Cooperative driver pathway discovery via fusion of multi-relational data of genes, miRNAs, and pathways

Jun Wang¹, Ziyi Yang¹, Carlotta Domeniconi², Xiangliang Zhang³, and Guoxian Yu¹, *

¹College of Computer and Information Sciences, Southwest University, Chongqing 400715, China.
²Department of Computer Science, George Mason University, Fairfax 22030, VA, USA.
³CEMSE, King Abdullah University of Science and Technology, Thuwal, SA.

Corresponding Author: gxyu@swu.edu.cn (Guoxian Yu).

Abstract

Discovering driver pathways is an essential step to uncover the molecular mechanism underlying cancer, and to explore precise treatments for cancer patients. However, due to the difficulties of mapping genes to pathways, and the limited knowledge about pathway interactions, most previous work focus on identifying individual pathways. In practice, two (or even more) pathways interplay and often cooperatively trigger cancer. In this study, we proposed a new approach called CDPathway to discover cooperative driver pathways. First, CDPathway introduces a driver impact quantification function to quantify the driver weight of each gene. CDPathway assumes genes with larger weights contribute more to the occurrence of the target disease and identifies them as candidate driver genes. Next, it constructs a heterogeneous network composed of genes, miRNAs, and pathways nodes based on the known intra(inter)-relations between them, and assigns the quantified driver weights to gene-pathway and gene-miRNA relational edges. To transfer driver impacts of genes to pathway interaction pairs, CDPathway collaboratively factorizes the weighted adjacency matrices of the heterogeneous network to explore the latent relations between genes, miRNAs, and pathways. After this, it reconstructs the pathway interaction network, and identifies the pathway pairs with maximal interactive and driver weights as cooperative driver pathways. Experimental results on the breast, uterine corpus endometrial carcinoma and ovarian cancer data from TCGA show that CDPathway can effectively identify candidate driver genes (AUROC ≥ 0.9) and reconstruct the pathway interaction network (AUROC > 0.9); and it uncovers much more known (potential) driver genes than other competitive methods. In addition, CDPathway identifies 150% more driver pathways and 60% more potential cooperative driver pathways than the competing methods. The code of CDPathway is available at http://mlda.swu.edu.cn/codes.php?name=CDPathway.

Key words: cooperative driver pathway, cancer gene, microRNA, data fusion, biological network

*Jun Wang is an Associate Professor of College of Computer and Information Sciences, Southwest University. Her research interests include cancer data mining, pathway analysis, GWAS and machine learning.
Ziyi Yang is a Master student of College of Computer and Information Sciences, Southwest University. Her research interests include cancer pathways analysis and biological data fusion.
Carlotta Domeniconi is an Associate Professor of Department of Computer Science, George Mason University. Her research interests include biological data mining and machine learning.
Xiangliang Zhang is an Associate Professor of Computer Science, King Abdullah University of Science and Technology, SA. Her research interests include data mining and machine learning.
Guoxian Yu is a Professor of College of Computer and Information Sciences, Southwest University. His research interests include gene function prediction, ontology analysis and data mining.
1 Introduction

Cancer is one of the most complex diseases that threaten human health. Systemic cancer genomics projects, such as the Cancer Genome Atlas (TCGA) [1] and the International Cancer Genome Consortium (ICGC) [2], have taken initial steps toward developing a blueprint of human cancer genomes by identifying, characterizing, and cataloguing alterations in thousands of tumor samples. A major challenge of interpreting the data in these projects is to distinguish the mutations that play a role in the initiation and progression of cancer, from the much larger number of passenger alterations that have little influence in cancer cell development.

Most of the early studies focus on detecting individual driver genes based on periodic mutation rates of genes in a large population of cancer patients [3, 4]. The useful implementation of this concept, however, faces a number of well known challenges. First, the vast majority of driver genes occur rarely, making them difficult to be detected statistically [5]. Second, different types of mutations exist and the amount of mutated genes is very large; as a consequence, the sets of genes involved in different studies typically have a small overlap. Therefore, it is difficult to establish a consistent causal mechanism for a given cancer [6]. Third, transformed cells are typically mutated in multiple members of a set of functionally related genes. Consequently, mutations that drive transformation, are best sought and understood in a pathway context, especially when the mutation rates are relatively low [7]. For these reasons, efforts have been moving toward identifying driver pathways [8, 9, 10].

One way to proceed is to identify known pathways that are enriched in genes carrying somatic mutations [1, 4]. More recent methods exploit genomic characteristics of mutations (i.e., mutual exclusivity that genes in the same pathway rarely mutate in the same sample), to identify oncogenic modules [11, 12]. Vandin et al. [13] proposed Dendrix (de novo driver exclusivity), which maximizes a novel weight function that considers the high coverage and mutual exclusivity of the driver gene set, to discover mutated driver pathways. To better address the maximum weight function, Zhao et al. [14] proposed an exact binary linear programming (BLP) model to maximize the weight function. Zhang et al. [15] proposed two optimization models to separately discover common driver gene sets among multiple cancer types (ComMDP), and specific driver gene sets of a particular cancer (SpeMDP). However, these de novo models focus on single pathways or modules, and they cannot shed light on how various cellular and physiological processes cooperatively alter during the initiation and progression of cancer. Besides mutation data, gene level knowledge including gene size bias [16, 17], gene-interaction networks, and expression levels [5, 7, 18] have been incorporated to uncover the significant mutations of a pathway. Copy number alterations [19] and other biological knowledge of mutational processes, such as transcript isoforms, variation in mutation type and redundancy of genetic code, were also integrated for analysis [20]. Approaches that infer patient specific pathways have also been developed [21, 22]. However, these methods ignore pathway interactions and cannot directly identify driver pathways. Furthermore, all the aforementioned methods may be biased towards super important genes, while ignoring other important driver genes with a low mutation rate.

