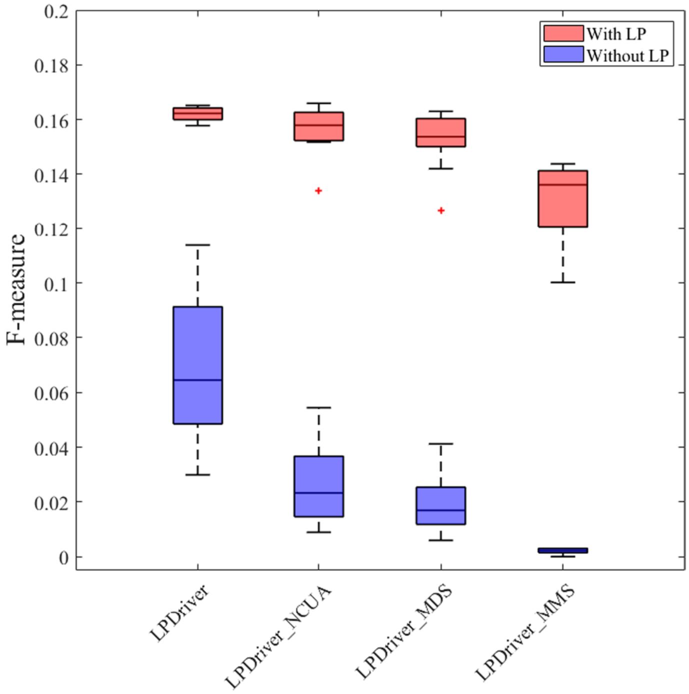


Fig. 5 F-measures of LPDriver, LPDriver_NCUA, LPDriver_MMS and LPDriver_MDS. "With LP" and "Without LP" denote that the LPDriver's modes are performed with and without using linear neighborhood propagation respectively. The x-axis shows four LPDriver's modes and the y-axis is for F-measure. For each mode, the F-measure value is the result of twelve cancer datasets including BRCA, COAD, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, THCA and UCEC

LPDriver's modes performed without using linear neighborhood propagation. These results illustrate the effectiveness of predicting driver genes via linear neighborhood propagation in LPDriver.

Discovering personalized driver genes

The personalized driver genes may vary for different patients due to cancer heterogeneity [3]. Hence, the genes that rarely mutated in a population could potentially drive the cancer development of each patient, and these genes are also called rare driver genes [3]. The rare driver genes are likely ignored by mutation frequency-based methods [3]. In this section, we discuss the rare driver genes predicted by LPDriver to learn the effectiveness of detecting personalized driver genes for LPDriver at the individual level.

Here, we define the genes that are mutated in no more than 5% of patients in a cancer dataset and is ranked top 100 of a patient's potential driver genes as the personalized rare driver genes. We used DAVID [41–43] tools to perform the functional enrichment analysis on these personalized rare driver genes identified by LPDriver against the pathway database Kyoto Encyclopedia of Genes and Genomes (KEGG) [44]. Taking BRCA as an example, the top 10 enriched KEGG pathways of the personalized rare driver genes are shown in Fig. 6. The identified rare driver genes and the top 10 enriched pathways of these driver genes for BRCA, COAD, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, THCA and UCEC cancers are given in the Additional file 4 and file 5 respectively. From these results, we can find that these rare driver genes are enriched in some

