OXFORD

*Briefings in Bioinformatics*, 2023, 1–12

https://doi.org/10.1093/bib/bbad003

Problem Solving Protocol

# DROEG: a method for cancer drug response prediction based on omics and essential genes integration

Peike Wu, Renliang Sun, Aamir Fahira, Yongzhou Chen, Huiting Jiangzhou, Ke Wang, Qiangzhen Yang, Yang Dai, Dun Pan, Yongyong Shi and Zhuo Wang

Corresponding author: Zhuo Wang, E-mail: zhuowang@sjtu.edu.cn; Tel: 86-21-62932779; Fax: 86-21-62933338

## Abstract

Predicting therapeutic responses in cancer patients is a major challenge in the field of precision medicine due to high inter- and intra-tumor heterogeneity. Most drug response models need to be improved in terms of accuracy, and there is limited research to assess therapeutic responses of particular tumor types. Here, we developed a novel method DROEG (Drug Response based on Omics and Essential Genes) for prediction of drug response in tumor cell lines by integrating genomic, transcriptomic and methylomic data along with CRISPR essential genes, and revealed that the incorporation of tumor proliferation essential genes can improve drug sensitivity prediction. Concisely, DROEG integrates literature-based and statistics-based methods to select features and uses Support Vector Regression for model construction. We demonstrate that DROEG outperforms most state-of-the-art algorithms by both qualitative (prediction accuracy for drug-sensitive/resistant) and quantitative (Pearson correlation coefficient between the predicted and actual IC50) evaluation in Genomics of Drug Sensitivity in Cancer and Cancer Cell Line Encyclopedia datasets. In addition, DROEG is further applied to the pan-gastrointestinal tumor with high prevalence and mortality as a case study at both cell line and clinical levels to evaluate the model efficacy and discover potential prognostic biomarkers in Cisplatin and Epirubicin treatment. Interestingly, the CRISPR essential gene information is found to be the most important contributor to enhance the accuracy of the DROEG model. To our knowledge, this is the first study to integrate essential genes with multi-omics data to improve cancer drug response prediction and provide insights into personalized precision treatment.

**Keywords:** cancer drug sensitivity, CRISPR essential genes, multi-omics integration, machine learning model, pan-gastrointestinal tumor

## Introduction

Patients may differently respond to the same medical treatment when suffering from the same cancer, which highlights the importance of personalized and precision medicine. Precision medicine can avoid this deficiency by offering targeted therapies and exploiting cancer-specific vulnerabilities [1, 2]. Recent studies suggested that genome-directed precision oncology can match only 11% of patients [3] in clinical trials, and only around 5% of patients benefit from precision oncology [4], which emphasizes the accuracy of predicting drug response in actual patients. Furthermore, genetic profiles have a significant impact on the therapeutic response of different cancer patients [5, 6]; thus, identifying drug responses of different cancer cell lines/patients is promising in precision medicine [7].

Based on the drug response prediction evaluation methods, the models can be roughly classified as qualitative and quantitative evaluation-based models. Quantitative methods estimate the drug response values and calculate Pearson correlation coefficient (PCC) with measured drug response data to evaluate the overall predicted efficacy. The early method is that Geeleher *et al.* [8] used whole-genome ridge regression model with baseline gene expression and drug IC50 values to obtain *in vivo* drug sensitivity prediction. A recent study by Jia *et al.* developed a deep variational autoencoder model (VAEN) to compress genes into low-dimensional latent vectors and can impute drug response with better accuracy controlling the overfitting problem [9].

Qualitative approaches aim to differentiate between medication responses that are sensitive and resistant, and then use classifier evaluation parameters to assess the model performance. Similarity-based network methods have been widely used in drug response prediction. Wang *et al.* [10] incorporated drug chemical structure similarity and gene expression similarity information

**Peike Wu** is a Master's Student in Bio-X Institutes at Shanghai Jiao Tong University. His research interests are in bioinformatics and machine learning.
**Renliang Sun** is a Research Assistant in Shanghai Institute of Nutrition and Health. His research interests are in omics-data integration and modeling.
**Aamir Fahira** is a Postdoc Scholar in Bio-X Institutes at Shanghai Jiao Tong University. His research interests are in genetics and genomics.
**Yongzhou Chen** is a Master's Student in School of Mathematics at Shanghai Jiao Tong University. His research interests are statistical learning.
**Huiting Jiangzhou** is a Master's Student in Bio-X Institutes at Shanghai Jiao Tong University. Her research interests are in bioinformatics and phenomics.
**Ke Wang** is a PhD candidate in Bio-X Institutes at Shanghai Jiao Tong University. Her research interests are in genetics and genomics.
**Qiangzhen Yang** is a PhD candidate in Bio-X Institutes at Shanghai Jiao Tong University. His research interests are in structural biology and drug discovery.
**Yang Dai** is a Master's Student in Bio-X Institutes at Shanghai Jiao Tong University. His research interests are in biological network reconstruction and analysis.
**Dun Pan** is an Associate Professor in Bio-X Institutes at Shanghai Jiao Tong University. His research interests are in nanobiotechnology.
**Yongyong Shi** is a Professor in Bio-X Institutes at Shanghai Jiao Tong University. His research interests are in genetics, genomics and bioinformatics.
**Zhuo Wang** is a Professor in Bio-X Institutes at Shanghai Jiao Tong University. Her research interests are in bioinformatics, omics integration, biological network and applications in precision medicine.

**Received:** August 18, 2022. **Revised:** December 6, 2022. **Accepted:** December 30, 2022

and then decomposed them into latent vectors to construct similarity-regularized matrix factorization (SRMF) to predict drug sensitivity. Zhang *et al.* [11] proposed a heterogeneous network-based method (HNMDRP) to efficiently predict cell line-drug associations by making use of cell line gene expression profile, drug chemical structure feature, drug target interaction and protein-protein interactions (PPIs) information. MOLI [12], a multi-omics late integration method based on deep neural networks, took gene expression, CNV and mutation as input and generate three sub-networks. Subsequently, MOLI concatenated these features into a final network and served as input for a classification sub-network to obtain the final drug response. DrugGCN [13] constructed a gene graph containing a PPI network graph and took gene expression values as graph signals, then learned a graph convolutional network with localized filters and got the final drug response. Lenhof *et al.* [14] proposed a method MERIDA using literature information to annotate and binarize gene expression, mutation and copy number variation (CNV) data as well as drug response data and subsequently applied integer linear programming (ILP) formulation to get a set of sensitivity/resistance-associated alterations for drug response prediction, whereas MOFGCN (Multi-Omics Fusion and Graph Convolution Network) [15] employed graph convolution operation on the heterogeneous network to extract features and then reconstructed the association matrix by linear correlation coefficient decoder to predict drug response.

Numerous studies have shown that gene expression data are the most informative and adequate data for drug response prediction [16–18]. Specific oncogenes in clinical practice can help the selection of targeted therapies [19, 20] e.g. BRAF V600E mutation [21]. Hence, mutational data (including single nucleotide variants (SNV) and CNV) can help the improvement of the drug response prediction model. Epigenetic modification has also been found to be directly/indirectly correlated to cancer [22, 23], thus some drug response prediction models have used methylation data and showed better prediction power [24, 25]. However, single-omic data is insufficient to accurately predict the real drug response of certain cancer. Therefore, drug response prediction models by integrating multi-omics information will show more advantages in revealing actual drug responses in cancer cell lines or patients.

Genetic vulnerabilities of genes in the specific cancer cell can be used to guide the development of treatment strategies and novel therapeutics [26]. The CRISPR–Cas9 system enables precise genome-scale identification of essential genes and cancer cell survival status [27, 28]. Anglada-Girotto *et al.* proposed a combined computational/experimental strategy to perform high-throughput de novo functional annotations of antibacterials by combining CRISPRi and non-targeted metabolomics [29]. Kuenzi *et al.* developed DrugCell, which is a Visual Neural Network that simulates the response of human cancer cells to therapeutic chemical compounds. Analysis of DrugCell mechanisms leads directly to the design of synergistic drug combinations, which were then validated by combinatorial CRISPR, drug–drug screening *in vitro* and patient-derived xenografts [30]. Here, we anticipate that the gene essentiality information by CRISPR will also play an important role in drug response learning and prediction for cancer cells, and therefore we developed a novel drug response prediction model by leveraging the integration of CRISPR gene essential information with multi-omics data, called DROEG (Drug Response based on Omics and Essential Genes). This method can predict both real drug response value (quantitative) and also make sensitive/resistant classification (qualitative) and has been validated on GDSC and CCLE datasets. We aimed to prove the important prediction power of CRISPR essential gene information in enhancing the drug response model. In addition, we applied the DROGE method for drug response prediction of Pan-gastrointestinal (Pan-GI) cancer at both cell line level and clinical patient level, to verify its effectiveness for specific cancer types, which provide insights into the drug response mechanism and promote the development of personalized precision medicine for gastrointestinal cancer.

