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

Eliseos J. Mucaki[1], Jonathan Zhao[1], Dan Lizotte[2,3], and Peter K. Rogan[1,2,3,4,5]

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

[§]**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 drugs that accurately predict chemotherapy response could improve cancer outcomes. We derive optimized gene signatures for response to common platinum-based drugs, cisplatin, carboplatin, and oxaliplatin, and respectively validate each with bladder, ovarian, and colon cancer patient data. Initially, using breast cancer cell line gene expression and growth inhibition (GI_{50}) data, we performed backwards feature selection with cross-validation to derive predictive gene sets in a supervised support vector machine (SVM) learning approach. These signatures were also verified in bladder cancer cell lines. Aside from published associations between drugs and genes, we also expanded these gene signatures using a systems biology approach. Signatures at different GI_{50} thresholds distinguishing sensitivity from resistance to each drug contrast the contributions of different genes at extreme vs. median thresholds. An ensemble machine learning technique combining different GI_{50} thresholds was used to create threshold independent gene signatures. The most accurate models for each platinum drug in cell lines consisted of 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;* and oxaliplatin: *BRAF*, *FCGR2A*, *IGF1*, *MSH2*, *NAGK*, *NFE2L2*, *NQO1*, *PANK3*, *SLC47A1*, *SLCO1B1*, *UGT1A1*. Recurrence in bladder urothelial carcinoma patients from the Cancer Genome Atlas (TCGA) treated with cisplatin after 18 months was 71% accurate (59% in disease-free patients). In carboplatin-treated ovarian cancer patients, predicted recurrence was 60.2% (61% disease-free) accurate after 4 years, while the oxaliplatin signature predicted disease-free colorectal cancer patients with 72% accuracy (54.5% for recurrence) after 1 year. The best performing cisplatin model best predicted outcome for non-smoking TCGA bladder cancer patients (100% accuracy for recurrent, 57% for disease-free; N=19), the median GI_{50} model (GI_{50} = 5.12) predicted

outcome in smokers with 79% with recurrence, and 62% who were disease free; N=35). Cisplatin and carboplatin signatures were comprised of overlapping gene sets and GI_{50} values, which contrasted with models for oxaliplatin response.

KEY WORDS. Chemotherapy response, SVM, gene signatures, 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). However, individual tumours vary in or evolve their resistance to these treatments, which has led to tiered sequential strategies for selection of agents, based on their overall efficacy (3). Prediction of response in newly diagnosed tumors from genomic changes would be beneficial for choice of initial treatment, possibly reduce adverse effects, and may improve patient outcomes. We and others have developed gene signatures aimed at predicting response to specific chemotherapeutic agents and minimizing chemoresistance (4–6).

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; [7]). 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 (8). Oxaliplatin can be up to two times as cytotoxic as cisplatin, but it forms fewer DNA adducts (9). 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 (11), organic cation transporters are oxaliplatin-specific and likely play a role in its efficacy in colorectal cancer cells where

these transporters are commonly overexpressed (12). The platinum drugs induce mismatch-repair proteins and DNA damage-recognition proteins (13). Oxaliplatin specifically plays a role in interfering with both DNA and RNA synthesis, unlike cisplatin which only infers with DNA (14). 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 for these drugs to predict these differences and to provide additional insight into these and other gene products, including their relative importance to the cellular response. The signatures are derived by training a machine to classify observations as resistant or sensitive to the drug. Traditionally, supervised learning algorithms have been used, including random forest models (15); support vector machine (SVM) models (16); neural networks (17); and linear regression models (5). Pathway and network analysis of gene expression have been used to indicate hundreds of genes potentially up- and down-regulated upon cisplatin treatment (18). Integrative approaches such as elastic net regression using inferred pathway activity (PARADIGM scores [19]) of bladder cancer cell line data have been applied to derive cisplatin-specific gene signatures (20). These methods implicate genes independent of the known literature. Supervised machine learning with biochemically-relevant genes has been useful for predicting drug response (6). An insufficient number of samples coupled to a large number of features in each sample during training of the classifier often leads to overfitting of the model to the data (21, 22). We reduce the number of dimensions by selecting genes biologically relevant to the drugs under observation (6,22). Additional selection criteria are necessary when the number of genes implicated in peer-reviewed reports is still prohibitively large compared to sample size.

The performance of biochemically-inspired gene signatures to predict patient treatment response has been promising. A paclitaxel machine learning signature based on tumor gene expression had a higher success predicting the pathological complete response rate (pCR; 23]) for sensitive patients (84% of patients with no / minimal residual disease) than models based on differential gene expression analysis (6). For gemcitabine, a signature derived from both expression and copy number data from breast cancer cell lines was derived, and subsequently applied to analysis of nucleic acids from patient archival material. Multiple other outcome measures used to validate gene signatures include prognosis (5), Miller-Payne response (24), and recurrence (25). The assignment of labels for binary SVM classifiers is difficult for continuous measures such as prognosis and recurrence. Binary classification of pCR is simpler to perform using a binary SVM model. Nevertheless, differences in clinical recurrence have been noted between patients demonstrated with pCR and those who do not exhibit disease pathology (23). This source of variability in defining patient response can confound transferability of SVM models between different datasets.

We apply biochemically-inspired machine learning to predict and compare the cellular and patient responses to cisplatin, carboplatin and oxaliplatin. Novel techniques are used to hone in on genes within the large biological feature space, based on systems biological relationships aimed at finding dependent genes that contribute to drug response. These gene signatures for each of these drugs illustrate similarities as well as distinct genomic differences resistance across various cancer types. Classifiers independent of arbitrary-defined class thresholds learn gene signatures, and assess potential overfitting. We use cancer cell line data to train models for classification of platin resistance in cancer, evaluating the models against patient subclasses, such as smoking history.

MATERIALS AND METHODS

Data and preprocessing

Microarray gene expression (GE) and growth inhibition (GI_{50}) measurements were available for breast cancer cell lines (N=39 treated with cisplatin, 46 with carboplatin, and 47 with oxaliplatin) from (15). Bladder cancer cell line GE and IC_{50} measurements were available for cisplatin from Genomics of Drug Sensitivity in Cancer project (http://www.cancerrxgene.org; N=17). However, all training was performed on breast cancer cell line data as the training set of bladder cancer cell lines (N=17) was prohibitively small. RNA-seq gene expression and survival measurements were downloaded from The Cancer Genome Atlas (TCGA) for bladder urothelial carcinoma (26), ovarian epithelial tumor (27) and colorectal adenocarcinoma (28). Additionally, microarray GE of cisplatin-treated patients of cell carcinoma of the urothelium (N=30) (29) and for oxaliplatin-treated colorectal cancer (CRC) patients (30) were downloaded. Drug treatment data 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 (31,32).

Initial gene sets for developing signatures for each drug were identified from previously published literature and databases, such as PharmGKB and DrugBase (33–40). The gene list was then cross-referenced to the Reactome gene network (31,41) to select additional genes which might contribute to the drug response, based on proximity to those genes in the same pathways. Genes were then eliminated if they showed no or little expression in TCGA bladder cancer patients (where RNA-seq by Expectation Maximization [RSEM] is < 5.0 for majority of individuals). The final gene sets were chosen using Multiple Factor Analysis (MFA) to analyze interactions between GE, copy

number (CN), and GI_{50} data for the drug of interest (42). Gene whose GE and/or CN showed a direct or inverse correlation with GI_{50} were selected for SVM training. This reduces the overall number of genes for SVM analysis, and thus helps to avoid a data to size sample imbalance. For cisplatin, MFA was repeated using IC_{50} values for 17 bladder cancer cell lines; however the available CN data was commonly invariant for these genes. Therefore, the available IC_{50} values for three other cancer drugs (doxorubicin, methotrexate and vinblastine) were used as a differentiating factor against the IC_{50} of cisplatin.

Applying an SVM model directly to patient data without a normalization approach is imprecise when training and testing data are not obtained using similar methodology (i.e. different microarray platforms). To compare the cell line GE microarray data and the patient RNA-seq GE datasets, expression values were normalized by conversion to z scores using MATLAB (43). Log2 microarray data was not available. However, RNA-seq GE and $\log_2$ GE values from microarray data are highly correlated (44).

Support Vector Machine Learning

Support vector machines (SVM) were trained with breast cancer cell line GE datasets (15) with the Statistics Toolbox in MATLAB (43) using Gaussian kernel functions (fitcsvm), and then tested with a leave-one-out cross-validation (using '*crossval'* and '*leaveout'* options). A greedy backwards feature selection algorithm was used to improve classification accuracy (45,46). The gene subset with the lowest misclassification is selected for subsequent testing with patient gene expression and clinical data.

Variable GI_{50} thresholded gene signatures.

We previously set a threshold distinguishing drug sensitivity from resistance based on the median concentration of drug that inhibited cell growth by 50%. By varying this threshold, we hypothesized that gene signatures could be derived that distinguished sensitivity vs. resistance at different levels of drug resistance. Machine learning experiments to determine the effect shifting the threshold for classifying resistance or sensitivity (according to GI_{50}) were conducted by generating optimized Gaussian SVM models at each new threshold and assessing the performance with patient data. This results in a series of SVM models which were derived over a range of response thresholds. A heat map which illustrates the frequency of genes selected with this method was created with the R language *hist2d* function.

A composite gene signature was created by ensemble averaging of all models generated at each resistance threshold. Ensemble averaging combines signatures through averaging the weighted accuracy of a set of related models (47,48). This allows multiple weak performing models to produce results with better performance than any of the individual models used to derive the average (48,49). The decision function for the ensemble classifier is the mean of the decision function scores of the component classifiers, weighted by the Area Under the Curve (AUC).

The potential for models to overfit the data was assessed by permutation analysis randomizing labels, and by randomly selecting sets of genes, as described previously (6). The methods determine whether overfitting occurs during training and/or feature selection.

