# Machine learning models to predict *in vivo* drug response via optimal dimensionality reduction of tumour molecular profiles

Linh Nguyen[1,2], Stefan Naulaerts[1], Alexandra Bomane[1], Alejandra Bruna[3], Ghita Ghislat[4] & Pedro J. Ballester[1]

[1] 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.

[2] Department of Pharmacological, Medical and Agronomical Biotechnology, University of Science and Technology of Hanoi, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Hanoi, Vietnam.

[3] Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge CB2 0RE, UK.

[4] Centre d'Immunologie de Marseille-Luminy, Inserm, U1104, CNRS UMR7280, Marseille, France.

* Correspondence to: Pedro Ballester, email: pedro.ballester@inserm.fr, Phone: +33 (0) 4 86 97 72 01

Keywords: biomarker discovery, machine learning, patient-derived xenograft, precision oncology, tumour profiling

## ABSTRACT

Inter-tumour heterogeneity is one of cancer's most fundamental features. Patient stratification based on drug response prediction is hence needed for effective anti-cancer therapy. However, lessons from the past indicate that single-gene markers of response are rare and/or may fail to achieve a significant impact in clinic. In this context, Machine Learning (ML) is emerging as a particularly promising complementary approach to precision oncology. Here we leverage comprehensive Patient-Derived Xenograft (PDX) pharmacogenomic data sets with dimensionality-reducing ML algorithms with this purpose. Results show that combining multiple gene alterations via ML leads to better discrimination between sensitive and resistant PDXs in 19 of the 26 analysed cases. Highly predictive ML models employing concise gene lists were found for three cases: Paclitaxel (breast cancer), Binimetinib (breast cancer) and Cetuximab (colorectal cancer). Interestingly, each of these ML models identify some responsive PDXs not harbouring the best actionable mutation for that case (such PDXs were missed by those single-gene markers). Moreover, ML multi-gene predictors generally have much fewer false negatives, i.e. these retrieve a much higher proportion of treatment-sensitive

PDXs than the corresponding single-gene marker. As PDXs often recapitulate clinical outcomes, these results suggest that many more patients could benefit from precision oncology if multiple ML algorithms were also applied to existing clinical pharmacogenomics data, especially those algorithms generating classifiers combining data-selected gene alterations.

## INTRODUCTION

It is now well-established that the efficacy of cancer drugs is strongly patient-dependent. Whereas analgesics such as Cox-2 inhibitors show efficacy in 80% of patients, on average only 25% of oncological patients actually respond to cancer drugs[1]. Consequently, there is a great need to find accurate ways to predict which cancer patients will respond to a given anti-cancer treatment. The predominant approach to date has been to identify a specific somatic mutation to act as single-gene biomarker discriminating between therapy responders and non-responders[2]. Such a predictive biomarker is commonly referred to as an actionable mutation (either a point mutation, deletion or amplification of a specific gene in the tumour sample). Despite being able to predict the response to some drugs[3,4], most patients cannot benefit from single-gene markers because these have simply not been found for the vast majority of drugs[5,6]. Moreover, drug markers are generally found predictive on a specific cancer type, which means that the marker might not be predictive of drug response on patients from other types[7].

Not only are these simple drug-gene associations rare, but they are also not strong predictors of drug response in some cases. For example, the mutational status of EGFR in Non-Small Cell Lung Cancer (NSCLC) is a FDA-approved marker of response to Erlotinib[2,8].The response rate in EGFR-mutant NSCLC tumours was found to be only 16% in this study[8], i.e. a 16% precision. Low precision may be due to the interplay of a range of confounding factors, acting on either the same gene (e.g. low expression of mutant EGFR) or other genes (e.g. resistance-

inducing mutations in the TP53 gene[9]). The same study unveiled that 67% of the responsive patients were not correctly identified as such, which corresponds to a 33% recall, due to their NSCLC tumours not being EGFR-mutant. This means that two-thirds of NSCLC patients responded to Erlotinib by molecular mechanisms that do not involve EGFR mutations. The Matthews Correlation Coefficient (MCC) summarises both types of error (false positives and false negatives, whose numbers are inversely proportional to precision and recall, respectively) into a single performance metric. This single-gene marker obtained a MCC of just 0.11, which is slightly better than random classification (MCC=0) and very far from perfect classification (MCC=1). Figure 1 clarifies the limitations of this single-gene marker further.

While precision and recall vary depending on the drug and its actionable mutation, the example in Figure 1 is representative in that the values of these metrics are generally quite modest[5,6,10]. Furthermore, it is widely believed that a suitable cancer treatment can only be predicted for those patients who have one of such actionable mutations[5]. We argue however that single-gene markers of drug response constitute only one possible approach to precision oncology. Consequently, more patients could benefit from taking an alternative approach that instead captures the interplay between multiple gene alterations that co-operatively control treatment response within tumours of a specific cancer type.

A promising complement to single-gene markers is the application of Machine Learning (ML)[11] to learn which combinations of gene alterations are most predictive of *in vivo* tumour response to a given treatment. ML algorithms can build *in silico* models with higher precision (i.e. fewer false positives) by learning which gene alterations, other than the single actionable mutation, influence drug response and how. Regarding increasing recall (i.e. fewer false negatives), ML can potentially learn all the different ways with which tumours of a given cancer type respond to a specific anti-cancer therapy. In that case, ML models would correctly

identify not only responders with the actionable mutation as the single-gene marker, but also the responders that are wild-type for that gene. Moreover, ML can provide predictive multi-gene models for some of the many drugs for which a single somatic mutation is simply not enough to predict tumour response[12,13].

Unfortunately, the limiting factor for the application of ML to this problem is the availability of relevant data. Although the public release of new clinical pharmacogenomics data sets to power precision oncology is often promised, drug response data is typically excluded from these sets or at the very least limited to a few drug treatments. For example, the first release of the AACR Project Genie[14] contained 19,000 molecularly-profiled tumour samples from various cancer types. However, the responses of the corresponding cancer patients to the administered treatments are still withheld to this date. Even if that information was revealed, cancer patients usually receive drugs in combination and/or several lines of therapy after sample collection, which constitute confounding factors in the discovery of new predictors of drug response. Deep molecular profiling, with no treatment response data, is only part of the puzzle.

In the last six years, data comprising thousands of molecularly-profiled cell lines treated with hundreds of cancer drugs have been made freely available, e.g. GDSC[15], CCLE[16] or CTRP[17]. Cell lines are relatively cheap, quick to grow and amenable to high-throughput experiments[18]. In addition, some cell lines have been shown to mimic sufficiently well primary tumours[18–20]. Given their suitability for high-throughput experiments[18], they are also the model for which more data is publicly available. Thus, such *in vitro* pharmacogenomics data sets are the only ones available to predict response on many drug-cancer type pairs. Despite these advantages, cell lines suffer from several inherent limitations. For example, intra-tumour heterogeneity, extracellular environment and immune system

response processes are not captured by cell lines. They are furthermore prone to divergence across passages[19].

In this context, Patient-Derived Xenograft (PDX) models are of great importance when no relevant clinical data is available[21–24]. Indeed, PDXs effectively capture patient-to-patient response variability to anti-cancer therapy[25,26]. In addition, these preclinical tools faithfully preserve the intra- and inter-tumoral heterogeneity observed in the originating cancer sample and the clinical population, respectively[27,28]. Taken together, these results support the use of these PDX models in guiding clinical therapeutic decisions for a more effective cancer treatment management[29,30]. PDX pharmacogenomics data represents an attractive opportunity to build ML models to predict tumour response in those treatment-cancer type pairs for which clinical data sets are not available. Prominent among such data sets stands the NIBR-PDXE[30] resource for its high number of PDXs and their comprehensive profiling. Over 1000 PDXs were established, with 40% of these molecularly-profiled at three levels: whole-exome single-nucleotide variants (SNVs), copy-number alterations (CNAs) and gene expression (GEX). Importantly, some of these PDXs were also evaluated with a panel of 60 treatments, which makes these data sets amenable to ML modelling.

Here we carry out a systematic study to investigate how ML can improve the prediction of *in vivo* drug response from tumour molecular profiles by combining multiple gene alterations. To the best of our knowledge, this is the first time that NIBR-PDXE data is analysed with this aim. Note that previous studies applying ML to this problem across multiple drugs have been based on pharmacogenomics data from *in vitro* cell lines, instead of PDXs. There are three additional aspects of this paper that are novel: 1) directly comparing the performance of ML classifiers against that of single-gene markers across multiple treatments, 2) introducing and applying a variant of RF intended to provide a more stringent Feature Selection (FS)[31] as a way to better

handle the high-dimensionality of data sets, and 3) adopting a non-competitive approach where the goal is to identify the most suitable profile and classifier type for each treatment rather than that with the highest average performance across treatments (the current "one-size-fits-all" approach).

Given that single-gene markers based on somatic mutation data are central to current genomic medicine[6,32], it is surprising that practically all studies evaluating multi-gene ML models have not included a direct comparison with single-gene markers. When we recently carried out such comparison *in vitro* across 127 drugs[13], we observed that ML models combining multiple gene alterations identified a higher proportion of drug-sensitive cell lines, i.e. had a higher recall, in 93% of the drugs. Here we will determine whether this is also the case *in vivo*. If confirmed *in vivo*, this will mean that many more patients could benefit from precision oncology if multi-gene ML methods are also applied to existing clinical pharmacogenomics data.

One major challenge to build any predictive model is the high dimensionality of pharmacogenomics data. While typically only tens of tumours have their response to the drug available, the molecular profiles of these tumours may easily aggregate over 50,000 features. To face this challenge, ML algorithms with built-in FS such as Elastic Nets[33–37], Ridge[34,37], LASSO[34,35,37,38] or Random Forest (RF)[13,34,36,37,39,40] have been used to model pharmacogenomics data from *in vitro* cell lines. For instance, RF ignores those features irrelevant for predicting drug response and thus has been able to tackle to some extent this challenge. However, these methods have also been found to be unable to provide predictive models for many drugs. Here we will employ a new strategy, Optimal Model Complexity (OMC), to complement the ability of RF to reduce the dimensionality of tumour molecular profiles. With OMC, only a very small proportion of the typically thousands of features in the

considered molecular profile will be employed by the resulting model. This is beneficial in that much fewer features would have to be experimentally determined in forthcoming tumours. In addition, in those cases where a few tumour features control drug response, predictions should tend to be more accurate because RF will no longer be considering the thousands of irrelevant features that promote model overfitting.

