Published OnlineFirst March 5, 2007; DOI: 10.1158/1535-7163.MCT-06-0650



Spotlight on Molecular Profiling

# Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study

**Uma T. Shankavaram,[1] William C. Reinhold,[1] Satoshi Nishizuka,[1] Sylvia Major,[1] Daisaku Morita,[1] Krishna K. Chary,[2] Mark A. Reimers,[1] Uwe Scherf,[1] Ari Kahn,[1] Douglas Dolginow,[3] Jeffrey Cossman,[3] Eric P. Kaldjian,[3] Dominic A. Scudiero,[4] Emanuel Petricoin,[5] Lance Liotta,[5] Jae K. Lee,[1] and John N. Weinstein[1]**

[1]Genomics and Bioinformatics Group, Laboratory of Molecular Pharmacology, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, Maryland; [2]Office of Information Technology, Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland; [3]Gene Logic, Inc., Gaithersburg, Maryland; [4]Science Applications International Corp.-National Cancer Institute-Frederick Cancer Research and Development Center, Frederick, Maryland; and [5]George Mason University, Manassas, Virginia

## Abstract

**To evaluate the utility of transcript profiling for prediction of protein expression levels, we compared profiles across the NCI-60 cancer cell panel, which represents nine tissues of origin. For that analysis, we present here two new NCI-60 transcript profile data sets (A based on Affymetrix HG-U95 and HG-U133A chips; Affymetrix, Santa Clara, CA) and one new protein profile data set (based on reverse-phase protein lysate arrays). The data sets are available online at http://discover.nci.nih.gov in the CellMiner program package. Using the new transcript data in combination with our previously published cDNA array and Affymetrix HU6800 data sets, we first developed a ''consensus set'' of transcript profiles based on the four different microarray platforms. Using that set, we found that 65% of the genes showed statistically significant transcript-protein correlation, and the correlations were generally higher than those reported previously for panels of mammalian cells. Using the predictive analysis of microarray nearest shrunken centroid algorithm for functional prediction of tissue of origin, we then found that (*a*) the consensus mRNA set did better than did data from any of the individual mRNA platforms and (*b*) the protein data seemed to do somewhat better ($P = 0.027$) on a gene-for-gene basis in this particular study than did the consensus mRNA data, but both did well. Analysis based on the Gene Ontology showed protein levels of structure-related genes to be well predicted by mRNA levels (mean $r = 0.71$). Because the transcript-based technologies are more mature and are currently able to assess larger numbers of genes at one time, they continue to be useful, even when the ultimate aim is information about proteins. [Mol Cancer Ther 2007;6(3):820–32]**

## Introduction

Microarrays and other high-throughput technologies have proved useful for transcript expression profiling in the characterization of biological processes, disease states, developmental stages, responses to drugs, and responses to genetic perturbations (1). However, most biological functions are executed by proteins rather than by mRNAs, and there continues to be a question as to how well transcript levels predict the corresponding protein levels. That question is an important one, given the number of studies being done at the RNA level to identify molecular target proteins and biomarker proteins for ''personalization'' of medicine. Lack of transcript-protein concordance across samples can arise for any of several reasons, including differences in regulatory controls, in rates of translation per transcript, in posttranslational modification, and/or in protein degradation. Because the technologies for transcript profiling are much more advanced than those for protein profiling, we have tended to look under the mRNA lamppost, even when our real interest is in the proteins.

Most of the current methods for proteomic profiling are best for quantitative analysis of many proteins in one or a pair of samples, rather than single proteins across samples. Those technologies include two-dimensional PAGE, antibody microarrays, isotope-coded affinity tag labeling, and other

Received 10/10/06; revised 12/21/06; accepted 2/1/07.

**Grant support:** Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. Cells for the various profiling studies were supplied under material transfer agreement by the NCI Developmental Therapeutics Program.

**Note:** Current address for S. Nishizuka: Molecular Therapeutics Program, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD; current address for D. Morita: Department of Pathology II, National Defense Medical College, Saitama, Japan; current address for U. Scherf: Division of Microbiology Devices, Center for Devices and Radiological Health, Food and Drug Administration, Rockville, MD; current address for J. Cossman: The Critical Path Institute, Rockville, MD and Tucson, AZ; current address for J.K. Lee: Department of Public Health Sciences, University of Virginia, Charlottesville, VA.

**Requests for reprints:** John N. Weinstein, Genomics and Bioinformatics Group, Laboratory of Molecular Pharmacology, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD 20892. Phone: 301-496-9571. E-mail: weinstein@dtpax2.ncifcrf.gov



doi:10.1158/1535-7163.MCT-06-0650

Published OnlineFirst March 5, 2007; DOI: 10.1158/1535-7163.MCT-06-0650



mass spectrometry–based techniques (2). Reverse-phase protein lysate arrays (RPLA), to the contrary, perform the more interesting function of comparing protein levels across many samples (3, 4). Hence, we wanted to use that technology to compare mRNA and protein expression patterns across a panel of cell types from different tissues of origin. For that purpose, we have studied the 60 human cancer cell lines (the NCI-60) used by the Developmental Therapeutics Program of the National Cancer Institute to screen more than 100,000 chemical compounds since 1990 (5–9).

The NCI-60 set includes leukemias, melanomas, and cancers of ovarian, renal, breast, prostate, colon, lung, and central nervous system (CNS) origin. Because the screening data proved rich in information on the mechanisms of action of tested compounds (6, 10–14), our laboratory and others have profiled the NCI-60 extensively at the DNA, RNA, protein, chromosomal, and functional levels for correlation with pharmacologic sensitivity of the cells (10, 15–19). Overall, the NCI-60 lines have been characterized more extensively than any other set of cells in existence, and the molecular databases on them are publicly available resources used for research and analysis by thousands of laboratories.

We and our collaborators have previously reported transcript profiling of the NCI-60 cell lines using 9,706-clone spotted cDNA arrays (15, 16, 20) and Affymetrix 6,800-feature set arrays (HU6800; refs. 20–23), and real-time reverse transcription-PCR (24). Here, we report additional profiling studies using the Affymetrix HG-U95A chip, part of a five-array set (~60,000 feature sets), and the Affymetrix HG-U133A chip, part of two-array set (~30,000 feature sets). We also report new RPLA data for 94 proteins, detected using a total of 162 monoclonal antibodies.

All of those databases (along with metadata files and external links) are freely available in query table, interoperable form at our website[6] in the CellMiner resource. In addition, although not used in calculations for this article, data from the HG-U95 (B–E) and HG-U133 (B) chips are included in CellMiner. Insofar as possible, aliquots from the same cell cultures were used for all four transcript platforms and for the proteomic arrays. When that was not possible, culture conditions (e.g., FCS lots), methods of harvest, and methods of purification were kept as constant as possible. For RNA, the time was kept to <1 min from incubator to stabilization of the preparation. Hence, it has been possible to cross-correlate results from all five platforms with a minimum of confounding factors. Here, we (*a*) present the new transcript and protein databases, (*b*) derive a consensus transcript profile data set based on integration of results from the four transcript expression data types, (*c*) test the ability of that consensus set to predict the corresponding protein expression profiles, (*d*) assess the performance of protein and RNA data in a functional prediction, in this case, prediction of tissue of origin, and (*e*) use a formal statistical analysis based on Gene Ontology (GO) categories to show that protein-transcript correlations are particularly high (mean, 0.71) for structural proteins.

## Materials and Methods

### Transcript Expression Profiling

For the present analysis, we used mRNA expression databases on the NCI-60 from four different microarray platforms (see Table 1), two of them previously published and the other two presented here for the first time. The first two were based on 9,706-clone spotted cDNA array (15, 16) and ~6,800-feature 25-mer Affymetrix oligonucleotide arrays (20, 21). The unpublished data were from the A-chip of the five-array (~60,000-feature set) HG-U95 Affymetrix set and the A-chip of the two-array (~30,000-feature set) HG-U133 Affymetrix set. All four data sets are available at our Web site,[6] in what we term the CellMiner database (U.T. Shankavaran, in preparation), a queryable relational system conducive to ''integromic'' analysis (25, 26). Also at the site is a set of tools (the ''Miner Suite'') that can be used for several integrative functions. The new data are also being deposited in the Gene Expression Omnibus and Developmental Therapeutics Program repositories.

