A gene expression database for the molecular pharmacology of cancer

Uwe Scherf[1,8], Douglas T. Ross[2], Mark Waltham[1], Lawrence H. Smith[1], Jae K. Lee[1], Lorraine Tanabe[1], Kurt W. Kohn[1], William C. Reinhold[1], Timothy G. Myers[4], Darren T. Andrews[1], Dominic A. Scudiero[5], Michael B. Eisen[3], Edward A. Sausville[6], Yves Pommier[1], David Botstein[3], Patrick O. Brown[2,7] & John N. Weinstein[1]

We used cDNA microarrays to assess gene expression profiles in 60 human cancer cell lines used in a drug discovery screen by the National Cancer Institute. Using these data, we linked bioinformatics and chemoinformatics by correlating gene expression and drug activity patterns in the NCI60 lines. Clustering the cell lines on the basis of gene expression yielded relationships very different from those obtained by clustering the cell lines on the basis of their response to drugs. Gene-drug relationships for the clinical agents 5-fluorouracil and *L*-asparaginase exemplify how variations in the transcript levels of particular genes relate to mechanisms of drug sensitivity and resistance. This is the first study to integrate large databases on gene expression and molecular pharmacology.

Introduction

Gene expression profiles can be assessed for human tumours, but from the pharmacological perspective, there is a problem: the associated treatment histories, if any, are generally complex, fragmentary and difficult to interpret. Here we describe studies using cDNA microarrays to assess gene expression profiles in a set of 60 human cancer cell (NCI60) lines that, in contrast to clinical tumours, have been characterized pharmacologically by treatment with more than 70,000 different agents, one at a time and independently. These cells are used by the Developmental Therapeutics Program (DTP) of the National Cancer Institute (NCI) to screen potential anticancer drugs[1–6]. Screening the compounds for activity also profiles the cells for sensitivity, offering us a unique opportunity to relate variations in gene expression to the molecular pharmacology of cancer. The accompanying report by Ross *et al.*[7] describes how gene expression profiles characterize patterns of phenotypic variation in the 60 cancer cell types; here we analysed gene expression patterns from the same experiments for their relationship to drug sensitivity. Note that the gene expression patterns are those for untreated cells, and that this study focuses on sensitivity to therapy rather than on the molecular consequences of therapy. This pharmacogenomic analysis is analogous to the assessment of molecular markers in the tumours of untreated patients. Analytical tools and data are available (http://discover.nci.nih.gov and http://genome-www.stanford.edu/nci60), as are additional data from the drug screen (http://dtp.nci.nih.gov).

The NCI60 set includes cell lines derived from cancers of colorectal, renal, ovarian, breast, prostate, lung and central nervous system origin, as well as leukaemias and melanomas. Growth inhibition is assessed from changes in total cellular protein after 48 hours of drug treatment using a sulphorhodamine B assay. The endpoint is non-specific, but patterns of drug activity across the cell lines provide information on mechanisms of drug action, resistance and modulation[8–12]. These patterns have been correlated with molecular structure descriptors of the tested compounds[13,14] and with molecular characteristics (for example, MDR1 levels and p53 status) of the test cells[8,15–26]. Previously, most cell characteristics were assessed one gene, gene product or molecular pathway at a time, but we have adopted a more comprehensive approach[27] that generates information on large numbers of gene products simultaneously. We first generated a protein-expression database using two-dimensional gel electrophoresis[28]. Here, and in the accompanying paper[7], we present the corresponding mRNA expression database of the cell lines generated using pin-spotted, PCR-amplified cDNA microarrays on glass slides[29–31].

A schematic view of our overall approach is shown (Fig. 1). Activity patterns in database A (>70,000 compounds tested against 60 cell lines) have been correlated with mRNA expression levels in database T_r (9,703 cDNAs representing ~8,000 unique genes in 60 cell lines). As signposts for interpretation of the gene expression profiles, we included in the analysis other molecular characteristics (termed 'targets') individually assessed by various laboratories, as represented in database T_i (40 targets in 60 cell lines, see http://dtp.nci.nih.gov/docs/dtp_data.html). But before exploring the drug-gene correlations, it will be necessary to examine gene expression and drug sensitivity relationships separately.

[1]*Laboratory of Molecular Pharmacology, Division of Basic Sciences, Building 37/5D-02, National Cancer Institute (NCI), National Institutes of Health (NIH), Bethesda, Maryland, USA.* [2]*Department of Biochemistry, Stanford University School of Medicine, Stanford, California, USA.* [3]*Department of Genetics, Stanford University School of Medicine, Stanford, California, USA.* [4]*Information Technology Branch, Developmental Therapeutics Program (DTP), Division of Cancer Treatment and Diagnosis (DCTD), NCI, NIH, Bethesda, Maryland, USA.* [5]*SAIC-NCI-Frederick Cancer Research and Development Center, Frederick, Maryland, USA.* [6]*Office of the Associate Director, DTP, DCTD, NCI, NIH, Bethesda, Maryland, USA.* [7]*Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California, USA.* [8]*Present address: Gene Logic Inc., Gaithersburg, Maryland, USA. Correspondence should be addressed to J.N.W. (e-mail: weinstein@dtpax2.ncifcrf.gov) or P.O.B. (e-mail: pbrown@cmgm.stanford.edu).*

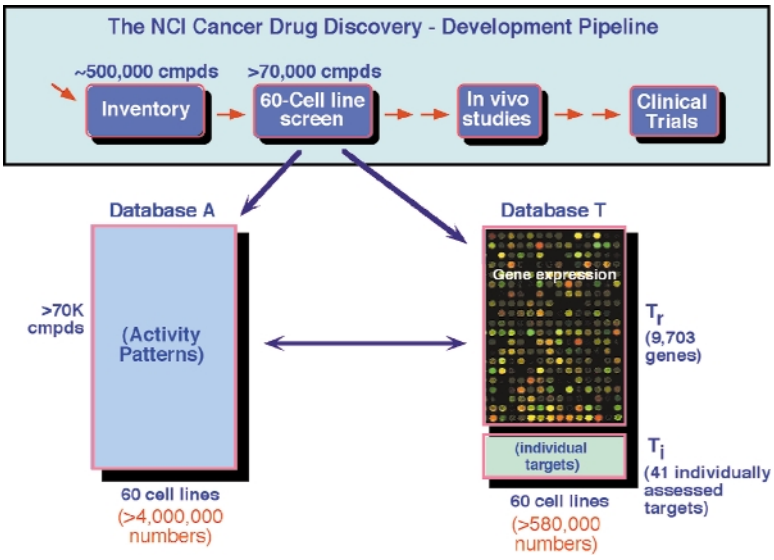

Fig. 1 Simplified schematic overview of database generation in relation to the NCI drug discovery program. Each row of the activity database (A) represents the pattern of activity of a particular compound across the 60 cell lines, and each column represents the pattern of sensitivities of a particular cell line to the compounds tested. The gene-expression database (T_r) contains fluorescence hybridization ratio values from two-colour cDNA microarray measurements on the 60 cell lines. The database of 40 individually assessed molecular targets (T_i) is the product of experiments in many different laboratories, as compiled at the DTP web site (http://dtp.nci.nih.gov). The union of T_r and T_i (as well as a protein database not considered here[28]) constitutes an overall database of molecular targets for analysis. Modified from ref. 8.

Results

Cell-cell correlations on the basis of gene expression profiles (T-matrix)

We applied selective filters to reduce the initial 9,703 gene spots to a 1,376-gene subset for the present analysis. These were the genes that showed strong patterns of variation among the cell lines and had less than or equal to 4 of 60 values excluded on the basis of visual quality control or low signal.

We performed cluster analyses using a variety of algorithms and metrics to organize the cell lines on the basis of gene expression pattern. The lines tended to cluster into groups that reflect their tissue of origin (Fig. 2*a*). With average linkage clustering and a correlation metric, the 1,376 genes, along with 40 individually assessed targets, yielded 11 distinct cell clusters differing in average inter-cluster correlation coefficient by more than 0.3. MDA-MB-435 (derived from the pleural effusion of a patient with breast cancer) and its Erb/B2 transfectant MDA-N expressed large numbers of genes characteristic of melanoma and clustered with the melanomas[7].

