Supplementary Materials for:

**Evaluating the molecule based prediction of clinical drug responses in cancer**

**Zijian Ding, Songpeng Zu and Jin Gu\***

MOE Key Laboratory of Bioinformatics, TNLIST Bioinformatics Division & Center for Synthetic and Systems Biology, Departmdent of Automation, Tsinghua University, Beijing 100084,China

Table of Contents:

**Supplementary Figures** (p.2 – p.6)

Figure S1. Clinical Responses to anti-cancer drugs across cancer types

Figure S2. Chronological orders between drug treatments and tumor resections, new tumor events respectively

Figure S3. Overview of the computational framework

Figure S4. The prognostic power of the two mRNA signatures of cisplatin-BLCA and cisplatin-CESC

Figure S5. The predictive performances for cisplatin or carboplatin responses in multiple cancer analysis

**Supplementary Materials and Methods** (p.7 – p.10)

**Supplementary Results** (p.11 – p.12)

**References** (p.13)

C:\Users\zding\workspace\projects\drug_sensitivity\data\drug\responses_all_cancers-01.tif

**Figure S1**. **Clinical Responses to anti-cancer drugs across cancer types**. The number in each bar represents the number of patients belonging to one of the four clinical responses. Please check the full names of each cancer acronyms in Supplementary Table S2.

C:\Users\zding\workspace\projects\drug_sensitivity\results\present_results\z_supplementary_figure\FigureS2.update_neoadjuvant-01.tif

**Figure S2.** **Chronological orders between drug treatments and tumor resections, new tumor events respectively**. (**A**) 99.4% Tumor resections were ahead of the drug treatments for the corresponding patients. “Tumor resections (days)” means that the days from the initial pathological diagnosis to the procurement of tumor samples from patients. A black star represents one patient’s drug start time and tumor resection time. The red circles represent those patients who received tumor resections prior to drug treatments. (**B**) 86.2% drug treatments were started without or before new tumor events. “New tumor events (days)” means the days from the initial pathological diagnosis to the time a new tumor occurred. A black star represents one patient’s drug start time and new tumor event time. In both (A) and (B), “Drug treatment starts (days)” means the days from the initial pathological diagnosis to the time a drug treatment started.

C:\Users\zding\workspace\projects\drug_sensitivity\results\present_results\paper_writing\manuscript\Figure1.tif

**Figure S3**. **Overview of the computational framework.** (**A**) The pipeline of the framework to evaluate the molecule-based predictions of clinical drug responses. (**B**) The construction of classifiers using elastic net with bootstrapping. “BS set” is short for bootstrapping dataset.

C:\Users\zding\workspace\projects\drug_sensitivity\results\present_results\z_supplementary_figure\FigureS5-01.tif

**Figure S4. The prognostic power of the two mRNA signatures of cisplatin-BLCA and cisplatin-CESC**. (A) The mRNA signature of cisplatin-BLCA significantly stratifies all BLCA patients into 2 or 3 groups. (B) The mRNA signature of cisplatin-LUAD fails to stratify all CESC patients into 2 or 3 groups.

C:\Users\zding\workspace\projects\drug_sensitivity\results\present_results\main_figure_4_pancancer\Figure4.0428-01.tif

**Figure S5**. **The predictive performances for cisplatin or carboplatin responses in multiple cancer analysis.** “CNA” is short for copy number alteration and “Methy” is short for DNA methylation. The black dots represent the AUC values. In each box plot, the middle line represents the mean value, the upper and lower lines represent the values of mean±sd. **(A)** Predictive performances across cancer types. The AUC values are based on predictions from different cancer types. (**B**) Predictive performances in single cancer types. “Multiple-cancer” means that the classifiers were trained on the datasets from multiple cancer types. “Single-cancer” means that the classifiers were trained on the datasets from the single cancer type. But, both classifiers were evaluated on the testing datasets from the same single cancer type.

**Supplementary Materials and Methods**

**Drug response acquisition**

We observed that different drugs have the same or very close treatment start and end time for some patients (see the full list of treatment time points in Supplementary Table S2). These entries may be recorded due to drug combinations. However, this information is hard to be checked based on the available annotations. Therefore, we can only ignore the possible effects of drug combinations in this study.

