Bioinformatics, 36(26), 2020, i583–i591
doi: 10.1093/bioinformatics/btaa797
ECCB2020

Genes

DriverGroup: a novel method for identifying driver gene groups

Vu V.H. Pham[1], Lin Liu[1], Cameron P. Bracken[2,3], Gregory J. Goodall[2,3], Jiuyong Li[1] and Thuc D. Le [1,*]

[1]UniSA STEM, University of South Australia, Mawson Lakes, SA, 5095, Australia, [2]Centre for Cancer Biology, an alliance of SA Pathology and University of South Australia, Adelaide, SA, 5000, Australia and [3]Department of Medicine, The University of Adelaide, Adelaide, SA 5005, Australia

*To whom correspondence should be addressed.

Abstract

Motivation: Identifying cancer driver genes is a key task in cancer informatics. Most existing methods are focused on individual cancer drivers which regulate biological processes leading to cancer. However, the effect of a single gene may not be sufficient to drive cancer progression. Here, we hypothesize that there are driver gene groups that work in concert to regulate cancer, and we develop a novel computational method to detect those driver gene groups.

Results: We develop a novel method named *DriverGroup* to detect driver gene groups by using gene expression and gene interaction data. The proposed method has three stages: (i) constructing the gene network, (ii) discovering critical nodes of the constructed network and (iii) identifying driver gene groups based on the discovered critical nodes. Before evaluating the performance of *DriverGroup* in detecting cancer driver groups, we firstly assess its performance in detecting the influence of gene groups, a key step of *DriverGroup*. The application of *DriverGroup* to DREAM4 data demonstrates that it is more effective than other methods in detecting the regulation of gene groups. We then apply *DriverGroup* to the BRCA dataset to identify driver groups for breast cancer. The identified driver groups are promising as several group members are confirmed to be related to cancer in literature. We further use the predicted driver groups in survival analysis and the results show that the survival curves of patient subpopulations classified using the predicted driver groups are significantly differentiated, indicating the usefulness of *DriverGroup*.

Availability and implementation: *DriverGroup* is available at https://github.com/pvvhoang/DriverGroup

Contact: Thuc.Le@unisa.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

It is important to identify cancer drivers and their regulatory mechanisms due to their critical role in the initialization and progression of cancer. Understanding cancer drivers is beneficial for the design of effective cancer treatments too. Thus, several computational methods have been developed to discover cancer drivers, for example OncodriveFM (Gonzalez-Perez and Lopez-Bigas, 2012), OncodriveCLUST (Tamborero *et al.*, 2013), ActiveDriver (Reimand and Bader, 2013), DawnRank (Hou and Ma, 2014) and CBNA (Pham *et al.*, 2019).

These methods, however, only identify single genes as cancer drivers, whereas there is evidence showing that genes work together to regulate the same targets and the regulation of individual genes might not have significant impacts (Cursons *et al.*, 2018; Karim *et al.*, 2016). Furthermore, researchers have started to conduct wet-lab experiments to investigate the regulation by groups of genes in biological processes (Cursons *et al.*, 2018). All these highlight the importance of studying biological components working in groups.

In this article, we introduce the concept of 'driver gene group', which is a set of genes that work in concert to regulate cancer or cancer markers. The driver gene groups are different from the gene modules studied by recent methods such as WeSME (Kim *et al.*, 2017), MEMo (Ciriello *et al.*, 2012) and iMCMC (Zhang et al., 2013). WeSME discovers cancer drivers by using statistical tests to evaluate the mutual exclusivity of mutations of gene pairs and the pairs whose mutations have a significant mutual exclusivity are considered as modular candidate drivers. Similar to WeSME, MEMo and iMCMC also use mutual exclusivity of gene mutations in detecting cancer drivers. However, instead of testing the mutual exclusivity of mutations of gene pairs, MEMo and iMCMC test the mutual exclusivity of mutations of genes in modules. The modules include genes which are recurrently altered in samples and likely to belong to the same pathway (in MEMo) or coherent subnetworks with large weights in both edges and nodes (in iMCMC).

Although the above methods detect modules of cancer drivers, members in each of the modules may not work jointly to regulate targets to drive cancer as the mutation in a single member of a module may have been sufficient to trigger cancer development (Kim *et al.*, 2017). However, the idea of driver gene groups is that all genes in a group collaboratively drive cancer. In addition, these

methods only deal with coding genes while cancer drivers may be non-coding genes since a large portion of mutations may exist in non-coding regions (Yang *et al.*, 2016a), and non-coding genes can regulate gene targets to drive cancer (Puente *et al.*, 2015; Weinhold *et al.*, 2014). Thus, there is a strong need for novel methods to identify driver groups of which the members work in concert to progress cancer while considering both coding and non-coding genes.

In this article, we propose a novel method named *DriverGroup* to identify both coding driver gene groups (i.e. driver groups including coding genes) and non-coding driver gene groups (i.e. driver groups including non-coding genes). As proliferation is associated with cancer development (Lopez-Saez *et al.*, 1998; Feitelson *et al.*, 2015) and proliferation genes are related to the prognosis of cancer patients (Li *et al.*, 2018), we identify driver gene groups by detecting groups of genes which collaboratively regulate proliferation genes.

Our method is based on the gene network and its critical nodes (i.e. nodes playing a central role in controlling the whole network) to identify driver gene groups. Because of the important role of critical nodes, we consider them as members of driver gene groups. Inspired by the Influence Maximization (IM) problem (Gong *et al.*, 2016; Yang *et al.*, 2016b) which identifies k-seed sets (i.e. sets have k seed nodes) with the maximum influence in a network, we develop novel algorithms to compute the influence of a group of critical nodes on the proliferation genes. At the end, a driver gene group is a maximal subset of critical nodes which have the maximum impact on the proliferation genes, i.e., adding or removing one critical node from the subset will decrease the impact of the subset.

Before evaluating *DriverGroup* in identifying driver gene groups, we firstly assess its ability in discovering the influence of the gene groups in a network using the DREAM4 data from the DREAM4 In Silico Network Challenge (Marbach *et al.*, 2010; Schaffter *et al.*, 2011). We then compare *DriverGroup* with jointIDA (Nandy *et al.*, 2017), a method used to estimate the joint effects of a group of variables on other variables, and the random method. Our method outperforms both jointIDA (Nandy *et al.*, 2017) and the random method in most cases. We then use the BRCA dataset for identifying driver gene groups and several members of the driver groups predicted by *DriverGroup* are confirmed to be related to cancer by literature, suggesting the biological meaning of the findings of the proposed method. The analysis of the driver groups predicted by *DriverGroup* in prognosis shows that the subtypes identified based on the predicted driver groups have significant prognostic values for survival analysis (i.e. P-values <0.05), indicating that the driver groups identified by *DriverGroup* may have important clinical implications for cancer treatment. We also apply *DriverGroup* to the study of synthetic lethality and miRNA driver groups of epithelial-mesenchymal transition (EMT). All the results show the potential of *DriverGroup* as a framework for studying molecular mechanisms of the progression of cancer.

