0026-895X/06/6903-737–748
MOLECULAR PHARMACOLOGY
U.S. Government work not protected by U.S. copyright
Mol Pharmacol 69:737–748, 2006

Vol. 69, No. 3
17061/3087588
Printed in U.S.A.

Differential Gene Expression as a Potential Classifier of 2-(4-Amino-3-methylphenyl)-5-fluorobenzothiazole-Sensitive and -Insensitive Cell Lines[S]

Anders Wallqvist, John Connelly, Edward A. Sausville,[1] David G. Covell, and Anne Monks

Science Applications International Corporation (A.W., J.C., A.M.) and Developmental Therapeutics Program (E.A.S, D.G.C.), National Cancer Institute–Frederick, Frederick, Maryland

Received July 21, 2005; accepted December 5, 2005

ABSTRACT

2-(4-Amino-3-methylphenyl)-5-fluorobenzothiazole (5F-203) is a candidate antitumor agent empirically discovered with the aid of the National Cancer Institute (NCI) Anticancer Drug Screen. In an effort to determine whether basal expression of genes could be used to classify cell sensitivity to this agent, serial analysis of gene expression (SAGE) libraries were generated for three sensitive and two insensitive human tumor cell lines. When the SAGE tags expressed within these cell line libraries were compared and evaluated for differences, several genes seemed more highly expressed in 5F-203-sensitive cell lines than in the insensitive cell lines. Constitutive expressions of 15 genes identified by the analysis were then measured by quantitative reverse-transcription polymerase chain reaction in the 60 cell lines of the NCI Anticancer Drug Screen. This generated a pattern of relative basal gene expression across the cell lines and also confirmed the differential expression of SAGE-discovered genes within the initial set of five cell lines. Further analyses of these expression data in 60 cell lines suggested that a smaller subset of genes could be used to classify tumor cell sensitivity to 5F-203. In contrast, the same set of genes did not predict with equivalent precision sensitivity to unrelated active drugs, and another set of genes could not better classify the cell lines in terms of 5F-203 sensitivity. These results suggest that constitutive gene expression profiles, in which the genes are not necessarily known to be related to the mechanism of action of a given drug, may be viewed as a general tool to extend and improve the concept of a single predictive gene to groups of predictive genes.

The advent of gene expression profiling raises the prospect of an exploitable association between constitutive gene expression and response to a chemotherapeutic agent. The candidate antitumor agent 2-(4-amino-3-methylphenyl)-5-fluorobenzothiazole (5F-203) (Chua et al., 2000; Bradshaw et al., 2002; Brantley et al., 2004) is selectively active in vitro against certain human tumor-derived cells (including some breast, renal, and ovarian cell lines) as shown in Fig. 1 and has demonstrated activity against breast (Fichtner et al., 2004) and ovarian xenograft models (Bradshaw et al., 1998). Selective metabolism seems to underlie the activity profile of this drug, because only drug-sensitive cell lines accumulate and biotransform 5F-203. This metabolic biotransformation in human cells involves P450 CYP1A1-induced C-oxidation of the benzothiazole nucleus to form an inactive 6-hydroxy metabolite (Chua et al., 2000) and other metabolites as yet undefined but capable of causing DNA damage. The parent benzothiazole itself is known to trigger aryl hydrocarbon receptor (AhR) translocation to the nucleus (Loaiza-Perez et al., 2002) and induce activation of xenobiotic response element-bearing promoters, resulting in increased CYP1A1 and CYP1B1 transcription in sensitive MCF-7 cells. Moreover, MCF-7 cells exposed to ^{14}C-labeled 5F-203 are found to co-

This project was funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract number N01-CO12400.

[1] Current affiliation: University of Maryland Marlene and Stewart Greenebaum Cancer Center, Baltimore, Maryland.

[S] The online version of this article (available at http://molpharm.aspetjournals.org) contains supplemental material.

Article, publication date, and citation information can be found at http://molpharm.aspetjournals.org.
doi:10.1124/mol.105.017061.

ABBREVIATIONS: 5F-203, 2-(4-amino-3-methylphenyl)-5-fluorobenzothiazole; 5I-203, 2-(4-amino-3-methylphenyl)-5-iodobenzothiazole; DTP, Developmental Therapeutics Program; NCI, National Cancer Institute; PCC, Pearson correlation coefficient; P450, cytochrome P450; Q-RT-PCR, quantitative reverse transcription-polymerase chain reaction; RT-PCR, reverse transcription-polymerase chain reaction; SAGE, serial analysis of gene expression; AhR, aryl hydrocarbon receptor; GI_{50}, concentration value at which the tested drug resulted in a 50% reduction in the net protein increase in control cells; PCR, polymerase chain reaction; bp, base pair(s); GAPDH, glyceraldehyde-3-phosphate dehydrogenase; ESR1, estrogen receptor 1; TFF1, trefoil factor; XBP1, X-box binding protein; ECM, extracellular matrix; ANXA1, annexin A1; CAV, caveolin; VIM, vimentin; 5-FU, fluorouracil; CTGF, connective tissue growth factor; S, sensitive; I, insensitive.

valently bind radioactivity to subcellular macromolecules in a manner not detected in drug-insensitive cells (Brantley et al., 2004), further implicating a metabolizing system unique to drug-sensitive cell lines. Efforts to determine the basis of tumor responsiveness to 5F-203 have centered on the apparent requirement for P450 metabolic activation of this drug (Loaiza-Perez et al., 2002; Hose et al., 2003). In an alternative approach aimed at discovering surrogate markers that might identify 5F-203-sensitive tumors, serial analysis of gene expression (SAGE) libraries (Velculescu et al., 1995; Yamamoto et al., 2001) were constructed from several 5F-203-sensitive and -insensitive cell lines. These data provided a basis for evaluating whether patterns of basal gene expression might be useful in the classification of drug-sensitive cell lines. Along these lines, a published report of an integrated analysis of gene expression and chemosensitivity profiles indicated that this type of approach could be useful in the development of systems to predict drug efficacies of cancer cells by examining the expression levels of particular genes (Dan et al., 2002). Moreover, constitutive gene expression patterns have historically proven useful for molecular classification of tumors, allowing for the identification of otherwise undetected and clinically significant subtypes of cancer (Alizadeh et al., 2000; Scherf et al., 2000; Wallqvist et al., 2002; Covell et al., 2003) and for understanding the underlying genetic basis for toxicology (Amin et al., 2002; Kramer and Kolaja, 2002; Thomas et al., 2002).

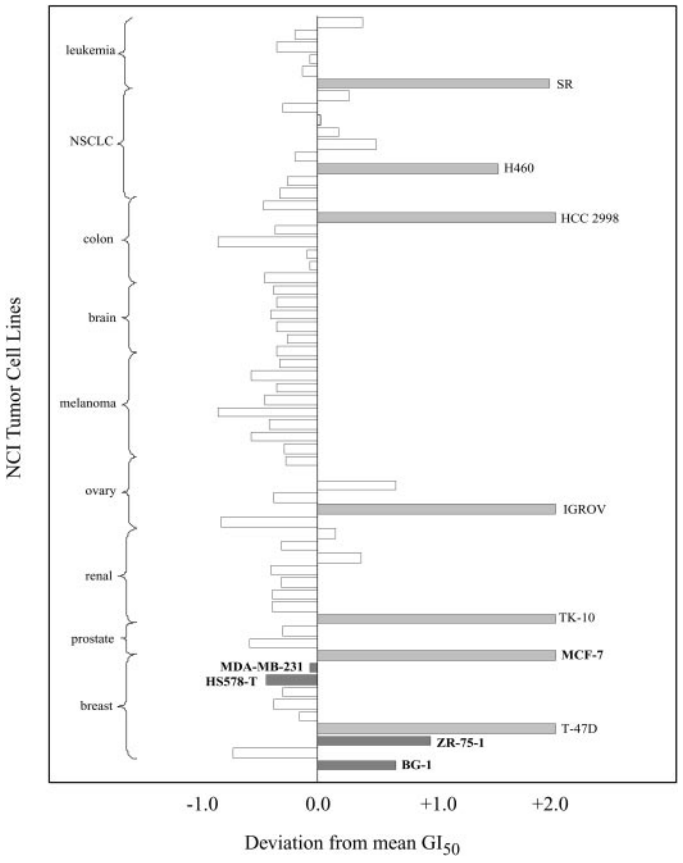

Fig. 1. Drug response is represented in a mean graph showing the relative sensitivity of each cell line after 48-h treatment with 100 to 0.01 μM 5F-203 on a logarithmic scale. To generate the mean graphs, the average 50% growth inhibitory concentration (using the negative $\log_{10}$ of the GI_{50}) across the 60 cell lines was subtracted from each individual cell line GI_{50} value. The center line represents the average $-\log 10(GI_{50})$ response; bars projecting to the right indicate higher than average drug sensitivity (relative to the length of the bar), and bars projecting to the left indicate lower than average sensitivity. Drug-sensitive cell lines with $GI_{50} < 10^{-7}$ M are shown in light gray, the cell lines that were SAGE-profiled are indicated with boldface cell names, and corresponding SAGE-profiled-insensitive cell lines are shown in dark gray.