# Materials and methods

## Omics and drug sensitivity data of cancer cell lines

This study mainly used two databases, GDSC (https://www.cancerrxgene.org/) and CCLE (https://sites.broadinstitute.org/ccle/). In GDSC, the drug sensitivity data are represented as a half-maximal inhibitory concentration (IC50) value for all screened cell-line/drug combinations. In the CCLE database, drug sensitivity data are represented as activity area (ActArea), which equals to 1-AUC and negatively correlates with the IC50 value in GDSC. The gene expression, CNV, somatic mutation and methylation data were downloaded from the GDSC and CCLE databases. We only consider the cell lines with $<20\%$ missing drug sensitivity values and having all four types of omics data and CRISPR CERES information available. We retrieved multi-omics data of 214 drugs for 495 cell lines from GDSC 1 + 2 dataset, 173 drugs for 413 cell lines from GDSC2 dataset and 24 drugs for 264 cell lines from CCLE dataset. The drug IC50 or ActArea values for each cell line are shown in Supplementary Table S1.

## CRISPR essential gene analysis

CRISPR gene effect data were obtained from the Depmap website (https://depmap.org/portal/) cell line dependency screening. The CERES was widely used in Depmap to calculate the gene knockout effect, since CERES decreased false-positive results and estimated sgRNA activity for most screens. In addition, the CERES computationally sums the effects of gene knockout and copies number impact to represent the observed sgRNA depletion, which can be used to identify cancer-type-specific vulnerabilities [31]. CERES infers the gene-knockout effects and all other parameters by fitting the model to the observed data via alternating least squares regression. The inferred gene-knockout effects are then scaled per cell line such that scores of 0 and $-1$ represent the median effects of nonessential genes and common core essential genes, respectively. CERES scores were used to summarize gene-level dependency which smaller values indicate greater sensitivity to gene knockout [32]. Moreover, CERES has been suggested in the selection of reagents for follow-up experiments [26]. Thus, we consider CRISPR essential genes may reflect the actual functions of genes and play vital roles in drug response prediction. We downloaded common essential genes list (21Q1_CRISPR_common_essentials.csv) from Depmap for cell lines analysis, and the CERES score of each cell line in GDSC and CCLE datasets is shown in Supplementary Tables S2 and S3.

## Principle and pipeline of the newly proposed DROEG model

For each drug, a machine learning model will be constructed by combining the multi-omics data with CRISPR gene effect information in corresponding cell lines. We randomly partitioned the cell lines into training set (80%) and test set (20%). Figure 1 describes the strategy and pipeline of the DROEG model, which mainly includes five steps.

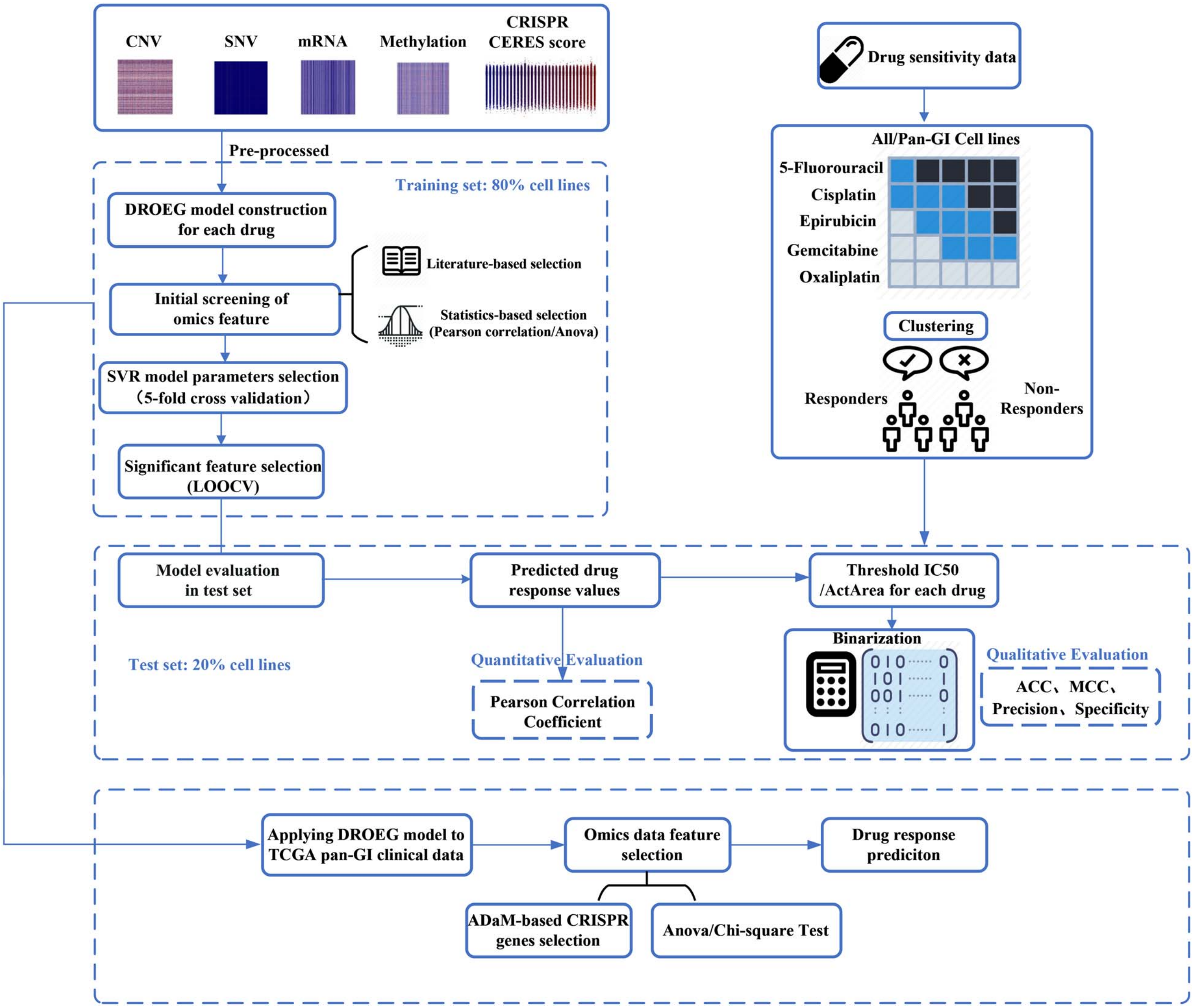

CNV
SNV
mRNA
Methylation
CRISPR CERES score
Pre-processed
Training set: 80% cell lines
DROEG model construction for each drug
Literature-based selection
Initial screening of omics feature
Statistics-based selection (Pearson correlation/Anova)
SVR model parameters selection (5-fold cross validation)
Significant feature selection (LOOCV)
Drug sensitivity data
All/Pan-GI Cell lines
5-Fluorouracil
Cisplatin
Epirubicin
Gemcitabine
Oxaliplatin
Clustering
Responders
Non-Responders
Model evaluation in test set
Predicted drug response values
Threshold IC50 /ActArea for each drug
Binarization
Test set: 20% cell lines
Quantitative Evaluation
Pearson Correlation Coefficient
Qualitative Evaluation
ACC、MCC、Precision、Specificity
Applying DROEG model to TCGA pan-GI clinical data
Omics data feature selection
Drug response prediciton
ADaM-based CRISPR genes selection
Anova/Chi-square Test

**Figure 1.** Design and workflow of the newly proposed DROEG method for drug response prediction at cell line and patient level.

## *Data preprocessing*

The input data into the model consist of mutation, CNV, gene expression, DNA methylation and CRISPR gene effect (CERES score) of all cell lines from the GDSC or CCLE dataset. Gene expression data were obtained from RNA data using RMA normalized basal expression level, whereas pre-processed CpG islands $\beta$-values were used as methylation data. Gene mutation data were binary data, and CNV data were ternary, where −1, 0 and 1 indicate loss, normal and gain of copy number genes, respectively.

## *Initial feature screening*

Due to the high dimension of features, we adopted a literature-based screen to shrink the feature size by using prior knowledge from cancer-related databases i.e. IntOGen, OncoKB, Methy-Cancer, CGI and Depmap, respectively. Next, we performed a core essential genes screening, which identified essential genes with CERES scores lower than the overall cell line average. After the first-step feature screening, a statistics-based screen for each drug was conducted. Concisely, we calculated the PCC between drug response and continuous feature variables including gene CERES, mRNA expression and DNA methylation, and calculated the Analysis of Variance (ANOVA) F value between drug response and discrete feature variables including SNV and CNV. After that, highly weighted features with PCC or F value greater than the mean value plus 2 standard deviations were selected and fed into the Support Vector Regression (SVR) model. Therefore, each drug has its unique set of features for the prediction of its response on a specific cell line.