It has been recognized that pathways often function cooperatively in carcinogenesis [23, 24]. For example, TGF-β (transforming growth factor-β) transactivates EGFR (epidermal growth factor receptor) through canonical Smad3 and ERK6/Sp1 signaling pathways, and then facilitates breast cancer migration and invasion [25]. Unfortunately, computational solutions for cooperative driver pathway discovery are still much limited, due to the difficulty of mapping genes to pathways and the incomplete pathway knowledge. Furthermore, it is difficult to collect the information of molecules within pathways and the associations among pathways. Yeang et al. [12] reported that there are significant combinatorial patterns of mutations occurring in the same patients (i.e., co-occurrence), whose driver genes usually function in different pathways, and genes in the same pathway rarely mutate in the same sample (i.e., mutual exclusivity). Leiseron et al. [26] proposed Multi-Dendrix, which uses a sum of multiple quantities for multiple pathways to identify driver pathways. Multi-Dendrix may obtain only sub-optimal results and weaken the relations between pathways, since it does not consider the functional interactions or co-occurrence of multiple pathways. Zhang et al. [27] proposed CoMDP, which extends Dendrix by maximizing the mutual exclusivity within a pathway and the co-occurrence of mutations between different pathways. But CoMDP can only obtain one pair of cooperative driver gene sets for a target disease. So CoMDP can be easily trapped, only identify super important genes with the highest coverage and miss other potential driver pathways. Dao et al. [28] proposed BeCo-WithMEFun to discover cooperative driver pathways. BeCo-WithMEFun pursues the mutual exclusivity and functional interactions within pathways, and the co-occurrence between pathways using Integer Linear Programming (ILP). However, when there are insufficient co-occurrences or functional interactions among genes, BeCo-WithMEFun cannot find a solution, or at best can only find few genes. This is because gene modules with joint mutual exclusivity, co-occurrence, and functional interaction are rare. Yang et al. [29] first integrated multiple prior knowledge data by matrix factorization and then applied a tri-random walk to identify cooperative driver pathways. This solution ignores mutation data, and it suffers from too many parameters and high time cost.

Existing cooperative driver pathway discovery methods suffer from at least one of the following issues: neglecting pathway knowledge; unable to directly identify cooperative driver pathways [26, 27, 28], limited information given by (noisy) single type of data [30]; difficulty in optimizing the problem [28]; or tedious parameter selection [29]. To address these issues, we develop a novel approach called CDPathway to leverage prior knowledge (i.e., gene-gene interactions and signal networks) and mutation profiles for cooperative driver pathways discovery. CDPathway identifies candidate driver genes based on a driver impact quantification function that integrates somatic mutation, signaling network, and gene-gene interaction (GGI) data. The quantification function measures the importance score of each gene in the mutation profile and prior functional networks, and then takes the score as driver weight of the gene. In this way, CDPathway can uncover driver genes with maximal weights that play important roles in the carcinogenesis and signaling network. Next, CDPathway imports the quantified driver weights, along with the multi-relational data of genes, miRNAs and pathways, to construct a heterogeneous network, in which the edges of gene-miRNA and gene-pathway are specified by the quantified driver weights. To map driver genes to driver pathways and to uncover the cooperation between these driver pathways, CDPathway collaboratively factorizes the weight adjacency matrices of the heterogeneous network. As such, CDPathway can transfer the influence of driver genes on disease to pathways, to reconstruct the pathway interaction network. After that, CDPathway identifies the interactive pathways with the maximal driver weights on disease as cooperative driver pathways. We apply CDPathway on two real TCGA datasets: somatic mutation and copy number variation profiles of breast cancer (BRCA) and of uterine corpus endometrial carcinoma (UCEC). After statistical and biological enrichment analysis (GO biological process terms and KEGG pathways), the genes identified by the driver impact quantification function show strong associations with disease (AUROC ≥ 0.9) and play crucial roles in important biological processes. In addition, these identified genes are tightly connected in signaling pathways, which suggests that cooperations do exist between pathways to cooperatively trigger cancer. We also evaluate the effectiveness of the collaborative matrix factorization via five fold cross validation, and obtain an AUROC > 0.9, which supports the
Cooperative driver pathway discovery

2 Methods

2.1 Step 1: Identify candidate driver genes via a driver impact quantification function

Mutations in a cancer genome tend to converge on a few biological pathways [31] and there are relatively complete mutation data on diseases. Therefore, we first identify candidate driver genes in this step and then map these candidates to pathways. In this way, the effects of the candidate driver genes on cancer can be transferred to pathways. In addition, pathway-based or network-based approaches [21, 22] for discovering cancer genes have shown that functional networks are helpful to identify cancer driver genes. Therefore, we import the gene-gene interaction (GGI) network and signaling network to identify driver genes. Several methods have been introduced to integrate mutation and prior knowledge data to identify driver genes [32, 33]. Inspired by MaxMIF [33], we introduce a new driver impact quantification function to quantify the driver weight of each gene and to identify candidate driver genes using the weights.

