

## SUPPLEMENTARY MATERIALS

### **Development of the “Co-eXpression Extrapolation” (COXEN) algorithm**

#### ***Drug activity and transcript expression profile data (Steps 1, 2, and 4)***

Publicly available drug sensitivity data, expressed in terms of 50% growth inhibition (GI50) for the NCI-60, were obtained from the NCI DTP web site, <http://dtp.nci.nih.gov>. For this study two independent sets of NCI-60 transcript expression profiling on HG-U133A GeneChip® arrays (Affymetrix, Santa Clara, CA, USA) were used, one from Novartis, Inc. (<http://wombat.gnf.org/index.html>) and the other from a collaboration of the Genomics & Bioinformatics Group at the National Cancer Institute (NCI) with Gene Logic, Inc. (Gaithersburg, MD, U.S.A.). BLA-40 transcript expression data were obtained using the HG-U133A chips (22). The BLA-40 cells and their growth conditions have been described previously and detailed in Supplementary Materials (22). We obtained and organized publicly available gene expression profiles for the clinical breast cancers, including HG-U95Av2 GeneChip® data for the 24 docetaxel trial (DOC-24) patients and 22,575-gene customized cDNA array data for the 60 tamoxifen trial (TAM-60) patients. We performed quality control checks on the Affymetrix array data for the NCI-60 and breast cancer patients and then analyzed them using the RMA algorithm(23) to obtain expression levels. We matched the gene identities of the cDNA array data with those of the HG-U133A arrays using annotation information such as gene accession numbers provided in the original study. All the microarray data were standardized (by subtracting the mean and dividing by standard deviation for each gene) within each set prior to our COXEN analysis, including log-ratio expression data from the cDNA microarrays in order to compare each gene’s expression values across different platforms.

#### ***Identification of candidate “chemosensitivity biomarkers” in the NCI-60 panel (Step 3)***

For each compound in the public NCI-60 drug database, we identified the 12 most sensitive and 12 most resistant NCI-60 cell lines. Using slightly different percent cutoffs did not change the ultimate results appreciably (data not shown). For concreteness in describing the COXEN algorithm and its results, we use the examples of cisplatin and paclitaxel, two drugs commonly used for clinical treatment of bladder

cancer (24). After selection of NCI-60 cells sensitive and resistant to each of the two compounds, we used the “Significance Analysis of Microarrays” (SAM) (13) test with false discovery rate (FDR) 0.1 to identify microarray probe sets expressed differentially between the two cell subsets. That procedure identified 191 probe sets for cisplatin and 105 for paclitaxel. Those probe sets can be thought of as candidate “chemosensitivity biomarkers” based on the NCI-60 data.

#### ***Identification of co-expression extrapolation signatures (Step 5)***

Starting with the set of candidate chemosensitivity probes for a given compound, we next identified a subset of those probes that showed concordant co-expression relationships between the NCI-60 and BLA-40 cancer cell line panels (25). In this step we tried to identify the common gene networks that are similarly coexpressed due to similar co-regulation between two independent profiled populations (i.e., NCI-60 vs. BLA-40 and NCI-60 vs. DOC-24 or TAM-60). Note that this step did not use any chemosensitivity information of the second set (i.e. BLA-40, DOC-24, or TAM-60) (**Figure S1A and detailed in Supplementary Materials**). For example, applying the procedure to the 191- and 105-probe sets for cisplatin and paclitaxel (**Figure 1A**), 18 and 13 probes, respectively, showed statistical significance (at  $p < 0.02$  one-tailed correlation distribution), (**Supplementary Table S1**).

#### ***Development of chemosensitivity prediction models for the NCI-60 panel (Step 6)***

After candidate biomarker probes were identified for each tested compound on the basis of significant differential expression for drug sensitivity in the NCI-60 and high co-expression between the NCI-60 and each of the target sets, we next searched among those candidate biomarkers for ones that would form optimal parsimonious models for prediction of the compound’s activity. For that purpose, we used the “Misclassification-Penalized Posterior” (MiPP) algorithm, which we introduced previously and summarize in Supplementary Materials (14). Note that in this chemosensitivity prediction discovery step, model training was performed strictly with the NCI-60, and the predictions for all target sets were made prospectively, purely based on the NCI-60-trained models. For the cross-platform applications of our prediction models to e.g., U95A and cDNA arrays, each gene’s intensity values (or ratio statistics of

cDNA arrays) were standardized prior to COXEN modeling. We found that standardization enabled us to apply the models derived from NCI-60 U133A data to any of the different platforms used in our current study. The main classification algorithm used in our current study was linear discriminant analysis (LDA), a widely-used classical method, which we applied under the MiPP algorithm for rigorous model training and validation and for model averaging across the several most parsimonious prediction models.

