# Bimodal Gene Expression in Patients with Cancer Provides Interpretable Biomarkers for Drug Sensitivity

Wail Ba-Alawi[1,2], Sisira Kadambat Nair[1], Bo Li[3], Anthony Mammoliti[2], Petr Smirnov[2], Arvind Singh Mer[1,2], Linda Z. Penn[1,2], and Benjamin Haibe-Kains[1,2,3,4]

## ABSTRACT

Identifying biomarkers predictive of cancer cell response to drug treatment constitutes one of the main challenges in precision oncology. Recent large-scale cancer pharmacogenomic studies have opened new avenues of research to develop predictive biomarkers by profiling thousands of human cancer cell lines at the molecular level and screening them with hundreds of approved drugs and experimental chemical compounds. Many studies have leveraged these data to build predictive models of response using various statistical and machine learning methods. However, a common pitfall to these methods is the lack of interpretability as to how they make predictions, hindering the clinical translation of these models. To alleviate this issue, we used the recent logic modeling approach to develop a new machine learning pipeline that explores the space of bimodally expressed genes in multiple large *in vitro* pharmacogenomic studies and builds multivariate, nonlinear, yet interpretable logic-based models predictive of drug response. The performance of this approach was showcased in a compendium of the three largest *in vitro* pharmacogenomic datasets to build robust and interpretable models for 101 drugs that span 17 drug classes with high validation rates in independent datasets. These results along with *in vivo* and clinical validation support a better translation of gene expression biomarkers between model systems using bimodal gene expression.

**Significance:** A new machine learning pipeline exploits the bimodality of gene expression to provide a reliable set of candidate predictive biomarkers with a high potential for clinical translatability.

## Introduction

Identifying reliable predictive biomarkers of drug response is a key step in precision oncology. Large-scale cancer pharmacogenomic studies have boosted the research for finding predictive biomarkers by profiling thousands of human cancer cell lines at the molecular level and screening them with hundreds of drugs (1–5). Genomic features, including gene mutations and copy-number variations (CNV), have been so far regarded as the state-of-the-art method for predicting patients' response to drugs in the clinic. However, it has been shown that most genomic biomarkers are found in small proportions of patients and within that subset, only a few have shown response to associated drugs (6).

Several studies have investigated alternative sources for predictive biomarkers of drug sensitivity in cancer pharmacogenomics (7, 8). These studies have shown that gene expression outperforms other molecular features such as mutations and CNVs in predicting drug response in human cancer cell lines (7, 8). Yet, a major criticism of gene expression as a source of predictive biomarkers is the lack of reproducibility due to dependency on profiling assays and batch effects. To overcome such limitations, several studies have focused their analyses on genes that have shown bimodal distribution of expression (9–11). An advantage of a bimodal gene as a biomarker is that its modes can be used to robustly classify samples into two distinct expression states, allowing for easier interpretation, reproducibility, and translation of the biomarker into the clinic. For example, estrogen receptor (ESR1) bimodal expression defines two biological states within patients with breast cancer. These states have been used to stratify breast cancer patients into the clinically relevant subtypes (ER±) and derive treatment decisions (12, 13). Another example in cancer genomics is the use of 73 bimodal genes within ovarian cancer to define molecular subtypes with distinct survival rates (14). We also have shown that epithelial-to-mesenchymal transition (EMT) related genes were found to be bimodal pan-cancer and predictive of response to statin class of drugs (15).

Most pharmacogenomic studies that tackled the challenge of finding reliable predictive biomarkers for drug sensitivity employed univariate models for simplicity and interpretability (1, 2, 16). However, such models do not account for dependencies between genes yielding suboptimal model predictions. Recent studies have applied more sophisticated machine learning techniques that capture dependencies between genes and produce more accurate biomarkers predictive of drug sensitivity (7, 17). However, it becomes hard to biologically interpret these predicted biomarkers due to the complexity of these models and how they define the dependencies between the genes. In this study, we developed a machine learning pipeline to explore the large space of bimodally expressed genes and build multivariate, nonlinear, yet interpretable logic-based models predictive of drug response in large *in vitro* pharmacogenomic studies (**Fig. 1A**). Following our proposed approach, we developed robust and interpretable models predictive of drug sensitivity for more than 100 drugs of different pharmacologic classes. Common models with two independent large test sets were validated and yielded high predictive rates (92% and 61% respectively).

[1]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [2]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [3]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [4]Ontario Institute of Cancer Research, Toronto, Ontario, Canada.

**Note:** Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

**Corresponding Author:** Benjamin Haibe-Kains, University Health Network, 101 College Street, Toronto, Ontario M5G 1L7, Canada. Phone: 416-581-7628; Fax: 416-581-8626; E-mail: Benjamin.Haibe-Kains@uhnresearch.ca

Cancer Res 2022;82:2378–87

**doi:** 10.1158/0008-5472.CAN-21-2395

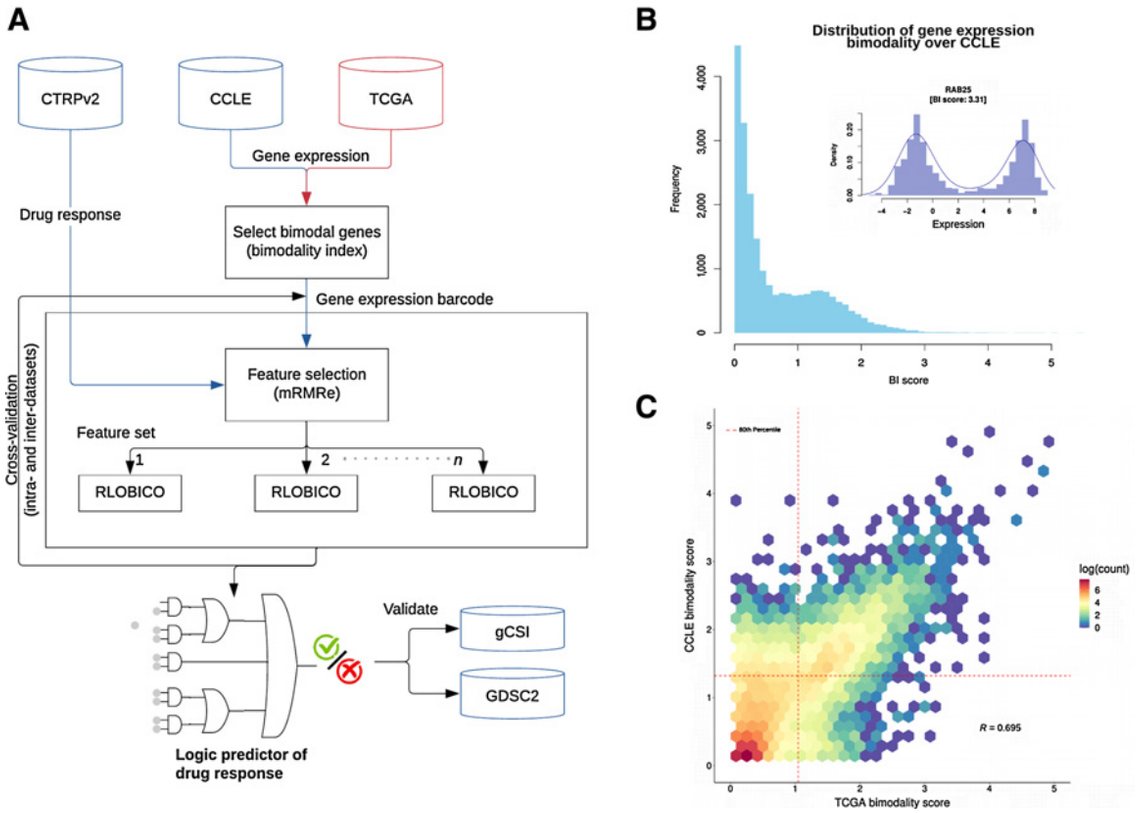


**Figure 1.**

**A,** Overview of the pipeline to create logic predictors of drug response. **B,** Distribution of bimodality index scores (BI) for all protein-coding genes based on RNA-seq gene expression profiles of cell lines in CCLE. **C,** Distribution of BI scores across CCLE and TCGA. Genes showing high bimodality (>80th percentile) in both datasets are chosen as global bimodal genes. Color represents frequency of genes ranging from high (red) to low (purple). *R*, Pearson correlation coefficient between common protein-coding genes in both CCLE and TCGA ($N = 21{,}903$).