Previous studies indicate that a driver gene set has the key property of covering a large number of samples (high coverage) [13, 28]. As such, we calculate the mutation score for each gene to measure the contribution of its mutations to cancer. The mutation data is stored in a binary mutation matrix $X$, in which the rows represent the genes, and the columns represent the cancer samples (patients). For a protein-coding gene $i$, $X(i, k) = 1$ if it has at least one nonsilent somatic mutation in sample $k$. For a non-protein-coding gene $i$, $X(i, k) = 1$ if it has at least one mutation; and $X(i, k) = 0$ otherwise. The mutation score $M(i)$ for each gene $i$ is computed as:

$$M(i) = \begin{cases} \sum_{k \in K_i} \frac{1}{N_i} & K_i \neq \emptyset \\ \frac{1}{N_i} & K_i = \emptyset \end{cases}$$

where $K_i$ is the set of samples in which gene $i$ mutates, $N_i$ is the total number of mutated genes in sample $k$, and $N_{max}$ is the maximal number of mutated genes in all the samples. If gene $i$ does not mutate in any sample, $M(i)$ is defined as a background mutation score (BMS), which is no larger than the mutation score of any other mutated genes. In this way, BMS can help identifying possible driver genes with an extremely low mutation rate, but that play important roles in functional networks. Therefore, driver genes with a small number of mutations can still be discovered.

The gravity model derived from Newton’s law of gravitation has been used in several fields to calculate the gravitation of two bodies, e.g., population migration [34]. To further import the biological functions of genes, we assume that the genetic interaction between genes $i$ and $j$ follows a gravity model. The gene gravity model assumes that if two genes have high mutation density and strong gene co-expression in a given cancer type,
they should have a higher score and related to a higher risk of inducing mutations to other genes [35, 36, 37, 38]. Then we measure the impact of interactions between two mutated genes based on biological functions, and calculate the functional mutation score of two genes i and j as follows:

$$F(i, j) = M(i)M(j)/r_{ij}^2$$  \hspace{1cm} (2)

where $M(i)$ is the mutation score of gene i computed by Eq. (1), and $r_{ij}$ is the ‘interaction distance’ between genes i and j; $W(i, j)$ is the average interaction weight between genes i and j in the GGI and signaling networks. The interaction weights in the GGI and signaling networks indicate the confidence or interaction strength of two genes. If the i-j gene pair is only available in the GGI or signaling network, the interaction weight between them in the unavailable network is 0. Thus, F can integrate mutation information and functional relationships between two genes. Two genes with high mutation rates and (or) interacting tightly can integrate mutation information and functional relationships between two mutated genes based on biological functions, mutations to other genes [35, 36, 37, 38]. Then we measure the impact of interactions between two genes, miRNAs, and pathways, three types of inter-relational networks, and three driver weight obtained in Step 1, and $D_{gg}(i, j) = 0 (j > 1)$. In this way, the impact of each candidate gene to the disease can be embedded into the inter-relational networks, and further transferred into pathways.

Nonnegative matrix factorization (NMF) and its variants can explore sub-structures of the data matrix and have been widely used to fuse multiple data types [41, 42, 43, 44]. To remedy the sparsity of the interaction data and to map genes’ influence on the carcinogenesis to pathways through the multiple relations among genes, miRNAs and pathways, we factorize the adjacency matrices of the heterogeneous network in a collaborative fashion as follows.

$$\min_{H,F} L(H, F) = \sum_{a \in \{g, m, p\}} ||R_{aa} - H_aS_aH_a^T||_F^2$$  \hspace{1cm} (6)

where $|| \cdot ||_F$ and $tr(\cdot)$ are the Frobenius norm and the matrix trace operator. $H_a \in R^{n_a \times k_a}$ is the low-rank representation of $n_a$ gene (miRNA or pathway) nodes, $S_a \in R^{n_a \times k_a}$ encodes the latent intra-relations between respective node types. $k_a \ll n_a$ is the low-rank size. $||R_{aa} - H_aS_aH_a^T||_F^2$ is the reconstruction loss of intra-relational networks of genes (miRNAs and pathways). $tr(H_a^T R_{pp} H_a)$ is introduced to enforce the inter-relations being preserved with respect to the low-rank representations of different objects. $H_a^T$ enforces the sparsity of $H_a$ and $\gamma$ is a scalar weight parameter. Eq. (6) can respect the intrinsic structure of the heterogeneous network and explore the latent relationships between genes, miRNAs, and pathways. In addition, the driver impacts embedded into the inter-relational networks can also coordinate the exploration of cooperation between pathways. The detailed optimization procedure for Eq. (6) is provided in Section 1 of the Supplementary file.

After obtaining the optimized $H_a$ and $S_{aa}$ in Eq. (6), we can reconstruct the pathway-pathway interaction subnetwork $R_{pp}$ as follows:

$$R_{pp} = H_aS_{pp}H_a^T$$  \hspace{1cm} (7)

Since the inter-relational gene-pathway and gene-miRNA networks import the driver weights to reconstruct the pathway-pathway interaction networks, pathways with interaction and driver influence on disease have high association values in $R_{pp}$, as our experimental results will manifest. Given this, we identify the top K pathway pairs with the highest value in $R_{pp}$ as cooperative driver pathways.

3 Results

3.1 Datasets and parameter investigation

To investigate the effectiveness of CDPathway, we carry out experiments on publicly available cancer data of breast cancer (BRCA) and uterine corpus endometrial carcinoma (UCEC) from TCGA [1]. Details about the
multiple data sources used for the experiments are listed in Table 1. Each of the interaction weight between two genes was extracted and standardized with a value ranging from 0 to 1, by divided to the largest interaction weight among all genes. Self-loop interactions were removed to simplify the networks. Multiple miRNAs’ expressions change during tumorigenesis and development, and cascade to downstream target gene expression in the signaling pathway, and thus affect tumor progression. Given that, we assume that an miRNA and a pathway are more likely to be related if they are associated with the same gene, so we construct the miRNA-pathway association network based on the known gene-pathway and gene-miRNA associations.

