11-9-2018

Predicting Response to Platin Chemotherapy Agents with Biochemically-inspired Machine Learning

Peter Rogan
The University of Western Ontario, progan@uwo.ca

Eliseos J. Mucaki
Western University, emucaki@uwo.ca

Dan Lizotte
Western University, dlizotte@uwo.ca

Signal Transduction and Targeted Therapy, in press

Predicting Response to Platin Chemotherapy Agents with Biochemically-inspired Machine Learning

Eliseos J. Mucaki[1], Jonathan Z.L. Zhao[1], Daniel J. Lizotte[2,3], and [§]Peter K. Rogan[1,2,3,4,5]

Running Title:

Predicting Responses to Platin Drugs by Machine Learning

Author Affiliations

[1]Department of Biochemistry, Schulich School of Medicine and Dentistry, Western University, London, Canada, N6A 2C1

[2]Department of Computer Science, Faculty of Science, Western University, London, Canada, N6A 2C1

[3]Department of Epidemiology & Biostatistics, Faculty of Science, Western University, London, Canada, N6A 2C1

[4]Cytognomix Inc. London, Canada N5X 3X5

[5]Department of Oncology, Schulich School of Medicine and Dentistry, Western University, London, Canada, N6A 2C1

Author Emails: emucaki@uwo.ca, jzhao293@uwo.ca, dlizotte@uwo.ca, progan@uwo.ca

[§]**Correspondence to:** Peter K. Rogan (progan@uwo.ca), Department of Biochemistry, Schulich School of Medicine and Dentistry, Western University, London, Ontario, Canada, N6A 2C1. 1 (519) 661-4255.

ABSTRACT.

Selection of effective genes that accurately predict chemotherapy response could improve cancer outcomes. We compare optimized gene signatures for cisplatin, carboplatin, and oxaliplatin response in the same cell lines, and respectively validate each with cancer patient data. Supervised support vector machine learning was used to derive gene sets whose expression was related to cell line GI_{50} values by backwards feature selection with cross-validation. Specific genes and functional pathways distinguishing sensitive from resistant cell lines are identified by contrasting signatures obtained at extreme vs. median GI_{50} thresholds. Ensembles of gene signatures at different thresholds are combined to reduce dependence on specific GI_{50} values for predicting drug response. The most accurate gene signatures for each platin are: cisplatin: *BARD1*, *BCL2*, *BCL2L1*, *CDKN2C*, *FAAP24*, *FEN1*, *MAP3K1*, *MAPK13*, *MAPK3*, *NFKB1*, *NFKB2*, *SLC22A5*, *SLC31A2*, *TLR4*, *TWIST1*; carboplatin: *AKT1*, *EIF3K*, *ERCC1*, *GNGT1*, *GSR*, *MTHFR*, *NEDD4L*, *NLRP1*, *NRAS*, *RAF1*, *SGK1*, *TIGD1*, *TP53*, *VEGFB*, *VEGFC;* oxaliplatin: *BRAF*, *FCGR2A*, *IGF1*, *MSH2*, *NAGK*, *NFE2L2*, *NQO1*, *PANK3*, *SLC47A1*, *SLCO1B1*, *UGT1A1*. TCGA bladder, ovarian and colorectal cancer patients were used to test cisplatin, carboplatin and oxaliplatin signatures (respectively), resulting in 71.0%, 60.2% and 54.5% accuracy in predicting disease recurrence and 59%, 61% and 72% accuracy in predicting remission. One cisplatin signature predicted 100% of recurrence in non-smoking bladder cancer patients (57% disease-free; N=19), and 79% recurrence in smokers (62% disease-free; N=35). This approach should be adaptable to other studies of chemotherapy response, independent of drug or cancer types.

KEY WORDS. Chemotherapy response, support vector machines, gene signatures, cancer, cisplatin, oxaliplatin, carboplatin, machine learning, bladder cancer, breast cancer, ovarian cancer

INTRODUCTION

Chemotherapy regimens are selected based on overall outcomes for specific types and subtypes of cancer pathology, progression to metastasis, other high-risk indications, and prognosis[1,2], and variability in tumor resistance has led to tiered sequential strategies for selection of agents based on their overall efficacy[3]. We and others have developed machine learning (ML)-based gene signatures (i.e. predictive models) aimed at predicting response to specific chemotherapeutic agents and minimizing chemoresistance based on inhibition of growth or drug targets (GI_{50} or IC_{50})[4–6]. In this study, we present integrated ML models of platin drug responses (cis-, carbo- and oxaliplatin), and evaluate them on clinical outcomes data that were not used to construct the signatures. Previous studies have reviewed the genes[7], gene products[8] and specific individual pathways that are activated and repressed by drugs[9], but lack comprehensive models of the global cellular response to drugs. We use integrated ML-based signatures based on expression of multiple genes to predict key responses to each of these platin agents, for the first time, at different resistance levels.

Cisplatin, carboplatin and oxaliplatin are each widely prescribed compounds for their antineoplastic effects. While each contains platinum to form adducts with tumour DNA, their effectiveness differs for specific types of cancers, such as bladder (cisplatin), ovarian (cisplatin and carboplatin) and colorectal cancer (oxaliplatin). Carboplatin differs in structure from cisplatin, exchanging the latter's dichloride ligands with a CBDCA (cyclobutane dicarboxylic acid) group, while oxaliplatin is paired with both a DACH (diaminocyclohexane) ligand and a bidentate oxalate group. These chelating ligands have greater stability and solubility to aqueous solutions, which lead to differences in drug toxicity compared to cisplatin[10]. Oxaliplatin can be up to two times as cytotoxic as cisplatin, but it forms fewer DNA adducts[11]. The large hydrophobic DACH ligand which

overlaps the major groove is thought to prevent binding of certain DNA repair enzymes such as the POL polymerases, and may contribute to the low cross-resistance between oxaliplatin and cisplatin and carboplatin[10]. While all three drugs can enter the cell via copper transporters, organic cation transporters are oxaliplatin-specific and likely play a role in its efficacy in colorectal cancer (CRC) cells where these transporters are commonly overexpressed[7]. Oxaliplatin specifically plays a role in interfering with both DNA and RNA synthesis, unlike cisplatin which only infers with DNA[12]. It is these intrinsic properties between the platinum drugs which lead to differences in their activity and resistance profiles, despite their similar mode of action.

We derived gene signatures to predict drug response at different sensitivity and resistance levels for each of these agents. We and others have used supervised learning algorithms, including random forest models[13]; support vector machine (SVM) models[6]; neural networks[14]; and linear regression models[5] to make these predictions. Pathway and network analysis of gene expression have been used to indicate hundreds of genes potentially up- and down-regulated upon cisplatin treatment[15]. Cisplatin-specific gene signatures have been developed with integrative approaches such as elastic net regression using inferred pathway activity of bladder cancer cell line data[16]. These methods have implicated genes that have not been described previously. Supervised ML with biochemically-relevant genes has also been useful for predicting drug response[6]. A concern with each of these ML approaches is that an insufficient number of samples coupled to a large number of features, i.e. gene expression changes, in each sample can result in overfitting of the model affecting its generalizability with other sources of data[17]. We therefore reduce the number of dimensions by selecting genes biologically relevant to the drugs under observation[6,17]. In this study, genes included in the final signatures have well-defined roles in their corresponding drug responses

100 (Supplementary References, Section A). Additional selection criteria are necessary
101 when the number of genes implicated in peer-reviewed reports is still prohibitively large
102 compared to sample size.

103 Biochemically-inspired gene signatures have shown good performance in
104 predicting treatment response. A paclitaxel ML signature based on tumor gene
105 expression (GE) had a higher success predicting the pathological complete response
106 rate (pCR [18]) for sensitive patients (84% of patients with no / minimal residual disease)
107 than gene signatures based on differential GE analysis[6]. For gemcitabine, a signature
108 derived from both expression and copy number (CN) data from breast cancer cell lines
109 was derived, and subsequently applied to analysis of nucleic acids from patient archival
110 material. Multiple other outcome measures used to validate gene signatures include
111 prognosis[5], Miller-Payne response[19], and disease recurrence. Binary SVM classifiers
112 based on discrete time thresholds have been used to classify continuous outcome
113 measures such as prognosis and recurrence. By contrast, pCR is simpler to interpret
114 with binary SVM models. Nevertheless, differences in clinical recurrence have been
115 noted between patients demonstrated with pCR and those who do not exhibit disease
116 pathology[18]. This source of variability in defining patient response can confound
117 transferability of SVM models between different datasets.

118 We apply biochemically-inspired ML to predict and compare the cellular and
119 patient responses to cisplatin, carboplatin and oxaliplatin. We train models and perform
120 model selection for classification of platin resistance with cancer cell line data, and
121 validate using patient GE and clinical outcome data. Our previous gene signatures
122 derived from cell line data were based on median GI_{50} for each drug[6]. Models (i.e. gene
123 signatures) learned and selected using the cell line data have not been re-trained prior to
124 application on the patient data, since GI_{50} values are not available in patient samples.

This has been a necessary compromise; however, in this study, we derive different signatures at the highest vs. the lowest levels of drug resistance. A series of candidate gene signatures are derived by shifting the GI_{50} thresholds that distinguish sensitivity from resistance. The frequency of genes selected at median vs. extreme thresholds highlights pathways that most likely define these responses among different patient subsets.

RESULTS

Selection of Platin Drug Related Genes

We documented genes in the peer-reviewed literature associated with drug effectiveness or response (Supplemental References, Section B). For cisplatin, carboplatin and oxaliplatin, this implicated 179, 90, and 288 genes, respectively (Suppl. Table S1). Multiple factor analysis (MFA) was used to determine which genes were correlated to GI_{50} in breast cancer cell lines through either GE and/or CN[13], significantly reducing the sizes of the gene sets for cisplatin (N=39), carboplatin (N=28), and oxaliplatin (N=55). Genes with significant relationships to GI_{50} and direction of correlation (positive or inverse) are indicated in Figure 1. The diverse functions of these genes included apoptosis, DNA repair, transcription, cell growth, metabolism, immune system, signal transduction and membrane transport. Analysis of IC_{50} and GE levels for cisplatin-treated bladder cancer cell lines confirmed these relationships evident from GI_{50} values of different breast cancer lines. IC_{50} values were related to GE for *CFLAR*, *FEN1*, *MAPK3*, *MSH2*, *NFKB1*, *PNKP*, *PRKAA2*, and *PRKCA*[20]. Similarly, separate bladder cell line IC_{50} values from the Genomics of Drug Sensitivity in Cancer project (http://www.cancerrxgene.org; N=17) were correlated with GE for *CFLAR*, *FEN1*, and *NFKB1*, in addition to *ATP7B*, *BARD1*, *MAP3K1*, *NFKB2*, *SLC31A2* and *SNAI1*.

We performed MFA on the GI_{50} values for cisplatin, carboplatin and oxaliplatin, without consideration of either GE or CN. Responses to cis- and carboplatin were directly correlated (a 6.2º separation between vectors), but neither was related to the oxaliplatin response (Figure 2). Previous studies have shown that cisplatin-resistant cell lines are generally sensitive to oxaliplatin[21–23].

