Available online at www.sciencedirect.com

SCIENCE DIRECT®

C. R. Biologies 326 (2003) 909–920

COMPTES RENDUS BIOLOGIES

Molecular biology and genetics

Transcriptomic analysis of the NCI-60 cancer cell lines

John N. Weinstein [*], Yves Pommier

Laboratory of Molecular Pharmacology, Center for Cancer Research, US National Cancer Institute, NIH, Department of Health and Human Services, 9000 Rockville Pike, Bethesda, MD 20892, USA

Received 13 March 2003; accepted 29 August 2003

Presented by François Gros

Abstract

Pharmacogenomics aims at molecular subsetting of patients for more effective therapy. Transcriptomic profiling of the 60 human cancer cell lines (the NCI-60) used by the US National Cancer Institute serves that aim because the cells have been treated with $>100,000$ chemical compounds over the last 13 years. Patterns of potency can be mapped into molecular structures of the compounds or into molecular characteristics of the cells. We discuss conceptual and experimental aspects of the profiling, as well as a number of bioinformatic computer programs that we have developed for biological interpretation of the profiles. ***To cite this article: J.N. Weinstein, Y. Pommier, C. R. Biologies 326 (2003).***

Résumé

Analyse du transcriptome dans les 60 lignées cellulaires de cancers du NCI. Un des objectifs de la pharmacogénomique est d'identifier des groupes de patients afin d'augmenter l'efficacité thérapeutique. Le profilage transcriptomique des 60 lignées cellulaires de l'Institut national du cancer des États-Unis répond à cet objectif, car la réponse de ces cellules à plus de 100 000 agents chimiques et chimiothérapeutiques a été caractérisée depuis 13 ans. Les profils de réponse peuvent être répertoriés en fonction des structures des agents chimiques et des caractéristiques moléculaires des cellules. Nous discutons les aspects conceptuels et expérimentaux appliqués au profilage, ainsi que les différents logiciels bioinformatiques que nous avons développés pour l'interprétation biologique de ces profilages. ***Pour citer cet article : J.N. Weinstein, Y. Pommier, C. R. Biologies 326 (2003).***

Keywords: bioinformatics; cancer; genomic; NCI-60 cell lines; pharmacology; proteomics

Mots-clés : bioinformatique ; cancer ; génomique ; lignées cellulaires NCI-60 ; protéomique ; pharmacologie

1. Introduction

The principal goal of pharmacogenomics is clear: Use information on the molecular profiles of tumor cells to individualize therapy for cancer or select more appropriate therapy for particular subgroups of pa-

[*] Corresponding author.
E-mail addresses: jw4i@nih.gov (J.N. Weinstein), pommier@nih.gov (Y. Pommier).

1631-0691/$ – see front matter © 2003 Académie des sciences. Published by Elsevier SAS. All rights reserved.
doi:10.1016/j.crvi.2003.08.005

tients. For the most part, that goal is being pursued through study of clinical tumors, but the task has proved more difficult than expected. There are several reasons [1]: (*i*) clinical tumors are difficult to study, given anesthesia effects, surgical trauma, and constraints (both logistical and ethical) on the design of clinical trials; (*ii*) clinical tumors often have complex, fragmentary histories. Demographic and clinical information may be difficult to obtain because of ethical or legal issues; (*iii*) clinical studies are expensive; (*iv*) clinical tumors are heterogeneous – a mixture of cancer cells and stromal components, including endothelial cells, fibroblasts, and infiltrating leukocytes. Any molecular profile obtained for a bulk tumor is a mixture of the characteristics of those components. Techniques such as laser capture microdissection [2] can be used to isolate tumor cells, for example in the pseudo-glandular epithelial structures of adenocarcinomas. But then an amplification method must generally be used to generate enough DNA or mRNA for study. Methods of amplification available include T7-viral amplification, rolling circle amplification, two-primer PCR, and single-primer PCR, but their fidelity is still a significant question [1].

In contrast, cell lines have the advantage of being homogeneous in cell lineage (though not in cell cycle state). They can be obtained in quantity; they are reproducible from experiment to experiment and year to year; they can be manipulated by transfection, knockout, selection for resistant forms, or exposure to siRNA, antisense RNA, drugs, or radiation. The problem, of course, is that they are not really representative of cancer cells in vivo. Even primary cultures of cancer cells have been removed from the influence of other cell types, cytokines, extracellular fluid, and the three-dimensional architecture of the tumor. They have been selected for growth on plastic in standard medium with relatively fast cell-cycling. Therefore, prediction forward from cultured cells toward the clinic is uncertain; we can, at best, obtain clues to formulate hypotheses to be validated in real tumors, either clinically or through pathological studies, for example using tissue arrays. When one extrapolates backwards from cell line studies to the basic biology or pharmacology, however, one is on reasonably sound ground. Most of our knowledge of the biology and pharmacology has, in fact, been obtained from cultured cells or else from molecular studies, not from clinical materials [1].

For pharmacological purposes, we would like to study cancer cell types that have been exposed to large numbers of potential drugs. The most prominent such cell set is the 60 human cancer line panel (the NCI-60) used by the Developmental Therapeutics Program (DTP) of the US National Cancer Institute (NCI) to test for potential anticancer agents. The cells have been characterized pharmacologically by exposure to more than 100,000 defined chemical compounds (plus a large number of natural product extracts), one at a time and independently.

2. The NCI-60 cancer cells and screen

As of 1985, the NCI was using P388 murine leukemia to screen compounds for anticancer activity. That strategy identified agents active against leukemias but was not thought to be effective in identifying activity against the common solid tumors of humans. Therefore, the decision was made to seek a different strategy for screening. The result after many competing factors were taken into account was the NCI-60 cell screen, which went into production mode in April of 1990. Since then, >100 000 chemically defined compounds (plus natural product extracts) have been screened. Since 1991, the 60 cell lines have included leukemias, melanomas, and cancer cells of renal, ovarian, colon, breast, prostate, lung, and central nervous system origin. That list is by no means complete, but it includes the most common human tumors. The guiding hypothesis was that selective activity against cancers from a particular tissue or organ would predict clinical activity against the same type of tumor. Such predictiveness has not been demonstrated, but the NCI-60 system took on a new role: increasingly, it has been used for secondary profiling of compounds already found to attack a defined molecule or pathway. Even more generally, it became a system for profiling both the compounds tested and the cell lines.

Fig. 1 shows the NCI-60 system in highly schematic form. Database (A) of activities can be mapped into a database (S) of structural characteristics of the compounds tested and a database (T) of molecular targets and other cell characteristics. This set of databases provides the conceptual architecture for the pharmacogenomic studies to be described here. The first topic to be discussed will be the screen itself.

