SCIENTIFIC REPORTS

OPEN

Adaptively Weighted and Robust Mathematical Programming for the Discovery of Driver Gene Sets in Cancers

Received: 24 July 2018
Accepted: 28 March 2019
Published online: 11 April 2019

Xiaolu Xu[1], Pan Qin[1], Hong Gu[1], Jia Wang[2] & Yang Wang[3]

High coverage and mutual exclusivity (HCME), which are considered two combinatorial properties of mutations in a collection of driver genes in cancers, have been used to develop mathematical programming models for distinguishing cancer driver gene sets. In this paper, we summarize a weak HCME pattern to justify the description of practical mutation datasets. We then present AWRMP, a method for identifying driver gene sets through the adaptive assignment of appropriate weights to gene candidates to tune the balance between coverage and mutual exclusivity. It embeds the genetic algorithm into the subsampling strategy to provide the optimization results robust against the uncertainty and noise in the data. Using biological datasets, we show that AWRMP can identify driver gene sets that satisfy the weak HCME pattern and outperform the state-of-arts methods in terms of robustness.

Driver mutations, which are the mutations responsible for cancer, are different from randomly occurring passenger mutations. Because driver mutations typically target genes involved in cellular signalling and regulatory pathways[1,2]. The examination of these mutations in the context of pathways and gene sets is an essential issue in cancer genome research. However, an exhaustive search for driver pathways is impossible due to the enormous number of gene set candidates. Thus, prior knowledge, such as mutation patterns, is often used as a constraint to limit the scale of the gene set candidates. In particular, high coverage and mutual exclusivity (HCME), two combinatorial properties of driver mutations in a cellular signalling pathway or regulatory pathway[2,3], are being used as important prior knowledge in *de novo* discovery methods for driver gene sets (i.e., groups of mutated driver genes)[4–19]. High coverage means that the members in the driver gene set recurrently occur in patient cohorts, and mutual exclusivity means that almost all the patients exhibit no more than one single driver mutation event in the driver gene set. For the developments of state-of-art discovery methods for cancer driver pathways, readers are referred to the latest survey by Zhang and Zhang[20].

The mathematical programming models for the *de novo* discovery of driver gene sets can be deduced from the HCME pattern. Vandin *et al.* developed the Dendrix algorithm, which investigates the optimal gene set by maximizing a HCME-derived score function[4]. The scoring function in Dendrix was formulated by the cardinalities of sets of patients and genes, and thus, the function was not sufficiently explicit for the optimization design. To this end, Zhao *et al.* further developed an explicit binary linear programming model and optimization framework, called MDPfinder, for the scoring system[5]. Zhao *et al.* initially introduced the genetic algorithm (GA)[21] for this problem[5]. Leiserson *et al.* generalized Dendrix for the batch discovery of multiple driver gene sets[6]. Zhang *et al.* developed CoMDP to identify co-occurring driver gene sets[7]. Zhang *et al.* proposed ComMDP and SpeMDP to investigate common and specific driver gene sets among multiple cancer types, respectively[8]. In addition to the mathematical programming based *de novo* discovery methods, several probabilistic and statistical approaches have also been proposed. For example, Constantinescu *et al.* proposed TiMEx, a probabilistic generative model for the identification of mutually exclusive patterns[17]. Leiserson *et al.* proposed CoMEt for the identification of

[1]Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China. [2]Department of Breast Surgery, Institute of Breast Disease, Second Hospital of Dalian Medical University, Dalian, China. [3]Institute of Cancer Stem Cell, Dalian Medical University, Dalian, China. Correspondence and requests for materials should be addressed to P.Q. (email: qp112cn@dlut.edu.cn) or J.W. (email: wangjia77@hotmail.com)

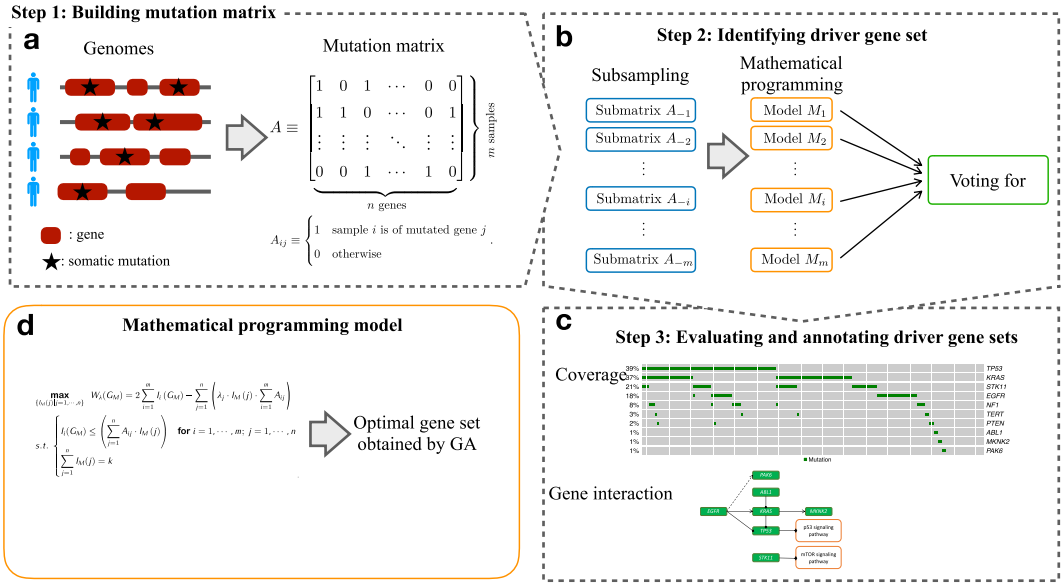


Figure 1. Overview of AWRMP. (**a**) We constructed binary-valued mutation matrices from mutation data files. (**b**) We used subsampling to make our method robust against the uncertainty and noise in the data. (**c**) The optimal gene set was evaluated based on coverage and exclusivity scores and annotated to analyse the gene interactions using DAVID. (**d**) We proposed a new mathematical programming model that uses adaptive weights to tune the balance between the coverage and mutual exclusivity.

genes exhibiting mutual exclusivity[18]. Kim *et al.* proposed WeSME, a computational cost saving method for the permutation test in the discovery of mutual exclusivity[19].

