

RESEARCH ARTICLE

Tissue-guided LASSO for prediction of clinical drug response using preclinical samples

Edward W. Huang¹, Ameya Bhope², Jing Lim¹, Saurabh Sinha^{1,3,4*}, Amin Emad^{2*}

1 Department of Computer Science, University of Illinois at Urbana-Champaign, Illinois, United States of America, **2** Department of Electrical and Computer Engineering, McGill University, Canada, **3** Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Illinois, United States of America, **4** Cancer Center at Illinois, University of Illinois at Urbana-Champaign, Illinois, United States of America

☞ These authors contributed equally to this work.

* sinhas@illinois.edu (SS); amin.emad@mcgill.ca (AE)



OPEN ACCESS

Citation: Huang EW, Bhope A, Lim J, Sinha S, Emad A (2020) Tissue-guided LASSO for prediction of clinical drug response using preclinical samples. *PLoS Comput Biol* 16(1): e1007607. <https://doi.org/10.1371/journal.pcbi.1007607>

Editor: Avner Schlessinger, Icahn School of Medicine at Mount Sinai, UNITED STATES

Received: August 6, 2019

Accepted: December 15, 2019

Published: January 22, 2020

Copyright: © 2020 Huang et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the manuscript, its Supporting Information files and on <https://github.com/emad2/TG-LASSO>.

Funding: This work was supported by McGill's Faculty of Engineering (AE), Natural Sciences and Engineering Research Council of Canada (NSERC) grant RGPIN-2019-04460 (AE), McGill Initiative in Computational Medicine (MiCM) and McGill Interdisciplinary Initiative in Infection and Immunity (MI4) (AE), and the research grant 1U54GM114838 awarded by NIGMS through

Abstract

Prediction of clinical drug response (CDR) of cancer patients, based on their clinical and molecular profiles obtained prior to administration of the drug, can play a significant role in individualized medicine. Machine learning models have the potential to address this issue but training them requires data from a large number of patients treated with each drug, limiting their feasibility. While large databases of drug response and molecular profiles of preclinical *in-vitro* cancer cell lines (CCLs) exist for many drugs, it is unclear whether preclinical samples can be used to predict CDR of real patients. We designed a systematic approach to evaluate how well different algorithms, trained on gene expression and drug response of CCLs, can predict CDR of patients. Using data from two large databases, we evaluated various linear and non-linear algorithms, some of which utilized information on gene interactions. Then, we developed a new algorithm called TG-LASSO that explicitly integrates information on samples' tissue of origin with gene expression profiles to improve prediction performance. Our results showed that regularized regression methods provide better prediction performance. However, including the network information or common methods of including information on the tissue of origin did not improve the results. On the other hand, TG-LASSO improved the predictions and distinguished resistant and sensitive patients for 7 out of 13 drugs. Additionally, TG-LASSO identified genes associated with the drug response, including known targets and pathways involved in the drugs' mechanism of action. Moreover, genes identified by TG-LASSO for multiple drugs in a tissue were associated with patient survival. In summary, our analysis suggests that preclinical samples can be used to predict CDR of patients and identify biomarkers of drug sensitivity and survival.

Author summary

Cancer is among the leading causes of death globally and prediction of the drug response of patients to different treatments based on their clinical and molecular profiles can enable individualized cancer medicine. Machine learning algorithms have the potential to play a

funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (SS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

significant role in this task; however, these algorithms are designed based on the premise that a large number of labeled training samples are available, and these samples are accurate representations of the profiles of real tumors. However, due to ethical and technical reasons, it is not possible to screen humans for many drugs, significantly limiting the size of training data. To overcome this data scarcity problem, machine learning models can be trained using large databases of preclinical samples (e.g. cancer cell line cultures). However, due to the major differences between preclinical samples and real tumors, it is unclear how accurately such preclinical-to-clinical computational models can predict the clinical drug response of cancer patients. Here, first we systematically evaluate a variety of different linear and nonlinear machine learning algorithms for this particular task using two large databases of preclinical (GDSC) and tumor samples (TCGA). Then, we present a novel method called TG-LASSO that utilizes a new approach for explicitly incorporating the tissue of origin of samples in the prediction task. Our results show that TG-LASSO outperforms all other algorithms and can distinguish resistant and sensitive patients for the majority of the tested drugs. Follow-up analysis reveal that this method can also identify biomarkers of drug sensitivity in each cancer type.

Introduction

Cancer is one of the leading causes of death globally and is expected to be the most important obstacle in increasing the life expectancy in the 21st century [1]. Individualized cancer medicine has the potential to revolutionize patient prognosis; however, two major challenges in this area include the prediction of the individual responses to different treatments and the identification of molecular biomarkers of drug sensitivity. While factors such as cancer type or its symptoms have been traditionally used to identify the treatment [2], the development of high throughput sequencing technologies [3] and sophisticated machine learning (ML) approaches present the possibility of individualizing treatment based on molecular ‘omics’ profiles of patients’ tumors [4]. However, due to the technical and ethical challenges of screening individuals against many drugs [5], such models are either trained for only a handful of drugs [6] or are trained using preclinical samples such as 2D cancer cell line cultures (CCLs) [7–10]. In spite of the success of these methods in predicting the drug response of left-out *preclinical* samples using models trained on *preclinical* samples, they have had limited success in predicting the CDR of real patients [9, 11], with some exceptions [12–14].

Various preclinical models of cancer have been developed to enable the study of cancer and its treatment in the laboratory. CCLs, which are 2D cell cultures developed from tumor samples, are one of the least expensive and most studied of these models. Recently, several large-scale studies have cataloged the molecular profiles of thousands of CCLs and their response to hundreds of drugs [15–17]. Although various computational models have been developed to predict the CCLs’ drug response using their molecular profiles [7–9], these models have shown limited success in predicting CDR in real patients. In spite of sporadic successes for a handful of drugs [12, 13], the current belief remains that developing an accurate computational ‘pre-clinical-to-clinical’ model is extremely difficult if not impossible [5]. Our goal in this study was to perform an unbiased systematic evaluation on a panel of drugs to determine 1) whether regression models trained on *in vitro* preclinical samples can predict the CDR of real patients for each drug and 2) what type of side information (e.g. interaction of the genes, the tissue of origin of samples) might improve the CDR prediction.

To this end, we first formed a computational framework to systematically evaluate the prediction accuracy of different computational methods. We obtained preclinical training samples from the Genomics of Drug Sensitivity in Cancer (GDSC) database [16] and obtained molecular profiles of tumor samples from The Cancer Genome Atlas (TCGA) [18]. We focused on drugs that were shared between these two datasets and utilized the gene expression profiles of samples to predict the drug response, since previous studies have demonstrated gene expression to be most informative for this task [7]. Our analysis showed that regularized linear regression models provide the best performance among various algorithms. In addition, we included prior information on the relationship among genes (in the form of gene interaction networks) using several algorithms; however, this prior information did not improve the prediction.

Next, we developed a novel approach called Tissue-Guided LASSO (TG-LASSO) to explicitly include information on the tissue of origin of samples in the regularized regression model. This method outperformed all other approaches evaluated. Using this method, we showed that the CDR of cancer patients can be predicted using preclinical CCL training samples, for the majority of drugs. More specifically, out of 12 drugs, TG-LASSO separated resistant patients from sensitive patients for 7 drugs. In addition, for each tissue type and drug, TG-LASSO identified a small set of genes that may be used as tissue-specific biomarkers of drug response for each drug. We showed that genes selected by TG-LASSO for prediction of drug response are informative of patient survival when used as a gene signature, and also provide pathway-level insights into mechanisms of drug action. These results emphasize the clinical relevance of molecular profiles of preclinical samples cataloged in large-scale databases and demonstrate the importance of properly including information on the lineage of samples in follow-up analyses.

Results

Prediction of clinical drug response of cancer patients using *in vitro* experiments on preclinical cancer cell lines

In this study, our first goal was to determine whether commonly used machine learning algorithms are capable of predicting the clinical drug response (CDR) in cancer patients using computational models trained only on cancer cell lines' (CCLs) basal gene expression profiles (i.e. before administration of the drug) and their drug response. For this purpose, we identified 23 drugs (Supplementary S1 Table) that were administered to patients of The Cancer Genome Atlas (TCGA) [18] and were also present in the Genomics of Drug Sensitivity in Cancer (GDSC) [16] database. We obtained the gene expression profiles of 531 primary tumor samples of TCGA patients (17 different cancer types) who were administered any of these drugs from the Genomic Data Commons [19] (see [Methods](#) and Supplementary S1 Table). We obtained the carefully collected and curated information on clinical drug response (CDR) of these patients from [6]. Similarly, we obtained the gene expression profiles and the logarithm of half-maximal inhibitory concentration ($\log(\text{IC}_{50})$) of 979 cancer cell lines (of 55 different tissues) from GDSC (see Supplementary S1 Table for the number of cell lines from each tissue).