The aim of this investigation was to identify gene expression patterns in untreated cells that could be interpreted to characterize or potentially predict cellular response to subsequent 5F-203 exposure. The initial part of the study used SAGE as a genetically unbiased tool to identify genes that may be related to the 5F-203-sensitive and -insensitive cellular phenotypes. A comprehensive SAGE profile was determined in five untreated but previously designated sensitive and insensitive cell lines. A search for genes that were expressed in sensitive but not insensitive cell lines, and vice versa, identified a set of genes whose commonality linked them to a latent stress-response network. These genes are not related to any known mechanism of action exerted by the drug 5F-203 itself. In an attempt to link these genes to drug sensitivity, we extended our investigation to the full complement of NCI's drug screening cell line panel, in which 5F-203 sensitivity (GI_{50}) has been characterized for these cell lines. Relative expression of 15 identified genes was performed using quantitative RT-PCR in all 60 cell lines. Examination of these data indicated that linking the gene expression profiles and corresponding GI_{50} values across all cell lines was nontrivial. A qualitative analysis in terms of an "up" and "down" pattern of the selected genes could not accurately account for the cellular GI_{50} response to 5F-203. Thus, parametric models were used that reliably accounted for the response with a high specificity. This allowed us to empirically capture and identify relevant basal gene expression data from untreated cells that can be used to classify the 5F-203 sensitivity of 60 human tumor cell lines.

The results of our analysis may have relevance for further development of members within the benzothiazole class. Previous investigations have indicated that basal levels of expression for AhR, aryl hydrocarbon receptor nuclear translocator, CYP1A1, and CYP1B1 are not correlated to sensitivity, but the ability to induce expression of the *CYP1A1* gene is (Loaiza-Perez et al., 2002; Hose et al., 2003). However, assessment of inducibility of a gene by a drug is difficult in the clinical arena; thus, it would be advantageous to define a complement of gene expression that could be determined before therapy and used to correlate with cell sensitivity and ultimately perhaps patient susceptibility to the agent.

Materials and Methods

SAGE. Five cell lines were selected for global gene expression analysis on the basis of drug response. These included three 5F-203 sensitive cell lines, MCF-7, BG-1, and ZR-75-1, in which the latter two cell lines are not part of the standard NCI-60 cell line panel, and two insensitive cell lines, MDA-MB-231 and HS578T. ZR-75-1 is an estrogen receptor-positive breast cancer cell line, and BG-1 is an estrogen receptor-expressing ovarian cell line. SAGE libraries were constructed according to published methods (Velculescu et al., 1995). In brief, mRNA was isolated from each cell line using the PolyAtract System 1000 (Promega, Madison, WI) and then was annealed with 5′-biotintylated oligo(dT) and converted to cDNA using Superscript Choice System's kit for cDNA synthesis (Gibco BRL, Carlsbad, CA). The cDNA was digested with NlaIII (anchoring enzyme), and the fragments were isolated with streptavidin-coated magnetic beads

(Dynal Biotech, Lake Success, NY) and split into two tubes. Each tube was ligated with one of the two annealed linker pairs. Tags were released from the restriction site (CATG) of each transcript by digestion with BsmFI (tagging enzyme) and then blunt-ended with Klenow. The two blunt-ended samples were ligated to form ditags that were amplified by PCR using primers designed by Integrated DNA Technologies (Coralville, IA) (primer 1, 5′ GGATTTGCTGGTGCAGTACA 3′; primer 2, 5′ CTGCTCGAATTCAAGCTTCT 3′). Amplification required 26 to 28 cycles of 30 s at 94°C followed by 1 min at 55°C, 1 min at 74°C, and one hold for 5 min at 74°C. The amplified ditags were isolated by acrylamide gel, extracted, and purified. After digestion with NlaIII, the linkers were removed by the addition of streptavidin-lined magnetic beads and subsequent gel electrophoresis. The resulting 26-bp bands were concatenated overnight with T4 Ligase (Gibco BRL) to form concatemers with a length between 600 and 2500 bp that were isolated by agarose gel. The resulting DNA fragments were cloned into an SphI-cleaved pZero-1 plasmid (Invitrogen, Carlsbad, CA) and transfected into ElectroMAX DH10Bs cells (Gibco BRL). Colonies were screened by PCR to select long inserts (>600 bp) for automated sequencing (ABI 3700DNA sequencer; Applied Biosystems, Foster City, CA). Raw, concatenated tag-sequencing data were analyzed through the SAGE software, and the abundance of tags was calculated. Tags were matched to genes using a download of the SAGE Tag to Unigene (Wheeler et al., 2003) Map (NlaIII, 10 bp), Human 03/01 (5263 kB) version from the SAGE Resources website. Where possible, an 11th base pair was identified and used to manually focus the identification of a tag to a specific gene in cases of multiple Unigene cluster assignments. Differentially expressed genes were selected as those with a ≥70% probability that tags have greater than 2-fold difference between drug-sensitive and drug-resistant cell lines; see below for further details.

RT-PCR for Gene Expression Confirmation. RNA samples of 60 human tumor cell lines from the NCI Anticancer Drug Screen, provided by Dr. Scudiero through the DTP Molecular Target Program, were used to measure selected target expression by real-time PCR (TaqMan). Six 1-μg samples of total RNA were reverse-transcribed for each cell line. RT-PCR reactions were measured with the ABI Prism 7700 Sequence Detection System in a 50-μl reaction volume using TaqMan SYBR green PCR master mix (Applied Biosystems). Primers were designed with Primer Express software (Applied Biosystems) using the GenBank sequence for the human gene. Triplicate wells were prepared for the target (at a primer concentration of 300 nM) and for the internal standard, GAPDH (at a primer concentration of 100 nM). Thermocycler parameters were 30 min at 48°C, 10 min at 95°C, and 40 PCR cycles of 15 s at 95°C and 1 min at 60°C. Data were analyzed using the comparative CT method (PerkinElmer Life and Analytical Sciences, Boston, MA), and normalized basal expression of each target was expressed relative to a calibrator cell line (EKVX) to provide a reference pattern of gene expression. Duplicate measurements were made for five genes using the same 60 samples but with newly prepared primer sets. The Pearson correlation coefficient (PCC) between the duplicate 60 cell line patterns was greater than 0.87 for 4 measurements, and PCC = 0.68 for the fifth. This indicates good reproducibility for the measurement of relative gene expression in the 60 cell lines.

DTP Resources of GI_{50} and Gene Microarray Expression Data. The NCI-60 drug discovery panel was developed as an in vitro tool to assess anticancer activity of compounds against a range of cell lines derived from different tumors, including lung, renal, colorectal, ovarian, breast, prostate, central nervous system, melanoma, and hematological malignancies (Monks et al., 1991). The data contain concentration values (GI_{50}) at which the tested drug resulted in a 50% reduction in the net protein increase in control cells during the 48-h drug incubation. The GI_{50} data vectors used in our analysis were log-transformed and selected to have a maximum of 20 missing data elements and a signal covariance of at least 0.02. Missing data elements were not included in any calculation. The concentration range at which the GI_{50} values were determined for 5F-203 ranged from a $-\log(GI_{50})$ of 4.0 to 8.0. The pattern of GI_{50} values across the tumor cell lines has proven useful for identifying mechanisms of action for some drug classes and aids in the classification of novel drugs submitted to the NCI's tumor screen (Paull et al., 1989; Keskin et al., 2000; Rabow et al., 2002).

The oligonucleotide microarray data on the constitutive gene expression in the NCI 60 cell lines used in this work are derived from triplicate measurements on Affymetrix U95Av2 chips (Affymetrix, Santa Clara, CA) extracted from the public data resources available at the DTP website (http://dtp.nci.nih.gov). Individual data vectors for each gene were constructed from the log of the median values, in which we only considered gene measurements that are associated with a p value of 0.05 or less.

Statistical Analysis and Classification Scheme. The significance of differentially expressed sequence tags was calculated from a Bayesian approach (Yamamoto et al., 2001) in which the posterior probability (P) that the concentration of a particular sequence tag in one cell line a (out of a total of A tags) is different from another cell line b (out of a total of B tags) by at least a factor of F is given by

$$P(x \geq L) = \frac{\int_{L}^{1} g(x)dx}{\int_{0}^{1} g(x)dx}$$

where $L = F/F + 1$, and

$$g(x) = f(x)\,\frac{x^a(1-x)^b}{[1 + (A/B - 1)x]^{a+b}}$$

$$f(x) = x^3(1-x)^3$$

The evaluation of the P values is carried out by first choosing a value for F (i.e., the desired differential factor between two gene sets) and then using the experimentally determined number of tags a and b to define $g(x)$. With the now-determined parameters in $g(x)$, the P value is calculated by evaluating the ratio of the integrals between the limits L to 1 and 0 to 1. Genes were considered differentially expressed in the sensitive and insensitive SAGE cell lines if their relative tag count was equal to or greater than a factor of 2, with a 70% probability that the observation is significant.

The Pearson or sample correlation coefficient (PCC) between two vectors $\vec{u}$ and $\vec{v}$ is defined as

$$\mathrm{PCC}(\vec{u}, \vec{v}) = \frac{\sum_i (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_i (u_i - \bar{u}) \cdot \sum_i (v_i - \bar{v})}}$$

where $\bar{u}$ denotes the average of all elements in $\vec{u}$ and similarly for the vector $\vec{v}$. The correlation coefficient measures the fidelity of a linear fit of the observables elements in $\vec{v}$ on the variables in $\vec{u}$, and takes on values between -1 and $+1$. A PCC value of 1 indicates that each vector is linearly dependent on the other; it does not mean that the vectors are exactly the same.

Standard least-squares fitting procedures were used to estimate the unknown parameters in the model used for predicting the 5F-203 GI_{50} pattern. Minimizing the residual error was done using the praxis procedure, a principal axis method requiring no gradients (Brent, 1973). To solve the system of linear equations, a conventional single-value decomposition method was used.

