Unearthing new genomic markers of drug response by improved measurement of discriminative power

Cuong C. Dang, Antonio Peón & Pedro J. Ballester*

Cancer Research Center of Marseille, INSERM U1068, F-13009 Marseille, France; Institut Paoli-Calmettes, F-13009 Marseille, France; Aix-Marseille Université, F-13284 Marseille, France; and CNRS UMR7258, F-13009 Marseille, France

* correspondence to: pedro.ballester@inserm.fr

Abstract

Oncology drugs are only effective in a small proportion of cancer patients. To make things worse, our current ability to identify these responsive patients before treatment is still very limited. Thus, there is a pressing need to discover response markers for marketed and research oncology drugs in order to improve patient survival, reduce healthcare costs and enhance success rates in clinical trials. Screening these drugs against a large panel of cancer cell lines has been recently employed to discover new genomic markers of *in vitro* drug response. However, while the identification of these markers among several thousands of candidate drug-gene associations is error-prone, an appraisal of the effectiveness of such detection task is currently lacking.

Here we propose a new approach that directly measures the discrimination power of a drug-gene association by posing each of these associations as a binary classification problem. The application of this methodology has led to the identification of 232 new genomic markers distributed across 81% of the analysed drugs, including 8 drugs without previously known markers, which were missed by the methodology initially applied to the Genomics of Drug Sensitivity in Cancer (GDSC) dataset.

Introduction

Cancer is a leading cause of morbidity and mortality in industrialised nations, with failed treatment being often life-threatening. While a wide range of drugs are now available to treat cancer patients, in practice only about 25% of them respond to these drugs[1]. To make things worse, our current ability to identify responsive patients before treatment is still very limited[2]. This situation has a negative impact on patient survival (the tumour keeps growing until an effective drug is administered), healthcare costs (very expensive drugs are ineffective, and thus wasted, on 75% of cancer patients[1,3]) and the success rates of oncology clinical trials (10% fall in Phase II studies, with the number of phase III terminations doubling in recent years[4]). Therefore, there is a pressing need to understand and predict this aspect of human variation to make therapy safer and more effective by determining which drugs will be more appropriate for any given patient.

The analysis of tumour and germline DNA has been investigated as a way to personalise cancer therapies for quite some time[5]. However, the recent and comprehensive flood of new data from much cheaper and faster Next Generation Sequencing (NGS) technologies along with the maturity of more established molecular profiling technologies represents an unprecedented opportunity to study the molecular basis of drug response. These data have shown that drug targets often present genomic alterations across patient tumours[6]. At the molecular level, these somatic mutations affect the abundance and function of gene products driving tumour growth and hence may influence disease outcome and/or response to therapy[7]. Therefore, there is opportunity for genetic information to aid the selection of effective therapy by relating the molecular profile of tumours to their observed sensitivity to drugs. Research on the identification of drug-gene associations that can be used as predictive biomarkers of drug response is carried out on human cancer tumour-derived cell lines[8–10]. Cell lines allow relatively quick and cheap experiments that are generally not feasible on more accurate disease models[11]. Here the molecular profile of the untreated cell line is determined and a phenotypic readout is made to assess the intrinsic cell sensitivity or resistance to the tested drug. In addition to biomarker discovery[8–10], these data sets have also been used to enable pharmacogenomic modelling[12–14], pharmacotranscriptomic modelling[15,16], QSAR modelling[17,18], drug repositioning[18,19] and molecular target identification[19–21], among other applications.

Our study focuses on the Genomics of Drug Sensitivity in Cancer (GDSC) data analysed by Garnett et al.[9] and publicly released after additional curation[22]. The released data set comprises 638 human tumour cell lines, representing a broad spectrum of common and rare cancer types. One benefit of looking at a large number of cell lines is that the pool of data becomes larger, which is beneficial for biomarker discovery. These authors profiled each cell line for various genetic abnormalities, including point mutations, gene amplifications, gene deletions, microsatellite instability, frequently occurring DNA rearrangements and changes in gene expression. Thereafter, the sensitivity of 130 drugs against these cell lines was measured with a cell viability assay in vitro (cell sensitivity to a drug was summarised by the half-maximal inhibitory concentration or IC_{50} of the drug-cell pair). A p-value was calculated for 8637 drug-gene associations using a MANOVA test (P_{MANOVA}), with 396 of those associations being above a FDR=20% Benjamini-Hochberg[23] adjusted threshold and thus deemed significant (full details in the Methods section). Overall, it was found that only few drugs had strong genomic markers, with no actionable mutations being identified for 14 drugs.

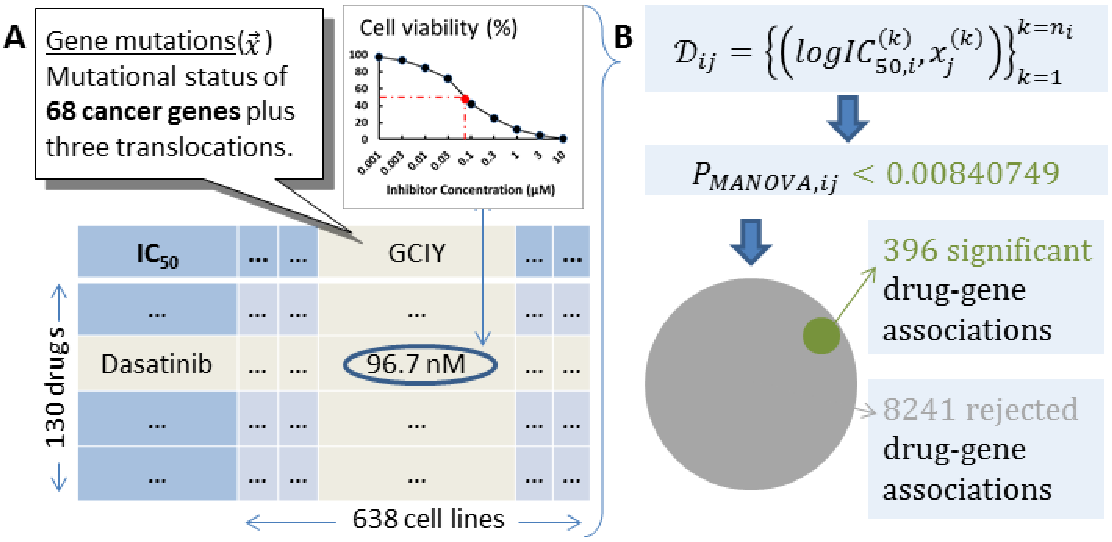

A
Gene mutations($\vec{x}$)
Mutational status of 68 cancer genes plus three translocations.
Cell viability (%)
100
80
60
40
20
0
0.001
0.003
0.01
0.03
0.1
0.3
1
3
10
Inhibitor Concentration (µM)
IC_{50}
...
GCIY
Dasatinib
96.7 nM
130 drugs
638 cell lines
B
$\mathcal{D}_{ij} = \left\{\left(logIC_{50,i}^{(k)}, x_j^{(k)}\right)\right\}_{k=1}^{k=n_i}$
$P_{MANOVA,ij} < 0.00840749$
396 significant drug-gene associations
8241 rejected drug-gene associations

Figure 1 | Released GDSC data. **(A)** Garnett et al. analysed a slightly different dataset than the one that was later released. In the released dataset, a panel of 130 drugs was tested against 638 cancer cell lines, leading to 47748 IC_{50} values (57.6% of all possible drug-cell pairs). For each cell line, 68 cancer genes were sequenced and their mutational status determined, plus three translocations and a microsatellite instability status. **(B)** A dataset D_{ij} can be compiled for each drug-gene combination comprising the n_i cell responses to the i^{th} drug (in our case, each response as the logarithm base 10 of IC_{50} in µM units), with $x_j^{(k)}$ being a binary variable indicating whether the j^{th} gene is mutated or not in the k^{th} cell line. Next, a p-value was calculated for each drug-gene pair using the MANOVA test. Those pairs with p-values below an adjusted threshold of 0.00840749 were considered statistically significant (396 of the 8637 drug-gene associations).

