Discovery of cancer common and specific driver gene sets

Junhua Zhang and Shihua Zhang

Abstract—Cancer is known as a disease mainly caused by gene alterations. Discovery of mutated driver pathways or gene sets is becoming an important step to understand molecular mechanisms of carcinogenesis. However, systematically investigating commonalities and specificities of driver gene sets among multiple cancer types is still a great challenge, but this investigation will undoubtedly benefit deciphering cancers and will be helpful for personalized therapy and precision medicine in cancer treatment. In this study, we propose two optimization models to *de novo* discover common driver gene sets among multiple cancer types (ComMDP) and specific driver gene sets of one certain or multiple cancer types to other cancers (SpeMDP), respectively. We first apply ComMDP and SpeMDP to simulated data to validate their efficiency. Then, we further apply these methods to 12 cancer types from The Cancer Genome Atlas (TCGA) and obtain several biologically meaningful driver pathways. As examples, we construct a common cancer pathway model for BRCA and OV, infer a complex driver pathway model for BRCA carcinogenesis based on common driver gene sets of BRCA with eight cancer types, and investigate specific driver pathways of the liquid cancer lymphoblastic acute myeloid leukemia (LAML) versus other solid cancer types. In these processes more candidate cancer genes are also found.

Index Terms—Bioinformatics, cancer genomics, pan-cancer study, driver pathway, mutual exclusivity

✦

1 INTRODUCTION

C ANCER is a complex and heterogeneous disease with diverse genetic and environmental factors involved in its etiology. With the advances of deep sequencing technology, huge volume cancer genomics data have been generated through several large-scale programs (e.g., The Cancer Genome Atlas (TCGA) [1], International Cancer Genome Consortium (ICGC) [2], and the Cancer Cell Line Encyclopedia (CCLE) [3]), which provide huge opportunities for understanding the molecular mechanisms and pathogenesis underlying cancer [4]. Currently, a crucial challenge in cancer genomics is to distinguish driver mutations and driver genes which contribute to cancer initiation and development from passenger ones which accumulate in cells but do not contribute to carcinogenesis [5], [6]. Most early efforts have been devoted to detect individual driver genes with recurrent mutations [7]. However, this kind of methods do not consider the complicated mutational heterogeneity in cancer genomes with diverse mutations in genes.

Although cancer patients exhibit diverse genomic alterations, many studies have demonstrated that driver mutations tend to affect a limited number of cellular signaling and regulatory pathways [1], [8], [9]. Therefore, a great deal of attention has been devoted to evaluate the recurrence of mutations in groups of genes derived from known pathways or protein-protein interaction networks [9], [10], [11]. These groups of genes are considered as candidate driver pathways, which may be frequently perturbed within tumor cells [12], [13] and can lead to the acquisition of carcinogenic properties such as cell proliferation, angiogenesis, or

• *J. Zhang and S. Zhang are with National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.*
E-mail: zjh@amt.ac.cn, zsh@amss.ac.cn.

metastasis [14], [15]. A main concern is that the human protein interaction network and biological pathways are far from being complete. It is necessary to develop new methods without relying on prior knowledge to discover novel mutated driver gene sets or pathways.

Previous studies indicate that a driver gene set has two key properties: (1) covering a large number of samples (high coverage); and (2) its mutations tend to exhibit mutual exclusivity (high mutual exclusivity), i.e., a single mutation is usually enough to disturb one pathway [8], [16], [17]. For example, the mutation of *TP53* and the copy number amplification of *MDM2* seldom appear simultaneously in glioblastoma multiforme (GBM) patients (p53 pathway) [1]. These rules have been frequently used to *de novo* discover mutated driver gene sets in recent years [18], [19], [20]. For example, Vandin *et al.* developed Dendrix by designing a weight function to combine the coverage and exclusivity of a gene set, and maximizing it via a Markov chain Monte Carlo (MCMC) approach to extract driver gene sets [18]. Zhao *et al.* further developed a binary linear programming (BLP) model [19] to get the exact solutions of the maximization problem, and designed a genetic algorithm to optimize variant weight functions and incorporate prior biological knowledge into it in a more flexible manner. However, these studies have all focused on a single pathway without considering the cooperativeness between pathways [18], [19], [20], [21].

In fact, a great deal of evidence has suggested that pathways often function cooperatively in cancer initiation and progression [15], [16], [22], [23]. Thus, exploring the complex collaboration among different biological pathways and functional modules may shed new lights on the understanding of the cellular mechanisms underlying carcinogenesis. Leiserson *et al.* [24] generalized Dendrix (Multi-Dendrix) to simultaneously identify multiple driver gene

sets in cancer. More importantly, the collaboration among different pathways means these gene sets are likely simultaneously mutated in a large cohort of patients. To this end, Zhang *et al.* [25] developed CoMDP to *de novo* discover co-occurring mutated driver gene sets in cancer by introducing a novel weight function and a mathematical programming model; Melamed *et al.* [26] introduced an information theoretic method GAMToC to identify combinations of genomic alterations in cancer; and Remy *et al.* [27] developed a logical model to explain mutually exclusive and co-occurring genetic alterations in bladder carcinogenesis.

On the other hand, different cancer types may have certain commonalities [28]. Investigating the similarities and differences among multiple cancer types may enhance the understanding of pathologies underlying cancers and provide new clues to efficient drug design and cancer treatment. The TCGA pan-cancer project surveyed multi-platform aberration data in cancer samples from thousands of cancer patients among 12 cancer types [29], which provides huge opportunities to make such investigations [30], [31]. For example, different histological cancers can be classified into the same clusters [30], [32], [33], which means that different cancers may be treated by the same drugs. Recently, Leiserson *et al.* [34] proposed a directed heat diffusion model (HotNet2) to identify pathways and protein complexes based on pan-cancer network analysis; Kim *et al.* [35] investigated different kinds of mutual exclusivity among multiple cancer types and designed statistical testing methods for driver gene set identification (MEMCover). Although recent pan-cancer studies revealed that some pairs of genes showing mutually exclusivity are common or specific for some cancer types [28], [36], there is still a lack of systematic investigation of commonalities and specificity in pathway level.

In this study, we develop two mathematical programming models (ComMDP and SpeMDP) to *de novo* identify cancer common and specific driver gene sets, respectively. For the former, we detect a set of genes which have significantly high mutual exclusivity and large coverage in two or more cancer types simultaneously. For the latter, we identify a driver gene set specific to one or a group of cancer types (say, S_1) versus another group of cancer types (say, S_2). In other words, we require the detected genes to have significantly high mutual exclusivity and large coverage in the group S_1 but not in S_2. We first apply ComMDP and SpeMDP to simulated data to validate their effectiveness. Then, we apply them to the mutation data of 12 cancer types from the pan-cancer project [29], [30] and obtain several biologically meaningful driver gene sets. For example, for breast carcinoma (BRCA) and ovarian carcinoma (OV), we identify their common driver gene sets as well as their individual specific driver gene sets relative to the other. Interestingly, the identified common gene sets are involved with distinct cancer pathways such as apoptosis pathway, *ErbB* signaling pathway, *PI3K-Akt* signaling pathway and *MAPK* signaling pathway, which enable us to construct a common cancer pathway model for BRCA and OV. Further, we construct a hypothetical mutated driver pathway model for BRCA carcinogenesis and progression based on eight common driver gene sets of BRCA with eight cancer types, indicating the complexity of BRCA carcinogenesis.
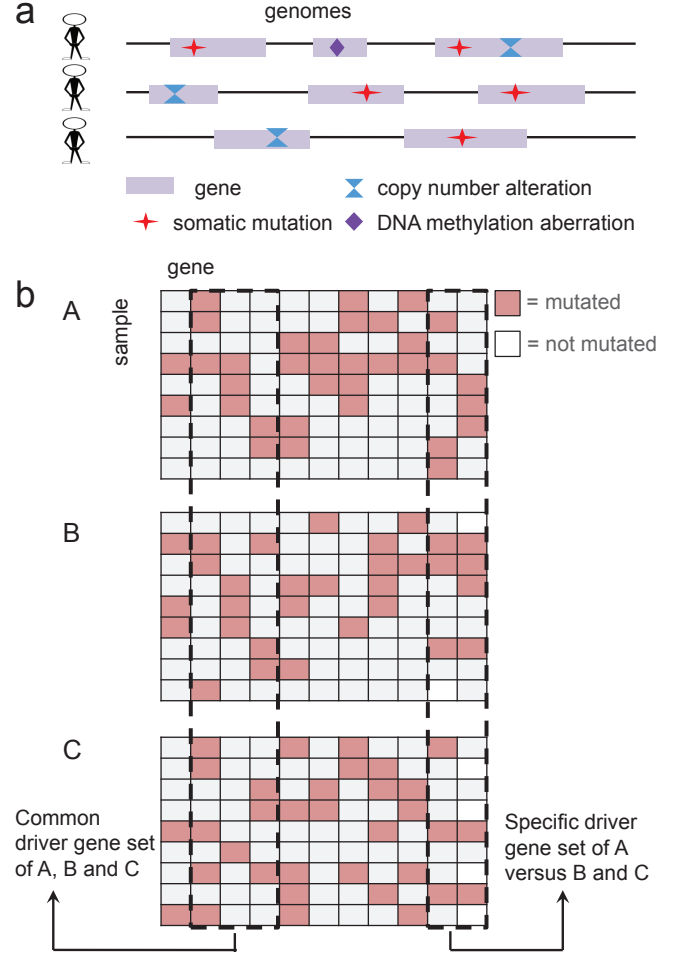

Fig. 1. Schematic illustration of the key idea of this study. (a) Obtain the mutation matrix from the sequencing data (referring to somatic mutations, copy number alterations (CNAs) and DNA methylation aberrations) [18]. (b) Identify the common and specific driver gene sets using ComMDP and SpeMDP.

In addition, we investigate specific driver pathways of the liquid cancer lymphoblastic acute myeloid leukemia (LAML) versus other solid cancer types, and identify mutations of *FLT3, IDH2, NRAS, IDH1, RUNX1, NPM1, TET2, KIT*, amplifications of *MLL, IGSF5*, and deletions of *TP53, GNAQ*, which are involved in proliferation, transcriptional deregulation, impaired hematopoietic differentiation, and so on. We expect the proposed methods can discover new commonalities and specificities among cancers and help to understand cancer initialization and progression further.