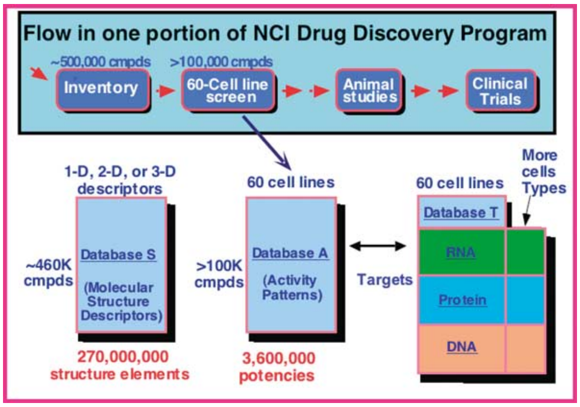

Fig. 1. Simplified schematic overview of an information-intensive approach to cancer pharmacogenomics and pharmacoproteomics based on the NCI-60 cancer cell lines. Each row of the activity (A) database represents the pattern of activity of a particular compound across the 60 cell lines. The A database can be mapped into a structure (S) database containing 2D or 3D chemical structure descriptors of the compounds and a target (T) database containing molecular profile information on the cells. The T database consists of data on individual molecules and omic data at the DNA, mRNA, protein, and functional levels. The bioinformatic challenge is to analyze and understand each of these databases separately, then to integrate them with each other and with public information resources for pharmacogenomic purposes. Modified from [25].

2.1. *Methodology of the NCI-60 screen*

The methodology of the NCI-60 screen has been described in detail elsewhere (see http://dtp.nci.nih.gov). Briefly, on day zero the human tumor cell lines are plated in 96-well microtiter format in RPMI 1640 medium with 5% fetal calf serum and 2 mM L-glutamine. On day 1, the drug (dissolved in DMSO) is added to achieve five concentrations at 10-fold intervals, plus a negative control. The usual concentration range is 10^{-8} to 10^{-4} M. After 48 h of drug exposure at 37 °C, the cells are fixed in situ with trichloroacetic acid. The supernatant is discarded (along with floating cells and cell fragments), and the plates are washed five times, then air-dried. Colorimetric measurement of sulforhodamine B (SRB) dye is used to quantitate the cell material remaining attached to the well at the end of the incubation period. The 50% growth inhibition (GI_{50}) is calculated as the concentration of drug required to inhibit cell growth by a factor of two. The fundamental parameter used as a measure of potency is $-\log_{10} GI_{50}$. The details of this protocol are important to an understanding of the meaning and limitations of the data from it.

2.2. *The Activity (A) database*

While analyzing data from pilot studies for the screen in the late 1980's, Kenneth D. Paull realized that the absolute potency of a compound gave much less information on its mechanisms of action and resistance than did the pattern of *relative* activities across the cell lines. He therefore subtracted out the log-mean over the 60 cell lines to obtain the very useful 'mean-graph' representation of activity data. The lack of information on mechanism in absolute potency values was later corroborated formally by principal components analysis [3–5].

The mean graph representation of patterns led to the COMPARE algorithm [6,7]. Given one compound as a 'seed', COMPARE searches the database of screened agents and compiles a list of those most similar to the seed in their patterns of activity against the NCI-60 panel. Similar patterns generally indicate similarity in mechanism of action, mechanism of resistance, and/or molecular structure. The similarity metric was initially taken as the Euclidean distance, later as the Pearson correlation coefficient. COMPARE has been applied productively to topoisomerase inhibitors [8–13], pyrimidine biosynthesis inhibitors [14], compounds with preferential effects against Nm23-expressing cells [15], anti-mitotics [16–19], and agents active against epidermal growth factor-expressing cells [20], among many other classes of compounds.

In 1992, we introduced feed-forward, back-propagation neural networks (with statistical analysis by cross-validation and sensitivity analysis) to discriminate among various possible mechanisms of drug action on the basis of activity patterns [21]. A large number of other statistical and artificial intelligence techniques have since then been applied to the relationship between pattern and mechanism. Among those methods have been principal components analysis [3–5] and Kohonen self-organizing maps [5,22,23]. Self-organizing maps in this context are used to represent the structural or functional similarities of compounds in the form of two-dimensional maps.

In 1994, we introduced clustering and 'clustered image maps' (CIMs) for analysis and visualization

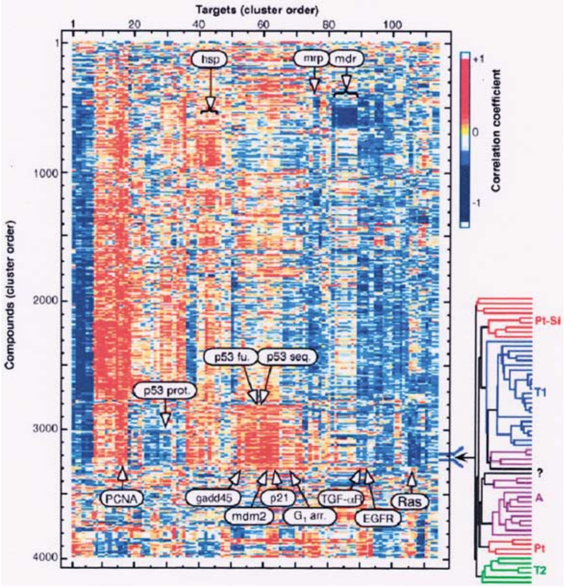

Fig. 2. Clustered Image Map of the relationship between compounds tested and molecular targets in the NCI-60 cells. This normalized $A{\cdot}T^T$ product matrix (where the superscript T indicates the matrix transpose) correlates target patterns with patterns of growth inhibition for a set of 3989 important compounds. A red or orange point (high positive Pearson correlation coefficient) indicates that the agent tends to be selectively active in the SRB assay against cells lines that express the target in large amounts (or in functional form). A dark blue point (high negative correlation) indicates the opposite. The 113 columns correspond to 76 distinct target molecules or functions, some represented multiple times in different mathematical transformations. Compounds and targets have been cluster-ordered by an average linkage algorithm to bring like together with like. To the right is shown one 61-leaf 'twig' of the overall 3989-leaf cluster tree of compounds. Symbols for mechanisms of action are as follows: T1, topoisomerase 1 inhibitors; T2, topoisomerase 2 inhibitors; A, alkylating agents; Pt, platinum compounds; Pt–Si, platinum agents containing a silane moiety; ?, mechanism unknown. The most prominent features are a red patch that indicates compounds (2802–3309) that tend to be active in the assay in cell lines with intact p53 function and a blue patch that indicates compounds (513–667) selectively inactive in Mdr-1/Pgp-expressing cells. Modified from [25].

of the pharmacological and molecular data [24,25]. Fig. 2 shows a CIM that correlates the activity patterns with molecular characteristics ('targets') of the cells, including gene expression data. CIMs have since become the most popular way to represent gene expression data sets visually (although they by no means capture all of the information available in those data).