However, a statistically significant drug-gene association is not necessarily a useful genomic marker of *in vitro* drug response. Indeed, significant p-values are merely intended to highlight potential discoveries among thousands of possibilities, but their practical importance still have to be evaluated for the problem at hand[24–26]. In this context, the latter means assessing how well the gene mutation discriminates between sensitive and resistant cells to a given drug, which is only approximated by its p-value. A related consideration is that the p-value comes from a selected statistical test, which in practice is more or less appropriate to carry out this detection task[24].

Consequently, a p-value may lead to two types of errors at the inter-association level, a false discovery (type I error or false positive) or a missed discovery (type II error or false negative). The latter can be evaluated by directly measuring the discrimination offered by a drug-gene association and comparing it to its corresponding p-value. Here, it is worth noting that a false negative can have very negative consequences (e.g. missing a genomic marker able to identify tumours sensitive to a drug for which no marker have been found yet). Thus, research intended to identify more appropriate statistical procedures for biomarker discovery on comprehensive pharmacogenomic resources such as GDSC is crucial to make the most out of these valuable data.

Here we propose a methodology that directly measures the discrimination power of a drug-gene association by posing each of these associations as a binary classification problem. This change of perspective is enabled by the definition of a meaningful criterion to determine the sensitivity threshold for each association. Furthermore, we adopted a suitable statistical test to evaluate how likely was the corresponding classification of cells to arise by chance, which determines the set of markers with statistically significant discriminative power. Lastly, we applied this methodology to identify genomic markers from GDSC data and compare the results against those arising from the MANOVA test[9].

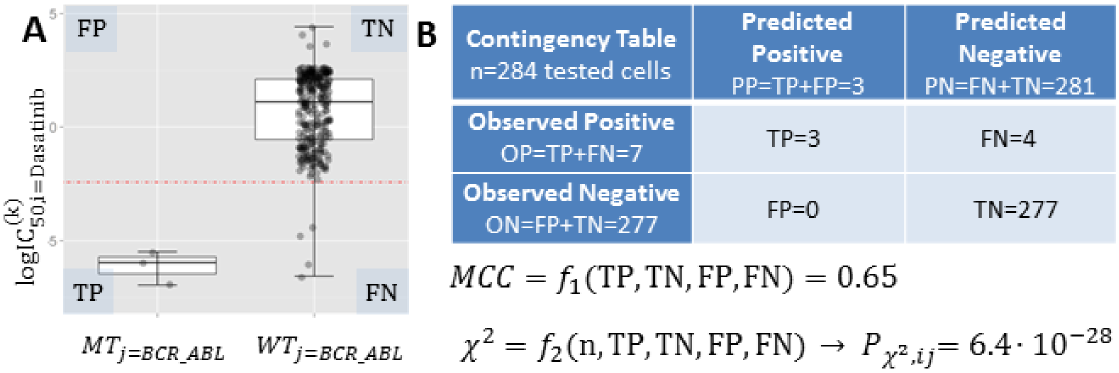
Results

Direct measurement of discriminative power

Response biomarkers aim at discriminating between sensitive and resistant tumours to a given drug. Consequently, direct measurement of such discriminative power requires the definition of an IC_{50} threshold above/below which tumours are considered to be resistant/sensitive to the drug, which effectively poses biomarker evaluation as a binary classification problem. The latter will permit to employ a number of common performance metrics that are better suited to estimate and assess the prospective performance of these biomarkers. The first issue that needs to be address is how to define the IC_{50} threshold. A fixed threshold for all drugs, e.g. 1µM, is not useful as drugs often have a very different distribution of IC_{50} measurements across cell lines. Instead, for each drug-gene association, we characterise each group of cells, i.e. those with the mutated gene and those with the wild-type (WT) gene, by its median IC_{50} and define the threshold as the mean of both medians (e.g. the dotted red line of the scatter plot in Figure 2A). In this way, the size of each group and their outliers do not distort the position of this decision boundary, which is equidistant to both classes and leads to an intuitive notion of class membership as distance from the threshold.

Once the IC_{50} threshold is calculated, the mutation-based prediction of drug response of a cell line can be categorised as a true positive (TP), true negative (TN), false positive (FP) or false negative (FN). From this contingency table at the intra-association level, the discrimination offered by a drug-gene association can be summarised by its Matthews Correlation Coefficient (MCC)[27]. Because the definition of a positive instance depends on whether the somatic mutation is sensitising or resistant (see the Methods section), MCC can only take values from 0 (gene mutation have absolutely no discriminative power) to 1 (gene mutation perfectly predicts whether cell lines are sensitive or resistant to the drug). Furthermore, since cells are now partitioned into four non-overlapping categories with respect to their response to a drug, the χ^2 test can be computed from this 2x2 contingency table to identify those drug-gene associations with statistically significant discriminative power (the χ^2 statistic measures how far is the contingency table obtained by the classification method from the values that would be expected by chance). The process is sketched in

Figure 2 and leads to an alternative set of p-values from the χ^2 test ($P_{\chi 2}$). To establish which associations are significant according to the χ^2 test, we also calculated for this case the FDR=20% Benjamini-Hochberg adjusted threshold (0.00940155). Overall, 403 statistically significant drug-gene associations were found using the χ^2 test from the same set of 8637 associations that were downloaded (i.e. seven significant associations more than with the MANOVA test). Importantly, only 171 associations are deemed statistically significant by both tests. These discrepancies will be investigated in the next section to unveil false and missed biomarkers.