We formed a computational framework to systematically evaluate the prediction capability of different algorithms (Fig 1). In this framework, we first normalize the data and remove batch effects to ensure that the gene expression profiles from these two datasets are comparable ([Methods](#)). This is particularly important since GDSC contains microarray gene expression values, while TCGA contains RNA-seq data. We used ComBat [20] for batch effect removal, which has been previously used to successfully remove the batch effect between RNA-seq and

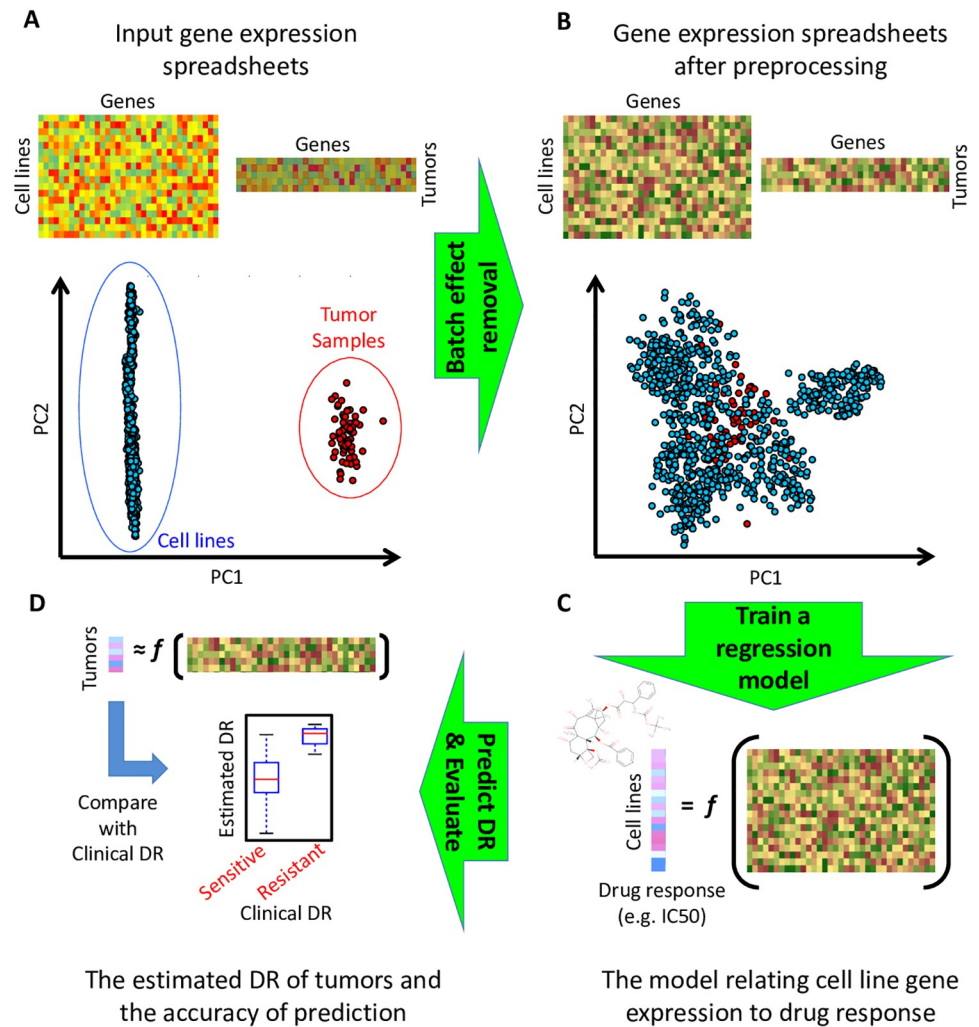


Fig 1. The pipeline used for prediction of clinical drug response of cancer patients using computational models trained on gene expression and drug response of preclinical cell line samples. The input gene expression data (A) corresponding to cancer cell lines (training set) and patients' tumors (test set) are first homogenized and their batch-effect is removed. The homogenized gene expression training data (B) and the cell lines' log (IC50) values are used to train a regression model (C). The trained model is applied to gene expression profiles of patients' tumors to predict their log (IC50) values, which are then used to evaluate the prediction performance (D).

<https://doi.org/10.1371/journal.pcbi.1007607.g001>

microarray data [21] (see Supplementary S1 Fig for the distribution of samples before and after batch effect removal). Note that the CDR of the test samples were not used during this step. Next, we trained a regression model to relate the gene expression profiles of CCLs to their log (IC50) values for a specific drug. Given this model, we then estimated log (IC50) values for different patient tumors using their gene expression profiles. Finally, we compared the estimated log (IC50) values to the true CDR of the tumors of patients treated with the same drug to determine the accuracy of prediction.

We used a one-sided nonparametric Mann Whitney U test to determine whether the estimated log(IC50) values of resistant tumors (those with CDR of 'clinical progressive disease' or 'stable disease') are significantly larger than sensitive tumors (those with CDR of 'partial response' or 'complete response'). One should note that due to the difference in the type of measured drug response in the training set (continuous-valued log(IC50)) and the test set

Table 1. The performance of different algorithms in predicting the CDR of patients using models trained on preclinical CCL samples. The second column shows the properties of the algorithm (linear versus nonlinear; single task versus multi-task learning). The third column shows the number of drugs for which a statistically significant discrimination between resistant and sensitive patients was obtained (one-sided Mann Whitney U test). The fourth column shows the total number of drugs included in the evaluation, and the fifth column shows the combined p-value (using Fisher’s method) for all the drugs in the analysis.

Algorithm	Properties	Drugs with P<0.05	Drugs	Combined P (Fisher)
LASSO	Linear, Single task	5	12	5.21E-09
ElasticNet	Linear, Single task	5	12	1.18E-08
MTL-LASSO	Linear, Multi-task	5	12	3.64E-06
Ridge	Linear, Single task	4	12	1.75E-05
MTL-ElasticNet	Linear, Multi-task	3	12	4.83E-06
SVR (Linear Kernel)	Linear, Single task	3	12	1.10E-05
SVR (Polynomial Kernel)	Nonlinear, Single task	3	12	1.82E-05
SVR (RBF kernel)	Nonlinear, Single task	3	12	2.92E-05
K-Nearest Neighbor	Nonlinear, Single task	3	12	8.26E-05
Multi-Layer Perceptron	Nonlinear, Single task	2	12	4.86E-02
Gelhar, et al. (2017)	Linear, Single task	1	11	1.85E-02
Random Forest	Nonlinear, Single task	1	12	0.19

<https://doi.org/10.1371/journal.pcbi.1007607.t001>

(categorical CDR), such an approach is necessary and other measures of performance such as concordance index or mean squared error are not suitable. In this evaluation, we only used 12 drugs that had at least 2 tumor samples in each category of resistant or sensitive and had at least 8 total samples with known CDR. Table 1 shows a summary of the performance of different methods. In this table, we used the combined p-value of all 12 drugs (using Fisher’s method to combine p-values) as a measure to summarize the results of different methods. Table 2 and Supplementary S2 Table contain the detailed performance of LASSO and all other methods, respectively, for prediction of the CDR of each drug. We focused on these methods as they have been previously used for this task (but for fewer drugs and using other datasets), with different degrees of success [12, 22, 23]. Recently, [5] reported a computational model based on ridge regression to predict the CDR of TCGA patients using GDSC training samples. Table 1 also includes the performance of this method using our evaluation, based on the predictions reported in the original paper.

Table 2. The performance of LASSO algorithm in predicting the CDR of patients using models trained on preclinical CCL samples. The second column shows the p-value (one-sided Mann Whitney U test) for the predicted log (IC50) values of sensitive and resistant tumors. The third and fourth columns show the number of resistant and sensitive tumors used in the statistical test.

Drug	P-value (one-sided)	Num Resistant (PD or SD)	Num Sensitive (CR or PR)
bicalutamide	0.34	3	14
bleomycin	0.10	4	46
cisplatin	6.67E-05	25	111
docetaxel	0.98	17	55
doxorubicin	3.42E-03	7	54
etoposide	7.57E-04	10	71
gemcitabine	0.14	43	37
paclitaxel	0.62	28	74
sorafenib	0.19	13	2
tamoxifen	8.82E-03	4	14
temozolomide	9.08E-02	84	11
vinorelbine	2.10E-03	6	23

<https://doi.org/10.1371/journal.pcbi.1007607.t002>

Table 3. The contingency table of predictions using LASSO for cisplatin. The predicted log(IC50) values were labeled as resistant or sensitive based on the threshold that obtained the highest oddsratio [24] (oddsratio = 10.5, $p < 0.001$).

	Predicted Resistant	Predicted Sensitive	Total
True Resistant	23	2	25
True Sensitive	58	53	111
Total	81	55	136

<https://doi.org/10.1371/journal.pcbi.1007607.t003>

These results suggest several important points. First, consistent with the reports in [12, 22], we observed that regularized linear models resulted in the best performance, with LASSO performing the best. Second, we observed that although the method proposed in [5] is based on ridge regression, its performance is inferior to the ridge regression utilized in our study. This is likely due to the difference between the preprocessing and batch effect removal approaches used in the two studies. More specifically, instead of using ComBat to homogenize the gene expression data in the preclinical and clinical samples (as was done in our study and also in [12]), they simply standardized the mean of each gene to zero and its variance to one. This point emphasizes the importance of data preprocessing in pharmacogenomics studies. Third, we observed that for some drugs, the CDR could be predicted independent of the method, while for others, the choice of the method is important. For example, the CDR of cisplatin could be predicted ($p < 0.05$) using ten out of the twelve methods above (as an example Table 3 shows that 92% of resistant patients are correctly designated using LASSO, while keeping precision at ~30% and specificity at ~50%). As another example, the majority of the methods could not predict the CDR of taxane-based chemotherapy agents (docetaxel and paclitaxel). We suspect that this lack of success is due to the existence of various parameters that influence their response, such as tissue dependence or microenvironmental factors [25, 26], which may not be captured using these simple methods trained on gene expression profiles of CCLs. In fact, we later show that including the tissue of origin *explicitly* in the predicting model using TG-LASSO can significantly improve the drug response prediction for paclitaxel.

