Spotlight on Molecular Profiling: Commentary

Spotlight on molecular profiling: ''Integromic'' analysis of the NCI-60 cancer cell lines

John N. Weinstein

Genomics and Bioinformatics Group, Laboratory of Molecular Pharmacology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland

''Our horizon is never quite at our elbows''
—Henry David Thoreau, *Walden* (1854)

In this issue, *Molecular Cancer Therapeutics* launches a brave new experiment in the publication of pharmacogenomic and pharmacoproteomic information: a series of invited, refereed articles justified by the broad interest and utility of the molecular profile databases they present, rather than by the testing of a particular biological or pharmacological hypothesis. The initial articles under the rubric ''Spotlight on Molecular Profiling'' focus on molecular profiling of the 60 human cancer cell lines (the NCI-60) used by the National Cancer Institute's Developmental Therapeutics Program (DTP) to screen >100,000 chemically defined compounds and natural product extracts since 1990 (1–4). In statistical and machine-learning analyses, the screening data have proved rich in information about drug mechanisms of action and resistance (5–8). The NCI-60 panel already constitutes by far the most comprehensively profiled set of cells in existence (4, 9), and much more molecular profile information on them is coming. The data have already yielded considerable biological and biomedical insight, but we have only scratched the surface thus far. The real value is realized when biomedical scientists with particular domain expertise are able to integrate and use the information fluently for hypothesis generation, hypothesis-testing, and what I would term ''hypothesis-enrichment.'' Given the large drug activity database, the NCI-60 cell line panel provides a unique opportunity for the enrichment of pharmacologic hypotheses and for advances toward the oft-cited goal of personalized medicine.

Why is there a need for a series of article like this? The broad, generic answer is clear. For almost half a century after Watson and Crick's brainstorm, the dominant paradigm of what might be called the ''pregenomic era'' was hypothesis-driven, R01-funded research focused on particular molecules or processes. That paradigm served us well. But now, thanks largely to technological advances in the ''post-genomic era'', we have access to information on 20,000 to 25,000 genes, >100,000 splice variants of those genes, an unknown number of regulatory RNAs, and perhaps a million protein states of possible functional significance if one counts posttranslation modifications such as the phosphorylations central to cell signaling. Then there are the many types of molecules that make up what have been termed the lipidome, glycome, metabolome, epigenome, immunome, and so forth. That multiplicity constitutes a challenge and an opportunity. To meet the challenge and take advantage of the opportunity, it will be necessary to create and exploit synergies between hypothesis-driven and ''omic'' modes of research (10, 11). Those synergies will be particularly important as researchers try to understand system level interactions among the molecules of what Eric Lander aptly calls ''the parts list'' of the cell.

Those who generate omic data (10) share a common experience: many scientists want access to the data but few are ready to pay for them in academic coin of the realm. That's a major public loss. For example, after a microarray study is done, it typically takes months to find a hypothesis-driven ''story'' in the data to justify publication and many more months to flesh out and validate the story with functional experiments. One consequence is delay in public availability of the data. Another is that the tail ends up wagging the dog; the data are given short shrift, and the article focuses on downstream hypothesis testing (11).

That process runs counter to the current emphasis on availability and interoperability of molecular data. Publishing standards now dictate that the data should be deposited in a public repository such as the Gene Expression Omnibus, ArrayExpress, or Center for Information Biology Gene Expression Database, and that they should meet standards of content and interoperability such as the Minimum Information About a Microarray Experiment protocols. Why, then, should criteria for publication not reflect those aims?

Most molecular databases don't deserve prominent publication in their own right, of course. The technical

Mol Cancer Ther 2006;5(11):2601–5

Received 10/16/06; accepted 10/17/06.

Grant support: The Genomics and Bioinformatics Group's research is supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

Requests for reprints: John N. Weinstein, Laboratory of Molecular Pharmacology, National Cancer Institute, 37 Convent Drive, Room 5056B, Bethesda, MD 20892. Phone: 301-496-9571.
E-mail: weinstein@dtpax2.ncifcrf.gov

doi:10.1158/1535-7163.MCT-06-0640

quality must be high, and, equally important, the data must be of more than parochial significance. The Genome Project sequence is at one end of the spectrum, important to almost every laboratory doing biological or biomedical research; sparse molecular characteristics of particular cell types are at the other end, often of value only to the investigators themselves. Molecular profiling data on the NCI-60 fall somewhere between the two extremes. The data are of interest to thousands of laboratories, both for their basic biological uses and for their connection to cancer therapeutics.

