BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Discovery of co-occurring driver pathways in cancer

Junhua Zhang*, Ling-Yun Wu, Xiang-Sun Zhang and Shihua Zhang*

Abstract

Background: It has been widely realized that pathways rather than individual genes govern the course of carcinogenesis. Therefore, discovering driver pathways is becoming an important step to understand the molecular mechanisms underlying cancer and design efficient treatments for cancer patients. Previous studies have focused mainly on observation of the alterations in cancer genomes at the individual gene or single pathway level. However, a great deal of evidence has indicated that multiple pathways often function cooperatively in carcinogenesis and other key biological processes.

Results: In this study, an exact mathematical programming method was proposed to *de novo* identify **co**-occurring **m**utated **d**river **p**athways (CoMDP) in carcinogenesis without any prior information beyond mutation profiles. Two possible properties of mutations that occurred in cooperative pathways were exploited to achieve this: (1) each individual pathway has high coverage and high exclusivity; and (2) the mutations between the pair of pathways showed statistically significant co-occurrence. The efficiency of CoMDP was validated first by testing on simulated data and comparing it with a previous method. Then CoMDP was applied to several real biological data including glioblastoma, lung adenocarcinoma, and ovarian carcinoma datasets. The discovered co-occurring driver pathways were here found to be involved in several key biological processes, such as cell survival and protein synthesis. Moreover, CoMDP was modified to (1) identify an extra pathway co-occurring with a known pathway and (2) detect multiple significant co-occurring driver pathways for carcinogenesis.

Conclusions: The present method can be used to identify gene sets with more biological relevance than the ones currently used for the discovery of single driver pathways.

Background

The pathogenesis of cancer in humans is still poorly understood. To improve the diagnosis and treatment of cancer patients, several large-scale cancer genomics projects (e.g., the Cancer Genome Atlas (TCGA) [1], and International Cancer Genome Consortium (ICGC) [2]) have been performed in recent years. Analyzing these high-throughput data provides valuable opportunities to understand the formation and progression of cancer [3,4].

Generally, a large number of mutations occur in cancer genomes (e.g., somatic mutations and copy number alterations (CNAs)). One crucial step in cancer research is to distinguish driver mutations and driver genes, which contribute to the progression of cancer from normal to malignant states, from passenger mutations and passenger genes, which accumulate in cells but do not contribute to cancer development [5,6]. Most early efforts were devoted to the detection of individual driver genes based on recurrent mutations of the genes in a large cohort of cancer patients [7].

Because of the mutational heterogeneity of cancer genomes, more attention has been paid to identify driver pathways and modules rather than individual genes in recent years [1,8,9]. It is noteworthy that most such methods involve the use of prior knowledge about pathways and/or protein interaction networks. For example, known pathways were analyzed for enrichment of somatic mutations [1,8,9], or were examined to find which ones are significantly disturbed across many patients [10,11]. On the other hand, several studies indicated that driver pathways often cover a large number of samples. More importantly, mutations of the genes in one pathway usually

*Correspondence: zjh@amt.ac.cn; zsh@amss.ac.cn
National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

exhibit mutual exclusivity, i.e., a single mutation is usually enough to disturb one pathway [12,13]. These rules have been frequently used to detect driver pathways and modules [14-16]. For example, Ciriello *et al.* proposed MEMo (Mutual Exclusivity Modules) to detect oncogenic network modules within a constructed network using gene mutation information and a human reference network (including protein interactions and signal transduction pathways) [14].

However, it is believed that the human protein interaction network is incomplete. There are many unknown protein interactions and a great deal of knowledge about biological pathways remains unclear. Many reverse engineering approaches were developed in recent years to infer complex biological regulatory relationships. For example, Acharya *et al.* proposed the Gene Set Gibbs Sampling (GSGS) method to reconstruct signaling pathway structures by sequentially inferring the information flows from the overlapping information flow gene sets [17]. It is necessary to develop new methods to discover mutated driver pathways or core modules without relying on prior information. Recently, Vandin *et al.* proposed an approach, called Dendrix (*de novo* driver exclusivity), to *de novo* discover mutated driver pathways using somatic mutation data [18]. In this method, a novel weight function was introduced by combining the coverage and exclusivity of the gene set. Maximization of the weight function is defined as the maximum weight submatrix problem. This was originally solved by the Markov chain Monte Carlo (MCMC) method [18], and was then addressed using an exact binary linear programming (BLP) model [19]. However, these studies for the identification of driver pathways or core modules have all focused on single pathways or modules [15,16,18,19]. How various cellular and physiological processes are coordinately altered during the initiation and progression of cancer, it is still a major challenge.

It is well known that multiple pathways with mutations are generally required for cancer [20]. In fact, it has been recently recognized that pathways often function cooperatively in carcinogenesis [13,21-23]. Based on mutation data from COSMIC [24] and six major cancer-associated pathways from previous studies, Yeang *et al.* demonstrated that there were significant combinatorial patterns of mutations occurring in the same patients (i.e., co-occurring), for which the corresponding genes usually function in different pathways, whereas mutations in genes functioning in the same pathway are rarely mutated in the same sample (i.e., mutually exclusive) [13]. Cui *et al.* identified 12 oncogene-signaling blocks from the integrated human signaling network [21]. They found that some of them (such as the *RAS* and *TP53* blocks in central nervous system, pancreas, skin, and blood tumors) would collaboratively promote cancer signaling and foster tumorigenesis. Using 18 pathways enriched with mutations in lung adenocarcinoma [8], Gu *et al.* investigated pathway cooperation in cancer cells in terms of superpathways, which are clusters of co-disrupted pathways whose significance is tested by the hypergeometric model [25]. More recently, Gu *et al.* devised a heuristic approach to detect cooperative functional modules in the glioblastoma multiforme (GBM) altered network which is obtained by mapping mutated genes onto a protein interaction network from the Pathway Commons database, and several pairs of significantly co-altered modules were identified which are involved in the main pathways known to be perturbed in GBM [26].

All these studies indicate that carcinogenesis is a complex process and the malignant transformation from a normal cell to a tumor is indeed a highly cooperative procedure involving synergy between pathways. Therefore, systematically exploring the complex collaboration among different biological pathways and functional modules is a crucial step, which will shed new lights on our understanding of the cellular mechanisms underlying tumorigenesis. Current studies have mainly focused on the utilization of prior knowledge to determine whether two or more pathways or modules are simultaneously perturbed in the same samples. Considering the incompleteness of the knowledge about pathways and protein interaction networks, *de novo* discovery of collaborative pathways playing driver roles in cancer initiation and development is of pressing need. Although iteratively performing Dendrix [18] or BLP [19] can obtain multiple pathways by removing the gene sets found in each previous iteration, however, such pathways are not guaranteed to be significantly co-disrupted in the same patients.

In this study, a mathematical programming approach to discover **co**-occurring **m**utated **d**river **p**athways (CoMDP) in cancer generation and progression was developed. The co-occurring pathways detected here possess two properties: first, each pathway is a set of mutated genes with high coverage and high exclusivity; second, the mutations between pathway genes exhibit a statistically significant co-occurrence in cancer samples. CoMDP is an exact method where the optimal set of pathways is obtained using an efficient algorithm. It does not require any prior information besides mutation profiles. To evaluate this method, we first applied it onto simulated data and compared it with the original BLP method. Then we applied it onto four biological datasets and several pathways which might play collaborative roles in carcinogenesis were identified. For example, for the glioblastoma tumor data and lung adenocarcinoma data, several significant co-occurring gene pathways were detected. Each pair interacts and regulates the cell survival and protein synthesis processes. In addition,

a modified form (named mod_CoMDP) was proposed in situations in which a certain pathway has been previously proven to play important roles in some cancers and one wants to know whether there are other pathways with cooperative effects in tumorigenesis. Furthermore, multiple co-occurring driver pathways can be discovered by combining previously detected pairs of gene sets and identifying others using mod_CoMDP. When applied to the ovarian carcinoma dataset, CoMDP and/or mod_CoMDP identified driver pathways related to *TP53* in the generation and progression of ovarian cancer. In summary, we developed a method for identifying mutated co-occurring driver pathways which can enhance the understanding of molecular mechanisms underlying tumorigenesis.

Methods

Brief introduction of the maximum weight submatrix problem

The Dendrix method was designed to *de novo* discover a single mutated driver pathway from somatic mutation data, where a weight function W was introduced by combining the coverage and exclusivity of the gene set [18]. Given a binary mutation matrix A with m rows (samples) and n columns (genes), the maximization of W was defined as the maximum weight submatrix problem [18], which means to find a submatrix M of size $m \times k$ in the mutation matrix A by maximizing the weight function W:

$$W(M) = |\Gamma(M)| - \omega(M) = 2|\Gamma(M)| - \sum_{g \in M} |\Gamma(g)|, \quad (1)$$

where $\Gamma(g) = \{i : A_{ig} = 1\}$ denotes the set of patients in which the gene g is mutated and $\Gamma(M) = \cup_{g \in M}\Gamma(g)$, $|\Gamma(M)|$ measures the coverage of M and $\omega(M) = \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|$ measures the coverage overlap of M.