The MDA-MB435/MDA-N pair provides evidence of the reproducibility of these expression profiles. Because MDA-N does not generally express much Erb/B2 under non-selective growth conditions, the two lines can be considered as replicates cultured separately and processed independently. As indicated by the cluster tree (Fig. 2*a*), they are by far the most similar pair of cell lines. The Pearson correlation coefficient was 0.97 (with a bootstrap two-tail 95% confidence interval of 0.879–0.998). In contrast, the average correlation over all pairs (60×59/2=1,740) of lines was 0.30. This modest correlation reflects factors common to expression patterns in tumour cell lines. The median difference per gene between the two cell lines over the 1,376 genes was 0.21 $\log_{10}$ units, a factor of 1.62. To further test the reproducibility of the patterns, RNA samples from two cell lines (MCF7 breast and K562 leukaemia) were collected on three different occasions (at different passage numbers), then labelled, hybridized and scanned independently. These replicates (labelled MCF7 I, II and III, and K562 I, II and III) clustered side by side[7], with approximately the same degree of similarity as shown by the MDA-MB435/MDA-N pair.

Cell-cell correlations on the basis of drug activity profiles (A-matrix)

From the overall database of more than 70,000 chemical compounds tested, we selected 1,400 compounds for this analysis that had been tested at least four times on all or most of the 60 cell lines. We included most of the drugs currently in clinical use for cancer treatment. The final data set used for calculations (that is, one GI_{50} value for each drug-cell pair) included 1.64% sporadic missing values, 5.92% values censored at the high-concentration end of the range and 6.86% censored at the low-concentration end. The mean $-\log GI_{50}$ potency was 5.71 with a standard deviation of 1.79 and the median was 5.72 with an interquartile range of 4.36–7.00

We clustered the 60 cell lines using an average-linkage algorithm and a metric based on the growth inhibitory activities (GI_{50}) of the 1,400 compounds[8] (Fig. 2*b*). Comparison of Figs 2*a* and 2*b* indicates that the clustering by organ of origin was not as strong on the basis of activity as it was on the basis of gene expression. We observed 15 distinct branches at an average inter-cluster correlation coefficient of more than or equal to 0.3. Only two cell types tended to cluster on single branches: leukaemia (6/6) and CNS (5/6) cells. MDA-N and MDA-MB-435 clustered with three of the melanoma lines (M14, UACC-62 and SK-MEL-5). Breast cancer lines HS 578T, MDA-MB-231 and BT-549 clustered together, but far from lines T-47D and MCF7, which are positive for the oestrogen receptor. Ovarian and colon lines were considerably more heterogeneous in sensitivity to drugs than in gene expression.

This difference in clustering (Fig. 2*a*,*b*) was probably due, at least in part, to the activity of genes important to drug sensitivity and resistance. For example, several tumour cell lines known to express the multi-drug resistance gene *ABCB1* (formerly *MDR1*) had closely related drug-activity profiles. HCT-15, with one of the highest levels of *ABCB1* expression, is a colon-derived line that clustered by gene expression pattern with other colon lines but by activity pattern with NCI/ADR-Res, an *ABCB1*-expressing line selected for adriamycin resistance[32]. Likewise, ACHN, UO-31 and CAKI1, three renal-cancer cell lines known to express high levels of *ABCB1*, clustered on the same branch (Fig. 2*b*).

For quantitative comparison of the clusterings (Fig. 2*a*,*b*), we derived a correlation of correlation parameter, *r*, defined as the mean Pearson correlation coefficient of the Pearson correlation coefficients relating all possible pairs of cells in terms of their response to drugs and in terms of their gene expression. For these data sets, *r* was only 0.21. If these clusterings (Fig. 2*a*,*b*) had been identical, *r* would have been unity; if there had been no relationship at all, *r* would have been 0.

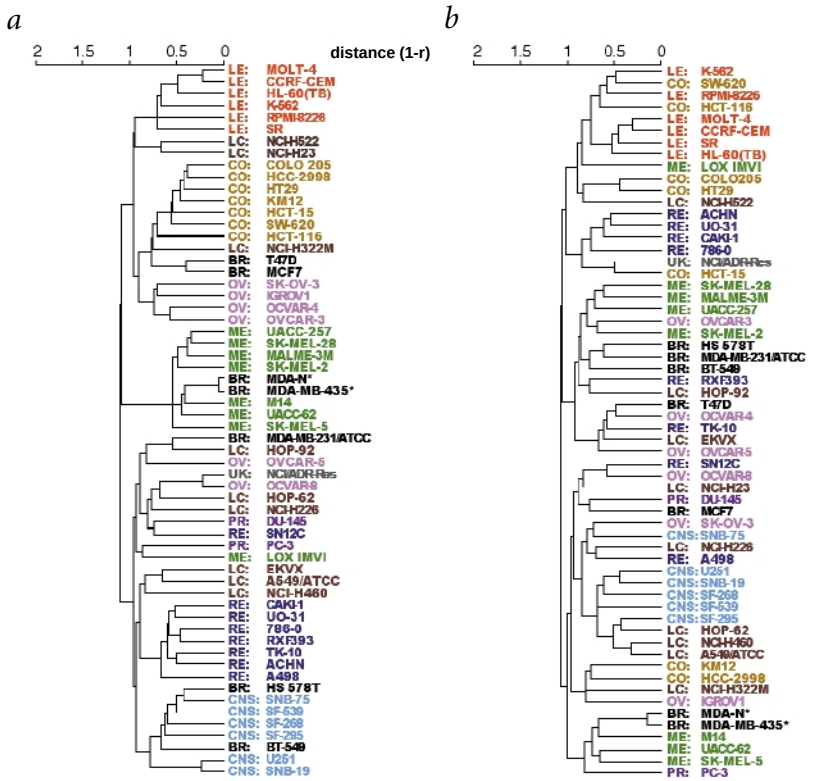

Fig. 2 Dendrograms showing average-linkage hierarchical clustering of human cancer cell lines. ***a***, Cluster tree of the 60 cell lines based on their gene expression profiles for 1,376 genes and 40 individual targets. All of the colon cancer lines (CO; 7/7), the CNS lines (CNS; 6/6) and the leukaemias (LE; 6/6) clustered together. Of eight melanoma lines (ME), seven clustered together, except the one reported to lack melanin production (LOX-IMVI; ref. 5). Of eight renal carcinoma lines (RE), seven clustered together, as did four of six ovarian lines (OV). Non-small-cell lung cancer cells (LC) clustered on two different branches, and those of breast origin (BR) appeared most heterogeneous. The breast cell lines positive for the oestrogen receptor, T-47D and MCF7, appeared together and grouped with the colon lines, whereas the breast cell lines negative for the oestrogen receptor, HS578T and BT-549, clustered with CNS malignancies. NCI/ADR-Res is of unknown origin (UK). ***b***, Cluster tree for the cells based on their patterns of sensitivity to 1,400 compounds tested. The colour of the cell line name indicates its assigned organ of origin classification. The distance metric used was (1–Pearson correlation coefficient). *Two cell lines (MDA MB435 and MDA-N) with the gene expression and drug sensitivity signatures of melanotic melanoma, but derived from a pleural effusion of a patient with breast cancer.

Relationship of drug-activity patterns to mechanism of action

Most of the compounds tested have unknown mechanisms of action, although their mechanisms can often be inferred from results obtained with the COMPARE program[11,12] or from clustering on the basis of their patterns of activity in the 60 cell lines[8,13,14]. For the analysis of mechanisms, we focused on a 118-drug subset (Table 1) of database A whose mechanisms of action are putatively understood. Some of these drugs are currently in routine clinical use; others have undergone clinical trials or are in late stages of drug development.

We generated an average-linkage dendrogram based on the activity patterns of the 118 drugs over the 60 cell lines (Fig. 3*a*). Five large, coherent clusters corresponded closely to mechanisms of action: DNA and DNA/RNA antimetabolites, tubulin inhibitors, DNA-damaging agents, topoisomerase 1 (Top1) inhibitors and topoisomerase 2 (Top2) inhibitors. The antimetabolite cluster included nine dihydrofolate reductase (DHFR) inhibitors (D_f). The only outlying compound was ftorafur, a 5-fluorouracil (5-FU) prodrug that was almost inactive in the two-day growth inhibition assay (Table 1).

This dendrogram contains information on many drug classes, but for illustration, we will focus on antimetabolites, antitubulins and topoisomerase inhibitors. 5-FU appeared with the RNA synthesis inhibitors (Rs) in a cluster next to dihydrofolate reductase inhibitors. 5-FU is known to act on DNA as well as on RNA. The fact that it clustered with RNA synthesis inhibitors suggests that RNA activity is its dominant mechanism of action.