RESULTS

Selection of Platin Drug Related Genes

Development of biochemically-inspired machine learning models of platin drug response began with collection of genes exhibiting peer-reviewed associations with the effectiveness or response. For cisplatin, this implicated 178 genes as well as others crucial for the function of those genes (Supplementary Table S1A). MFA analysis of breast cancer cell line data (15) reduced this to 39 genes where GI_{50} was correlated with GE and/or CN. Figure 1 illustrates the diverse roles of some of the most significant gene products in cisplatin response and the type of GI_{50} correlation (GE and/or CN; positive or negative). Genes with insufficient expression in TCGA bladder urothelial carcinoma datasets were removed (*MT1A, MT3, MT4, SLC22A10, SLC22A11, SLC22A13, SLC22A16*). The remaining 32 genes used to derived SVM gene signature models were: *ATP7B, BARD1, BCL2, BCL2L1, CDKN2C, CFLAR, ERCC2, ERCC6, FAAP24, FEN1, FOS, GSTO1, GSTP1, MAP3K1, MAPK13, MAPK3, MSH2, MT2A, NFKB1, NFKB2, PNKP, POLD1, POLQ, PRKAA2, PRKCA, PRKCB, SLC22A5, SLC31A2, SNAI1, TLR4, TP63,* and *TWIST1*.

Further MFA analyses for cisplatin were performed with a different bladder cancer cell line dataset UBC-40 with GI_{50} data (50). The genes which exhibited significant relationships to GI_{50} in the breast cancer lines were again subjected to MFA using the bladder cancer cell line data (N=12). This data showed cisplatin IC_{50} was related to gene expression for *CFLAR*, *FEN1*, *MAPK3*, *MSH2*, *NFKB1*, *PNKP*, *PRKAA2*, and *PRKCA*. Similarly, bladder cell line IC_{50} values from cancerRxgene were related to the gene expression of *CFLAR*, *FEN1*, and *NFKB1*, in addition to *ATP7B*, *BARD1*, *MAP3K1*, *NFKB2*, *SLC31A2* and *SNAI1*. An SVM generated using the combination of these gene sets, with median IC_{50} dividing sensitive and resistant cells, had a classification error of 11.8% (*PNKP* and *PRKCA*). When including the low level expressed gene *SLC22A11*, which also correlated with cisplatin IC_{50}, an alternate SVM-

based signature with an equivalent error rate was selected (consisting of *ATP7B*, *CFLAR*, *FEN1*, *MAPK3*, *NFKB1* and *SLC22A11*).

In some instances, cisplatin-treated patients were also treated with carboplatin (in combination with other drugs). From an initial set of 90 genes which play a role in carboplatin sensitivity and resistance, 28 showed correlation to GI_{50} with MFA analysis (Supplementary Table S1B). From these 28 genes, the following carboplatin SVM was generated (cross-validation misclassification rate of 10.4%; median GI_{50} threshold): *AKT1*, *ATP7B*, *EGF*, *EIF3I*, *ERCC1*, *GNGT1*, *HRAS*, *MTR*, *NRAS*, *OPRM1*, *RAD50*, *RAF1*, *SCN10A*, *SGK1*, *TIGD1*, *TP53*, and *VEGFB*. For oxaliplatin, the initial set of 288 genes was reduced to 55 using MFA (Suppl. Table S1C). To avoid overfitting, the SVM analysis was limited to the 35 genes where GE was correlated with GI_{50}. With these genes, the median GI_{50} model for oxaliplatin was derived with a 2.1% misclassification rate: *AGXT*, *APOBEC2*, *BRAF*, *CLCN6*, *FCGR2A*, *IGF1*, *MPO*, *MSH2*, *NAGK*, *NAT2*, *NFE2L2*, *NOTCH1*, *PANK3*, *PRSS1*, and *UGT1A1*.

Although platinum-based drugs conjugate with DNA by a common mechanism, responses of breast cancer cell lines to cisplatin and carboplatin appeared to be more similar to each other, than either was to oxaliplatin. MFA of GI_{50} values for cisplatin, carboplatin and oxaliplatin indicated a direct correlation between the cis- and carboplatin drugs, but oxaliplatin did not correlate with either (Figure 2). The response to oxaliplatin suggests that its sensitivity and effectiveness is different in these cell lines, which is consistent with previous studies showing that cisplatin-resistant cell lines are generally sensitive to oxaliplatin (51–53).

GI_{50} Threshold Independent Modeling

Previously, we established a convention where the threshold between drug resistance and sensitivity for cell line data was set to the median GI_{50} value (6). While a valid starting point (5,6), it is arbitrary. We've therefore developed a threshold independent machine learning approach, in addition to approaches for identification and selection of predictive genes. We performed machine learning experiments by shifting the GI_{50} threshold to which cell lines are labeled resistant. Optimized Gaussian SVM models were created over a range of response thresholds, followed by an ensemble averaging approach to create a threshold independent machine learning technique for cell line data.

Across all the GI_{50} thresholded models derived for cisplatin, certain genes were more often included in the final gene set (Figure 3A). Kinases (*MAPK3*, *MAP3K1*) and apoptotic family members (*BCL2*, *BCL2L1*) were most common, with genes associated with error-prone and base-excision DNA repair also consistently represented (Supplementary Table S2A). The kinases more commonly appear at lower thresholds for sensitivity, while *BCL2* and *BCL2L1* are more ubiquitous. Of the DNA repair genes, *POLD1* and *POLQ* are more frequently represented in models with lower sensitivity thresholds, while *FEN1* is generally selected at high cisplatin GI_{50} thresholds. Thresholded models 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 gene was less common when the threshold for sensitivity was decreased (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). The mismatch repair (MMR) gene *MSH2* is commonly found in gene signatures when the GI_{50} threshold is set to high levels of resistance, which is of interest as MMR

deficiency has been shown to be predictive for oxaliplatin resistance (54). The autoimmune disease-associated gene *SIAE*, which has been previously shown to have a strong negative correlation to oxaliplatin response in advanced CRC patients (55), was selected for the vast majority of threshold oxaliplatin models. The gene *BCL2*, which was commonly selected for cisplatin, was rarely selected for oxaliplatin.

The transcription profile of the breast cancer cell lines and bladder cancer test patients will obviously have significant differences. To be able to compare these differences in relation to gene priority, SVM models were also built using the cisplatin and/or carboplatin-treated TCGA bladder urothelial carcinoma patients (N=72) using a series of post-treatment time to recurrence as resistance thresholds: 6 months (mo.), 9 mo., 12 mo., 18 mo., 2 years, 3 years and 4 years (Supplementary Table S3). Similar trends to the cell line SVMs are apparent; *POLQ* is frequently selected when the resistance threshold is set at a longer time period, while *FEN1* appears when it is shorter. However, *BCL2*, which is present in a majority of breast cancer cell line SVMs is present only once in all models derived from TCGA data. Similarly, *MSH2* was rarely selected using cell line data, yet appears in nearly all TCGA patient derived SVMs with > 1 year recurrence.

GI_{50} threshold models for each platin drug were performed with the breast cancer cell line data, produced 70 cisplatin, 83 carboplatin, and 83 oxaliplatin SVM models across varying drug resistance thresholds. Each of these models were then applied against available platin-treated patient datasets: 1) 72 TCGA bladder urothelial carcinoma patient data and 30 additional patients with urothelial cell carcinoma from Als et al. (29) treated with cisplatin; 2) 410 TCGA ovarian epithelial tumor patients treated with carboplatin; and 3) 99 oxaliplatin-treated TCGA colorectal adenocarcinoma (CRC) patients, and an additional 83 CRC patients from Tsuji et al (30). The chemotherapy

response metadata provided for these studies were not always consistent. Als et al. reported survival post-treatment and Tsuji et al. categorized patients as responders and non-responders, while TCGA provided multiple data points which could better illustrate chemotherapy response. Two different measures provided by TCGA were assessed for predictive accuracy in our models – chemotherapy response and disease free survival. Accuracy is similar using either measure (Supplementary Table S4A); however recurrence and disease free survival was used as the primary measure of response as it provides a higher numbers of observations in all three TCGA data sets tested. Patients from Als et al. with a ≥ 5 year survival post-treatment were labeled as sensitive to treatment. The difference between the data used to assign drug sensitivity may, in part, account for the differences in the prediction accuracy of the thresholded SVM models.

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. The reverse is true for models built using lower resistance thresholds (high GI_{50}). We therefore state SVMs generated at these extreme thresholds are not very useful at predicting patient data. When used to predict recurrence in the TCGA datasets, sensitivity and specificity appears to be maximized in models where the GI_{50} threshold for resistance was set near (but not necessarily at) the median (Supplemental Figure 1; Suppl. Tables S4A to 4C). While this pattern holds true for Tsuji et al (30), oxaliplatin models where GI_{50} thresholds were set above the median could better separate primary and metastatic CRC patients (best model predicting 92.6% metastatic and 60.7% primary cancers; Suppl. Table S4C). While less consistent, cisplatin models generated with thresholds above median GI_{50} performed better when evaluating the Als *et al.* (29) patient dataset (Suppl. Figure 2).

