Published in final edited form as:
IEEE J Biomed Health Inform. 2015 July ; 19(4): 1264–1270. doi:10.1109/JBHI.2015.2412522.

Optimal Drug Prediction from Personal Genomics Profiles

Jianting Sheng[†],
NCI Center for Modeling Cancer Development, Department of Systems Medicine and Bioengineering, Houston Methodist Research Institute, Weill Cornell Medical College, Houston, TX 77030, USA

Fuhai Li[†], and
NCI Center for Modeling Cancer Development, Department of Systems Medicine and Bioengineering, Houston Methodist Research Institute and Houston Methodist Cancer Center, Houston Methodist Hospital, Weill Cornell Medical College, Houston, TX 77030, USA

Stephen T.C. Wong[*]
NCI Center for Modeling Cancer Development, Department of Systems Medicine and Bioengineering, Houston Methodist Research Institute and Houston Methodist Cancer Center, Houston Methodist Hospital, Weill Cornell Medical College, Houston, TX 77030, USA; Houston Methodist Cancer Center, Houston Methodist Hospital, Houston, TX 77030, USA

Jianting Sheng: jsheng@houstonmethodist.org; Fuhai Li: fli@houstonmethodist.org

Abstract

Cancer patients often show heterogeneous drug responses such that only a small subset of patients is sensitive to a given anti-cancer drug. With the availability of large-scale genomic profiling via next generation sequencing (NGS), it is now economically feasible to profile the whole transcriptome and genome of individual patients in order to identify their unique genetic mutations and differentially expressed genes, which are believed to be responsible for heterogeneous drug responses. Although subtyping analysis has identified patient subgroups sharing common biomarkers, there is no effective method to predict the drug response of individual patients precisely and reliably. Herein, we propose a novel computational algorithm to predict the drug response of individual patients based on personal genomic profiles, as well as pharmacogenomic and drug sensitivity data. Specifically, more than 600 cancer cell lines (viewed as individual patients) across over 50 types of cancers and their responses to 75 drugs were obtained from the Genomics of Drug Sensitivity in Cancer (GDSC) database. The drug-specific sensitivity signatures were determined from the changes in genomic profiles of individual cell lines in response to a specific drug. The optimal drugs for individual cell lines were predicted by integrating the votes from other cell lines. The experimental results show that the proposed drug prediction algorithm can be used to improve greatly the reliability of finding optimal drugs for individual patients and will thus form a key component in the precision medicine infrastructure for oncology care.

[*] Corresponding author: stwong@houstonmethodist.org.
[†] Authors contributed equally to this work

Index Terms

personalized medicine; drug sensitivity; personal genomics; drug response prediction

I. Introduction

Traditionally, cancer types have been defined by their anatomical origin. Molecular analysis now shows that cancers of different organs may share similar features while cancers from the same organ are often quite distinct, as evidenced by the observation that cancer patients with the same cancer type often show heterogeneous responses to the same therapy. It is believed that complex and distinct genomic instability is responsible for the diversity in drug response [1, 2]. For example, one non-small cell lung cancer (NSCLC) patient was found to be responsive to Dasatinib dramatically and remained cancer-free four years later [3]. In the consequent genomic analysis, an inactivating BRAF mutation (Y472C) was found in the patient's tumor while no inactivating BRAF mutations were found in the non-responding tumors [3]. The Pan-Cancer project [4], which aims to analyze the molecular aberrations in cancer cells across a range of tumor types, combines all of the data from different tumor types and enables researchers and scientists to find new patterns of genomic aberrations.

To support studies on uncovering the relationship between the genomic aberrations and diverse drug responses, many valuable large-scale genomic data sets funded by federal agencies have been generated for public access. For example, Cancer Cell Line Encyclopedia (CCLE) [5] and Genomics of Drug Sensitivity in Cancer (GDSC) [6] are two projects that generate such databases and aim to systemically study the causal correlations between drug sensitivity and genomic biomarkers in hundreds of cancer cell lines. Moreover, The Connectivity Map (CMAP) generated gene expression signature data of different cell lines under various perturbations by more than 1,000 small molecules [7]. These data have since been scaled over 1,000 times and is publicly available at the Library of Integrated Network-based Cellular Signatures (LINCS) (http://www.lincscloud.org/). These large-scale data sets have been used to enable drug repositioning [8, 9], predict drug combinations [10, 11] and delineate mechanisms of action (MoA) [12] and are becoming an important component in drug discovery [13, 14]. Meanwhile, Gene Expression Omnibus (GEO) [15], a public database repository, also archives millions of microarray data sets of various types of cancer under different treatments and perturbations.