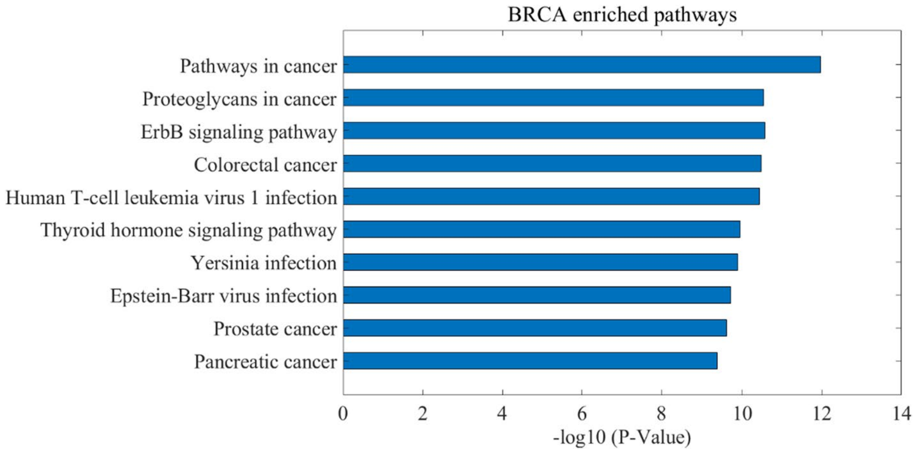


Fig. 6 Top 10 enriched KEGG pathways of the personalized rare driver genes on BRCA cancer dataset. The Y-axis indicates the name of enriched KEGG pathway. The X-axis represents the opposite value of log transformed *p* value. The larger value of X-axis indicates that the genes are more significantly enriched in the pathway

critical pathways related with cancer [41]. It is noted that, in the functional enrichment analysis [2, 3, 24, 45], the higher -log(*p* value) value of the pathway is, the better enriched significance of the pathway is. Notably, we chose three rare driver genes *GSK3B*, *SP1* and *XRCC6*, which have the minimum occurrence frequency as mutated genes in BRCA cancer dataset, to analyze.

Taking the gene *GSK3B* as example, it ranks the top 2.3% of the potential driver genes for the patient TCGA-BH-A0DL and ranks the top 2.8% of the potential driver genes for the patient TCGA-BH-A18M. Even though the mutation frequency of gene *GSK3B* in BRCA cancer dataset is only 0.504%, our LPDriver still ranks *GSK3B* in top 10% of all personalized driver genes with mutation in BRCA cancer dataset. Additionally, it was reported that the overexpression of *GSK3B* promotes the development of multiple cancers [46]. Similar results can also be observed in the patients with genes *SP1* and *XRCC6*. *SP1* plays a critical role in the development of pancreatic cancer [47] and *XRCC6* is a risk allele for breast cancer [48]. These two rare mutation genes *SP1* and *XRCC6* are also obviously ranked ahead in the personalized driver genes for other patients in BRCA cancer dataset. Even when some driver genes rarely mutate in a cohort, LPDriver still uncovers and promotes these genes for personalized therapies. In short, the above results demonstrate that LPDriver is able to detect personalized driver genes by their network characteristics even if the mutation profiles of the genes are hidden.

Statistic analysis of the identified driver genes

LPDriver constructs personalized gene interaction network by using a group of tumor reference samples and the size of tumor reference samples may influence the effect of identifying driver genes. In order to learn the impact of tumor reference sample size to the performance of LPDriver, we apply LPDriver to identify driver genes on twelve cancer datasets (i.e. BRCA, COAD, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, THCA and UCEC) by using different sizes of tumor reference samples. The average F-measures of LPDriver for each size of tumor reference samples across the twelve datasets are shown in Fig. 7. As can be seen in Fig. 7, the F-measure values keep stable with different sizes of tumor reference samples. This result indicates that the performance of LPDriver is not greatly affected by the size of tumor reference samples to some extent.

Detecting potentially novel driver genes

In order to assess the ability of LPDriver on detecting potentially novel driver genes, we first used LPDriver to search for the top-ranked 100 driver genes from the breast cancer dataset BRCA while not in NCG and CGC, and obtained 42 potentially novel driver genes (see Additional file 6). We then used DAVID [41] tools to perform functional enrichment analysis on these 42 obtained driver genes against Genetic Association Database (GAD) which records the genes associated with diseases [49]. Interestingly, 36(85.7%) of these 42 genes are involved in GAD, and 28 (66.7%) genes are related with cancer, 18 genes are enriched for "Breast Cancer" (p value $= 9.3 \times 10^{-4}$, FDR $= 6.6 \times 10^{-2}$). Particularly, it has been confirmed that *ACTB* (actin beta) (ranked the 3th in patient TCGA-E2-A158 and the 5th in patient TCGA-BH-A1EU), is distinctly associated with the metastatic ability of human colon adenocarcinoma cells [50] and accumulating evidence demonstrates that *ACTB* is irregularly expressed in a variety of cancers and affects the metastasis and invasiveness of tumors [51].