Models were further evaluated for their accuracy to TCGA patients using various recurrence times post-treatment to classify resistant and sensitive patients (0.5 - 5 years; Suppl. Table S5A-C). The best performing cisplatin model (*BARD1, BCL2, BCL2L1, CDKN2C, FAAP24, FEN1, MAP3K1, MAPK13, MAPK3, NFKB1, NFKB2, SLC22A5, SLC31A2, TLR4, TWIST1*; GI_{50} of 5.11; hereby identified as **Cis1**) was able to accurately predict 71.0% of bladder cancer patients who recurred after 18 mo. (N=31; 58.5% accurate for disease-free patients [N=41]). The best performing carboplatin model (*AKT1*, *EIF3K*, *ERCC1*, *GNGT1*, *GSR*, *MTHFR*, *NEDD4L*, *NLRP1*, *NRAS*, *RAF1*, *SGK1*, *TIGD1*, *TP53*, *VEGFB*, *VEGFC*; designated **Car1** [Suppl. Table S5B]) was developed using a GI_{50} threshold of 4.42, predicting recurrence of ovarian cancer after 4 years at an accuracy of 60.2% (N=302; 61.0% accurate for disease-free patients [N=108]). For oxaliplatin, the best performing model (*BRAF*, *FCGR2A*, *IGF1*, *MSH2*, *NAGK*, *NFE2L2*, *NQO1*, *PANK3*, *SLC47A1*, *SLCO1B1*, *UGT1A1*; GI_{50} of 5.10; designated **Oxa1** [Suppl. Table S5C]) accurately predicted 71.6% of the disease-free TCGA colorectal cancer patients after one year (N=88; 54.5% accuracy predicting recurrence [N=11]). These models, as well as the previously mentioned SVMs based on bladder cell line data, were added to the online web-based SVM calculator (http://chemotherapy.cytognomix.com; introduced in [6]) which can be used to predict platin response using normalized gene expression values.

All models generated for each individual platin drug were evaluated in a threshold independent manner through ensemble machine learning, which involves the averaging of hyperplane distances for each model to generate a composite score for each TCGA patient tested. Hyperplane distances across all cisplatin models were close in range, with a mean score of -0.22 and a standard deviation of 3.5 hyperplane units across the set of patient data. The resulting cisplatin ensemble model combines

predictions from 70 different models, each optimized for a particular resistance threshold. The ensemble model classified disease-free bladder cancer patients with 59% accuracy and those with recurrent disease with 47% accuracy. Despite their better performance on an individual basis, limiting ensemble averaging to only cisplatin models generated at a moderate GI_{50} threshold (ranging from 5.10 to 5.50) did not show a significant improvement (44% accuracy for disease-free and 66% for recurrent patients; Supplementary Table S6A). Ensemble averaging predictions for carboplatin were never significantly better than random, regardless of the GI_{50} threshold interval selected (Suppl. Table S6B), despite a similar distribution of hyperplane distances (-0.11 +/- 3.9 hyperplane units). Ensemble averaging for oxaliplatin (-0.12 +/- 2.7 hyperplane units) showed to be most accurate after 1 year (60% accuracy for disease-free and 73% for recurrent patients; Suppl. Table S6C). Similar to cisplatin, limiting this analysis to moderate oxaliplatin GI_{50} models significantly increased accuracy. In each case, the accuracy of the averaged model is comparable to, albeit lower than that of the model which performed best across all models generated.

Model Prediction for Bladder Urothelial Carcinoma Patients with Variable Responses

To evaluate the consistency in the prediction of TCGA bladder cancer patients treated with cisplatin, the predicted distance from the hyperplane for all SVMs generated were plotted for each patient with a short recurrence time (<6 mo., N=10; Supplementary Figure 3A). Despite showing similar levels of resistance to treatment, patterns 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 models. 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 3B).

As the cisplatin models generated consist of an optimized gene set which minimized misclassification, their removal can only decrease cross-validation accuracy. To determine their independent impact on misclassification, each gene within every SVM model was excluded, and model accuracy was reassessed (Supplementary Table S2A; S2B and S2C contain carbo- and oxaliplatin models, respectively). Genes which consistently cause significant increases in misclassification (averaging > 16% increase) in moderate threshold SVMs (GI_{50} thresholds set from 5.1 to 5.5) include *ERCC2*, *POLD1*, *BARD1*, *BCL2*, *PRKCA* and *PRKCB*. *ERCC2* and *POLD1* perform nucleotide and base excision repair, respectively, and are both apoproteins lacking 4Fe-4S (www.reactome.org). *PRKCA* and *PRKCB* are paralogs with significant roles in signal transduction. *BARD1* has been shown to reduce *BCL2* in the mitochondria when overexpressed (56), and *BARD1* does play a role in apoptosis (57). Commonly selected genes with a high variance in increased misclassification between different models include *NFKB1*, *NFKB2*, *TWIST1*, *TP63*, *PRKAA2*, and *MSH2*. 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 paired together in 7 SVMs generated at a moderate GI_{50} threshold. There is evidence of possible epistatsis 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*, when paired with *NFKB2*, is over one-third lower compared to models where *NFKB2* is absent.

Predicting cisplatin response in patients based on smoking history

Tobacco smoking is known as the highest risk factor for the development of bladder cancer (58). We therefore subdivided the patients based on their smoking history and tested the thresholded models (Supplementary Tables S7 and S8). 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 model (**Cis18**; Suppl. Table S7) had performed equally as well for non-smokers, and these two models share 7 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 model built at the median GI_{50} threshold. 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. Genes in this SVM which are not present in the two models which performed well for non-smokers include *CFLAR* and *PRKAA2*.

Tobacco smoking history has been shown to have a significant impact on methylation levels in the genome (59). In a subset of the TCGA bladder urothelial carcinoma patients tested, CpG island methylation and smoking pack years was reported (26). We suspected that the level of methylation measured in the genes in SVMs which best performed for smoking and non-smoking patients might be different, and with possible concomitant effects on gene expression. When ranking each gene from **Cis1** by highest methylation and gene expression, only 88 patient-gene pairs

(methylation and expression levels for a specific gene in a particular patient; N=1080 total pairs) showed the expected inverse correlation with methylation levels and gene expression (i.e. high methylation and low gene expression). However, instances of an inverse correlation of methylation and gene expression levels was more common than direct correlation (i.e. high methylation and high expression; N=17). When evaluating these instances based on smoking history, direct correlation was more common in patients with a recent smoking history (70.5%). This pattern was also observed in the median GI_{50} model (**Cis2**; second best performing model; Supplementary Table S5A), which best predicted recurrence in smokers. It may be possible that in these cases of direct correlation of methylation and gene expression, smoking is instead altering expression by mutational effects rather than methylation.

To determine which genes in these high-performing models were leading to discordant predictions of patient outcome, the expression for each gene was gradually altered (individually) until the prediction was corrected. If the gene expression value required to cross this threshold went beyond the highest or lowest expression level for that gene, the gene was considered to have a low contribution to the prediction. Genes which, when altered, commonly corrected discordant predictions included *MAP3K1*, *MAPK3*, *SLC22A5* and *SLC31A2*, indicating that these genes may have a large impact on discordant predictions. Genes which commonly could not alone correct a discordant prediction (required expression change ≥ 3-fold the highest/lowest expression of that gene) included *PRKAA2*, *NFKB1*, *NFKB2* and *TWIST1*. Altering *BCL2L1* expression was more likely to correct the discordant predictions of **Cis1** (4 out of 5) than with **Cis2** (2 out of 4).

Model Portability

Backwards feature selection is prone to overfitting and can get stuck in a local minimum (60). Potential to overfit was assessed for each model by permutation analysis with both randomly selected genes and with randomized labels. If the misclassification rate these permutations are commonly equal to or better than the model being compared, the model is considered overfit. Random gene selection was tested using the median cisplatin GI_{50} as the resistance threshold for 15 genes; the number of genes within models **Cis1**, **Car1** and **Oxa1.** Models generated with a random set of genes had a rate of misclassification exceeding the lowest median SVM models (error rate of 5.1%). After 10000 iterations, 2 random gene signatures had an error rate of 7.7%. Thus, signatures of randomly-selected genes exhibited higher misclassification errors than the best performing SVM models. Randomization of cell line labels as either sensitive or resistant to the drug based on GI_{50} thresholds were also used to test significance of the best gene signatures. After 10000 iterations, cisplatin, carboplatin and oxaliplatin gene expression data for these label combinations respectively generated 8, 1 and 1 signatures with lower error rates than the best biochemically-inspired signatures.

DISCUSSION

Using gene expression signatures, we derived both GI_{50} threshold-dependent and -independent machine learning models which evaluate the chemotherapy response for cisplatin, carboplatin and oxaliplatin. The cisplatin model **Cis1** (best performing cisplatin model; Supplementary Table S5A) most accurately predicted response in bladder cancer patients after 18 months, and **Car1** (best performing carboplatin model; Suppl. Table S5B) best predicted response in ovarian cancer patients after 4 years. Both models better predicted recurrent patients compared to disease-free, while **Oxa1** (best oxaliplatin model; Suppl. Table S5C) more accurately predicted disease-free patients than recurrent at a one year threshold. The thresholds which best represented time to

recurrence are different between the platin drugs in each cancer type. Not surprisingly, smoking history affects the accuracy of cisplatin models, as certain models had noticeably improved performance when this factor was taken into account.

Despite their similarities in mode of action, the three platin drugs analyzed in this study result in distinctly different gene signature models. Initial gene sets selected from the literature contains some overlap between platin drugs (N=67 between any two platins), but very few of the genes shared showed MFA correlation for multiple platin genes (N=3; *ATP7B*, *BCL2* and *MSH2*). Despite the close similarity between cisplatin and carboplatin GI_{50} (see Figure 2), only one gene (*ATP7B*) showed MFA correlation for both. *BCL2* and *MSH2* correlated with both cisplatin and oxaliplatin GI_{50}. The increase in misclassification caused by the removal of *MSH2* from either drug is often significant, such as the oxaliplatin model **Oxa21** (Supplementary Table S5C) and cisplatin model **Cis14** (Suppl. Table S5A) at +19.1% and +28.2% misclassification, respectively. As so few genes are correlated with GI_{50} among all three drugs, the overall gene signatures differ significantly. These differences may reflect the spectrum of activity, sensitivity, and toxicity of the three platin genes (51–53).