## Materials and Methods

### Datasets

Data from Cancer Cell Line Encyclopedia (CCLE; ref. 16), Cancer Therapeutics Response Portal (CTRPv2; refs. 5, 18), Genentech Cell Line Screening Initiative (gCSI, released in 2018; refs. 19, 20), Genomics of Drug Sensitivity in Cancer (GDSC2, released in 2019; refs. 2, 3) and The Cancer Genome Atlas (TCGA; ref. 21) were all processed using the same pipeline using the PharmacoGx R package pipeline (22–24). Gene expression profiles were generated using Kallisto pipeline (25) with GRCh38 as human reference. Patient-derived xenograft encyclopedia (PDXE) dataset (26) was downloaded and processed using Xeva R package (27). Researchers can access HARTWIG Medical Foundation's (HMF) clinical and genomic data by applying for a data request at https://www.hartwigmedicalfoundation.nl/applying-for-data.

### Bimodality of gene expression profiles

Gene expression profiles, obtained from CCLE dataset, were used to characterize the bimodality feature of each gene in the set by fitting its distribution into a mixture of two Gaussian distributions. For those genes with a good fit, a bimodality score was calculated using the following formula:

$$\textit{Bimodality score } (BI) = \sqrt{\pi * (1-\pi)} * \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\nu_1+\nu_2}{2}}},$$

where $\pi$ is the proportion of samples in one group, $\mu_1$ and $\mu_2$ are the means of the expression level of the two modes, and $\nu_1$ and $\nu_2$ are the variances. Similar characterization was done using the TCGA dataset. Genes, then, were ranked according to their bimodality scores and the common protein-coding genes in the top 80th percentile of bimodality scores distribution in both CCLE and TCGA were chosen as top bimodal genes feature set. A binarization cutoff for each gene distinguishing relatively low versus high expression was calculated by taking the average point between the modes of the two fitted Gaussian distributions. To assess the robustness of these bimodal genes and their ability to reproduce expression status (High/Low) in other datasets, we used the same cutoffs generated from CCLE data and applied them on the corresponding genes for the same cell lines in gCSI dataset that are in common with CCLE. We then compared that to binarizing the rest of the genes in the transcriptome using average expression as a cutoff to assess if bimodal genes have better reproducibility than using the rest of the genes in the transcriptome (Supplementary Fig. S1A). Aiming to determine if any bimodal gene is a surrogate of a tissue type, we assessed the ability of each bimodal gene to predict each tissue type, where tissue types are hot-encoded, i.e., a vector of a tissue type will have the value of 1 for samples of that tissue type and 0 for the rest of the samples. Precision and F1 metrics were used to assess these predictions (Supplementary Fig. S6).

### Logic-based models

Logic-based models are machine learning models aiming at constructing Boolean logic functions that model the relationship between a binary set of features and a class label. Interpretability of the modeled associations is a key advantage of these types of models in comparison with other traditional machine learning models, which is an important feature for clinical translation of biomarkers. We developed RLOBICO, which is an R implementation of Logic Optimization for Binary Input to Continuous Output (LOBICO) method (28), to find binary rules that predict sensitivity of samples to different drugs. Our proposed pipeline starts with a binarized expression matrix followed by a feature selection method [minimum redundancy maximum relevance (mRMRe)] to choose highly relevant and complementary features that are then fed into RLOBICO to search the space of possible rules and associate these rules with a drug effect.

For each drug, we create a binarized expression matrix based on top bimodal gene features set and represent the effect of the drug on samples using the area above the dose–response curve (AAC) metric. AAC represents the area above the dose–response curve where cell viability decreases with increasing drug concentration. This metric is a measure of overall efficacy across different doses or in other words an average of both potency and maximal efficacy (29, 30). This is in contrast to the $IC_{50}$, which is a measurement of potency. In previous studies, we have shown that AAC, in general, is a more consistent metric across datasets (29, 30) and a better metric for training drug sensitivity models using *in vitro* pharmacogenomic datasets (29). We, therefore, selected the AAC metrics to train our set of logical models. LOBICO requires binarizing the effect of the drug and so we chose AAC of 0.2(30) as a threshold classifying samples to be either resistant (AAC < 0.2) or sensitive (AAC > 0.2) to each drug. However, the continuous values of AACs are still used as weights to optimize the modeling step such that a higher penalty would be incurred if a highly sensitive sample was misclassified as resistant. Generated rules by LOBICO are described using the disjunctive normal form, which is a standard notation to express logic functions. The disjunctive normal form is parameterized by two parameters: $K$, the number of disjuncts; and $M$, the number of terms per disjunct. We varied $K$ and $M$ to represent models of different complexities, i.e., from single predictors ($K = 1$, $M = 1$) to more complex models [($K$, $M$): (1, 2) (1, 3) (1, 4) (2, 2) (2, 1) (3, 1) (4, 1)]. We used mRMRe to limit the search space of all possible logical combinations of features to the top ten highly relevant and complementary features to control the risk of overfitting and facilitate interpretation of the model. Supplementary Figure S9 shows that increasing the number of features increases the runtime of training the model. We then apply RLOBICO to find the best rule predicting sensitivity of samples to drugs. Finally, to achieve more robust results, we create an ensemble rule based on a majority vote from rules generated by three different mRMRe feature sets followed by RLOBICO. All of these steps are performed in a 5-fold cross-validation setting to ensure no leak of information between folds. For evaluation of models, we used a modified version of the concordance index (CI; https://github.com/bhklab/wCI; ref. 31), in which we compare the concordance between measured AACs for a particular drug and the predictions of the associated model for the same samples. This modification accounts for noise in the drug screening assays as we found that repeating the same drug-cell line experiment in CTRPv2 resulted in inconsistencies in terms of measured drug response (AAC). We further investigated this observation and found that 95% of the replicates of the same drug and cell line experiments showed differences ($\Delta$ AAC = | $AAC_{replicate1}$ - $AAC_{replicate2}$ |) within 0.2 range (Supplementary Fig. S8). Hence, we remove the pairs of AACs that have $\Delta$ AAC < 0.2 from the calculation of the regular CI as they can flip directions within that range randomly. To avoid any information leak between the training and test sets, we have considered fully independent datasets from our PharmacoDB database (22–24).

### Research reproducibility

CCLE, CTRPv2, gCSI, and GDSC2 can be downloaded using PharmacoGx R package (22). Code to reproduce the results and figures is available at https://github.com/bhklab/Gene_Expression_Bimodality. We also provide a complete software environment through Code Ocean containing all necessary data and code to reproduce the analysis and figures described in this manuscript under the DOI https://codeocean.com/capsule/8205812/tree/v1. RLOBICO R package was used to generate the logic-based models (github.com/bhklab/RLOBICO).