Including information on gene interactions does not improve CDR prediction

Various studies have suggested that including information on the interaction of the genes (and their protein products) or their involvement in different pathways can improve the accuracy of different bioinformatics tasks [27] such as gene prioritization [28], gene function prediction [29], gene set characterization [30], and tumor subtyping [31]. Since the genes (and their protein products) involved in a drug's mechanism of action biochemically and functionally interact with each other, we sought to determine whether including these interactions could improve CDR prediction. Since linear models provided the best performance in our preliminary analyses (Table 1), we focused on methods that incorporate gene interaction networks into linear predictive models. These included Generalized Elastic Net (GELnet) [32], Network-Induced Classification Kernels (NICK) [33], Sparse Group LASSO (SGL) [34], as well as a method based on LASSO combined with single sample gene set enrichment analysis (ssGSEA) [35] (see Methods). In all cases, we used four gene interaction networks: an experimentally verified network of protein-protein and genetic interactions, a gene co-expression network, and a network built based on text mining from the STRING database [36], as well as the HumanNet integrated network [37] (see Methods and Supplementary S1 Table for details). Table 4 summarizes the results and Supplementary S3 Table provides the details of the evaluations. These results suggest that in this application, incorporating network information using

Table 4. The performance of network-based algorithms in predicting the CDR of patients using models trained on preclinical CCL samples. The third column shows the number of drugs for which a statistically significant discrimination between resistant and sensitive patients was obtained (one-sided Mann Whitney U test). The fourth column shows the total number of drugs included in the evaluation, and the fifth column shows the combined p-value (using Fisher’s method) for all the drugs in the analysis. As a point of comparison, LASSO without the use of any network yielded p-value < 0.05 for five of 12 drugs, with combined p-value of 5.21E-09 (Table 1).

Network	Algorithm	Drugs with P<0.05	Drugs	Combined P (Fisher)
STRING PPI	NICK	5	12	1.79E-06
	GELnet	5	12	1.21E-04
	SGL	3	12	1.75E-05
	ssGSEA-LASSO	3	12	6.94E-06
STRING Co-Expression	NICK	5	12	2.40E-06
	GELnet	5	12	1.07E-04
	SGL	2	12	2.62E-04
	ssGSEA-LASSO	4	12	1.40E-2
STRING Text Mining	NICK	5	12	2.14E-06
	GELnet	5	12	1.23E-04
	SGL	3	12	7.95E-05
	ssGSEA-LASSO	3	12	2.57E-02
HumanNet Integrated Network	NICK	5	12	1.09E-06
	GELnet	5	12	1.21E-04
	SGL	3	12	7.69E-04
	ssGSEA-LASSO	4	12	6.06E-07

<https://doi.org/10.1371/journal.pcbi.1007607.t004>

these methods does not improve the prediction compared to linear models (e.g. LASSO) that do not incorporate such information (Table 1). This was in spite of the fact that some of these network-guided methods (e.g. NICK with STRING Text Mining) do improve the performance of within-dataset cross-validation (using only GDSC samples) compared to LASSO (see Supplementary Methods in S1 Text).

Incorporating the tissue of origin to improve CDR prediction

Up to this point, we only used the tissue of origin of the preclinical and clinical samples *implicitly* (through their gene expression profiles) by training a single model for a drug on all CCLs of different lineages, and then using this global model to predict the response of patients with different cancer types. However, due to the importance of the tissue of origin in the efficacy of anticancer drugs observed in various studies [38–40] and its inclusion as auxiliary information in some drug response prediction studies [9, 41, 42], we sought to determine whether *explicitly* including the tissue of origin would improve the prediction of CDR, and if so, the best method for this inclusion. For our analysis, we focused on variations of LASSO (without including gene interactions), which previously yielded the best performance among all the tested algorithms (Table 1). We matched the lineage of the CCLs with those of cancer patients, identifying 13 shared tissue types.

One of the most common methods of including the tissue of origin in regression analysis is introducing new binary features to each sample, representing whether the sample belongs to that tissue (‘1’) or not (‘0’) [15]. We included 13 such binary features in the analysis (‘method 1’). However, the prediction results of this approach were almost identical to the results of LASSO when not including any tissue information. This is not surprising, since in this application the number of one type of features (i.e. genes) is much larger than the number of the other type of features (i.e. tissue types). As a result, the predicted drug response values will be highly biased by the influence of gene expression data and the tissue of origin’s influence will be overlooked. As an alternative, we trained different LASSO models for each tissue type by

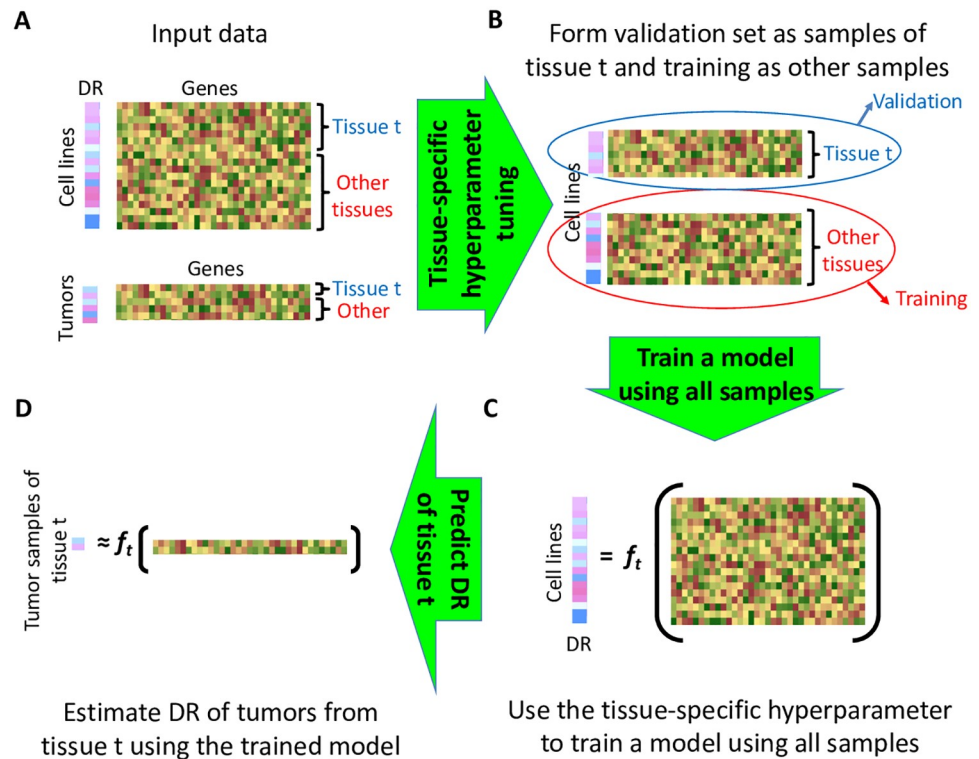


Fig 2. The TG-LASSO pipeline. The steps of TG-LASSO are depicted for one tissue type. These steps are repeated for each tissue type. To predict the drug response of tumors corresponding to tissue t, the cell lines of the same lineage are identified (A). These cell lines are used as the validation set, while cell lines of all other lineages are used as the training set for hyperparameter tuning (B). The identified hyperparameter is used to train a tissue-dependent model using all the CCLs (C). The trained model is used to predict the drug response of tumors from tissue t (D). Since the hyperparameter is tuned in a tissue-dependent manner (B and C), the models trained for each tissue type are distinct (C and D).

<https://doi.org/10.1371/journal.pcbi.1007607.g002>

restricting the training (CCL) and test (tumor) samples to those originating from the same tissue of interest (‘method 2’). For tumor samples without CCLs with matching tissue, we used all CCLs to train the model. This method resulted in poor performance, with only one drug having a significant p-value and a combined p-value (Fisher’s method) of 0.16. The reason behind this poor performance is the small number of samples in training each model: due to the tissue-specificity condition imposed above, only a small fraction of the total samples are used in training each model, which results in poor generalizability of the models.

To overcome these issues, while explicitly incorporating information on the samples’ tissue of origin, we devised a new approach called Tissue-Guided LASSO (TG-LASSO). The idea behind this approach is to use *all* CCLs originating from different tissue types in training the LASSO model, but choose the hyperparameter of the LASSO model, α , in a tissue-dependent manner (Fig 2). This avoids the issues caused by the small number of training samples in Method 2, while adding a tissue-dependent aspect to the training of the model. Since α controls the number of features (i.e. genes) used by the LASSO model, this approach allows us to optimally select the number of predictive genes for each tissue type (see [Methods](#) for details) yet use all CCLs to train these tissue-dependent regression models.

