CellMiner: A Web-Based Suite of Genomic and Pharmacologic Tools to Explore Transcript and Drug Patterns in the NCI-60 Cell Line Set

William C. Reinhold1, Margot Sunshine1,3, Hongfang Liu1,4, Sudhir Varma1,5, Kurt W. Kohn1, Joel Morris2, James Doroshow1,2, and Yves Pommier1

Abstract

High-throughput and high-content databases are increasingly important resources in molecular medicine, systems biology, and pharmacology. However, the information usually resides in unwieldy databases, limiting ready data analysis and integration. One resource that offers substantial potential for improvement in this regard is the NCI-60 cell line database compiled by the U.S. National Cancer Institute, which has been extensively characterized across numerous genomic and pharmacologic response platforms. In this report, we introduce a CellMiner (http://discover.nci.nih.gov/cellminer/) web application designed to improve the use of this extensive database. CellMiner tools allowed rapid data retrieval of transcripts for 22,379 genes and 360 microRNAs along with activity reports for 20,503 chemical compounds including 102 drugs approved by the U.S. Food and Drug Administration. Converting these differential levels into quantitative patterns across the NCI-60 clarified data organization and cross-comparisons using a novel pattern match tool. Data queries for potential relationships among parameters can be conducted in an iterative manner specific to user interests and expertise. Examples of the in silico discovery process afforded by CellMiner were provided for multidrug resistance analyses and doxorubicin activity; identification of colon-specific genes, microRNAs, and drugs; microRNAs related to the miR-17-92 cluster; and drug identification patterns matched to erlotinib, gefitinib, afatinib, and lapatinib. CellMiner greatly broadens applications of the extensive NCI-60 database for discovery by creating web-based processes that are rapid, flexible, and readily applied by users without bioinformatics expertise. Cancer Res; 72(14); 3499–511. ©2012 AACR.

Introduction

Access to bioinformatics frequently acts as a choke point in the flow of information between large-scale technologies and the researchers who have the expertise to assess the data. Difficulty in fluid data access leads to restricted ability to integrate diverse data types, reducing understanding of complex biologic and pharmacologic systems. One such large-scale information set with multiple genomic and drug response platforms is the NCI-60 cancer cell line database. These cell lines, due to the extensive pharmacology and genomic data available, are prime candidates for data integration and broad public access.

The NCI-60 cell line panel was initially developed as an anticancer drug efficacy screen by the Developmental Therapeutics Program (DTP; http://dtp.nci.nih.gov/) of the U.S. National Cancer Institute (NCI). Many thousands of compounds have been and continue to be applied to the NCI-60 (1, 2). In parallel, multiple platforms have been used to characterize the cells including (i) array comparative genomic hybridization (aCGH; ref. 3) and karyotypic analysis (4), (ii) DNA mutational analysis (5), (iii) DNA fingerprinting (6), (iv) microarrays for microRNA expression (9, 10), and (vi) protein reverse-phase lysate microarrays (11).

An emphasis within our group is integration and open-access dissemination of molecular biology and molecular pharmacology information. One form of data integration we have developed over the last several years is the combination of multiple transcript microarray platforms. This integration saves time by preventing researchers from having to review data from each platform individually and improves the accuracy and reliability of the results. Starting with a 3-platform integration (12), we next tested the use of z-score averages for probes to facilitate integration of results of different platforms done at different times (13). A z-score is a mathematical transformation that for each probe measurement, for example, ABCB1 gene expression across the NCI-60, subtracts the mean (to center the data) and then divides by the SD (to normalize the range). This approach has recently been expanded to integrate 5 platforms (14)...
has proven to be both reliable and informative (15–18). Here, we show this approach can be adopted for the DTP drug activity as well.

In the current study, we introduce for noninformaticists a set of web-based tools accessible through our CellMiner web application (19) that allows rapid access to and comparison of transcript expression levels of 22,379 genes, 360 micro-RNAs, and 20,503 compounds including 102 Food and Drug Administration (FDA)-approved drugs. The tools allow easy identification of drugs with similar activity profiles across the NCI-60. The gene and drug assessments, having been derived from widely varying numbers of probes or experiments, include all probe or experimental results that pass quality control, allowing the assessment of data reliability. In addition, we introduce our Pattern Comparison tool, which rapidly searches for robust connections between these parameters, as well as any independent pattern of interest, and allows the user to mine data not only for specific genes or drugs but also for systems biology and systems pharmacology investigations.