Many approaches have been developed to reposition or repurpose known drugs to new indications [16, 17], based on the assumption that drugs sharing similar chemical structures would have same targets. With the advance of large-scale genomic profiling techniques, it is economically feasible to profile the whole transcriptome and genome (e.g., microarray, RNAseq, DNAseq and somatic copy number alterations (SCNAs)) [18-20] of individual patients and identify their unique genetic mutations and differentially expressed genes, which in turn can be used as biomarkers to predict drug sensitivity for those patients. For example, The Cancer Genome Atlas (TCGA) [21, 22] database repository has collected comprehensive genomics profiles of over ten thousand patient samples covering more than 30 types of cancer. Also, the International Cancer Genome Consortium (ICGC) [23] was formed to coordinate the generation and management of large-scale genomic data sets from over fifty cancer types or subtypes collected from all over the world. These large-scale and comprehensive genomic profiling data sets are invaluable in identifying predictive biomarkers of an individual's drug response, these unique biomarkers can be used for personalized medicine [15, 24-26].

Though subtyping analysis of cancer patients has identified patient subgroups sharing common biomarkers, the genomic instability patterns among the subtypes of a cancer are complex and not easily distinguishable [18, 22, 27-29]. Thus, these genomic patterns derived from the subtyping analysis often fail to offer insight into drug sensitivity and identify reliable predictive biomarkers of individual subtypes. Herein, we propose a novel computational algorithm to predict the optimal drugs for individual patients based on personal genomic profiles, as well as available pharmacogenomics and drug sensitivity information across cancer types. Specifically, 624 cancer cell lines (viewed as individual patients) and their responses to 75 drugs were obtained from GDSC. The drug-specific sensitivity signatures were obtained with the genomics profile changes of individual cell lines. Then, the optimal drug treatments for individual cell lines were predicted by integrating the drug response information from the other cell lines. To test the accuracy of our drug prediction algorithm, we performed two-fold cross-validation on GDSC dataset and further downloaded 480 cell lines and their responses to 11 drugs from CCLE for validation. The results show that our prediction algorithm can significantly improve the chance of finding sensitive drugs for individual patients.

II. Data and Methodology

2.1 Data

The simplified molecular-input line-entry system (SMILES) data that describe the structure of compounds using short ASCII strings were downloaded from PubChem (version 2.0.1, 17 Oct, 2011) [30]. Gene expression data of cell lines, and the IC50 data of drugs on the cell lines were downloaded from GDSC (release 4, march 2013) and CCLE (version 2.17, August 2014). Only drugs that have both SMILES and IC50 information, and cell lines that have both IC50 and gene expression profile data were used in our analysis. In summary, IC50 data of 75 drugs on 624 cell lines in GDSC and 24 drugs on 504 cell lines in CCLE were obtained.

2.2 Algorithm overview

Figure 1 shows an overview of the proposed drug prediction algorithm. Let D be the set of known drugs in GDSC, C be the cell line set, and G be the drug-related gene signature consisting of differentially expressed genes between the drug-specific resistant and sensitive cell lines, where $n_d = |D|$, $n_c = |C|$, $n_g = |G|$ denoting the number of drugs, cell lines, and genes respectively. Let $A = (a_{ij})$ be the cell line expression matrix, where a_{ij} is the expression value of gene $g_i \in G$ in cell line $C_j \in C$. Let $B = (b_{jk})$ be the IC50 value matrix, where b_{jk} is the normalized IC50 value of cell line $C_j \in C$ treated by drug $d_k \in D$. For each drug-cell line pair, we first calculate similarity scores for pairwise drugs based on their chemical structures and then select most similar drugs to the given one. A gene signature is calculated based on the IC50 values of the selected drugs and gene expression profiles of the cell lines. A flexible similarity score between cell lines was calculated based on the gene signature and finally we assign a drug-cell line effect score by combining drug-similarity scores and cell line similarity scores. Using this method, we predicted the top candidate drugs that have the best sensitivity for the given cell line.