The protocol for cell harvests in the two new studies (using Affymetrix HG-U95 and HG-U133 chips) was as follows. Seed cultures of the 60 cell lines were drawn from aliquoted stocks used for ongoing assays in the NCI-60 screen. The cells were then passaged once in T-162 flasks and monitored frequently for degree of confluence. The medium (30 mL for attached cells; 40 mL for leukemias) was RPMI 1640 with phenol red, 2 mmol/L glutamine, and 5% FCS. For compatibility with our other profiling studies, all FCS was obtained from the same large batches as were used by the Developmental Therapeutics Program. No antibiotics were included in the medium. One day before harvest, the cells were re-fed with the original amount and composition of medium. Attached cells were harvested at ~80% confluence, as assessed for each flask by phase microscopy. Suspended cells (leukemias) were harvested at $\sim 0.5 \times 10^6$ cells/mL. In pilot studies, samples of medium showed no appreciable change in pH between refeeding and harvest, and no color change in the medium was seen in any of the flasks harvested. The time from incubator to stabilization of the preparation was kept to <1 min. Total RNA was purified using the Qiagen (Valencia, CA) RNeasy Midi kit according to the manufacturer's instructions. The RNA was then quantitated spectrophotometrically and aliquoted for storage at −80°C.

### Probe Labeling

RNA quality was evaluated using an Agilent 2100 BioAnalyzer (Agilent Technologies, Palo Alto, CA). Five to 40 μg total RNA was used for cDNA synthesis. Briefly, RNA was reverse transcribed with a T7-(dT)24 oligo primer with SuperScript II (Invitrogen Corp., Carlsbad, CA) followed by second strand synthesis with *Escherichia coli* DNA Polymerase I as recommended by Affymetrix. Double-stranded cDNA was purified using phenol/chloroform and Phase Lock gel. Following ethanol precipitation

[6] http://discover.nci.nih.gov

Published OnlineFirst March 5, 2007; DOI: 10.1158/1535-7163.MCT-06-0650



**Table 1. Overview of the transcript and protein data sets**

| Data set | Description | No. feature sets or clones (genes) | Reference |
|---|---|---|---|
| mRNA expression | | | |
| cDNA microarray | Pin-spotted cDNA array | 9,706 Clones (7,190 genes) | (15, 16) |
| Affymetrix HU6800 | *In situ* synthesized oligonucleotide array | 6,800 Feature sets (5,562 genes) | (20, 21) |
| Affymetrix HG-U95A | *In situ* synthesized oligonucleotide array (5 chips) | ~12,000 Feature sets (8,978 genes) | This publication |
| Affymetrix HG-U133A | *In situ* synthesized oligonucleotide array (2 chips) | ~22,000 Feature sets (13,032 genes) | This publication |
| mRNA consensus set | Consensus of the four individual platforms (see Materials and Methods) | See Materials and Methods (75 genes) | This publication |
| Protein expression | | | |
| RPLA | 640 Cell lysate spots in 10 dilutions (2-fold) for each of the 60 cell lines plus controls consisting of a pool of all 60 | 162 Antibodies (94 proteins) | This publication for much of the data(4) |

and resuspension, the cDNA was used as a template for *in vitro* transcription using biotinylated CTP, UTP, and unlabeled nucleotide triphosphates. The labeled cRNA was collected and purified using an RNeasy column. The quantity and purity of the cRNA were determined, respectively, by measuring absorbance at 260 nm and the 260:280 nm absorbance ratio. Quality of the cRNA was evaluated by assessing size distribution on a 1% agarose gel, and cRNA of good quality was fragmented according to Affymetrix recommendations. Quality of the fragmented cRNA was evaluated on an Agilent 2100 BioAnalyzer.

### Affymetrix Chip Hybridization and Scanning

Fragmented cRNA was combined with hybridization mix to yield a concentration of 10 μg of biotin-labeled, fragmented cRNA per 200 μL hybridization mix, applied to HG-U95A or HG-U133A microarrays, and hybridized overnight at 45°C. Arrays were washed and stained according to Affymetrix recommendations and scanned on an Affymetrix GS2500 scanner. The MAS5 algorithm was used to generate signal intensities for HG-U95A and HG-U133A arrays. Expression values were normalized to a mean target level of 100.

### Protein Expression Profiling

Protein levels in the cells were assessed using high-density RPLA (3, 4), essentially as described in ref. 4. Protein expression data are available online.[6] Briefly, cells at ~80% confluence were harvested into lysis buffer containing 9 mol/L urea (Sigma, St. Louis, MO), 4% CHAPS (Calbiochem, La Jolla, CA), 2% Pharmalyte (pH 8.0–10.5; Amersham Pharmacia Biotech, Piscataway, NJ), and 65 mmol/L DTT (Amersham Pharmacia Biotech). Lysate from each cell line was robotically spotted onto nitrocellulose-coated glass slides in 10 serial 2-fold dilutions. The first dilution (4-fold) was made with buffer containing 5 mol/L urea, 2% Pharmalyte (pH 8–10.5), and 65 mmol/L DTT. The remaining dilutions were then made with buffer containing 6 mol/L urea, 1% CHAPS, 2% Pharmalyte (pH 8–10.5), and 65 mmol/L DTT. Hence, only the lysate concentration changed along each dilution series. The urea concentration was kept at 6 mol/L, and the CHAPS concentration at 2%, to keep proteins in their denatured forms. Because DTT is nonvolatile, reducing conditions were maintained.

Overall, each array consisted of 640 spots corresponding to the 60 cell lines plus four replicates of a pool consisting of equal amounts of each of the 60 lines. Protein was quantitated on the arrays using the Catalyzed Signal Amplification System (DAKO, Carpinteria, CA). Each slide was washed manually with deionized water to remove urea. Then, in an Autostainer Universal Staining System (DAKO), it was blocked with I-block (Tropix, Bedford, MA) and incubated with primary and secondary antibodies. Also in the Autostainer, it was incubated with streptavidin-biotin complex, biotinyl tyramide (for amplification) for 15 min, streptavidin-peroxidase for 15 min, and 3,3′-diaminobenzidine tetrahydrochloride chromagen for 5 min. Between steps, each slide was washed with Catalyzed Signal Amplification buffer. The signal was scanned using a Perfection 1200S Scanner (Epson America, Long Beach, CA) with 256-shade gray scale at 600 dots per inch. Spot images were converted to raw pixel values by a modified version of the P-SCAN (Peak quantification with Statistical Comparative Analysis) software.[7] Before use on the arrays, each monoclonal antibody was tested for specificity by Western blot against the NCI-60 lysate pool. Only antibodies that gave a single predominant band at the expected molecular weight were used for array measurements. Total protein on the array was quantitated with SYPRO

[7] http://abs.cit.nih.gov/index.html

Published OnlineFirst March 5, 2007; DOI: 10.1158/1535-7163.MCT-06-0650



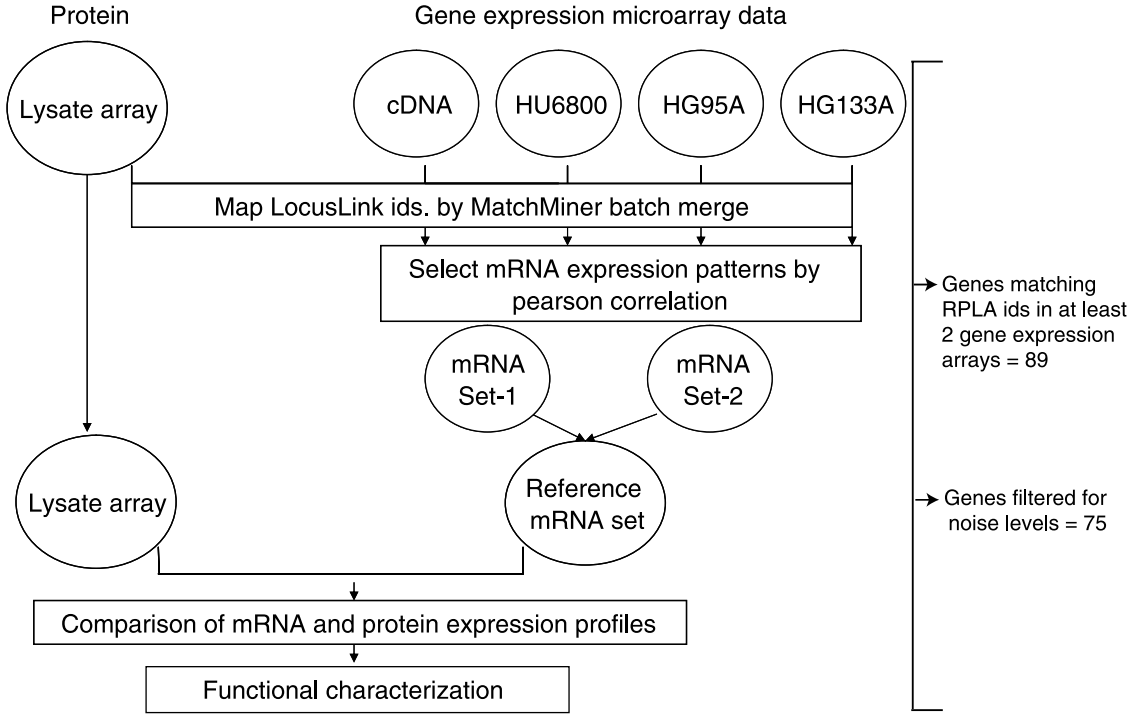


**Figure 1.** Schematic flow diagram of the methodology used for comparative analysis of mRNA and protein expression profiles.