SVM-based signatures were initially derived for each platin drug from breast cancer cell line GE data. A 13-gene signature for cisplatin that predicts whether observed growth inhibition is above or below the median GI_{50} threshold (5.2% cross-validation misclassification rate) consisted of *BARD1, BCL2L1, FAAP24, CFLAR, MAP3K1, MAPK3, NFKB1, POLQ, PRKAA2, SLC22A5, SLC31A2, TLR4,* and *TWIST1*. A similarly derived carboplatin signature included *AKT1*, *ATP7B*, *EGF*, *EIF3I*, *ERCC1*, *GNGT1*, *HRAS*, *MTR*, *NRAS*, *OPRM1*, *RAD50*, *RAF1*, *SCN10A*, *SGK1*, *TIGD1*, *TP53*, and *VEGFB* (10.4% cross-validation misclassification). For oxaliplatin, the final SVM gene signature consisted of *AGXT*, *APOBEC2*, *BRAF*, *CLCN6*, *FCGR2A*, *IGF1*, *MPO*, *MSH2*, *NAGK*, *NAT2*, *NFE2L2*, *NOTCH1*, *PANK3*, *PRSS1*, and *UGT1A1* (2.1% cross-validation misclassification). A cisplatin SVM generated from 17 bladder cancer cell lines in cancerRxgene resulted in 2 equally accurate signatures (with 11.8% cross-validation misclassification) consisting of either *PNKP* and *PRKCA* or *ATP7B*, *CFLAR*, *FEN1*, *MAPK3*, *NFKB1* and *SLC22A11*. These gene signatures were not useful for predicting patient outcomes due to the limited size of the training set.

GI_{50}-Threshold Independent Modeling

In our previous studies, we set median GI_{50} value as the threshold to distinguished drug resistance and sensitivity[5,6]. An important question is whether the genes contributing to drug response are consistent among different cell lines, each with

their own unique GI_{50} values. Different ML gene signatures were obtained by shifting the GI_{50} threshold, which changed the labels of resistant vs. sensitive cell lines. After feature selection, the compositions of the corresponding gene signatures for each threshold were compared. Finally, ensemble averaging of all of these optimized SVMs with Gaussian kernels were derived for different GI_{50} thresholds was used to create a single aggregated, threshold-independent, ML-based predictive model, comprised of all genes that were selected in any of the threshold-specific models (i.e. a composite gene signature).

Kinase (*MAPK3*, *MAP3K1*) genes and apoptotic family members (*BCL2*, *BCL2L1*) were most the common in the cisplatin signatures at different GI_{50} thresholds, with consistent representation of error-prone and base-excision DNA repair genes as well (Figure 3A; Supplementary Table S2A). The kinases are more concentrated in signatures with lower drug sensitivity thresholds, whereas *BCL2* and *BCL2L1* are more ubiquitous at all levels. The error prone polymerases, *POLD1* and *POLQ,* are more frequent in gene signatures with lower sensitivity thresholds, while the flap endonuclease *FEN1* tends to be present at high levels of resistance. Thresholded gene signatures for carboplatin-related genes commonly contained the apoptotic family member *AKT1*, transcription regulation genes *ETS2* and *TP53*, as well as cell growth factors *VEGFB* and *VEGFC*, although the latter was less common at lower sensitivity thresholds (Figure 3B). Common oxaliplatin-related genes included transporters *SLCO1B1* and *GRTP1* (but not *SLC47A1*), transcription genes *NFE2L2*, *PARP15* and *CLCN6*, as well as multiple metabolism-related genes (Figure 3C).

GI_{50} thresholded ML models were also derived using the log-loss function to evaluate whether an alternative loss function (for classification) would differ significantly to the misclassification-based gene signatures (by both the distribution of selected genes and

198 by model accuracy to patient data). Log-loss penalizes false classifications, whose value
199 ranges from zero (or completely accurate), to 1 (or completely inaccurate;
200 Supplementary Table S3). The overall distribution of genes across GI_{50} thresholds has
201 many distinct similarities with the gene signatures derived by misclassification. For both
202 sets of cisplatin gene signatures, *BCL2*, *BCL2L1* and *FEN1* are common in low-to-
203 moderate GI_{50} thresholds, while *NFKB1* is enriched at high thresholds (Figure 3A; Suppl.
204 Figure 1A). For carboplatin, *AKT1*, *VEGFB* and *VEGFC* are similarly distributed across
205 GI_{50} thresholds with both methods, although *VEGFB* is less densely represented in log-
206 loss based gene signatures at low GI_{50} values (Figure 3B; Suppl. Figure 1B). In both
207 sets of oxaliplatin gene signatures, *SIAE* and *SLC47A1* are represented at high densities
208 across all GI_{50} thresholds, whereas *ABCG2* is present less frequently (<50% inclusion;
209 Figure 3C and Suppl. Figure 1C). There are differences between signatures selected by
210 minimizing log-loss and misclassification rates. *EGF* and *ERCC1* were selected at a
211 greater frequency at a moderate carboplatin GI_{50} with log-loss, rather than by
212 misclassification. Similarly, oxaliplatin signature genes, *APOBEC2*, *HLA-B*, *LTA*, and
213 *MPO,* were selected considerably more often by log-loss. Therefore, while the
214 misclassification and log-loss based gene signatures are not interchangeable, overall,
215 they are quite similar to one another.

216 Log-loss gene signatures were initially constructed either by (a) a modified
217 version of the misclassification-based method, or (b) using the backwards feature
218 selection (BFS) software described in Zhao *et al.* (2018)[25]. Multiple signatures with low
219 log-loss values can have different compositions, consistent with the possibility that there
220 may be various diverse gene combinations that can give rise to signatures with
221 satisfactory performance. However, these signatures often contain a larger number of
222 gene features than the misclassification-based signatures, and raised concerns that they

might be more prone to overfitting. The log-loss minimized gene signatures generated by both methods had comparable compositions. The median GI_{50} thresholded cisplatin gene signature generated by the log-loss modified software [*ATP7B*, *BCL2L1*, *CDKN2C*, *CFLAR*, *ERCC2*, *ERCC6*, *FAAP24*, *FOS*, *GSTO1*, *GSTP1*, *MAP3K1*, *MAPK13*, *MAPK3*, *MSH2*, *MT2A*, *PNKP*, *POLD1*, *POLQ*, *PRKAA2*, *PRKCA*, *PRKCB*, *SLC22A5*, *SLC31A2*, *SNAI1*, *TLR4*, *TP63*] shares 15/19 genes with the signature generated by the BFS software[25] [*ATP7B*, *BARD1*, *BCL2*, *BCL2L1*, *ERCC2*, *FAAP24*, *FEN1*, *FOS*, *MAP3K1*, *MAPK13*, *MAPK3*, *MSH2*, *MT2A*, *NFKB1*, *PNKP*, *POLQ*, *PRKCB*, *SLC22A5*, *SNAI1*]).

Impact of Features in Gene Signatures

To determine the contribution of individual genes on overall cross-validation accuracy of a gene signature, each gene was excluded (independently), and model accuracy was reassessed within every SVM signature (Supplementary Tables S2A; S2B and S2C contain cis-, carbo- and oxaliplatin gene signatures, respectively). Elimination of *ERCC2*, *POLD1*, *BARD1*, *BCL2*, *PRKCA* and *PRKCB* consistently significantly increase misclassification error (average > 16% increase) in moderate threshold cisplatin SVMs (GI_{50} thresholds: 5.1 to 5.5). *ERCC2* and *POLD1* perform critical functions in nucleotide and base excision repair, respectively. *PRKCA* and *PRKCB* are paralogs with significant roles in signal transduction. *BARD1* has been shown to reduce apoptotic *BCL2* in the mitochondria[26], and has a key role in genomic stability through its association with *BRCA1*. The genes *NFKB1*, *NFKB2*, *TWIST1*, *TP63*, *PRKAA2*, and *MSH2* show a high variance in increased misclassification between different gene signatures. The variance of these genes may be due to epistatic interactions with other biological components, including the other genes in the SVM. For example, *NFKB1* and *NFKB2* are jointly included in 7 SVMs generated at a moderate GI_{50} threshold. There is evidence of possible epistasis in that the removal of either of these genes, but not

necessary both, will have a large impact in model misclassification rates (≥ 18.0% increase). The misclassification variance of *NFKB1* with *NFKB2*, is significantly lower than in SVM gene signatures lacking *NFKB2*.

Derivation of Gene Signatures from Bladder Carcinoma Patient Data

Gene signatures derived with cell line data are to be validated on cancer patient data. To explore the similarities of the gene signatures to said patient data, we also developed SVMs using the cisplatin and/or carboplatin-treated TCGA (The Cancer Genome Atlas) bladder urothelial carcinoma patients, using post-treatment time to relapse as a surrogate criterion for different GI_{50} resistance thresholds (as performed in Mucaki *et al.* [2017][24]; Supplementary Table S4). Similar trends to cell line SVMs are apparent: *POLQ* is frequently included in gene signatures with recurrence threshold of longer duration, while *FEN1* is a marker of resistance, when time to relapse is shorter. However *BCL2*, which is present in a majority of breast cancer cell line SVMs, is present in only one gene signature derived from TCGA data. Similarly, *MSH2* was rarely selected using cell lines, yet appears in nearly all patient derived SVMs with > 1 year recurrence. However, independently derived patient SVMs could not be used for any other analyses.

Validation of Cell-line based Models against Cancer Patient Data

GI_{50}-thresholded modeling for each platin drug, generated with the breast cancer cell line data, produced 70 cisplatin, 83 carboplatin, and 83 oxaliplatin SVM gene signatures, respectively. In order to understand how the choice of GI_{50} threshold for training on cell line data impacts predictive accuracy when the resulting gene signatures are applied to patient outcomes, each of the thresholded gene signatures was applied to available platin-treated patient datasets[27–31]. In this study, cisplatin gene signatures were

validated on bladder cancer patient data, carboplatin signatures were validated on ovarian cancer patient data, and oxaliplatin signatures were validated on colorectal cancer patient data. While the available data did contain the necessary GE information, the clinical response metadata differed between studies. The response of bladder cancer patients to cisplatin was provided as survival post-treatment by Als *et al.*[30], whereas colorectal cancer patients treated with oxaliplatin were categorized as responders and non-responders by Tsuji *et al* [31]. TCGA provided two different measures which were used to assess predictive accuracy in our gene signatures – clinical response to chemotherapy and disease-free survival. Signature accuracy was found to be similar using either measure (Supplementary Table S5A); however recurrence and disease-free survival was used as the primary measure of response, as it was more consistently recorded among the TCGA data sets tested. Patients from Als *et al.* [30] with a ≥ 5 year survival post-treatment were labeled as sensitive to treatment. The differences between these metadata may, in part, contribute to differences in the prediction accuracy of the thresholded SVM gene signatures.

At higher resistance thresholds for any platin drug (low GI_{50}), where more cell lines are labeled sensitive, the positive class (disease-free survival) is correctly classified, while the negative class (recurrence) is highly misclassified (Suppl. Figures 2 and 3). The reverse is true for gene signatures derived using lower resistance thresholds (high GI_{50}). For these reasons, SVMs generated at these extreme thresholds were not very useful at predicting patient outcomes. When used to predict recurrence in the TCGA datasets, sensitivity and specificity appears to be maximized in gene signatures where the GI_{50} threshold for resistance was set near (but not necessarily at) the median (Suppl. Figure 2; Suppl. Tables S5A to 5C). While this pattern holds true for data from Tsuji *et al.*[30], oxaliplatin gene signatures where GI_{50} thresholds were set above the

297 median could better separate primary and metastatic CRC patients (best signature
298 predicting 92.6% metastatic and 60.7% primary cancers; Suppl. Table S5C). Although
299 less consistent, cisplatin gene signatures generated with thresholds above median GI_{50}
300 performed better when evaluating the patient dataset from Als *et al.*[30] (Suppl. Figure 3).