Lastly, the application of a ML method achieving slightly better performance than a previous method on average across drugs is commonly reported. However, the performance of ML methods with similar average performance across drugs can be very different on a particular drug. We therefore adopt a non-competitive approach instead, where the goal is to identify the most suitable profile and classifier type for each treatment. This approach should be collectively more predictive than using the method with the highest average performance across treatments. Another expected advantage is that we will be able to find out which molecular profile is most predictive for a given drug, which is valuable to reduce the time and cost that would be associated to determining the rest of profiles on patients treated with that drug.

## RESULTS

We started the analysis by determining which drug-cancer type pairs in NIBR-PDXE[30] have sufficient data to be likely to lead to predictive models (see the Methods section). We identified 13 of such treatments in Breast Cancer (BRCA) and another set of 13 treatments in Colorectal Cancer (CRC). All but one of these 26 treatment-cancer type pairs had at least 35 PDXs, each PDX with treatment-response, SNV, CNA and GEX profiles.

### Establishing the best multi-gene predictor for each treatment and cancer type pair

To perform this task, we trained and evaluated two ML algorithms on each data set using leave-one-out cross-validation (LOOCV) as detailed in the Methods section. The first algorithm is RF

using all available features (RF-all), whereas the second is an OMC variant of RF to identify the most predictive features in each case (RF-OMC). Both algorithms are used to build binary classification models. To account for their stochastic nature, we trained each algorithm on each LOOCV training fold 10 times, thus obtaining 10 estimations of each performance metric per case. Note that this study always reports the median performances of each algorithm on held-out PDXs that were not used to train or select the model providing the prediction (e.g. each reported MCC is the median of 10 MCC determinations from 10 independent nested LOOCVs). To assess which cases are better predicted by a particular gene, we also performed the standard single-gene analysis to evaluate which genes sensitise PDXs to the treatment when an actionable SNV is detected. In particular, we identified the sensitive marker with the lowest p-value and reported its LOOCV performance. Full results can be found in the "single-gene_markers" tab of the Supplementary Tables.

Figure 2 shows the validation results for each of the 26 cases. We found out that the accuracy in predicting treatment response on left-out PDXs strongly depends on the considered treatment, molecular profile and classifier type in both cancer types. The performances of the best predictors from each case range from being slightly above purely random classification (MCC=0) to high in the context of this problem (MCC=0.57). In addition, a large variability is often obtained across the four molecular profiles within the ML algorithm. For each ML algorithm, we show the performance of the most predictive molecular profile for that case. As CNA is just a binarisation of the real-valued copy number (CN) profile, information-richer CN is used much more often than CNA in the best models across cases (19 vs 3 models in Figure 2, respectively). Performance also varies strongly across the three model types within a given case.

RF-OMC not only leads to more accurate predictors in 13 of the 26 cases, but these predictors merely require a very small subset of all gene alterations to operate (these concise gene lists are reported in tab RF_predictors of Supplementary Tables). By contrast, solely 6 of the 26 cases were better predicted by RF-all. Figure 2 shows that the MCCs of RF-OMC across the 13 treatments were not better than those of RF-all in CRC (P=0.47 from a one-sided paired t-test, both algorithms obtaining an average MCC of 0.19). However, RF-OMC was found to outperform RF-all in BRCA (P=0.009 from a one-sided paired t-test, average MCCs of 0.32 and 0.16, respectively). The best predictor in the remaining 7 cases was a single-gene marker based on SNV data (a comparison between single-gene and multi-gene classifiers restricted to SNV data can be found in the γ plots in Supplementary Figures 7 and 8). Overall, these results stress the importance of considering several model types and profiles to predict *in vivo* treatment response.

For further validation, we have also compared each of these two ML classifiers against a random model where response class is predicted using the prior probability of a PDX to be sensitive. Supplementary figures 2 and 3 show that RF-OMC predicts 21 of the 26 cases better than random (P<0.05; one-sided paired t-test). Highly significant p-values strongly suggest that the corresponding lists of selected features have high predictive value. By contrast, supplementary figures 4 and 5 show that RF-all is only able to predict 16 of the 26 cases at this level (P<0.05; one-sided paired t-test).

Each of the models in this study is evaluated using independent data by LOOCV (i.e. the measured response of any test PDX was not used in any way to train or select the model). CV is a standard way to assess predictive accuracy not only in this topic[13,41–43], but in any problem where there are few data instances (here PDXs) to train, select and test a machine-learning model[44].

Additional experiments using data from a second PDX resource would be in principle appealing. However, we are not aware of another resource where each PDX was profiled for both drug response and these molecular profiles. Furthermore, even if these experiments were possible, inconsistencies arising from the lack of standardisation in such high-throughput efforts would be likely to be present, as it has been the case with in vitro pharmaco-omic resources[45]. Such additional test sets would not only evaluate the predictors, but also the impact of differences in the experimental setup generating the data.

In the next three subsections, we describe the three predictors found to have the highest accuracies in more detail. These three are the most useful models to predict which forthcoming PDXs will be responsive.

**Predicting BRCA PDX response to Binimetinib**

The best multi-gene predictor for BRCA PDXs treated with Binimetinib, a MEK1/2 inhibitor, was RF-OMC applied to GEX data comprising the expression values of 22,665 genes. Our analysis discarded the other three molecular profiles for Binimetinib-BRCA (SNV, CN and CNA), as they were substantially less predictive than the GEX profile in this case. RF-OMC practically offered the same performance as RF-all (MCC of 0.57 vs 0.56, respectively). However, RF-OMC identified 14 out of these 22,665 genes as the more informative to predict BRCA PDX response to Binimetinib. The resulting RF-OMC predictions are hence optimal combinations of the expression values of only these 14 genes, whereas RF-all was trained on all 22,665 GEX features.

Figure 3 displays the performance of this multi-gene predictor compared to that of the best single-gene marker for Binimetinib-BRCA (the mutational state of PABPC3 with P=0.02 from a two-sided Fisher's exact test). The multi-gene predictor achieves a more substantial

discrimination between sensitive and resistant markers than the PABPC3 marker. This is also indicated by a higher MCC (0.57 vs 0.24). To help to understand what MCC values represent in terms of achieved discrimination, we also indicated FP and FN errors from both classifiers.

**Predicting BRCA PDX response to Paclitaxel**

The best multi-gene predictor for BRCA PDXs treated with Paclitaxel was RF-OMC applied to somatic mutation data comprising the presence or absence of a SNV in 15,232 genes. Our analysis discarded the other three molecular profiles for Paclitaxel-BRCA (GEX, CN and CNA), for being less predictive than the SNV profile in this case. The resulting RF-OMC model employs two out of these 15,232 genes (MUC20 and UPK3BL). RF-based combination of these two mutational states provide strongly better prediction than a RF model using the mutational states of all the genes (MCCs of 0.49 and -0.07, respectively).

Incidentally, we would like to highlight that RF-all constitutes an embedded feature selection technique[46]. RF-OMC performs generally better because it promotes a much more stringent feature selection, which has been found to be more suitable in most cases. We illustrate this point with paclitaxel-BRCA-SNV as an example. RF-OMC with only two SNV features achieves a MCC of 0.49. By contrast, with all features and the same data, RF-all obtains a far worse MCC of practically zero (MCC=-0.07). Out of these 15,232 features, only 3374 are actually used by any of the 1000 trees forming the RF-all model, as each RF tree only uses those features providing the best discrimination among the $m_{try}$ randomly chosen at each node. The poor performance indicates that 3374 features are still far too many for this particular case, which is much better predicted by a RF model employing the two most predictive features.

Figure 4 visualises the high performance of this two-gene predictor on Paclitaxel-BRCA. The performance of the best single-gene marker, also shown, is very poor. This marker is the

mutational state of HYDIN, a gene coding for Protein Phosphatase 1's Regulatory Subunit 31. While we found this sensitising mutation to be the most strongly associated with the cytotoxic drug Paclitaxel in BRCA (P=0.04; two-sided Fisher's exact test), its performance in left-out PDXs suggests that it is a spurious correlation.

**Predicting CRC PDX response to Cetuximab**

The best multi-gene predictor for CRC PDXs treated with Cetuximab was RF-OMC applied to somatic mutation data. The other three molecular profiles for Cetuximab-CRC (GEX, CN and CNA), were discarded for being less predictive than the SNV profile in this case. We identified four out of these 15,232 genes whose combined mutational states provide better prediction than a RF model using the mutational states of all genes (MCCs of 0.47 and 0.39, respectively).

Figure 5 visualises the higher performance of this four-gene predictor on Cetuximab-CRC. The performance of the best single-gene marker, also shown, is even higher in this case. This marker is the mutational state of ACR (Acrosin), whose association to this targeted drug is the strongest across the 26 cases (P=0.0003; two-sided Fisher's exact test). Unlike RF-OMC, most prediction errors correspond to sensitive PDXs that were not correctly identified as such (FN=7 vs FP=1).

**Multi-gene predictors generally offer substantially higher recall than single-gene markers**

In the three cases analysed in Figures 3-5, multi-gene predictors exhibit a substantially higher recall than the corresponding best single-gene markers. More concretely, recalls of 0.91, 0.88 and 0.81 (multi-gene) versus recalls of 0.68, 0.00 and 0.56 (single-gene). Figure 6 shows that this is actually a strong general trend: 23 out of 26 studied cases have a higher proportion of correctly predicted sensitive PDXs using the multi-gene markers. We recently observed a

similar trend when using the standard version of RF to predict *in vitro* drug response from SNV data[13], which is here confirmed *in vivo*.

In the three cases that do not follow this trend, a slightly higher proportion of correctly predicted sensitive PDXs was found using single-gene markers. The first case is BGJ398-BRCA and its best single-gene marker is the drug-gene association BGJ398-KIF20B. The other two cases are in CRC and have as best markers the following drug-gene associations: BYL719-TMEM184A and BYL719+LJM716-LSR.

## DISCUSSION

Single-gene markers such as actionable somatic mutations are not strong predictors of treatment response in general[5,6,10,12,13]. This situation fuels the need to complement single-gene markers with multi-gene models, especially those built by dimensionality-reducing ML algorithms. In this context, the following novel insights have been gleaned from our analyses:

1) Multi-gene RF models generally retrieve a much higher proportion of in vivo treatment-responsive PDXs than the best single-gene marker for the considered case (see Figure 6). This means that many responders without the marker are now correctly identified as such owing to effective combination of multiple gene alterations.