In addition to the stochastic MCMC search procedure [18], a BLP model has been introduced to solve this problem exactly [19]. Inspired by the BLP model, a binary linear programming model CoMDP to discover co-occurring driver pathways was developed here (Figure 1). The focus was placed on finding possible cooperative driver pathways in carcinogenesis. For example, in Figure 1, the gene set D not C can be detected using MCMC or BLP for $k = 2$. This is because gene set D has higher mutation score W than that of gene set C. Mutations in the gene set B and the gene set C occurred simultaneously among a cohort of patients. CoMDP can successfully identify such co-occurring gene sets which may have been missed by previous approaches.

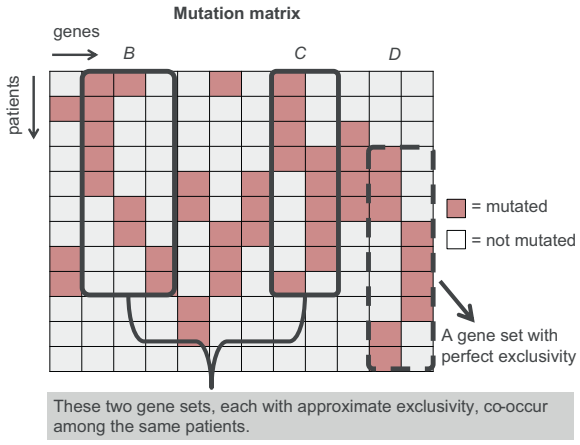

Figure 1 Schematic illustration of driver gene sets (pathways) and their co-occurrence in the mutation matrix. Three gene sets, *B*, *C*, and *D*, with high coverage and high exclusivity are shown. Gene sets *B* and *C* occur simultaneously among the same patients.

CoMDP: a binary linear programming model for the identification of co-occurring driver pathways

For the mutation matrix A, let us consider two submatrices M and N (which correspond to two gene sets or pathways S and T). Given the coverage $\Gamma(M)$ and $\Gamma(N)$ of the two gene sets (sometimes called individual coverage in this study), we define (1) the common coverage $c(M,N) = |\Gamma(M) \bigcap \Gamma(N)|$; (2) the union coverage $b(M,N) = |\Gamma(M) \bigcup \Gamma(N)|$. We further define the non-shared coverage $d(M,N) = b(M,N) - c(M,N)$, which describes the extent of the mutation co-occurrence between the two gene sets: the smaller the value d, the larger the co-occurrence is. As suggested before, $\omega(M)$ and $\omega(N)$ reflect the exclusivity of M and N respectively.

To identify co-occurring gene sets with large coverage and high exclusivity, we introduce the following weight function H:

$$H(M,N) = c(M,N) - d(M,N) - \omega(M) - \omega(N). \quad (2)$$

To maximize this weight function, a binary linear programming model is introduced as follows:

$$\begin{aligned} \max\ G(x,y,z,u,v) = \lambda \sum_{i=1}^{m} z_i + \eta \sum_{i=1}^{m} (x_i + y_i - 2z_i) + \sum_{i=1}^{m} x_i \\ - \sum_{j=1}^{n} \left(u_j \cdot \sum_{i=1}^{m} a_{ij} \right) + \sum_{i=1}^{m} y_i \\ - \sum_{j=1}^{n} \left(v_j \cdot \sum_{i=1}^{m} a_{ij} \right), \end{aligned} \quad (3)$$

$$\text{s.t.}\begin{cases}\frac{1}{n}\sum_{j=1}^{n} a_{ij}u_j \le x_i \le \sum_{j=1}^{n} a_{ij}u_j, & i=1,\ldots,m,\\ \frac{1}{n}\sum_{j=1}^{n} a_{ij}v_j \le y_i \le \sum_{j=1}^{n} a_{ij}v_j, & i=1,\ldots,m,\\ x_i+y_i-1 \le 2z_i \le x_i+y_i, & i=1,\ldots,m,\\ u_j+v_j \le 1,\ j=1,\ldots,n,\\ \sum_{j=1}^{n}(u_j+v_j)=k,\\ x_i,y_i,z_i,u_j,v_j \in \{0,1\},\ i=1,\ldots,m,\ j=1,\ldots,n,\end{cases} \tag{4}$$

where u_j and v_j are indicators whether column j of A falls into the submatrx M or N, so all the columns j's with $u_j = 1$ and $v_j = 1$ constitute M and N respectively; x_i and y_i are indicators whether the entries of row i of M and N are not all zeros, so $\sum_{i=1}^{m} x_i$ and $\sum_{i=1}^{m} y_i$ represent the coverage of M and N (i.e., $|\Gamma(M)|$ and $|\Gamma(N)|$ respectively; z_i is the indicator whether both x_i and y_i equal to 1, so $\sum_{i=1}^{m} z_i$ represents the overlap between the coverage of M and N (i.e., the common coverage $c(M,N)$). k is the total number of genes within S and T; and finally, λ and η are two parameters controlling the common coverage $c(M,N)$ and the non-shared coverage $d(M,N)$ of the two gene sets.

Note that $\sum_{i=1}^{m} z_i$ and $\sum_{i=1}^{m}(x_i + y_i - 2z_i)$ in model (3) are always nonnegative according to the constraints in (4). One can properly set λ and η to be positive or negative to obtain gene sets with specific characteristics. For example, if $\lambda < 0$ and $\eta > 0$, the model tends to detect gene sets with large non-shared coverage but small common coverage under the maximization of $G(x,y,z,u,v)$. Certainly, $\lambda > 0$ and $\eta < 0$ must be set if one wants to identify co-occurring driver pathways by maximizing the function H in (2), which is the main focus of this study. More discussion on the behavior of the model with λ and η can be referred to **Simulation study** below.

mod_CoMDP: Finding a pathway that co-occurs with a known one

In some cases, some prior information is known for a disease. For example, a certain pathway may have been previously proven to play important roles in cancer. The problem is determining whether another pathway with a cooperative effect on tumorigenesis exists. CoDMP can be modified to answer this question to some extent. For a known pathway or a gene set C, a possible co-occurring pathway D can be identified by the following modified optimization problem:

$$\begin{aligned}\max\ \ G_C(y,z,v) = \lambda\sum_{i=1}^{m} z_i + \eta\sum_{i=1}^{m}(y_i-2z_i) + \sum_{i=1}^{m} y_i \\ -\sum_{j=1}^{n}\left(v_j\cdot\sum_{i=1}^{m} a_{ij}\right),\end{aligned} \tag{5}$$

$$\text{s.t.}\begin{cases}\frac{1}{n}\sum_{j=1}^{n} a_{ij}v_j \le y_i \le \sum_{j=1}^{n} a_{ij}v_j, & i=1,\ldots,m,\\ x_i+y_i-1 \le 2z_i \le x_i+y_i, & i=1,\ldots,m,\\ u_j+v_j \le 1,\ j=1,\ldots,n,\\ \sum_{j=1}^{n} v_j = r,\\ y_i,z_i,v_j \in \{0,1\},\ \ i=1,\ldots,m,\ \ j=1,\ldots,n,\end{cases} \tag{6}$$

where x_i and u_j are indicators whether the entries of row i in gene set C are not all zeros and whether the gene corresponding to column j of A falls into C, respectively; y_i, z_i, v_j and the parameters λ and η have the same meaning as in (3) and (4); r is the size of the desired gene set D.

Generally, a branch-and-bound algorithm or others can be used to produce an optimal exact solution for CoMDP (also for mod_CoMDP). In this study, an IBM ILOG CPLEX Optimizer was used to test the effectiveness of the model. The experiments were performed on a 2.50 GHz Core i5-2520M CPU PC. For each given k, CoMDP can automatically identify two gene sets when the sum of their sizes equals k. Although the problem was NP-hard, it can still be solved efficiently due to the sparsity of the mutation matrix.

Statistical significance

A permutation test was used to assess the significance of the results. As in a previous study [18], the weight W in (1) served as a statistic to test the significance of the exclusivity and coverage of each identified gene set (called individual significance). We employed the co-occurrence ratio, which is defined as the ratio of the common coverage to the union coverage, as the statistic to test the significance of the co-occurrence of these two gene sets (called co-occurrence significance).

Simulation data

Three datasets were constructed to illustrate the properties of the proposed method. The first set of simulated data, Sim_data1, was generated as in a previous study [19]. First, an empty m (samples) $\times$ n (genes) matrix was given ($m = 500, n = 1{,}000$ were used). Then, gene sets M_i ($i = 1, \cdots, I$; and each set has 10 genes) with a mutation probability p_i were embedded in the matrix ($p_i = 1 - i\cdot\Delta$, $\Delta = 0.05$, and $I = 10$ were used here). For each sample, a gene uniformly chosen from M_i with p_i was mutated, and once one gene was mutated, the other genes in M_i had a probability p_0 to be mutated (here $p_0 = 0.04$ was used). Finally, the genes not in M_i were mutated in at most three samples. The second dataset, Sim_data2, was generated using the strategy described above for noisy probability p_0 from 0.04 to 0.24 in steps of 0.02.

The third dataset, Sim_data3, was generated as follows. Starting with an empty $r \times s$ matrix (here $r = 600, s = 1,000$ were used), we embedded J gene sets $N_1, N_2, \cdots, N_J$ ($J \geq 1$, here $J = 9$ was used) into it. N_i has size $m_i \times n_0$ (in this study $n_0 = 5$ and $m_i = [m/2] + 2^{i-1}$ were used where $[m/2]$ denotes the integer part of $m/2$). N_i was constructed according to the strategy like that for M_i stated above. Similarly, we mutated the genes not in N_i at most in three samples.