Prediction Models. The simultaneous variation in constitutive gene expression found for the genes in the five cell lines investigated using SAGE was identified on the basis of known cellular 5F-203 sensitivity and insensitivity. To proceed with the exploration of connections between gene expression and cellular response, we make two assumptions: 1) the identified pool of genes carry *gene*-based

information on 5F-203 sensitivity and insensitivity for *any* tumor cell type; and 2) the constitutive expression pattern within these genes contains *sufficient* information to characterize cellular sensitivity. These assumptions exclude considerations of alternative genes or factors not identified here that could also serve the same purpose (i.e., the genes identified here may not be unique). These assumptions are accompanied by three caveats: 1) we do not know a priori the minimum number of genes necessary to characterize sensitivity; 2) the mechanism by which this pattern of gene expressions confers sensitivity to these cells is not known; and 3) there is no a priori preconception about how to interpret expression patterns. These assumptions and caveats pose several challenges to applying genetic markers as predictors of response; for example, the use of only a few genes as indicators might undersample the global gene expression pool and thus contribute to the possibility that a precise pattern of over- and underexpressed genes may be appropriate indicators of response for only a fraction of cell types. Although restrictive, these assumptions allow us to quantify our observations into a model predictive of sensitivity. Our immediate aim is to use as few genes as possible to correctly classify the largest fraction of NCI's 60 tumor cells according to 5F-203 sensitivity. We accomplish this aim by imposing a parametric interpretation of the gene expression. Therefore, it is outside of our scope to propose the exact biological processes by which any selected gene's expression pattern ultimately governs cellular sensitivity to 5F-203.

The two models considered are a linear and a nonlinear model, in which the sensitivity S of a cell is either given as

$$S_{\text{linear}}(\vec{e}) = \sum_i c_i e_i$$

or

$$S_{\text{nonlinear}}(\vec{e}) = \frac{\sum_i c_i e_i}{1 - \sum_i d_i e_i}$$

and the vector $\vec{e}$ consists of RT-PCR-measured gene expression data relative to EKVX in each particular cell line, the summation runs over the number of selected genes, and the set of parameters c and d are to be determined. By setting all d coefficients to 0 in $S_{\text{nonlinear}}$, the linear version is recovered. The range of values for S reflects the concentrations used in the dose-response measurements, and values outside of these limits are set to the boundary values of $-\log(\text{GI}_{50})$ (i.e., 4 and 8). The linear equation S_{linear} can be solved using standard methods of linear algebra for a system of equations, whereas the coefficients of the nonlinear equation $S_{\text{nonlinear}}$ have to be determined via parametric optimization.

Probability of Sensitive Response from Gene Expression Patterns. To simplify the description of gene expression, we can introduce a qualitative measure denoting three states: up, down, or no change. Because these three states correspond to a range of expression values, a gene would be considered up-regulated if its expression was larger by a factor of 2, the same if the expression ranged between a factor of 0.5 to 2, and down-regulated if the expression was less than a factor of 0.5 from the reference standard. In the case of 5F-203, a sensitive response is characterized as requiring a maximum concentration of 10^{-7} M to achieve 50% growth inhibition. If we were considering a gene expression pattern with only two genes, there would be nine (3^2) possible combinations of up ($\uparrow$), down ($\downarrow$), and same ($\circ$) [i.e., ($\uparrow\uparrow$), ($\uparrow\circ$), ($\uparrow\downarrow$), ($\circ\uparrow$), ($\circ\circ$), ($\circ\downarrow$), ($\downarrow\uparrow$), ($\downarrow\circ$), and ($\downarrow\downarrow$)], whereas for seven genes there are 3^7 (2187) possible combinations. To calculate the probability of response, p_r, we can enumerate all possible combinations of genes and assign the three expression states from the sensitivity model then and calculate p_r for each combination.

Data and Program Availability. All data and program sources are freely available upon request. Contact the corresponding author for details.

Results

Cellular Sensitivity (GI_{50}) Profile of 5F-203. Figure 1 displays individual GI_{50} values of 5F-203 across the NCI tumor panels. The GI_{50} values show considerable cell line variability, ranging from insensitive ($\text{GI}_{50} > 50$ μM) to sensitive ($\text{GI}_{50} < 0.1$ μM) cell lines, reflecting a 4 order of magnitude difference. The different growth inhibitory concentrations for each cell line in this assay point to specific variations in response to 5F-203. There is no apparent specificity in the response according to the tissue of origin, as shown by the intratumor panel variations and the lack of any consistent intertumor panel difference. The breast panel has two sensitive cell lines in this assay, MCF-7 and T-47D, and includes the insensitive cell lines BT-549, HS578T, MDA-MB-435, and MDA-MB-231. The other sensitive cell lines occur in colon, leukemia, lung, ovarian, and renal tumor panels. There are a few cell lines of intermediate sensitivity in the concentration range of 10^{-5} to 10^{-6} M, but the bulk of the cells are relatively insensitive, with GI_{50} values greater than 10 μM (10^{-5}M). Variations in GI_{50} response partly reflect the underlying genetic characteristics of the cells, either via constitutively expressed genes or via genes induced by the drug insult. Our analysis postulates that the identification of constitutively expressed gene patterns may reveal a genetic signature that can be used to classify GI_{50} sensitivity and, ultimately for some agents, the possibility of patient sensitivity.

Differential Gene Expression. SAGE libraries were constructed from the drug-sensitive cell lines BG1, ZR-75-1, and MCF-7, plus a drug-treated MCF-7, and for the drug-insensitive cell lines MDA-MB-231 and HS578T. The drug-treated MCF-7 cell line showed an increased expression of tags associated with cytochrome P450 genes *CYP1A1* and *CYP1B1* (32 and 14 tags, respectively) compared with control (1 and 0), confirming results reported previously using microarrays and quantitative RT-PCR (Hose et al., 2003). The chance probability of rejecting these counts as corresponding to a 2-fold or greater increase in expression is less than 0.04. CYP1A1 and CYP1B1 gene products are believed to be essential for metabolizing 5F-203 to an active form. It has been proposed that the induction of CYP1A1 in response to 5F-203 treatment could be used as a potential surrogate marker for sensitivity to 5F-203, because there is a direct correlation between sensitivity and cellular inducibility of this gene in response to drug exposure (Hose et al., 2003). However, constitutive expression of these metabolizing genes could not be directly related to drug response, and the feasibility of measuring ex vivo gene induction in patient samples may be questionable. Consequently, we used the current approach, making SAGE libraries (probing the expression of all genes in the cell and not just restricted to gene sets available on a printed array) as a paradigm for exploring the significance of genetic differentials in constitutive gene expression patterns before drug exposure between drug-sensitive and insensitive cell lines. The SAGE tag counts generated (>44,000/cell) for constitutive gene expressions were compared between the set of sensitive cell lines (BG1, ZR-75-1, and MCF-7) and the set

of insensitive cell lines (MDA-MB-231 and HS578T). Using the relative difference in genes assigned to the SAGE tag counts between the two sets of cells, a small group of genes was identified that seems to be uniquely related to 5F-203 sensitivity and insensitivity. Within this group, two subgroups of genes were found: genes whose expression in the drug-insensitive cell lines was greater than a factor of 2 compared with sensitive cells, and those genes whose expression in the sensitive cell lines was greater than a factor of 2 compared with insensitive cells. The selection of each gene is thus based on six comparisons between sensitive (S) and insensitive (I) cells, I1/S1, I1/S2, I1/S3, I2/S1, I2/S2, and I2/S3. Because each of the six possible differences between sensitive and insensitive cells is tested at the 70% level, the chance occurrence of any one such pattern of six comparisons is less than $(1.00 - 0.70)^6$ to $7.3 \cdot 10^{-4}$. The combination of the chance of selecting one gene into a pattern of six is thus miniscule. These genes and their corresponding tag counts are listed in Tables 1 and 2. Most of these genes have unique tags that directly identify the corresponding gene sequence. The genes identified as differentially expressed in drug-resistant or -sensitive cell lines do not regulate their own expression in response to the drug, because gene abundance in an 5F-203-treated cell line was the same as the basal expression measured in the control MCF-7 cell lines, as shown in column D1 of these tables. This raises the hypothesis that genes identified in Tables 1 and 2 may be regarded as a reflection of a constitutive expression pattern that is characteristic of the sensitivity of these few cell lines to the toxic insult of 5F-203.

To extend this hypothesis, 15 SAGE-selected genes (plus three genes selected on the basis of information about the proposed mechanism of action or drug-sensitive cell populations), were measured with quantitative RT-PCR (Q-RT-PCR) in the same set of cells plus an additional 57 human tumor cell lines of the NCI Anticancer Drug Screen (Table 3). For the most part, these data confirmed the differential expressions measured in the five SAGE-characterized cell lines, although annexin 2, HSPD1, and TRA1 could not unambiguously be differentiated between the three sensitive and two insensitive cell lines. Thus, the SAGE data for differentially expressed genes in 5 cell lines was verified in the same 5 cell lines for 12 of the 15 genes (12/15 = 0.8) when measured by Q-RT-PCR. This is in agreement with the

TABLE 1

Genes expressed in sensitive but not in insensitive cell lines as determined from SAGE

The specific tag linked to the gene is given in the first column, followed by the tag count in the SAGE libraries. The two insensitive cell lines HS578T and MDA-MB-231 are designated I1 and I2 (columns 2 and 3), and the sensitive cell lines BG1, ZR-75–1, and MCF-7 are designated S1 to S3 (columns 4–6). The 5F-203-treated MCF-7 cell line is indicated as D1. The genes in the sensitive cell lines are present in the drug-exposed MCF-7 cell line.