2) Substantially better prediction is achieved if a new variant of RF with enhanced feature selection (RF-OMC) is employed instead of the standard RF algorithm (see Supplementary Figure 6). Additional advantages of RF-OMC: it only requires profiling a small subset of genes to operate and such gene list can be used as starting point for mechanistic studies.

3) Considering multiple types of classifiers and molecular profiles collectively improve the prediction of in vivo tumour response across treatments and cancer types (see Figure 2 as well as Supplementary Figures 7 and 8).

We found that combining multiple gene alterations via ML has resulted in better discrimination between sensitive and resistant PDXs in 19 of the 26 analysed cases. More importantly, this ML approach has determined the most predictive classifier type and molecular profile for each treatment – cancer type pair. Despite training on data from practically the same PDXs within a cancer type, we found that some treatments can be predicted much better than others (this is true for both cancer types). The results show that an effective way to improve the prediction of a given case is to evaluate several model types. For instance, the collective prediction of these 26 cases would have been much worse if we had only considered the standard RF algorithm (see for instance the difference between the MCCs of square and triangle signs for drug LKA136 in Figure 2). Indeed, the higher the number of classifiers considered, the higher the average accuracy across treatment is (as shown by plots α to γ in Supplementary Figures 7 and 8 with up to three classifiers). We therefore expect that considering additional classifiers will improve results further.

Another effective way to improve the prediction of a given case is to consider all the available molecular profiles. It is noteworthy that the results would have also been worse if only one profile had been employed. For instance, 21 of the 26 cases were better predicted by profiles other than GEX (see MCCs of blue signs in Figure 2). This new knowledge is displayed in Figure 2 and fully reported in the Supplementary Tables. We did not investigate the integration of all the profiles because this would be too time- and cost-intensive in a clinical setting. A secondary reason is that that integrating multiple omics profiles may provide no benefit, despite the higher information content, due to the harder challenge of tackling even higher dimensional

problems. In practice, when a benefit has been achieved, this has typically been incremental[47].We have also seen that even ML algorithms with built-in FS can often struggle to provide predictive classification models. To further mitigate the problem of overfitting caused by high-dimensional data[48], we have introduced and evaluated RF-OMC. OMC ranks features by their individual ability to discriminate between resistant and sensitive tumours (e.g. via a t-test), with only the highest ranked being used to train the ML model. Thus, such univariate filters may miss co-operativity effects among features. In principle, wrapper FS techniques such as Recursive Feature Elimination (RFE)[49] may improve model accuracy by capturing these effects at the risk of identifying a local optimum instead of the globally best solution. However, at least in the related problem of cancer prognosis prediction[50], univariate filters generally outperform RFE despite the computational cost of RFE being much higher. In this same study, embedded (built-in) FS techniques such as LASSO did not generally outperform univariate filters either. Instead of using a fixed predetermined cutoff, e.g. the 100 top-ranked genes as in that study[50], a novel aspect of OMC is that the complexity of the ML model is optimised for the drug, cancer type, molecular profile and available data. As the dimensionality of the employed data is optimally reduced for the considered case, thousands of less informative gene alterations are not included in model building. This is often an advantage, as the least informative of these features are probably irrelevant and hence harm classifier performance.

In practice, we have found that OMC complements the standard version of RF on this type of problems (16 of the 26 cases were better predicted by RF-OMC). More importantly, RF-OMC predicts 11 cases with an MCC of at least 0.3, whereas RF-all only predicts 4 cases with at least that accuracy (see Supplementary Figure 6). Furthermore, while RF-OMC predicts 21 of the 26 cases better than random, this is only the case for 16 of the 26 cases predicted by RF-all.

Interestingly, unlike here with multi-omics features of tumours, FS did not generally result in more accurate ML models when using chemical features of drug molecules in the related problem of QSAR[51]. In our study, RF-all only outperforms RF-OMC once among the most predictive cases (i.e. those with an MCC of at least 0.3), when predicting BRCA PDX response to HDM201 based on SNV profiles. This probably due to the limitations of this initial version of RF-OMC, where features are pre-ranked without taking into account their possible synergies with the rest of features and the cut-off might be too tight. These exceptions also serve as a reminder of the importance of considering multiple ML algorithms.

The results also show that the accuracy in predicting *in vivo* tumour response strongly depends on the treatment regardless of cancer type. The boxplots in Supplementary Figures 7 and 8 show that this variability is strongly reduced as we consider more types molecular profiles and classifiers. Still, MCCs across treatments vary from slightly above 0 to about 0.6 (rightmost boxplot), despite practically the same numbers of profiled PDXs being available for each treatment. Thus, a possible source for this variability is that the optimal profile and/or classifier could not have been found yet for the treatments with worse predictions. For example, profiles probing non-coding genes, miRNAs, DNA methylation or proteomics, not among the available NIBR-PDXE profiles, might provide the basis to predict the response to some treatments better. Another source of variability is that responsive and non-responsive PDXs may form two well-separated clusters in feature space in some treatments, but not in others (the former case being easier for classification than the latter). This topology is controlled by multiple convoluted factors such as data biases (e.g. inter-tumour molecular and response heterogeneity), the choice of representation (e.g. GEX) or how complex is the relationship between response and profile (e.g. drug polypharmacology). Lastly, experimental errors might be treatment-dependent. For

instance, a different response would be measured if two mice were engrafted with the same tumour but they metabolised the drug differently.

As anticipating which classifiers will work best on a case is currently not possible, we benchmarked three algorithms building models of varying complexity. In this way, their relative performance can shed some light into the intrinsic non-observable complexity of each case. For instance, only a given list of genes could be controlling drug response for tumours of that type and that control could be preferentially exerted at a given omics level. If a single-gene marker works best, it is likely that such subset is actually reduced down to that single gene and the omics level is SNV (e.g. the mutational status of ACR as a predictor of CRC PDX response to Cetuximab in Figure 2 right). However, it could also be that mutations in other genes influence drug response as well, but we do not have sufficient data to detect more complex cooperativity patterns. In cases where RF-OMC works best, a few genes are likely to be controlling drug response via the most predictive molecular profile (e.g. the mutational status of NR1H2, TLK2 and CTSA to predict BRCA PDX response to LKA136 in Figure 2 left), although again more data could result in more complex gene interactions being exploited and thus models with higher predictive accuracy. In cases where RF-all obtains the best performance (e.g. an RF model trained on all SNV features to predict BRCA PDX response to HDM201 in Figure 2 left), this suggests that more features than those considered by the OMC strategy must be positively contributing to the prediction of drug response.

Indeed, an important advantage of RF-OMC over standard RF models is that only a few genes need to be profiled to predict whether a PDX is responsive or not. Concise gene lists in a highly predictive model are valuable for interpretation and clinical application purposes. Take for instance the 14 genes forming part of the Binimetinib-BRCA GEX predictor (Figure 3). RF-OMC unveiled that this gene list is a promising starting point for mechanistic studies. Such

studies would aim at explaining how the nonlinear interplay between the expression values of these genes accurately predicts BRCA PDX response to Binimetinib. On the other hand, concise gene lists permit cheaper and faster clinical implementation. For example, instead of carrying out three whole-exome molecular profiles per tumour sample, we now know that it suffices to determine the mutational status of just two genes to predict BRCA tumour response to Paclitaxel (Figure 4). This example also highlights that the most responsive tumours to a cytotoxic drug can also be accurately predicted. The latter indicates that the applicability of precision medicine in current standard of care oncological therapeutic regimes is not restricted to targeted agents, but also includes cytotoxic chemotherapy[13,52]. While this is not the aim of the study, we have commented on the cancer relevance of the genes selected to predict treatment response for the three best RF-OMC predictors. This literature review can be found in the supplementary information file (pages 7-14).

We have also discovered that multi-gene predictors of *in vivo* drug response generally have higher recall than single-gene markers, which confirms *in vivo* previous findings *in vitro*[13,40]. The recall of a single-gene marker will be necessarily poor in all cases in which the prevalence of the mutation is much lower than the response rate. It is therefore not surprising that the three cases where markers have higher recall than RF-OMC in Figure 6 are based on genes with high to very high prevalence (KIF20B, TMEM184A and LSR with respective prevalence across tumours of 47%, 61% and 85%). Albeit exceptions, this general trend makes sense because a marker by construction can only detect those responsive tumours with the actionable mutation. In other words, the marker is blind to responsive tumours arising from alternative molecular mechanisms as illustrated by the example in Figure 1. By contrast, a ML algorithm can implicitly learn all the mechanisms captured by the training data.

While the generalisation of these outcomes to patient data is still to be confirmed, the fact that they have been observed in both *in vitro* and *in vivo* preclinical models is promising. If such generalisation was observed, without generating any additional data, many more patients would benefit from precision oncology by applying multiple ML algorithms to existing clinical pharmacogenomics data (our study has also made a set of data modelling recommendations that can be applied to the analysis of any similar data sets). Also, as PDXs capture the diversity and complexity of their originating tumours[8], the translational potential of the generated predictors in a clinical setting is likely. Improved predictors for Paclitaxel and Cetuximab, both of which are standards of care for breast and colon cancers respectively, could hence have an impact on cancer treatment effectiveness in the clinic. Our approach could also be applied to generating improved predictors of response to drugs in development (e.g. Binimetinib), supporting the use of ML for patient selection in clinical trials. Beyond these potential clinical applications, our study shows that ML can be employed to improve our understanding of cancer biology. On the one hand, ML could anticipate which PDXs not harbouring the actionable mutation are responsive (these would have therefore been missed by existing single-gene markers). On the other hand, the OMC strategy provides a concise list of gene alterations that control treatment response in the considered cancer type (these alterations have been individually linked to cancer hallmarks, therapeutic response and prognosis as explained in pages 7-10 of the supplementary information file). Thus, an alternative polygenic hypothesis explaining treatment response can be generated by combining both streams of information.

## METHODS

### NIBR-PDXE data

The NIBR- PDXE data set[30] is publicly available as the Supplementary Table 1 at http://www.nature.com/nm/journal/v21/n11/full/nm.3954.html#supplementary-information. This Excel file has five tabs named RNASeq_fpkm, copy_number, pdxe_mut_and_cn2, PCT_raw_data and PCT_curve_metrics. The first three tabs contain three molecular profiles of the xenografted tumours. The RNASeq_fpkm tab contains gene expression values. The copy number tab contains the actual copy number of each gene. Copy number is also available as a categorical variable at the pdxe_mut_and_cn2 tab (this table also contains detected mutations per gene). Around 400 PDX models were profiled at each of these omic levels. The other two tabs are for treatment response data. The raw response data tab (PCT_raw_data) includes the percentage of tumour volume change relative to tumour volume at the start of treatment (%ΔTVol) of each treated PDX recorded every 3-4 days. Lastly, the processed response data tab (PCT_curve_metrics) includes the categorisation of PDX responses into one of four classes calculated from raw response data. Further information about how this data set was generated can be found in the original study[30].