## Results

### Bimodality of gene expression

To comprehensively explore the space of bimodal gene expressions, we performed a genome-wide characterization of gene expression distribution in large sets of patient tumors and immortalized cancer cell lines. Using the gene expression data from the CCLE (945 cell lines from 23 tissue types; 16), we determined the expression bimodality of a given gene by fitting a mixture of two Gaussian distributions across all samples and then calculating the bimodality index (**Fig. 1B**; ref. 9). We restricted this analysis to solid tumors as hematopoietic and lymphoid cell lines have distinctive molecular profiles and are generally more sensitive to chemical perturbations in comparison with solid tumors (18). Similarly, we computed the bimodality index for all genes using the gene expression of the solid tumors in TCGA (10,534 tumors from 30 tissue types; ref. 21). We subsequently selected the protein-coding genes that showed high bimodality index (>80th percentile) in both cancer cell lines and patient tumors (2,816 of 21,903 genes; **Fig. 1C**; Supplementary Table S1). This set of bimodal genes showed, in general, better reproducibility of expression status (high *vs.* low) than using the rest of the genes (Supplementary Fig. S1A; see Materials and Methods). Pathway enrichment analysis revealed a significant association of bimodal genes with G protein–coupled receptor signaling (GPCR) related pathways (Supplementary Fig. S1B), which are involved in the modulation of PI3K pathway, MAPK proteins, cAMP-dependent protein kinases, and cellular Ca2+ (32). Further characterization of these strongly bimodal genes revealed a low correlation (median Matthews correlation coefficients, MCC = 0.03, IQR = 0.08) between their mRNA expression and hence low redundancy in the information they carry (Supplementary Fig. S1C).

### Development of interpretable models predictive of drug sensitivity

We implemented a machine learning approach based on logic-based models to identify reliable and interpretable biomarkers of sensitivity to different drugs, building upon the recent LOBICO approach (28). Logic-based models offer logic formulas using the 'AND,' 'OR,' and 'NOT' operators to build multivariate, nonlinear, yet interpretable predictive models. They overcome the limitations of univariate models that do not account for versatile gene dependencies. To make such models broadly available, we developed RLOBICO, which is an R implementation of LOBICO method (28), to find binary rules that predict sensitivity of samples to different drugs. Several studies have shown that a reduced feature space improves the predictivity and interpretability of drug sensitivity models (33–35). To

reduce the feature space and subsequent computational cost, we used the ensemble mRMRe feature selection strategy (**Fig. 1A**; ref. 36). The resulting models were represented as logic formulas including ≤ 10 genes to control the risk of overfitting and facilitate interpretation of the models. We assessed the predictive value of the logic models using the CI (see Materials and Methods; refs. 31, 37).

To fit the logic-based models, we used the pharmacogenomic data from the CTRP by the Broad Institute, which represents one of the largest sets of drug response data publicly available to date (version 2, including 544 drugs; refs. 5, 18), extracted from PharmacoGx (version 1.14.0; ref. 22). We excluded drugs for which less than 10% of tested cancer cell lines are sensitive [area above the drug dose–response curve (AAC) ≥ 0.2; see Materials and Methods for AAC definition]. Supplementary Figure S2 depicts the different filtration steps applied on the pharmacogenomic datasets. On the basis of our approach, we were able to build models yielding a CI greater than 0.6 ($P$ value < 0.05) in a 5-fold cross-validation setting for 39.9% of the drugs in CTRPv2 (**Fig. 2A**). CI > 0.6 was chosen as a threshold to define robust models based on the evaluation of CI scores between 11 drugs of different drug classes and their associated known biomarkers (Supplementary Fig. S3; Supplementary Table S2). The models cover a wide spectrum of drug classes such as EGFR signaling inhibitors and RTK signaling inhibitors (**Fig. 2B**; Supplementary Table S3) supporting the generalizability of the predictive value of bimodal genes. Interestingly, we found that among all models chosen by our method to predict drug sensitivity using the bimodal genes, 87% of multivariate models had a higher predictive value (CI) than univariate ones (Supplementary Fig S4). Here, for each drug, we compared the best univariate model ($K = 1$ AND $M = 1$; see Materials and Methods) to the best multivariate model ($K \,! = 1$ OR $M \,! = 1$; see Materials and Methods) on the same data. This supports observations by other studies that multivariate models have, in general, better predictive power than univariate ones (1, 7, 38, 39). However, it is not always the case as some drugs are greatly influenced by specific gene alterations, for example, dabrafenib is mainly predicted by BRAF status.

The top-performing predictors included drugs targeting growth factor receptors such as EGFR, ERBB2, and VEGFR. As mentioned earlier, the bimodal genes are enriched for several GPCR-related pathways. GPCRs are involved in cross-talks with growth factor receptors. Transactivation of EGFR in cancer cell lines by GPCRs such as chemokine and angiotensin II receptors has been reported extensively (40). Persistent transactivation of EGFR and ErbB2/HER2 by protease-activated receptor-1, a GPCR activated by extracellular proteases, has been shown to promote breast carcinoma cell invasion (41). In addition, a strong complex formation between VEGFR2, another major growth factor, and the GPCR β2–adrenoceptor has been reported, resulting in VEGFR2 activation (42). Among our top-performing models, we found that higher expression of FGF-binding protein 1 (FGFBP1) was correlated with increased sensitivity to erlotinib (**Fig. 2C**). FGFBP1 is a secreted chaperone that helps release

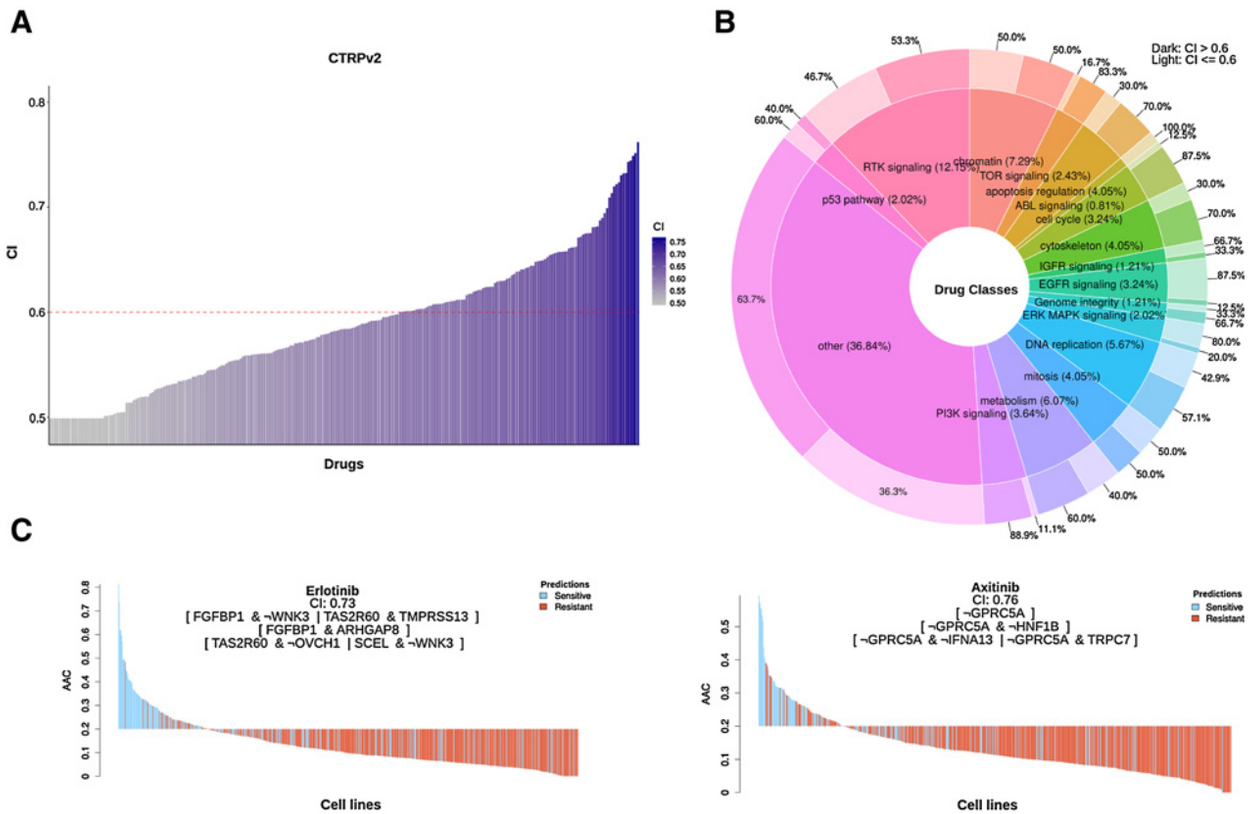


**Figure 2.**

**A,** Cross-validation performance of developed logical models for each drug in CTRPv2. Red-dashed line represents cutoff for good and bad models. **B,** Distribution of good (CI > 0.6; dark color) and bad (CI ≤ 0.6; light color) models (outer ring) for each drug class in CTRPv2 and distribution of drug classes in CTRPv2 (inner ring). **C,** Examples of top-performing trained logical models along with the rules predicted to assess sensitivity to the respective drugs.