This is not the place for a full-scale review of the NCI-60 panel and its molecular profiles. But a brief summary will be useful to motivate what follows. The panel was initially assembled in the late 1980s by Michael Boyd and colleagues at the DTP, under the aegis of Division Director Bruce Chabner to provide a tissue-specific screening capability (1). Largely through pioneering analyses by the late Kenneth Paull (5), it soon developed a second personality—as a system for profiling the compounds and natural product extracts tested against it. Studies in the laboratories of Tito Fojo, Susan Bates, and Robert Shoemaker added molecular characterization of the cells with respect to MDR1 and other drug resistance transporters (12–14). Broad omic profiling of the cells had its inception in a discussion in Bruce Chabner's office. I challenged him to list the molecules he would most like to see profiled in the cells. To my astonishment, he provided a list the next week. Our opening salvo, in the mid-1990s, was a two-dimensional gel study with Leigh Anderson that produced a database of 1,014 spots indexed over all 60 cell lines. The data were integrated through clustered heat map visualizations of the type that have since then become the ever-present visual icon of post-genomic biology (15). When we submitted the article to a prominent (non-AACR) journal that shall remain nameless, it was promptly rejected without review by an editor who asked, ''Where's the hypothesis?'' We later published the study elsewhere (16). Numerous such experiences that we and others have had highlight the need for the current spotlight series. Four of our database-heavy publications on the NCI-60 over the last decade have thus far accumulated >400 literature citations each (15, 17–19), but all four were initially dismissed by journals, largely because they were viewed as lacking a hypothesis.

Figure 1 shows a schematic we use to organize our thinking about the NCI-60 databases (15). The assay run by DTP produces a database (A) of activities, which can be mapped into molecular structures of the compounds tested (S) or into molecular targets and other characteristics of the cells (T). If other cell or tissue types—e.g., transfectants, knock-downs, knock-outs, clinical tumors—are profiled in a compatible way, then it's possible to extrapolate the phenomenal pharmacologic characterization of the NCI-60 panel to the additional sample types without actually doing the assays in those samples. Often, the additional assays

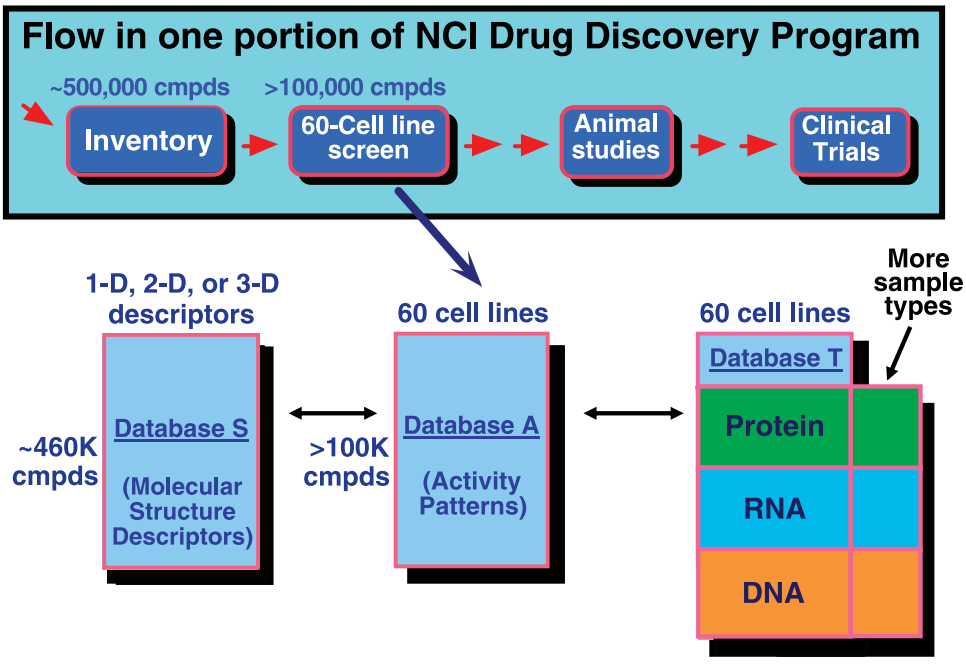

Figure 1. Schematic overview of the NCI-60 databases. Each row of the activity (**A**) database represents the pattern of activity of a particular compound across the 60 cell lines. The **A** database can be mapped into a structure (**S**) database containing one-dimensional, two-dimensional, and/or three-dimensional chemical structure descriptors of the compounds or into a target (**T**) database containing molecular profile data on the cells. The T-database consists of data on individual molecules as well as omic data at the DNA, RNA, protein, and functional levels. The drug activity data can be extrapolated to additional cell or tumor types that have been profiled in the same way as the NCI-60. The bioinformatic challenge is to analyze and understand each of the databases separately, then to integrate them with each other and with public information resources. Modified from ref. (15).