2 Datasets and methods

2.1 Datasets

In this study, we use the BRCA dataset of TCGA (The Cancer Genome Atlas Research Network *et al.*, 2013). This dataset contains the expression data of miRNAs, TFs and mRNAs of tumour/normal samples. The tumour samples are used to identify edges of the gene network, and the normal samples are used to compute node weights. The TF list, which is used to detect which genes are TF genes in the expression dataset, is obtained from Lizio *et al.* (2017). We also use interaction data (i.e. target binding information), including PPIs (Vinayagam *et al.*, 2011), miRNA-TF/mRNA interactions [miRTarBase 6.1 (Chou *et al.*, 2016), TarBase 7.0 (Vlachos *et al.*, 2015), miRWalk 2.0 (Dweep and Gretz, 2015) and TargetScan 7.0 (Agarwal *et al.*, 2015)], and TF-miRNA interactions [TransmiR 2.0 (Wang *et al.*, 2010)] to refine the built gene network. In addition, to evaluate the performance of *DriverGroup* in detecting the influence of gene groups in a network, we use DREAM4 data obtained from the DREAM4 In Silico Network Challenge (Marbach *et al.*, 2010; Schaffter *et al.*, 2011). We also use the SynLethDB synthetic lethality database (Guo *et al.*, 2016) for identifying synthetic lethality, the EMT signatures (Tan *et al.*, 2014) and the EMT miRNAs (Cursons *et al.*, 2018) for discovering EMT driver groups. More details of these datasets will be introduced in the following sections. All these datasets are available at https://github.com/pvvhoang/DriverGroup.

2.2 Drivergroup

2.2.1 Overview

An overview of *DriverGroup*, the proposed method for identifying driver gene groups, is shown in Figure 1. *DriverGroup* includes three stages: (i) constructing the miRNA-TF-mRNA network, (ii) discovering critical nodes in the constructed network and (iii) identifying driver gene groups. Particularly, we firstly construct the network using the matched expression data of mRNAs, transcription factors (TFs) and miRNAs of a given cohort of cancer patients. Then the directed PPI network (Vinayagam *et al.*, 2011) and the target binding information are used to refine the network by removing those interactions not supported by these databases. Next, we discover critical nodes of the network by applying control theory (Kalman, 1963) and the Network Control method (Liu *et al.*, 2011). The critical nodes play a central role in controlling the whole network. Finally, based on the network and its critical nodes, we identify driver gene groups. The detail of *DriverGroup* is described in the following sections.

2.2.2 Procedure for identifying driver gene groups

Stage (1) Constructing the miRNA-TF-mRNA network. To identify driver gene groups, we detect groups of miRNAs and coding genes which jointly impact on the proliferation genes in a gene regulatory network. Since we evaluate both miRNA cancer driver groups and coding cancer driver groups, we construct the network which includes both miRNAs and coding genes (i.e. TFs and mRNAs). It is called the miRNA-TF-mRNA network in this article. In the first stage, we build the miRNA-TF-mRNA network through the three following steps.

- Step 1a: Prepare the miRNA/TF/mRNA expression data. We obtain the miRNA/TF/mRNA expression data of matched samples from the BRCA dataset (The Cancer Genome Atlas Research *et al.*, 2013). For coding genes, we select genes which are in the PPI network (Vinayagam *et al.*, 2011) or in the proliferation gene list. We select the PPI network as it contains a large amount of cancer driver genes and has been used to identify driver genes by Vinayagam *et al.* (2016). The proliferation genes are retrieved from the biological process of cell population proliferation (GO: 0008283). We then use the TF list to detect which genes are TFs in the selected genes. As a result, we have 5273 mRNAs and 850 TFs. For miRNAs, we select all 1719 miRNAs from the BRCA dataset. Finally, we extract the expression data of these 1719 miRNAs, 850 TFs and 5273 mRNAs for 747 tumours and 76 normal samples.
- Step 1b: Build the miRNA-TF-mRNA network. We firstly build a miRNA-TF-mRNA network based on the miRNA/TF/mRNA expression data of tumour samples. A node of the network is a miRNA, a TF, or a mRNA. The node weight is the absolute difference of average expression of that node between tumour and normal states. The node weight indicates the cost to change the state of a node from normal to tumour. The bigger the weight of a node is, the higher the cost required to change between the states. An edge between two nodes is added if the absolute Pearson correlation coefficient between them (calculated based on expression data) is larger than or equal to a threshold, which is the average of the absolute pairwise Pearson correlation coefficients of all node pairs. The edge weight is the absolute value of the correlation coefficient between the two nodes. Edge directions are determined according to the motif shown in Figure 2. Particularly, miRNAs can regulate TFs/mRNAs, TFs can regulate miRNAs/mRNAs and TFs/mRNAs can regulate other TFs/mRNAs, respectively.

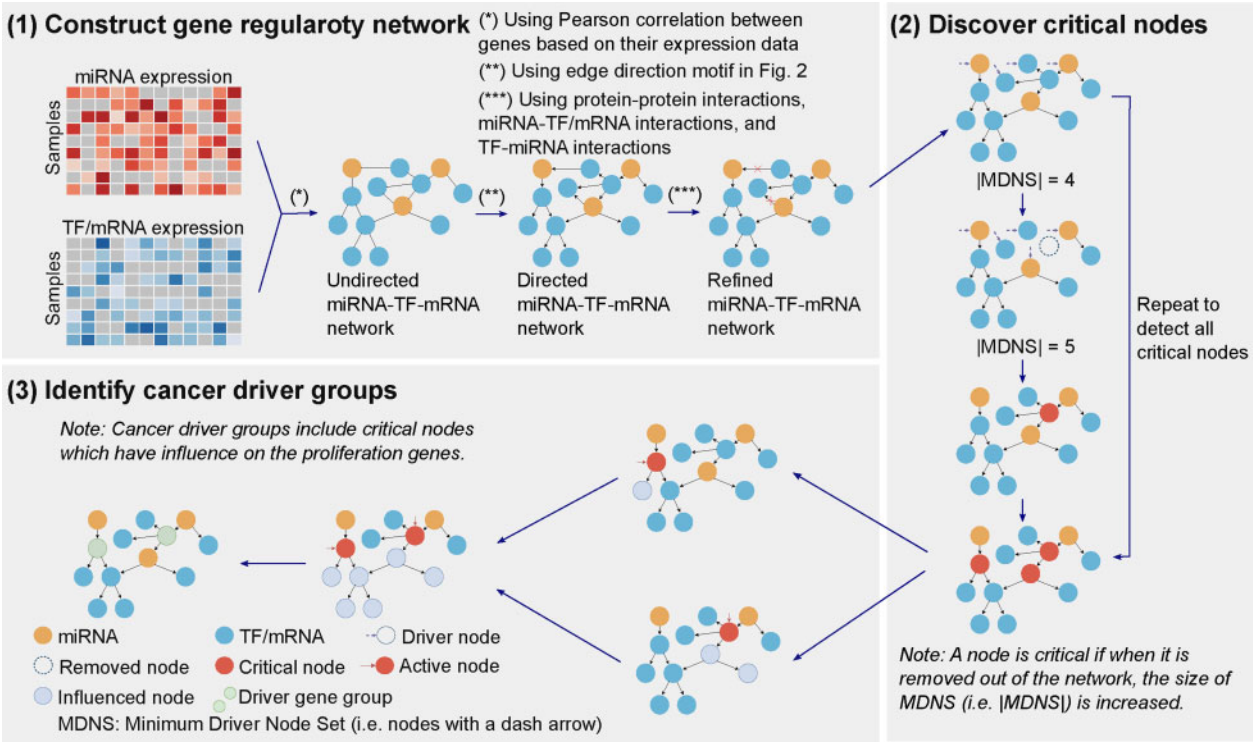

Fig. 1 An illustration of *DriverGroup*. (1) Build the gene regulatory network by combining the gene network constructed from the gene expression data with the protein-protein interactions and other existing databases, including miRTarBase 6.1, TarBase 7.0, miRWalk 2.0, TargetScan 7.0 and TransmiR 2.0, (2) Discover critical nodes by evaluating the increase of the size of Minimum Driver Node Set (MDNS) (i.e. the minimal set of nodes which can control the whole network) when a node is removed and (3) identify driver gene groups by detecting groups of critical nodes which have influence on the proliferation genes.