Moreover, we performed enrichment analysis on these 42 genes against three pathway databases Kyoto Encyclopedia of Genes and Genomes (KEGG) [52], Gene Ontology (GO) [53, 54] and Reactome [55], the results show that these 42 genes are enriched for "Viral carcinogenesis" (KEGG pathway, p value $= 4.8 \times 10^{-11}$, FDR $= 6.3 \times 10^{-9}$), "R-HSA-2894862" (Reactome pathway, p value $= 2.3 \times 10^{-9}$, FDR $= 2.5 \times 10^{-7}$), and "viral process" (GO biological process, p value $= 3.4 \times 10^{-8}$, FDR $= 2.1 \times 10^{-5}$). Specifically, *RBPJ* (ranked 8th in patient TCGA-BH-A0BZ and 9th

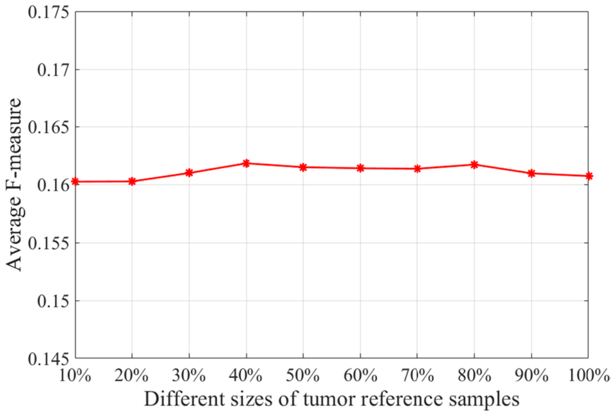


Fig. 7 Average F-measures of LPDriver for different sizes of tumor reference samples. The y-axis indicates the average F-measures across the twelve datasets (i.e. BRCA, COAD, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, THCA and UCEC). The x-axis shows the size of tumor reference samples used to construct gene interaction network, e.g. 10% indicates 10% of the tumor reference samples in the cancer data set are used to construct gene interaction network

in patient TCGA-BH-A0BJ) was reported as a potential oncogene in certain cancers and the upregulation of *RBPJ* can induce pancreatic cancer [56], although *RBPJ* has not been listed in NGC and CGC under current version yet.

Driver genes may play different roles in cancer. Based on the identified driver genes from BRCA, we accordingly categorized some of these potentially novel driver genes that were recently reported to associate with cancer into four types based on their roles in the development of cancer [1]. The four types are the direct driver gene, induced driver gene, target driver gene and biomarker driver gene. These categorized genes are summarized in Additional file 7. In the following we discuss some of these categorized genes to learn the effectiveness of identifying potentially novel driver genes for LPDriver. The direct driver gene is the driver gene that was reported to directly cause the cancer [57]. For example, *CDK2*(ranked 16th in patient TCGA-BH-A18M) was recently discovered to be essential for the proliferation of prostate cancer cell [58]. Moreover, accumulating evidence indicates that *MED23* (ranked 9th in patient TCGA-E2-A1LH) plays an oncogenic role in the development of NSCLC (a non-small cell lung cancer) and influences the invasiveness and development of tumors [59].

The induced driver gene can exert its action on other genes or proteins to cause cancer [60]. For instance, *PCAF* (ranked 11th in patient TCGA-E9-A1RF), is recently reported to have positive role for inducing the acetylation of Glycerol 3-phosphate dehydrogenase (an enzyme in glycolysis) to promote cell proliferation in liver tumor [61]. Besides, due to the decrement of expression level, the attenuated function of gene *SIN3A* (ranked 93th in patient TCGA-E9-A1N6) may lead to the epigenetic deregulation of the growth-associated genes, which results in the oncogenesis of lung cancer cells [62]. Basically, the direct and induced driver genes predicted by LPDriver could help us to study the cause of cancer on genomic level.

On the other hand, the target driver gene could serve as the therapy target for curing cancer [63]. For example, recently, the overexpression of *SKIP* (ranked 10th in patient TCGA-BH-A1FD) was included in the pathogenesis and diagnosis of breast cancer, which could possibly serve as a future therapeutic target for breast cancer [64]. Besides, *SH3KBP1*(ranked 10th in patient TCGA-BH-A1FD) was reported to serve as a new regulator of carcinogenic EGFR (Epidermal Growth Factor Receptor), and it could also serve as a potential therapy target for GBM (Glioblastoma multiforme, a kind of cancer) patients with EGFR activation [65].