Ruby Protein Stain (Molecular Probes, Eugene, OR). Levels of specific protein were calculated from the arrays, with an adjustment for total protein, using the 25% ''dose-interpolation'' (DI25) algorithm described previously (4). The overall coefficient of variation for protein expression measurement was 17%.

### "Meta-profiling" and Statistical Analysis

We used Pearson correlation to construct a consensus mRNA expression data set based on levels measured using the four different microarray types. As shown in Fig. 1 (see Supplementary Fig. S1 for more details),[8] we first found Entrez Gene IDs for the mRNA and protein expression arrays and then matched IDs among the platforms to find those that were common to the protein arrays and at least two of the mRNA platforms. Our method for selecting the best mRNA measures from the four microarray types was as follows. We separated the probe sets into three groups depending on whether they were found on all four arrays (*a*), any three arrays (*b*), or any two arrays (*c*). For condition *a*, we computed all six (i.e., $4 \times 3 / 2$) pairwise Pearson correlation coefficients for the four expression patterns across the 60 cell lines for each row; for condition *b*, we calculated all three pairwise Pearson correlation coefficients; and for condition *c*, we computed the one Pearson correlation coefficient. For each group, we then chose the two patterns per gene that were maximally correlated between array types. The average of the final pair of expression patterns formed the consensus set and was used to compute pairwise Pearson correlation coefficients with protein expression.

The decision to choose the best-correlated two platforms for the consensus set was heuristic; it was not based on any particular statistical theory or model. Empirically, it did improve predictiveness, but we did not compare its results with any other rule for choosing a consensus set of values. Hence, there was no need for a multiple comparisons correction when we compared performance of the consensus and individual transcript expression databases. SDs as measures of gene expression pattern across the 60 cell lines were calculated, and vectors across the 60 were filtered out if they did not meet the criteria SD >0.3 and Pearson correlation between any two of the four mRNA data sets >0.3. The significance of the correlations between consensus and protein expression data was assessed by bootstrap confidence limits with 1,000 iterations and without bias correction (because bias-corrected bootstrapped confidence limits for correlation coefficients have been reported to do worse than do uncorrected ones; ref. 27).

One further technical point deserves mention: if any platform included multiple probe sets or RPLA detection antibodies for a given gene, we chose the values that yielded maximal correlation between mRNA and protein. That procedure resulted in a mean RNA-protein Pearson correlation coefficient of +0.42 for the 89 genes. We chose the highest correlation for development of an optimal set because a value highly correlated with protein is less likely

[8] Supplementary material for this article are available at Molecular Cancer Therapeutics Online (http://mct.aacrjournals.org/).

Published OnlineFirst March 5, 2007; DOI: 10.1158/1535-7163.MCT-06-0650



to be spurious. Statistically, there are more ways that a spurious value can lead to an artifactually low correlation than ways that it can lead to an artifactually high one. However, that choice was, in principle, biasing toward higher RNA-protein correlation, so we did the calculation in two other, nonbiased ways. First, we calculated the mean Pearson correlation coefficient over all possible pairs of mRNA and protein values. For example, when there were two probe sets and three antibodies for a given gene, we averaged the six possible correlation coefficients. For the consensus set, that algorithm yielded a mean correlation coefficient of +0.40. Second, instead of the maximum or the mean, we used a randomly selected probe set and a randomly selected antibody for the given gene. For the consensus set, that algorithm yielded a mean correlation of +0.41. Statistically, there was no significant difference among the three results (+0.42, +0.40, and +0.41), so we used the maximal correlation values for further analyses. Supplementary Fig. S2 shows the very similar distributions of RNA-protein correlation coefficients calculated in the three different ways.[8] In Results, we describe division of the genes into groups 1 and 2 based on their RNA-protein correlation. For group 1, the mean correlation coefficients calculated in the three different ways were +0.71, +0.69, and +0.70; for group 2, +0.28, +0.26, and +0.26. In other words, the algorithm used to deal with multiple probe sets and/or RPLA antibodies per gene made little difference.

As an index of global concordance, we used the ''correlation of correlations'' ($r_c$), a variable we introduced previously for transcript expression analysis (15, 20). Conceptually, $r_c$ for cells can be explained as follows. For one of the mRNA or protein data sets, visualize the 60 cell lines as nodes linked in all possible pairwise combinations by $60 \times (60 - 1) / 2 = 1{,}770$ connections each of which is the Pearson correlation coefficient over transcript or protein levels between the two nodes. Do the same for a second array data set, obtaining another set of 1,770 correlation coefficients for the same set of genes. The correlation of correlations is the Pearson correlation coefficient of those 1,770 pairs of values. Mathematically, $r_c$ was calculated as follows: let $U_{ij}$ denotes the correlation of cells $i$ and $j$ (for $i$ and $j$, from 1 to $n$) based on their expression patterns in the first array data set, and let $V_{ij}$ denotes the correlation of cells $i$ and $j$ based on their expression patterns in the second array data set. For example, if $X_{di}$ denotes the expression level of gene $d$ (for $d$, from 1 to $D$) in cell $i$, and then the Pearson correlation coefficient for cells $i$ and $j$ based on gene expression is given by the expression

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

**Table 2. Summary data on gene and protein identifier matching and construction of the consensus transcript expression set**

| | Category | cDNA chip (Clone ID) | HU6800 (Affymetrix ID) | HG-U95A (Affymetrix ID) | HG-U133A (Affymetrix ID) | RPLA (Antibodies) |
|---|---|---|---|---|---|---|
| 1 | Total no. mRNA or protein features | 9,706 | 6,810 | 12,386 | 22,283 | 162 |
| 2 | No. unique locuslink identifiers | 7,190 | 5,562 | 8,978 | 13,032 | 94 |
| No. features from each microarray in comparison with RPLA | | | | | | |
| 3 | No. microarrays used to compare = all 4 | 62 | 62 | 62 | 62 | 62 (49 genes) |
| 4 | No. microarrays used to compare = any 3 | 18 | 58 | 72 | 74 | 74 (29 genes) |
| 5 | No. microarrays used to compare = any 2 | 11 | 3 | 7 | 15 | 18 (11 genes) |
| 6 | No. microarrays used to compare = any 1 | 1 | — | — | 5 | 6 (3 genes) |
| 7 | No. features on RPLA matched in 0 microarrays | — | — | — | — | 2 (2 genes) |
| 8 | No. features that match if only one gene expression array is chosen (rows 3,4,5, and 6) | 92 | 123 | 141 | 156 | |
| 9 | Total no. features matched in more than two gene expression microarrays (rows 3, 4, and 5) | 91 (63 genes) | 123 (72 genes) | 141 (82 genes) | 151 (88 genes) | 154 (89 genes) |
| 10 | Percentage genes contributed to consensus mRNA set | 8% | 12% | 39% | 41% | |