We divided the original data by 7:1:2 into training data, validation data and test data, and selected the parameters using the validation data. The driver weight threshold \( \theta \) is an important parameter for filtering useless genes in Step 2. We split useless genes with driver weight less that \( \theta \) to reconstruct relational networks of genes, miRNAs and pathways. As a result, the reconstructed gene-gene network used in Step 2 is a sub-network of the original gene-gene interaction (GGI) network used in Step 1. We show the distribution of driver weight in Figure 2, the majority of genes’ driver weight are less than 0.01 and 0.001 for BRCA and UCEC datasets, indicating that these genes are passenger genes, which do not contribute to the carcinogenesis. To study the sensitivity of \( \theta \), based on the distribution of driver weights in Figure 2, we fix low-rank size \( k_\gamma = k_m = k_p = k_o \) as 50 and penalty coefficient \( \gamma \) as 100, and increase \( \theta \) from 0 to 0.1 in BRCA dataset and from 0 to 0.01 in UCEC dataset. Figure 3 reports the the area under the receiver operating characteristic curve (AUROC) and the area under the precision recall curve (AUPRC) under different input values of \( \theta \) in reconstructing pathway-pathway associations.

Both the AUROC value and AUPRC value vary no more than 0.2 as the increase of \( \theta \), and reach to the highest when \( \theta \approx 0.07 \) on the BRCA dataset and \( \theta \approx 0.006 \) on the UCEC dataset. Particularly, when \( \theta = 0 \), both the AUROC and AUPRC are the lowest, indicating that filtering passenger genes is helpful for reconstructing the pathway-pathway interaction network. In practice, we also tested with a larger threshold, like 0.5, 1 and 5, the AUROC and AUPRC values drop to lower than when \( \theta = 0 \), since important driver genes are excluded with such a large threshold. Given these observations, we adopt \( \theta = 0.07 \) and \( \theta = 0.006 \) for experiments on BRCA and UCEC datasets, respectively.

The low-rank size \( k_o \) and penalty coefficient \( \gamma \) are important parameters for reconstructing pathway-pathway interaction network, and tuning them altogether is time consuming. For simplicity, we simply fix \( k_p = k_m = k_o = k_w \), and then study how \( k_o \) and \( \gamma \) affect the reconstruction of pathway-pathway interaction network. We increase \( k_o \) from 10 to 110 and \( \gamma \) from \( 10^{-6} \) to \( 10^6 \) on BRCA and UCEC datasets. Figure 4 reports the AUROC and AUPRC under different input values of \( k_o \) and \( \gamma \).

Both the AUROC and AUPRC increase as the increase of \( k_o \) and reach to the plateau when \( k_o \geq 40 \) on the two datasets. That is because the intra-relations cannot be preserved when the size \( (k_o) \) of low-rank representation \( H_o \) is too small. Similarly, both the AUROC and AUPRC increase as the rise of \( \gamma \) and reach to a plateau when \( \gamma \geq 10 \) on both datasets. The values of AUROC and AUPRC manifest a significant reduce when \( \gamma < 1 \). That is because \( \sum_{o \in \{g,m,p\}} H_o \) encourages the sparsity of \( H_o \) and thus to reduce the false positive reconstructed edges between pathway nodes. A too low \( \gamma \) cannot enforce the sparsity.

Based on the above analysis, we adopt \( k_o = 90 \) and \( k_o = 100 \), \( \gamma = 10^3 \) and \( \gamma = 10^4 \) for experiments on BRCA and UCEC datasets, respectively. Based on these selected parameters, CDPathway again performs well on the test data with an AUROC of 0.8952 and 0.9041, an AUPRC of 0.5714 and 0.5846 on the BRCA and UCEC dataset, respectively.

3.2 The driver weight quantification function can effectively identify driver genes

Known driver genes can be identified by our driver weight quantification function with statistic significance. To study the effectiveness of identifying known driver genes (reported in [53]), we specify the number of identified candidate driver genes in Step 1 from one to the number of genes in the mutation profiles (13965 in BRCA and 18421 in UCEC), and plot the ROC and PR curves in Figure 5.

For both BRCA and UCEC datasets, the AUROC values are over 0.9, and AUPRC values are over 0.2. The reason why the AUPRC is relatively low compared to the AUROC is due to that CDPathway identifies genes which drive the carcinogenesis of the target cancer, while the known genes are drivers for various diseases. Another cause is that AUPRC is more sensitive to the class-imbalance task and identifying known driver gene in essence is a class-imbalance task. As can be seen, the Precision is very high at the beginning, which indicates that top ranked genes are important driver genes. The genes with weights less than \( \theta \) should be excluded from the heterogeneous network, and they drag down the Precision for a high Recall and cause a low AUPRC value. Overall, CDPathway can effectively identify known driver genes.
Fig. 3. AUROC and AUPRC vs. $\theta$.

Fig. 4. The AUROC and AUPRC of CDPathway under different low-rank size $k_o$ and sparsity coefficient $\gamma$.

Table 1. Details of collected datasets.