The biomarker driver gene could serve as biomarker for detecting the existence of cancer cell [66]. For example, a recent report showed that *RPA1* (ranked 8th in patient TCGA-E2-A15M) works as a presumed oncogene in tumorigenesis and serves as a prognosticative biomarker for colorectal cancer [67]. Also, the upregulation of *HnRNPM* (ranked 13th in patient TCGA-BH-A0H9) is contained in the human colorectal epithelial tumorigenesis and could serve as a tumor biomarker for colorectal cancer. These biomarker and target driver genes detected by LPDriver could help people to find the existence of cancer cell and provide potential therapy target for curing cancer.

In summary, the abovementioned results demonstrate that LPDriver is also an effective method for detecting potentially novel cancer driver genes.

Discussion and conclusion

In this work, we propose LPDriver, a novel computational method for predicting personalized cancer driver genes. LPDriver offers several distinct advantages over its counterparts. First, LPDriver innovatively explores gene-patient associations within personalized gene networks to uncover functionally similar genes among patients. A key differentiator in this step is that LPDriver does not rely on known driver genes or mutation data to detect driver genes. This novelty accelerates and simplifies the identification process of cancer driver genes. Meanwhile, unlike other network-based methods, LPDriver identifies potential driver genes by selecting genes from the maximum matching set of personalized gene networks. This maximizes the coverage of gene network edges while preserving room for further exploration of driver genes, and strikes a balance between comprehensiveness and specificity in identifying candidate driver genes.

We have conducted comprehensive experiments on multiple cancer datasets from TCGA, benchmarking LPDriver against the state-of-the-art methods. LPDriver's superior performance in the experimental comparison demonstrates its effectiveness in detecting cancer driver genes. Notably, LPDriver excels in identifying known cancer driver genes, while also revealing potentially novel driver genes that are documented to be cancer-related. LPDriver thus can serve as an effective and valuable complement to the existing toolkit for identifying cancer driver gene, ultimately contributing to a comprehensive understanding of cancer genetics.

Despite the effectiveness of LPDriver in identifying cancer driver genes, some limitations remain. LPDriver constructs personalized gene networks upon the same known gene interaction networks and the gene expression data of a specific cancer. The information of specific cancer could be lost in the construction of personalized gene networks. A further extension is to utilize the gene interactions that are specific to a cancer under consideration to initiate the construction of personalized gene networks for a specific cancer. Moreover, as a future work, a variety of the biological features of genes and cancers, such as sequence profiles of genes or miRNA, could be incorporated to further promote the identification performance. Furthermore, for simplicity, LPDriver sets the edge weight of the bipartite graph of gene network to 1 to find the maximum matching in the bipartite graph for inferring potential gene-patient associations. In the future, more gene interaction information could be integrated to enrich the edge weight of the bipartite graph of gene network, which may help to find more accurate gene-patient associations.

Abbreviations

TCGA	The cancer genome atlas
CGC	Cancer gene census
NCG	Network of Cancer Genes
LP	Linear neighborhood propagation
PCC	Pearson correlation coefficient
BRCA	Breast invasive carcinoma
COAD	Colon adenocarcinoma
HNSC	Head and neck squamous cell carcinoma
KICH	Kidney chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cellcarcinoma

PRAD	Prostate adenocarcinoma
THCA	Papillary thyroid carcinoma
UCEC	Uterine corpus endometrial carcinoma
KEGG	Kyoto Encyclopedia of Genes and Genomes
GO	Gene ontology

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05662-4.

Additional file 1: The gene interaction network.

Additional file 2: The list of Cancer Genes Census (CGC) and Network of Cancer Genes (NCG).

Additional file 3: The prediction overlap of the remaining 11 datasets.

Additional file 4: The identified rare driver genes.

Additional file 5: The top 10 enriched pathways of the identified rare driver genes.

Additional file 6: The 42 potentially novel driver genes.

Additional file 7: The categorized genes identified by LPDriver.

Acknowledgements

The authors thank the editor and anonymous reviewers for their constructive comments and suggestions, which greatly help us improve our manuscript. The experimental work was supported by the high-performance computing center of Guangxi University.