Tags	I1	I2	S1	S2	S3	D1	Gene	Description
AAGAATTTGA	1	0	11	21	14	7	*GAD1*	Glutamate decarboxylase 1 (brain, 67 kD)
AAGAATTTGA	1	0	11	21	14	7	*NDUFB1*	NADH dehydrogenase (ubiquinone) 1β
CAATTAAAAG	7	7	23	68	43	44	*XBP1*	X-box binding protein 1
CCTCCAGCTA	4	39	379	159	725	580	*KRT8*	Keratin 8
CTGGCCCTCG	0	0	26	13	14	14	*TFF1*	Trefoil factor 1
TACCAGTGTA	1	1	11	11	13	11	*HSPD1*	Heat shock 60 kD protein 1 (chaperonin)
TACCCCACCC	5	6	31	21	21	20	*MAZ*	MYC-associated zinc finger protein
TGTGGGTGCT	0	0	22	32	9	9	*CDH1*	Cadherin 1, E-cadherin (epithelial)

TABLE 2

Genes expressed in insensitive but not in sensitive cell lines as determined from SAGE

The specific tag linked to the gene is given in the first column, followed by the tag count in the SAGE libraries. The two insensitive cell lines HS578T and MDA-MB-231 are designated I1 and I2, and the sensitive cell lines BG1, ZR-75–1, and MCF-7 are designated S1 to S3. The 5F-203-treated MCF-7 cell line is indicated as D1. The genes in the sensitive cell lines are not present in the drug-exposed MCF-7 cell line.

Tags	I1	I2	S1	S2	S3	D1	Gene	Description
AAATGCCACA	16	21	3	3	2	0	*RTN4*	Reticulon 4
AGAAAGATGT	17	39	1	1	0	0	*ANXA1*	Annexin A1
AGAACCTTCC	9	33	0	0	0	1	*EEF1A1*	Major histocompatibility complex, class I, A
AGGAATGCTT	16	16	3	2	2	4	*TARS*	Threonyl-tRNA synthetase
AGTGTCTGTG	17	27	0	0	0	0	*CR61*	Cysteine-rich, angiogenic inducer, 61
ATATGTATAT	11	13	0	1	1	1	*ZNF6*	Zinc finger protein 6 (CMPX1)
CATATCATTA	113	72	0	2	1	2	*IGFBP7*	Insulin-like growth factor binding protein 7
CTGACCTGTG	10	22	0	1	0	0	*HLA-B*	Major histocompatibility complex, class I, B
CTTCCAGCTA	24	17	3	3	2	4	*ANXA2*	Annexin A2
GCCATAAAAT	14	13	0	0	0	0	*PRG1*	Proteoglycan 1, secretory granule
GCCCCCAATA	190	84	17	3	19	7	*LGALS1*	Lectin, galactoside-binding, soluble, 1 (galectin 1)
GGAGTGTGCT	51	7	0	0	0	0	*MYL9*	Myosin regulatory light chain 2
GGTTATTTTG	15	11	0	0	0	0	*SERPINE1*	Serine (cysteine) proteinase inhibitor, E (nexin)
TAAAAATGTT	44	21	0	0	0	0	*SERPINE1*	Serine (cysteine) proteinase inhibitor, E (nexin)
TAGAAACCAG	12	17	0	1	0	0	*CNN3*	Calponin 3, acidic
TATACCAATC	10	12	1	0	1	0	*DDAH1*	Dimethylarginine dimethylaminohydrolase 1
TCCAAATCGA	145	15	0	0	0	0	*VIM*	Vimentin
TCCTGTAAAG	37	15	1	1	0	0	*CAV*	Caveolin 1 caveolae protein, 22 kD
TCTTGTGCAT	62	48	13	8	12	3	*LDHA*	Lactate dehydrogenase A
TGTATAAAAA	43	67	7	10	9	4	*TRA1*	Tumor rejection antigen (gp96) 1
TGTCATCACA	53	17	0	0	0	0	*LOXL2*	Lysyl oxidase-like 2
TTCTATTTCA	7	16	0	0	0	0	*MSN*	Moesin
TTGAAAGGTT	13	11	1	0	1	3	*UAP1*	UDP-*N*-acteylglucosamine pyrophosphorylase 1
TTGCCCCCGT	7	30	0	0	0	0	*AXL*	AXL receptor tyrosine kinase
TTTGCACCTT	123	24	0	0	1	0	*CTGF*	Connective tissue growth factor

chosen ≥70% level of significance used in evaluating tag statistics.

The constitutive expression of this set of 18 genes measured by Q-RT-PCR across the 60 cell lines in the NCI cancer tumor panel overall showed a good correlation with the publicly available constitutive gene expression profiles for these cell lines, measured via Affymetrix oligonucleotide microarray. The median correlation was 0.8, whereas the actual values ranged from 0.34 (TRA1) to 0.92 (vimentin, VIM), indicating, in some instances, potential large discrepancies between individual gene expression measurements.

Although the SAGE-selected gene profiles in the extended cell line panel were not as clearly associated with cell line sensitivity as the initial five cell lines, to determine whether any of these individual genes might be a surrogate marker for drug sensitivity, the measured gene expression profiles were correlated with the GI_{50} response profile from the DTP drug database (Table 3). In general, none of the individual expression profiles correlated well with the growth inhibition pattern of 5F-203; the strongest correlation was seen in the CTGF expression pattern with a PCC value of −0.37. Thus, analogous with the constitutive CYP1A1 gene expression pattern, none of these genes seems a priori to be related directly to the mechanism of growth inhibition exhibited by 5F-203. Of significance, however, was the observation that the SAGE data indicated that this group of genes, in contrast to the CYP1A1 gene, was differentially expressed within the initial set of five sensitive and insensitive cell lines. The question thus arises as to whether the combinations of these constitutively differential gene expression profiles could be used as an indicator of cellular response to 5F-203.

To study the influence of genes that may be tangentially associated with 5F-203 sensitivity, on the basis of an a priori proposed mechanism of action or observation of sensitive cell lines, the expressions of estrogen receptor 1 (ESR1) (Hose et al., 2003), the AhR (Loaiza-Perez et al., 2002), and the CD44 antigen were also measured. CD44 is a cell-surface glycoprotein that has been found to be correlated to both estrogen receptor status (Durst et al., 2001; Sorbello et al., 2003) and AhR signaling (Esser et al., 2004). However, as expected, none of the measured expression profiles correlated with the 5F-203 GI_{50} data vector.

Differentially Expressed Genes Are Part of the Stress-Response Network. The genes from the sensitive and insensitive cell lines that were identified via the SAGE libraries as overexpressed in one cell type but not the other and examined by Q-RT-PCR are listed in Table 3. This set includes mostly structural/cytoskeletal and stress-related proteins. E-cadherin (CDH1) (Fig. 2) is an epithelial cell-cell adhesion glycoprotein that has been identified as a putative invasion suppressor molecule (Vleminckx et al., 1991) whose absence is believed to contribute to cancer progression and other disease states. Cytokeratin (KRT8) is an intermediate filament protein functioning in cytoskeletal and biogenic processes and is reported to provide resistance to fas-mediated apoptosis in hepatocytes (Gilbert et al., 2001). Trefoil factor (TFF1) is a protein expressed only in human breast cancers and is regulated by estrogen (Sun et al., 2005); one of its functions is to protect against insults to the epithelium. The X-box binding protein (XBP1) is a transcription factor that has been shown to be induced via an endoplasmic reticulum stress response. Loss of XBP1 has also been associated with increased sensitivity of transformed cells to killing by hypoxia and impaired tumor growth (Romero-Ramirez et al., 2004). These genes are part of one or more latent stress-response pathways associated with cytoskeletal and extracellular matrix (ECM) components. Stress signaling and mediation through the ECM and surface adhesion interactions can be linked to the expression of AHR, CYP1A1, and CYP1B1 (Larsen et al., 2004), which, in turn, seem jointly to be critical for the efficacy of 5F-203 (Brantley et al., 2004).

The SAGE-identified genes expressed at a higher level in the insensitive compared with the sensitive set of cells (Table 3) also share the common theme of stress-related signaling via the ECM. Annexin A1 (ANXA1) is located on the cytosolic face of the plasma membrane and can be involved in the

TABLE 3

Genes investigated using Q-RT-PCR across all 60 NCI cancer cell lines

All genes, except AHR, ESR1, and CD44, were indicated by SAGE as being specific either to 5F-203-insensitive or -sensitive cell lines. The differential expression pattern in these lines was also confirmed with RT-PCR measurements, except for the ANXA2, CTGF, HSPD1, and TRA1 genes. The expression pattern across the 60 cell lines were also correlated with previous Affymetrix microarray data and the GI_{50} pattern of 5F-203. The seven genes ultimately selected as those that represent a pattern that may aid in the classification of 5F-203-sensitive cell lines are shown in bold.