We note that the average mutation rate for genes in a dataset in current simulation study is comparable to those of real datasets. For example, for Sim_data2 with the noisy probabilities of 0.04 and 0.24, each gene has an average mutation rate 0.0142 and 0.0274 respectively. For the four biological datasets (GBM1, GBM2, lung cancer and ovarian cancer) introduced in the following subsection, each dataset has an average mutation rate of 0.0658, 0.0416, 0.0206, and 0.0134 for each gene, respectively.

Biological data

To assess the proposed methods for practical applications, four biological datasets were collected due to their popularity and abundant prior knowledge. Note that the CoMDP can be easily applied to other cancer mutation datasets.

The glioblastoma multiforme data 1 (GBM1) and the lung adenocarcinoma dataset were obtained directly from a previous study [18]. These sets contain mutations in 178 genes across 84 GBM patients (samples) and 356 genes in 163 lung cancer patients, respectively. The GBM2 and ovarian carcinoma datasets were obtained from another previous study [16]. These sets contain CNAs for 1269 genes spanning 169 GBM patients, somatic mutations for 343 genes across 135 GBM patients, CNAs in 966 genes across 559 ovarian patients, and somatic mutations in 8431 genes across 320 ovarian cancer patients. For the last two datasets, somatic mutations and CNAs were first integrated by merging the genes on the common patients. Finally, a binary mutation matrix A was obtained for each of the four datasets. The genes that are mutated in the same samples were combined into a gene set which was named as a metagene in this study. Note that the definition of a metagene differs from that defined based on the matrix factorization method.

Results and discussion

Simulation study

CoMDP can get the optimal solution of the original maximum weight submatrix problem. Like the BLP model [19], CoMDP can detect all the embedded gene sets in Sim_data1 with $k = 10$, $\eta = 1$ and $\lambda < 0$ (here $\lambda = -10$ was used). In the current situation, CoMDP usually degenerates to find one gene set which corresponds to the optimal solution of BLP. Sometimes it can produce two gene sets with no common coverage. For example, with Sim_data1 where ten gene sets were embedded in the 500 (samples) $\times$ 1000 (genes) matrix, we identified two sets with two and eight genes respectively, which are mutated in 74 and 340 samples respectively. But their common coverage is 0. In fact, these two sets constitute one of the embedded gene sets which can be found using the BLP method, so they can be viewed as the same driver gene set as obtained using BLP directly.

Note that as p_0 increases, the exclusivity among the genes in M_i decreases, so the detection of the embedded gene sets M_i becomes more and more difficult. Let $k = 10$. CoMDP ($\lambda = -10$, $\eta = 1$) and BLP were applied on Sim_data2. Both were able to precisely identify all ten embedded gene sets when $p_0 \leq 0.10$. For BLP, the average number of detected embedded gene sets decreased sharply as p_0 increased. However, by properly choosing the parameter η, CoMDP can obtain more accurate and robust results than BLP at high values of p_0. For example, CoMDP with $\lambda = -10$ and $\eta = 2$ has much higher identification accuracy than BLP for $p_0 \geq 0.20$ (Figure 2A).

CoMDP can identify co-occurring gene sets efficiently. We further applied CoMDP to Sim_data3 to demonstrate its effectiveness, and assessed the effect of λ and η. We found that the results are robust with the selection of these two parameters. CoMDP can always get the embedded gene sets with the largest co-occurrence ratio 0.8696 for λ ranging from 6 to 24 in step of 2 and η ranging from -10 to -1 in step of 1. The performance of CoMDP was also demonstrated on different η with $\lambda = 10$ (Figure 2B). For example, when $\eta \leq 3$, the two detected 5-gene sets mutated almost in the same samples (the individual coverage of these two sets was 286 and 273 and common coverage was 260). When η became larger the coverage difference of the two sets increased, and the common coverage became smaller. When $\eta \geq 12$, CoMDP detected one gene set with 10 genes, which had the coverage 437 and so the co-occurrence ratio was 0. Generally speaking, we found that CoMDP has similar performance when λ equals 2 or 10 (Figure 2C). A high co-occurrence ratio (i.e., 0.8696) was obtained when $\eta \leq 0.5$, and a ratio of 0 was obtained when $\eta \geq 3.5$. The simulation study confirmed that $\lambda > 0$ and $\eta < 0$ are the proper selections for identifying co-occurring gene sets. In the following biological applications, without loss of generality, $\lambda = 10$ and $\eta = -2$ were used.

Applications to biological data

In this section, CoMDP was used on four biological datasets (i.e., GBM1, lung cancer, GBM2, and ovarian carcinoma datasets) to identify the co-occurring driver pathways with $k = 4 \sim 10$. We also demonstrated that

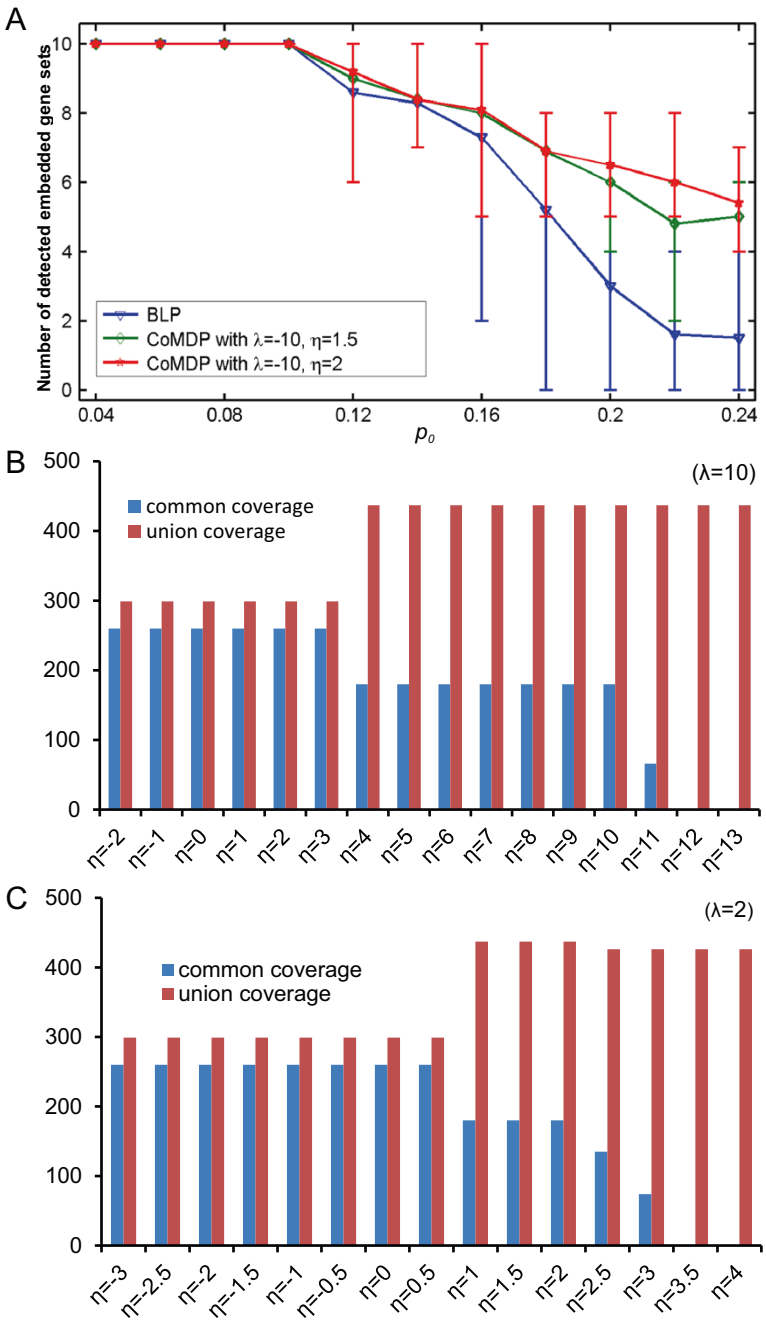

Figure 2 Results of CoMDP applied to simulation data. (A) Comparison of CoMDP with BLP for the ability to identify the embedded gene sets. Display of co-occurrence of the gene sets detected by CoMDP for different η when $\lambda > 0$: **(B)** $\lambda = 10$, and **(C)** $\lambda = 2$.

mod_CoMDP (model (5) and (6)) was applied onto the ovarian carcinoma data to detect more driver pathways co-occurred with *TP53* in carcinogenesis and to find multiple significant co-occurring driver pathways. Each run for GBM1 and GBM2 datasets takes less than two seconds, and each run for the lung cancer dataset takes less than four seconds.

GBM1 dataset

For $k = 4$, two gene sets were detected: {*CDKN2A*, MG_1} and {*MTAP, CYP27B1*} (MG_1 is a metagene consisting of *CDK4, FAM119B, MARCH9, TSFM, CENTG1, METTL1* and *TSPAN31*) with individual significance $p_1 = 0.0207$, $p_2 = 0.0058$, co-occurrence significance $p_{1,2} < 0.0001$, and co-occurrence ratio $r_{1,2} = 0.9412$ (Table 1).