### **NCI-60 panel and drug activity data**

The NCI-60 panel consists of 60 cancer cell lines representing different types of human cancer including breast, colon, central nervous system, leukemia, non-small cell lung, melanoma, ovarian, prostate, and renal. The *in vitro* drug screening potency data on the NCI-60 provide information-rich pharmacological profiles of the compounds in terms of 60 potency values for each compound. The GI50 (50% growth inhibitory concentration) for each compound in each cell line is calculated from five-point dose-response curves with duplicate wells and positive and negative controls.. The assay uses sulforhodamine B as a measure of the amount of protein present in the well after an *in vitro* 48 hr microtiter plate assay (see <http://dtp.nci.nih.gov>). For this study we used the public NCI-60 drug potency database updated in September 2005. It comprises log(GI50) values on 45,545 compounds.

### **NCI-60 expression profiling**

For this study two independent sets of NCI-60 transcript expression profiling on HG-U133A GeneChip® arrays (Affymetrix, Santa Clara, CA, USA) were used. One was from Novartis, Inc. (<http://wombat.gnf.org/index.html>), and the other from a collaboration of the National Cancer Institute (NCI) with GeneLogic, Inc. (Gaithersburg, MD, U.S.A.), which is available to investigators from the NCI Genomics & Bioinformatics Group. It is planned that the Gene Logic data set will be made publicly available (J.W. personal communication). The protocols for cell culture, cell harvests, and RNA purification, and microarray studies have been described in detail elsewhere (Shankavaram, et al., *Molec. Cancer Therapeutics*, in press). Briefly, seed cultures of the 60 cell lines were drawn from aliquoted stocks, passaged once in T-162 flasks, and monitored frequently for degree of confluence. The

medium was RPMI-1640 with phenol red, 2 mM glutamine, and 5% fetal bovine serum. For compatibility with our other profiling studies, all fetal bovine serum was obtained from the same large batches as were used by DTP for the drug screen. One day before harvest, the cells were re-fed. Attached cells were harvested at ~80% confluence, as assessed for each flask by phase microscopy. Suspended cells were harvested at  $\sim 0.5 \times 10^6$  cells/mL. In pilot studies, samples of medium showed no appreciable change in pH between re-feeding and harvest, and no color change in the medium was seen in any of the flasks harvested. The time from incubator to stabilization of the preparation was kept to <1 min. Total RNA was purified using the Qiagen (Valencia, CA) RNeasy Midi Kit according to manufacturer's instructions. The RNA was then quantitated spectrophotometrically and aliquoted for storage at  $-80^\circ\text{C}$ . The samples were labeled and hybridized to HG-U133A GeneChip<sup>®</sup> microarrays according to standard procedures by GeneLogic, Inc. (<http://www.genelogic.com/docs/pdfs/dataGenProductSheet.pdf>).

#### **BLA-40 expression profiling and drug activity assay**

We recently collected 40 commonly used human bladder cancer cell lines, here designated the “BLA-40 cell panel.” For activity assays on the BLA-40, cisplatin (Sigma, St. Louis, MO) was dissolved in Dulbecco's phosphate-buffered saline and aliquoted in 1-mg/ml stocks. Paclitaxel (Sigma) was dissolved in dimethyl sulphoxide and aliquoted in 1-mM stocks. Gemcitabine (University of Virginia Medical Center Pharmacy) was dissolved in PBS and aliquoted in 0.1-M stocks. All cell lines except CRL2169 (SW780) were maintained in appropriate medium in a humidified atmosphere containing 5 % CO<sub>2</sub> in air. CRL2169 (SW780) was cultured without CO<sub>2</sub>, as required for its growth. Cell lines were subcultured in an aqueous solution of 0.05% trypsin (Difco, 1:250) and 0.016% EDTA. Cell lines were seeded in 96-well cell culture plates (Costar) at a density of 1000 cells/well. 24 hours later, the cells were exposed to drug diluted in RPMI-1640 medium containing 10% FBS, at a total volume of 200  $\mu\text{L}$ . Each drug dose was plated in triplicate, and the experiment was repeated four to seven times. The doses for Cisplatin were 200, 400, 800, 1600, 3200, and 6400 ng/ml; for Paclitaxel, 0.1, 1, 2, 5, 10, and 100 nM. After incubation with medium or drug, growth inhibition was assessed using reduction of the fluorescence of Alamar Blue aqueous dye (<http://dtp.nci.nih.gov>) with excitation at 545 nm and emission at 590 nm. Cisplatin and

