Published in final edited form as:

Clin Cancer Res. 2015 September 1; 21(17): 3841–3852. doi:10.1158/1078-0432.CCR-15-0335.

Using CellMiner 1.6 for Systems Pharmacology and Genomic Analysis of the NCI-60

William C. Reinhold[1], **Margot Sunshine**[1,2], **Sudhir Varma**[1,2,3], **James H. Doroshow**[1,4], and **Yves Pommier**[1]

[1]Developmental Therapeutics Branch and Laboratory of Molecular Pharmacology, Center for Cancer Research, NCI, NIH, Bethesda, Maryland [2]Systems Research and Applications Corp., Fairfax, Virginia [3]HiThru Analytics LLC, Laurel, Maryland [4]Developmental Therapeutics Program, DCTD, NCI, NIH, Bethesda, Maryland

Abstract

The NCI-60 cancer cell line panel provides a premier model for data integration and systems pharmacology being the largest publicly available database of anticancer drug activity,, genomic, molecular, and phenotypic data. It comprises gene expression (25,722 transcripts), microRNAs (360 miRNAs), whole genome DNA copy number (23,413 genes), whole exome sequencing (variants for 16,568 genes), protein levels (94 genes), and cytotoxic activity (20,861 compounds). Included are 158 Food and Drug Administration (FDA)-approved drugs and 79 that are in clinical trials. To improve data accessibility to bioinformaticists and non-bioinformaticists alike, we have developed the CellMiner web-based tools. Here we describe the newest CellMiner version, including integration of novel databases and tools associated with whole exome sequencing and protein expression, and review the tools. Included are i) "Cell line signature" for DNA, RNA, protein and drugs, ii) "Cross correlations" for up to 150 input genes, microRNAs, and compounds in a single query, iii) "Pattern comparison" to identify connections among drugs, gene expression, genomic variants, microRNA and protein expressions, iv) "Genetic variation versus drug visualization" to identify potential new drug:gene DNA variant relationships, and v) "Genetic variant summation", designed to provide a synopsis of mutational burden on any pathway or gene group for up to 150 genes. Together, these tools allow users to flexibly query the NCI-60 data for potential relationships between genomic, molecular and pharmacological parameters in a manner specific to the user's area of expertise. Examples for both gain- (RAS) and loss- (PTEN) of-function alterations are provided.

Corresponding Authors: William C. Reinhold, NIH, 9000 Rockville Pike, Building 37, Room 5041, Bethesda, MD 20892; wcr@mail.nih.gov, and Yves Pommier, NIH, 9000 Rockville Pike, Building 37, Room 5068, Bethesda, MD 20892; pommier@nih.gov.
W.C. Reinhold and Y. Pommier share senior authorship.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

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

Introduction

This review provides a synopsis of both the use and novel features of the CellMiner web application. CellMiner is designed specifically for the purpose of facilitating integration of pharmacological with molecular data from the NCI-60 cell lines. Its provision of “Cell line signatures” for both drug activity and multiple forms of molecular data, in which many of the preprocessing steps have already been done, allow a broad segment of the scientific community to make rapid and meaningful explorations into pharmacological-molecular relationships. The cell line models will always form the basis for studies of this type, due to their obvious advantages in providing testable models. Observations and hypotheses with translational importance made with these models will increasingly form the intellectual basis for a more specific and logical application of treatments, based on a patient’s disease’s specific molecular characteristics.

Data and Tools Available through CellMiner

CellMiner is a web-based application that provides both the data for the NCI-60 cancer cell lines, and the tools to mine those data (1, 2). It is appropriate for both the novice as well as the expert in the field. The application is accessed using the URL in ref. 3, and, in the current version, is organized into the seven tabs shown in Fig. 1A (3).

The “Home” tab (top left, Fig. 1A) provides: i) a general description of the site, ii) references, iii) recent press releases, iv) a description of each of the other six tabs, and v) links to the Discover, and Developmental Therapeutics Program websites (4, 5).

The “NCI-60 Analysis Tools” tab provides visualizations and patterns for quality controlled molecular and pharmacological data, as well as tools for data integration. Both the data sets and integration tools available in this tab are focused on in this publication.