2.3. *The Structure (S) database*

The chemical structures in S can be coded in terms of any set of one-, two-, or three-dimensional descriptors. Useful structural codings can be found at the DTP's web site. Analyses that relate the S and A databases can be thought of as generalizations of the Q-SAR ('quantitative structure-activity' relationships) paradigm. A number of studies have highlighted various aspects of these relationships for the NCI-60 [4,26–30]. Genetic function approximation [31] (an amalgam of genetic algorithm for variable selection and regression splines for data fitting) proved a useful approach [4,26,28]. Since structural descriptors are available for $>500\,000$ compounds [32], it has been possible to map interesting patterns of activity into the S database and develop abstract pharmacophore templates with which to search the $>400\,000$ compounds not yet screened and bring candidate compounds into the testing process.

2.4. *The Target (T) database*

2.4.1. *Miscellaneous molecular targets*

The first molecular target analyzed experimentally and analytically was the drug resistance transporter *P*-glycoprotein (Pgp), encoded by the multi-drug resistance gene Mdr-1 [33–36]. Fig. 2, a clustered image map obtained by combining information from the T and A databases, shows the importance of Pgp/Mdr-1 to the pattern of drug sensitivities of the cell lines. The dark blue patch for compounds 513–667 indicates that those compounds are negatively correlated with targets 81 to 88, which are the indices of Pgp/Mdr-1 expression and function. The statistics were impressive. We analyzed NCI-60 data for a set of 35 compounds of diverse structure and mechanism that had been reported previously, on the basis of transport assays, to be Mdr-1 substrates [33,35,37,38]. Of those, 18 (51%) fell within the blue patch, whereas only 4% would have been expected to do so by chance. The probability (exact binomial) of such an extreme enrichment being found by chance is < 0.0001. Eighteen of the 35 reported substrates fell within the blue patch, whereas 0 of 12 compounds studied and reported *not* to be substrates [33,35,37,38] did so ($P = 0.001$ by Fisher's exact test). As might have been expected from the known pharmacophoric properties of Pgp substrates, compounds 513–667 were highly enriched for natural products of high molecular weight, often cationic. By linear discriminant analysis, we found that those three factors could predict with a specificity of 84% and sensitivity of 78% which compounds would be found in the blue patch ($P < 0.0001$). Columns 76 and 77 in Fig. 2 are indices of mRNA expression for Mrp-1, another transporter molecule associated with multidrug resistance [34]. There was little overlap between compounds sensitive to Mrp-1 and those sensitive to Pgp/Mdr-1. These calculations provided a proof of principle for the pattern recognition process [25]. Various other molecular targets have been assessed in the NCI-60 system, most prominently a set of molecular characteristics associated with p53 function [39]. Data on miscellaneous targets can be found at http://dtp.nci.nih.gov.

2.4.2. 'Omic' profiling

To complement studies in our laboratory and many others of individual targets in the NCI-60, we have taken an 'omic' approach [40,41], characterizing DNA, mRNA, and protein species in the cells in aggregate. The result is the richest, most varied profiling of any set of cells that we know of.

2.4.2.1. Proteomic and DNA-level profiling. We began with proteins in the early 1990's, doing 2-D gel electrophoresis [42] and developing a MALDI-TOF mass spectroscopic protocol for identifying proteins on the gels [43]. However, by that time it was clear that identification of hundreds or thousands of proteins was not a job for a small academic laboratory. Hence, we decided to wait for the proteomic technologies to improve and, meanwhile, dropped back to the transcript level, where the task appeared easier. Those studies will be described in the next sections.

In parallel, with the transcriptomic studies, we have undertaken collaborations with a number of laboratories for profiling at the DNA level: with the laboratory of Ilan Kirsch and Anna Roschke (NCI) for spectral karyotyping (SKY) and comparative genomic hybridization (CGH) ([44]; also Roschke, et al., in preparation); with that of Kenneth Buetow (NCI) using Affymetrix SNP chips for single nucleotide polymorphisms (Alexander, et al., in preparation); with that of Joe Gray at the University of California Cancer Center for CGH based on BAC arrays (Chen, et al. and Bussey, et al., in preparation); with the laboratories of David Munroe (NCI) and Andrew Feinberg (Johns Hopkins University) for detailed sequence analysis of cytosine methylation in the promoter regions of cancer-related genes (Reinhold, et al. and Maunakea, et al., in preparation).

Most recently, at the protein level, with Lance Liotta (NCI) and Emanuel Petricoin (Food and Drug Administration), we have developed high-density 'reverse-phase' protein lysate microarray for proteomic profiling of the 60 lines without the need for spot identification ([45] and Nishizuka, et al., in preparation). For validation of hypotheses directed toward the clinic, we have used tissue arrays produced by the TARP (Tissue Array Program) Consortium at the NCI [45]. The arrays consist of cores from 503 human tumors of disparate types plus 62 normal human tissues. Although this article focuses on the transcriptome, it is worth noting that a major part of our effort is devoted to understanding, and capitalizing on, the relationship among the various types of data. Not entirely in jest, we refer to this enterprise as 'integromics'.

2.4.2.2. Transcriptomic profiling. In part, the challenge at the mRNA level appeared easier because there are 'only' 30–60 000 independent transcripts and perhaps 200,000 splice variants of those transcripts, rather than the 500 000–2 000 000 functional protein states. We were able to generate transcript profiles for the NCI-60 using four different platforms: a 7907-clone cDNA array with the Brown/Botstein laboratory [46,47], a 6800-gene Affymetrix oligonucleotide chip (Hu6800) with the Golub/Lander group [48], and the Hu95 and Hu133 Affymetrix oligonucleotide chips with Uwe Scherf at Gene Logic, Inc. Versions of the first two data sets used for our calculations are available at http://discover.nci.nih.gov.

Transcript expression profiling by cDNA array. The methods used in this study have been described in detail elsewhere [46,47]. Very briefly, cells were harvested (with less than 1 minute from incubator to stabilization of the preparation) at approximately 80% confluence. Total RNA was stored and then further purified to obtain poly-A mRNA shortly prior to hybridization with microarrays (Synteni, Inc.; now Incyte, Inc.) consisting of robotically spotted, PCR-amplified cDNAs on coated glass slides [49].

The 9703 DNA elements on the array were cDNAs from the Washington University/Merck IMAGE set, obtained from Research Genetics, Inc. The array included 3700 named genes, 1900 human genes homologous to those of other organisms, and 4104 ESTs of unknown function but defined chromosome map location. For each hybridization, cDNA from the test cell's mRNA was labeled by incorporation of Cy5-dNTP during reverse transcription. cDNA synthesized from pooled mRNA of 12 highly diverse cell lines out of the 60 [47] was analogously labeled by incorporation of Cy3-dNTP. Cells for the pool were selected to satisfy 3 criteria [47]: (*i*) at least one cell line from each organ of origin; (*ii*) diversity of growth rates; (*iii*) diversity in terms of protein expression pattern, based on prior two-dimensional gel studies [42]. After appropriate filtering, we settled on a data set of 1376 clones for detailed analysis and added forty miscellaneous cancer-related targets from the DTP database.