Author contributions

YH and FC conceived of the study and designed the methodology. FC carried out the source code and program. YH, FC and CZ performed the experiments and the analysis of the results. YH, HS, FC and CZ wrote and revised the manuscript. All authors read and approved the final manuscript.

Funding

This work is supported by the National Natural Science Foundation of China (No. 62362004), the Natural Science Foundation of Guangxi Province (No. 2020GXNSFAA159074) and the National Natural Science Foundation of China (No. 61862006).

Availability of data and code

The source code and data are available at https://github.com/hyr0771/LPDriver.

Declarations

Ethics approval and consent to participate

All the methods were performed in accordance with the relevant guidelines and regulations.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 29 August 2023 Accepted: 18 January 2024

Published online: 22 January 2024

References

1. Cho A, Shim JE, Kim E, Supek F, Lehner B, Lee I. MUFFINN: cancer gene discovery via network analysis of somatic mutation data. Genome Biol. 2016;17(1):129.
2. Pham VVH, Liu L, Bracken CP, Nguyen T, Goodall GJ, Li J, Le TD. pDriver: a novel method for unravelling personalized coding and miRNA cancer drivers. Bioinformatics. 2021;37(19):3285–92.
3. Zhang T, Zhang S-W, Li Y. Identifying driver genes for individual patients through inductive matrix completion. Bioinformatics. 2021;37(23):4477–84.
4. Peng YZ, Lin Y, Huang Y, Li Y, Luo G, Liao J. GEP-EpiSeeker: a gene expression programming-based method for epistatic interaction detection in genome-wide association studies. BMC Genomics. 2021;22(1):910.
5. Huang Y, Xie Y, Zhong C, Zhou F. Finding branched pathways in metabolic network via atom group tracking. PLoS Comput Biol. 2021;17(2):e1008676.
6. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol. 2016;17(1):128.

7. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214–8.
8. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Res. 2012;40(21):e169–e169.
9. Han Y, Yang J, Qian X, Cheng W-C, Liu S-H, Hua X, Zhou L, Yang Y, Wu Q, Liu P. DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies. Nucleic Acids Res. 2019;47(8):e45–e45.
10. Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol. 2013;9(1):637.
11. Pham VV, Liu L, Bracken CP, Goodall GJ, Long Q, Li J, Le TD. CBNA: a control theory based method for identifying coding and non-coding cancer drivers. PLoS Comput Biol. 2019;15(12):e1007538.
12. Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, Huntsman DG, Caldas C, Aparicio SA, Shah SP. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol. 2012;13(12):1–14.
13. Song J, Peng W, Wang F. A random walk-based method to identify driver genes by integrating the subcellular localization and variation frequency into bipartite graph. BMC Bioinf. 2019;20(1):1–17.
14. Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. 2012;22(2):398–406.
15. Huang Y, Wu Z, Lan W, Zhong C: Predicting disease-associated N7-methylguanosine(m7G) sites via random walk on heterogeneous network. In: IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2023.
16. Huang Y, Bin Y, Zeng P, Lan W, Zhong C. NetPro: neighborhood interaction-based drug repositioning via label propagation. IEEE/ACM Trans Comput Biol Bioinf. 2023;20(3):2159–69.
17. Huang Y, Zhong C. Detecting list-colored graph motifs in biological networks using branch-and-bound strategy. Comput Biol Med. 2019;107:1–9.
18. Hou JP, Ma J. DawnRank: discovering personalized driver genes in cancer. Genome Med. 2014;6(7):1–16.
19. Guo W-F, Zhang S-W, Liu L-L, Liu F, Shi Q-Q, Zhang L, Tang Y, Zeng T, Chen L. Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. Bioinformatics. 2018;34(11):1893–903.
20. Dinstag G, Shamir R. PRODIGY: personalized prioritization of driver genes. Bioinformatics. 2020;36(6):1831–9.
21. Erten C, Houdjedj A, Kazan H, Taleb Bahmed AA. PersonaDrive: a method for the identification and prioritization of personalized cancer drivers. Bioinformatics. 2022;38(13):3407–14.
22. Erten C, Houdjedj A, Kazan H. Ranking cancer drivers via betweenness-based outlier detection and random walks. BMC Bioinformatics. 2021;22(1):1–16.
23. Cheng X, Amanullah M, Liu WG, Liu Y, Pan XQ, Zhang HH, Xu HM, Liu PY, Lu Y. WMDSnet: a network control framework for identifying key players in transcriptome programs. Bioinformatics. 2023;39(2):btad071.
24. Guo W-F, Zhang S-W, Zeng T, Li Y, Gao J, Chen L. A novel network control model for identifying personalized driver genes in cancer. PLoS Comput Biol. 2019;15(11):e1007520.
25. Liu Y-Y, Slotine J-J, Barabási A-L. Controllability of complex networks. Nature. 2011;473(7346):167–73.
26. Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, Ally A, Balasundaram M, Birol I, Butterfield YSN, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113–20.
27. Repana D, Nulsen J, Dressler L, Bortolomeazzi M, Venkata SK, Tourna A, Yakovleva A, Palmieri T, Ciccarelli FD. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol. 2019;20(1):1–12.
28. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer. 2004;4(3):177–83.
29. Kuijjer ML, Tung MG, Yuan G, Quackenbush J, Glass K. Estimating sample-specific regulatory networks. Iscience. 2019;14:226–40.
30. Kamburov A, Wierling C, Lehrach H, Herwig R. ConsensusPathDB—a database for integrating human functional interaction networks. Nucleic Acids Res. 2008;37(suppl_1):D623–8.
31. Barel G, Herwig R. NetCore: a network propagation approach using node coreness. Nucleic Acids Res. 2020;48(17):e98–e98.
32. Chewi S, Yang F, Ghosh A, Parekh A, Ramchandran K: Matching observations to distributions: Efficient estimation via sparsified hungarian algorithm. In: *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton): 2019*. IEEE, p. 368–75.
33. Grinman A. The Hungarian algorithm for weighted bipartite graphs. Massachusetts Institute of Technology; 2015.
34. Li G, Luo J, Xiao Q, Liang C, Ding P. Predicting microRNA-disease associations using label propagation based on linear neighborhood similarity. J Biomed Inform. 2018;82:169–77.
35. Wang F, Zhang C. Label propagation through linear neighborhoods. IEEE Trans Knowl Data Eng. 2007;20(1):55–67.
36. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science. 2000;290(5500):2323–6.
37. Jorge N, Stephen JW. Numerical optimization. Spinger; 2006.
38. Liu XP, Wang YT, Ji HB, Aihara K, Chen LN. Personalized characterization of diseases using sample-specific networks. Nucleic Acids Res. 2016;44:22.
39. Hwang S, Kim CY, Yang S, Kim E, Hart T, Marcotte EM, Lee I. HumanNet v2: human gene networks for disease research. Nucleic Acids Res. 2019;47(D1):D573–80.
40. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(D1):D447–52.
41. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13.
42. Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA. The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. 2007;8(9):1–16.

43. Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 2007;35(suppl_2):W169–75.
44. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. Nucleic Acids Res. 2019;47(D1):D590–5.
45. Hou Y, Gao B, Li G, Su Z. MaxMIF: a new method for identifying cancer driver genes through effective data integration. Adv Sci. 2018;5(9):1800640.
46. Chen L, Zuo Y, Pan R, Ye Z, Wei K, Xia S, Li W, Tan J, Xia X. GSK-3β regulates the expression of P21 to promote the progression of chordoma. Cancer Manag Res. 2021;13:201.
47. Malsy M, Graf B, Almstedt K. The active role of the transcription factor Sp1 in NFATc2-mediated gene regulation in pancreatic cancer. BMC Biochem. 2019;20(1):1–11.
48. Willems P, De Ruyck K, Van den Broecke R, Makar A, Perletti G, Thierens H, Vral A. A polymorphism in the promoter region of Ku70/XRCC6, associated with breast cancer risk and oestrogen exposure. J Cancer Res Clin Oncol. 2009;135(9):1159–68.
49. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet. 2004;36(5):431–2.
50. Nowak D, Skwarek-Maruszewska A, Zemanek-Zboch M, Malicka-Błaszkiewicz M. Beta-actin in human colon adenocarcinoma cell lines with different metastatic potential. Acta Biochim Pol. 2005;52(2):461–8.
51. Gu Y, Tang S, Wang Z, Cai L, Lian H, Shen Y, Zhou Y. A pan-cancer analysis of the prognostic and immunological role of β-actin (ACTB) in human cancers. Bioengineered. 2021;12(1):6166–85.
52. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2016;45(D1):D353–61.
53. Consortium TGO. Gene ontology consortium: going forward. Nucleic Acids Res. 2014;43(D1):D1049–56.
54. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9.
55. Fabregat A, Sidiropoulos K, Viteri G, Forner O, Marin-Garcia P, Arnau V, D'Eustachio P, Stein L, Hermjakob H. Reactome pathway analysis: a high-performance in-memory approach. BMC Bioinf. 2017;18(1):142.
56. Onishi H, Yamasaki A, Kawamoto M, Imaizumi A, Katano M. Hypoxia but not normoxia promotes Smoothened transcription through upregulation of RBPJ and Mastermind-like 3 in pancreatic cancer. Cancer Lett. 2016;371(2):143–50.
57. Hu H, Zhang Y, Zou M, Yang S, Liang X-Q. Expression of TRF1, TRF2, TIN2, TERT, KU70, and BRCA1 proteins is associated with telomere shortening and may contribute to multistage carcinogenesis of gastric cancer. J Cancer Res Clin Oncol. 2010;136(9):1407–14.
58. Flores O, Wang Z, Knudsen KE, Burnstein KL. Nuclear targeting of cyclin-dependent kinase 2 reveals essential roles of cyclin-dependent kinase 2 localization and cyclin E in vitamin D-mediated growth inhibition. Endocrinology. 2010;151(3):896–908.
59. Shi J, Liu H, Yao F, Zhong C, Zhao H. Upregulation of mediator MED23 in non-small-cell lung cancer promotes the growth, migration, and metastasis of cancer cells. Tumor Biol. 2014;35(12):12005–13.
60. Samoylenko A, Vynnytska-Myronovska B, Byts N, Kozlova N, Basaraba O, Pasichnyk G, Palyvoda K, Bobak Y, Barska M, Mayevska O. Increased levels of the HER1 adaptor protein Ruk l/CIN85 contribute to breast cancer malignancy. Carcinogenesis. 2012;33(10):1976–84.
61. Hirano G, Izumi H, Kidani A, Yasuniwa Y, Han B, Kusaba H, Akashi K, Kuwano M, Kohno K. Enhanced expression of PCAF endows apoptosis resistance in cisplatin-resistant cells. Mol Cancer Res. 2010;8(6):864–72.
62. Suzuki H, Ouchida M, Yamamoto H, Yano M, Toyooka S, Aoe M, Shimizu N, Date H, Shimizu K. Decreased expression of the SIN3A gene, a candidate tumor suppressor located at the prevalent allelic loss region 15q23 in non-small cell lung cancer. Lung Cancer. 2008;59(1):24–31.
63. Domoto T, Uehara M, Bolidong D, Minamoto T. Glycogen synthase kinase 3β in cancer biology and treatment. Cells. 2020;9(6):1388.
64. Chen S, Zhang J, Duan L, Zhang Y, Li C, Liu D, Ouyang C, Lu F, Liu X: Identification of HnRNP M as a novel biomarker for colorectal carcinoma by quantitative proteomics. *American Journal of Physiology-Gastrointestinal and Liver Physiology* 2014, 306(5):G394-G403.
65. Song H, Wang Y, Shi C, Lu J, Yuan T, Wang X. SH3KBP1 promotes glioblastoma tumorigenesis by activating EGFR signaling. Front Oncol. 2021;10:3155.
66. Ju Q, Zhao Y, Li X-M, Zhang H. BRCA1 Associated Protein is a prognostic biomarker and correlated with immune infiltrates in liver hepatocellular carcinoma: a pan-cancer analysis. Front Mol Biosci. 2020;7:279.
67. Li S, Xu K, Gu D, He L, Xie L, Chen Z, Fan Z, Zhu L, Du M, Chu H. Genetic variants in RPA1 associated with the response to oxaliplatin-based chemotherapy in colorectal cancer. J Gastroenterol. 2019;54(11):939–49.

Publisher's Note