Gene signature models derived from cell lines and tested on patients differ in their outcome measures; while plausible, survival data in patients after drug treatment has not been proven to be related to growth inhibition of cell lines with the same drug. Further, 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 (61,62). In this study, we developed GI_{50}-threshold independent machine learning models for platin drugs which predict drug response without relying on arbitrary GI_{50} thresholds for sensitivity and resistance. The ensemble developed using this approach is more accurate than the majority of individual models making up the ensemble. For cisplatin,

ensemble averaging over models generated on different resistance thresholds shows a slightly increased accuracy over most models, better representing the sensitive, disease-free class (eg. 59% accuracy). Interestingly, ensemble averaging of only the models 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 the latter model having a greater overall impact on the ensemble prediction. When limiting ensemble averaging to only those models with the highest AUC at each resistant threshold, differences in the subsequent predictions were negligible. Ensemble machine learning can potentially avoid problems with poor performance and overfitting by combining models that individually only do slightly better than chance to increase the accuracy (48).

While it is possible to find accurate gene signatures without any of the features being related to chemoresistance, this is difficult to reconcile with tumor biology. From a discovery standpoint, it is preferable to explore pathways related to genes in the signature to improve accuracy, identify potential targets for further study into chemoresistance, and expand the model parameters to take into account alternate resistance states other than those captured in the original signature (63). In addition to improving classification accuracy, our resistance thresholding approach may reveal potentially important genes and pathways associated with cisplatin resistance. At the highest levels of resistance (lower GI_{50} thresholds), the cisplatin models 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 (35,39,64) and these specific genes may be explored in subsequent work to identify the contribution to chemotherapy response in a biochemical context.

Our methodology can be applied to the derivation of gene signatures for other drugs. This is crucial, as patients of most cancer types will be treated with a combination of anti-cancer drugs. Along with cisplatin and/or carboplatin, all TCGA bladder urothelial carcinoma patients analyzed in this study were also treated with gemcitabine. Thus, patient response may be independent of treatment with the platinum-based drugs. However, gemcitabine performed even more poorly on the patients in this study (AUC = 0.50, data not shown) so analysis was suspended for the time being. Not included in our analysis were signatures for methotrexate, vinblastine, and doxorubicine (part of the MVAC cocktail) in bladder cancer. This was due primarily to a lack of patients treated with this drug in the TCGA bladder dataset (N=11). Methotrexate and doxorubicin models were derived using GE data from 84 breast cancer patients (Molecular Taxonomy of Breast Cancer International Consortium; METABRIC) who were treated with both combination chemotherapy and hormone therapy (65). The doxorubicin model (*ABCC2*, *ABCD3*, *CBR1*, *FTH1*, *GPX1*, *NCF4*, *RAC2*, *TXNRD1*) and the methotrexate model (*ABCC2*, *ABCG2*, *CDK2*, *DHFRL1*) could accurately predict mortality for 75% and 71.4% of patients, respectively, though it must be noted that patient chemotherapy drug combinations were not consistent nor fully described across the cohort which could negatively contribute to model accuracy. In the future, development and testing of predictions of a combination chemotherapy approach could be applied that could better evaluate patient data.

The predictive accuracy for the same model could differentiate highly between the two datasets. **Cis3** (Supplemental Table S5A), which was generated at a high resistance threshold (GI_{50} of 5.60) had a high predictive accuracy for TCGA bladder cancer patients, with the highest AUC calculated across all models (0.64). However, the AUC calculated was low when applied to the Als *et al.* (29) dataset (0.18). Patient

metadata for the latter only reports patient survival times, while we base the expected TCGA patient outcome on time to recurrence. As the basis of our expected outcome differs between datasets, these differences may be acting as a confounding factor to model accuracy. The datasets also differ in methodology for expression measurement (microarray and RNA-seq). The relevance of models based on training and testing data from different platforms can affect the accuracy of validation, which may not be improved by normalization approaches. Currently, GE values of each dataset were subject to zscore normalization, as divergent experimental conditions, batch effects, and different microarray platforms make direct comparison highly inaccurate. In subsequent studies, other techniques to correct for some of these effects have been described and could be used (66).

We have also developed a number of methods which can detect overfitting during model building. One method of detection is to reserve a portion of the dataset for use as a validation cohort (67) which we do in this work by reserving the patient data outside of the training and feature selection algorithm. However, when training on cell lines and testing on patients, it can be difficult to determine if the performance is the result of overfitting or a result of genetic differences between cell lines and patients, measurement uncertainty, or related to uncertainty as to whether a particular patient can be labeled resistant to the chemotherapy or not.

In addition, bladder cancer cell line panels with drug sensitivity data are limited. We recently acquired bladder cancer IC_{50} data and gene expression data from The Genomics of Drug Sensitivity in Cancer project (68,69) (N=18). In preliminary testing the resulting model was not predictive of outcome. However, with only 18 observations, it is a prohibitively small training set, and is likely insufficient in size for robust patient testing.

Additional testing is planned once an adequately sized bladder cancer dataset becomes available.

CONCLUSIONS

We have generated and validated a novel approach to predict chemotherapy response to platin agents in cancer patients. Using SVMs, we developed a threshold independent machine learning model which can predict chemotherapy response without relying on arbitrary GI_{50} thresholds for sensitivity and resistance. The ensemble developed using this approach is more accurate than the majority of single models making up the ensemble. This approach also allowed us to identify genes which may be involved in cisplatin resistance, including those that may have greater impact dependant on patient smoking history. The methodology described here can be adapted to other work looking at prediction of chemoresistance, independently of the drug or cancer type described. By create a range of models for many 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 (70).

Author Contributions

PKR designed the methodology with input from DL, and PKR supervised the project. EJM ran MFA analysis for gene selection, performed the original resistance thresholding experiments. JZ contributed to development of gene signatures. EJM and PKR wrote the manuscript.

Funding

PKR is supported by Natural Sciences and Engineering Research Council of Canada (Discovery Grant RGPIN-2015-06290), Canadian Foundation for Innovation, Canada Research Chairs Secretariat, and CytoGnomix Inc..

Acknowledgements

We acknowledge Katherina Baranova, who contributed to the early development of cisplatin gene signatures in this paper, and Dimo Angelov who developed the automated feature selection algorithm. Compute Canada and Shared Hierarchical Academic Research Computing Network (SHARCNET) providing high performance computing and storage facilities, which were used for some aspects of this project.

References

1. Cardoso F, Harbeck N, Fallowfield L, Kyriakides S, Senkus E, on behalf of the ESMO Guidelines Working Group. Locally recurrent or metastatic breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2012 Oct 1;23(suppl 7):vii11-vii19.

2. Oostendorp LJ, Stalmeier PF, Donders ART, van der Graaf WT, Ottevanger PB. Efficacy and safety of palliative chemotherapy for patients with advanced breast cancer pretreated with anthracyclines and taxanes: a systematic review. Lancet Oncol. 2011 Oct;12(11):1053–61.

3. Alfarouk KO, Stock C-M, Taylor S, Walsh M, Muddathir AK, Verduzco D, et al. Resistance to cancer chemotherapy: failure in drug response from ADME to P-gp. Cancer Cell Int. 2015;15:71.

4. Gąsowska-Bodnar A, Bodnar L, Dąbek A, Cichowicz M, Jerzak M, Cierniak S, 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. 2014 May;24(4):687–96.

5. Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. JAMA. 2011 May 11;305(18):1873–81.

6. Dorman SN, Baranova K, Knoll JHM, Urquhart BL, Mariani G, Carcangiu ML, et al. Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning. Mol Oncol. 2016 Jan;10(1):85–100.

7. Ali I, Wani WA, Saleem K, Haque A. Platinum compounds: a hope for future cancer chemotherapy. Anticancer Agents Med Chem. 2013 Feb;13(2):296–306.

8. Wheate NJ, Walker S, Craig GE, Oun R. The status of platinum anticancer drugs in the clinic and in clinical trials. Dalton Trans Camb Engl 2003. 2010 Sep 21;39(35):8113–27.

9. Woynarowski JM, Faivre S, Herzig MCS, Arnett B, Chapman WG, Trevino AV, et al. Oxaliplatin-Induced Damage of Cellular DNA. Mol Pharmacol. 2000 Nov 1;58(5):920–7.

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. 2008 Jan 28;14(4):1330–41.

11. Howell SB, Safaei R, Larson CA, Sailor MJ. Copper Transporters and the Cellular Pharmacology of the Platinum-Containing Cancer Drugs. Mol Pharmacol. 2010 Jun;77(6):887–94.

12. Zhang S, Lovejoy KS, Shima JE, Lagpacan LL, Shu Y, Lapuk A, et al. Organic Cation Transporters Are Determinants of Oxaliplatin Cytotoxicity. Cancer Res. 2006 Sep 1;66(17):8847–57.

13. Wang D, Lippard SJ. Cellular processing of platinum anticancer drugs. Nat Rev Drug Discov. 2005 Apr;4(4):307.

14. 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. 1989 Jan 1;43(4):251–60.

15. Daemen A, Griffith OL, Heiser LM, Wang NJ, Enache OM, Sanborn Z, et al. Modeling precision treatment of breast cancer. Genome Biol. 2013;14(10):R110.

16. Han M, Dai J, Zhang Y, Lin Q, Jiang M, Xu X, et al. Support vector machines coupled with proteomics approaches for detecting biomarkers predicting chemotherapy resistance in small cell lung cancer. Oncol Rep. 2012 Dec;28(6):2233–8.

17. Yuan Y, Tan C-W, Shen H, Yu J-K, Fang X-F, Jiang W-Z, 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. 2012 Dec;59(120):2461–5.

18. 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. 2008 Feb 26;9:99.

19. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinforma Oxf Engl. 2010 Jun 15;26(12):i237-245.

20. Nickerson ML, Witte N, Im KM, Turan S, Owens C, Misner K, et al. Molecular analysis of urothelial cancer cell lines for modeling tumor biology and drug response. Oncogene. 2016 Jun 6;

21. Oommen T, Misra D, Twarakavi NKC, Prakash A, Sahoo B, Bandopadhyay S. An Objective Analysis of Support Vector Machine Based Classification for Remote Sensing. Math Geosci. 2008 Mar 14;40(4):409–24.