Fig. 3a shows a cluster tree that represents the patterns of gene expression across the cell lines. As indicated by the accompanying annotations, there is considerable, but not complete, regularity by organ of origin. Fig. 3b shows the strikingly different tree obtained when the same cells are clustered on the basis of drug activity. The 'correlation of correlations' [47] between the two trees was only +0.21. The correlation of correlations, r_c, is a parameter we developed to quantitate the similarity of two clusterings. In the present context, r_c is the mean Pearson correlation coefficient of the Pearson correlation coefficients relating all 1770 possible pairs of cell types in terms of their response to drugs and in terms of their gene expression. More generally, r_c can be used to quantitate the similarity of any two distance matrices such as those used in hierarchical clustering. For example, we have used it to compare different distance metrics applied to one data set and to compare the data obtained from different microarray platforms [50]. Values of 1, 0, and −1 indicate perfect similarity, no similarity, and perfect inverse similarity, respectively. We should perhaps have expected the low correlation of correlations between the drug- and gene-based clusterings, but we did not. The reason for it appears to be that certain gene products, most prominently Pgp, have a disproportionate effect on activity profiles that cuts across organ of origin distinctions.

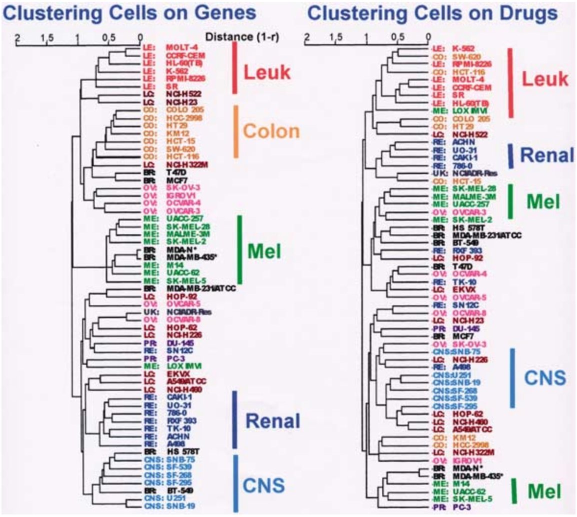

Fig. 3. Dendrograms showing average-linkage hierarchical clustering of human cancer cell lines. (**a**) Cluster tree of the 60 cell lines based on their gene expression profiles for 1376 genes and 40 individual targets. 100% of the colon cancer lines (CO) (7/7) the central nervous system lines (CNS) (6/6), and the leukemias (LE) (6/6) clustered together. Seven out of 8 melanoma lines (ME) clustered together, the exception being the one reported to lack melanin production (LOX-IMVI). Seven out of 8 renal carcinoma lines (RE) clustered together, as did four out of 6 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 estrogen receptor-positive breast lines, T-47D and MCF7, appeared together and grouped with the colon lines, whereas the estrogen receptor-negative 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 1400 compounds tested. The color of the cell line name indicates its assigned organ of origin classification. The distance metric used was (1 – Pearson correlation coefficient). * Indicates 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. Modified from [47].

Fig. 4 shows a CIM that summarizes all possible pairwise relationships between the gene database and a set of 118 drugs of putatively known mechanism of action. Each patch of color represents a story – which may be causally interesting, epiphenomenal, or statistical coincidence. There is clearly not sufficient statistical power to eliminate most of the false positive associations without losing most of the true positive ones. Hence, we must generally consult the literature and public databases for clues to determine which relationships are worth pursuing.

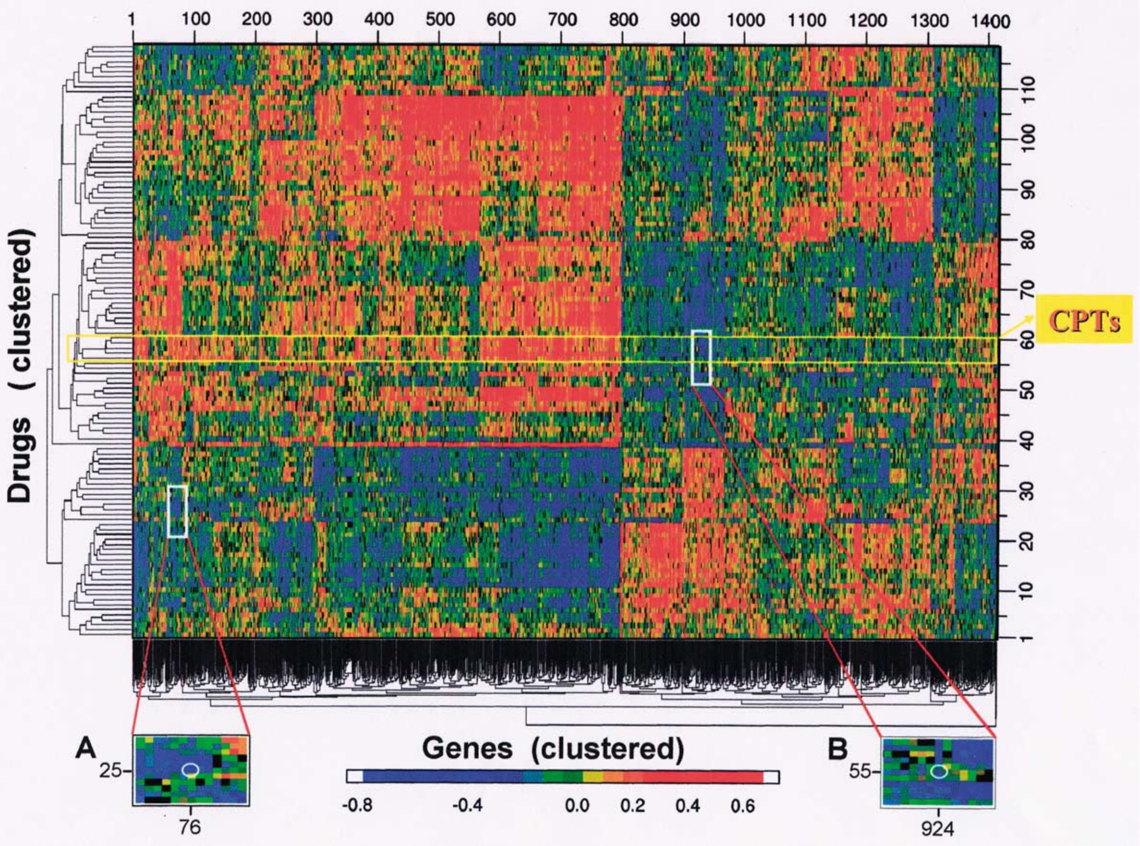

Fig. 4. Clustered image map (CIM) relating activity patterns of 118 tested compounds to the expression patterns of 1376 genes in the 60 cell lines. Included in addition to the gene expression levels are data for 40 molecular targets assessed one at a time in the cells. A red point (high positive Pearson correlation coefficient) indicates that the agent tends to be more active (in the two-day 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). Sharp edges of the colored patches reflect deep forks in the corresponding cluster tree. The position of the topoisomerase 1 inhibitor camptothecin (and its analogues) is indicated. Insert **A** shows a magnified view of the region around the point (white circle) representing the correlation between the dihydropyrimidine dehydrogenase gene and 5-fluorouracil. Insert **B** is an analogous magnified view for the asparagine synthetase gene and the drug L-asparaginase. Modified from [47].