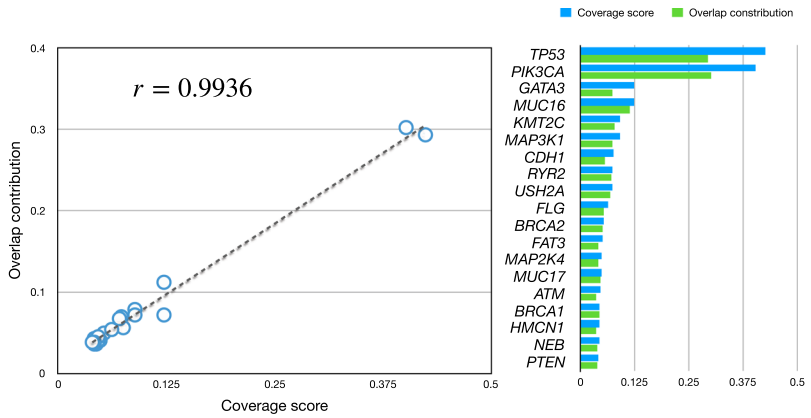
The assumption of mutual exclusivity implies that a patient exhibits no more than one crucial mutation event. Thus, this assumption is strong for the discovery of the driver gene sets from the mutation data of a cohort of patients. As indicated by[16], the application of such a strong assumption can lead to a highly unbalanced pattern, in which a single frequently mutated gene is coupled to several other infrequently mutated genes to satisfy the assumption of mutual exclusivity. By observing the mutation patterns in critical cancer driver pathways (Supplementary Fig. 1), we found that a gene that is mutated in many patients always overlaps with other genes. The coverage of an individual gene is positively correlated with its overlap with other genes in a pathway. On the basis of this fact, we proposed the following weak HCME pattern for discovering a driver gene set from a cohort of patients: (a) the members in the driver gene set recurrently occur in a patient cohort; (b) the members in the driver gene set approximately satisfy mutual exclusivity and (c) the overlaps should be adequately permitted and the members that cover many patients can have relatively more overlaps than the rarely mutated members. On the other hand, the mutation datasets used in the *de novo* discovery methods are commonly sparse, i.e., the total number of patients (samples) is smaller than that of genes (variables). Similar to other data-driven inference methods, the sparseness of datasets presents another challenge for ensuring the robustness of *de novo* discovery methods.

Here, we introduce adaptively weighted and robust mathematical programming (AWRMP) for identifying driver gene sets that satisfy the weak HCME pattern. We constructed mathematical programming models using the cardinalities of sets of patients associated with the mutated genes as adaptive weights, for tuning the balance of importance between coverage and mutual exclusivity, to construct mathematical programming models. Motivated by[5], GA[21] was used as the basic optimization solver to efficiently solve the optimization problem. GA was embedded in a systematic subsampling strategy to obtain robust solutions against uncertainty and noise in the mutation data. Additionally, the subsampling approach can identify a parsimonious gene set, whose dimension can be considered a low bound for the dimension of the driver gene set in the sense of robustness. We applied our method to several biological datasets, and the results showed that our method identified rational driver gene sets. We then tested the significance of mutual exclusivity on our results using CoMEt[18] and TiMEx[17], and proved the robustness of AWRMP through a disturbance test.

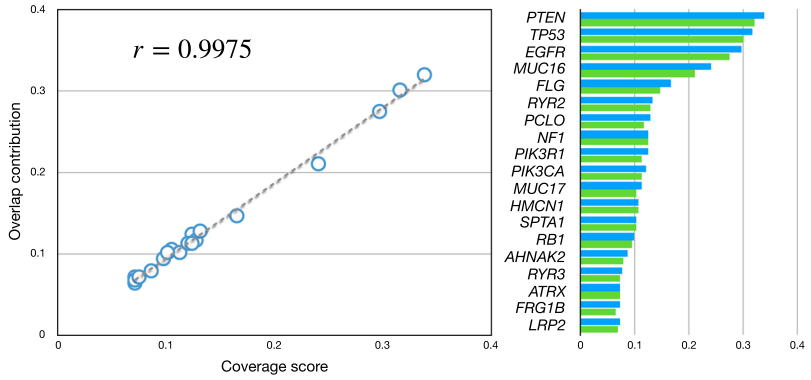
Results

AWRMP workflow. The AWRMP procedure can be divided into three modules as follows (Fig. 1). We first converted the mutation data into a binary-valued matrix A with m rows (samples) and n columns (genes). Each element $A_{ij} \in \{0, 1\}$ of A was defined as

$$A_{ij} = \begin{cases} 1 & \text{gene } j \text{ is mutated in sample } i \\ 0 & \text{otherwise} \end{cases}. \quad (1)$$



(a) Apoptosis pathway in BC



(b) ERBB pathway in GBM

Figure 2. Scatter plots of the coverage score against the overlap contribution for (**a**) the apoptosis pathway of BC and (**b**) the ERBB pathway of GBM.

We constructed a binary integer programming model on the basis of weak HCME, which is used to investigate optimal submatrix of A. Compared with Dendrix and its extensions, we embedded the adaptive weights to tune the balance between coverage and mutual exclusivity. GA was used as the optimization solver. We further integrated GA with a systematic subsampling strategy[22] to eliminate the uncertainty and noise in the mutation data, and then annotated and evaluated the identified gene sets using DAVID[23].

Correlations between coverage score and overlap contribution. By observing the critical cancer driver pathways, we found that the coverage score of a mutated gene defined by formula (19) and its overlap contribution defined by formula (22) are highly positively correlated. For example, Supplementary Fig. 1 illustrates coMut plots of the somatic mutations in the apoptosis pathway obtained from the breast cancer (BC) mutation data[24] and in the ErbB pathway obtained from glioblastoma (GBM) mutation data[25]. The two plots showed that the mutated genes approximately satisfied HCME. However, the genes with high coverage scores showed many overlaps with other genes, such as *TP53*, *PIK3CA*, *EGFR*, and *PTEN* in the two pathways. Figure 2 illustrates two scatter plots of the coverage score against the overlap contribution for all the genes in the two pathways. The correlation coefficient corresponding to the apoptosis pathway for BC was 0.9936; and that corresponding to the ErbB pathway for GBM was 0.9975. Therefore, the proper overlaps should be considered to identify the driver gene sets from mutation data from a cohort of patients.

Identified driver gene set for lung adenocarcinoma. Lung adenocarcinoma (LUAD) is the most common histological type of lung cancer. To illustrate the performance of AWRMP, we applied AWRMP to LUAD mutation data[26], which was also previoustly used to test Dendrix[4]. The variable k denotes the gene set dimension that is pre-defined to be identified by AWRMP. The gene sets obtained from the LUAD data with $k = 2, 3, \ldots, 10$ were investigated (Fig. 3). *TP53*, *KRAS*, *EGFR*, and *STK11*, which have relatively high mutation frequencies, were always included in the identified gene sets obtained with k values larger than 4 (Fig. 3(a)). The identified gene sets became nested with increasing values of k (Fig. 3(b)).