Table 1 Co-occurring gene sets identified by applying CoMDP to GBM1

k	Gene set 1	Gene set 2	p_1	p_2	n_1	n_2	$r_{1,2}$	$p_{1,2}$
4	*CDKN2A*, MG_1	*MTAP, CYP27B1*	0.0207	0.0058	50	49	0.9412	< 0.0001
5	*CDKN2A, TP53,* MG_1	*CDKN2B, CYP27B1*	0.0003	0.0018	68	57	0.7606	< 0.0001
6	*CDKN2A, PTEN,* *CYP27B1*	*CDKN2B, TP53,* MG_1	0.0002	0.0003	69	71	0.8182	< 0.0001
7	*CDKN2A, PTEN,* *CYP27B1*	*CDKN2B, RB1,* *TP53*, MG_1	0.0002	0.0001	69	74	0.8571	< 0.0001
8	*CDKN2A, PTEN,* *NF1, CYP27B1*	*CDKN2B, RB1,* MG_1, *ERBB2*	0.0015	< 0.0001	72	70	0.8933	< 0.0001
9	*CDKN2A, PTEN, NF1,* *CYP27B1, KDR*	*CDKN2B, RB1,* MG_1, *ERBB2*	0.0006	< 0.0001	74	70	0.9200	< 0.0001
10	*CDKN2A, PTEN, CYP27B1,* *KDR*, MG_2	*CDKN2B, NF1, RB1,* MG_1, *ERBB2*	< 0.0001	< 0.0001	72	73	0.9333	< 0.0001

Here p_1 and p_2 are the *p*-values of the individual significance of two identified gene sets, $p_{1,2}$ represents the *p*-value of their co-occurrence significance, n_1 and n_2 denote their respective coverage, and $r_{1,2}$ is the ratio of the common coverage to their union coverage (i.e., co-occurrence ratio). There are same meanings in the following tables. MG_1 is a metagene including seven genes: *CDK4, FAM119B, MARCH9, TSFM, CENTG1, METTL1, TSPAN31*. MG_2 is a metagene including four genes: *WT1, SLC1A2, PAX6, ABCC4*.

The two genes *MTAP* and *CDKN2A* were found to be frequently co-deleted [27,28]. They are both located on chromosome 9p21, a typical tumor suppressor region whose deletion is related to many different types of cancers. *CYP27B1* and the metagene MG_1 were mutated in the same patients with one exception: a single-nucleotide mutation was recorded in one additional patient for *CYP27B1* (Figure 3A). Previous studies have suggested that *CDK4* is the target of a common CNA in the corresponding patients [29]. Two protein products of *CDKN2A*, *INK4A* (also known as *p16*) and *ARF* (also known as *p14*), are involved in the *p53* and *RB* tumor suppressor pathways (Figure 3B). It has been shown that any error disrupting these pathways causes tumor formation [30]. *CDKN2A* and *CDK4* are considered part of the *RB* pathway. Both *MTAP* and *CYP27B1* encode important enzymes. The enzymes encoded by *MTAP* play a major role in polyamine metabolism and those encoded by *CYP27B1* play a role in calcium metabolism and tissue differentiation.

For $k = 5$, two gene sets including {*CDKN2A, TP53,* MG_1} and {*CDKN2B, CYP27B1*} were detected with $p_1 = 0.0003$, $p_2 = 0.0018$, $p_{1,2} < 0.0001$ and $r_{1,2} = 0.7606$. *CDKN2B* encoding *INK4B* (also known as *p15*) also locates in chromosome 9p21 homozygous deletion region, and *CDKN2B* is usually co-deleted with *CDKN2A*. This disrupts the *p53* and *RB* pathways. For this reason, combinatorial inactivation of *CDKN2A* and *CDKN2B* is frequently observed in these tumors. The cross-talk between the *p53* and *RB* pathways (Figure 3B) suggests that *CDKN2A, TP53* and *CDK4* are in the same pathway (Figure 3C(a) or 3C(b)).

For $k = 10$, two gene sets {*CDKN2A, PTEN, CYP27B1, KDR,* MG_2} and {*CDKN2B, NF1, RB1,* MG_1, *ERBB2*} with p_1, p_2 and $p_{1,2}$ less than 0.0001 and $r_{1,2} = 0.9333$ were identified (Table 1 and Figure 4A). The first gene set was found to be involved in the *p53* and *PI3K/Akt* signaling pathways and the second in the *RB* and *RTK/RAS/ERK* signaling pathways. The *RTK/RAS/PI3K* signaling pathway can also be induced by the mutations in these two gene sets (Figure 4B). These pathways are implicated in biological processes associated with cell survival, cell cycle, protein synthesis, and cell proliferation. *p53*, *RB*, and *RTK/RAS/PI3K* have been previously reported to contribute to GBM pathogenesis in original TCGA GBM studies [1]. Five well-known tumor suppressors (*CDKN2A, CDKN2B, PTEN, NF1, and RB1*) are involved in these two gene sets. Besides the co-occurrence of *CDKN2A* and *CDKN2B*, *NF1* and *RB1* in the second gene set have exclusive mutations, which are co-occurrent with mutations of *PTEN* in the first set (Figure 4A). Recently, several studies have shown the cooperativity of tumor suppressors in carcinogenesis [33-35]. For example, Rahrmann *et al.* demonstrated that co-occurring mutations in *PTEN* and *NF1* cooperate in the development of grade 3 PNSTs (peripheral nerve sheath tumors) in mice, suggesting that they may cooperate in human MPNST (malignant PNST) progression [33]. Another study by Chow *et al.* [34] showed that cooperativity among *PTEN*, *TP53*, and *RB1* can cause high-grade astrocytoma in mouse adult brain, in which the majority of glioblastomas arise. For another two almost simultaneously mutated oncogenes *CYP27B1* and *CDK4* (Figure 4A), Beckner *et al.* have demonstrated their cooperative amplification and

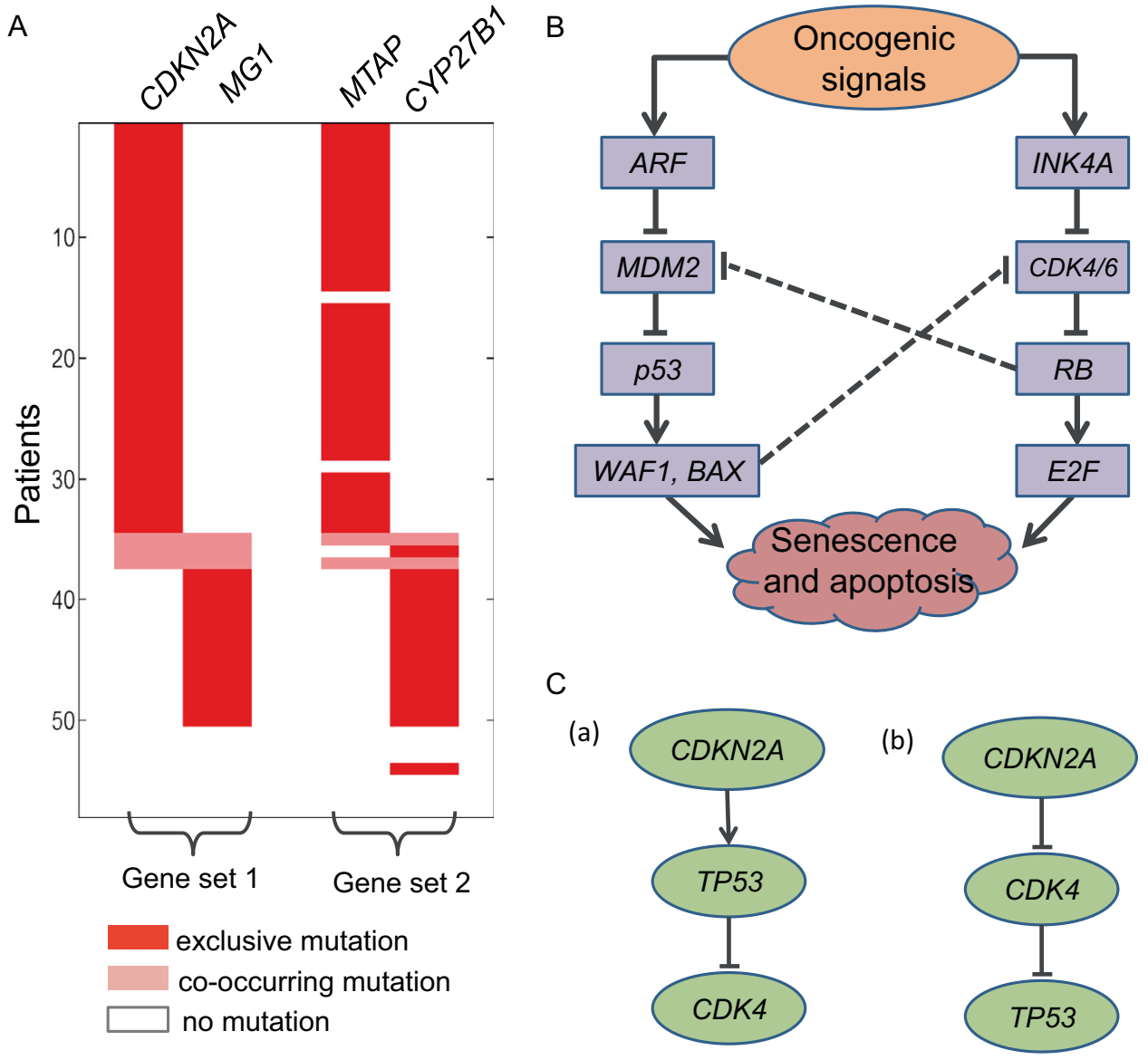

Figure 3 The results of CoMDP on the GBM1 dataset with $k = 4$ and $k = 5$, and the related *p53* and *RB* pathways. (A) The co-occurrence between the two gene sets {*CDKN2A*, MG_1} and {*MTAP*, *CYP27B1*} identified by CoMDP with $k = 4$ in the GBM1 dataset. **(B)** The *TP53* and *RB* tumor suppressor pathways. *ARF* and *INK4A* are two alternatively spliced transcripts of *CDKN2A*. This figure was adapted/extracted according to [31,32]. **(C)** The alterative gene regulatory pathway ((a) or (b)) involving *CDKN2A*, *CDK4* and *TP53*.

co-expression for potential modulation of vitamin D in glioblastomas [36].