Contingency Table n=284 tested cells	Predicted Positive PP=TP+FP=3	Predicted Negative PN=FN+TN=281
Observed Positive OP=TP+FN=7	TP=3	FN=4
Observed Negative ON=FP+TN=277	FP=0	TN=277

$$MCC = f_1(\mathrm{TP},\mathrm{TN},\mathrm{FP},\mathrm{FN}) = 0.65$$

$$\chi^2 = f_2(\mathrm{n},\mathrm{TP},\mathrm{TN},\mathrm{FP},\mathrm{FN}) \rightarrow P_{\chi^2,ij} = 6.4\cdot 10^{-28}$$

Figure 2 | Measuring the discriminative power of a genomic marker with MCC and the χ^2 test. (A) Scatter plot showing the $\log IC_{50}$ of n=284 cell lines tested with the marketed drug Dasatinib. The left boxplot shows BCR_ABL positive cell lines, whereas the boxplot on the right shows cell lines without this mutation (the median of each group appears as a black horizontal line within the boxplot). The red dotted line is the IC_{50} threshold, which is defined as the mean of both medians. **(B)** Contingency table showing the number of cell lines in each of the four non-overlapping categories (TP, FN, FP, TN), where positives are cell lines below the threshold in the case of a sensitising mutation (above the threshold if the mutation induces resistance). MCC and χ^2 are functions of these metrics and summarise binary classification performance, as further described in the Methods section. BCR_ABL is a very strong marker of Dasatinib sensitivity as shown in the scatter plot and highlighted by both statistical tests ($P_{MANOVA}=1.4\cdot 10^{-10}$, $P_{\chi 2}=6.4\cdot 10^{-28}$), offering unusually high discrimination between sensitive and resistant cell lines (MCC=0.65).

A last aspect to discuss about the proposed methodology is the duality of MCC and χ^2. In statistics, MCC is known as the φ coefficient, which was introduced[28] by Yule in 1912 and later rediscovered[27] by Matthews in 1975 as the MCC (interestingly, despite being more recent, the MCC has become a much more popular metric for binary classification than the φ coefficient[29–34]). As $\chi^2 = n \cdot \varphi^2$ holds[28], so does $\chi^2 = n \cdot MCC^2$ with n being the number of tested cell lines for the considered drug and thus MCC will be highly correlated with $P_{\chi 2}$. Figure 3A presents the number of drug-gene associations for each number of tested cell lines, from which it is observed that each drug has only been tested on a subset n of the 638 cell lines (i.e. gene associations for a given drug will be all evaluated on the same number of cell lines n). Two distinctive groups of drugs emerge: those tested on around 300 cell lines (red bars) and those tested around 450 cell lines (black bars). Figure 3B shows that MCC and -log$P_{\chi 2}$ are highly correlated even across different n (for associations with the same n, a perfect Pearson and Spearman correlation is obtained as expected – data not shown). Given the observed distribution of n values, all markers with an MCC of about 0.15 or more are too discriminative to have arisen by chance. This connexion is useful in that MCC is widely used[29–34] but without establishing its statistical significance for the tackled problem instance.

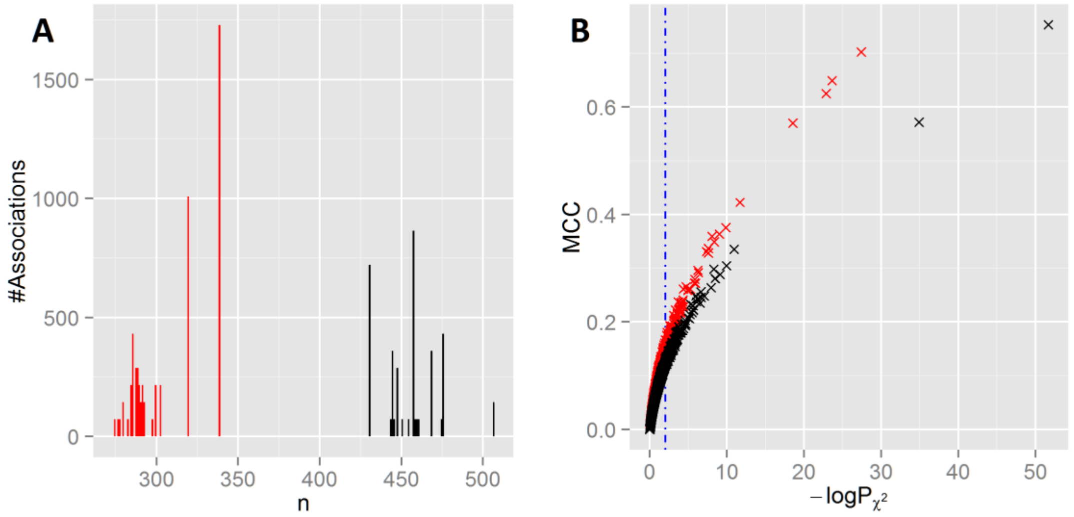


Figure 3 | MCC vs. -logP$_{\chi2}$ across all the 8637 drug-gene associations from GDSC. **(A)** Number of drug-gene associations for each number of tested cell lines (n). Two distinctive groups of drugs emerge: those tested on around 300 cell lines (red bars) and those tested around 450 cell lines (black bars). **(B)** MCC versus -logP$_{\chi2}$ across the drug-gene associations (same colour code). The Spearman and Pearson correlations between both metrics are 0.99 and 0.82, respectively. The plot shows that all markers with an MCC of around 0.15 or more are too discriminative to have arisen by chance (above an MCC of 0.12 if we restrict to the markers evaluated with more data shown as black crosses).

False-positive and false-negative markers

We have shown that the introduction of a meaningful threshold for sensitivity permits the direct measurement of the discriminative power of a drug-gene association using the MCC along with its significance using P$_{\chi2}$. We analyse next those associations with the largest discrepancies between the p-values from both statistical tests against the ground truth provided by their scatter plots. First, we identified the association with the largest difference between P$_{MANOVA}$ and P$_{\chi2}$ among those not significant by the χ^2 test, which should therefore be an error of the MANOVA test (a false positive). Indeed, the left scatter plot in Figure 4 evidences that this drug-gene association (GW441756-FLT3) discriminates poorly between cell lines despite a very low

P_{MANOVA}~10^{-10}. In contrast, a high $P_{\chi2}$~10^{-1} is obtained which means that the χ^2 test correctly rejected this false positive of the MANOVA test.