301 Gene signatures were individually evaluated for their accuracy in TCGA patients
302 using various recurrence times post-treatment to classify resistant and sensitive patients
303 (0.5 - 5 years; Supplemental Table S6A-C). The best performing cisplatin signature
304 (hereby identified as **Cis1**; Table 1) was able to accurately predict 71.0% of bladder
305 cancer patients who recurred after 18 mo. (N=31; 58.5% accurate for disease-free
306 patients [N=41]). The best performing carboplatin gene signature (designated **Car1**
307 [Table 1]) predicted recurrence of ovarian cancer after 4 years at an accuracy of 60.2%
308 (N=302; 61.0% accurate for disease-free patients [N=108]). For oxaliplatin, the best
309 performing gene signature (designated **Oxa1** [Table 1]) accurately predicted 71.6% of
310 the disease-free TCGA CRC patients after one year (N=88; 54.5% accuracy predicting
311 recurrence [N=11]). These gene signatures (based on GE measured by Affymetrix Gene
312 Chip Human Exon 1.0 ST arrays), TCGA sample expression data, as well as SVMs
313 based on bladder cell line data (based on expression measured by Affymetrix U133A
314 microarray), were added to the online web-based SVM calculator
315 (http://chemotherapy.cytognomix.com; introduced in Dorman *et al.* [2016][6]) to predict
316 platin response.

317 The TCGA bladder cancer data set contained 19 patients treated with carboplatin
318 (but not cisplatin), which enabled evaluation of the specificity of cisplatin models relative
319 to patients not treated with this drug. The cisplatin model which best predicted outcomes
320 of carboplatin-treated TCGA bladder patients was not **Cis1** (the best performing cisplatin
321 model) but rather **Cis12** at two years post-treatment (80% accurate for responding

patients [N=5]; 93% for recurrent patients [N=14]). **Cis12** contains 9 genes not present in **Cis1** including *ATP7B*, which is a gene found in many of our carboplatin models. The presence of this gene may have a significant impact on the overall accuracy of **Cis12** to the carboplatin-treated bladder cancer patients. We also evaluated these 19 patients to the carboplatin-specific gene signatures, and found the signature which best predicted the response of these patients (**Car73**) was 84% accurate for patients after 1 year of treatment (100% for responding patients [N=11]; 62.5% accuracy for recurrent [N=8]). Interestingly, **Car73** shares the same *ATP7B* gene with **Cis12**. Two additional carboplatin gene signatures are tied for overall accuracy (84%; **Car9** and **Car51**), but more successfully predict non-responsive patients (87.5%; 82% accuracy for responding patients). *AKT1, ETS2, GNGT1*, and *VEGFB* were shared among these carboplatin gene signatures.

To evaluate the consistency in the response prediction of TCGA bladder cancer patients treated with cisplatin, distances from the hyperplane for all SVMs generated were determined for patients with short recurrence time (<6 mo., N=10; Supplementary Figure 4). Despite showing similar levels of resistance to treatment, distances differed between patients. While these patients would be expected to be indicated as highly cisplatin resistant (hyperplane distance < 0), two patients (TCGA-XF-A9SU and TCGA-FJ-A871) were predicted sensitive across nearly all SVM gene signatures. Similar variation was also seen in patients with either a long recurrence time (>4 years) or no recurrence at all after 6 years (Suppl. Figure 5).

An aggregate, threshold-independent model was generated for each individual platin drug at different GI_{50} thresholds through ensemble ML, which involves the averaging of hyperplane distances for each model to generate a composite score for each TCGA patient tested (i.e. a composite gene signature). Hyperplane distances

across all 70 cisplatin gene signatures were similar, with a mean score of -0.22 and a standard deviation of 3.5 hyperplane units (hu) across the set of patient data. The ensemble model classified disease-free bladder cancer patients with 59% accuracy and those with recurrent disease with 47% accuracy. Limiting ensemble averaging to only cisplatin gene signatures generated at a moderate GI_{50} threshold (ranging from 5.10 to 5.50) did not significantly improve accuracy (44% for disease-free and 66% for recurrent patients; Suppl. Table S7A). For carboplatin, ensemble ML did not produce significantly better predictions than random, regardless of the GI_{50} threshold interval selected (Suppl. Table S7B) or the similar mean hyperplane distances (-0.11 +/- 3.9 hu). For oxaliplatin, the ensemble ML model (mean = -0.12 +/- 2.7 hu) was most accurate after 1 year (60% accuracy for disease-free and 73% for recurrent patients; Suppl. Table S7C). As in cisplatin, limiting this analysis to oxaliplatin SVM gene signatures with moderate GI_{50} thresholds did not significantly increase accuracy.

K-Fold Cross-Validation

The misclassification-based cisplatin, carboplatin and oxaliplatin gene signatures were also evaluated by k-fold cross-validation on TCGA bladder, ovarian and colorectal cancer patient data, respectively. This cross-validation is independent of cell line data; that is, the genes and hyper-parameters of signatures are used, but the GE data used is exclusively from patients. Patients were evenly distributed in 5 groups with an equal (or near-equal) ratio of disease-free and recurrent patients. The majority of the cisplatin gene signatures showed an overall accuracy > 50%. The cisplatin gene signature which performed best under the k-fold analysis (6-resistance level; *BARD1*, *BCL2*, *BCL2L1*, *PRKAA2*, *PRKCA*, *PRKCB*, *TWIST1*) showed an overall accuracy of 71.2% (84.4% accurate for sensitive and 53.9% accurate for resistant patients). The accuracy of the carboplatin and oxaliplatin gene signatures did not exceed 60%. In general, treating the

patient data as a held-out test set yielded higher performance estimates than training and evaluating the models on the patient data using k-fold cross-validation.

Predicting cisplatin response in patients based on smoking history

Tobacco smoking is known as the highest risk factor for the development of bladder cancer[32]. Head and neck cancer patients who smoke while undergoing cisplatin and radiotherapy treatment have been shown to have a shorter overall survival rate[33]. We therefore subdivided the patients based on their smoking history and tested the thresholded gene signatures (Supplementary Tables S8 and S9). When testing patients who were lifelong non-smokers, the prediction accuracy of **Cis1** predicted all non-smoking patients who were recurrent after 18 months as cisplatin-resistant (N=5). Prediction accuracy for disease-free patients was 57.1% (N=14). Another gene signature (**Cis18**; Suppl. Table S8) had performed equally as well for non-smokers, and these two gene signatures share the genes *BCL2*, *BCL2L1*, *FAAP24*, *MAP3K1*, *MAPK13*, *MAPK3*, and *SLC31A2*. Threshold independent analysis predicted disease-free equally well, but recurrence was less accurate (66.7%). Note that non-smokers make up a small subset of the patients tested (N=19). Threshold-independent prediction of recurrence in patients with a smoking history was 46% accurate (N=13), while disease-free patients were correctly predicted at a rate of 58% (N=19). Recurrence in these patients was best predicted by a gene signature built at the median GI_{50} threshold (**Cis2**). Accuracy improved for both disease-free (57.7% -> 61.9%) and recurrent patients (76.0% -> 78.6%) when excluding patients who quit smoking more than 15 years before diagnosis. This SVM includes *CFLAR* and *PRKAA2*, genes which are not present in the two gene signatures which performed well for non-smokers.

To determine which genes in these gene signatures led to discordant predictions of patient outcome, we gradually altered the expression of each signature gene until the misclassification was corrected. Expression of *MAP3K1*, *MAPK3*, *SLC22A5* and *SLC31A2,* when altered, corrected discordant predictions of patient outcome. Altering *BCL2L1* expression was more likely to correct the discordant predictions of **Cis1** (4 out of 5) than with **Cis2** (2 out of 4). If the change exceeded ≥ 3-fold the highest/lowest expression of that gene and the prediction was still unchanged between different patients, the effect of that gene was considered to be minor. Expression of *PRKAA2*, *NFKB1*, *NFKB2* and *TWIST1* could not be altered sufficiently to correct a discordant prediction.

Cytosine Methylation Levels of Genes in Cisplatin Models

Tobacco smoking has a significant impact on cytosine methylation levels in the genome[34]. CpG island methylation has been associated with smoking pack years in a subset of the TCGA bladder urothelial carcinoma patients[26]. We suspected that the level of methylation measured in the SVMs which performed best for smoking and non-smoking patients might differ, and with possible concomitant effects on GE. When ranking each gene from **Cis1** by highest methylation and GE, 88 of 1080 patient: gene combinations showed the expected inverse correlation between methylation levels and GE (i.e. high methylation and low GE). Inverse correlation of methylation and GE was more common than direct correlation (i.e. high methylation and high GE; N=17). However, direct correlation was more common in patients with a recent smoking history (70.5%). This pattern was also observed for **Cis2**, which best predicted recurrence in smokers. In cases where methylation and GE are directly correlated, we propose that smoking may alter expression by other effects, e.g. mutagenic, rather solely than by epigenetic inactivation through methylation.

DISCUSSION

Using gene expression signatures, we derived both GI_{50} threshold-dependent and -independent ML models which predict the chemotherapy responses for cisplatin, carboplatin and oxaliplatin, respectively. The cisplatin gene signature **Cis1** (Supplementary Table S6A) most accurately predicted response in bladder cancer patients after 18 months, and **Car1** (Suppl. Table S6B) best predicted response in ovarian cancer patients after 4 years. **Oxa1** (Suppl. Table S6C) more accurately predicted disease-free patients than recurrent disease at the one year treatment threshold. The thresholds which best represented time-to-recurrence differed between the platin drugs in each cancer type. Cisplatin gene signatures had noticeably improved performance when smoking history was taken into account.

The three platin drugs produce distinctly different gene signatures. Initial gene sets exhibited some overlap between platin drugs (N=67 between any two platins), but very few of these were correlated by MFA of GI_{50} with multiple platin drugs (*ATP7B*, *BCL2* and *MSH2*). *BCL2L1, GSTP1, MAP3K1, MAPK3, MT1A,* and *MT2* were genes common to multiple platin drugs whose expression was correlated with cisplatin GI_{50} values but not with carboplatin and/or oxaliplatin values. Similarly, genes correlating only to carboplatin GI_{50} included *AKT1, EGF, ERCC1, KRAS, LIG3, MTHFR, MTR, RAD50, TP53*, while genes correlating to only oxaliplatin GI_{50} included *ATM, BCL2, CLCN6, ERCC2, ERCC6,* and *UGT1A1*. Despite the close similarity between cisplatin and carboplatin GI_{50} response (see Figure 2), only one gene (*ATP7B*) was related by MFA to GI_{50} levels of both drugs. *BCL2* and *MSH2* correlated with both cisplatin and oxaliplatin GI_{50} (*BCL2* did not correlate with carboplatin GI_{50}). The increase in misclassification caused by the elimination of *MSH2* from any gene signature in which it was present was significant; for example, misclassification of **Cis14** and **Oxa21** (Table 1) were increased

by 28.2% and 19.1%, respectively (Suppl. Tables S2A and S2C). These differences may reflect the spectrum of activity, sensitivity, and toxicity of these signature genes[21–23,35,36].