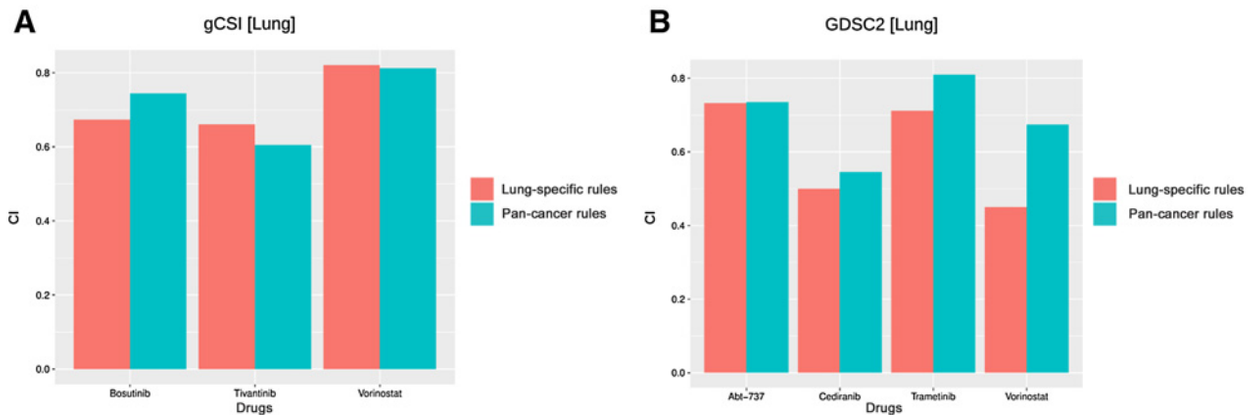


**Figure 3.**
Comparing lung-specific rules versus pan-cancer rules in predicting drug response within lung samples in gCSI (**A**) and GDSC2 (**B**).

fibroblast-binding factors (FGF), stored in the extracellular matrix, and presents them to their cognate receptors, thereby enhancing FGF signaling. FGFBP1 mediated carcinogenesis has been implicated in many studies (43). According to Verbist and colleagues (44), FGFBP1 gene expression is downregulated by erlotinib, resulting in decreased cell proliferation in cancer. These studies support our findings that high expression of FGFBP1 might be imparting sensitivity to erlotinib via the inhibition of FGFBP1-FGF signaling axis. EGFR expression, a known biomarker for erlotinib, was excluded from our set of bimodal genes because its expression was not sufficiently bimodal in the TCGA cohort. Yet, we found a significant correlation between predictions based on rules that our method generated for erlotinib and EGFR expression (PCC, 0.34; *P* value, 8.03E-18) suggesting that our method was able to find a surrogate mimicking EGFR association with erlotinib response. Moreover, rules that our method generated for erlotinib had better association with erlotinib response (PCC, 0.36; *P* value, 1.63E-20) than EGFR expression (PCC, 0.29; *P* value, 2.17E-13). Another example of the top-performing models is for the selective VEGFR inhibitor axitinib, in which low expression of GPCR, class C, group 5, member A (GPRC5A) was shown to be predictive of response (**Fig. 2C**). GPRC5A, also known as retinoic acid-induced gene 3 has been shown to elicit tissue-specific oncogenic and tumor-suppressive functions and is involved in the regulation of major cancer-related signaling pathways such as cAMP, NF-κB, and STAT3 (45, 46). Besides STAT3 and NF-κB signaling, GPRC5A is reported to impact cell-cycle genes such as FEN1, MCM2, CCND1, and UBE2C in lung adenocarcinoma (47). Knockout of GPRC5A has been reported to reduce proliferation and migration ability of PaCa cell lines and suppress the chemotherapy drug resistance of gemcitabine, oxaliplatin, and fluorouracil in PaCa cells (48). Knockdown of GPRC5A has also been found to negatively impact FAK/Src activation, and RhoA GTPase activity, the key mediators of VEGF signaling in cancer cell lines (49). These findings support a possible mechanism for axitinib sensitivity imparted by low expression of GPRC5A, via VEGF-activated signaling intermediates. All trained models (CI > 0.6) from CTRPv2 and their associated predictive rules are shared in the supplementary data (Supplementary File 1).

The prospect of tissue specificity for drug response predictions constitutes another layer of complexity. We investigated whether bimodal genes within a specific tissue could generate a more accurate predictor of sensitivity for samples of that tissue type. One challenge is the low number of samples within tissues, which will impact the general applicability of the generated bimodal genes (Supplementary Fig. S5; Supplementary Table S4). Lung cancer was chosen as a case study, as the number of samples available in both CCLE (173 samples) and TCGA (1,122 samples) is sufficient to extract reliable bimodal genes. We applied our pipeline to these samples and developed logic-based models with minimum predictive value (CI > 0.6) for about 30% of drugs in CTRPv2. ABT-737, a selective inhibitor of BCL-2 used in lung cancer therapy, was among the best performing models we found (CI = 0.78). In addition, we validated our predicted rules for this drug on an external dataset of lung cancer samples from the GDSC2 (2019; refs. 2, 3), CI = 0.73. Furthermore, we compared the pan-cancer and tissue-specific models on lung samples in the gCSI (2018; refs. 19, 20) dataset and GDSC2 and found that both feature sets yielded similar associations with response (**Fig. 3A** and **B**). These results suggest that both sets of rules can be predictive of drug response and provide different levels of biomarker granularity. The variation in defining bimodal genes is mostly due to the difference in the distributions of genes within tissues and across different cancer types.

## Bimodality of gene expressions is a rich source of predictive biomarkers

To test whether the gene expression of the top bimodal genes composes a richer feature set for predicting drug response than other data types such as tissue of origin, mutation, and CNV, we systematically analyzed all the data types by running them through the same computational pipeline used for bimodal genes. Our results indicate that the expression of bimodal genes significantly outperformed the other data types (mutations and CNV) in 72% of the drugs (**Fig. 4A** and **B**). We also show that bimodal genes are not surrogates to tissue types in two ways. First, for each drug, we compared drug sensitivity predictions based on RNA sequencing (RNA-seq) models to those based on tissue models and found a very low correlation between them (median MCC, 0.11; IQR, 0,09; **Fig. 4C**). Second, we assessed the ability of each bimodal gene to predict any tissue type and found that no bimodal gene was able to fully predict a tissue type (See Materials and Methods; Supplementary Fig. S6). Tissue type of the sample was found to be the best model predicting sensitivity to 16% of the drugs, suggesting a strong specificity of drug response (**Fig. 4D**; ref. 50). Dabrafenib, for example, is an inhibitor of BRAF serine-threonine kinase, which was predicted by our model to show a high association with skin cancer (**Fig. 4D**). This drug is indeed approved by FDA as a single agent for the treatment of patients with unresectable or