This approach resulted in the best performance among all the methods tested, with 7 (out of 12) drugs showing significant discrimination between resistant and sensitive tumors ($p < 0.05$) and a combined p-value (Fisher’s method for all 12 drugs) of $2.25E-10$ (Fig 3, Table 5

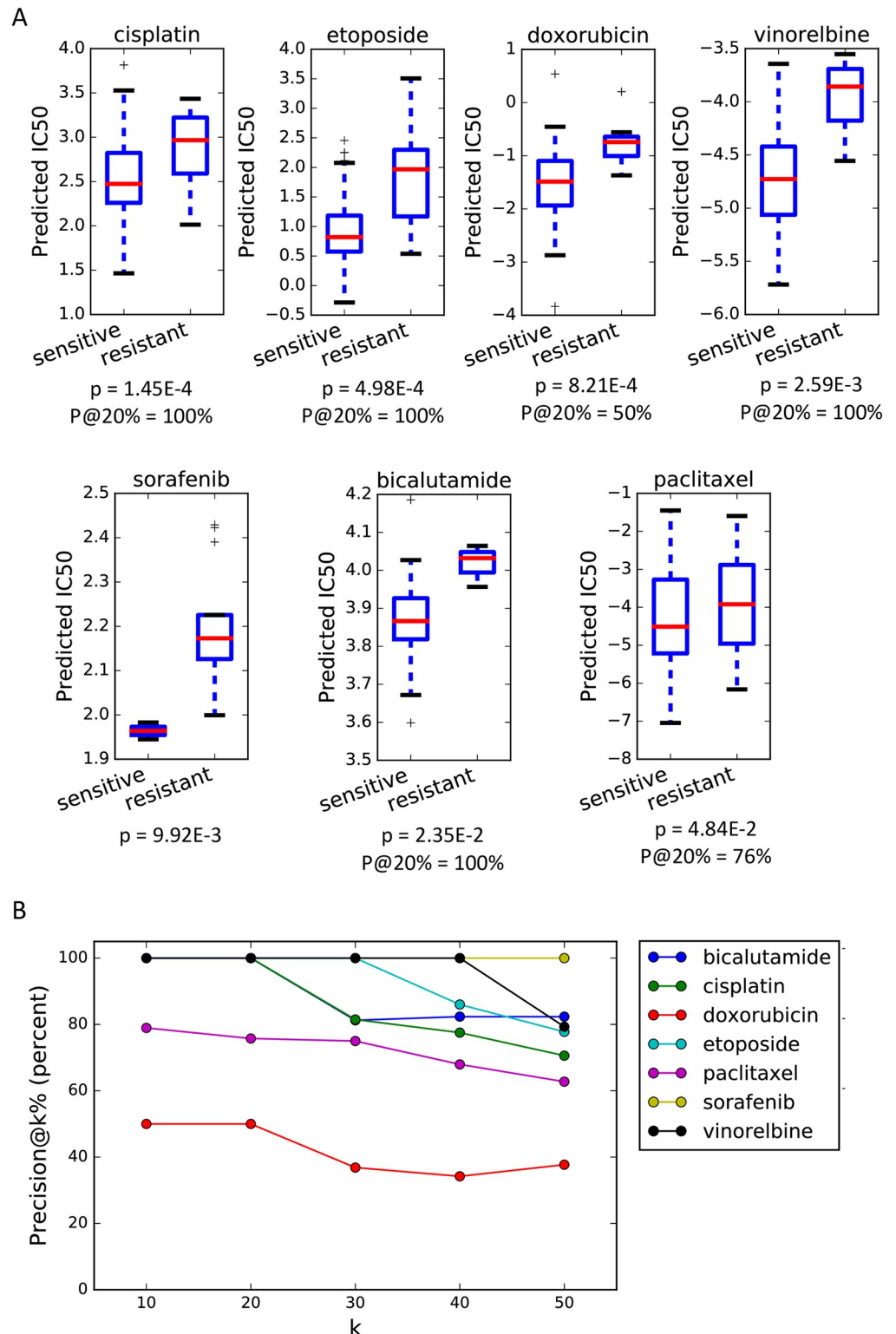


Fig 3. The drug response prediction performance of seven drugs for which TG-LASSO predictions separated sensitive patients from resistant ($p < 0.05$). A) The box plots reflect the distribution of estimated log (IC50) values using TG-LASSO for each group of resistant or sensitive patients. The p-values correspond to a one-sided Mann-Whitney U test. The Precision@20% (written as P@20%) is the precision of the method when samples with the predicted log (IC50) above 80th percentile of the training log(IC50) values are declared as resistant and those below the 20th percentile are declared as sensitive. B) The Precision@k% as a function of k.

<https://doi.org/10.1371/journal.pcbi.1007607.g003>

Table 5. The prediction performance of different approaches that incorporate information on tissue of origin in LASSO. The second column shows which subset of the training samples were used for training. The third column shows how tissue information was used. The fourth column shows the number of drugs for which a statistically significant discrimination between resistant and sensitive patients was obtained (one-sided Mann Whitney U test). The fifth column shows the total number of drugs included in the evaluation, and the sixth column shows the combined p-value (using Fisher’s method) for all the drugs in the analysis.

Algorithm	Training Samples	Tissue information	Drugs with P<0.05	Drugs	Combined P (Fisher)
TG-LASSO	All samples	Used during hyperparameter tuning	7	12	2.25E-10
Method 1	All samples	Used as new binary features	5	12	5.21E-09
Method 2	Only samples matching the test samples’ tissue	Used to identify relevant training samples	1	12	0.16

<https://doi.org/10.1371/journal.pcbi.1007607.t005>

and Supplementary S4 Table). These results not only show that including the tissue of origin can improve CDR prediction using preclinical samples, but also suggest that the method of utilizing this information has a significant influence on the performance.

To better assess the ability of TG-LASSO in predicting whether a drug should be administered to a patient or not, we defined a measure which we called Precision@k%. Intuitively, this measure represents the precision of the method when test samples with predicted log(IC50) above the (100–k)th percentile of the training samples’ log(IC50)s are labeled as resistant and those below the kth percentile are labeled as sensitive (see Methods for details). Five drugs (bicalutamide, cisplatin, etoposide, vinorelbine and temozolomide) had a Precision@20% equal to 100% and the average value of this measure for all drugs was equal to 80% (see Fig 3B and Supplementary S4 Table). These results suggest that the 20th and 80th percentiles of the training samples’ log (IC50) may be good thresholds for deciding whether a patient is sensitive or resistant to these drugs.

One interesting observation was that paclitaxel, the response of which could not be predicted accurately with the majority of methods reported in Table 1, showed a significant improvement in the response prediction with TG-LASSO (p = 0.048, one-sided Mann Whitney U test), suggesting a prominent role for the tissue of origin in its drug response. On the other hand, the CDR prediction of docetaxel did not improve (p = 0.99), even though docetaxel is also a taxane, like paclitaxel, and these two drugs have a statistically significant correlation in their CCL responses (Spearman rank correlation = 0.38, p = 1.7E-13). We suspected that this difference between the performance of TG-LASSO for docetaxel and paclitaxel is related to how well the CCL panel used for training represents the tumor samples of patients to whom these drugs were administered. To evaluate this, we calculated the similarity between the gene expression profiles of tumor samples to those of CCLs from the same tissue of origin for these drugs. This analysis showed a lower similarity between the docetaxel-administered tumors and CCLs (average cosine similarity = 0.07) compared to paclitaxel-administered tumors and CCLs (average cosine similarity = 0.11). These results provide evidence in favor of our hypothesis that the difference in the performance of TG-LASSO is related to how well the CCLs represent the profile of tumors to which these two drugs were administered.

Since some of the drugs used in our study were administered in combination with other drugs, we asked how well TG-LASSO predicts the CDR in such cases of treatment with drug combinations. For this purpose, we evaluated its CDR prediction for a drug only on patients for whom that drug was administered over a period overlapping their treatment with at least one other drug. We limited our analysis to 9 drugs with at least two samples (patients) in each group (sensitive and resistant) and with at least 8 samples in total. Supplementary S5 Table shows that, consistent with our previous results, TG-LASSO outperforms all other methods, capable of predicting the CDR of 6 (out of 9) drugs (p<0.05, one-sided Mann Whitney U test).

Next, we sought to evaluate the effect of batch-effect removal and preprocessing on the performance of TG-LASSO. For this purpose, we did not perform ComBat data homogenization or z-score normalization on the gene expression data. As expected, the performance of both TG-LASSO and LASSO deteriorated, with the former resulting in 4 drugs with $p < 0.05$ and the latter with only 3 (Supplementary S6 Table). In spite of this, TG-LASSO still performed better than LASSO, emphasizing the importance of including the tissue information.

Characterization of genes identified by TG-LASSO

During its training phase, TG-LASSO automatically selects a subset of genes to be used in the regression model by tuning the hyperparameter α introduced above. The number of genes selected in this manner depends on the drug and tissue type for which the model is trained to make response predictions and was found to range between 9 and 808 genes with a median of 174 genes. The genes identified by TG-LASSO included many direct targets of each drug. (For these analyses we used all 23 drugs shared between TCGA and GDSC and not just those with a large number of samples in TCGA). For example, EGFR, which is a direct target of both cetuximab and gefitinib [43], was selected by this algorithm when trained to predict response of these drugs in each of the 13 tissue types (Supplementary S7 Table). Similarly, FLT3, a target of the drugs sorafenib and sunitinib [43], was selected by TG-LASSO for predicting response to these drugs in 13 and 12 tissues, respectively. In addition to direct targets, many of the identified genes have been shown to be indirect targets of these drugs and to be involved in their mechanism of action. For example DNER, a gene identified by TG-LASSO for all tissue types for cisplatin (but was not identified by LASSO for this drug), has been shown to be significantly upregulated in response to this drug in NCI-H526 cell lines [44].