We also identified the patients who received drug treatments prior to tumor resections, as the molecular data of these patients might be interfered by these drug treatments. We collected the time points of drug treatments and tumor resections (recorded in “clinical\_drug\_cancer.txt” and "biospeciman\_cqcf\_cancer.txt", respectively) and also the history of neo-adjuvant therapies (recorded in “nationwidechildrens.org\_clinical\_patient\_cancer.txt”). We found 16 drug-patient pairs (0.4% of all records) who 1) had earlier drug starts time than tumor resections time or 2) had records of neo-adjuvant therapies (Supplementary Figure S2A). These patients were not involved in the evaluation of drug response prediction. We further identified the patients who had new tumor events prior to drug treatments. We collected the records of new tumor events (according to "clinical\_nte\_cancer.txt") and compared to the drug treatments. 86.2% of drug treatments were started without or before new tumor events (Supplementary Figure S2B).

**Molecular data collection**

For CNAs we used the level-4 non-discretized gene-level focal alterations computed by GISTIC2.0 (Mermel *et al.*, 2011) (Affymetrix Genome-wide Human SNP Array 6.0). Other data types were from level-3 processed data: gene mutations were detected by exome-sequencing as gene-level 0/1 binary values; for DNA methylations we used the probes annotated with known genes (Illumina HumanMethylation450 BeadChip); mRNA and miRNA expressions were calculated from sequencing data, as log2-transformed and quantile-normalized RSEM values for mRNAs and log2-transformed RPMs (reads per million mapped) for miRNAs; protein activities were profiled by reverse phase protein array (RPPA).

The building process of the 21 core datasets was based on the available patients with both drug responses and molecular data: 1) for one drug-cancer pair, the number of non-responders was at least 5 and the total number of patients was at least 30, since too few non-responders or all patients made it difficult to generate reliable results; 2) one specific drug-cancer pair was chosen if at least three different kinds of molecular data meet the requirement 1). The gene mutation and protein expression data did not meet the above two requirements and were not applied in our following analysis. We deleted the genes or probes with missing values (NA) in more than 10% patients for mRNA, miRNA and methylation data, and deleted the genes with CNAs in less than 10% patients. In this study, each molecular data type for one drug-cancer pair was used as a separate “dataset”.

**Molecule based prediction of clinical drug responses**

The computational framework was implemented to estimate performances of individual data types on predicting clinical drug responses (Supplementary Figure S3). The steps are as follows:

**Molecular feature pre-selection**. As the features far outnumbered the patients in each dataset, we firstly filtered the features with poor association with the responses (univariate logistic regression, p-value > 0.05). In the multiple cancer type analysis, this univariate model also took the cancer type indictor into consideration.

**Molecular signature identification**. We further identified a group of predictive features from the pre-selected candidates via elastic net classification (Zou and Hastie, 2005). However, since the numbers of responders are much more than non-responders, the training processes need to deal with the imbalanced data problem which made a model perform poorly in predicting the minority class (He and Garcia, 2009). Bootstrapping was used to solve this problem, via generating 100 balance datasets by under-sampling responders and over-sampling non-responders, meanwhile keeping the sizes of each bootstrapping dataset equal to that of the original dataset. Based on each bootstrapping dataset, an elastic net regularized multivariate logistic regression was fitted:

This objective function consisted of the negative log-likelihood function and the elastic net penalty. The likelihood function fit the unknown parameters according to the drug response and molecular features of the th patient and a total of patients. The elastic net penalty dealt with the problem that features far outnumbered patients via making most of shrink to zero. In this penalty, controls the balance of L1 and L2 penalties and controls the overall strength. We used the glmnet package (Friedman *et al.*, 2010) to solve the objective function in order to obtain the best sparse linear model. Firstly, was set to be 0~1 with the step of 0.1 and was set by glmnet's grid method. Then under each value of , glmnet's cross-validation was applied to find the best . The final best model was the one with the largest AUC under a pair of and .

We then identified a group of predictive features for drug responses, i.e. a molecular signature, in which each feature was recurrently selected by above steps. A feature score for the *p*th coefficient was defined as. In this definition, represented the percentages of coefficients larger or smaller than 0. The differential frequency favored the features which are more consistently and recurrently selected. represented the average of the *p*th coefficient in *K* elastic net models. The features with scores at least from the average of all scores were selected as the final molecular signatures.