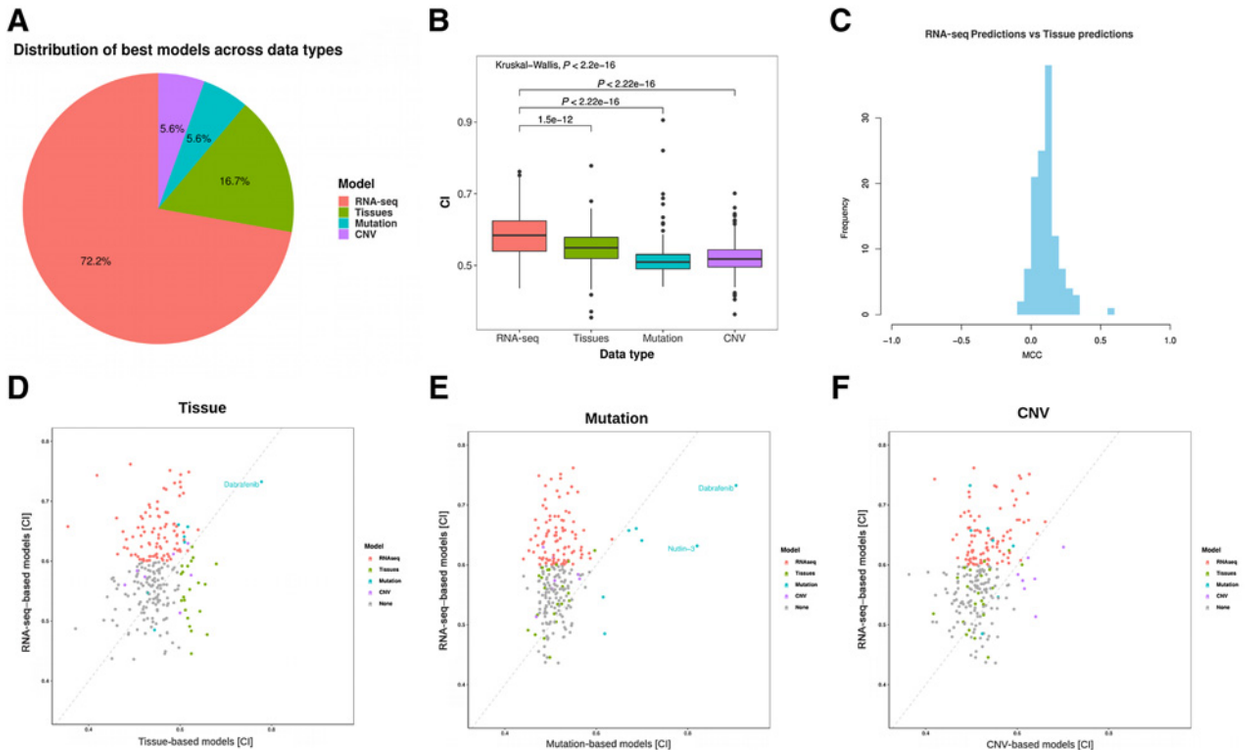


**Figure 4.**

**A,** Distribution of best models across data types. **B,** Statistical comparison between models across data types. *P* values are based on Wilcoxon signed-rank test and Kruskal–Wallis test was used to compare between all groups. **C,** Comparison between RNA-seq–based predictions and tissue-based predictions (median, 0.11; IQR, 0,09). **D–F,** Comparing RNA-seq–based models with tissues (**D**), mutation (**E**), CNV (**F**). Color indicates best models across all data types

metastatic melanoma with BRAF V600E (51). Mutation and CNV features were found to be the best in predicting sensitivity in 11.2% of the drugs (**Fig. 4E** and **F**). An example of these drugs is Nutlin-3A, an MDM2 inhibitor that activates wild-type p53 (52). TP53 wild-type was predicted by our approach to indicate sensitivity to Nutlin-3A (**Fig. 4E**). This outperformance of expression data in comparison with other data types conforms with previous studies and community efforts that investigated the relevance of different data types to predict drug sensitivity and showed that gene expression has more rich information and predictive power than other data types in general (7). Nevertheless, some drugs' responses are still better predicted by other data types such as mutations, CNVs, and tissue types. These results also suggest that combining these different data types in a multi-omics model could improve the resultant predictors given the heterogeneity of the chosen feature sets we observed for different drugs (**Fig. 4A**).

## Validation of drug response predictors

Recognizing that large-scale pharmacogenomic studies employ complex, potentially noisy experimental protocols (20, 53), it is crucial to validate the performance of our new predictors in fully independent datasets to assess their generalizability. Therefore, we validated our models on two large pan-cancer pharmacogenomic datasets, namely gCSI and GDSC2, both included in our PharmacoGx package (22). Among all the models in common with gCSI, our models achieved 92% validation rate (12 drugs of 13 in common with CTRPv2 for which training CI > 0.6; **Fig. 5A**; Supplementary Fig. S2). On GDSC2, our models achieved a validation rate of 61% (16 drugs of 26 in common with CTRPv2 for which training CI > 0.6; **Fig. 5B**; Supplementary Fig. S2). To put these rates in perspective, Supplementary Fig. S3 shows the CI scores between 11 drugs of different drug classes and their associated known biomarkers (Supplementary Table S2). If we consider each biomarker as a different model and consider CI > 0.6 to be a validated model, we would achieve 19% validation in this small set of known biomarkers models.

There were 7 of 9 (78%) predictive models that were validated on both external datasets (**Fig. 5**), strongly supporting the generalizability of the logic rules predictive of drug response. Dasatinib, whose predictive logic model was also validated in GDSC2 and gCSI, showed association with several genes including high mobility group AT-Hook 2 (HMGA2). HMGA2 is a member of the HMG protein family that binds to the DNA minor groove at sequences rich in A and T nucleotides, and acts as a transcriptional regulator. Apart from its role as a transcriptional co-regulator, HMGA2 has been found to induce EMT in lung cancer (54). HMGA2 also functions as a positive regulator of cell proliferation and its expression is implicated as a prospective diagnostic biomarker in the assessment of endometrial serous cancer (55). According to Turkson J. and colleagues (56), nuclear Src and p300 associate with HMGA2 promoter and regulate its gene expression in PDAC patient samples. Src inhibition by dasatinib might negatively impact HMGA2 mediated cell oncogenesis, resulting in sensitivity in cancers with high HMGA2 expression as predicted in our study. Among the other top-performing drugs, the sensitivity of gefitinib, an EGFR inhibitor has been attributed to the expression of ARHGAP8, a gene implicated in EGFR-mediated ERK1/2 phosphorylation and oncogenesis (57, 58). The expression of other bimodal genes associated with lapatinib sensitivity such as