On the other hand, *WT1* encodes a transcription factor that plays an essential role in cellular development and cell survival [37]. It regulates the expression of numerous target genes, including the famous tumor suppressor *TP53* and the *Wnt* signaling pathway [38]. *KDR* encodes a *VEGF* (vascular endothelial growth factor) receptor. *VEGF* plays a crucial role in angiogenesis and progression of malignant brain tumors. *ERBB2* encoding the protein *HER2* (human epidermal growth factor receptor 2) is a member of the epidermal growth factor receptor (*EGFR*) family, and it has been shown to play an important role in the pathogenesis and progression of many different types of cancer. *WT1*, *KDR* and *ERBB2* may drive the carcinogenesis of GBM, indicating that CoMDP can identify low-frequency candidate driver genes that play important roles in cancer initiation and development.

Lung cancer

In this case, the significant results were obtained with $k = 4, 5, 10$ (Table 2). For $k = 4$ the co-occurring gene sets

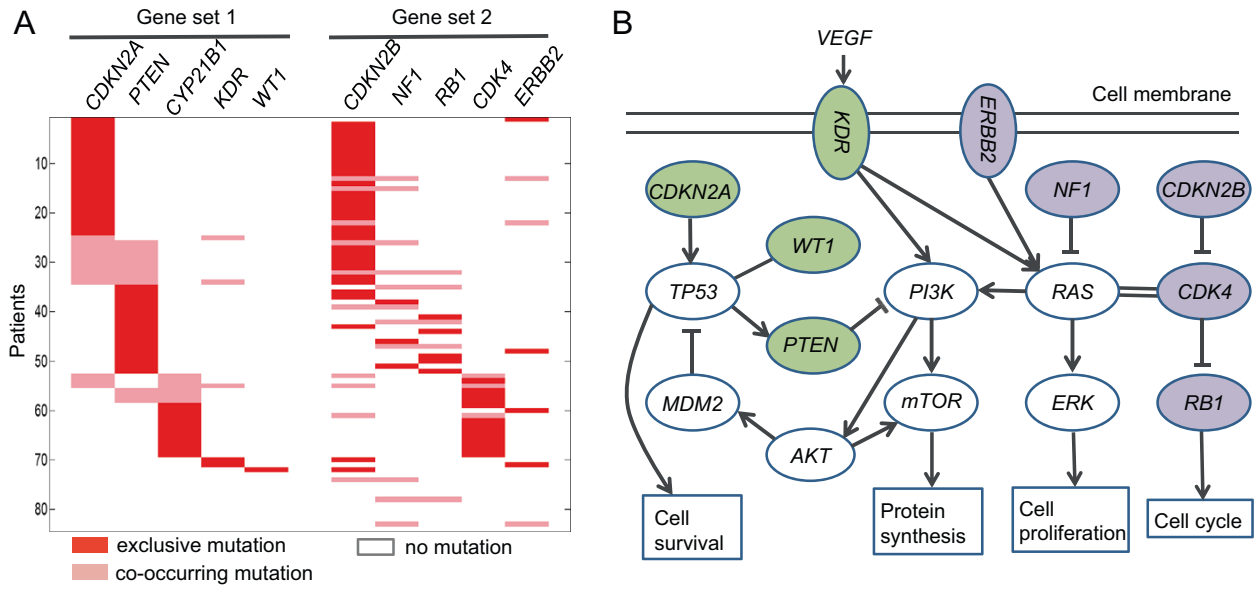

Figure 4 The results of CoMDP on the GBM1 dataset with $k = 10$. (A) The co-occurrence between the two gene sets {*CDKN2A, PTEN, CYP27B1, KDR, WT1*} and {*CDKN2B, NF1, RB1, CDK4, ERBB2*}. **(B)** Several pathways are involved, including the *p53*, *PI3K/Akt*, *RTK/RAS/ERK*, *RB* and *RTK/RAS/PI3K* signaling pathways. The line between *p53* and *WT1* denotes they are physically interacted, whereas the double lines between *KRAS* and *CDK4* represents sometimes there is a synthetic lethal interaction between them. The green and blue nodes denote the genes involved in the two detected co-occurring gene sets. The regulatory relations are extracted from the KEGG database and related literature.

are {*ATM, TP53*} and {*EGFR, KRAS*}, and for $k = 5$ they are {*ATM, TP53*} and {*EGFR, KRAS, STK11*}. As stated in a previous study [18], *ATM* and *TP53* interact directly and are involved in the cell cycle checkpoint control [39]. *EGFR, KRAS* and *STK11* are all involved in the regulation of the *mTOR* signaling pathway, whose dysregulation has been reported to be important to lung adenocarcinoma [8]. However, the gene set {*ATM, TP53*} can only be obtained by removing the mutations of {*EGFR, KRAS, STK11*} from the dataset in the previous study by Vandin [18]. Here, these two gene sets were identified simultaneously and found to show significant co-occurrence. *ATM, TP53* are involved in the regulation of cell apoptosis and *EGFR, KRAS, STK11* are related to protein synthesis, indicating that the cooperativity of these two processes for the generation and progression of lung cancer.

For $k = 10$, {*STK11, ATM, TP53, PAK4*} and {*KRAS, NTRK3, EGFR, GNAS, EPHA3, NRAS*} were identified. All four genes *STK11, ATM, TP53, PAK4* have been demonstrated to be closely related to the *p53* signaling pathway in lung cancer [40,41]. Two members of the *RAS* subfamily, *KRAS* and *NRAS* function as binary molecular switches controlling the intracellular signaling networks that regulate several key cancer-related processes, such as proliferation, differentiation, cell adhesion, apoptosis, and cell migration. *GNAS* is a guanine nucleotide-binding protein (G protein). It acts as a modulator or transducer in various transmembrane signaling systems. *GNAS* may interact with *MDM2*, which may lead to *MDM2*-mediated degradation of *TP53*. Solomon *et al.* found that many kinds of *TP53* mutations can regulate *RAS* in different ways, inducing a cancer-related gene signature [42]. Kosaka *et al.* demonstrated that *TP53, EGFR* and *KRAS* may cooperatively determine the prognosis of the patients in lung adenocarcinoma [43].

GBM2 dataset

We observed that some new co-occurring gene sets were identified for GBM2 compared to GBM1 (Table 3). For $k = 9$, we identified {*CDKN2A, TP53, MG*$_3$*, PIK3R1,*

Table 2 Co-occurring gene sets identified by applying CoMDP to the lung cancer data

k	Gene set 1	Gene set 2	p_1	p_2	n_1	n_2	$r_{1,2}$	$p_{1,2}$
4	*ATM, TP53*	*KRAS, EGFR*	0.0121	< 0.0001	76	90	0.3833	0.0129
5	*ATM, TP53*	*STK11, KRAS, EGFR*	0.0125	< 0.0001	76	110	0.4091	0.0220
10	*STK11, ATM, TP53, PAK4*	*KRAS, NTRK3, EGFR, GNAS, EPHA3, NRAS*	0.0379	< 0.0001	98	108	0.5489	0.0001

Table 3 Co-occurring gene sets identified by applying CoMDP to GBM2 with somatic mutations and CNAs

k	Gene set 1	Gene set 2	p_1	p_2	n_1	n_2	$r_{1,2}$	$p_{1,2}$
4	*CDKN2A, MG*$_3$	*CDKN2B, CYP27B1*	0.0045	0.0056	60	63	0.9219	< 0.0001
5	*CDKN2A, CYP27B1, COL1A2*	*CDKN2B, MG*$_3$	0.0072	0.0018	62	63	0.9531	< 0.0001
6	*CDKN2A, CYP27B1, COL1A2*	*CDKN2B, MG*$_3$*, ERBB2*	0.0073	0.0003	62	65	0.9538	< 0.0001
7	*CDKN2A, TP53, MG*$_3$*, TAF1*	*CDKN2B, RB1, CYP27B1*	0.0001	0.0001	77	70	0.8375	< 0.0001
8	*CDKN2A, TP53, MG*$_3$*, TAF1*	*CDKN2B, RB1, CYP27B1, PRNP*	0.0002	< 0.0001	77	71	0.8500	< 0.0001
9	*CDKN2A, TP53, TAF1, MG*$_3$*, PIK3R1*	*CDKN2B, RB1, CYP27B1, SYNE1*	0.0002	< 0.0001	79	72	0.8642	< 0.0001
10	*CDKN2A, TP53, TAF1, MG*$_3$*, PIK3R1*	*CDKN2B, RB1, SYNE1, CYP27B1, MG*$_4$	0.0001	< 0.0001	79	73	0.8765	< 0.0001

MG$_3$ is a metagene including three genes: *CDK4, MARCH9, TSPAN31*. *MG*$_4$ is a metagene including 168 genes.