Previous validation of patient data for other drugs validated with other datasets[6,24] using biochemically inspired machine learning have had better performance than those reported here. We investigated the possibility that disease and molecular heterogeneity in platin-treated patients may have affected the accuracy of our results. Model predictions were reevaluated after stratifying clinical features such as time-to-disease recurrence, cancer stage, and metastatic lymph node count. Breast cancer patients with advanced disease (stage III and IV) were analyzed separately from those with earlier stage diagnoses (stage I and II). Cisplatin gene signature **Cis1** performed best on stage IV patients (overall accuracy 72.4% at a 2 year recurrence threshold), while **Oxa1** similarly performed best in predicting late stage cancers (74.5% accurate for stage III and 71.4% accurate for stage IV at a 2 year recurrence threshold). **Cis5** was also more accurate for later stage cancer patients (72.4% overall accuracy at 18 months). The accuracies of gene signatures were similar across all stages (e.g. **Car1** ranged from 58-74%). Cisplatin-treated, TCGA bladder cancer patients and oxaliplatin-treated TCGA colorectal cancer patients were also stratified by Lymph Node status (N0, N1, and N2 [bladder cancer patient data set comprised of only two N3 patients, which were included with the analysis of N2 patients; N3 was not represented in colorectal cancer]). In TCGA bladder cancer patients, **Cis1** exhibited ~60% accuracy across all categories; however it performed better in sensitive N0 and N1 patients relative to N2. **Cis2** was less accurate for N2 patients than for N0 and N1. Sensitive N2 patients were more likely to be misclassified (<40%) than relapsed N2 patients. In TCGA colorectal cancer patients, **Oxa1** was 88% accurate in N2 patients (95% accurate for sensitive N2 patients [n=19], and 67% accurate for relapsed N2 patients [n=6]). Oxaliplatin gene signatures were less

accurate for N1 patients compared to N0 and N2. Thus, heterogeneity in disease stage as well as metastatic phenotypes adversely confounds the overall accuracies of our predictions.

Gene signature models derived from cell lines and tested on patients differ in their outcome measures. The exact GI_{50} cell line threshold that is most predictive of patient outcome is not known, and different groups use different methods to discretize GI_{50} values[37,38]. Therefore, we developed ML models for platin drugs which predict drug response without relying on arbitrary GI_{50} thresholds. For cisplatin, SVM ensemble averaging generated on different resistance thresholds shows a small increase in accuracy over most gene signatures, better representing the sensitive, disease-free class (59% accuracy). Interestingly, ensemble averaging of only the gene signatures built using a moderate GI_{50} thresholds yielded results which better represented the resistance class. This result closer matches the accuracy of **Cis1**, and may be due to **Cis1** having a greater overall impact on the ensemble prediction. When limiting ensemble averaging to only those gene signatures with the highest area under the curve (AUC) at each resistance threshold, differences in predictions were negligible. Ensemble ML can potentially avoid problems with poor performance and overfitting by combining gene signatures that individually perform slightly better than chance[39].

It is difficult to reconcile gene signatures without features known to be related to chemoresistance with tumor biology. Our thresholding approach may reveal potentially important genes and pathways associated with platin resistance. It would be preferable to explore pathways related to signature genes to improve accuracy, identify potential targets for further study of chemoresistance, and expand the model parameters to take into account alternate states besides those captured in the original signature[40]. Signatures for resistance may be useful for developing targeted intervention to re-

sensitize tumours. For example, the mismatch repair (MMR) gene *MSH2* is commonly present in gene signatures at high resistance levels for oxaliplatin, which is of interest, as MMR deficiency has been shown to be predictive for oxaliplatin resistance[36]. Indeed, *MLH1*, *MSH2* and *MSH6*-deficient cells are more susceptible to oxaliplatin, despite MMR-deficiency being associated with cisplatin resistance[35]. The autoimmune disease-associated gene *SIAE*, which has been previously shown to have a strong negative correlation to oxaliplatin response in advanced CRC patients[41], was selected in the majority of thresholded oxaliplatin gene signatures (Supplementary Table S2C). The gene *BCL2*, which was commonly selected for cisplatin (Figure 3A), was rarely selected for oxaliplatin (Figure 3C). At the highest levels of resistance to cisplatin, gene signatures were enriched for genes belonging to DNA repair, anti-oxidative response, apoptotic pathways and drug transporters (Figure 3A). These gene pathways are known to be involved in cisplatin resistance[42,43] and these specific genes may be explored in subsequent work to identify the contribution to chemotherapy response in a biochemical context.

Log-loss evaluates the accuracy of a classifier by penalizing erroneous classifications, and is relevant in cases where data is imbalanced and/or have an unequally distributed error cost. We assessed whether ML gene signatures based on log-loss minimization could improve accuracy to predicting patient response (Supplementary Table S3) and compared them to gene signatures generated by minimizing cell line misclassification. When gene signatures generated by both methods were highly similar (generated at the same GI_{50} threshold, consist of a similar number of genes and consist of ≥ 80% shared genes), prediction accuracy of TCGA cancer patient outcomes were nearly indistinguishable, as accuracy can vary over different relapse thresholds. Where significant differences in predictions were seen, the misclassification-based gene signatures were more accurate overall (**Cis1**, **Cis17** and the “12-Resistant”

carboplatin gene signature were +8.3%, +5.6% and +3.9% more accurate compared to the log-loss gene signature, respectively). Oxaliplatin gene signatures were dissimilar across all GI_{50} thresholds, as the log-loss minimized ML gene signatures often contain increased numbers of genes compared to the misclassification-based gene signatures. Many of these larger gene signatures were less accurate in patients compared to gene signatures which minimized misclassification rates consistent that this evaluation and model selection method is more prone to overfitting. This pattern was also noted for gene signatures generated at extreme GI_{50} thresholds for all three platin drugs in which response was, by definition, somewhat imbalanced.

It may be feasible to predict responses to combination chemotherapy with the gene signatures described here. Not included in the present analysis were signatures for methotrexate, vinblastine, and doxorubicin, which comprise the MVAC cocktail used to treat bladder cancer. This was due primarily to a lack of patients treated with this drug combination in the TCGA bladder dataset (N=11). Individual signatures for several of these drugs have been derived and analyzed using the patient data from METABRIC (Molecular Taxonomy of Breast Cancer International Consortium)[24]. A reasonable approach to predicting combination chemotherapy would first determine the probability of sensitivity or resistance to individual drugs, accounting for the misclassification rate by each (defined as d_1, …, d_k). The ML classifiers output these probabilities, analogous to their misclassification rates in a set of patients treated identically. If the model predicts that the patient is sensitive to drug d_1 with 90% probability, and sensitive to drug d_2 with 5% probability, and the errors are independent, then the probability of sensitivity to the combination is 1 - (1 - 0.9)*(1 - 0.05) = 90.5%, and the probability of resistance is 9.5%, assuming no synergistic effects between drugs. If interaction or dependence among errors is suspected, the combined probability of resistance to the pair d_{12} could be estimated based on the features that are shared by the signatures of both drugs. The

probability of sensitivity to the combination would then be given by $1 - (1 - d_{12})*(1 - d_3)*\ldots*(1 - d_k)$.

The predictive accuracy for the same gene signature could differentiate highly between the two datasets. **Cis3** (Supplemental Table S6A) had an AUC of 0.64 when validated against TCGA bladder cancer patients. However, the AUC was lower when applied to the Als *et al.*[29] dataset (AUC=0.18). Patient metadata in the latter study only indicated patient survival times, while we base the expected TCGA patient outcome on time to disease recurrence. As the basis of our expected outcome differs between datasets, these differences may be acting as a confounding factor to determine accuracy of gene signatures. The datasets also differ in how expression was measured (microarray vs. RNA-seq). The relevance of gene signatures based on training and testing data from different platforms can affect the accuracy of validation, which might not be improved by data normalization. In this study, datasets were subjected to z-score normalization. In subsequent studies, other techniques to correct for some of these effects have been described and could be applied[44].

In summary, we describe GI_{50}- or IC_{50}-threshold-independent ML gene signatures to predict chemotherapy response to platin agents in cancer patients. Ensemble machine learning produced combined signatures that were more accurate than most individual gene signatures generated with different thresholds. Genes associated cisplatin response included those which exacerbate resistance in patients with a history of smoking. The methodology described here should be adaptable to other drugs and cancer types. With a range of gene signatures for multiple drugs, it may be possible to improve the efficacy of treatment by tailoring treatment to a patient's specific tumour biology, and reduce treatment duration by limiting the number of different therapeutic regimens prescribed before achieving a successful response[45].

MATERIALS AND METHODS

Data and preprocessing

Cell-line Data Sets

Microarray GE and data were from breast cancer cell lines were used to train ML-based gene signatures of drug response based on respective growth or target inhibition data (GI_{50} or IC_{50}). Cell lines were treated with either cisplatin (N=39), carboplatin (N=46), or oxaliplatin (N=47)[13]. Bladder cancer cell line GE and IC_{50} measurements for cisplatin were obtained from cancerRxgene (N=17). However, all models (gene signatures) used to evaluate patient data were trained on breast cancer cell line data, because the number of bladder cancer cell lines was insufficient to produce accurate signatures.

Cancer Patient Data Sets

RNA-seq GE and survival measurements were downloaded from TCGA for bladder urothelial carcinoma (N=72 patients treated with cisplatin)[26], ovarian epithelial tumor (N=410 treated with carboplatin)[27] and colorectal adenocarcinoma (N=99 treated with oxaliplatin)[28]. GE of cisplatin-treated patients of cell carcinoma of the urothelium (N=30)[29] and for oxaliplatin-treated CRC patients (N=83)[30] were obtained from the Gene Expression Omnibus. Clinical metadata and GE for TCGA patients were obtained from Genomic Data Commons (https://gdc.cancer.gov/), while methylation HM450 (Illumina) data for these patients was downloaded from cBioPortal[46].

Development and Pre-Processing of Biochemically-Inspired Gene Sets

Initial gene sets for developing signatures for each drug were identified from previously published literature (see Supplemental References, Section B) and

595 databases, such as PharmGKB and DrugBank[47,48]. The evidence supporting each gene
596 contained in the final signatures is independent scientific evidence that the genes
597 selected are not the result of spurious associations. The final gene sets were chosen
598 using MFA with the breast cancer cell-line data to analyze interactions between GE, CN,
599 and GI_{50} data for the drug of interest[49]. Genes whose GE and/or CN showed a direct or
600 inverse correlation with GI_{50} were selected for SVM training. Because the number of
601 genes related to GI_{50} for oxaliplatin exceeded the number of cell lines available for
602 training, we limited the input to the oxaliplatin ML model to those genes whose GE were
603 related to GI_{50}. Similarly, the number correlating genes in cisplatin treated cells
604 exceeded the number of cell lines. For cisplatin, genes whose expression correlated with
605 GI_{50} were eliminated if they showed no or little expression in TCGA bladder cancer
606 patients (i.e. RNA-seq counts by Expectation Maximization [RSEM] were < 5.0 for
607 majority of individuals). This reduces the overall number of genes for SVM analysis, and
608 thus helps to avoid a data to size sample imbalance. For cisplatin, MFA was repeated
609 using IC_{50} values for 17 bladder cancer cell lines; however, the available CN data
610 generally showed a lack of variation in the cell lines for these genes. Instead, the
611 available IC_{50} values for three other cancer drugs (doxorubicin, methotrexate and
612 vinblastine) were compared with the IC_{50} of cisplatin by MFA.

613 Applying an SVM model directly to patient data without a normalization approach
614 is imprecise when training and testing data are not obtained using similar methodology
615 (i.e. different microarray platforms). To compare the cell line GE microarray data and the
616 patient RNA-seq GE datasets, expression values were normalized by conversion to z-
617 scores using MATLAB[50]. Although Log2 intensity values from microarray data were not
618 available for TCGA samples, RNA-seq based GE and $\log_2$ intensities from microarray
619 data are highly correlated[51].

