Discovery of cancer common and specific driver gene sets

Junhua Zhang1,* and Shihua Zhang1,2,*

1National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China and 2School of Mathematics Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

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.

INTRODUCTION

Cancer 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 (9–11). 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 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: (i) covering a large number of samples (high coverage); and (ii) 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 (18–20). 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 (18–21).

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 specificities in pathway level.

In this study, we develop two mathematical programming models (CoMDP 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 CoMDP 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. 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, IDHI, 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.

### MATERIALS AND METHODS

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

#### The maximum weight submatrix problem

Given a binary mutation matrix $A$ with $m$ rows (samples) and $n$ columns (genes), Vandijk 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 $W$:

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

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 $M$.

#### CoMDP 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
We propose the following BLP model to maximize it:

\[
B_t \text{maximize } F_m \left( x, u \right) = \sum_{r=1}^{R} \left[ 2 \sum_{i=1}^{m_r} x_i^{(r)} - \sum_{j=1}^{n} \left( u_j \cdot \sum_{i=1}^{m_r} a_{ij}^{(r)} \right) \right] \tag{3}
\]

\[
\text{s.t. } \begin{align*}
x_i^{(r)} & \leq \sum_{j=1}^{n} a_{ij}^{(r)} u_j, \quad i = 1, \ldots, m_r, \quad r = 1, \ldots, R, \\
\sum_{j=1}^{n} u_j & = K, \\
x_i^{(r)}, u_j & \in \{0, 1\}, \quad i = 1, \ldots, m_r, \quad j = 1, \ldots, n, \quad r = 1, \ldots, R,
\end{align*}
\tag{4}
\]

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

**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 \( B_t = (b_{ij}^{(t)}) \) (\( t = 1, \ldots, T \)) to denote corresponding mutation matrices, respectively. We introduce the weight function \( S_m \):

\[
S_m(M) = \frac{1}{R} \sum_{r=1}^{R} [K|\Gamma_A(M)| - \sum_{g \in M} |\Gamma_A(g)|] - \frac{1}{T} \sum_{t=1}^{T} [K|\Gamma_B(M)| - \sum_{g \in M} |\Gamma_B(g)|].
\tag{5}
\]

We maximize \( S_m \) by the following BLP model:

\[
\text{max } G_m(x, y, u) = \frac{1}{R} \sum_{r=1}^{R} \left[ K \sum_{j=1}^{n} y_{ij}^{(r)} - \sum_{j=1}^{n} \left( u_j \cdot \sum_{i=1}^{m_r} a_{ij}^{(r)} \right) \right] - \frac{1}{T} \sum_{t=1}^{T} \left[ K \sum_{k=1}^{n} y_{kt}^{(t)} - \sum_{k=1}^{n} \left( u_k \cdot \sum_{i=1}^{m_t} b_{ik}^{(t)} \right) \right],
\tag{6}
\]

\[
\left\{ \begin{align*}
x_i^{(r)} & \leq \sum_{j=1}^{n} a_{ij}^{(r)} u_j, \quad i = 1, \ldots, m_r, \quad r = 1, \ldots, R, \\
\frac{1}{n} \sum_{j=1}^{n} b_{ij}^{(t)} u_j \leq y_{ij}^{(t)} \leq \sum_{j=1}^{n} b_{ij}^{(t)} u_j, \quad i = 1, \ldots, m_r, \quad j = 1, \ldots, n, \quad r = 1, \ldots, R, \\
\sum_{j=1}^{n} u_j & = K, \\
x_i^{(r)}, u_j & \in \{0, 1\}, \quad i = 1, \ldots, m_r, \quad j = 1, \ldots, n, \quad r = 1, \ldots, R, \quad t = 1, \ldots, T.
\end{align*} \right.
\tag{7}
\]

where \( x = \{ x^{(1)}, \ldots, x^{(R)} \}, \ y = \{ y^{(1)}, \ldots, y^{(T)} \} \). 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 \( \frac{1}{n} \sum_{j=1}^{n} b_{ij}^{(t)} u_j \leq y_{ij}^{(t)} \leq \sum_{j=1}^{n} b_{ij}^{(t)} u_j \) to ensure that \( \sum_{k=1}^{n} y_{kt}^{(t)} \) 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.