2 MATERIALS AND METHODS

We first briefly describe the maximum weight submatrix problem, where the coverage and exclusivity of a gene set is combined to form a weight function for discovering driver gene sets in a single mutation data [18], [19]. Then we propose ComMDP and SpeMDP to *de novo* discover cancer common and specific mutated driver gene sets among multiple cancer types, respectively (Fig. 1).

2.1 The maximum weight submatrix problem

Given a binary mutation matrix A with m rows (samples) and n columns (genes), Vandin *et al.* introduced a weight function W and defined the maximum weight submatrix problem [18]. Specifically, it is designed to find a submatrix M of size $m \times k$ in matrix A by maximizing 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 samples in which the gene g is mutated, $\Gamma(M) = \cup_{g \in M} \Gamma(g)$ measures the coverage of M, and $\omega(M) = \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|$ measures the coverage overlap of M.

2.2 ComMDP for identifying common mutated driver gene sets among two or multiple cancer types

Considering R ($R \geq 2$) cancer types, for each we have the mutation matrix $A_r = (a_{ij}^{(r)})$ with m_r samples and the same n mutated genes ($r = 1, \ldots, R$). To find a common mutated driver gene set M with large coverage and high exclusivity, we introduce a weight function C_m:

$$C_m(M) = \sum_{r=1}^{R} \Big[2|\Gamma_{A_r}(M)| - \sum_{g \in M} |\Gamma_{A_r}(g)| \Big]. \quad (2)$$

We propose the following BLP model to maximize it:

$$\max \quad F_m(\mathbf{x}, u) = \sum_{r=1}^{R} \Big[2 \sum_{i=1}^{m_r} x_i^{(r)} - \sum_{j=1}^{n} \Big(u_j \cdot \sum_{i=1}^{m_r} a_{ij}^{(r)} \Big) \Big] \quad (3)$$

$$\text{s.t.} \quad \begin{cases} x_i^{(r)} \leq \sum_{j=1}^{n} a_{ij}^{(r)} u_j, \ i = 1, \ldots, m_r, \ r = 1, \ldots, R, \\ \\ \sum_{j=1}^{n} u_j = K, \\ \\ x_i^{(r)}, u_j \in \{0,1\}, i = 1, \ldots, m_r, \ j = 1, \ldots, n, \\ \qquad\qquad\qquad r = 1, \ldots, R, \end{cases}$$
$$(4)$$

where u_j indicates whether column j of the mutation matrices falls into submatrx M or not, and all the columns j's with $u_j = 1$ constitute M; $\mathbf{x} = \{x^{(1)}, \ldots, x^{(R)}\}$, and $x_i^{(r)}$ indicates whether the entries of row i are zeros or not in A_r ($r = 1, \ldots, R$). Thus, $\sum_{i=1}^{m_r} x_i^{(r)}$ represents the coverage of M in A_r (i.e., $|\Gamma_{A_r}(M)|$); K is the total number of genes within M.

2.3 SpeMDP for identifying a certain or multiple cancer specific driver gene sets

Suppose we want to find the specific mutated driver gene sets for R cancer types relative to other T ones ($R \geq 1, T \geq 1$). We use $A_r = (a_{ij}^{(r)})$ ($r = 1, \ldots, R$) and $B_t = (b_{kj}^{(t)})$ ($t = 1, \ldots, T$) to denote corresponding mutation matrices, respectively. We introduce the weight function S_m:

$$S_m(M) = \frac{1}{R} \sum_{r=1}^{R} \Big[K|\Gamma_{A_r}(M)| - \sum_{g \in M} |\Gamma_{A_r}(g)| \Big]$$
$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad (5)$$
$$- \frac{1}{T} \sum_{t=1}^{T} \Big[K|\Gamma_{B_t}(M)| - \sum_{g \in M} |\Gamma_{B_t}(g)| \Big].$$

We maximize S_m by the following BLP model:

$$\max \ G_m(\mathbf{x}, \mathbf{y}, u) = \frac{1}{R} \sum_{r=1}^{R} \Big[K \sum_{i=1}^{m_r} x_i^{(r)} - \sum_{j=1}^{n} \Big(u_j \cdot \sum_{i=1}^{m_r} a_{ij}^{(r)} \Big) \Big]$$
$$- \frac{1}{T} \sum_{t=1}^{T} \Big[K \sum_{k=1}^{l_t} y_k^{(t)} - \sum_{j=1}^{n} \Big(u_j \cdot \sum_{k=1}^{l_t} b_{kj}^{(t)} \Big) \Big], \quad (6)$$

$$\text{s.t.} \quad \begin{cases} x_i^{(r)} \leq \sum_{j=1}^{n} a_{ij}^{(r)} u_j, \quad i = 1, \ldots, m_r, \ r = 1, \ldots, R, \\ \\ \frac{1}{n} \sum_{j=1}^{n} b_{kj}^{(t)} u_j \leq y_k^{(t)} \leq \sum_{j=1}^{n} b_{kj}^{(t)} u_j, \\ \qquad\qquad k = 1, \ldots, l_t, \ t = 1, \ldots, T, \\ \\ \sum_{j=1}^{n} u_j = K, \\ x_i^{(r)}, y_k^{(t)}, u_j \in \{0,1\}, i = 1, \ldots, m_r, \ j = 1, \ldots, n, \\ \qquad\qquad r = 1, \ldots, R, \ k = 1, \ldots, l_t, \ t = 1, \ldots, T, \end{cases}$$
$$(7)$$

where $\mathbf{x} = \{x^{(1)}, \ldots, x^{(R)}\}, \mathbf{y} = \{y^{(1)}, \ldots, y^{(T)}\}$. As stated above, the constraint $x_i^{(r)} \leq \sum_{j=1}^{n} a_{ij}^{(r)} u_j$ in Eq. (7) ensures that $\sum_{i=1}^{m_r} x_i^{(r)}$ is the coverage of M in A_r. In Eq. (5) or Eq. (6), because of the subtraction of the weights of B_t from that of A_r, we use the restrictions $\frac{1}{n} \sum_{j=1}^{n} b_{kj}^{(t)} u_j \leq y_k^{(t)} \leq \sum_{j=1}^{n} b_{kj}^{(t)} u_j$ to ensure that $\sum_{k=1}^{l_t} y_k^{(t)}$ is the coverage of M in B_t, and we use the coefficient K to ensure the weights of A_r and B_t are all non-negative.

2.4 Statistical significance

We perform a permutation test to assess the significance of results. We permutate the mutations independently among samples to preserve the mutation frequency of each gene. Two kinds of significance are calculated: (1) individual one measuring the significance of a gene set in a certain mutation matrix, where the weight W in Eq. (1) is used as the statistic; (2) overall one measuring the significance of a gene set by viewing all the mutation matrices as a whole, where the weight C_m in Eq. (2) and the weight S_m in Eq. (5) are used as the statistics for ComMDP and SpeMDP, respectively.

2.5 Simulated data

To assess the performance of the proposed methods on a variety of data, we construct eight datasets, sd1, $\cdots$, sd8, for simulation study. For convenience of description, in the following we use A_r or B_r to denote the mutation matrices, $M_i^{(r)}$ to denote the i-th embedded submatrix (or gene set) in A_r or B_r for which the proposed methods are used to identify, and $p_i^{(r)}$ to denote the gene mutation rate in $M_i^{(r)}$ ($1 \leq r \leq R, 1 \leq i \leq I$).

The datasets sd1 and sd2 are generated to illustrate the performance of ComMDP for identifying common driver gene sets among multiple cancer types. The difference is that in sd1 for each r the $M_i^{(r)}$ have a constant mutation rate ($1 \leq i \leq I$), but in sd2 they have varying ones, to investigate the possible impact of mutation rates in the gene

sets on the discovery accuracy. sd1 is constructed as follows. First, we have three empty matrices A_r with the same sizes: m (samples) $\times$ n (genes) (here $m = 500$, $n = 900$). Then, we embed I submatrices $M_i^{(r)}$ with a mutation rate $p^{(r)}$ into each matrix A_r ($r = 1, \cdots, 3$; $i = 1, \cdots, I$; $I = 9$; $p^{(1)} = 0.80$, $p^{(2)} = 0.85$, $p^{(3)} = 0.90$), where for each r, $M_i^{(r)}$ contains $i + 1$ genes ($i = 1, \cdots, I$), and for each i, these submatrices $M_i^{(r)}$ occupy the same columns in the corresponding A_r ($r = 1, \cdots, 3$). For each sample in A_r, a gene uniformly chosen from $M_i^{(r)}$ is mutated with rate $p^{(r)}$, and once one gene is mutated, the other genes in $M_i^{(r)}$ have a rate p_0 to be mutated ($p_0 = 0.04$). Finally, the genes not in $M_i^{(r)}$ are mutated in at most three samples, which can be viewed as the background mutation rate in the simulated data. The dataset sd2 is constructed in a similar way, the difference is that each gene set has 9 genes ($K = 9$), and $M_i^{(r)}$ has a mutation rate $p_i^{(r)} = 1 - i * \delta(r)$ ($r = 1, \cdots, 3; i = 1, \cdots, 9$), where $\delta(1) = 0.03, \delta(2) = 0.04, \delta(3) = 0.05$.

The simulated datasets sd3-sd7 are generated to demonstrate the performance of SpeMDP for identifying specific driver gene sets of one or several cancer type(s) versus other cancers. In this case, the datasets are constructed to contain two kinds of embedded gene sets. The first kind of gene sets have mutations with (approximately) mutual exclusivity, like those in sd1 (called the first manner of embedding); but for the second ones, we randomly select 60% samples for which two genes are randomly chosen to be mutated with proper mutation rates, ensuring they are not exclusive (called the second manner of embedding). For the details of the construction of sd3-sd7, please refer to the Supplementary Data. We use these five datasets sd3-sd7 to investigate different aspects for SpeMDP applications. Specially,

- sd3 for discovering a certain cancer specific driver gene sets

- sd4 for investigating in which case SpeMDP can identify specific driver gene sets

- sd5 for investigating the impact of diverse mutation rates on the results

- sd6 for investigating the method performance in different mutation generation manners

- sd7 for discovering multiple cancer specific driver gene sets

We construct dataset sd8 to see if the previous individual cancer type approaches can also identify cancer common and specific driver gene sets (e.g., BLP in MDPFinder [19] or Dendrix [18]). The construction of sd8 is similar to that of sd1. Seven groups of submatrices $M_i^{(r)}$ with 6 genes in each are embedded into A_r with $m = 500, n = 700, r = 1, \cdots, 3$. The first group of submatrices are constructed in the first embedding manner (thus corresponding to a common driver gene set); each of the 2nd to the 4th groups contains two with the first embedding manner, one with background mutations (corresponding to neither common nor specific driver gene sets); each of the last three groups contains one with the first embedding manner, two with background mutations (each corresponding to a specific driver gene set).