22. Yuryev A. Gene expression profiling for targeted cancer treatment. Expert Opin Drug Discov. 2015 Jan;10(1):91–9.

23. Sataloff DM, Mason BA, Prestipino AJ, Seinige UL, Lieber CP, Baloch Z. Pathologic response to induction chemotherapy in locally advanced carcinoma of the breast: a determinant of outcome. J Am Coll Surg. 1995 Mar;180(3):297–306.

24. Ogston KN, Miller ID, Payne S, Hutcheon AW, Sarkar TK, Smith I, et al. A new histological grading system to assess response of breast cancers to primary chemotherapy: prognostic significance and survival. Breast Edinb Scotl. 2003 Oct;12(5):320–7.

25. Kreike B, Halfwerk H, Kristel P, Glas A, Peterse H, Bartelink H, et al. Gene expression profiles of primary breast carcinomas from patients at high risk for local recurrence after breast-conserving therapy. Clin Cancer Res Off J Am Assoc Cancer Res. 2006 Oct 1;12(19):5705–12.

26. Weinstein JN, Akbani R, Broom BM, Wang W, Verhaak RGW, McConkey D, et al. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014 Jan 29;507(7492):315–22.

27. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011 Jun 29;474(7353):609–15.

28. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012 Jul 18;487(7407):330–7.

29. Als AB, Dyrskjøt L, von der Maase H, Koed K, Mansilla F, Toldbod HE, et al. Emmprin and survivin predict response and survival following cisplatin-containing chemotherapy in patients with advanced bladder cancer. Clin Cancer Res Off J Am Assoc Cancer Res. 2007 Aug 1;13(15 Pt 1):4407–14.

30. Tsuji S, Midorikawa Y, Takahashi T, Yagi K, Takayama T, Yoshida K, et al. Potential responders to FOLFOX therapy for colorectal cancer by Random Forests analysis. Br J Cancer. 2012 Jan 3;106(1):126–32.

31. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013 Apr 2;6(269):pl1.

32. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discov. 2012 May 1;2(5):401–4.

33. Basu A, Krishnamurthy S. Cellular responses to Cisplatin-induced DNA damage. J Nucleic Acids. 2010;2010.

34. Galluzzi L, Senovilla L, Vitale I, Michels J, Martins I, Kepp O, et al. Molecular mechanisms of cisplatin resistance. Oncogene. 2012 Apr 12;31(15):1869–83.

35. Borst P, Rottenberg S, Jonkers J. How do real tumors become resistant to cisplatin? Cell Cycle Georget Tex. 2008 May 15;7(10):1353–9.

36. Choi W, Porten S, Kim S, Willis D, Plimack ER, Hoffman-Censits J, et al. Identification of Distinct Basal and Luminal Subtypes of Muscle-Invasive Bladder Cancer with Different Sensitivities to Frontline Chemotherapy. Cancer Cell. 2014 Feb;25(2):152–65.

37. Poisson LM, Munkarah A, Madi H, Datta I, Hensley-Alford S, Tebbe C, et al. A metabolomic approach to identifying platinum resistance in ovarian cancer. J Ovarian Res. 2015 Mar 26;8:13. doi: 10.1186/s13048-015-0140-8.

38. von Stechow L, Ruiz-Aracama A, van de Water B, Peijnenburg A, Danen E, Lommen A. Identification of Cisplatin-Regulated Metabolic Pathways in Pluripotent Stem Cells. Santos J, editor. PLoS ONE. 2013 Oct 16;8(10):e76476.

39. Wernyj R, Morin P. Molecular mechanisms of platinum resistance: still searching for the Achilles? heel. Drug Resist Updat. 2004 Oct;7(4–5):227–32.

40. Mo D, Fang H, Niu K, Liu J, Wu M, Li S, et al. Human Helicase RECQL4 Drives Cisplatin Resistance in Gastric Cancer by Activating an AKT-YB1-MDR1 Signaling Pathway. Cancer Res. 2016 May 15;76(10):3057–66.

41. Milacic M, Haw R, Rothfels K, Wu G, Croft D, Hermjakob H, et al. Annotating cancer variants and anti-cancer therapeutics in reactome. Cancers. 2012;4(4):1180–211.

42. Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev Comput Stat. 2010 Jul;2(4):433–59.

43. MATLAB and Statistics Toolbox Release 2012b, The MathWorks, Inc., Natick, Massachusetts, United States.

44. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008 Sep;18(9):1509–17.

45. Bermingham ML, Pong-Wong R, Spiliopoulou A, Hayward C, Rudan I, Campbell H, et al. Application of high-dimensional feature selection: evaluation for genomic prediction in man. Sci Rep. 2015 May 19;5:10312.

46. Dash M, Liu H. Feature Selection for Classification. Intell Data Anal. 1997;(3):131–156.

47. Naftaly U, Intrator N, Horn D. Optimal ensemble averaging of neural networks. Netw Comput Neural Syst. 1997 Jan 1;8(3):283–96.

48. Clemen RT. Combining forecasts: A review and annotated bibliography. Int J Forecast. 1989 Jan;5(4):559–83.

49. Tan AC, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Appl Bioinformatics. 2003;2(3 Suppl):S75-83.

50. Earl J, Rico D, Carrillo-de-Santa-Pau E, Rodríguez-Santiago B, Méndez-Pertuz M, Auer H, et al. The UBC-40 Urothelial Bladder Cancer cell line index: a genomic resource for functional studies. BMC Genomics. 2015;16:403.

51. Rixe O, Ortuzar W, Alvarez M, Parker R, Reed E, Paull K, 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. 1996 Dec 24;52(12):1855–65.

52. Mehmood RK. Review of Cisplatin and oxaliplatin in current immunogenic and monoclonal antibody treatments. Oncol Rev. 2014 Sep 23;8(2):256.

53. Kweekel DM, Gelderblom H, Guchelaar H-J. Pharmacology of oxaliplatin and the use of pharmacogenomics to individualize therapy. Cancer Treat Rev. 2005 Apr 1;31(2):90–105.

54. Alex AK, Siqueira S, Coudry R, Santos J, Alves M, Hoff PM, et al. Response to Chemotherapy and Prognosis in Metastatic Colorectal Cancer With DNA Deficient Mismatch Repair. Clin Colorectal Cancer. 2017 Sep;16(3):228-239.

55. Li X-X, Peng J-J, Liang L, Huang L-Y, Li D-W, Shi D-B, et al. RNA-seq identifies determinants of oxaliplatin sensitivity in colorectal cancer cell lines. Int J Clin Exp Pathol. 2014 Jun 15;7(7):3763–70.

56. Tembe V, Martino-Echarri E, Marzec KA, Mok MTS, Brodie KM, Mills K, et al. The BARD1 BRCT domain contributes to p53 binding, cytoplasmic and mitochondrial localization, and apoptotic function. Cell Signal. 2015 Sep;27(9):1763–71.

57. Feki A, Jefford CE, Berardi P, Wu J-Y, Cartier L, Krause K-H, et al. BARD1 induces apoptosis by catalysing phosphorylation of p53 by DNA-damage response kinase. Oncogene. 2005 May 26;24(23):3726–36.

58. Freedman ND, Silverman DT, Hollenbeck AR, Schatzkin A, Abnet CC. Association between smoking and risk of bladder cancer among men and women. JAMA. 2011 Aug 17;306(7):737–45.

59. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic Signatures of Cigarette Smoking. Circ Cardiovasc Genet. 2016 Sep 20;

60. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007 Oct 1;23(19):2507–17.

61. Sos ML, Michel K, Zander T, Weiss J, Frommolt P, Peifer M, et al. Predicting drug susceptibility of non-small cell lung cancers based on genetic lesions. J Clin Invest. 2009 Jun;119(6):1727–40.

62. Laderas TG, Heiser LM, Sönmez K. A Network-Based Model of Oncogenic Collaboration for Prediction of Drug Sensitivity. Front Genet. 2015 Dec 23;6:341.

63. Airley R. Cancer chemotherapy. Chichester, UK; Hoboken, NJ: Wiley-Blackwell; 2009.

64. von Stechow L, Ruiz-Aracama A, van de Water B, Peijnenburg A, Danen E, Lommen A. Identification of Cisplatin-Regulated Metabolic Pathways in Pluripotent Stem Cells. Santos J, editor. PLoS ONE. 2013 Oct 16;8(10):e76476.

65. Mucaki EJ, Baranova K, Pham HQ, Rezaeian I, Angelov D, Ngom A, et al. Predicting Outcomes of Hormone and Chemotherapy in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) Study by Biochemically-inspired Machine Learning. F1000Research. 2017 May 12;5:2124.

66. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostat Oxf Engl. 2007 Jan;8(1):118–27.

67. Breiman L. Heuristics of instability and stabilization in model selection. Ann Stat. 1996 Dec;24(6):2350–83.

68. 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. 2013 Jan;41(Database issue):D955-961.

69. Yang X, Xiong Y, Duan H, Gong R. Identification of genes associated with methotrexate resistance in methotrexate-resistant osteosarcoma cell lines. J Orthop Surg Res. 2015 Sept 4;10:136.

70. Akamatsu N, Nakajima H, Ono M, Miura Y. Increase in acetyl CoA synthetase activity after phenobarbital treatment. Biochem Pharmacol. 1975 Sep 15;24(18):1725–7.

FIGURES

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 multifactor analysis (MFA) of breast (GI_{50}) and bladder (IC_{50}) cancer cell line data.

Figure 2: MFA **c**orrelation circle of GI_{50} values for cisplatin, carboplatin and oxaliplatin from 49 breast cancer cell lines, reveal a direct correlation between cisplatin and carboplatin growth inhibition in different cell lines. Oxaliplatin GI_{50} values were not related to either the cisplatin or carboplatin response.