**Statistical significance**

We perform a permutation test to assess the significance of results. We permute the mutations independently among samples to preserve the mutation frequency of each gene. Two kinds of significance are calculated: (i) 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; (ii) 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.

**Simulated data**

To assess the performance of the proposed methods on a variety of data, we construct eight datasets, sd1, ..., 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 $p_{ij}^{(r)}$ 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 $p_{ij}^{(r)}$ into each matrix $A_r$ ($r = 1, \ldots, 3; i = 1, \ldots, I; I = 9; p_{ij}^{(1)} = 0.80, p_{ij}^{(2)} = 0.85, p_{ij}^{(3)} = 0.90$), where for each $r$, $M_i(r)$ contains $i + 1$ genes ($i = 1, \ldots, I$), and for each $i$, these submatrices $M_i(r)$ occupy the same columns in the corresponding $A_r$ ($r = 1, \ldots, 3$). For each sample in $A_r$, a gene uniformly chosen from $M_i(r)$ is mutated with rate $p_{ij}^{(r)}$, and once one gene is mutated, the other genes in $M_i(r)$ have a rate $p_{ij}^0$ to be mutated ($p_{ij}^0 = 0.04$). 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

$$p_{ij}^{(r)} = 1 - i \cdot \delta(r)(r = 1, \ldots, 3; i = 1, \ldots, 9),$$

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,

- 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 six genes in each are embedded into $A_r$, with $m = 500$, $n = 700$, $r = 1, \ldots, 3$. The first group of submatrices are constructed in the first embedding manner (thus corresponding to a common driver gene set); each of the second to the fourth 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).

**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 multiforme (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.

**RESULTS**

**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).

**Common driver gene set discovery.** For sd1 ComMDP can identify the embedded gene sets for all the ten runs when the number of genes $K \leq 8$ (Figure 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 (Figure 2B).
Figure 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: (i) two kinds of exclusive gene sets corresponding to a common driver pathway, so it cannot be found; (ii) exclusive (red) versus nonexclusive (brown) gene sets can be found; (iii) 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 $I = 9$.

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 (Figure 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 (Figure 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 (Figure 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 $i = 2$ (Figure 2F). Lastly, the result on sd7 indicates that SpeMDP is also effective to identify specific gene sets for multiple cancer types (Figure 2G), where the dataset is simulated in a similar way to that of sd6.

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 Figure 3. Then we apply ComMDP and SpeMDP to identify the common and specific driver gene sets among all the $A_r$s, 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 gene sets in the second and third mutation matrices ($A_2$ and $A_3$) are not specific (Figure 3).
Figure 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).

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 (Supplementary Table S1). The mutation rates of $TP53$ in LUSC and OV are very high (164/182 and 405/445, respectively). We distinguish these 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, etc. Based on known KEGG pathway knowledge (Figure 4A), we propose a common mutated pathway model for cancer initiation and progression in both BRCA and OV (Figure 4B). We show the heat map of the alterations of the gene set for $K = 10$ (Supplementary Figure S1) and see that $TP53$ has a very high mutation rate in OV (as stated 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 set exploration helps to identify driver pathways with low mutation frequency (Figure 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 (Figure 4B) can be found by BLP for each of these two cancers (Figure 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 (Figure 4B) and are closely related to the carcinogenesis of BRCA and OV (37–40). 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 (Figure 4), whereas ComMDP can integrate information from different cancers and imply a more biologically reasonable common driver pathway.

Importantly, identifying all significant common driver gene sets of BRCA with certain cancer types will help
Figure 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 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 \(7 \leq K \leq 10\)) and (b) OV, according to the detected genes in Table 2.

Table 1. Significant common driver gene sets between BRCA and OV identified by ComMDP

| \(K\) | Common gene set | \(p_1\) | \(p_2\) | \(p\) |
|-------|----------------|--------|--------|------|
| 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.

