

Supplementary Information

Methods:

NCI-60 predictors. The $[-\log_{10}(M)]$ GI50/IC50, TGI (Total Growth Inhibition dose) and LC50 (50% cytotoxic dose) data was used to populate a matrix with MATLAB software, with the relevant expression data for the individual cell lines. Where multiple entries for a drug screen existed (by NCS number), the entry with the largest number of replicates was included. Incomplete data were assigned as Nan (not a number) for statistical purposes. To develop an in vitro gene expression based predictor of sensitivity/resistance from the pharmacologic data used in the NCI-60 drug screen studies, we chose cell lines within the NCI-60 panel that would represent the extremes of sensitivity to a given chemotherapeutic agent (mean GI50 \pm 1SD). Furthermore, since the TGI and LC50 dose also represent the cytostatic and cytotoxic levels of any given drug; the log transformed TGI and LC50 dose of the sensitive and resistant subsets was then correlated with the respective GI50 data to ascertain consistency between the TGI, LC50 and GI50 data. Cell lines with low GI50 (< 1 SD of mean) also needed to have a low LC50, and TGI concentration to be considered sensitive. Likewise, those with the highest GI50 (> 1 SD of mean), TGI and LC50 concentration for a given drug, were considered resistant. Our hypothesis was that such a rigorous selection would identify cell lines that represent the extremes of sensitivity to a given drug. As is seen in supplementary figure 1, the curated samples demonstrate significant differences in the means of the sensitive and resistant populations. For all correlation analyses, because the GI50 data is non-Gaussian, a variance fixed t-test was used to calculate significance in analyses of correlation. Relevant expression data (updated data available on the Affymetrix U95A2 GeneChip) for the solid tumor cell lines and the respective pharmacological data for the chemotherapeutics docetaxel, paclitaxel (Taxol), topotecan, 5-fluorouracil (5-FU), adriamycin (doxorubicin), cyclophosphamide (Cytoxan), and etoposide was downloaded from the NCI website (http://dtp.nci.nih.gov/docs/cancer/cancer_data.html). The individual drug sensitivity and

resistance data from the selected solid tumor NCI60 cell lines was then used in a supervised analysis using binary regression methodologies, as described previously²⁵, to develop models predictive of chemotherapeutic response. Also, importantly, the drug screening data on the NCI-60 panel in conjunction with matched expression data does not always yield predictors of chemotherapy response. In the approach we use, the presence of high quality dose response data is critical. For instance, there need to be enough samples representing the extremes of sensitivity to any given agent. Thus, if a drug screening experiment did not result in variable GI50 and/or LC50 data, the generation of a genomic predictor is not possible (e.g. cis/carboplatin and methotrexate).

Identification of docetaxel response data in patients. Chang and colleagues have published expression (GEO accession numbers: GSE349, GSE350, GSE360) data and objective response information to docetaxel¹⁰. Of the 24 patients reported in their study, there were 13 patients with docetaxel sensitivity and 11 patients with resistance. This dataset was used to validate the in vitro predictive model and generate a complementary in vivo model of docetaxel sensitivity using analyses described below.

Breast cancer tumor validation data. An independent dataset of expression data with corresponding clinical and pathologic response information on 51 patients treated with neoadjuvant chemotherapy (paclitaxel, 5-flourouracil, adriamycin and cyclophosphamide) was obtained (courtesy Pusztai et al) from the M.D. Anderson Cancer Center bioinformatics website. There were 13 responders (complete pathologic response) and 38 non-responders. All gene array data was on an Affymetrix U133A GeneChip, from core biopsies of patients' tumor.

In addition, we accessed a large well-defined collection of early stage breast cancer samples (GSE3143) with expression data and relevant clinical information. Forty five patients in this cohort received 5-flourouracil, adriamycin and cyclophosphamide (FAC) adjuvant chemotherapy and had

accurate response data. Responders (n = 34) were defined as those patients in whom there was no evidence of disease recurrence for at least five years after chemotherapy. Non-responders (n = 11) were patients in whom a pathologic disease recurrence had occurred within 1-year of starting FAC chemotherapy.

Human ovarian cancer samples. We measured expression of 22,283 genes in 13 ovarian cancer cell lines and 119 advanced (FIGO stage III/IV) serous epithelial ovarian carcinomas using Affymetrix U133A GeneChips. All ovarian cancers were obtained at initial cytoreductive surgery from patients treated at H. Lee Moffitt Cancer Center & Research Institute or Duke University Medical Center. Patients that experienced progressive or recurrent disease following initial platinum-based therapy received single agent topotecan, adriamycin, docetaxel, or paclitaxel salvage chemotherapy. All tissues were collected under the auspices of respective institutional (Duke University Medical Center and H. Lee Moffitt Cancer Center) IRB approved protocols involving written informed consent.