Tubulin inhibitors formed the most coherent cluster. Drugs inhibiting tubulin monomer polymerization (vinca alkaloids and colchicines) clustered on one drug branch, and drugs inhibiting depolymerization (taxanes) on another. Geldanamycin and bisantrene were also in the cluster. The presence of geldanamycin might be due to its capacity to induce G1 cell-cycle arrest, as has been observed for taxanes. Why bisantrene, thought to be a Top2 inhibitor[33], clustered with the antitubulins remains unclear, but the grouping did not appear to be due to experimental noise.

A 'supercluster' included both the Top1 and Top2 branches. The Top1 inhibitor camptothecin (CPT) and all of its derivatives formed a very tight cluster. These CPTs (refs 34,35) clustered next to a group of DNA synthesis inhibitors (Ds). This observation was consistent with the DNA-replication dependence of camptothecin cytotoxicity, which has been proposed to result from damage to DNA by formation of 'replication fork encounter lesions'[36]. The Top2 inhibitors, except for etoposide and teniposide, bind to DNA, generally by intercalation[34,37]. In addition to their action on Top2, they may therefore act on DNA in other ways. Because most of the DNA-binding Top2 inhibitors clustered together and were in the same cluster as etoposide and teniposide, the Top2 activity was probably the dominant mechanism of action for these compounds (including derivatives of doxorubicin, mitoxantrone and amsacrine). These observations show how databases of activity in cells can generate new hypotheses with respect to drug mechanisms of action.

Gene-drug correlations on the basis of gene expression and drug activity (AT-matrix clustering)

We analysed expression profiles of the 1,376 genes plus 40 individually assessed targets in relation to the activity profiles of the 118 drugs with known mechanisms of action (Fig. 3*b*). The drugs were clustered on the basis of Pearson correlation coefficients that related their activity patterns across the 60 cell lines to the expression patterns of genes over the 60 cell lines. These correlation coefficients were calculated for each combination of a gene and a drug by taking the (normalized) level of expression of the gene in each cell line, multiplying it by the corresponding (normalized) sensitivity of the cell to the drug, summing the results over all of the cell lines and renormalizing. This yielded 1,376 + 40 correlation coefficients (one for each gene and target) for each

Table 1 • Database of drugs analysed

Mechanism of action*	Drug	NSC no	Mean –log GI50	s.d.	No. expts	Average no. lines/expt	Mechanism of action*	Drug	NSC no	Mean –log GI50	s.d.	No. expts	Average no lines/expt
A2	mitomycin	26980	6.11	0.56	137	42.9	Db	cyanomorpholino doxorubicin	357704	10.29	0.32	11	46.1
A2	porfiromycin	56410	5.43	0.61	13	43.2	Db	hycanthone	142982	5.10	0.20	9	31.1
A6	carmustine (BCNU)	409962	4.15	0.22	136	42.5	Db	morpholino-adriamycin	354646	7.73	0.32	8	49.0
A6	chlorozotocin	178248	3.21	0.40	10	45.8	Db	N-N-dibenzyl-daunomycin	268242	4.83	0.47	15	41.6
A6	clomesone	338947	3.72	0.38	15	43.1	Db	pyrazoloacridine	366140	6.56	0.29	15	43.4
A6	lomustine (CCNU)	79037	4.35	0.31	56	38.9	Di	5-6-dihydro-5-azacytidine	264880	4.63	0.75	16	40.0
A6	mitozolamide	353451	3.93	0.31	15	43.0	Di	α-2′-deoxythioguanosine	71851	3.80	0.38	15	43.3
A6	PCNU	95466	3.68	0.44	15	42.1	Di	azacytidine	102816	6.11	0.29	15	44.2
A6	semustine (MeCCNU)	95441	4.37	0.18	15	39.9	di	β-2′-deoxythioguanosine	71261	5.94	0.50	15	44.7
A7	asaley	167780	5.30	0.44	16	38.3	di	thioguanine	752	5.91	0.46	135	43.3
A7	busulfan	750	3.22	0.36	4	54.8	Df	aminopterin	132483	6.18	1.39	3	44.3
A7	carboplatin	241240	3.88	0.27	59	39.8	Df	aminopterin-derivative	134033	6.60	1.23	3	43.0
A7	chlorambucil	3088	4.22	0.40	130	42.8	Df	aminopterin-derivative	184692	6.67	1.55	3	47.7
A7	cisplatin	119875	5.38	0.37	127	41.4	Df	an-antifol	623017	7.01	1.40	2	39.0
A7	cyclodisone	348948	4.41	0.26	14	44.1	Df	an-antifol	633713	8.17	0.80	2	50.5
A7	diaminocyclohexyl-Pt-II	271674	5.51	0.50	16	44.0	Df	Baker's-soluble-antifolate	139105	6.24	1.47	5	50.6
A7	dianhydrogalactitol	132313	4.33	0.51	16	40.4	Df	methotrexate	740	6.94	1.28	4	53.2
A7	diaziridinylbenzoquinone	182986	5.50	0.42	52	39.0	Df	methotrexate-derivative	174121	8.05	0.87	3	51.0
A7	fluorodopan	73754	3.46	0.23	10	39.4	Df	trimetrexate	352122	8.58	1.11	4	50.8
A7	hepsulfam	329680	3.67	0.38	14	43.4	Dr	guanazole	1895	2.23	0.24	15	44.1
A7	iproplatin	256927	4.45	0.31	16	40.0	Dr	hydroxyurea	32065	3.14	0.42	55	39.9
A7	mechlorethamine	762	5.52	0.57	56	39.9	Dr	pyrazoloimidazole	51143	2.59	0.39	15	43.6
A7	melphalan	8806	4.56	0.38	56	38.2	Ds	aphidicolin-glycinate	303812	5.02	0.78	14	42.4
A7	piperazine mustard	344007	3.97	0.51	15	42.8	Ds	cyclocytidine	145668	4.73	1.37	17	38.4
A7	piperazinedione	135758	6.11	0.57	16	41.6	Ds	cytarabine (araC)	63878	4.82	1.49	132	39.7
A7	pipobroman	25154	4.16	0.28	56	38.8	Ds	floxuridine (FUdR)	27640	6.39	1.13	4	54.5
A7	spiromustine	172112	3.82	0.29	12	33.2	Ds	fluorouracil (5FU)	19893	4.63	0.73	1149	53.6
A7	teroxirone	296934	4.90	0.47	15	43.3	Ds	ftorafur	148958	2.67	0.34	4	51.8
A7	tetraplatin	363812	5.91	0.52	13	43.8	Ds	thiopurine (6MP)	755	5.31	0.67	134	42.4
A7	thiotepa	6396	4.09	0.46	131	42.8	Rs	acivicin	163501	5.50	0.48	16	39.4
A7	triethylenemelamine	9706	5.20	0.47	136	43.3	Rs	dichloroallyl-lawsone	126771	4.97	0.50	16	41.9
A7	uracil mustard	34462	4.56	0.51	56	40.1	Rs	DUP785 (brequinar)	368390	5.80	1.07	10	42.5
A7	yoshi-864	102627	2.90	0.31	15	44.0	Rs	L-alanosine	153353	5.06	0.74	16	39.8
T1	camptothecin	94600	7.40	0.58	9	38.3	Rs	N-phosphonoacetyl-L-aspartic-acid	224131	3.35	0.75	15	39.5
T1	camptothecin,7-Cl	249910	7.42	0.83	5	48.8	Rs	pyrazofurin	143095	5.26	1.03	12	43.6
T1	camptothecin,9-MeO	176323	7.10	0.97	4	52.0	TU	colchicine	757	7.26	1.17	7	45.9
T1	camptothecin,9-NH2 (RS)	629971	7.36	0.74	5	51.2	TU	colchicine-derivative	33410	7.58	0.93	7	47.3
T1	camptothecin,9-NH2 (S)	603071	7.43	0.66	6	49.8	TU	dolastatin-10	376128	9.53	0.42	4	47.0
T1	camptothecin,10-OH	107124	7.51	0.56	7	35.7	TU	halichondrin B	609395	8.93	0.48	4	47.8
T1	camptothecin,11-formyl (RS)	606172	5.69	0.69	3	50.3	TU	maytansine	153858	8.23	0.33	5	52.2
T1	camptothecin,11-HOMe (RS)	606173	5.43	0.60	2	46.5	TU	trityl-cysteine	83265	6.01	0.51	15	42.7
T1	camptothecin,20-ester (S)	606497	6.51	0.75	4	50.5	TU	vinblastine-sulphate	49842	9.04	1.00	134	39.0
T1	camptothecin,20-ester (S)	606985	7.42	0.79	2	51.5	TU	vincristine-sulphate	67574	6.82	0.65	60	37.8
T1	camptothecin,20-ester (S)	610456	6.84	0.74	4	51.5	TU	taxol (paclitaxel)	125973	7.35	0.59	14	55.2
T1	camptothecin,20-ester (S)	618939	7.19	0.75	3	51.7	TU	taxol analogue	600222	5.65	0.68	2	54.5
T2	amonafide	308847	5.49	0.21	16	39.9	TU	taxol analogue	656178	5.66	0.75	2	49.5
T2	amsacrine	249992	6.32	0.70	135	42.5	TU	taxol analogue	658831	5.43	0.83	2	50.0
T2	anthrapyrazole-derivative	355644	6.68	0.68	9	48.2	TU	taxol analogue	661746	6.86	0.60	2	51.5
T2	bisantrene	337766	6.76	0.67	11	39.4	TU	taxol analogue	664402	6.85	0.71	2	49.5
T2	daunorubicin	82151	7.10	0.58	78	45.8	TU	taxol analogue	664404	7.80	1.11	2	51.0
T2	deoxydoxorubicin	267469	7.34	0.55	7	49.3	TU	taxol analogue	666608	7.00	0.72	2	54.0
T2	doxorubicin	123127	6.84	0.56	1171	54.8	TU	taxol analogue	671867	7.59	0.93	2	52.5
T2	etoposide	141540	5.36	0.65	43	37.6	TU	taxol analogue	671870	6.11	0.59	2	55.5
T2	menogaril	269148	6.07	0.62	15	43.6	TU	taxol analogue	673187	6.43	0.85	2	56.0
T2	mitoxantrone	301739	7.19	0.71	13	40.1	TU	taxol analogue	673188	7.30	0.97	2	54.5
T2	oxanthrazole (piroxantrone)	349174	5.83	0.44	14	43.0	P90	geldanamycin	330500	6.26	0.60	12	42.4
T2	teniposide	122819	6.35	0.65	13	42.6	Uk	3-hydropicolinaldehyde-thiosemicarbazone	95678	5.79	0.40	15	43.3
T2	zorubicin (rubidazone)	164011	6.59	0.49	16	40.6	Uk	5-hydroxypicolinaldehyde-thiosemicarbazone	107392	5.01	0.45	14	43.0
Pi	L-asparaginase	109229	-0.35	0.64	104	40.6	Uk	inosine-glycodialdehyde	118994	3.54	0.33	16	38.4