Gene expression profiling by Affymetrix oligonucleotide chip. The methods used have been described previously [48]. Very briefly, mRNA was obtained from the cells [47] and used to prepare biotinylated cDNA, which was hybridized to Hu6800 arrays (Affymetrix, Santa Clara, CA). The resulting cell clusters generally reflected what we had found with the cDNA arrays. We then cross-compared the oligonucleotide and cDNA array data to generate a robust database of >1600 transcripts for which results from the two very different technologies are reasonably concordant across the 60 cell types [50]. That ‘mutually validated’ database has proved particularly useful when we want a firm statistical basis for further analyses.

2.4.2.3. The bioinformatics of transcript profiling. Anyone who does gene expression profiling (or similar omic experiments) for molecular targets finds that most of the time and energy are spent *after* the experiment – in statistical analysis of the data and then in biological interpretation. The problems are particularly acute in integromic studies because we are trying to integrate so many types of information – at the DNA, RNA, protein, functional, and pharmacological levels. Motivated by the needs of our experimental program, we have developed a number of algorithms and computer program packages to assist in the analysis and interpretation steps. These programs, publicly available at http://discover.nci.nih.gov, are proving useful to others as well.

CIM-Miner generates color-coded Clustered Image Maps (CIMs) (also called clustered heat maps) to represent 'high-dimensional' data sets such as gene expression profiles. We introduced CIMs in the mid-1990's for data on drug activities, target expression levels, gene expression values, and proteomic profiles [24,25,51]. The clustering of both axes (or sometimes only one if there is another organizing principle for the second axis) puts like together with like to create patterns of color. A program for producing CIMs can be found at http://discover.nci.nih.gov. Each patch of color in a CIM (e.g., in Fig. 4) represents a possible story. But how can we determine whether a patch represents a causally interesting story, an epiphenomenal correlation (which still may identify a useful molecular marker), or statistical coincidence? As noted in the last section, the usual answer is that we must consult the biomedical literature and public databases. Since that can be a tedious process, we developed a program package called MedMiner for efficient searching and organization of the literature on complex gene, gene–gene, and gene–drug relationships.

MedMiner [52] publicly available at http://discover.nci.nih.gov) uses a combination of GeneCards from the Weizmann Institute, PubMed from the National Library of Medicine (NLM), syntactic analysis, truncated-keyword filtering of relationals, and user-controlled sculpting of Boolean queries to identify key sentences from pertinent abstracts. Those sentences are then organized so that the user can access the most pertinent ones directly by clicking on a relational relevance-term. Whole abstracts of interest can then be accessed quickly through a direct link to PubMed and dropped into a 'shopping basket' for display or for automated entry into a library under EndNote (ISI ResearchSoft, Berkeley, CA) or other bibliographic software. Experienced users have estimated that MedMiner speeds up 5- to 10-fold the process of capturing and organizing the literature from PubMed searches on gene–gene and gene-drug relationships.

MatchMiner [53] publicly available at http://discover.nci.nih.gov provides a solution to the major problem of translating among various gene identifier types for lists of hundreds or thousands of genes. Currently included are GenBank accession numbers, IMAGE clone ids, common gene names, gene symbols, UniGene clusters, FISH-mapped BAC clones, Affymetrix identifiers, and chromosome locations. The LookUp function in MatchMiner makes such translations, providing the user with diagnostics that indicate how the translation was done. The Merge function finds the intersection of two lists of genes, which may be designated by either the same or different identifiers. This functionality is particularly important to our 'integromic' efforts to meld information from the variety of different data types on the NCI-60.

GoMiner [54] publicly available at http://discover.nci.nih.gov provides an answer to the vexing question, "Now that I've done the gene expression experiment and identified a set of 'interesting' genes, what do those genes mean biologically?" To address that question, GoMiner batch-processes and organizes lists of thousands or tens of thousands of genes and provides two fluent, robust visualizations of the genes embedded within the framework of the Gene Ontology hierarchy. One is a tree-like structure; the other is a 'directed acyclic graph'. GoMiner calculates summary statistics indicating for each GO category whether it is enriched with, or depleted of, 'interesting' genes and gives p-values with which to assess the statistical robustness of the enrichment or depletion.

LeadScope/LeadMiner$^{\text{TM}}$ [55] provides a firm link between molecular markers and the drug discovery process. More precisely, it links gene expression profiles for the NCI-60 (or other cell panels used for screening) to a set of 27 000 chemical substructure descriptors of the compounds tested against the cells. One can use it, for example, to identify substructure classes that are found in compounds active in the screen against cell types that express large amounts of a particular gene. That is precisely what a medicinal chemist or researcher designing a directed combinatorial library would like to be able to do in pursuing pharmacogenomic goals.

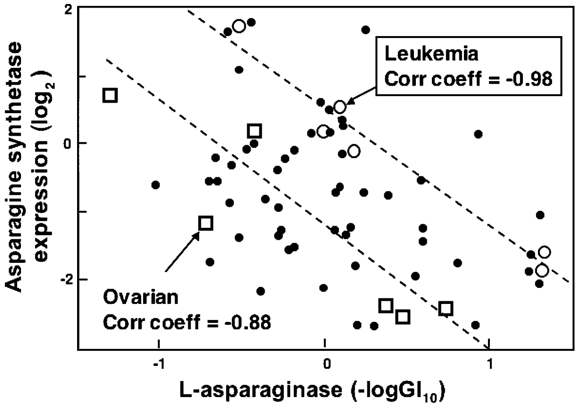

Fig. 5. Relationship between asparagine synthetase expression levels and chemosensitivity of the NCI cell lines to L-asparaginase. Main effects have been removed for both cells and drugs. Hence, a negative $\log(GI_{50})$ value of 1 for sensitivity indicates a 10-fold higher than average sensitivity of the cell line to the agent. The asparagine synthetase expression level is plotted as the abundance of the asparagine synthetase transcript, relative to its abundance in the reference pool of 12 cell lines. A value of +2 indicates 4-fold higher expression than in the reference pool. The large circles indicate leukemia cell lines. The linear regression line (correlation coefficient = −0.98; *P* value <0.01) was fitted to the leukemia data. Modified from [47].