PCC is defined as follows:

$$PCC = \frac{\mathrm{cov}\,(X, Y)}{\sigma_X \sigma_Y} \tag{1}$$

Anova F-statistics is defined as follows:

$$\mathrm{MSE} = \frac{\sigma_1^2 + \sigma_2^2 + \ldots \sigma_k^2}{k}, \mathrm{MSB} = \mathrm{n}\sigma_\mu^2, \mathrm{F} = \frac{MSB}{MSE} \tag{2}$$

## *Model fitting and parameters selection*

Next, we chose Gaussian RBF (Radical Basis Function) as the SVR model kernel function and applied 5-fold cross-validation to grid search gamma ($\gamma$) and penalty factor C for each model of each drug. The equation of Gaussian RBF is:

$$k\left(x_i, x_j\right) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right), for\ \gamma > 0, \gamma = 1/2\sigma^2 \tag{3}$$

And SVR solves the following primal problem:

$$\min_{w,b,\zeta,\zeta^*} 1/2\omega^{T}\omega + C\sum_{i=1}^{n}\left(\zeta_i+\zeta_i^*\right) \tag{4}$$

subject to $y_i - \omega^{T}\phi\left(x_i\right) - b \leq \varepsilon + \zeta_i$,

$$\begin{gathered}\omega^{T}\phi\left(x_i\right) + b - y_i \leq \varepsilon + \zeta_i^*,\\ \zeta_i, \zeta_i^* \geq 0, i = 1, \ldots, n\end{gathered} \tag{5}$$

### *Model prediction in training test*

After optimizing the SVR model parameters for each drug, we applied leave one out cross-validation in training set to further select significant omic features for predicting the cell line response value to the drug.

### *Model evaluation in test set*

Finally, we evaluated the model performance in the independent 20% test set. We calculated the PCC between predicted and measured drug response values as a quantitative evaluation. On the other hand, the DROGE model can also make the qualitative evaluation, which first classifies each drug response in each cell line into sensitive/resistant groups using K-means clustering [33].

$$J = \sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}\left\|x^{(i)} - \mu_j\right\|^2 \tag{6}$$

Then we used the clustering threshold to binarize the predicted drug response values and further compared them with the actual drug sensitivity status to get the evaluation metric including accuracy, precision, sensitivity, specificity, F1 score and Matthews correlation coefficient (MCC).

## Drug response data of cell lines in Pan-GI cancer

We also apply the DROGE method for drug response prediction of Pan-GI cancer cell line, to verify its effectiveness for specific cancer types. Pan-GI cancer includes esophageal carcinoma (ESCA), stomach adenocarcinoma (STAD), colon adenocarcinoma (COAD), rectum adenocarcinoma (READ), liver hepatocellular carcinoma (LIHC) and pancreatic adenocarcinoma (PAAD). The drugs with IC50 or ActArea values for the six cancer cell lines from GDSC or CCLE database are shown in Supplementary Table S4. There are total 173 drugs for 110 Pan-GI cell lines from GDSC2 and 24 drugs for 67 Pan-GI cell lines from CCLE dataset.

## TCGA Pan-GI cancer patient level analysis

To prove the remarkable ability of the DROEG method in predicting drug response of clinical cancer patients, we screened the drugs that can treatment more than two types of gastrointestinal cancers and have at least 20 patients clinical data in The Cancer Genome Atlas (TCGA) database. Oxaliplatin, Gemcitabine, Epirubicin, Cisplatin and 5-Fluorouracil are used for Pan-GI TCGA drug response performance evaluation. Only those patients who have complete clinical data and omics data were included, all datasets were downloaded from GDC (https://portal.gdc.cancer.gov/). Clinical data include drug response data, OS (overall survival) months and OS status data, whereas omics data consist of mutation, CNV, mRNA expression and DNA methylation data. Table 1 shows the number of clinical patients with full omics data for each Pan-GI drug.

When applying DROEG model to TCGA Pan-GI patients data, given the lack of CRISPR essential gene information for real patients, we adopted the Adaptive daisy model (ADaM) [34] algorithm to filter Pan-GI essential genes and integrate it with other four omics data to predict TCGA Pan-GI drug response. ADaM is a heuristic algorithm for the identification of essential genes, which is generalized from the Daisy Model [35] but adaptively determines cell line number m* in a driven way and adaptively defines the number of cancer-types k for which a gene can be regarded as essential genes. For a given Pan-GI cancer type *T* in *M* screened cell lines, ADaM computes fuzzy intersections of genes $I_m$, for each m = 1,…, *M*, which includes genes that are significantly depleted in at least m cell lines. Then a true positive rate TPR(*m*) for each m is calculated by considering as true positives the genes included in a priori known essential gene set E:

$$\mathrm{TPR}(m) = \mid E \cap I_m \mid / \mid E \cap G \mid, \tag{7}$$

where G is the set of all screened genes and the deviation of $|I_m|$ from its expectation $\pi_m$ is defined as follows:

$$D(m) = \log_{10}\left(|I_m|/\pi_m\right), \tag{8}$$

$\pi_m$ is estimated as the average value of the $I_m$ computed across the 1000 randomized versions of the binary depletion scores, whereas m* is defined as the maximal value of m providing the best trade-off between TPR(m) and D(m). Finally, we apply ADaM to calculate ADaM threshold m* for Pan-GI cancer cell lines and obtain corresponding Pan-GI essential genes.

After ADaM feature selection, we choose ANOVA/Chi-square to select omics features of the model because the patients' drug response type is binary. ANOVA was used to screen drug response-mRNA pairs and drug response-methylation pairs, whereas Chi-square was used to screen drug response-mutation pairs and drug response-CNV pairs.

Furthermore, we choose the SVM Classification method to predict the drug response of TCGA Pan-GI patients, as the drug response data are discrete, namely complete response, partial response, stable disease and clinical progressive disease. We then acknowledge clinical progressive disease as drug-resistant, whereas the other three types of response as drug-sensitive.

## Prognostic predictive markers identification and survival analysis

For DROEG application in TCGA Pan-GI patient response prediction, we selected the gene expression features by the DROEG model, which are important to account for drug response. Then we used the cross association to test the predictivity and descriptivity between drug response and CRISPR shRNA (CERES score), and selected the significant genes associated with drug response. Cross association quantifies the difference and reflects the bidirectional relationship between groups, and can describe the association when variables are unidirectionally related [36]. Q-omics provide a handy and user-friendly software to plot cross association between drug response and CRISPR gene efficacy [37]. We used Q-omics based on the cell line drug response data from the GDSC2 dataset (Supplementary Table S1), and the CERES score of

**Table 1.** Pan-GI drugs and the clinical patient numbers with complete omics information of TCGA database

| | COAD | READ | STAD | PAAD | LIHC | ESCA | Total samples | Samples with complete omics |
|---|---|---|---|---|---|---|---|---|
| 5-Fluorouracil | 50 | 31 | 53 | 18 | 1 | 4 | 157 | 77 |
| Cisplatin | 0 | 0 | 38 | 1 | 1 | 17 | 57 | 41 |
| Epirubicin | 0 | 0 | 26 | 0 | 0 | 1 | 27 | 20 |
| Gemcitabine | 0 | 0 | 0 | 61 | 3 | 1 | 65 | 45 |
| Oxaliplatin | 47 | 20 | 21 | 10 | 1 | 3 | 102 | 43 |

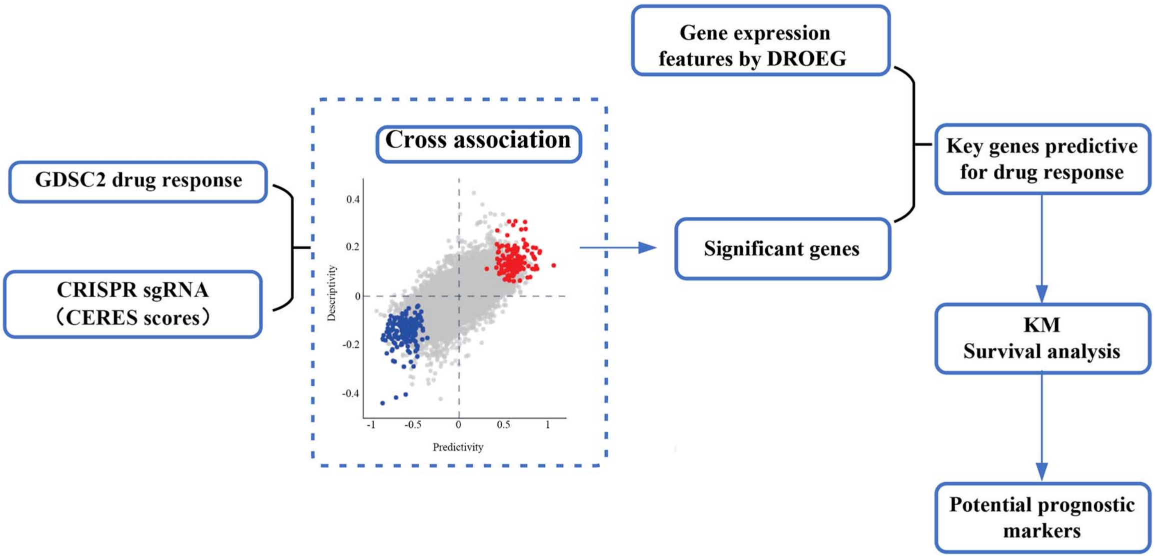

Gene expression features by DROEG
Cross association
GDSC2 drug response
CRISPR sgRNA (CERES scores)
Key genes predictive for drug response
Significant genes
KM Survival analysis
Potential prognostic markers
0.4
0.2
0
-0.2
-0.4
Descriptivity
-1
-0.5
0
0.5
1
Predictivity