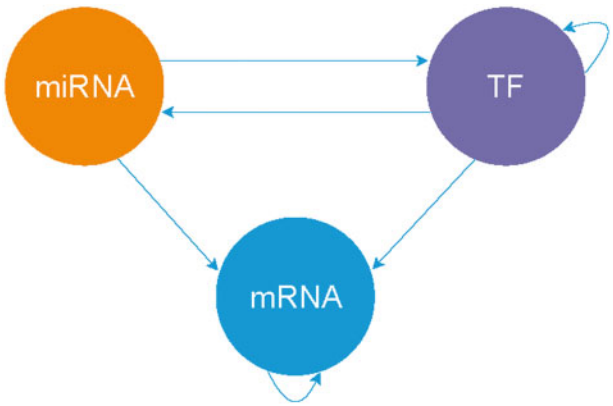

Fig. 2 Motif of the edge directions of the miRNA–TF–mRNA regulatory network. In the miRNA–TF–mRNA regulatory network, miRNAs can regulate TFs/mRNAs, TFs can regulate miRNAs/mRNAs, TFs/mRNAs can regulate other TFs/mRNAs, respectively.

- Step 1c: Refine the gene network. We use the PPIs to refine the expression network. If a TF–TF/mRNA or mRNA–mRNA interaction is in the expression network but not in the PPI network, we remove it from the expression network. We continue to refine the obtained network by using the existing databases. Particularly, TF–miRNA interactions are refined with TransmiR, and miRNA–TF/mRNA interactions are refined with miRTarBase, TarBase, miRWalk and TargetScan. Since miRTarBase and TarBase include experimentally validated miRNA–target interaction information, they help to retain true miRNA–target interactions, but they may have false negatives. Thus, we also use miRWalk and TargetScan, which include predicted miRNA–target interaction information, to obtain potential interactions which may not be included in miRTarBase and TarBase. Because the obtained network is based on both the expression data of a particular cancer type and the existing databases, it is more reliable and specific to that cancer type. The final network includes 7842 nodes (1719 miRNAs, 850 TFs and 5273 mRNAs) and 171 459 edges (23 037 miRNA–TF, 105 019 miRNA–mRNA, 30 096 TF–miRNA, 1235 TF–TF, 815 TF-mRNA and 11 257 mRNA–mRNA) (see Section 3 of the Supplementary Material for the numbers of edges which are dropped in different edge types during the refinement).

Stage (2) Discovering critical nodes in the network. According to the Network Control method (Liu *et al.*, 2011), any network can be controlled fully by a minimum set of nodes of the network, called a Minimum Driver Node Set (MDNS). The detail of the Network Control method is discussed in Section 2.2.3. Applying this Network Control idea, we discover a MDNS of the miRNA–TF–mRNA network built in Stage (1) above. Based on the discovered MDNS, we then detect critical nodes of the network. A critical node is a node whose absence increases the size of the MDNS. In other words, when a critical node is removed from the network, we need a bigger MDNS to fully control the network. Thus, critical nodes play the central role in the network, and we consider them as members of potential driver gene groups. This stage is illustrated in Part (2) in Figure 1.

Stage (3) Identifying driver gene groups. In the last stage, we identify driver groups with the steps below.

- Step 3a: Estimate the influence of groups of critical nodes on the proliferation genes. This step includes the following two substeps.
  1. Form k-way combinations of the selected critical nodes. As we aim to detect nodes which have high influence on the

proliferation genes in the network, we focus on nodes with higher out degrees. Out degree of a node is the number of edges going out from that node. We firstly rank critical nodes of the miRNA–TF–mRNA network in descending order of node out degree. We select top n nodes from the ranked list then define k-way combinations of these top n nodes ($k \in \{1, \ldots, n\}$).

2. Evaluate influence of the k-way combinations on the proliferation genes. Influence is indicated by the number of proliferation nodes. Adopting the idea of Influence Maximization (IM) (Gong *et al.*, 2016; Yang *et al.*, 2016b) for detecting k-seed sets having the maximum impact in a network, we propose a novel algorithm to assess the impact of a group of critical nodes on the proliferation genes. The detail of the proposed algorithm is discussed in Section 2.2.4. Before using the proposed algorithm to evaluate the influence of the k-way combinations, we firstly normalize the node weights and the edge weights of the network so that the weight of a node and the total weight of edges going into a node are in the range from 0 to 1 to make possible to compare the weights in the algorithm. The normalized weight of a node is equal to the original node weight divided by the largest node weight. For edge weights, we firstly find each node's total incoming edge weight, then find the largest among all these total weights. We normalize an edge weight by dividing it by the largest total weight found. We apply the proposed algorithm to evaluate the influence of each of the k-way combinations of critical nodes. The output of this step is the number of proliferation nodes for each k-way combination ($k \in \{1, \ldots, n\}$).

- Step 3b: Identify driver gene groups. In this step, we identify the maximal k-way combinations and regard the identified maximal combinations are the driver gene groups. A k-way combination g ($k \in \{1, \ldots, n\}$) is maximal if the $(k+1)$-way combination obtained by adding to g a critical node has the same or lower influence than g. More details are in Section 2.2.4.

2.2.3 Controllability of complex networks

According to the Network Control method (Liu *et al.*, 2011), any directed network can be controlled by a subset of nodes in the network, known as driver nodes of the network. The method to identify driver nodes is described as follows.

Suppose that we have a directed network with N nodes $x_1, \ldots, x_N$. The matrix $A_{N\times N}$ which captures the interaction strength between nodes can be represented as:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{pmatrix}, \quad (1)$$

where a_{ij} indicates the interaction strength of node j on node i ($i, j \in \{1, \ldots, N\}$), and a_{ij} is 0 if there is not an edge from j to i.

Let the matrix $B_{N\times M}$ represent the interaction of an external controller on M nodes ($M \leq N$) in the network:

$$B = \begin{pmatrix} b_1 & 0 & \cdots & 0 \\ 0 & b_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & b_M \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}, \quad (2)$$

where b_i is the strength of the interaction between the external controller and node i ($i \in \{1, \ldots, M\}$) in the network.

Let $C_{N\times NM}$ be the controllability matrix:

$$C = (B, AB, A^2B, \ldots, A^{N-1}B). \quad (3)$$

According to Kalman's controllability condition (Kalman, 1963), the network represented by matrix A is controllable through the M nodes in matrix B if the controllability matrix C satisfies the condition:

$$rank(C) = N. \quad (4)$$

The M nodes are called driver nodes of the network. Intuitively, the rank of the controllability matrix C being N implies that all N nodes of the network can be controlled.

Given the network, we may discover various sets of nodes which satisfy the condition 4, i.e. a network can have multiple sets of driver nodes. In this article, we focus on the driver node set with the smallest number of driver nodes, called the Minimum Driver Node Set (MDNS). In Stage (2) of *DriverGroup*, we use the condition 4 to detect the MDNS of the miRNA–TF–mRNA network. We then identify critical nodes of the network by removing one node at a time from the network, and if the MDNS of the network with the node removed is bigger than the MDNS of the original network, the removed node is a critical node.