Machine Learning

SVMs were trained with breast cancer cell line GE datasets[13] with the Statistics Toolbox in MATLAB[50] similar to Dorman et al (2016)[6]. Rather than a linear kernel, we used a Gaussian kernel function (fitcsvm), and then tested with leave-one-out cross-validation (using the options '*crossval'* and '*leaveout'*). A greedy backwards feature selection algorithm was used to improve classification accuracy[52]. BFS leaves out individual genes from the initial MFA-qualified gene set, then trains a cross validated Gaussian kernel SVM on the training samples, removing the gene with the highest misclassification rate. The procedure is repeated until all genes have been evaluated. The gene subset with the lowest misclassification rate[6] or log-loss statistic[25] based on cross-validation is selected as the gene signature for subsequent testing with patient GE and clinical data. K-fold cross validation of the misclassification-based gene signatures was performed using MATLAB software described in Zhao et al. (2018)[25].

SVMs minimized according to the log-loss classification function were also generated with both software described in Zhao *et al.* (2018; uses multiclass compatible 'fitcecoc' function)[25], and with a modified version of the software described above (using 'fitSVMPosterior' to compute posterior probabilities). Computed probabilities differ between 'fitSVMPosterior' and 'fitcecoc' (range: 0.02-0.04), thus the resultant gene signatures will differ between the two programs. When given unbalanced data (e.g. lower resistance thresholds), 'fitSVMPosterior' will warn that some classes are not represented, and thus those folds will not predict the labels for those missing classes. The log-loss gene signatures described in this manuscript were generated with the multiclass compatible 'fitcecoc' function software[25].

Derivation of gene signatures for different drug resistance thresholds

644 We have previously set a conventional GI_{50} threshold distinguishing sensitivity
645 from resistance at the *median* of the range of drug concentrations that inhibited cell
646 growth by 50%[6]. We hypothesized that different gene signatures could be derived for
647 different levels of drug resistance by varying this threshold. ML experiments for
648 classifying resistance or sensitivity at GI_{50} values generated a series of optimized
649 Gaussian SVM gene signatures whose performance were assessed with patient
650 expression data for each signature. A heat map which illustrates the frequencies of
651 genes appearing in these gene signatures was created with the R language *hist2d*
652 function.

653 A composite gene signature was created by ensemble averaging of all gene
654 signatures generated at each resistance threshold. Ensemble averaging combines
655 signatures through averaging the weighted accuracy of a set of related models[39]. The
656 decision function for the ensemble classifier is the mean of the decision function scores
657 of the component classifiers, weighted by the AUC.

658 *Significance of cell line-derived gene signatures*

659 The significance of the derived SVMs (whether the observed performance of the
660 gene signatures could have arisen by chance) was first assessed by permutation
661 analysis with randomized cell line labels and with random sets of genes, as described
662 previously[6]. Using the median cisplatin GI_{50} as the resistance threshold, 10,000 gene
663 signatures based on random gene selection (15 genes) had higher rates of
664 misclassification than the best median SVM gene signatures (2 signatures with 7.7%
665 misclassification). Cisplatin, carboplatin and oxaliplatin GE data for random cell line label
666 combinations (n=10,000) generated only 8, 1 and 1 signatures, respectively, with lower
667 error rates than the best biochemically-inspired signatures. When minimizing for log-loss

(rather than misclassification), random gene analysis (10,000 iterations; median cisplatin GI_{50} threshold) resulted only in gene signatures with a higher log-loss than the signature generated with the initial cisplatin gene set. Log-loss based random label analysis (n=2000 combinations) resulted in 3.4% of random label gene signatures resulted in a lower log-loss than the cisplatin signature at the same GI_{50} threshold (5.27).This was not entirely surprising, since this result depends on the GI_{50} threshold used for labeling. The differences between GI_{50} values for cell lines close to the median GI_{50} used in this analysis are almost negligible (e.g. 5.11 vs 5.12) and likely within the measurement error for these values.

Regarding the specificity of the cisplatin gene signatures, the best performing cisplatin gene signatures (**Cis1** and **Cis2**) were used to evaluate participants who were treated with other drugs (using an 18 months post-treatment threshold). In such patients, 36.5% of those who were disease-free were predicted accurately with the **Cis1** signature (N=178; 22% less accurate than platin-treated patients), and 62.9% accurate for those with recurrent disease (N=70; 8.1% less accurate). **Cis2** was 43.8% accurate for disease-free non-platin treated patients (N=178; 12.3% lower accuracy), and 60.0% of those who relapsed (N=70; 2.9% less accurate). GE changes in patients treated with platin drugs are better modeled by cancer cell-line based predictors than in patients receiving other drug treatments.

ACKNOWLEDGEMENTS

Katherina Baranova contributed to the cisplatin gene signatures and Dimo Angelov developed automated feature selection. We thank Murray Junop for commenting on the manuscript. Compute Canada and Shared Hierarchical Academic Research Computing Network (SHARCNET) provided high performance computing and storage facilities.

CONFLICTS OF INTEREST

PKR cofounded CytoGnomix Inc., which hosts the interactive resource described in this study for prediction of responses to chemotherapy agents. The other authors have no conflicts of interest.

AUTHOR CONTRIBUTIONS

PKR and DL designed the methodology. EJM and JZ performed analyses. EJM and PKR wrote the manuscript.

FUNDING

PKR is supported by NSERC (RGPIN-2015-06290), Canadian Foundation for Innovation, Canada Research Chairs, and CytoGnomix.

Supplementary information accompanies the manuscript on the *Signal Transduction and Targeted Therapy* website http://www.nature.com/sigtrans.

REFERENCES

1. Cardoso, F. *et al.* Locally recurrent or metastatic breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **23,** vii11–vii19 (2012).
2. Oostendorp, L. J., Stalmeier, P. F., Donders, A. R. T., van der Graaf, W. T. & Ottevanger, P. B. Efficacy and safety of palliative chemotherapy for patients with advanced breast cancer pretreated with anthracyclines and taxanes: a systematic review. *Lancet Oncol.* **12,** 1053–1061 (2011).

3. Alfarouk, K. O. *et al.* Resistance to cancer chemotherapy: failure in drug response from ADME to P-gp. *Cancer Cell Int.* **15,** 71 (2015).
4. Gąsowska-Bodnar, A. *et al.* Survivin Expression as a Prognostic Factor in Patients With Epithelial Ovarian Cancer or Primary Peritoneal Cancer Treated With Neoadjuvant Chemotherapy: *Int. J. Gynecol. Cancer* **24,** 687–696 (2014).
5. Hatzis, C. *et al.* A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* **305,** 1873–1881 (2011).
6. Dorman, S. N. *et al.* Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning. *Mol. Oncol.* **10,** 85–100 (2016).
7. Zhang, S. *et al.* Organic Cation Transporters Are Determinants of Oxaliplatin Cytotoxicity. *Cancer Res.* **66,** 8847–8857 (2006).
8. Poisson, L. M. *et al.* A metabolomic approach to identifying platinum resistance in ovarian cancer. *J. Ovarian Res.* **8,** (2015).
9. Cadoná, F. C. *et al.* Guaraná a Caffeine-Rich Food Increases Oxaliplatin Sensitivity of Colorectal HT-29 Cells by Apoptosis Pathway Modulation. *Anticancer Agents Med. Chem.* **16,** 1055–1065 (2016).
10. Kasparkova, J., Vojtiskova, M., Natile, G. & Brabec, V. Unique Properties of DNA Interstrand Cross-Links of Antitumor Oxaliplatin and the Effect of Chirality of the Carrier Ligand. *Chem. – Eur. J.* **14,** 1330–1341 (2008).
11. Woynarowski, J. M. *et al.* Oxaliplatin-Induced Damage of Cellular DNA. *Mol. Pharmacol.* **58,** 920–927 (2000).
12. Tashiro, T., Kawada, Y., Sakurai, Y. & Kidani, Y. Antitumor activity of a new platinum complex, oxalato (trans-l-1,2-diaminocyclohexane)platinum (II): new experimental data. *Biomed. Pharmacother.* **43,** 251–260 (1989).

13. Daemen, A. *et al.* Modeling precision treatment of breast cancer. *Genome Biol.* **14,** R110 (2013).
14. Yuan, Y. *et al.* Identification of the biomarkers for the prediction of efficacy in first-line chemotherapy of metastatic colorectal cancer patients using SELDI-TOF-MS and artificial neural networks. *Hepatogastroenterology.* **59,** 2461–2465 (2012).
15. L'Espérance, S., Bachvarova, M., Tetu, B., Mes-Masson, A.-M. & Bachvarov, D. Global gene expression analysis of early response to chemotherapy treatment in ovarian cancer spheroids. *BMC Genomics* **9,** 99 (2008).
16. Nickerson, M. L. *et al.* Molecular analysis of urothelial cancer cell lines for modeling tumor biology and drug response. *Oncogene* (2016). doi:10.1038/onc.2016.172
17. Yuryev, A. Gene expression profiling for targeted cancer treatment. *Expert Opin. Drug Discov.* **10,** 91–99 (2015).
18. Sataloff, D. M. *et al.* Pathologic response to induction chemotherapy in locally advanced carcinoma of the breast: a determinant of outcome. *J. Am. Coll. Surg.* **180,** 297–306 (1995).
19. Ogston, K. N. *et al.* A new histological grading system to assess response of breast cancers to primary chemotherapy: prognostic significance and survival. *Breast Edinb. Scotl.* **12,** 320–327 (2003).
20. Earl, J. *et al.* The UBC-40 Urothelial Bladder Cancer cell line index: a genomic resource for functional studies. *BMC Genomics* **16,** 403 (2015).
21. Rixe, O. *et al.* Oxaliplatin, tetraplatin, cisplatin, and carboplatin: Spectrum of activity in drug-resistant cell lines and in the cell lines of the national cancer institute's anticancer drug screen panel. *Biochem. Pharmacol.* **52,** 1855–1865 (1996).
22. Mehmood, R. K. Review of Cisplatin and oxaliplatin in current immunogenic and monoclonal antibody treatments. *Oncol. Rev.* **8,** 256 (2014).

760 23. Kweekel, D. M., Gelderblom, H. & Guchelaar, H.-J. Pharmacology of oxaliplatin and the use
761 of pharmacogenomics to individualize therapy. *Cancer Treat. Rev.* **31,** 90–105 (2005).
762 24. Mucaki, E. J. *et al.* Predicting Outcomes of Hormone and Chemotherapy in the Molecular
763 Taxonomy of Breast Cancer International Consortium (METABRIC) Study by Biochemically-
764 inspired Machine Learning. *F1000Research* **5,** 2124 (2017).
765 25. Zhao, J. Z. L., Mucaki, E. J. & Rogan, P. K. Predicting ionizing radiation exposure using
766 biochemically-inspired genomic machine learning. *F1000Research* **7,** 233 (2018).
767 26. Tembe, V. *et al.* The BARD1 BRCT domain contributes to p53 binding, cytoplasmic and
768 mitochondrial localization, and apoptotic function. *Cell. Signal.* **27,** 1763–1771 (2015).
769 27. Robertson, A. G. *et al.* Comprehensive Molecular Characterization of Muscle-Invasive
770 Bladder Cancer. *Cell* **171,** 540-556.e25 (2017).
771 28. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian
772 carcinoma. *Nature* **474,** 609–615 (2011).
773 29. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon
774 and rectal cancer. *Nature* **487,** 330–337 (2012).
775 30. Als, A. B. *et al.* Emmprin and survivin predict response and survival following cisplatin-
776 containing chemotherapy in patients with advanced bladder cancer. *Clin. Cancer Res. Off. J.*
777 *Am. Assoc. Cancer Res.* **13,** 4407–4414 (2007).
778 31. Tsuji, S. *et al.* Potential responders to FOLFOX therapy for colorectal cancer by Random
779 Forests analysis. *Br. J. Cancer* **106,** 126–132 (2012).
780 32. Freedman, N. D., Silverman, D. T., Hollenbeck, A. R., Schatzkin, A. & Abnet, C. C. Association
781 between smoking and risk of bladder cancer among men and women. *JAMA* **306,** 737–745
782 (2011).