For $k = 2$, the gene set (*KRAS, TP53*) was identified by AWRMP with a subsampling rate of 1 obtained using Eq. (13). In contrast, the set (*KRAS, EGFR*) was identified by Dendrix[4]. For $k = 3$, the triplet (*EGFR*, *KRAS*, *TP53*) was the unique optimal gene set selected by AWRMP with a subsampling rate 1 calculated using Eq. (13) in Methods. This gene set was mutated in 119 patients with an overlap score of 0.7059, which was obtained using

Figure 3. Nested gene sets identified by AWRMP for the gene set dimensions $k = 2, 3, 4, \ldots, 10$. (**a**) Subsampling rates and coverage scores of genes obtained using Eqs (13) and (19) in Methods, respectively. (**b**) Elements of gene sets obtained with $k = 2, 3, \ldots, 10$.

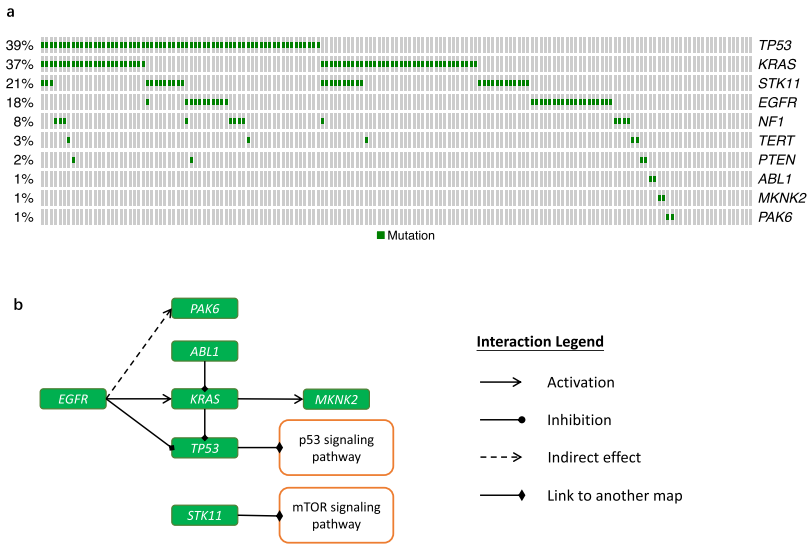


Figure 4. Optimal gene set identified by AWRMP for the LUAD dataset ($k = 10$). (**a**) Coverage plot of the optimal gene set. (**b**) Interaction of the genes in the optimal set annotated by knowledge of known pathways.

Eq. (20) in Methods. Mutations in *EGFR*, *KRAS*, and *TP53* are vital in lung cancer biology, and the molecular alterations associated with these mutation profiles have been widely investigated[27]. Note that the triplet (*EGFR*, *KRAS*, *STK11*) was obtained with Dendrix. This difference was obtained because *TP53* overlapped with the other two genes, and Dendrix ignored *TP53* to ensure mutual exclusivity in its programming model.

For $k = 10$, we found that the gene set (*ABL1*, *EGFR*, *KRAS*, *MKNK2*, *NF1*, *PAK6*, *PTEN*, *STK11*, *TERT*, *TP53*) was mutated in 145 patients (Fig. 4(a)). Through annotation using DAVID[23], these genes were found to be involved in the ErbB, MAPK, and PI3K-Akt signalling pathways, which are known to be critical in LUAD. Based on the knowledge of these pathways, we observed that most genes in this set involve interactions (Fig. 4(b)). The subset (*KRAS*, *EGFR*, *STK11*, *PTEN*, *TP53*) covering 133 patients is a subset of the PI3K-Akt signalling pathway, and PI3K-Akt pathway mutations involved in tumourigenesis have been reported for LUAD[28]. Various treatments aiming to inhibit lung cancer cell proliferation, migration and invasion through the PI3K-Akt pathway have been developed[29]. The subset (*KRAS*, *MKNK2*, *EGFR*, *NF1*, *TP53*), which constitutes a subset of the MAPK pathway, plays a pivotal role in cell proliferation, differentiation and survival[30]. MAPK signal amplification contributes to the rapid progression of established adenomas to LUAD and takes effect during both malignant progression and tumour initiation[31,32]. The subset (*KRAS*, *MKNK2*, *EGFR*, *NF1*) overlapped in five patients, whereas the subset (*KRAS*, *MKNK2*, *EGFR*, *NF1*, *TP53*) overlapped in 44 patients. This finding indicated that *TP53* showed little mutual exclusivity with the other four genes. Whereas the remaining genes *KRAS*, *MKNK2*, *EGFR*, and *NF1*

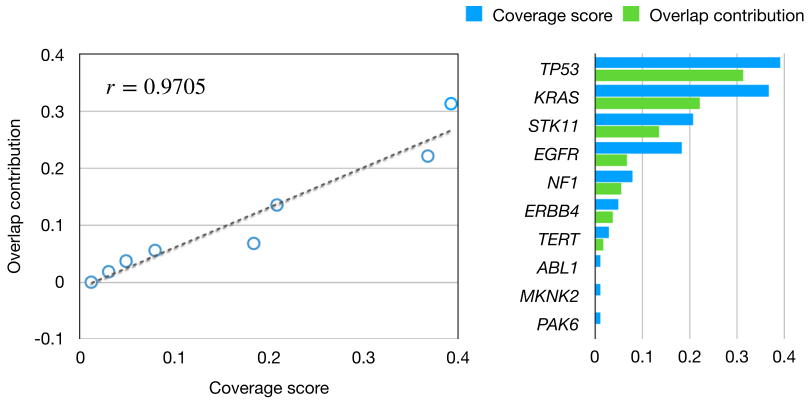


Figure 5. Scatter plot of coverage score against overlap contribution for the optimal gene set.

exhibited highly mutual exclusivity. The subset (*ABL1*, *EGFR*, *KRAS*, *PAK6*) which was mutated in 94 patients, forms part of the ErbB signalling pathway, which involves a family of tyrosine kinases and has been confirmed to be vital for the development of LUAD[33,34]. All the genes in this subset exhibit highly mutual exclusivity scores. Although *TERT* was not annotated in the aforementioned pathways, it has been found to be the most frequent genetic event in the early stages of non-small cell lung cancer[35].

To date, no explicit method has been developed to determine the dimension of the driver gene set identified by *de novo* discovery method. However, based on the subsampling strategy in AWRMP, we calculated the subsampling rate of each gene using Eq. (16) in Methods. Consequently, according to the subsampling rates, the subset (*EGFR, KRAS, TP53, STK11, NF1*) can be considered a parsimonious set that shows robustness against the uncertainty and noise in the data. The dimension of the parsimonious set can be considered a lower bound for the dimension of the driver gene set.

Performance of AWRMP. Figure 5 shows a scatter plot of the coverage score against the overlap contribution for the optimal gene set. As shown, the optimal gene set identified by AWRMP shows a similar pattern with the mutation pattern of the well-known cancer driver pathways illustrated in Fig. 2. This fact confirmed that our adaptive weights in Eq. (8) worked well, and this adaptiveness allowed us to identify useful overlaps. For example, the co-mutation (overlaps) of *TP53* and *NF1* has been known to be the feature of the PI subtype of LUAD[28].