*Alkylating agents: A2, A7, alkylating at N-2, N-7 position of guanine, respectively; A6, alkylating at O-6 position of guanine; T1, topoisomerase I inhibitor; T2, topoisomerase II inhibitor; Db, DNA binder; Di, DNA incorporation; Df, antifols; Dr, ribonucleotide reductase inhibitor; Ds, DNA synthesis inhibitor; Rs, RNA synthesis inhibitor; Tu, tubulin-active antimitotic agents; Pi, protein synthesis inhibitor; P90, hsp90 binder; Uk, unknown.

of the 118 drugs. We then clustered the 118 drugs on the basis of these correlation coefficients.

Comparison of Fig. 3*b* with Fig. 3*a* indicates that analysis on the basis of gene-drug correlation changed the clustering of many, but not all, mechanistic classes of compounds. The antimetabolite and alkylating agent clusters changed in ways not clearly linked to known structural or mechanistic features. The antitubulin cluster did not change, but the topoisomerase inhibitors rearranged in a manner that revealed mechanistic distinctions among subclasses of compounds.

The antimetabolites appeared in five distinct clusters, moderately changed relative to their clustering based on activity alone. We found the antifols (Df) in a large, coherent branch, which also included two RNA synthesis inhibitors, DUP785 and dichloroallyl-lawsone. All of the purine analogues (Di and Ds) appeared together on a small branch. The pyrimidine analogues, which formed a single branch on the basis of activities, separated into two groups (Fig. 3*b*), one composed of aphidicolin-glycinate and floxuridine, and the other, of cyclocytidine and cytarabine (Ara C).

The alkylating agents separated into several clusters. N-7 nitrogen mustards and ethylenimines formed one branch. The nitrosoureas (carmustin, lomustin, fluorodopan and semustin) formed a tight group by themselves. The three alkyl alkane sulfonates (yoshi-864, hepsulfam and busulfan) clustered together, but also with pipobroman and pyrazoloimidazole.

The five most active Top1 inhibitors (CPT, CPT,7-CL, CPT,9-MeO, CPT,9-NH2(RS) and CPT,9-NH2(S)), which do not require activation, clustered together, whereas the prodrugs (CPT,20-esters and CPT,11-formyl) clustered in a separate group. One CPT stood out as an exception: CPT,10-OH. Preliminary evidence indicates that this compound may be glucuronidated (unpublished data).

The Top2 inhibitors clustered in two distinct groups, one composed of anthracyclines (deoxydoxorubicin, daunorubicin, zorubicin and doxorubicin) and teniposide (VM-26), the second composed of mitoxantrone, oxanthrazole and an anthrapyrazole derivative. The latter clustered next to the bioreductive compounds porfiromycin and mitomycin, suggesting that their ability to produce double-strand breaks in DNA is a major determinant of the correlation between their activity and gene expression. Etoposide (VP-16) clustered paradoxically with the alkylating agents, perhaps implying that drug metabolism rather than mechanism of action is an important feature of the activity-expression correlation.

AT-matrix clustered image map

The AT-clustered image map (CIM; Fig. 4) summarizes the relationship between drug activity and gene expression. CIMs offer a convenient way to visualize patterns of similarity and difference in large sets of high-dimensional data. We have previously used CIMs to visualize relationships among drug activities, individual targets, protein expression patterns and gene expression patterns[8,28,38,39]. The algorithm in the form used here has been described[8,38]. In this CIM, the cluster tree of drugs (Fig. 3*b*) is represented on the *y* axis, and genes and individually assessed targets (n=1,376 genes+40 individually assessed targets) are clus-