The “Query Genomic Data” tab provides access to all molecular data, queryable by gene identifier, chromosomal or genomic location, or platform specific identifier for 17 platforms. It does not provide a synopsis result by gene or have internal requirements for consistency or range (within a single gene) as within the “NCI-60 Analysis Tools”. Data are received in both Excel (.xls) and text (.txt) format. New data sets and organization improvements continue to be added to this tab.

The “Query Drug Data” tab provides access to the data for growth inhibition 50% (GI50) compound activities measured by the Developmental Therapeutics Program (5, 6). The curated version of these data with synopsis results by drug and internal requirements for consistency and range (within a single drug) are in the “NCI-60 Analysis Tools” (see below). Compounds are queryable using one of six options (Fig. 1B, Step 1). Data are received in both Excel and text format. Data updates and organization improvements have been and will continue to be incorporated in this tab.

The “Download Data Sets” tab allows one to download entire datasets in either raw or normalized formats (dependent on dataset). Available in this section is also our cell line fingerprinting data, a useful resource for identification of each of the 60 cell lines (7). Also, introduced and included here in the lower “Download Normalized Data Set” section of the page is our “RNA: 5 Platform Gene Transcript” compilation of quality controlled transcript data from microarrays. This is the all gene version of the information provided by “NCI-60 Analysis Tools”\“Cell line signature”\“Gene transcript z scores”. The data sets within this tab will primarily be of use to the bioinformatician. New data sets and organization improvements continue to be incorporated to this tab.

The “Cell Line Metadata” tab provides background information on the cell lines, including tissue of origin, age and sex of patient, prior treatment (when known), histology, ploidy, TP53 status, multi-drug resistance function, and doubling time. A link to each cell lines fingerprinting data is included.

The “Data Set Metadata” tab provides background information on the types of data, platform information, the principal collaborators, and some description of the data. Currently there are 22 datasets described here, with more being added.

The “Query Genomic Data” and “Query Drug Data” Tools

These two tools give access to the specific data in the absence of the additional quality control assumptions applied within the “NCI-60 Analysis Tools” section. The data in this form may be preferential dependent on the question being asked, and allows users flexibility to apply their own judgment and assumptions.

“Query Genomic Data” functions as the unfiltered data query tool for the molecular data sets (Fig. 1A). In Step 1, users select the type of queries that best fit their needs. The query options include i) gene name, ii) RefSeq (mRNA or protein), iii) Entrez identifier, iv) chromosome number, v) chromosome location, vi) cytoband, or vii) four types of platform specific identifier. In Step 2, users input these identifiers either as a list or as an uploaded text (.txt) or Excel (.xls) file. In Step 3, users select from among 17 datasets that provide various types of information at the DNA, RNA, or protein level described previously (2, 3, 8–13). In Step 4, users provide their E-mail address, and click “Get data”. There are currently 25,722 transcripts, including genes, pseudogenes, and open reading frames with data in this format.

“Query Drug Data” functions as the unfiltered data query tool for the compound activity data set (Fig. 1B). In Step 1, users may select the type of query: i) NSC, ii) compound name, iii) molecular formula-exact match, iv) molecular formula-element match, v) molecular weight range, or vi) mechanism of action (introduced here). The “Molecular formula-element match” allows the used to search based on specific elements in the compound, such as Zn or Se. In Step 2, the user inputs these identifiers either as a list or as an uploaded text (.txt) or Excel (.xls) file. In Step 3, users provide their E-mail address, and click “Get data”. There are currently 52,269 compounds with activity data in this format.

NCI-60 Analysis Tools

These tools (Fig. 2A) provide synopsis information for five forms of molecular as well as compound activity data, in addition to offering several forms of data integration. Data in this form will be used most frequently when one wishes to make comparisons across platforms in a more systems biological fashion without having to engage in time consuming and detailed analysis for each data set. A potential drawback for using data of this type is that quality control requirements are applied, such as data reproducibility and minimal probe and experimental range requirements, and these might eliminate meaningful data in specialized cases. In those cases where data are unavailable, the user will receive a "USER_ERROR_MESSAGE". This file will detail why data were not received, including the input gene or compound failed quality control, that we have no data for that input, or there was some problem with your input.