Gene	Name	PCC	
		with Affymetrix Microarray	with 5F-203 GI_{50} Profile
AHR	Aryl hydrocarbon receptor	−0.18	0.10
ANXA1	Annexin A1	0.46	0.09
ANXA2	Annexin A2	0.36	−0.15
AXL	AXL receptor tyrosine kinase	0.77	−0.05
CAV1	Caveolin 1	0.69	0.06
CD44	CD44 antigen	0.84	0.00
CDH1	Cadherin 1	0.80	−0.11
CTGF	Connective tissue growth factor	0.84	−0.37
ESR1	Estrogen receptor 1	N.A.	−0.04
HSPD1	60-kDa Heat shock protein	0.34	0.13
IGFBP7	Insulin-like growth factor binding protein 7	0.86	−0.19
KRT8	Keratin 8	0.89	0.20
LGALS1	Galectin 1	0.68	−0.16
SERPINE1	Serpin E1	0.77	−0.20
TFF1	Trefoil factor 1	0.88	0.21
TRA1	Tumor rejection antigen (gp96) 1	0.34	0.02
VIM	Vimentin	0.92	−0.21
XBP1	X-box binding protein 1	0.75	0.15

N.A., not available.

mediation of anti-inflammatory responses. It has also been reported to be down-regulated in breast cancer (Shen et al., 2005) and in head and neck tumors, in which the loss was associated with a more advanced stage of disease (Garcia Pedrero et al., 2004). AXL is a receptor tyrosine kinase that mediates signals from the ECM used to regulate proliferation and has been associated with a proliferative phenotype (Chung et al., 2003) and resistance to apoptosis (Melaragno et al., 2004). Caveolin (CAV) is the main scaffold protein found in caveolae plasma membranes and is a tumor suppressor gene candidate (Massimino et al., 2002). IGFBP7 can regulate insulin-like growth factors and has been identified as a potential tumor suppressor protein (Mutaguchi et al., 2003). LGALS1 (lectin) belongs to a family of β-galactoside binding proteins believed to affect cell-cell and cell-matrix interactions and may act as an autocrine negative growth factor that regulates cell proliferation. SERPINE1 (nexin) belongs to the plasminogen activator inhibitor class and has, among other things, been related to thrombosis and wound healing (Erickson et al., 1990). VIM is a cytoskeletal element within microfilaments and microtubules that stabilizes the architecture of the cytoplasm; secretion of this protein into the extracellular space has been linked to the immune response (Mor-Vaknin et al., 2003). Together, these genes are part of a latent, intermingled network of stress response and signaling pathways, and they were found to be differentially expressed in our initial subset of cell lines showing distinctive sensitivity to 5F-203. Although we cannot directly link these genes to the known mechanisms underlying the activity of this drug, we believe some or all of them may present a profile in which the relationship of their expression to each other carries characteristics of 5F-203-sensitive cell lines. It is possible that although the response to 5F-203 is believed to be critically dependent on competent aryl hydrocarbon receptor and cytochrome P450 pathways necessary to produce the active metabolite, these pathways may be coordinately linked to other stress-response networks that could act as surrogate markers of sensitivity to this agent. Because no constitutively expressed single gene from this set is capable of predicting sensitivity, we hypothesize that more complex gene relations, comprising a combination of expression profiles of constitutively expressed genes, might serve as a predictor of response.

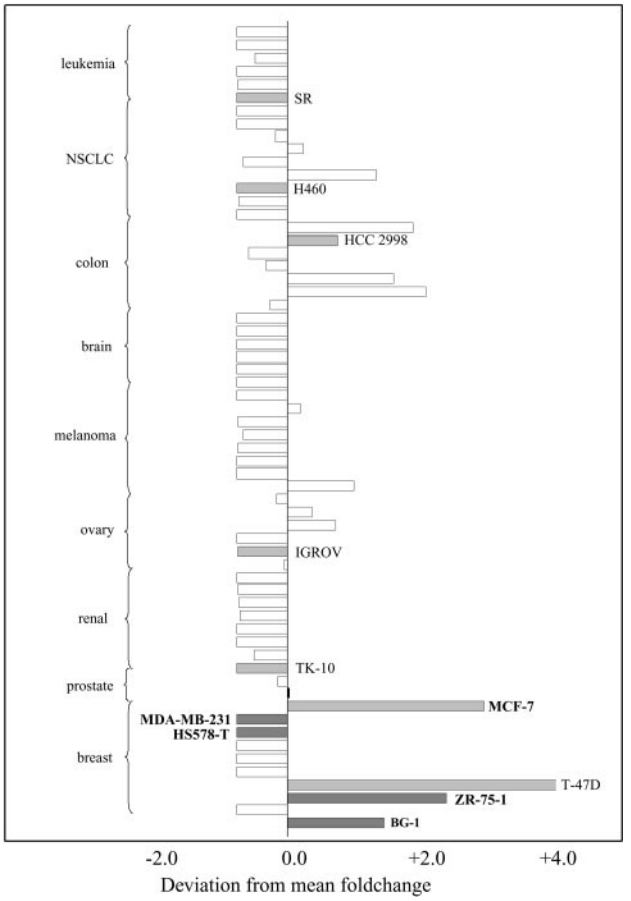

Fig. 2. Gene expression of E-cadherin (CDH1) is represented in a mean graph showing the relative basal expression of CDH1 in each cell line. To generate the mean graphs, the average CDH1 expression was calculated from the Q-RT-PCR data generated for each cell line using GAPDH as an internal standard, and then expressed as a function of the U251 CDH1 expression as an internal standard. The center line represents the average CDH1 expression; bars projecting to the right indicate higher than average expression (relative to the length of the bar), and bars projecting to the left indicate lower than average expression. Drug-sensitive cell lines with $GI_{50} < 10^{-7}$ M are shown in light gray, the cell lines that were SAGE-profiled are indicated with boldface cell names, and corresponding SAGE-profiled-insensitive cell lines are shown in dark gray.

Model Assessment. Using the assumption that the underlying difference in sensitivity of cancer cells to 5F-203 is a reflection of the cell's genetic makeup, the identification of a set of genes that are significantly differentially expressed in the sensitive and insensitive population may be used to classify the outcome of drug-treated cells. Even though these genes do not seem to relate directly to the critical activation mechanisms involving the aryl hydrocarbon receptor and the induction of cytochrome P450 genes, they may implicate an inherent stress-response pathway that characterizes 5F-203 sensitivity from yet-to-be-discovered mechanisms. The differentially expressed genes identified via the SAGE procedure were quantitatively measured using RT-PCR techniques and were normalized with GAPDH as an internal standard, and then they were referenced to the control EKVX lung-cancer cell line. The drug concentrations at which cellular growth is inhibited by 50% across the 60 immortalized tumor cell lines were obtained from the DTP data repository and were used as a measure of drug sensitivity (Fig. 1). Among the most sensitive cell lines in this panel are breast cancer cell lines T-47D and MCF-7, renal TK-10, ovarian IGROV1, and colon HCC-2998. Figure 3 gives the RT-PCR expression data for seven of the investigated genes (columns), in which the cells (rows) are ordered according to descending 5F-203 sensitivity. The patterns formed by these seven genes represent but one part of the total gene expression pattern that governs response. Our initial analysis indicated that none of these genes was individually associated with sensitivity at a sufficient level of significance (Table 3), and no single pattern of genes up- or down-regulated seemed predictive. This is visually evident by inspecting Fig. 3, showing that there is no coherent pattern associated with the most sensitive cell lines (at the top), from the bulk of the insensitive cell lines (at the bottom). Taken together, these results suggest that an analysis of more complex associations between genes, allowing for the identification of a small subset whose expression pattern relative to each other can be characterized, might support a model useful for identification of drug-sensitive cell lines.

The sensitivity of a tumor cell line to 5F-203 was then

modeled as a parametric function depending on the relative gene expression for selected genes. In practice, one would like to achieve a robust prediction using as few genes as possible by selecting the best subgroup of genes from Table 3. Included in the set of genes considered were members that were not differentially expressed but were believed to be associated with or involved in the activation and subsequent action of 5F-203. At first, all 18 genes in Table 3 were considered candidates; however, by assessing all possible combinations of two and three genes, the six genes that occurred least often in optimized solutions were deselected. Thus AHR, ANXA2, CD44, CTGF, ESR1, and VIM were not considered further. It is interesting to note that the three genes not selected by SAGE but included because of preconceived associations in 5F-203 mechanism of action, AHR, ESR1, and CD44, were all eliminated as predictive candidate genes. Using the 12 remaining gene data sets, all possible combinations of smaller subgroups of genes were considered and were used to model the GI_{50} pattern of 5F-203. The PCC between the modeled and the actual GI_{50} values was used to monitor a model's performance. Figure 4 shows the resultant correlation coefficient for the optimal selection of genes for each gene group containing fewer than 12 members. For the linear model, both the algebraic solution and a fitted optimal solution are given. The optimal solution for reproducing the GI_{50} pattern is given by a combination of seven genes in the $S_{\text{nonlinear}}$ model, albeit not dramatically better than the fitted linear model. Given other considerations, such as complexity and the number of model parameters, the linear fit may be a viable alternative. In the following, we consider only the optimal $S_{\text{nonlinear}}$ model. The genes selected from the optimal values and the actual parameter values are noted in Table 4, and their corresponding expression patterns, as measured by RT-PCR, are shown in Fig. 3. The actual as well as modeled prediction values of sensitivity are also indicated in the last two columns of this figure

The differential measured by SAGE in the selected genes was confirmed with Q-RT-PCR, with the exception of the *TRA1* gene, which, in contrast to the SAGE data, was measured as more highly expressed in the ZR-75-1 cell line (sensitive) than the HS578T cell line (insensitive). The statistically selected genes divide into one group of four genes that were overexpressed in the insensitive cell lines (ANXA1, TRA1, AXL1, and IGFBP7) and one group of three genes that were overexpressed in sensitive cell lines (KRT8, XBP1, and CDH1).