2.3 IC50 value normalization

As suggested by GDSC, cell lines with IC50 values greater than the maximum concentration of a drug are likely to be resistant to that drug while those with IC50 values smaller than maximum concentration are likely to be sensitive. So we divide the raw IC50 value by the maximum concentration of $d_k \in D$ and then log2-transform it as normalized IC50 value. For drug-cell line pairs that lack IC50 values from GDSC, we set their normalized IC50 to 0. Cell lines with a normalized IC50 value greater than 0 are considered to be resistant cell lines to the drug while others with normalized IC50 value smaller than 0 were considered to be sensitive. As the range of IC50 values of sensitive cell lines differs from the range of resistant cells, we further scaled the normalized IC50 value to the range of [-1,1] for the following analysis.

2.4 Drug-drug similarity

Given two drugs $d_{k1} \in D$ and $d_{k2} \in D$, the ‘rcdk’ R package [31] is used to retrieve their fingerprint information fp_{k1} and fp_{k2} from the SMILES data. The similarity between the two drugs, $S(d_{k1}, d_{k2})$, is calculated using Tanimoto coefficient:

$$S(d_{k_1}, d_{k_2}) = \frac{c}{a+b+c}$$

where c is the number of fingerprints that appear in both fp_{k1} (figureprints of d_{k1} by using the ‘fingerprint’ R package) and fp_{k2} (figureprints of d_{k2}), a is the number of fingerprints that only appear in fp_{k1}, and b is the number of fingerprints that only appear in fp_{k2}.

2.5 Patient-patient genomics similarity

Even if two cell lines (or patient cells) share similar overall gene expression signatures, they might still have different drug-treatment response due to the variability in gene expressions of the drug-resistance related genes. Here, we propose a new method of calculating the cell line similarity scores based on drug-specific related genes, i.e., different similarity scores are given to each pair of cell lines for different drugs as follows.

For a given drug $d_{given} \in D$, to calculate cell line similarity, we first calculate the drug-drug similarities, $S_k = S(d_{given}, d_k)$, between d_{given} and all other drugs $d_k \in D$. Then, we rank d_k in descending order S_k. Top r_d drugs, which show the most similarity to the given drug, are selected as $D_{d_{given}}$. The next step is to find a gene signature for $D_{d_{given}}$ and calculate cell line similarity score based on this signature.

For each drug $d_t \in D_{d_{given}}$, we separate cell lines into two groups according to their normalized IC50 values. If b_{jt}, which is the IC50 value of cell lines C_j treated by drug d_t, is greater than 0, C_j is assigned to group 1 (the resistant group); whereas the others with $b_{jt} < 0$

are assigned to group 2 (the sensitive group). For each gene g_i, we calculate its fold change between the mean expression levels of the two groups,

$\mathrm{Fold}(\mathrm{g_i}) = \frac{\mathrm{mean}\{\mathrm{a_{ij}}|\mathrm{b_{jt}}>0, j=1,\ldots,\mathrm{n_c}\}}{\mathrm{mean}\{\mathrm{a_{ij}}|\mathrm{b_{jt}}<0, j=1,\ldots,\mathrm{n_c}\}}$. The top 1% genes with the largest or smallest fold change are selected as the gene signature of d_t, denoted by G_{d_t}. Then, signatures of all drugs in $D_{d_{given}}$ are combined together as the signature of $D_{d_{given}}$, denoted by $G(d_{given}) = \cup_{d_t \in D_{d_{given}}} G_{d_t}$.

Using the *k* signature genes in $G(d_{given})$ as the features, we define the similarity values $S(c_{j1}, C_{j2})$ between two cell lines $c_{j1}, c_{j2} \in C$ under drug d_{given} as the Pearson Correlation Coefficient between $<a_{i1j1},a_{i2j1}, \ldots, a_{ikj1}>(a_{ij_1})$ and $<a_{i1j2},a_{i2j2}, \ldots, a_{ikj2}>(a_{ij_2})$, where $g_i \in G(d_{given})$. To be consistent to the drug-drug similarity score, we scale $S(c_{j1}, c_{j2})$ to [0,1]. Given a cell line c_{given}, the top r_c cell lines that are most similar to it are selected as $C_{c_{given}}$.