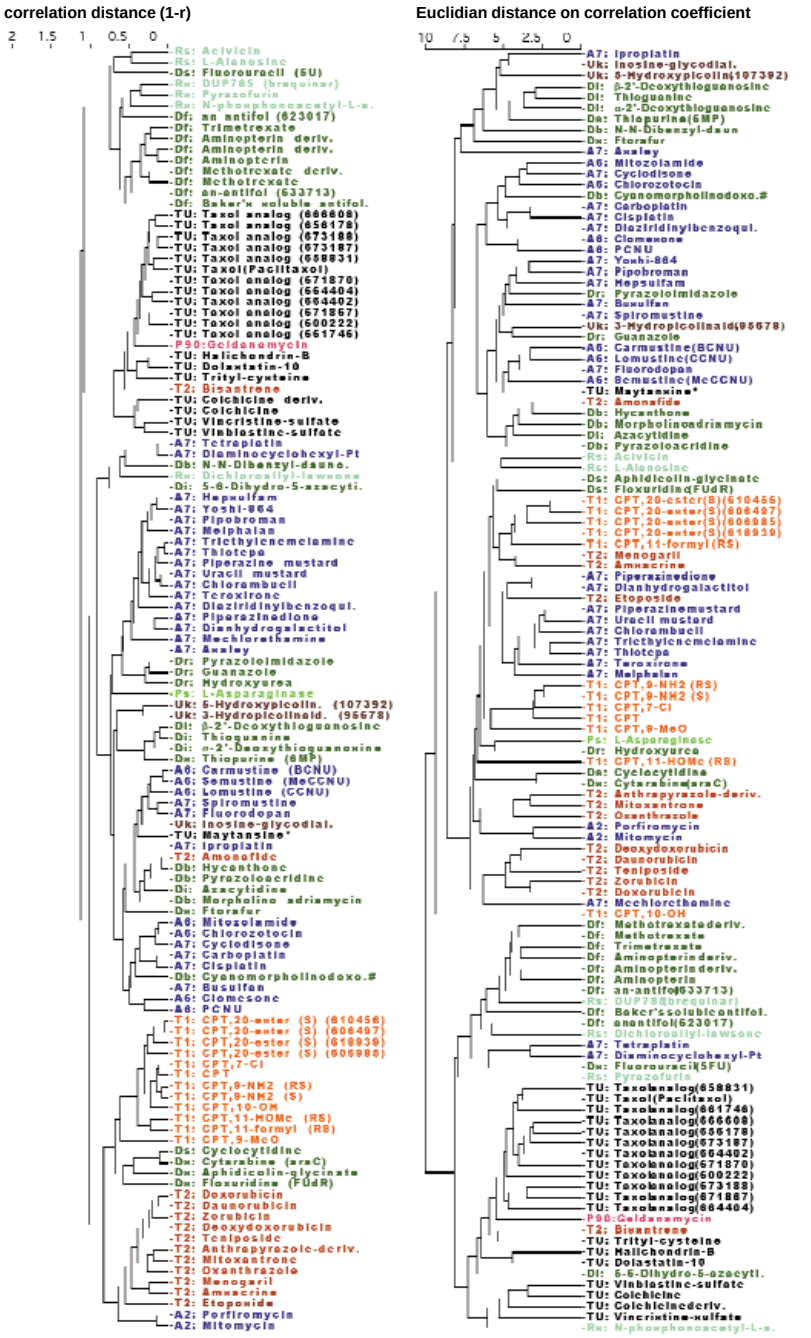

Fig. 3 Dendrograms showing average-linkage hierarchical clustering of 118 'mechanism of action' drugs. ***a***, Cluster tree of 118 drugs with putatively known mechanisms of action based on their activity patterns across the 60 cell lines. ***b***, Cluster tree of the 118 drugs based on the correlation of their activity patterns with expression patterns of the genes. The distance metric used for (*a*) was 1–*r*, where *r* is the Pearson correlation coefficient. The distance metric used in (*b*) was the Euclidean distance between Pearson correlation coefficients for the gene-drug combinations. The data clustered were $-\log_{10}(GI_{50})$ values, with main effects removed for both cells and drugs. The distance metric used was (1–Pearson correlation coefficient). See Table 1 for definitions of mechanism of action abbreviations.

tered on the *x* axis. Each block of red or blue represents a high positive or negative correlation between a cluster of genes and a cluster of drugs. The data and a full-resolution version of the figure are available (http://discover.nci.nih.gov).

Examples of causally related gene-drug pairs

The antimetabolite 5-FU, commonly used to treat colorectal and breast cancer, can inhibit both RNA processing and thymidylate synthesis. Dihydropyrimidine dehydrogenase (DPYD, encoded by *DPYD*), the rate-limiting enzyme in uracil and thymidine catabolism, is also rate-limiting in 5-FU catabolism. High DPYD levels would be expected to decrease exposure of cells to the active phosphorylated forms of 5-FU. Consistent with this hypothesis, we found a highly significant negative correlation (–0.53) between *DPYD* expression and 5-FU potency against the 60 cell lines (Fig. 4, inset A). On closer examination, we found that 14 of 18 cell lines with low expression of DPYD (less than 25% of the reference pool level) are sensitive or highly sensitive to 5-FU. Perhaps not coincidentally, given the clinical use of 5-FU against colon cancer, all of the colon-derived cell lines (7/7) fall into that category. DPYD enzyme activity has been assessed[40,41], and the results in clinical materials have been inconsistent[41]. The data presented here suggest that further study of DPYD as a clinical marker is warranted.

Certain malignant cells, including those of many acute lymphoblastic leukaemias (ALL), lack asparagine synthetase (ASNS, encoded by *ASNS*) and therefore depend on exogenous L-asparagine[42]. This dependence is exploited by treating ALL and other lymphoid malignancies with L-asparaginase, which depletes extracellular L-asparagine[43]. We found a moderately high negative correlation (–0.44) between expression of *ASNS* and L-asparaginase sensitivity in the 60 cell lines (Fig. 4, inset B). The two-tailed 95% bootstrap[44] confidence interval was –0.593 to –0.248; for comparison, that calculated from Fisher's z-transform was very similar, –0.620 to –0.204. When we stratified the data by subtracting the mean log values for drug sensitivity and gene expression within each organ of origin group of cells, the correlation was stronger (–0.55). For the subpanel of leukaemic lines (Fig. 5), the correlation coefficient was much higher, –0.98 (with a bootstrap confidence interval of –1.00 to –0.928). The *P* value, calculated from 1,000 bootstrap samples for the null hypothesis of zero correlation, was 0.005. This value is statistically significant even if a Bonferroni correction is applied. The two ALL lines (MOLT-4 and CCRF-CEM) expressed the lowest levels of *ASNS* mRNA and were the most sensitive to L-asparaginase. K-562, a chronic myelogenous leukaemia line, had the highest expression of *ASNS* and was the least sensitive to L-asparaginase.

There were also suggestive correlations between expression of *ASNS* and L-asparaginase sensitivity for the ovarian lines (–0.88; bootstrap confidence limits –0.231 to –0.987). The correlation for all cell types, other than leukaemia and ovarian, was –0.32 (confidence interval –0.044 to –0.557). Early clinical trials done with solid tumours have shown occasional responses to L-asparaginase in melanoma, chronic granulocytic leukaemia, lymphosarcoma and reticulum cell sarcoma[43], but not in other tumour types. Because newer polyethylene glycol-modified forms of L-asparaginase[45] appear to show much better pharmacokinetic properties and much less immunosuppression than the native form of the enzyme, our findings support the possible use of *ASNS* expression as a marker for clinical decisions regarding L-asparaginase therapy as well as a closer look at the use of L-asparaginase therapy for solid tumours.

Discussion

We have described the pharmacological implications of gene-expression profiling studies of the NCI60 cell lines. Because the gene expression patterns were determined in untreated cells, our data relates to sensitivity to therapy, rather than to the molecular consequences of therapy. In that sense, our study is analogous to an assessment of clinical tumours for markers that predict sensitivity to therapy. Our essential aims were to understand molecular pharmacology, to aid in the process of drug discovery and to provide a rationale for selection of therapy on the basis of molecular characteristics of a patient's tumour.

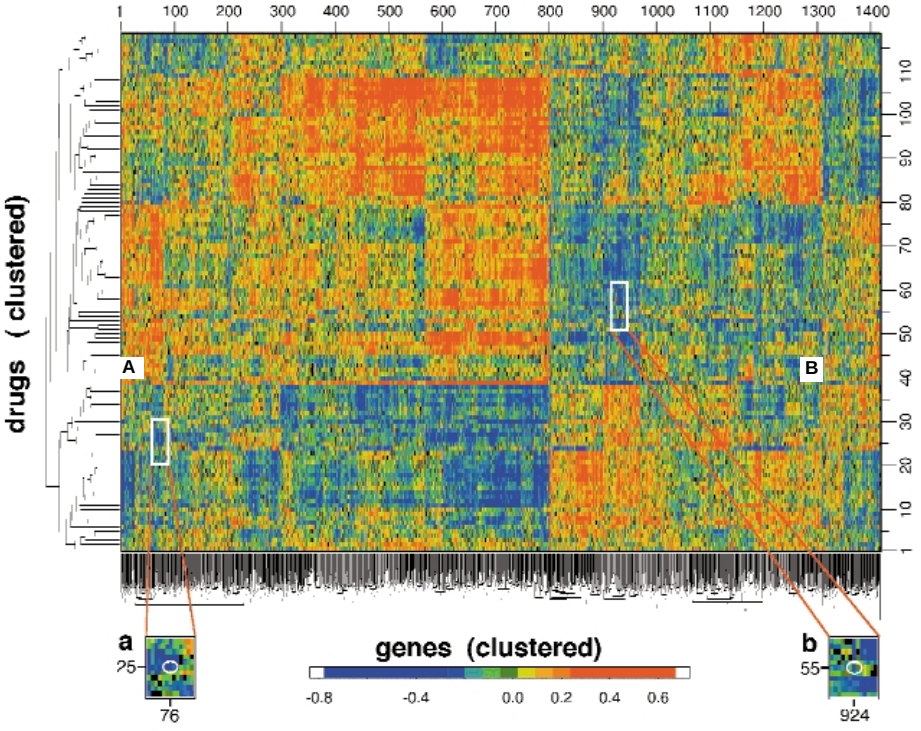

Fig. 4 CIM relating activity patterns of 118 tested compounds to the expression patterns of 1,376 genes in the 60 cell lines. Included, in addition to the gene expression levels, are data for 40 molecular targets assessed one at a time in the cells. A red point (high positive Pearson correlation coefficient) indicates that the agent tends to be more active (in the two-day SRB assay) against cell lines that express more of the gene; a blue point (high negative correlation) indicates the opposite tendency. Genes were cluster-ordered on the basis of their correlations with drugs (mean-subtracted, average-linkage clustered with correlation metric); drugs were clustered on the basis of their correlations with genes (mean-subtracted, average-linkage clustered with correlation metric). The drug cluster tree is the same as that in Fig. 3*b*, which can be consulted to identify individual drugs. A larger version of this $A \cdot T^T$ clustered correlation (ClusCorr) CIM (with the drug and gene names and the cluster trees; refs 8,28,38) is available (http://discover.nci.nih.gov). Inset A shows a magnified view of the region around the point (white circle) representing the correlation between *DPYD* (76) and 5-FU (25). Inset B is an analogous magnified view for *ASNS* (924) and the drug L-asparaginase (55).