To illustrate the robustness of AWRMP, we artificially disturbed some elements A_{ij}s of the mutation matrix A, by randomly turning the value 0 to 1 or randomly turning the value 1 to 0, to generate 100 new mutation matrices, and AWRMP was then performed for each disturbed mutation matrix. Consequently, the numbers of times that the candidate genes were selected by AWRMP with all 100 disturbed mutation matrices were used to evaluate the robustness of the proposed method (Fig. 6).

We conducted the disturbance test for $d = 10$ and 50, where d denotes the total number of disturbed elements in the mutation matrix A. The same optimal gene set (*ABL1*, *EGFR*, *KRAS*, *MKNK2*, *NF1*, *PAK6*, *STK11*, *TERT*, *PTEN*, *TP53*) was always identified for $d = 10$ in the both disturbance schemes (Fig. 6(a)). Increasing the value of d to 50, decreased the percentage of the 100 disturbed mutation matrices obtained by tuning values of 1 to 0 that yielded the optimal gene set to 27% and the percentage of the 100 disturbance schemes obtained by tuning values of 0 to 1 that yielded the optimal gene set to 30%. As expected, the robustness of AWRMP degenerated with increases in d. By observing the number of the times that ten genes of the optimal gene set were identified in the 100 runs of the disturbance test, we found that the subset (*EGFR*, *KRAS*, *STK11*, *TP53*, *NF1*) was always identified, even for $d = 50$ (Fig. 6(b)). This subset was thus the parsimonious set identified according to the subsampling rates. The total numbers of samples harboring these five genes were 30, 60, 34, 64, and 13, respectively. Thus the genes with relatively high coverage endured the disturbance. Furthermore, *TERT* showed the most sensitivity to the disturbance, even though it did not show the lowest observed mutation frequency. Moreover, *TERT* was not involved in any pathway detected by AWRMP (Fig. 4). This finding implies that *TERT* is slightly different from the other nine genes due to its weak HCME pattern. The results of the disturbance tests for other related methods are shown in Supplementary Fig. 2.

In addition to the robustness analysis, we also performed statistical significance tests using CoMEt[18] and TiMEx[17], and the results are depicted in Table 1. The optimal gene set identified by AWRMP can be considered to be significant for mutual exclusivity.

Parsimonious sets identified from breast cancer and glioblastoma datasets. In addition to the LUAD data, we also applied AWRMP to mutation datasets, including BC mutation data[24] and GBM mutation data[25].

For the BC mutation data, AWRMP identified the parsimonious set (*AKT1, BRCA2, GATA3, MAP3K1, PIK3CA TP53, RGS1(A)*, where "(A)" refers to amplification) with a high coverage score of 0.86 and a low overlap score of *0.45* (Supplementary Fig. 3). Among these genes, *BRCA2* truncating mutations have been associated with an increased risk of BC[36]. *GATA3* has been identified as a prognostic marker for BC[37]. The genes (*AKT1, MAP3K1, PIK3CA*) are associated with the abrogation of JUN kinase signalling, which occurs in approximately half of BC patients[38]. The biological consequences of a reduction in JUN kinase activity in response to stress might

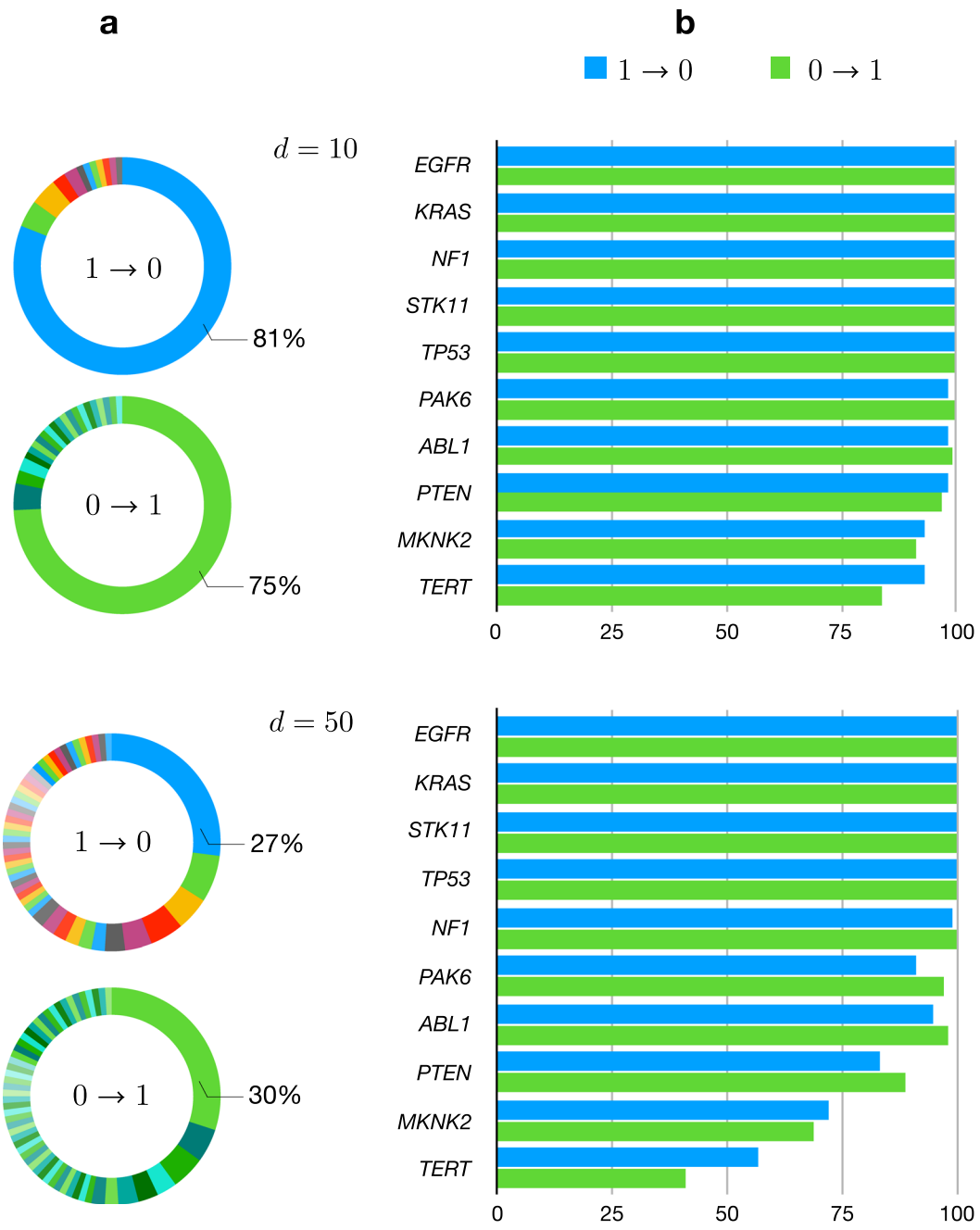


Figure 6. Number of times that the optimal gene set (*ABL1*, *EGFR*, *KRAS*, *MKNK2*, *NF1*, *PAK6*, *PTEN*, *STK11*, *TERT*, *TP53*) and its elements were identified by AWRMP with the 100 disturbed mutation matrices. (**a**) Percentages of the 100 disturbed mutation matrices obtained by tuning values of 1 into 0 (blue) and by tuning values of 0 into 1 (green). (**b**) Number of times that the 10 genes in the optimal gene set were identified with all 100 disturbed mutation matrices.

include destabilization and consequent inactivation of *TP53* and thereby disruption of pro-apoptotic cellular signalling[39]. Thus, the co-mutations in the parsimonious set obtained by the adaptiveness of AWRMP are reasonable. The relation between *RGS1* mutation and BC has been discovered in[40].