TAF1} and {*CDKN2B, CYP27B1, RB1, SYNE1*} (*MG*$_3$ is a metagene including *CDK4, MARCH9* and *TSPAN31*) with $p_1 = 0.0002$, $p_2 < 0.0001$, $p_{1,2} < 0.0001$ and $r_{1,2} = 0.8642$ (Figure 5A and Table 3). In addition to the cooperative effects between *CDKN2A* and *CDKN2B*, *TP53* and *RB1* have been reported to have frequently co-occurring mutations related to several cancers, including the central nervous system tumor [13]. Recently, the collaboration of *TP53* and *CDKN2B* was also studied with respect to cell apoptosis and aneurysm formation [44]. On the other hand, for the two detected low-frequency mutated genes *TAF1* (2/170) and *SYNE1* (3/170), *TAF1* encoding a transcription initiation factor phosphorylates *TP53* during G1 cell-cycle progression, so *TAF1* may be a member of the *p53* signaling pathway; *SYNE1* was found to be associated with the GBM patients' lifetime, and was therefore considered to be an important biomarker of glioblastoma survival [45]. Our studies indicated that

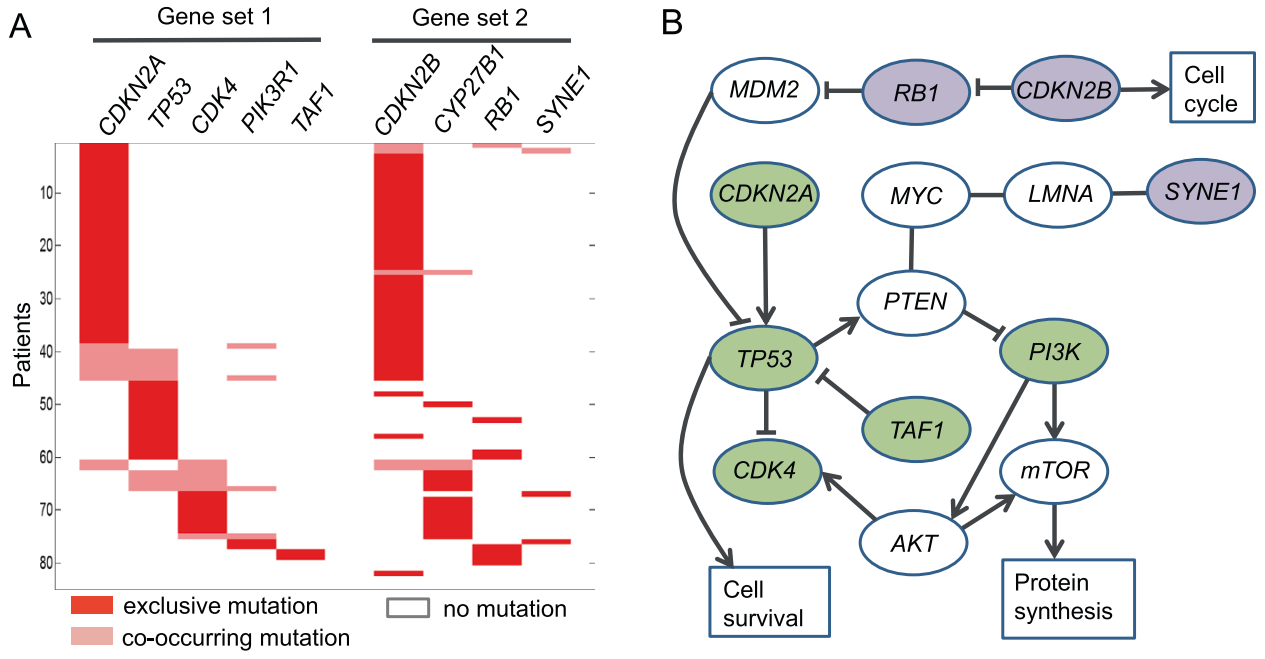

Figure 5 The results of CoMDP on the GBM2 dataset with $k = 9$. (A) The significant co-occurrence mutations between the two gene sets {*CDKN2A, TP53, MG*$_3$*, PIK3R1, TAF1*} and {*CDKN2B, CYP27B1, RB1, SYNE1*}. **(B)** The relevant pathways of the two gene sets.

the *p53*, *RB*, and the *PI3K*-related signaling pathways may collaboratively contribute to carcinogenesis in GBM via combined genetic alterations (Figure 4B and Figure 5B).

Ovarian cancer

The mutation distribution among genes in the ovarian carcinoma data is quite nonuniform. Among all the 314 samples, *TP53* is mutated in 251 of them and all the other genes are mutated in less than 26% of samples. This indicates that *TP53* plays a crucial role in the carcinogenesis of ovarian cancer (*TTN* was removed in the present analysis because of the possible artifacts of its mutations [46]). Determining whether there are other driver genes or pathways collaborating with *TP53* will be helpful for understanding the pathogenesis of this cancer.

We applied CoMDP to the ovarian cancer data with $k = 4 \sim 10$ (Table 4). The first three rows in Table 4 showed significantly co-occurring gene sets with *TP53*. For $k = 5$ we identified {*MYC, CCNE1, NINJ2,* MG_5} (MG_5 is a metagene including *CHKB* and *KLHDC7B*). *MYC* and *CCNE1* are two important proto-oncogenes involved in cell cycle progression. The functional correlation of *MYC* and *TP53* in the carcinogenic progression of ovarian carcinoma and other cancers have been evaluated in several studies [47-49]. Recently, Kuhn *et al.* discovered that molecular genetic aberrations of *CCNE1* together with those of the *p53* and *PI3K* pathways are major mechanisms in the development of uterine serous carcinoma [50]. Both *CHKB* and *KLHDC7B* are located on chromosome 22q13.33, where *KLHDC7B* is involved in breast cancer and lymph node metastasis in cervical cancer and *CHKB* encodes choline kinase (*ChoK*) beta. de Molina *et al.* demonstrated that *ChoK* acts as a link connecting phospholipid metabolism and cell cycle regulation [51]. It is here supposed that *TP53* and *CHKB* may regulate CDK4/6 collaboratively to suppress the progression of ovarian cancer.

To identify other driver gene sets coupled to *TP53*, we applied mod_CoMDP with $r = 1 \sim 10$ to the ovarian cancer data and significant results were obtained for $r = 3 \sim 10$ (Additional file 1: Table S1). For example, for $r = 10$ we identified {*MYC, CCNE1, NINJ2,* MG_5, *USH2A, NF1, HMCN1, ZNF596, USP35,* MG_6} (MG_6 is a metagene including four genes: *STMN3, SLC2A4RG, ZGPAT, RTEL1*) with $ra = 0.6563$ (the co-occurrence ratio with *TP53*). Frequent somatic mutations in *NF1* have been previously shown to co-occur with *TP53* mutations in ovarian carcinomas [52,53]. *STMN3* and *NF1* have been demonstrated to be involved in the *MAPK* signaling pathway [19]. Furthermore, to discover possible collaborations of multiple driver pathways with *TP*53, we combined *TP53* and the aforementioned 10-gene set into one nominal gene, which was considered mutated in a sample if both sets were mutated in that sample. Then we applied mod_CoMDP to identify gene sets significantly co-occurring with the nominal gene. For $r = 1 \sim 10$ we identified *PPP2R2A*, which is generally implicated in the negative control of cell growth and division. Kalev *et al.* revealed that *PPP2R2A* plays a critical role in DNA double-strand break repair through modulation of *ATM* phosphorylation [54]. Youn and Simon recently studied mutator alterations relevant to ovarian cancer [55].

Table 4 Co-occurring gene sets identified by applying CoMDP to the ovarian carcinoma dadaset

k	Gene set 1	Gene set 2	p_1	p_2	n_1	n_2	$r_{1,2}$	$p_{1,2}$
4	*TP53*	*MYC, CCNE1, NINJ2*	1.0000	< 0.0001	251	155	0.4397	0.0410
5	*TP53*	*MYC, CCNE1, NINJ2,* MG_5	1.0000	0.0100	251	169	0.4894	0.0250
6	*TP53*	*MYC, CCNE1, NINJ2, ZNF596, USH2A*	1.0000	< 0.0001	251	183	0.5228	0.0120
7	*TP53, LYRM5*	*MYC, CCNE1, NINJ2, ZNF596, USH2A*	0.0270	0.0030	264	183	0.5629	< 0.0001
8	*TP53, LYRM5*	*MYC, CCNE1, NINJ2, BRD4, ZNF596, USH2A*	0.0230	0.0210	264	197	0.6007	0.001
9	*TP53, LYRM5*	*MYC, CCNE1, NINJ2, BRD4, ZNF596, USH2A, HMCN1*	0.0340	0.0160	264	206	0.6263	< 0.0001
10	*TP53, LYRM5*	*MYC, CCNE1, NINJ2, NF1, ZNF596, USH2A, HMCN1, TPD52L2*	0.0390	0.001	264	211	0.6493	< 0.0001

MG_5 is a metagene including two genes: *CHKB, KLHDC7B*. For $k = 4 \sim 6$ because of only one gene contained in the gene set 1, the corresponding *p*-value equals 1.0000.