2.6 Biological data

We use mutation data from the pan-cancer project [30], [33] to assess our methods for practical applications. The 12 types of cancer include bladder carcinoma (BLCA), breast carcinoma (BRCA), colon adenocarcinoma (COAD), glioblastoma multiformae (GBM), head and neck squamous carcinoma (HNSC), kidney renal clear-cell carcinoma (KIRC), lymphoblastic acute myeloid leukemia (LAML), lung adenocarcinoma (LUAD), lung squamous carcinoma (LUSC), ovarian carcinoma (OV), rectal adenocarcinoma (READ), and uterine cervical and endometrial carcinoma (UCEC). Here colon adenocarcinoma and rectal adenocarcinoma are combined into one type denoted as COADREAD.

3 RESULTS

3.1 Simulation study

We first apply ComMDP to the simulated datasets sd1, sd2 and apply SpeMDP to sd3 - sd7 to assess their performance. We run each method ten times for each dataset. We further apply them to sd8 and compare them with driver gene set discovery approaches for individual cancer types (BLP [19] is used here).

3.1.1 Common driver gene set discovery.

For sd1 ComMDP can identify the embedded gene sets for all the ten runs when the number of genes $K \leq 8$ (Fig. 2a). When $K = 9$, it can detect the embedded gene set for five runs, and it has a wrong one for each of other five runs. When $K = 10$, each detection of eight runs contains nine correct genes plus a wrong one, and each of other two runs has two wrong ones. We also investigate the possible impact of varying mutation rates in the gene sets on the discovery accuracy. For the embedded nine gene sets of $K = 9$ with diverse mutation rates in sd2, each of ten runs can identify at least eight correct genes in each gene set (Fig. 2b).

3.1.2 Specific driver gene set discovery.

First, we consider the situation of one cancer specific driver gene sets. In sd3, SpeMDP can correctly detect the embedded gene sets for $K = 2$ to 10 (Fig. 2c). The results on sd4 demonstrate that when the gene set is exclusive in kind 1 (kind 2) set but not in kind 2 (kind 1) set, or one is exclusive and the other takes background mutations, SpeMDP can successfully identify it (Fig. 2d). In the first case, the gene set is approximately exclusive in both kinds 1 and 2 sets which corresponds to a common driver pathway, so SpeMDP cannot find it. In sd5, the mutation rates in kind 1 and kind 2 sets simultaneously get smaller and smaller, so the mutation coverage will get small (so does the weight W) in kind 1 set along with i gets large ($1 \leq i \leq I, I = 9$), and more exclusive mutation in kind 2 dataset will become possible. Therefore, the performance to detect the embedded gene set will decrease when i gets large (Fig. 2e). We further validate this on sd6. The mutation rate of sd6 in kind 2 dataset gets larger and that in kind 1 set gets smaller along with i becomes large ($1 \leq i \leq I, I = 9$). In this case, we successfully identify all the embedded gene sets except the one corresponding to $i = 2$ (Fig. 2f). Lastly, the result on sd7 indicates that SpeMDP is also effective to identify specific gene sets for multiple cancer types (Fig. 2g), where the dataset is simulated in a similar way to that of sd6.
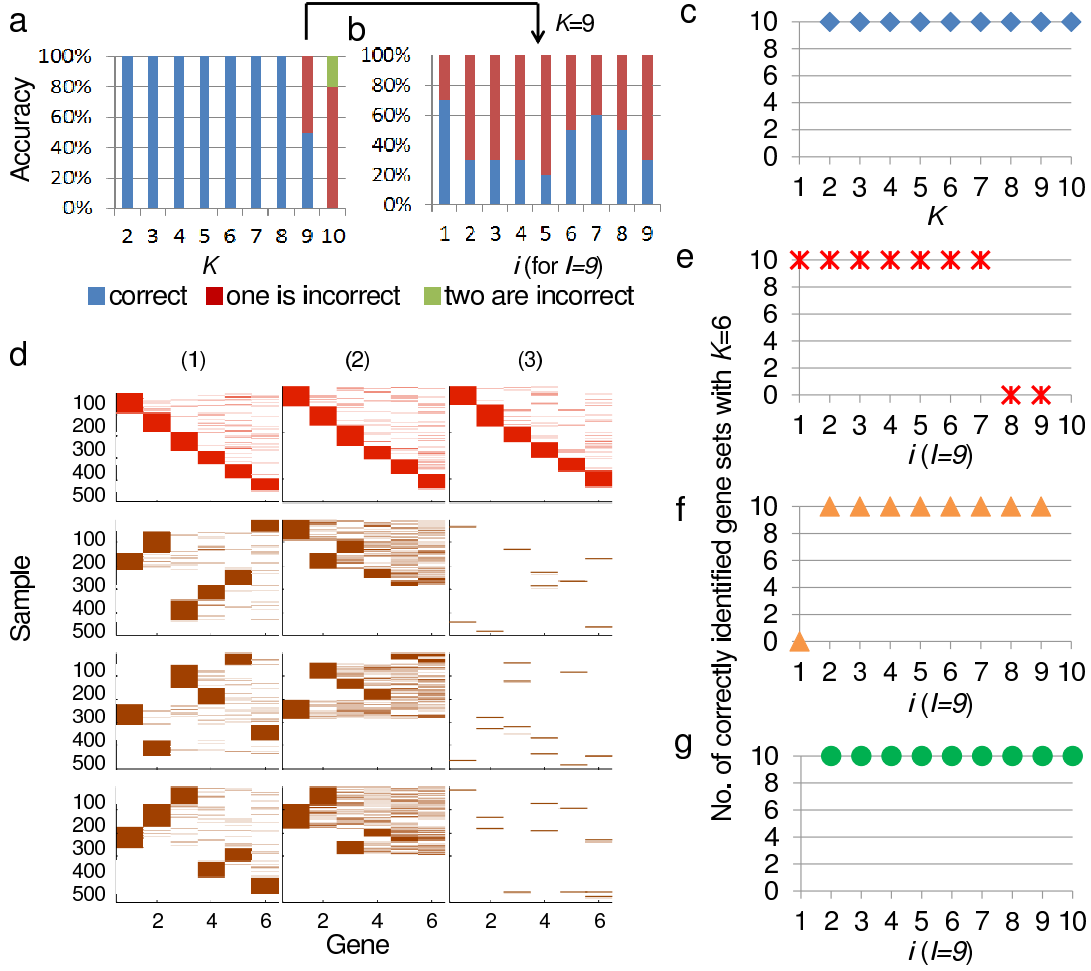
Fig. 2. The results of simulation study. Accuracy of the common driver gene set discovery: (a) for $K = 2$ to 10 with a constant mutation rate and (b) for fixed $K = 9$ in which the embedded nine gene sets ($I = 9$) have decreasing mutation rates. (c) Accuracy of the specific driver gene set discovery for $K = 2$ to 10 with a constant mutation rate. (d) Illustration of the situation for specific driver gene set discovery: (1) two kinds of exclusive gene sets corresponding to a common driver pathway, so it cannot be found; (2) exclusive (red) versus nonexclusive (brown) gene sets can be found; (3) exclusive (red) versus random (brown) gene sets can be found. Numbers of correctly identified specific driver gene sets for ten runs with $K = 6$ and $I = 9$: (e) both kinds of gene sets have decreasing mutation rates and (f) one has increasing and the other has decreasing mutation rates. (g) Numbers of correctly identified specific driver gene sets for ten runs about multiple cancer types with $K = 6$ and $I = 9$.

3.1.3 Individual driver gene set discovery approaches cannot detect common and specific driver gene sets well.

For sd8, we first use the BLP model in MDPFinder [19] to identify individual driver gene sets in each A_r which contains seven embedded submatrices, and we get the ones marked by ellipses in Fig. 3. Then we apply ComMDP and SpeMDP to identify the common and specific driver gene sets among all the A_rs, and we obtain those marked by the rectangle and dotted rectangles, respectively. Note that the detected individual and common driver gene sets do not have any overlap. Moreover, the detected individual driver pathways in the second and third sets (A_2 and A_3) are not specific (Fig. 3).

3.2 Applications to biological data

We investigate common driver gene sets among all the pair mutation data of the 11 cancer types with $K = 2$ to 10. We summarize all the significant driver gene sets with both individual and overall significance less than 0.05 (Suppl. Table S1). The mutation rates of *TP53* in LUSC and OV are very

high (164/182 and 405/445, respectively). We distinguish this situations with or without *TP53* when relating to these two cancer types.

