Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data

Linh Nguyen, Cuong C Dang, Pedro J Ballester

RESEARCH ARTICLE

REVISED Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data [version 2; referees: 2 approved]

Linh Nguyen[1-4], Cuong C Dang [1-4], Pedro J. Ballester [1-4]

[1]Cancer Research Center of Marseille UMR7258, Marseille, France
[2]Aix-Marseille Université, Marseille, France
[3]Institut Paoli-Calmettes, Marseille, France
[4]Cancer Research Center of Marseille, INSERM U1068, Marseille, France

v2 **First published:** 28 Dec 2016, **5**(ISCB Comm J):2927 (doi: 10.12688/f1000research.10529.1)
Latest published: 14 Mar 2017, **5**(ISCB Comm J):2927 (doi: 10.12688/f1000research.10529.2)

Abstract

Background: Selected gene mutations are routinely used to guide the selection of cancer drugs for a given patient tumour. Large pharmacogenomic data sets, such as those by Genomics of Drug Sensitivity in Cancer (GDSC) consortium, were introduced to discover more of these single-gene markers of drug sensitivity. Very recently, machine learning regression has been used to investigate how well cancer cell line sensitivity to drugs is predicted depending on the type of molecular profile. The latter has revealed that gene expression data is the most predictive profile in the pan-cancer setting. However, no study to date has exploited GDSC data to systematically compare the performance of machine learning models based on multi-gene expression data against that of widely-used single-gene markers based on genomics data. *Methods:* Here we present this systematic comparison using Random Forest (RF) classifiers exploiting the expression levels of 13,321 genes and an average of 501 tested cell lines per drug. To account for time-dependent batch effects in IC_{50} measurements, we employ independent test sets generated with more recent GDSC data than that used to train the predictors and show that this is a more realistic validation than standard k-fold cross-validation. *Results and Discussion:* Across 127 GDSC drugs, our results show that the single-gene markers unveiled by the MANOVA analysis tend to achieve higher precision than these RF-based multi-gene models, at the cost of generally having a poor recall (i.e. correctly detecting only a small part of the cell lines sensitive to the drug). Regarding overall classification performance, about two thirds of the drugs are better predicted by the multi-gene RF classifiers. Among the drugs with the most predictive of these models, we found pyrimethamine, sunitinib and 17-AAG. *Conclusions:* Thanks to this unbiased validation, we now know that this type of models can predict *in vitro* tumour response to some of these drugs. These models can thus be further investigated on *in vivo* tumour models. R code to facilitate the construction of alternative machine learning models and their validation in the presented benchmark is available at http://ballester.marseille.inserm.fr/gdsc.transcriptomicDatav2.tar.gz.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
REVISED version 2 published 14 Mar 2017		
version 1 published 28 Dec 2016	✓ report	✓ report

1 **Marc Poirot**, INSERM (French National Institute of Health and Medical Research) , France

2 **Ronnie Alves**, Graduate Program in Computer Science, Brazil
Instituto Tecnologico Vale, Brazil

Discuss this article

Comments (0)

This article is included in the International Society for Computational Biology Community Journal gateway.

This article is included in the Machine learning: life sciences collection.

Corresponding author: Pedro J. Ballester (pedro.ballester@inserm.fr)

Competing interests: No competing interests were disclosed.

How to cite this article: Nguyen L, Dang CC and Ballester PJ. **Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data [version 2; referees: 2 approved]** *F1000Research* 2017, **5**(ISCB Comm J):2927 (doi: 10.12688/f1000research.10529.2)

Grant information: This work has been carried out thanks to the support of a A*MIDEX grant (#ANR-11-IDEX-0001-02) funded by the French Government 'Investissements d'Avenir' programme, and the 911 Programme PhD scholarship from Vietnam National International Development.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 28 Dec 2016, **5**(ISCB Comm J):2927 (doi: 10.12688/f1000research.10529.1)

REVISED **Amendments from Version 1**

In response to the feedback of the reviewers (see our responses online), the following changes have been made: • We have now defined the abbreviation "GDSC" in the abstract too. • Figure 3's caption has been clarified. • What constitutes an acceptable performance prediction has been clarified in page 8. • In page 6, we now state that RF is also robust to overfitting, as evidenced in Figure 1. • In page 3, reasons for differences in predictive performance of single-gene vs multi-gene markers across drugs are clarified. • The abstract is also modified to indicate how to get the R code, which was requested by one of the reviewers.

See referee reports

Introduction

Personalised approaches to the diagnosis and treatment of cancer is a strong current trend, often based on the analysis of tumour DNA[1]. Somatic DNA mutations can affect the abundance and function of a range of gene products, including those involved in the response of the tumour to anticancer therapy[2]. Therefore, the genomic profile of a tumour is usually valuable for predicting its sensitivity to a certain drug[3,4]. Thus, a number of studies have profiled tumours using single-nucleotide variants or copy-number alterations to use them as input features to predict which tumours will be sensitive to a given drug. In addition, transcriptomic data has also been proven to be an informative molecular profile[5], as the expression levels of genes have led to the identification of cancer subtypes, prognosis prediction and drug sensitivity prediction[6].

Human-derived cancer cell lines, especially immortalised cancer cell lines, play an important role in preclinical research for the discovery of genomic markers of drug sensitivity[5,7–9]. This type of tumour model permits experiments to be implemented quickly and with a relatively low cost[10,11], unlike more patient-relevant models, such as *ex vivo* tumour cultures[12,13] or patient-derived xenografts[14,15] (in contrast to these advantages, cell lines have also well-known limitations that have to be kept in mind[10]). The molecular profiles of such cell lines are often used as input features for drug sensitivity prediction[5,8] via the development of both single-gene markers and other models, like pharmacogenomics[16–18], pharmacotranscriptomics[19–21], multi-task learning[16,17,22–25] and quantitative structure-activity relationship (QSAR) models[26,27]. Recently, several consortia have generated large pharmacogenomic data sets, which consist of both molecular and drug sensitivity profiles of several hundreds of cancer cell lines, e.g. Genomics of Drug Sensitivity in Cancer (GDSC)[8], Cancer Cell Line Encyclopedia (CCLE)[9], and profiling a panel of 60 cancer cell lines from the National Cancer Institute (NCI-60)[7]. Among them, GDSC data is currently offering the highest number of cell lines tested per drug[8].

To date, predictive models based on GDSC data have been mostly restricted to single-gene markers of drug sensitivity[8] (i.e. statistically significant drug-gene associations). However, multi-gene elastic net models have also been used for a related purpose, namely estimating the importance of somatic mutations in drug sensitivity prediction[8]. Some researchers have also investigated the performance of multi-gene machine learning models exploiting GDSC data[16]. Nevertheless, we and others[9,17,18] have not studied how well multi-gene markers compare to single-gene markers. Such analysis is essential to understand the benefits of modelling multiple gene alterations. Very recently, machine learning models have been used to compare the predictive value of various molecular profiles in drug sensitivity modelling[5], but without comparing such models to single-gene markers. An important outcome of that study revealed that gene expression data was the most predictive molecular profile in the pan-cancer setting. Beyond this research area, multivariate machine learning models are also starting to be advocated for genomic-based prediction of other complex phenotypic traits[28].

In practice, it is entirely possible that models based on one feature (single-gene markers) are more predictive than those based on more than one feature (multi-variate classifiers). In part, this is due to the high-dimensionality of training data (in the present study, the number of gene expression values is much higher than that of cell lines treated with the considered drug), which poses a challenge to classifiers. In addition, while both models look at the same drug sensitivity data, each model employs a different set of features (genomic vs transcriptomic). Therefore, a single gene mutation might sometimes be more predictive of drug sensitivity than a model based on gene expression values. Furthermore, only a very small subset of features might be predictive of cell line sensitivity to a given drug, a case that could be challenging for a model using all the transcriptomic features. Moreover, the size of training and test sets varies because each drug was tested with a different number of cell lines (thus, class imbalances in training set and test set are also different depending on the drug). Taking all these convoluted factors into account, the relative performance of these models should be drug-dependent and hard to anticipate prior to the corresponding numerical experiments.

This leads to a key question: for which drugs are multi-variate markers more predictive of cell line sensitivity than univariate markers. Very recently, this question has been finally investigated using large-scale GDSC data[5], although there are several limitations in this analysis. First, this study considered LOBICO logic models with up to four features because searching for more complex models was not feasible with the underlying integer linear programming solver[5]; however, a drug can have many more than four informative gene alterations. Second, machine learning models were only used to establish which molecular profiles were more informative on average across all drugs. Hence, the performances of these models were not compared against those of single-gene markers (this was only done with logic models). Third, both logic model selection and its classification performance assessment were performed using the same data folds in the adopted cross-validation procedure. Therefore, these cross-validated results represent an overoptimistic performance assessment of LOBICO models and hence it is not clear how predictive the resulting markers really are.