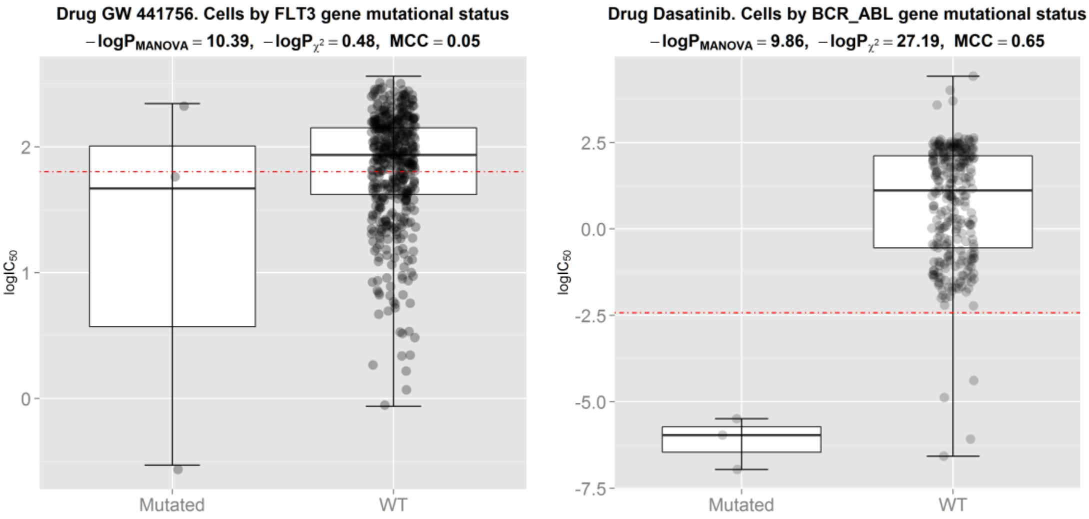


Figure 4 | False-positive marker correctly rejected by the χ^2 test. (**left**) The scatter plot for the drug-gene association (GW441756-FLT3) with the largest -logP_{MANOVA} among those not significant according to the χ^2 test. Hence, mutated-FLT3 is a marker of sensitivity to the experimental drug GW441756 according to the MANOVA test, but not according to the χ^2 test. However, this marker offers practically no discriminative power as indicated by a MCC of just 0.05 and evidenced by the scatter plot. Therefore, the χ^2 test correctly rejected this false positive of the MANOVA test. (**right**) Conversely, to assess whether the χ^2 test could also lead to strong false positives, we searched for the drug-gene association with largest -log$P_{\chi2}$ among those with a similar -logP_{MANOVA} to that of GW441756-FLT3, which is Dasatinib-BCR_ABL. Whereas the p-value for Dasatinib-BCR_ABL is of the same magnitude as that for GW441756-FLT3 using the MANOVA test (P_{MANOVA}~10^{-10}), the p-values for the same associations using the χ^2 test differ is almost 27 orders of magnitude. Thus, unlike the MANOVA test, the χ^2 test correctly detects the very high difference in discriminative power offered by these two drug-gene associations. Indeed, the BCR_ABL translocation is a highly discriminative marker of Dasatinib sensitivity (MCC=0.65), as also evidenced by the scatter plot.

Conversely, to assess whether the χ^2 test also leads to strong false positives, we searched for the drug-gene association with smallest $P_{\chi2}$ among those with a similar P_{MANOVA} to that of GW441756-FLT3, which is Dasatinib-BCR_ABL (Figure 4 right). If this association was a false positive of the χ^2 test, the scatter plot would offer poor discrimination between BCR_ABL positive and WT cell lines. However, the opposite is observed, as the BCR_ABL translocation is a highly discriminative marker of Dasatinib sensitivity (MCC=0.65). Note that, whereas the p-value for Dasatinib-BCR_ABL is of the same magnitude as that for GW441756-FLT3 using the MANOVA test (P_{MANOVA}~10^{-10}), the p-values for the same associations using the χ^2 test differ is almost 27 orders of magnitude. Thus, unlike the MANOVA test, the χ^2 test correctly detects the very high difference in discriminative power offered by these two drug-gene associations.

The next experiment consists in searching for the largest discrepancy in the opposite direction. First, we identified the association with the largest difference between P_{MANOVA} and $P_{\chi2}$, this time among those not significant by the MANOVA test. Thus, this is expected to be an error of the MANOVA test (a false negative). The left scatter plot in Figure 5 evidences that this drug-gene association (Dasatinib-CDKN2a.p14) is actually a false negative of the MANOVA test, as it offers good discrimination between mutant and WT cell lines despite a high P_{MANOVA}~10^{-1}. In contrast, a low $P_{\chi2}$~10^{-9} is obtained which means that the χ^2 test correctly detected this false negative of the MANOVA test. Conversely, to assess whether the χ^2 test could also lead to strong false negatives, we searched for the drug-gene association with smallest P_{MANOVA} among those with a similar $P_{\chi2}$ to that of Dasatinib-CDKN2a.p14, which is SB590885-BRAF (Figure 5 right). Whereas the p-values for Dasatinib-CDKN2a.p14 and SB590885-BRAF differ 27 orders of magnitude using the MANOVA test, the p-values for the same associations have similar p-values using the χ^2 test ($P_{\chi2}$~10^{-9}). Thus, unlike the MANOVA test, the χ^2 test correctly detected that both markers have similar discriminative power as also highlighted by the MCC (SB590885-BRAF has a MCC of 0.29 for 0.35 of Dasatinib-CDKN2a.p14).

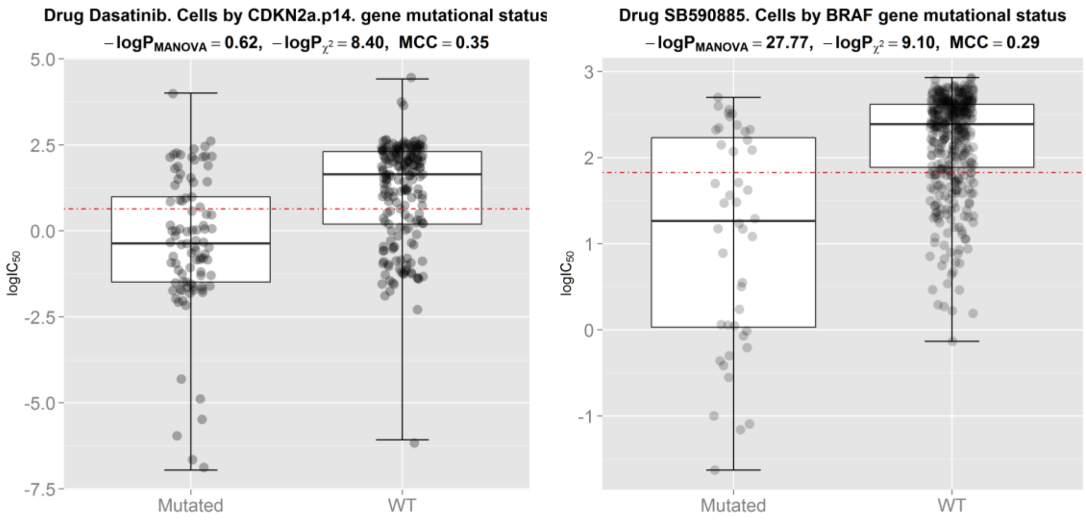


Figure 5 | False-negative marker correctly detected by the χ^2 test. (**left**) The scatter plot for the drug-gene association (Dasatinib-CDKN2a.p14) with the largest -log$P_{\chi2}$ among those not significant according to the MANOVA test. Hence, mutated-CDKN2a.p14 is a marker of sensitivity to the marketed drug Dasatinib according to the χ^2 test, but not according to the MANOVA test. However, this marker is highly discriminative as proven by a MCC of 0.35 and evidenced by the scatter plot. Therefore, the χ^2 test correctly detected this false negative of the MANOVA test. (**right**) Conversely, to assess whether the χ^2 test could also lead to strong false negatives, we searched for the drug-gene association with largest -logP_{MANOVA} among those with a similar -log$P_{\chi2}$ to that of Dasatinib-CDKN2a.p14, which is SB590885-BRAF. Whereas the p-values for Dasatinib-CDKN2a.p14 and SB590885-BRAF differ 27 orders of magnitude using the MANOVA test, the p-values for the same associations have similar p-values using the χ^2 test ($P_{\chi2}$~10^{-9}). Thus, unlike the MANOVA test, the χ^2 test correctly detected that both markers have similar discriminative power (SB590885-BRAF has a MCC of 0.29 for 0.35 of Dasatinib-CDKN2a.p14).