Common mutated driver gene sets among two or multiple cancer types

Previous studies indicate that BRCA and OV have similar phenotypes to some extent. Interestingly, we indeed obtain significant common driver gene sets between them by ComMDP for $K = 7$ to 10 (Table 1), and reveal 10 genes *TP53, PIK3CA, MAP3K1, MAP2K4, PIK3R1, LPA, KRAS, ERBB2, FGFR2, TNXB* in total. These genes are enriched in several signaling pathways relating to apoptosis, ErbB signaling pathway, PI3K-Akt signaling pathway, MAPK signaling pathway, etc. Based on known KEGG pathway knowledge (Fig. 4A), we propose a common mutated pathway model for cancer initiation and progression in both BRCA and OV (Fig. 4B). We show the heat map of the alterations of the gene set for $K = 10$ (Suppl. Figure S1) and see that *TP53* has a very high mutation rate in OV (as stated
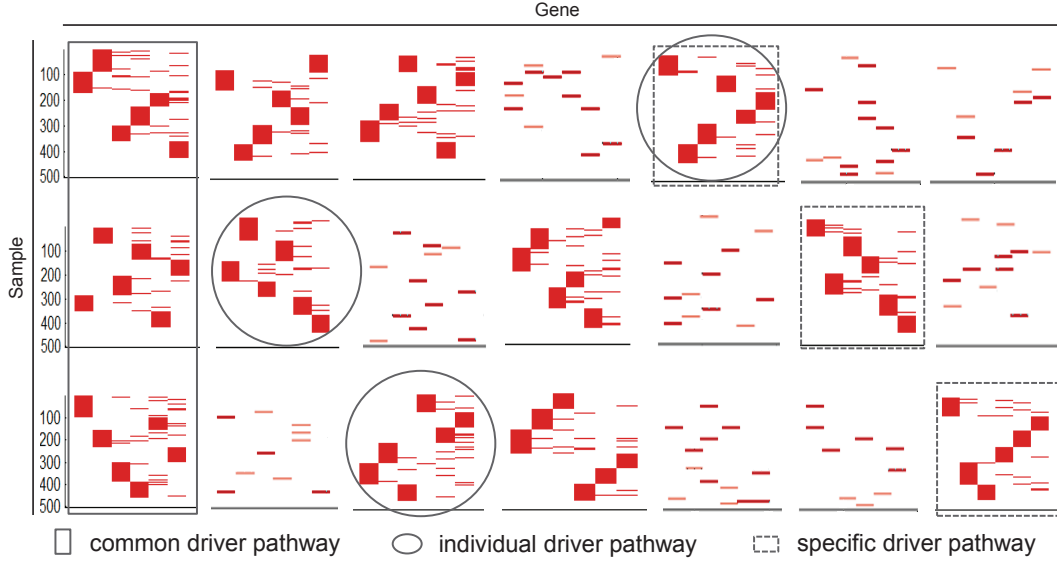
Fig. 3. ComMDP and SpeMDP can exactly identify the embedded common driver gene sets (rectangle) and specific driver gene sets (dotted rectangles), respectively. The BLP of MDPFinder [19] can identify the individual driver gene sets in each dataset (ellipses).

TABLE 1
Significant common driver gene sets between BRCA and OV identified by BLP

K	Common pathways	p_1	p_2	p
7	TP53, PIK3CA, MAP3K1, MAP2K4, PIK3R1, LPA, KRAS	0	0.003	0
8	TP53, PIK3CA, MAP3K1, MAP2K4, PIK3R1, LPA, KRAS, ERBB2	0	0.004	0
9	TP53, PIK3CA, MAP3K1, MAP2K4, PIK3R1, LPA, KRAS, ERBB2, FGFR2	0	0	0
10	TP53, PIK3CA, MAP3K1, MAP2K4, PIK3R1, LPA, KRAS, ERBB2, FGFR2, TNXB	0	0	0

p_1 and p_2 denote the p-values of the common gene sets in BRCA and OV, respectively. p represents the overall significance.

above). The mutation rates of other nine genes are very low. It implies that *TP53* mutation plays a dominant role in this pathway in OV, indicating that the common driver pathway exploration helps to identify driver ones with low mutation frequency (Fig. 4).

We also employ BLP [19] to identify individual driver gene sets in BRCA and OV (Table 2), respectively. For each K in Table 2, there is only one common gene *TP53* between the identified gene sets for these two cancers. For other genes in the common gene sets in Table 1, *PIK3CA, MAP3K1, MAP2K4* and *PIK3R1* only appear in the gene sets of BRCA, and *KRAS, FGFR2* and *LPA* appear only in those of OV (Table 2). Thus, only a local path of the common mutated pathways (Fig. 4B) can be found by BLP for each of these two cancers (Fig. 4C).

ComMDP has distinct advantages over both the gene-centric frequency-based approaches and the individual driver gene set based approaches. First, in the identified common gene sets (Table 1), some genes have very low mutation frequency. For example, *TNXB, LPA* and *FGFR2* all have less than five mutations in 466 BRCA samples and 445 OV samples, respectively. With such low frequency, these genes cannot be discovered by the gene-centric frequency-based approaches. But all the three genes have important biological functions (Fig. 4B) and are closely related to the carcinogenesis of BRCA and OV [37], [38], [39], [40]. For

instance, Hu *et al.* validated *TNXB* as a promising biomarker for early metastasis of breast cancer [37]; Kim *et al.* demonstrated *TNXB* might be helpful to predict the prognosis of patients with stage III serous ovarian cancer through differential expression analysis [38]; *LPA* and its receptors play an important role in mediating malignant behaviors in various cancers and recent studies [39], [41] suggested they could be potential diagnostic biomarkers for BRCA and OV, respectively; *FGFR2* were suggested as candidate targets for therapeutics in clinical trial for BRCA and OV [40], [42]. Second, some of the ten important common genes (Table 1) cannot be identified by the driver gene set identification approaches for individual cancer type (Table 2). Especially, *TNXB* and *ERBB2* are not identified for any cancer by BLP. Actually, *ERBB2* is a well-known cancer gene, and it plays a crucial role for certain subtypes of BRCA and OV patients [43], [44]. Third, it is important to note that the individual cancer type approach can only discover a small part of the common gene set for each cancer type (Fig. 4), whereas ComMDP can integrate information from different cancers and imply a more biologically reasonable common driver pathway.

Importantly, identifying all the significant common driver gene sets of BRCA with certain cancer types will help to understand various aspects of BRCA carcinogenesis. Besides OV, other cancer types include BLCA, COADREAD,
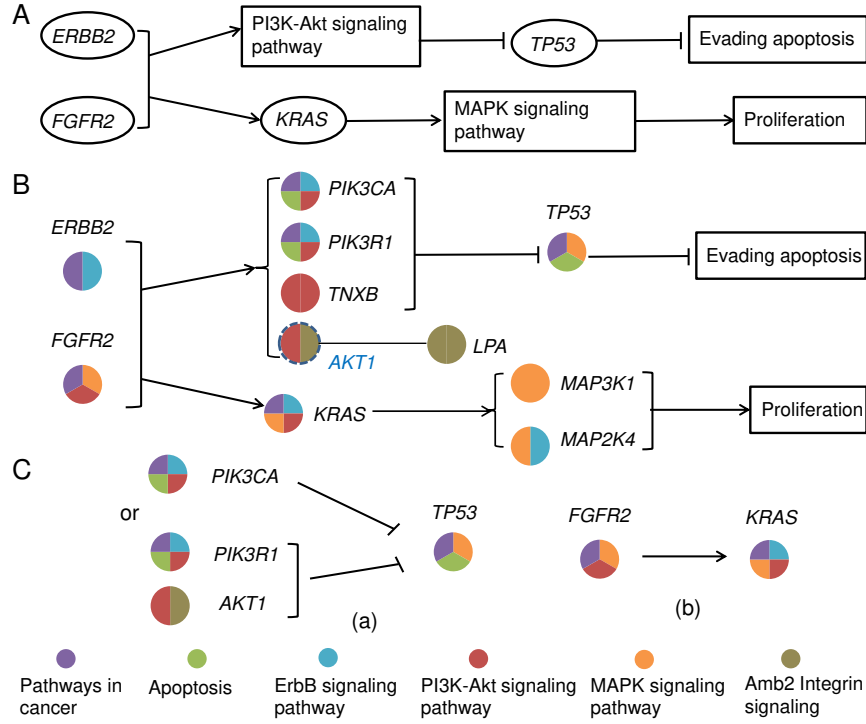
Fig. 4. (A) Known KEGG pathway knowledge in cancer. (B) A common mutated pathway model for BRCA and OV initiation and progression. It is inferred based on the identified common gene sets and their participant pathway knowledge. The gene *AKT1* relating to LPA with the PI3K-Akt signaling pathway does not appear in the identified common gene sets. (C) Only local parts of the common mutated pathway in (B) can be found by the individual cancer type approach BLP: (a) BRCA (up for $2 \leq K \leq 6$, below for $7 \leq K \leq 10$), (b) OV.

TABLE 2
Significant individual driver gene sets in BRCA and OV

K	Driver pathway in BRCA	Driver pathway in OV
2	TP53, PIK3CA	TP53, KRAS
3	TP53, PIK3CA, GATA3	TP53, KRAS, IDI2
4	TP53, PIK3CA, GATA3, CDH1	TP53, KRAS, FGFR2, PIGV
5	TP53, PIK3CA, GATA3, CDH1, CTCF	TP53, KRAS, IDI2, PIGV, BRAF
6	TP53, PIK3CA, GATA3, CDH1, CTCF, MACROD2	TP53, KRAS, IDI2, BRAF, LPA, EGFR
7	TP53, GATA3, CDH1, MACROD2, AKT1, MAP3K1, MAP2K4	TP53, KRAS, IDI2, BRAF, PIGV, EGFR, C4orf45
8	TP53, GATA3, CDH1, MACROD2, AKT1, MAP3K1, MAP2K4, PIK3R1	TP53, KRAS, FGFR2, C4orf45, EPHA3, PPID, ETFDH, FNIP2
9	TP53, GATA3, CDH1, MACROD2, AKT1, MAP3K1, MAP2K4, PIK3R1, POLD4	TP53, KRAS, FGFR2, PIGV, EGFR, C4orf45, EPHA3, PPID, FNIP2
10	TP53, GATA3, CDH1, MACROD2, AKT1, MAP3K1, MAP2K4, PIK3R1, POLD4, ARID1A	TP53, KRAS, IDI2, BRAF, LPA, C4orf45, EPHA3,PPID, ETFDH, FNIP2

Here the p-values are all less than 0.0001.

GBM, HNSC, KIRC, LUAD and UCEC (Suppl. Table S1). In total, we discover 38 different genes in all the eight significant gene sets (Suppl. Figure S2). These genes are involved in many important signaling pathways (Suppl. Figure S3) and relate to diverse cancers (such as prostate cancer, endometrial cancer, pancreatic cancer, lung cancer, glioma, colorectal cancer, etc) (searched by DAVID [45]). It is known that cancer is a very complex disease. We integrate prior pathway knowledge and all the common driver gene sets of BRCA with the eight cancer types to explore more details about BRCA carcinogenesis (Fig. 5). Compared to Fig. 4, we find some new paths and more

genes involving in the important hallmarks of cancer in Fig. 5, such as [*IFNA6–cytokineR/JAK–*PI3K-Akt signaling pathway–*TP53–Fas–CASP8*] leading to apoptosis, [*ARID1A–NF1–KRAS–*MAPK signaling pathway] leading to proliferation, [*GATA3–MAPK14*] leading to cell survival, etc.