2.4.3. Pharmacogenomic use of NCI-60 transcript profiles: An example

The white rectangle on the gene expression vs. drug sensitivity CIM in Fig. 4 points to a story with likely causal significance on the basis of literature information. That story [47] involves the gene asparagine synthetase and the bacterial enzyme-drug L-asparaginase. Many acute lymphoblastic leukemias (ALL) lack asparagine synthetase and therefore must scavenge exogenous L-asparagine to survive. This dependence is exploited by treating ALL with bacterial L-asparaginase, which depletes extracellular L-asparagine and selectively starves the cancer cells. Fig. 5 shows the relationship between L-asparaginase activity and asparagine synthetase expression across the NCI-60. As might have been predicted on the basis of the above mechanism, there was a statistically robust negative correlation (−0.44; bootstrap 95% confidence interval −0.59 to −0.25) between expression of the asparagine synthetase gene and L-asparaginase sensitivity in the 60 cell lines [47]. Although statistically robust, the correlation was only moderately strong. We knew, however, to focus specifically on the leukemic subpanel, and in that case the correlation was a striking −0.98 (bootstrap 95% confidence interval −1.00 to −0.93). This value survived even a Bonferroni correction for statistical multiple comparisons. Furthermore, the two ALL-derived lines expressed the lowest levels of asparagine synthetase mRNA and were the most sensitive to L-asparaginase, as might have been predicted. These results supported the possible use of asparagine synthetase as a marker for clinical decisions about L-asparaginase therapy [47].

We then asked whether any other cell line panel showed similar correlation. The answer was ‘yes’, though not as strongly. The correlation coefficient for the ovarian lines was −0.88 (confidence interval −0.23 to −0.99) [47]. Early clinical trials done with an assortment of solid tumors showed occasional responses to L-asparaginase in melanoma, chronic granulocytic leukemia, lymphosarcoma, and reticulum cell sarcoma but not in other tumor types (see [47] for references). The microarray findings, however, support a closer look at L-asparaginase therapy for solid tumors, particularly for a subset of ovarian cancers low in asparagine synthetase. Further studies of this correlation are underway in collaboration with D. von Hoff (Arizona Cancer Center). The preferred material for a clinical trial would be the polyethylene glycol-modified forms of L-asparaginase, which shows much better pharmacokinetic and immunological properties than does the native bacterial form of the enzyme.

3. Concluding remarks

Pharmacogenomic profiling – or, in accord with the title of this contribution, should we call it ‘pharmacotranscriptomic profiling – holds undeniable promise for molecular subsetting of patients and for individualization of therapy. Much of the research to realize those aims can be done with clinical materials, rather than cultured cells, if a number of purely technical challenges are overcome. But the limitation of clinical tumors that cannot be overcome, is this: they have not been exposed to large numbers of chemical compounds one at a time and independently under well-defined experimental control.

Acknowledgements

We are grateful to E.A. Sausville, A. Monks, D.A. Scudiero, K.D. Paull, and others in the NCI DTP, whose work over the years and in collaboration with us has made these studies of the NCI-60 cell lines possible. We thank D.A. Ross and others in the Brown/Botstein laboratory at Stanford University and J. Staunton and others in the Golub/Lander group at the Whitehead Institute for collaborations that generated the transcript profiles discussed here. We are most grateful to P. Blower and colleagues at LeadScope, Inc. for collaborating in development of the LeadScope/LeadMiner program package; to M. Wang's group at the Georgia Institute of Technology and Emory University for major contributions to the GoMiner program; to J. Riss and J.C. Barrett at the NCI, also for contributions to GoMiner. We particularly thank the many members of the NCI Laboratory of Molecular Pharmacology, past and present, and SRA International staff (on contract for software development) whose efforts in recent years have made the work described here possible. These developments have been made possible largely by funding from the Center for Cancer Research, NCI.

References

[1] J.N. Weinstein, Linking drugs and genes: Pharmacogenomics, pharmacoproteomics, bioinformatics, and the NCI-60, in: C. Brenner, D.J. Duggan (Eds.), Oncogenomics: Molecular approaches to cancer, New York (in press).

[2] M.R. Emmert-Buck, R.F. Bonner, P.D. Smith, R.F. Chuaqui, Z. Zhuang, S.R. Goldstein, R.A. Weiss, L.A. Liotta, Laser capture microdissection, Science 274 (1996) 998–1001.

[3] A. Koutsoukos, L. Rubinstein, D. Faraggi, S. Kalyandrug, J.N. Weinstein, K.D. Paull, K.W. Kohn, R.M. Simon, Discrimination techniques applied to the NCI in vivo antitumor drug screen: Predicting biochemical mechanism of action, Stat. Med. 13 (1994) 719–730.

[4] L.M. Shi, J.K. Lee, Y. Fan, M. Waltham, D.T. Andrews, U. Scherf, K.D. Paull, J.N. Weinstein, Mining and visualizing large anticancer drug discovery databases, J. Chem. Inf. Comput. Sci. 40 (2000) 367–379.

[5] O. Keskin, I. Bahar, R.L. Jernigan, J.A. Beutler, R.H. Shoemaker, E.A. Sausville, D.G. Covell, Characterization of anticancer agents by their growth inhibitory activity and relationships to mechanism of action and structure, Anticancer Drug Des. 15 (2000) 79–98.

[6] M.R. Boyd, The future of new drug development, in: J.E. Neiderhuber (Ed.), Current Therapy in Oncology, Philadelphia, 1992.

[7] K.D. Paull, R.H. Shoemaker, L. Hodes, A. Monks, D.A. Scudiero, L. Rubinstein, J. Plowman, M.R. Boyd, 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 (1989) 1088–1092.

[8] M. Gupta, A. Fujimori, Y. Pommier, DNA topoisomerase I, Biochim. Biophys. Acta 1262 (1995) 1–14.

[9] F. Leteurtre, D.L. Sackett, J. Madalengoitia, G. Kohlhagen, T. Macdonald, E. Hamel, K.D. Paull, Y. Pommier, Azatoxin derivatives with potent and selective action on topoisomerase II, Biochem. Pharmacol. 49 (1995) 1283–1290.

[10] F. Leteurtre, G. Kohlhagen, K.D. Paull, Y. Pommier, Topoisomerase II inhibition by anthrapyrazoles, DuP 937 & DuP 941 (Losoxanthrone) and cytotoxicity in the NCI cell screen, J. Natl Cancer Inst. 86 (1994) 1239–1244.

[11] E. Solary, F. Leteurtre, K.D. Paull, D. Scudiero, E. Hamel, Y. Pommier, Dual inhibition of topoisomerase II and tubulin polymerization by azatoxin, a novel cytotoxic agent, Biochem. Pharmacol. 45 (1993) 2449–2456.

[12] H.M. Koo, A. Monks, A. Mikheev, L.V. Rubinstein, M. Gray-Goodrich, M.J. McWilliams, W.G. Alvord, H.K. Oie, A.F. Gazdar, K.D. Paull, H. Zarbl, G. Vande Woude, Enhanced sensitivity to 1-beta-D-arabinofuranosylcytosine and topoisomerase II inhibitors in tumor cell lines harboring activated ras oncogenes, J. Natl Cancer Inst. 56 (1996) 5211–5216.