218 new markers found for 97 of the 116 drugs with previously known markers

Having established that the χ^2 test is able to identify the false negatives of the MANOVA test in theory and practice, the rest of the study will focus on unearthing these missed discoveries. Indeed, these new genomic markers constitute additional knowledge that can be extracted from existing data, i.e. without requiring any further experiment. In the data released by the GDSC, the 396 genomic markers from the

MANOVA test were distributed among 116 drugs, leaving the remaining 14 drugs without any maker. In this subsection, we analyse these 116 drugs with the χ^2 test, whereas the next subsection will deal with the 14 drugs currently without markers.

The χ^2 test found a total of 218 new makers for 97 drugs from these 116 drugs (S1 File). Figure 6 shows two examples. The scatter plot at the top left presents the EWS_FLI1 translocation as a new marker of sensitivity to Mitomycin C, which was missed by the MANOVA test. This marker offers substantially more discrimination than some of the previously known Mitomycin C markers suggested by the MANOVA test, which are actually false positives (e.g. the scatter plot for the MET gene mutation on the right). The second example is shown at the bottom of Figure 6. The EWS_FLI1 translocation is also a new response marker for the drug BMS-754807, which was again missed by the MANOVA test. This marker offers substantially more discrimination than some of the previously known BMS-754807 markers suggested by the MANOVA test, which are false positives as well (e.g. the scatter plot for the PTEN gene mutation on the right). Furthermore, in 45 of these 116 drugs, the new marker offers better discrimination than the best previously known marker for the drug (S2 File).

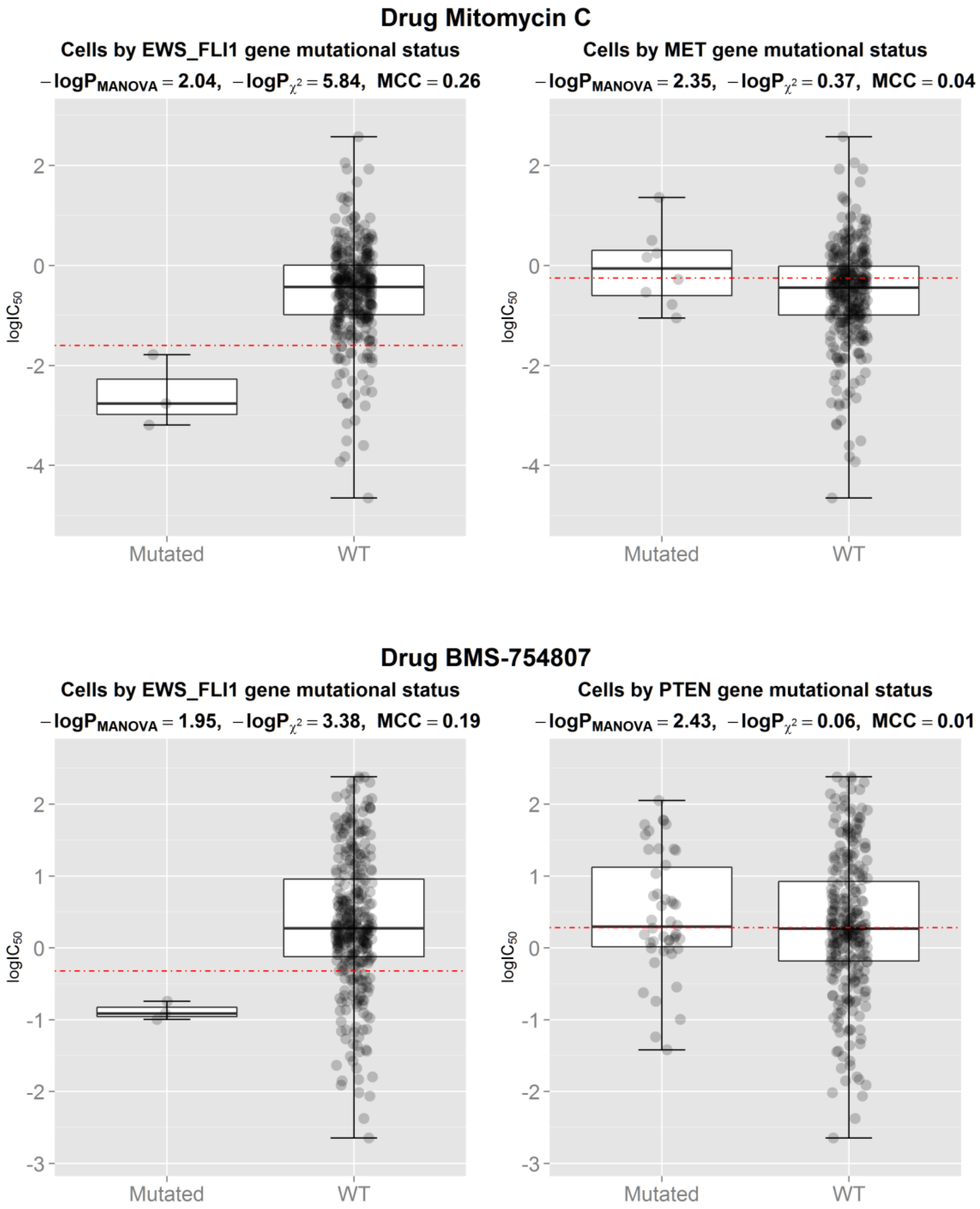


Figure 6 | Examples of new genomic markers for drugs with previously known markers. **(top)** On the left, the EWS_FLI1 translocation is found to be the most discriminative biomarker for the approved drug Mitomycin C (MCC=0.26, $P_{\chi 2}$=$1.5 \cdot 10^{-6}$), which was missed by the MANOVA test (P_{MANOVA}=$9.1 \cdot 10^{-3}$). The latter contrasts with MANOVA-significant association with MET gene mutation (P_{MANOVA}=$4.5 \cdot 10^{-3}$), which is barely discriminative (MCC=0.04) and thus rejected by the χ^2 test ($P_{\chi 2}$=0.4). **(bottom)** On the left, the EWS_FLI1 translocation is found to be the most discriminative biomarker for the development drug BMS-754807 (MCC=0.19, $P_{\chi 2}$=$4.1 \cdot 10^{-4}$), which was also missed by the MANOVA test (P_{MANOVA}=0.011). The latter contrasts with MANOVA-significant association with PTEN gene mutation (P_{MANOVA}=3.7·10^{-3}), which offers practically no discrimination (MCC=0.01) and thus strongly rejected by the χ^2 test ($P_{\chi2}$=0.86). For both drugs, the two plots on the left show two false negatives of the MANOVA test, whereas the two plots on the right illustrates two false positives of the same statistical test.

14 new markers found for 8 of the 14 drugs without previously known markers

New genomic markers are particularly valuable in those drugs for which no marker is known yet. By applying the χ^2 test to the same dataset, we have identified 14 new markers for the 8 drugs for which the MANOVA test did not find any marker (S3 File). Figure 7 shows two of these markers. On the right, the mutational status of the SMAD4 gene is the most discriminative biomarker for the development drug BI-2536. On the left, the EWS_FLI1 translocation is found to sensitise cell lines to Gemcitabine.