Materials and Methods

Quantitation of gene transcript expression levels in the NCI-60 using five microarray platforms

Transcript expression for each gene was determined through the integration of all pertinent probes from five platforms. From Affymetrix (Affymetrix Inc.), we used the Human Genome U95 Set (HG-U95, GEO accession GSE5949; ref. 8); the Human Genome U133 (HG-U133, GEO accession GSE5720; ref. 8); the Human Genome U133 Plus 2.0 Arrays (HG-U133 Plus 2.0, GEO accession GSE32474; ref. 13); and the GeneChip Human Exon 1.0 ST array (GH Exon 1.0 ST, GEO accession GSE29682; ref. 14). From Agilent (Agilent Technologies, Inc.), we used the Whole Human Genome Oligo Microarray (WHG, GEO accession GSE29288; ref. 9). Affymetrix microarrays were normalized by GCRMA (20). All WHG mRNA probes detected in at least 10% of cell lines were normalized using GeneSpring GX by (i) setting gProcessedSignal values less than 5 to 5, (ii) transforming gProcessedSignal or gTotalGeneSignal to log2, and (iii) normalizing per array to the 75th percentile (9). Data for these microarrays are accessible at our web-based data

Figure 1. Snapshot of the NCI-60 analysis web site, a suite of web-based tools designed to facilitate rapid pharmacologic and genomic bioinformatics for the NCI-60 cell lines. A, these tools are accessible at the CellMiner web site (http://discover.nci.nih.gov/cellminer/) by clicking on the NCI-60 Analysis Tools tab. B, the analysis of interest (Z score determination, Mean centered graphs for microRNAs, or Pattern comparisons) is selected in step 1 using the check boxes. The specific identifier or pattern of interest is selected in step 2, either by typing in an identifier using the "Input list" function, or by uploading a file using the "Upload file" function. A maximum of 150 identifiers (genes, microRNAs, or drugs) can be input at once. The results are e-mailed to the address entered in step 3. Multiple check boxes may be selected for a single input. Radio buttons (circles) for an analysis type are mutually exclusive.
Quality control for genes is done as follows. For every probe for that gene, the intensity range across the NCI-60 is determined, and all probes with a range of $\leq 1.2$ log$_2$ are dropped. The number of probes that pass this criterion is determined and 25% of that number calculated (keeping a minimum of 2 and a maximum of 253). For the remaining probes, Pearson correlations are determined for all probe/probe combinations. The average correlation for each probe is determined compared with all other (remaining) probes. Probes whose average correlations are less than 0.30 ($P < 0.02$ for 60 cell lines in the absence of multiple comparisons correction) and not correlated to any other individual probes at $P < 0.03$ are dropped. The average correlation for remaining probes is then recalculated. Next, for probes with average correlations less than 0.60 ($P < 0.0000004$ for 60 cell lines), the lowest of these are dropped, and average correlations recalculated for all remaining probe combinations. Probes are dropped until all are $\geq 0.60$ or the 25% (of probes that passed the 1.2 log$_2$ range criteria) level is reached. This ensures significant pattern match across probes.

Probe intensity values that pass these quality controls are transformed to z-scores (21). Average z-scores were determined for each gene for each cell line. Probes with only one experiment that passes the $\leq 1.2$ log$_2$ range test are included as they are potentially informative but must be considered less reliable. For the 25% (of experiments that passed the 1.2 log$_2$ range criteria) level is reached. This serves to eliminate nonresponsive and out of proper range data. The number of experiments that pass these criteria is determined and 25% of that number calculated (keeping a minimum of 2 and a maximum of 253). Pearson correlations are determined for all remaining possible experiment/experiment combinations. Experiments whose average correlations are less than 0.334 ($P < 0.05$ for the 35 cell line minimum in the absence of multiple comparisons correction) and were not correlated to individual probes at $P < 0.334$ are dropped. For the remaining experiments with average correlations less than 0.60 ($P < 0.00014$ for the 35 cell line minimum), the lowest is dropped, and the correlations recalculated for all remaining possible experiment/experiment combinations. Experiments are dropped in this fashion until either all are $\geq 0.60$ or the 25% (of experiments that passed the 1.2 log$_2$ range criteria) level is reached. This ensures significant pattern match across experiments. Drugs with only one experiment but that pass the $\leq 1.2$ log$_2$ range test are included as output as they are potentially informative but must be considered less reliable. For the purpose of the manuscript, compounds that have not yet undergone clinical trials, as well as those that have, will sometimes be referred to as drugs.

MicroRNA expression levels

MicroRNA expression levels were determined as described previously (9) for the Agilent Technologies 15k feature Human miRNA Microarray (V2) following the manufacturer’s recommendations and are available at CellMiner (http://discover.nci.nih.gov/cellminer/) as well as at GEO (accession GSE22821).

Pattern comparisons

The comparisons of gene transcript expression, microRNA expression, and drug activity across the NCI-60 use the data prepared as described in the previous 3 sections. Pearson correlations between these parameters were calculated using Java 5.0.

Results

Web-based bioinformatics tool access for the NCI-60

Several important forms of bioinformatics analyses, made to be easily accessible for the noninformaticist, are now available at the Genomics and Bioinformatics Group’s CellMiner (http://discover.nci.nih.gov/cellminer/) site (Fig. 1A). To access those, click the "NCI-60 Analysis Tools" tab. These tools (Fig. 1B) allow the user to rapidly obtain information for the NCI-60 that would otherwise require lengthy data retrieval, compilation, and assessment. Included are data for relative transcript and microRNA levels, as well as drug activity. The tools include pattern comparison functionality that enables the identification of relationships between these and other parameters, as driven by the user’s interest.