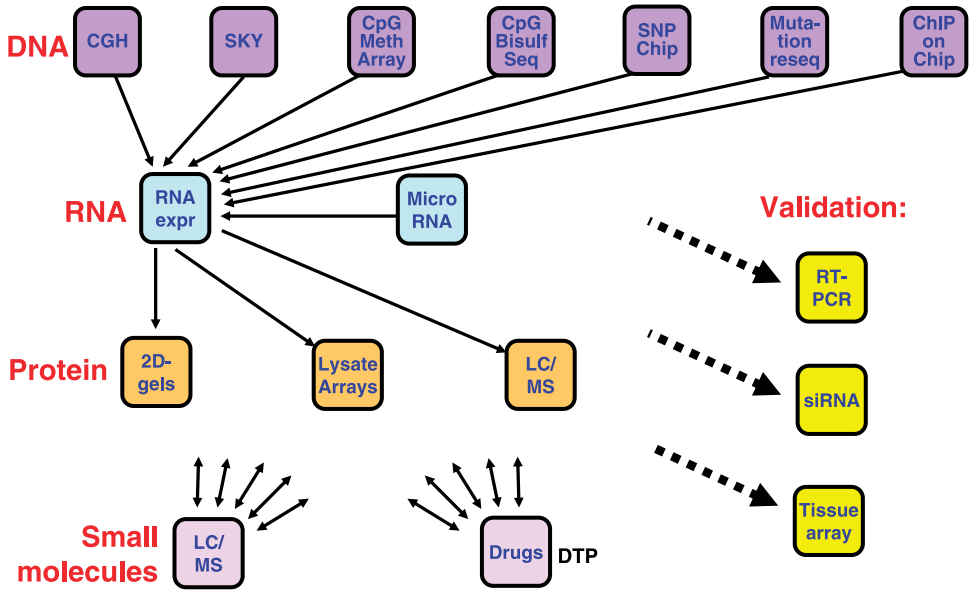

Figure 2. Conceptual schema for molecular profiling of the NCI-60 (or other cancer cells). Some of the profiling studies have already been completed, some are in progress, and some are being piloted currently. Also shown are three levels of validation explained in the text. CGH, comparative genomic hybridization (for DNA copy number); SKY, spectral karyotyping (for chromosomal aberrations); Meth, methylation; Bisulf Seq, bisulfite sequencing; reseq, resequencing; ChIP, chromatin immunoprecipitation; LC/MS, liquid chromatography/mass spectrometry.

would be impossible to do, especially if materials are limited, as they generally are with clinical tumors.

Figure 2 indicates the multiple types of molecular profiling studies that our research group and collaborators have done, are doing, or are piloting in NCI-60 cells. To maximize the consistency and interoperability of the data sets, we adopted standard operating procedures that include the use of the same or matched serum batches, harvesting at a particular percentage of confluence, and minimization of the time from incubator to stabilization of the materials. For consistency, over a period of years, the cell cultures, harvests, and almost all purifications have been done by a single scientist, William Reinhold. The profile data run the gamut from DNA to RNA to protein to small-molecule in focus. The aim is an extensive inventory of those classes of molecules in the cells. That inventory can then be coupled bioinformatically with, or spliced into, the growing body of information on how the parts fit together to form systems. Because of the DTP's activity data on >100,000 compounds and natural product extracts, the link to molecular pharmacology and therapeutics is obvious.

Our new molecular profile databases (and most of the legacy data) will appear in this or future issues of Molecular Cancer Therapeutics. The data are available for download at http://discover.nci.nih.gov in the form of a searchable, relational database (CellMiner). Also at that site will be metadata on the NCI-60 cells and computational tools (the Miner Suite) for integrating the various types of data with each other. The data will also be available at DTP's web site, http://dtp.nci.nih.gov, and in the Gene Expression Omnibus, SKYWEB, or other data repositories, as appropriate to the type of data.