From the GBM mutation data, AWRMP identified the parsimonious set (*EGFR, NF1, PIK3CA, PIK3R1, PTEN, GABRA6, TP53*) with a coverage score of 0.70 and an overlap score of 0.30 (Supplementary Fig. 4). Among these genes, *NF1* is a human glioblastoma suppressor gene[41], and patients harbouring *NF1* mutation or deletion tended to show decreased PKC pathway activity and elevated MAP kinase activity[25]. *GABRA6*, an inhibitory neurotransmitter in the mammalian brain, contributes to coding for a transmembrane polymorphic antigen glycoprotein[25]. The subset (*EGFR, PIK3CA, PIK3R1, PTEN, TP53*) is part of the PI3K signalling pathway, and 62% of the glioblastoma samples harboured at least one genetic event associated with this subset. The PI3K-Akt

Genes	Pathway (q-value)	CoMEt	TiMEx
KRAS, EGFR, TP53 MKNK2, NF1	MAPK signalling pathway (2.00e-3)	0.02	6.45e-7
KRAS, EGFR, ABL1 PAK6	ErbB signalling pathway (2.10e-3)	4.27e-8	1.57e-7
KRAS, EGFR, STK11PTEN, TP53	PI3K-Akt signalling pathway (4.70e-3)	0.05	3.34e-6

Table 1. Pathway enrichment analysis and assessment of the statistical significance of the optimal gene set for LUAD identified by AWRMP from LUAD mutation data.

signalling pathway plays an important role in the regulation of signal transduction, which mediates various biological processes, including cell proliferation, apoptosis, metabolism, motility and angiogenesis in GBM[42].

Discussion

By observing the mutation patterns in cancer driver pathways from practical mutation datasets, we found the following: (a) the HCME pattern was approximately satisfied by the genes in the driver pathways and (b) overlaps were always observed, particularly among the genes with high coverage scores. For this reason, we proposed that the HCME pattern should be weakened by allowing proper overlaps in the discovery of driver gene sets. We developed AWRMP to identify the driver gene sets in cancer from mutation data. Ultimately, the goal of this approach is to investigate the gene sets that adaptively satisfy the weak HCME pattern. Moreover, by considering the sparsity of the mutation data, AWRMP can endure the potential uncertainty and noise in the data using the subsampling method. Here, we tested the performance of AWRMP using several biological datasets.

Driver mutations have often been investigated by observing the recurrence of individual genes[43,44]. However, mutational heterogeneity complicates the identification of functional mutations due to the recurrence of individual genes across many samples. As an alternative, an investigation of the putative driver gene set found across patients, has been proven to be another feasible approach. It is obvious that increases in the dimension of gene sets increases the monotonic coverage. For this reason, it becomes necessary to utilise constraints derived from biological knowledge. Notably, the mutual exclusivity of the pathways was used in combination with coverage to investigate driver gene sets. As noted by[6], the driver pathways exhibiting the HCME patterns are generally smaller and more focused than most pathways annotated in the databases.

Figure 2 shows two examples of mutation patterns in cancer driver pathways, and these show that the coverage scores of the gene members are positively correlated with the overlap contributions. The information provided in Supplementary Fig. 6 suggests that this positive correlation can be generally observed in all mutated gene sets, not just in cancer driver pathways. Thus, when investigating cancer driver gene sets, the genes covering many patients should be allowed to exhibit more overlaps with other genes. For this reason, we claimed that the weak HCME pattern is more feasible for describing the mutation patterns in cancer driver pathways. According to the weak HCME pattern, we proposed the use of adaptive weights in AWRMP. Because of the adaptive weights included in our programming model, our results were different from those obtained with Dendrix[4] (Supplementary Table 1), MDPfinder[5] (Supplementary Table 2), Mutex[13] (Supplementary Table 3), and CoMDP[7] (Supplementary Table 4), all of which assign identical weights to all gene candidates. The analysis of LUAD mutation data using our method included *TP53* with a high coverage score in the final result. Because CoMEt and TiMEx were proposed based on the rigorously mutual exclusivity, these four related methods showed better scores than AWRMP (Supplementary Tables 7–10). However, the optimal gene set obtained by AWRMP still passed the permutation test of mutual exclusivity performed using CoMEt and TiMEx. In other words, the gene set identified by our method satisfied the mutual exclusivity, although our method permits more overlaps than other related methods. Furthermore, the overlaps identified by our AWRMP can be useful, like the overlaps between *TP53* and *NF1* identified for the LUAD data set. We do not claim that our method is better than other related approaches for the identification of *TP53*. After all, frequently mutated genes individuals can be identified using MutsigCV[43]. Our proposal is that the results obtained by AWRMP are more concordant to the objective mutation pattern, i.e., weak HCME, as demonstrated in Figs 2 and 5. Supplementary Fig. 5 shows the correlation between the coverage score and the overlap contribution of the optimal gene sets obtained by the other four methods, and these findings showed that these four methods did not satisfy the weak HCME pattern as well as our method. Note that ComMDP can also identify genes with high mutation frequencies, such as *TP53* and *PIK3CA*[8]. However, ComMDP was proposed for the identification of the common driver gene set across several types of cancer by combining their mutation matrices. Based on the mathematical programming model[8], ComMDP is identical to MDPfinder for a single type of cancer.