Here we study the performance of machine learning exploiting gene expression profiles. In addition, we compare the performance of these multi-gene machine learning models to that of single-gene markers. For each drug, this analysis is conducted by selecting its best single-gene marker and its multi-gene model on a training set representing the data available at model selection time. Thereafter, we test both models in an unbiased manner using a time-stamped

independent test set, i.e. data that was generated after the training data and not used for model building or selection. The advantages of using a time-stamped data partition instead of K-fold cross-validation are that this mimics a blind test, the same data is not used for both model selection and performance assessment (thus avoiding performance overestimation) and real-world issues like time-dependent batch effects[29] are taken into account. On the other hand, since transcriptomic data has been found to be the most predictive in the pan-cancer setting[5], our study focuses on the exploitation of transcriptomic data. In particular, the predictive performance of pan-cancer markers of drug sensitivity on an independent test set is most relevant to help stratify patients for basket trials[30], where patients with any type of cancer are included if their tumours are predicted to be sensitive to the investigated treatment. Another reason to limiting the scope to transcriptomic-based machine learning models is that models integrating data from multiple molecular profiling technologies would be less amenable for clinical implementation, due to much higher requirements in cost, time and resources per patient. Therefore, there is a need to understand for which drugs models combining gene expression values provide better cell line sensitivity prediction than standard single-gene markers.

Methods

GDSC pharmacogenomic data

From the Genomics of Drug Sensitivity in Cancer (GDSC) ftp server (ftp://ftp.sanger.ac.uk/pub4/cancerrxgene/releases/), the following files from the first data release (release 1.0) were downloaded:gdsc_manova_input_w1.csv and gdsc_manova_output_w1.csv. There are 130 unique drugs in gdsc_manova_input_w1.csv, as camptothecin was tested twice (drug IDs 195 and 1003), and thus we only kept the instance that was more widely tested (i.e. drug ID 1003 on 430 cell lines). Hence, the data represent a panel of 130 drugs tested against 638 cancer cell lines resulting in a total of 47748 IC_{50} values (57.6% of all possible drug-cell pairs). In addition, we downloaded new data from release 5.0 (gdsc_manova_input_w5.csv), which is the latest release using the same experimental techniques to generate pharmacogenomic data, and considering the same genes as in the first release. Release 5.0 contains 139 drugs tested on 708 cell lines comprising 79,401 IC_{50} values (80.7% of all possible drug-cell pairs). Hence, the majority of the new IC_{50} values came from previously untested drug-cell pairs formed by drugs and cell lines in common between both releases. The downloaded IC_{50} values are actually the natural logarithm of IC_{50} in μM units, so negative values came from drug responses more potent than 1μM. Each of these values were converted into their logarithm base 10 in μM units, denoted as $logIC_{50}$ (e.g. $logIC_{50}$=1 corresponds to IC_{50}=10μM). In this way, differences between the two drug response values are expressed as orders of magnitude in the molar scale.

gdsc_manova_input_w1.csv also contains genetic mutation data for 68 cancer genes (these were selected as the most frequently mutated cancer genes[8] and their mutational statuses characterise each of the 638 cell lines). For each gene-cell pair, a 'x::y' description is provided, where 'x' specifies a coding variant and 'y' states copy number information from SNP6.0 data. As usual[8], a gene for which a mutation is not detected in a given cell line is annotated as wild-type (wt). A gene mutation is annotated if a) a protein sequence variant is detected (x ≠{wt,na}) or b) a gene deletion/amplification is detected. The latter corresponds to a copy number (cn) range that is different from the wt value of y=0<cn<8. Furthermore, three genomic translocations were considered (BCR_ABL, MLL_AFF1 and EWS_FLI1) by the GDSC. For each of these gene fusions, cell lines are either identified as a not-detected fusion or the identified fusion is stated (i.e. wt or mutated with respect to the gene fusion, respectively). The microsatellite instability (msi) status of each cell line is also determined and provided. Further details can be found in the original publication by Garnett *et al.*[8].

GDSC pharmacotranscriptomic data

Gene expression data was generated using Affymetrix Human Genome U219 Array Chip and was normalized with the Robust Multi-Array Average method. The number of cell lines with gene expression data in releases 1.0 and 5.0 of the GDSC are 571 and 624, respectively. In terms of data in common, both releases contain the expression level of 13,321 genes across 624 cancer cell lines. These genes consist of 12,644 protein coding genes, 47 pseudogenes, 29 non-coding RNA genes and 601 uncharacterized genes according to the HUGO Gene Nomenclature Committee (HGNC).

Non-overlapping training and test sets by partitioning data in chronological order

There are 127 drugs in common between both releases. Three drugs are exclusively included in the first release (A-769662, Metformin and BI-D1870), whereas release 5.0 contains 12 additional drugs (TGX221, OSU-03012, LAQ824, GSK-1904529A, CCT007093, EHT 1864, BMS-708163, PF-4708671, JNJ-26854165, TW 37, CCT018159 and AG-014699).

Regarding genomic features, cell lines from both releases have been profiled for 71 common gene alterations in cancer. In addition to the three translocations and msi status, the mutational statuses of 67 genes could be considered (i.e. those for the 68 selected genes in the first release except for the mutational status of the WT1 gene, which was not included in the subsequent 5.0 release). To ensure that we are using exactly the same drug-gene associations as in the GDSC study, we directly employed the associations and their p-values as downloaded from release 1.0.

There are two non-overlapping data sets per drug. The training set contains the cell lines tested with the drug and gene expression data in release 1.0 (the minimum, average and maximum numbers of cell lines across training data sets are 237, 330 and 467, respectively), along with their IC_{50}s for the considered drug. The test set contains the new cell lines tested with the drug and with gene expression data in release 5.0 (the minimum, average and maximum numbers of cell lines in the test data sets are 42, 171 and 306, respectively). Thus, a total of 254 pharmacotranscriptomic data sets were assembled and analysed for this study.

Measuring predictive performance of a classifier on a given data set

The pharmacotranscriptomic data for the i^{th} drug (D_i) can be represented as follows:

$$D_i = \left\{\left(logIC_{50,i}^{(k)}, \boldsymbol{x}^{(k)}\right)\right\}_{k=1}^{k=n_i}$$

in which, the sensitivity of cancer cell lines against the i^{th} drug has been tested on n_i cell lines. **x** is the vector with 13,321 gene expression values. The data can act as a training set, cross-validation fold or test set of any of the tested drugs.

First, a cell line sensitivity threshold is defined to distinguish between those resistant or sensitive to a given drug. For each drug, we calculated the median of all the $logIC_{50}$ values from training set cell lines and fix it as the threshold. Cell lines with $logIC_{50}$ below the threshold are therefore sensitive, while those with $logIC_{50}$ above the threshold are resistant to the drug of interest.

Upon using the model to make predictions in a given data set, two different sets of cell lines will be obtained for each drug: those predicted to be sensitive and those predicted to be resistant. Then, using the threshold for the drug, we can assess classification performance by calculating the number of cell lines in each of these four categories: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). These can be summarised by the Matthews Correlation Coefficient (MCC):

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FN)\cdot(FN+TN)\cdot(TN+FP)\cdot(FP+TP)}}$$

MCC takes values from -1 to 1. A MCC value of 0 means that the tested model has no predictive value, MCC lower than 0 means that the tested model predicts drug sensitivity worse than random and an MCC value equal to 1 indicates that the tested model perfectly predicts the sensitivity of the cell lines against the drug of interest.

In addition to MCC, we also investigated precision (PR), recall (RC) and F1-scores (F1) of the model for each drug to provide a more comprehensive comparison of multi-gene models to single-gene markers. Precision and recall are the two measures of performance for binary classifier, which can be calculated as follows:

$$PR = \frac{TP}{TP+FP} \qquad RC = \frac{TP}{TP+FN}$$

Both metrics can take values from 0 to 1. Precision and recall equal to 0 means that TP = 0, the model fails to identify any cell line sensitive to the drug. By contrast, PR and RC equal to 1 means that FP and FN are equal to 0, respectively. In these cases, either the model does not predict any resistant cell line as sensitive (FP = 0) or it does not misclassify sensitive cell lines as resistant (FN = 0), respectively.

The F1-score is another measure combining PR and RC. F1-score can be computed as:

$$F1 = 2\frac{PR \cdot RC}{PR + RC}$$

The F1-score is at most 1 (when both PR and RC = 1) and minimum value equal to 0 (RC =0 regardless of the PR value or *vice versa*). High F1 scores mean that both precision and recall are high for the classifier.