Table 2. Significant individual driver gene sets in BRCA and OV by BLP

| \(K\) | Driver gene set in BRCA | Driver gene set 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, BRAF, LPA, EGFR |
| 6     | TP53, PIK3CA, GATA3, CDH1, CTCF, MACROD2 | TP53, KRAS, IDI2, BRAF, LPA, EGFR, C4orf45 |
| 7     | TP53, GATA3, CDH1, MACROD2, AKT1, MAP3K1, MAP2K4 | TP53, KRAS, FGFR2, C4orf45, EPHA3, PPID, ETFDH, FNIP2 |
| 8     | TP53, GATA3, CDH1, MACROD2, AKT1, MAP3K1, MAP2K4, PIK3R1 | TP53, KRAS, FGFR2, C4orf45, EPHA3, PPID, FNIP2 |
| 9     | TP53, GATA3, CDH1, MACROD2, AKT1, MAP3K1, MAP2K4, PIK3R1, POLD4 | TP53, KRAS, FGFR2, PIGV, EGFR, C4orf45, EPHA3, PPI, ETFDH, 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.
to understand various aspects of BRCA carcinogenesis. Besides OV, other cancer types include BLCA, COADREAD, GBM, HNSC, KIRC, LUAD and UCEC (Supplementary Table S1). In total, we discover 38 different genes in all eight significant gene sets (Supplementary Figure S2). These genes are involved in many important signaling pathways (Supplementary 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 (Figure 5). Compared to Figure 4, we find some new paths and more genes involving in the important hallmarks of cancer in Figure 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–MAPK4] 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 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), PI3K 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, TXN1B, NOTCH2 (Supplementary 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 (46–48); 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 (Supplementary Table S1). For BRCA stated above, some genes frequently appear in many common gene sets (Supplementary Figure S2). But for BLCA, the common gene sets are not necessarily the same, such as those with BRCA (Supplementary Table S2), COADREAD (Supplementary Table S3), GBM (Supplementary Table S4) and LUSC (Supplementary Table S5). For example, we identify two different sets of genes for BRCA and COADREAD, whereas their functional annotations are quite similar (Supplementary 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 $K = 4$ (Figure 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 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, WHSC1LI, CCND1, PLK1, RPL4A, DDAH1, many of which have been suggested to be closely related with breast cancer initiation and progression by a number of studies (54–57). 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 gene sets relative to the liquid cancer LAML with $K = 9, 10$ (Supplementary 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
Table 3. Significant common driver gene sets among BRCA, OV, LUAD and GBM

| K  | Common gene set                              | $p_1$  | $p_2$  | $p_3$  | $p_4$  | $p$     |
|----|---------------------------------------------|--------|--------|--------|--------|---------|
| 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, EF300    | 0      | 0.0080 | 0.0040 | 0      | 0       |
| 7  | TP53, IFNA6, PIK3CA, KRAS, MAP3K1, EF300, PIK3R1 | 0      | 0.0040 | 0      | 0      | 0       |
| 8  | TP53, IFNA6, PIK3CA, KRAS, MAP3K1, EF300, PIK3R1, TNXB | 0      | 0.0020 | 0.0010 | 0      | 0       |
| 9  | TP53, IFNA6, PIK3CA, KRAS, MAP3K1, EF300, PIK3R1, TNXB, KDM6A | 0      | 0.0040 | 0.0020 | 0      | 0       |
| 10 | TP53, IFNA6, PIK3CA, KRAS, MAP3K1, EF300, 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.