Published OnlineFirst March 5, 2007; DOI: 10.1158/1535-7163.MCT-06-0650



**Table 3. Global comparisons of HU6800, HG-U95A, HG-U133A, consensus mRNA, and RPLA protein data sets**

| | cDNA | HU6800 | HG-U95A | HG-U133A | Consensus |
|---|---|---|---|---|---|
| (A) $r_c$ for cells | | | | | |
| HU6800 | 0.48 | | | | |
| HG-U95A | 0.54 | 0.81 | | | |
| HG-U133A | 0.54 | 0.82 | 0.82 | | |
| Consensus | 0.54 | 0.70 | 0.70 | 0.83 | |
| Protein | 0.48 | 0.58 | 0.51 | 0.58 | 0.63 |
| (B) $r_c$ for genes | | | | | |
| HU6800 | 0.38 | | | | |
| HG-U95A | 0.35 | 0.39 | 0.61 | | |
| Consensus | 0.52 | 0.55 | 0.7 | 0.86 | |
| Protein | 0.20 | 0.24 | 0.24 | 0.31 | 0.34 |
| (C) $P$ | | | | | |
| HU6800 | 0.30 | | | | |
| HG-U95A | 0.77 | 0.45 | | | |
| HG-U133A | 1.00 | 0.23 | 0.72 | | |
| Consensus | 0.55 | 0.027 | 0.074 | 0.45 | |
| Protein | 0.0055 | 0.00041 | 0.0019 | 0.0098 | 0.027 |

NOTE: A, Pearson correlation of correlation coefficients ($r_c$) for cells (over genes) calculated for all 15 possible pairs of the six data sets. B, Pearson values for genes (over cells) for all 15 possible pairs of the six data sets. C, two-tailed McNemar $P$ values for all 15 possible pairs of data sets in the PAM shrunken centroid functional class prediction. As explained in the text, the prostate lines were omitted because there were only two of them.

and similarly for $V_{ij}$. The Pearson correlation of $U_{ij}$ and $V_{ij}$ then gives a measure of the global similarity in distributions between the two expression data sets. That correlation is given by

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

where the sums are over all distinct pairs of cells $i$ and $j$, there being $n\ (n\ -\ 1)\ /\ 2$ such pairs. The correlation of correlations for genes was calculated identically except that the roles of genes and cells in the calculation were reversed. A nonparametric (Spearman) version can be defined similarly, but that alternative was not used here. A program for calculating $r_c$ can be found at our Web site[6] under ''tools.''

We used the predictive analysis of microarray (PAM) shrunken centroid algorithm and software package implemented in R[9] (28) for class prediction (of tissue of origin). Briefly, the method of nearest shrunken centroids identifies subsets of genes that best characterize each class. Here, we used 10-fold cross validation to compare protein, consensus, and four mRNA array data sets using 49 genes common to all of the data sets. PAM stratifies the data so that each of the 10 partitions contains approximately the same distribution of classes. The statistical significance of misclassification discrepancies among the different data sets was assessed using McNemar's $\chi^2$ test (two-tailed $P$ values) in the R statistical package.

## Results

### Matching mRNA and Protein Expression Data

The number of features and summary statistics for the various data sets used in the study are shown in Tables 1 and 2. In summary, we were able to quantitate the levels of 94 distinct Entrez Gene–identified proteins from the RPLA based on detection by 162 antibodies. The numbers of those same 94 genes identifiable on the transcript expression microarrays included 63 from the cDNA arrays, 72 from the HU6800 arrays, 82 from the HU95A arrays, and 88 from the HU133A arrays.

### A Consensus mRNA Expression Database

To build a consensus mRNA expression data set with minimal noise and error, we compared and combined gene expression profiles from the four different microarray platforms (cDNA arrays and Affymetrix HU6800, HG-U95A, and HG-U133A arrays). For each gene, we first used our MatchMiner program package[6] (29) to translate Image Clone and Affymetrix transcript expression identifiers in the four data sets into Entrez Gene identifiers (29). The latter have the advantage that they assign both gene and protein sequences to a common entity (30, 31).

To select the best consensus mRNA measurements, we next computed all pairwise Pearson correlation coefficients ($r$) as described in Materials and Methods. For each gene, we chose the two such patterns that were maximally correlated with each other (Fig. 1). That procedure yielded 89 gene expression profile pairs. The between-measure correlations for the 89 genes in the resulting data sets ranged from −0.164 to 0.988. In any system with noise, the correlation between two measures is biased downward for variables whose level of noise is comparable with the variation in signal. That statistical artifact has been underappreciated in previous studies of concordance (e.g., ref. 32). Accordingly, we excluded genes that showed low pattern as measured by SD across the 60 cell lines and low transcript-transcript concordance ($r < 0.3$) as in ref. 20. Note that the filtering algorithm above did not involve protein data and therefore was not biasing with respect to our comparison of the databases for assessment of mRNA-protein concordance (see below). That filtering algorithm resulted in a 75-gene set for further analysis. It may, however, have tended to bias the results in favor of mRNA over protein in the functional classification (see below) because we did not similarly filter the protein data. We next computed confidence limits for the 75 mRNA-protein correlations using bootstrap without bias correction. Fifty-eight (77% of the genes) mRNA/protein pairs had 95% confidence limits that excluded zero (see Supplementary Table S1).[8] To ask whether the degree of transcript-protein correlation was a function of signal intensity on the microarrays, we calculated the Pearson correlation between signal amplitude and the Pearson correlation of transcript-protein expression for each of the six data sets used in the analysis (one protein, four mRNA

[9] http://www-stat.stanford.edu/~tibs/PAM/index.html

Published OnlineFirst March 5, 2007; DOI: 10.1158/1535-7163.MCT-06-0650



plus consensus data set). None of the six correlation values calculated were statistically significant and the overall mean of those correlations was $0.003 \pm 0.180$, indicating that there was no significant dependence of the correlation on amplitude (i.e., on protein or transcript abundance once the filtering had removed genes with low expression from calculations).

### Analyzing the Global Concordance of mRNA and Protein Expression

To compare the global concordance of expression among the consensus mRNA, individual-platform mRNA, and RPLA protein data sets, we used the correlation of correlations coefficient, $r_c$, described in Materials and Methods (15, 20). The results are shown in Table 3A and B. The