**Processing treatment response data for modelling**

For each treated PDX, we retrieved its category from the processed response data. We also calculated its category from raw response data as indicated by Gao et al.[30]. This calculation was based on the variables Best Response (the minimum value of %ΔTVol for $t \geq 10$ days) and Best Average Response (the minimum value of the set of average responses spanned by all t values with $t \geq 10$ days). In particular, CR (Complete Response) was assigned if Best Response < -95% and Best Average Response <-40%; PR (Partial Response) if $-95\% \leq$ Best Response < -50% and $-40\% \leq$ Best Average Response <-20%; SD (Stable Disease) if $-50\% \leq$ Best Response < 35% and $-20\% \leq$ Best Average Response < 30%; otherwise PD (Progressive Disease) was assigned as the response category. Retrieved and calculated response categories differ in 277 of

the 4758 PDX-treatment pairs. Although such discrepancies were small (mostly swaps between contiguous categories) and not numerous (5.8% of the cases), we decided to use the calculated categories in these cases so that all PDX-treatment pairs were categorised following the same set of rules. Gao et al.[30] further subdivided PDX response into two classes: responders as those PDXs exhibited some level of sensitivity to the treatment (CR, PR or SD) and non-responders (PD) as those resistant to the treatment.

**Processing molecular profiling data for modelling**

For each gene, the Single-Nucleotide Variant (SNV) feature for that gene is assigned a value of 1 if at least one SNV was detected in this gene region (i.e. reported in the pdxe_mut_and_cn2 tab). If no SNV was detected in this gene, the gene was labelled as Wild Type (WT) and the SNV feature was assigned a value of 0. This encoding scheme is commonly used[53–55], as it has the advantage of leading to a much less sparse data instances vs features matrix than if a binary feature was defined for each SNV (a gene typically has several SNVs). In this way, each PDX profiled at the SNV level was characterised by a set of 15,232 binary features (one feature per gene). On the other hand, the copy_number tab contains the actual copy number (CN) of each gene as determined by Gao et al.[30]. This is the CN profile composed by 23,853 real-valued features (one per gene) and was not analysed by Gao et al. Instead, these authors categorised these measurements as follows: Amp5 if the gene is moderately amplified with copy number in the range >= 5 and < 8, Amp8 if the gene is strongly amplified with copy number >= 8 and Del0.8 if the gene is deleted with copy number <=0.8 (these per-gene CN categories were reported in pdxe_mut_and_cn2). As previously done elsewhere[53,54], we binarised copy-number data to generate less sparse features: the Copy-Number Alteration (CNA) feature of a gene has a value of 1 for aberrant copy number (Amp8, Amp5 or Del0.8) and 0 otherwise. Thus, each PDX profiled at the CNA level is described by a set of 21,534

binary features (one per gene). By contrast, the fourth molecular profile, Gene EXpression (GEX), is directly the data provided in the RNASeq_fpkm tab. Thus, each PDX profiled at the GEX level is characterised by 22,665 real-valued features (one per gene as well).

**Processed data sets for modelling**

Only a part of the PDX models from NIBR-PDXE have been both treatment-response and molecularly profiled. A previous cancer pharmacogenomics modelling study showed that it is possible to predict drug response of held-out tumours with a ML model trained on just 35 tumours[47]. As we are not aware of successful studies using smaller training sets, we focused on the two cancer types with the highest numbers of profiled PDXs per treatment, Breast Cancer (BRCA) and Colorectal Cancer (CRC), where all but one of these 26 treatment-cancer type pairs had at least 35 PDXs. With these settings, a set of 13 treatments were administered to BRCA PDX models and another set of 13 treatments were administered to CRC PDX models. The first two tabs of Supplementary Table 1 state the numbers of sensitive and resistant PDXs per treatment for BRCA and CRC, respectively.

Note that no claim about the optimality of the 35 PDXs threshold is made. Although less likely, some cases with fewer PDXs might be predictive too. In any case, a threshold is required to make the analysis manageable. We run two ML algorithms for two cancer types (BRCA and CRC), 13 treatments per cancer type, 4 types of molecular profiles for each treatment-cancer type pair and 10 replicates. Thus, a total 2,080 leave-one-out cross-validations (LOOCVs) have been carried out. A less restrictive threshold would increase further the number of trained and tested ML models as well as the subsequent analysis, which is already extensive for a single study.

**Measuring the predictive performance of a classifier**

The pharmacogenomics data set for a given cancer type, molecular profile and the $i^{th}$ treatment can be represented as

$$\mathcal{D}_i = \left\{\left(class_i^{(k)}, \boldsymbol{x}^{(k)}\right)\right\}_{k=1}^{k=n_i}$$

Where $\boldsymbol{x}$ is a high-dimensional vector with the features from the considered profile and the $i^{th}$ treatment has been administered to $n_i$ PDX models. In these binary classification problems, positive data instances are PDXs sensitive to the considered treatment (class=sensitive), whereas negatives are resistant PDXs (class=resistant). Note that, while they have slightly different meanings, we use the terms responder and sensitive PDX interchangeably as it is customary (same applies to the terms non-responder and resistant PDX).

Each of these data sets is employed to train classifiers to predict the class of a PDX from its corresponding molecular profile. Predictive performance is always reported on PDXs not used to train the classifier making the predictions. In particular, classifiers not employing model selection are evaluated with standard LOOCV. Moreover, classifiers employing model selection are evaluated with nested LOOCV to avoid overestimating their performance[56,57]. It is worth noting that nested LOOCV is nothing but a standard LOOCV where the model optimised (selected) with the training set of a given fold is applied to the test set of that fold (i.e. instead of training and testing the same model per fold).

Once observed classes are compared to predicted classes, PDXs in the considered data set can be broken down into true positives, true negatives, false positives and false negatives (their numbers being TP, TN, FP and FN, respectively). Thus, the discrimination offered by a classifier can be summarised by the Matthews Correlation Coefficient (MCC)[58]

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

MCC can take values from -1 to 1, where 1 means that the classifier provides perfect agreement between observed and predicted classes, -1 indicates a perfect disagreement and 0 means that the classifier performance is equivalent to that of predicting the class at random.

To investigate how the two sources of error contribute to the overall predictive performance represented by MCC, we also calculate Precision (PR) and Recall (RC) for each predictor. PR and RC are two classical metrics[59] whose definitions are:

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

In this study, a PR value of 0 would mean that all the PDXs predicted sensitive by the classifier are actually resistant, whereas a PR value of 1 would indicate that all the PDXs predicted responsive were experimentally confirmed to be responsive. On the other hand, RC is 0 when none of the sensitive PDXs is correctly identified as such, whereas RC is 1 if no sensitive PDX is missed by the classifier.

The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) were calculated using the R package ROCR version 1.0-7. The ROC curve was built based on the response class predicted by the model (RF-OMC or SG) and the actual observed class of each test PDX. For a given model, the ROC curve shows all different pairs of values for the metrics True Positive Rate (TPR) and False Positive Rate (FPR) resulting from varying the operating threshold from 0 to 1 to cutoff the probability of the PDX to be a positive (responsive). RF-based models assign each test PDX with a class probability value between 0 and 1, thus their ROC curves have one point for each threshold leading to a different TPR-FPR pair. By contrast, SG markers predict the class of each test PDX as either resistant/non-responsive (0) or sensitive/responsive (1), thus all thresholds between 0 and 1 lead to the same TPR-FPR pair (i.e. a single intermediate point in the ROC curve apart from the two extreme values). The

AUC is the area under the ROC curve. AUC ranges from 0 to 1. AUC=1 denotes a perfect classification, whereas AUC=0.5 corresponds to a random-level prediction of classes.

**Multi-gene classifiers with built-in Feature Selection (FS)**

Some ML algorithms can construct classifiers with built-in FS to mitigate the impact of the high dimensionality of data on their generalisation to unseen data. Random Forest (RF)[60] is one of these algorithms, as it generates trees that ignore irrelevant features by construction and thus is often found effective in modelling high-dimensional omic data[61]. We used the recommended values for RF hyperparameters (1000 for the number of trees and the square root of the number of considered features for $m_{try}$). We preferred this to tuning these hyperparameters for each training set, as RF tuning generally results in just marginal improvements at the cost of being much more computationally expensive[39,62,63]. As no model selection was carried out for this algorithm, standard LOOCV was performed to estimate the performance of RF using all the features (RF-all) on each data set (treatment-cancer type-molecular profile). To assess the variability introduced by the stochastic character of RF, we perform 10 repetitions of LOOCV per case, each using a different random seed (see corresponding boxplots in supplementary figures 4 and 5).

The proportion of responsive and non-responsive PDXs changes from case to case (see tabs data_by_treatment_BRCA and data_by_treatment_CRC in the Supplementary Tables). Although class imbalances are not strong, these could still introduce some loss in performance. Thus, we enabled class weighting in the RF algorithm (R package 'randomForest' version 4.6-12), which counterbalances class imbalances by putting a heavier penalty on misclassifying the minority class. The misclassification penalty of the minority class was set to the proportion of the majority class, thus promoting RF trees that are equally accurate regardless of the class.

**Multi-gene classifiers with Optimal Model Complexity (OMC)**

An effective way to improve predictive performance is to reduce the dimensionality of the data. Here data dimensionality can be defined as the number of considered features over the number of PDXs. One route to reduce dimensionality is hence to use more training data, but these are usually not available. An alternative route is to only consider the most informative features in the data, thus typically discarding the many thousands of less informative features (hence strongly reducing data dimensionality while retaining most the initial information content).