In the application of the computational framework, the significance threshold of pre-selecting candidates, namely p value smaller than 0.05, was sometimes too stringent to select any drug response associated features. This was often specific to molecular data type, e.g. copy number alteration, or occasionally in several pairs of training and testing sets. To deal with this problem, we relaxed the p value threshold by 0.01 step by step to identify at least 2 features. This minimum considered the proper running of the following molecular signature identification procedure. The upper limit was set to be 0.1, if still no features was selected then the top 2 significant features were directly used in the following analysis. If two features were used, then the final signature contained these two features. We recorded the practical p value used and numbers of pre-selected candidate features.

Sometimes there were more than 2 pre-selected features. But the number of these candidates cannot fulfill the threshold to select final features in molecular signatures. This often happened to miRNA datasets, since there were relatively fewer miRNA transcripts for pre-selection which led to a small number of miRNA candidates. In practice, we relax the threshold to , while the other three data types used . Occasionally the threshold was still stringent for some training sets, therefore we further relaxed the threshold by 0.1 step by step, until 2 features can be selected. We also recorded the actual threshold used and the number of features in final signatures. We recorded all these statistics as a result of applying the framework to each core dataset in Supplementary Table S4.

**Predictive model construction**. Based on the identified molecular signature, an ensemble model of multivariate logistic regression was built. Each base classifier was fitted to one of the 100 bootstrapping datasets. The predicted response was the majority votes of the 100 base classifiers. The ROC curve was obtained by adjusting the thresholds to classify non-responders and responders. The area under the curve (AUC) was calculated to represent the predictive performance.

**Predictive performance evaluation**. To estimate the predictive performances, each molecular dataset was randomly split into 5 folds for 20 times, resulting in 100 pairs of training and testing sets. Each fold was ensured to have both responders and non-responders. Four folds were used to train a predictive model and the left one was used to evaluate the predictive performances. In the multiple cancer analysis, the data splitting procedure also kept the proportion of each cancer approximate to the complete dataset in training sets. The predictive performances could be biased due to the small sample size of each dataset in single cancer analysis or different drug response distributions of different cancer types in the multiple cancer analysis, therefore we implemented permutation tests by randomly permuting sample labels in training sets to estimate the null distributions of the AUCs for each dataset. The null distributions represented the performances of random classifiers and were compared to the true performances to assess the statistical significances. We shuffled the drug responses in each training set for 100 times and calculating the AUCs on corresponding testing sets. Finally the Mann-Whitney U test was applied to check if the real AUCs were significantly larger than the AUCs from permutated datasets. In the multiple cancer analysis, the shuffling procedure was applied to each single cancer types in order to keep the cancer specific distributions of drug responses.

**Estimating the statistical significance of features in each of the three signatures**

We identified the 3 mRNA or miRNA signatures using each complete dataset without data splitting. Therefore there was no other data to validate the predictive performances of the three signatures. We instead implemented a permutation test to assess each feature’s statistical significance. Thus, we proposed a hypothesis test. Since the feature score represented both the differential frequency and association coefficients, a true response associated feature had large score and an irrelevant feature had score as zero. Therefore, the hypothesis test was as follows:

Null hypothesis: ; alternative hypothesis: or

Here is the score of one feature. Since it was hard to analytically identify the true background distribution of under the null hypothesis, we shuffled the drug response labels 10,000 times for the elastic net models with bootstrapping after pre-selecting candidates. The 10,000 scores for each features represented the null distribution. The significance of one feature was calculated as follows:

In practice, we did not apply the permutation tests to the miRNA signature. There were only 10 pre-selected miRNAs, in which each miRNA may be selected too often by accident. For the two mRNA signatures, there were more than six hundred mRNA candidates, which can avoid this problem.

**Supplementary Results**

**Survival analysis of the signatures genes of cisplatin-BLCA and cisplatin-CESC**

A molecular signature here is defined to be a group of features that are recurrently selected by the predictive models. A feature can be a gene, a miRNA or a probe of methylation depending on the molecular data. In the signature of cisplatin-BLCA or cisplatin-CESC, a signature gene refers to an mRNA transcript in the corresponding molecular signature. We applied the survival analysis to identify the prognostic values of these signature genes.