[13] G. Kohlhagen, K. Paull, M. Cushman, P. Nagafufuji, Y. Pommier, Protein-linked DNA strand breaks induced by NSC 314622, a non-camptothecin topoisomerase I poison, Mol. Pharmacol. 54 (1998) 50–58.

[14] E.S. Cleaveland, A. Monks, A. Vaigro-Wolff, D.W. Zaharevitz, K. Paull, K. Ardalan, D.A. Cooney, H. Ford Jr., Site of action of two novel pyrimidine biosynthesis inhibitors accurately predicted by the compare program, Biochem. Pharmacol. 49 (1995) 947–954.

[15] J.M. Freije, J.A. Lawrence, M.G. Hollingshead, A. de la Rosa, V. Narayanan, M. Grever, E.A. Sausville, K. Paull, P.S. Steeg, Identification of compounds with preferential inhibitory activity against low-Nm23-expressing human breast carcinoma and melanoma cell lines, Nature Med. 3 (1997) 395–401.

[16] R. Bai, K.D. Paull, C.L. Herald, L. Malspeis, G.R. Pettit, E. Hamel, B. Halichondrin, homohalichondrin B, marine natural products binding in the vinca domain of tubulin. Discovery of tubulin-based mechanism of action by analysis of differential cytotoxicity data, J. Biol. Chem. 266 (1991) 15882–15889.

[17] E. Hamel, C.M. Lin, J. Plowman, H.K. Wang, K.H. Lee, K.D. Paull, Antitumor 2,3-dihydro-2-(aryl)-4(1H)-quinazolinone derivatives. Interactions with tubulin, Biochem. Pharmacol. 51 (1996) 53–59.

[18] S.C. Kuo, H.Z. Lee, J.P. Juang, Y.T. Lin, T.S. Wu, J.J. Chang, D. Lednicer, K.D. Paull, C.M. Lin, E. Hamel, et al., Synthesis and cytotoxicity of 1,6,7,8-substituted 2-(4′-substituted phenyl)-4-quinolones and related compounds: identification as antimitotic agents interacting with tubulin, J. Med. Chem. 36 (1993) 1146–1156.

[19] K.D. Paull, C.M. Lin, L. Malspeis, E. Hamel, Identification of novel antimitotic agents acting at the tubulin level by computer-assisted evaluation of differential cytotoxicity data, Cancer Res. 52 (1992) 3892–3900.

[20] K. Wosikowski, D. Schuurhuis, K. Johnson, K.D. Paull, T.G. Myers, J. Weinstein, S.E. Bates, Identification of epidermal growth factor receptor and c-erbB2 pathway inhibitors by correlation with gene expression patterns, J. Natl Cancer Inst. 89 (1997) 1505–1513.

[21] J.N. Weinstein, K.W. Kohn, M.R. Grever, V.N. Viswanadhan, L.V. Rubinstein, A.P. Monks, D.A. Scudiero, L. Welch, A.D. Koutsoukos, A.J. Chiausa, K.D. Paull, Neural computing in cancer drug development: predicting mechanism of action, Science 258 (1992) 447–451.

[22] W.W. van Osdol, T.G. Myers, K.D. Paull, K.W. Kohn, J.N. Weinstein, Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents, J. Natl Cancer Inst. 86 (1994) 1853–1859.

[23] W.W. van Osdol, T.G. Myers, J.N. Weinstein, Neural network techniques for the informatics of cancer drug discovery, Methods in Enzymology 321 (2000) 369–395.

[24] J.N. Weinstein, T.G. Myers, J.K. Buolamwini, K. Raghavan, W. van Osdol, J. Licht, V.N. Viswanadhan, K.W. Kohn, L.V. Rubinstein, A.D. Koutsoukos, A.P. Monks, D.A. Scudiero, N.L. Anderson, D. Zaharevitz, B.A. Chabner, M.R. Grever, K.D. Paull, Predictive statistics and artificial intelligence in the US National Cancer Institute's drug discovery program for cancer and AIDS, Stem Cells 12 (1994) 13–22.

[25] J.N. Weinstein, T.G. Myers, P.M. O'Connor, S.H. Friend, A.J. Fornace, K.W. Kohn, T. Fojo, S.E. Bates, L.V. Rubinstein, N.L. Anderson, J.K. Buolamwini, W.W. van Osdol, A.P. Monks, D.A. Scudiero, E.A. Sausville, D.W. Zaharevitz, B. Bunow, V.N. Viswanadhan, G.S. Johnson, R.E. Wittes, K.D. Paull, An information-intensive approach to the molecular pharmacology of cancer, Science 275 (1997) 343–349.

[26] Y. Fan, L.M. Shi, T.G. Myers, K.W. Kohn, Y. Pommier, J.N. Weinstein, Quantitative structure-antitumor activity relationships of camptothecins: cluster analysis and genetic algorithm-based studies, J. Med. Chem. 44 (2001) 3254–3263.

[27] L.M. Shi, Y. Fan, T.G. Myers, P.M. O'Connor, K.D. Paull, S.H. Friend, J.N. Weinstein, Mining the NCI anticancer drug discovery databases: Genetic function approximation for the QSAR study of anticancer ellipticine analogues, J. Chem. Inf. Comput. Sci. 38 (1998) 189–199.

[28] L.M. Shi, Y. Fan, T.G. Myers, M. Waltham, K.D. Paull, J.N. Weinstein, Mining the anticancer activity database generated by the US National Cancer Institute's drug discovery program using statistical and artificial intelligence techniques, J. Chem. Inf. Comput. Sci. (1998).

[29] L.M. Shi, T.G. Myers, Y. Fan, P.M. O'Connor, K.D. Paull, S.H. Friend, J.N. Weinstein, Mining the National Cancer Institute Anticancer Drug Discovery Database: cluster analysis of ellipticine analogs with p53-inverse and central nervous system-selective patterns of activity, Mol. Pharmacol. 53 (1998) 241–251.

[30] A.A. Rabow, R.H. Shoemaker, E.A. Sausville, D.G. Covell, Mining the National Cancer Institute's tumor-screening database: identification of compounds with similar cellular activities, J. Med. Chem. 45 (2002) 818–840.

[31] D. Rogers, A.J. Hopfinger, Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships, J. Chem. Inf. Comput. Sci. 34 (1994) 854–866.

[32] G.W.A. Milne, M.C. Nicklaus, J.S. Driscoll, S. Wang, D. Zaharevitz, National Cancer Institute drug information system 3D database, J. Chem. Inf. Comput. Sci. 34 (1994) 1219–1224.

[33] M. Alvarez, K.D. Paull, C. Hose, J.S. Lee, J.N. Weinstein, M. Grever, S. Bates, T. Fojo, 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. Investig. 95 (1995) 2205–2214.