Single-gene markers built from the training dataset

We downloaded gdsc_manova_output_w1.csv containing 8701 drug-gene associations with their corresponding p-values computed by MANOVA test. Then, we kept those associations involving the 127 common drugs leading to a set of 8330 drug-gene associations, of which 386 were significant (i.e. p-value smaller than a FDR 20% Benjamini-Hochberg adjusted threshold of 0.00840749). As in previous studies[5,8], each statistically significant drug-gene association is regarded as a single-gene marker of *in vitro* drug response.

The best single-gene marker for a drug was identified as its drug-gene association with the lowest p-value. This constitutes a binary classifier with a single independent variable, built using training data alone and fixed at this model selection stage. These drug-lowest p-values were not statistically significant in 15 of the 127 drugs, with the highest of these being P = 0.0354067. In these cases, we still selected them as the best available for these single-gene classifiers. Otherwise, multi-gene markers would be directly better than the single-gene approach for these drugs.

After the model selection step, the single-gene marker for each drug is assessed on the corresponding independent test set. This form of external validation is particularly demanding since the test data is completely separate from training data and constitutes future data from the model training perspective. In 27 drugs, none of the cell lines in the test set harbour the marker mutation and hence TP=FP=0. Therefore, no prediction is provided by these markers and thus MCC and PR are assigned a zero value.

Multi-gene transcriptomic markers built from the training data set

For each of the 127 drugs, we built a Random Forest (RF) classification model[31] using exactly the same pharmacological data for training as the corresponding single-gene marker. However, while single-gene markers leverage genomic data, these RF models exploit transcriptomic data instead. All the 13321 gene expression values are used as features (RF_all). Each RF model was built using 1000 trees and the recommended value of its control parameter m_{try} (the square root of the number of considered features, thus m_{try}=115 here). All the described modelling was implemented in R language, using Microsoft R Open (MRO) version 3.2.5.

Results and discussion

Comparing single-gene markers and transcriptomic-based models on the same test sets

A single-gene marker is a classifier considering the mutational status of a given gene as its only independent variable (i.e. whether the gene is wild-type or mutated). As the gene used as a marker arises from the analysis of which drug-gene associations are statistically significant based on the training data, external validation of such markers is not carried out. In this sense, machine learning represents a different culture[32], where the validity of the predictor is only demonstrated if its prediction is better than random on a test set independent of the employed training set. In this study, we use the same test set to compare the performance of both single-gene markers and multi-gene transcriptomic-based RF models.

For each drug, there were two data sets generated with non-overlapping sets of cancer cell lines. The first data set was the training set, which contains cell lines that were tested prior to the release of release 1.0 of the GDSC data, each with its IC_{50} values for the drug and its gene expression profile. The second data set was the test set, including the new cell lines from release 5.0 (i.e. new data not included in the first release). The median $logIC_{50}$ in μM units of all cell lines in the training set defines the sensitivity threshold for both the training set and the test set. The next step was evaluating the performance of both methods in both data sets by calculating the Matthews Correlation Coefficient (MCC), Precision (PR), Recall (RC) and F1-score (F1). The Methods section provides further details on performance evaluation.

Random Forest (RF)[31] is a machine learning technique that is robust to overfitting and works well on high-dimensional data[33], including GDSC data[16]. Therefore, without making any claim about its optimality for this class of problems, we constructed a RF classification model on the same training data set as the single-gene marker. This permits a direct comparison of the two models. Each RF model was built using 1000 trees, with the default value of the control parameters m_{try} (the square root of the number of considered features to split a tree node). The built RF model was subsequently tested on the corresponding test set. Figure 1 displays the results for the drug pyrimethamine as an example. Pyrimethamine targets dihydrofolate reductase in the DNA replication pathway[34] and its strongest association is to the BRAF gene (P=0.002) leading to a moderate level of prediction in this training set (Figure 1A). The prediction of this single-gene marker on the test set (Figure 1B) is worse than random (MCC=-0.03), with its recall being particularly poor (RC=0.03) and average precision at 0.50. Unsurprisingly, RF prediction on the training set is perfect due to intense overfitting[35] arising from the high dimensionality of the problem (Figure 1C). Nevertheless, it is important to note that this overfitted model achieves a substantially better test set performance than that of the best single-gene marker (compare Figures 1D and 1B, respectively).

Large inter-drug variability in the response rate of cell lines predicted to be sensitive

To assess the proportion of cell lines predicted to be sensitive that are actually sensitive to a drug by each model, we calculated their precision (PR) on the test set. Figure 2 shows the comparison between test set precision of single-gene markers and that of multi-gene models across 127 drugs. The precision of each method is highly drug-dependent and 61 drugs had their best single-gene marker leading to higher precision than the corresponding multi-gene model, whereas the other 66 drugs had the multi-gene model with better precision (see Supplementary File 2). In other words, the sensitivity of cancer cell lines against 66 drugs can be predicted with higher precision when exploiting multi-variate gene expression data rather than a single gene mutation. In particular, the multi-gene model provides better precision for all the drugs for which the best single-gene marker involves a relatively rare mutation (i.e. those for

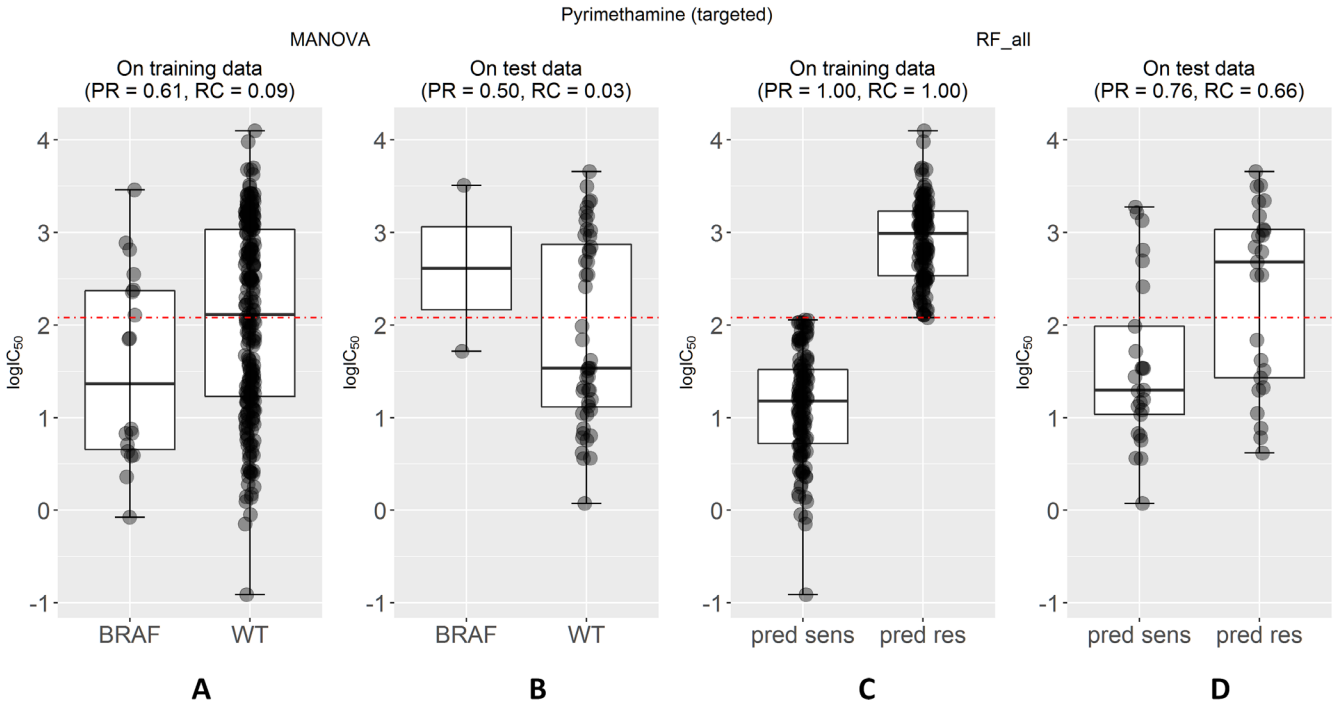

Figure 1. Predictive performance of markers of cell sensitivity to the approved drug pyrimethamine. (**A**) The single-gene marker with the lowest p-value on the training set was the pyrimethamine-BRAF sensitising association (P=0.002)[8]. (**B**) The boxplots show the sensitivity of cell lines on the independent test set for pyrimethamine depending on whether these harbour mutations in the BRAF gene or not (WT). Using this marker, BRAF-mutant cell lines are predicted to be sensitive to this drug (i.e. below the threshold in red established with training data), but the prediction is worse than random (Matthews Correlation Coefficient (MCC)=-0.03) with its recall being particularly poor (RC=0.03) and average precision (PR)=0.50. (**C**) The multi-gene marker RF_all was built using Random Forests (RF) and all the gene expression values on exactly the same drug-cell pairs as the single-gene markers. (**D**) On the test set, the RF classifier achieves a substantially better performance than single-gene markers (MCC=0.36 vs -0.03) with PR=0.76 and RC=0.66.