**Figure 2.** The pipeline of prognostic prognostic markers prediction for TCGA Pan-GI patient samples.

corresponding cell lines from DepMap (Supplementary Table S2), and calculated the predictivity and descriptivity as Equations (9) and (10). The high or low *responses* of cell lines are determined by the threshold of the median IC50 values to a particular drug. The high or low *CRISPR_efficacy* is determined by the threshold of the median CERES scores.

$$Predictivity = avg\,(CERES\ in\ high.response) - avg\,(CERES\ in\ low.response) \quad (9)$$

$$Descriptivity = avg\,(IC50\ in\ high.CRISPR\_efficacy) - avg\,(IC50\ in\ low.CRISPR\_efficacy) \quad (10)$$

The significant genes associated with drug response were selected according to the threshold of *Predictivity P-value* $< 0.05$ *and Descriptivity P-value* $< 0.05$.

Then, by intersecting DROEG-selected gene expression features and cross-association significant genes, we obtained a list of highly significant genes as potential predictive markers. Furthermore, each potential gene for each drug was used to conduct survival analysis using the R package (survival, survminer) to test whether the gene expression level influenced the patients OS time, thus filtered the significant survival associated genes as potential prognostic markers for specific cancer. The flowchart of prognostic markers identification based on TCGA Pan-GI patient drug response is illustrated in Figure 2.

# Results

## Performance of DROEG by comparison with prior prediction methods

To compare the newly proposed DROEG model with other methods, GDSC full datasets (GDSC1 + 2) are used as a basis. Qualitative (Binarization evaluation) and quantitative (PCC) methods are both applied in our model performance evaluation, respectively. The binarization method refers to Vidhi's [33] cluster method, and we calculated accuracy, precision, sensitivity, F1 score and MCC to test DROEG model performance. For external validation, we choose the CCLE dataset and use the drug ActArea value for drug sensitivity prediction. The performance comparison (Table 2) demonstrates that DROEG has better or comparable prediction efficacy, and also greater interpretability.

For quantitative PCC evaluation, DROEG shows better results compared with the VAEN model [9], with 87.85% of drugs having PCC > 0.5 in GDSC dataset and all drugs having PCC 0.55–0.73 in CCLE dataset, whereas VAEN having 80.88% drugs with PCC >0.5 in GDSC dataset and PCC between 0.38 and 0.77 in CCLE dataset, as shown in Table 2 and Figure 3. We also calculated the Spearman correlation coefficient, which also demonstrated the better performance of DROEG than VAEN (Supplementary Figure S1). For qualitative evaluation, DROEG performs better than most methods in the GDSC full dataset in terms of accuracy, precision, sensitivity, F1 score and MCC. The reason why DROEG did not outperform MOFGCN is that the cell line numbers fulfilling the complete omics and CRISPR gene essentiality information are relatively less in DROEG model construction. Notably, DROEG exhibits the best performance in the CCLE dataset in terms of accuracy, precision

**Table 2.** Comparison of DROEG model performance with other methods

| Methods | Input data | Datasets | Qualitative evaluation | | | | | Quantitative evaluation(PCC) |
|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | Precision | Sensitivity | F1 score | MCC | |
| DROEG | CNV, mRNA, mutation, methylation, CRISPR | GDSC1 + 2 | 0.7359 | 0.6985 | 0.5963 | 0.6273 | 0.3517 | 188/214(87.85%) > 0.5 |
| | | CCLE | 0.8045 | 0.8215 | 0.8775 | 0.8474 | 0.4606 | 0.55–0.73 |
| MOFGCN | CNV, mRNA, mutation | GDSC1 + 2 | 0.7726 | 0.7402 | 0.8411 | 0.7872 | 0.5509 | |
| | | CCLE | 0.7836 | 0.76 | 0.8335 | 0.794 | 0.5719 | |
| HNMDRP | mRNA, drug structure, PPI | GDSC1 + 2 | 0.6302 | 0.5890 | 0.8662 | 0.7008 | 0.2959 | |
| | | CCLE | 0.6238 | 0.5856 | 0.9219 | 0.7153 | 0.3281 | |
| SRMF | SDF, mRNA | GDSC1 + 2 | 0.5587 | 0.5358 | 0.9078 | 0.6731 | 0.1615 | |
| | | CCLE | 0.6792 | 0.619 | 0.9336 | 0.7441 | 0.4182 | |
| VAEN | CNV, mRNA, mutation | GDSC1 + 2 | | | | | | 203/251(80.88%) > 0.5 |
| | | CCLE | | | | | | 0.38–0.77 |

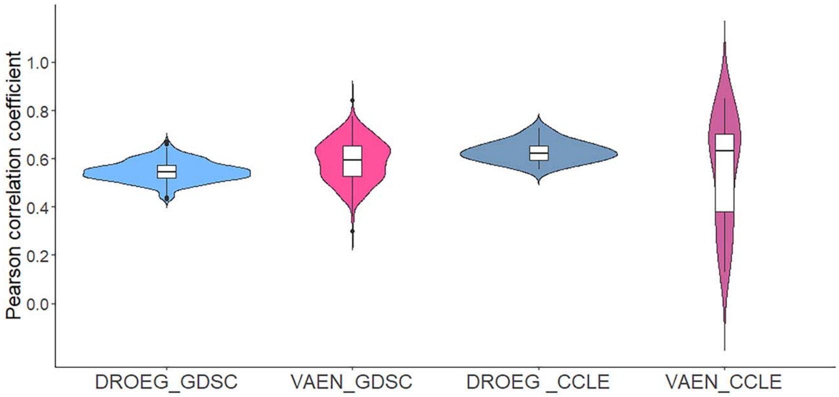


**Figure 3.** The comparison of PCC predicted by DROEG and VAEN method.

and F1 score, indicating DROEG is more effective than other state-of-the-art methods for drug response prediction (Table 2).

## CRISPR essential genes enhance the predictive performance of the DROEG model

As aforementioned, CRISPR essential genes play vital roles in determining cell growth and proliferation, and thus greatly influence drug response prediction accuracy. The GDSC2 dataset have certain improvements in the screen design, we construct DROEG in GDSC2 dataset and adopt this model for further analysis. For all 192 drugs in the GDSC2 dataset, there are 55 drugs (28.64%) targeting essential genes from CRISPR screening, whereas for approved 94 drugs recorded in Drugbank, the targets of 35 drugs (37.23%) are CRISPR essential genes. This result is similar to a previous study reported that 26% of the drugs are directly phenocopied by the CRISPR killing pattern of the known drug target (with a pathway is 48%), which means CRISPR essential genes can be a new way to identify correlated CRISPR and drug-killing profiles [38].

To test whether and to what extent CRISPR essential genes affect the model predictive performance, we first compare the two models constructed with or without CRISPR essential genes for GDSC2 (173 drugs after filtering from 192) and CCLE (24 drugs) datasets, respectively. For the DROEG model with CRISPR essential genes incorporated, the accuracy, precision, specificity and sensitivity are significantly better than the model without CRISPR essentiality information (Figure 3A). We also evaluated how essential genes influence the prediction results on the Pan-GI cancer cell line dataset (Figure 4A), and found DROEG model performed better in both GDSC Pan-GI and CCLE Pan-GI datasets than in corresponding full datasets (the average drug prediction accuracy is 0.8080 and 0.8663, respectively), which also validated that CRISPR essential genes could improve the prediction performance for drug response.

We also test whether the number of CRISPR essential genes would affect the model prediction performance by including complete essential genes (2134) or only high correlation-weighted essential genes (544) in the DROEG model. We found the models with limited and without essential gene information revealed lower accuracy compared with the original DROEG model constructed with complete essential genes (Figure 4B).