In each case (Fig. 1B), users select the tool in step 1. Identifiers or patterns are typed in if one chooses "Input list," or uploaded as a .txt or .xls file if "Upload file" is chosen in step 2. Next, users enter their e-mail address for receipt of the data in step 3, and click "Get data" to receive data (as an Excel file). Data are generally available within a few minutes. Error files are returned if no results are provided, stating the reason. These tools are described with examples in the subsequent figures.

Relative transcript levels tool

To obtain high-definition, easily interpretable transcript-level data, select the "Z score determination" tool and "Gene transcript level" shown in step 1 of Fig. 2A. Check "Input list" and input your gene of interest in step 2 (Fig. 2A and B). Official gene names are required [see the Human Genome Organization, HUGO (http://www.genenames.org)]. In the example, ABCB1 is the HUGO name of the gene encoding the P-glycoprotein MDR (P-gp) drug efflux transporter. The tool compiles all probe information available for the selected gene from 5 microarray platforms, the Affymetrix HG-U95, HG-133, HG-U133 Plus 2.0, and the GeneChip Human Exon 1.0 ST array. Quality control is included on the basis of single probe variation, as well as probe-to-probe correlation to eliminate poor quality, dead, or background-level probes.

The tool output includes relative transcript intensity presented as z-scores and visualized as a bar graph (Fig. 2C). The bars for each cell line are color-coded by tissue of origin. Parameters of the Affymetrix probe intensities (range, minimum, maximum, average, and SD) are included (Fig. 2D) to provide an untransformed reflection of transcript-level
Figure 2. Relative transcript expression and drug activity levels in the NCI-60. A, check the "Z score determination" check-box and then choose either the "Gene transcript level" or the "Drug activity" radio button in step 1. In those instances in which multiple genes or drugs are entered, a cross-correlation of all...
variation. Chromosomal location is also included (Fig. 2D). In the example shown in Fig. 2C, NCI-ADR-RES and HCT-15 are the top 2 expressers of ABCB1. Notably, most cell lines appear to the left of the mean with uniformly low values. Those low values all approach 2.79 log₂ intensity units (Fig. 2D), which is within background and indicates that all those cell lines have undetectable or low ABCB1 transcript levels.

In each data file for individual genes, the user also receives (i) the total number of probes for the gene (including those that failed), (ii) all intensity values (for those probes that passed quality control) to a maximum of 122, (iii) the z-score transforms for each probe, (iv) the platform that each probe originated from, (v) the exon to which the probe hybridizes, (vi) the average z-score for each cell line, and (vii) the distribution of the individual cell line average intensities presented as a histogram. Currently information on 22,379 genes is available.

The "Z score determination" tool also allows multiple gene entries in the input box (Fig. 2B; with each gene on a separate line). Each gene is then returned as separate Excel file with its own plots and data. In the case of multiple genes being entered, a cross-correlation table of the resultant z-scores can be generated by clicking the "Include Cross-correlations" box in Fig. 2A, step 1.

Relative drug activities tool
To obtain curated GI₅₀ data (50% growth inhibition), select the "Z score determination" tool shown in Fig. 1A (top right) and then select "Drug activity" Specify the compound using the National Service Center number (NSC) as the input (Fig. 2E, the NSC of doxorubicin is 123127). This tool will compile all experimental information available from the DTP (http://dtp.nci.nih.gov/) for the selected compound(s). Quality control is included on the basis of single experiment variation as well as experiment-to-experiment correlation to eliminate out of appropriate concentration, weak or invariant response, and irreproducible experiments. The compound must have a minimum of 35 cell lines with activity information to be included. The tool output includes relative compound activity presented as z-scores and visualized as a bar graph (Fig. 2F). Parameters of the compound activities (range, minimum, maximum, average, and SD) are included (Fig. 2G) to provide the user with an untransformed reflection of compound activity. The distribution of the individual cell line average responses is presented as a histogram. In the example shown, NCI-ADR-RES and HCT-15 are the two cell lines most resistant to doxorubicin, which is consistent with the fact they overexpress ABCB1, whose gene product P-gp pumps the drug from the cell (23, 24).

In the case of multiple drugs being entered, a cross-correlation table of the resultant z-scores for each drug may also be generated by clicking the "Include Cross-correlations" box in Figs. 2A and 1B, step 1.

The user also receives (i) the total number of experiments done for the compound (including those that failed), (ii) whether the compound has been FDA-approved, (iii) all activity values (for those experiments that passed quality control, to a maximum of 253), (iv) the z-score transforms for each experiment, and (v) the average z-score for each cell line. Currently, 50,703 compounds have passed through these filters, resulting in data for 20,503 compounds. NSC numbers are the required input (download the list of currently available NSCs at "List of NSC numbers available for analysis"; Fig. 1B). NSC numbers can be cross-referenced to other identifying parameters, including their chemical structures (http://dtp.nci.nih.gov/dtpstandard/dwindex/index.jsp).