In AWRMP, the optimization solver GA was embedded into the subsampling strategy to ensure the robustness of the algorithm. Prior to this study, the robustness of *de novo* discovery methods has seldom been considered. Nevertheless, the mutation matrices used as the inputs in these methods were always derived from high-throughput sequencing data, which are well known to be noisy. Furthermore, the total number of samples is notably much smaller than the number of genes. The use of sparse data always leads to statistical inference that is not robust to noise and uncertainty. The disturbance tests of Dendrix, MDPfinder, CoMDP, and, Mutex (Supplementary Tables 5 and 6, and Supplementary Fig. 2) revealed that a single run of the MCMC method and integer linear programming method were not robust to the disturbance. Because the subsampling strategy is always applied to estimate the precision of sample statistics, we adopted the subsampling method to compute the probabilities of gene sets obtained by the optimization solver. Consequently, the gene sets with high probabilities can be considered robust results. Because the adaptive weight defined by Eq. (8) is a nonlinear function of $I_M(j)$

defined by the Eq. (3), the programming model (6) is no longer a linear programming model. Motivated by[5], the heuristic GA was used in AWRMP. As a type of combinatorial optimization model, the mathematical programming model defined by formula (8) often consists of multiple solutions. AWRMP can offer the robustness level for each solution based on the subsampling strategy.

Through AWRMP, we propose that the gene candidates should be assigned different levels of importance based on the weak HCME pattern. In addition to the weights derived from the coverage scores obtained by AWRMP, the covariates associated with mutations, such as the expression level of genes and the DNA replication time of genes used in MutsigCV[43], can also be considered weights. The application of subsampling can assuredly increase the computational cost. However, we insist that the robust results obtained from sparse data need to be cautiously investigated.

Methods

Cancer genetic data and mutation matrix. We directly used the mutation matrix derived from LUAD mutation data by Dendrix[4], which included 163 patients with at least one mutated gene and 356 genes mutated in at least one patient.

The BC and GBM mutation datasets (maf files) were downloaded from The Cancer Genome Atlas Data Portal (http://tcga-data.nci.nih.gov), and these datasets consider point mutations and copy number alterations (CNAs). Somatic point mutations were identified with MutsigCV[43]. The corresponding entry in the mutation matrix was assigned a value of 1 to indicate significant point mutation. Using the approach described in[16], if a CNA event is concordant with the expression data, the corresponding entry in the mutation matrix is 1. After pre-processing, 487 samples and 274 genes were included in the BC mutation matrix and 282 samples and 308 genes were included in the GBM mutation matrix.

Previous methods. For the mutation matrix A defined by Eq. (1), which has m rows (samples) and n columns (gene candidates), Dendrix initially proposed the following programming model for the identification of an $m \times k$ optimal submatrix M that satisfies the HCME pattern

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

where G_M denotes the set of genes corresponding to the mutation matrix M, $\Gamma(g) \equiv \{i: A_{ig} = 1\}$ denotes the set of patients who presented mutations in gene g. $g, g' \in G_M$ are mutually exclusive, if $\Gamma(g) \cap \Gamma(g') = \varnothing$. The sum of the cardinalities $\sum_{g \in G_M} |\Gamma(g)|$ denotes the total number of mutation events in M. $\Gamma(G_M) \equiv \cup_{g \in G_M} \Gamma(g)$ is the set of patients with mutations in the genes in M, and its cardinality $|\Gamma(G_M)|$ can be further used to measure the coverage of the submatrix M. Thus, the coverage overlap $\omega(G_M) \equiv \sum_{g \in G_M} |\Gamma(g)| - |\Gamma(G_M)|$ can be used to measure exclusivity.

By noticing that the formula (2) is not easy for developing the optimization strategy, Zhang *et al.*[5] initially defined two indicator functions

$$I_M(j) \equiv \begin{cases} 1 & j \in G_M \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

for $j = 1, 2, \ldots, n$ and

$$I_i(G_M) \equiv \begin{cases} 1 & \text{genes in } G_M \text{ are mutated in patient } i \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

for $i = 1, 2, \ldots, m$, and reformulated the maximization of $W(G_M)$ as MDPfinder, which is a binary linear programming (BLP) problem:

$$\max_{\{I_M(j) | j=1,\cdots,n\}} \quad W(G_M) = 2\sum_{i=1}^{m} I_i(G_M) - \sum_{j=1}^{n} \left(I_M(j) \cdot \sum_{i=1}^{m} A_{ij} \right)$$
$$s.t. \begin{cases} I_i(G_M) \le \left(\sum_{j=1}^{n} A_{ij} \cdot I_M(j) \right), & \text{for } i = 1, \cdots, m;\ j = 1, \cdots, n \\ \sum_{j=1}^{n} I_M(j) = k \end{cases} \tag{5}$$

Mathematical programming model of AWRMP. In the mathematical programming model (5), $W(G_M)$ is divided into two parts: $\sum_{i=1}^{m} I_i(G_M)$ measures the coverage using the sum with respect to patient i and the second term $\sum_{j=1}^{n} \left(I_M(j) \cdot \sum_{i=1}^{m} A_{ij} \right)$ is the total number of mutation events (i.e., entries with a value of "1") in the mutation matrix. The latter term indicates that MDPfinder assigns identical weights to all the genes. As we mentioned before, coverage is more important than exclusivity for the genes involved in multiple pathways. Consequently, we improve the mathematical programming model by assigning different weights to the genes contained in G_M, i.e.,

$$\begin{aligned} W_\lambda(G_M) &\equiv |\Gamma(G_M)| - \omega_\lambda(M) \\ &= \sum_{i=1}^{m} I_i(G_M) - \left(\sum_{j=1}^{n}\left(\lambda_j \cdot I_M(j) \cdot \sum_{i=1}^{m} A_{ij}\right) - \sum_{i=1}^{m} I_i(G_M)\right) \\ &= 2\sum_{i=1}^{m} I_i(G_M) - \left(\sum_{j=1}^{n}\left(\lambda_j \cdot I_M(j) \cdot \sum_{i=1}^{m} A_{ij}\right)\right) \end{aligned} \tag{6}$$

where

$$\lambda_j \equiv \begin{cases} \dfrac{\exp(-|\Gamma(j)|)}{\sum_{r\in G_M}\exp(-|\Gamma(r)|)} & j \in G_M \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

is the weight assigned to gene j for $j = 1, 2, \ldots, n$. For all $j \in G_M$, $\lambda_j \in (0, 1)$ and $\sum_{j\in G_M}\lambda_j = 1$. For gene j, $\lambda_j \in (0, 1)$ makes coverage slightly more important than mutual exclusivity, and introduces overlaps with other genes. $\sum_{j\in G_M}\lambda_j = 1$ allows the frequently mutated genes to have more overlaps than the rarely mutated genes in a gene set. In the case of $\lambda_j \to 1$ for $|\Gamma(j)|$, mutual exclusivity is tuned to be as important as coverage. Using this approach, the balance between coverage and exclusivity can be adaptively adjusted for various genes with respect to the cardinality $|\Gamma(j)|$. For this reason, λ_j is called as adaptive weight. Consequently, the AWRMP programming model can be summarized as the following

$$\begin{aligned} &\max_{\{I_M(j)|j=1,\cdots,n\}} \quad W_\lambda(G_M) = 2\sum_{i=1}^{m} I_i(G_M) - \sum_{j=1}^{n}\left(\lambda_j \cdot I_M(j) \cdot \sum_{i=1}^{m} A_{ij}\right) \\ &s.t. \begin{cases} I_i(G_M) \le \left(\sum_{j=1}^{n} A_{ij} \cdot I_M(j)\right) & \text{for } i = 1, \cdots, m;\ j = 1, \cdots, n \\ \sum_{j=1}^{n} I_M(j) = k \end{cases} \end{aligned} \tag{8}$$