Besides the well-known mutator gene *TP53*, they identified *PPP2R2A* and the chromosomal region 22q13.33 as the new mutator candidates. We find that these so called mutator genes, which increase genomic instability when altered, may be collaboratively involved in the processes of DNA synthesis and repair, chromosome segregation, damage surveillance, cell cycle checkpoints, and apoptosis. The discovered driver patterns here may provide new information to enhance our understanding of the ovarian carcinoma pathogenesis, and further explorative analysis is needed to verify their biological relevance.

Conclusions

In this study, we proposed a method CoMDP for the *de novo* identification of co-occurring driver pathways in cancer. It considers two types of optimization simultaneously: First, it makes the maximization of the weight W for each individual pathway, i.e., high coverage and high exclusivity. Second, it ensures that the maximization of the inter-overlap between the pathway pair. Simulation study indicated that for a range of values of the parameters λ and η, CoMDP can always get the exact solution. It was here demonstrated that CoMDP has the following characteristics: (1) It can identify individual driver gene sets as BLP [19] or Dendrix [18]. (2) It obtains more accurate and robust results when the noise increases. (3) It uses no prior information such as the incomplete knowledge about the pathways and protein interaction networks. (4) CoMDP is an exact method and the procedure is quite fast.

When the project approximated to the end, we noticed that Leiserson *et al.* proposed a method for the simultaneous identification of multiple driver pathways [56]. The present study is related to Leiserson's but in many ways quite different. First, the so-called Dendrix$_{ILP}$ in [56] is the same as the BLP method [19]. Second, the weight function used by their Multi-Dendrix algorithm does not explicitly incorporate co-occurrence of mutations between genes in different pathways [56]. The Multi-Dentrix of the two gene sets was found to be a special form of the present model with $\lambda = 2$ and $\eta = 1$. Our simulation study demonstrated that, in this case, the coverage of the two sets detected was 286 and 331 respectively. Although the union coverage got larger (i.e., 437), a lower co-occurrence ratio 0.4119 was obtained because of the smaller common coverage. This also indicates that the multiple driver pathways (gene sets) identified by Multi-Dendrix cannot be guaranteed to be co-occurring.

We note that the heterogeneity among tumors can affect the findings of the current method. Investigating combinatorial patterns of driver pathways in different subtypes will be helpful for understanding the molecular mechanisms of carcinogenesis and designing efficient treatments for cancer patients. It will be interesting to explore the effect of heterogeneity among tumors in the future studies.

In summary, we have developed a method to identify co-occurring driver pathways, which may reveal the functional cooperation of different driver pathways during carcinogenesis. The results of this study show that the present method will be a powerful tool to explore the collaborative effects among mutated driver pathways and enhance our understanding of the molecular mechanisms.

Additional file

Additional file 1: The results by applying mod_CoMDP to the ovarian carcinoma dadaset.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JZ and SZ conceived and designed this study; JZ implemented the algorithm and carried out the experiment; JZ, LYW and SZ analyzed the data and wrote the paper. XSZ supervised this project and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This project was supported by the National Natural Science Foundation of China, No. 61379092 and 11131009, the Foundation for Members of Youth Innovation Promotion Association, CAS, The Outstanding Young Scientist Program of CAS, the Scientific Research Foundation for ROCS, SEM, and the Key Laboratory of Random Complex Structures and Data Science, CAS.

Received: 2 March 2014 Accepted: 1 August 2014
Published: 9 August 2014

References

1. The Cancer, Genome Atlas Research Network: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature* 2008, **455:**1061–1068.
2. International Cancer Genome Consortium: **International network of cancer genome projects.** *Nature* 2010, **464:**993–998.
3. Zhang S, Liu CC, Li W, Shen H, Laird P, Zhou XJ: **Discovery of multi-dimensional modules by integrative analysis of cancer genomic data.** *Nucleic Acids Res* 2012, **40:**9379–9391.
4. Liu Z, Zhang XS, Zhang S: **Breast tumour subgroups reveal diverse clinical predictive power.** *Sci Rep* 2014, **4:**4002.
5. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, *et al.*: **Patterns of somatic mutation in human cancer genomes.** *Nature* 2007, **446:**153–158.
6. Stratton MR, Campbell PJ, Futreal PA: **The cancer genome.** *Nature* 2009, **458:**719–724.
7. Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, Vivanco I, Lee JC, Huang JH, Alexander S, Du J, Kau T, Thomas RK, Shah K, Soto H, Perner S, Prensner J, Debiasi RM, Demichelis F, Hatton C, Rubin MA, Garraway LA, Nelson SF, Liau L, Mischel PS, Cloughesy TF, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, *et al.*: **Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma.** *Proc Natl Acad Sci* 2007, **104:**20007–20012.
8. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, *et al.*: **Somatic mutations affect key pathways in lung adenocarcinoma.** *Nature* 2008, **455:**1069–1075.