We note that, based on the common driver gene sets of BRCA with multiple cancer types, we can get distinct new discoveries versus previous work. For example, the authors in [28] identified 127 significantly mutated genes (SMGs) from diverse signaling and enzymatic processes, and calculated the most frequently mutated genes in the pan-cancer cohort for each cancer type. Especially, for
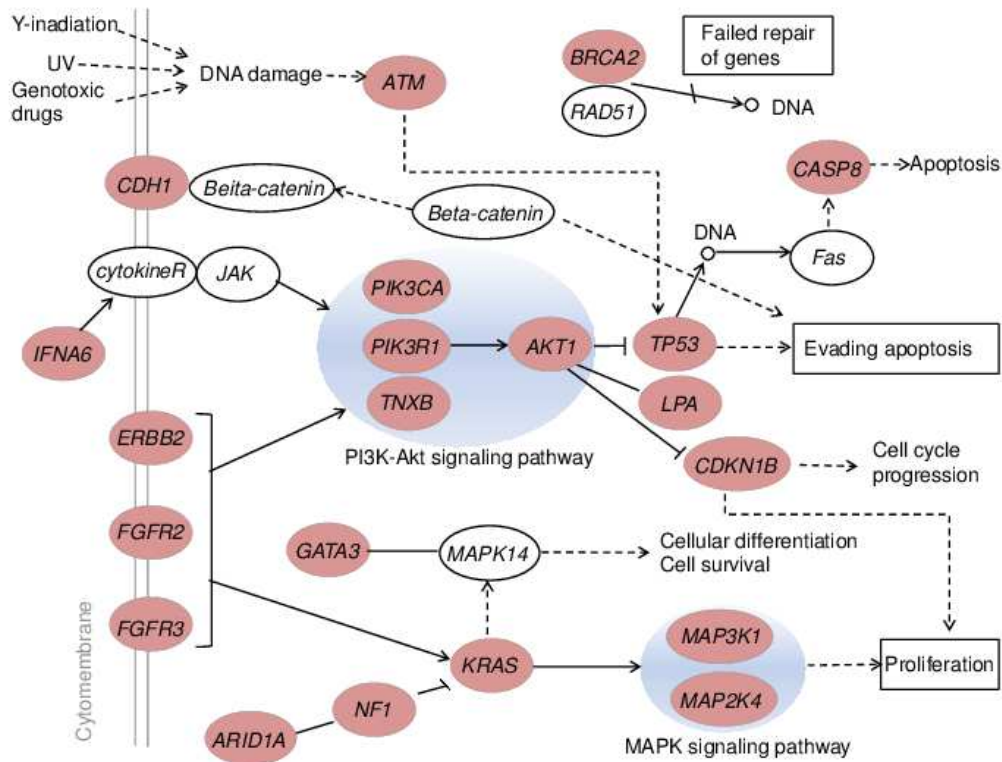
Fig. 5. Hypothetical driver pathways for BRCA carcinogenesis and progression. It is inferred based on the common driver gene sets between BRCA and other eight cancer types. The dotted arrows denote indirect effects, and a line represent a known interaction between them or co-occurrence in a known signaling pathway. The ComMDP discovered genes are in red.

BRCA they obtained eight genes (*TP53, PIK3CA, MAP3K1, MAP2K4, GATA3, AKT1, CDH1, CBFB*), seven of which belong to our 38 genes except *CBFB*. On one hand, we find more genes involved in the cellular processes that the above eight genes relating to: transcription factor/regulator (*CTCF*), genome integrity (*ATM, BRCA2*), MAPK signaling (*KRAS, NF1*), PI(3)K signaling (*PTEN, PIK3R1*). On the other hand, we detect several genes involved in other important biological processes in cancer: histone modifier (*ARID1A, PBRM1, KDM6A*), RTK signaling (*FGFR2, FGFR3*), cell cycle (*CDKN1B*). This indicates that these biological processes may also contribute to the carcinogenesis of BRCA. More importantly, we identify some other genes that are not included in the 127 selected genes in [28], but they play also crucial roles, such as *CASP8, IFNA6, ERBB2, TNXB, NOTCH2* (Suppl. Figure S3). For example, *CASP8* is involved in the programmed cell death induced by *Fas* and various apoptotic stimuli, and there are many studies relating to its biological functions [46], [47], [48]; *NOTCH2* plays a role in a variety of developmental processes by controlling cell fate decisions, and has close relationship with BRCA progression [49], [50]; *IFNA6* belongs to the family of interferon, although it has not been well studied, this kind of immune-associated genes may be worth paying great attention for immunotherapy of cancers [51].

Note that BLCA has common significant driver gene sets with all the other 10 cancer types (Suppl. Table S1). For BRCA stated above, some genes frequently appear in many common gene sets (Suppl. Figure S2). But for BLCA,

the common gene sets are not necessarily the same, such as those with BRCA (Suppl. Table S2), COADREAD (Suppl. Table S3), GBM (Suppl. Table S4) and LUSC (Suppl. Table S5). For example, we identify two different sets of genes for BRCA and COADREAD, whereas their functional annotations are quite similar (Suppl. Figure S4). All these are closely related to cancer generation and progression.

Furthermore, we also investigate the common mutated driver gene sets among multiple cancer types. For example, we find that BRCA, OV, LUAD and GBM have common significant gene sets with $K = 4$ to 10 (Table 3), which relate to the mutations of genes *TP53, PIK3CA, KRAS, MAP3K1, EP300, PIK3R1, TNXB, KDM6A, LPA* and deletion of gene *IFNA6*. As an example, we show the heat map of the alterations of the gene sets in these four cancer types for $K = 4$ (Fig. 6). These gene alterations are approximately mutually exclusive in all the four cancer types. Compared to the situation of only considering BRCA and OV (Table 1), it covers three new genes. Besides *IFNA6* stated above, two others are *EP300* and *KDM6A*. *EP300* interacting with *TP53* is a transcriptional coactivator to mediate many transcriptional events including DNA repair [52]. It also functions as a histone acetyltransferase to regulate transcription via chromatin remodeling. Gene *KDM6A* is associated with chromatin organization and transcriptional misregulation in cancer [53]. Indeed, this investigation can help one to reveal common characteristics among diverse cancers.

Mutated driver gene sets specific to one cancer or multiple

TABLE 3
Significant common driver gene sets among BRCA, OV, LUAD and GBM

K	Common pathway	p_1	p_2	p_3	p_4	p
4	*TP53, IFNA6, PIK3CA, KRAS*	0.0010	0.0040	0.0020	0	0
5	*TP53, IFNA6, PIK3CA, KRAS, MAP3K1*	0	0.0020	0.0040	0	0
6	*TP53, IFNA6, PIK3CA, KRAS, MAP3K1, EP300*	0	0.0080	0.0040	0	0
7	*TP53, IFNA6, PIK3CA, KRAS, MAP3K1, EP300, PIK3R1*	0	0.0040	0	0	0
8	*TP53, IFNA6, PIK3CA, KRAS, MAP3K1, EP300, PIK3R1, TNXB*	0	0.0020	0.0010	0	0
9	*TP53, IFNA6, PIK3CA, KRAS, MAP3K1, EP300, PIK3R1, TNXB, KDM6A*	0	0.0040	0.0020	0	0
10	*TP53, IFNA6, PIK3CA, KRAS, MAP3K1, EP300, PIK3R1, TNXB, KDM6A, LPA*	0	0.0010	0.0080	0	0

p_1, p_2, p_3 and p_4 denote the p-values of the common gene sets in BRCA, OV, LUAD and GBM, respectively. p represents the overall significance.
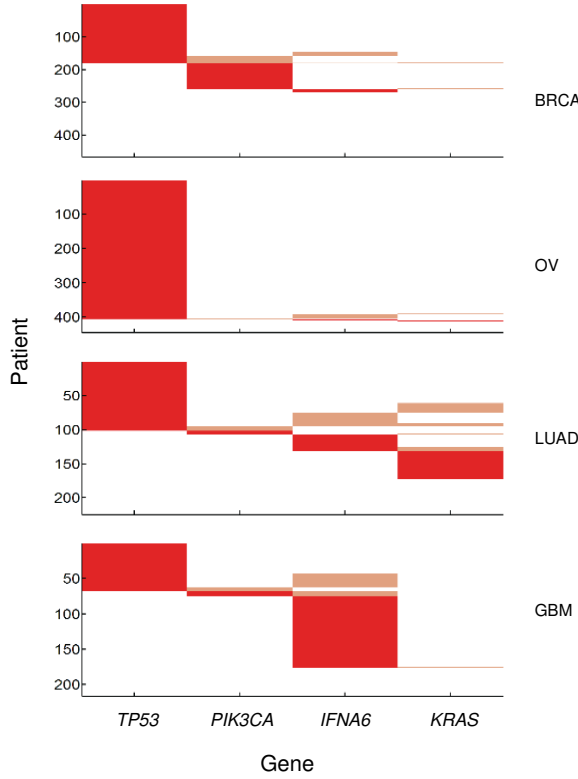

Fig. 6. The heat map of the alterations in the common driver gene set (*TP53, PIK3CA, IFNA6, KRAS*) among the four cancer types including BRCA, OV, LUAD and GBM.

cancer types

We apply SpeMDP to the mutation data without common driver genes and identify several significant BRCA specific driver gene sets relative to OV with $K = 3, 4, 9, 10$ (Table 4). These gene sets relate to the mutations of *GATA3, CDH1, AKT1, CTCF* and amplifications of *ERBB2, WHSC1L1, CCND1, PLK1, RFPL4A, DDAH1*, many of which have been suggested to be closely related with breast cancer initiation and progression by a number of studies [54], [55], [56], [57]. For example, *GATA3* plays a specific role in the differentiation of breast luminal epithelial cells, and has particular diagnostic utility in the setting of triple-negative breast carcinomas [58]; the tumor suppressor *CDH1* has

been shown to be a potential drug target in breast cancer [54]; and epigenetic silencing of *HOXA10* by *CTCF* in breast cancer cells is related to tumorigenesis [55]. Similarly, we also identify significant OV specific driver gene sets relative to BRCA with $K = 2$ to 10 (Table 4), and significant BRCA and OV specific driver pathways relative to the liquid cancer LAML with $K = 9, 10$ (Suppl. Table S6).

The liquid cancer LAML has significant common or specific driver gene sets compared to solid cancer types

LAML is the only liquid cancer in the current study. Interestingly, it has some common driver gene sets with solid cancers. Specifically, by using ComMDP we identify LAML has a significant common driver gene set with COADREAD, GBM and BLCA for $K = 5$, which includes deletion of *IFNA6*, and mutations of *TP53, IDH1, WT1, SDK1*.