2.6 Drug-patient similarity

Given the gene expression profile of a patient, we rank the effects of all drugs in GDSC on this patient by defining the drug-patient effect score $S(d_{given}, c_{given})$. We first find $D_{d_{given}}$ to include top r_d drugs that are most similar to d_{given} and $C_{c_{given}}$. to include top r_c cell lines that are most similar to c_{given}. Then the drug-patient effect score is defined by:

$$\mathrm{S}(\mathrm{d_{given}}, \mathrm{c_{given}}) = \frac{1}{r_d r_c} \sum_{\substack{\mathrm{d_k} \in \mathrm{D_{d_{given}}} \\ \mathrm{c_j} \in \mathrm{C_{c_{given}}}}} \mathrm{b_{jk}} \sqrt{\mathrm{S}(\mathrm{d_{given}}, \mathrm{d_k}) \times \mathrm{S}(\mathrm{c_{given}}, \mathrm{c_j})}.$$

III. Results

The raw IC50 values for 75 drugs on 624 cell lines from GDSC and their cell line expression profiles were collected, and normalized as described in Method part. Figure 2 shows the heatmap of the normalized IC50 values (column scaled). The normalized IC50 values were further scaled to [-1,1] (-1 means the most sensitive, and 1 means the most resistant) for the subsequent analysis. We selected the top 1% genes (about 300 genes) with the largest or smallest fold change between the resistant group and sensitive group as the gene expression signature of individual drugs. Figure 3 shows the distinct gene expression signatures of AICAR and Dasatinib. To predict the effectiveness between a given cell line and a given drug, the top r_d most similar drugs and r_c cell lines were tested. Finally, we assigned a similarity score for each drug-cell line (patient) pair. Given a new cell line, the top ranked drugs were considered to be the best candidates.

We have conducted two new validation strategies to reduce the bias in validation and for selecting optimal values for r_d and r_c. The first one is to use two-fold cross-validation for modeling GDSC dataset, and the second one is to use GDSC dataset as the training set and CCLE dataset as the independent testing set.

First, we apply two-fold cross-validation on GDSC dataset, i.e., randomly divide the cell lines into two sub-groups, with equal size and each time use one as training set and the other as testing set. The testing results are then averaged for each r_d and r_c. The drug response on the testing set were estimated using drug treated IC50 values on the training set. Top ten prioritized drugs for each testing cell line were selected. All the drugs were ranked in increasing order based on their normalized IC50 values on the given cell line. We then calculated the enrichment score of the ten drugs in the ordered drug list using Gene Set Enrichment Analysis (GSEA) [32], denoted as $\mathrm{ES_{cl}}$. The enrichment score, $\mathrm{ES_{cl}}$, which ranges from -1 to 1, and would be close to 1 if the selected drugs were found at the top of the ranked drug list and vice versa. To test whether our selected drugs for the given cell line ranked higher than random selection, we randomly selected ten drugs for each cell line for five thousand times and recalculated the enrichment score, denoted by $\mathrm{ES_i}$, i = 1,..,5000. The p value was calculated as: $\frac{\sum_{\mathrm{i=1}}^{5000}\mathrm{f_i}}{5000}$, where $\mathrm{f_i}=\begin{cases} 1, \mathrm{ES_i} \geq \mathrm{ES_{cl}} \\ 0, \mathrm{ES_i}<\mathrm{ES_{cl}} \end{cases}$. Small p value (here we set the threshold as p<0.05) means that the prediction can discover the more sensitive drugs to the given cell line, compared with the random selection. The percentage of cell lines with corresponding p value<0.05 were then averaged (two-fold cross-validation) to for each r_d = [1, …,4] and r_c = [1,…,10]. The results were shown in Figure 4. As can be seen, the prediction power of single drug (r_d=1) decreases when more similar cell lines being added (noise sensitive). Whereas, the prediction power with multiple similar drugs increases with more similar cell lines (noise robust). Especially, the parameter setting, r_d =3 and r_c =9, generates the best prediction power.

Moreover, the CCLE data sets were also collected for the validation. There are 11 drugs in both GDSC and CCLE database. We use GDSC data as the training set, and download drug treated IC50 values for 480 cell lines of the 11 drugs from CCLE database as the testing dataset. We then predict the drug responses on 480 cell lines in CCLE for 11 drugs. Since there are only 11 drugs available, it is hard to find sensitive drugs for each cell line. We thus select top 10 drug-cell line pairs predicted using our model and compare their IC50 values in CCLE with all other drug-cell line IC50 values. Then *p* values were calculated for each r_d and r_c using GSEA as described above. The results were shown in Table 1. As we can see, the p values are smaller than 0.05 in almost all the cases, which means the top 10 drug-cell line pairs predicted by our model have a much lower IC50 value compared with random selection and thus the cell lines are more sensitive to these drugs. Note that the *p* value for r_d=3 and r_c =9 was smaller than 4e-04. By considering both validation method, we suggest to use r_d=3 and r_c=9 in our model.

