Applications of Machine Learning to Predict Cisplatin Resistance in Lung Cancer

Yanan Gao
School of Biomedical Engineering, Southern Medical University, Guangzhou, Guangdong, China. Department of Information, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong, China

Qiong Lyu
Department of Oncology, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong, China

Rui Zhou
School of Biomedical Engineering, Southern Medical University, Guangzhou, Guangdong, China. Department of Information, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong, China

Peng Luo
Department of Oncology, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong, China

Jian Zhang
Department of Oncology, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong, China

Qingwen Lyu (gzbeer@smu.edu.cn)
Department of Information, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong, China https://orcid.org/0000-0002-6490-5461

Research

Keywords: Lung cancer, Cisplatin, Machine learning, SVMs, Biomarkers

Posted Date: March 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-262425/v1

Applications of machine learning to predict cisplatin resistance in lung cancer

Yanan Gao[1,2†], Qiong Lyu[3†], Rui, Zhou[1,2], Peng Luo[3*], Jian Zhang[3*], Qingwen Lyu[2*]

[1] School of Biomedical Engineering, Southern Medical University, Guangzhou, Guangdong, China.

[2] Department of Information, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong, China.

[3] Department of Oncology, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong, China.

[†]**These authors have contributed equally to this work and share first authorship**

[*] **Correspondence:**

Qingwen Lyu, Department of Information, Zhujiang Hospital, Southern Medical University; Email: gzbeer@smu.edu.cn; Telephone numbers: 0086-13925035628; Fax number: +86 020-61643579.

Jian Zhang, Department of Oncology, Zhujiang Hospital, Southern Medical University; Email: blacktiger@139.com.

Peng Luo , Department of Oncology, Zhujiang Hospital, Southern Medical University; Email: luopeng@smu.edu.cn.

Qingwen Lyu will handle correspondence at all stages of refereeing and publication, also post-publication.

Running title: Predicting the cisplatin resistance in lung cancer

ABSTRACT

Background: Lung cancer, mainly including lung adenocarcinoma, lung squamous cell carcinoma and small cell lung cancer, is the cancer with the highest incidence and cancer-related mortality in the world. Platinum-based chemotherapy plays an important role in the treatment of various lung cancer subtypes, but not all patients can benefit from it, so it is worth identifying lung cancer patients who are resistant or insensitive.

Method: The drug response and sequencing data of 170 lung cancer cell lines were downloaded from the Genomics of Drug Sensitivity in Cancer (GDSC) database, and support vector machines (SVMs) and beam search were used to select an optimal gene panel that can predict the sensitivity of cell lines to cisplatin. Then, we used the available cell line data to explore the potential mechanisms.

Result: In this study, SVMs and beam search were used to screen a 9-gene panel related to lung cancer cell line resistance to cisplatin, with an area under the curve (AUC) of 0.873 ± 0.004. The natural logarithm of the half maximal inhibitory concentration (lnIC50) values of the panel-MT group were significantly higher than those of the panel-WT group, regardless of whether lung cancer subtype was considered. In addition, we found that the differentially expressed pathways between the two groups may explain the difference.

Conclusion: In this study, we found that a panel including nine genes (PLXNC1, KIAA0649, SPTBN4, SLC14A2, F13A1, COL5A1, SCN2A, PLEC, and ALMS1) can accurately predict sensitivity to cisplatin, which may provide individualized

43 treatment recommendations to improve the prognosis of patients with lung cancer.

44 **Key words: Lung cancer; Cisplatin; Machine learning; SVMs; Biomarkers**

1. INTRODUCTION

Lung cancer is the most commonly diagnosed cancer worldwide and the leading cause of cancer-related deaths(1, 2). According to histological classification, approximately 85% of lung cancer cases are non-small cell lung cancer (NSCLC). The current pathological classification mainly includes 3 histological subtypes: adenocarcinoma, squamous cell carcinoma and small cell carcinoma(3). Small cell lung cancer (SCLC), a unique type different from NSCLC, is a highly aggressive tumor, accounting for 13%-15% of lung cancer cases(4). For patients with limited-stage SCLC, in addition to surgical resection, platinum-based neoadjuvant or adjuvant therapy could improve patient outcomes. For patients with extensive-stage SCLC, systemic treatment with cisplatin combined with etoposide is the most widely used(5). For NSCLC patients, except for stage IIA patients who can be treated with platinum-based adjuvant therapy after complete tumor resection, platinum-based chemotherapy is still the main recommendation in the first-line treatment of stage IIB patients and above(6-8).

Human cancer cell lines originating from tumor tissues and retaining most of the characteristics of tumor tissues(9) are the simplest experimental model and are widely used in the development of antitumor drugs. Large-scale drug sensitivity screening data and genomics data from cancer cell lines have been used to explore the interactions between drugs and genes(10-13). Following the NCI-60 cell line screening(14), the Genomics of Drug Sensitivity in Cancer (GDSC) project (https://www.cancerrxgene.org/) has also made a great contribution to exploring the

relationship between drug sensitivity and genomic data to discover therapeutic biomarkers that can be used to identify patients mostly likely to benefit from anticancer drugs(15). This new release of the GDSC database contains drug response data for nearly 1,000 cell lines, representing common tumor types. There are 518 drugs, including both cytotoxic drugs and targeted drugs. In addition, nearly every cell line has corresponding genomics data, including whole-exome sequencing (WES), gene expression, copy number alteration, DNA methylation, gene fusion and microsatellite instability data. In summary, the large-scale amount of drug response data and genomics data contained in the GDSC database provides the opportunity to explore potential biological indicators of drug responsiveness.

Mutations in key genes such as oncogenes may drive tumorigenesis to influence the responsiveness of cell lines to drugs, which can be validated in a clinical cohort. For example, BCR-ABL rearrangement mutations are significantly related to the efficacy of ABL inhibitors, and ABL inhibitors have been approved for chronic myeloid leukemia (CML) patients with BCR-ABL fusion mutations(16). In addition, BRAF mutations are related to the efficacy of BRAF, MEK1, and MEK2 inhibitors. The inhibitor vemurafenib (B-Raf V600E) can prolong the survival of patients with BRAF mutation-positive melanoma in clinical trials(17). ERBB amplification(18), oncogene EGFR(19) and FLT3 mutations(20) are also sensitive to their target inhibitors. Recent studies have found that new gene mutations are associated with chemotherapy sensitivity and patient prognosis(21, 22).

Machine learning, driven by computing power and massive data, has made

outstanding achievements in the health and medical fields(23). Using preclinical research models with genomics and drug response data, machine learning or deep learning methods may identify genomic features(24) or transcriptomic features(23, 25) related to drug response to help clinicians choose suitable drugs for individual patients. Based on the expression data of tumor organoids and their drug response data, researchers have used machine learning to identify pathway features related to the half maximal inhibitory concentration (IC50) that were consistently verified in colon cancer and bladder cancer patient cohorts(26).

In this study, we conducted a sophisticated machine learning scheme, including beam search and classification(27), to select an optimal gene panel to predict resistance to cisplatin and tried to explain the potential underlying mechanism by analyzing the sequencing and drug response data of 170 lung cancer cell lines from the GDSC database. Support vector machines (SVMs) were used to construct the model(28), and 10 times stratified 5-fold cross-validation was performed to ensure the robustness and reliability of our results. Our research has great potential to provide individualized treatment recommendations to improve the prognosis of patients with lung cancer.

2. MATERIALS AND METHODS

2.1 Drug response, gene expression and mutation data

The natural logarithm of the half maximal inhibitory concentration (lnIC50) values of all selected cell lines treated with cisplatin were downloaded from GDSC. Robust

Multichip Average (RMA) normalized the expression data from the Affymetrix Human Genome U219 array, and gene mutation information found in cell lines by Illumina HiSeq 2000 WES were also obtained from GDSC.

2.2 Identification of cisplatin-sensitive and cisplatin-resistant cell lines

The sensitivity of cancer cell lines to drugs is mainly expressed in terms of the IC50 value, which refers to the concentration of the drug needed to kill half of the tumor cells in vitro. Because the drug concentration was diluted by one-tenth or one-hundredth, we used the lnIC50 to distinguish resistant or sensitive cell lines. Based on the GDSC 8.1 database (updated on October, 2019), a total of 170 lung cancer cell lines have cisplatin drug sensitivity data, WES mutation data and RNA sequencing (RNA-Seq) data. To distinguish between resistant and sensitive cell lines, we analyzed the distribution of lnIC50 values and performed binary Gaussian fitting to fit the distribution. Finally, the k-means clustering algorithm was used to determine whether a cell line was sensitive or resistant based on the lnIC50 value(29).

2.3 Feature normalization

The Z-score normalization method was used to normalize the data to improve the accuracy and reliability of the model(30). Specifically, the normalization of each feature was as follows:

$$X^{*} = \frac{X - \mu(X)}{\sigma(X)},$$

where X denotes measurements of a specific feature, $\mu(X)$ denotes the mean value of X, $\sigma(X)$ denotes the standard deviation of X, and X^* denotes the normalized feature value. We used the same normalization procedure in both the training and test sets to make the experiments more precise.