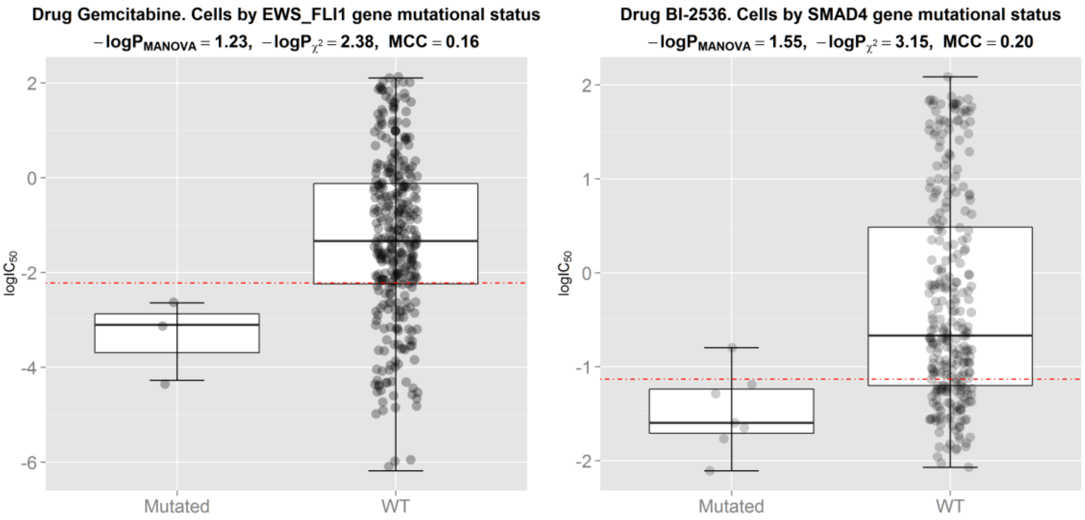


Figure 7 | Examples of new genomic markers for drugs without previously known markers. **(left)** The EWS_FLI1 translocation is found to be the most discriminative biomarker for the approved drug Gemcitabine (MCC=0.16, $P_{\chi2}$=4.2·10^{-3}), which was missed by the MANOVA test (P_{MANOVA}=0.059). **(right)** The SMAD4 gene mutation is found to be the most discriminative biomarker for the development drug BI-2536 (MCC=0.2, $P_{\chi2}$=7.1·10^{-4}), which was also missed by the MANOVA test (P_{MANOVA}=0.028).

Discussion

The assessment of MANOVA-significant drug-gene associations against the ground truth shown in scatter plots has been revealing. For instance, it is now evident that selecting PTEN-mutated tumour cells do not result in a substantially different response to BMS-754807, as the distribution of IC_{50}s across PTEN-mutated cell lines is essentially the same as that for cell lines with WT PTEN (Figure 6 bottom right). Thus, mutated PTEN is not useful as a predictive biomarker of response to BMS-754807 despite its statistically significant p-value from the MANOVA test.

To improve the search of genomic markers of drug response, we have proposed a new approach that directly measures the discrimination power of a drug-gene association by posing each of these associations as a binary classification problem. Here discrimination is measured with χ^2 statistic and its significance with $P_{\chi2}$, which has resulted in a better alignment of the statistical and biological significance of a drug-gene association. Furthermore, we have shown that, since MCC is linked to χ^2, the significance of an MCC value can also be calculated with the χ^2 test. This is useful in that MCC is widely used but without establishing its statistical significance for the problem at hand.

Next, the χ^2 test has been applied to the identification of genomic markers from GDSC data and these markers compared to those arising from the MANOVA test[9]. The largest discrepancies arising from both sets of p-values have been discussed in detail. Figures 4 and 5 provide examples of false positives and false negatives of the MANOVA test, which illustrates the theoretical adequacy of the χ^2 test for this task in practice. In addition, we have carried out a systematic comparison across 8637 drug-gene associations for which a p-value from the MANOVA test had been calculated in the GDSC study[9]. The MANOVA test highlighted 396 of these associations as statistically significant, for 403 from the χ^2 test looking at the same data. However, only 171 associations were deemed statistically significant by both tests. Having established the χ^2 statistic as a direct measure of discrimination both in theory and practice, it follows that the 225 associations that are only significant by the MANOVA test are false positives at the inter-association level. From a translational perspective, these false discoveries can potentially lead to wasting resources on follow-up experiments on more accurate disease models, although one can always visualise the corresponding scatter plot prior to decision-making (Figure 4 left).

On the other hand, there are 232 associations that were only detected by the χ^2 test and hence are false negatives of the MANOVA test. These missed discoveries are easy to miss as it is not possible to visualise thousands of rejected markers, which highlights the value of the proposed approach. 218 of these new 232 drug response markers were found in 97 of the 116 drugs with known markers (see examples in Figure 6), which represent markers that could have a higher translational potential than those already known for a drug. The remaining 14 markers were for 8 of the 14 drugs without previously known markers (see examples in Figure 7), which are hence particularly valuable. Overall, we have identified new genomic markers for 105 of the 130 drugs (81%). In 53 of these 105 drugs, the genomic marker was more discriminative than the best among the previously known for the drug.

Predictive biomarkers are increasingly important tools in drug development and clinical research[45,46]. During the development of methods for cancer diagnosis and treatment, a vast amount of cancer genomics data is now being generated[47] and thus there is an urgent need for their accurate analysis[48]. Therefore, this study is important in a number of ways. First, these new genomic markers of *in vitro* drug response represent testable hypothesis that can now be evaluated on more relevant disease models to humans. Second, they may also constitute further evidence supporting newly proposed oncology targets[44]. Third, beyond the exploitation of these results, the widespread application of this methodology should lead to the discovery of new predictive biomarkers on existing data as it has been the case with the GDSC. Indeed, this new approach has been demonstrated on a large-scale drug screening against human cancer cell lines, but it can also be applied to other biomarker discovery problems such as those adopting more accurate disease models (e.g. primary tumours[35,36], patient-derived xenografts[37] or patients[38,39]), those employing other molecular profiling data (e.g. secretome proteomics[40], epigenomics[41] or single-cell genomics[42]) or those involving drug combinations[43]. Looking more broadly, the methodology can also be applied to large-scale drug screening against human or non-human molecularly-profiled pathogen cultures, such as those in antibacterial or agricultural research.

Methods

GDSC data

From the Genomics of Drug Sensitivity in Cancer (GDSC) ftp server[22], we downloaded the following data files: gdsc_manova_input_w1.csv and gdsc_manova_output_w1.csv.