Paclitaxel were purchased from Sigma (St. Louis, MO). The cell lines were seeded in 96-well plates (Costar) at a density of 1000 cells/well. 24 hours later, the cells were exposed to drug at various concentrations in triplicate, and the experiments were independently replicated two or three times. After 72 hours, growth inhibition was assessed using Alamar Blue aqueous dye (<http://dtp.nci.nih.gov>) (BioSource International, Inc., Camarillo, CA). Expression profiling for the BLA-40 was done using HG-U133A arrays on duplicate samples generated from independent cell cultures as described previously. When the image files of the NCI-60 and BLA-40 cell lines passed quality-control checks, they were analyzed using the RMA analysis software for GeneChip<sup>®</sup> data to obtain expression levels.

### **Hierarchical clustering**

To examine the overall expression patterns of the co-expression extrapolation probe sets, we co-clustered (1) the NCI-60 data with BLA-40 data and co-clustered the NCI-60 data with clinical breast cancer data. As shown in **Figure 2B** for cisplatin, the NCI-60 and BLA-40 the cells clustered largely according to their sensitivity or resistance, not according to their organ of origin or whether they were from the NCI-60 or BLA-40 panel. That visual result strongly indicates that the probes selected to form the co-expression signature are better markers for response to cisplatin than they are to the other variables, such as histological subtype for example. In stark contrast, the NCI-60 and BLA-40 cell types still separated almost completely, irrespective of cisplatin response, when they were hierarchically clustered on the basis of probe profiles not selected with relation to co-expression extrapolation, but only to drug sensitivity. Results similar to those in **Figure 2A,B** were obtained for paclitaxel on the BLA-40 cell lines (**Figures S2A-B**) and the docetaxel clinical trial on the DOC-24 patients (**Figures S4A-B**).

### **Identification of co-expression extrapolation signatures (Step 5)**

To parameterize each probe's co-expression relationships between two studies mathematically, we calculated a "co-expression extrapolation coefficient (CEEC),"  $rc(j)$ , for each probe  $j$  as follows: Using the probe expression data, we constructed two correlation matrices (of dimension  $n \times n$ ) for the set of  $n$  candidate chemosensitivity probes. The two correlation matrices, one for the NCI-60, the other for the

BLA-40, were evaluated as  $U = [U_{ij}]_{n \times n}$  and  $V = [V_{ij}]_{n \times n}$ , where  $U_{ij}$  and  $V_{ij}$  are the correlation coefficients between probes  $i$  and  $j$  in the NCI-60 and BLA-40, respectively. Then,  $rc(j)$  is defined as

$$rc(j) = \frac{\sum_{k=1}^n (U_{kj} - \bar{U}_k)(V_{kj} - \bar{V}_k)}{\sqrt{\sum_{k=1}^n (U_{kj} - \bar{U}_k)^2} \sqrt{\sum_{k=1}^n (V_{kj} - \bar{V}_k)^2}}$$

where  $\bar{U}_k$  and  $\bar{V}_k$  are the mean correlation coefficients of the row- $k$  correlation coefficient vectors for the NCI-60 and BLA-40.  $rc$  reflects the degree of co-expression extrapolation of probe  $j$  with the set of  $n$  probes between the NCI-60 and BLA-40 cell lines. If  $rc(j)$  exceeded a cut-off criterion (e.g., 98th percentile of the corresponding random distribution generated by randomly shuffling the probe identities between the two sets), probe  $j$  was selected as a probe for co-expression extrapolation between the two panels. Since probe  $j$  was selected from the set of  $n$  candidate chemosensitivity predictors, it had that pharmacological characteristic as well. Note that CEEC will be high if a probe's co-expression network relationships with the other probes on the first set (i.e. NCI-60) are concordant with those of the second set (i.e. BLA-40).

### **Development of chemosensitivity prediction models for the NCI-60 panel (Step 6)**

We searched among those candidate biomarkers obtained from steps 1-5 for ones that would form optimal parsimonious models for prediction of the compound's activity. For that purpose, we used the "Misclassification-Penalized Posterior" (MiPP) algorithm, which we introduced previously (2). In brief, MiPP is based on stepwise incremental classification modeling for discovery of the most parsimonious prediction models. It includes double cross-validated evaluation for each trained prediction model. Model training can be performed using any of several different classification algorithm such as linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), support vector machine (SVMs) learning, or logistic regression. In the current study we used LDA for most of the applications. In double cross-validation, the first cross-validation is based on random splitting of the whole data set into a training set and an independent test set for external model validation; the second is an  $n$ -fold cross-