2.4 Selection of mutated resistance gene panel

A total of 170 lung cancer cell lines in the GDSC database were identified as cisplatin-sensitive or cisplatin-resistant cell lines with the aforementioned method. A total of 1693 genes with mutation frequencies above 5% were selected as candidate genes. SVMs were used to construct the classifiers, and the average area under the curve (AUC), accuracy, sensitivity, and specificity were used to evaluate the performance of the classifiers. We aimed to select the optimal gene panel by combining SVMs and beam search(27, 28). The entire procedure of our workflow is shown in Fig. 1b; within each loop, the mutation data of each mutation panel were used as the input data of the model to predict the resistance to cisplatin. Ten times stratified 5-fold cross-validation was used to select the stable gene panels. Specifically, within each iteration, a fold was used as the test set, and all remaining folds were used as the training set. This classification procedure was repeated 5 times, and the average result of 5 folds was regarded as the cross-validation result of the model. The average results of 10 repeats of cross-validation were calculated to evaluate each gene panel, and all gene panels were ranked by AUC. First, 1693 single genes were used as a 1-gene panel cohort with a size of 1693. Then, we traversed this cohort to classify the cell lines and took the gene panel with the highest AUC as the priority gene panel and, similarly, the gene panels with the top 100 AUCs as the priority cohort. Finally, we

exhaustively added a gene to the panels in the priority cohort to form a new gene panel cohort. This process was repeated until the accuracy of the priority gene panel of the next loop no longer increased significantly to obtain the optimal gene panel. The Scikit-learn (version: 0.23.1) software package was used to conduct the experiments in this study(31).

2.5 TMB and DDR

The nonsynonymous mutations of lung cancer cell lines were taken as the raw mutation count and divided by 38 Mb to quantify the tumor mutation burden (TMB)(32). The R package ComplexHeatmap(33) was used to visualize the top 20 mutated genes in the sample and the gene panels identified by SVM. DNA damage repair (DDR) pathway-related gene sets were downloaded from the Molecular Signatures Database (MSigDB) of the Broad Institute(34). These gene sets were used to evaluate the number of nonsynonymous mutations in the DDR pathway and compare the difference between the panel-MT(panel-MT) group and the panel-WT type(panel-WT) group.

2.6 Differential gene expression analysis and gene set enrichment analysis

The R package limma was used to perform differential analysis on the gene expression data downloaded from GDSC(35). The R package clusterProfiler was used to perform gene set enrichment analysis (GSEA)(36), among which $P<0.05$ in Gene Ontology (GO) terms, Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome was considered significant. The gene sets in GSEA were downloaded from MSigDB of the Broad Institute(34).

2.7 Statistical analysis

The differences in drug response data and TMB between the panel-MT and panel-WT groups in GDSC were examined using the Mann–Whitney U test, and the associations between the panel status and the top 20 recurrently mutated genes and genes in the panel were examined using Fisher's exact test. $P < 0.05$ was considered statistically significant, and all tests were two-sided. All statistical tests and visualizations were performed with R software (version 3.6.1) and R studio (Version 1.2.1335). In addition, the R package ggpubr was used to create boxplots(37).

3. RESULTS

3.1 Identification of cisplatin-sensitive and cisplatin-resistant cell lines

The workflow of our entire study is shown in Fig. 1a. The distribution of lnIC50 was analyzed to distinguish cisplatin-sensitive cell lines from cisplatin-resistant cell lines. As shown in Fig. 2a, based on this distribution, binary Gaussian fitting was performed, and the goodness-of-fit coefficient ($R^2 = 0.9958$) indicated that the curve fit very well. Therefore, cisplatin-sensitive and cisplatin-resistant phenotypes can be characterized by a binary Gaussian distribution. From the binary Gaussian distribution, we know that the cell lines that correspond to lnIC50 in the left (blue) curve indicate cisplatin-sensitive cell lines, and those in the right (pink) curve indicate cisplatin-resistant cell lines. Based on the characteristics of the binary Gaussian distribution, the k-means clustering algorithm was used to perform 100 thousand iterations to determine whether each cell line is resistant or sensitive. As shown in Fig. 2b, 104 cell lines were identified as cisplatin-sensitive cell lines (cluster 0), and the

196 remaining 66 cell lines were identified as cisplatin-resistant cell lines (cluster 1).

197 **3.2 Classification performance and optimal gene panel**

198 As shown in Fig. 3a, when using a beam search to select the optimal gene panel, the
199 AUC of the priority gene panel gradually increased as the number of genes increased.
200 When 9 genes were selected, the accuracy of the priority gene panel no longer
201 increased significantly. Therefore, 9 genes were selected as an optimal gene panel to
202 predict resistance to cisplatin. The corresponding genes were PLXNC1, KIAA0649,
203 SPTBN4, SLC14A2, F13A1, COL5A1, SCN2A, PLEC, and ALMS1. As shown in
204 Table 1, the model achieved an AUC of 0.873 and an overall accuracy of 84.71%
205 when using mutations of the optimal gene panel. In addition, it had an accuracy of
206 84.68% in correctly identifying resistant cell lines (i.e., sensitivity) and an accuracy of
207 84.61% in identifying cell lines sensitive to cisplatin (i.e., specificity).

208 **3.3 The gene panel can predict the responsiveness of lung cancer cell lines to**
209 **cisplatin**

210 To further demonstrate the accuracy of the trained model, we grouped the cell lines
211 according to the mutation characteristics of the gene panel to demonstrate whether the
212 features selected by SVMs can accurately classify cisplatin-sensitive cell lines and
213 cisplatin-resistant cell lines. It is expected that in the panel-MT group containing any
214 mutation in the gene panel, the lnIC50 value of the overall lung cancer cell line will
215 be higher ($p < 0.001$) (Fig. 3b). In addition, when considering The Cancer Genome

Atlas (TCGA) tumor label provided in the GDSC drug screening data, these lung cancer cell lines were grouped into lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), SCLC, mesothelioma (MESO) and cell lines that could not be classified. The lnIC50 value of the panel-MT group was significantly higher than that of the panel-WT group except for MESO (Fig. 3c).

In addition, we explored the correlation between the labels by k-means clustering and the optimal gene panel selected by SVMs and beam search. We found that more panel-WT cell lines were significantly enriched in cluster 0, while more panel-MT cell lines were significantly enriched in cluster 1 (Fig. 3d). This finding also verified that the gene panel containing 9 genes identified by machine learning can be used as a marker for the drug response of lung cancer cell lines treated with cisplatin.

3.4 The predictive ability of the gene panel for other drugs

To explore whether the optimal gene panel has a similar predictive ability for other drugs, we also compared other drug response data between the panel-WT group and the panel-MT group (Supplemental Table 1). Not surprisingly, we also found the same predictive ability for some other chemotherapeutics (Fig. 4). Although the targets of these drugs are different, we found that there were some drugs that are similar to cisplatin through our manual review. Chemical drugs targeting DNA synthesis include cytarabine, bleomycin, etoposide, camptothecin, YK-4-279, 5-fluorouracil, and CX-5461. In addition, some drugs target the cell cycle, including docetaxel, AZD7762,

vinblastine, RO-3306, THZ-2-102-1, PHA-793887, epothilone B, IOX2, NPK76-II-72-1, temsirolimus, and Genentech Cpd 10.

3.5 Differences in gene mutation load between the panel-WT and panel-MT groups

To determine the potential mechanism by which the gene panel can predict the response of lung cancer cell lines to cisplatin, we combined the available sequencing data for subsequent analysis. By calculating the TMB of each group, we found that the panel-MT group had a significantly higher TMB than the panel-WT group (Mann-Whitney U test, $p < 0.05$) (Fig. 5a). Because DNA is the target of cisplatin, we also explored the differences in the frequency of gene mutations in the DDR pathway. Overall, the median frequency of mutations in the DDR pathway in the panel-MT group was higher than that of the panel-WT group. In detail, the frequency of mutations in panel-MT homologous recombination (HR) pathways was also significantly higher (Fig. 5b).

Next, we explored the differences between the recurrently mutated genes and the genes in the panel. Fig. 5c shows the recurrently mutated genes and the 9 genes in the optimal gene panel grouped based on k-means clustering. We found that among the lung cancer cell lines included in the study, the most recurrently mutated genes were TP53, TTN, MUC16 and RYR2. For mutations in TP53, missense mutations were most common, which may be related to their inactivation status. In contrast to TP53, TTN and MUC16 mainly contained missense mutations and multiple mutations. In

addition to the most common mutant genes, we found genes with different mutation frequencies in the two groups. Among the top 20 mutant genes, the mutation frequency of XIRP2 in the cluster 1 group was higher at 41%, while the mutation frequency in cluster 0 was 26%. Additionally, the mutation frequency of ALMS1 in the panel was also significantly higher in the cluster 1 group (26% in cluster 1; 10% in cluster 0). Moreover, we found that some recurrently mutated genes with different mutation frequencies in the two groups (Supplemental Table 2), such as CDH10, ENSG00000121031, SCN1A, WDFY4, and NLRP5, had higher mutation frequencies in cluster 1 (27%, 23%, 23%, 23%, and 21%, respectively), while ABCA1, ZFAT, PCDHG_cluster, RP11-551L14.1, VCX, HEPHL1 and LHCGR had higher mutation frequencies in cluster 0 (16%, 13%, 12%, 12%, 11%, 11%, and 11%, respectively). Among the 9 genes in the panel, the mutation frequencies of COL5A1 and F13A1 in cluster 1 were significantly higher than those in cluster 0 (COL5A1 17% vs 1%; F13A1 14% vs 3%).