More importantly, LAML is expected to have some specific mutation patterns. We investigate LAML specific driver pathways relative to other 10 solid cancers, and discover significant driver gene sets with $K = 2$ to 10 except $K = 5$ (Table 5). These gene alterations include mutations of *FLT3, IDH2, NRAS, IDH1, RUNX1, NPM1, TET2, KIT*, amplifications of *MLL, IGSF5* and deletions of *TP53, GNAQ*. We show the heat map of the alterations of the gene sets in all the 11 cancer types for $K = 10$ (Fig. 7) and see that the alterations display significant mutual exclusivity in LAML, but not in other ten cancer types. Most of these identified genes have been previously reported to be related to LAML [59]. For example, eight genes of them are involved with six functional categories associated with LAML carcinogenesis (Fig. 8). Many large-scale studies have confirmed that *FLT3* can activate mutations in LAML occurrence and disease progression and thus plays an important role in the pathogenesis of LAML [59], [60]; *NPM1* is thought to be involved in several processes including centrosome duplication, cell proliferation and regulation of the *ARF/TP53* pathway and its mutations are associated with LAML supported by various studies [61], [62], [63]; *KIT* confers unfavorable prognosis for LAML patients [59].

Moreover, we can predict the potential implication of *GNAQ* with LAML based on its appearance in the LAML specific driver gene set even with very low mutation frequency in LAML (2/164). *GNAQ* has been considered as one

TABLE 4
BRCA and OV specific mutated driver gene sets relative to each other

Type	K	Specific pathway	p	q	P
BRCA/OV	3	*ERBB2, GATA3, CDH1*	0.0110	1	0.0100
	4	*ERBB2, GATA3, CDH1, WHSC1L1*	0.0360	0.9510	0.0300
	9	*ERBB2, GATA3, CDH1, WHSC1L1, CCND1, AKT1, CTCF, PLK1, RFPL4A*	0.0410	0.7810	0.0420
	10	*ERBB2, GATA3, CDH1, WHSC1L1, CCND1, AKT1, CTCF, PLK1, RFPL4A, DDAH1*	0.0160	0.7550	0.0280
OV/BRCA	2	*BRCA1, BRCA2*	0	0.6000	0
	3	*BRCA1, BRCA2, CACNA1A*	0	0.6570	0
	4	*BRCA1, BRCA2, CACNA1A, WT1*	1.0000e-03	0.6770	0
	5	*BRCA1, BRCA2, CACNA1A, CASC1, GUSBP3*	0	1	0
	6	*BRCA1, BRCA2, CACNA1A, CASC1, GUSBP3, HUS1B*	0	1	0
	7	*BRCA1, BRCA2, WT1, ADPRHL2, METTL17, DNM2, COX4I2*	0	0.9970	0
	8	*BRCA1, BRCA2, WT1, ADPRHL2, METTL17, DNM2, COX4I2, SRP19*	0	1	0
	9	*BRCA1, BRCA2, WT1, ADPRHL2, METTL17, DNM2, COX4I2, SRP19, PARP8*	0	1	0
	10	*BRCA1, BRCA2, WT1, ADPRHL2, METTL17, DNM2, COX4I2, SRP19, PARP8, PRPS2*	0	1	0

p and q denote the p-values of the gene set in BRCA relative to OV (BRCA/OV) or vice versa (OV/BRCA), respectively. P represents the overall significance. Here the identified gene set is significant means that p and P are both less than 0.05, but q is larger than 0.05.

TABLE 5
LAML specific mutated driver gene sets relative to BRCA, HNSC, KIRC, LUSC, BLCA, GBM, LUAD, COADREAD, OV and UCEC

K	Specific pathway	p	$q_1, \cdots, q_{10}$	P
2	*FLT3, IDH2*	0.0150	1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 0.9520, 0.9890, 1.0000, 1.0000	0.0170
3	*FLT3, IDH2, NRAS*	< 0.0001	1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 0.9220, 0.7220, 0.9960, 0.8160	0.0010
4	*FLT3, IDH2, NRAS, IDH1*	0.0020	1.0000, 1.0000, 1.0000, 1.0000, 0.9480, 0.9340, 0.8970, 0.5450, 0.9960, 0.8450	0.0020
6	*FLT3, MLL, IGSF5, RUNX1, NPM1, TP53*	0.0180	0.5110, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 0.9920, 0.9830, 0.9840, 0.9640	0.0140
7	*FLT3, IDH2, IGSF5, RUNX1, NPM1, TP53, TET2*	0.0430	0.4720, 0.9940, 0.9930, 1.0000, 0.9630, 1.0000, 0.9930, 0.9930, 0.9800, 0.9410	0.0240
8	*FLT3, IDH2, IGSF5, MLL, RUNX1, NPM1, TP53, KIT*	0.0050	0.5020, 0.9950, 1.0000, 1.0000, 1.0000, 1.0000, 0.9630, 0.8860, 0.9640, 0.9870	0.0050
9	*FLT3, IDH2, IGSF5, MLL, RUNX1, NPM1, TP53, KIT, TET2*	0.0130	0.4890, 0.9900, 0.9960, 1.0000, 0.9710, 1.0000, 0.9650, 1.0000, 0.9680, 0.9950	0.0050
10	*FLT3, IDH2, IGSF5, MLL, GNAQ, RUNX1, NPM1, TP53, KIT, TET2*	0.0090	0.3820, 0.9780, 0.9980, 1.0000, 0.9750, 1.0000, 0.9790, 1.0000, 0.9740, 0.9880	0.0030

$p, q_1, \cdots, q_{10}$ denote the p-values of the gene set in LAML, BRCA, HNSC, KIRC, LUSC, BLCA, GBM, LUAD, COADREAD, OV and UCEC, respectively. P represents the overall significance. For each K, p and P are less than 0.05, but $q_1, \cdots, q_{10}$ are all larger than 0.05.

of uveal melanoma driver genes [64], and a prognostic factor for mucosal melanoma [65]. Another study [66] indicates that variations of *GNAQ* tend to occur in childhood LAML patients. On the other hand, mutational-driven comparison with other cancer types showed that uveal melanoma is very similar to pediatric cancers, characterized by very few somatic insults and, possibly, important epigenetic changes [64]. Thus, we suggest that *GNAQ* might be a candidate driver gene for childhood LAML patients and its function in LAML carcinogenesis and progression worth further exploration.

4 DISCUSSION

In this study, we develop ComMDP and SpeMDP to identify cancer common and specific mutated driver gene sets among two or multiple cancer types, respectively. We first apply them to a set of simulated data with diverse mutation rates and pathway sizes to demonstrate their effectiveness. We further apply them to real biological data from TCGA, and obtain a set of cancer common and specific gene sets which are involved in several key biological processes or signaling pathways. This suggests that the identified common or specific driver gene sets may play crucial roles and worth to be further explored.
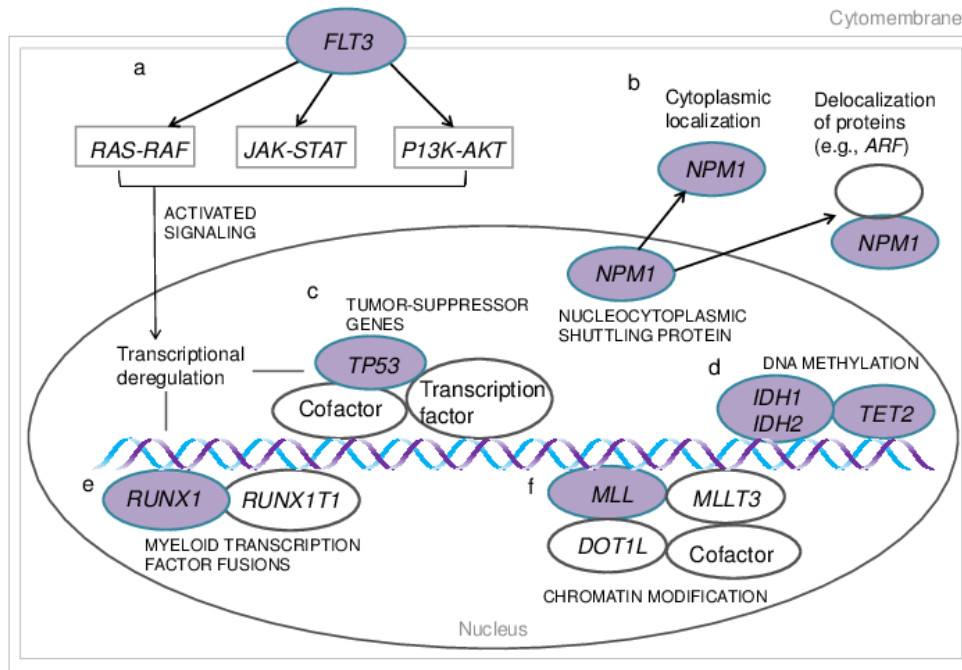
Fig. 8. Specific driver pathways or biological processes of LAML relative to the 10 solid cancer types. (a) Mutations in *FLT3* confer a proliferative advantage through the *RAS-RAF*, *JAK-STAT*, and *PI3K-AKT* signaling pathways. (b) Mutations in *NPM1* result in the aberrant cytoplasmic localization of *NPM1* and *NPM1*-interacting proteins. (c) Deletions of tumor suppressor genes, such as *TP53*, lead to transcriptional deregulation and impaired degradation through *MDM2* and *PTEN*. (d) *DNMT3A* and *TET2* mutations, as well as *IDH1* and *IDH2* mutations, can lead to the deregulation of DNA methylation. (e) Mutations in myeloid transcription factors such as *RUNX1* and transcription factor fusions by chromosomal rearrangements lead to transcriptional deregulation and impaired hematopoietic differentiation. (f) Mutations of genes involved in the epigenetic homeostasis of cells, such as mutations of *ASXL1* and *EZH2*, lead to deregulation of chromatin modification as well as *MLL-MLLT3* gene fusion, which can impair other methyltransferases. Note: the genes in purple represent they appear in the identified specific driver gene sets. (Referring to [59], [67])

Applications of ComMDP and SpeMDP to real data show their advantages over both gene-centric frequency-based approaches and individual driver gene set based approaches. For example, we identified *TNXB, LPA, FGFR2, CASP8, NOTCH2* for BRCA, all of which are mutated with very low frequency (less than five mutations in 466 patients), but have critical biological functions in carcinogenesis of BRCA. All these genes cannot be discovered by the gene-centric frequency-based approaches [28]. We also find that some of the identified important common genes (Table 1) cannot be detected by the driver gene set identification approaches for individual cancer types [19]. Moreover, the individual cancer type approaches can only discover a small part of the common pathway for each cancer type (Fig. 4), whereas ComMDP can integrate information from different cancers and give a more biologically reasonable common driver pathway. Furthermore, in the specific driver gene sets of LAML relative to solid cancer types by SpeMDP, *GNAQ* (with only two mutations in 164 LAML patients) has showed potential implication with LAML carcinogenesis and progression, but it cannot be detected by gene-centric approaches.