2.2.4 Influence of groups of nodes

Influence maximization (IM) finds a k-seed set that has the maximum influence in a network (Gong *et al.*, 2016; Yang *et al.*, 2016b). IM is usually used to identify influential users in online social networks (Kempe *et al.*, 2003) as described below.

Given a network G with N nodes and a budget k, IM is to find a set S containing k nodes of G (called a k-seed set) which maximizes the influence spread over G. The influence spread is the number of nodes influenced by a k-seed set and it is denoted as $\sigma(S)$. That is:

$$S = \underset{|S|=k}{argmax}(\sigma(S)). \quad (5)$$

Inspired by IM, we propose the following described method to calculate the influence spread of a k-way combination of critical nodes on the proliferation genes in a miRNA–TF–mRNA network, and then find the maximal combinations as driver groups. In our problem, we do not fix the budget k and we evaluate the influence of k-way combinations on the proliferation genes instead of over the whole network.

In general, diffusion models are used to resolve the IM problem (Gong *et al.*, 2016). The diffusion models identify the influence spread of a k-seed set over a network by considering that nodes in the k-seed set are active and proposing strategies to activate other nodes in the network. The larger the number of active nodes a k-seed set creates, the more influence it has in the network. These models employ the rules below:

- A node can be active or inactive.
- During the diffusion process, inactive nodes can be activated but active nodes cannot be inactivated.
- The process terminates if no more nodes can be activated.

Independent cascade (IC) and linear threshold (LT) are two popular diffusion models (Kempe *et al.*, 2003; Ko *et al.*, 2018). With IC, a node is activated based on the active neighbouring nodes *independently* (i.e. considering the effect of each edge on the node separately). On the other hand, with LT, each node has a threshold (i.e. node weight) and it is activated if the *sum* of the weights of the edges pointing from its active neighbour nodes to this node is larger than its threshold.

To evaluate the collaboration of nodes in a network, we propose a novel algorithm to evaluate the influence of a k-seed set (i.e. k-way combination of critical nodes in our problem) on a target set (i.e. the proliferation genes) based on LT. Instead of identifying the influence spread of a k-seed set over the whole network as for solving the IM problem, we compute the influence of a k-way combination of critical nodes on a particular set of nodes in the network (i.e. all the

proliferation genes in the constructed network). The detailed algorithm is illustrated in Algorithm 1 in Section 1 of the Supplementary Material.

After applying Algorithm 1 to get the influence of k-way combinations of the top n critical nodes selected in Step 3a of Stage (3) on the proliferation genes, we rank the k-way combinations in descending order of their influence. We then use the other proposed algorithm (Algorithm 2) to retain only the maximal combinations. The detail of Algorithm 2 is shown in Section 2 of the Supplementary Material.

2.2.5 Algorithms

We have developed two algorithms: Algorithm 1 for evaluating the influence of a k-seed set on a target set and Algorithm 2 for refining k-way combinations. The details of these two algorithms are in Section 1 and Section 2 of the Supplementary Material, respectively.

2.2.6 Implementation

The R source code of the implementation and scripts to reproduce the experiments are available at https://github.com/pvvhoang/DriverGroup.

3 Results

Due to the lack of the ground truth for predicted driver gene groups, we have used several strategies to evaluate *DriverGroup*. We assess the ability of *DriverGroup* in discovering regulatory effects of gene groups in a network in Section 3.1. We evaluate the performance of *DriverGroup* in discovering driver gene groups in Section 3.2. We analyse biological implications of the predicted driver groups by using them in prognosis analysis (Section 3.3) and analysing their target genes (Section 3.4). We also use *DriverGroup* to study synthetic lethality (Section 3.5) and miRNA driver groups of EMT (Section 3.6).

3.1 *Drivergroup* is effective in detecting group-based regulatory effects

Before evaluating the performance of *DriverGroup* in detecting cancer driver groups, we firstly assess its performance in detecting the regulation of gene groups in a network, a key step (Step 3a) of *DriverGroup*. We use the DREAM4 data obtained from the DREAM4 In Silico Network Challenge (Marbach *et al.*, 2010; Schaffter *et al.*, 2011). The dataset includes five subsets and each subset contains the data of 100 genes, including wild-type data, knockout data (considered as expression data), dual knockout index data (i.e. indexes of 20 gene pairs which are knocked out simultaneously), dual knockout data (i.e. expression data corresponding to dual knockout index data) and network data (The detailed experiment setting is described in Section 4 of the Supplementary Material). Given each of the five sub datasets, we can identify the list of genes affected by the 20 knocked out gene pairs in the network and they are considered as the gold standard of the experiment.

For each of the five sub datasets, we are looking at whether each method can find the targets of each knocked out pairs. We compare *DriverGroup* with jointIDA (Nandy *et al.*, 2017) and the random method. jointIDA is also used to estimate the joint effects of a group of variables on other variables. However, it estimates the joint effects of variables on a target by knocking down all variables at the same time. In the random method, we randomly pick target genes for each the knocked out gene pairs 100 times. We validate the results of each method with the gold standard above. Figure 3 shows the precisions achieved by the three methods.

In Figure 3, we see that *DriverGroup* outperforms jointIDA in four out of the five cases and achieves similar precision as jointIDA in the case of network 5. Both *DriverGroup* and jointIDA outperform the random method in all cases.

To have a detailed evaluation, we compare the results of the 3 methods (i.e. jointIDA, *DriverGroup*, and the random method) and

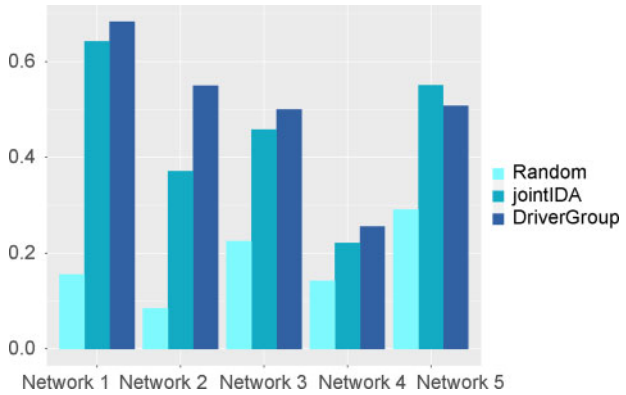

Fig. 3 Comparison of precision for the target genes predicted by the random method, jointIDA and *DriverGroup* in five networks. The target genes predicted by each method are validated against the gold standard. Each bar indicates the *Precision* of each method.

the combination of jointIDA and *DriverGroup* (i.e. jointIDA_DriverGroup) for all the 20 gene pairs of the 5 networks. For each network, we validate the predicted target genes of the knocked-out gene pairs against the gold standard. We then compute the accumulated number of validated target genes of all the 20 knocked-out gene pairs. The result is shown in Figure 4. We can see that *DriverGroup* outperforms the random method and jointIDA in the first four networks and it is comparable to jointIDA in the fifth network. Furthermore, the combination of jointIDA and *DriverGroup* outperforms the other three methods in all the cases.