3.6 High enrichment of DNA repair-related pathways in the panel-MT group may promote cell resistance to cisplatin

Next, we performed gene differential expression analysis (DEA) and GSEA to identify molecules or pathways that may explain the differences in the responses of the two groups to cisplatin. The DEA results showed that a total of 7 genes were upregulated in the panel-MT group and 14 genes were upregulated in the panel-WT group when p value <0.05 and fold change (FC)>3/2 or FC <2/3 (Fig. 6a,

Supplemental Table 3).

Pathways representing several function-related genes can achieve specific biological functions. In contrast, dysfunction pathways are related to the occurrence and development of diseases. In this study, we found that the pathways enriched in the panel-MT group and the panel-WT group were different (Fig. 6b, Supplemental Table 4). Based on the pathways filtered by $p<0.05$, we found that pathways related to telomerase maintenance and cell cycle-related pathways were enriched in the panel-MT group. In addition, we found that the DNA synthesis involved in DNA repair and interstrand crosslink repair pathways were significantly enriched in the panel-MT group.

4. DISCUSSION

In this study, we found that a combination of mutations and machine learning can accurately predict resistance to cisplatin. Furthermore, we selected a 9-gene panel that may be highly associated with resistance to cisplatin and an efficient biomarker for resistance to cisplatin in lung cancer cells. In this paper, we innovatively applied beam search and machine learning for the prediction of resistance to cisplatin in lung cancer cell lines. First, we performed binary Gaussian fitting on the drug susceptibility data of lung cancer and used k-means clustering to identify the cisplatin-sensitive and cisplatin-resistant cell lines. Second, we applied a beam search to select the optimal gene panel. In addition to selecting 1693 genes of larger magnitude as candidate genes, we also traversed as many gene panels as possible to evaluate their prediction ability for resistance to cisplatin in lung cancer cell lines to select an optimal gene panel that

can accurately predict resistance to cisplatin. Moreover, 10 times stratified 5-fold cross-validation was employed to obtain stable and reliable observation results. Finally, classification was based on the mapping relationship between the features and labels, so the optimal gene panel we selected may implicitly indicate the correlation between these genes and resistance to cisplatin, which can help us to explain the potential mechanism.

The 9-gene panel included PLXNC1, KIAA0649, SPTBN4, SLC14A2, F13A1, COL5A1, SCN2A, PLEC, and ALMS1. Except F13A1, which has only been reported in benign tumors(38), the remaining genes have been reported to be associated with malignant tumors. Among them, highly expressed plexin c1 (PLXNC1)(39), KIAA0649(40), SCN2A(41) and SCN2A(42, 43) are related to malignant tumor progression, metastasis or chemotherapy resistance.

The lung cancer cell lines in GDSC were grouped into panel-MT and panel-WT groups, and the accuracy of the panel in classifying sensitive and resistant cell lines was verified in our study. We found that in all cell lines, the lnIC50 values of the panel-MT group were significantly higher than those of the panel-WT group ($p<0.001$). In addition, with the exception of MESO (malignant tumors derived from the pleura), cell lines from different lung cancer subtypes in the panel-MT group were less responsive to cisplatin, and the corresponding lnIC50 values were higher, all of which were statistically significant ($p<0.05$). The above results suggest that the mutation status of the gene panel selected by SVM can predict the response of lung cancer cell lines to cisplatin well.

Cisplatin-based treatment regimens play a very important role in each subtype of lung cancer(5-8). The main mechanism by which cisplatin suppresses tumors is interacting with DNA to form covalent adducts with purine DNA bases, causing DNA

damage and disrupting DNA replication and transcription(44). Theoretically, tumors with damaged DDR pathways are more sensitive to cisplatin because they cannot recover the DNA damage caused by cisplatin in a timely manner(45), which is found in many tumor types(46, 47). It is worth mentioning that in the DDR sub-pathway, the HR pathway repairs DNA double-strand breaks, and abnormalities in its function cause significant damage to tumor cells. However, other previous studies have found that colon cancer cells and endometrial cancer cells with defects in DNA damage repair are resistant to cisplatin and carboplatin, respectively(48). The possible mechanism may be that the normal function of the mismatch repair system (MMR) after DNA replication can induce cell apoptosis and increase the sensitivity of cisplatin to damaged DNA, while MMR-deficient cells can lead to decreased apoptosis and cell resistance(49). In our study, the overall number of DDR mutations in the panel-MT group was significantly higher than that in the panel-WT group ($p<0.05$).

Previous studies suggested that tumors with mutations in the DDR pathway show higher TMB because of a greater accumulation of unrepaired DNA damage in cells(50). Similarly, we also found that the TMB of the panel-MT group was higher, which is consistent with the higher mutations in the DDR pathway in the panel-MT group. Immunotherapy has made remarkable achievements in the treatment of solid tumors, including lung cancer(51), but only a small subset of the population benefits, and there is an urgent need to identify patients who are likely to benefit from immune checkpoint inhibitors (ICIs). The KEYNOTE-158 pan-cancer study recently promoted the Food and Drug Administration (FDA) approval of pembrolizumab for the treatment of patients with tumors with high TMB (> 10 mutations/Mb)(51) (https://www.fda.gov/drugs/ drug-approvals-and-databases/ fda-approves

-pembrolizumab -adults-and-children-tmb-h-solid-tumors). The above research indicates that we can also use the panel genes found in our research to predict TMB. In addition, our previous study demonstrated that cancer cells with high TMB are associated with higher IC50 values(21), which is consistent with our current conclusion, suggesting that high TMB may also be a mechanism of cisplatin resistance.

In addition, we explored genes with a high frequency of mutations in lung cancer cells. Among the top 20 frequently mutated genes, the mutation frequency of XIRP2 in cluster 1 was significantly higher (41% vs 26%). This gene has been reported in breast cancer(53) and gastric cancer(54) in clinical samples, but it has not been reported in other cancers. This suggests the need to increase the number of clinical samples to discover new mutant genes and provide opportunities for subsequent mechanistic research. Among the genes identified by SVMs, ALMS1 (26% vs 10%), COL5A1 (17% vs 1%) and F13A1 (14% vs 3%) had mutation frequencies in cluster 1 that were significantly higher than those in cluster 0. Except for the mechanisms of COL5A1 in tumors, the mechanisms of the remaining two genes in malignant tumors have not been reported. Previous studies suggested that high expression of COL5A1 is associated with poor prognosis in breast cancer(42) and metastasis of lung adenocarcinoma(43). The above results suggest that our algorithm can identify new molecular markers related to chemotherapy, which should be validated further.

Studies on the mechanism of cisplatin, which is a kind of cell cycle-specific drug, have shown that cisplatin is mainly cross-linked with DNA during replication to affect

372 the function of DNA and cause cell death(55). In addition, enhanced DNA damage
373 repair capabilities can prevent the accumulation of lethal DNA damage induced by
374 platinum-based treatment, leading to chemotherapy resistance(56). In our research, we
375 found that telomere pathways, such as telomere maintenance, extension, and C-chain
376 lagging synthesis, and cell cycle pathways, such as meiotic cell cycle process, DNA
377 replication initiation and DNA replication, were enriched in the panel-MT group,
378 suggesting that cells in the panel-MT group may be more sensitive to cisplatin.
379 However, the IC50 values of the panel-MT group were significantly higher than those
380 of the panel-WT group, indicating that there may be other factors influencing the
381 response of cell lines to cisplatin. In our research, we also found that pathways related
382 to DNA repair during DNA synthesis were also enriched in the panel-MT group,
383 suggesting that cells in the panel-MT group have a stronger ability to repair DNA
384 damage, thereby reducing the formation of damaged DNA induced by cisplatin. The
385 latter factor may play a major role in explaining why cells in the panel-MT group are
386 more likely to be resistant to cisplatin.

387 The panel features identified by the SVMs algorithm have the same predictive
388 ability for the response of lung cancer cell lines to other chemotherapeutics. It is
389 worth mentioning that among the drugs we identified, there are many drugs that have
390 the same mechanism as cisplatin, interacting with DNA and preventing DNA
391 synthesis, including cytarabine, bleomycin, etoposide, camptothecin, YK-4-279,
392 5-fluorouracil, and CX-5461. In addition, we found some drugs that target the cell
393 cycle, including docetaxel, AZD7762, vinblastine, RO-3306, THZ-2-102-1,

PHA-793887, epothilone B, IOX2, NPK76-II-72-1, temsirolimus, and Genentech Cpd 10. The above cytotoxic drugs are cell cycle-specific. The results suggest that the panel features identified by SVMs can predict not only the sensitivity of lung cancer cell lines to cisplatin but also the drug response to the same or similar mechanism.