In summary, the proposed prediction algorithm can improve the effective drug selection significantly, and the evaluation results indicate the possibility of selecting the optimal drugs among all anti-cancer drugs based on their genomic profiles to achieve better therapeutic outcomes.

IV. Discussion

It is difficult to determine “sensitive” responses to drugs. We use normalized IC50 score to measure the cell line sensitivity based on the documents of GDSC which suggest that cell lines with acute drug response, IC50<max drug concentration and small confidence intervals are supposed to be sensitive to drugs. In the latest release of GDSC (release 5, June 2014), a

Z-score is calculated for each cell line to define "sensitive" and "resistance" to drugs. Both measurements should be fine and people can also use other drug response data to replace the normalized IC50 values in our algorithm.

There also exist other algorithms to find differentially expressed genes between sensitive groups and resistant groups. We first tested if t test could be used to find drug signatures. We did Shapiro-Wilk test on the expression values of genes across cell lines and found that the expression value of most of the genes (over 10,000) did not form normal distribution. Thus t test and other similar methods such as SAMR were not suitable here. Then we tested Wilcoxon rank sum test which was a nonparametric statistic test followed by the Benjamini and Hochberg method (Journal of the Royal Statistical Society Series B 57, 289–300, 1995) to control false discovery rate. Genes with adjusted p value <0.05 were selected as drug signature. However, the number of differentially expressed genes varies significantly from tens to thousands; furthermore, the drug prediction result is worse than using the top 1% fold change genes. One possible reason is that the sample size of sensitive cell lines for each drug is usually too small (less than 10) to provide sufficient power for statistic testing.

V. Conclusion

Patients with the same type of cancer often respond differently to the same drug treatment, and only a small subset of patients has profound sensitivity to the given drugs. The complex and distinct genomic instability of individual cancer patients is believed to be responsible for that diversity in drug response. Due to recent advances made in the next generation sequencing, large-scale pharmacogenomic databases are now publicly available while genomic profiling of individual patients is becoming economically affordable. It is timely to leverage such information to predict the drug sensitivity for each patient in order to achieve optimal personalized treatment. However, computational analysis is the bottleneck in associating the complex genomic profiles reliably with heterogeneous drug responses.

Motivated by the public availability of informative data resources, this study contributed a computational algorithm for predicting the best drugs for individual cell lines (representing individual patients) based on their genomic profiles and drug response data. Our cross-validation showed that the best drug prediction results on over 40% cell lines are improved significantly compared to random selection. Moreover, the results from CCLE validation showed that the top ten most sensitive drug-cell line pairs predicted by our model using GDSC database have significant lower IC50 values than others. Our research finding suggests the potential of selecting optimal drugs among anti-cancer drugs for individual patients based on their genomic profiles, which will impact the cancer clinical regimen profoundly. Our ongoing work of integrating the drug prediction algorithm into the precision medicine framework being developed at Houston Methodist Cancer Center to investigate personal genomic profiles from patients, evaluate the prediction results on patient-derived xenograft models, and track the predictive performance on selected patients under treatment at Houston Methodist Hospital.

Acknowledgments

This work is supported by CPRIT RP110532, NIH U54 CA149196, John S Dunn Research Foundation, and Ting Tsung and Wei Fong Chao Foundation. The authors would like to thank members of Systems Biomedicine lab and Data Science lab of Department of Systems Medicine and Bioengineering at HMRI, Weill Cornell Medical College for their valuable discussions.

This work was supported by CPRIT RP110532, NIH U54 CA149196, and John S Dunn Research Foundation.