9. Jones S, Zhang X, Parsons DW, Lin JCH, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, Hong SM, Fu B, Lin MT, Calhoun ES, Kamiyama M, Walter K, Nikolskaya T, Nikolsky Y, Hartigan J, Smith DR, Hidalgo M, Leach SD, Klein AP, Jaffee EM, Goggins M, Maitra A, Iacobuzio-Donahue C, Eshleman JR, Kern SE, Hruban RH, *et al.*: **Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.** *Science* 2008, **321:**1801–1806.
10. Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G: **Patient-oriented gene set analysis for cancer mutation data.** *Genome Biol* 2010, **11:**R112.
11. Efroni S, Ben-Hamo R, Edmonson M, Greenblum S, Schaefer CF, Buetow KH: **Detecting cancer gene networks characterized by recurrent genomic alterations in a population.** *PLoS ONE* 2011, **6:**e14437.
12. Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control.** *Nat Med* 2004, **10:**789–799.
13. Yeang CH, McCormick F, Levine A: **Combinatorial patterns of somatic gene mutations in cancer.** *FASEB J* 2008, **22:**2605–2622.
14. Ciriello G, Cerami E, Sander C, Schultz N: **Mutual exclusivity analysis identifies oncogenic network modules.** *Genome Res* 2012, **22:**398–406.
15. Miller CA, Settle SH, Sulman EP, Aldape KD, Milosavljevic A: **Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors.** *BMC Med Genomics* 2011, **4:**34.
16. Zhang J, Zhang S, Wang Y, Zhang XS: **Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data.** *BMC Syst Biol* 2013, **7:**S4.
17. Acharya LR, Judeh T, Duan Z, Rabbat MG, Zhu D: **GSGS: a computational approach to reconstruct signaling pathway structures from gene sets.** *IEEE/ACM Trans Comput Biol Bioinformatics* 2012, **9:**438–450.
18. Vandin F, Upfal E, Raphael BJ: **De novo discovery of mutated driver pathways in cancer.** *Genome Res* 2012, **22:**375–385.
19. Zhao J, Zhang S, Wu LY, Zhang XS: **Efficient methods for identifying mutated driver pathways in cancer.** *Bioinformatics* 2012, **28:**2940–2947.
20. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation.** *Cell* 2011, **144:**646–674.
21. Cui Q, Ma Y, Jaramillo M, Bari H, Awan A, Yang S, Zhang S, Liu L, Lu M, O'Connor- McCourt M, Purisima EO, Wang E: **A map of human cancer signaling.** *Mol Syst Biol* 2007, **3:**152.
22. Klijn C, Bot J, Adams DJ, Reinders M, Wessels L, Jonkers J: **Identification of networks of co-occurring, tumor-related DNA copy number changes using a genome-wide scoring approach.** *PLoS Comput Biol* 2010, **6:**e1000631.
23. Kumar N, Rehrauer H, Cai H, Baudis M: **Cdcoca: a statistical method to define complexity dependence of co-occuring chromosomal aberrations.** *BMC Med Genom* 2011, **4:**21.
24. Forbes S, Clements J, Dawson E, Bamford S, Webb T, Dogan A, Flanagan A, Teague J, Wooster R, Futreal PA, Stratton MR: **COSMIC 2005.** *Br J Cancer* 2006, **94:**318–322.
25. Gu Y, Zhao W, Xia J, Zhang Y, Wu R, Wang C, Guo Z: **Analysis of pathway mutation profiles highlights collaboration between cancer-associated superpathways.** *Hum Mutat* 2011, **32:**1028–1035.
26. Gu Y, Wang H, Qin Y, Zhang Y, Zhao W, Qi L, Zhang Y, Wang C, Guo Z: **Network analysis of genomic alteration profiles reveals co-altered functional modules and driver genes for glioblastoma.** *Mol Biosyst* 2013, **9:**467–477.
27. Gursky S, Olopade OI, Rowley JD: **Identification of a 1.2 Kb cDNA fragment from a region on 9p21 commonly deleted in multiple tumor types.** *Cancer Genet Cytogenet* 2001, **129:**93–101.
28. Christopher SA, Diegelman P, Porter CW, Kruger WD: **Methylthioadenosine phosphorylase, a gene frequently codeleted with p16(cdkN2a/ARF), acts as a tumor suppressor in a breast cancer cell line.** *Cancer Res* 2002, **62:**6639–6644.
29. Wikman H, Nymark P, Vayrynen A, Jarmalaite S, Kallioniemi A, Salmenkivi K, Vainio-Siukola K, Husgafvel-Pursiainen K, Knuutila S, Wolf M, Anttila S: **CDK4 is a probable target gene in a novel amplicon at 12q13.3-q14.1 in lung cancer.** *Gene Chromosome Cancer* 2005, **42:**193–199.
30. Mays-Hoopes LL: **Aging and cell division.** *Nat Educ* 2010, **3:**55.
31. Campisis J: **Cancer and ageing: rival demons?** *Nat Rev Cancer* 2003, **3:**339–349.
32. Kabbarah O, Chin L: **Advances in malignant melanoma: genetic insights from mouse and man.** *Front Biosci* 2006, **11:**928–942.
33. Rahrmann EP, Watson AL, Keng VW, Choi K, Moriarity BS, Beckmann DA, Wolf NK, Sarver A, Collins MH, Moertel CL, Wallace MR, Gel B, Serra E, Ratner N, Largaespada1 DA: **Forward genetic screen for malignant peripheral nerve sheath tumor formation identifies new genes and pathways driving tumorigenesis.** *Nat Genet* 2013, **45:**756–766.
34. Chow LML, Endersby R, Zhu X, Rankin S, Qu C, Zhang J, Broniscer A, Ellison DW, Baker SJ: **Cooperativity within and among Pten, p53, and Rb pathways induces high-grade astrocytoma in adult brain.** *Cancer Cell* 2011, **19:**305–316.
35. Xing F, Persaud Y, Pratilas CA, Taylor BS, Janakiraman M, She QB, Gallardo H, Liu C, Merghoub T, Hefter B, Dolgalev I, Viale A, Heguy A, De Stanchina E, Cobrinik D, Bollag G, Wolchok J, Houghton A, Solit DB: **Concurrent loss of the PTEN and RB1 tumor suppressors attenuates RAF dependence in melanomas harboring (V600E)BRAF.** *Oncogene* 2012, **31:**446–457.
36. Beckner ME, Patil S, LeBlanc JF, Katira K, Nanda A, Martin SS, Brunson LE, Truong LN, Nordberg ML: **MLPA and REMBRANDT data predict potential modulation of vitamin D via increased CYP27B1 in aggressive primary brain tumors.** *Cancer Res* 2010, **70**(8 Suppl):Abstract 1573.
37. Chau YY, Hastie ND: **The role of Wt1 in regulating mesenchyme in cancer, development, and tissue homeostasis.** *Trends Genet* 2012, **28:**515–524.
38. Kim MKH, McGarry TJ, Broin PO, Flatow JM, Golden AAJ, Licht JD: **An integrated genome screen identifies the Wnt signaling pathway as a major target of WT1.** *Proc Natl Acad Sci* 2009, **106:**11154–11159.
39. Chehab NH, Malikzay A, Appel M, Halazonetis TD: **Chk2/hCds1 functions as a DNA damage checkpoint in G1 by stabilizing p53.** *Genes Dev* 2000, **14:**278–288.
40. Greulich H: **The genomics of lung adenocarcinoma: opportunities for targeted therapies.** *Genes cancer* 2010, **1:**1200–1210.
41. Kesanakurti D, Chetty C, Rajasekhar Maddirela D, Gujrati M, Rao JS: **Functional cooperativity by direct interaction between PAK4 and MMP-2 in the regulation of anoikis resistance, migration and invasion in glioma.** *Cell Death Dis* 2012, **3:**e445.
42. Solomon H, Buganim Y, Kogan-Sakin I, Pomeraniec L, Assia Y, Madar S, Goldstein I, Brosh R, Kalo E, Beatus T, Goldfinger N, Rotter V: **Various p53 mutant proteins differently regulate the Ras circuit to induce a cancer-related gene signature.** *J Cell Sci* 2012, **125:**3144–3152.
43. Kosaka T, Yatabe Y, Onozato R, Kuwano H, Mitsudomi T: **Prognostic implication of EGFR, KRAS, and TP53 gene mutations in a large cohort of Japanese patients with surgically treated lung adenocarcinoma.** *J Thorac Oncol* 2009, **4:**22–29.
44. Leeper NJ, Raiesdana A, Kojima Y, Kundu RK, Cheng H, Maegdefessel L, Toh R, Ahn GO, Ali ZA, Anderson DR, Miller CL, Roberts SC, Spin JM, de Almeida PE, Wu JC, Xu B, Cheng K, Quertermous M, Kundu S, Kortekaas KE, Berzin E, Downing KP, Dalman RL, Tsao PS, Schadt EE, Owens GK, Quertermous T: **Loss of CDKN2B promotes p53-dependent smooth muscle cell apoptosis and aneurysm formation.** *Arterioscler Thromb Vasc Biol* 2013, **33:**e1–e10.
45. Serão NVL, Delfino KR, Southey BR, Beever JE, Rodriguez-Zas SL: **Cell cycle and aging, morphogenesis, and response to stimuli genes are individualized biomarkers of glioblastoma progression and survival.** *BMC Med Genomics* 2011, **4:**49.
46. The Cancer Genome Atlas Research Network: **Integrated genomic analyses of ovarian carcinoma.** *Nature* 2011, **474:**609–615.
47. Plisiecka-Halasa J, Karpinska G, Szymanska T, Ziolkowska I, Madry R, Timorek A, Debniak J, Ulanska M, Jedryka M, Chudecka-Glaz A, Klimek M, Rembiszewska A, Kraszewska E, Dybowski B, Markowska J, Emerich J, Pluzanska A, Goluda M, Rzepka-Gorska I, Urbanski K, Zielinski J, Stelmachow J, Chrabowska M, Kupryjanczyk J: **$P21^{WAF1}$, $P27^{KIP1}$, TP53 and C-MYC analysis in 204 ovarian carcinomas treated with platinum-based regimens.** *Ann Oncol* 2003, **14:**1078–1085.
48. Yeung SJ, Pan J, Lee MH: **Roles of p53, Myc and HIF-1 in regulating glycolysis - the seventh hallmark of cancer.** *Cell Mol Life Sci* 2008, **65:**3981–3999.
49. Calcagno DQ, Freitas VM, Leal MF, de Souza CRT, Demachki S, Montenegro R, Assumpcao PP, Khayat AS, Smith MAC, dos Santos AKCR, Burbano RR: ***MYC, FBXW7* and *TP53* copy number variation and expression in Gastric Cancer.** *BMC Gastroenterol* 2013, **13:**141.

50. Kuhn E, Wu RC, Guan B, Wu G, Zhang J, Wang Y, Song L, Yuan X, Wei L, Roden RBS, Kuo KT, Nakayama K, Clarke B, Shaw P, Olvera N, Kurman RJ, Levine DA, Wang TL, Shih IM: **Identification of molecular pathway aberrations in uterine serous carcinoma by genome-wide analyses.** *J Natl Cancer Inst* 2012, **104:**1503–1513.
51. de Molina AR, Gallego-Ortega D, Sarmentero-Estrada J, Lagares D, del Pulgar TG, Eva Bandrés E, García-Foncillas J, Lacal JC: **Choline kinase as a link connecting phospholipid metabolism and cell cycle regulation: Implications in cancer therapy.** *Int J Biochem Cell Biol* 2008, **40:**1753–1763.
52. Sangha N, Wu R, Kuick R, Powers S, Mu D, Fiander D, Yuen K, Katabuchi H, Tashiro H, Fearon ER, Cho KR: **Neurofibromin 1 (NF1) defects are common in human ovarian serous carcinomas and co-occur with TP53 mutations.** *Neoplasia* 2008, **10:**1362–1372.
53. Zhang J, Shi Y, Lalonde E, Li L, Cavallone L, Ferenczy A, Gotlieb WH, Foulkes WD, Majewski J: **Exome profiling of primary, metastatic and recurrent ovarian carcinomas in a BRCA1-positive patient.** *BMC Cancer* 2013, **13:**146.
54. Kalev P, Simicek M, Vazquez I, Munck S, Chen L, Soin T, Danda N, Chen W, Sablina A: **Loss of PPP2R2A inhibits homologous recombination DNA repair and predicts tumor sensitivity to PARP inhibition.** *Cancer Res* 2012, **72:**6414–6424.
55. Youn A, Simon R: **Using passenger mutations to estimate the timing of driver mutations and identify mutator alterations.** *BMC Bioinformatics* 2013, **14:**363.
56. Leiserson MDM, Blokh D, Sharan R, Raphael BJ: **Simultaneous identification of multiple driver pathways in cancer.** *PLoS Comput Biol* 2013, **9:**e1003054.

doi:10.1186/1471-2105-15-271
Cite this article as: Zhang *et al.*: Discovery of co-occurring driver pathways in cancer. *BMC Bioinformatics* 2014 **15**:271.