To investigate whether the selected genes carried any true specificity for reproducing the GI_{50} profile of 5F-203, the same set of gene expression profiles was used to optimize the $S_{\text{nonlinear}}$ model to other drug toxicity profiles. These results are given in Table 5. Three drugs that are believed to inhibit cellular growth by mechanisms other than 5F-203 were used, including paclitaxel (an antitubulin agent), 5-FU (a DNA antimetabolite), and bleomycin (a DNA-damaging agent). Whereas our subset of seven genes could capture a substantial part of these GI_{50} profiles, the information content ($\sim r^2$) is higher, up to a factor of 2, for the 5F-203 data. On the other hand, a close analog of 5F-203, the iodine-substituted compound showed an equally good fit to that of the 5F-203 data. Aminoflavone, a compound that is also activated through the AhR/CYP1 mechanism (Kuffel et al., 2002), scored slightly higher in the optimized fit than those with totally unrelated mechanisms of action. The goodness of fit is closely reflected in the PCC between the GI_{50} profiles of the drugs to that of 5F-203. Because these GI_{50} profiles themselves are a biolog-

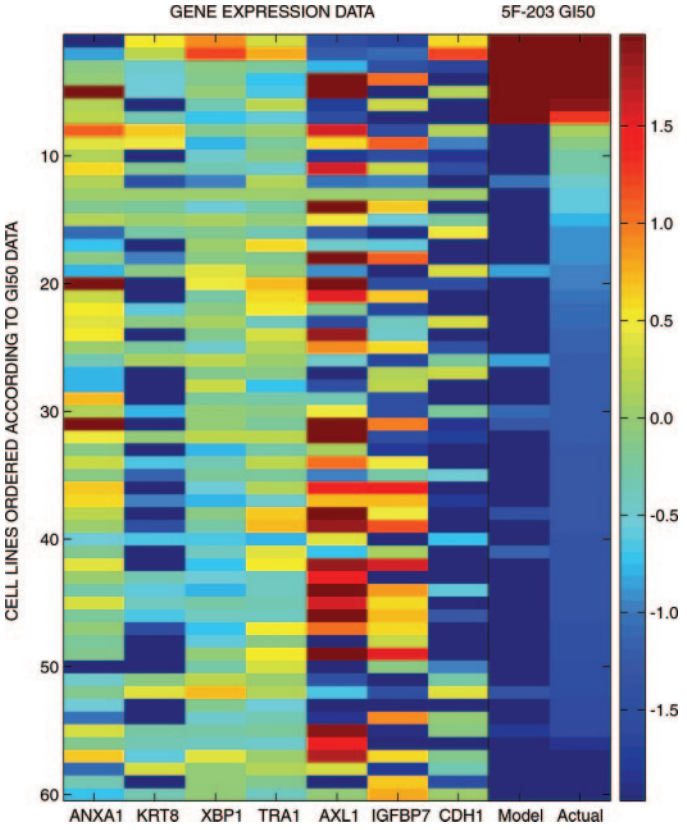

Fig. 3. Differential expression matrix for seven genes (columns) determined via RT-PCR for NCI's 60 tumor cell lines (rows) color-coded relative to the same internal standard, where red indicates overexpression and blue, underexpression. The cell lines have been ordered according to their actual sensitivity to 5F-203 given in the last column, with the predicted sensitivity according to the model being shown in the penultimate column labeled "model". The color bar at the right refers to the relative expression values, which are given on a logarithmic scale.

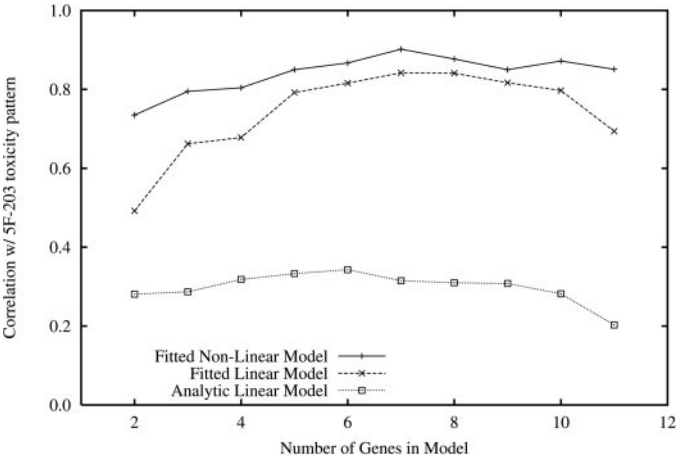

Fig. 4. Properties of optimally fitted gene combinations to reproduce the GI_{50} pattern of 5F-203. At each fixed total number of genes, we fit the model using all possible combinations of genes from the total set of 12 genes. There are M!/(M − N)!N! possible combinations of selecting a unique set of N genes from a pool of M genes. The best correlation of the model with the GI_{50} profile of 5F-203 is shown. The correlations have a maximum value around six to seven genes; for fewer or greater number of genes, the quality of the fit is either similar or decreased slightly.

ical readout of growth inhibition mechanism from drug action, it is plausible to infer that the selected gene set may be connected to the cellular response of 5F-203, although these mechanisms remain to be determined.

A further test of whether the selected gene set is somehow characteristic for the 5F-203 responses uses the RT-PCR data for genes that were deselected at the beginning of this section, AHR, ANXA2, CD44, CTGF, ESR1, SERPINE1, and VIM. Because AHR, CD44, and ESR1 are not differentially expressed in sensitive and insensitive cell lines, and ANXA2 could not be confirmed via RT-PCR, this gene set should perform more poorly in modeling the GI_{50} profile of 5F-203. These results are given in the last column of Table 5 and do indeed reflect a diminished capacity of the gene expression profiles to model 5F-203 sensitivity. This gene set seems to predict equally well for the entire set of investigated drug profiles, except for 5-FU, and thus shows no specificity to 5F-203. This is consistent with the notion that because these genes are not constitutively differentially expressed between sensitive and insensitive cell lines, they should be less able to differentiate specific cellular sensitivity to 5F-203. Even though the cell lines used to develop and identify differentially expressed genes were derived from breast and ovarian cancer tumors, these genes may have the capability of predicting sensitivity of other tumor cells from different tissues of origin. Further statistical assessment of the model is given as Supplemental Data S1 and in Supplemental Figure S1.

Probability of Sensitive Response from Gene Expression Patterns. Although this particular model remains applicable only to cell lines, we can use it to develop alternative perspectives about evaluating a population of cells for 5F-203 sensitivity. Thus, an extension of our approach would be to ask whether a particular pattern of more or less highly expressed genes has predictive power (e.g., can we assign a probability of response, p_r, for the case of a cell having up-regulated ANXA1 and XBP1 expressions and down-regulation of CDH1 expression?). In its simplest form, we can interrogate the model by assigning limits to what we can realistically call underexpression ($\downarrow$), no change in expression ($\circ$), or overexpression ($\uparrow$) relative to the standard used in the RT-PCR experiments. With this simplified model, each of the seven genes identified during our parametric model development and given in Table 4 can be characterized by these three states, up, down, or no change (see *Materials and Methods*). To calculate the probability of a response for a particular set of genes across these three possibilities, we use our developed model to estimate the number of sensitive and insensitive responses for all combinations of gene expressions seen in the RT-PCR data. The best combination when using all seven genes is ANXA1($\uparrow$), KRT8($\downarrow$), XBP1($\circ$), TRA1($\downarrow$), AXL1($\downarrow$), IGFBP7($\circ$), CDH1($\downarrow$), which is associated with a p_r value of 0.52; that is, if we identified all cells from a large population of cells that have this expression pattern and selected them, 52% of those cells would show sensitivity to 5F-203. Other combinations of expression patterns for these seven genes are less predictive of sensitivity; for example, the original pattern identified from the SAGE data as carrying possible genetic information to discriminate sensitive and insensitive cell lines is associated with a p_r value of only 0.20.

We can also look at combinations of fewer than seven genes. If we use a maximum of six genes (e.g., by excluding AXL1), we can again calculate the fraction of sensitive and insensitive cells for all combinations of the six remaining genes. The largest of these fractions (i.e., the best predictor of 5F-203 sensitivity), max p_r, is given in Table 6 as a function of the number of genes considered. Thus, for seven genes, we recover the result of 52% discussed above. In the consideration of six genes, the best pattern to retrieve sensitive cell lines is (ANXA1($\uparrow$), KRT8($\downarrow$), XBP1($\circ$), TRA1($\downarrow$), IGFBP7($\circ$), CDH1($\downarrow$), with a 50% success rate, which is not significantly different from the 52% retrieved with the best pattern of seven. The difference between the patterns is just the removal of the *AXL1* gene for consideration, and all other up ($\uparrow$), down ($\downarrow$), or same ($\circ$) expression criteria remains the same. Further reduction of the number of genes assayed decreases the ability of the best pattern to successfully recognize 5F-203 sensitive cells. At the minimum of deriving a probability from the status of one gene, *XBP1*, only 10% of the cells that overexpress this gene are sensitive to 5F-203.

TABLE 4

Model parameters

The model is fit to the measured GI_{50} concentration of 60 cell lines using the RT-PCR-measured relative gene expression of seven genes in these same cell lines. Thus, the fitting procedure used 420 data points using 14 parameters to describe the 60 possible outcomes.