In addition, we investigated the importance of each omic information on the model prediction outcomes by running DROEG excluding each type of the input omics data, respectively. Based on the paired *t*-test on the prediction results of the complete DROEG model and that with each data type excluded, we found all partial models with one-omics-excluded perform worse than the complete DROEG model in GDSC2 full and Pan-GI datasets, except for one model excluding CNV in GDSC2 full dataset (Figure 5). The contributions of different omic levels to model accuracy are ranked as CRISPR gene effect, mRNA expression, mutation, methylation and CNV, in both GDSC full and Pan-GI datasets, indicating that CRISPR information has the most significant power to enhance the model performance. The second important data type is mRNA expression, which has been verified to be more powerful in drug sensitivity prediction, compared with SNV, CNV and methylation levels [16–18] in previous studies. We also compared the performance differences among partial models by

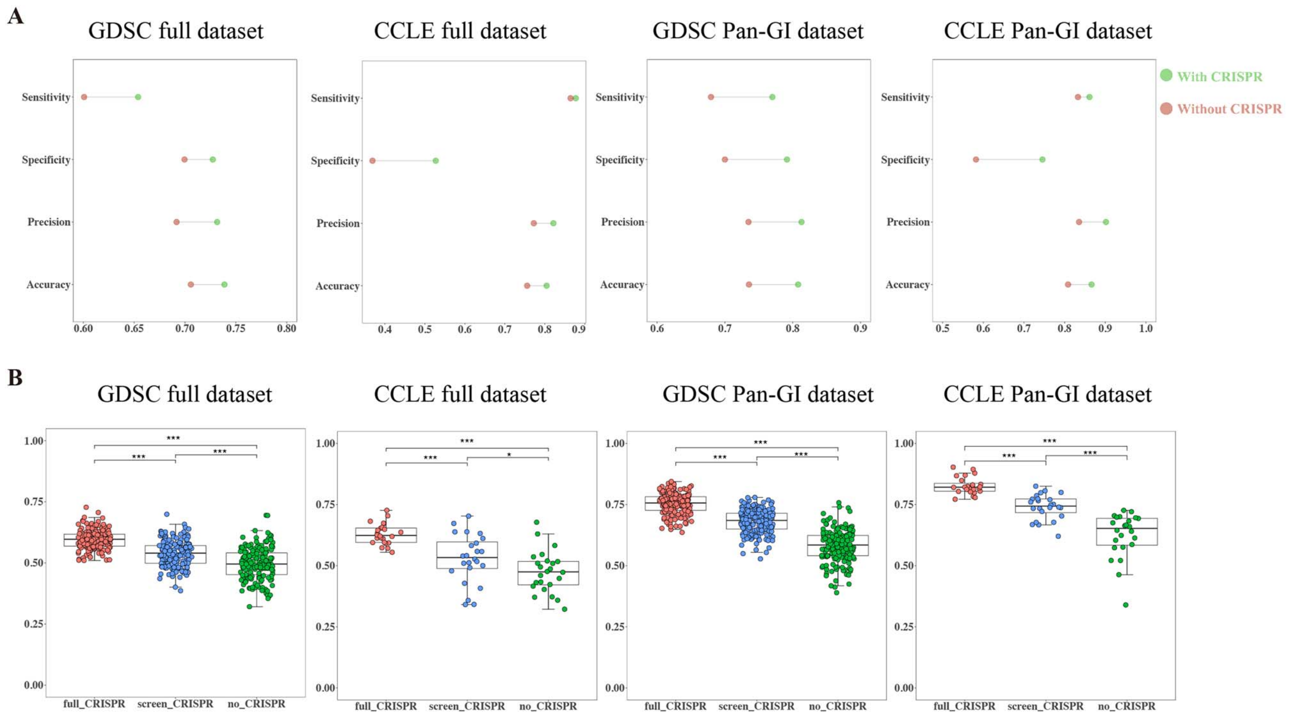


**Figure 4.** CRISPR essential genes enhance DROEG model performance. **A**. The different prediction results by models with and without CRISPR essential genes in GDSC/CCLE datasets and Pan-GI datasets. **B**. The different prediction results by models with different numbers of CRISPR essential genes incorporated.

excluding each omics type, as shown in Supplementary Table S5. In the GDSC2 full dataset, the accuracy between models with no_CRISPR essential genes and models excluding any other omics shows a significant difference ($P < 0.05$). The accuracy between the no_mRNA model and that with no_methylation and no_CNV is also significantly different, but there is no difference with the no_mutation model. Moreover, the accuracy between no_mutation and no_methylation models shows no significant difference. All results have proven that CRISPR essential genes take the most important place in drug response model construction. In summary, we incorporated the CRISPR essential genes information for drug response prediction for the first time and demonstrated its decisive contribution to the prediction results, which might be because the essential genes can account for nearly half of the mechanism of agents (MoA), and other MoAs may also indirectly related with gene synthetic lethal information.

## Validation of DROEG model efficacy on pan-GI drugs at the cell line level

Gastrointestinal cancers are the most common malignant cancers with a high mortality rate and a lack of effective drug treatment [39]. Previous study used integrative omics analysis to identify molecular stratification and potential prognostic markers in Pan-GI cancers [7]. Herein, constructing a high-accuracy drug response model for Pan-GI can help the advance of precision oncology treatment. We selected drugs that can treatment two or more than two types of gastrointestinal cancers (approved or experimental) as Pan-GI drugs. There are eight overlapped Pan-GI drugs in TCGA samples and GDSC2 dataset, and one common drug Sorafenib for liver cancer [40]. We performed the DROEG model for nine Pan-GI drugs, and found the PCCs between the predicted and actual IC50 values for the nine drugs are all higher than 0.7 (Figure 6), illustrating the robust predictive power of the DROEG method again.

## Validation of DROEG model efficacy on pan-GI drugs at the clinical patient level

Due to the lack of CRISPR gene essential information in TCGA patients, we incorporate ADaM [34] screened essential genes into four omics data (mRNA, methylation, CN, and mutation) to test model performance for the screened five Pan-GI drugs (see detail in the Methods section), including Epirubicin, Oxaliplatin, Cisplatin, Gemcitabine and 5-Fluorouracil. The ADaM results for each type of gastrointestinal cancer are shown in Supplementary Figures S2–S6. Among the Pan-GI drugs, Epirubicin and Oxaliplatin have been used for treatment of all types of Pan-GI cancers. Cisplatin can treat PAAD, STAD, ESCA and LIHC patients. Gemcitabine can treat PAAD, ESCA and LIHC patients, and Epirubicin can treat STAD and ESCA patients. All these five drugs show excellent performance in classifying sensitive and insensitive drug responses, as shown in Figure 7. The prediction accuracy of Oxaliplatin, Gemcitabine, Epirubicin, Cisplatin and 5-Fluorouracil is 90.7, 80.0,100, 80.5 and 84.4%, respectively, and all MCC values are higher than 0.5.

We retrieved significant genes between drugs and sgRNAs using cross-association analysis from Q-omics software (Supplementary Figures S7–S11 for 5 Pan-GI drugs) and then overlapped with the gene expression features to obtain candidate predictive markers affecting drug response. Moreover, we identified gene *WDR7* in PAAD samples responds to Cisplatin, gene *FAU* in ESCA samples responds to Cisplatin and genes *RPL37* and *PSMC3* in STAD samples respond to Epirubicin, those genes can distinguish the corresponding cancer samples into subtypes with significantly different OS (Figure 8), and therefore they might become potential prognostic markers. Particularly, *RPL37* and