Setting up of GA. According to Eq. (7), λ_j is a nonlinear function of $I_M(j)$, which indicates that the AWRMP optimization model (8) is a nonlinear programming (NLP) model. According to the MDPfinder solver[5], we used the metaheuristic GA method as the NLP solver. The settings of the GA are as follows:

GA search space. The genes in A were labeled as $1, 2, \ldots, n$. According to Eqs (3) and (8), a binary-valued vector $\boldsymbol{x} \equiv [x_1, x_2, \ldots, x_n]^T$ is used as an individual of a population, in which $x_i \in \{0, 1\}$ characterizes the i-th gene in submatrix M. Thus, the GA search space is as follows:

$$S = \left\{ x \middle| x_i \in \{0,\ 1\} \text{ for } i = 1,\ 2, \cdots,\ n,\ \sum_{i=1}^{n} x_i = |G_M| \right\}. \tag{9}$$

GA fitness function. In a GA, the fitness function is used to evaluate the quality of individual $s_j \in S$. In AWRMP, we ranked each individual solution s_j with respect to $W_\lambda(G_{M_j})$ obtained by the programming model (8), in which M_j is the submatrix corresponding to s_j. The ranked result, denoted by r_j, is used to evaluate the fitness of s_j.

GA operations. Selection, crossover, and mutation are three basic operators of GA. To distinguish from the above-mentioned mutation, we denoted the 'mutation' operator as 'GA_mutation'. For individual s_j and rank r_j of each individual s_j based on the fitness value, the selection probability was defined as

$$p_j = \frac{2r_j}{n(n+1)} \tag{10}$$

where n is the population size.

The detailed GA procedure is provided in the supplementary information.

Integrating GA with subsampling. Robustness means that the algorithm can give identical results for various datasets with high probability. Through the use of subsampling, AWRMP investigates probabilities of the gene sets selected by the GA. We used a leave-one-out subsampling strategy to obtain n subsamples A_{i-} for $i = 1, 2, \ldots, m$, in which A_{i-} was obtained by removing the ith row of A. For all subsamples $\{A_{i-}\}$ and a given k, m runs of the GA were conducted to select the optimal gene sets. $\{G_k^{SS} | k = 1, 2, \cdots,\ m^{SS}\}$ denotes the selected gene sets obtained by m runs of the GA. Note that the possible multiple solutions of the optimization model (8) can lead to $m^{SS} > m$. For G_k^{SS}, we defined

$$m_k^{SS} \equiv \sum_{i=1}^{m} I_i(G_k^{SS}) \tag{11}$$

with

$$I(G_k^{SS}) \equiv \begin{cases} 1 & G_k^{SS} \text{ is selected with } i\text{th subsample } A_{i-} \\ 0 & \text{otherwise} \end{cases}. \tag{12}$$

m_j^{SS} is the total number of times that G_k^{SS} was selected in all m runs of the GA. Consequently, the probability of G_k^{SS} being selected as the optimal gene set can be obtained by

$$SSR_{G_k^{SS}} \equiv Pr(G_k^{SS} \text{ is selected}) = \frac{m_k^{SS}}{m} \tag{13}$$

which is called the subsampling rate (SSR) in this study. Moreover, the subsampling rate of a gene can also be calculated by Eq. (13), which denotes the probability of a gene being included in the optimal gene set. To test the significant robustness of G_k^{SS}, the null hypothesis was set up as follows: the distribution of m_j^{SS} was assumed to be a binomial distribution Bin(p, m). By taking the uncertainty of data into consideration, p is further assumed to obey a Beta distribution Beta($p_0 m$, m) where $p_0 \in (0, 1)$ is a user-defined hyper-parameter. In this study, $p_0 = 0.1$. Note that the Beta distribution is a conjugate distribution of the binomial distribution and the Beta-binomial distribution is the corresponding posterior distribution. Consequently, the following statistics

$$Q_k \equiv 1 - \sum_{r=0}^{m_k^{SS}} H(r, m, p_0 m, m) \tag{14}$$

is calculated. H is the Beta-binomial probability mass function

$$H(m_1, M_1, m_2, M_2) = \binom{M_1}{m_1} \frac{B(m_1 + a, M_1 - m_1 + b)}{B(a, b)} \tag{15}$$

where $B(\cdot)$ is the Beta function, $a = m_2 + 1$, and $b = M_2 - m_2 + 1$. The G_k^{SS} that satisfies $Q_j \leq 0.05$ was considered to form the driver gene set. We further defined the subsampling rate for gene g as follows:

$$SSR_g \equiv \Pr(g \text{ is selected in the driver gene set}) = \frac{\sum_{i=1}^{m} I_i(g)}{m} \tag{16}$$

with

$$I_i(g) \equiv \begin{cases} 1 & g \text{ is selected in the ith subsampling run} \\ 0 & \text{otherwise} \end{cases}. \tag{17}$$

Based on SSR_g, we define a parsimonious set as follows:

$$\text{Parsimonious set} \equiv \left\{ g | SSR_g = 1 \right\}, \tag{18}$$

which indicates the most robust result obtained by AWRMP.

Evaluation of the gene set *G*. The coverage, mutual exclusivity, and optimal performance of the gene set G were evaluated by the coverage score, overlap score, and total score, respectively as follows:

$$\text{Coverage score} \equiv \frac{1}{m} |\Gamma(G)| \tag{19}$$

$$\text{Overlap score} \equiv \frac{1}{m} \omega(G) \tag{20}$$

$$\text{Totals core} \equiv \frac{1}{m} W_\lambda(G). \tag{21}$$

We further define the overlap contribution for gene $g \in G$ as follows:

$$\text{Overlap contribution of gene } g \equiv \frac{1}{m} \left(\omega(G) - \omega\left(G_{g-}\right) \right) \tag{22}$$

where G_{g-} is the gene set obtained by subtracting gene g from gene set G, and this analysis is used to measure how gene g affects the overlap score of G.