33. Fortin, A., Wang, C. S. & Vigneault, E. Influence of smoking and alcohol drinking behaviors on treatment outcomes of patients with squamous cell carcinomas of the head and neck. *Int. J. Radiat. Oncol. Biol. Phys.* **74,** 1062–1069 (2009).
34. Joehanes, R. *et al.* Epigenetic Signatures of Cigarette Smoking. *Circ. Cardiovasc. Genet.* (2016). doi:10.1161/CIRCGENETICS.116.001506
35. Raymond, E., Faivre, S., Chaney, S., Woynarowski, J. & Cvitkovic, E. Cellular and Molecular Pharmacology of Oxaliplatin1. *Mol. Cancer Ther.* **1,** 227–235 (2002).
36. Alex, A. K. *et al.* Response to Chemotherapy and Prognosis in Metastatic Colorectal Cancer With DNA Deficient Mismatch Repair. *Clin. Colorectal Cancer* (2016). doi:10.1016/j.clcc.2016.11.001
37. Sos, M. L. *et al.* Predicting drug susceptibility of non-small cell lung cancers based on genetic lesions. *J. Clin. Invest.* **119,** 1727–1740 (2009).
38. Laderas, T. G., Heiser, L. M. & Sönmez, K. A Network-Based Model of Oncogenic Collaboration for Prediction of Drug Sensitivity. *Front. Genet.* **6,** (2015).
39. Clemen, R. T. Combining forecasts: A review and annotated bibliography. *Int. J. Forecast.* **5,** 559–583 (1989).
40. Airley, R. *Cancer chemotherapy*. (Wiley-Blackwell, 2009).
41. Li, X.-X. *et al.* RNA-seq identifies determinants of oxaliplatin sensitivity in colorectal cancer cell lines. *Int. J. Clin. Exp. Pathol.* **7,** 3763–3770 (2014).
42. Borst, P., Rottenberg, S. & Jonkers, J. How do real tumors become resistant to cisplatin? *Cell Cycle Georget. Tex* **7,** 1353–1359 (2008).
43. Wernyj, R. & Morin, P. Molecular mechanisms of platinum resistance: still searching for the Achilles heel. *Drug Resist. Updat.* **7,** 227–232 (2004).

44. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostat. Oxf. Engl.* **8,** 118–127 (2007).
45. Akamatsu, N., Nakajima, H., Ono, M. & Miura, Y. Increase in acetyl CoA synthetase activity after phenobarbital treatment. *Biochem. Pharmacol.* **24,** 1725–1727 (1975).
46. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6,** pl1 (2013).
47. Whirl-Carrillo, M. *et al.* Pharmacogenomics Knowledge for Personalized Medicine. *Clin. Pharmacol. Ther.* **92,** 414–417 (2012).
48. Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42,** D1091–D1097 (2014).
49. Abdi, H. & Williams, L. J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2,** 433–459 (2010).
50. MATLAB and Statistics Toolbox Release 2012b, The MathWorks, Inc., Natick, Massachusetts, United States.
51. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18,** 1509–1517 (2008).
52. Bermingham, M. L. *et al.* Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Sci. Rep.* **5,** 10312 (2015).

Tables

Table 1: Gene Signatures Which Best Predicted Response in TCGA Cancer Patients

Gene Signature ID	Cancer Type Tested	GI50 Threshold	Signature (C, σ*)
Cis1 (Cisplatin)	Bladder	5.11	*BARD1, BCL2, BCL2L1, CDKN2C, FAAP24, FEN1, MAP3K1, MAPK13, MAPK3, NFKB1, NFKB2, SLC22A5, SLC31A2, TLR4, TWIST1* (100000, 100)
Cis2 (Cisplatin)	Bladder	5.12	*BARD1, BCL2L1, CFLAR, FAAP24, MAP3K1, MAPK3, NFKB1, POLQ, PRKAA2, SLC22A5, SLC31A2, TLR4, TWIST1* (10000, 100)
Cis3 (Cisplatin)	Bladder	5.60	*BCL2, CFLAR, ERCC2, ERCC6, FAAP24, FEN1, MAP3K1, NFKB1, NFKB2, PNKP, POLQ, PRKCB, SLC22A5, SNAI1, TLR4* (100000, 100)
Cis12 (Cisplatin)	Bladder	5.40	*ATP7B, BCL2, BCL2L1, CDKN2C, ERCC2, FAAP24, GSTO1, MAP3K1, MAPK3, MT2A, NFKB1, NFKB2, POLD1, POLQ, PRKCB, SNAI1, TLR4, TP63 (*10000, 100)
Cis14 (Cisplatin)	Bladder	5.16	*BARD1, BCL2, BCL2L1, CDKN2C, FAAP24, FEN1, FOS, GSTP1, MAP3K1, MAPK13, MAPK3, MSH2, NFKB1, POLD1, POLQ, PRKAA2, PRKCB, SLC22A5, SLC31A2, SNAI1, TWIST1* (10000, 100)
Cis17 (Cisplatin)	Bladder	5.10	*ATP7B, BCL2, BCL2L1, FEN1, GSTP1, MAP3K1, MAPK3, MT2A, NFKB1, PNKP, POLQ, PRKAA2, PRKCB, SLC31A2, TLR4, TP63* (100000, 100)
Car1 (Carboplatin)	Ovarian	4.22	*AKT1*, *EIF3K*, *ERCC1*, *GNGT1*, *GSR*, *MTHFR*, *NEDD4L*, *NLRP1*, *NRAS*, *RAF1*, *SGK1*, *TIGD1*, *TP53*, *VEGFB*, *VEGFC* (100000, 100)
Car9 (Carboplatin)	Ovarian	4.32	*AKT1, ATP7B, EIF3I, ETS2, GNGT1, HRAS, KRAS, LIG3, MTHFR, MTR, NRAS, RAD50, SCN10A, TIGD1, TP53, VEGFB* (10000, 100)
Car51 (Carboplatin)	Ovarian	4.34	*AKT1, EGF, EIF3I, ERCC1, ETS2, GNGT1, KRAS, MTHFR, MTR, NEDD4L, NLRP1, NRAS, RAD50, RAF1, SGK1, TIGD1, TP53, VEGFB, VEGFC* (10000, 100)
Car73 (Carboplatin)	Ovarian	4.09	*AKT1, ATP7B, ETS2, GNGT1, HRAS, NLRP1, SCN10A, VEGFB* (100000, 1000)
Oxa1 (Oxaliplatin)	Colorectal	5.10	*BRAF*, *FCGR2A*, *IGF1*, *MSH2*, *NAGK*, *NFE2L2*, *NQO1*, *PANK3*, *SLC47A1*, *SLCO1B1*, *UGT1A1* (10, 10)
Oxa21 (Oxaliplatin)	Colorectal	5.10	*BRAF, IGF1, IGF1R, KLF3, MSH2, NAT2, NFE2L2, NQO1, PANK3, PRSS1, SIAE, SLC47A1, SLCO1B1, UGT1A1* (1000, 100)

*C - The box-constraint. σ – the kernel-scale ("sigma"). Bolded gene signatures are those that best overall performance against TCGA cancer patient gene expression data.

FIGURE LEGENDS

Figure 1. Schematic of platinum drug sensitivity and resistance genes which showed MFA correlation for GI_{50} of A) cisplatin, B) carboplatin, and C) oxaliplatin. The genes used to derive the SVM are shown in context of their effect in the cell and role in cisplatin mechanisms of action. GE and CN correlation with inhibitory drug concentration by MFA of breast (GI_{50}) and bladder (IC_{50}) cancer cell line data.

Figure 2: GI_{50} values for cell lines treated with the three platin drugs were plotted in order of ascending oxaliplatin GI_{50}. For most cell lines, there is a visible trend between the GI_{50} for cisplatin and carboplatin, reflecting the correlation between the two drugs seen by MFA. Despite this correlation, carboplatin shows a much smaller variance (0.22) compared to cisplatin (0.37; oxaliplatin variance is 0.34).

Figure 3. The variation in gene composition of misclassification-based SVMs at different GI_{50} thresholds for A) cisplatin, B) carboplatin, and C) oxaliplatin. GI_{50} intervals are indicated on the left, with the number of cell lines with GI_{50} values within said intervals in brackets. Each box represents the density of genes appearing in optimized Gaussian SVM gene signatures in those functional categories, with darker grey indicating frequent genes in indicated GI_{50} threshold intervals, while lighter grey indicates less commonly selected genes. The number of thresholded gene signatures used to derive the density plot within each interval is equal (or greater, in the case of multiple equally performing gene signatures) to the number of cell lines within that GI_{50} interval.

Supplementary Figure 1. The variation in gene composition of log-loss based SVMs at different GI_{50} thresholds for A) cisplatin, B) carboplatin, and C) oxaliplatin. Each box represents the density of genes appearing in optimized Gaussian log-loss SVM gene

signatures in those functional categories, with darker grey indicating frequent genes in indicated GI_{50} threshold intervals, while lighter grey indicates less commonly selected genes.

Supplementary Figure 2. Classification accuracy of gene signatures on TCGA bladder cancer patients treated with cisplatin and/or carboplatin as the resistance threshold is varied. Recurrence and disease-free survival are used as a binary measure to assess performance. The x-axis indicates movement of the resistance threshold, with more cell lines labeled sensitive on the left and more labeled resistant on the right. Maximal AUC is indicated by the downward arrows.

Supplementary Figure 3. Classification accuracy of SVM gene signatures for cisplatin, at a range of response thresholds, were assessed using gene expression data for cisplatin-treated bladder cancer patients from Als *et al.* [29]. Patients with a ≥ 5 year survival post-treatment were labeled sensitive. Red arrows indicate the SVM gene signatures with the highest positive predictive value (PPV) in the accuracy of classification of patient outcome.

Supplementary Figure 4. Hyperplane distance calculated by all thresholded SVMs for recurrent (<6 months) TCGA patients. Each diagram represents the predictions of all SVMs for all patients who had recurrence less than 6 months after treatment (N=10). Each point represents an SVM, where the x-axis represents the number of cell lines set to resistant (in order of lowest to highest GI_{50}), and the y-axis represents the calculated hyperplane distance. A negative hyperplane distance would represent a prediction of resistance to cisplatin. Despite this, some patients show a strong preference towards predictions of sensitivity (i.e. TCGA-XF-A9SU).

Supplementary Figure 5. Hyperplane distance calculated by all thresholded SVMs for sensitive TCGA patients. Each diagram represents the predictions of all SVMs for all patients who had recurrence > 4 years after treatment (top; N=3), or patients who showed no recurrence after 6 years (bottom; N=6). Each point represents an SVM, where the x-axis represents the number of cell lines set to resistant (in order of lowest to highest GI_{50}), and the y-axis represents the calculated hyperplane distance. A positive hyperplane distance would represent a prediction of sensitivity to cisplatin.

Supplementary References. A) Experimental evidence supporting inclusion of genes using references relating expression to platinum drug efficacy. A subset of genes are shown with consistent significant increase in misclassification error. B) The initial peer-reviewed literature used to develop gene signatures associated with cis-, carbo- and oxaliplatin response.