The tool of interest is selected in Step 1 using the check boxes. The current choices are i) "Cell line signature", ii) "Cross-correlation of transcripts, drugs, and microRNAs, and drugs", iii) "Pattern comparison", iv) "Exome sequencing (DNA) graphical synopsis by gene", and v) "Genetic variant versus drug visualization". Two footnotes are included; the first "[1]Available identifiers and drug mechanisms of action definitions" provides all identifiers for i) drugs (and compounds), ii) gene transcripts, iii) microRNAs, iv) proteins, v) amino acid changing genetic variants, vi) protein function affecting genetic variants, vii) DNA copy number, and viii) the definitions for the drug mechanism of action categories and abbreviations. The second footnote, "[2]Pattern comparison input template", provides the input file to which numerical values may be added for uploading into "Pattern comparison". A maximum of ten patterns may be entered in this fashion in a single query.

In Step 2 you select your input by checking either "Input list" or "Upload file". Using "Input list" allows you to type up to 150 identifiers (as provided in footnote 1 into the "Input the identifier(s)" box). Using "Upload file" allows uploading of these same identifiers as either a text (.txt) or Excel (.xls) file, or to upload your "Pattern comparison input template". In Step 3, enter your e-mail address, and click "Get data".

"Cell Line Signature" Data Compilation and Outputs

"Cell line signature" (Fig. 2B–C) has been developed to i) integrate the multiple probes or experiments that exist for a single gene or drug, ii) provide a "best" synopsis of several forms of molecular and pharmacological data, facilitating their cross-comparison, and iii) avoid the necessity for users to have expertise in the integration of multiple platform types for systems biological and pharmacological studies. This tool provides several data types in a stereotypical format. The six current profiles types are shown in Fig. 2C, and include i) "Gene transcript z scores" for 25,722 genes, pseudogenes, and open reading frames (from five microarray platforms, recorded in the National Center for Biotechnology Information, Gene Expression Omnibus (GEO) with accession numbers GSE5949, GSE5720, GSE32474, GSE29682, and GSE29288) ii) "microRNA mean values" for 360 microRNAs (GEO accession number GSE22821), iii) "Drug activity z scores" for 20,861 compounds, iv) "Gene DNA copy number" for 23,413 genes (from four microarray platforms, with GEO accession numbers GPL11068, GPL13786, GPL3812, and GPL6983) v) "Genetic variant summation" for 12,705 amino acid changing and 9,142 protein function affecting genes and vi) "Protein mean values" for 94 genes and 162 antibodies (GEO accession number GSE5501). Introduced here are help pages designed to give the new user a quick synopsis of what the tool does for each of the signatures. These are accessed by clicking on the text for the signature-of-interest (Fig. 2B).

To use the tool, in Step 1 check the “Cell line signature” box, and then select the radio button for the data type of interest. Input the appropriate identifiers in “Step 2” (provided in the Step 1 footnote 1). Five of these signatures of have been introduced previously (2, 8, 14). Examples of the bar graph outputs for each of the six signature types are presented in Fig. 2C.

The “Protein mean values” signature is introduced in this manuscript with the release of the CellMiner version 1.6. Data were generated using reverse phase protein lysate arrays (15). The data set is high quality, having met stringent specificity requirements. Details of the antibodies used in its generation are described in our AbMiner web-based tool (16, 17). The tool output includes relative protein levels as tabular data, in both normalized and normalized mean centered forms. The mean centered version is visualized as a bar graph (example in Fig. 2C for p53), with the bars for each cell line color-coded by tissue of origin (2). Range, minimum, maximum, average, and standard deviation for the normalized data are included. The user also receives the distribution of the normalized data as a histogram.

The “Cross-Correlations” Tool: Combining and Comparing Different Data

To enable cross-comparison of several forms of data, we have developed “Cross correlations of transcripts, microRNAs and drugs” (Supplementary Fig. S1A). Upgraded from its introduction as an option for either gene transcripts or drug activities (2) under the old “z score determinations” tool, it is now a stand-alone tool (introduced in this manuscript), allowing direct comparison of any combination of from 2 to 150 gene transcript levels, microRNA expression levels, and compound activities in the same query. Identifiers available for input are listed in the “Identifiers and drug mechanism of action” footnote 1 (Fig. 2A, Step 1). The output provides all cross-correlations for the selected identifiers.

In the example given in Supplementary Fig. S1B, the “Cell line signature” pattern of SLFN11 transcript expression, a gene recently identified in the DNA damage response pathway (18, 19) was used as input for a “Pattern comparison” analysis, which identified significantly correlated drugs. The cross-correlations output between SLFN11’s transcript expression level and the activities of these drugs, organized by mechanism of action type, is shown (mechanism of action headers have been added to clarify the figure).