There were several potential limitations in our study. First, our sample size was limited, and there were only 170 lung cancer cell lines with cisplatin drug sensitivity data, mutation data and transcription data. However, full validation strategies were performed to ensure the reliability and robustness of the observations. Second, there are currently no suitable large-sample clinical data to directly support our conclusion, and further relevant clinical studies are needed to verify our conclusion.

5. CONCLUSION

In conclusion, we analyzed the drug response data and sequencing data of 170 lung cancer cell lines and established a 9-gene panel related to cisplatin sensitivity. Targeted sequencing containing these 9 genes helps predict the responsiveness of lung cancer patients to cisplatin and may provide personalized guidance for patient management.

6. List of abbreviations

Abbreviation	Abbreviate from
GDSC	Genomics of Drug Sensitivity in Cancer
SVMs	support vector machines
AUC	area under the curve

NSCLC	non-small cell lung cancer
SCLC	small cell lung cancer
WES	whole-exome sequencing
CML	chronic myeloid leukemia
RMA	Robust Multichip Average
RNA-Seq	RNA sequencing
TMB	tumor mutation burden
DDR	DNA damage repair
MSigDB	Molecular Signatures Database
GSEA	gene set enrichment analysis
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
TCGA	The Cancer Genome Atlas
LUAD	lung adenocarcinoma
LUSC	lung squamous cell carcinoma
MESO	mesothelioma
DEA	differential expression analysis
MESO	malignant tumors derived from the pleura
MMR	mismatch repair system
FDA	Food and Drug Administration

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The datasets generated and/or analysed during the current study are available in the [the Genomics of Drug Sensitivity in Cancer] repository, [https://www.cancerrxgene.org/].

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable

Authors' contributions

P.L., J.Z and Q.W.L. contributed to conception of our research project. P.L. organized the study. Y.N.G. executed the research project. Q.L. and R.Z. performed the statistical analysis. Q.L. performed Differential gene expression analysis and gene set enrichment analysis. Y.N.G. and Q.L. wrote the first draft of manuscript. P.L., J.Z and Q.W.L. reviewed and critiqued the manuscript. Y.N.G. and Q.L. wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Acknowledgments

Not applicable.

References:

1.Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. *NATURE*.(2018)553: 446-454. doi:10.1038/nature25183

2.Torre LA, Siegel RL, Ward EM, Jemal A. Global Cancer Incidence and Mortality Rates and Trends--An Update. *Cancer Epidemiol Biomarkers Prev*.(2016)25: 16-27. doi:10.1158/1055-9965.EPI-15-0578

3.Scagliotti G, von Pawel J, Novello S, Ramlau R, Favaretto A, Barlesi F, et al. Phase III Multinational, Randomized, Double-Blind, Placebo-Controlled Study of Tivantinib (ARQ 197) Plus Erlotinib Versus Erlotinib Alone in Previously Treated Patients With Locally Advanced or Metastatic Nonsquamous Non-Small-Cell Lung Cancer. *J CLIN ONCOL*.(2015)33: 2667-74. doi:10.1200/JCO.2014.60.7317

4.Sabari JK, Lok BH, Laird JH, Poirier JT, Rudin CM. Unravelling the biology of SCLC: implications for therapy. *NAT REV CLIN ONCOL*.(2017)14: 549-561. doi:10.1038/nrclinonc.2017.71

5.Farago AF, Keane FK. Current standards for clinical management of small cell lung cancer. *Transl Lung Cancer Res*.(2018)7: 69-79. doi:10.21037/tlcr.2018.01.16

6.Pignon JP, Tribodet H, Scagliotti GV, Douillard JY, Shepherd FA, Stephens RJ, et

al. Lung adjuvant cisplatin evaluation: a pooled analysis by the LACE Collaborative Group. *J CLIN ONCOL*.(2008)26: 3552-9. doi:10.1200/JCO.2007.13.9030

7.Strauss GM, Herndon JN, Maddaus MA, Johnstone DW, Johnson EA, Harpole DH, et al. Adjuvant paclitaxel plus carboplatin compared with observation in stage IB non-small-cell lung cancer: CALGB 9633 with the Cancer and Leukemia Group B, Radiation Therapy Oncology Group, and North Central Cancer Treatment Group Study Groups. *J CLIN ONCOL*.(2008)26: 5043-51. doi:10.1200/JCO.2008.16.4855

8.Arriagada R, Bergman B, Dunant A, Le Chevalier T, Pignon JP, Vansteenkiste J. Cisplatin-based adjuvant chemotherapy in patients with completely resected non-small-cell lung cancer. *N Engl J Med*.(2004)350: 351-60. doi:10.1056/NEJMoa031644

9.Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. *CELL*.(2016)166: 740-754. doi:10.1016/j.cell.2016.06.017

10.Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *NATURE*.(2012)483: 603-607. doi:10.1038/nature11003

11.Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *CELL*.(2013)154: 1151-1161. doi:10.1016/j.cell.2013.08.003

12.Dempster JM, Pacini C, Pantel S, Behan FM, Green T, Krill-Burger J, et al.

475 Agreement between two large pan-cancer CRISPR-Cas9 gene dependency data
476 sets. *NAT COMMUN*.(2019)10: 5817. doi:10.1038/s41467-019-13805-y
477 13.Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al.
478 Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset.
479 *CANCER DISCOV*.(2015)5: 1210-23. doi:10.1158/2159-8290.CD-15-0235
480 14.Chabner BA. NCI-60 Cell Line Screening: A Radical Departure in its Time. *J Natl*
481 *Cancer Inst*.(2016)108. doi:10.1093/jnci/djv388
482 15.Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al.
483 Systematic identification of genomic markers of drug sensitivity in cancer cells.
484 *NATURE*.(2012)483: 570-5. doi:10.1038/nature11005
485 16.Druker BJ, Guilhot F, O'Brien SG, Gathmann I, Kantarjian H, Gattermann N, et al.
486 Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *N*
487 *Engl J Med*.(2006)355: 2408-17. doi:10.1056/NEJMoa062867
488 17.Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, et al.
489 Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N*
490 *Engl J Med*.(2011)364: 2507-16. doi:10.1056/NEJMoa1103782
491 18.Konecny GE, Pegram MD, Venkatesan N, Finn R, Yang G, Rahmeh M, et al.
492 Activity of the dual kinase inhibitor lapatinib (GW572016) against
493 HER-2-overexpressing and trastuzumab-treated breast cancer cells. *CANCER*
494 *RES*.(2006)66: 1630-9. doi:10.1158/0008-5472.CAN-05-1182
495 19.Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW,
496 et al. Activating mutations in the epidermal growth factor receptor underlying

497 responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med*.(2004)350:
498 2129-39. doi:10.1056/NEJMoa040938
499 20.Daver N, Wei AH, Pollyea DA, Fathi AT, Vyas P, DiNardo CD. New directions
500 for emerging therapies in acute myeloid leukemia: the next chapter. *BLOOD*
501 *CANCER J*.(2020)10: 107. doi:10.1038/s41408-020-00376-1
502 21.Li M, Lin A, Luo P, Shen W, Xiao D, Gou L, et al. DNAH10 mutation correlates
503 with cisplatin sensitivity and tumor mutation burden in small-cell lung cancer.
504 *Aging (Albany NY)*.(2020)12: 1285-1303. doi:10.18632/aging.102683
505 22.Qiu Z, Lin A, Li K, Lin W, Wang Q, Wei T, et al. A novel mutation panel for
506 predicting etoposide resistance in small-cell lung cancer. *Drug Des Devel*
507 *Ther*.(2019)13: 2021-2041. doi:10.2147/DDDT.S205633
508 23.Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al
509 A guide to deep learning in healthcare. *NAT MED*.(2019)25: 24-29.
510 doi:10.1038/s41591-018-0316-z
511 24.Snow O, Lallous N, Ester M, Cherkasov A. Deep Learning Modeling of Androgen
512 Receptor Responses to Prostate Cancer Therapies. *INT J MOL SCI*.(2020)21.
513 doi:10.3390/ijms21165847
514 25.Steiner MC, Gibson KM, Crandall KA. Drug Resistance Prediction Using Deep
515 Learning Techniques on HIV-1 Sequence Data. *Viruses*.(2020)12.
516 doi:10.3390/v12050560
517 26.Kong J, Lee H, Kim D, Han SK, Ha D, Shin K, et al. Network-based machine
518 learning in colorectal and bladder organoid models predicts anti-cancer drug

519 efficacy in patients. *NAT COMMUN*.(2020)11: 5485.
520 doi:10.1038/s41467-020-19313-8
521 27.Sabuncuoglu I, Bayiz M. Job shop scheduling with beam search. *EUR J OPER*
522 *RES*.(1999)118: 390-412. doi:https://doi.org/10.1016/S0377-2217(98)00319-1
523 28.Hsu CW, Lin CJ. A comparison of methods for multiclass support vector
524 machines. *IEEE Trans Neural Netw*.(2002)13: 415-25. doi:10.1109/72.991427
525 29.Xu R, Wunsch DN. Survey of clustering algorithms. *IEEE Trans Neural*
526 *Netw*.(2005)16: 645-78. doi:10.1109/TNN.2005.845141
527 30.Manmatha R, Rath T, Feng F. Modeling Score Distributions for Combining the
528 Outputs of Search Engines., (2001) doi:10.1145/383952.384005
529 31.Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al,
530 Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine
531 Learning in Python. *J MACH LEARN RES*.(2011)12: 2825–2830.
532 32.Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, et al. Analysis of
533 100,000 human cancer genomes reveals the landscape of tumor mutational burden.
534 *GENOME MED*.(2017)9: 34. doi:10.1186/s13073-017-0424-2
535 33.Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in
536 multidimensional genomic data. *BIOINFORMATICS*.(2016)32: 2847-9.
537 doi:10.1093/bioinformatics/btw313
538 34.Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et
539 al. Gene set enrichment analysis: A knowledge-based approach for interpreting
540 genome-wide expression profiles. *Proceedings of the National Academy of*

Sciences.(2005)102: 15545-15550. doi:10.1073/pnas.0506580102

35.Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *NUCLEIC ACIDS RES*.(2015)43: e47. doi:10.1093/nar/gkv007

36.Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*.(2012)16: 284-7. doi:10.1089/omi.2011.0118

37.Kassambara A. (2018): ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.1.7. Available at: https://CRAN.R-project.org/package=ggpubr.