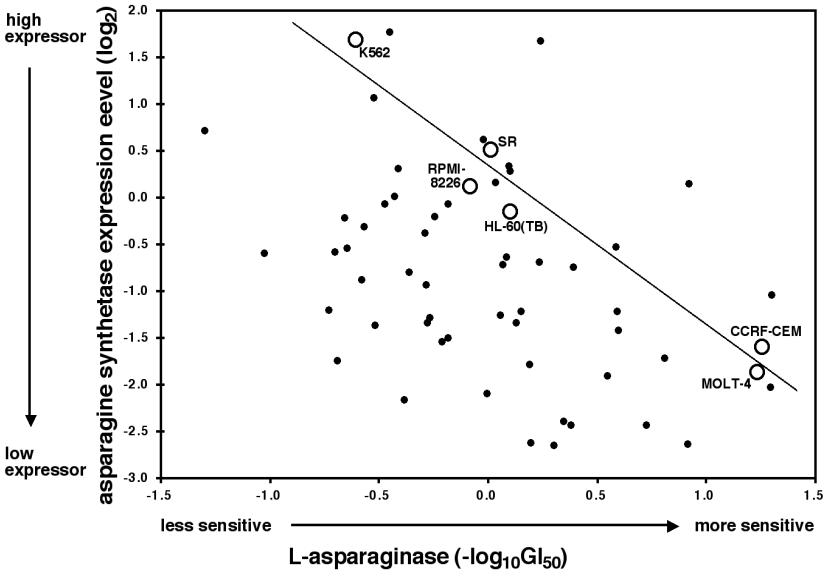

Fig. 5 Relationship between *ASNS* expression levels and chemosensitivity of the NCI cell lines to L-asparaginase. The main effects have been removed for both cells and drugs. Hence, a negative log(GI_{50}) value of 1 for sensitivity indicates a tenfold higher than average sensitivity of the cell line to the agent. The *ASNS* level is plotted as the abundance (log_2) of the *ASNS* transcript, relative to its abundance in the reference pool of 12 cell lines. A value of +2 indicates fourfold higher expression than in the reference pool. The large circles indicate leukaemia cell lines. The linear regression line (correlation coefficient=–0.98; *P* value<0.01) was fitted to the leukaemia data.

This approach has several limitations[8]. First, cell lines differ from tumour cells, particularly as they have been removed from their *in vivo* environment and selected for growth characteristics in culture. They should therefore be considered as surrogates that may contain information on the molecular cell biology and molecular pharmacology of cancer. Second, we generated our activity database using only one assay end-point, an index of short-term growth inhibition and cytotoxicity. Third, the relationships established between drug activities and gene expression levels are correlative, not causal, and they generate hypotheses that must now be tested. Fourth, there are only 60 cell lines. Finally, fewer than 10% of all human genes (although a much larger percentage of those in any given cell type) are represented on the arrays.

Our analysis of the gene expression profiling and pharmacological studies relies conceptually on the pair of database matrices. The A-matrix expresses the relationship between tested compounds and the 60 cell types. The T-matrix relates cells to their molecular characteristics, mRNA expression levels (T_r) and individual targets (T_i). The inner product of A and T (normalized to produce Pearson correlation coefficients[8,28,38]) yields a set of relationships between tested compounds and measured gene expression levels. The gene expression profiles show considerable coherence, in that cells clustered on the basis of their expression profiles for 1,376 genes and 40 targets tend to sort themselves by organ of origin. This generalization most clearly holds for the leukaemias, melanomas and carcinomas of renal, CNS and colon origin.

The activity patterns of 1,400 compounds did not group the cells as well by organ of origin as did the gene expression profiles. The reason is clear: an individual gene can have a major impact on the activities of a large number of drugs but, being just 1 gene out of 1,376, it can have little effect on clustering by gene expression pattern. For example, *ABCB1* has a large impact on drugs that it can transport out of the cell[8,15,21–24]. Because *ABCB1* is expressed at significant levels in at least one cancer cell line from each of several different organs (that is, renal, breast, colon and lung), it tends to confound grouping by organ of origin on the basis of drug-activity profiles. This observation may largely explain the unexpectedly low correlation (r=0.21) that we found between the grouping of cells on the basis of gene expression and that on the basis of drug activity.

Drugs clustered according to their patterns of activity show generally good correlations with presumed mechanism of action, but there are exceptions. Bisantrene, for example, was not expected to cluster with the antitubulin agents. Cyanomorpholinodoxorubicin was presumptively classified as a DNA binder, but clusters with the alkylating agents, suggesting that alkylation by the cyano-moiety is the dominant mechanism of action.

Exceptions to expected clustering relationships can, in principle, be explained on the basis of the following: (i) experimental variability; (ii) the effect of dimensionality reduction, which occurs during compression of 60-dimensional activity data into one dimension and results in a loss of information; and (iii) incorrect or incomplete assignment of mechanism of action. Drugs with the same primary mechanism of action may have secondary mechanisms that differ, and they may be susceptible to different pharmacological factors (for example, efflux mediated by MDR1). Despite these possibilities, there was a high degree of coherence for most mechanisms of action, consistent with previous observations for various drug data sets[8–11].

In the ClusCorr CIM, each block of colour represents an association between a cluster of genes and a cluster of drugs. The block is red if the gene and drug clusters are positively correlated, blue if the gene-drug correlation is negative, and yellow or green if there is little correlation. Where the cluster tree for genes or drugs has a deep fork, the block of colour tends to have a sharp boundary. Each block of red or blue may represent a causal correlation, an epiphenomenal association or a statistical artefact. Appropriate randomization studies can often rule out statistical artefact, but the more difficult distinction to make is that between epiphenomenon and causal association. This must generally be done by searching the literature and available databases for clues, or by carrying out additional experiments. To search the literature on gene-gene and gene-drug relationships more rapidly and flexibly, we developed a web-based program, MedMiner[46] (http://discover.nci.nih.gov). MedMiner uses the Weizmann Institute's GeneCards and the National Library of Medicine's PubMed to extract literature information and then organize it in a way that reduces five- to tenfold the time required to explore complex relationships.

By combining genome-wide expression profiling with drug activity data, we are exploring a large set of possible gene-gene, gene-drug and drug-drug relationships simultaneously. Our aim is exploratory: we obtain clues, generate hypotheses and establish context rather than testing a particular biological hypothesis in the classical manner[27,47]. At present, however, we can interpret only a small proportion of the relationships. The DPYD/5-FU

and ASNS/L-asparaginase correlations are cases in which we knew enough to recognize a likely causal nexus (with clinical implications), and in which the gene expression data provided considerable added value. The most interesting relationships are presumably those we cannot yet recognize. To facilitate exploration of this data resource over the coming years, both the analysis tools and data are available (http://discover.nci.nih.gov and http://genome-www.stanford.edu/nci60).

A final limitation of the study is that pharmacologically interesting behaviours are not always reflected at the transcriptional level. It will be necessary to assess differences among cells at the DNA and protein levels as well. An overall aim of this enterprise[27], then, is to combine the three levels of experiment and analysis. Toward that end, we have collected DNA and protein in parallel with the RNA for cross-indexable characterizations with respect to all three types of molecules (unpublished data).

Methods

Assay for drug activity. The drug profiling protocols of the NCI have been described[1,3,6]. Briefly, the cells were grown in 96-well microtitre plates and exposed to the test compound for 48 h. Growth inhibition, assessed by the sulphorhodamine B assay for cellular protein, can be expressed in terms of the quantity $-\log(GI_{50})$, where GI_{50} is the concentration required to inhibit cell growth by 50% in comparison with untreated controls. The activity profile of a compound is composed of 60 such activity values, one for each cell line.