Relative microRNA levels tool
To rapidly obtain the log₂ intensity values for microRNAs, the "microRNA mean centered graphs" tool is selected as shown in step 1 of Fig. 3A. Input your microRNA(s) of interest in step 2 using the name from the provided "List of microRNA identifiers available for analysis" accessed by clicking [download] in Step 1. The microRNA tool provides access to the average log₂ intensity values for the Agilent Technologies Human miRNA Microarray (V2; ref. 9). The z-scores are not used as this is a single platform analysis done at a single time. The tool output includes relative transcript intensities visualized as a bar graph. In the example shown in Fig. 3B, we queried the data for a microRNA from the miR-17-92 oncogenic cluster (25). Other microRNAs from the same cluster showed the same pattern of expression, with high levels in leukemia and colon cancer lines (data not shown; see Discussion). Parameters of probe intensities (range, minimum, maximum, average, and SD), as well as the chromosomal location are included (Fig. 3C). The distribution of the individual cell line average intensities is presented as a histogram. Currently information on 360 microRNAs is available.

Pattern comparison tools
To make comparisons between an input pattern of interest (including expression of a gene or microRNA, activity of a drug, or any pattern of interest) and (i) gene expression, (ii) microRNA expression, and (iii) drug activity levels, start by selecting the "Pattern comparison" tool shown in step 1 of Fig. 4A, and 1 of the 4 possible types of input, "Gene symbol," "microRNA," "Drug NSC=", or "Pattern in 60 element
Figure 3. Relative microRNA transcript expression levels in the NCI-60. A, input of data. Check the "microRNA mean centered graphs" check box in step 1. To access the list of microRNAs names to use, go to footnote 2, "List of microRNA identifiers" and click "download". In step 2, the user chooses whether to type in the input or upload it as a file (.txt or .xls) by selecting "Input list" or "Upload file," respectively. If typing in the input, input the microRNA name(s) in the "Input the identifiers" box. In step 3, enter the e-mail address to send the results to. B, the output includes a plot of the mean-centered average log2 intensities along with their numerical values (data not shown). C, mean log2 intensity values for range, minimum, maximum, average, and SD for all cell lines are included to assist in data interpretation, as well as a histogram of the cell lines average intensities. For the histogram, the x-axis is the average intensity for the cell lines and the y-axis the frequency at which they occur.
### Step 1: Select analysis type:
- **Pattern comparison**
  - Gene symbol (input HUGO name)
  - microRNA
  - Drug NSC#
  - Pattern in 60 element array

1. List of NSC numbers available for analysis (download).
2. List of microRNA identifiers available for analysis (download).
3. Pattern comparison template file (download). Just file in the values next to the appropriate cell line.

### Step 2 - Identifiers may be input as a list or a file (maximum 150 names). Select input format:
- Input list
- Upload file

Input the identifier(s):

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Input Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCB1</td>
<td></td>
</tr>
</tbody>
</table>

### Step 3: Your E-mail Address

youremail@org

---

### Significant gene correlations

<table>
<thead>
<tr>
<th>Name</th>
<th>Correlations</th>
<th>Annotations</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCB1</td>
<td>1.000</td>
<td>Efflux</td>
<td>7q21.12</td>
</tr>
<tr>
<td>RUNDC3B</td>
<td>0.750</td>
<td>na</td>
<td>7q21.12</td>
</tr>
<tr>
<td>DNAJC5G</td>
<td>0.691</td>
<td>na</td>
<td>2p23.3</td>
</tr>
<tr>
<td>RG87BP</td>
<td>0.688</td>
<td>na</td>
<td>5q12.3</td>
</tr>
<tr>
<td>SLC13A5</td>
<td>0.649</td>
<td>Drug uptake</td>
<td>17p13.2</td>
</tr>
<tr>
<td>RASIP1</td>
<td>0.639</td>
<td>na</td>
<td>19q13.31</td>
</tr>
<tr>
<td>RPL17P4</td>
<td>0.629</td>
<td>na</td>
<td>14q32.33</td>
</tr>
<tr>
<td>ZCWPW2</td>
<td>0.615</td>
<td>na</td>
<td>3p24.1</td>
</tr>
<tr>
<td>SMG7</td>
<td>-0.419</td>
<td>na</td>
<td>1q25.3</td>
</tr>
<tr>
<td>EDEM3</td>
<td>-0.429</td>
<td>na</td>
<td>1q31.1</td>
</tr>
<tr>
<td>B2M</td>
<td>-0.430</td>
<td>na</td>
<td>15q21.1</td>
</tr>
<tr>
<td>PDCD10</td>
<td>-0.447</td>
<td>DNA damage</td>
<td>3q26.1</td>
</tr>
<tr>
<td>MRI1</td>
<td>-0.460</td>
<td>na</td>
<td>19p13.12</td>
</tr>
</tbody>
</table>