38.Supsrisunjai C, Hsu CK, Michael M, Duval C, Lee J, Yang HS, et al. Coagulation Factor XIII-A Subunit Missense Mutation in the Pathobiology of Autosomal Dominant Multiple Dermatofibromas. *J INVEST DERMATOL*.(2020)140: 624-635.e7. doi:10.1016/j.jid.2019.08.441

39.Balakrishnan A, Penachioni JY, Lamba S, Bleeker FE, Zanon C, Rodolfo M, et al. Molecular profiling of the "plexinome" in melanoma and pancreatic cancer. *HUM MUTAT*.(2009)30: 1167-74. doi:10.1002/humu.21017

40.Yang L, Zhao J, Lü W, Li Y, Du X, Ning T, et al. KIAA0649, a 1A6/DRIM-interacting protein with the oncogenic potential. *Biochem Biophys Res Commun*.(2005)334: 884-90. doi:10.1016/j.bbrc.2005.06.179

41.Sun J, Wang C, Zhang Y, Xu L, Fang W, Zhu Y, et al. Genomic signatures reveal DNA damage response deficiency in colorectal cancer brain metastases. *NAT COMMUN*.(2019)10: 3190. doi:10.1038/s41467-019-10987-3

563 42.Wu M, Sun Q, Mo CH, Pang JS, Hou JY, Pang LL, et al. Prospective molecular
564 mechanism of COL5A1 in breast cancer based on a microarray, RNA sequencing
565 and immunohistochemistry. *ONCOL REP*.(2019)42: 151-175.
566 doi:10.3892/or.2019.7147
567 43.Liu W, Wei H, Gao Z, Chen G, Liu Y, Gao X, et al. COL5A1 may contribute the
568 metastasis of lung adenocarcinoma. *GENE*.(2018)665: 57-66.
569 doi:10.1016/j.gene.2018.04.066
570 44.Kelland L. The resurgence of platinum-based cancer chemotherapy. *NAT REV*
571 *CANCER*.(2007)7: 573-84. doi:10.1038/nrc2167
572 45.Park S, Lee H, Lee B, Lee SH, Sun JM, Park WY, et al. DNA Damage Response
573 and Repair Pathway Alteration and Its Association With Tumor Mutation Burden
574 and Platinum-Based Chemotherapy in SCLC. *J THORAC ONCOL*.(2019)14:
575 1640-1650. doi:10.1016/j.jtho.2019.05.014
576 46.Plimack ER, Dunbrack RL, Brennan TA, Andrake MD, Zhou Y, Serebriiskii IG, et
577 al. Defects in DNA Repair Genes Predict Response to Neoadjuvant Cisplatin-based
578 Chemotherapy in Muscle-invasive Bladder Cancer. *EUR UROL*.(2015)68: 959-67.
579 doi:10.1016/j.eururo.2015.07.009
580 47.Teo MY, Bambury RM, Zabor EC, Jordan E, Al-Ahmadie H, Boyd ME, et al.
581 DNA Damage Response and Repair Gene Alterations Are Associated with
582 Improved Survival in Patients with Platinum-Treated Advanced Urothelial
583 Carcinoma. *CLIN CANCER RES*.(2017)23: 3610-3618.
584 doi:10.1158/1078-0432.CCR-16-2520

585 48.Fink D, Nebel S, Aebi S, Zheng H, Cenni B, Nehmé A, et al. The role of DNA
586 mismatch repair in platinum drug resistance. *CANCER RES*.(1996)56: 4881-6.
587 49.Stewart DJ. Mechanisms of resistance to cisplatin and carboplatin. *Crit Rev Oncol*
588 *Hematol*.(2007)63: 12-31. doi:10.1016/j.critrevonc.2007.02.001
589 50.Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, et al. Analysis of
590 100,000 human cancer genomes reveals the landscape of tumor mutational
591 burden. *GENOME MED*.(2017)9: 34. doi:10.1186/s13073-017-0424-2
592 51.Wu YL, Lu S, Cheng Y, Zhou C, Wang J, Mok T, et al. Nivolumab Versus
593 Docetaxel in a Predominantly Chinese Patient Population With Previously
594 Treated Advanced NSCLC: CheckMate 078 Randomized Phase III Clinical Trial. *J*
595 *THORAC ONCOL*.(2019)14: 867-875. doi:10.1016/j.jtho.2019.01.006
596 52.Marabelle A, Fakih M, Lopez J, Shah M, Shapira-Frommer R, Nakagawa K, et al.
597 Association of tumour mutational burden with outcomes in patients with advanced
598 solid tumours treated with pembrolizumab: prospective biomarker analysis of the
599 multicohort, open-label, phase 2 KEYNOTE-158 study. *LANCET*
600 *ONCOL*.(2020)21: 1353-1365. doi:10.1016/S1470-2045(20)30445-9
601 53.Paul MR, Pan TC, Pant DK, Shih NN, Chen Y, Harvey KL, et al. Genomic
602 landscape of metastatic breast cancer identifies preferentially dysregulated
603 pathways and targets. *J CLIN INVEST*.(2020)130: 4252-4265.
604 doi:10.1172/JCI129941
605 54.Li X, Wu WK, Xing R, Wong SH, Liu Y, Fang X, et al. Distinct Subtypes of
606 Gastric Cancer Defined by Molecular Characterization Include Novel Mutational

607 Signatures with Prognostic Capability. *CANCER RES*.(2016)76: 1724-32.
608 doi:10.1158/0008-5472.CAN-15-2443
609 55.Rottenberg S, Disler C, Perego P. The rediscovery of platinum-based cancer
610 therapy. *NAT REV CANCER*.(2020). doi:10.1038/s41568-020-00308-y
611 56.Perez RP. Cellular and molecular determinants of cisplatin resistance. *EUR J*
612 *CANCER*.(1998)34: 1535-42. doi:10.1016/s0959-8049(98)00227-5

Figure legends

613 Fig. 1.

614 **(a)** Flowchart of the selection of the mutated resistance gene panel. N indicates the
615 sample size. AUC, area under the curve; SVMs, support vector machines.
616 **(b)** Work flow of this paper. SVMs, support vector machines; WES, whole-exome
617 sequencing.

618 Fig. 2. **IC50 distribution of cisplatin in lung cancer cells.**
619 **(a)** Fit curve displaying the distribution of lnIC50 values in 170 lung cancer cell lines.
620 **(b)** Scatter plot of the IC50 distribution of cisplatin in 170 lung cancer cells. The first

red dotted line shows the maximum screening concentration of 10.0 μM, and the second red dotted line at the bottom shows the minimum screening concentration of 0.0391 μM. The red dots correspond to the predicted cisplatin-resistant cell lines by the k-means method, and the blue dots correspond to the predicted cisplatin-sensitive cell lines.

Fig. 3. The identified features can distinguish sensitive cell lines from drug-resistant cell lines.

(a) Comparison of SVMs algorithms containing 1-9 characteristic genes. When more gene features are included, the accuracy and sensitivity of the SVMs algorithm can be substituted, as with AUC.

(b) Regardless of the subtype of lung cancer, the lnIC50 values of cell lines containing any mutations in the 9 genes are significantly higher, and these cell lines are resistant to cisplatin.

(c) Considering the subtypes of lung cancer, with the exception of EMSO, cell lines containing mutations in the 9 genes have significantly higher lnIC50 values and are resistant to cisplatin.

(d) Correlation analysis of cluster labels between SVMs and the k-means method. Cluster 0 by k-means is significantly enriched in more panel-MT cell lines.

Fig. 4. **The identified features can be extended to other chemotherapy drugs in the GDSC database.**

(a) Cell lines containing any mutations in the 9 genes are resistant to some common drugs that target DNA synthesis, such as cytarabine, bleomycin, etoposide, camptothecin, YK-4-279, 5-fluorouracil, and CX-5461.

(b) Cell lines containing any mutations in the 9 genes are resistant to some common cell cycle-targeting drugs, such as docetaxel, AZD7762, vinblastine, RO-3306, THZ-2-102-1, PHA-793887, epothilone B, IOX2, NPK76-II-72-1, temsirolimus, and Genentech Cpd 10.