References

1. Hahn, W. C. & Weinberg, R. A. Modelling the molecular circuitry of cancer. *Nat. Rev. Cancer* **2**(5), 331–341 (2002).
2. Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat. Med.* **10**(8), 789–799 (2004).
3. Yeang, C. H., McCormick, F. & Levine, A. Combinatorial patterns of somatic gene mutations in cancer. *FASEB J.* **22**(8), 2605–2622 (2008).
4. Vandin, F., Upfal, E. & Raphael, B. J. De novo discovery of mutated driver pathways in cancer. *Genome Res.* **22**(2), 375–385 (2012).
5. Zhao, J. F., Zhang, S. H., Wu, L. Y. & Zhang, X. S. Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics* **28**(22), 2940–2947 (2012).
6. Leiserson, M. D., Blokh, D., Sharan, R. & Raphael, B. J. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput. Biol.* **9**(5), e1003054 (2013).
7. Zhang, J. H., Wu, L. Y., Zhang, X. S. & Zhang, S. H. Discovery of co-occurring driver pathways in cancer. *BMC Bioinformatics* **15**(1), 271 (2014).
8. Zhang, J. H. & Zhang, S. H. Discovery of cancer common and specifc driver gene sets. *Nucleic Acids Res.* **45**(10), e86 (2017).
9. Zhang, J. H., Zhang, S. H., Wang, Y. & Zhang, X. S. Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data. *BMC Syst. Biol.* **7**(2), S4 (2013).
10. Lu, S. *et al*. Identifying driver genomic alterations in cancers by searching minimumweight, mutually exclusive sets. *PLoS Comput. Biol.* **11**(8), e1004257 (2015).
11. Miller, C. A., Settle, S. H., Sulman, E. P., Aldape, K. D. & Milosavljevic, A. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med. Genomics* **4**(1), 34 (2011).
12. Kim, Y. A., Cho, D. Y., Dao, P. & Przytycka, T. M. MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. *Bioinformatics* **31**(12), i284–i292 (2015).
13. Babur, Ö. *et al*. Systematic identification of cancer driving signalling pathways based on mutual exclusivity of genomic alterations. *Genome Biol.* **16**(1), 45 (2015).
14. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22**(2), 398–406 (2012).
15. Hua, X. *et al*. MEGSA: A powerful and exible framework for analyzing mutual exclusivity of tumor mutations. *Am. J. Hum. Genet.* **98**(3), 442–455 (2016).
16. Szczurek, E. & Beerenwinkel, N. Modeling mutual exclusivity of cancer mutations. *PLoS Comput. Bio.* **10**(3), e1003503 (2014).
17. Constantinescu, S., Szczurek, E., Mohammadi, P., Rahnenfhrer, J. & Beerenwinkel, N. TiMEx: a waiting time model for mutually exclusive cancer alterations. *Bioinformatics* **32**(7), 968–975 (2015).
18. Leiserson, M. D., Wu, H. T., Vandin, F. & Raphael, B. J. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol.* **16**(1), 160 (2015).
19. Kim, Y. A., Madan, S. & Przytycka, T. M. WeSME: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics* **33**(6), 814–821 (2016).
20. Zhang, J. & Zhang, S. The discovery of mutated driver pathways in cancer: Models and algorithms. *IEEE ACM T. Comput. Bi.* **15**(3), 988–998 (2018).
21. Goldberg, D. E. Genetic algorithms in search optimization and machine learning *Addison-Wesley Pub. Co., New Jersey* (1989).
22. Politis, D. N. & Romano, J. P. Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Stat.* **22**(4), 2031–2050 (1994).
23. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protoc.* **4**(1), 44–57 (2009).
24. Koboldt, D. C. *et al*. Comprehensive molecular portraits of human breast tumours. *Nature* **490**(7418), 61–70 (2012).
25. Brennan, C. W. *et al*. The somatic genomic landscape of glioblastoma. *Cell* **155**(2), 462–477 (2013).
26. Ding, L. *et al*. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**(7216), 1069–1075 (2008).
27. Bjaanaes, M. M. *et al*. Genome-wide DNA methylation analyses in lung adenocarcinomas: Association with EGFR, KRAS and TP53 mutation status, gene expression and prognosis. *Mol. Oncol.* **10**(2), 330–343 (2016).
28. Collisson, E. A. *et al*. Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network. *Nature* **511**(7511), 543–550 (2014).
29. Xia, M. *et al*. Tramadol regulates proliferation, migration and invasion via PTEN/PI3K/AKT signalling in lung adenocarcinoma cells. *Eur. Rev. Med. Pharmacol. Sci.* **20**(12), 2573–2580 (2016).
30. Chang, L. F. & Karin, M. Mammalian MAP kinase signalling cascades. *Nature* **410**(6824), 37–40 (2001).
31. Cicchini, M. *et al*. Context-dependent effects of amplified MAPK signalling during lung adenocarcinoma initiation and progression. *Cell Rep.* **18**(8), 1958–1969 (2017).
32. Gao, X. *et al*. MAP4K4 is a novel MAPK/ERK pathway regulator required for lung adenocarcinoma maintenance. *Mol. Oncol.* **11**(6), 628–639 (2017).
33. Kato, Y. *et al*. 476. Highly enhanced ErbB signalling pathway was unveiled in lepidic predominant invasive lung adenocarcinoma. *Eur. J. Surg. Oncol.* **9**(42), S171 (2016).
34. Hoque, M. O. *et al*. Genetic and epigenetic analysis of erbB signalling pathway genes in lung cancer. *J. Thorac. Oncol.* **5**(12), 1887–1893 (2010).
35. Kang, J. U., Koo, S. H., Kwon, K. C., Park, J. W. & Kim, J. M. Gain at chromosomal region 5p15. 33, containing TERT, is the most frequent genetic event in early stages of non-small cell lung cancer. *Cancer Genet Cytogenet* **182**(1), 1–11 (2008).
36. Easton, D. F. *et al*. A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. *Am. J. Hum. Genet.* **81**(5), 873–883 (2007).
37. Mehra, R. *et al*. Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis. *Cancer Res.* **65**(24), 11259–11264 (2005).
38. Stephens, P. J. *et al*. The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**(7403), 400–404 (2012).
39. Wu, G. S. The functional interactions between the MAPK and p53 signalling pathways. *Cancer Biol. Ther.* **3**(2), 156–161 (2004).
40. Volik, S. *et al*. Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res.* **16**(3), 394–404 (2006).
41. Mclendon, R. E. *et al*. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**(7216), 1061–1068 (2008).
42. Zhao, H. F. *et al*. Recent advances in the use of PI3K inhibitors for glioblastoma multiforme: current preclinical and clinical development. *Mol. cancer* **16**(1), 100 (2017).
43. Lawrence, M. S. *et al*. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**(7454), 214–218 (2013).
44. Dees, N. D. *et al*. MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* **22**(8), 1589–1598 (2012).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61633006, 61502074, 81602309, 81422038, 81872247, 91540110, and 31471235 to Y.W.). We thank Pi Xu Liu and Hailing Cheng for useful discussion.

Author Contributions

X.X. and P.Q. processed the data, designed the algorithm and the programming codes, and written the manuscript. X.X. and P.Q. contributed equally to this work. Y.W. supported result interpretation and manuscript writing. J.W and H.G. supervised the project and contributed to writing the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-42500-7.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.