In addition, the overlap of the gold standard and the target genes predicted by jointIDA and *DriverGroup* is shown in Figure 5. In the figure, the target genes identified by jointIDA and *DriverGroup* of all 20 gene pairs of each network are validated against the gold standard. In all the five networks, although there are some target genes uncovered by both jointIDA and *DriverGroup*, there are a large amount of validated target genes discovered only by *DriverGroup*. Since the results of the two methods are complementary, it would be beneficial if they could be used together in predicting targets of groups of genes.

3.2 Identifying driver groups

We apply *DriverGroup* to the BRCA dataset to identify driver groups (i.e. groups of coding RNAs/miRNAs which have an impact on the proliferation genes). We also categorize the identified groups into additional groups and enhanced groups. Additional groups regulate target genes which are in the union of the target genes of individuals in the groups. Enhanced groups regulate genes in and outside the union of the target genes of individuals in the groups. We identify 82 coding cancer driver groups and 36 miRNA cancer driver groups. We sort these groups based on their influence on the proliferation genes (i.e. The larger number of proliferation genes a group impacts on, the higher it is in the ranking list). The top 10 driver groups discovered by our method are presented in Table 1 for coding genes and Table 2 for miRNAs. We see that most of the identified groups are enhanced groups, indicating that members in the identified groups work collaborately to increase the effects on the proliferation genes.

The driver groups predicted by *DriverGroup* are promising as some members of the predicted groups are confirmed to be related to breast cancer. Among the genes in the top 10 coding cancer driver groups predicted by *DriverGroup*, GATA1, TCF3, JUN and MYB are in the Cancer Gene Census (CGC) from the COSMIC database (Forbes *et al.*, 2015). In addition, other genes, including FOS, MBD3, E2F6 and SPI1, are also previously proved to be related to breast cancer. Specifically, FOS is critical to the growth of MCF-7 breast cancer cell (Lu *et al.*, 2005) and its family plays an important role in the biological function of breast tumours (Langer *et al.*, 2006). There is a relationship between MBD3 and human breast

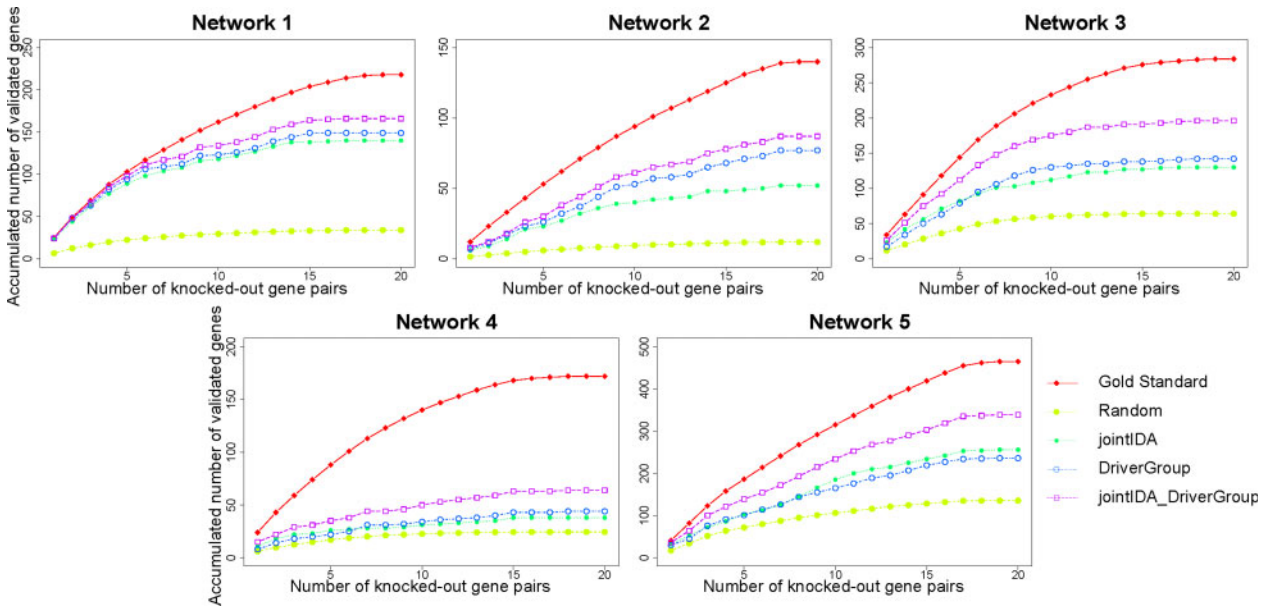

Fig. 4 Comparison of performance of the random method, jointIDA, *DriverGroup*, the combination of jointIDA and *DriverGroup* (i.e. jointIDA_DriverGroup). There are five networks in total and each chart shows the results for a network. In each chart, the x-axis indicates the number of knocked-out gene pairs. The y-axis is the accumulated number of validated genes predicted by the random method, jointIDA, *DriverGroup*, the combination of jointIDA and *DriverGroup*. The red line is the gold standard and it shows the true numbers of genes affected by gene pairs. In the first four cases, *DriverGroup* outperforms the random method and jointIDA, and it is comparable to jointIDA in the last case. Furthermore, the combination of jointIDA and *DriverGroup* outperforms the other three methods in all the cases.

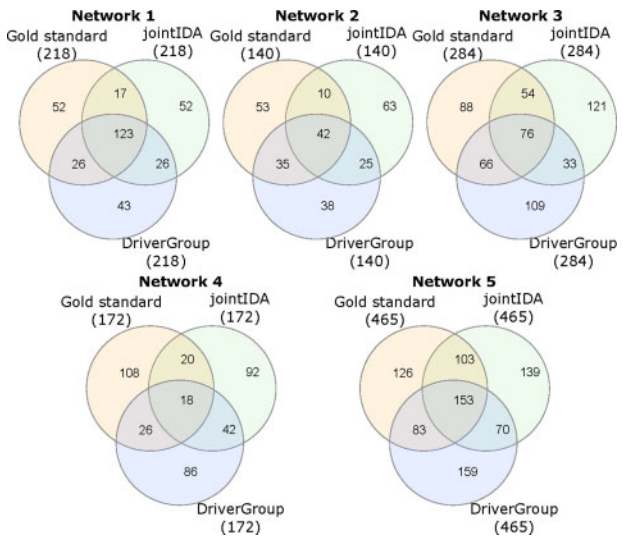

Fig. 5 Overlap between jointIDA, *DriverGroup* and the gold standard. The diagram shows the overlap of the gold standard and the target genes predicted by jointIDA and *DriverGroup* in the five networks. In all the five networks, *DriverGroup* can detect large amounts of target genes which are not discovered by jointIDA.

Table 1 Coding BRCA driver groups predicted by *DriverGroup*

Group	Predicted driver groups	Size of group	Type
1	FOS, MBD3, JUN, E2F6, MYB, SPI1	6	Enhanced
2	GATA1, FOS, MBD3, JUN, MYB, SPI1	6	Enhanced
3	TCF3, FOS, MBD3, JUN, MYB, SPI1	6	Enhanced
4	GATA1, FOS, MBD3, JUN, SPI1	5	Enhanced
5	GATA1, TCF3, FOS, MBD3, JUN, MYB	6	Enhanced
6	GATA1, FOS, MBD3, SPI1	4	Enhanced
7	FOS, MBD3, JUN, SPI1	4	Enhanced
8	TCF3, FOS, MBD3, JUN, E2F6, MYB	6	Enhanced
9	GATA1, MBD3, JUN, MYB	4	Enhanced
10	MBD3, JUN, MYB, SPI1	4	Enhanced

The top 10 coding driver groups are enhanced groups whose members work in concert to increase the influence on the proliferation genes.

cancer cells (Shimbo *et al.*, 2016). E2F6 regulates BRCA1 negatively in human cancer cells (Oberley *et al.*, 2003) and SPI1 can be used for prognosis in breast cancer (Wang *et al.*, 2007).