Figure 3. The variation in gene composition of 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 SVM models in those functional categories, with darker grey indicating frequent genes in indicated GI_{50} threshold intervals, while lighter grey indicates less commonly selected genes.

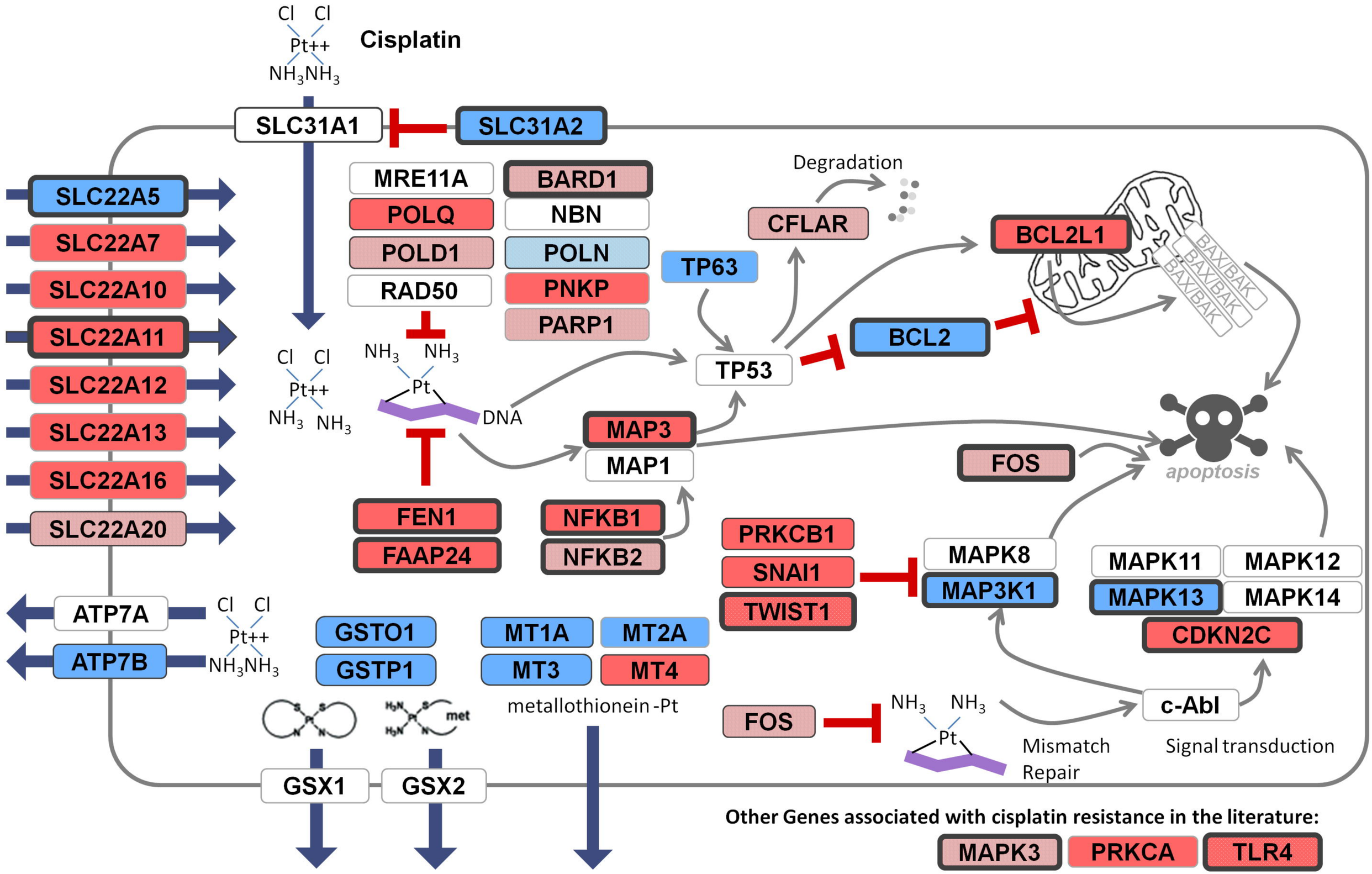
Supplementary Figure 1. Classification accuracy of models on bladder cancer patients treated with cisplatin and/or carboplatin as the resistance threshold is varied. Recurrence and disease free survival are used as a binary measure to assess performance. The x-axis indicates movement of the resistance 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 2. Classification accuracy of SVM models for cisplatin, at a range of response thresholds, were assessed using gene expression data for cisplatin-treated bladder cancer patients. Patients with a ≥ 5 year survival post-treatment were

labeled sensitive. Red arrows indicate the SVM models with the highest positive predictive value (PPV) in the accuracy of classification of patient outcome.

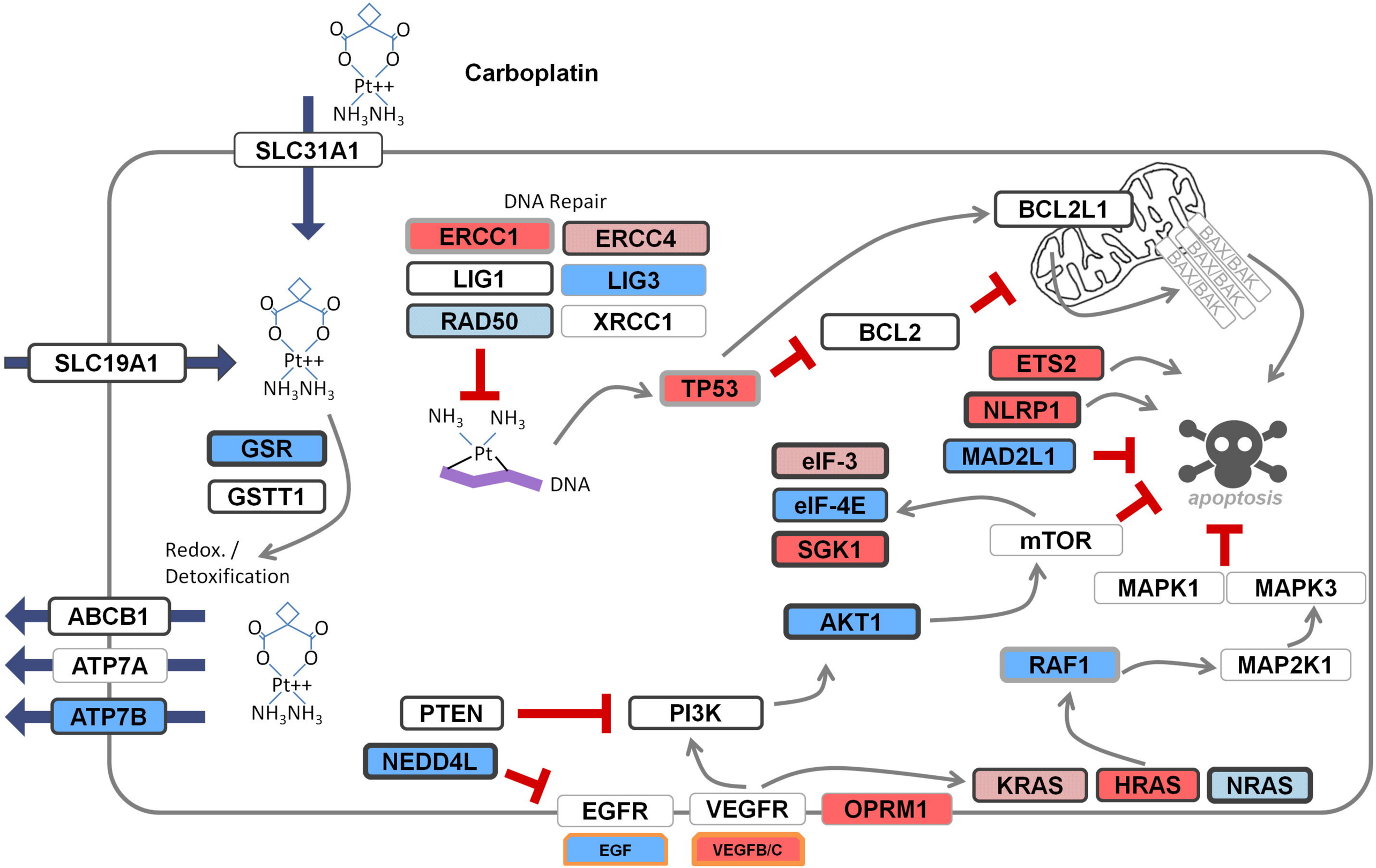
Supplementary Figure 3A. Hyperplane distance calculated by all thresholded SVMs for recurrent (< 6 months) TCGA patients. Each diagram represents the predictions of all SVMs for all patients who had recurrence less than 6 months after treatment (N=10). Each point represents an SVM, where the x-axis represents 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 3B. 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.