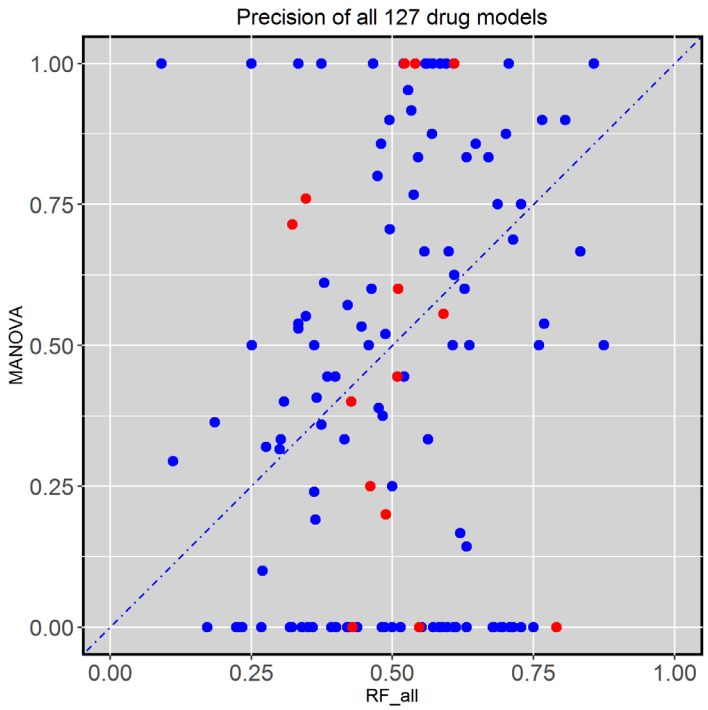

Figure 2. Test set precision of MANOVA single-gene markers versus RF transcriptomic models across the 127 drugs. A large variability is observed, with 66 drugs obtaining better precision with Random Forest (RF) classifiers using all transcriptomic features. Cytotoxic drugs are in red and targeted drugs are in blue.

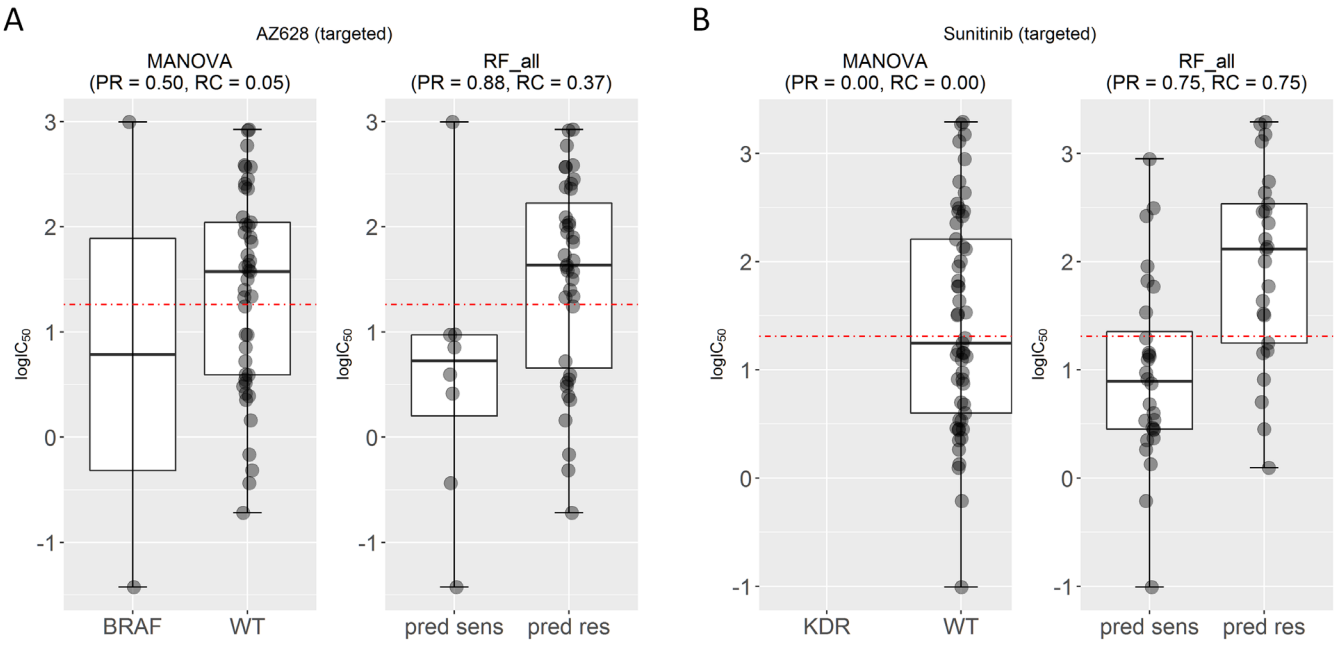

Figure 3. Examples of drugs for which transcriptomic markers predict their cell line sensitivity with better precision than that of single-gene markers on the test set. (**A**) Test set precision obtained by the AZD628-BRAF marker is moderate (PR=0.50) despite being a strong drug-gene association ($P=3{\cdot}10^{-15}$). By contrast, the multi-gene marker for AZD628 achieves a substantially higher precision (PR=0.88). (**B**) The sunitinib-Kinase Insert Domain Receptor (KDR) association (P=0.0002) offers no precision in the test set, since none of the test cell lines harbour mutations in the KDR gene. By contrast, the transcriptomic marker achieves a much higher precision (PR=0.75). Interestingly, both multi-gene markers achieve much better recall (RC=0.37 for AZD628 and RC=0.75 for sunitinib) than their corresponding single-gene markers (RC=0.05 and RC=0.00), which means that a substantially higher proportion of sensitive cell lines are correctly predicted as sensitive.

which no test set cell line is mutated with respect to the marker gene and thus are unable to provide any level of precision).

Next, we present two examples of drugs for which the test set precision generated by the multi-gene model is higher than that of the single-gene model (Figure 3). AZD628 is a b-raf inhibitor, which plays a regulatory role in the MAPK/ERK pathway[36]. This drug is associated with the mutations in the BRAF gene ($P=3{\cdot}10^{-15}$), which codes for the b-raf kinase. In total, 50% of BRAF-mutant cell lines are sensitive to this drug, while using the RF model combining all 13,321 transcriptomic features results in 88% of cell lines predicted to be sensitive being actually sensitive to this drug. The second example is the prediction of sensitivity to sunitinib, which targets multiple receptor tyrosine kinases regulating different aspects of cell signaling[37]. The most strongly associated gene to sunitinib is Kinase Insert Domain Receptor (KDR) (P=0.0002). As no KDR mutation was found in any test cell lines, the single-gene marker could not predict the sensitivity of any cell line to sunitinib (PR=0). In contrast, the multi-gene model provides a much better precision for this drug (PR=0.66). The multi-gene models of both drugs generate a higher recall than their corresponding single-gene model, which is investigated in the following section.

Multi-gene markers generally achieve much higher recall than single-gene markers

Figure 3 shows that the test set recall is much higher for multi-gene markers than for single-gene markers of AZD628 and sunitinib. To examine whether this is a general trend, Figure 4A plots test set recall across all the drugs. There is indeed a clear trend: 119 out of 127 drugs obtain a higher proportion of correctly predicted sensitive cell lines with the multi-gene markers.

Figure 4B shows the test set F-score (F1) for the same drugs. High F1 values highlight markers achieving both high precision and high recall in the test set. Notably, the multi-gene classifiers lead to better recall and F1-scores in all the cytotoxic drugs. We have selected two drugs with high F1 values by the multi-gene marker, BAY-61-3606 and 17-AAG, in order to analyse them further (Figure 5).