However, the optimal number of features and their identities depend on various factors (treatment, profile, cancer type and data set). Consequently, we designed OMC as a strategy to build ML models employing only the most relevant features. In a nutshell, OMC is made of three modules: one to rank features according to their relevance to treatment response, another to train a ML model per considered subset of features and third one to select the optimal model among those trained. Regarding ranking features, we used p-values from two-sided Fisher's exact tests to rank binary features within a given binary profile (SNV or CNA) and p-values from two-sided unpaired t-tests to rank real-valued features (GEX or CN). For each profile, treatment and cancer type, we considered n/2 subsets of features (n is the number of PDXs available for that case): the subset with the top 2 features, that with the top 3 features, …, that with top n/2 features and finally all features for that profile. This limit of n/2 ensures that the ML algorithm will not be challenged by high-dimensional data (i.e. all trained models will have at least two data points per considered feature), except for the run using all features intended to find out whether the case requires more than the top n/2 features. Lastly, the best among these n/2 models is selected as that with the highest LOOCV MCC. To estimate its performance, nested LOOCV MCC is calculated with model selection in the inner loop[56,57] (the same values of the hyperparameters used for RF-all were also used here for RF-OMC). Sometimes

the best RF-OMC model is that only employing the top 2 or the top 3 features (i.e. more complex models are not more predictive). As the RF model reduces in these cases to a set of redundant shallow trees, we are probably wasting computer time, as running RF with 10-50 trees should do just as well.

**Random model based on the prior probability of each case**

The following protocol was followed for each case. First, the proportion of sensitive PDXs in each LOOCV training fold was taken as an estimate of the probability of a PDX to be sensitive. Second, a random number between 0 and 1 was generated to decide whether the held-out PDX in the LOOCV test fold was predicted sensitive or not according to the estimated prior probability. Third, the process was iterated over all LOOCV folds. Once the response class of every PDX had been predicted in this way, a single MCC value was computed from predicted and actual classes of the PDXs. Lastly, we repeated this process 10 times with the same 10 different random seeds employed with RF-OMC and RF-all. From the 10 MCCs for the case and the other 10 MCCs for the RF-model (either RF-OMC or RF-all), we carried out a one-sided paired t-test to determine whether the performance of the RF model was better than that of this random model ($P<0.05$). Results can be found in supplementary figures 2 to 5.

**Single-gene markers**

We identified the best single-gene marker for each of the 26 treatment-cancer type pairs using exactly the same data as RF-all and RF-OMC via LOOCV. The SNV profile was used as the source of detected somatic mutations, as there is currently a strong interest in using them as pharmacogenomic markers in oncology[5]. For each fold, the response of the PDX held-out in the test fold was predicted with the most significant sensitive marker (i.e. predicted sensitive if a SNV is detected in the marker gene, predicted resistant otherwise). Such marker was

determined by calculating two-sided Fisher's exact tests across training fold PDXs, one per gene, leading to p-values and effect sizes ($\phi$[64]) for 15,232 genes. The gene with the lowest p-value among those constituting sensitive mutations ($\phi>0$) was identified. The operation was repeated for each fold resulting in a LOOCV predicted class for each treated PDX and thus the LOOCV MCC for the best marker. After evaluating its predictive performance, we recalculated each best single-gene marker using now all the data so that these markers are ready to be used on forthcoming tumours. These results are in the Supplementary Tables (single-gene_marker tab).

## ABBREVIATIONS

FN: number of False Negatives

FP: number of False Positives

PDX: Patient-Derived tumour Xenograft model

NIBR-PDXE: Novartis Institutes for Biomedical Research - PDX encyclopedia

MCC: Matthews Correlation Coefficient

PR: PRecision

RC: ReCall

SG: Single-gene

RF: Random Forest

OMC: Optimal Model Complexity

TN: number of True Negatives

TP: number of True Positives

WT: Wild-Type

LOOCV: Leave-One-Out Cross-Validation

CN: Copy Number

CNA: Copy-Number Alteration

SNV: Single-Nucleotide Variant

GEX: Gene EXpression

## FIGURE CAPTIONS

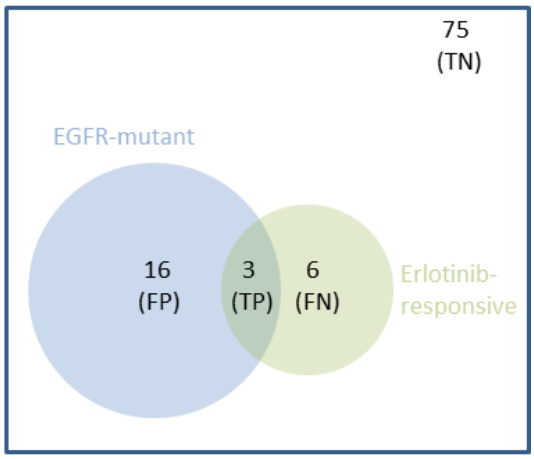



**Figure 1: Venn diagram showing the performance of a representative single-gene marker.** 100 NSCLC patients were treated with Erlotinib[8], 19 of them harbouring EGFR-mutant tumours. However, despite being a FDA-approved genomic marker[2], the mutational status of EGFR was a modest predictor of NSCLC tumour response to Erlotinib. Indeed, 84% (16/19) of EGFR-mutant tumours did not respond. This false positive (FP) rate also means that only 16% (3/19) of these tumours turned out to be responsive (precision of 0.16). Furthermore, 7% (6/81) of NSCLC tumours with wild-type EGFR actually responded to the drug. This false negative (FN) rate means that only 33% (3/9) of the responsive tumours were recalled by this predictor (recall of 0.33). Thus, Erlotinib-NSCLC-EGFR.snv is representative of single-gene markers in that their performance is generally quite modest[5,6]. Intuitively, one can see how combining multiple gene alterations could result in a predictor (blue circle) having a better overlap with the set of responsive tumours (green circle) by decreasing FPs and FNs. The overlap corresponds to the number of true positives (TP). Lastly, patients who do not fall into any of these three non-overlapping categories are true negatives (TN).

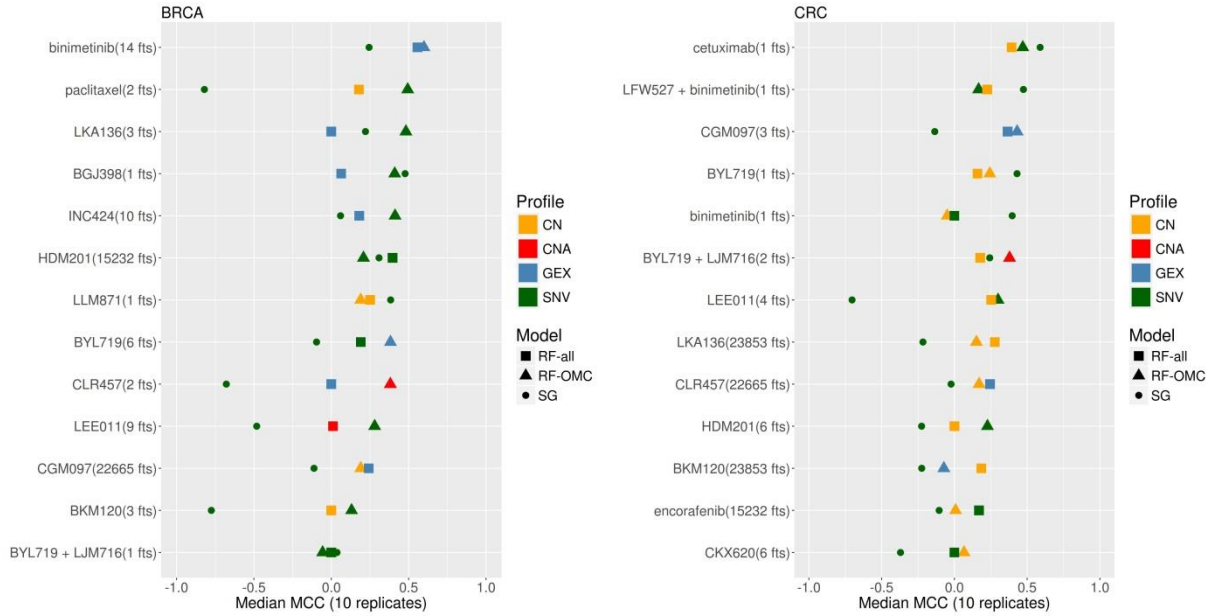



**Figure 2: Predictive performance of the best single-gene (SG) marker, Random Forest (RF) with Optimal Model Complexity (RF-OMC) and RF using all features (RF-all). (left)** Best predictor per treatment and classification model type on BRCA PDXs. Each row shows the results for one treatment and shows the number of top features with which the best classifier for that treatment was trained on. Furthermore, the colour and shape of each classifier indicates the employed molecular profile and model type, respectively ("*fts*" stands for "features"). For instance, paclitaxel (2 fts) appears as a green triangle, which means that Paclitaxel had RF-OMC-SNV as the classifier with the largest median MCC on LOOCV held-out PDXs and this classifier employed the top 2 features from the SNV profile (MUC20 and UPK3BL). All treatments have at least one predictor performing above random level (MCC=0), with the accuracy in predicting BRCA PDX response being strongly treatment-dependent. Importantly for clinical implementation, the best classifier per treatment is usually a model that only requires a handful of features to operate. **(right)** Best predictor per treatment and classifier type on CRC PDXs. Strong predictors of treatment response were also found for this cancer type. However, there were fewer of these predictive models in CRC than in BRCA (five treatments were predicted with MCC>0.4 in BRCA, but only two treatments were predicted at this level in CRC). Top models are more frequently associated with CN profiles in CRC than in BRCA. It is also clear that CNA profiles, using CN as a binary feature (altered/wild type), leads to less predictive models than real-valued CN.