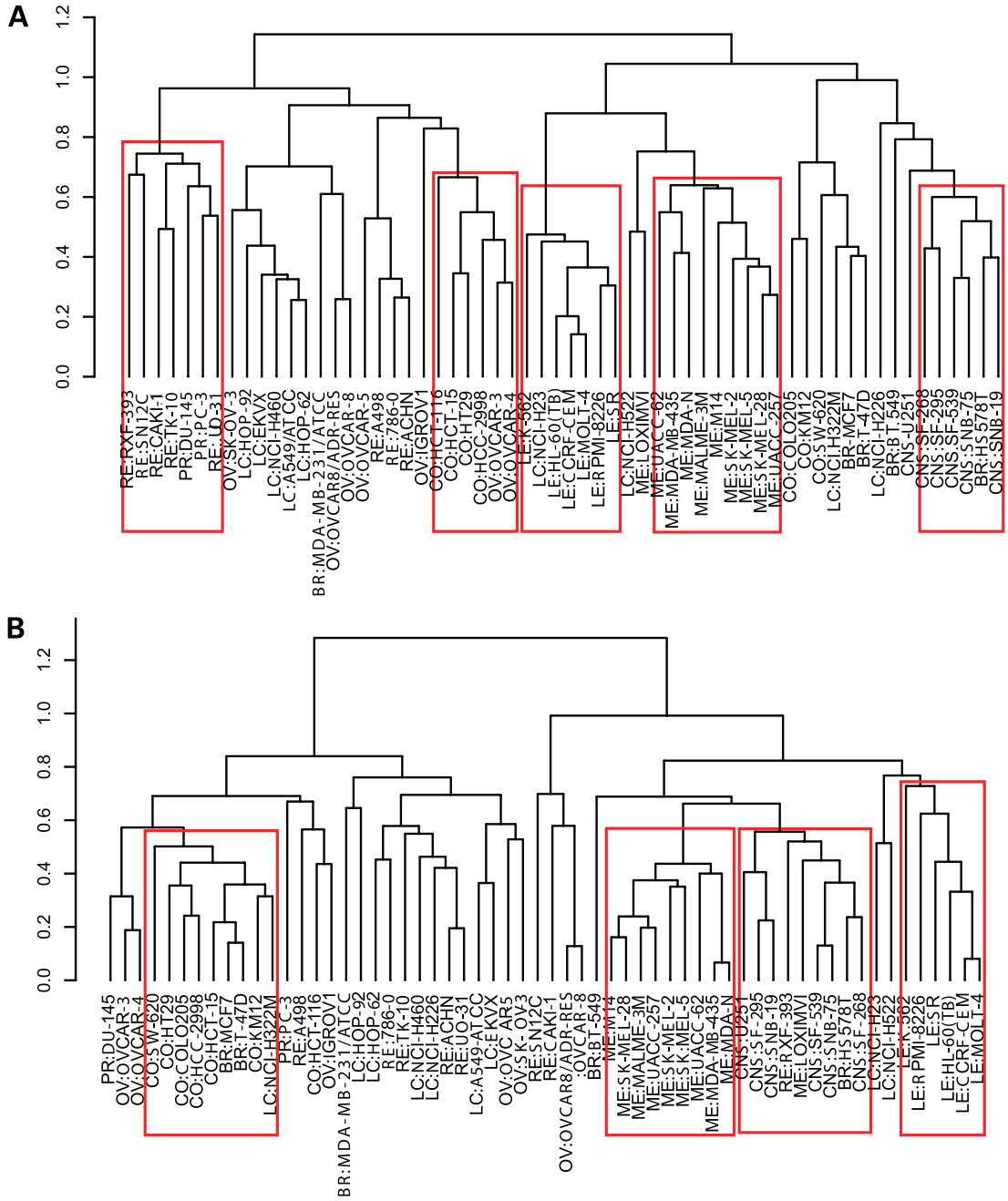


**Figure 2.** Comparison of cluster trees (average linkage algorithm, Pearson correlation metric) for the 60 cell lines using protein expression (**A**) and consensus mRNA expression (**B**) data sets for 75 genes, after quality control filtering as detailed in the text. In each case, there were four to five clusters with major membership of cell lines belonging to the same tissue of origin. The tissue type abbreviations are: BR, breast cancer; CNS, central nervous system; CO, colon; LC, lung cancer; LE, leukemia; ME, melanoma; OV, ovarian; PR, prostate; RE, renal.

Published OnlineFirst March 5, 2007; DOI: 10.1158/1535-7163.MCT-06-0650



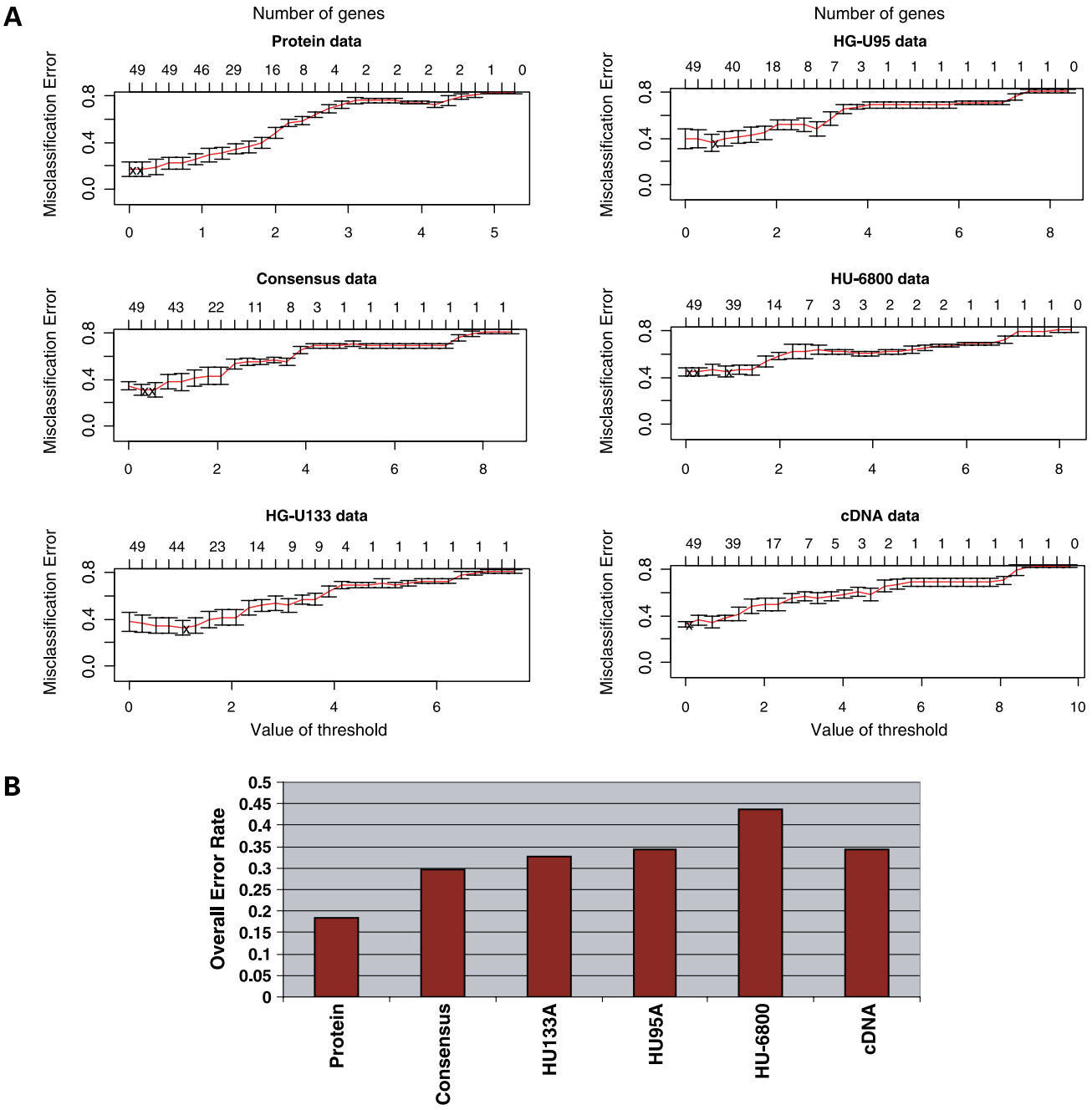


**Figure 3.** Functional prediction of tissue of origin using the mRNA and protein data. **A,** misclassification results using the PAM shrunken centroid algorithm. *X,* the thresholds at which the minimum numbers of misclassifications were predicted. **B,** the overall misclassification error rate of each data set.

consensus mRNA set showed the highest concordance with protein data; the HG-U133A set was second. Next, we wanted to evaluate the ability of profiles from various platforms to cluster cell lines by tissue of origin. As shown in Fig. 2, hierarchical clustering using (1-Pearson correlation coefficient) as a distance metric and the average linkage algorithm indicated that cells belonging to five of the nine tissue types were reasonably well clustered based on protein data. Four or five of the nine were reasonably well clustered based on consensus transcript expression data. Cluster dendrograms based on the individual mRNA expression data sets did not show quite such strong clustering by tissue of origin (data not shown).

### Functional Prediction on the Basis of Protein and mRNA Expression Profiles

For that test, we excluded the prostate lines (because there are only two of them in the NCI-60) and used the PAM algorithm in R (28) to predict tissue of origin for