Fig. 5. **Differences in gene mutations between the panel-MT and panel-WT groups.**

(a) The TMB in the panel-MT group was significantly higher ($p < 0.001$).

(b) The number of mutations of the overall DDR and HR pathways in the panel-MT group was significantly higher than that in the panel-WT group ($p < 0.05$).

(c) The top 20 mutant genes and genes in the panel are grouped by the results of k-means clustering. Fisher's exact test was used to test the associations between the panel status and the mutated genes. TMB, tumor mutation burden. DDR, DNA damage repair. HR, homologous recombination.

Fig. 6. **Differences in molecular and pathway expression between the panel-MT and panel-WT groups.**

(a) Differentially expressed genes between the panel-MT and panel-WT groups. A total of 14 genes were upregulated in the panel-WT group when p value <0.05 and

FC > 3/2 or FC < 2/3.FC, fold change.

(b-d) The GSEA results show the significantly enriched pathways in the panel-MT group. Pathways related to telomerase maintenance (B) and the cell cycle (D) were enriched in the panel-MT group. The DNA synthesis involved in DNA repair and interstrand crosslink repair pathways were significantly enriched in the panel-MT group **(c).** GSEA, gene set enrichment analysis.

Figures

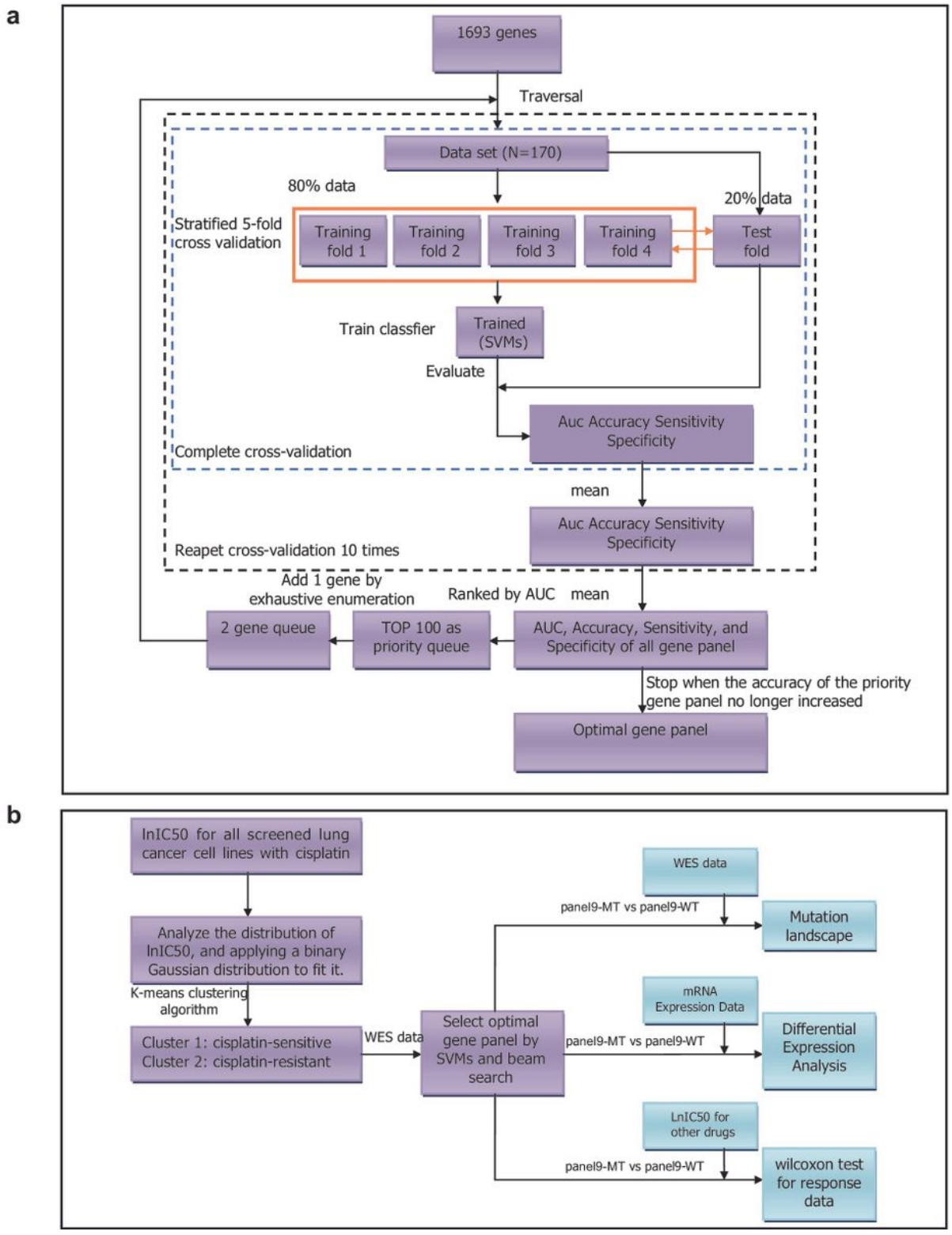

Figure 1

(a) Flowchart of the selection of the mutated resistance gene panel. N indicates the sample size. AUC, area under the curve; SVMs, support vector machines. (b) Work flow of this paper. SVMs, support vector machines; WES, whole-exome sequencing.

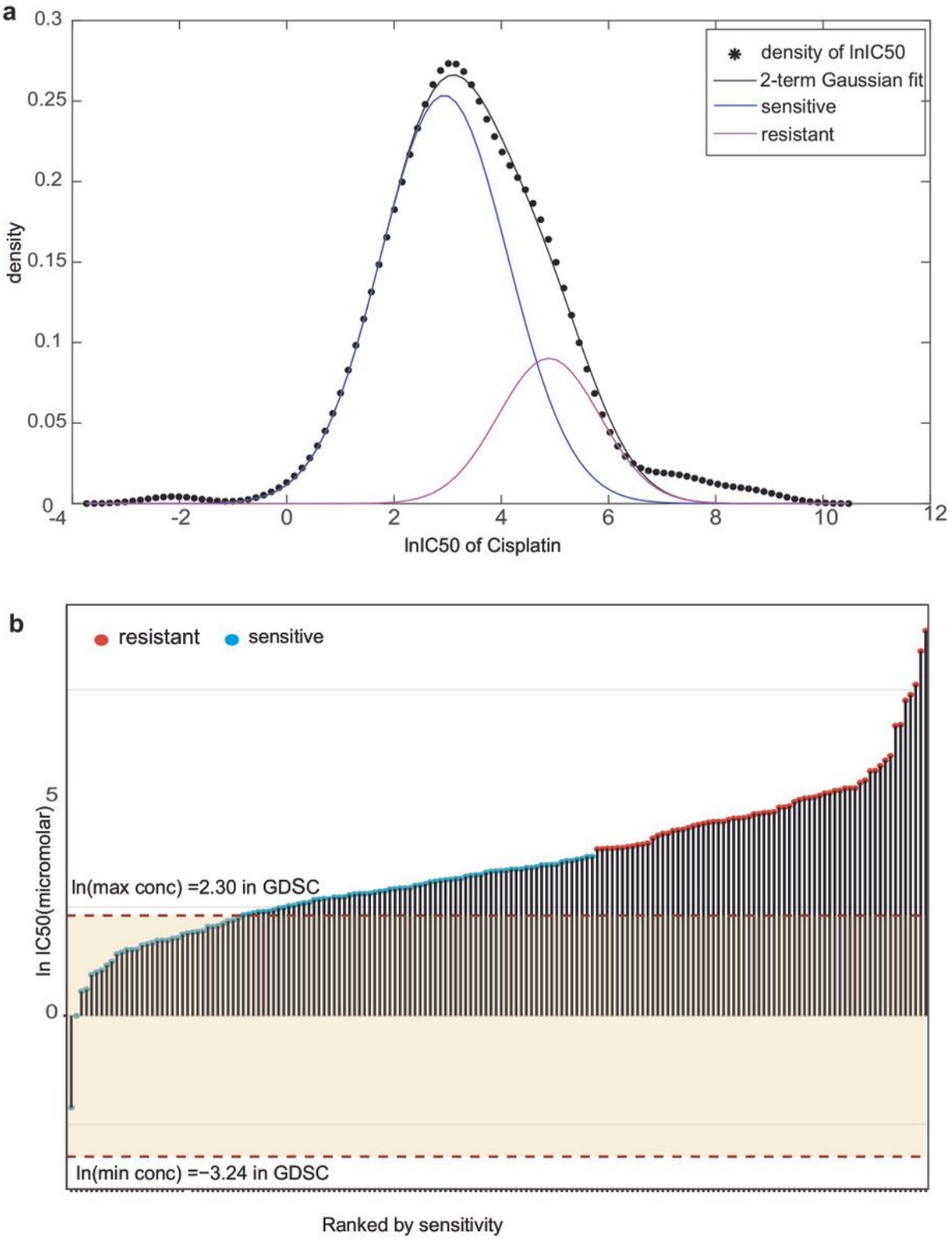

Figure 2

IC50 distribution of cisplatin in lung cancer cells. (a) Fit curve displaying the distribution of lnIC50 values in 170 lung cancer cell lines. (b) Scatter plot of the IC50 distribution of cisplatin in 170 lung cancer cells. The first red dotted line shows the maximum screening concentration of 10.0 μM, and the second red dotted line at the bottom shows the minimum screening concentration of 0.0391 μM. The red dots

correspond to the predicted cisplatin-resistant cell lines by the k-means method, and the blue dots correspond to the predicted cisplatin-sensitive cell lines.