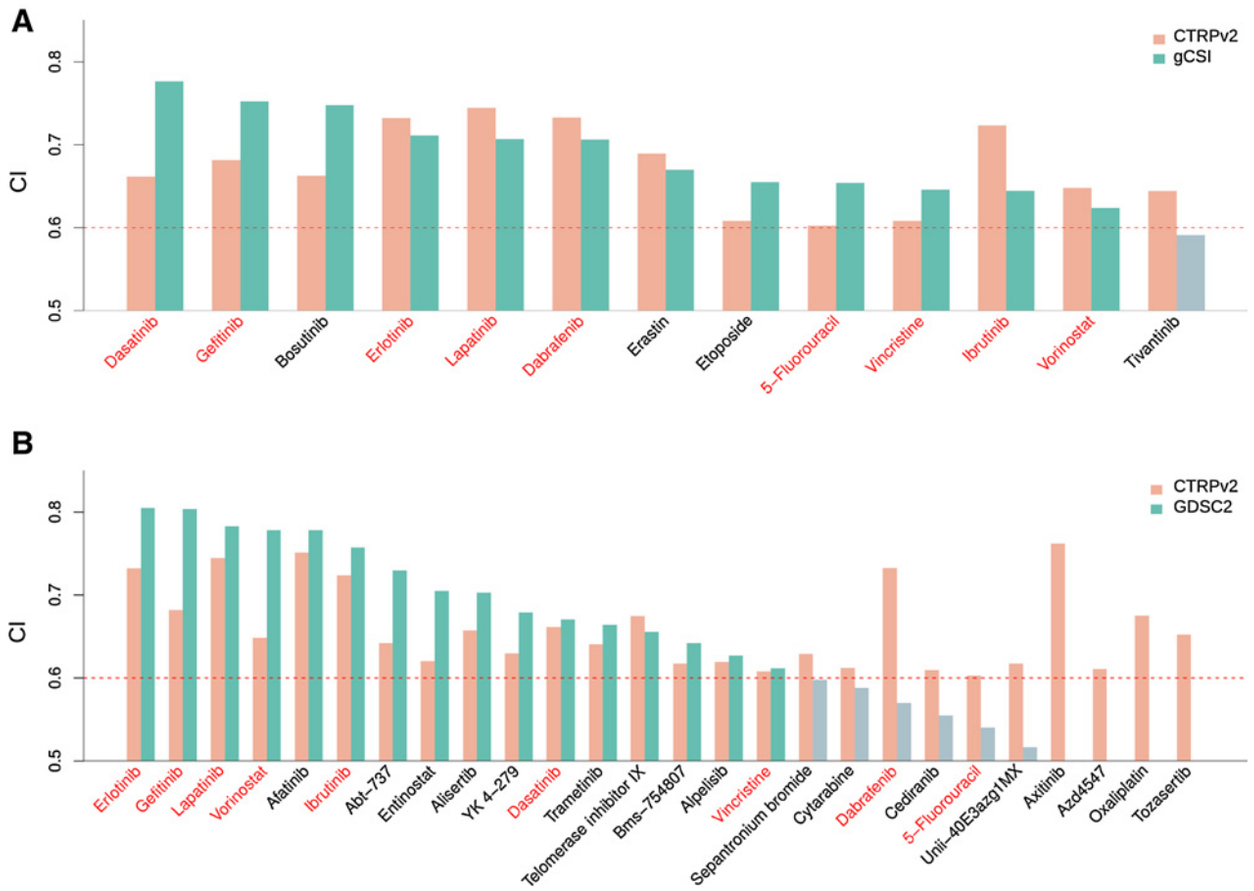


**Figure 5.**

Validating developed logic models on external datasets. **A,** gCSI. **B,** GDSC2. Red colored drugs are common between GDSC2 and gCSI. Dark green bars represent models with CI > 0.6 and light green bars represent models with CI < 0.6.

MARVELD3 and EPN3 has been reported to promote migration and invasion of cancer cells (59–61). These results provide high support for the reliability of our predicted biomarkers, which could improve the existing poor clinical performance for many drugs such as the limited performance of EGFR inhibitors (62–65).

The logic model predictive of erlotinib response described previously (**Fig. 2C**) yielded a high predictive value in both independent *in vitro* datasets (CI of 0.79 and 0.73 in GDSC2 and gCSI, respectively), suggesting high confidence in the generated rule-based biomarkers. We then sought to validate the erlotinib logic model in *in vivo* setting on lung cancer PDX from the Novartis PDXE dataset (26, 27). We used the angle between the curves of the control and treated PDXs as the response metric. Our results showed a significant validation of the erlotinib logic model on PDXE lung samples (CI, 0.63 and *P* value, 1.12E-02; **Fig 6**). Growth curves and angle scores for the PDXs treated with erlotinib can be found in Supplementary Fig S7. These results suggest that predictors based on bimodal gene expressions can be translated from *in vitro* to *in vivo* cancer models.

One of the major bottlenecks of biomarker predictions is its translatability to the clinic. The ultimate aim of this work is to provide models and biomarkers that can aid in treatment decisions. In addition to the cross-validation on preclinical data, we sought to assess the extent of translatability of our models to patient data. We used a rich clinical genomics dataset from the HMF (https://www.hartwigmedicalfoundation.nl), where molecular data of patients of over 30 cancer types is coupled with therapy response measured in clinical trial settings. We found that three drugs (pazopanib, doxorubicin, gemcitabine) with at least one positive outcome in HARTWIG dataset intersect with our set of predictive models with a lowered CI threshold of 0.55. Among these models, we were able to validate two drugs (**Fig. 7**). In this analysis, for each drug, we applied the *in vitro* generated rules to classify HMF patients into respondent/nonrespondents and compared that to the RECIST response on those patients (**Fig. 7**; Supplementary Table S5). Collectively, these results support the reliability of our predictive models paving the way for more efficient designs of molecular targets and betterment of biomarkers used by clinicians.

## Discussion

Bimodality of gene expression represents an interesting phenomenon associated with several biological processes. One of the advantages of bimodal genes as biomarkers is that they can be used to robustly classify samples into two distinct expression states based on their modes, allowing for easier interpretation and translation of the biomarker into the clinic. In this study, we showed that top bimodal genes are mostly associated with extracellular membrane pathways, which have a downstream effect on important cancer-related processes

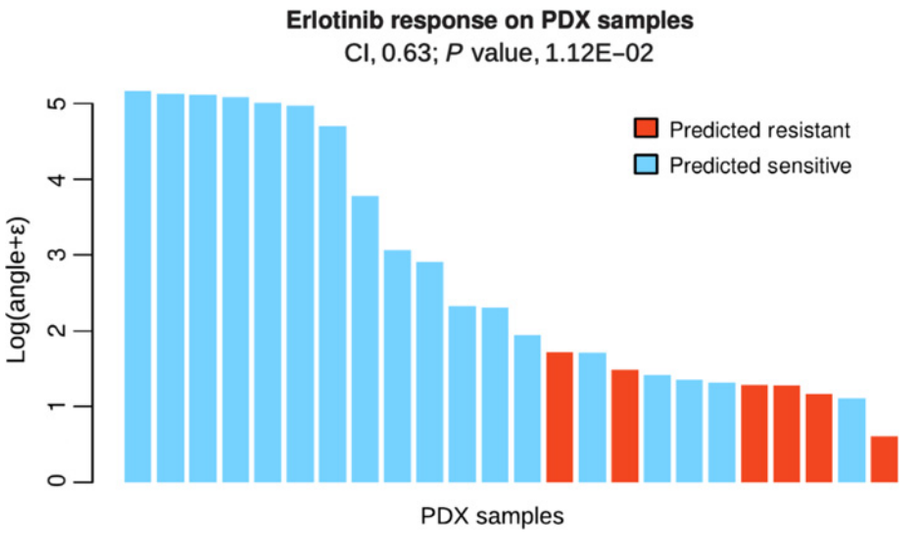


**Figure 6.**
Erlotinib's logic model validation on PDX samples (lung cancer). *Y*-axis is erlotinib response based on angle between control and treated PDXs (higher angle represents higher response). ε is (1-minimum angle) and normalization is used to show differences at low angles. *P* value is based on Wilcoxon signed-rank test.

such as MEK and PI3K signaling. We introduced the largest comprehensive set of bimodal genes derived from a large panel of cancer cell lines tested against hundreds of drugs and patient data from TCGA. In addition, we found a high correlation between the bimodality scores of the corresponding genes within the cell line and patient datasets (PCC, 0.695; $P < 2.2e\text{-}16$), which showcases the reliability of the chosen genes to be globally bimodal within cancer. We found a subset of genes that exhibited a bimodal distribution in one data but not in the other probably due to differences in tissue distribution of samples, or to intrinsic transcriptional differences between the *in vitro* models and the patient tumors.

Although the bimodality of expression provides multiple advantages in biomarker discovery, restricting the modeling to only bimodal genes filters out many known drug biomarkers because their expressions do not follow a bimodal distribution. EGFR expression, for example, is a known biomarker for erlotinib. However, it is not bimodal in TCGA, which excluded it from our set of bimodal genes that we used for training the models. Yet, as shown earlier, we found a high concordance between predictions based on rules our method generated for erlotinib and EGFR expression, suggesting that bimodal genes could still capture information related to those genes that were filtered out. Despite the constraint on the number of bimodal genes we use, we have shown that this set of features along with our novel method of applying logic-based models was able to predict sensitivity to 101 drugs from 17 different drug classes, suggesting global utility of these features (**Fig. 2B**).

We also showed that, in general, bimodal genes outperformed other data types, mutations and CNV, in predicting sensitivity to different drugs. Though, few drugs are still better predicted by genomic features. An interesting follow-up to this analysis would be to investigate the complementary effect of merging these data types in building more accurate models. Challenges that we anticipate are the availability of data types across datasets, data normalization and computational complexity to search the larger search space for candidate rules.