### Significant microRNA correlations

<table>
<thead>
<tr>
<th>Name</th>
<th>Correlations</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-miR-517c</td>
<td>0.411</td>
<td>19q13.33</td>
</tr>
<tr>
<td>hsa-miR-521</td>
<td>0.381</td>
<td>19q13.33</td>
</tr>
<tr>
<td>hsa-miR-522</td>
<td>0.360</td>
<td>19q13.33</td>
</tr>
<tr>
<td>hsa-miR-516a-5p</td>
<td>0.337</td>
<td>19q13.33</td>
</tr>
<tr>
<td>hsa-miR-516a-5p</td>
<td>0.315</td>
<td>19q13.33</td>
</tr>
<tr>
<td>hsa-miR-106b</td>
<td>-0.273</td>
<td>7q22.1</td>
</tr>
<tr>
<td>hsa-miR-598</td>
<td>-0.274</td>
<td>8p23.1</td>
</tr>
<tr>
<td>hsa-miR-362-3p</td>
<td>-0.282</td>
<td>Xp11.22</td>
</tr>
<tr>
<td>hsa-miR-93</td>
<td>-0.284</td>
<td>7q22.1</td>
</tr>
<tr>
<td>hsa-miR-16-2*</td>
<td>-0.285</td>
<td>3q25.33</td>
</tr>
<tr>
<td>hsa-miR-362-5p</td>
<td>-0.294</td>
<td>Xp11.22</td>
</tr>
<tr>
<td>hsa-miR-25</td>
<td>-0.319</td>
<td>7q22.1</td>
</tr>
<tr>
<td>hsa-miR-652</td>
<td>-0.368</td>
<td>Xq22.3</td>
</tr>
</tbody>
</table>

### Significant drug correlations

<table>
<thead>
<tr>
<th>NSC number</th>
<th>Name</th>
<th>Mechanism</th>
<th>Correlations</th>
<th>FDA status</th>
</tr>
</thead>
<tbody>
<tr>
<td>665724</td>
<td>na</td>
<td>na</td>
<td>0.707</td>
<td>Not approved</td>
</tr>
<tr>
<td>661908</td>
<td>na</td>
<td>na</td>
<td>0.706</td>
<td>Not approved</td>
</tr>
<tr>
<td>673294</td>
<td>na</td>
<td>na</td>
<td>0.690</td>
<td>Not approved</td>
</tr>
<tr>
<td>698964</td>
<td>na</td>
<td>na</td>
<td>0.687</td>
<td>Not approved</td>
</tr>
<tr>
<td>710857</td>
<td>na</td>
<td>na</td>
<td>0.652</td>
<td>Not approved</td>
</tr>
<tr>
<td>39688</td>
<td>na</td>
<td>na</td>
<td>0.650</td>
<td>Not approved</td>
</tr>
<tr>
<td>57969</td>
<td>na</td>
<td>na</td>
<td>0.641</td>
<td>Not approved</td>
</tr>
<tr>
<td>123127</td>
<td>Doxorubicin</td>
<td>T2</td>
<td>-0.551</td>
<td>FDA approved</td>
</tr>
<tr>
<td>630176</td>
<td>Romidepsin</td>
<td>HDAC</td>
<td>-0.854</td>
<td>FDA approved</td>
</tr>
<tr>
<td>58514</td>
<td>Chromomycin</td>
<td>na</td>
<td>-0.855</td>
<td>Not approved</td>
</tr>
<tr>
<td>76411</td>
<td>na</td>
<td>na</td>
<td>-0.864</td>
<td>Not approved</td>
</tr>
<tr>
<td>259968</td>
<td>Bouvardin</td>
<td>na</td>
<td>-0.872</td>
<td>Not approved</td>
</tr>
<tr>
<td>685703</td>
<td>na</td>
<td>na</td>
<td>-0.918</td>
<td>Not approved</td>
</tr>
</tbody>
</table>

---

Figure 4. Pattern comparison to transcript expression, microRNA expression, and drug activity levels in the NCI-60. A, choose your input type by checking the “Pattern comparison” check box and then choose either the “Gene symbol,” “microRNA,” “Drug NSC#,” or “Pattern in 60 element array” radio buttons in step 1. In step 2, the user chooses whether to type in the input or upload it as a file by selecting “Input list” or “Upload file,” respectively (use the same identifiers as in Figs. 2 and 3). To input your own template, use the “Pattern comparison template file” download from footnote 3, with your numerical values and “na” for missing or to be ignored values. B, significant gene correlation output, given for all genes that match your input pattern at a significance level of \( P < 0.05 \). C, significant microRNA correlations output, given for all microRNAs that match your input pattern at a significance level of \( P < 0.05 \). D, significant drug correlations output, given for all compounds that match your input pattern at a significance level of \( P < 0.05 \). Only the top and bottom of the lists are shown in each case.
### Significant gene correlations