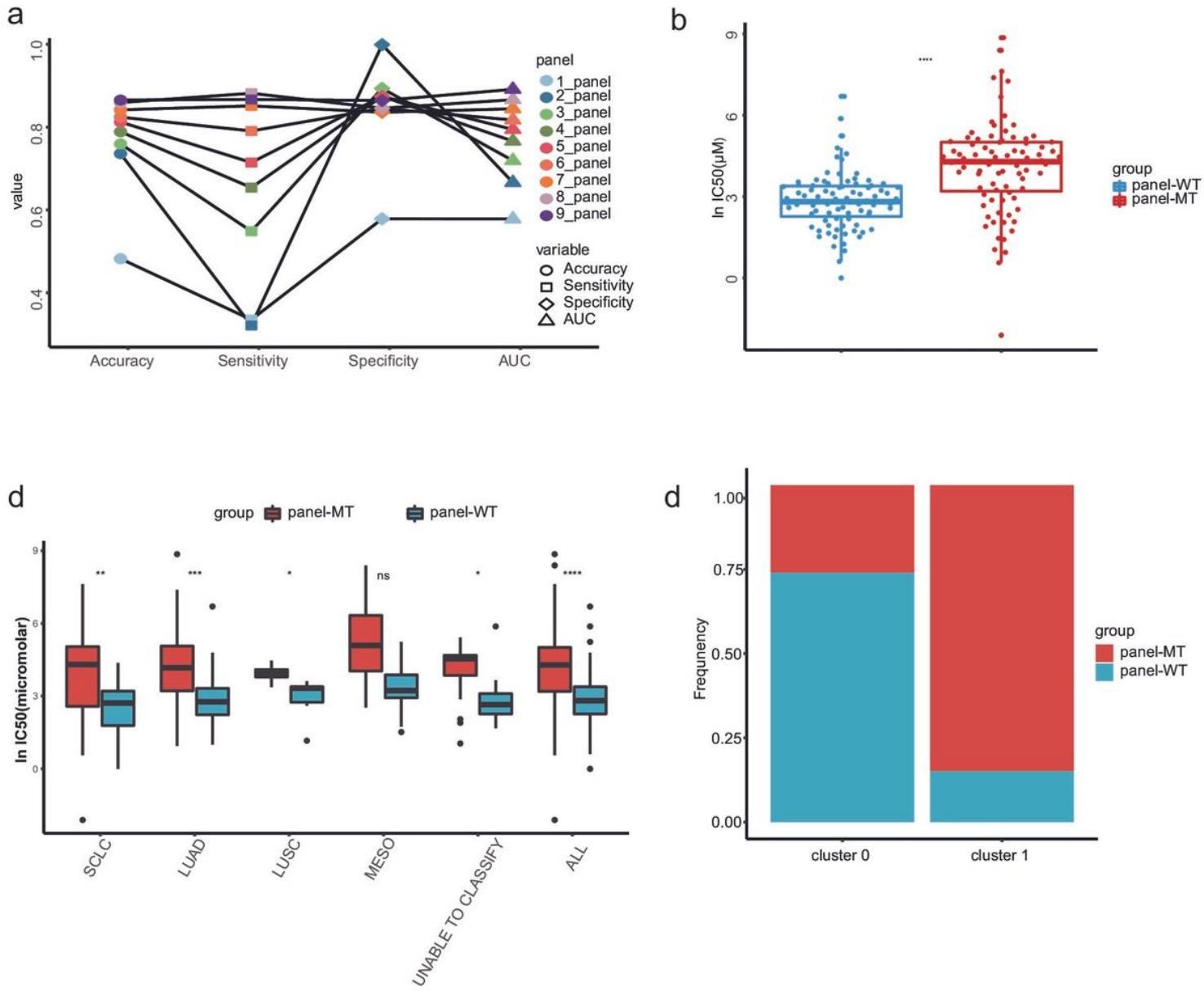

Figure 3

The identified features can distinguish sensitive cell lines from drug-resistant cell lines. (a) Comparison of SVMs algorithms containing 1-9 characteristic genes. When more gene features are included, the accuracy and sensitivity of the SVMs algorithm can be substituted, as with AUC. (b) Regardless of the subtype of lung cancer, the lnIC50 values of cell lines containing any mutations in the 9 genes are significantly higher, and these cell lines are resistant to cisplatin. (c) Considering the subtypes of lung cancer, with the exception of EMSO, cell lines containing mutations in the 9 genes have significantly higher lnIC50 values and are resistant to cisplatin. (d) Correlation analysis of cluster labels between SVMs and the k-means method. Cluster 0 by k-means is significantly enriched in more panel-MT cell lines.

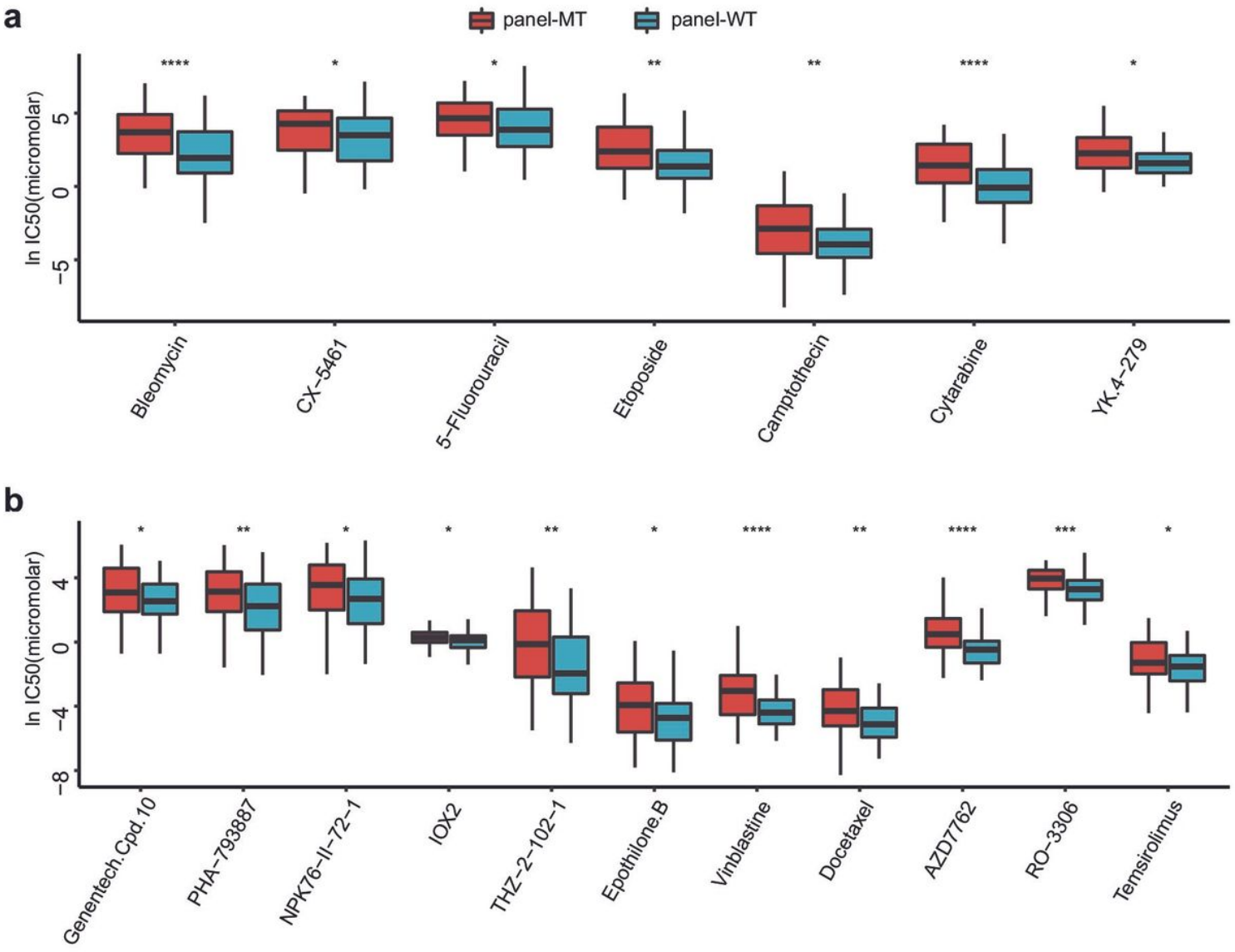

Figure 4

The identified features can be extended to other chemotherapy drugs in the GDSC database. (a) Cell lines containing any mutations in the 9 genes are resistant to some common drugs that target DNA synthesis, such as cytarabine, bleomycin, etoposide, camptothecin, YK-4-279, 5-fluorouracil, and CX-5461. (b) Cell lines containing any mutations in the 9 genes are resistant to some common cell cycle-targeting drugs, such as docetaxel, AZD7762, vinblastine, RO-3306, THZ-2-102-1, PHA-793887, epothilone B, IOX2, NPK76-II-72-1, temsirolimus, and Genentech Cpd 10.