The value of molecular profile inventories is central to The Cancer Genome Atlas project, which is just getting under way as a joint 3-year pilot project of the NCI and the National Human Genome Research Institute. Originally, that project was to focus on the resequencing of a large number of human tumors. It was then realized that the value would be increased enormously by the inclusion of other types of profiling, at least at the genomic, epigenomic, and transcriptomic levels. Currently, the Cancer Genome Atlas's plans are much less expansive than those in Fig. 2, but they must function in the much more difficult context of clinical tumors (lung cancer, ovarian cancer, and glioma for the pilot project). In a sense, the NCI-60 profiling enterprise can be thought of as a stalking-horse for the Cancer Genome Atlas—doing in cell lines what will be much harder to do in clinical tumors.

Everyone knows the limitations of cell lines as surrogates for clinical cancers. Even primary cultures have been removed from the influence of cytokines, hormones, three-dimensional architecture, and the community of other cell types in a tumor. Furthermore, cell lines have been adapted or selected for survival and rapid proliferation on plastic. The NCI-60, in particular, have the disadvantage (or advantage) that they represent diverse lineages, and there are only 60 of them—more than enough for some analyses but too few for others. On the other side of the ledger, the lines are homogeneous in lineage, available in unlimited numbers, manipulable (e.g., by transfection), and useful for high-throughput drug assays. Furthermore, they make it possible to step into the same stream over and over again, and the screening data have a major legacy value. Thankfully, cell lines don't tend to raise

issues of informed consent, and they rarely sue for intellectual property rights. The Cancer Genome Atlas does face issues of confidentiality and intellectual property, as well as the inevitably difficult decisions about how best to use the finite, irreplaceable clinical materials. It also faces the inhomogeneity of cell type in clinical tumors, a problem that will become even more acute if it turns out that we really want information on rare stem cells in a tumor or on cells at the proliferating, invading margin, or information on well-oxygenated cells near blood vessels.

In my view, then, the cell lines should be considered, first and foremost, as instances of biology in their own right. Most of our knowledge about cell biology, physiology, and pharmacology has come from a study of the cell lines. In that context, the NCI-60 metadata and molecular profiles often prove useful when one wants to choose a parental cell type with particular characteristics for transfection or other experimental manipulations. When we're predicting toward the clinic, however, it's caveat emptor. As with any model system, there will be leads and there will be mis-leads. It's necessary to find clues that generate testable hypotheses without worrying too much about the clues that don't work out. As usual in science, one has to find a personally comfortable balance between following up the most improbable observation (which may be the most important) and following up the prosaic ones that are more likely to bear fruit.

Figure 2 includes three levels of validation studies. The first, exemplified by real-time reverse transcription-PCR, simply tests the technical accuracy of microarray data. More substantively, the second level, small interfering RNA knock-down, provides a way to turn correlative information from the NCI-60 into causal information, and the third level, use of tissue arrays, tests in real tumors the hypotheses that arise from NCI-60 data. We most often derive biomedically useful knowledge from the NCI-60 by integrating the various data types with each other—the integromic analysis (4, 5)—and then working back and forth iteratively between those data, the validation data, and information on clinical cancers. That process is often ''seat-of-the-pants'' more than it is statistical; we scramble for clues to formulate new hypotheses, we try to corroborate old ones, or we find ways in which the old ones can be enriched.

To illustrate those ways of approaching and using the data, I'll briefly mention several published instances in which we and our collaborators have made that sort of extrapolation from the NCI-60 in the context of molecular therapeutics. Because this commentary isn't intended as a comprehensive review, with apologies I won't try to do justice to the many others around the world who have made excellent use of the information.