Published OnlineFirst March 5, 2007; DOI: 10.1158/1535-7163.MCT-06-0650



**Table 4. Confusion matrices from the PAM algorithm for class prediction of NCI-60 tissue of origin**

| | BR | CNS | CO | LC | LE | ME | OV | RE | Class error rate |
|---|---|---|---|---|---|---|---|---|---|
| Protein | | | | | | | | | |
| BR | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 1 | 0.8 |
| CNS | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CO | 0 | 0 | 6 | 0 | 0 | 0 | 1 | 0 | 0.14 |
| LC | 0 | 0 | 1 | 6 | 1 | 0 | 1 | 0 | 0.33 |
| LE | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 |
| ME | 0 | 1 | 0 | 0 | 0 | 9 | 0 | 0 | 0.1 |
| OV | 0 | 1 | 0 | 0 | 0 | 0 | 5 | 1 | 0.29 |
| RE | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 |
| Consensus | | | | | | | | | |
| BR | 0 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 1 |
| CNS | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CO | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 |
| LC | 0 | 1 | 1 | 4 | 0 | 0 | 2 | 1 | 0.56 |
| LE | 0 | 2 | 0 | 0 | 4 | 0 | 0 | 0 | 0.33 |
| ME | 0 | 1 | 0 | 0 | 0 | 9 | 0 | 0 | 0.1 |
| OV | 0 | 0 | 2 | 1 | 0 | 0 | 4 | 0 | 0.43 |
| RE | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 6 | 0.25 |
| HU133A | | | | | | | | | |
| BR | 0 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 1 |
| CNS | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CO | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 |
| LC | 0 | 1 | 1 | 4 | 0 | 0 | 2 | 1 | 0.56 |
| LE | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 1 | 0.17 |
| ME | 0 | 1 | 0 | 0 | 0 | 9 | 0 | 0 | 0.1 |
| OV | 0 | 0 | 1 | 5 | 0 | 0 | 1 | 0 | 0.86 |
| RE | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 6 | 0.25 |
| HU95A | | | | | | | | | |
| BR | 0 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 1 |
| CNS | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CO | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 |
| LC | 0 | 2 | 1 | 3 | 0 | 0 | 0 | 3 | 0.67 |
| LE | 0 | 1 | 0 | 0 | 4 | 1 | 0 | 0 | 0.33 |
| ME | 0 | 1 | 0 | 0 | 0 | 9 | 0 | 0 | 0.1 |
| OV | 0 | 0 | 3 | 0 | 0 | 0 | 2 | 2 | 0.71 |
| RE | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 6 | 0.25 |
| HU6800 | | | | | | | | | |
| BR | 0 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 1 |
| CNS | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 1 | 0.17 |
| CO | 0 | 0 | 6 | 1 | 0 | 0 | 0 | 0 | 0.14 |
| LC | 0 | 2 | 1 | 2 | 0 | 0 | 2 | 2 | 0.78 |
| LE | 0 | 1 | 0 | 0 | 5 | 0 | 0 | 0 | 0.17 |
| ME | 0 | 1 | 0 | 0 | 0 | 9 | 0 | 0 | 0.1 |
| OV | 0 | 0 | 3 | 4 | 0 | 0 | 0 | 0 | 1 |
| RE | 0 | 1 | 2 | 0 | 0 | 0 | 1 | 4 | 0.5 |
| cDNA | | | | | | | | | |
| BR | 0 | 1 | 2 | 1 | 0 | 0 | 0 | 1 | 1 |
| CNS | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 1 | 0.17 |
| CO | 0 | 0 | 6 | 0 | 0 | 0 | 1 | 0 | 0.14 |
| LC | 0 | 2 | 0 | 3 | 1 | 0 | 2 | 1 | 0.67 |
| LE | 1 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0.17 |
| ME | 0 | 1 | 0 | 0 | 0 | 9 | 0 | 0 | 0.1 |
| OV | 0 | 0 | 1 | 3 | 0 | 0 | 3 | 0 | 0.57 |
| RE | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 6 | 0.25 |

NOTE: Rows indicate ''true'' classifications; columns indicate ''predicted'' classifications.

Abbreviations: BR, breast cancer; CO, colon; LC, lung cancer; LE, leukemia; ME, melanoma; OV, ovarian; RE, renal.

cells in the other eight classes. PAM, which uses a nearest shrunken centroid algorithm, was applied to each mRNA and protein data set to obtain the number of classification errors after 10-fold cross-validation. The data sets were directly comparable in that each one contained the same set of genes and same cell lines. Figure 3 and Table 4 show the results. As shown graphically in Fig. 3B, the overall misclassification error rate was lowest for the protein data (0.184) followed by the consensus (0.297), HG-U133A (0.328), HG-U95A (0.344), cDNA (0.344), and lastly HU6800 (0.437) data sets. That is, the HG-U133A and HG-U95A platforms (labeling and hybridizations done at Gene Logic, Inc., Gaithersburg, MD) were the two best transcript level performers and, not surprisingly, contributed the most expression levels (41% and 39%, respectively) to the consensus set (Table 2). The threshold value was tuned on the test set, but that was true for all of the data sets, so it provided no obvious bias toward any of them. To compare the predictiveness of each pair of data sets statistically, we used the McNemar test, which focuses on discrepancies in class prediction. For that test, we scored the tissue-of-origin predictions as either correct or incorrect. Table 3C shows two-tailed McNemar $P$ values calculated for each pair of data sets under the null hypothesis that there was no difference in number of incorrect predictions. The protein database seemed to do somewhat better than any of the individual RNA databases ($P = 0.00041–0.0098$) or the consensus set ($P = 0.027$), but all did reasonably well. Possible confounding factors are described in Discussion.

### GO Structural Genes Show High mRNA-Protein Correlations

The distribution of correlations between protein and consensus mRNA data sets for the filtered data on 75 genes is bimodal (Fig. 4C), indicating two distinct populations of genes: group 1 (mean $r = 0.71$) and group 2 (mean $r = 0.28$). To identify functional associations with the two groups, we mapped the 75 genes to GO functions using the National Center for Biotechnology Information Entrez Gene database annotations. A clustered image map of the correlations is shown in Fig. 4A. The X-axis and Y-axis reflect protein and consensus RNA data, respectively. Both axes were clustered based on the protein data (i.e., to bring together genes of similar protein expression pattern). Note that the matrix would have been symmetrical around the diagonal if the protein and mRNA profiles had been exactly concordant. The clustered image map shows distinct clusters of genes (e.g., the two highlighted clusters). The two are enriched with group 1 (red) and group 2 (blue) genes, respectively. Genes in the GO structural category were most significantly associated with high correlations (mean $r = 0.71 \pm 0.08$ SD), whereas genes in the GO signal transduction category were significantly associated with lower correlations (mean $r = 0.39 \pm 0.15$; Fig. 4B). To put those results in more detailed terms, the unpaired two-tailed $t$ test $P$ value for the null hypothesis that GO structural genes (in red) show the same mean correlation coefficient (versus consensus mRNA data) as all of the other