We obtain the common driver gene sets of all pairs of the 11 cancer types with $K = 2$ to 10 (Suppl. Table S1), and note that the significance of common driver gene sets has no transitivity. For example, there are significant common driver gene sets between LAML and COADREAD as well as LAML and LUAD for $K = 3$ to 10 (Suppl. Table S7 and S8). But there are no significant ones between COADREAD and LUAD for $K = 2$ to 10. In contrast, there are no significant common gene sets between GBM and HNSC as well as HNSC and OV for $K = 2$ to 10, but there are significant ones between GBM and OV for $K = 3$ to 10 (Suppl. Table S9).

In this study, to identify common driver gene sets, we first select the genes which have mutations in all the examined cancer types for further analysis. In fact, this model can be generalized to include the genes which have no mutations in some of the considered cancer types. We may add some constrains to ensure that the number of non-mutation cancer types is not more than a preassigned number for any considered gene. Moreover, it can also be used to investigate the commonalities and specificities among different subtypes within a certain cancer. We expect that our methods can provide crucial information for understanding the molecular mechanism of cancer generation and progression.

5 AVAILABILITY

The methods are implemented in the MATLAB code and are available upon request.

6 SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

7 ACKNOWLEDGEMENTS

8 FUNDING

REFERENCES

[1] The Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061-1068.

[2] International Cancer Genome Consortium. (2010) International network of cancer genome projects. *Nature*, **464**, 993-998.

[3] Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehár,J., Kryukov,G.V., Sonkin,D., *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603-607.

[4] Zhang,S., Liu,C.C., Li,W., Shen,H., Laird,P.W. and Zhou,X.J. (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.*, **40**, 9379-9391.

[5] Greenman,C., Stephens,P., Smith,R., Dalgliesh,G.L., Hunter,C., Bignell,G., Davies,H., Teague,J., Butler,A., Stevens,C., *et al.* (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446**, 153-158.

[6] Stratton,M.R., Campbell,P.J. and Futreal,P.A. (2009) The cancer genome. *Nature*, **458**, 719-724.

[7] Beroukhim,R., Getz,G., Nghiemphu,L., Barretina,J., Hsueh,T., Linhart,D., Vivanco,I., Lee,J.C., Huang,J.H., Alexander,S., *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci.*, **104**, 20007-20012.

[8] Vogelstein,B. and Kinzler,K.W. (2004) Cancer genes and the pathways they control. *Nat. Med.*, **10**, 789-799.

[9] Ding,L., Getz,G., Wheeler,D.A., Mardis,E.R., McLellan,M.D., Cibulskis,K., Sougnez,C., Greulich,H., Muzny,D.M., Morgan,M.B., *et al.* (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, **455**, 1069-1075.

[10] Jones,S., Zhang,X., Parsons,D.W., Lin,J.C.H., Leary,R.J., Angenendt,P., Mankoo,P., Carter,H., Kamiyama,H., Jimeno,A., *et al.* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, **321**, 1801-1806.

[11] Ciriello,G., Cerami,E., Sander,C. and Schultz,N. (2012) Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.*, **22**, 398-406.

[12] Boca,S.M., Kinzler,K.W., Velculescu,V.E., Vogelstein,B. and Parmigiani,G. (2010) Patient-oriented gene set analysis for cancer mutation data. *Genome Biol.*, **11**, R112.

[13] Efroni,S., Ben-Hamo,R., Edmonson,M., Greenblum,S., Schaefer,C.F. and Buetow,K.H. (2011) Detecting cancer gene networks characterized by recurrent genomic alterations in a population. *PLoS ONE*, **6**, e14437.

[14] Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57-70.

[15] Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646-674.

[16] Yeang,C.H., McCormick,F. and Levine,A. (2008) Combinatorial patterns of somatic gene mutations in cancer. *FASEB J.*, **22**, 2605-2622.

[17] Zhang,J. and Zhang,S. (2016) The discovery of mutated driver pathways in cancer: models and algorithms. *IEEE/ACM Trans Comput Biol Bioinform*, doi: 10.1109/TCBB.2016.2640963.

[18] Vandin,F., Upfal,E. and Raphael,B.J. (2012) *De novo* discovery of mutated driver pathways in cancer. *Genome Res.*, **22**, 375-385.

[19] Zhao,J., Zhang,S., Wu,L.Y. and Zhang,X.S. (2012) Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics*, **28**, 2940-2947.

[20] Zhang,J., Zhang,S., Wang,Y. and Zhang,X.S. (2013) Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data. *BMC Syst. Biol.*, **7**, S4.

[21] Miller,C.A., Settle,S.H., Sulman,E.P., Aldape,K.D. and Milosavljevic,A. (2011) Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med. Genomics*, **4**, 34.

[22] Cui,Q., Ma,Y., Jaramillo,M., Bari,H., Awan,A., Yang,S., Zhang,S., Liu,L., Lu,M., O'Connor-McCourt,M., *et al.* (2007) A map of human cancer signaling. *Mol. Syst. Biol.*, **3**, 152.

[23] Klijn,C., Bot,J., Adams,D.J., Reinders,M., Wessels,L. and Jonkers,J. (2010) Identification of networks of co-occurring, tumor-related DNA copy number changes using a genome-wide scoring approach. *PLoS Comput. Biol.*, **6**, e1000631.

[24] Leiserson,M.D.M., Blokh,D., Sharan,R. and Raphael,B.J. (2013) Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput. Biol.*, **9**, e1003054.

[25] Zhang,J., Wu,L.Y., Zhang,X.S. and Zhang,S. (2014) Discovery of co-occurring driver pathways in cancer. *BMC Bioinformatics*, **15**, 271.

[26] Melamed,R.D., Wang,J., Iavarone,A. and Rabadan,R. (2015) An information theoretic method to identify combinations of genomic alterations that promote glioblastoma. *J. Mol. Cell Biol.*, **7**, 203-213.

[27] Remy,E., Rebouissou,S., Chaouiya,C., Zinovyev,A., Radvanyi,F. and Calzone,I. (2015) A modelling approach to explain mutually exclusive and co-occurring genetic alterations in bladder tumorigenesis. *Cancer Res.*, **75**, 4042-4052.

[28] Kandoth,C., McLellan,M.D., Vandin,F., Ye,K., Niu,B., Lu,C., Xie,M., Zhang,Q., McMichael,J.F., Wyczalkowski,M.A., *et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333-339.

[29] The Cancer Genome Atlas Research Network, *et al.* (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113-1120.

[30] Ciriello,G. Miller,M.L., Aksoy,B.A., Senbabaoglu,Y., Schultz,N. and Sander,C. (2013) Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.*, **45**, 1127-1133.

[31] Liu,Z. and Zhang,S. (2014) Toward a systematic understanding of cancers: a survey of the pan-cancer study. *Front Genet.*, **5**, 194.

[32] Hofree,M., Shen,J.P., Carter,H., Gross,A. and Ideker,T. (2013) Network-based stratification of tumor mutations. *Nat. Methods*, **10**, 1108-1115.

[33] Liu,Z. and Zhang,S. (2015) Tumor characterization and stratification by integrated molecular profiles reveals essential pan-cancer features. *BMC Genomics*, **16**, 503.

[34] Leiserson,M.D.M., Vandin,F., Wu,H.T., Dobson,J.R., Eldridge,J.V., Thomas,J.L., Papoutsaki,A., Kim,Y., Niu,B., McLellan,M., *et al.* (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.*, **47**, 106-114.

[35] Kim,Y.A., Cho,D.Y., Dao,P. and Przytycka,T.M. (2015) MEMCover: Integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. *Bioinformatics*, **31**, i284-i292.

[36] Szczurek,E. and Beerenwinkel,N. (2014) Modeling mutual exclusivity of cancer mutations. *PLoS Comput. Biol.*, **10**, e1003503.

[37] Hu, X., Zhang, Y., Zhang, A., Li, Y., Zhu, Z., Shao, Z., Zeng, R. and Xu, L.X. (2009) Comparative serum proteome analysis of human lymph node negative/positive invasive ductal carcinoma of the breast and benign breast disease controls via label-free semiquantitative shotgun technology. *OMICS*, **13**, 291-300.

[38] Kim, Y.S., Hwan, J.D., Bae, S., Bae, D.H. and Shick, W.A. (2010) Identification of differentially expressed genes using an annealing control primer system in stage III serous ovarian carcinoma. *BMC Cancer*, **10**, 576.

[39] Wang, J., Sun, Y., Qu, J., Yan, Y., Yang, Y. and Cai, H. (2016) Roles of *LPA* receptor signaling in breast cancer. *Expert Rev. Mol. Diagn.*, **16**, 1103-1111.

[40] Campbell, T.M., Castro, M.A., de Santiago, I., Fletcher, M.N., Halim, S., Prathalingam, R., Ponder, B.A. and Meyer, K.B. (2016) *FGFR2* risk SNPs confer breast cancer risk by augmenting oestrogen responsiveness. *Carcinogenesis*, **37**, 741-750.

[41] Jesionowska, A., Cecerska-Heryc, E., Matoszka, N. and Dolegowska, B. (2015) Lysophosphatidic acid signaling in ovarian cancer. *J. Recept. Signal Transduct. Res.*, **35**, 578-584.

[42] Cole, C., Lau, S., Backen, A., Clamp, A., Rushton, G., Dive, C., Hodgkinson, C., McVey, R., Kitchener, H. and Jayson, G.C. (2010) Inhibition of *FGFR2* and *FGFR1* increases cisplatin sensitivity in ovarian cancer. *Cancer Biol. Ther.*, **10**, 495-504.