validation on the training set to avoid the pitfalls of a large-screening search and to obtain the most parsimonious optimal prediction models. Independent splits of the data result in multiple prediction models. The multiple models are then re-evaluated using a large number (e.g., 1000) of random splits of test and training sets to obtain confidence bounds on the accuracy of prediction. On the basis of those confidence bounds, the prediction performance and mean misclassification error rates (ER) are obtained for each of the candidate prediction models. The final prediction of a cell line as “sensitive” or “resistant” is determined by the cell’s (posterior) classification probability (CP) from each LDA prediction model. If  $CP > 0.5$  based on the top three to five prediction models, the cell line is considered sensitive; if not, it is considered resistant. We found that optimal MiPP prediction models were often obtained with only a small number of probes, e.g. three or four probes, among the ones identified in Step 5. The original publications provide technical details and comprehensive descriptions. (2, 3) For the cross-platform application of our prediction models to e.g., U95A and cDNA arrays, each gene’s intensity values (or red-green channel ratios for cDNA arrays) were standardized prior to our COXEN modeling. Standardization was required for applicability of models derived from the NCI-60 U133A data to any of the other platforms used in the study.

### **Functional Relationships among Genes in COXEN models**

To explore the functional relationships among the genes with altered expression in the COXEN models we used Ingenuity Pathway Analysis (IPA)<sup>™</sup> (Ingenuity, Inc., Mountain View, CA, USA).

**SUPPLEMENTARY TABLES**

**Table S1. Co-expression extrapolation signatures for prediction of cisplatin and paclitaxel activity in the BLA-40 panel.** Eighteen probes for cisplatin and 13 for paclitaxel identified on the basis of statistically significant differential expression between NCI-60 sensitive and resistant cell lines and high co-expression extrapolation coefficients between the NCI-60 and BLA-40 cell line panels. Desmoplakin (DSP) has been reported as a biomarker for meningiomas (4). Loss of claudin 4 (CLDN4) has been observed in breast and other cancers (5). TACSTD1 is tumor-associated calcium signal transducer. Ladinin (1 (LAD1) is found to be one of the most highly expressed genes in ovarian serous papillary carcinomas (OSPCs) (6). Periplakin (PPL) was recently reported to be significantly downregulated in esophageal cancer (7), COXEN-identified genes for paclitaxel, include defective in sister chromatid cohesion homolog 1 (DCC1) required in efficient repair of yeast cell damage for the collision between a topoisomerase I-DNA intermediate and an advancing replication fork (8). The deficiency of dyskeratosis congenita 1 (DKC1) was found to result in multiple abnormalities including premature ageing, bone marrow failure and cancer (9).

Affymetrix ID	Gene symbol	Locus ID	Gene acc. number	Description
<b>Cisplatin</b>				
200606_at	DSP	1832	NM_004415	Desmoplakin
201428_at	CLDN4	1364	NM_001305	claudin 4
201839_s_at	TACSTD1	4072	NM_002354	tumor-associated calcium signal transducer 1
203287_at	LAD1	3898	NM_005558	ladinin 1
203407_at	PPL	5493	NM_002705	Periplakin
203713_s_at	LLGL2	3993	NM_004524	lethal giant larvae homolog 2 (Drosophila)
205709_s_at	CDS1	1040	NM_001263	CDP-diacylglycerol synthase1
206722_s_at	EDG4	9170	NM_004720	lysophosphatidic acid G-protein-coupled receptor, 4
209873_s_at	PKP3	11187	AF053719	Plakophilin 3
210058_at	MAPK13	5603	BC000433	mitogen-activated protein kinase 13
210059_s_at	MAPK13	5603	BC000433	mitogen-activated protein kinase 13
210480_s_at	MYO6	4646	U90236	myosin VI
210761_s_at	GRB7	2886	AB008790	growth factor receptor-bound protein 7
218780_at	HOOK2	29911	NM_013312	hook homolog 2 (Drosophila)
218966_at	MYO5C	55930	NM_018728	myosin VC
219395_at	RBM35B	80004	NM_024939	RNA binding motif protein 35A
219513_s_at	SH2D3A	10045	NM_005490	SH2 domain containing 3A
31846_at	RHOD	29984	AW003733	ras homolog gene family, member D
<b>Paclitaxel</b>				
201478_s_at	DKC1	1736	U59151	dyskeratosis congenita 1, dyskerin
201479_at	DKC1	1736	NM_001363	dyskeratosis congenita 1, dyskerin
203221_at	TLE1	7088	AI758763	Transducin-like enhancer of split 1
203625_x_at	SKP2	6502	BG105365	S-phase kinase-associated protein 2 (p45)
203895_at	PLCB4	5332	AL535113	phospholipase C, beta 4
203896_s_at	PLCB4	5332	NM_000933	phospholipase C, beta 4
204767_s_at	FEN1	2237	BC000323	flap structure-specific endonuclease 1
204768_s_at	FEN1	2237	NM_004111	flap structure-specific endonuclease 1
209654_at	KIAA0947	23379	BC004902	NA
211651_s_at	LAMB1	3912	M20206	laminin, beta 1
213918_s_at	NIPBL	25836	BF221673	Nipped-B homolog (Drosophila)
218979_at	C9orf76	80010	NM_024945	chromosome 9 open reading frame 76
219000_s_at	DCC1	79075	NM_024094	defective in sister chromatid cohesion homolog 1