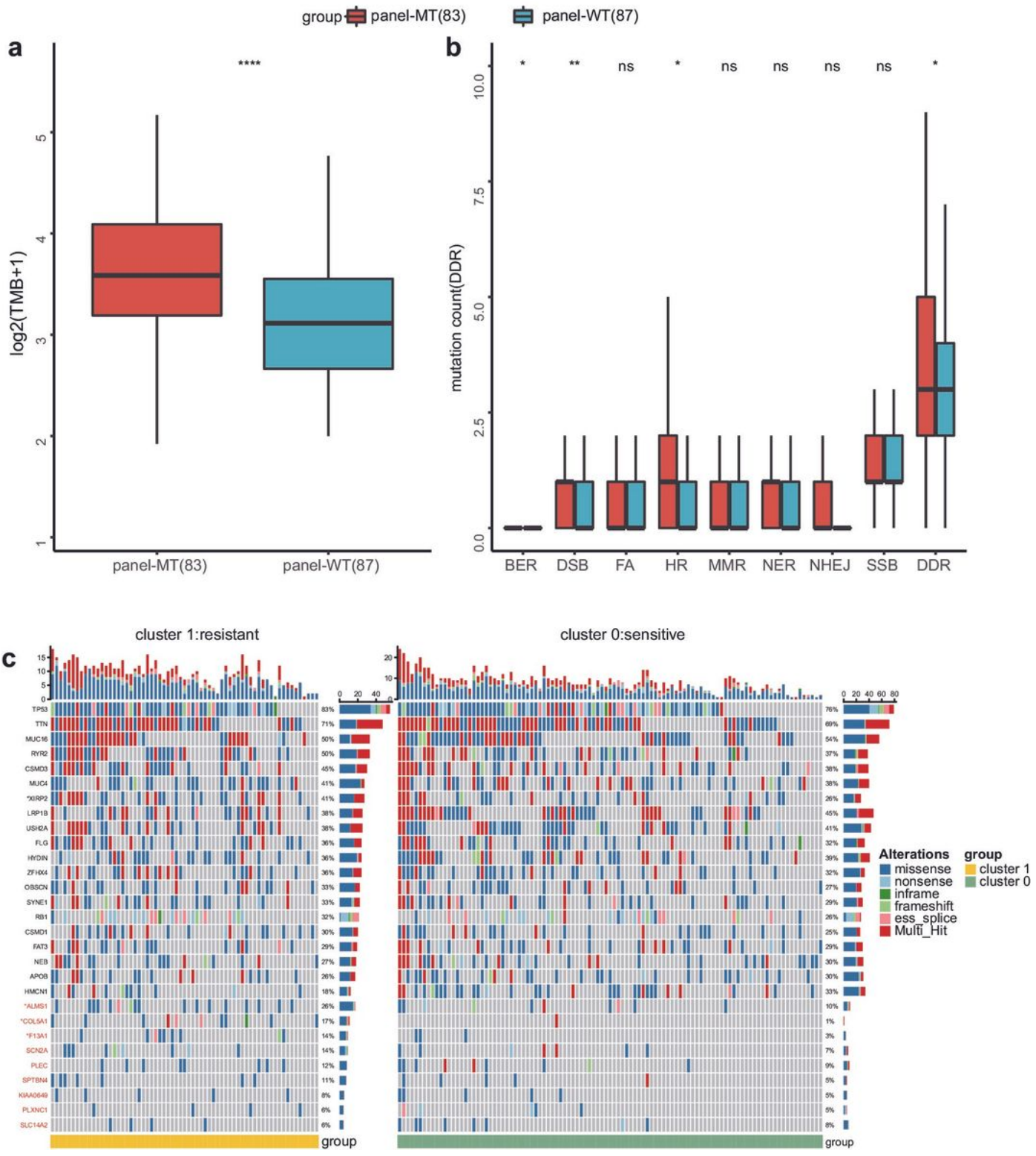

Figure 5

Differences in gene mutations between the panel-MT and panel-WT groups. (a) The TMB in the panel-MT group was significantly higher ($p < 0.001$). (b) The number of mutations of the overall DDR and HR pathways in the panel-MT group was significantly higher than that in the panel-WT group ($p < 0.05$). (c) The top 20 mutant genes and genes in the panel are grouped by the results of k-means clustering.

Fisher's exact test was used to test the associations between the panel status and the mutated genes. TMB, tumor mutation burden. DDR, DNA damage repair. HR, homologous recombination.

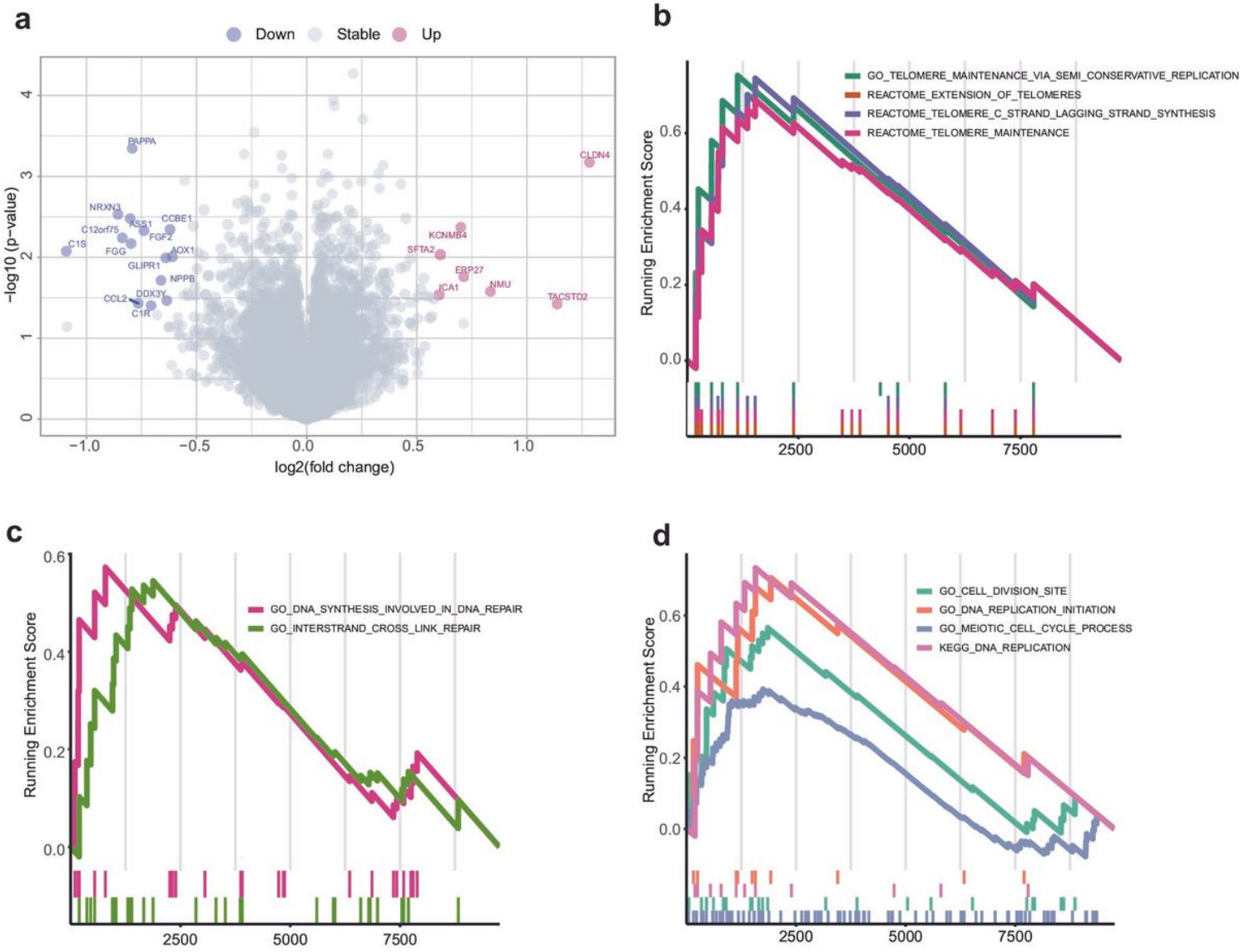

Figure 6

Differences in molecular and pathway expression between the panel-MT and panel-WT groups. (a) Differentially expressed genes between the panel-MT and panel-WT groups. A total of 14 genes were upregulated in the panel-WT group when p value <0.05 and FC > 3/2 or FC < 2/3.FC, fold change. (b-d) The GSEA results show the significantly enriched pathways in the panel-MT group. Pathways related to telomerase maintenance (B) and the cell cycle (D) were enriched in the panel-MT group. The DNA synthesis involved in DNA repair and interstrand crosslink repair pathways were significantly enriched in the panel-MT group (c). GSEA, gene set enrichment analysis.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalTable1.csv
- SupplementalTable2.csv
- SupplementalTable3.csv
- SupplementalTable4.csv