Classification of salvage chemotherapy response in tumors. Response to therapy was evaluated using standard criteria for patients with measurable disease, based upon WHO guidelines⁹. CA-125 was used to classify responses only in the absence of a measurable lesion; CA-125 response criteria were based on established guidelines^{10,11}. A complete response (CR) was defined as a complete disappearance of all measurable and assessable disease or, in the absence of measurable lesions, a normalization of the CA-125 level following salvage therapy. A partial response (PR) was considered a 50% or greater reduction in the product obtained from measurement of each bi-dimensional lesion for at least 4 weeks or a drop in the CA-125 by at least 50% for at least 4 weeks. Disease progression (progressive disease, PD) was defined as a 50% or greater increase in the product from any lesion documented within 8 weeks of initiation of therapy, the appearance of any new lesion within 8 weeks of initiation of therapy, or a doubling in the CA-125 from baseline. For the purposes of our analysis, a

clinically beneficial response (i.e. “responder”) included CR or PR. A patient who did not demonstrate a CR or PR was considered a “non-responder”.

Lung cancer cell culture. Total RNA was extracted and oncogenic pathway predictions was performed similar to the methods described previously²⁶. All liquid media as well as the Thiazolyl Blue Tetrazolium Bromide were purchased from Sigma Aldrich (St. Louis, MO). LY-294002 (a phosphatidylinositol 3-kinase (PI3K) inhibitor) was purchased from Calbiochem (San Diego, CA). The non-small cell carcinoma cell lines were grown as recommended by the supplier (ATCC, Rockville, MD and DSMZ, Braunschweig, Germany). All tissue culture reagents were obtained from Sigma (UK).

Ovarian cancer cell culture. All liquid media as well as the Thiazolyl Blue Tetrazolium Bromide were purchased from Sigma Aldrich (St. Louis, MO). The Src inhibitor SU6656 was purchased from Calbiochem (San Diego, CA). The ovarian cancer cell lines, OV-90, OVCAR-5, TOV-21G, and TOV-112D were grown as recommended by the supplier (ATCC, Rockville, MD). FUOV-1, a human ovarian carcinoma, was grown according to the supplier (DSMZ, Braunschweig, Germany). Seven additional cell lines (C13, OV2008, A2780-CP, A2780S, IGROV-1, T8, IMCC3, A2008) were provided by Dr. Patricia Kruk, College of Medicine (University of South Florida, FL). All of those seven cell lines were grown in RPMI 1640, supplemented with 10% Fetal Bovine Serum, 1% Sodium pyruvate, and 1% non essential amino acids. All tissue culture reagents were obtained from Sigma (UK).

Cell and RNA preparation. Full details of the methods used for RNA extraction and development of gene expression signatures representing deregulation of oncogenic pathways in the lung (n = 91) and breast tumors (n = 171) are described in our recent publication²⁶. Briefly, total RNA was extracted

using the Qiashredder and Qiagen Rneasy Mini kits. Quality of the RNA was checked by an Agilent 2100 Bioanalyzer. The targets for Affymetrix DNA microarray analysis were prepared according to the manufacturer's instructions. Biotin-labeled cRNA, produced by *in vitro* transcription, was fragmented and hybridized to the Affymetrix U133A GeneChip arrays (www.affymetrix.com/products_arrays_specific_Hu133A.affx) at 45^o C for 16 hr and then washed and stained using the GeneChip Fluidics. The arrays were scanned by a GeneArray Scanner and patterns of hybridization detected as light emitted from the fluorescent reporter groups incorporated into the target and hybridized to oligonucleotide probes. All analyses were performed in a MIAME (minimal information about a microarray experiment)-compliant fashion, as defined in the guidelines established by MGED (www.mged.org). The expression data for the tumor samples and cell lines and the genelists that constitute the individual chemotherapy predictors is available as GEO accession number GSE3151 and at <http://data.cgt.duke.edu/Combo1.php>

Cross-platform Affymetrix Gene Chip comparison. To map the probe sets across various generations of Affymetrix GeneChip arrays, we utilized an in-house program, Chip Comparer (<http://tenero.duhs.duke.edu/genearray/perl/chip/chipcomparer.pl>) as described previously²⁶.

Cell proliferation assays. Growth curves for cells were produced by plating 500-10,000 cells per well in 96-well plates. The growth of cells at 12hr time points (from t =12 hrs) was determined using the CellTiter 96 Aqueous One 23 Solution Cell Proliferation Assay Kit by Promega, which is a colorimetric method for determining the number of growing cells²⁷. The growth curves plot the growth rate of cells vs. each concentration of drug tested against individual cell lines. Cumulatively, these experiments determined the concentration of cells to use for each cell line, as well as the dosing range of the inhibitors. The final dose-response curves in our experiments plot the percent of cell population responding to the chemotherapy vs. the concentration of the drug for each cell line.

Sensitivity to docetaxel and a phosphatidylinositol 3-kinase (PI3 kinase) inhibitor (LY-294002)²⁸ in 17 lung cell lines, and topotecan and a src inhibitor (SU6656) in 13 ovarian cell lines was determined by quantifying the percent reduction in growth (versus DMSO controls) at 96 hrs using a standard MTT colorimetric assay²⁹. Concentrations used ranged from 1-10nM for docetaxel, 300nM-10 μ M (SU6656), and 300nM-10M for LY-294002. All experiments were repeated at least three times.