More importantly, the knockdown or overexpression of many of the identified genes has been shown to influence the sensitivity of cancer cells to these drugs. For example, the shRNA knockdown of CHI3L1, a gene identified for etoposide and cisplatin response in every tissue (but was not identified using LASSO for any of these drugs), has been shown to sensitize glioma cells to these two drugs, while its overexpression reduced their sensitivity [45]. As another example, the knockdown of SALL4 (identified in all tissues) in cancer cell lines has been shown to increase the sensitivity of lung cancer cells [46] and esophageal squamous cell carcinoma cells [47] to cisplatin. Supplementary S8 Table summarizes some of the evidence we curated from literature for the role of different genes identified by TG-LASSO in all tissue types for cisplatin (and whether they could have been identified using LASSO or not), as an illustration. These examples show the fact that the genes utilized by TG-LASSO in prediction of CDR of patients not only include targets of respective drugs, but also include genes whose expression has been experimentally shown to predict the sensitivity of these drugs: a property necessary for any predictive model of drug response. In addition, many of these genes could not be identified by simply using LASSO.

Genes identified for multiple drugs in a tissue are associated with patient survival

We hypothesized that genes that were identified by TG-LASSO as response predictors of many drugs in a single tissue (Supplementary S9 Table) may be able to predict the survival of patients who have cancer that originated from that tissue, as they may play a significant role in the development and progress of the disease. To test this, we obtained gene expression values of 4908 primary tumors from 10 different cancer types (corresponding to the tissue types in our study) from TCGA, requiring the data to include at least 170 patients and 20 incidents of deaths for each cancer type (Supplementary S10 Table). Then, we clustered the primary

tumors of each cancer type into two groups based on the expression of genes identified by TG-LASSO for more than 5 different drugs in the tissue corresponding to that cancer type. We used hierarchical clustering with cosine similarity. Kaplan-Meier survival analysis showed that this clustering approach could separate patients with poor survival from those with better survival (log-rank test, $p < 0.05$) for 6 out of the 10 cancer types (Fig 4, Supplementary S2 Fig, Supplementary S10 Table).

Next, we repeated the analysis above using genes identified by LASSO for more than 5 drugs as a benchmark (Supplementary S10 Table). The Kaplan-Meier survival analysis was significant for only 3 cancer types (log-rank test, $p < 0.05$), with the Lower Grade Glioma (LGG) clusters having the smallest p-value equal to $p = 0.013$ (compare to 6 significant cases using TG-LASSO and the log-rank p-value of the LGG clusters equal to $p = 7.61E-13$). These results provide further evidence in favor of the role of the genes identified by TG-LASSO in the progress of the corresponding cancer type.

Functional and pathway enrichment analysis of LGG related genes

Since Kaplan-Meier analysis of LGG clusters obtained using TG-LASSO genes resulted in the smallest p-value (log-rank test, $p = 7.61E-13$), we sought to further characterize the identified genes that resulted in this significant patient stratification using functional and pathway enrichment analysis. For this purpose, we used the KnowEnG's gene set characterization pipeline [27] and identified 20 GO terms and two pathways enriched ($FDR < 0.05$) in this gene set (Supplementary S11 Table).

Several of the most significantly enriched GO terms were related to extracellular matrix (ECM), which plays an important role in the infiltration of glioma cells into the brain [48, 49]. Another important GO term was neutrophil degranulation ($FDR = 2.1E-3$). Neutrophils are the most abundant type of white blood cells and the number of infiltrating neutrophils has been shown to be associated with the malignancy of glioma and its drug resistance [50]. In addition, it has been shown that in patients with glioblastoma, neutrophil degranulation is associated with peripheral cellular immunosuppression [51]. Another noteworthy GO term was integrin binding ($FDR = 0.037$). Integrins are transmembrane proteins that mediate cell adhesion, play an important role in promoting the invasiveness of glioma cells [52], and have been suggested as potential targets with diagnostic and prognostic value in glioma [53]. Several enriched GO terms were related to the activity of endopeptidases and collagen. It has been shown that the level of collagen in glioma patients is increased, and it also plays a key role in promoting the tumor progression [54]. Matrix metalloproteinases (MMPs) are one important class of endopeptidases that are responsible for regulating the turnover of collagens, and their expression and activity has been associated with the progression of human glioma [54, 55]. Finally, 'response to drug' was another enriched GO term, which reflects the relevance of the identified genes to the general mechanisms of drug response in a cell.

The enriched pathways included miRNA targets in ECM and membrane receptors ($FDR = 2.0E-3$) and Syndecan-1-mediated signaling ($FDR = 0.04$). Syndecan-1 is a cell surface heparan sulfate proteoglycan and its expression has been shown to be correlated with tumor cell differentiation in various cancers [56]. In addition, its knockdown has been shown to inhibit glioma cell proliferation and invasion and has been suggested as a therapeutic target for glioma [57]. These results support our expectation that the LGG-related gene set not only involves drug response related genes, but also includes those that play important roles in glioma and may act as diagnostic biomarkers or therapeutic targets.

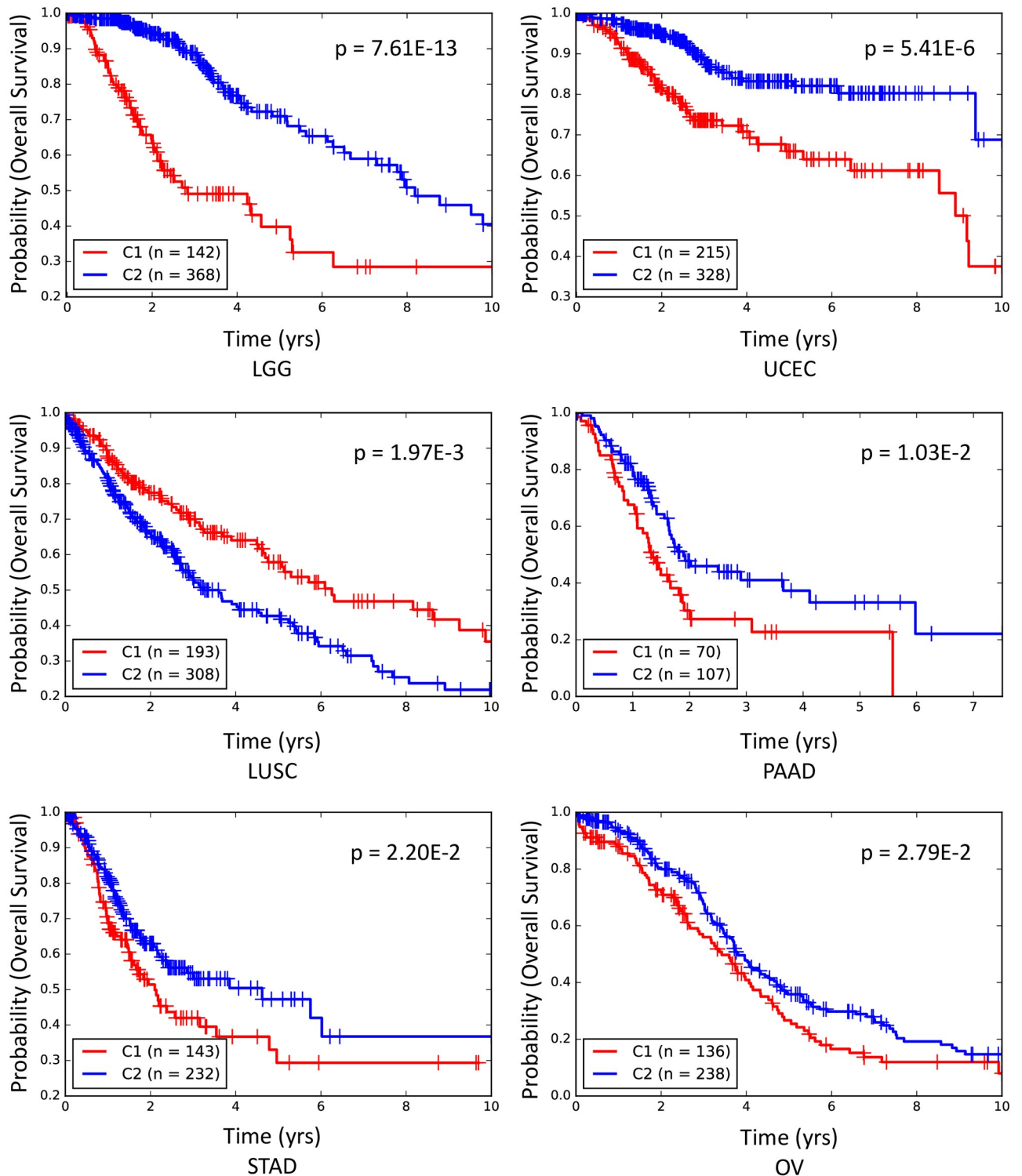


Fig 4. The Kaplan Meier survival analysis results for six cancer types. Patients were clustered based on the expression of genes that were identified by TG-LASSO for more than 5 drugs in the corresponding tissue. The p-value was calculated using a log-rank test.

<https://doi.org/10.1371/journal.pcbi.1007607.g004>

Discussion

Ideally, a predictive model of CDR should be trained on data obtained directly from patients. Similarly, identification of biomarkers of drug sensitivity has the most potential clinical impact when based on patient data. However, since in practice most patients only receive the ‘standard of care’ treatment based on their specific cancer type, CDR data is scarcely available for the newly approved drugs or drugs that have not yet passed the clinical trial, limiting our ability to decipher the mechanisms of drug sensitivity for these drugs. An alternative approach is to train ML models on preclinical samples (e.g. CCLs) to predict the CDR of patients, then use these predictions to discover novel biomarkers and druggable targets.