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

| Type         | K  | Specific gene set                          | $p$     | $q$     | $P$     |
|--------------|----|-------------------------------------------|---------|---------|---------|
| 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, CTNF, PLK1, RPPL4A | 0.0410  | 0.7810  | 0.0420  |
|              | 10 | ERBB2, GATA3, CDH1, WHSC1L1, CCND1, AKT1, CTNF, PLK1, RPPL4A, 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               | 1.000e-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, COX42 | 0       | 0.9970  | 0       |
|              | 8  | BRCA1, BRCA2, WT1, ADPRHL2, METTL17, DNM2, COX42, SRP19 | 0       | 1       | 0       |
|              | 9  | BRCA1, BRCA2, WT1, ADPRHL2, METTL17, DNM2, COX42, SRP19, PARP8 | 0       | 1       | 0       |
|              | 10 | BRCA1, BRCA2, WT1, ADPRHL2, METTL17, DNM2, COX42, 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 gene set            | $p$     | $q_1$, $q_2$, $q_3$, $q_4$, $q_5$, $q_6$, $q_7$, $q_8$, $q_9$, $q_{10}$ | $P$     |
|----|------------------------------|---------|-----------------------------------------------|---------|
| 2  | FLT3, IDH2                   | 0.0150  | 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 0.9980, 0.9990, 1.0000, 1.0000 | 0.0170  |
| 3  | FLT3, IDH2, NRAS             | <0.0001 | 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000 | 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, 0.9690, 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, 1.0000, 1.0000, 1.0000, 1.0000 | 0.0130  |
| 7  | FLT3, IDH2, IGSF5, RUNX1, NPM1, TP53, KIT | 0.0340  | 0.4720, 0.9940, 0.9930, 1.0000, 1.0000, 0.9360, 0.9410, 1.0000, 1.0000, 1.0000 | 0.0240  |
| 8  | FLT3, IDH2, IGSF5, MLL, RUNX1, NPM1, TP53, KIT, TET2 | 0.0050  | 0.5020, 0.9950, 0.1000, 1.0000, 1.0000, 0.9980, 0.9980, 0.9980, 0.9980, 0.9980 | 0.0050  |
| 9  | FLT3, IDH2, IGSF5, MLL, RUNX1, NPM1, TP53, KIT, TET2 | 0.0130  | 0.4890, 0.9900, 0.9960, 0.9960, 0.9960, 0.9960, 0.9960, 0.9960, 0.9960, 0.9960 | 0.0050  |
| 10 | FLT3, IDH2, IGSF5, MLL, GNAQ, RUNX1, NPM1, TP53, KIT, TET2 | 0.0090  | 0.3820, 0.9780, 0.9980, 0.9980, 0.9980, 0.9980, 0.9980, 0.9980, 0.9980, 0.9980 | 0.0030  |

$p$, $q_1$, $q_2$, $q_3$, $q_4$, $q_5$, $q_6$, $q_7$, $q_8$, $q_9$, $q_{10}$ 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 $q_1$, $q_2$, $q_3$, $q_4$, $q_5$, $q_6$, $q_7$, $q_8$, $q_9$, $q_{10}$ are all larger than 0.05.
Figure 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 represents a known interaction between them or occurrence of the mutations in a known signaling pathway. The ComMDP discovered genes are in red.

of FLT3, IDH2, NRAS, IDH1, RUNX1, NPM1, TET2, KIT, amplifications of MLL, IGF5 and deletions of TP53, GNAQ. We show the heat map of the alterations of the gene sets in all the 11 cancer types for $K = 10$ (Figure 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 (Figure 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 (61–63); 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 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 is worth further exploration.

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 are worthy to be further explored.

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
Figure 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.

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 (Figure 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 (Supplementary 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 (Supplementary Tables 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 (Supplementary Table S9).

Figure 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.
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 constraints 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.

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

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
We would like to thank Dr Yong Wang from the Academy of Mathematics and Systems Science, CAS for his suggestion.

FUNDING
National Natural Science Foundation of China [61379092, 61422309, 61621003 and 11661141019]; Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) [XDB13040600]; Outstanding Young Scientist Program of CAS, CAS Frontier Science Research Key Project for Top Young Scientist [QYZDB-SSW-SYS008]; Key Laboratory of Random Complex Structures and Data Science, CAS [2008DP173182]. Funding for open access charge: National Natural Science Foundation of China. Conflict of interest statement. None declared.

REFERENCES
1. McLendon,R., Friedman,A., Bigner,D., Van Meir,E.G., Brat,D.J., Mastrogiannakis,G.M., Olson,J.J., Mikkelsen,T., Lehman,N., Aldape,K. et al. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, 455, 1061–1068.
2. Hudson,T.J., Anderson,W., Artez,A., Barker,A.D., Bell,C., Bernabé,R.R, Bhan,M.K., Calvo,F., Eerola,I., Gerhard,D.S. et al. (2010) International network of cancer genome projects. Nature, 464, 993–998.
3. Barretina,J., Caponigro,G., Stranks,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.

Figure 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 IDHI 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))