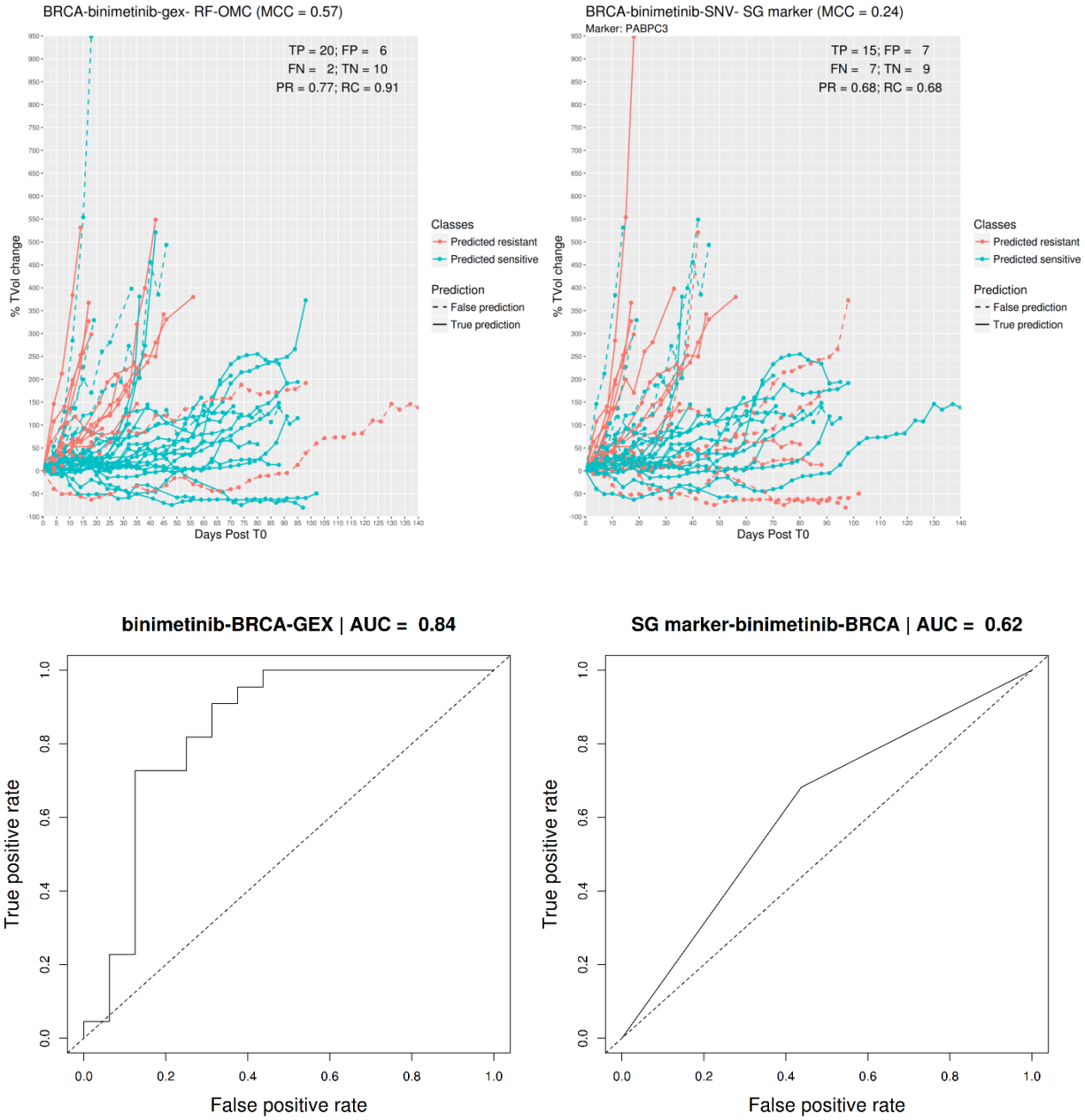
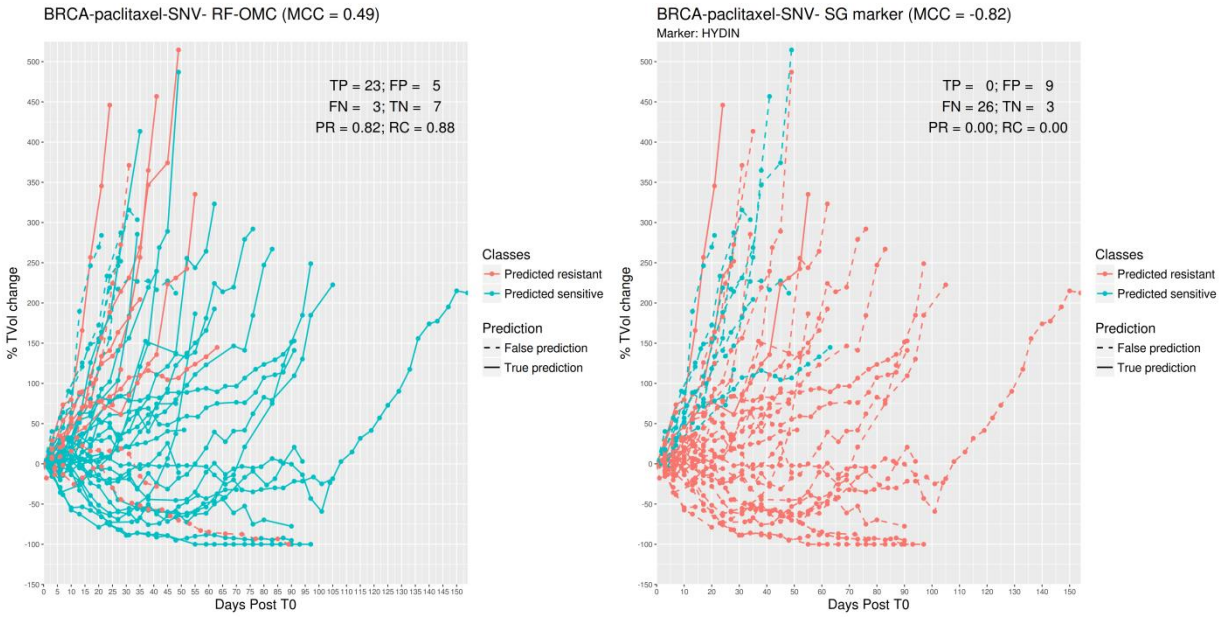



**Figure 3: Predicting BRCA PDX response to the MAPK inhibitor Binimetinib. (A)** Visualisation of tumour response prediction for test PDXs. Each line represents one PDX. The vertical axis shows the % of tumour change relative to the tumour volume at time zero (T0; i.e. immediately before the first administration of the drug) and the horizontal axis shows when measurement time in days after T0. From the legend, red discontinuous lines represent false negatives (responsive PDXs that were predicted to be non-responsive) and blue discontinuous lines represent false positives (non-responsive PDXs that were predicted responsive). The better the predictor is, the higher the proportions of lines at the bottom appear in blue (higher recall) and at the top in red (higher precision). **(left)** *Binimetinib (14 fts)*: RF-OMC model predicts BRCA tumour response to Binimetinib by optimally combining the

expression levels of only 14 genes (CRB3, NDUFA1, MPG, ECI1, ING2, KIF9, TSTD1, FAM100A, TCEAL3, HAGH, PEX11G, SNORA72, SNORA70 and PIN1). A high level of predictive accuracy was achieved on PDXs not used to train the model: MCC=0.57 (PR=0.77 and RC=0.91). **(right)** *Binimetinib (1 fts)*: Using the same input data and evaluation protocol, the best single-gene marker of Binimetinib sensitivity was the mutational state of PABPC3. The RF-OMC model obtained a substantially higher level of prediction than that of this standard single-gene procedure: MCC=0.24 (PR=0.68 and RC=0.68). **(B)** ROC curves and their AUC values to compare the discrimination offered by both models across all possible operating thresholds. The corresponding AUC value was indicated on the top of the ROC curve. This alternative performance metric also assigns higher predictive performance to the multi-gene RF-OMC model.



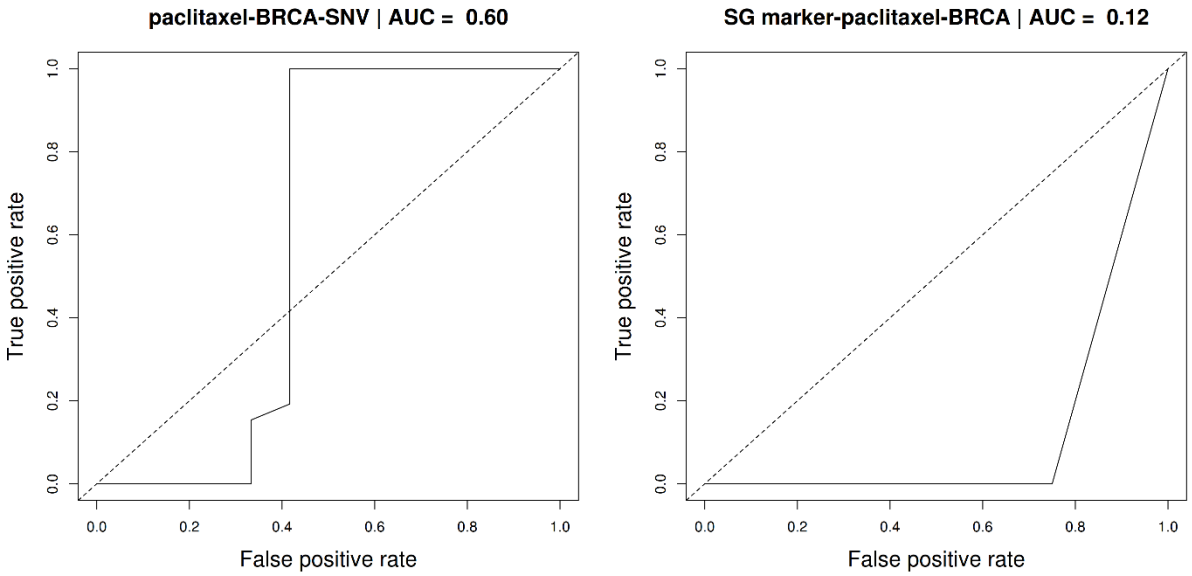



**Figure 4: Predicting BRCA PDX response to the tubulin inhibitor Paclitaxel. (A)** Visualisation of tumour response prediction for test PDXs. **(left)** *Paclitaxel (2 fts)*: RF-OMC predicts BRCA tumour response to Paclitaxel by optimally combining the mutational states of 2 genes (MUC20 and UPK3BL). A high level of predictive accuracy was achieved on PDXs not used to train the model: MCC=0.49 (PR=0.82 and RC=0.88). **(right)** *Paclitaxel (1 fts)*: Using the same input data and evaluation protocol, the best single-gene marker of Paclitaxel sensitivity was the mutational state of HYDIN. The RF-OMC model obtained a much higher level of prediction than that of this standard single-gene procedure: MCC=-0.82 (PR=0.00 and RC=0.00). **(B)** ROC curves and their AUC values to compare the discrimination offered by both models across all possible operating thresholds. The corresponding AUC value was indicated on the top of the ROC curve. This alternative performance metric also assigns higher predictive performance to the multi-gene RF-OMC model.