	ANXA1	KRT8	XBP1	TRA1	AXL1	IGFBP7	CDH1
c_t	+1.04	−0.10	−0.45	−1.19	+0.02	−0.46	−0.37
d_t	+0.12	−0.17	+0.69	+0.37	−0.06	−0.45	−0.22

TABLE 5

Properties of the nonlinear model to reproduce GI_{50} patterns for some selected drugs

The table gives the correlation between the fitted predictions of cellular GI_{50} values to the true GI_{50} values and to the individual drug GI_{50} values in the second and third column. Compounds (5I-203 and aminoflavone) that show a significant correlation with the 5F-203 toxicity profile can be fitted better using the seven selected genes than those with low or insignificant correlations. The last entries show the correlation value of the true GI_{50} values to the nonlinear model predictions when using only the seven genes that were originally deselected as unsuitable from the gene set shown in Table 3.

Drug	PCC		
	Fitted $S(\vec{e})$ with GI_{50} Profile r (r^2)	GI_{50} Profile with 5F-203	Fitted $S(\vec{e})$ with GI_{50} Profile Using Deselected Genes r (r^2)
5F-203	0.90 (0.80)	1.00	0.65 (0.42)
5I-203	0.90 (0.80)	0.63	
Paclitaxel	0.73 (0.53)	0.04	0.44 (0.19)
5-FU	0.65 (0.42)	0.21	0.29 (0.08)
Bleomycin	0.73 (0.53)	−0.04	0.66 (0.44)
Aminoflavone	0.77 (0.59)	0.38	0.68 (0.46)

Increasing the number of cells considered in the assay thus greatly improves the ability to discern sensitive and insensitive cells. Because the status of the seven genes used here is not known a priori in the population of all cell types, and because all combinations can not be considered equally probable, it is hard to gauge whether a reductionist approach of using a simplified expression status (up, down, and same) to describe a heterogeneous gene population is inherently worse than using the full model, which requires accurate RT-PCR measurements for each gene.

Discussion

Genetic markers for disease and disease states commonly refer to the presence or absence of functional proteins due to genetic causes, a result that is often inferred to reflect the abnormal overexpression of a gene as the cause of excessive appearance of the fully functional gene product. Accepting this viewpoint, measurements of gene expression are currently being used in many aspects of classifying disease state and prognosis. Conversely, cell lines representing different diseases can be classified according to their expression patterns or profiles. As an example, within a panel of immortalized tumor cell lines, characteristics of each disease type's gene expression profile provide a biological readout of the biochemical processes relevant to the maintenance of the tumor cell (Alizadeh et al., 2000; Scherf et al., 2000; Covell et al., 2003). The absence of cellular mRNA is typically not a result of specific DNA mutations preventing translation but rather is a reflection of many factors affecting gene regulation within the cell. As one might expect, cells observed under different conditions will also have different gene expression profiles. Moreover, although tumor cell lines represent a viable foundation for understanding the complexities of genetic interactions, ultimately it will be important to extend these studies to clinical tumor material and its normal counterpart to determine whether this type of approach could aid in predicting differential host toxicity.

In the specific case of cell line-based 5F-203 sensitivity, the five cell types queried by SAGE on the basis of gene expressions were found to have a consistent pattern of over- and underexpressed genes. This observation is not likely to be a result of genetic mutations in each cell's DNA, but rather reflects differences in gene regulation particular to each cell. The genes that were observed to be differentially expressed (with a tolerance) by either the 5F-203 sensitive or insensitive cell lines are thus postulated to be regulated differently within the cellular environment of sensitive or insensitive cell lines. Because we do not know the entire suite of processes that are necessary to determine sensitivity, we cannot postulate that all other cell lines are regulated in the same manner. For example, other control processes may link two of the identified genes to a third gene that remains hidden and whose status and effect may be different in other cell lines. Hidden and higher-order correlations among gene control elements do not lend themselves to easy interpretations. In fact, the original pattern used in the SAGE analysis to identify consistently regulated genes has only 20% predictability when considering the complete NCI-60 tumor cell lines, and the maximum achievable prediction with a single gene is only 10%. These observations reinforce the notion that we are truly looking at a complex phenomenon. If the expression status of the genes does contain information on cellular sensitivity and insensitivity to 5F-203, facile ways will be needed to connect these properties without full knowledge of the true processes. This necessitates a drastic reduction in complexity and the introduction of a model that mirrors our understanding and captures the true processes reflective of 5F-203 sensitivity and insensitivity. A minimal feature of such a model is the ability to simultaneously consider more than one gene at a time. Optimizing parameters for these models based on GI_{50} values and relative gene expression of those genes identified by SAGE exhibit nontrivial properties. In fact, the proposed models show that increasing the number of genes improves the ability to classify response, but improvement plateaus at only five to six gene expression profiles. A perfect model that reflects all cellular processes would show a continuous improvement because one includes more gene expression information. Our model does contain the expected initial improvement, but it cannot effectively use more information beyond that contained in approximately five to six genes. This defines the limit of the model, per se. The ability to classify 5F-203 sensitivity is dependent on the genes selected in the model development phase; the selection of genes not identified by SAGE as a characteristic component of cellular sensitivity was found to be less accurate in predicting response. In addition, other cytotoxicity patterns derived from drugs that were not related to the mechanism of 5F-203 could not predict with equivalent fidelity when using the selected gene set, as identified by SAGE. As an internal check of our model effort, cytotoxicity patterns from a drug with a mechanism of action closely related to 5F-203 could be modeled equally well as 5F-203. These properties of gene and drug specificity are commensurate with a model that takes advantage of specific genes to model a specific outcome. As

TABLE 6

Probability of achieving a sensitive response to 5F-203

The maximum probability, max p_r, of achieving a sensitive response to 5F-203 as a function of the number of genes investigated from the selected gene set. Three states of gene expression were considered: expression larger than standard (↑), expression less than standard (↓), and expression the same as standard (○). An asterisk indicates that the expression status of that gene is not relevant for max p_r at the specified number of genes (N). Thus, the best combination of gene expression patterns for four genes is ANXA1 overexpressed, TRA1 and IGFBP7 underexpressed, and XBP1 expression unchanged relative to standard. If this combination is observed, 33% of the cell population with that expression pattern will have a sensitive response.

No	ANXA1	KRT8	XBP1	TRA1	AXL1	IGFBP7	CDH1	max p_r
1	*	*	↑	*	*	*	*	0.10
2	↓	*	*	↑	*	*	*	0.14
3	↑	*	*	↓	*	↓	*	0.21
4	↑	*	○	↓	*	↓	*	0.33
5	↑	*	○	↓	*	↓	○	0.38
6	↑	↓	○	↓	*	○	↓	0.50
7	↑	↓	○	↓	↓	○	↓	0.52

such, this approach is able to capture the connection between gene expression and cellular sensitivity to 5F-203, but of course it does not explain the connection, and it is not a unique, model solution to the problem. An application of a typical nonparametric model or decision tree to predict outcome could use, for example, a particular pattern of "up" and "down" expression signatures. One such pattern would be the particular up/down signature used in analyzing the SAGE data to identify genes characteristic of 5F-203 sensitivity. We saw that this gene pattern could only predict sensitivity with 20% accuracy, and the best possible pattern had a maximum predictability of 52%.

To summarize, SAGE libraries were constructed for five cell lines representing drug-sensitive and insensitive cells, plus an additional drug-treated, sensitive cell line. From examining differential SAGE tag expression, 32 genes were identified as significantly expressed by a 2-fold more (8) or less (24) in the drug-sensitive cell lines than in the drug-resistant cell lines. Based on those genes with the most extensive knowledge of function and role in the cell, 15 of them were selected for Q-RT-PCR analysis of expression in the cell lines of the NCI Anticancer Drug Screen for the development of a more detailed expression profile. None of the patterns of expression of these individual genes correlated significantly with drug-sensitivity profiles. A function of normalized RT-PCR expression of marker genes, however, was capable of predicting cell line sensitivity to 5F-203. This model used optimally weighted, exhaustively sampled combinations of normalized gene expression data to predict cellular sensitivity to 5F-203. Our procedure selected a set of seven genes whose gene products are primarily involved in stress response. The proposed model showed that those gene expressions were specific to the 5F-203-induced growth inhibition, because other GI_{50} drug profiles could not be modeled as well as that of 5F-203, and random assemblies of genes could not achieve the same specificity. The expression profiles of these genes have not before been implicated in the process of transporting or activating 5F-203.

SAGE provided a tool to determine differentially expressed genes in five cell lines with opposing drug-sensitivity profiles. Expression of 15 genes in NCI's 60 tumor cell lines provided preliminary results to attempt predictions of drug sensitivity based on patterns of gene expression. By using seven of these genes patterns, we were able to successfully model the 5F-203 response of both sensitive and resistant cell lines. This result supports the hypothesis that constitutive gene expression patterns, not necessarily associated with a known mechanism of action of a drug, can be successfully mined to characterize drug-responsive and nonresponsive cell lines.

References

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature (Lond)* **403:**503–511.

Amin RP, Hamadeh HK, Bushel PR, Bennett L, Afshari CA, and Paules RS (2002) Genomic interrogation of mechanism(s) underlying cellular responses to toxicants. *Toxicology* **181–182:**555–563.

Bradshaw TD, Bibby MC, Double JA, Fichtner I, Cooper PA, Alley MC, Donohue S, Stinson SF, Tomaszewjski JE, Sausville EA, et al. (2002) Preclinical evaluation of amino acid prodrugs of novel antitumor 2-(4-amino-3-methylphenyl)benzothiazoles. *Mol Cancer Ther* **1:**239–246.

Bradshaw TD, Shi DF, Schultz RJ, Paull KD, Kelland L, Wilson A, Garner C, Fiebig HH, Wrigley S, and Stevens MF (1998) Influence of 2-(4-aminophenyl)benzothiazoles on growth of human ovarian carcinoma cells in vitro and in vivo. *Br J Cancer* **78:**421–429.