Figure 5A compares the test set performance between single-gene and multi-gene models for the drugs BAY-61-3606 and 17-AAG, respectively. BAY-61-3606 is an inhibitor for the spleen tyrosine kinase, with key roles in adaptive immune receptor signalling, as well as regulation of cellular adhesion and vascular development[38]. The single-gene model generates poor precision and recall for this drug (PR = RC = 0), as the only cell line that harbours the actionable mutation was incorrectly predicted as resistant (TP = 0). By contrast, the multi-gene model achieves high performance in terms of both precision and recall (PR = 0.68 and RC = 0.83). On the other hand (Figure 5B), 17-AAG specifically inhibits HSP90, a protein that chaperones the folding of proteins required for tumour growth[39]. The multiple-gene model provides much higher PR (PR = 0.61) and RC (RC = 0.75) compared with its best single-gene marker (PR = 0.50 and RC = 0.03). This case exemplifies a common problem with single-gene markers: often only a small proportion of tumours harbour the actionable mutation[40]. This translates to very low recall, which in a clinical setting would mean that only a small proportion of patients

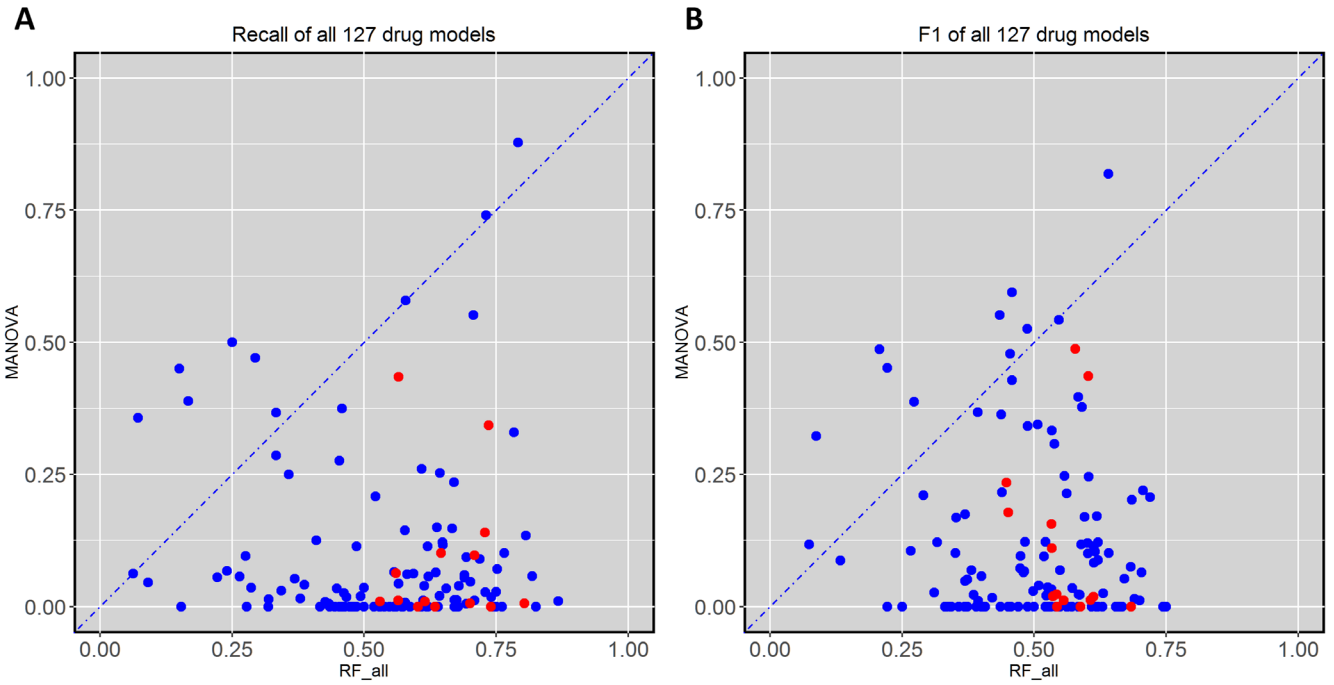

Figure 4. Test set recall and F-scores of single-gene and transcriptomic models across the 127 drugs. (**A**) Transcriptomic markers achieve much higher recall than single-gene markers in 117 of the 127 drugs. (**B**) Similarly, multi-gene markers achieve higher F-scores in 117 of the 127 drugs. In each plot, cytotoxic drugs are in red and targeted drugs are in blue. All cytotoxic drugs have better recall and F-scores by the Random Forest (RF) transcriptomic models.

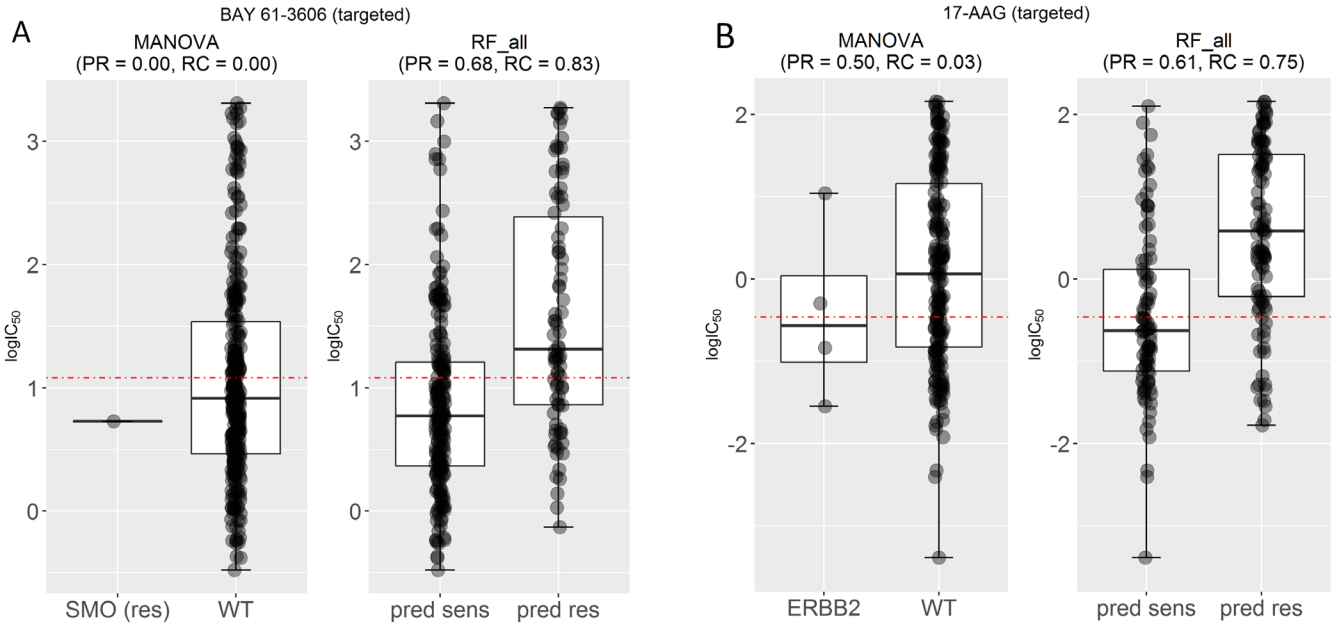

Figure 5. Examples of drugs that have high recall and F-scores. (**A**) Mutated Smoothened, Frizzled Class Receptor (SMO) was the most significant single-gene marker for BAY-61-3606 resistance (P=0.03) using training data. On the test set, this marker obtained no precision and no recall because the only SMO-mutant test set cell line was misclassified. By contrast, the corresponding multi-gene marker, built with the same training data, obtained a high precision (PR=0.68) and better recall (RC=0.83) on the same test data. (**B**) Mutated receptor tyrosine-protein kinase erbB-2 (ERBB2) is the most significant single-gene marker of 17-AAG sensitivity (P=0.008), but its test set recall is poor (RC=0.03). By contrast, the multi-gene marker achieves a much higher precision (PR=0.61) and recall (RC=0.75).

responsive to the drug would be treated with it because of it being missed by its marker.

The importance of using independent test sets to benchmark markers of drug sensitivity

After separately analysing the two sources of classification error via precision and recall, we analysed both types of error together in order to assess which predictors are better than random classification (i.e. MCC = 0)[41]. In the context of this study, we are interested in those RF models offering a MCC value on the test set higher than that provided by the best single-gene marker of the same drug. Note that the higher the test set MCC of the model, the better its ability to discriminate between unseen cell lines.

The classification of both models can in principle be assessed in three ways across the considered drugs (Figure 6). Figure 6A evaluates the MCC of both predictors on the training data, which is common practice with single-gene markers. Figure 6C presents the evaluation of MCC on the non-overlapping test sets. Single-gene markers perform better on the training set than on the test set (on average, $MCC_{training}$=0.11 vs MCC_{test}=0.05; Figures 6A and C), which is due at least in part to the identification of chance correlations in the training set. Unsurprisingly, multi-gene models perform much better on the training set due to intense over-fitting (on average across drugs, $MCC_{training}$=1 vs MCC_{test}=0.12). However, despite overfitting, it is important to note that these models provide on average better test set performance than single-gene markers (MCC_{test}=0.12 vs MCC_{test}=0.05). This is a well-known characteristic of the RF technique, which is robust to overfitting, in that it is able to provide competitive generalisation to other data sets despite overfitting (this behaviour has also been observed in analogous applications of RF[43]).