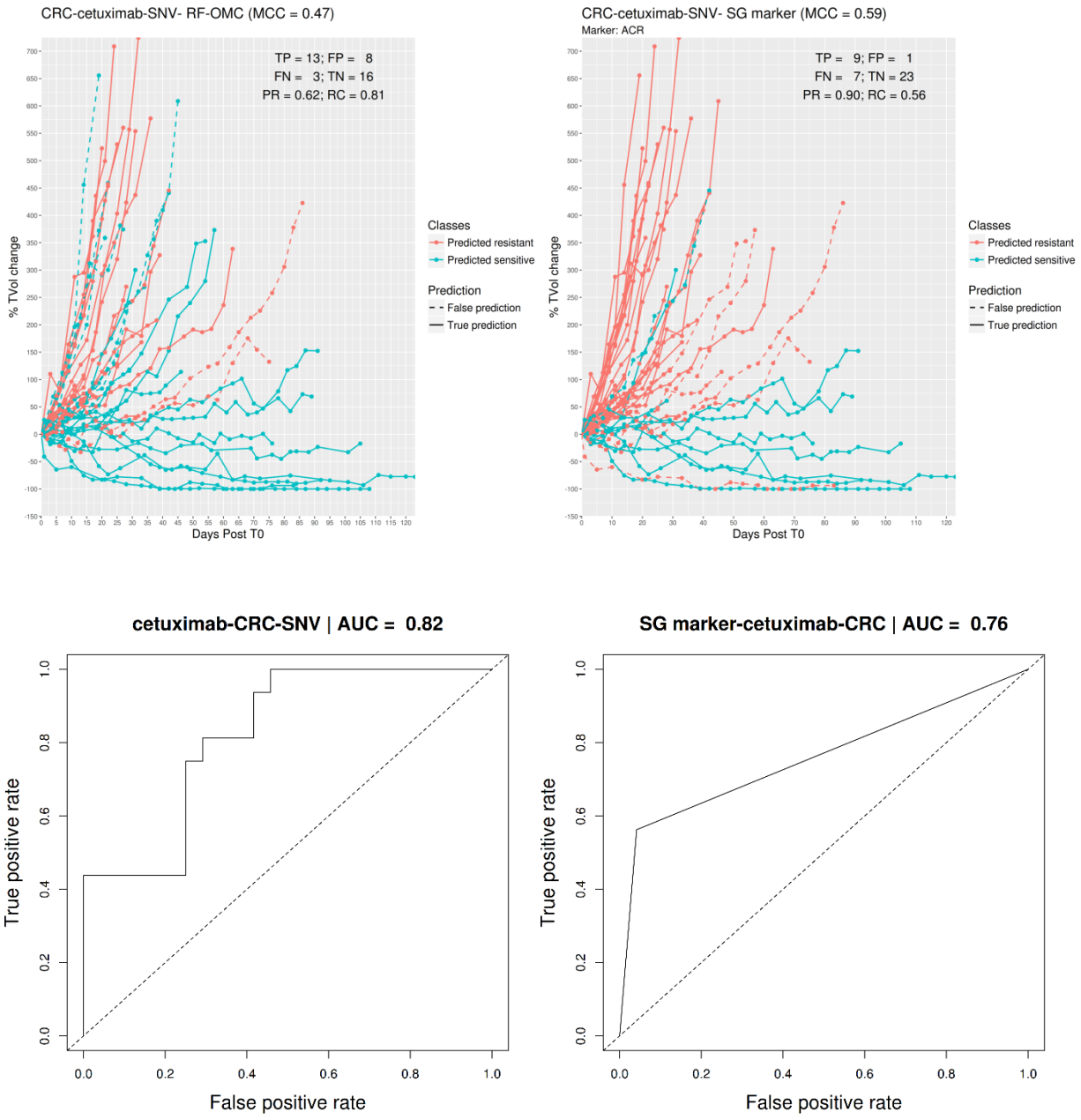



**Figure 5: Predicting CRC PDX response to the EGFR inhibitor Cetuximab. (A)** Visualisation of tumour response prediction for test PDXs. **(left)** *Cetuximab (4 fts)*: RF-OMC predicts CRC tumour response to Cetuximab by optimally combining the mutational states of 4 genes (ACR, DENND4B, NOTCH1 and RPL22). Again, a high level of predictive accuracy was achieved on PDXs not used to train the model: MCC=0.47 (PR=0.62 and RC=0.81). **(right)** *Cetuximab (1 fts)*: Using the same input data and evaluation protocol, the best single-gene marker of Cetuximab sensitivity was the mutational state of ACR. The RF-OMC model provided a slightly lower level of prediction than that of this standard single-gene procedure: MCC=0.59 (PR=0.90 and RC=0.56). **(B)** ROC curves and their AUC values to compare the discrimination offered by both models across all possible operating

thresholds. The corresponding AUC value was indicated on the top of the ROC curve. This alternative performance metric assigns now a slightly higher predictive performance to the multi-gene RF-OMC model. As the opposite outcome was obtained with MCC with the default 0.5 threshold, it is expected that optimising this threshold would lead to a more predictive RF-OMC model.

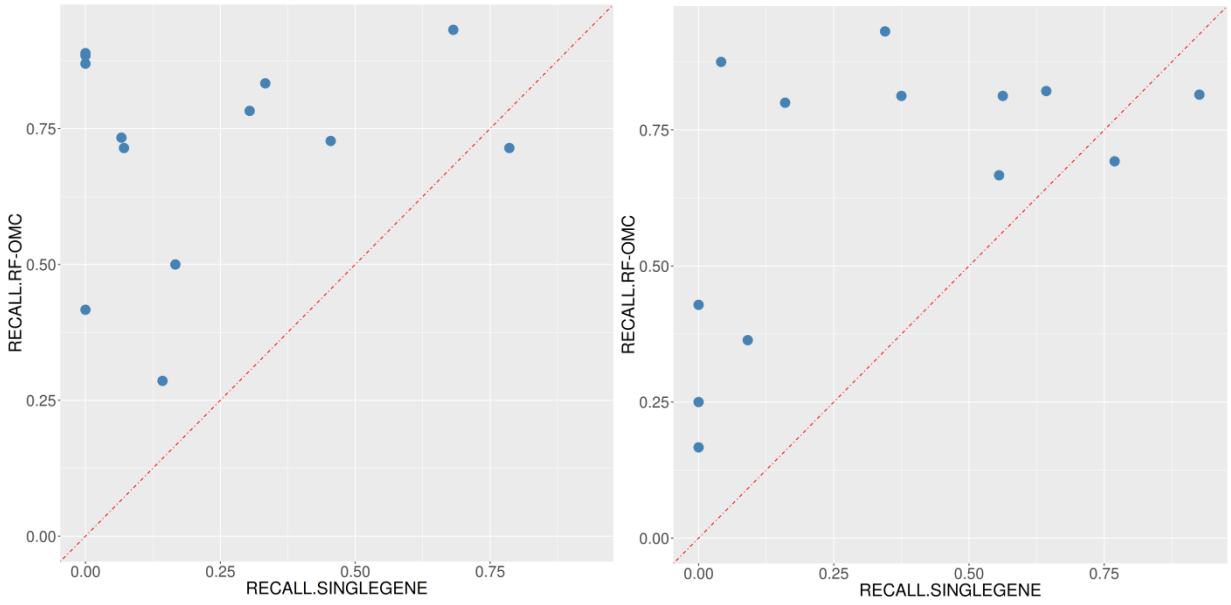



**Figure 6: ML multi-gene classifiers exhibit a higher recall than single-gene markers in 23 of the 26 treatment-cancer type pairs. (left)** In 12 of the 13 treatments for BRCA, multi-gene markers achieve a higher recall than the single-gene marker. The exception was the BGJ398-KIF20B association. **(right)** In 11 of the 13 treatments for CRC, multi-gene markers achieve a higher recall than the single-gene marker. These two single-gene markers with higher recall are the association BYL719-TMEM184A and the association BYL719+LJM716-LSR. Note that these three genes (KIF20B, TMEM184A and LSR) all have very high prevalence (47%, 61% and 85%, respectively) in addition to having the lowest p-value as sensitive marker in their respective cancer types.

## SUPPLEMENTARY FILES

**Supplementary Tables:** Description of processed NIBR-PDXE data and the discovered multi-omics predictors of *in vivo* treatment response.

**Supplementary Information:** Visualisation of the criteria for categorising PDX responses, p-values of ML models with respect to a null distribution, comparison RF-OMC vs RF-all across cases and plots showing how

considering multiple classifiers and molecular profiles collectively improve predictive accuracy. A literature review of the cancer relevance of the selected features from the three most predictive RF-OMC models.

## AUTHOR CONTRIBUTIONS

P.J.B. conceived the study and designed the experiments. P.J.B wrote the manuscript with the assistance of all authors. L.N. carried out the numerical experiments. All authors analysed the results and contributed to their discussion.

## ACKNOWLEDGMENTS

This work has been carried out thanks to the support of the 911 Programme PhD scholarship from Vietnam National International Development (L.N.), IPC PhD scholarship (A.B.), Horizon 2020 work program of the Marie Curie Actions (G.G.), Canceropôle PACA (S.N.) and INSERM (P.J.B).

## CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

## DATA AVAILABILITY

This study did not generate any data.

## CODE AVAILABILITY

The code to reproduce all the results is available upon request.

## REFERENCES

[1] Spear BB, Heath-Chiozzi M, Huff J. Clinical application of pharmacogenetics. Trends Mol Med 2001;7:201–4. doi:10.1016/S1471-4914(01)01986-4.

[2] Rodríguez-Antona C, Taron M. Pharmacogenomic biomarkers for personalized cancer treatment. J Intern Med 2015;277:201–17. doi:10.1111/joim.12321.

[3] Sosman JA, Kim KB, Schuchter L, Gonzalez R, Pavlick AC, Weber JS, et al. Survival in BRAF V600–Mutant Advanced Melanoma Treated with Vemurafenib. N Engl J Med 2012;366:707–14. doi:10.1056/NEJMoa1112302.

[4] Ascierto PA, Minor D, Ribas A, Lebbe C, O'Hagan A, Arya N, et al. Phase II Trial

(BREAK-2) of the BRAF Inhibitor Dabrafenib (GSK2118436) in Patients With Metastatic Melanoma. J Clin Oncol 2013;31:3205–11. doi:10.1200/JCO.2013.49.8691.

[5] Prasad V. Perspective: The precision-oncology illusion. Nature 2016;537:S63–S63. doi:10.1038/537S63a.

[6] Huang M, Shen A, Ding J, Geng M. Molecularly targeted cancer therapy: some lessons from the past decade. Trends Pharmacol Sci 2014;35:41–50. doi:10.1016/j.tips.2013.11.004.

[7] Prahallad A, Sun C, Huang S, Di Nicolantonio F, Salazar R, Zecchin D, et al. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. Nature 2012;483:100–3. doi:10.1038/nature10868.