[34] M.A. Izquierdo, R.H. Shoemaker, M.J. Flens, G.L. Scheffer, L. Wu, T.R. Prather, Overlapping phenotypes of multidrug resistance among panels of human cancer-cell lines, Int. J. Cancer 65 (1996) 230–237.

[35] J.S. Lee, K.D. Paull, M. Alvarez, C. Hose, A. Monks, M. Grever, A.T. Fojo, S.E. Bates, Rhodamine efflux patterns predict P-glycoprotein substrates in the National Cancer Institute drug screen, Molec. Pharmacol. 46 (1994) 627–638.

[36] L. Wu, A.M. Smythe, S.F. Stinson, L.A. Mullendore, A. Monks, D.A. Scudiero, K.D. Paull, A.D. Koutsoukos, L.V. Rubinstein, M.R. Boyd, R.H. Shoemaker, Multidrug-resistant phenotype of disease-oriented panels of human tumor cell lines used for anticancer drug screening, Cancer Res. 52 (1992) 3029–3034.

[37] K.V. Chin, I. Pastan, M.M. Gottesman, Function and regulation of the human multidrug resistance gene, Adv. Cancer Res. 60 (1993) 157–180.

[38] M.M. Gottesman, I. Pastan, Biochemistry of multidrug resistance mediated by the multidrug transporter, Annu. Rev. Biochem. 62 (1993) 385–427.

[39] P.M. O'Connor, J. Jackman, I. Bae, T.G. Myers, S. Fan, M. Mutoh, D.A. Scudiero, A. Monk, E.A. Sausville, J.N. Weinstein, S. Friend, J. Fornace, K.W. Kohn, 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 (1997) 4285–4300.

[40] J.N. Weinstein, Fishing expeditions, Science 282 (1998) 627–628.

[41] J.N. Weinstein, 'Omic' and hypothesis-driven research in the molecular pharmacology of cancer, Curr. Opin. Pharmacol. 2 (2002) 361–365.

[42] T.G. Myers, M. Waltham, G. Li, J.K. Buolamwini, D.A. Scudiero, L.V. Rubinstein, K.D. Paull, E.A. Sausville, N.L. Anderson, J.N. Weinstein, A protein expression database for the molecular pharmacology of cancer, Electrophoresis 18 (1997) 647–653.

[43] G. Li, M. Waltham, E. Unsworth, A. Treston, J. Mushine, N.L. Anderson, K.W. Kohn, J.N. Weinstein, Rapid protein identification from two-dimensional polyacrylamide gels by MALDI mass spectrometry, Electrophoresis 18 (1997) 647–653.

[44] W.C. Reinhold, H. Kouros-Mehr, K.W. Kohn, A.K. Maunakea, S. Lababidi, A. Roschke, K. Stover, J. Alexander, P. Pantazis,

L. Miller, E. Liu, I.R. Kirsch, Y. Urasaki, Y. Pommier, J.N. Weinstein, Apoptotic susceptibility of cancer cells selected for camptothecin resistance: Gene expression profiling, functional analysis, and molecular interaction mapping, Cancer Res. 63 (1993) 1000–1011.

[45] S. Nishizuka, S.-T. Chen, F.G. Gwadry, J. Alexander, U. Scherf, W.C. Reinhold, M. Waltham, L. Charboneau, L. Young, K.J. Bussey, S. Kim, S. Lababidi, J.K. Lee, S. Pittaluga, P.J. Munson, E. Petricoin, L.A. Liotta, S.M. Hewitt, M. Raffeld, J.N. Weinstein, Diagnostic markers that distinguish colon and ovarian adenocarcinomas: Identification by genomic, proteomic, and tissue array profiling, Cancer Res., in press.

[46] D.T. Ross, U. Scherf, M.B. Eisen, C.M. Perou, C. Rees, P. Spellman, V. Iyer, S.S. Jeffre, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J.C.F. Lee, D. Lashkari, D. Shalon, T.G. Myers, J.N. Weinstein, D. Botstein, P.O. Brown, Systematic variation in gene expression patterns in human cancer cell lines, Nat. Genet. 24 (2000) 227–235.

[47] U. Scherf, D.T. Ross, M. Waltham, L.H. Smith, J.K. Lee, L. Tanabe, K.W. Kohn, W.C. Reinhold, T.G. Myers, D.T. Andrews, D.A. Scudiero, M.B. Eisen, E.A. Sausville, Y. Pommier, D. Botstein, P.O. Brown, J.N. Weinstein, A gene expression database for the molecular pharmacology of cancer, Nat. Genet. 24 (2000) 236–244.

[48] J.E. Staunton, D.K. Slonim, H.A. Coller, P. Tamayo, M.J. Angelo, J. Park, U. Scherf, J.K. Lee, J.N. Weinstein, J.P. Mesirov, E.S. Lander, T.R. Golub, Chemosensitivity prediction by transcriptional profiling, Proc. Natl Acad. Sci. USA 98 (2001) 10787–10792.

[49] D. Shalon, S.J. Smith, P.O. Brown, A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization, Genome Res. 6 (1996) 639–645.

[50] J.K. Lee, U. Scherf, K.J. Bussey, F.G. Gwadry, W.C. Reinhold, G. Riddick, J.N. Weinstein, Comparing cDNA and oligonucleotide array data: Concordance of gene expression across platforms for the NCI-60 cancer cell lines, Proc. Natl Acad. Sci. USA (submitted).

[51] T.G. Myers, J.N. Weinstein, K. Raghavan, J.K. Buolamwini, N.L. Anderson, P. O'Connor, K.W. Kohn, D.A. Scudiero, A.P. Monks, S. Friend, D. Zaharevitz, L.V. Rubinstein, K.D. Paull, An "information intensive" strategy for drug discovery in cancer and AIDS: relating cell cycle factors to patterns of drug activity, Proc. Am. Assoc. Cancer Res. 36 (1995) A1813.

[52] L. Tanabe, L.H. Smith, J.K. Lee, U. Scherf, L. Hunter, J.N. Weinstein, MedMiner: An internet tool for mining information, with application to gene expression profiling, BioTechniques 27 (1999) 1210–1217.

[53] K.J. Bussey, D. Kane, M. Sunshine, S. Narasimhan, S. Nishizuka, W.C. Reinhold, B. Zeeberg, Ajay and Weinstein J.N., MatchMiner: A tool for batch navigation among gene and gene product identifiers, Genome Biol., in press.

[54] B. Zeeberg, W. Feng, G. Wang, M.D. Wang, A.T. Fojo, D.W. Kane, W.C. Reinhold, J.N. Weinstein, GoMiner: A resource for biological interpretation of genomic and proteomic data, Genome Biol., in press.

[55] P.E. Blower, C. Yang, M.A. Fligner, J.S. Verducci, L. Yu, S. Richman, J.N. Weinstein, Pharmacogenomic analysis: correlating molecular substructure classes with microarray gene expression data, Pharmacogenomics J. (Nature) 2 (2002) 259–271.