Figure 6B shows the comparison between performance of the single-gene markers on the test set and the 10-fold cross-validated performance of the multi-gene markers on the training set. The latter provides a more optimistic performance assessment (average MCC=0.18 and 84.2% of drugs better predicted by the multi-gene models). This is likely due to effects between different batches of culture medium that are known to affect drug sensitivity measurements[10,42]. As expected, testing the models on the independent test sets generates worse results than on the training test or the cross-validation set.

Conclusions

To the best of our knowledge, this is the first systematic comparison of single-gene markers versus transcriptomic-based machine learning models of cell line sensitivity to drugs. This is important as transcriptomic data has been shown to be the most predictive data type in the pan-cancer setting[5]. A closely related analysis was included in a very recent study[5]. However, this analysis is based on logic classifiers that can only exploit up to four

features instead of fully-featured machine learning classifiers. Furthermore, the performance results in that study are based on cross-validations, thus leading to overoptimistic performance, due to batch effects as we have seen here. The latter would be exacerbated if the same cross-validation is also used for model selection, as it was the case in the previous study[5]. Despite these limitations, these new logic classifiers are very valuable as they can potentially explain why a particular cell line is sensitive to the drug, something that machine learning classifiers are not suitable for.

Although single-gene markers were able to predict the sensitivity of cancer cell lines to anti-cancer drugs with generally high test set precision (Figure 2), very poor precision and a very low recall was provided for other drugs, especially those that are best associated with relatively rare actionable mutation. On the other hand, multi-gene classifiers obtained a much better recall, also known as sensitivity, for most of the drugs (Figure 4). This result is in line with criticism of single-gene markers, which lead to an extremely small proportion of patients that can benefit[40]. In this sense, one could argue that there is a need for not only precision oncology, but for precision and recall oncology, and that multivariate classifiers have the potential to identify all the responsive patients, not only a subset of those with an actionable mutation.

While no strong single-gene markers of sensitivity were found for cytotoxic drugs[8], the multi-gene machine learning models perform better than the single-gene markers in 12 of the 14 cytotoxic drugs (Figure 6C), with all cytotoxic drugs having better recall (Figure 4A). This suggests that the sensitivity to cytotoxic drugs has a stronger multi-factorial nature, which is thus better captured by multi-gene models. Although much less developed to date, personalised oncology approaches have already been suggested for cytotoxic drugs[44,45].

The study of molecular markers for drug sensitivity is currently of great interest. This endeavour is not limited to improve personalised oncology, it is also important for drug development and clinical research[46,47]. As a part of cancer diagnosis and treatment research, a vast amount of tumour molecular profiling data is typically generated[48] and thus there is an urgent need for their optimal exploitation[49]. Here we propose a method to exploit transcriptomic data of cancer cell lines to classify them into sensitive and resistant groups. Our study has found that cancer cell sensitivity to two thirds of the studied drugs, including 12 of the 14 cytotoxic drugs, are better predicted with multi-variate transcriptomic-based RF classifiers. These models are particularly useful in those drugs where their best genomic markers

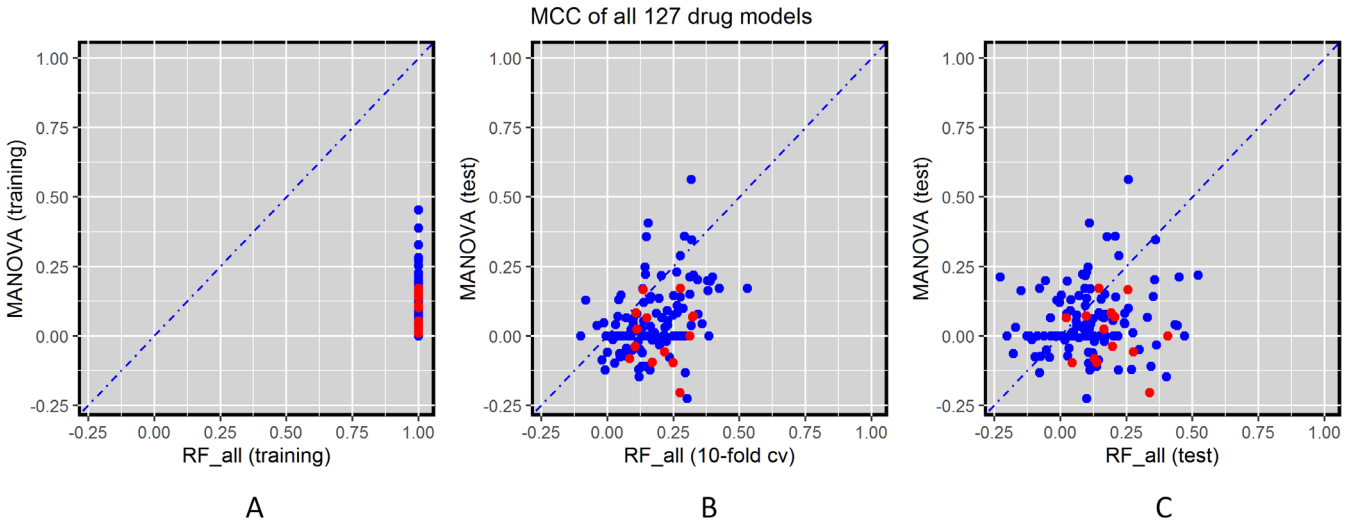

Figure 6. Global performance assessment of single-gene markers versus transcriptomic markers across all 127 drugs. (**A**) Performance assessment on the training data would be strongly biased towards multi-gene markers due to intense overfitting (given the high dimensionality of training data, multi-gene markers obtain maximum (Matthews Correlation Coefficient) MCC for all drugs). (**B**) The performance of single-gene markers on the test set is compared to the 10-fold cross-validated performance of multi-gene markers using training data. The cross-validation is not used for model selection as there is only one Random Forest (RF) model per drug (i.e. no RF control parameter is tuned because the recommended m_{try} is used, due to the high dimensionality of each of the 127 classification problems). However, cross-validation results are substantially better than those from the test set with more recent GDSC data (MCC of 0.18 averaged over the drugs), which suggests time-dependent batch effects[10,42]. (**C**) Using all the comparable data released after the initial GDSC release as a time-stamped test set, 66.1% of drugs are better predicted by the transcriptomic features (this figure is 84.2% using cross-validation). This is the most realistic form of retrospective performance assessment, which leads to the worse results on this challenging problem (MCC of 0.12 averaged over the drugs).

are based on rare mutations. Another contribution of this study is in the proposal of a more realistic performance assessment of markers, which leads to less spectacular, but more robust results. Beyond this proof-of-concept study across 127 drugs, there are several important avenues for future work, which are far too extensive to be incorporated here. For instance, there is a plethora of feature selection techniques that can be applied to reduce the dimensionality of the problem prior to training the classifier for a given drug. Furthermore, the predictive performance of these models can be evaluated on more data or integrated with other molecular profiles. Lastly, we have used a robust classifier technique, RF, but there are many others available and some of these may be more appropriate depending on the analysed drug.

Data availability

The Genomics of Drug Sensitivity in Cancer data sets used in the present study can be found at: ftp://ftp.sanger.ac.uk/pub4/cancerrxgene/releases/release-1.0/

ftp://ftp.sanger.ac.uk/pub4/cancerrxgene/releases/release-5.0/

Author contributions

P.J.B. conceived the study, designed its implementation and wrote the manuscript with the help of L.N. L.N. and C.C.D. implemented the software and carried out the numerical experiments. All authors discussed results and commented on the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

This work has been carried out thanks to the support of a A*MIDEX grant (#ANR-11-IDEX-0001-02) funded by the French Government 'Investissements d'Avenir' programme, and the 911 Programme PhD scholarship from Vietnam National International Development.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material

Supplementary File 1: For each analysed drug, the performances of the best MANOVA-based single-gene marker and Random Forest (RF)-based multi-gene marker on the same test set (both methods were in addition trained on the same data set) are provided. Furthermore, the 10-fold cross-validated performance of the RF-based multi-gene marker is included.

Click here to access the data.