Recent large-scale studies that have cataloged the molecular profiles of thousands of CCLs and their response to hundreds of drugs [15–17] are great resources to achieve this goal. In this study, we adopted such an approach and systematically assessed a variety of linear and non-linear single-task and multi-task ML algorithms. Our analyses showed that the CDR of many drugs can be predicted using ML models (especially, regularized linear models) trained on CCLs. However, by evaluating a variety of methods that include auxiliary information (e.g. interaction of the genes, the tissue of origin, etc.), we observed that improving the performance beyond what is achievable using linear models is extremely difficult and requires careful modeling and novel computational techniques. It appears that the way by which auxiliary information is utilized has a large impact: for example, several methods that include the tissue of origin did not improve the results obtained by LASSO, and only TG-LASSO could improve the performance. Additionally, we showed that TG-LASSO identifies tissue-dependent gene sets for each drug that include various targets of the drug, genes involved in the drug’s mechanism of action, and genes whose under- or over-expression could sensitize cancer cells to the drug. Moreover, these sets include genes that are involved in cancer progression and are associated with patient survival. These results suggest that in addition to a superior drug response prediction performance, TG-LASSO can identify biomarkers of patient survival and drug sensitivity.

Another important factor that played an important role in the performance of the ML models was data homogenization and batch-effect removal. The performance of TG-LASSO and LASSO both deteriorated when we did not remove the existing batch-effect between the training dataset and the test dataset. In spite of this, TG-LASSO could distinguish between resistant and sensitive patients for four drugs, when applied to non-homogenized data. This suggests two approaches when dealing with scenarios in which new test samples arrive. The first approach is to simply use the model trained on non-homogenized preclinical samples and accept the worse performance. The alternative is to retrain the model every time a new test sample arrives. This allows for training and prediction on homogenized data, but significantly increases the computational cost. An alternative could be developing a new data homogenization and batch effect-removal method that only transforms the gene expression of the test samples (keeping the gene expression profiles of training samples unchanged) by mapping them to the subspace spanned by the training samples. However, the development of such a method is beyond the scope of this study.

We note that due to the major differences between CCLs and tumors (e.g. the greater heterogeneity of cells in a tumour compared to CCLs), obtaining more accurate results based on classical ML techniques may not be possible. The reason is that classical ML methods assume that the training samples and the test samples are drawn from the same or similar distributions. While batch-effect removal and other homogenization and normalization techniques help to alleviate this issue, more realistic preclinical models of cancer are necessary to significantly improve these results. Recent advances in developing human derived xenografts [58]

and 3D human organoids [59] may enable developing a more accurate predictive model of CDR in cancer. However, due to the current high cost of these models, a more practical approach is developing computational methods that explicitly model these differences. Such methods must go beyond utilizing bulk gene expression data and take advantage of multi-omics analysis of bulk and single-cell sequencing profiles of samples. Due to the rapid advances in these domains, we expect that large databases of single-cell multi-omics profiles of preclinical and clinical samples and their drug response will become available in the near future.

Methods

Datasets, preprocessing and batch effect removal

We obtained the gene expression profiles (FPKM values) of 531 primary tumor samples of TCGA patients who were administered any of the 23 drugs mentioned earlier. First, we removed genes that contained missing values. We also removed any gene that was not expressed (i.e. $FPKM < 1$) for more than 90% of the samples. Then, we performed a log-transformation and obtained $\log_2(FPKM + 0.1)$ values for each gene. The resulting gene expression matrix contained 19,437 genes and 531 samples. We obtained the CDR of these patients from the supplementary files of [6] (see the original paper for their approach in curating this data from TCGA). Similarly, we obtained the Robust Multi-array Average (RMA)-normalized basal gene expression profiles and the logarithm of half maximal inhibitory concentration, $\log(IC_{50})$, of 979 cancer cell lines from GDSC (Supplementary S1 Table) for 17,737 genes.

To homogenize the gene expression data from these two datasets, we first removed genes not present in both datasets as well as genes with low variability across all the samples (standard deviation < 0.1), resulting in a total of 13,942 shared genes. Then, we used ComBat [20] for batch effect removal to homogenize the gene expression data from TCGA (RNA-seq) and GDSC (microarray). This approach, which has been previously used to successfully homogenize these two data types [21], removed the batch effect present in the gene expression datasets (see Supplementary S1 Fig). For all follow-up analysis, we performed z-score normalization on each gene across all the samples to ensure a mean of zero and a standard deviation equal to one.

For the network-guided analyses, we downloaded four networks of gene interactions in humans from the KnowEnG's knowledgebase of genomic networks [27] (https://github.com/KnowEnG/KN_Fetcher/blob/master/Contents.md). The details of each network including the number of nodes and edges are provided in Supplementary S1 Table.

Machine learning regression models

The baseline models (Table 1) were all implemented using Scikit-learn [60] in Python and the hyperparameters were selected using cross validation (using only CCL samples from GDSC). The multi-task learning methods (MTL-LASSO and MTL-ElasticNet) require availability of training drug response values for all drugs and all samples. Since the training drug response matrix contains missing values, to obtain the results reported in Table 1 we used the 'average strategy' (average of the drug response values across cell lines of a single drug) to impute the missing values. We also tried a different approach (i.e. dropping cell lines with missing values), but since the results were much worse, we did not include them in Table 1.

For the network-based algorithms (Table 4), we used four networks summarized in Supplementary S1 Table. We used the normalized graph Laplacian of these networks to run GELnet [32]. This method forces neighboring genes in the graph to have similar weights in order to guide drug response prediction. Specifically, it defines a regularization penalty $R(w)$ for the

standard linear model.

$$R(w) = \lambda_1 \sum_j d_j |w_j| + \frac{\lambda_2}{2} w^T P w$$

where d and P are additional penalty weights for individual features and pairs of features, respectively. Our basic GELnet implementation sets $P = L$ and $d = 0$. Furthermore, we used Network-Induced Classification Kernels (NICK), a method closely related to GELnet. The NICK framework is actually a special case of the GELnet, with $P = (I + \beta L)$ for some $\beta \geq 0$ and $d = 0$. The parameter β provides a trade-off between graph-driven regularization and the traditional ridge regression penalty of the SVMs.

In addition to the above methods that utilize the graph Laplacian of each network in the regression algorithm, we used sparse group LASSO (SGL). This method takes a collection of pathways as input and induces sparsity at both the pathway and the gene level to generate the input. We performed community detection on each of the networks in Table 4 by maximizing the modularity using the Louvain heuristics [61] to identify gene sets to be used in the SGL algorithm. We then ran SGL by fitting a regularized generalized linear model with group memberships of genes as deemed by the community detection to predict drug response.

Finally, we developed a heuristic method based on ssGSEA [35] followed by LASSO. In this method, we used ssGSEA to assign a score to each sample for the enrichment of its gene expression profile in communities of each network, obtained earlier. These scores were then used as features to train a LASSO model for prediction of CDR.

Methods for including tissue of origin in CDR prediction

In the first approach (Method 1 in Table 5), we augmented the gene expression profile of each sample (both CCLs and tumors) with binary features corresponding to different tissues of origin shared between the TCGA and GDSC samples (a total of 13 features). For each sample its tissue of origin was assigned a value of '1', and other tissues were assigned a value of '0'. Then, the LASSO algorithm was used to train a drug response model on CCLs and predict the CDR of tumors.

In the second approach (Method 2 in Table 5), we trained different LASSO models for each drug-tissue pair (23 drugs and 13 tissue types). More specifically, to predict the CDR of drug d in a tumor of tissue t , we trained a LASSO regression model using the log (IC50) of drug d in only cell lines corresponding to tissue t (i.e. a subset of the training samples). For tumors originating in tissues without matching training CCLs, we used all the CCLs to train the model.

Prediction of CDR in cancer tumors using Tissue-guided LASSO

TG-LASSO is a method for predicting the CDR of tumors using the information in *all* training samples (originating from different tissue lineages), while incorporating information on the tissue of origin of the samples. By utilizing all the training samples, it overcomes the lack of generalizability stemming from limited number of CCLs from each tissue type, a major issue in Method 2 above. In addition, by incorporating the information on the tissue of origin of the samples in the training step, it improves the performance of tissue-naïve regression methods, such as those in Table 1.

During training, LASSO minimizes the objective function $\frac{1}{2n} \|y - Xw\|_2^2 + \alpha \|w\|_1$, where n is the number of training samples, y is the response vector of length n , X is an $n \times m$ feature matrix (m is the number of features), $\| \cdot \|_2$ denotes the L2 vector norm, $\| \cdot \|_1$ denotes the L1 vector norm, and α is the hyperparameter that determines the sparsity of the model (i.e. number

of features used in training). The hyperparameter tuning is usually achieved independent of the structure of the training samples (e.g. their tissue of origin), for example using random cross-validation or a regularization path. However, we and others [62] have shown that including the group structure of data in selecting the hyperparameter is important in assessing the generalizability of regression models. Motivated by these results, even though TG-LASSO utilizes the gene expression and the drug response of all CCLs in training, the hyperparameter α is selected in a tissue- and drug-specific manner, as explained below.

Let D be the set of all drugs and T be the set of all tissues in the test set (i.e. the TCGA dataset). To train a model to predict the CDR of tumor samples from tissue $t \in T$ to drug $d \in D$, we identify all the training CCLs corresponding to tissue t and use them as the validation set. In addition, we use all other CCLs as the training set. Then, the hyperparameter α is selected as the one that obtains the best accuracy on predicting the log(IC50) values of the samples of tissue t in the validation set. Designing the hyperparameter-tuning step such that the validation and the test sets have the same tissues of origin ensures that the value of α is selected so as to generalize well to the test set. The obtained value of α is then used with *all* CCLs (including those from tissue t) to fit a model minimizing the LASSO objective function. In the prediction step, this fitted model is then used with the gene expression of tumor samples from tissue t to predict their CDR.