In addition, to see the reason why the driver groups predicted by *DriverGroup* may cause cancer, we evaluate the genomic aberrations of genes in these driver groups. Interestingly, most of genes in these driver groups are mutated in breast cancer patients. For instance, genes MYB, SPI1, E2F6 and GATA1 are mutated in 12, 3, 2 and 2 patients in the BRCA data, respectively. Furthermore, both gene E2F6 and gene SPI1 in the first predicted driver group are mutated in patient TCGA-AN-A046, both gene GATA1 and gene SPI1 in the second, fourth and sixth predicted driver groups are mutated in patient TCGA-A8-A09Z. These findings indicate that the predicted driver groups may play a significant role in developing the disease in breast cancer patients.

Among the miRNAs in the top 10 miRNA cancer driver groups predicted by *DriverGroup*, there are 3 miRNAs (hsa-miR-22-5p, hsa-miR-342-5p and hsa-miR-34a-5p) involved in tumourigenesis of breast cancer, which are confirmed by OncomiR (Wong *et al.*, 2018), a database for studying pan-cancer miRNA dysregulation. Out of these three miRNAs, hsa-miR-342-5p is proved to be a regulator of the development of breast cancer cells in another work (Lindholm *et al.*, 2019) as well. Another three miRNAs, hsa-miR-130a-5p, hsa-miR-146a-5p and hsa-miR-223-5p, are also confirmed to be related to breast cancer. Specifically, hsa-miR-130a-5p targets FOSL and upregulates ZO-1 to suppress breast cancer cell migration (Chen *et al.*, 2018), hsa-miR-146a-5p has an over expression in breast cancer cells (Sandhu *et al.*, 2014), and hsa-miR-223-5p is a coordinator of breast cancer (Pinatel *et al.*, 2014).

Table 2 miRNA BRCA driver groups predicted by *DriverGroup*

Group	Predicted driver groups	Size of group	Type
1	hsa-miR-96-5p, hsa-miR-342-5p	2	Enhanced
2	hsa-miR-130a-5p, hsa-miR-34a-5p, hsa-miR-22-5p, hsa-miR-222-5p, hsa-miR-223-5p	5	Enhanced
3	hsa-miR-130a-5p, hsa-miR-22-5p, hsa-miR-222-5p, hsa-miR-223-5p, hsa-miR-6797-5p	5	Enhanced
4	hsa-miR-130a-5p, hsa-miR-34a-5p, hsa-miR-22-5p	3	Enhanced
5	hsa-miR-130a-5p, hsa-miR-22-5p, hsa-miR-146a-5p	3	Enhanced
6	hsa-miR-130a-5p, hsa-miR-22-5p, hsa-miR-6797-5p	3	Enhanced
7	hsa-miR-34a-5p, hsa-miR-22-5p, hsa-miR-222-5p	3	Enhanced
8	hsa-miR-34a-5p, hsa-miR-22-5p, hsa-miR-223-5p	3	Enhanced
9	hsa-miR-34a-5p, hsa-miR-22-5p, hsa-miR-6797-5p	3	Enhanced
10	hsa-miR-130a-5p, hsa-miR-34a-5p, hsa-miR-222-5p, hsa-miR-342-5p, hsa-miR-6797-5p	5	Enhanced

The top 10 miRNA driver groups are enhanced groups whose members work in concert to increase the influence on the proliferation genes.

3.3 Predicted driver groups are useful in predicting survival

As the predicted driver groups likely cause carcinogenesis, they could be promising biomarkers for tumour classification. To explore this concept, we use the driver gene groups predicted by *DriverGroup* to stratify breast cancer patients. We obtain the BRCA gene expression data from Zhang et al. (2019), which includes clinical data, for survival analysis. We use the first predicted coding driver groups in Table 1, including FOS, MBD3, JUN, E2F6, MYB and SPI1, and the Similarity Network Fusion (SNF) method (Xu *et al.*, 2017; Wang *et al.*, 2014) to cluster cancer patients (see Section 5 of the Supplementary Material for the results with the second and the third driver groups). SNF takes expression of these genes (i.e. 6 genes in this case) as input and outputs subtypes of cancer patients. We then evaluate the survival outcomes of patients in the classified subtypes. The results show that the survivals of patients in different subtypes are significantly different (P-value = 0.0152) as in Figure 6. In addition, the clustering display indicates the similarity of samples in each subtype and the silhouette plot shows a high quality clustering with a large average silhouette width (i.e. 0.77).

3.4 Members of predicted driver groups regulating common target genes

To see the functional association among the members of driver groups predicted by *DriverGroup*, we check if they regulate common target genes. We use the TransmiR database of TF–miRNA interactions to identify target genes of the members of predicted coding driver groups and use the miRTarBase, TarBase, miRWalk and TargetScan databases of miRNA–TF/mRNA interactions to identify target genes of predicted miRNA driver groups. We observe that for the top 10 predicted driver groups in the both cases of coding and non-coding, all participants in each group regulate some common target genes, indicating the functional link of the members in driver groups identified by our proposed method.

3.5 Detecting synthetic lethality with *DriverGroup*

Two genes have a synthetic lethal (SL) interaction if the perturbation of both genes simultaneously is lethal but a perturbation that affects either gene alone is viable (Lord *et al.*, 2015; O'Neil *et al.*, 2017). It means that in cancer patients, the collaboration of two genes in a SL interaction results in the loss of viability. To validate the ability of *DriverGroup* in discovering SL interaction, we apply it to BRCA data to detect only the driver gene groups of size 2. Since the existing synthetic lethality database SynLethDB (Guo *et al.*, 2016) only includes SL coding genes, we apply *DriverGroup* to identify coding driver groups only in this case. We validate the top 1000 predicted SL gene pairs against SynLethDB, and 6 of them have been confirmed by SynLethDB, which are NFKB1-TP53, FOS-MAPK1, CASP3-JUN, JUN-MAPK1, JUN-SMAD3 and E2F3-RB1. Based on the hypergeometric test, the overlap between the predicted SL gene pairs and the gold standard is significant, with a P-value of 0.00027. Furthermore, most of these genes are reported to be related to breast cancer, including NFKB1 (Kim *et al.*, 2018), TP53 (Ungerleider *et al.*, 2018), FOS (Lu *et al.*, 2005; Langer *et al.*, 2006), JUN (Langer *et al.*, 2006), SMAD3 (Petersen *et al.*, 2010), E2F3 (Lee *et al.*, 2015) and RB1 (Jones *et al.*, 2016).