**Table S2. Co-expression extrapolation signatures for prediction of paclitaxel (DOC-24) and tamoxifen (TAM-60) activity in clinical breast cancer.** Fourteen probe sets for paclitaxel and eight for tamoxifen identified on the basis of statistically significant differential expression between sensitive and resistant NCI-60 cell lines and high co-expression extrapolation coefficients between the NCI-60 and each of the two clinical samples. Sucrase-isomaltase (SI) gene is an enterocyte-specific differentiation marker which is distinctly downregulated after methotrexate treatment (10). The expression of metallothionein proteins including 1X (MT1X) was found to be overexpressed in human pancreatic cancer drug-resistant cell lines (11). stress-associated endoplasmic reticulum protein 1 (SERP1) was reported to be suppressed in papillary thyroid cancer (12). Causal mutations in programmed cell death 10 (PDCD10) have been frequently demonstrated in various cancers (13). Defects in translational regulation by eukaryotic initiation factor 2 (eIF2) alpha may represent a common hallmark of tumorigenesis (14). COXEN-identified genes differentially expressed between responder and nonresponder TAM-60 patients treated with tamoxifen include probe 207881\_at (gene accession number AF050199) , a novel gene factor whose function has not been reported.

Affymetrix ID	Gene symbol	Locus ID	Gene acc. Number	Description
<b>Paclitaxel*</b>				
211915_s_at	TUBB4Q	56604	U83110	tubulin, beta polypeptide 4, member Q
216022_at	WNK1	65125	AL049278	WNK lysine deficient protein kinase 1
208387_s_at	MMP24	10893	NM_006690	matrix metalloproteinase 24 (membrane-inserted)
202312_s_at	COL1A1	1277	NM_000088	collagen, type I, alpha 1
210738_s_at	SLC4A4	8671	AF011390	solute carrier family 4
214133_at	MUC6	4588	AI611214	mucin 6, gastric
209995_s_at	TCL1A	8115	BC003574	T-cell leukemia/lymphoma 1A
214589_at	FGF12	2257	AL119322	fibroblast growth factor 12
209552_at	PAX8	7849	BC001060	paired box gene 8
204505_s_at	EPB49	2039	NM_001978	erythrocyte membrane protein band 4.9 (dematin)
212974_at	DENND3	22898	AI808958	DENN/MADD domain containing 3
215904_at	MLLT4	4301	AL049698	myeloid/lymphoid or mixed-lineage leukemia
213560_at	GADD45B	4616	AV658684	growth arrest and DNA-damage-inducible, beta
211886_s_at	TBX5	6910	U80987	T-box 5
<b>Tamoxifen</b>				
200970_s_at	SERP1	27230	AL136807	Stress-associated endoplasmic reticulum (ER) protein 1
201632_at	EIF2B1	1967	NM_001414	eukaryotic translation initiation factor 2B, subunit 1 alpha
204326_x_at	MT1L	4500	NM_002450	metallothionein 1L
206664_at	SI	6476	NM_001041	sucrase-isomaltase (alpha-glucosidase)
208581_x_at	MT1X	4501	NM_005952	metallothionein 1X
208869_s_at	GABARAPL1	23710	AF087847	GABA(A) receptor-associated protein like 1
210907_s_at	PDCD10	11235	BC002506	programmed cell death 10
212730_at	DMN	23336	AK026420	Desmuslin

\* The reason for using paclitaxel instead of docetaxel is explained in the text

**Table S3. Prospective evaluation of the top three MiPP classification models for in vitro prediction of chemosensitivity of bladder cancer cell lines to cisplatin and paclitaxel .** The top three MiPP models, trained strictly on the NCI-60 cisplatin and paclitaxel responses, were evaluated for their ability to predict the sensitivity of independent BLA-40 cell lines to cisplatin and paclitaxel. Predictions were confirmed prospectively in independent experimental assays of the two drugs on the BLA-40 panel.