References

1. Zlotta AR. Words of wisdom: Re: Genome sequencing identifies a basis for everolimus sensitivity. Eur Urol. 2013; 64(3):516. [PubMed: 23915465]
2. Iyer G, et al. Genome sequencing identifies a basis for everolimus sensitivity. Science. 2012; 338(6104):221. [PubMed: 22923433]
3. Sen B, et al. Kinase-impaired BRAF mutations in lung cancer confer sensitivity to dasatinib. Sci Transl Med. 2012; 4(136):136ra70.
4. Weinstein JN, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013; 45(10):1113–20. [PubMed: 24071849]
5. Barretina J, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483(7391):603–7. [PubMed: 22460905]
6. Garnett MJ, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012; 483(7391):570–5. [PubMed: 22460902]
7. Lamb J, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006; 313(5795):1929–35. [PubMed: 17008526]
8. Dudley JT, et al. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. Sci Transl Med. 2011; 3(96):96ra76.
9. Sirota M, et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Transl Med. 2011; 3(96):96ra77.
10. Huang L, et al. DrugComboRanker: drug combination discovery based on target network analysis. Bioinformatics. 2014; 30(12):i228–i236. [PubMed: 24931988]
11. Ji-Hyun, Lee, et al. CDA: Combinatorial Drug Discovery Using Transcriptional Response Modules. PLoS one. 2012; 7(8):e42573.10.1371/journal.pone.0042573 [PubMed: 22905152]
12. Iorio F, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci U S A. 2010; 107(33):14621–6. [PubMed: 20679242]
13. van Noort, Vv, et al. Novel drug candidates for the treatment of metastatic colorectal cancer through global inverse gene expression profiling. Cancer Research. 2014
14. Jahchan NS, et al. A Drug Repositioning Approach Identifies Tricyclic Antidepressants as Inhibitors of Small Cell Lung Cancer and Other Neuroendocrine Tumors. Cancer Discovery. 2013
15. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30(1):207–10. [PubMed: 11752295]
16. Campillos M, et al. Drug Target Identification Using Side-Effect Similarity. Science. 2008; 321(5886):263–266. [PubMed: 18621671]
17. Kinnings SL, et al. Drug Discovery Using Chemical Systems Biology: Repositioning the Safe Medicine Comtan to Treat Multi-Drug and Extensively Drug Resistant Tuberculosis. PLoS Comput Biol. 2009; 5(7):e1000423. [PubMed: 19578428]
18. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489(7417): 519–25. [PubMed: 22960745]
19. Wilkerson MD, et al. Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. Clin Cancer Res. 2010; 16(19):4864–75. [PubMed: 20643781]
20. Li F, et al. Conditional random pattern model for copy number aberration detection. BMC Bioinformatics. 2010; 11:200. [PubMed: 20412592]
21. TCGA-web, https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp.

22. Wang L, et al. A Computational Method for Clinically Relevant Cancer Stratification and Driver Mutation Module Discovery Using Personal Genomics Profiles. BMC Genomics. 2014 To appear.
23. Hudson TJ, et al. International network of cancer genome projects. Nature. 2010; 464(7291):993–8. [PubMed: 20393554]
24. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet. 2010; 11(10):685–696. [PubMed: 20847746]
25. Lawrence MS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499(7457):214–218. [PubMed: 23770567]
26. Li YY, Jones SJ. Drug repositioning for personalized medicine. Genome Med. 2012; 4(3):27. [PubMed: 22494857]
27. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490(7418):61–70. [PubMed: 23000897]
28. Kandoth C, et al. Integrated genomic characterization of endometrial carcinoma. Nature. 2013; 497(7447):67–73. [PubMed: 23636398]
29. Wang, L., et al. A Computational Method for Clinically Relevant Cancer Stratification and Driver Mutation Module Discovery Using Personal Genomics Profiles. International Conference on Intelligent Biology and Medicine (ICIBM), 2014; San Antonio. December 4∼6, 2014;
30. Rajarshi G. fingerprint: Functions to operate on binary fingerprint data. R package, version 3.5.2. 2013
31. Guha R. Chemical Informatics Functionality in. R Journal of Statistical Software. 2007
32. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102(43):15545–50. [PubMed: 16199517]