- Nishizuka and colleagues (20) integrated information from cDNA arrays, Affymetrix oligonucleotide arrays, resequencing, reverse-phase protein lysate arrays, and tissue arrays to identify promising candidate biomarkers for distinguishing colon from ovarian tumors of unknown origin.
- Ludwig et al. (21) and Szakacs and colleagues (22) used correlative information from the NCI-60 screen and real-time reverse transcription-PCR profiling to identify what we term ''MDR1-inverse'' compounds, which paradoxically are more active in cells that express large amounts of the drug-resistance transporter MDR1-Pgp. Those correlative results were then corroborated in tet-regulated transfectants, siRNA knock-downs, and selected resistant cell lines. *In vivo* testing is under way.
- Reinhold and colleagues (23) analyzed our gene expression data from the NCI-60, in combination with extensive flow cytometry studies of apoptosis and analysis of molecular interaction maps (24) to formulate the two-step ''Permissive Apoptosis-Resistance'' model for acquisition of drug resistance.
- In 1993, Fojo and colleagues (25) organized the NCI-60 data on hundreds of platinum compounds that had been screened. The resulting clustered heat maps revealed 12 distinct families of compounds on the basis of activity patterns, and 11 of those classes turned out to be structurally homogeneous as well. One of the families, the diaminocyclohexyl group, was relatively more potent in the colon cancer lines. Those analyses, coupled with suggestive results from a clinical trial in France, convinced the company that owned oxaliplatin, a diaminocyclohexyl compound, to proceed with clinical development. Oxaliplatin is now a standard-of-care agent for treatment of primary and recurrent colorectal cancer.

Each of those brief descriptions indicates how partial information can be put together from multiple sources, including the NCI-60 data, for basic molecular pharmacology, drug discovery, or biomarker identification. Two further examples are published in this issue as flagship articles for the Spotlight on Molecular Profiling series:

- Ikediobi and colleagues (26) comprehensively resequenced the exons and exon-intron splice junctions of a range of cancer-related genes in the NCI-60 to identify single nucleotide polymorphisms and sporadic somatic mutations. The article focuses on the resequencing itself and on primary analysis of the results. In accord with the philosophy of the Spotlight on Molecular Profiling series, this data-centric publication will make the data publicly available even as downstream pharmacological hypotheses generated by them are being pursued.
- Lorenzi and colleagues (27) have spearheaded studies that suggest a new, broader indication for an old anticancer agent, the enzyme-drug L-asparaginase, which has been used since the 1970s to treat acute lymphoblastic leukemia. There was previously a known relationship between L-asparaginase activity and the enzyme asparagine synthetase. On the basis of our data from four different transcript expression platforms and a comparative genomic hybridization

platform, we developed a rationale for the possible use of L-asparaginase against ovarian cancers. As described in the article, that rationale is making its way from the NCI-60 cell lines to clinical materials and trials.

Biology, as exemplified by the 18th century taxonomics of Linnaeus, was once a primarily observational science. In the 19th century, the most influential insight in the history of science had its origin in observational studies of finch beaks in the Galapagos. In the 20th century, the pendulum (particularly in biomedical research) then swung decisively toward the hypothesis-driven. Now, in the 21st century, it's time for the pendulum to swing back toward the center. Comprehensive understanding of biological systems—and application of that understanding to biomedical problems—will require a synergistic combination of hypothesis-driven and omic, discovery-based research strategies (10).

The pendulum is indeed swinging, but slowly. Large institutions and scientific fields don't change their cultures overnight. Most editors, reviewers, study sections, and site visitors in the academic world are still addicted to the hypothesis-driven paradigm as a standard of judgment. That remains true despite the obvious practical and conceptual importance of the Genome Project and its aftermath. This innovative Spotlight on Molecular Profiling series nudges the pendulum back toward equilibrium. The articles in it will include various proportions of hypothesis-driven and omic research. But always, the emphasis will be on high quality and early availability of the data so that other researchers can search or mine the molecular profiles according to their interests and domain expertise. Most particularly, the promise is that the molecular profiling data highlighted will promote the overall goals of 21st century personalized medicine.

Acknowledgments

Past and present staff of the NCI DTP deserve the research community's gratitude for establishing and conducting the NCI-60 screen over the years. I particularly want to remember the contributions of Michael Boyd and Bruce Chabner, who originated the screen, and the late Kenneth Paull, who pioneered informatic analysis of the screen data. I also want to thank our many collaborators and other scientists around the world for the molecular profile databases they've generated on the NCI-60.

References

1. Boyd MR, Paull KD. Some practical considerations and applications of the National Cancer Institute *in vitro* anticancer drug discovery screen. Drug Dev Res 1995;34:91 – 109.

2. Holbeck SL. Update on NCI *in vitro* drug screen utilities. Eur J Cancer 2004;40:785 – 93.

3. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer 2006;6:813 – 23.