<table>
<thead>
<tr>
<th>Data type</th>
<th>Size(BRCA)</th>
<th>Size(UCEC)</th>
<th>Direction</th>
<th>Weight</th>
<th>Sources</th>
<th>Version or Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>mutation profile</td>
<td>15863 × 776</td>
<td>18621 × 344</td>
<td>No</td>
<td>No</td>
<td>TCGA [1]</td>
<td>2018.11</td>
</tr>
<tr>
<td>GGE</td>
<td>16607 × 16907</td>
<td>16607 × 16907</td>
<td>Undirected</td>
<td>Weighted</td>
<td>STRING [58]</td>
<td>2017.05</td>
</tr>
<tr>
<td>signaling network</td>
<td>12456 × 12456</td>
<td>12456 × 12456</td>
<td>Undirected</td>
<td>Weighted</td>
<td>STRING [58]</td>
<td>2017.05</td>
</tr>
<tr>
<td>gene-gene ($R_{gg}$)</td>
<td>1631 × 1631</td>
<td>1631 × 1631</td>
<td>Undirected</td>
<td>Weighted</td>
<td>STRING [58]</td>
<td>2018.11</td>
</tr>
<tr>
<td>gene-miRNA ($R_{gm}$)</td>
<td>1631 × 559</td>
<td>1631 × 559</td>
<td>Undirected</td>
<td>Weighted</td>
<td>STRING [58]</td>
<td>2018.05</td>
</tr>
<tr>
<td>gene-pathway ($R_{gp}$)</td>
<td>1631 × 212</td>
<td>1631 × 212</td>
<td>Undirected</td>
<td>Weighted</td>
<td>STRING [58]</td>
<td>2018.05</td>
</tr>
<tr>
<td>miRNA-miRNA ($R_{mm}$)</td>
<td>559 × 559</td>
<td>559 × 559</td>
<td>Undirected</td>
<td>Weighted</td>
<td>STRING [58]</td>
<td>2018.05</td>
</tr>
<tr>
<td>miRNA-pathway ($R_{mp}$)</td>
<td>559 × 212</td>
<td>559 × 212</td>
<td>Undirected</td>
<td>Weighted</td>
<td>STRING [58]</td>
<td>2018.05</td>
</tr>
<tr>
<td>pathway-pathway ($R_{pp}$)</td>
<td>212 × 212</td>
<td>212 × 212</td>
<td>Undirected</td>
<td>Weighted</td>
<td>STRING [58]</td>
<td>2018.05</td>
</tr>
</tbody>
</table>
Cooperative driver pathway discovery

Fig. 5. Receiver operating characteristics (ROC) and precision recall (PR) curves of identifying known driver genes for BRCA and UCEC datasets. The corresponding AUROC and AUPRC values are also included in the legend. CDPathway-noSig: CDPathway without the signaling network. CDPathway-noGGI: CDPathway without the GGI network. CDPathway-noNetwork: CDPathway without the signaling and GGI networks.

In order to investigate the contribution of GGI and signaling networks, we also compare CDPathway with CDPathway without the GGI or the signaling networks. When a single functional network is disregarded, the AUROC and AUPRC values are reduced by 1.7% and 16.7% on BRCA, and by 1.8% and 13.1% on UCEC, respectively. When both networks are removed, the AUROC (AUPRC) value drops sharply by 15.1% (62.8%) and 16.3% (53.6%) on BRCA and UCEC datasets, respectively. Therefore, we can conclude that both prior knowledge networks contribute to the identification of driver genes, and disregarding any of the two will compromise the performance. In addition, we also study how CDPathway perform when the functional networks are not complete. We randomly remove 5%-50% edges in the GGI or signaling network and record the AUROC and AUPRC values. As can be seen, the AUROC and AUPRC manifest a decrease as more edges removed. Generally, CDPathway relies to the completeness of functional networks, but the AUROC (AUPRC) drops less than 10% (25%) when even 50% of edges in the GGI or signaling network are removed. That is because we also integrate the somatic mutation data in Step 1, and the mutation rate of genes (i.e., coverage) are helpful to identify potential driver genes. Therefore, we can infer that the driver weights of genes are not so dependent on the completeness of the functional networks.

After filtering useless genes, we further apply a hypergeometric test on the identified candidate driver genes of the two real datasets to study the statistical significance of the candidate driver genes. The \( p \)-values of the identified known driver genes are 9.68E-62 and 1.67E-54 for BRCA and UCEC datasets, respectively. Therefore, we can conclude that CDPathway can identify known driver genes with statistical significance. In addition, we select the top 30 candidate driver genes in Step 1, and do functional and pathway enrichment analysis using the widely-used DAVID toolkit [54]. As Table 2 shows, there are 17 and 14 known driver genes among the 30 candidate driver genes for BRCA and UCEC datasets, respectively. Besides the known driver genes, CDPathway can also identify several novel driver genes. For the BRCA dataset, it is highlighted that FOXA1 (Forkhead box protein A1) and GATA3 (GATA binding protein 3) co-regulate the expression of genes, which is essential for the luminal mammary epithelial cell development [55, 56]. Another example is that mutations in MAP2K4 produce perturbations similar to MAP3K1 loss [57]. For UCEC dataset, TP53 (tumor protein p53) is involved with DNA repair, growth arrest, and apoptosis. In particular, mutations in TP53 can lead to uncontrolled proliferation and invasive growth. On the other hand, DCC (Deleted in Colorectal Carcinoma) is reported to have an anti-metastatic role [58], meaning that it may only contribute to cancer in the context of a pre-existing condition. We conjecture that the mutations in DCC may contribute to cancer progression for patients with defective mismatch repair and/or impaired TP53 functionality. A network based analysis also provides supporting evidence of TTN mutations as a disease marker [60].