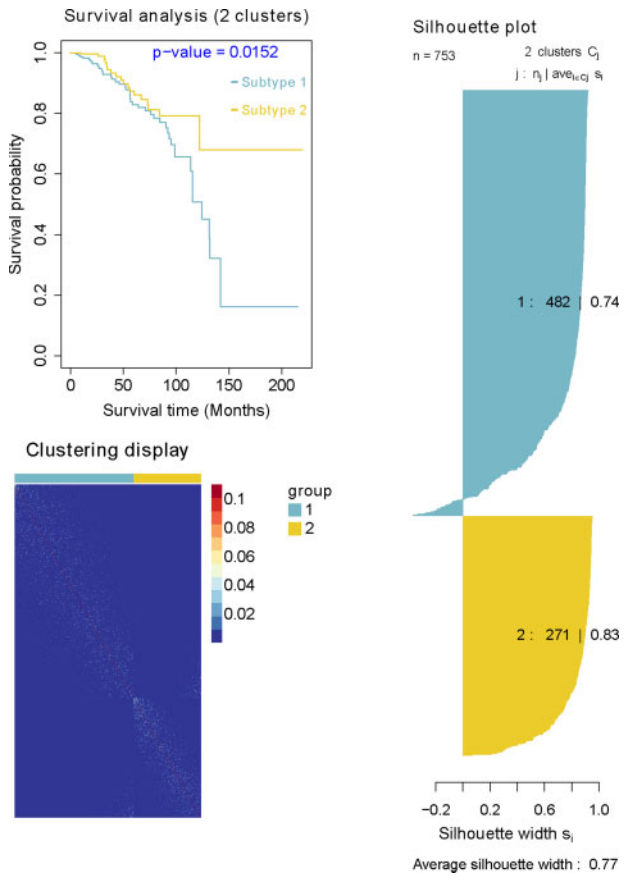

Fig. 6 Survival curves, silhouette plot and clustering display. Survival curves, silhouette plot and clustering display of cancer subtypes identified by using *the first predicted coding driver groups* (*including FOS, MBD3, JUN, E2F6, MYB and SPI1*) indicate that the survivals of patients are significantly different in the two subtypes and the clustering is highly qualified with a large average silhouette width and the similarity of samples in each subtype.

3.6 Detecting driver groups of EMT

Metastasis is a process where cancer cells migrate from the primary tumour to distant locations in the body. It is the major cause of death of cancer patients. EMT is one of the processes which create these metastatic cells (Park *et al.*, 2008). EMT is promoted by coding genes (Lee *et al.*, 2018) and/or non-coding genes (Gregory *et al.*, 2008). In this section, we apply *DriverGroup* to the BRCA dataset to discover driver groups for the EMT of breast cancer patients by identifying miRNA groups which have maximum influence on the EMT signatures (Tan *et al.*, 2014). As *DriverGroup* detects miRNA

Table 3 miRNA driver groups of EMT predicted by *DriverGroup*.

Group	Predicted driver groups	Size of group
1	hsa-miR-130a-5p, hsa-miR-34a-5p, hsa-miR-22-5p, hsa-miR-223-5p, hsa-miR-99a-5p	5
2	hsa-miR-130a-5p, hsa-miR-34a-5p, hsa-miR-22-5p, hsa-miR-99a-5p	4
3	hsa-miR-130a-5p, hsa-miR-34a-5p, hsa-miR-22-5p	3
4	hsa-miR-130a-5p, hsa-miR-22-5p, hsa-miR-223-5p, hsa-miR-99a-5p	4
5	hsa-miR-130a-5p, hsa-miR-34a-5p, hsa-miR-222-5p, hsa-miR-223-5p, hsa-miR-99a-5p	5
6	hsa-miR-130a-5p, hsa-miR-34a-5p, hsa-miR-223-5p, hsa-miR-99a-5p, hsa-miR-6797-5p	5
7	hsa-miR-130a-5p, hsa-miR-22-5p, hsa-miR-222-5p, hsa-miR-342-5p, hsa-miR-99a-5p	5
8	hsa-miR-130a-5p, hsa-miR-22-5p, hsa-miR-99a-5p	3
9	hsa-miR-130a-5p, hsa-miR-34a-5p, hsa-miR-342-5p, hsa-miR-99a-5p	4
10	hsa-miR-130a-5p, hsa-miR-22-5p, hsa-miR-222-5p, hsa-miR-342-5p	4

groups which regulate EMT signatures, the detected miRNA groups are expected to drive the EMT transition in breast cancer patients. We identify 61 miRNA driver groups for EMT and we sort these groups based on their influence on the EMT signatures (i.e. The larger number of EMT signatures a group impacts on, the higher it is in the ranking list). The list of top 10 miRNA driver groups for EMT in breast cancer is shown in Table 3. Among these miRNAs, hsa-miR-130a-5p and hsa-miR-223-5p are EMT miRNAs (Cursons *et al.*, 2018), indicating the potential of *DriverGroup* in detecting driver groups for different biological processes such as EMT.

4 Conclusion

Since there is evidence showing that genes work in concert to regulate targets and progress cancer, several methods have been developed to identify these genes. However, current methods only discover mutated modules. Only one mutated gene in a module is sufficient to progress cancer. Thus, members in these mutated modules do not collaborate in driving cancer and these mutated modules are not truly driver gene groups. In addition, current methods only identify coding drivers while non-coding genes can also regulate targets to drive cancer. Therefore, novel methods are required to identify driver gene groups to elucidate their regulatory mechanism.

To overcome the limitations of existing methods, in this article, we have developed a novel method, *DriverGroup*, to uncover driver groups. We have evaluated the effectiveness of *DriverGroup* with various experiments. The results have demonstrated that *DriverGroup* can explore promising driver gene groups. Predicted coding driver groups can be used to classify cancer patients into subtypes and the survivals of patients in different subtypes are significantly different. Furthermore, *DriverGroup* can also detect synthetic lethal gene pairs and EMT driver groups. All these results show that the findings of *DriverGroup* can provide new insights into molecular regulatory mechanisms of cancer initialization and progression, and *DriverGroup* has the potential to contribute to the development of effective cancer treatments.

As a future work, to improve *DriverGroup*, we will consider the role of other ncRNAs, e.g. long non-coding RNAs, and their sponge activities with miRNAs in developing cancer. We also plan to apply *DriverGroup* to the study of multiple cancer types by using the predicted driver groups for subtype classification and survival analysis.

Acknowledgements

This research is supported by the Australian Government Research Training Program (RTP) Scholarship and the Vice Chancellor & President's Scholarship offered by the University of South Australia.

Funding

This work has been supported by the ARC DECRA [200100200] and the Australian Research Council Discovery [DP170101306].

Conflict of Interest: The authors have declared that no conflict of interests exist.