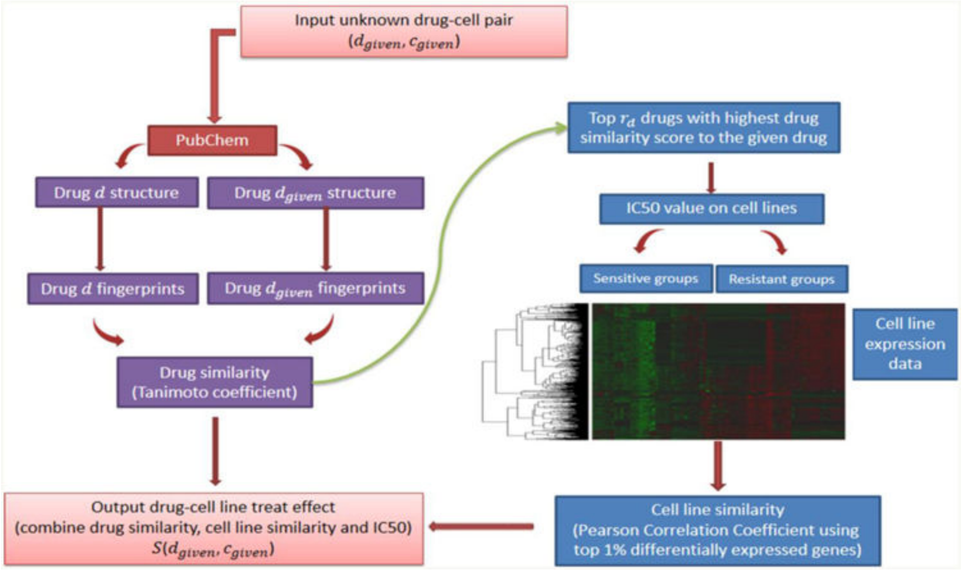
Input unknown drug-cell pair
(d_{given}, c_{given})
PubChem
Drug d structure
Drug d_{given} structure
Drug d fingerprints
Drug d_{given} fingerprints
Drug similarity
(Tanimoto coefficient)
Top r_d drugs with highest drug similarity score to the given drug
IC50 value on cell lines
Sensitive groups
Resistant groups
Cell line expression data
Cell line similarity
(Pearson Correlation Coefficient using top 1% differentially expressed genes)
Output drug-cell line treat effect
(combine drug similarity, cell line similarity and IC50)
$S(d_{given}, c_{given})$

Fig. 1.

An overview of the drug prediction algorithm.

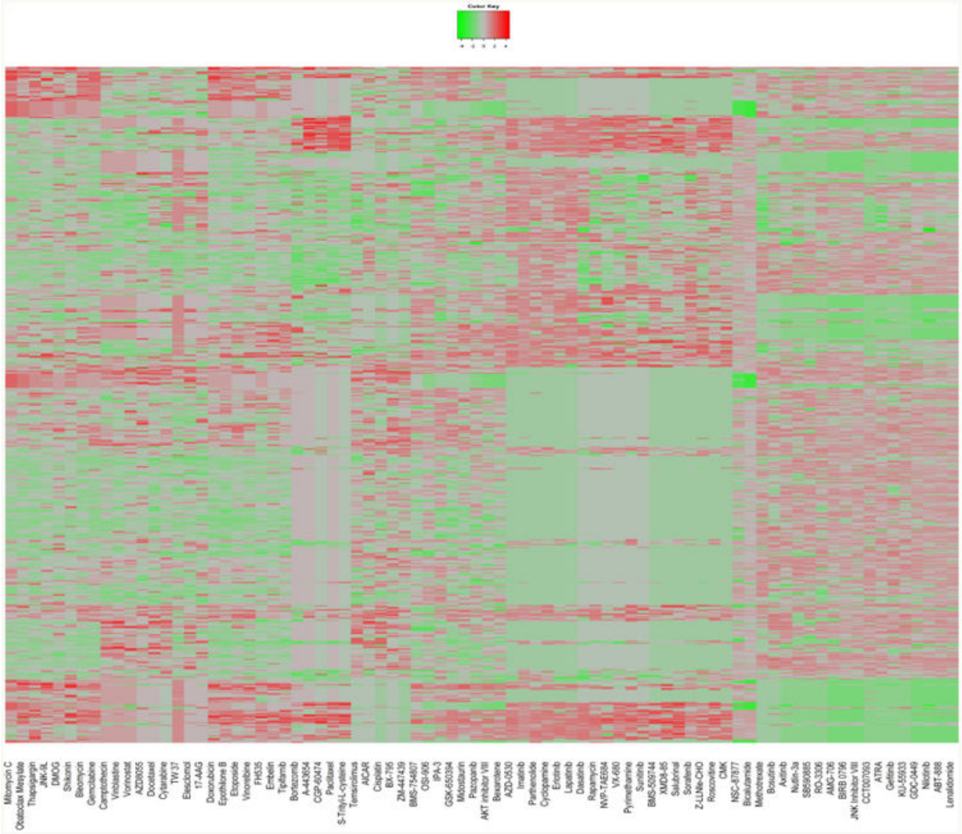

Fig. 2.

Heatmap for normalized IC50 values of 75 drugs (columns) on 624 cell lines (rows). Green means the most sensitive, red means the most resistant.