Published OnlineFirst March 5, 2007; DOI: 10.1158/1535-7163.MCT-06-0650



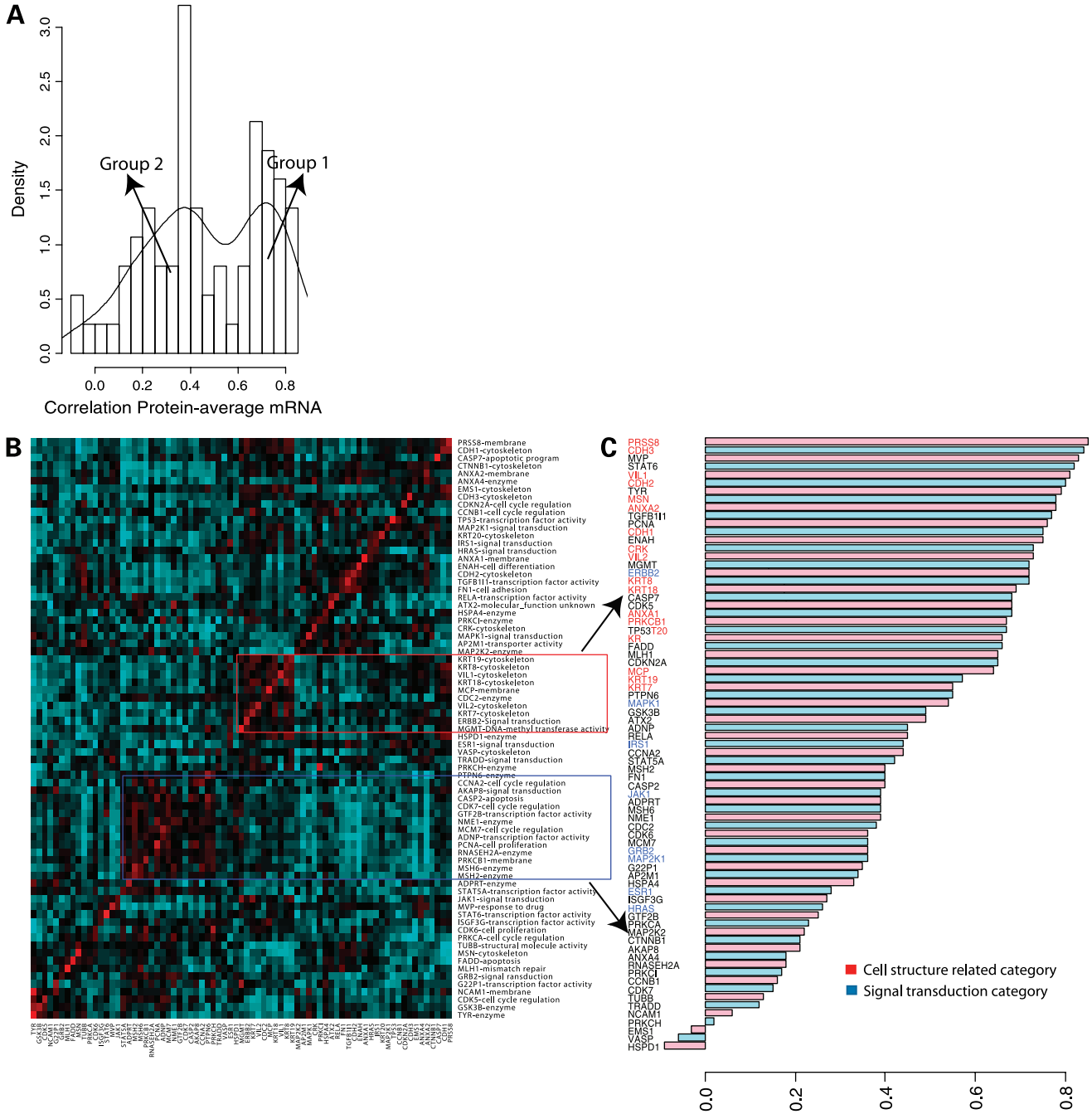


**Figure 4.** Distribution of Pearson correlation coefficient measurements between consensus mRNA and protein data sets. **A,** clustered image map (heat map; ref 10) relating the protein and consensus mRNA data (1-correlation metric, average linkage algorithm). *Red,* high correlation; *blue,* low correlation. Data were ordered on both axes according to clustering of the protein data. *Highlighted boxes,* gene clusters enriched with particular GO functional categories (*red box*, structural; *blue box*, signal transduction). **B,** distribution of protein-consensus mRNA correlation coefficients. The red labels representing structure-related genes are clearly more highly correlated than are the blue labels indicating signal transduction genes. **C,** histogram showing a bimodal distribution of correlation coefficients.

genes in Fig. 4B was $9.42 \times 10^{-13}$; the corresponding nonparametric *P* value (two-tailed Wilcoxon rank-sum) was $1.89 \times 10^{-06}$. The analogous values for the GO signal transduction genes (blue) were 0.043 and 0.070, respectively. Those relationships are apparent visually in Fig. 4B: the protein-mRNA correlations are strikingly higher than average for the structural genes and marginally lower than average for the signal transduction genes. Despite the

Published OnlineFirst March 5, 2007; DOI: 10.1158/1535-7163.MCT-06-0650



marginal statistical significance (even without multiple comparisons correction), we show calculations here for the signal transduction category because (*a*) it is an important category that includes a considerable number of the genes assayed and (*b*) a previous study (26) reported the signal transduction category as showing among the highest protein-mRNA correlations, contrary to our findings reported here.

## Discussion

Because transcript profiling is technologically easier than protein profiling, it developed more rapidly. Hence, one commonly considers transcript profiles, explicitly or implicitly, as surrogates for the protein profiles. In this study, we have assessed the similarities and differences between mRNA and protein expression patterns across the NCI-60 cancer cell panel. We have not attempted to ask, however, whether complex gene signatures generated by transcript profiling can be used clinically at the protein level or by immunohistochemistry. That might be possible in some cases, but it would take a different experimental design to address that question directly. We profiled the cells in the baseline, untreated state of the cells without cell cycle synchronization, thereby avoiding the transient phenomena that complicate most forms of perturbation experiments (e.g., those with drugs). For interpretation of mRNA-protein relationships in drug perturbation experiments, it would be necessary to model the kinetics of translation and posttranslational processes, as well as the distribution of lifetimes of the mRNA and protein species. In the steady state (per cell) represented by log-phase growth, to the contrary, those kinetic factors fall out of the calculations, as we showed mathematically in a previous publication (33).

There have been several other studies, in which both gene and protein profiling technologies have been used to compare the same samples (2, 4, 34, 35). Comparison of transcript and protein expression for 19 liver-related molecules (34) showed a moderate correlation of 0.48. A study of lung adenocarcinoma samples (36) showed a range of correlation coefficients from −0.467 to 0.442 in a comparison between mRNA expression levels and protein levels obtained from 165 protein spots on two-dimensional gels. In a more elaborate study of mammalian cells, both steady-state and drug response levels of mRNA and proteins showed a correlation of 0.59 (37). In those reports, protein abundance was assessed using two-dimensional gels and/or mass spectrometry. Those technologies are best suited to measuring multiple protein levels in a single sample or a pair of samples. But for many functional studies, we want to measure protein expression in multiple samples for one molecular species at a time. Hence, for the present study, we used RPLA, which does that on a single glass slide.

Using two high-throughput microarray platforms (cDNA and HU6800 Affymetrix arrays), we previously compared mRNA and protein profiles in the NCI-60 for a limited set of 19 genes (4). The results led anecdotally to the hypothesis that protein-mRNA correlations are higher for structural proteins. In the present study, we generated experimentally and then analyzed a greatly expanded RPLA data set based on 154 antibodies corresponding to 89 genes. We also generated an additional two mRNA expression databases (Affymetrix HG-U95 and HG-U133 arrays) and present those data for the first time. The new data sets substantially improved the functional predictiveness of the mRNA profiles and, hence, have been useful for the mRNA-protein comparison. From the four mRNA expression databases, we also developed a consensus mRNA expression data set, and it did better than did any of the individual ones in the functional prediction of tissue of origin (Fig. 3; Table 3). Basing statistical analyses on multiple mRNA expression platforms has the additional advantage that it may lessen somewhat the effects of different splice variants, clone contaminations, oligonucleotide misidentifications, and other sources of technical error. In our comparison of 104 antibodies corresponding to 75 genes (after gene filtering), 77% of them showed statistically significant correlations between consensus mRNA and protein profiles. The correlations for all 89 gene pairs ranged from −0.09 to 0.848. Overall, the correlations were considerably higher than those obtained previously across a panel of mammalian cells (−0.467–0.442; ref. 36). Three possible reasons are that RPLA is a more appropriate technology for cross-sample functional analysis, that our gene selection criteria provided more reliable data from the mRNA microarray experiments, and that the comparison across tissues of origin simply reflects a different type of relationship.

Returning to the important question, ''Which type of data are better at predicting function,'' what we needed was a test case for which we knew the answer. Such test cases (beyond the level of individual genes or gene families) have not been easy to find. However, one good choice based on the NCI-60 was prediction of the tissues of origin of cell lines. In a sense, that class prediction task subsumes a large set of functional predictions. Our first approach, cluster analysis (Fig. 2), suggested that the consensus and protein data do well, with a slight edge to the protein. However, unsupervised clustering gives only a qualitative answer. Hence, we used the PAM shrunken centroid classifier to address the question more formally. As indicated in Fig. 3 and Table 3C, both transcript and protein data did well at minimizing the misclassification of cell lines. To assess the statistical significance of any differences in predictiveness of the databases, we used the two-tailed McNemar test, which focuses on discrepancies in prediction using the two data sets for the same set of genes. The protein data do better than any of the individual RNA data sets ($P$ = 0.00041–0.0098) and somewhat better than the consensus RNA set ($P$ = 0.027). To our knowledge, this is the first attempt to assess the effectiveness of mRNA and protein data in a functional prediction task for which the answer is independently known.

That analysis should be interpreted with several cautions. (*a*) It provides only a first example. (*b*) We are not formally able to distinguish true biological predictiveness from

Published OnlineFirst March 5, 2007; DOI: 10.1158/1535-7163.MCT-06-0650



differences in noise level and/or bias in the respective measurement methods. RPLA has favorable characteristics, but the transcript expression technologies are much better developed. (*c*) The transcript data sets were filtered more stringently than the protein set to eliminate data with low signal-to-noise ratios, but the genes for analysis were initially selected based on good signal in the RPLA measurements. Hence, there could have been a net selection bias in either direction (but perhaps more likely toward the protein). (*d*) Oligonucleotide probes may have been incorrectly mapped, cDNA clones may have been misidentified, and/or antibodies used in RPLA may not have been optimally specific. (*e*) Splice variation may have entered into the analysis. (*f*) The tissues of origin assigned to the cell lines might not be entirely correct. In the past, we found apparent misclassifications and corrected them: MDA-MB435 and MDA-N were reclassified as melanoma, rather than breast (15, 16), and MCF7/ADR was reclassified as the agnostic NCI/ADR and later OVCAR8/ADR (18). Those misidentifications were resolved (independently of the expression data sets) by sequence, DNA copy number, spectral karyotyping, and drug sensitivity data. It is also worth noting that some of the tissue-of-origin cell groups (colon, renal, melanotic melanoma, CNS, and leukemia) are quite distinct in their molecular properties, whereas others are more heterogeneous and less distinguishable from each other. But inhomogeneity of cells within a tissue-of-origin category (or undiscovered errors in ''gold standard'' classification) would have made it harder, rather than easier, to obtain statistically significant differences in predictiveness of the mRNA and protein classifiers. Factors such as the ones enumerated here would have to be borne in mind in *any* attempt to compare such different technologies as those for profiling of transcripts and proteins.