Supplementary Tables. Details of gene signatures, validation, and accuracy are indicated in Tables: S1A) Genes Selected for MFA of Expression/Copy Number to Cisplatin GI50; S1B) Genes Selected for MFA of Expression/Copy Number to Carboplatin GI50; S1C) Genes Selected for MFA of Expression/Copy Number to Oxaliplatin GI50; S2A) SVM Models by Varying Resistance Thresholds and Impact on Misclassification, Categorized by Gene Function for Cisplatin; S2B) SVM Models by Varying Resistance Thresholds and Impact on Misclassification, Categorized by Gene Function for Carboplatin; S2C) SVM Models by Varying Resistance Thresholds and Impact on Misclassification, Categorized by Gene Function for Oxaliplatin; S3A) Cisplatin SVM Models Derived Over a Range of Response Thresholds Using Log-Loss Minimization; S3B) Carboplatin SVM Models Derived Over a Range of Response Thresholds Using Log-Loss Minimization; S3C) Oxaliplatin SVM Models Derived Over a Range of Response Thresholds Using Log-Loss Minimization; S4) SVM Models

Generated by Bladder Cancer Patient Data at Various Time to Recurrence thresholds, Categorized by Gene Function for Cisplatin; S5A) Cisplatin SVM Models Derived Over a Range of Response Thresholds for TCGA Bladder Cancer Patients using Misclassification; S5B) Carboplatin SVM Models Derived Over a Range of Response Thresholds for TCGA ovarian epithelial tumor patients using Misclassification; S5C) Oxaliplatin SVM Models Derived Over a Range of Response Thresholds for TCGA colorectal adenocarcinoma patients using Misclassification; S6A) Accuracy of SVMs for TCGA Bladder Cancer Patients, in relation to Time to Recurrence Post-Treatment With Cisplatin; S6B) Accuracy of SVMs for TCGA Bladder Cancer Patients, in relation to Time to Recurrence Post-Treatment With Carboplatin; S6C) Accuracy of SVMs for TCGA Bladder Cancer Patients, in relation to Time to Recurrence Post-Treatment With Oxaliplatin; S7A) Accuracy of Threshold Independent Analysis (Ensemble Averaging) for Cisplatin; S7B) Accuracy of Threshold Independent Analysis (Ensemble Averaging) for Carboplatin; S7C) Accuracy of Threshold Independent Analysis (Ensemble Averaging) for Oxaliplatin; S8) Accuracy of SVMs for Non-Smoking TCGA Bladder Cancer Patients; S9) Accuracy of SVMs for Smoking TCGA Bladder Cancer Patients (within 15 years of diagnosis).

A.

Cisplatin

SLC31A1 | SLC31A2

SLC22A5
SLC22A7
SLC22A10
SLC22A11
SLC22A12
SLC22A13
SLC22A16
SLC22A20

ATP7A
ATP7B

MRE11A | BARD1
POLQ | NBN
POLD1 | POLN
RAD50 | PNKP
PARP1

FEN1
FAAP24

NFKB1
NFKB2

GSTO1
GSTP1

MT1A | MT2A
MT3 | MT4
metallothionein -Pt

GSX1 | GSX2

Degradation

CFLAR

TP63

TP53

BCL2

BCL2L1

BAX/BAK
BAX/BAK
BAX/BAK

MAP3
MAP1

FOS

apoptosis

PRKCB1
SNAI1
TWIST1

MAPK8
MAP3K1

MAPK11 | MAPK12
MAPK13 | MAPK14

CDKN2C

FOS

c-Abl

Mismatch Repair

Signal transduction

Other Genes associated with cisplatin resistance in the literature:

MAPK3 | PRKCA | TLR4

MFA:	Expression	Copy Number	Both
Higher Levels Correlate with Higher GI50 & IC50			
Lower Levels Correlate with Higher GI50 & IC50			

No MFA Correlation

B.

Carboplatin

SLC31A1

SLC19A1

GSR

GSTT1

Redox. / Detoxification

ABCB1

ATP7A

ATP7B

DNA Repair

ERCC1 | ERCC4

LIG1 | LIG3

RAD50 | XRCC1

DNA

TP53

BCL2

BCL2L1

BAX/BAK

ETS2

NLRP1

MAD2L1

apoptosis

eIF-3

eIF-4E

SGK1

mTOR

MAPK1 | MAPK3

AKT1

RAF1

MAP2K1

PTEN

PI3K

NEDD4L

KRAS | HRAS | NRAS

EGFR | VEGFR | OPRM1

EGF

VEGFB/C

Other Genes associated with carboplatin resistance in the literature:

GNGT | MTR | MTHFR

OPRM1 | SCN10A | TIGD1

C.

Oxaliplatin

SLC31A1

DNA Repair

ERCC2 | ERCC6 | MSH2 | MSH6 | XRCC1 | XRCC3

AGXT | MT1A | MPO | NAT2

GSTs | NQO1 | UGT1A1

Detoxification

NFE2L2

KLF3 | NOTCH1 | PANK3 | PARP15 | SGPP2 | SIAE

Pt

DNA

TP53

BCL2L1

BCL2

BAX

BAK

CSMD1 | LTA | PROC

apoptosis

HIF1A

Angiogenesis

PTGS2 | VEGFA | KISS1

ERK1/2 | MEK1/2 | BRAF | KRAS

IGF1R | IGF1

KIT

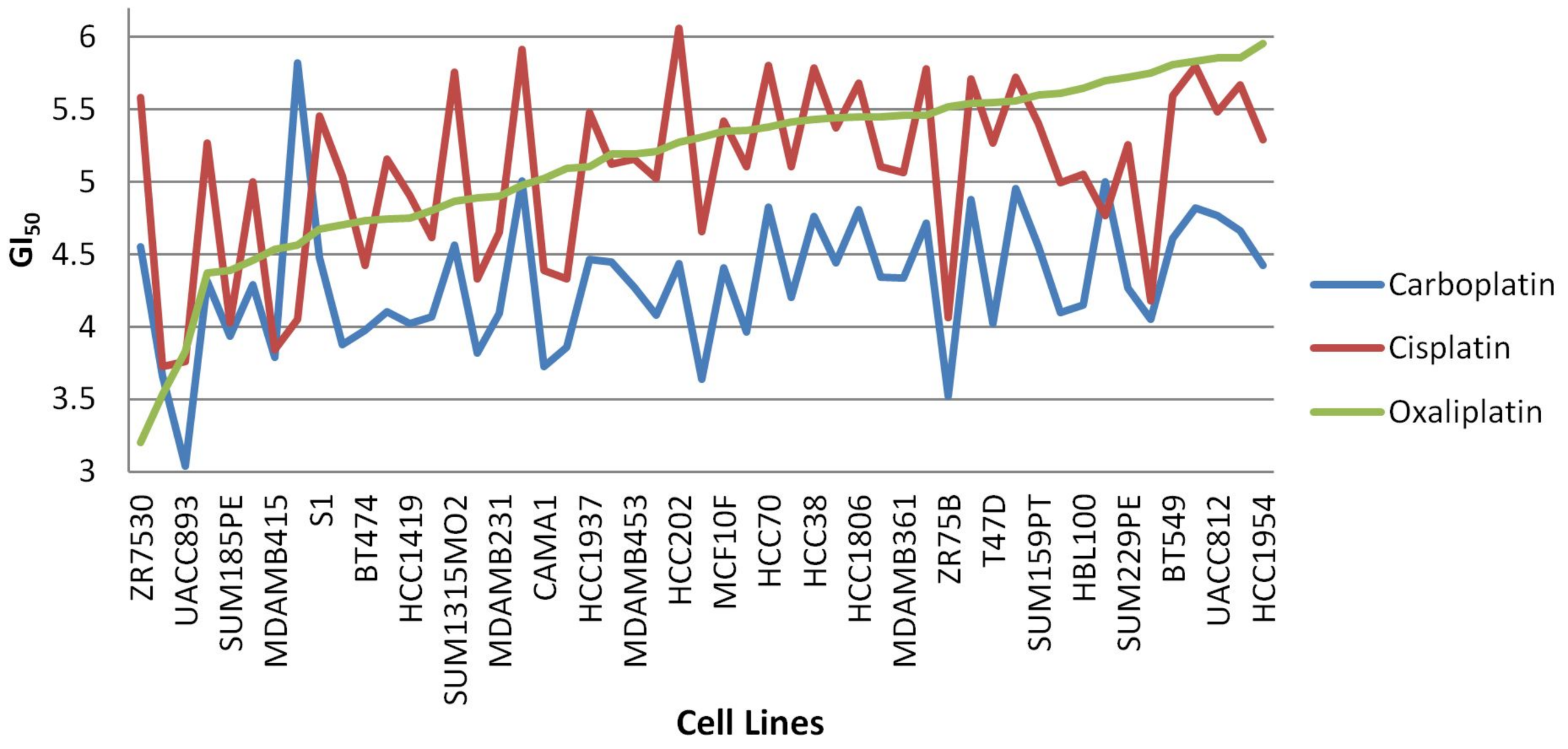
ATP7B | SLC47A1

ABCG2 | ABCC2 | ABCB1

Other Genes associated with oxaliplatin resistance in the literature:

APOBEC2 | CLCN6 | GRTP1 | FCGR2A | ICAM5

HLA-B | KLF3 | NAGK | PRSS1 | SLCO1B1


6
5.5
5
4.5
4
3.5
3
GI50
ZR7530
UACC893
SUM185PE
MDAMB415
S1
BT474
HCC1419
SUM1315MO2
MDAMB231
CAMA1
HCC1937
MDAMB453
HCC202
MCF10F
HCC70
HCC38
HCC1806
MDAMB361
ZR75B
T47D
SUM159PT
HBL100
SUM229PE
BT549
UACC812
HCC1954
Cell Lines
Carboplatin
Cisplatin
Oxaliplatin

A.

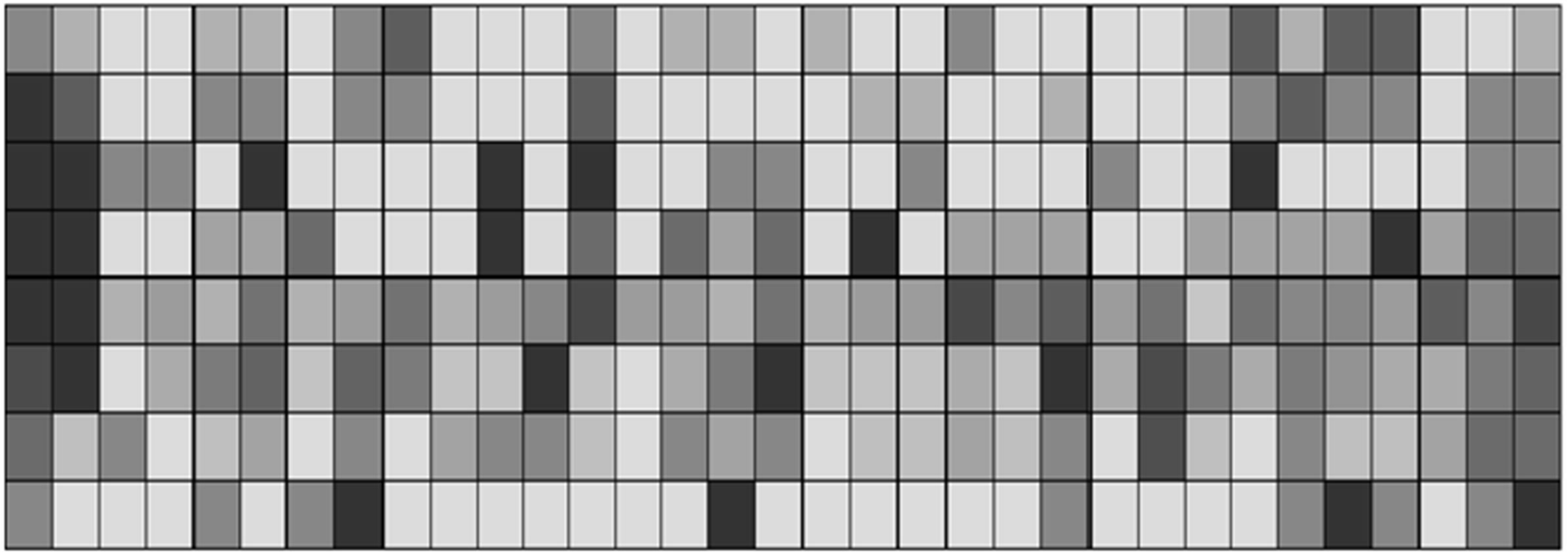
Apoptosis
Cell Signaling
Cell Growth
DNA Repair
Metabolism
Metal Binding
Signal Transduction
Transcription
Membrane Transport