Table 2 also shows some enriched function terms. There are several overlapping important biological processes enriched in both BRCA and UCEC datasets, including transcriptional activator activity, central carbon metabolism in cancer, apoptosis, etc. The candidate driver genes are also enriched in the target disease and cancer related pathways. These results demonstrate that CDPathway can effectively identify candidate driver genes engaging in the cancer related pathways and biological processes.
3.3 Collaborative matrix factorization can effectively reconstruct the pathway-pathway interaction network

To investigate the effectiveness of collaborative matrix factorization in Step 2, we randomly divide the original pathway-pathway interaction network into five folds for cross validation. Next, we plot the ROC and PR curves of CDPathway in Figure 7. We run five fold cross validation for ten independent rounds, and report the average results. We can see that CDPathway achieves high AUROC values of 0.9419 and 0.9318, AUPRC values of 0.7838 and 0.7707 on BRCA and UCEC datasets, respectively. The high AUROC and AUPRC values indicate that CDPathway can accurately reconstruct pathway-pathway associations.
To investigate the contribution of the quantified driver gene weights obtained in Step 1, we test the collaborative matrix factorization without importing driver weights. As shown in Figure 7, both AUROC and AUPRC drop by 12.6% and 8.3%, 27.2% and 23.4% on BRCA and UCEC datasets, respectively, indicating that the imported weights are helpful in reconstructing the pathway-pathway interaction network. In addition, we also test CDPathway without prior functional networks. Without any functional network, both the AUROC and AUPRC values drop more seriously. The integration of the signaling and GGI networks can jointly improve the performance. In addition, the GGI and signaling networks show equal importance (no more than 3% difference) for the reconstruction of the pathway-pathway interaction network. We can conclude that the GGI and signaling networks complement each other and both contribute to the reconstruction of the pathway-pathway interaction network.

3.4 CDPathway can identify significant cooperative driver pathways

We select 10 pathway pairs with top 10 highest values in the reconstructed pathway-pathway interaction network as cooperative driver pathways. As shown in Table 3, by checking the member genes of these pathways, we find that each pathway contains at least one known driver gene, which indicates that all pathways play important roles in carcinogenesis by undertaking the biological activities of the driver genes. In addition, these pathways also confirm that the driver effects of genes on the disease are correctly transferred to pathways by the collaborative matrix factorization on the heterogeneous network in Step 2.

We further check their cooperation and find that all the pathway pairs are associated in the collected pathway-pathway interaction network. We visualize pathway pair 1 of BRCA dataset in Figure 8 as an example. Other examples are provided in the Supplementary file. We observe that the pathway pair 1 interacts with each other to undertake biological activities. These two pathways also share common known driver genes: PIK3CA, PIK3R1, and KRAS. These genes have interactions and may cooperate with each other to drive the target disease. The shared gene NCOA3 does not have evident functional edges with any other gene, but it was predicted to have connections with the driver gene EGFR [59], which also supports the capability of CDPathway in identifying cooperative driver pathways.

There are 5 known pan-cancer related pathways identified by CDPathway, including Class I PI3K signaling events mediated by Akt, mTOR signaling pathway, TGF-β receptor signaling, Canonical Wnt signaling pathway, and Notch signaling pathway [61]. Breast cancer studies have shown that the abnormality of Internalization of ErbB1, Canonical Wnt signaling pathway, Alternative NF-kappaB pathway, Class...
Fig. 8. Pathway pair 1 of BRCA dataset in the GGI network from String [59]. The rose red lines indicate interactions from curated databases and blue lines indicate interactions have been experimentally determined. Genes in the ‘Internalization of ErbB1’ pathway are circled, and genes in ‘Trk receptor signaling mediated by PI3K and PLC-gamma’ pathway are enclosed by rectangles. These two pathways share common known driver genes of PIK3CA, PIK3R1, and KRAS.

I PI3K signaling events mediated by Akt, Notch signaling pathway and C-MYC pathway promote the carcinogenesis [62, 63, 64, 65, 66, 67]. For UCEC, Trk receptor signaling mediated by PI3K and PLC-gamma, TGFBeta receptor signaling, FOXM1 transcription factor network and mTOR signaling pathway are reported to have tight associations with the target disease [68, 69, 70]. We can also find that identified driver pathway pairs play important roles in the life activities related to the initiation of cancer. For example, in pathway pair 3 (Figure S12 of Supplementary file) on the UCEC dataset, ATM pathway provides the crucial link between DNA damage, cell cycle progression and cell death by first sensing double stranded DNA breaks, and subsequently phosphorylating and activating other downstream proteins functioning in DNA damage repair, cell cycle arrest and apoptotic pathways [71]. Therefore, the identified pathways’ pairs cooperate with each other and play important roles in biological activities. It’s quite possible that their cooperation drives the occurrence of the target disease.

3.5 Comparison with other methods

We compare CDPathway against six related and competitive methods, including two individual driver pathways identification methods: Dendrix [13] and BLP [14], and four multiple driver pathways identification methods: CoMDP [27], Multi-Dendrix [26], BeCo-WithMEFun [28], and CoDP [29]. We discussed these approaches in the Introduction. We set the parameter ranges of the comparing methods as described in the original papers, and select pathways that contain the most known driver genes with statistic significance. Since these methods, except CoDP, do not identify exact driver pathways, we perform a comparison at the gene level, by mapping pathways identified by CDPathway and CoDP to genes using known gene-pathway association networks. On the other hand, we compare CDPathway with these methods at the pathway level by doing a KEGG enrichment test for the driver gene set identified by comparing methods via DAVID [54].

As shown in Table 4, CDPathway outperforms other methods by identifying on average twice more known driver genes. Biologically more genes make less sense, but CDPathway uncovers more driver genes instead of passenger genes than other methods. In fact, CDPathway uncovers how driver pathways cooperatively trigger cancer by including these driver genes to affect important biological activities. Therefore, CDPathway provides driver pathways with diversity. As shown in Table 4, the known driver genes are identified with statistical significance (Hypergeometric test P-value <0.05), and has better statistical significance and F1-Score than other comparing methods, except getting a similar F1-Score with CoMDP on the UCEC dataset. Therefore CDPathway does not identify more known driver genes by only identifying more pathways, it balances the Precision and Recall of identifying cooperative driver pathways. In contrast, the comparing methods identify limited driver genes. As a result, they also provide limited diversity of driver pathways in different datasets. These results indicate that CDPathway can better capture the heterogeneity of different datasets than comparing methods.