References

1. Wheeler HE, Maitland ML, Dolan ME, *et al.*: **Cancer pharmacogenomics: strategies and challenges.** *Nat Rev Genet.* 2013; **14**(1): 23–34.
PubMed Abstract | Publisher Full Text | Free Full Text
2. McLeod HL: **Cancer pharmacogenomics: early promise, but concerted effort needed.** *Science.* 2013; **339**(6127): 1563–1566.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Azuaje F: **Computational models for predicting drug responses in cancer research.** *Brief Bioinform.* 2016; bbw065.
PubMed Abstract | Publisher Full Text
4. Covell DG: **Data Mining Approaches for Genomic Biomarker Development: Applications Using Drug Screening Data from the Cancer Genome Project and the Cancer Cell Line Encyclopedia.** *PLoS One.* 2015; **10**(7): e0127433.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Iorio F, Knijnenburg TA, Vis DJ, *et al.*: **A Landscape of Pharmacogenomic Interactions in Cancer.** *Cell.* 2016; **166**(3): 740–754.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Rapin N, Bagger FO, Jendholm J, *et al.*: **Comparing cancer vs normal gene expression profiles identifies new disease entities and common transcriptional programs in AML patients.** *Blood.* 2014; **123**(6): 894–904.
PubMed Abstract | Publisher Full Text
7. Abaan OD, Polley EC, Davis SR, *et al.*: **The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology.** *Cancer Res.* 2013; **73**(14): 4372–82.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Garnett MJ, Edelman EJ, Heidorn SJ, *et al.*: **Systematic identification of genomic markers of drug sensitivity in cancer cells.** *Nature.* 2012; **483**(7391): 570–575.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Barretina J, Caponigro G, Stransky N, *et al.*: **The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.** *Nature.* 2012; **483**(7391): 603–307.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Weinstein JN: **Drug discovery: Cell lines battle cancer.** *Nature.* 2012; **483**(7391): 544–5.
PubMed Abstract | Publisher Full Text
11. Majumder B, Baraneedharan U, Thiyagarajan S, *et al.*: **Predicting clinical response to anticancer drugs using an *ex vivo* platform that captures tumour heterogeneity.** *Nat Commun.* 2015; **6**: 6169.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Pemovska T, Kontro M, Yadav B, *et al.*: **Individualized Systems Medicine Strategy to Tailor Treatments for Patients with Chemorefractory Acute Myeloid Leukemia.** *Cancer Discov.* 2013; **3**(12): 1416–29.
PubMed Abstract | Publisher Full Text
13. Azzam D, Volmar CH, Hassan AA, *et al.*: **A Patient-Specific *Ex Vivo* Screening Platform for Personalized Acute Myeloid Leukemia (AML) Therapy.** *Blood.* 2015; **126**(23): 1352.
Reference Source
14. Hidalgo M, Amant F, Biankin AV, *et al.*: **Patient-derived xenograft models: an emerging platform for translational cancer research.** *Cancer Discov.* 2014; **4**(9): 998–1013.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Gao H, Korn JM, Ferretti S, *et al.*: **High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response.** *Nat Med.* 2015; **21**(11): 1318–25.
PubMed Abstract | Publisher Full Text

16. Menden MP, Iorio F, Garnett M, *et al.*: **Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties.** *PLoS One.* 2013; **8**(4): e61318.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Ammad-ud-din M, Georgii E, Gönen M, *et al.*: **Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization.** *J Chem Inf Model.* 2014; **54**(8): 2347–59.
PubMed Abstract | Publisher Full Text

18. Cortés-Ciriano I, van Westen GJ, Bouvier G, *et al.*: **Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel.** *Bioinformatics.* 2016; **32**(1): 85–95.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Riddick G, Song H, Ahn S, *et al.*: **Predicting *in vitro* drug sensitivity using Random Forests.** *Bioinformatics.* 2011; **27**(2): 220–224.
PubMed Abstract | Publisher Full Text | Free Full Text

20. 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.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Kim S, Sundaresan V, Zhou L, *et al.*: **Integrating Domain Specific Knowledge and Network Analysis to Predict Drug Sensitivity of Cancer Cell Lines.** *PLoS One.* 2016; **11**(9): e0162173.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Wang Y, Fang J, Chen S: **Inferences of drug responses in cancer cells from cancer genomic features and compound chemical and therapeutic properties.** *Sci Rep.* 2016; **6**: 32679.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Yuan H, Paskov I, Paskov H, *et al.*: **Multitask learning improves prediction of cancer drug sensitivity.** *Sci Rep.* 2016; **6**: 31619.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Ammad-Ud-Din M, Khan SA, Malani D, *et al.*: **Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization.** *Bioinformatics.* 2016; **32**(17): i455–i463.
PubMed Abstract | Publisher Full Text

25. Zhang N, Wang H, Fang Y, *et al.*: **Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model.** *PLoS Comput Biol.* 2015; **11**(9): e1004498.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Lee AC, Shedden K, Rosania GR, *et al.*: **Data mining the NCI60 to predict generalized cytotoxicity.** *J Chem Inf Model.* 2008; **48**(7): 1379–88.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Kumar R, Chaudhary K, Singla D, *et al.*: **Designing of promiscuous inhibitors against pancreatic cancer cell lines.** *Sci Rep.* 2014; **4**: 4668.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Okser S, Pahikkala T, Airola A, *et al.*: **Regularized machine learning in the genetic prediction of complex traits.** *PLoS Genet.* 2014; **10**(11): e1004754.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Weinstein JN, Lorenzi PL: **Cancer: Discrepancies in drug sensitivity.** *Nature.* 2013; **504**(7480): 381–3.
PubMed Abstract | Publisher Full Text

30. Redig AJ, Jänne PA: **Basket trials and the evolution of clinical trial design in an era of genomic medicine.** *J Clin Oncol.* 2015; **33**(9): 975–977.
PubMed Abstract | Publisher Full Text

31. Breiman L: **Random Forests.** *Mach Learn.* 2001; **45**(1): 5–32.
Publisher Full Text

32. Breiman L: **Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author).** *Stat Sci.* 2001; **16**(3): 199–231.
Publisher Full Text

33. Chen X, Ishwaran H: **Random forests for genomic data analysis.** *Genomics.* 2012; **99**(6): 323–329.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Tommasino C, Gambardella L, Buoncervello M, *et al.*: **New derivatives of the antimalarial drug Pyrimethamine in the control of melanoma tumor growth: an *in vitro* and *in vivo* study.** *J Exp Clin Cancer Res.* 2016; **35**(1): 137.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Lever J, Krzywinski M, Altman N: **Points of Significance: Model selection and overfitting.** *Nat Methods.* 2016; **13**: 703–704.
Publisher Full Text

36. Anderson DJ, Durieux JK, Song K, *et al.*: **Live-cell microscopy reveals small molecule inhibitor effects on MAPK pathway dynamics.** *PLoS One.* 2011; **6**(8): e22607.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Shukla S, Robey RW, Bates SE, *et al.*: **Sunitinib (Sutent, SU11248), a small-molecule receptor tyrosine kinase inhibitor, blocks function of the ATP-binding cassette (ABC) transporters P-glycoprotein (ABCB1) and ABCG2.** *Drug Metab Dispos.* 2009; **37**(2): 359–65.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Pamuk ON, Tsokos GC: **Spleen tyrosine kinase inhibition in the treatment of autoimmune, allergic and autoinflammatory diseases.** *Arthritis Res Ther.* 2010; **12**(6): 222.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Whitesell L, Lindquist SL: **HSP90 and the chaperoning of cancer.** *Nat Rev Cancer.* 2005; **5**(10): 761–772.
PubMed Abstract | Publisher Full Text

40. Huang M, Shen A, Ding J, *et al.*: **Molecularly targeted cancer therapy: some lessons from the past decade.** *Trends Pharmacol Sci.* 2014; **35**(1): 41–50.
PubMed Abstract | Publisher Full Text

41. Lever J, Krzywinski M, Altman N: **Points of Significance: Classification evaluation.** *Nat Methods.* 2016; **13**: 603–604.
Publisher Full Text

42. Haibe-Kains B, El-Hachem N, Birkbak NJ, *et al.*: **Inconsistency in large pharmacogenomic studies.** *Nature.* 2013; **504**(7480): 389–93.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Li H, Leung KS, Wong MH, *et al.*: **Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets.** *Mol Inform.* 2015; **34**(2–3): 115–126.
PubMed Abstract | Publisher Full Text

44. Felip E, Martinez P: **Can sensitivity to cytotoxic chemotherapy be predicted by biomarkers?** *Ann Oncol.* 2012; **23**(Suppl 10): x189–92.
PubMed Abstract | Publisher Full Text

45. Ejlertsen B, Jensen MB, Nielsen KV, *et al.*: ***HER2*, *TOP2A*, and TIMP-1 and responsiveness to adjuvant anthracycline-containing chemotherapy in high-risk breast cancer patients.** *J Clin Oncol.* 2010; **28**(6): 984–90.
PubMed Abstract | Publisher Full Text

46. de Gramont AA, Watson S, Ellis LM, *et al.*: **Pragmatic issues in biomarker evaluation for targeted therapies in cancer.** *Nat Rev Clin Oncol.* 2015; **12**(4): 197–212.
PubMed Abstract | Publisher Full Text

47. Tran B, Dancey JE, Kamel-Reid S, *et al.*: **Cancer genomics: technology, discovery, and translation.** *J Clin Oncol.* 2012; **30**(6): 647–60.
PubMed Abstract | Publisher Full Text