The “Cross correlations” output identifies their interrelationships, identifying significant positive correlations between SLFN11 and DNA damaging drugs (alkylating agents targeting guanine N7, A7; DNA synthesis inhibitors, Ds; topoisomerase I inhibitors, T1; topoisomerase II inhibitors, T2). Among the FDA-approved or clinical trial drugs, tubulin inhibitors and serine-threonine kinase inhibitors (STK), including RO-5126766 had significant negative correlation. The importance of SLFN11 expression for the DNA-targeting classes of drugs has been documented (18, 19). An additional point made obvious by viewing the data is that the A7, Ds, T1, and T2 drugs have significant positive correlated to one another, presumably due to their common DNA damaging effects and the common features determining cellular response to these drugs.

These relationships between SLFN11 and drugs from these mechanism-of-action categories, originally detected bioinformatically, have been shown to be causal, and illustrate the discovery potential of these databases (18–20).

Pattern Comparison

To compare any pattern of interest for the NCI-60 to the complete lists of multiple forms of data, we have developed "Pattern comparison" (Supplementary Fig. S1C). The "Pattern comparison" output has previously provided correlations to the input pattern for: i) the activities of compounds (20,861 currently), ii) the transcript expression of genes (26,065 currently) and iii) 365 microRNAs. Included within the compounds are 158 FDA-approved and 79 clinical trial drugs. Upgraded and reintroduced in this manuscript, "Pattern comparison" now also provides correlations to: i) amino acid changing genetic variants (from exome sequencing) for 12,705 genes, ii) putative protein function affecting variants absent in either the 1000 Genomes or ESP5400 (non-cancerous genomes from 5400 patients) for 9,143 genes, iii) protein levels for 94 genes (as measured by 162 antibodies), and iv) 24 phenotypic parameters, currently focused on genomic instability, epithelial versus mesenchymal status, and pharmacological response (21, 22). The two forms of genetic variants are as obtained using the "Genetic variant summation" cell line signature (Fig. 2B). Format changes include a split of the results into two worksheets, with the statistically significant correlations in the "Significant" worksheet, and the complete (significant and insignificant) results within the "All" worksheet. For comparisons derived from "60 element pattern" inputs, the outputs also include a "Pattern input" worksheet that records the input pattern used (as an organizational help). Patterns may be entered using NA values to exclude cell lines from the analysis when there is scientific reason to do so, such as when one wishes to consider only those cell lines with wild type TP53.

Direct identification of input parameters is available for: i) gene transcript, ii) microRNA transcript, iii) drug activity, and, iv) protein expression levels (introduced here). All these identifiers are available within the footnote 1 identifiers download (Fig. 2A). Any other pattern, such as for a phenotype, characteristic, combination of molecular events, or tissue-of-origin may be entered using the "60 element pattern" option. An input template for this option is available using the "Pattern comparison input template" download from footnote 2 (Supplementary Fig. S1C). This template has also been updated to allow up to ten patterns to be submitted simultaneously.

In the example given in Supplementary Fig. S1C and D, a melanoma tissue-specific "60 element pattern" was uploaded (Supplementary Fig. S1D, "Input"). The top two correlated gene transcript patterns, plus CTLA4 are shown under "Outputs". Calpain-3 (CAPN3) and dopachrome tautomerase (DCT) have prior association with invasiveness, and assessment of melanomas, respectively (23, 24). CTLA4 is a target for ipilimumab in patients with non-resectable and malignant melanoma, however its method of action is thought to occur through effects on the patient's endogenous immune system so its unexpected expression within 5/10 melanomas is of interest (25, 26). Of the next five most highly correlated genes from this list, four have previously established connections to melanoma (SOX10, TYR, S100A1, and MLANA). The top two significant microRNA correlations, hsa-miR-146a and

211 have prior association to melanoma initiation and progression, and to invasiveness, respectively (27, 28). Among the FDA-approved or clinical trial drugs, the top two correlated drugs, vemurafenib and selumetinib both have prior report of efficacy in melanoma patients (29, 30). The pattern matches between the melanoma input pattern and these already established molecular and pharmacological patterns illustrate the quality and informative nature of these databases. The drug target CTLA4, microRNA 514, and drug hypothemycin are all novel correlations, and illustrate the potential for of Pattern Comparison as a discovery tool.