Precision@k%

To further assess the performance of TG-LASSO, we defined a measure called Precision@k% (motivated by Precision@k in information retrieval). To define Precision@k%, we first used the log(IC50) values of the preclinical cell lines from GDSC to find the Kth percentile ($K \leq 50$) and the (100-K)th percentile of each drug (separately), denoted as t_K and t_{100-K} , respectively. Then, given the predicted log(IC50) values of the tumors and their annotation as ‘sensitive’ or ‘resistant’ (based on their known CDR), we defined

$$\text{Precision@k\%} = \frac{R_{100-K} + S_K}{N_{100-K} + N_K},$$

where R_{100-K} is the number of resistant tumors whose predicted log(IC50) is larger than t_{100-K} , S_K is the number of sensitive tumors whose predicted log(IC50) is smaller than t_K , N_{100-K} is the total number of tumors whose predicted log(IC50) is larger than t_{100-K} , and N_K is the total number of tumors whose predicted log(IC50) is smaller than t_K . Intuitively, this measure shows the precision of predicting the tumors with predicted log(IC50) values larger than t_{100-K} as resistant and those with predicted log(IC50) values smaller than t_K as sensitive. Note that due to this definition of Precision@k%, for some values of k, the denominator may be equal to 0 and the measure may not be defined.

Gene ontology and pathway enrichment analysis

We used the gene set characterization pipeline of KnowEnG analytical platform [27] for this analysis, which utilizes Fisher’s exact test to determine the significance of enrichments. We excluded GOs or pathways with too few genes, focusing only on those with more than 10 members. For the pathway analysis, we used the ‘Enrichr’ pathways [63] available on KnowEnG. All p-values were corrected for multiple hypothesis testing using Benjamini-Hochberg false discovery rate, available as part of the python module [64].

Software availability

An implementation of TG-LASSO in python, with appropriate documentation and input files, is available at: <https://github.com/emaad2/TG-LASSO>.

Supporting information

S1 Text. The supplementary methods, all the supplementary figures and their captions, and the caption of supplementary tables.

(DOCX)

S1 Table. The summary of data used in this study.

(XLSX)

S2 Table. The detailed drug response prediction performance of baseline methods for each drug.

(XLSX)

S3 Table. The detailed drug response prediction performance of network-based methods for each drug.

(XLSX)

S4 Table. The detailed drug response prediction performance of tissue-based methods for each drug.

(XLSX)

S5 Table. The detailed drug response prediction performance of various methods for drugs whose administration period overlapped with at least another drug (i.e. administered as a combination).

(XLSX)

S6 Table. The detailed drug response prediction performance of TG-LASSO and LASSO applied to non-homogenized data.

(XLSX)

S7 Table. The list of genes identified by TG-LASSO for each drug, ranked based on the number of tissues in which they were identified.

(XLSX)

S8 Table. The evidence curated from literature for genes identified by TG-LASSO in all tissues for cisplatin.

(XLSX)

S9 Table. The list of genes identified by TG-LASSO for each tissue, ranked based on the number of drugs for which they were identified.

(XLSX)

S10 Table. The results of Kaplan Meier survival analysis on TCGA samples using genes identified by TG-LASSO and LASSO.

(XLSX)

S11 Table. Gene ontology and pathway enrichment results for the top genes identified by TG-LASSO for brain, used in the survival analysis of LGG samples.

(XLSX)

S1 Fig. The distribution of preclinical as well as tumor samples before (A) and after (B) batch effect removal, depicted using principal component analysis (PCA) of their gene expression.

(TIF)

S2 Fig. The Kaplan Meier survival analysis corresponding to BRCA, COAD, LUAD, and LIHC samples from TCGA. Patients were clustered into two groups using the expression of

genes identified by TG-LASSO for more than 5 drugs in each tissue type. (TIF)

Author Contributions

Conceptualization: Saurabh Sinha, Amin Emad.

Formal analysis: Edward W. Huang, Ameya Bhope, Jing Lim, Amin Emad.

Funding acquisition: Saurabh Sinha, Amin Emad.

Investigation: Amin Emad.

Methodology: Edward W. Huang, Ameya Bhope, Saurabh Sinha, Amin Emad.

Software: Edward W. Huang, Ameya Bhope, Jing Lim, Amin Emad.

Supervision: Saurabh Sinha, Amin Emad.

Visualization: Amin Emad.

Writing – original draft: Edward W. Huang, Ameya Bhope, Jing Lim, Saurabh Sinha, Amin Emad.

Writing – review & editing: Saurabh Sinha, Amin Emad.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2018; 68(6):394–424. Epub 2018/09/13. <https://doi.org/10.3322/caac.21492> PMID: 30207593
2. Verma M. Personalized medicine and cancer. *J Pers Med.* 2012; 2(1):1–14. <https://doi.org/10.3390/jpm2010001> PMID: 25562699
3. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. *Molecular cell.* 2015; 58(4):586–97. <https://doi.org/10.1016/j.molcel.2015.05.004> PMID: 26000844
4. Deo RC. Machine learning in medicine. *Circulation.* 2015; 132(20):1920–30. <https://doi.org/10.1161/CIRCULATIONAHA.115.001593> PMID: 26572668
5. Geeleher P, Zhang Z, Wang F, Gruener RF, Nath A, Morrison G, et al. Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. *Genome Res.* 2017; 27(10):1743–51. <https://doi.org/10.1101/gr.221077.117> PMID: 28847918
6. Ding Z, Zu S, Gu J. Evaluating the molecule-based prediction of clinical drug responses in cancer. *Bioinformatics.* 2016; 32(19):2891–5. <https://doi.org/10.1093/bioinformatics/btw344> PMID: 27354694
7. Costello JC, Heiser LM, Georgii E, Gonen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol.* 2014; 32(12):1202–12. <https://doi.org/10.1038/nbt.2877> PMID: 24880487
8. Jiang P, Sellers WR, Liu XS. Big Data Approaches for Modeling Response and Resistance to Cancer Drugs. *Annual Review of Biomedical Data Science.* 2018; 1(1):1–27.
9. Ali M, Aittokallio T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys Rev.* 2018.
10. Qian C, Sidiropoulos ND, Amiridi M, Emad A, editors. From Gene Expression to Drug Response: A Collaborative Filtering Approach. ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019: IEEE.
11. Gillet JP, Varma S, Gottesman MM. The clinical relevance of cancer cell lines. *J Natl Cancer Inst.* 2013; 105(7):452–8. <https://doi.org/10.1093/jnci/djt007> PMID: 23434901
12. 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. <https://doi.org/10.1186/gb-2014-15-3-r47> PMID: 24580837
13. Falgreen S, Dybkær K, Young KH, Xu-Monette ZY, El-Galaly TC, Laursen MB, et al. Predicting response to multidrug regimens in cancer patients using cell line experiments and regularised regression models. *BMC cancer.* 2015; 15(1):235.

14. Huang HH, Dai JG, Liang Y. Clinical Drug Response Prediction by Using a Lq Penalized Network-Constrained Logistic Regression Method. *Cell Physiol Biochem*. 2018; 51(5):2073–84. Epub 2018/12/07. <https://doi.org/10.1159/000495826> PMID: 30522095
15. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012; 483(7391):603–7. <https://doi.org/10.1038/nature11003> PMID: 22460905
16. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*. 2013; 41(Database issue):D955–61. <https://doi.org/10.1093/nar/gks1111> PMID: 23180760
17. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol*. 2016; 12(2):109–16. <https://doi.org/10.1038/nchembio.1986> PMID: 26656090
18. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013; 45(10):1113–20. <https://doi.org/10.1038/ng.2764> PMID: 24071849
19. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med*. 2016; 375(12):1109–12. <https://doi.org/10.1056/NEJMp1607591> PMID: 27653561
20. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007; 8(1):118–27. <https://doi.org/10.1093/biostatistics/kxj037> PMID: 16632515
21. Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol*. 2016; 17(1):174. <https://doi.org/10.1186/s13059-016-1028-7> PMID: 27549193
22. Falgreen S, Dybkaer K, Young KH, Xu-Monette ZY, El-Galaly TC, Laursen MB, et al. Predicting response to multidrug regimens in cancer patients using cell line experiments and regularised regression models. *BMC Cancer*. 2015; 15:235. <https://doi.org/10.1186/s12885-015-1237-6> PMID: 25881228
23. Azuaje F. Computational models for predicting drug responses in cancer research. *Brief Bioinform*. 2017; 18(5):820–9. <https://doi.org/10.1093/bib/bbw065> PMID: 27444372
24. Szumilas M. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry*. 2010; 19(3):227–9. Epub 2010/09/16. PMID: 20842279
25. Magadoux L, Isambert N, Plenchette S, Jeannin JF, Laurens V. Emerging targets to monitor and overcome docetaxel resistance in castration resistant prostate cancer (review). *Int J Oncol*. 2014; 45(3):919–28. <https://doi.org/10.3892/ijo.2014.2517> PMID: 24969394
26. Son B, Lee S, Youn H, Kim E, Kim W, Youn B. The role of tumor microenvironment in therapeutic resistance. *Oncotarget*. 2017; 8(3):3933–45. <https://doi.org/10.18632/oncotarget.13907> PMID: 27965469
27. Blatti C, Emad A, Berry MJ, Gatzke L, Epstein M, Lanier D, et al. Knowledge-guided analysis of 'omics' data using the KnowEnG cloud platform. to appear in *PLOS Biology*, bioRxiv. 2019:642124.
28. Emad A, Cairns J, Kalari KR, Wang L, Sinha S. Knowledge-guided gene prioritization reveals new insights into the mechanisms of chemoresistance. *Genome Biol*. 2017; 18(1):153. <https://doi.org/10.1186/s13059-017-1282-3> PMID: 28800781
29. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*. 2010; 38(Web Server issue):W214–20. <https://doi.org/10.1093/nar/gkq537> PMID: 20576703
30. Blatti C, Sinha S. Characterizing gene sets using discriminative random walks with restart on heterogeneous biological networks. *Bioinformatics*. 2016; 32(14):2167–75. <https://doi.org/10.1093/bioinformatics/btw151> PMID: 27153592
31. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. *Nat Methods*. 2013; 10(11):1108–15. <https://doi.org/10.1038/nmeth.2651> PMID: 24037242
32. Sokolov A, Carlin DE, Paull EO, Baertsch R, Stuart JM. Pathway-Based Genomics Prediction using Generalized Elastic Net. *PLoS Comput Biol*. 2016; 12(3):e1004790. <https://doi.org/10.1371/journal.pcbi.1004790> PMID: 26960204
33. Lavi O, Dror G, Shamir R. Network-induced classification kernels for gene expression profile analysis. *J Comput Biol*. 2012; 19(6):694–709. <https://doi.org/10.1089/cmb.2012.0065> PMID: 22697242
34. Silver M, Chen P, Li R, Cheng CY, Wong TY, Tai ES, et al. Pathways-driven sparse regression identifies pathways and genes associated with high-density lipoprotein cholesterol in two Asian cohorts. *PLoS Genet*. 2013; 9(11):e1003939. <https://doi.org/10.1371/journal.pgen.1003939> PMID: 24278029

35. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*. 2009; 462(7269):108–12. <https://doi.org/10.1038/nature08460> PMID: 19847166
36. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. 2015; 43(Database issue):D447–52. <https://doi.org/10.1093/nar/gku1003> PMID: 25352553
37. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res*. 2011; 21(7):1109–21. Epub 2011/05/04. <https://doi.org/10.1101/gr.118992.110> PMID: 21536720
38. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*. 2016; 166(3):740–54. <https://doi.org/10.1016/j.cell.2016.06.017> PMID: 27397505
39. Schneider G, Schmidt-Suppran M, Rad R, Saur D. Tissue-specific tumorigenesis: context matters. *Nat Rev Cancer*. 2017; 17(4):239–53. <https://doi.org/10.1038/nrc.2017.5> PMID: 28256574
40. Yao F, Madani Tonekaboni SA, Safikhani Z, Smirnov P, El-Hachem N, Freeman M, et al. Tissue specificity of in vitro drug sensitivity. *J Am Med Inform Assoc*. 2018; 25(2):158–66. Epub 2017/10/11. <https://doi.org/10.1093/jamia/ocx062> PMID: 29016819
41. Turki T, Wei Z, Wang JTL. A transfer learning approach via procrustes analysis and mean shift for cancer drug sensitivity prediction. *J Bioinform Comput Biol*. 2018; 16(3):1840014. Epub 2018/06/28. <https://doi.org/10.1142/S0219720018400140> PMID: 29945499
42. Ding MQ, Chen L, Cooper GF, Young JD, Lu X. Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics. *Mol Cancer Res*. 2018; 16(2):269–78. Epub 2017/11/15. <https://doi.org/10.1158/1541-7786.MCR-17-0378> PMID: 29133589
43. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*. 2018; 46(D1):D1074–D82. <https://doi.org/10.1093/nar/gkx1037> PMID: 29126136
44. Olszewski U, Ulsperger E, Geissler K, Hamilton G. Comparison of the effects of the oral anticancer platinum(IV) complexes oxoplatin and metabolite cis-diammine-tetrachlorido-platinum(IV) on global gene expression of NCI-H526 cells. *J Exp Pharmacol*. 2011; 3:43–50. <https://doi.org/10.2147/JEP.S13630> PMID: 27186109
45. Ku BM, Lee YK, Ryu J, Jeong JY, Choi J, Eun KM, et al. CHI3L1 (YKL-40) is expressed in human gliomas and regulates the invasion, growth and survival of glioma cells. *Int J Cancer*. 2011; 128(6):1316–26. <https://doi.org/10.1002/ijc.25466> PMID: 20506295
46. Yanagihara N, Kobayashi D, Kuribayashi K, Tanaka M, Hasegawa T, Watanabe N. Significance of SALL4 as a drugresistant factor in lung cancer. *Int J Oncol*. 2015; 46(4):1527–34. <https://doi.org/10.3892/ijo.2015.2866> PMID: 25646965
47. He J, Zhou M, Chen X, Yue D, Yang L, Qin G, et al. Inhibition of SALL4 reduces tumorigenicity involving epithelial-mesenchymal transition via Wnt/beta-catenin pathway in esophageal squamous cell carcinoma. *J Exp Clin Cancer Res*. 2016; 35(1):98. <https://doi.org/10.1186/s13046-016-0378-z> PMID: 27329034
48. Gladson CL. The extracellular matrix of gliomas: modulation of cell function. *J Neuropathol Exp Neurol*. 1999; 58(10):1029–40. <https://doi.org/10.1097/00005072-199910000-00001> PMID: 10515226
49. Ferrer VP, Moura Neto V, Mentlein R. Glioma infiltration and extracellular matrix: key players and modulators. *Glia*. 2018; 66(8):1542–65. <https://doi.org/10.1002/glia.23309> PMID: 29464861
50. Liang J, Piao Y, Holmes L, Fuller GN, Henry V, Tiao N, et al. Neutrophils promote the malignant glioma phenotype through S100A4. *Clin Cancer Res*. 2014; 20(1):187–98. <https://doi.org/10.1158/1078-0432.CCR-13-1279> PMID: 24240114
51. Sippel TR, White J, Nag K, Tsvankin V, Klaassen M, Kleinschmidt-DeMasters BK, et al. Neutrophil degranulation and immunosuppression in patients with GBM: restoration of cellular immune function by targeting arginase I. *Clin Cancer Res*. 2011; 17(22):6992–7002. <https://doi.org/10.1158/1078-0432.CCR-11-1107> PMID: 21948231
52. Paolillo M, Serra M, Schinelli S. Integrins in glioblastoma: Still an attractive target? *Pharmacol Res*. 2016; 113(Pt A):55–61. <https://doi.org/10.1016/j.phrs.2016.08.004> PMID: 27498157
53. Malric L, Monferran S, Gilhodes J, Boyrie S, Dahan P, Skuli N, et al. Interest of integrins targeting in glioblastoma according to tumor heterogeneity and cancer stem cell paradigm: an update. *Oncotarget*. 2017; 8(49):86947–68. <https://doi.org/10.18632/oncotarget.20372> PMID: 29156849
54. Payne LS, Huang PH. The pathobiology of collagens in glioma. *Mol Cancer Res*. 2013; 11(10):1129–40. <https://doi.org/10.1158/1541-7786.MCR-13-0236> PMID: 23861322

55. Wang M, Wang T, Liu S, Yoshida D, Teramoto A. The expression of matrix metalloproteinase-2 and -9 in human gliomas of different pathological grades. *Brain Tumor Pathol.* 2003; 20(2):65–72. <https://doi.org/10.1007/bf02483449> PMID: 14756443
56. Szatmari T, Otvos R, Hjerpe A, Dobra K. Syndecan-1 in Cancer: Implications for Cell Signaling, Differentiation, and Prognostication. *Dis Markers.* 2015; 2015:796052. <https://doi.org/10.1155/2015/796052> PMID: 26420915
57. Shi S, Zhong D, Xiao Y, Wang B, Wang W, Zhang F, et al. Syndecan-1 knockdown inhibits glioma cell proliferation and invasion by deregulating a c-src/FAK-associated signaling pathway. *Oncotarget.* 2017; 8(25):40922–34. <https://doi.org/10.18632/oncotarget.16733> PMID: 28422726
58. Tentler JJ, Tan AC, Weekes CD, Jimeno A, Leong S, Pitts TM, et al. Patient-derived tumour xenografts as models for oncology drug development. *Nat Rev Clin Oncol.* 2012; 9(6):338–50. <https://doi.org/10.1038/nrclinonc.2012.61> PMID: 22508028
59. Weeber F, Ooft SN, Dijkstra KK, Voest EE. Tumor Organoids as a Pre-clinical Cancer Model for Drug Discovery. *Cell Chem Biol.* 2017; 24(9):1092–100. <https://doi.org/10.1016/j.chembiol.2017.06.012> PMID: 28757181
60. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *Journal of machine learning research.* 2011; 12(Oct):2825–30.
61. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment.* 2008; 2008(10):P10008.
62. Tabe-Bordbar S, Emad A, Zhao SD, Sinha S. A closer look at cross-validation for assessing the accuracy of gene regulatory networks and models. *Sci Rep.* 2018; 8(1):6620. <https://doi.org/10.1038/s41598-018-24937-4> PMID: 29700343
63. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 2016; 44(W1):W90–7. <https://doi.org/10.1093/nar/gkw377> PMID: 27141961
64. Seabold S, Perktold J, editors. *Statsmodels: Econometric and statistical modeling with python*. Proceedings of the 9th Python in Science Conference; 2010: Scipy.