We first applied univariate cox regression to every gene based on the whole-genome gene expression data. Here we expanded to all BLCA or CESC patients regardless of drug treatments. Since the 5-year survival is more concerned in the clinic, we right censored the patients who lived longer than 5 years as 5 years and alive. The statistical significances of all genes were first calculated by log-rank tests and then corrected by the Benjamini-Hochberg method. We also recorded the rankings of every gene according to their statistical significances. Finally, we chose those signature genes according to two criteria: 1) the p value was smaller than 0.01 and 2) the ranking is among the top 5% of all genes in the whole genome. As a result, the significant genes in the signature of cisplatin-BLCA include DDB1 (p value as 0.0013, fdr as 0.127, ranking 157 among 15,455 genes), INTS5 (p value as 0.0010, fdr as 0.119, ranking 131/15,455), and in the mRNA signature of cisplatin-CESC include DDL4 (p value as 0.00018, fdr as 0.023, ranking 120/15,655), HNRNPA3 (p value as 0.0057, fdr as 0.093, ranking 735/15655) and pseudogene HNRNPA3P1 (p value as 0.00025, fdr as 0.026, ranking 147/15,655) (See the list for more details of cox regression in Supplementary Table S5). For each of the five genes, we grouped the patients into three clusters (high/medium/low expression groups) and performed the Kaplan-Meier analysis (Figure 2B).

After the survival analysis of each individual gene, we investigated the prognostic power of each signature as a whole. Taking the signature of cisplatin-BLCA as an example, firstly we applied the k-means method to group all patients into 2 clusters based on the expression values of genes in the signature. Then we used the Kaplan-Meier analysis to identify whether the 2 clusters significantly stratify these patients into high risk and low risk groups. We also checked the results when patients were grouped into 3 clusters. Results show that the mRNA signature of cisplatin-BLCA is able to significantly stratify patients into groups with high, medium and low risks (Supplementary Figure S4A), while the signature of cisplatin-CESC is not able to stratify the CESC patients (Supplementary Figure S4B). This is not surprising since the signatures are correlated with cisplatin responses, they are not necessarily related to patient overall survival.

**Evaluation of the molecule-based drug response predictions across cancer types**

We compared the performances in certain single cancer types between two kinds of classifiers, namely the multiple-cancer classifier and single-cancer classifier. The multiple-cancer classifiers were trained on data from multiple cancer types, and single-cancer classifiers were trained on data of exactly the tested single cancer types. That is to say, these two classifiers have the same testing data, and the only difference is their training data. Four cancer types were chosen, including BLCA, CESC, LUAD and UCEC, since they have relatively more patients than all other cancer types.

Out of the 15 comparisons as shown in Figure S5B, eight were significantly different (paired sample t test p value<0.05), including the methylation, miRNA expression and mRNA expression data for cisplatin-BLCA and cisplatin-LUAD, and methylation and mRNA expression data for carboplatin-UCEC (p values in Supplementary Table S8). Among these eight comparisons, only miRNA datasets consistently show significant better performances based on the multiple-cancer classifier for cisplatin treatments. In BLCA, the multiple-cancer classifier achieved the average AUC as 0.610 and the single cancer classifier achieved the average AUC as 0.457. The paired sample t test resulted in the p value as 3.13e-5. For LUAD, the multiple cancer classifier got the average AUC as 0.457 and the single cancer classifier as 0.33, with p value as 5.95e-6. Therefore, the multiple-cancer classifier of miRNA expression significantly improves performances on predicting cisplatin responses on several single cancer types, including BLCA and LUAD. We noticed that although the multiple-cancer classifier improved the performances in LUAD, the average AUC is still not larger than 0.5. But, the performances in BLCA is significantly better than random guessing (permutation test p value 1.62e-5). Therefore, the multiple-cancer classifier of miRNA expression significantly improves performances on predicting cisplatin responses in several single cancer types, including BLCA and LUAD.

**References**

Friedman,J. *et al.* (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.*, **33**, 1–22.

He,H. and Garcia,E. a. (2009) Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.*, **21**, 1263–1284.

Mermel,C.H. *et al.* (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.*, **12**, R41.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **67**, 301–320.