Exome Sequencing (DNA) Graphical Synopsis by Gene

To provide the user a visual summary of the genetic variants that occur within a gene as identified in our exome sequencing study, we developed "Exome sequencing (DNA) graphical synopsis by gene" (10). This tool is selected by checking its box as shown in Fig. 3A. Identifier input and data retrieval are as described for Fig. 2A. The output includes both a synopsis of all variants within the NCI-60, as seen in Fig. 3B, as well as individual visualizations for all cell lines (not shown). The types of variants are identified as defined in Fig. 3B. The version of the gene used for visualization is as defined by the denoted NCBI accession number (in the html output). The two examples shown include a tumor suppressor (TP53), and an oncogene (KRAS). Note that in general, inactivating mutations tend to be more widespread for tumor suppressor genes as seen for TP53, while activating mutations tend to congregate for oncogenes, as seen for KRAS codon 12, with mutations in seven cell lines. This may be expected to be the case for the important, although uncommon, neomorphic (gain of novel gene function) mutations as well.

Genetic Variant Versus Drug Visualization

Designed to explore potential gene-drug relationships, "Genetic variant versus drug visualization" provides a visualization that compares variants for a given gene to shifts in activity for a given compound (14). This tool is selected as shown in Fig. 3C.

Using this tool, previously recognized (proof-of-principle) relationships may be observed, as for the protein kinase BRAF and the MEK (MAP2K1, 3, and 6)-ERK (MAPK1, 3, 4 and 15) inhibitor hypothemycin (10). Novel plausible relationships may also be discovered, such as between the DNA repair gene MUS81 and the DNA synthesis inhibitor clofarabine. Another example of this type is the pro-apoptotic STAT2 and the alkylating agent uracil mustard. STAT2 has had prior reported association with DNA synthesis inhibitors (31); however genes that affect apoptosis might be expected to influence the response to multiple drug types. In addition, one can identify unstudied compounds that may target potentially pharmacologically useful genes, such as the epithelial-mesenchymal associated tight junction protein 3 (TJP3).

Introduced here, this tool also allows the user to query a single gene versus all compounds (example input, *:BRAF) or a single drug versus all genes (example input, 123127:*). These inputs will identify all correlated drugs for the *:BRAF example, or genes for the 123127:* (doxorubicin) example. Queried in this fashion, BRAF identifies 38 correlated drugs, and doxorubicin identifies four genes. The criteria for inclusion using this function include a

Mathew's correlation coefficients (MCC) of ≥ 0.596 ($p \leq 0.0002$ for $n = 35$), and are the gene-compound pairings in Table 3 and Supplementary Tables S4A and B of our prior manuscript (14). Outputs of this type with more than 150 matches will be received as a summary sheet.

Consideration of Inter-parameter Relationships: RAS Activation

In addition to the consideration of individual molecular events involved in either cancer progression or pharmacological response, CellMiner provides an opportunity to examine the contribution of multiple molecular events and forms of data simultaneously. A simple example that also illustrates the complexity of considering overlapping molecular events is provided by consideration of whether RAS is activated across these cancer cell lines. The three forms of RAS, H, K, and N, are well-studied, important oncogenes (32).

Three forms of data provided by the "Cell line signature" tool (Fig. 4A) are informative for this purpose, DNA copy number, gene transcript levels, and activating mutations (from the "Genetic variant summation" tool). Reviewing shifts in copy number from 2N from the "Gene DNA copy number" option, in its "Graphical Output" worksheet, one finds amplifications for all three forms of RAS. Examples are given for each gene in Fig. 4B. Reviewing those data systematically, one, six, and five of the cell lines appear to have DNA amplifications for HRAS, KRAS, and NRAS, respectively (Fig. 4C). Next, by observing the transcript levels using the "Gene transcript z scores" option, it is apparent that each of these RAS amplified cell lines also have up-regulated mRNA expression. In addition, one and four additional cell lines (without DNA amplifications) are found to have up-regulated transcript levels in HRAS and NRAS, respectively (Fig. 4C). Finally, by determining the presence of activating mutations (those with amino acid changes at 12, 13, and 61) using the "Genetic variant summation" option, one, ten, and two of the cell lines appear to be activated genetically in HRAS, KRAS, and NRAS, respectively (Fig. 4C). Viewing these data in aggregate (Fig. 4D), 3, 14, and 10 of the cell lines appear activated for HRAS, KRAS, and NRAS, respectively. So overall, 26/60 of the cell lines have indication of RAS activation, including 4/5 breast, 4/6 leukemia, and 5/9 lung cell lines.