GI_{50} (# cell lines in interval)

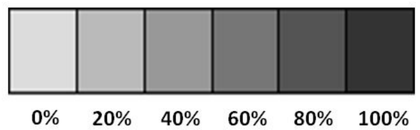
Resistant

4.0–4.25 (4)
4.25-4.5 (4)
4.5-4.75 (2)
4.75-5.0 (3)
5.0-5.25 (8)
5.25 – 5.5 (7)
5.5 – 5.75 (6)
5.75 – 6.0 (2)

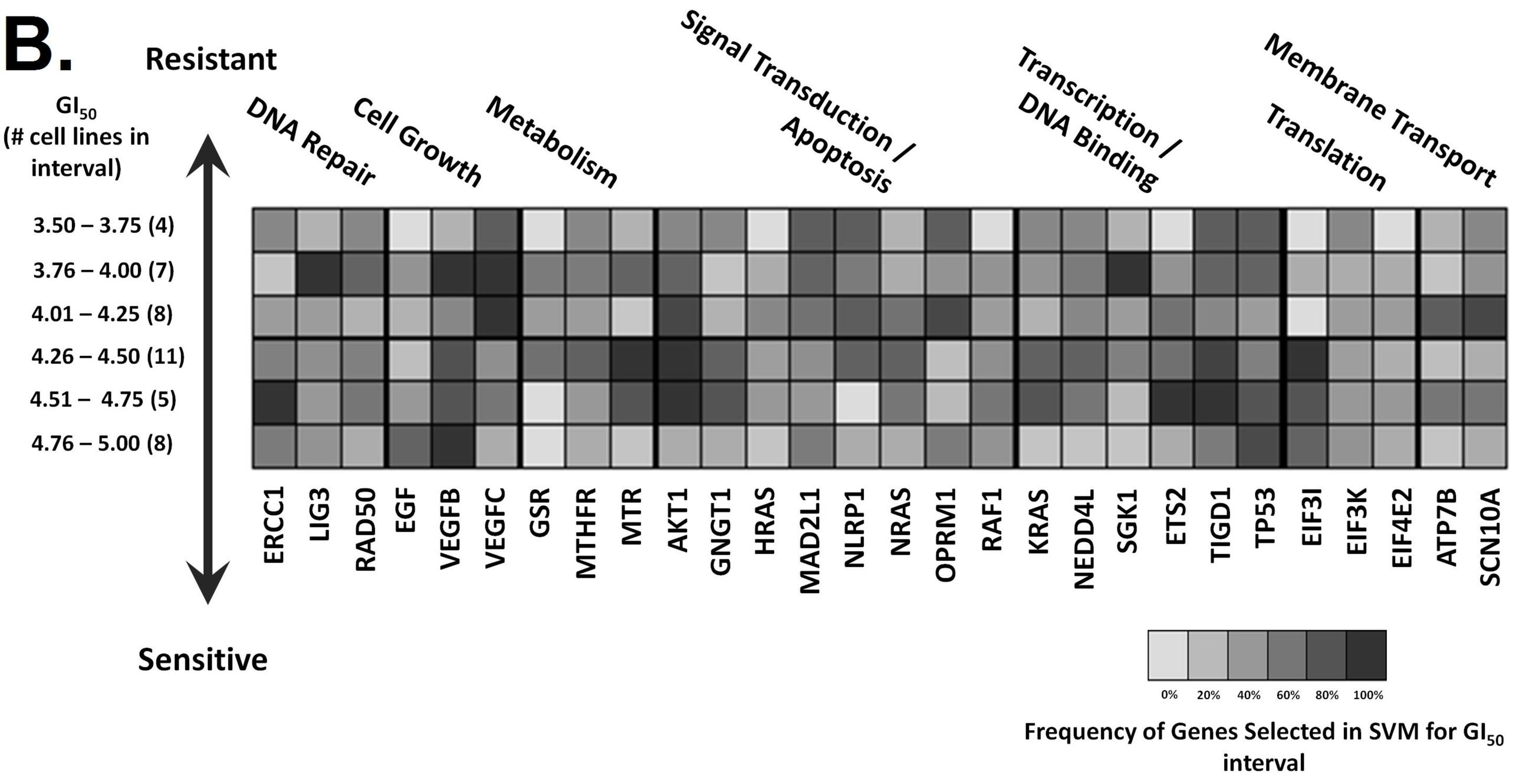
Sensitive

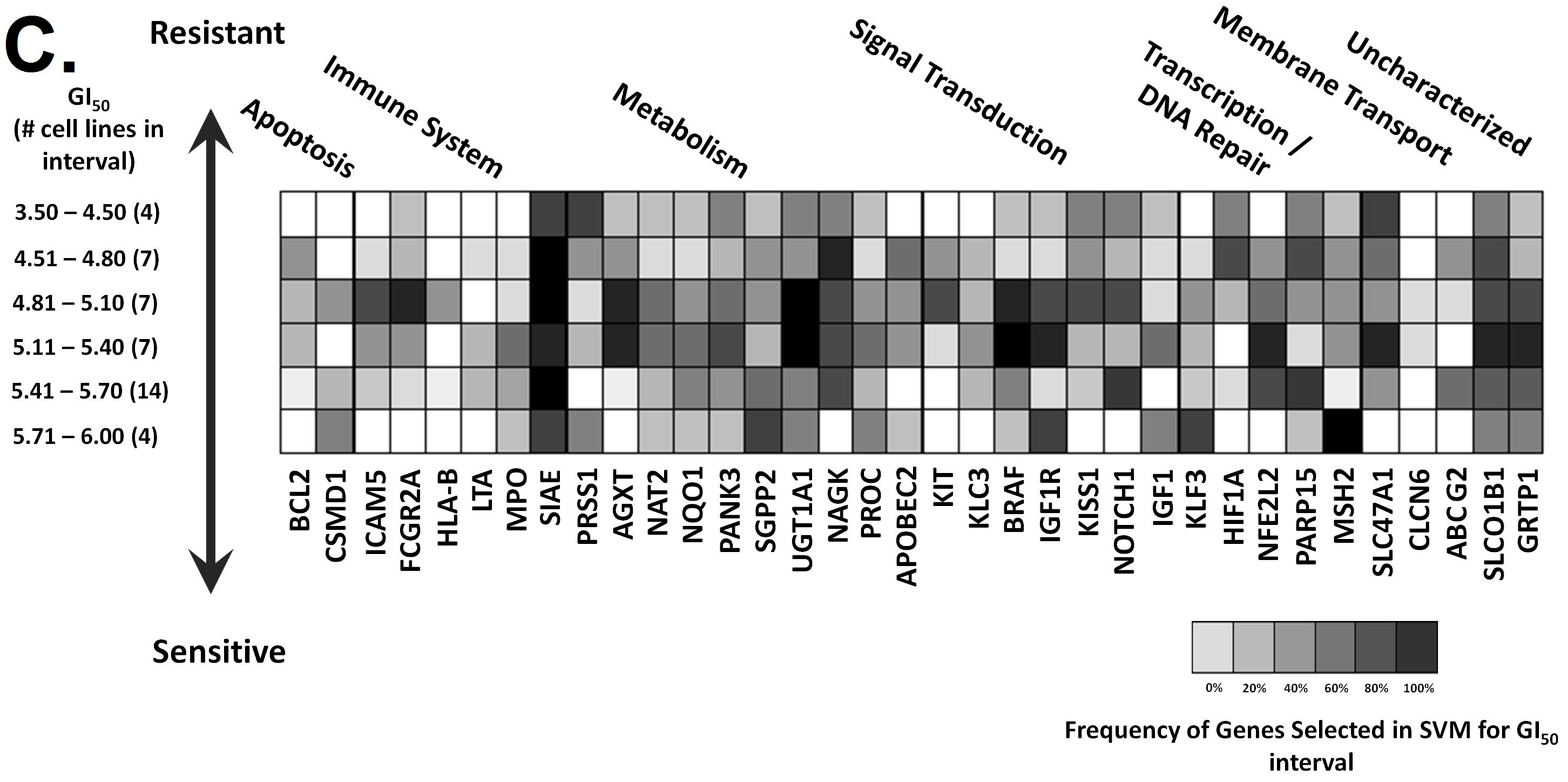


BCL2
BCL2L1
CFLAR
FOS
PRKCA
PRKCB
CDKN2C
SNAI1
BARD1
ERCC2
ERCC6
FAAP24
FEN1
MSH2
PNKP
POLD1
POLQ
GSTO1
GSTP1
MT2A
MAP3K1
MAPK13
MAPK3
FOS
NFKB1
NFKB2
PRKAA2
TLR4
TP63
TWIST1
ATP7B
SLC22A5
SLC31A2



Frequency of Genes Selected in SVM for GI_{50} interval


B.
Resistant
GI50 (# cell lines in interval)
3.50 – 3.75 (4)
3.76 – 4.00 (7)
4.01 – 4.25 (8)
4.26 – 4.50 (11)
4.51 – 4.75 (5)
4.76 – 5.00 (8)
Sensitive
DNA Repair
Cell Growth
Metabolism
Signal Transduction / Apoptosis
Transcription / DNA Binding
Translation
Membrane Transport
ERCC1
LIG3
RAD50
EGF
VEGFB
VEGFC
GSR
MTHFR
MTR
AKT1
GNGT1
HRAS
MAD2L1
NLRP1
NRAS
OPRM1
RAF1
KRAS
NEDD4L
SGK1
ETS2
TIGD1
TP53
EIF3I
EIF3K
EIF4E2
ATP7B
SCN10A
0%
20%
40%
60%
80%
100%
Frequency of Genes Selected in SVM for GI50 interval


C.
Resistant
Sensitive
GI50 (# cell lines in interval)
3.50 – 4.50 (4)
4.51 – 4.80 (7)
4.81 – 5.10 (7)
5.11 – 5.40 (7)
5.41 – 5.70 (14)
5.71 – 6.00 (4)
Apoptosis
Immune System
Metabolism
Signal Transduction
Transcription / DNA Repair
Membrane Transport
Uncharacterized
BCL2
CSMD1
ICAM5
FCGR2A
HLA-B
LTA
MPO
SIAE
PRSS1
AGXT
NAT2
NQO1
PANK3
SGPP2
UGT1A1
NAGK
PROC
APOBEC2
KIT
KLC3
BRAF
IGF1R
KISS1
NOTCH1
IGF1
KLF3
HIF1A
NFE2L2
PARP15
MSH2
SLC47A1
CLCN6
ABCG2
SLCO1B1
GRTP1
0%
20%
40%
60%
80%
100%
Frequency of Genes Selected in SVM for GI50 interval