<table>
<thead>
<tr>
<th>Name</th>
<th>Correlations</th>
<th>Annotations</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRIM15</td>
<td>0.901</td>
<td>–</td>
<td>6p21.33</td>
</tr>
<tr>
<td>RNF43</td>
<td>0.868</td>
<td>–</td>
<td>17q23.2</td>
</tr>
<tr>
<td>VIL1</td>
<td>0.844</td>
<td>–</td>
<td>2q35</td>
</tr>
</tbody>
</table>

### Colon-specific input pattern

- BR:MCF7
- CNS:SF_268
- CNS:SF_295
- CNS:SF_539
- CNS:SNB_19
- CNS:SNB_75
- CNS:U251
- CO:COLO205
- CO:HCC_2998
- CO:HCT_116
- CO:HCT_15
- CO:HT29
- CO:KM12
- CO:SW_620
- LE:CCR_F_CEM
- LE:HL_60
- LE:K_562
- LE:MOLT_4
- LE:RPMI_8226
- LE:SR
- RE:UO_31

### Drugs

#### Significant drug correlations

<table>
<thead>
<tr>
<th>NSC number</th>
<th>Name</th>
<th>Correlations</th>
<th>FDA status</th>
</tr>
</thead>
<tbody>
<tr>
<td>732298</td>
<td>–</td>
<td>0.653</td>
<td>Not approved</td>
</tr>
<tr>
<td>366140</td>
<td>Pyrazoloacridine</td>
<td>0.425</td>
<td>Clinical trial</td>
</tr>
<tr>
<td>741078</td>
<td>Selumetinib</td>
<td>0.367</td>
<td>Clinical trial</td>
</tr>
<tr>
<td>750690</td>
<td>Sunitinib</td>
<td>0.403</td>
<td>Clinical trial</td>
</tr>
<tr>
<td>732298</td>
<td>Selumetinib</td>
<td></td>
<td></td>
</tr>
<tr>
<td>741078</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### microRNAs

- *mir-215* $r = 0.745$
- *mir-194* $r = 0.739$
array.” Input your identifier(s) in step 2, using a gene, microRNA, drug, or pattern. The first 3 of these are the same 22,379 gene expression levels, 20,503 drugs activity levels, and 360 microRNA expression levels, described above (see Figs. 1–3). The fourth option is to input any pattern across the NCI-60 that is of interest to the user. A “Pattern comparison template file” for this option is accessible for download in step 1, footnote 3 (Fig. 4A).

The output is returned as a single Excel spreadsheet file that includes 6 main sections: “All” and “Significant gene correlations” (Fig. 4B), “All” and “Significant microRNA correlations” (Fig. 4C), and “All” and “Significant drug correlations” (Fig. 4D; the 3 All columns are not shown in B–D). The top and bottom portions of these sets from the input of ABCB1 are shown in Fig. 4. The “all” versions of these outputs contain the same columns as do the “significant,” with the exception of the FDA status column in Fig. 4D but include all possible correlations and are in either alphabetical (for genes and microRNAs) or numerical (for drugs by NSC number) order. The significant correlations outputs, are based on statistical significance ($P < 0.05$) in the absence of multiple comparisons correction and are ordered by descending correlations with color-coding (red- and blue-bold fonts for positive and negative significant correlations, respectively). The data used for the correlation calculations for the genes, drugs, and microRNAs are those generated in the gene transcript level, drug activity, and microRNA tools in Figs. 2 and 3, respectively. Information included for the genes sections will be Gene Ontology annotations accessed in Figs. 2 and 3, respectively. Information included for the drug sections are name (where available, see the Name column, Fig. 4D), mechanism of action (for 355 drugs currently, see the Mechanism column, Fig. 4D), and for the significantly correlated section, the FDA approval status (see the FDA status column, Fig. 4D). A footnotes page is included with the output to provide details on the definitions, approaches, abbreviations, and background.

Additional uses of the pattern tools (starting with a pre-defined pattern or a given drug) are presented as examples in the Discussion (see Figs. 5 and 6).

**Discussion**

The sheer quantity of data generated by current high-throughput platforms have encumbered their use and access. However, because of the emergence of new "omic" technologies and the questions to be asked in dealing with issues of human disease, it is extremely important to open such data to clinicians, molecular biologists, and others with insights into aspects of disease. This is clearly the case in the difficult and complex case of cancer.

Our web-based tools provide assistance in this area for the NCI-60. They allow rapid determination of (i) a composite "best" gene transcript expression level pattern from 5 microarray platforms, (ii) drug activity from all experimental repeats done by the DTP, and (iii) transcript levels of microRNAs from duplicate microarrays (ref. 9; Figs. 2 and 3). The "Pattern comparison" tools (Figs. 4–6) allow rapid, global exploration of relationships between these parameters and any input pattern of interest.