Consideration of Inter-parameter Relationships and Their Relationship to Pharmacology: PTEN Knockdown

An extension of the consideration of the contributions of multiple molecular events simultaneously is how it might influence pharmacology. An example of this is provided by the consideration of PTEN knockdown. PTEN is an important tumor suppressor that antagonizes the PIK3 (PIK3CB, C3, and R5)-AKT (AKT1, 2, and 3) pathway and is commonly deleted in cancer (33, 34).

The same three forms of data used in Fig. 4 are again provided by the "Cell line signature" tool, querying the database for PTEN DNA copy number, gene transcript levels, and deactivating mutations, as shown in Fig. 5A. Reviewing shifts in DNA copy number from 2N from the "Gene DNA copy number" option, in the "Graphical Output" worksheet, one finds PTEN deletions for four cell lines. Three of these are shown in Fig. 5B. By observing the transcript levels using the “Gene transcript z scores” option, it is apparent that each of these PTEN deleted cell lines also has down-regulated expression, and that these cell lines are the four lowest PTEN expressers in the NCI-60 (Fig. 5C). By determining the presence of predicted function-affecting mutations, as detailed in the “Genetic variant summation” tool, four different cell lines have indication of being genetic hypomorphs with loss-of-function (Fig. 5C). Of these, both BT-549 and SF-295 have nonsense (premature stops), MOLT-4 a frameshift, and SK-MEL28 the T167A mutation, which has been shown previously to result in a 50% reduction in function (35). Taking these three molecular parameters as a composite, a pattern of eight cell lines with apparent PTEN inactivation can be derived (Fig. 5C, right).

Using this PTEN composite pattern as input for the “Pattern comparison” tool, using the “60 element pattern” option (Fig. 5A), four drugs with either FDA-approval or clinical trial status are found to have significant correlation (p< 0.05, Fig. 5D). Of these, fenretinide (NSC 374551) has been shown to reactivate PTEN previously, affirming the relationship to the gene (36). PX-316 (NSC 710297) also has pathway connection, being an AKT-inhibitor. Thus 2/4 (50%) of the drugs with significant correlation to the input PTEN knockdown pattern have obvious connection to PTEN, whereas only 4.8% of the total known mechanism of action drugs present in CellMiner do. This enrichment was found to be significant (p<0.01) by binomial distribution, providing evidence for the saliency of the molecular data when compared to the pharmacological.

Discussion

In addition to providing a review of the preexisting CellMiner tools, the current manuscript also introduces new tools and upgrades to current ones, as well as providing examples of how the tools and databases may be used for systems biology and systems pharmacology. Examples are given that provide both proof-of-principle as well as novel findings (Supplementary Fig. S1A–D, Fig. 3C–D, Fig. 4A–D, and Fig. 5A–D).

New and upgraded tools include i) the introduction of the “Protein mean values” cell line signature (Fig. 2B and C), ii) the “Cross-correlations of transcripts, microRNAs, and drugs” as an upgraded stand-alone tool (Supplementary Fig. S1A and B), iii) four additional data types for the “Pattern comparison” output- “Amino acid changing” and “Protein function affecting” genetic variants, “Protein levels”, and “Miscellaneous phenotypic parameters”, and iv) introduction of the star feature (*) for the “Genetic variant versus drug visualization” tool, allowing the user to rapidly identify all compounds with significant correlation to a single gene, or vice versa.