References

Agarwal,V. *et al.* (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, e05005.
Chen,X. *et al.* (2018) MicroRNA-130a suppresses breast cancer cell migration and invasion by targeting FOSL1 and upregulating ZO-1. *J. Cell Biochem.*, **119**, 4945–4956.
Chou,C.H. *et al.* (2016) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.*, **44**, D239–D247.
Ciriello,G. *et al.* (2012) Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.*, **22**, 398–406.
Cursons,J. *et al.* (2018) Combinatorial targeting by microRNAs co-ordinates post-transcriptional control of EMT. *Cell Syst.*, **7**, 77–91.e7.
Dweep,H. and Gretz,N. (2015) miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nat. Methods*, **12**, 697–697.
Feitelson,M.A. *et al.* (2015) Sustained proliferation in cancer: mechanisms and novel therapeutic targets. *Semin. Cancer Biol.*, **35**, S25–S54. **Suppl**
Forbes,S.A. *et al.* (2015) Cosmic: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.
Gong,M. *et al.* (2016) Influence maximization in social networks based on discrete particle swarm optimization. *Inf. Sci.*, **367-368**, 600–614.
Gonzalez-Perez,A. and Lopez-Bigas,N. (2012) Functional impact bias reveals cancer drivers. *Nucleic Acids Res.*, **40**, e169.
Gregory,P.A. *et al.* (2008) The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat. Cell Biol.*, **10**, 593–601.
Guo,J. *et al.* (2016) SynLethDB: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. *Nucleic Acids Res.*, **44**, D1011–D1017.
Hou,J.P. and Ma,J. (2014) Dawnrank: discovering personalized driver genes in cancer. *Genome Med.*, **6**, 56.
Jones,R.A. *et al.* (2016) RB1 deficiency in triple-negative breast cancer induces mitochondrial protein translation. *J. Clin. Investig.*, **126**, 3739–3757.
Kalman,R. (1963) Mathematical description of linear dynamical systems. *J. Soc. Indus. Appl. Math. Ser. A*, **1**, 152–192.
Karim,S.M.M. *et al.* (2016) Identification of miRNA-mRNA regulatory modules by exploring collective group relationships. *BMC Genomics*, **17**, 7.
Kempe,D. *et al.* (2003). Maximizing the spread of influence through a social network. In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, p.137–146. ACM, New York, NY, USA.
Kim,G.-C. *et al.* (2018) Upregulation of Ets1 expression by NFATc2 and NFKB1/RELA promotes breast cancer cell invasiveness. *Oncogenesis*, **7**, 91.
Kim,Y.A. *et al.* (2017) WeSME: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics*, **33**, 814–821.
Ko,Y.-Y. *et al.* (2018) Influence maximisation in social networks: a target-oriented estimation. *J. Inf. Sci.*, **44**, 671–682.
Langer,S. *et al.* (2006) Jun and Fos family protein expression in human breast cancer: correlation of protein expression and clinicopathological parameters. *Eur. J. Gynaecol. Oncol.*, **27**, 345–352.
Lee,G.H. *et al.* (2018) FYN promotes mesenchymal phenotypes of basal type breast cancer cells through STAT5/NOTCH2 signaling node. *Oncogene*, **37**, 1857–1868.
Lee,M. *et al.* (2015) Silencing of E2F3 suppresses tumor growth of Her2+ breast cancer cells by restricting mitosis. *Oncotarget*, **6**, 37316–37334.
Li,P. *et al.* (2018) Proliferation genes in lung development associated with the prognosis of lung adenocarcinoma but not squamous cell carcinoma. *Cancer Sci.*, **109**, 308–316.

Lindholm,E. *et al.* (2019) miR-342-5p as a potential regulator of HER2 breast cancer cell growth. *Microrna*, **8**, 155–165.
Liu,Y.-Y. *et al.* (2011) Controllability of complex networks. *Nature*, **473**, 167–173.
Lizio,M. *et al.* (2017) Update of the FANTOM web resource: high resolution transcriptome of diverse cell types in mammals. *Nucleic Acids Res.*, **45**, D737–d743.
Lopez-Saez,J.F. *et al.* (1998) Cell proliferation and cancer. *Histol. Histopathol.*, **13**, 1197–1214.
Lord,C.J. *et al.* (2015) Synthetic lethality and cancer therapy: lessons learned from the development of PARP inhibitors. *Annu. Rev. Med.*, **66**, 455–470.
Lu,C. *et al.* (2005) cFos is critical for MCF-7 breast cancer cell growth. *Oncogene*, **24**, 6516–6524.
Marbach,D. *et al.* (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. USA*, **107**, 6286–6291.
Nandy,P. *et al.* (2017) Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *Ann. Stat.*, **45**, 647–674.
Oberley,M.J. *et al.* (2003) E2F6 negatively regulates BRCA1 in human cancer cells without methylation of histone H3 on lysine 9. *J. Biol. Chem.*, **278**, 42466–42476.
O'Neil,N.J. *et al.* (2017) Synthetic lethality and cancer. *Nat. Rev. Genet.*, **18**, 613.
Park,S.M. *et al.* (2008) The miR-200 family determines the epithelial phenotype of cancer cells by targeting the e-cadherin repressors ZEB1 and ZEB2. *Genes Dev.*, **22**, 894–907.
Petersen,M. *et al.* (2010) Smad2 and smad3 have opposing roles in breast cancer bone metastasis by differentially affecting tumor angiogenesis. *Oncogene*, **29**, 1351–1361.
Pham,V.V.H. *et al.* (2019) CBNA: a control theory based method for identifying coding and non-coding cancer drivers. *PLoS Comput. Biol.*, **15**, e1007538.
Pinatel,E.M. *et al.* (2014) miR-223 is a coordinator of breast cancer progression as revealed by bioinformatics predictions. *PLoS One*, **9**, e84859.
Puente,X.S. *et al.* (2015) Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*, **526**, 519–524.
Reimand,J. and Bader,G.D. (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.*, **9**, 637–637.
Sandhu,R. *et al.* (2014) Overexpression of miR-146a in basal-like breast cancer cells confers enhanced tumorigenic potential in association with altered p53 status. *Carcinogenesis*, **35**, 2567–2575.
Schaffter,T. *et al.* (2011) Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, **27**, 2263–2270.
Shimbo,T. *et al.* (2016) High-quality ChIP-seq analysis of MBD3 in human breast cancer cells. *Genomics Data*, **7**, 173–174.
Tamborero,D. *et al.* (2013) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, **29**, 2238–2244.
Tan,T.Z. *et al.* (2014) Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol. Med.*, **6**, 1279–1293.
The Cancer Genome Atlas Research Network *et al.* (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113.
Ungerleider,N.A. *et al.* (2018) Breast cancer survival predicted by TP53 mutation status differs markedly depending on treatment. *Breast Cancer Res.*, **20**, 115.
Vinayagam,A. *et al.* (2011) A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal*, **4**, rs8–rs8.
Vinayagam,A. *et al.* (2016) Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proc. Natl. Acad. Sci. USA*, **113**, 4976–4981.
Vlachos,I.S. *et al.* (2015) DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res.*, **43**, D153–D159.
Wang,B. *et al.* (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.
Wang,J. *et al.* (2010) TransmiR: a transcription factor-microRNA regulation database. *Nucleic Acids Res.*, **38**, D119–22.
Wang,X.B. *et al.* (2007) Expression and prognostic value of transcriptional factor sp1 in breast cancer. *Ai Zheng*, **26**, 996–1000.
Weinhold,N. *et al.* (2014) Genome-wide analysis of non-coding regulatory mutations in cancer. *Nat. Genet.*, **46**, 1160–1165.
Wong,N.W. *et al.* (2018) OncomiR: an online resource for exploring pan-cancer microRNA dysregulation. *Bioinformatics*, **34**, 713–715.
Xu,T. *et al.* (2017) CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification, validation and visualization. *Bioinformatics*, **33**, 3131–3133.
Yang,W. *et al.* (2016a) Predicting the recurrence of noncoding regulatory mutations in cancer. *BMC Bioinformatics*, **17**, 492.
Yang,Y. *et al.* (2016b) Continuous Influence Maximization: What Discounts Should We Offer to Social Network Users? In: *Proceedings of the 2016 International Conference on Management of Data (SIGMOD'16).* pp. 727–741. Association for Computing Machinery, New York, NY, USA.
Zhang,J. *et al.* (2013) Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data. *BMC Syst. Biol.*, **7**, S4.
Zhang,J. *et al.* (2019) Identifying miRNA synergism using multiple-intervention causal inference. *BMC Bioinformatics*, **20**, 613.