This study has potential limitations. First, the computational cost of the CPLEX optimizer used to train the logical models is limiting the development of predictors including more than a few dozen mutations, CNV and bimodal genes; this limitation could be overcome by the development of new heuristic approaches in the future. Second, a better translation of the bimodal gene expression features from *in vitro* to the clinic could be achieved by incorporating data from system models such as organoids and PDXs. However, there is yet no large compendium of such data that can be used for this purpose. Still, our study provides a way to translate the bimodal genes extracted from *in vitro* samples to the clinic by intersecting them with patient data from TCGA. Third, our study shows the potential of tissue-specific models to predict drug sensitivity. However, we could not showcase many tissues due to the low number of samples within tissues in the datasets used in this study. Finally, with more pharmacogenomic studies being generated, we will have more common drugs across datasets, providing richer data for independent validation, which is crucial to ensure robustness and accuracy of the generated models.

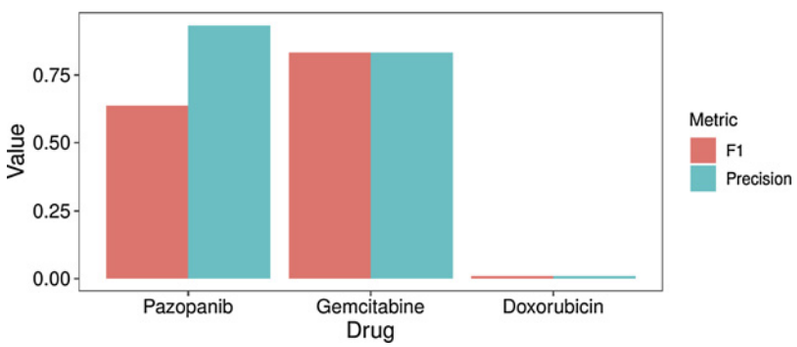


**Figure 7.**
Performance of predictive models (CI > 0.55) on clinical data provided by the HARTWIG Foundation.

Finding reliable and interpretable biomarkers that can predict patients' response to anticancer drugs remains a formidable challenge. We showed that bimodal genes represent a rich set of features for biomarker discovery and that they cover important cancer-associated pathways. Our results, using logic-based models to generate rules that predict sensitivity to drugs, show that genes exhibiting bimodal expression can be used to robustly predict drug response across datasets. These bimodal predictive biomarkers have a high potential of clinical translatability given the clear separation they would provide between patients responder and nonresponder cohorts, and the practicality of measuring a few genes for treatment planning using various targeted assays instead of whole-genome sequencing.

## Authors' Disclosures

B. Li reports personal fees from University Health Network during the conduct of the study. B. Haibe-Kains reports personal fees from Code Ocean Inc. outside the submitted work and is a part of the SAB of the Break Through Cancer Foundation. No disclosures were reported by the other authors.

## Authors' Contributions

**W. Ba-Alawi:** Conceptualization, data curation, software, formal analysis, validation, investigation, methodology, writing–original draft. **S. Kadambat Nair:** Data curation, validation, writing–review and editing. **B. Li:** Software, writing–review and editing. **A. Mammoliti:** Resources, data curation, writing–review and editing. **P. Smirnov:** Resources, data curation, writing–review and editing. **A.S. Mer:** Resources, data curation, writing–review and editing. **L.Z. Penn:** Conceptualization, validation, writing–review and editing. **B. Haibe-Kains:** Conceptualization, supervision, funding acquisition, validation, writing–review and editing.

## Acknowledgments

The authors would like to thank the investigators of the GDSC, the CCLE, Genentech (gCSI), the CTRP, the HMF, and the Center of Personalized Cancer Treatment who have made their valuable pharmacogenomic data available to the scientific community. This work was supported by the Terry Fox Research Institute, Canadian Institutes of Health Research, the Princess Margaret Cancer Foundation, and a Stand Up To Cancer Canada - Canadian Breast Cancer Foundation Breast Cancer Dream Team Research Funding, with supplemental support of the Ontario Institute for Cancer Research through funding provided by the Government of Ontario (Funding Award Number: SU2C-AACR-DT-18-15). Research funding is administered by the American Association for Cancer Research International - Canada, the scientific partner of SU2C Canada.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received July 27, 2021; revised February 24, 2022; accepted May 6, 2022; published first May 10, 2022.

## References

1. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A landscape of pharmacogenomic interactions in cancer. Cell. 2016;166:740–54.
2. 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.
3. 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;41:D955–61.
4. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol 2016;12:109–16.
5. Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell 2013;154:1151–61.
6. Prasad V Perspective: the precision-oncology illusion. Nature 2016;537:S63.
7. 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.
8. Menden MP, Wang D, Mason MJ, Szalai B, Bulusu KC, Guan Y, et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. Nat Commun 2019;10:2674.
9. Wang J, Wen S, Symmans WF, Pusztai L, Coombes KR The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. Cancer Inform 2009;7:199–216.
10. Bessarabova M, Kirillov E, Shi W, Bugrim A, Nikolsky Y, Nikolskaya T. Bimodal gene expression patterns in breast cancer. BMC Genomics 2010;11(Suppl 1):S8.
11. Ertel A. Article Commentary: Bimodal Gene expression and Biomarker Discovery. Cancer Inform 2010;9:CIN.S3456.
12. Muftah AA, Aleskandarany M, Sonbul SN, Nolan CC, Rodriguez MD, Caldas C, et al. Further evidence to support bimodality of estrogen receptor expression in breast cancer. Histopathology 2017;70:456–65.
13. Kim C, Tang G, Pogue-Geile KL, Costantino JP, Baehner FL, Baker J, et al. Estrogen receptor (ESR1) mRNA expression and benefit from tamoxifen in the treatment and prevention of estrogen receptor–positive breast cancer. J Clin Oncol 2011;29:4160–7.
14. Kernagis DN, Hall AHS, Datto MB. Genes with bimodal expression are robust diagnostic targets that define distinct subtypes of epithelial ovarian cancer with different overall survival. J Mol Diagn 2012;14:214–22.
15. Yu R, Longo J, van Leeuwen JE, Mullen PJ, Ba-Alawi W, Haibe-Kains B, et al. Statin-Induced cancer cell death can be mechanistically uncoupled from prenylation of RAS family proteins. Cancer Res 2018;78:1347–57.
16. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603–7.
17. Safikhani Z, Smirnov P, Thu KL, Silvester J, El-Hachem N, Quevedo R, et al. Gene isoforms as expression-based biomarkers predictive of drug response *in vitro*. Nat Commun 2017;8:1126.
18. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. Cancer Discov 2015;5:1210–23.
19. Klijn C, Durinck S, Stawiski EW, Haverty PM, Jiang Z, Liu H, et al. A comprehensive transcriptional portrait of human cancer cell lines. Nat Biotechnol 2015;33:306–12.
20. Haverty PM, Lin E, Tan J, Yu Y, Lam B, Lianoglou S, et al. Reproducible pharmacogenomic profiling of cancer cell line panels. Nature 2016;533:333–7.
21. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas pan-cancer analysis project. Nat Genet 2013;45:1113–20.
22. Smirnov P, Safikhani Z, El-Hachem N, Wang D, She A, Olsen C, et al. PharmacoGx: an R package for analysis of large pharmacogenomic datasets. Bioinformatics 2016;32:1244–6.
23. Feizi N, Nair SK, Smirnov P, Beri G, Eeles C, Esfahani PN, et al. PharmacoDB 2.0: improving scalability and transparency of *in vitro* pharmacogenomics analysis. Nucleic Acids Res 2022;50:D1348–57.
24. Smirnov P, Kofia V, Maru A, Freeman M, Ho C, El-Hachem N, et al. PharmacoDB: an integrative database for mining *in vitro* anticancer drug screening studies. Nucleic Acids Res 2018;46:D994–1002.
25. Bray NL, Pimentel H, Melsted P, Pachter L Erratum: near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 2016;34:888.
26. Poh A Novartis compiles mouse avatar "Encyclopedia. Cancer Discov 2016;6:5–6.
27. Mer AS, Ba-Alawi W, Smirnov P, Wang YX, Brew B, Ortmann J, et al. Integrative pharmacogenomics analysis of patient-derived xenografts. Cancer Res 2019;79:4539–50.
28. Knijnenburg TA, Klau GW, Iorio F, Garnett MJ, McDermott U, Shmulevich I, et al. Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy. Sci Rep 2016;6:36812.
29. Sharifi-Noghabi H, Jahangiri-Tazehkand S, Smirnov P, Hon C, Mammoliti A, Nair SK, et al. Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. Brief Bioinform 2021;22:bbab294.