	<b>Sensitive*</b> <b>N=10</b>	<b>Resistant*</b> <b>N=10</b>	<b>Overall</b> <b>N=20</b>	<b>Overall</b> <b>(p-value**)</b>
<b><i>Cisplatin</i></b>				
Model 1	9/10	8/10	85% (17/20)	0.002
Model 2	9/10	8/10	85% (17/20)	0.002
Model 3	9/10	8/10	85% (17/20)	0.002
<b><i>Paclitaxel</i></b>				
Model 1	8/10	8/10	80% (16/20)	0.012
Model 2	9/10	7/10	80% (16/20)	0.012
Model 3	8/10	7/10	75% (15/20)	0.041

\* Denominator: the 10 BLA-40 cell lines most or least sensitive to the drug (see **Figures S6C** and **S6D** for the cell identified in the cases of cisplatin and paclitaxel, respectively). Numerator: Number of the denominator cells correctly classified as sensitive or resistant to the drug.

\*\*From the exact binomial test with the null hypothesis that prediction is based on a random coin tossing.

**Table S4. Evaluation of predictive performance of the top three MiPP classification models for response of breast cancer patients in the docetaxel (DOC-24) and tamoxifen (TAM-60) trials.** The top three MiPP models, trained strictly on the NCI-60 drug responses, were evaluated for the prediction of chemotherapeutic responses of DOC-24 and TAM-60. Predictions were confirmed prospectively by comparing actual pathologic tumor regression and disease-free survival time data.

<b>Docetaxel</b>	<b>Responder* N=11</b>	<b>Nonresponder* N=13</b>	<b>Overall N=24</b>	<b>Overall (p-value**)</b>
Model 1	10/11	8/13	75% (18/24)	0.022
Model 2	11/11	7/13	75% (18/24)	0.022
Model 3	10/11	8/13	75% (18/24)	0.022
<b>Tamoxifen</b>	<b>Responder^ N=11</b>	<b>Nonresponder^ N=16</b>	<b>Overall N=27</b>	<b>Overall (p-value**)</b>
Model 1	7/11	13/16	74% (20/27)	0.019
Model 2	6/11	13/16	70% (19/27)	0.052
Model 3	7/11	12/16	70% (19/27)	0.052

\* Fraction of patients correctly classified according to outcome reported in the original study (15)

^ Fraction correctly classified according to criteria shown in **Figure S5A** and described in Results.

\*\* From the exact binomial test with the null hypothesis that prediction is based on a random coin tossing.

## SUPPLEMENTARY FIGURES

**Figure S1: (A) Schematic illustration of co-expression extrapolation.** In this artificial five-probe example, Probes 1 and 3 in Cell Set 1 (e.g., the NCI-60) show essentially the same patterns of co-expression correlation with other probes as do Probes 1 and 3 in Cell Set 2 (e.g., the BLA-40). Probes 2, 4, and 5 show different patterns of co-expression correlation in the two Cell Sets. Therefore, Probes 1 and 3 (but not 2, 4, and 5) might be selected by the “co-expression extrapolation” algorithm (Step 5) for inclusion in the prediction signature for Step 6. Note: The co-expression correlations here are those calculated across cell types for a given pair of probes. **Prospective evaluation of models for prediction of chemotherapeutic response in the BLA-40 panel.** Performance of the top three MiPP prediction models is shown based on continuous, predicted classification probabilities on the BLA-40 cells. **(B)** on the BLA-40 sensitive and resistant cell lines for cisplatin. Red and blue bars indicate sensitive and resistant bladder cell lines respectively, based on the level of growth inhibition in response to drug in vitro. A cell line with the larger classification probability implies a higher probability as being a predicted sensitive (i.e., predicted as sensitive if its classification probability (CP)  $>0.5$  (dotted line)). **(C)** on the BLA-40 sensitive and resistant cell lines for paclitaxel. **(D)** Direct comparison between COXEN prediction scores and experimentally measured cisplatin activities in the BLA-40 cell lines. The activity here and elsewhere was expressed as  $\log(\text{GI}_{50})$ , where  $\text{GI}_{50}$  is the drug concentration leading to 50% growth inhibition of cells compared with control. The cell lines were ordered in the figure based on their  $\log(\text{GI}_{50})$  values. COXEN scores and  $\text{GI}_{50}$  values were standardized by subtracting the overall mean and dividing by the standard deviation across the BLA-40. The statistical significance of the set of predictions (two-tailed  $p$ -value=0.016) on all 40 cells of the BLA-40 was assessed by Spearman correlation.