This accession of relative levels of transcript expression (Fig. 2C) provides several advantages. The tool (i) compiles all probes for a single gene, (ii) incorporates built-in quality control, and (iii) includes probe remapping based on HG-19, eliminating the need to spend time or have expertise in those areas. Allowing the user to see the input probes for a single gene allows assessment of reliability and accuracy. For example, an expression pattern for a gene with 46 highly correlated probes (as is the case for ABCB1 in Fig. 2C) shows both high reliability and accuracy. Conversely, a gene with 2 probes that barely pass quality control should be considered both less reliable and accurate. Availability of reliable patterns of relative transcript expression in turn facilitates comparison with other data, such as drug activity and microRNA expression. For ABCB1 and doxorubicin (Fig. 2), there is obvious correlation between high ABCB1 expression and doxorubicin resistance, consistent with prior results (23, 24).

During development of the tools derived for gene transcript z-score patterns, we have validated and exploited them to elucidate novel gene regulation mechanisms for MYC, TOP1, and CHEK2 (15–17). The cross-correlation function illustrated in Fig. 2A, step1, identified transcriptional coregulation among kinetochore genes (17) and across genes driving cell migration and adhesion (26).

That regulatory elements may be identified using these tools is shown by the example of the cell–cell adhesion factor CDH1 (E-cadherin). Using the pattern comparison tool with CDH1 as input, one finds that each of transcriptional repressors ZEB1 (TCF8), SNAI2 (SLUG), ZEB2 (SIP1), and TWIST1 has significant negative correlation to CDH1 transcript levels, at $–0.63$, $–0.47$, $–0.37$, and $–0.47$, respectively. These repressors have been reported previously to negatively regulate CDH1 (27–31). In
addition, they show robust positive correlations to hsa-miR-200c, -200a, -200b, -200b*, and -200a*, with values of 0.73, 0.59, 0.55, 0.48, and 0.41, respectively. The microRNA-200 family has been previously associated with E-cadherin regulation through ZEB1 and ZEB2 targeting (32, 33).

The microRNA mean-centered graph tool (Fig. 3A and B) simply provides raw data and mean-centered graphical representation of the data determined previously (9). In the example shown, miR-18a exhibits notable specificities for colon and especially leukemia cell lines. miR-18a has previously been shown to be part of a polycistronic oncogenic microRNA cluster, which includes 6 consecutive mature microRNAs, miR-17-5p, -19a, -19b-3p, -20a, and -92a co-expressed as a single primary transcript (25, 34). Notably, the pattern comparison tool for miR-18a retrieves the 5 other microRNAs at the top of the significantly correlated microRNA list (with Pearson correlation coefficients between 0.96 and 0.77) and MYC ranked 37th in the gene list with a correlation coefficient of 0.61. The
MYC correlation is consistent with the fact that the miR-17-92 cluster is a transcriptional MYC target (35). The ability to rapidly compare patterns of transcript expression, microRNA expression, and drug activity (Figs. 2 and 3) both with themselves and other patterns of interest provides a powerful and flexible tool to the user. The tool automatically (i) allows the input of multiple types of data, (ii) calculates the correlation to 3 types of data, and (iii) identifies those correlations that are statistically significant. Allowing the user to see the correlation level and order in which genes, microRNAs, or compounds are ranked for any single input provides criteria by which to identify potential relationships between disparate parameters and to access the robustness of these relationships. Of the named drugs in Fig. 4D other than doxorubicin, romidepsin (FK228) has previously been shown to undergo efflux by ABCB1 (36) and bouvardin to have cross-resistance with doxorubicin (37). To the best of our knowledge, chromomycin had not been previously associated with ABCB1. Significant microRNA expression versus gene expression, gene expression versus drug activity, and microRNA expression versus drug activity correlations have previously been identified in this fashion (14, 17).

A novel example is provided in Fig. 5, in which a colon-specific pattern is entered into the pattern comparison tool. The input was 1 for all non-colon cell lines and 5 for all colon cell lines. The top 3 genes from the significant gene correlations output are shown. The top gene, TRIM15, is a little studied tripartite motif family gene, but its high correlation (0.901) to the colon-specific input makes it a candidate for being a colon cancer--specific marker. RNF43 is an ubiquitin ligase known to be upregulated in colon cancer (38). VIL1 is an actin-binding protein previously identified as a diagnostic marker for colon cancer (39). The bar graphs for VIL1 and RNF43, derived from the "Z score determination" tool (Fig. 2) illustrate their colon specificity. The bar plots for the top 2 microRNAs from the significant microRNA correlations are shown next. Both miR-215 and -194 (with correlations of 0.745 and 0.739, respectively) have previously been described as prognostic indicators for colorectal cancer (40, 41). The bar graphs, from the microRNA mean centered graphs" tool from Fig. 3, illustrate their colon specificity. Four of the top 7 drugs from the significant drug correlations list are also shown. Of these, 3 have some clinical trial history. Pyrazoloacridine has been tested in colon cancer with modest success (42). Selumetinib (AZD6244) has been proposed as a potential therapy for colorectal cancer (43). From our bar graph, it appears to have specificity for melanoma in addition to colon cancers. Sunitinib is in clinical trial for colorectal cancer (44). The additional compound presented as a bar graph, NSC732298, has better specificity for colon that any of these (r = 0.653) but has not gone through clinical trials. The bar graphs for selumetinib and 732298 from the "Z score determination" tool from Fig. 2 illustrate their colon specificity. Together, these genes, microRNAs, and drugs illustrate how one can, from a single pattern input, rapidly obtain multiple types of both known and novel information relevant to a users’ area of interest.