Statistical analysis methods. Prior to statistical modeling, gene expression data is filtered to exclude probesets with signals present at background noise levels, and for probesets that do not vary significantly across samples. Each signature summarizes its constituent genes as a single expression profile, and is here derived as the top principal components of that set of genes. When predicting the chemosensitivity patterns or pathway activation of cancer cell lines or tumor samples, gene selection and identification is based on the training data, and then metagene values are computed using the principal components of the training data and additional cell line or tumor expression data. Bayesian fitting of binary probit regression models to the training data then permits an assessment of the relevance of the metagene signatures in within-sample classification²¹, and estimation and uncertainty assessments for the binary regression weights mapping metagenes to probabilities. To guard against over-fitting given the disproportionate number of variables to samples, we also performed leave-one-out cross validation analysis to test the stability and predictive capability of our model. Each sample was left out of the data set one at a time, the model was refitted (both the metagene factors and the partitions used) using the remaining samples, and the phenotype of the held out case was then predicted and the certainty of the classification was calculated. Given a training set of expression vectors (of values across metagenes) representing two biological states, a binary probit regression model, of predictive probabilities for each of the two states (resistant vs. sensitive) for each case is estimated using Bayesian methods. Predictions of the relative oncogenic pathway status and chemosensitivity of the validation cell lines or tumor samples are then evaluated using methods

previously described ^{16,21} producing estimated relative probabilities – and associated measures of uncertainty – of chemosensitivity/oncogenic pathway deregulation across the validation set.

Further details of the analysis of expression data are previously described. The statistical analysis involved in generating predictive models indicative of chemotherapeutic sensitivity uses standard binary regression models combined with singular value decompositions SVDs, also referred to as singular factor decompositions, and with stochastic regularization using Bayesian analysis. It is beyond the scope here to provide full technical details, so the interested reader is referred to manuscripts that are available at the Duke web site, url www.isds.duke.edu/~mw. Some key details are elaborated here. Assume n tumors and p genes, and write X for the pxn matrix of expression values, with rows as genes and columns as tumors. Column i of X is the p -vector x_i of expression levels of all genes on tumor i . Singular factor decomposition of the set of expression measures on the sample of tumors has the form $X=ADF$, where D is a diagonal nxn matrix of non-negative singular values, A is a pxn matrix with orthogonal columns, and F is an nxn orthonormal matrix of (metagene) factor values. Column i of F , denoted by f_i , is the n -vector of values of all n factors on tumor i , and we have $x_i=ADf_i$. In the current study as an example, write $p(x_i)$ for the probability tumor/cell line i is chemosensitive versus chemoresistant. A probit regression sets $p(x_i)=P(b'x_i)$ where P is the standard normal distribution function and $b'x_i$ is a linear combination of expression levels based on the p -vector of regression parameters b . Then $p(x_i)=F(g'f_i)$ where $g=DA'b$ is a n -vector of regression coefficients for the factors. Hence regression on genes reduces to regression on “metagene” factors, a much lower dimensional inference problem.

The resulting Bayesian analysis may be easily implemented using standard iterative Markov chain Monte Carlo (MCMC) simulation methods of Bayesian analysis (1-5) to impute sets of simulated parameter values whose distributions are summarized to produce point and interval estimates of model parameters g as well as of probabilities of chemotherapy sensitivity in both the

training set and validation samples. This involves the standard method of imputing the latent normal variates implicit in the probit function as part of the simulation analysis. The orthogonality of the factor design implied by the orthonormality of F leads to the use of independent Student T prior distributions on the elements of the factor regression parameter vector g , and model fitting involves representing the T distributions as scale mixtures of normal priors and includes estimation of the implicit scale factors in the MCMC analysis, again a standard technique. Analyses reported are based on Student T priors with 2 degrees of freedom, providing relatively vague prior forms.

In addition to posterior samples for the factor parameters g , the MCMC approach leads directly to the calculations required for prediction of chemotherapy sensitivity for any given predictor. Most importantly, the Bayesian SVD regression framework allows direct inversion to infer the parameters b from g , to provide inferences about which genes are important in defining $p(x)$, and how subsets of genes interact (3). Specifically, new theory in (2) shows that the relevant inversion is simply the least-norm generalized inverse $b=AD^{-1}g$. Hence posterior sample values for g are trivially mapped to corresponding sample values of b , and summarized to produce posterior estimates of b .

It is pertinent to explore comparisons of the chosen binary regression, using the probit form, with alternatives such as the standard logistic. We have done this, making repeat analysis using models in which P is a Student T distribution rather than normal, and with varying the degrees of freedom within which the Student T with 8 or 9 degrees of freedom very closely approximates the standard logistic function. Following, the MCMC analysis of the probit is trivially extended to Student T models. In these studies, predictive results and interpretation of the cell line and tumor response data are not altered significantly, indicating robustness to the assumed form in this setting.