[8] Tsao M-S, Sakurada A, Cutz J-C, Zhu C-Q, Kamel-Reid S, Squire J, et al. Erlotinib in Lung Cancer — Molecular and Clinical Predictors of Outcome. N Engl J Med 2005;353:133–44. doi:10.1056/NEJMoa050736.

[9] Ulivi P, Delmonte A, Chiadini E, Calistri D, Papi M, Mariotti M, et al. Gene mutation analysis in EGFR wild type NSCLC responsive to erlotinib: are there features to guide patient selection? Int J Mol Sci 2014;16:747–57. doi:10.3390/ijms16010747.

[10] Eckhardt SG, Lieu C. Is Precision Medicine an Oxymoron? JAMA Oncol 2018. doi:10.1001/jamaoncol.2018.5099.

[11] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet 2015;16:321–32. doi:10.1038/nrg3920.

[12] Geeleher P, Loboda A, Lenkala D, Wang F, LaCroix B, Karovic S, et al. Predicting Response to Histone Deacetylase Inhibitors Using High-Throughput Genomics. J Natl Cancer Inst 2015;107:djv247. doi:10.1093/jnci/djv247.

[13] Naulaerts S, Dang CC, Ballester PJ. Precision and recall oncology: combining multiple gene mutations for improved identification of drug-sensitive tumours. Oncotarget 2017;8. doi:10.18632/oncotarget.20923.

[14] AACR Project GENIE Consortium TAPG. AACR Project GENIE: Powering Precision Medicine through an International Consortium. Cancer Discov 2017;7:818–31. doi:10.1158/2159-8290.CD-17-0151.

[15] Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res 2012. doi:10.1093/nar/gks1111.

[16] 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:e0127433. doi:10.1371/journal.pone.0127433.

[17] Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price E V., Gill S, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol 2016;12:109–16. doi:10.1038/nchembio.1986.

[18] Goodspeed A, Heiser LM, Gray JW, Costello JC. Tumor-derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. Mol Cancer Res 2015:1541-7786.MCR-15-0189-. doi:10.1158/1541-7786.MCR-15-0189.

[19] Domcke S, Sinha R, Levine DA, Sander C, Schultz N. Evaluating cell lines as tumour models by comparison of genomic profiles. Nat Commun 2013;4. doi:10.1038/ncomms3126.

[20] Vincent KM, Findlay SD, Postovit LM. Assessing breast cancer cell lines as tumour models by comparison of mRNA expression profiles. Breast Cancer Res 2015;17. doi:10.1186/s13058-015-0613-0.

[21] Hidalgo M, Amant F, Biankin A V, Budinská E, Byrne AT, Caldas C, et al. Patient-derived xenograft models: an emerging platform for translational cancer research. Cancer Discov 2014;4:998–1013. doi:10.1158/2159-8290.CD-14-0001.

[22] Stewart E, Federico SM, Chen X, Shelat AA, Bradley C, Gordon B, et al. Orthotopic patient-derived xenografts of paediatric solid tumours. Nature 2017;549:96–100. doi:10.1038/nature23647.

[23] Cho S-Y, Kang W, Han JY, Min S, Kang J, Lee A, et al. An Integrative Approach to Precision Cancer Medicine Using Patient-Derived Xenografts. Mol Cells 2016;39:77–86. doi:10.14348/molcells.2016.2350.

[24] Krepler C, Sproesser K, Brafford P, Beqiri M, Garman B, Xiao M, et al. A Comprehensive Patient-Derived Xenograft Collection Representing the Heterogeneity of Melanoma. Cell Rep 2017;21:1953–67. doi:10.1016/j.celrep.2017.10.021.

[25] Izumchenko E, Paz K, Ciznadija D, Sloma I, Katz A, Vasquez-Dunddel D, et al. Patient-derived xenografts effectively capture responses to oncology therapy in a heterogeneous cohort of patients with solid tumors. Ann Oncol 2017;28:2595–605. doi:10.1093/annonc/mdx416.

[26] Li J, Ye C, Mansmann UR. Comparing Patient-Derived Xenograft and Computational Response Prediction for Targeted Therapy in Patients of Early-Stage Large Cell Lung Cancer. Clin Cancer Res 2016;22:2167–76. doi:10.1158/1078-0432.CCR-15-2401.

[27] Bruna A, Rueda OM, Greenwood W, Batra AS, Callari M, Batra RN, et al. A Biobank of Breast Cancer Explants with Preserved Intra-tumor Heterogeneity to Screen Anticancer Compounds. Cell 2016;167:260–274.e22. doi:10.1016/j.cell.2016.08.041.

[28] Kopetz S, Lemos R, Powis G. The promise of patient-derived xenografts: the best laid plans of mice and men. Clin Cancer Res 2012;18:5160–2. doi:10.1158/1078-0432.CCR-12-2408.

[29] Struss WJ, Black PC. Using PDX for Biomarker Development, Humana Press, Cham; 2017, p. 127–40. doi:10.1007/978-3-319-55825-7_9.

[30] Gao H, Korn JM, Ferretti S, Monahan JE, Wang Y, Singh M, et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. Nat Med 2015;21:1318–25. doi:10.1038/nm.3954.

[31] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in Bioinformatics. Bioinformatics 2007;23:2507–17. doi:10.1093/bioinformatics/btm344.

[32] Deyati A, Younesi E, Hofmann-Apitius M, Novac N. Challenges and opportunities for oncology biomarker discovery. Drug Discov Today 2013;18:614–24. doi:10.1016/j.drudis.2012.12.011.

[33] Ding MQ, Chen L, Cooper GF, Young JD, Lu X. Precision Oncology beyond Targeted

Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics. Mol Cancer Res 2017:molcanres.0378.2017. doi:10.1158/1541-7786.MCR-17-0378.

[34] 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:R47. doi:10.1186/gb-2014-15-3-r47.

[35] Fang Y, Qin Y, Zhang N, Wang J, Wang H, Zheng X. DISIS: Prediction of Drug Response through an Iterative Sure Independence Screening. PLoS One 2015;10:e0120408. doi:10.1371/journal.pone.0120408.

[36] Berlow N, Haider S, Wan Q, Geltzeiler M, Davis LE, Keller C, et al. An Integrated Approach to Anti-Cancer Drug Sensitivity Prediction n.d. doi:10.1109/TCBB.2014.2321138.

[37] Ammad-ud-din M, Khan SA, Wennerberg K, Aittokallio T. Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression. Bioinformatics 2017;33:i359–68. doi:10.1093/bioinformatics/btx266.

[38] Sun Y, Zhang W, Chen Y, Ma Q, Wei J, Liu Q, et al. Identifying anti-cancer drug response related genes using an integrative analysis of transcriptomic and genomic variations with cell line-based drug perturbations. Oncotarget 2016;7:9404–19. doi:10.18632/oncotarget.7012.

[39] Cortés-Ciriano I, van Westen GJP, Bouvier G, Nilges M, Overington JP, Bender A, et al. Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. Bioinformatics 2016;32:85–95. doi:10.1093/bioinformatics/btv529.

[40] Nguyen L, Dang CC, Ballester PJ. Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data. F1000Research 2017;5:2927. doi:10.12688/f1000research.10529.2.

[41] Zhang L, Chen X, Guan N-N, Liu H, Li J-Q. A Hybrid Interpolation Weighted Collaborative Filtering Method for Anti-cancer Drug Response Prediction. Front Pharmacol 2018;9:1017. doi:10.3389/fphar.2018.01017.

[42] Liu H, Zhao Y, Zhang L, Chen X. Anti-cancer Drug Response Prediction Using Neighbor-Based Collaborative Filtering with Global Effect Removal 2018. doi:10.1016/j.omtn.2018.09.011.

[43] Azuaje F. Computational models for predicting drug responses in cancer research. Brief Bioinform 2016:bbw065. doi:10.1093/bib/bbw065.

[44] Lever J, Krzywinski M, Altman N. Points of Significance: Model selection and overfitting. Nat Methods 2016;13:703–4. doi:10.1038/nmeth.3968.

[45] Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJWL, et al. Inconsistency in large pharmacogenomic studies. Nature 2013;504:389–93. doi:10.1038/nature12831.

[46] Hira ZM, Gillies DF. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. Adv Bioinformatics 2015;2015:198363. doi:10.1155/2015/198363.

[47] Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A

community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol 2014;32:1202–12. doi:10.1038/nbt.2877.

[48] Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer 2008;8:37–49. doi:10.1038/nrc2294.

[49] Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Mach Learn 2002;46:389–422. doi:10.1023/A:1012487302797.

[50] Haury A-C, Gestraud P, Vert J-P. The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures. PLoS One 2011;6:e28210. doi:10.1371/journal.pone.0028210.

[51] Eklund M, Norinder U, Boyer S, Carlsson L. Choosing Feature Selection and Learning Algorithms in QSAR. J Chem Inf Model 2014;54:837–43. doi:10.1021/ci400573c.

[52] Felip E, Martinez P. Can sensitivity to cytotoxic chemotherapy be predicted by biomarkers? Ann Oncol 2012;23 Suppl 1:x189-92. doi:10.1093/annonc/mds309.

[53] Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 2012;483:570–5. doi:10.1038/nature11005.

[54] Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, et al. Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. PLoS One 2013;8:e61318. doi:10.1371/journal.pone.0061318.

[55] Dang CC, Peón A, Ballester PJ. Unearthing new genomic markers of drug response by improved measurement of discriminative power. BMC Med Genomics 2018;11. doi:10.1186/s12920-018-0336-z.

[56] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 2006;7:91. doi:10.1186/1471-2105-7-91.

[57] Cawley GC, Talbot NLC. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. J Mach Learn Res 2010;11:2079–107.

[58] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta - Protein Struct 1975;405:442–51. doi:10.1016/0005-2795(75)90109-9.

[59] Van Rijsbergen CJ, Van CJ. Information retrieval. Butterworths; 1979.

[60] Breiman L. Random Forests. Mach Learn 2001;45:5–32. doi:10.1023/A:1010933404324.

[61] Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics 2012;99:323–9. doi:10.1016/j.ygeno.2012.04.003.

[62] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 2003;43:1947–58. doi:10.1021/ci034160g.

[63] Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. Bioinformatics 2010;26:1169–

75.

[64] Chedzoy OB. Phi-Coefficient. Encycl. Stat. Sci., John Wiley & Sons, Inc.; 2006.

[65] Nguyen L, Naulaerts S, Bomane A, Bruna A, Ghislat G, Ballester P. Machine learning models to predict in vivo drug response via optimal dimensionality reduction of tumour molecular profiles. bioRxiv 2018:277772. doi:10.1101/277772.