Brantley E, Trapani V, Alley MC, Hose CD, Bradshaw TD, Stevens MF, Sausville EA, and Stinson SF (2004) Fluorinated 2-(4-amino-3-methylphenyl)benzothiazoles induce CYP1A1 expression, become metabolized and bind to macromolecules in sensitive human cancer cells. *Drug Metab Dispos* **32:**1392–1401.

Brent R (1973) *Algorithms for Finding Zeros and Extrema of Functions without Calculating Derivatives.* Prentice-Hall, NJ.

Chua MS, Kashiyama E, Bradshaw TD, Stinson SF, Brantley E, Sausville EA, and Stevens MF (2000) Role of Cyp1A1 in modulation of antitumor properties of the novel agent 2-(4-amino-3-methylphenyl)benzothiazole (DF 203, NSC 674495) in human breast cancer cells. *Cancer Res* **60:**5196–5203.

Chung BI, Malkowicz SB, Nguyen TB, Libertino JA, and McGarvey TW (2003) Expression of the proto-oncogene Axl in renal cell carcinoma. *DNA Cell Biol* **22:**533–540.

Covell DG, Wallqvist A, Rabow AA, and Thanki N (2003) Molecular classification of cancer: Unsupervised self-organizing map analysis of gene expression microarray data. *Mol Cancer Ther* **2:**317–332.

Dan S, Tsunoda T, Kitahara O, Yanagawa R, Zembutsu H, Katagiri T, Yamazaki K, Nakamura Y, and Yamori T (2002) An integrated database of chemosensitivity to 55 anticancer drugs and gene expression profiles of 39 human cancer cell lines. *Cancer Res* **62:**1139–1147.

Durst B, Sorg RV, Roder G, Betz B, Beckmann MW, Niederacher D, Bender HG, and Dall P (2001) The influence of hormones on CD44 expression in endometrial and breast carcinomas. *Oncol Rep* **8:**987–993.

Erickson LA, Fici GJ, Lund JE, Boyle TP, Polites HG, and Marotti KR (1990) Development of venous occlusions in mice transgenic for the plasminogen activator inhibitor-1 gene. *Nature (Lond)* **346:**74–76.

Esser C, Temchura V, Majora M, Hundeiker C, Schwarzler C, and Gunthert U (2004) Signaling via the AHR leads to enhanced usage of CD44v10 by murine fetal thymic emigrants: possible role for CD44 in emigration. *Int Immunopharmacol* **4:**805–818.

Fichtner I, Monks A, Hose C, Stevens MF, and Bradshaw TD (2004) The experimental antitumor agents Phortress and doxorubicin are equiactive against human-derived breast carcinoma xenograft models. *Breast Cancer Res Treat* **87:**97–107.

Garcia Pedrero JM, Fernandez MP, Morgan RO, Herrero Zapatero A, Gonzalez MV, Suarez Nieto C, and Rodrigo JP (2004) Annexin A1 down-regulation in head and neck cancer is associated with epithelial differentiation status. *Am J Pathol* **164:**73–79.

Gilbert S, Loranger A, Daigle N, and Marceau N (2001) Simple epithelium keratins 8 and 18 provide resistance to Fas-mediated apoptosis. The protection occurs through a receptor-targeting modulation. *J Cell Biol* **154:**763–773.

Hose CD, Hollingshead M, Sausville EA, and Monks A (2003) Induction of CYP1A1 in tumor cells by the antitumor agent 2-[4-amino-3-methylphenyl]-5-fluorobenzothiazole: a potential surrogate marker for patient sensitivity. *Mol Cancer Ther* **2:**1265–1272.

Keskin O, Bahar I, Jernigan RL, Beutler JA, Shoemaker RH, Sausville EA, and Covell DG (2000) Characterization of anticancer agents by their growth inhibitory activity and relationships to mechanism of action and structure. *Anticancer Drug Des* **15:**79–98.

Kramer JA and Kolaja KL (2002) Toxicogenomics: an opportunity to optimise drug development and safety evaluation. *Expert Opin Drug Saf* **1:**275–286.

Kuffel MJ, Schroeder JC, Pobst LJ, Naylor S, Reid JM, Kaufmann SH, and Ames MM (2002) Activation of the antitumor agent aminoflavone (NSC 686288) is mediated by induction of tumor cell cytochrome P450 1A1/1A2. *Mol Pharmacol* **62:**143–153.

Larsen MC, Brake PB, Pollenz RS, and Jefcoate CR (2004) Linked expression of Ah receptor, ARNT, CYP1A1 and CYP1B1 in rat mammary epithelia, in vitro, is each substantially elevated by specific extracellular matrix interactions that precede branching morphogenesis. *Toxicol Sci* **82:**46–61.

Loaiza-Perez AI, Trapani V, Hose C, Singh SS, Trepel JB, Stevens MF, Bradshaw TD, and Sausville EA (2002) Aryl hydrocarbon receptor mediates sensitivity of MCF-7 breast cancer cells to antitumor agent 2-(4-amino-3-methylphenyl) benzothiazole. *Mol Pharmacol* **61:**13–19.

Massimino ML, Griffoni C, Spisni E, Toni M, and Tomasi V (2002) Involvement of caveolae and caveolae-like domains in signalling, cell survival and angiogenesis. *Cell Signal* **14:**93–98.

Melaragno MG, Cavet ME, Yan C, Tai LK, Jin ZG, Haendeler J, and Berk BC (2004) Gas6 inhibits apoptosis in vascular smooth muscle: role of Axl kinase and Akt. *J Mol Cell Cardiol* **37:**881–887.

Monks A, Scudiero D, Skehan P, Shoemaker R, Paull K, Vistica D, Hose C, Langley J, Cronise P, Vaigro-Wolff A, et al. (1991) Feasibility of a high-flux anticancer drug screen using a diverse panel of cultured human tumor cell lines. *J Natl Cancer Inst* **83:**757–766.

Mor-Vaknin N, Punturieri A, Sitwala K, and Markovitz DM (2003) Vimentin is secreted by activated macrophages. *Nat Cell Biol* **5:**59–63.

Mutaguchi K, Yasumoto H, Mita K, Matsubara A, Shiina H, Igawa M, Dahiya R, and Usui T (2003) Restoration of insulin-like growth factor binding protein-related protein 1 has a tumor-suppressive activity through induction of apoptosis in human prostate cancer. *Cancer Res* **63:**7717–7723.

Paull KD, Shoemaker RH, Hodes L, Monks A, Scudiero DA, Rubinstein L, Plowman J, and Boyd MR (1989) Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J Natl Cancer Inst* **81:**1088–1092.

Rabow AA, Shoemaker RH, Sausville EA, and Covell DG (2002) Mining the National Cancer Institute's tumor-screening database: identification of compounds with similar cellular activities. *J Med Chem* **45:**818–840.

Romero-Ramirez L, Cao H, Nelson D, Hammond E, Lee AH, Yoshida H, Mori K, Glimcher LH, Denko NC, Giaccia AJ, et al. (2004) XBP1 is essential for survival under hypoxic conditions and is required for tumor growth. *Cancer Res* **64:**5943–5947.

Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, et al. (2000) A gene expression database for the molecular pharmacology of cancer. *Nat Genet* **24:**236–244.

Shen D, Chang HR, Chen Z, He J, Lonsberry V, Elshimali Y, Chia D, Seligson D, Goodglick L, Nelson SF, et al. (2005) Loss of annexin A1 expression in human breast cancer detected by multiple high-throughput analyses. *Biochem Biophys Res Commun* **326:**218–227.
Sorbello V, Fuso L, Sfiligoi C, Scafoglio C, Ponzone R, Biglia N, Weisz A, Sismondi P, and De Bortoli M (2003) Quantitative real-time RT-PCR analysis of eight novel estrogen-regulated genes in breast cancer. *Int J Biol Markers* **18:**123–129.
Sun JM, Spencer VA, Li L, Yu Chen H, Yu J, and Davie JR (2005) Estrogen regulation of trefoil factor 1 expression by estrogen receptor alpha and Sp proteins. *Exp Cell Res* **302:**96–107.
Thomas RS, Rank DR, Penn SG, Zastrow GM, Hayes KR, Hu T, Pande K, Lewis M, Jovanovich SB, and Bradfield CA (2002) Application of genomics to toxicology research. *Environ Health Perspect* **110 (Suppl 6):**919–923.
Velculescu VE, Zhang L, Vogelstein B, and Kinzler KW (1995) Serial analysis of gene expression. *Science (Wash DC)* **270:**484–487.
Vleminckx K, Vakaet L Jr, Mareel M, Fiers W, and van Roy F (1991) Genetic manipulation of E-cadherin expression by epithelial tumor cells reveals an invasion suppressor role. *Cell* **66:**107–119.
Wallqvist A, Rabow AA, Shoemaker RH, Sausville EA, and Covell DG (2002) Establishing connections between microarray expression data and chemotherapeutic cancer pharmacology. *Mol Cancer Ther* **1:**311–320.
Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, et al. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res* **31:**28–33.
Yamamoto M, Wakatsuki T, Hada A, and Ryo A (2001) Use of serial analysis of gene expression (SAGE) technology. *J Immunol Methods* **250:**45–66.

Address correspondence to: Dr. Anders Wallqvist, Science Applications International Corporation, NCI-Frederick, P.O. Box B, Frederick, MD 21702. E-mail: wallqvis@ncifcrf.gov