The ability to access relative drug activity levels and patterns (Fig. 2F) provides advantages similar to those for relative transcript levels. The tool automatically (i) compiles all experiments for a single drug from all experiments done and (ii) incorporates quality control into the approach. It allows the user to see the input experiments, allowing assessment of reliability and accuracy. This is especially important for the activity data, as there are many compounds with small numbers of experiments and some with reduced numbers of cell lines tested. An example of a highly reliable activity pattern is that for doxorubicin, which has 253 highly correlated experiments (Fig. 2F). The identification of a reliable pattern of relative compound activity facilitates its comparison with other types of data. Drug patterns derived in this fashion have been used previously for comparison with mRNA and microRNA expression (14).

A final example of the flexibility of the new tools is depicted in Fig. 6. As illustrated, the pattern comparison tools can be readily adapted to do COMPARE-like (1, 45) drug analyses, in which the user can query the 20,503 drugs and chemicals from the DTP database to identify similar drugs. In Fig. 6, the input is erlotinib (NSC 718781), an inhibitor of the EGF receptor (EGFR) tyrosine kinase. Pleasingly, the 2 other FDA-approved EGFR-targeted drugs gefitinib and lapatinib ranked fifth and sixth, and afatinib (BIBW2992), which is in advanced clinical trials and also targets EGFR-ERB kinase, ranked third among the 20,503 drugs (Fig. 6B). Compound 693255, which ranked second is a tyrphostin derivative, a class of drugs shown to inhibit tyrosine kinase including EGFR (46). The plots shown in Fig. 6C were obtained using the z-score tool (see Fig. 2F). They show pattern similarity among the 4 highly correlated drugs. Conversely, the piperidinium compounds (618757, 636676, 638634, and 630602 with r = −0.587, −0.536, −0.532, and −0.504, respectively) have consistently high inverse correlation. That is, they work well when the EGFR inhibitors work poorly. This example illustrates the usefulness of tools in the comparison of drugs with similar mechanisms of action, including those that are in clinical development (such as afatinib), as well as identifying novel compounds that either might work in a similar fashion, or are inverse to it.

The ease, rapidity, and flexibility of the new tools provide users with an important new data integration capacity. One of our goals is to update and enhance these tools in an ongoing fashion. Our next 2 additional tools will be the Comparative Genomic Hybridization tool and database (aCGH), which will assist in interpretation of DNA copy number using our Roche NimbleGen 385k CGH array, and the Whole Exome Sequencing tool and database (WES), which will provide access to the whole genome sequences for all exons across the NCI-60 (http://discover.nci.nih.gov/cellminer/). Together, these databases and tools provide publicly available and unique opportunities for systems biology and systems pharmacology investigations.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.
Authors’ Contributions

Conception and design: W.C. Reinhold, M. Sunshine, H. Liu, Y. Pommier
Development of methodology: W.C. Reinhold, M. Sunshine, H. Liu, Y. Pommier
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): W.C. Reinhold, J. Morris, J. Doroshow
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): W.C. Reinhold, H. Liu, S. Varma, Y. Pommier
Writing, review, and/or revision of the manuscript: W.C. Reinhold, M. Sunshine, S. Varma, K.W. Kohn, J. Doroshow, Y. Pommier

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): M. Sunshine, S. Varma

Study supervision: W.C. Reinhold, Y. Pommier

References


Grant Support

This work is supported by the Center for Cancer Research, the intramural program of NCI, and the DTP, Division of Cancer Treatment and Diagnosis (DCTD), NCI.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received April 16, 2012; accepted April 27, 2012; published July 16, 2012.
43. Yoon J, Koo RH, Choi KY. MEK1/2 inhibitors AS703026 and AZD6244 may be potential therapies for KRAS mutated colorectal cancer that is resistant to EGFR monoclonal antibody therapy. Cancer Res 2011;71:445–53.
CellMiner: A Web-Based Suite of Genomic and Pharmacologic Tools to Explore Transcript and Drug Patterns in the NCI-60 Cell Line Set


Updated version
Access the most recent version of this article at:
http://cancerres.aacrjournals.org/content/72/14/3499

Cited articles
This article cites 45 articles, 24 of which you can access for free at:
http://cancerres.aacrjournals.org/content/72/14/3499.full.html#ref-list-1

Citing articles
This article has been cited by 6 HighWire-hosted articles. Access the articles at:
/content/72/14/3499.full.html#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.