[43] Elizalde, P., Cordo Russo, R.I., Chervo, M.F. and Schillaci, R. (2016) *ErbB-2* nuclear function in breast cancer growth, metastasis, and resistance to therapy. *Endocr. Relat. Cancer*, pii: ERC-16-0360.

[44] Hodeib, M., Serna-Gallegos, T. and Tewari, K.S. (2015) A review of *HER2*-targeted therapy in breast and ovarian cancer: lessons from antiquity - CLEOPATRA and PENELOPE. *Future Oncol.*, **11**, 3113-3131.

[45] Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nat. Protoc.*, **4**, 44-57.

[46] Kim,M., Hernandez,L. and Annunziata,C.M. (2016) *Caspase 8* expression may determine the survival of women with ovarian cancer. *Cell Death Dis.*, **7**, e2045.

[47] Park,H.L., Ziogas,A., Chang,J., Desai,B., Bessonova,L., Garner,C., Lee,E., Neuhausen,S.L., Wang,S.S., Ma,H., *et al.* (2016) Novel polymorphisms in *caspase-8* are associated with breast cancer risk in the California Teachers Study. *BMC Cancer*, **16**, 14.

[48] Sagulenko,V., Lawlor,K.E. and Vince,J.E. (2016) New insights into the regulation of innate immunity by caspase-8. *Arthritis Res. Ther.*, **18**, 4.

[49] Kim, R.K., Kaushik, N., Suh, Y., Yoo, K.C., Cui, Y.H., Kim, M.J., Lee, H.J., Kim, I.G. and Lee, S.J. (2016) Radiation driven epithelial-mesenchymal transition is mediated by *Notch* signaling in breast cancer. *Oncotarget*, doi: 10.18632/oncotarget.10802. [Epub ahead of print]

[50] Sehrawat, A., Sakao, K. and Singh, S.V. (2014) *Notch2* activation is protective against anticancer effects of zerumbone in human breast cancercells. *Breast Cancer Res. Treat.*, **146**, 543-555.

[51] Niwakawa,M., Hashine,K., Yamaguchi,R., Fujii,H., Hamamoto,Y., Fukino,K., Tanigawa,T. and Sumiyoshi,Y. (2012) Phase I trial of sorafenib in combination with interferon-alpha in Japanese patients with unresectable or metastatic renal cell carcinoma. *Invest. New Drugs*, **30**, 1046-1054.

[52] Hasan,S., Hassa,P.O., Imhof,R. and Hottiger,M.O. (2001) Transcription coactivator p300 binds PCNA and may have a role in DNA repair synthesis. *Nature*, **410**, 387-391.

[53] Su,Y., Subedee,A., Bloushtain-Qimron,N., Savova,V., Krzystanek,M., Li,L., Marusyk,A., Tabassum,D.P., Zak,A., Flacker,M.J., *et al.* (2015) Somatic Cell Fusions Reveal Extensive Heterogeneity in Basal-like Breast Cancer. *Cell Rep.*, **11**, 1549-1563.

[54] Huang,R., Ding,P. and Yang,F. (2015) Clinicopathological significance and potential drug target of *CDH1* in breast cancer: a meta-analysis and literature review. *Drug Des. Devel. Ther.*, **9**, 5277-5285.

[55] Mustafa,M., Lee,J.Y. and Kim,M.H. (2015) *CTCF* negatively regulates *HOXA10* expression in breast cancer cells. *Biochem. Biophys. Res. Commun.*, **467**, 828-834.

[56] Asp,N., Kvalvaag,A., Sandvig,K. and Pust,S. (2016) Regulation of *ErbB2* localization and function in breast cancer cells by *ERM* proteins. *Oncotarget*, **7**, 25443-25460.

[57] Irish,J.C., Mills,J.N., Turner-Ivey,B., Wilson,R.C., Guest,S.T., Rutkovsky,A., Dombkowski,A., Kappler,C.S., Hardiman,G. and Ethier,S.P. (2016) Amplification of *WHSC1L1* regulates expression and estrogen-independent activation of *ER* in *SUM-44* breast cancer cells and is associated with *ER* over-expression in breast cancer. *Mol. Oncol.*, **10**, 850-865.

[58] Asch-Kendrick,R. and Cimino-Mathews,A. (2016) The role of *GATA3* in breast carcinomas: a review. *Hum. Pathol.*, **48**, 37-47.

[59] Döhner,H., Weisdorf,D.J. and Bloomfield,C.D. (2015) Acute Myeloid Leukemia. *N. Engl. J. Med.*, **373**, 1136-1152.

[60] Gale,R.E., Green,C., Allen,C., Mead,A.J., Burnett,A.K., Hills,R.K., Linch,D.C. and Medical Research Council Adult Leukaemia Working Party. (2008) The impact of *FLT3* internal tandem duplication mutant level, number, size, and interaction with *NPM1* mutations in a large cohort of young adult patients with acute myeloid leukemia. *Blood*, **111**, 2776-2784.

[61] Hefazi,M., Siddiqui,M., Patnaik,M., Wolanskyj,A., Alkhateeb,H., Zblewski,D., Elliott,M., Hogan,W., Litzow,M. and Al-Kali,A. (2015) Prognostic impact of combined *NPM1+/FLT3-* genotype in patients with acute myeloid leukemia with intermediate risk cytogenetics stratified by age and treatment modalities. *Leuk. Res.*, **39**, 1207-1213.

[62] Alpermann,T., Schnittger,S., Eder,C., Dicker,F., Meggendorfer,M., Kern,W., Schmid,C., Aul,C., Staib,P., Wendtner,C.M., *et al.* (2016) Molecular subtypes of *NPM1* mutations have different clinical profiles, specific patterns of accompanying molecular mutations and varying outcomes in intermediate risk acute myeloid leukemia. *Haematologica*, **101**, pp. e55-8.

[63] Chiu,Y.C., Tsai,M.H., Chou,W.C., Liu,Y.C., Kuo,Y.Y., Hou,H.A., Lu,T.P., Lai,L.C., Chen,Y., Tien,H.F., *et al.* (2016) Prognostic significance of *NPM1* mutation-modulated microRNA-mRNA regulation in acute myeloid leukemia. *Leukemia*, **30**, 274-284.

[64] Royer-Bertrand, B., Torsello, M., Rimoldi, D., El Zaoui, I., Cisarova, K., Pescini-Gobert, R., Raynaud, F., Zografos, L., Schalenbourg, A., Speiser, D., *et al.* (2016) Comprehensive genetic landscape of uveal melanoma by whole-genome sequencing. *Am. J. Hum. Genet.*, pii: S0002-9297(16)30388-3. doi: 10.1016/j.ajhg.2016.09.008.

[65] Sheng, X., Kong, Y., Li, Y., Zhang, Q., Si, L., Cui, C., Chi, Z., Tang, B., Mao, L., Lian, B., *et al.* (2016) *GNAQ* and *GNA11* mutations occur in 9.5% of mucosal melanoma and are associated with poor prognosis. *Eur. J. Cancer*, **65**, 156-163.

[66] Marjanovic, I., Kostic, J., Stanic, B., Pejanovic, N., Lucic, B., Karan-Djurasevic, T., Janic, D., Dokmanovic, L., Jankovic, S., Vukovic, N.S., *et al.* (2016) Parallel targeted next generation sequencing of childhood and adult acute myeloid leukemia patients reveals uniform genomic profile of the disease. *Tumour Biol.*, [Epub ahead of print]

[67] The Cancer Genome Atlas Research Network. (2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.*, **368**, 2059-2074.
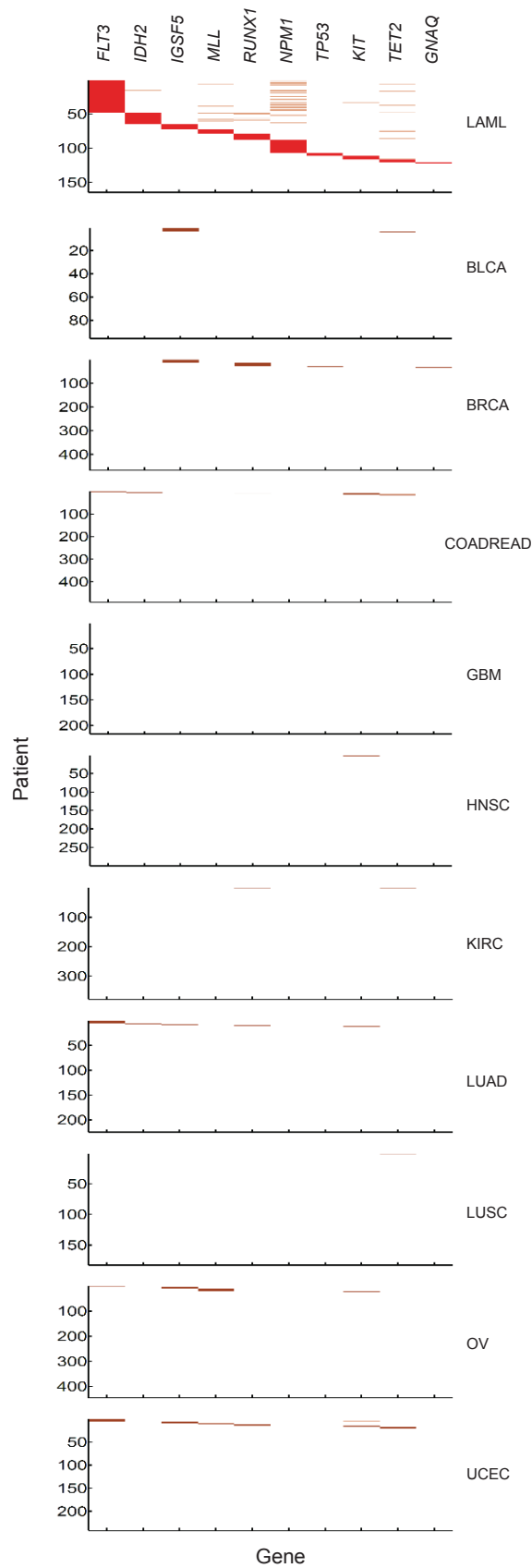
Fig. 7. The heat map of the alterations in the LAML specific driver gene set (*FLT3, IDH2, IGSF5, MLL, RUNX1, NPM1, TP53, KIT, TET2, GNAQ*) relative to other ten cancer types including BLCA, BRCA, COADREAD, GBM, HNSC, KIRC, LUAD, LUSC, OV and UCEC.