In gdsc_manova_input_w1.csv, there are 130 unique drugs as camptothecin was tested twice, drug ids 195 and 1003, and thus we only kept the instance that was more broadly tested (i.e. drug ID 1003 on 430 cell lines). Thus, effectively a panel of 130 drugs was tested against 638 cancer cell lines, leading to 47748 IC_{50} values (57.6% of all possible drug-cell pairs). Downloaded "IC_{50}" values are more precisely the natural logarithm of IC_{50} in µM units (i.e. negative values represent drug responses more potent than 1µM). We converted each of these values into their logarithm base 10 in µM units, which we denote as $logIC_{50}$ (e.g. $logIC_{50}$=1 corresponds to IC_{50}=10µM), as in this way differences between two drug response values are directly given as orders of magnitude in the molar scale.

gdsc_manova_input_w1.csv also contains genetic mutation data for 68 cancer genes, which were selected as the most frequently mutated cancer genes[9], characterising each of the 638 cell lines. For each gene-cell pair, a 'x::y' description was provided by the GDSC, where 'x' identifies a coding variant and 'y' indicates copy number information from SNP6.0 data. As in Garnett et al.[9], a gene for which a mutation is not detected is considered to be wild-type (wt). A gene mutation is annotated if: a) a protein sequence variant is detected ($x \neq \{wt,na\}$) or b) a deletion/amplication is detected. The latter corresponds to a copy number (cn) variation different from the wt value of $y=0<cn<8$. Furthermore, three translocations were considered (BCR_ABL, MLL_AFF1 and EWS_FLI1). For each of these gene fusions, cell lines are identified as fusion not-detected or the identified fusion is given (i.e. wt or mutated with respect to the gene fusion, respectively). The microsatellite instability (msi) status of each cell line was also determined. Full details can be found in the original publication[9].

Statistically significant drug-gene associations with the MANOVA test

gdsc_manova_output_w1.csv contains 8701 drug-gene associations with p-values. As we are considering all those involving the 130 unique drugs (i.e. removing the camptothecin duplicate), we are left with 8637 drug-gene associations with p-values of which 396 were above a FDR=20% Benjamini-Hochberg adjusted threshold (0.00840749) and thus deemed significant according to this test. Each statistically significant drug-gene association was considered to be a genomic marker of *in vitro* drug response[9].

Measuring the discriminative power of a genomic marker with MCC

Let the data for the association between the i^th^ drug and the j^th^ gene be

$$\mathcal{D}_{ij} = \left\{\left(logIC_{50,i}^{(k)}, x_j^{(k)}\right)\right\}_{k=1}^{k=n_i}$$

and the sets of mutated and wt cell lines with respect to the j^th^ gene, MT_j and WT_j, be

$$MT_j = \left\{k \mid x_j^{(k)} = 1\right\} \qquad WT_j = \left\{k \mid x_j^{(k)} = 0\right\}$$

Then, the $logIC_{50}$ threshold is defined as the mean of the median responses from each set (see subsection "Direct measurement of discriminative power"):

$$thres_{ij} = mean\left(median\left(\left\{logIC_{50,i}^{(k)}\right\}_{k\,\in MT_j}\right) + median\left(\left\{logIC_{50,i}^{(k)}\right\}_{k\,\in WT_j}\right)\right)$$

Now if the median response of the MT_j set is lower (i.e. more sensitive to the drug) than that of the WT_j set in the considered drug-gene association, then cell lines with $logIC_{50}$ values lower than the threshold (by this definition, cell lines sensitive to the drug) are positives and those with $logIC_{50}$ vales higher or equal than the threshold (i.e. cell lines resistant to the drug) are negatives. Conversely, if the median of the WT_j set is the lowest, then the positives are resistant cell lines and the negatives are sensitive cell lines. These cases correspond to candidate genomic markers of drug sensitivity and resistance, respectively.

At this point, the set of all the cell lines tested with a given drug can be partitioned into four categories as defined in Figure 2: true positive (TP), true negative (TN), false positive (FP) or false negative (FN). From this contingency table, the discrimination offered by a drug-gene association can be summarised by the Matthews Correlation Coefficient (MCC)[27]

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

By the above definition of positives and negatives, MCC can only take values from 0 (gene mutation have absolutely no discriminative power) to 1 (gene mutation perfectly predicts whether cell lines are sensitive or resistant to the drug). Also, note that both the definition of the logIC$_{50}$ threshold and the existence of mutated and wt cell lines in every association guarantees a non-zero value of the denominator in the MCC formula and thus MCC is always defined in this study.

Statistically significant drug-gene associations with the χ^2 test

For each of the 8637 drug-gene associations, the χ^2 test was computed from the 2x2 contingency table[49] to identify those drug-gene associations with statistically significant discriminative power. The formula to compute the χ^2 statistic is

$$\chi^2 = \sum_{l=1}^{2}\sum_{m=1}^{2}\frac{(O_{lm} - E_{lm})^2}{E_{lm}}$$

where O$_{lm}$ are the four categories in the table (TP,TN,FN,FP) and E$_{lm}$ are the corresponding expected values under the null hypothesis that this partition has arisen by chance. Thus, expected values are calculated with

$$E_{11} = E(TP) = PP\cdot\frac{OP}{n} \qquad E_{12} = E(FN) = PN\cdot\frac{OP}{n}$$

$$E_{21} = E(FP) = PP\cdot\frac{ON}{n} \qquad E_{22} = E(TN) = PN\cdot\frac{ON}{n}$$

For instance, the expected value of TP, E(TP), is the number of predicted positives (PP) times the probability of a cell being a positive given as the proportion of observed positives (OP) in the n tested cells.

This χ^2 statistic follows a χ^2 distribution with one degree of freedom and thus each p-value was computed with the R package *pchisq* from its corresponding χ^2 value, χ_0^2, as

$$P_{\chi^2} = pdf_{\chi^2}(\chi_0^2, df = 1)$$

The process is sketched in Figure 2 and leads to an alternative set of p-values from the χ^2 test ($P_{\chi 2}$). To establish which associations are significant according to the χ^2 test, we also calculated for this case the FDR=20% Benjamini-Hochberg adjusted threshold (0.00940155), that is

$$P_{\chi^2, ij} < 0.00940155$$

To facilitate reproducibility and the use of this methodology to analyse other pharmacogenomics data sets, the R script to calculate MCC, χ^2 and $P_{\chi 2}$ from gdsc_manova_input_w1.csv is available on request.

Author contributions

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

Acknowledgements

We thank Gustavo Stolovitzky (IBM Research, NY, USA) for feedback on the study. This work has been carried out thanks to the support of the A*MIDEX grant (n°

ANR-11-IDEX-0001-02) funded by the French Government «Investissements d'Avenir» program awarded to P.J.B.

References

1. Spear, B. B., Heath-Chiozzi, M. & Huff, J. Clinical application of pharmacogenetics. *Trends Mol. Med.* **7,** 201–204 (2001).

2. Huang, M., Shen, A., Ding, J. & Geng, M. Molecularly targeted cancer therapy: some lessons from the past decade. *Trends Pharmacol. Sci.* **35,** 41–50 (2014).

3. Luengo-Fernandez, R., Leal, J., Gray, A. & Sullivan, R. Economic burden of cancer across the European Union: a population-based cost analysis. *Lancet. Oncol.* **14,** 1165–74 (2013).

4. Deyati, A., Younesi, E., Hofmann-Apitius, M. & Novac, N. Challenges and opportunities for oncology biomarker discovery. *Drug Discov. Today* **18,** 614–624 (2013).