Cell collection and mRNA purification. Briefly, we took seed cultures of the cell lines from stocks used for ongoing assays in the DTP screen. They were then passaged once in T-162 flasks and monitored frequently for degree of confluence. We used RPMI-1640 medium (30 ml for attached cells; 40 ml for leukaemias) with phenol red, glutamine (2 mM) and 5% fetal calf serum. For compatibility with the drug-profiling regimen, we obtained all fetal calf serum from a large batch (BioWhittaker) used by DTP. No antibiotics were used. One day before collection, the cells were re-fed with the original amount and composition of medium. We collected cells at ~80% confluence, as assessed for each flask by phase microscopy and documented by photomicroscopy for two flasks of each cell type at each collection. Samples of medium showed no change in pH between refeeding and collection, and no colour change in the medium was seen in any of the flasks. Cells were collected in parallel for RNA, DNA and protein. For RNA, the interval from incubator to stabilization of the preparation was kept to <1 min. We purified total RNA using the RNeasy kit (Qiagen) according to the manufacturer's instructions. The RNA was then quantitated spectrophotometrically and aliquoted for storage at –70 °C. As needed, poly(A) mRNA was obtained from total RNA using the Oligotex kit (Qiagen). Purified message was routinely quality-controlled on formaldehyde agarose gels.

cDNA microarrays. We assessed gene expression patterns using microarrays (Synteni, Inc.; now Incyte, Inc.) consisting of robotically spotted, PCR-amplified cDNAs on coated glass slides[48]. The 9,703 DNA elements on the array were cDNAs from the Washington University/Merck IMAGE set (Research Genetics). The cDNAs on this array included 3,700 named genes, 1,900 human genes homologous to those of other organisms and 4,104 ESTs of unknown function but defined chromosome map location. For each hybridization, cDNA synthesized from the mRNA of test cells was labelled by incorporation of Cy5-dNTP during reverse transcription. We analogously labelled cDNA synthesized from pooled mRNA of 12 highly diverse cell lines of the 60 by incorporation of Cy3-dNTP. Cells for the pool were selected to satisfy three criteria: (i) at least one cell line from each organ of origin; (ii) diversity of growth rates; and (iii) diversity in terms of protein expression pattern, based on prior two-dimensional gel studies[28]. Inclusion of all 60 cell types would have insured non-zero values for all mRNA transcripts expressed in any of the cells, but would have been logistically difficult and hard to replicate at a later time. Cells included in the pool were leukaemias HL-60(TB) and K-562; non-small cell lung cancer NCI-H226; colon cancer COLO 205; CNS cancer SNB-19; melanoma LOX-IMVI; ovarian cancers OVCAR-3 and OVCAR–4; renal cancer CAKI-1; prostate cancer PC-3; and breast cancers MCF7 and HS 578T.

Genes. We selected genes for analysis from the 9,704 on the array on the basis of three layers of quality control. First, we visually examined the individual chips. Values from spots contaminated with dust or fluorescent specks were treated as missing. Second, we examined the intensities and ratio for each individual spot. Values from spots with raw intensity in both red and green channels lower than 1.5 times the local background were considered as missing. If the spot was 1.5-fold higher than the local background for one channel (for example, red), but not the other (for example, green), the difference between raw intensity and background was thresholded at 100 intensity units (~1/10 of background) for the low channel. Third, genes were included if and only if 4 or fewer measurements were excluded out of the 60 and 4 or more cell lines had red-green ratios >2.6 or <0.38. These filters resulted in selection of 1,376 genes.

As of December 1999, the DTP web site listed data for 41 published targets assessed individually in all or most of the 60 cell lines by laboratories at the NIH or elsewhere. Of these, 40 targets were added (in log transformation, with appropriate thresholding) to the gene expression data to provide signposts for the analysis. The forty-first was omitted because it had too many missing values.

The drug database. The >70,000 DTP-tested chemical compounds were winnowed to a final database of 1,400 for analysis by applying a series of filters based on the number of times a compound had been tested, the number of missing values and the number of cell lines for which the GI_{50} value fell within the range of concentrations tested. The smaller set of 118 included so-called "mechanism of action" drugs[9,10,38] and 10 additional Taxol analogues. The number of independent experiments conducted by DTP per compound ranged from 2 to 1,176 for the set of 118, with a median of 15 and an interquartile range of 3 to 23. The mean number of cell lines tested and yielding GI_{50} values that passed quality control for a given experiment on a given compound was 46.5. To arrive at GI_{50} values for use in analysis, we calculated medians of the individual values obtained in experiments performed over the best concentration range.

Data analysis. Most statistical analyses were carried out using the S-Plus statistical package (StatSci Division, MathSoft). S-Plus scripts were written to generate suitably formatted HTML documents, which were invoked by a CGI program written in C and subsequently delivered to the analysts' web browsers. The graphics generated and tools of analysis used are available (http://discover.nci.nih.gov). For exploratory analyses, we used a variety of clustering algorithms, metrics, data transformations and visualization techniques. We settled on average linkage clustering with a correlation metric. Except where otherwise indicated, all P values and confidence intervals quoted are two-tail 95%, calculated by Efron's bootstrap re-sampling method[46] without small-sample correction. To calculate the degree of similarity between cell clustering on the basis of drugs and on the basis of genes, we derived a 'correlation of correlation' parameter r as follows: let U_{ij} denote the correlation of cells i and j (for i and j from 1 to n) based on their drug activities, and let V_{ij} denote the correlation of cells i and j based on their gene expression. For example, if X_{di} denotes the activity of drug d (for d from 1 to D) against cell i, then the Pearson correlation coefficient for cells i and j based on drug activity is given by the formula

$$U_{ij} = \frac{\sum_{d=1}^{D} X_{di}X_{dj} - \frac{1}{D}\sum_{d=1}^{D} X_{di}\sum_{d=1}^{D} X_{dj}}{\sqrt{\sum_{d=1}^{D} X_{di}^2 - \frac{1}{D}\left(\sum_{d=1}^{D} X_{di}\right)^2}\sqrt{\sum_{d=1}^{D} X_{dj}^2 - \frac{1}{D}\left(\sum_{d=1}^{D} X_{dj}\right)^2}},$$

and similarly for V_{ij}. The Pearson correlation of U_{ij} and V_{ij} gives a measure of the similarity in the distributions of drug activity and gene expression. The formula is given by

$$r = \frac{\sum_{i<j} U_{ij}V_{ij} - \frac{2}{n(n-1)}\sum_{i<j} U_{ij}\sum_{i<j} V_{ij}}{\sqrt{\sum_{i<j} U_{ij}^2 - \frac{2}{n(n-1)}\left(\sum_{i<j} U_{ij}\right)^2}\sqrt{\sum_{i<j} V_{ij}^2 - \frac{2}{n(n-1)}\left(\sum_{i<j} V_{ij}\right)^2}},$$

where the sums are over all distinct pairs of cells i and j, there being $n(n–1)/2$ such pairs.

Because the L-asparaginase/*ASNS* pair was initially selected on the basis of its pharmaceutical significance, not on the basis of its correlation, the statistical multiple comparisons problem does not involve all of the hundreds of thousands of correlations being assessed in this study. Rather, it can more accurately be framed in terms of the following two questions. (i) What is the probability of obtaining a correlation coefficient so far from zero if the null hypothesis of zero correlation holds for the cancer cell class that is clinically treated with L-asparaginase (that is, leukaemia)? In this case, the null hypothesis can be rejected on the basis of the *P* value of 0.005. (ii) What is the probability of obtaining a correlation coefficient so far from zero for at least 1 of the 8 cancer cell types in the panel if the null hypothesis of zero correlation holds for all 8 cell types (excluding prostate, for which there are only two cell lines)? Applying the Bonferroni correction (which assumes independence and is, in fact, too stringent) to the latter case, the critical value would be approximately 0.05/8=0.006. That number is slightly higher than the value of 0.005 obtained, so, formally speaking, the null hypothesis of zero correlation could be rejected despite the Bonferroni correction. No small-sample correction has been made in the bootstrap algorithm, however, and the result should, in any case, be considered as indicative, not definitive. Note that experimental noise would tend to decrease the magnitude of the observed correlation, not increase it, and would make it harder to reject the null hypothesis of zero correlation.

Clustered image map (CIM). Calculations to derive visualizable relationships between drugs and targets in the form of a "clustered correlation" CIM were performed as described[8,28,38]. In brief, we normalized each element in the activity matrix (A) by subtracting its row-wise mean and dividing by its row-wise standard deviation; normalized each element in the target matrix (T) by subtracting its row-wise mean and dividing by its row-wise standard deviation; took the inner product of the normalized A and the transpose of the normalized T matrix; and divided each element in the resulting matrix by N–1, where N is 60 minus the number of components for which one or both vectors had a missing value. The resulting matrix ($A \cdot T^T$), where T^T is the transpose of T, contains Pearson correlation coefficients relating a pattern of drug activities to a pattern of target expression. A program for making clustered correlation CIMs (as well as other types of CIMs) is available (http://discover.nci.nih.gov).