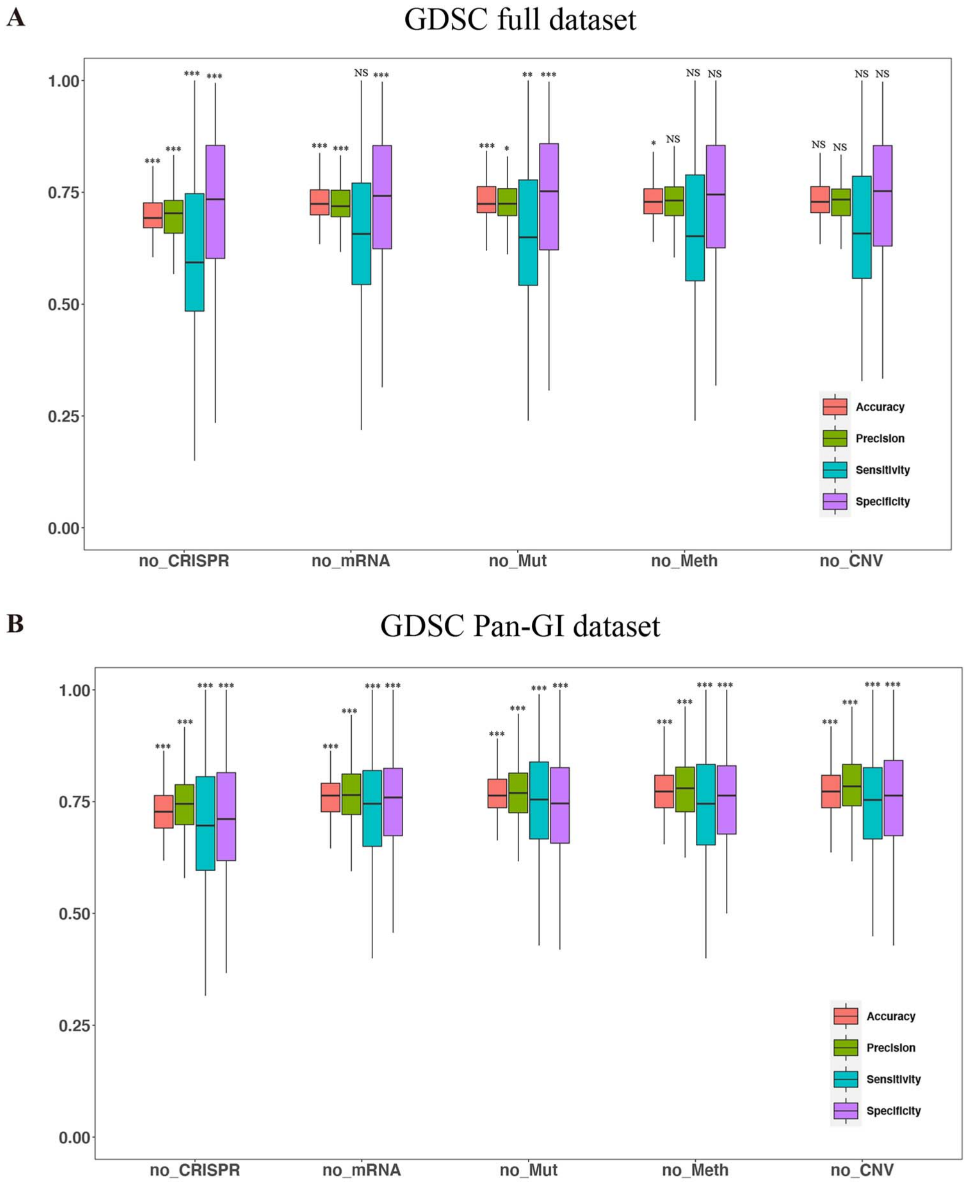


**Figure 5.** The contributions of different omic levels information to DROEG model performance in GDSC full dataset (**A**) and GDSC Pan-GI dataset (**B**). Statistics symbols indicate the paired *t*-test results between prediction results of the complete DROEG model and that with each data type excluded. **P* value <0.05, ** *P* <0.01, ****P* value <0.001,while NS means no significant difference.

*PSMC3* have been reported as prognostic markers (unfavorable) in renal cancer by the Human Protein Atlas, and Circular *PSMC3* (*CircPSMC3*) is also identified as a tumor suppressor in gastric cancer [41]. It is consistent with our finding that the STAD samples with lower expression of *RPL37* and *PSMC3* have better survival.

## Discussion

We proposed a new and effective method DROEG for drug response prediction at both cell line and patient levels, based on the integration of essential genes and multi-omics data. Our approach firstly screened omic-features from CNV, mutation, gene expression, methylation and CRISPR essential genes based on prior literature knowledge and statistical method, and then used SVR model with 5-fold cross-validation to predict the drug response and compare it with the measured value. DROEG shows excellent prediction results in the GDSC1 + 2 dataset in terms of quantitative and qualitative evaluation, and even outperforms all state-of-the-art methods in the CCLE dataset. Notably, our model can be more easily implemented and interpreted than the deep learning approach. We also tried to use XGBoost instead of SVR model to construct the prediction model (80% in training set and 20% in test set randomly) and evaluated the performance. We found the accuracy and sensitivity are similar with that by SVR; however, the precision and Matthew correlation coefficient are better by SVR (Supplemental Table S6). Therefore, we still choose SVR by considering both performance and computational complexity.

Concisely, DROEG is the first method to incorporate CRISPR essential gene features into model construction, and the essential gene information has proven to be the most effective factor with the highest predictive power in drug response prediction, which gives us insight into taking good advantage of CRISPR information. The reason CRISPR information is so important may