5. Wheeler, H. E., Maitland, M. L., Dolan, M. E., Cox, N. J. & Ratain, M. J. Cancer pharmacogenomics: strategies and challenges. *Nat. Rev. Genet.* **14,** 23–34 (2013).

6. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **464,** 993–8 (2010).

7. McLeod, H. L. Cancer Pharmacogenomics: Early Promise, But Concerted Effort Needed. *Science (80-.).* **339,** 1563–1566 (2013).

8. Abaan, O. D. *et al.* The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. *Cancer Res.* **73,** 4372–82 (2013).

9. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483,** 570–575 (2012).

10. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–307 (2012).

11. Weinstein, J. N. Drug discovery: Cell lines battle cancer. *Nature* **483,** 544–5 (2012).

12. Menden, M. P. *et al.* Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS One* **8,** e61318 (2013).

13. Ammad-ud-din, M. *et al.* Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization. *J. Chem. Inf. Model.* **54,** 2347–59 (2014).

14. Cortés-Ciriano, I. *et al.* Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. *Bioinformatics* btv529 (2015). doi:10.1093/bioinformatics/btv529

15. Riddick, G. *et al.* Predicting in vitro drug sensitivity using Random Forests. *Bioinformatics* **27,** 220–224 (2011).

16. Geeleher, P., Cox, N. J. & Huang, R. S. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol.* **15,** R47 (2014).

17. Lee, A. C., Shedden, K., Rosania, G. R. & Crippen, G. M. Data mining the NCI60 to predict generalized cytotoxicity. *J. Chem. Inf. Model.* **48,** 1379–88 (2008).

18. Kumar, R., Chaudhary, K., Singla, D., Gautam, A. & Raghava, G. P. S. Designing of promiscuous inhibitors against pancreatic cancer cell lines. *Sci. Rep.* **4,** 4668 (2014).

19. Holbeck, S. L., Collins, J. M. & Doroshow, J. H. Analysis of Food and Drug Administration-approved anticancer agents in the NCI60 panel of human tumor cell lines. *Mol. Cancer Ther.* **9,** 1451–60 (2010).

20. Füllbeck, M., Dunkel, M., Hossbach, J., Daniel, P. T. & Preissner, R. Cellular Fingerprints: A Novel Approach Using Large-Scale Cancer Cell Line Data for the Identification of Potential Anticancer Agents. *Chem. Biol. Drug Des.* **74,** 439–448 (2009).

21. Cheng, T., Wang, Y. & Bryant, S. H. Investigating the correlations among the chemical structures, bioactivity profiles and molecular targets of small molecules. *Bioinformatics* **26,** 2881–8 (2010).

22. Genomics of Drug Sensitivity in Cancer. at <ftp://ftp.sanger.ac.uk/pub4/cancerrxgene/releases/release-1.0/>

23. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57,** 289 – 300 (1995).

24. Malley, J. D., Dasgupta, A. & Moore, J. H. The limits of p-values for biological data mining. *BioData Min.* **6,** 10 (2013).

25. Motulsky, H. J. Common misconceptions about data analysis and statistics. *J. Pharmacol. Exp. Ther.* **351,** 200–5 (2014).

26. Nuzzo, R. Scientific method: Statistical errors. *Nature* **506,** 150–152 (2014).

27. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta - Protein Struct.* **405,** 442–451 (1975).

28. Chedzoy, O. B. in *Encycl. Stat. Sci.* (John Wiley & Sons, Inc., 2006).

29. Papadatos, G. *et al.* A document classifier for medicinal chemistry publications trained on the ChEMBL corpus. *J. Cheminform.* **6,** 40 (2014).

30. Vihinen, M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics* **13,** S2 (2012).

31. Smusz, S., Kurczab, R. & Bojarski, A. J. The influence of the inactives subset generation on the performance of machine learning methods. *J. Cheminform.* **5,** 17 (2013).

32. Klepsch, F., Vasanthanathan, P. & Ecker, G. F. Ligand and structure-based classification models for prediction of P-glycoprotein inhibitors. *J. Chem. Inf. Model.* **54,** 218–29 (2014).

33. Kolchinsky, A., Abi-Haidar, A., Kaur, J., Hamed, A. A. & Rocha, L. M. Classification of protein-protein interaction full-text documents using text and citation network features. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **7,** 400–11

34. Poil, S.-S. *et al.* Integrative EEG biomarkers predict progression to

Alzheimer’s disease at the MCI stage. *Front. Aging Neurosci.* **5,** 58 (2013).

35. Pemovska, T. *et al.* Individualized Systems Medicine Strategy to Tailor Treatments for Patients with Chemorefractory Acute Myeloid Leukemia. *Cancer Discov.* CD–13–0350 (2013). doi:10.1158/2159-8290.CD-13-0350

36. Kamiyama, H. *et al.* Personalized Chemotherapy Profiling Using Cancer Cell Lines from Selectable Mice. *Clin. Cancer Res.* **19,** 1139–1146 (2013).

37. Williams, S. A., Anderson, W. C., Santaguida, M. T. & Dylla, S. J. Patient-derived xenografts, the cancer stem cell paradigm, and cancer pathobiology in the 21st century. *Lab. Invest.* **93,** 970–82 (2013).

38. Simon, R. & Roychowdhury, S. Implementing personalized cancer genomics in clinical trials. *Nat. Rev. Drug Discov.* **12,** 358–369 (2013).

39. Majumder, B. *et al.* Predicting clinical response to anticancer drugs using an ex vivo platform that captures tumour heterogeneity. *Nat. Commun.* **6,** 6169 (2015).

40. Makridakis, M. & Vlahou, A. Secretome proteomics for discovery of cancer biomarkers. *J. Proteomics* **73,** 2291–305 (2010).

41. Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **32,** 1202–1212 (2014).

42. Potter, N. E. *et al.* Single-cell mutational profiling and clonal phylogeny in cancer. *Genome Res.* (2013). doi:10.1101/gr.159913.113

43. Al-Lazikani, B., Banerji, U. & Workman, P. Combinatorial drug therapy for cancer in the post-genomic era. *Nat. Biotechnol.* **30,** 679–92 (2012).

44. Patel, M. N., Halling-Brown, M. D., Tym, J. E., Workman, P. & Al-Lazikani, B. Objective assessment of cancer genes for drug discovery. *Nat. Rev. Drug Discov.* **12,** 35–50 (2013).

45. de Gramont, A. A. *et al.* Pragmatic issues in biomarker evaluation for targeted therapies in cancer. *Nat. Rev. Clin. Oncol.* **advance on,** (2014).

46. Tran, B. *et al.* Cancer genomics: technology, discovery, and translation. *J. Clin. Oncol.* **30,** 647–60 (2012).

47. Ahmed, J. *et al.* CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge. *Nucleic Acids Res.* **39,** D960–D967 (2011).

48. Boutros, P. C., Margolin, A. A., Stuart, J. M., Califano, A. & Stolovitzky, G. Toward better benchmarking: challenge-based methods assessment in cancer genomics. *Genome Biol.* **15,** 462 (2014).

49. Sheskin, D. J. Handbook of Parametric and Nonparametric Statistical Procedures. (2007). at <http://dl.acm.org/citation.cfm?id=1529939>