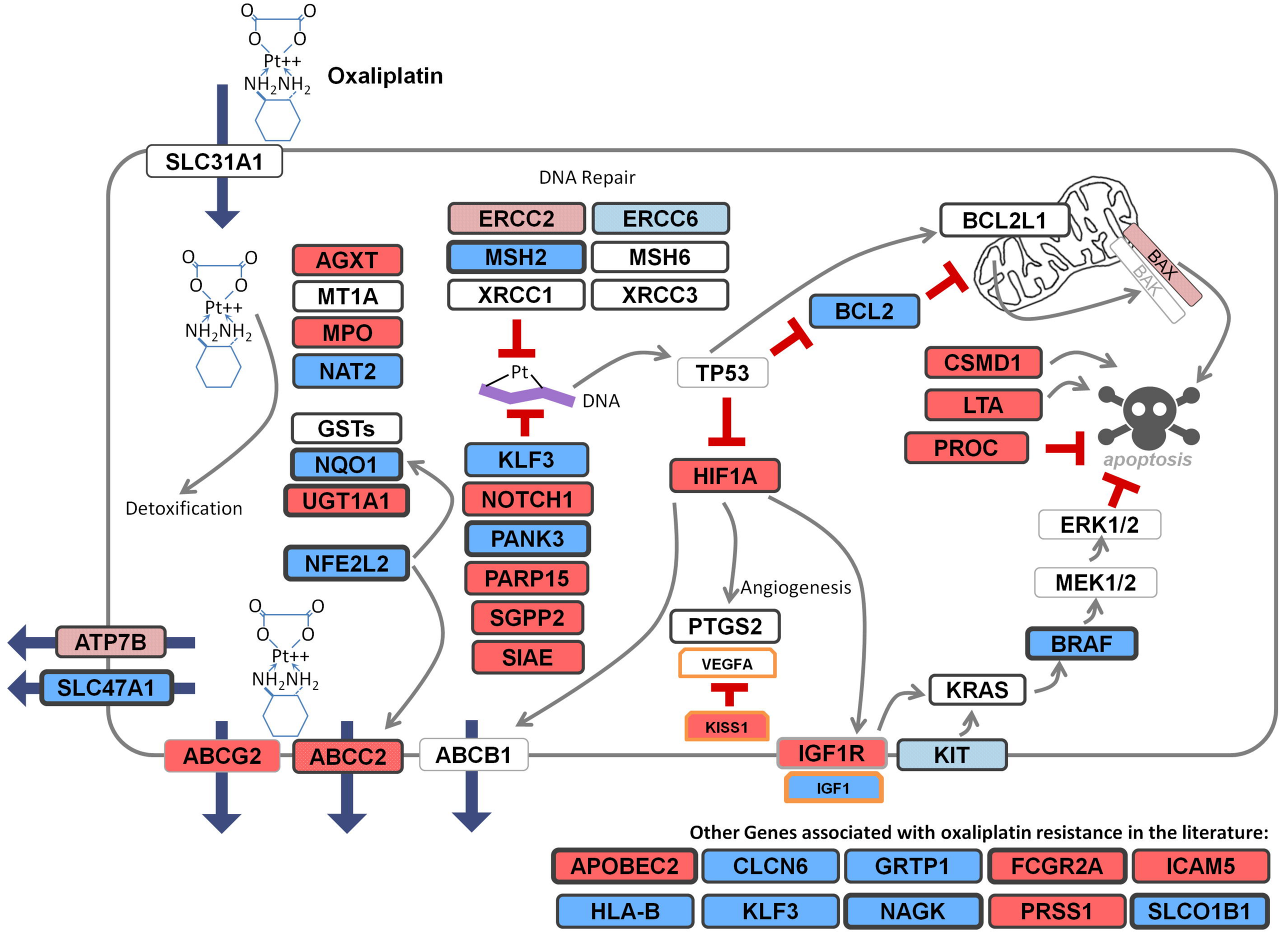
MFA:	Expression	Copy Number	Both
Higher Levels Correlate with Higher GI50 & IC50			
Lower Levels Correlate with Higher GI50 & IC50			