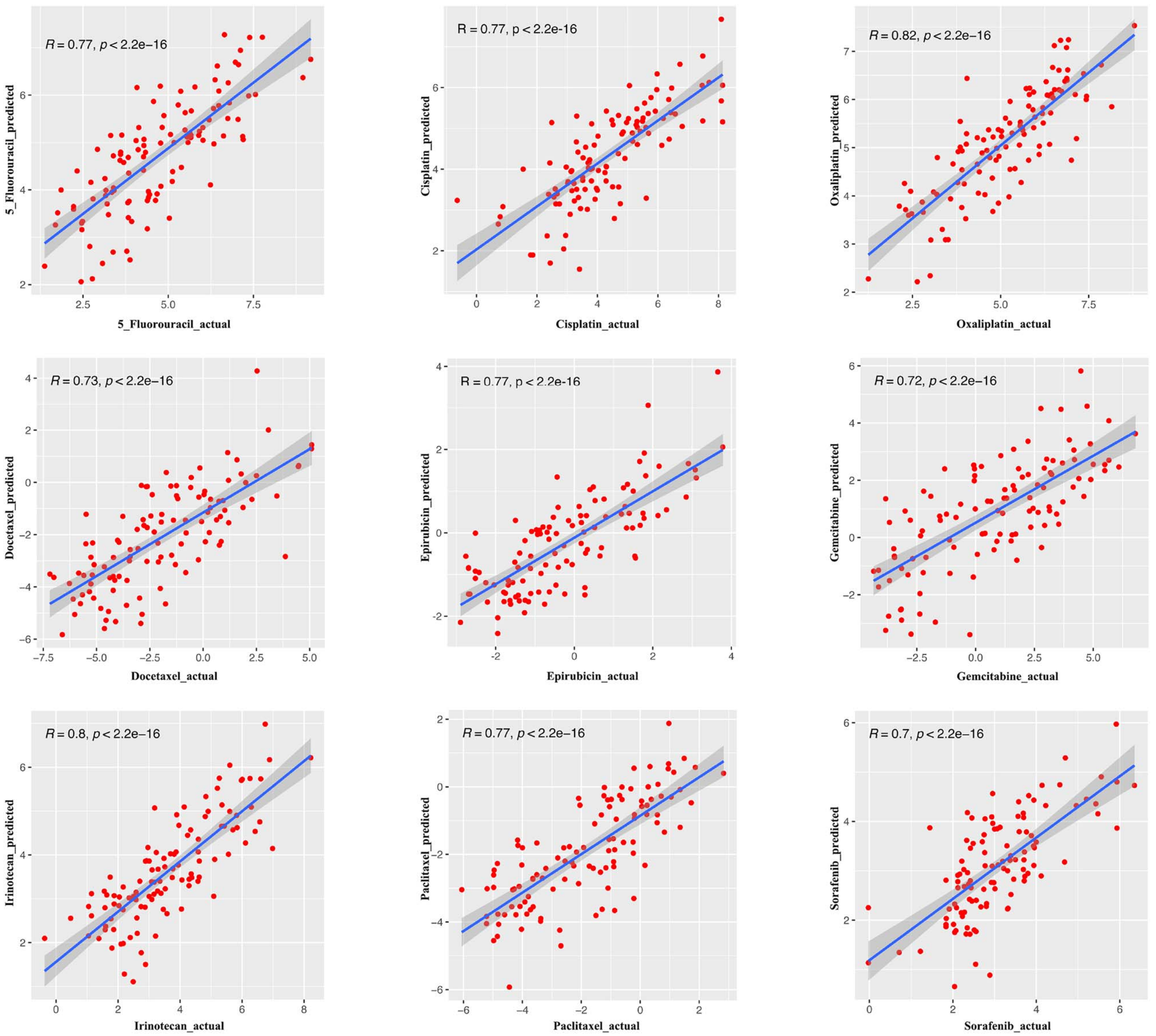


**Figure 6.** Scatter plot of PCCs between the predicted and actual response of nine Pan-GI drugs.

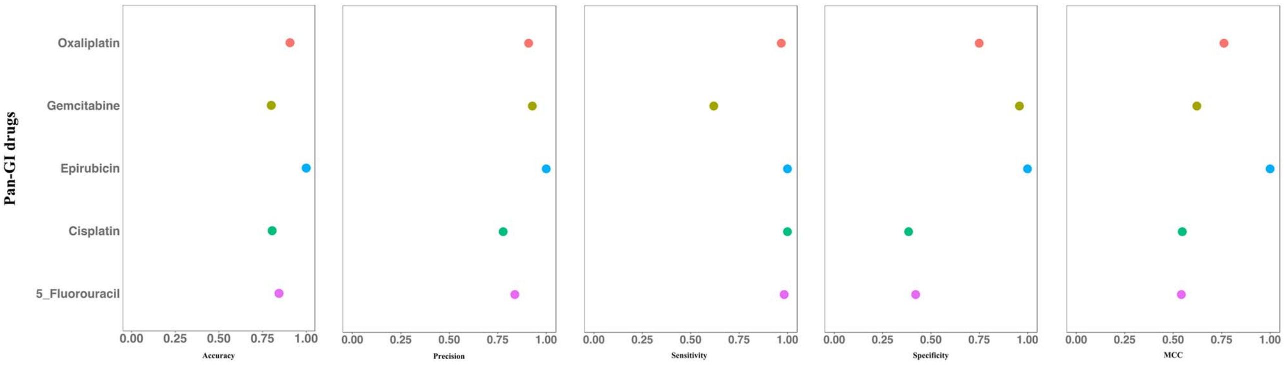
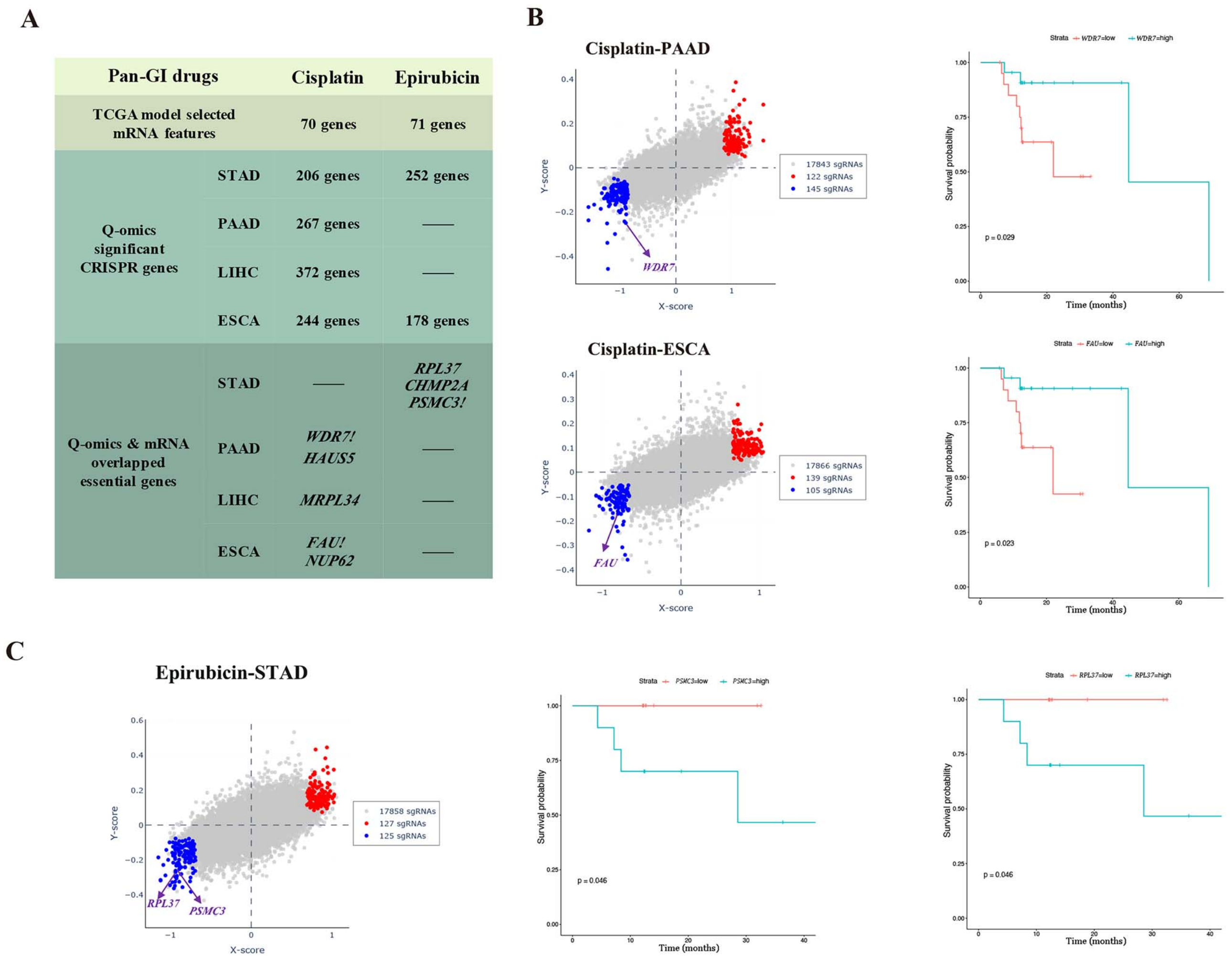


**Figure 7.** Prediction results of DROEG model for Pan-GI drugs on TCGA patient samples.

be that the tumor cell proliferation and growth can be mapped by essential genes, whereas CRISPR-Cas9 can perform systematic loss-of-function screening in well-annotated cell line levels representing the heterogeneity of tumors and thus establishing a cancer dependencies map [42]. Meanwhile, in the process of cancer drugs development, the lack of molecular entities for new targets and deficiency of targets efficacy are two major obstacles [43]. CRISPR screening has been widely used as targets related studies [26, 44] and proven to offer quite useful drug targets information [45]. The significance of CRISPR essential gene information has been extensively proven by our DROEG model. As shown in Figure 4A, when including CRISPR essential



| Pan-GI drugs | | Cisplatin | Epirubicin |
|---|---|---|---|
| TCGA model selected mRNA features | | 70 genes | 71 genes |
| Q-omics significant CRISPR genes | STAD | 206 genes | 252 genes |
| | PAAD | 267 genes | —— |
| | LIHC | 372 genes | —— |
| | ESCA | 244 genes | 178 genes |
| Q-omics & mRNA overlapped essential genes | STAD | —— | *RPL37*<br>*CHMP2A*<br>*PSMC3!* |
| | PAAD | *WDR7!*<br>*HAUS5* | —— |
| | LIHC | *MRPL34* | —— |
| | ESCA | *FAU!*<br>*NUP62* | —— |

**Figure 8.** Potential prognostic markers of Pan-GI cancers by the intersection of DROEG model selected features and cross-association significant genes. **A**. The number of overlapped essential genes by DROEG model mRNA features and Q-omics cross-association significant genes for drugs Cisplatin and Epirubicin. **B**. The survival analysis by high or low expression of *WDR7* and *FAU* in PAAD and ESCA patients with Cisplatin treatment. **C**. The survival analysis by high or low expression of *RPL37* and *PSMC3* in STAD patients with Epirubicin treatment.

genes into the DROEG model, the accuracy, precision, sensitivity and specificity in GDSC and CCLE datasets demonstrate the remarkable power of CRISPR essential genes in boosting model performance. Furthermore, the number of CRISPR essential genes may also influence the model performance, which suggests that incorporating more useful CRISPR essential genes for a specific drug may give better results (Figure 4B).

We also validated the effectiveness of the DROEG model on Pan-GI cancer as a case study. Since Pan-GI cancers are highly incident and mortal with heterogeneous treatment response, the precise prediction of Pan-GI patients' drug response is in great need. We have further investigated Pan-GI cancers drug response at both cell line and patient levels, and the DROEG model generated excellent predictive performance with PCC higher than 0.7 for all nine Pan-GI drugs in corresponding cell lines (Figure 6). Moreover, when applying the DROEG model to the patients in TCGA dataset, five Pan-GI drugs reveal delightful prediction results (Figure 7) in accuracy and precision (higher than 0.75), as well as MCC (higher than 0.5). We also identified some potential prognostic markers, such as *RPL37* and *PSMC3* could differentiate survival of gastric cancer samples, by cross-association between drugs and sgRNAs (Figure 8). Previous work usually focused on cell-line level drug response model construction and seldom emphasizing drug response at the patient level, whereas it is important to construct a suitable drug response model for patients with specific tumor types when administering a specific drug. Our work provides new insights into predicting drug response at the clinical patient level and explores some potential Pan-GI prognostic markers, which highlights the value of the drug response model in precision cancer medicine.

## Conclusion

In this study, we developed a new and effective model for cancer drug response prediction based on the integration of multi-omics data and CRISPR essential gene information (DROEG) and evaluated the performance of DROEG in both qualitative and quantitative methods. DROEG model showed excellent performance in terms of accuracy, precision and PCC. To our knowledge, DROEG is the first work to apply CRISPR essential genes into model construction and prove the key role of essential gene information in drug response. Further focusing on highly prevalent and mortal Pan-GI tumors, the effectiveness of the DROEG prediction model was demonstrated in both cell line and clinical sample datasets, and potential prognostic markers for specific drugs were identified, which is of great theoretical significance and practical value to promote personalized and precise medicine in cancer.