We also compare CDPathway with comparing methods on identifying known driver genes. We fix the number of identified driver genes from 1 to the number of genes in the somatic mutation profile, and show the ROC and PR curves in Fig. 9. Since BeCo-WithMEFun cannot get a solution when the number of identified genes is larger than 40, its ROC and PR curves cannot be obtained. As shown in Fig. 9, CDPathway outperforms the comparing methods with respect to AUROC and AURPC by at least 15%, since it makes use of somatic mutation data and prior knowledge.
of gene–gene interaction network to identify potential driver genes. We observe that multiple driver pathway identification methods outperform individual driver pathway identification methods, since the former methods uncover driver pathways with diversity, while the latter ones only uncover very limited driver genes. In addition, the PR curves of these comparing methods drop sharply with the increased number of identified genes. That is because these individual driver pathway identification methods can only identify one or two gene sets (i.e., pathway), and they cannot provide meaningful results when identifying more than 10 genes within only one or two pathways.

CDPathway can also uncover novel driver genes, which play important roles in important biological activities. For example, we obtain EGFR (Epidermal growth factor) in the BRCA dataset, and EGFR has tight associations with the survival rate of breast cancer patients [72]. The reason why CDPathway outperforms other methods is that CDPathway integrates both mutation and GGI data, which help reflecting the importance of why CDPathway outperforms other methods is that CDPathway integrates both mutation and GGI data, which help reflecting the importance of gene–gene interactions among pathways identified by CoDP (most effective method among the comparing methods) is supported by the pathway–pathway interaction network, while that of CDPathway is 100%. This observation indicates that CDPathway is more effective in identifying pathways with both cooperation and functional significance. CDPathway can accurately identify important genes in Step 1 and Step 2, can retain the driver effects of genes on interacting pathways, and thus more accurately identify cooperative driver genes. Because of the driver genes, these identified pathways not only have interaction with each other, but also cooperatively carry out disease related activities.

4 Discussion
Cooperative driver pathways discovery helps understanding how multiple pathways function cooperatively in cancer initiation and progression, and provides more precise therapy to patients. For this discovery, we propose a two-stage based approach called CDPathway. CDPathway first uses a driver impact quantification function to identify candidate driver genes. Next, CDPathway collaboratively factorizes the adjacency matrices of a heterogeneous network of genes, miRNAs, and pathway

---

**Table 3. Cooperative driver pathways identified by CDPathway**

<table>
<thead>
<tr>
<th>No.</th>
<th>BRCA</th>
<th>UCEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AKT1, AKT2, EGFR, BRAF, MET, JAK2, MAP2K1, PIK3CA, PTEN</td>
<td>AKT1, AKT2, EGFR, BRAF, MET, JAK2, MAP2K1, PIK3CA, PTEN</td>
</tr>
<tr>
<td>2</td>
<td>PIK3CA, PIK3R1, CTNNB1, AKT1, PTPN11, BRAF, CASP8, MAP3K1</td>
<td>PIK3CA, PIK3R1, CTNNB1, AKT1, PTPN11, BRAF, CASP8, MAP3K1</td>
</tr>
<tr>
<td>3</td>
<td>PIK3CA, PIK3R1, CTNNB1, AKT1, PTPN11, BRAF, CASP8, MAP3K1</td>
<td>PIK3CA, PIK3R1, CTNNB1, AKT1, PTPN11, BRAF, CASP8, MAP3K1</td>
</tr>
<tr>
<td>4</td>
<td>PIK3CA, PIK3R1, CTNNB1, AKT1, PTPN11, BRAF, CASP8, MAP3K1</td>
<td>PIK3CA, PIK3R1, CTNNB1, AKT1, PTPN11, BRAF, CASP8, MAP3K1</td>
</tr>
<tr>
<td>5</td>
<td>PIK3CA, PIK3R1, CTNNB1, AKT1, PTPN11, BRAF, CASP8, MAP3K1</td>
<td>PIK3CA, PIK3R1, CTNNB1, AKT1, PTPN11, BRAF, CASP8, MAP3K1</td>
</tr>
</tbody>
</table>

**Table 4. Known driver gene sets of BRCA and of UCEC identified by CDPathway and other comparing methods.**

<table>
<thead>
<tr>
<th>Method</th>
<th>Driver genes of BRCA</th>
<th>Driver genes of UCEC</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDPathway</td>
<td>PIK3CA, PIK3R1, CTNNB1, AKT1, PTPN11, BRAF, CASP8, MAP3K1</td>
<td>PIK3CA, PIK3R1, CTNNB1, AKT1, PTPN11, BRAF, CASP8, MAP3K1</td>
<td>6.28E-26</td>
</tr>
<tr>
<td>CoDP</td>
<td>PIK3CA, PIK3R1, CTNNB1, AKT1, PTPN11, BRAF, CASP8, MAP3K1</td>
<td>PIK3CA, PIK3R1, CTNNB1, AKT1, PTPN11, BRAF, CASP8, MAP3K1</td>
<td>1.23E-03</td>
</tr>
<tr>
<td>CoMDP</td>
<td>PIK3CA, PIK3R1, CTNNB1, AKT1, PTPN11, BRAF, CASP8, MAP3K1</td>
<td>PIK3CA, PIK3R1, CTNNB1, AKT1, PTPN11, BRAF, CASP8, MAP3K1</td>
<td>1.52E-23</td>
</tr>
</tbody>
</table>
Fig. 9. Comparison of ROC and PR curves of identifying known driver genes for BRCA and UCEC datasets.