4. Weinstein JN. Integromic analysis of the NCI-60 cancer cell lines. Breast Dis 2004;19:11 – 22.

5. Paull KD, Shoemaker RH, Hodes L, 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 1989;81:1088 – 92.

6. Weinstein JN, Kohn KW, Grever MR, et al. Neural computing in cancer drug development: predicting mechanism of action. Science 1992;258: 447 – 51.

7. van Osdol WW, Myers TG, Paull KD, Kohn KW, Weinstein JN. Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents. J Natl Cancer Inst 1994;86:1853 – 9.

8. Rabow AA, Shoemaker RH, Sausville EA, Covell DG. Mining the National Cancer Institute's tumor-screening database: identification of compounds with similar cellular activities. J Med Chem 2002;45:818 – 40.

9. Weinstein JN, Pommier Y. Transcriptomic analysis of the NCI-60 cancer cell lines. C R Biol 2003;326:909 – 20.

10. Weinstein JN. Fishing expeditions. Science 1998;282:627 – 8.

11. Weinstein JN. 'Omic' and hypothesis-driven research in the molecular pharmacology of cancer. Curr Opin Pharmacol 2002;2:361 – 5.

12. Alvarez M, Paull KD, Hose C, 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 1995;95:2205 – 14.

13. Izquierdo MA, Shoemaker RH, Flens MJ, Scheffer GL, Wu L, Prather TR. Overlapping phenotypes of multidrug resistance among panels of human cancer-cell lines. Int J Cancer 1996;65:230 – 7.

14. Lee J-S, Paull KD, Alvarez M, et al. Rhodamine efflux patterns predict P-glycoprotein substrates in the National Cancer Institute drug screen. Mol Pharmacol 1994;46:627 – 38.

15. Weinstein JN, Myers TG, O'Connor PM, et al. An information-intensive approach to the molecular pharmacology of cancer. Science 1997;275:343 – 9.

16. Myers TG, Anderson NL, Waltham M, et al. A protein expression database for the molecular pharmacology of cancer. Electrophoresis 1997; 18:647 – 53.

17. Ross DT, Scherf U, Eisen MB, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 2000;24: 227 – 35.

18. Scherf U, Ross DT, Waltham M, et al. A gene expression database for the molecular pharmacology of cancer. Nat Genet 2000;24:236 – 44.

19. O'Connor PM, Jackman J, Bae I, et al. Characterization of the p53 tumor suppressor pathway in cell lines of the National Cancer Institute anticancer drug screen and correlations with the growth-inhibitory potency of 123 anticancer agents. Cancer Res 1997;57:4285 – 300.

20. Nishizuka S, Chen S-T, Gwadry FG, et al. Diagnostic markers that distinguish colon and ovarian adenocarcinomas: identification by genomic, proteomic, and tissue array profiling. Cancer Res 2003;65:5243 – 50.

21. Ludwig JA, Szakacs G, Martin SE, et al. Selective toxicity of NSC73306 in MDR1-positive cells as a new strategy to circumvent multidrug resistance in cancer. Cancer Res 2006;66:4808 – 15.

22. Szakacs G, Annereau JP, Lababidi S, et al. Predicting drug sensitivity and resistance: profiling ABC transporter genes in cancer cells. Cancer Cell 2004;6:129 – 37.

23. Reinhold WC, Kouros-Mehr H, Kohn KW, et al. Apoptotic susceptibility of cancer cells selected for camptothecin resistance: gene expression profiling, functional analysis, and molecular interaction mapping. Cancer Res 2003;63:1000 – 11.

24. Kohn KW, Aladjem MI, Weinstein JN, Pommier Y. Molecular interaction maps of bioregulatory networks: a general rubric for systems biology. Mol Biol Cell 2006;17:1 – 13.

25. Fojo T, Farrell N, Ortuzar W, Tanimura H, Weinstein J, Myers TG. Identification of non-cross-resistant platinum compounds with novel cytotoxicity profiles using the NCI anticancer drug screen and clustered image map visualizations. Crit Rev Oncol Hematol 2005;53:25 – 34.

26. Ikediobi ON, Edkins S, Stevens C, et al. DNA sequence analysis of 32 known cancer genes in the NCI-60 cell lines. Mol Cancer Ther, this issue.

27. Lorenzi PL, Reinhold WC, Rudelius M, et al. Asparagine synthetase as a causal, predictive biomarker for L-asparaginase activity in ovarian cancer cells. Mol Cancer Ther, this issue.