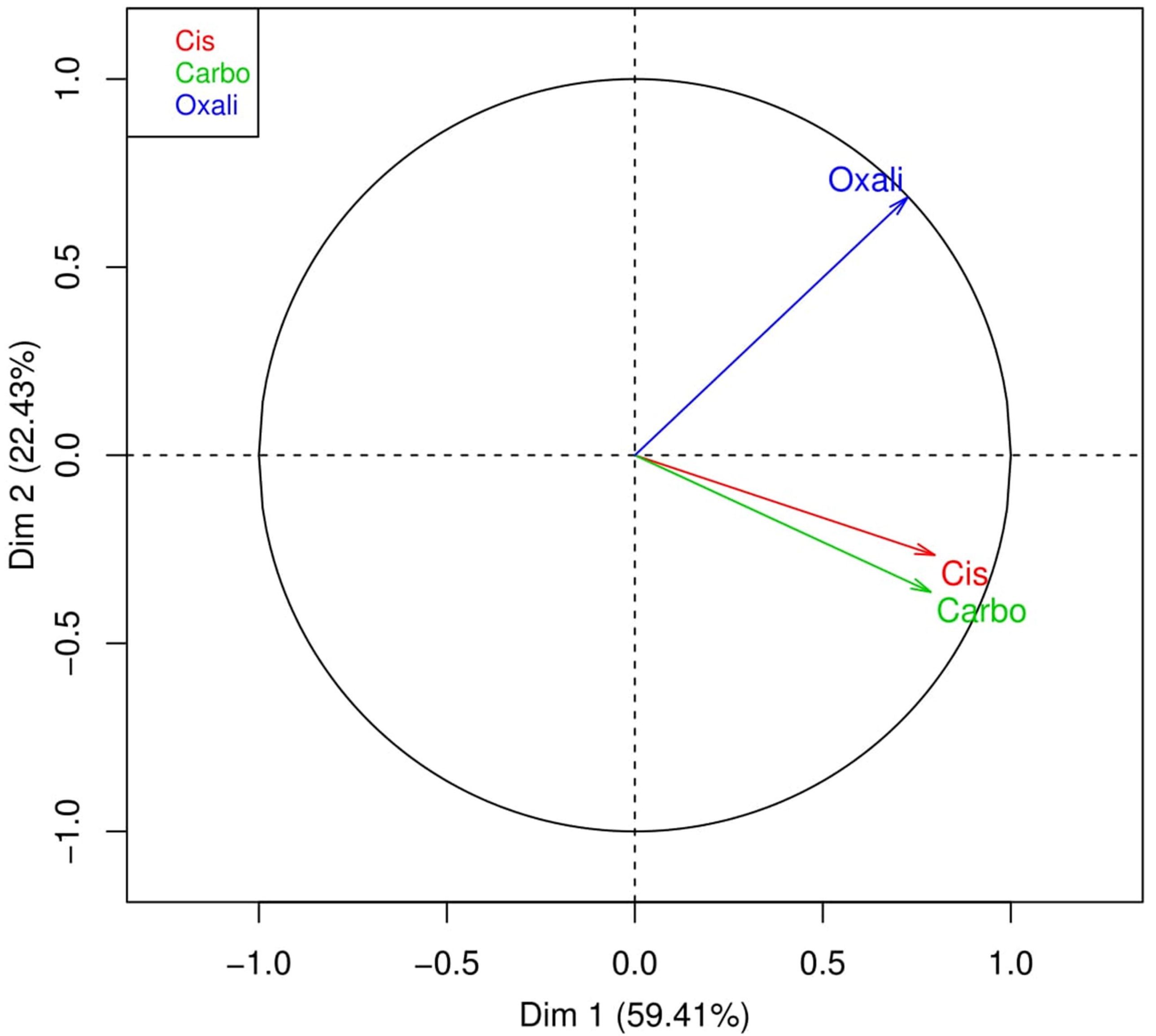
No MFA Correlation

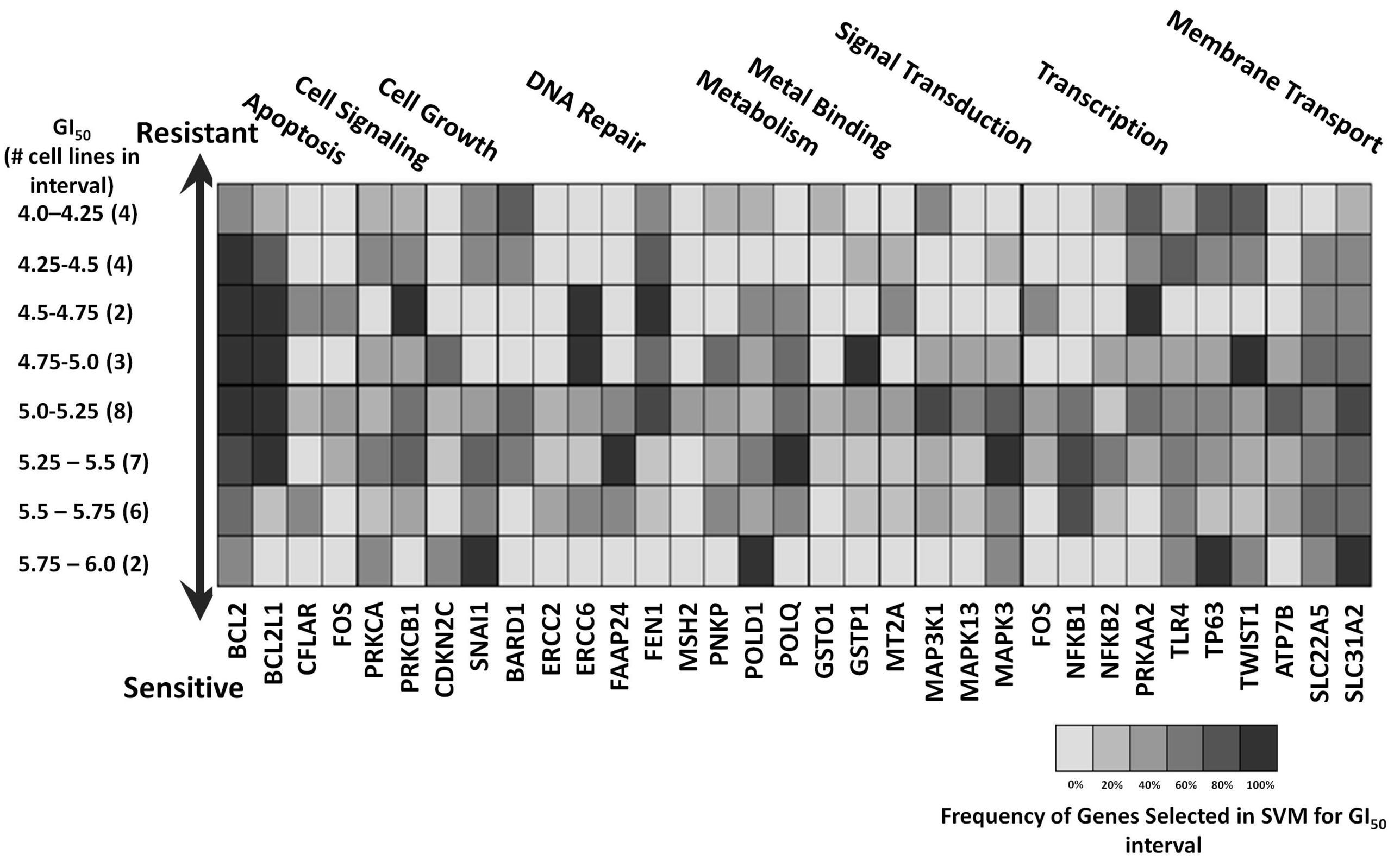


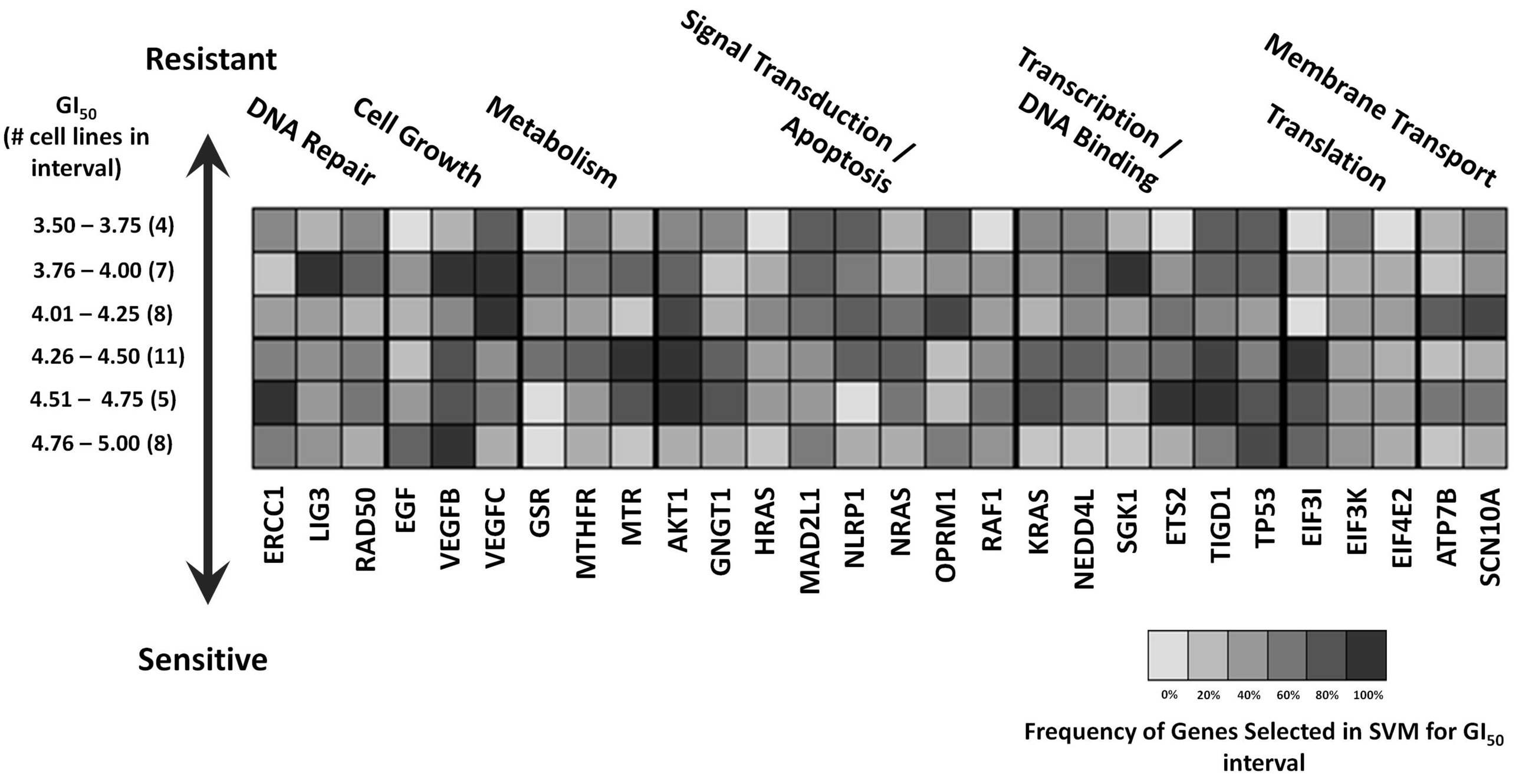
Other Genes associated with carboplatin resistance in the literature:

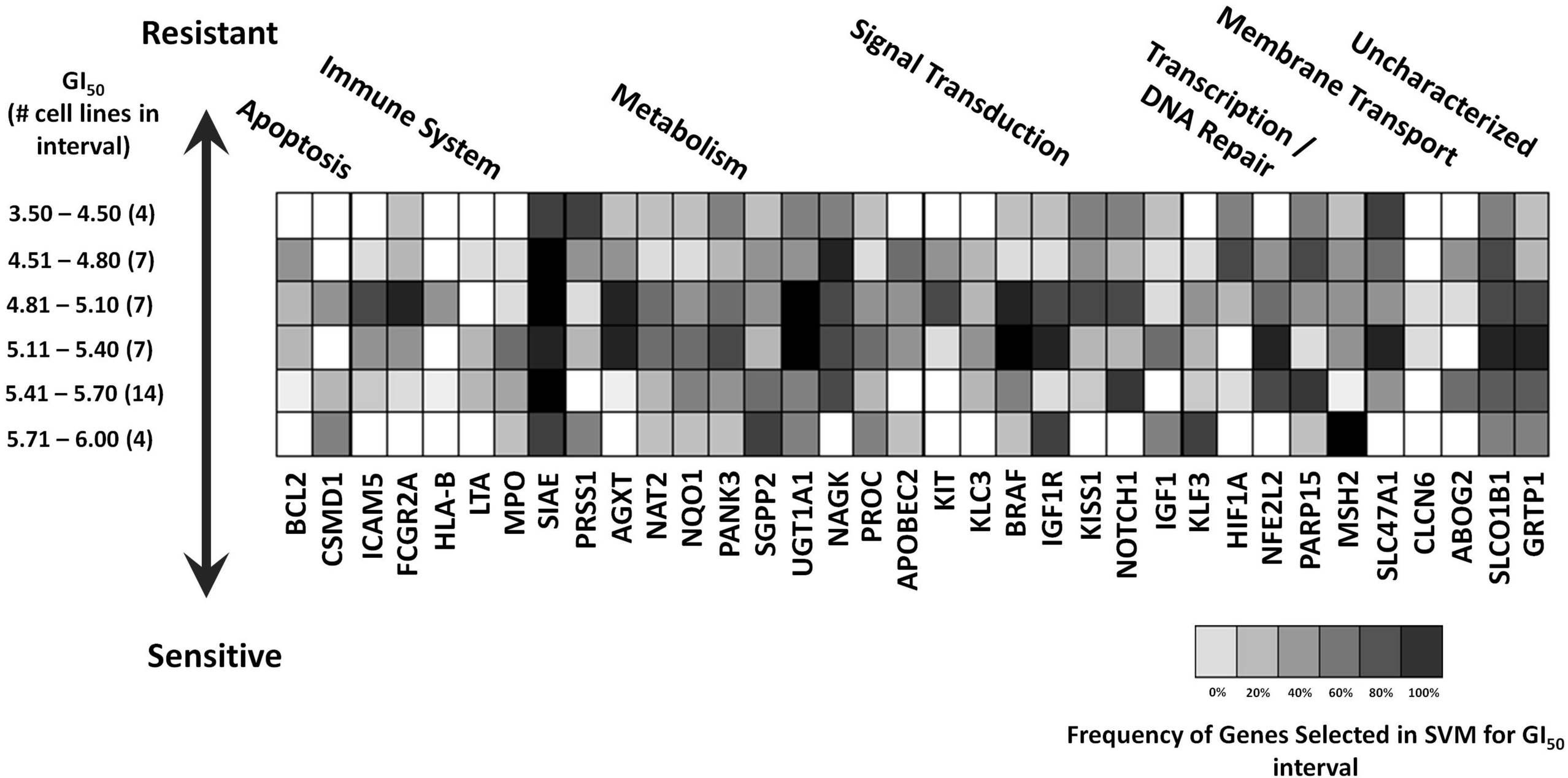
GNGT	MTR	MTHFR
OPRM1	SCN10A	TIGD1


Oxaliplatin
SLC31A1
DNA Repair
ERCC2
ERCC6
MSH2
MSH6
XRCC1
XRCC3
AGXT
MT1A
MPO
NAT2
GSTs
NQO1
UGT1A1
Detoxification
NFE2L2
Pt
DNA
KLF3
NOTCH1
PANK3
PARP15
SGPP2
SIAE
TP53
HIF1A
BCL2
BCL2L1
BAX
BAK
CSMD1
LTA
PROC
apoptosis
ERK1/2
MEK1/2
BRAF
KRAS
KIT
IGF1R
IGF1
Angiogenesis
PTGS2
VEGFA
KISS1
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


Cis
Carbo
Oxali
Oxali
Cis
Carbo
1.0
0.5
0.0
−0.5
−1.0
Dim 2 (22.43%)
−1.0
−0.5
0.0
0.5
1.0
Dim 1 (59.41%)


Apoptosis
Cell Signaling
Cell Growth
DNA Repair
Metabolism
Metal Binding
Signal Transduction
Transcription
Membrane Transport
GI50 (# 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
PRKCB1
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
0%
20%
40%
60%
80%
100%
Frequency of Genes Selected in SVM for GI50 interval


Resistant
Sensitive
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)
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


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
ABOG2
SLCO1B1
GRTP1
0%
20%
40%
60%
80%
100%
Frequency of Genes Selected in SVM for GI50 interval