**Figure S2. Co-clustering Clustered Image Maps (CIMs) for chemosensitivity signature probes and for COXEN signature probes in the NCI-60 and BLA-40 cell panels.** (A) Co-clustering of the NCI-60 and BLA-40 cell lines using the first 50 probes of the entire differentially expressed chemosensitivity

probe sets for paclitaxel. Red, black, and green indicate high, intermediate, and low expression, respectively. Red and blue in the upper bar indicates sensitive cells and resistant cells, respectively. Yellow and cyan in the lower bar indicate NCI-60 and the BLA-40 lines, respectively. Most cell lines clustered based on their panel of origin (NCI-60 or BLA-40), not on the basis of sensitivity or resistance. **(B)** CIM for the NCI-60 and the BLA-40 cell lines using the final 13 COXEN probes for paclitaxel after Step 5 in the COXEN algorithm. Cells clustered primarily on the basis of sensitivity and resistance, rather than cell panel. **(C-E)** Illustrated is the top-scoring pathway as defined by the Ingenuity analysis tool (details given in text). Each pathway member is depicted by a symbol. Red symbols indicate those genes with down-regulated expression, green represents the genes with increased expression in the analysis, white symbols identifies pathway members not found altered in the tumor cells. **(C)** Ingenuity generated interaction pathways of the identified COXEN biomarkers of response for the DOC-24 breast clinical trial of docetaxel. **(D)** Ingenuity generated interaction pathways of the identified COXEN biomarkers of response for the human bladder cancer cell lines (BLA-40) to paclitaxel. **(E)** Ingenuity generated interaction pathways of the identified COXEN biomarkers of response for the human bladder cancer cell lines (BLA-40) to cisplatin.

**Figure S3: Models for prediction of chemotherapeutic response in the in the docetaxel (DOC-24) (15, 16) breast cancer trial:** Performance of the top three MiPP prediction models is shown based on continuous, predicted classification probabilities on patients in the docetaxel trial (DOC-24) . Red and blue bars indicate responder and non-responder patients of the DOC-24, respectively, based on the chemotherapeutic responses (i.e. residual tumor sizes) reported in the original study.. A patient with the larger posterior classification probability implies a higher probability as being a predicted responder (i.e., predicted as responder if its classification probability (CP) >0.5 (dotted line)).

**Figure S4. Co-clustering Clustered Image Maps (CIMs) for chemosensitivity signature probes and for COXEN signature probes in the NCI-60 and breast cancer clinical trials (DOC-24) (15, 16): (A)** Co-clustering of the NCI-60 and DOC-24 patients using the first 50 probes of the entire differentially

expressed chemosensitivity probe sets for paclitaxel. Red, black, and green indicate high, intermediate, and low expression, respectively. Red and blue in the upper bar indicates sensitive cells and resistant cells, respectively. Yellow and cyan in the lower bar indicate NCI-60 and the DOC-24 patients, respectively. Most cell lines clustered based on their panel of origin (NCI-60 or DOC-24 patients), not on the basis of sensitivity or resistance. **(B)** The same as S4A but using the 14 COXEN Step 5 probes for DOC-24 trial. Cells clustered primarily on the basis of sensitivity and resistance, rather than cell panel.

**Figure S5. (A) Classification of responder and non-responder patients in the tamoxifen (TAM-60) (15, 16) trial:** Patients with recurrent disease recurred within a relatively short time (<50 months) after the tamoxifen treatment, whereas no patient with durable survival did so in that time period. Hence, we made the assumption (presumably only partially true) that the early-recurrence patients were tamoxifen non-responders (16 patients), whereas the patients with long-term survival (>130 months) were considered responders (11 patients). **(B) Models for prediction of chemotherapeutic response in the tamoxifen (TAM-60) trial:** Performance of the top three MiPP prediction models based on continuous, predicted classification probabilities is shown on responder and non-responder patients of the TAM-60 trial. Red and blue bars indicate responder and non-responder patients of the TAM-60, respectively, based on the determination of responders and non-responders prior to our COXEN modeling. A patient with the larger posterior classification probability implies a higher probability as being a predicted responder (i.e., predicted as responder if its classification probability (CP) >0.5 (dotted line)).