Table 5. Cooperative driver pathways identified by other comparing methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>BRCA</th>
<th>UCEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoDP</td>
<td>Beta1 integrin-cell surface interactions,</td>
<td>p53 pathway,</td>
</tr>
<tr>
<td></td>
<td>Syndecan-1-mediated signaling events;</td>
<td>Class I PDK signaling events mediated by Akt,</td>
</tr>
<tr>
<td></td>
<td>Validated transcriptional targets of API family members Fas and FasL</td>
<td>T cell receptor signaling pathway,</td>
</tr>
<tr>
<td></td>
<td>Signaling mediated by p85-alpha and p85-beta;</td>
<td>B cell receptor signaling pathway,</td>
</tr>
<tr>
<td></td>
<td>Signaling events mediated by VEGF-R1 and VEGF-R2;</td>
<td>mTOR Signaling Pathway,</td>
</tr>
<tr>
<td></td>
<td>Ceramide signaling pathway;</td>
<td>IL2 signaling events mediated by STAT5,</td>
</tr>
<tr>
<td></td>
<td>Direct p53 effector;</td>
<td>mTOR signaling pathway,</td>
</tr>
<tr>
<td></td>
<td>Validated nuclear estrogen receptor alpha network;</td>
<td>IL4-mediated signaling events;</td>
</tr>
<tr>
<td>BeCo-WithMEEFun</td>
<td>Downstream signaling in naive CD8 T cells,</td>
<td>TCR signaling in naive CD4 T cells,</td>
</tr>
<tr>
<td></td>
<td>Downstream signaling in naive CD8 T cells;</td>
<td>Validated targets of C-MYC transcriptional repression;</td>
</tr>
<tr>
<td>Dendrix</td>
<td>Thyroid cancer</td>
<td>TrkA Receptor Signaling Pathway,</td>
</tr>
<tr>
<td>BLP</td>
<td>Small cell lung cancer</td>
<td>mTOR Signaling Pathway,</td>
</tr>
<tr>
<td>CoMPD</td>
<td>Thyroid cancer, p53 pathway</td>
<td>B cell receptor signaling pathway,</td>
</tr>
<tr>
<td></td>
<td>ESR1 receptor signaling pathway,</td>
<td>B cell receptor signaling pathway,</td>
</tr>
<tr>
<td></td>
<td>mTOR Signaling Pathway,</td>
<td>Apoptosis,</td>
</tr>
<tr>
<td></td>
<td>B cell receptor signaling pathway,</td>
<td>B cell receptor signaling pathway,</td>
</tr>
</tbody>
</table>

Experimental results on public breast, endometrial and ovarian cancer datasets show that CDPathway not only uncovers known driver genes with better statistical significance than competitive methods, but also discovers more novel driver genes. We observe that integrating mutation data and prior functional networks can better differentiate driver genes from passenger genes, because CDPathway uncovers genes with high mutation rate, which are hub nodes in the functional networks. The collaborative matrix factorization correctly transfers these driver genes to pathways, and thus the cooperative driver pathways identified by CDPathway contain more known driver genes than those of existing methods. Additionally, less known and less frequently altered genes with well characterized cancer drivers, suggest a mechanism of action. For example, on BRCA dataset, FOXA1 and GATA3 co-regulate the expression of genes, which are essential for the luminal mammary epithelial cell development. On the pathway level, CDPathway can identify more potential biological cooperation among pathways. On both BRCA and UCEC datasets, each identified pathway contains at least one known driver gene, and we found...
the cooperation of Internalization of ErbB1 and Trk receptor signaling mediated by PI3K and PLC-gamma may be important to the occurrence of disease. We can infer that because of the tight associations within and between genes, miRNAs and pathways, transferring the driver impact from genes to pathways through miRNAs helps to uncover cooperative driver pathways.

We plan to extend this work to other biological networks. Given that there are many other types of molecules important to biological activities, such as LncRNAs, genes and drugs. Our method can be further applied on multi-type related biological networks to uncover more disease related molecules. The challenge of extending our method is to build such relational biological networks and to choose a type of molecules with trigger influence on the disease.

Key Points
• Pathways undertake important biological activities and are robust by providing reliable gene regulatory relationships. Biological pathways have been applied to explore the pathology involved with cancer occurrence, diagnosis and prognosis. In practice, multiple pathways cooperatively drive the occurrence and progression of cancer, but scanty computational models can identify cooperative driver pathways.
• We propose a novel model CDPathway to identify cooperative driver pathways. CDPathway combines the mutation data and prior knowledge of genes (i.e., gene-gene interaction network) to capture the influences of genes on the occurrence of cancers. CDPathway further exploits these influences to explore cooperative driver pathways via tightly associated biological networks of genes, miRNAs and pathways.
• CDPathway shows a good prognostic performance in breast, endometrial and ovarian cancers. CDPathway can identify driver pathways with significance and diversity. It not only identifies known driver genes already explored by existing methods, but also finds out novel driver genes. In addition, CDPathway uncovers the cooperations between driver pathways, which better reflects how biological activities trigger cancer than existing methods.

Funding
This work is supported by Natural Science Foundation of China (61873214 and 61873200), Fundamental Research Funds for the Central Universities (XDJK2019B024), Natural Science Foundation of CQ CSTC (cstc2018jcyjAX0228).

References