Acknowledgements

We thank the staff of the NCI DTP, particularly K.D. Paull, whose efforts over the years have resulted in the pharmacological databases used in this study. This study was supported in part by NCI grant CA77097 and by the Howard Hughes Medical Institute. D.T.R. is a Walter and Iden Berry Fellow. P.O.B. is an associate investigator of the Howard Hughes Medical Institute. The work of U.S. and J.N.W. was supported in part by a grant from the NCI intramural Breast Cancer Think Tank.

Received 19 July 1999; accepted 25 January 2000.

1. Boyd, M.R. & Paull, K.D. Some practical considerations and applications of the National Cancer Institute in vitro anticancer drug discovery screen. *Drug Dev. Res.* **34**, 91–109 (1995).
2. Alley, M.C. *et al.* Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay. *Cancer Res.* **48**, 589–601 (1988).
3. Monks, A. *et al.* 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 (1991).
4. Grever, M.R., Schepartz, S.A. & Chabner, B.A. The National Cancer Institute: cancer drug discovery and development program. *Semin. Oncol.* **19**, 622–638 (1992).
5. Stinson, S.F. *et al.* Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen. *Anticancer Res.* **12**, 1035–1053 (1992).
6. Boyd, M.R. in *Anticancer Drug Development Guide: Preclinical Screening, Clinical Trials, and Approval* (ed. Teicher, B.A.) 23–42 (Humana Press, Totowa, 1997).
7. Ross, D.T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.* **24**, 227–235 (2000).
8. Weinstein, J.N. *et al.* An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343–349 (1997).
9. Weinstein, J.N. *et al.* Neural computing in cancer drug development: predicting mechanism of action. *Science* **258**, 447–451 (1992).
10. van Osdol, W.W., Myers, T.G., Paull, K.D., Kohn, K.W. & Weinstein, J.N. Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents. *J. Natl Cancer Inst.* **86**, 1853–1859 (1994).
11. Paull, K.D., Hamel, E. & Malspeis, L. Prediction of biochemical mechanism of action from the in vitro antitumor screen of the National Cancer Institute. in *Cancer Chemotherapeutic Agents* (ed. Foye, W.E.) 1574–1581 (American Chemical Soc. Books, Washington, DC, 1993).
12. Paull, K.D. *et al.* 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 (1989).
13. Shi, L.M., Fan, Y., Myers, T.G., Paull, K.D. & Weinstein, J.N. Mining the NCI anticancer drug discovery databases: genetic function approximation for the quantitative structure-activity relationship study of anticancer ellipticine analogs. *J. Chem. Inf. Comput. Sci.* **38**, 189–199 (1998).
14. Shi, L.M. *et al.* Mining the National Cancer Institute's anticancer drug screen database: cluster analysis of ellipticine analogs with p53-inverse and central nervous system-selective patterns of activity. *Mol. Pharmacol.* **53**, 241–251 (1998).
15. Alvarez, M. *et al.* Generation of a drug resistance profile by quantitation of MDR-1/P-glycoprotein expression in the cell lines of the NCI anticancer drug screen. *J. Clin. Invest.* **95**, 2205–2214 (1995).
16. Izquierdo, M.A. *et al.* Overlapping phenotypes of multidrug resistance among panels of human cancer-cell lines. *Int. J. Cancer* **65**, 230–237 (1996).
17. O'Connor, P.M. *et al.* Characterization of the p53-tumor suppressor pathway in cells of the National Cancer Institute anticancer drug screen and correlations with the growth-inhibitory potency of 123 anticancer agents. *Cancer Res.* **57**, 4285–4300 (1997).
18. Freije, J.M. *et al.* Identification of compounds with preferential inhibitory activity against low-Nm23-expressing human breast carcinoma and melanoma cell lines. *Nature Med.* **3**, 395–401 (1997).
19. Koo, H.-M. *et al.* Enhanced sensitivity to 1-β-D-arabinofuranosylcytosine and topoisomerase II inhibitors in tumor cell lines harboring activated ras oncogenes. *J. Natl Cancer Inst.* **56**, 5211–5216 (1996).
20. Wosikowski, K. *et al.* Identification of epidermal growth factor receptor and c-erbB2 pathway inhibitors by correlation with gene expression patterns. *J. Natl Cancer Inst.* **89**, 1505–1513 (1997).
21. Bates, S.E. *et al.* Reversal of multidrug resistance. *Prog. Clin. Biol. Res.* **389**, 33–37 (1994).
22. Bates, S.E. *et al.* Molecular targets in the National Cancer Institute drug screen. *J. Cancer Res. Clin. Oncol.* **121**, 495–500 (1995).
23. Lee, J.-S. *et al.* Rhodamine efflux patterns predict P-glycoprotein substrates in the National Cancer Institute drug screen. *Mol. Pharmacol.* **46**, 627–638 (1994).
24. Wu, L. *et al.* Multidrug-resistant phenotype of disease-oriented panels of human tumor cell lines used for anticancer drug screening. *Cancer Res.* **52**, 3029–3034 (1992).
25. Kitada, S. *et al.* Expression and location of pro-apoptotic Bcl-2 family protein BAD in normal human tissues and tumor cell lines. *Am. J. Pathol.* **152**, 51–61 (1998).
26. Monks, A., Scudiero, D.A., Johnson, G.S., Paull, K.D. & Sausville, E.A. The NCI anticancer drug screen: a smart screen to identify effectors of novel targets. *Anticancer Drug Des.* **12**, 533–541 (1997).
27. Weinstein, J.N. Fishing expeditions. *Science* **282**, 627 (1998).
28. Myers, T.G. *et al.* A protein expression database for the molecular pharmacology of cancer. *Electrophoresis* **18**, 647–653 (1997).
29. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
30. Schena, M. *et al.* Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc. Natl Acad. Sci. USA* **93**, 10614–10619 (1996).
31. DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.* **14**, 457–460 (1996).
32. Scudiero, D.A., Monks, A. & Sausville, E.A. Cell line designation change: multidrug-resistant cell line in the NCI anticancer screen. *J. Natl Cancer Inst.* **90**, 862 (1998).
33. Capranico, G. *et al.* Mapping drug interactions at the covalent topoisomerase II-DNA complex by bisantrene/amsacrine congeners. *J. Biol. Chem.* **273**, 12732–12739 (1998).
34. Chen, A.Y. & Liu, L.F. DNA topoisomerases: essential enzymes and lethal targets. *Annu. Rev. Pharmacol. Toxicol.* **94**, 194–218 (1994).
35. Pommier, Y., Tanizawa, A. & Kohn, K.W. Mechanism of topoisomerase I inhibition by anticancer drugs. *Adv. Pharmacol.* **29B**, 73–92 (1993).
36. Shao, R.-G. *et al.* Replication-mediated DNA damage by camptothecin induces phosphorylation of RPA by DNA-dependent protein kinase and dissociates RPA:DNA-PK complexes. *EMBO J.* (in press).
37. Pommier, Y. DNA topoisomease II inhibitors. in *Cancer Therapeutics: Experimental and Clinical Agents* (ed. Teicher, B.A.) 153–174 (Humana Press, Totowa, 1997).
38. Weinstein, J.N. *et al.* Predictive statistics and artificial intelligence in the U.S. National Cancer Institute's drug discovery program for cancer and AIDS. *Stem Cells* **12**, 13–22 (1994).
39. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
40. Fischel, J.L. *et al.* Dihydropyrimidine dehydrogenase: a tumoral target for fluorouracil modulation. *Clin. Cancer Res.* **1**, 991–996 (1995).
41. McLeod, H.L. *et al.* Characterization of dihydropyrimidine dehydrogenase in human colorectal tumours. *Br. J. Cancer* **77**, 461–465 (1998).
42. Cooney, D.A. & Handschumacher, R.E. L-asparaginase and L-asparagine metabolism. *Annu. Rev. Pharmacol.* **10**, 421–440 (1970).
43. Capizzi, R.L., Bertino, J.R. & Handschumacher, R.E. L-Asparaginase. *Annu. Rev. Med.* **21**, 433–444 (1970).
44. Efron, B. & Gong, G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am. Statistician* **37**, 36–48 (1983).
45. Wada, H. *et al.* Antitumor enzyme: polyethylene glycol-modified asparaginase. *Ann. NY Acad. Sci.* **613**, 95–108 (1990).
46. Tanabe, L. *et al.* MedMiner: an internet tool for mining the biomedical literature, with application to gene expression profiling. *Biotechniques* **27**, 1210–1217 (1999).
47. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* **21** (suppl.), 33–37 (1999).
48. Shalon, D., Smith, S.J. & Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* **6**, 639–645 (1996).