48. Ahmed J, Meinel T, Dunkel M, *et al.*: **CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge.** *Nucleic Acids Res.* 2011; **39**(Database issue): D960–D967.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Boutros PC, Margolin AA, Stuart JM, *et al.*: **Toward better benchmarking: challenge-based methods assessment in cancer genomics.** *Genome Biol.* 2014; **15**(9): 462.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✔ ✔

Version 1

Referee Report 13 February 2017

doi:10.5256/f1000research.11347.r19182

✔ **Ronnie Alves** [1,2]

[1] Federal University of Pará , Graduate Program in Computer Science, Belém, Brazil

[2] Instituto Tecnologico Vale, Belém, Brazil

The authors present an empirical modeling approach, highlighting the pros and cons of using a robust machine learning strategy to pan-cancer cell line prediction solely based on gene expression profiles. Single-gene models may not provide an efficient solution on a such high dimensional data, therefore, multi-gene models have been used as a potential alternative to handle the combinatorial space (many candidate gene alterations). Even though the authors introduce quite well the GDSC pharmacotranscriptomic data, I would suggest the addition of more information regarding the baseline / benchmark regarding this classification problem. What would be an acceptable performance prediction? The evolution from single-gene to multi-gene classification could be improved along the feature engineering adopted within these classification strategies. The authors could motivate more the choice of the Random Forest (RF) technique. MANOVA has the problems of handling correlations among dependent variables, and effect size of these correlations. RF provides some improvement on MANOVA's limitations, however, it might suffer from feature subsampling selection and consequently, can overestimate the classification. The author could take a look at this work (Impact of subsampling and pruning on random forests) by Roxane Duroux and Erwan Scornet.[1]

Decision tree models are quite tight on training data. Given that the authors used the R language, there are many possibilities of tuning parameters along RF. Importance plots and partial plots could allow to expose features that can help to understand key features of a multi-gene model based on RF. Even though RF has a good performance, one may observe (Figure 4.A and 4.B) that there are some instances where MANOVA is better. The authors could share some light on these observations. Why are those ones hard to classify for RF? Regarding the GDSC data, it is not clear, while splitting the data, whether the data is well balanced along all drugs or not. The authors did well in keeping an independent test data, and it would be interesting to share more information regarding class (127 drugs) distribution along training and test data.

It would be great of the authors to provide the data and model, so other researchers are able to fully reproduce this study, as well as, devise other robust ensemble learning techniques that might be as good as RF.

This is an original work and it can be the first, indeed, proposing a benchmark on the estimation of the importance of somatic mutations in drug sensitivity classification.

References
1. Duroux R, Scornet E: Impact of subsampling and pruning on random forests. *HAL CCSD*. 2016.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 08 Mar 2017

Pedro Ballester, INSERM, France

The authors present an empirical modeling approach, highlighting the pros and cons of using a robust machine learning strategy to pan-cancer cell line prediction solely based on gene expression profiles. Single-gene models may not provide an efficient solution on a such high dimensional data, therefore, multi-gene models have been used as a potential alternative to handle the combinatorial space (many candidate gene alterations). Even though the authors introduce quite well the GDSC pharmacotranscriptomic data, I would suggest the addition of more information regarding the baseline / benchmark regarding this classification problem. What would be an acceptable performance prediction?
A drug sensitivity model with MCC>0 on the independent test set can be regarded as an acceptable performance because the performance of classifying cell lines at random is MCC=0. In the context of this study, we are interested in those acceptable models with a MCC value higher than that provided by the best single-gene marker of the same drug (i.e. models with positive MCC in the lower triangular part of Figure 6C). This is now stated in page 8.

The evolution from single-gene to multi-gene classification could be improved along the feature engineering adopted within these classification strategies. The authors could motivate more the choice of the Random Forest (RF) technique. MANOVA has the problems of handling correlations among dependent variables, and effect size of these correlations. RF provides some improvement on MANOVA's limitations, however, it might suffer from feature subsampling selection and consequently, can overestimate the classification. The author could take a look at this work (Impact of subsampling and pruning on random forests) by Roxane Duroux and Erwan Scornet. Decision tree models are quite tight on training data. Given that the authors used the R language, there are many possibilities of tuning parameters along RF.
Thanks for the suggestion. In page 5, we now state that RF is also robust to overfitting, as evidenced in Figure 1. In our experience, going deeper into the tuning control parameters for RF only brings marginal improvements in performance, although it could certainly be interesting from a theoretical point of view.

Importance plots and partial plots could allow to expose features that can help to understand key features of a multi-gene model based on RF.
We agree with the reviewer. However, we think that properly looking at the feature selection/importance question for each of the 127 drugs would require a separated study.

Even though RF has a good performance, one may observe (Figure 4.A and 4.B) that there are some instances where MANOVA is better. The authors could share some light on these observations. Why are those ones hard to classify for RF?
This is certainly an interesting question. For example, Figure 6C shows that 33.9% of the drugs are

harder to classify by a multi-variate RF model in the sense that a univariate model performs better. In page 3, we explained that the high dimensionality of the training data sets poses a challenge to classifiers and that these difficulties are drug-dependent. This is due to a number of convoluted factors. First of all, while both models look at the same data in each drug, each model employs a different set of features (genomic vs transcriptomic). Therefore, a single gene mutation might be more predictive of drug sensitivity than a model based on gene expression values in some cases. Second, only a very small subset of features might be predictive of cell line sensitivity to a given drug, which could be challenging for a RF using all the transcriptomic features. Third, the size of training and test sets varies because each drug was tested with a different number of cell lines. Consequently, class imbalances in training set and test set are also different depending on the drug. We are now stating these factors in page 3.

Regarding the GDSC data, it is not clear, while splitting the data, whether the data is well balanced along all drugs or not. The authors did well in keeping an independent test data, and it would be interesting to share more information regarding class (127 drugs) distribution along training and test data.
We completely agree with the reviewer in that it is essential to keep an independent test set to avoid overestimating performance, as standard k-fold cross-validation has been used for both model selection and performance assessment. Full information about the proportion of sensitive and resistant cell lines for each drug can be found in the data sets output by the released software (see below). One can see that training and/or test sets are not well balanced for some drugs and therefore more predictive RF models are likely to be obtained by using strategies to correct for class imbalances. However, the composition of training and test sets should not be altered, as this arise from a time-stamped partition and thus permit a realistic assessment of the performance that can be expected on future data sets (perhaps class imbalanced).

It would be great of the authors to provide the data and model, so other researchers are able to fully reproduce this study, as well as, devise other robust ensemble learning techniques that might be as good as RF.
We have now released the requested R script that was used to facilitate the construction of alternative machine learning models and their validation in the presented benchmark. This is available at http://ballester.marseille.inserm.fr/gdsc.transcriptomicDatav2.tar.gz. We hope that this release will facilitate further improvements on this class of problems.

This is an original work and it can be the first, indeed, proposing a benchmark on the estimation of the importance of somatic mutations in drug sensitivity classification.
We thank the reviewer for his positive assessment of this study. The released software implements this benchmark comprising 127 binary classification problems, one per drug. As drug response data is continuous, it is also possible to use the software to benchmark regression models. Furthermore, the software outputs the results of our study and hence these can be employed as a performance baseline for comparison to the results obtained by the benchmarked models.

Competing Interests: No competing interests were disclosed.

Referee Report 08 February 2017

doi:10.5256/f1000research.11347.r20033

Marc Poirot
The Cancer Research Center of Toulouse, INSERM (French National Institute of Health and Medical Research) , Toulouse, France

This is an original research article reporting a machine learning approach exploiting gene expression profiles to predict pan-cancer cell line sensitivity to drugs.
The title is appropriate and the abstract represent a suitable summary of the work
I would however suggest that the authors define the abbreviation “GDSC”.
The paper is well written, the experimental design is good and the conclusions fit with the data.

Minor point: page 7 figure 3 legend, RC=0.37 is not of the figure: this must be corrected

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 08 Mar 2017

Pedro Ballester, INSERM, France

This is an original research article reporting a machine learning approach exploiting gene expression profiles to predict pan-cancer cell line sensitivity to drugs.
The title is appropriate and the abstract represent a suitable summary of the work
I would however suggest that the authors define the abbreviation “GDSC”.
The paper is well written, the experimental design is good and the conclusions fit with the data.
We thank the reviewer for his positive appraisal of this article. As suggested, we have now defined the abbreviation “GDSC” in the abstract too.

Minor point: page 7 figure 3 legend, RC=0.37 is not of the figure: this must be corrected
Thanks for the observation. RC=0.37 was in mentioned in the part B of the caption, but referred to the part A of the figure. As this is confusing, it has been rewritten to specify that RC=0.37 is the test set recall of the AZ628 (part A) and it compared to that of Sunitinib (RC=0.75) in part B.

Competing Interests: No competing interests were disclosed.