Our next question was whether particular classes of genes would show higher mRNA-protein correlation than others. We found that the protein-consensus mRNA correlations fall into two distinct groups (Fig. 4C). The mean correlation of group 1 is 0.71 and that of group 2 is 0.39. We then wanted to test the hypothesis (4) that structural proteins are associated with high protein-mRNA correlations. The clustered image map (heat map; Fig. 4A) and bar graph of correlations (Fig. 4B) indicate qualitatively that, indeed, the structural category is enriched with group 1 genes; quantitative GO analysis yielded a $t$ test $P$ value of $9.42 \times 10^{-13}$ (Wilcoxon rank-sum $P$ value of $1.89 \times 10^{-6}$). Protein levels can be controlled at the transcriptional, translational, and/or posttranslational levels. It has been reported independently that particular structural proteins (cytokeratin 8, villin, and moesin) are regulated transcriptionally (38–41), consistent with our finding of a higher transcript-protein correlation for structural genes. For genes in the GO signal transduction category, the correlation between consensus mRNA levels and protein levels was 0.38 ($n = 8$), somewhat lower than the overall average ($P = 0.043$) and in the middle of the range for group 2 genes. As shown in Fig. 4B, the only signal transduction gene appearing in group 1 was ERBB2, an intrinsic membrane protein. If ERBB2 were omitted, the signal transduction category would be definitively low-correlation.

In conclusion, the RNA-protein correlation was generally higher for GO structural proteins than for other GO biological categories. In this particular analysis, the protein measurements seemed be somewhat better overall functional predictors (of tissue of origin) on a gene-for-gene basis, but it would be very hard in this, or any other such study, to rule out all of the possible confounding factors, given the differences between technologies and the differences between algorithms for analysis of data. Important to note is that both the protein and the consensus RNA data did quite well. In terms of utility, the transcript level technologies are more advanced and are richer in information because of the sheer numbers of transcripts assessable currently. But most molecular targets for therapy and most biomarkers are assessed clinically at the protein level (e.g., by immunohistochemistry or enzyme-linked immunoassay). The implication is clear: for the present at least, both types of measurements have major roles to play in the further development of molecular targets and biomarkers for personalization of medicine.

**Acknowledgments**

We thank the many Developmental Therapeutics Program staff members who make such studies possible; the late Kenneth D. Paull for his pioneering work on analysis of NCI-60 data; Susan Holbeck, Daniel Zaharevitz, and Robert Shoemaker for their continued work on the screen and its data; and Patrick Brown, David Botstein, Douglas Ross, Eric Lander, Todd Golub, Jane Staunton, and colleagues in their laboratories for their collaboration in producing the legacy cDNA and Affymetrix HU6800 data sets used in the calculations presented here.

**References**

1. Young RA. Biomedical discovery with DNA arrays. Cell 2000;102:9–15.

2. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 1999;17:994–9.

3. Paweletz CP, Charboneau L, Bichsel VE, et al. Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. Oncogene 2001;20:1981–9.

4. Nishizuka S, Charboneau L, Young L, et al. Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. Proc Natl Acad Sci U S A 2003;100:14229–34.

5. Shoemaker R. The NCI60 human tumour cell line screen. Nat Rev Cancer 2006;6:813–23.

6. 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.

7. Shoemaker RH, Monks A, Alley MC, et al. Development of human tumor cell line panels for use in disease-oriented drug screening. Prog Clin Biol Res 1988;276:265–86.

8. Grever MR, Schepartz SA, Chabner BA. The National Cancer Institute: cancer drug discovery and development program. Semin Oncol 1992;19:622–38.

9. 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.

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

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

12. Monks A, Scudiero D, Skehan P, et al. Feasibility of a high-flux

Published OnlineFirst March 5, 2007; DOI: 10.1158/1535-7163.MCT-06-0650



anticancer drug screen using a diverse panel of cultured human tumor cell lines. J Natl Cancer Inst 1991;83:757 – 66.

**13.** Stinson SF, Alley MC, Kopp WC, et al. Morphological and immunocytochemical characteristics of human tumor-cell lines for use in a disease-oriented anticancer drug screen. Anticancer Res 1992;12:1035 – 54.

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

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

**16.** 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.

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

**18.** Bussey KJ, Chin K, Lababidi S, et al. Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. Mol Cancer Ther 2006;5:853 – 67.

**19.** Weinstein JN. Spotlight on molecular profiling: ''integromic'' analysis of the NCI-60 cancer cell lines. Mol Cancer Ther 2006;5:2601 – 5.

**20.** Lee JK, Bussey KJ, Gwadry FG, et al. Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells. Genome Biol 2003;4:R82.

**21.** Staunton JE, Slonim DK, Coller HA, et al. Chemosensitivity prediction by transcriptional profiling. Proc Natl Acad Sci U S A 2001; 98:10787 – 92.

**22.** Huang Y, Anderle P, Bussey KJ, et al. Membrane transporters and channels: role of the transportome in cancer chemosensitivity and chemoresistance. Cancer Res 2004;64:4294 – 301.

**23.** Annereau JP, Szakacs G, Tucker CJ, et al. Analysis of ATP-binding cassette transporter expression in drug-selected cell lines by a microarray dedicated to multidrug resistance. Mol Pharmacol 2004;66:1397 – 405.

**24.** 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.

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

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

**27.** Shao J, Tu DS. The jackknife and bootstrap. New York: Springer-Verlag; 1995.

**28.** Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002;99:6567 – 72.

**29.** Bussey KJ, Kane D, Sunshine M, et al. MatchMiner: a tool for batch navigation among gene and gene product identifiers. Genome Biol 2003; 4:R27.

**30.** Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 2005;33:D54 – 8.

**31.** Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 2001;29:137 – 40.

**32.** Chen G, Gharib TG, Huang CC, et al. Proteomic analysis of lung adenocarcinoma: identification of a highly expressed set of proteins in tumors. Clin Cancer Res 2002;8:2298 – 305.

**33.** Zhou Y, Gwadry FG, Reinhold WC, et al. Transcriptional regulation of mitotic genes by camptothecin-induced DNA damage: microarray analysis of dose- and time-dependent effects. Cancer Res 2002;62:1688 – 95.

**34.** Anderson L, Seilhamer J. A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 1997;18:533 – 7.

**35.** Celis JE, Kruhoffer M, Gromova I, et al. Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. FEBS Lett 2000;480:2 – 16.

**36.** Chen G, Gharib TG, Huang CC, et al. Discordant protein and mRNA expression in lung adenocarcinomas. Mol Cell Proteomics 2002; 1:304 – 13.

**37.** Tian Q, Stepaniants SB, Mao M, et al. Integrated genomic and proteomic analyses of gene expression in mammalian cells. Mol Cell Proteomics 2004;3:960 – 9.

**38.** Barila D, Murgia C, Nobili F, Perozzi G. Transcriptional regulation of the ezrin gene during rat intestinal development and epithelial differentiation. Biochim Biophys Acta 1995;1263:133 – 40.

**39.** Calnek D, Quaroni A. Differential localization by *in situ* hybridization of distinct keratin mRNA species during intestinal epithelial cell development and differentiation. Differentiation 1993;53:95 – 104.

**40.** Lankes WT, Furthmayr H. Moesin: a member of the protein 4.1-talin-ezrin family of proteins. Proc Natl Acad Sci U S A 1991;88:8297 – 301.

**41.** Bretscher A, Weber K. Fimbrin, a new microfilament-associated protein present in microvilli and other cell surface structures. J Cell Biol 1980;86:335 – 40.

AACR American Association for Cancer Research

Molecular Cancer Therapeutics

# Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study

Uma T. Shankavaram, William C. Reinhold, Satoshi Nishizuka, et al.

*Mol Cancer Ther* 2007;6:820-832. Published OnlineFirst March 5, 2007.