**Key Points**

- A novel drug response prediction method DROEG by integrating genomic, transcriptomic and methylomic data along with CRISPR essential genes.
- DROEG outperforms most state-of-the-art algorithms by both qualitative and quantitative evaluation.
- CRISPR essential gene information is the most important contributor to enhance the drug response prediction accuracy.
- DROEG is efficiently used for the Pan-GI tumor and discover potential prognostic biomarkers.

## Data availability

All datasets used in this study can be available from github at https://github.com/joyzhuowang/DROEG_dataset

## Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

## Funding

This work was supported by the National Key R&D Program of China (2021YFC2100600, 2019YFA0905400, 2021YFC2702100, 2022FYC2503300), Natural Science Foundation of China (32070679, U1804284, 81871055), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), Taishan Scholar Program of Shandong Province (tsqn201812153) and Natural Science Foundation of Shandong Province (ZR2019YQ14).

The computations in this paper were run on the $\pi$ 2.0 cluster supported by the Center for High Performance Computing at Shanghai Jiao Tong University.

## Code availability

All source codes are available at https://github.com/joyzhuowang/DROEG_code

## References

1. Ashley EA. Towards precision medicine. *Nat Rev Genet* 2016;**17**(9):507–22.
2. Lee JK, Liu ZQ, Sa JK, *et al.* Pharmacogenomic landscape of patient-derived tumor cells informs precision oncology therapy. *Nat Genet* 2018;**50**(10):1399–411.
3. Cheng ML, Berger MF, Hyman DM, *et al.* Clinical tumour sequencing for precision oncology: time for a universal strategy. *Nat Rev Cancer* 2018;**18**(9):527–8.
4. Marquart J, Chen EY, Prasad V. Estimation of the percentage of US patients with cancer who benefit from genome-driven oncology. *JAMA Oncol* 2018;**4**(8):1093–8.
5. Garnett MJ, Edelman EJ, Heidorn SJ, *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 2012;**483**(7391):570–5.
6. Sheng JT, Li FH, Wong STC. Optimal drug prediction from personal genomics profiles. *IEEE J Biomed Health Inform* 2015;**19**(4):1264–70.
7. Jiangzhou HT, Zhang H, Sun RL, *et al.* Integrative omics analysis reveals effective stratification and potential prognosis markers of pan-gastrointestinal cancers. *Iscience* 2021;**24**(8):102824.
8. Geeleher P, Cox NJ, Huang RS. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol* 2014;**15**(3):R47.
9. Jia PL, Hu RF, Pei GS, *et al.* Deep generative neural network for accurate drug response imputation. *Nat Commun* 2021;**12**(1):1740.
10. Wang L, Li XZ, Zhang LX, *et al.* Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC Cancer* 2017;**17**:513.
11. Zhang F, Wang MH, Xi JN, *et al.* A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Sci Rep* 2018;**8**:3355.
12. Sharifi-Noghabi H, Zolotareva O, Collins CC, *et al.* MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* 2019;**35**(14):I501–9.
13. Kim S, Bae S, Piao Y, *et al.* Graph convolutional network for drug response prediction using gene expression data. *Mathematics* 2021;**9**(7):772.
14. Lenhof K, Gerstner N, Kehl T, *et al.* MERIDA: a novel Boolean logic-based integer linear program for personalized cancer therapy. *Bioinformatics* 2021;**37**(21):3881–8.
15. Peng W, Chen TL, Dai W. Predicting drug response based on multi-omics fusion and graph convolution. *IEEE J Biomed Health Inform* 2022;**26**(3):1384–93.
16. Costello JC, Heiser LM, Georgii E, *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol* 2014;**32**(12):1202–U57.
17. Ding ZJ, Zu SP, Gu J. Evaluating the molecule-based prediction of clinical drug responses in cancer. *Bioinformatics* 2016;**32**(19):2891–5.
18. Iorio F, Knijnenburg TA, Vis DJ, *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* 2016;**166**(3):740–54.
19. Bhullar KS, Lagaron NO, McGowan EM, *et al.* Kinase-targeted cancer therapies: progress, challenges and future directions. *Mol Cancer* 2018;**17**:48.
20. Kim P, Jia PL, Zhao ZM. Kinase impact assessment in the landscape of fusion genes that retain kinase domains: a pan-cancer study. *Brief Bioinform* 2018;**19**(3):450–60.
21. Turski ML, Vidwans SJ, Janku F, *et al.* Genomically driven tumors and actionability across histologies: BRAF-mutant cancers as a paradigm. *Mol Cancer Ther* 2016;**15**(4):533–47.
22. Mohammad HP, Barbash O, Creasy CL. Targeting epigenetic modifications in cancer therapy: erasing the roadmap to cancer. *Nat Med* 2019;**25**(3):403–18.
23. Zhao ZB, Shilatifard A. Epigenetic modifications of histones in cancer. *Genome Biol* 2019;**20**(1):245.
24. Lv WH, Zhang XD, Dong HL, *et al.* Exploring effects of DNA methylation and gene expression on pan-cancer drug response by mathematical models. *Exp Biol Med* 2021;**246**(14):1626–42.
25. Yuan R, Chen SL, Wang YC. Computational prediction of drug responses in cancer cell lines from cancer omics and detection of drug effectiveness related methylation sites. *Front Genet* 2020;**11**:917.
26. Meyers RM, Bryan JG, McFarland JM, *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat Genet* 2017;**49**(12):1779.
27. Tzelepis K, Koike-Yusa H, De Braekeleer E, *et al.* A CRISPR dropout screen identifies genetic vulnerabilities and therapeutic targets in acute myeloid Leukemia. *Cell Rep* 2016;**17**(4):1193–205.

28. Wang T, Yu HY, Hughes NW, *et al.* Gene essentiality profiling reveals gene networks and synthetic lethal interactions with oncogenic Ras. *Cell* 2017;**168**(5):890.
29. Anglada-Girotto M, Handschin G, Ortmayr K, *et al.* Combining CRISPRi and metabolomics for functional annotation of compound libraries. *Nat Chem Biol* 2022;**18**(5):482–91.
30. Kuenzi BM, Park J, Fong SH, *et al.* Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* 2020;**38**(5):672–684.e6.
31. Aguirre AJ, Meyers RM, Weir BA, *et al.* Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. *Cancer Discov* 2016;**6**(8):914–29.
32. Li HX, Ning SY, Ghandi M, *et al.* The landscape of cancer cell line metabolism. *Nat Med* 2019;**25**(5):850.
33. Malik V, Kalakoti Y, Sundar D. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. *BMC Genomics* 2021;**22**(1):214.
34. Behan FM, Iorio F, Picco G, *et al.* Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature* 2019;**568**(7753):511.
35. Hart T, Chandrashekhar M, Aregger M, *et al.* High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* 2015;**163**(6):1515–26.
36. Jeong E, Lee Y, Kim Y, *et al.* Analysis of cross-association between mRNA expression and RNAi efficacy for predictive target discovery in colon cancers. *Cancer* 2020;**12**(11):3091.
37. Lee J, Kim Y, Jin S, *et al.* Q-omics: smart software for assisting oncology and cancer research. *Mol Cells* 2021;**44**(11):843–50.
38. Vazquez F, Boehm JS. The cancer dependency map enables drug mechanism-of-action investigations. *Mol Syst Biol* 2020;**16**(7):e9757.
39. Liu Y, Sethi NS, Hinoue T, *et al.* Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer Cell* 2018;**33**(4):721–735.e8.
40. Sun RL, Xu YZ, Zhang H, *et al.* Mechanistic modeling of gene regulation and metabolism identifies potential targets for hepatocellular carcinoma. *Front Genet* 2020;**11**:595242.
41. Rong DW, Lu C, Zhang B, *et al.* CircPSMC3 suppresses the proliferation and metastasis of gastric cancer by acting as a competitive endogenous RNA through sponging miR-296-5p. *Mol Cancer* 2019;**18**:25.
42. Tsherniak A, Vazquez F, Montgomery PG, *et al.* Defining a cancer dependency map. *Cell* 2017;**170**(3):564.
43. Hay M, Thomas DW, Craighead JL, *et al.* Clinical development success rates for investigational drugs. *Nat Biotechnol* 2014;**32**(1): 40–51.
44. Koike-Yusa H, Li YL, Tan EP, *et al.* Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat Biotechnol* 2014;**32**(3):267–73.
45. Goncalves E, Segura-Cabrera A, Pacini C, *et al.* Drug mechanism-of-action discovery through the integration of pharmacological andCRISPRscreens. *Mol Syst Biol* 2020;**16**(7):e9405.