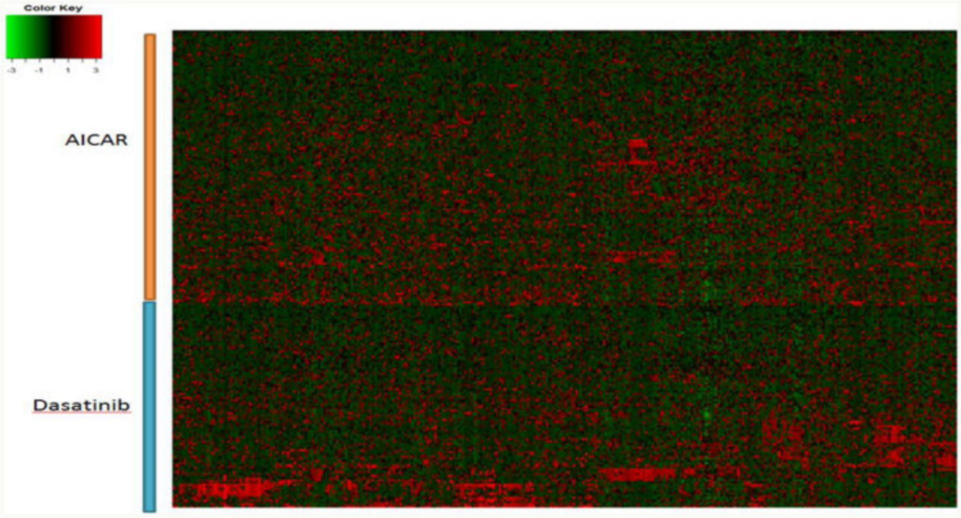

Fig. 3.

Heatmap for mRNA expression of gene signatures (rows) of two drugs (AICAR and Dasatinib) on 624 cell lines (columns).

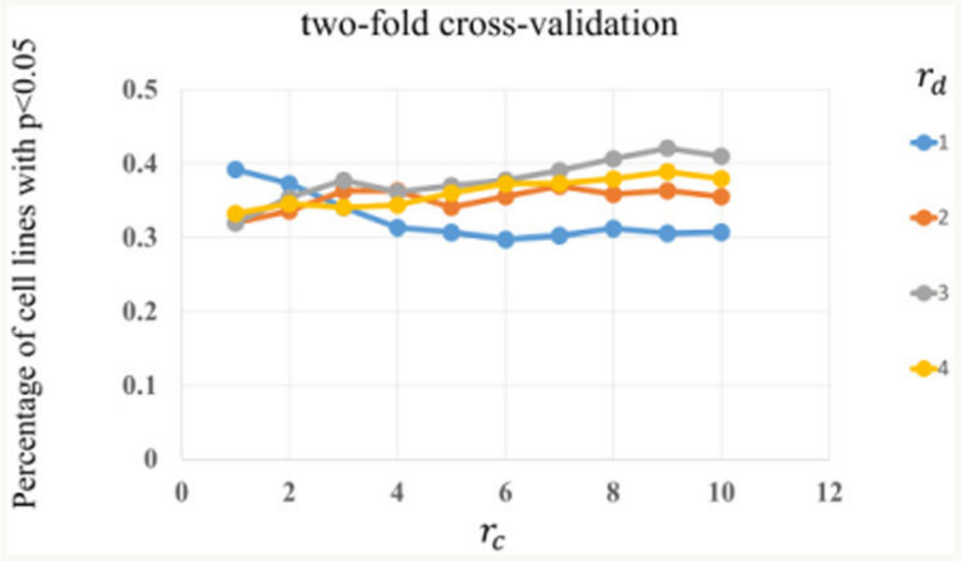

Fig. 4.
Two-fold cross-validation results for different r_d and r_c values.

Table. 1

Prediction results using CCLE as testing set.

***P*_value**		**r_d**			
		1	**2**	**3**	**4**
r_c	1	0.013	0.2652	0.098	0.0054
	2	0.006	0.0226	0.1114	0.0932
	3	0.0006	0.0054	<4e-04	0.8668
	4	0.0024	0.0006	<4e-04	0.8668
	5	0.0008	0.043	0.0008	0.8668
	6	0.0008	0.0698	<4e-04	0.0384
	7	0.0028	0.0494	<4e-04	0.005
	8	<4e-04	0.0068	<4e-04	0.0384
	9	<4e-04	0.0068	<4e-04	0.0384
	10	<4e-04	0.0116	0.1858	0.0384