**Figure S6. *In vitro* chemosensitivity of NCI-60 and BLA-40 cell lines. (A)** The NCI-60 cell lines ordered by their sensitivity to cisplatin (i.e., on the basis of log(GI50) values). Red diamond, blue circle, and black plus symbols represent sensitive, resistant, and intermediate cell lines. Top 20% most sensitive cell lines and 20% least sensitive (or most resistant) cell lines were selected as sensitive and resistant cell lines. **(B)** Same as (A) for the NCI-60 cell lines ordered by their sensitivity to paclitaxel. Top 20% most sensitive cell lines and 8 least sensitive (or most resistant) cell lines were selected as sensitive and resistant cell lines. The selection of the resistant cell line group was optimized based on the

separation of these cell lines from the others in their GI50 values. **(C)** The BLA-40 cell lines ordered by their sensitivity to cisplatin. Top 25% most sensitive cell lines and 25% least sensitive (or most resistant) cell lines were pre-determined as sensitive and resistant cell lines before validation on the COXEN prediction models. **(D)** Same as (C) for the BLA-40 cell lines ordered by their sensitivity to paclitaxel. **(E)**. The top COXEN scoring compound identified by the automated computational screening algorithm. NSC number and chemical structures were retrieved from <http://dtp.nci.nih.gov>.

## Reference

1. Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., Scudiero, D. A., Eisen, M. B., Sausville, E. A., Pommier, Y., Botstein, D., Brown, P. O. & Weinstein, J. N. (2000) *Nat Genet* **24**, 236-44.
2. Soukup, M., Cho, H. & Lee, J. K. (2005) *Bioinformatics* **21 Suppl 1**, i423-i430.
3. Soukup, M. & Lee, J. K. (2004) *J Bioinform Comput Biol* **1**, 681-94.
4. Baia, G. S., Slocum, A. L., Hyer, J. D., Misra, A., Sehati, N., VandenBerg, S. R., Feuerstein, B. G., Deen, D. F., McDermott, M. W. & Lal, A. (2006) *J Neurooncol* **78**, 113-21.
5. Wu, X., Chen, H., Parker, B., Rubin, E., Zhu, T., Lee, J. S., Argani, P. & Sukumar, S. (2006) *Cancer Res* **66**, 9527-34.
6. Santin, A. D., Zhan, F., Bellone, S., Palmieri, M., Cane, S., Bignotti, E., Anfossi, S., Gokden, M., Dunn, D., Roman, J. J., O'Brien, T. J., Tian, E., Cannon, M. J., Shaughnessy, J., Jr. & Pecorelli, S. (2004) *Int J Cancer* **112**, 14-25.
7. Nishimori, T., Tomonaga, T., Matsushita, K., Oh-Ishi, M., Kodera, Y., Maeda, T., Nomura, F., Matsubara, H., Shimada, H. & Ochiai, T. (2006) *Proteomics* **6**, 1011-8.
8. Redon, C., Pilch, D. R. & Bonner, W. M. (2006) *Genetics* **172**, 67-76.
9. Marrone, A., Walne, A. & Dokal, I. (2005) *Curr Opin Genet Dev* **15**, 249-57.
10. de Koning, B. A., Lindenbergh-Kortleve, D. J., Pieters, R., Rings, E. H., Buller, H. A., Renes, I. B. & Einerhand, A. W. (2006) *Cancer Chemother Pharmacol* **57**, 801-10.
11. Xie, Y., Zhao, Y. P., Chen, G., Yuan, C. H. & Li, L. J. (2005) *Zhongguo Yi Xue Ke Xue Yuan Xue Bao* **27**, 619-23.
12. Puskas, L. G., Juhasz, F., Zarva, A., Hackler, L., Jr. & Farid, N. R. (2005) *Cell Mol Biol (Noisy-le-grand)* **51**, 177-86.
13. Felbor, U., Sure, U., Grimm, T. & Bertalanffy, H. (2006) *Zentralbl Neurochir* **67**, 110-6.
14. Balachandran, S. & Barber, G. N. (2004) *Cancer Cell* **5**, 51-65.
15. Chang, J. C., Wooten, E. C., Tsimelzon, A., Hilsenbeck, S. G., Gutierrez, M. C., Elledge, R., Mohsin, S., Osborne, C. K., Chamness, G. C., Allred, D. C. & O'Connell, P. (2003) *Lancet* **362**, 362-9.
16. Ma, X. J., Wang, Z., Ryan, P. D., Isakoff, S. J., Barmettler, A., Fuller, A., Muir, B., Mohapatra, G., Salunga, R., Tuggle, J. T., Tran, Y., Tran, D., Tassin, A., Amon, P., Wang, W., Enright, E., Stecker, K., Estepa-Sabal, E., Smith, B., Younger, J., Balis, U., Michaelson, J., Bhan, A., Habin, K., Baer, T. M., Brugge, J., Haber, D. A., Erlander, M. G. & Sgroi, D. C. (2004) *Cancer Cell* **5**, 607-16.