30. Safikhani Z, Smirnov P, Freeman M, El-Hachem N, She A, Rene Q, et al. Revisiting inconsistency in large pharmacogenomic studies. F1000Res. 2016;5: 2333.
31. Smirnov P, Smith I, Safikhani Z, Ba-Alawi W, Khodakarami F, Lin E, et al. Evaluation of statistical approaches for association testing in noisy drug screening data. arXiv [stat.AP]. 2021.
32. Goldsmith ZG, Dhanasekaran DN G protein regulation of MAPK networks. Oncogene 2007;26:3122–42.
33. Koras K, Juraeva D, Kreis J, Mazur J, Staub E, Szczurek E. Feature selection strategies for drug sensitivity prediction. Sci Rep 2020;10:9377.
34. Naulaerts S, Menden MP, Ballester PJ. Concise polygenic models for cancer-specific identification of drug-sensitive tumors from their multi-omics profiles. Biomolecules 2020;10:963.
35. Parca L, Pepe G, Pietrosanto M, Galvan G, Galli L, Palmeri A, et al. Modeling cancer drug response through drug-specific informative genes. Sci Rep 2019;9: 15222.
36. De Jay N, Papillon-Cavanagh S, Olsen C, El-Hachem N, Bontempi G, Haibe-Kains B. mRMRe: an R package for parallelized mRMR ensemble feature selection. Bioinformatics 2013;29:2365–8.
37. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA Evaluating the yield of medical tests. JAMA 1982;247:2543–6.
38. Nguyen L, Dang CC, Ballester PJ Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data. F1000Res 2016;5:ISCB Comm J-2927.
39. Naulaerts S, Dang CC, Ballester PJ Precision and recall oncology: combining multiple gene mutations for improved identification of drug-sensitive tumors. Oncotarget 2017;8:97025–40.
40. Köse M GPCRs and EGFR - Cross-talk of membrane receptors in cancer. Bioorg Med Chem Lett 2017;27:3611–20.
41. Arora P, Cuevas BD, Russo A, Johnson GL, Trejo J Persistent transactivation of EGFR and ErbB2/HER2 by protease-activated receptor-1 promotes breast carcinoma cell invasion. Oncogene 2008;27:4434–45.
42. Kilpatrick LE, Alcobia DC, White CW, Peach CJ, Glenn JR, Zimmerman K, et al. Complex Formation between VEGFR2 and the β2-Adrenoceptor. Cell Chem Biol 2019;26:830–41.
43. Schmidt MO, Garman KA, Lee YG, Zuo C, Beck PJ, Tan M, et al. The role of fibroblast growth factor-binding protein 1 in skin carcinogenesis and inflammation. J Invest Dermatol 2018;138:179–88.
44. Verbist B, Klambauer G, Vervoort L, Talloen W, QSTAR Consortium, Shkedy Z, et al. Using transcriptomics to guide lead optimization in drug discovery projects: lessons learned from the QSTAR project. Drug Discov Today 2015;20:505–13.
45. Deng J, Fujimoto J, Ye X-F, Men T-Y, Van Pelt CS, Chen Y-L, et al. Knockout of the tumor suppressor gene Gprc5a in mice leads to NF-κB activation in airway epithelium and promotes lung inflammation and tumorigenesis. Cancer Prev Res 2010;3:424–37.
46. Zhou H, Rigoutsos I The emerging roles of GPRC5A in diseases. Oncoscience 2014;1:765–76.
47. Fujimoto J, Kadara H, Men T, van Pelt C, Lotan D, Lotan R Comparative functional genomics analysis of NNK tobacco-carcinogen induced lung adenocarcinoma development in Gprc5a-knockout mice. PLoS One 2010; 5:e11847.
48. Liu B, Yang H, Pilarsky C, Weber GF The effect of GPRC5a on the proliferation, migration ability, chemotherapy resistance, and phosphorylation of GSK-3β in pancreatic cancer. Int J Mol Sci 2018;19:1870.
49. Chen XL, Nam J-O, Jean C, Lawson C, Walsh CT, Goka E, et al. VEGF-induced vascular permeability is mediated by FAK. Dev Cell 2012;22:146–57.
50. Yao F, Madani Tonekaboni SA, Safikhani Z, Smirnov P, El-Hachem N, Freeman M, et al. Tissue specificity of *in vitro* drug sensitivity. J Am Med Inform Assoc 2018;25:158–66.
51. Duffy MJ, Crown J Companion biomarkers: paving the pathway to personalized treatment for cancer. Clin Chem 2013;59:1447–56.
52. Kucab JE, Hollstein M, Arlt VM, Phillips DH Nutlin-3a selects for cells harboring TP53 mutations. Int J Cancer 2017;140:877–87.
53. 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.
54. Gao X, Dai M, Li Q, Wang Z, Lu Y, Song Z HMGA2 regulates lung cancer proliferation and metastasis. Thorac Cancer 2017;8:501–10.
55. Wei L, Liu X, Zhang W, Wei Y, Li Y, Zhang Q, et al. Overexpression and oncogenic function of HMGA2 in endometrial serous carcinogenesis. Am J Cancer Res 2016;6:249–59.
56. Paladino D, Yue P, Furuya H, Acoba J, Rosser CJ, Turkson J A novel nuclear Src and p300 signaling axis controls migratory and invasive behavior in pancreatic cancer. Oncotarget 2016;7:7253–67.
57. Jiang T, Pan CQ, Low BC BPGAP1 spatially integrates JNK/ERK signaling cross-talk in oncogenesis. Oncogene 2017;36:3178–92.
58. Ravichandran A, Low BC SmgGDS antagonizes BPGAP1-induced Ras/ERK activation and neuritogenesis in PC12 cell differentiation. Mol Biol Cell 2013;24: 145–56.
59. Qian H, Tao Y, Jiang L, Wang Y, Lan T, Wu M, et al. PKG II effectively reversed EGF-induced protein expression alterations in human gastric cancer cell lines. Cell Biol Int 2018;42:435–42.
60. Steed E, Elbediwy A, Vacca B, Dupasquier S, Hemkemeyer SA, Suddason T, et al. MarvelD3 couples tight junctions to the MEKK1-JNK pathway to regulate cell behavior and survival. J Cell Biol 2014;204:821–38.
61. Wang Y, Song W, Kan P, Huang C, Ma Z, Wu Q, et al. Overexpression of Epsin 3 enhances migration and invasion of glioma cells by inducing epithelial□-mesenchymal transition. Oncol Rep 2018;40:3049–59.
62. Yu HA, Arcila ME, Hellmann MD, Kris MG, Ladanyi M, Riely GJ Poor response to erlotinib in patients with tumors containing baseline EGFR T790M mutations found by routine clinical molecular testing. Ann Oncol 2014;25:423–8.
63. Margolin KA, Moon J, Flaherty LE, Lao CD, Akerley WL 3rd, Othus M, et al. Randomized phase II trial of sorafenib with temsirolimus or tipifarnib in untreated metastatic melanoma (S0438). Clin Cancer Res 2012;18:1129–37.
64. Sullivan RJ, Flaherty KT Resistance to BRAF-targeted therapy in melanoma. Eur J Cancer 2013;49:1297–304.
65. Eisen T, Ahmad T, Flaherty KT, Gore M, Kaye S, Marais R, et al. Sorafenib in advanced melanoma: a Phase II randomized discontinuation trial analysis. Br J Cancer 2006;95:581–6.