CellMiner previously has enabled us to: i) identify promoter-proximal transcriptional pausing in human genes (37, 38), ii) discover the helicase SLFN11 as a causal determinant of response to DNA-damaging agents (18), iii) recognize the regulation of MYC expression by miR-375 (39), iv) recognize the importance of MYC as a driver of mitochondrial genes (40), v) reveal genetic inactivation or endogenous activation of CHEK2 across the NCI-60 (40); vi) link USP7 and Daxx to taxane resistance (41), vi) link TP53 wild type status, Mdm2 transcript level, and miR-34a transcript level with nutlin activity (10), viii) reveal the interrelationship between RAS (H, K, and NRAS)-BRAF-PTEN mutational status, EGFR expression, and ERBB2 expression with erlotinib activity (10), ix) demonstrate the strong correlation between ABCB1 expression and doxorubicin activity (2), x) recognize both known and novel genes expression levels, microRNA expression levels, and drug activities with a colon-specific pattern input to "Pattern comparison" (2), xi) identify predominant co-regulation among cell migration genes (42), xii) identify co-regulation among kinetochore genes, their prospective regulatory elements, and their association with genomic instability (43), xiii) show the connection between accumulation of mass homozygotes in the cancer cell lines as compared to non-cancerous HapMap trios (44), xiv) identify the drug Ro5-3335 as a candidate treatment for core binding factor leukemias (45), xv) associate CDKN2A DNA copy number and expression to mitoxantrone activity (8), xvi) define an epithelial gene expression signature (46), and xvii) recognize the composite relationship between the mutational status of multiple genes from the EGFR-ERBB2 pathway and drug response, including the directionality of that influence as a function of molecular pathway considerations (14). The diversity among these observations gives an indication of the boundless scope and range of the types of possible discoveries that can be made using the NCI-60 database and CellMiner set of tools.

In addition to being a resource for generating or providing tests for hypotheses, such as those described above, CellMiner also provides a template for making "omic" data of this type accessible and usable for a broad portion of the scientific public. Access of this type remains a serious shortcoming for the field currently, and improved access to and connectivity between the multiple cell line and clinical databases that have either already been done or are in progress should be a goal. A synopsis of the types of data available across the three major cell line screens, the NCI-60, the Cancer Cell Line Encyclopedia (CCLE), and the Cancer Genome Project (CGP), is presented in Supplementary Table S1. However, currently the barriers to data integration and interrogation remain daunting, restricting access primarily to bioinformaticians, statisticians, and those with computer expertise. Due to the multi-faceted nature of the disease information that needs to considered in the cancer context, there is certainly need for the input of molecular biologists, clinicians and all others with pertinent domain expertise as well.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Shankavaram UT, Varma S, Kane D, Sunshine M, Chary KK, Reinhold WC, et al. CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. BMC Genomics. 2009; 10:277. [PubMed: 19549304]
2. Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, et al. CellMiner: a Web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. Cancer Res. 2012; 72:3499–511. [PubMed: 22802077]
3. CellMiner. Bethesda (MD): National Cancer Institute; database on the Internet[cited 2015 Apr 13]. Available from: http://discover.nci.nih.gov/cellminer/
4. Genomics and Bioinformatics Group. Bethesda (MD): National Cancer Institute; [homepage on the Internet][cited 2015 Apr 13]. Available from: http://discover.nci.nih.gov/

5. Developmental Therapeutics Program. Bethesda (MD): National Cancer Institute; [homepage on the Internet][cited 2015 Apr 13]. Available from: http://dtp.nci.nih.gov/
6. Rubinstein LV, Shoemaker RH, Paull KD, Simon RM, Tosini S, Skehan P, et al. Comparison of in vitro anticancer-drug-screening data generated with a tetrazolium assay versus a protein assay against a diverse panel of human tumor cell lines. J Natl Cancer Inst. 1990; 82:1113–8. [PubMed: 2359137]
7. Lorenzi P, Reinhold W, Varma S, Hutchinson A, Pommier Y, Chanock S, et al. DNA fingerprinting of the NCI-60 cell line panel. Mol Cancer Ther. 2009; 8:713–24. [PubMed: 19372543]
8. Varma S, Pommier Y, Sunshine M, Weinstein JN, Reinhold WC. High resolution copy number variation data in the NCI-60 cancer cell lines from whole genome microarrays accessible through CellMiner. PLoS One. 2014; 9:e92047. [PubMed: 24670534]
9. Reinhold W, Reimers M, Maunakea A, Kim S, Lababidi S, Scherf U, et al. Detailed DNA methylation profiles of the E-cadherin promoter in the NCI-60 cancer cells. Mol Cancer Ther. 2007; 6:391–403. [PubMed: 17272646]
10. Abaan OD, Poley EC, Davis SR, Zhu YJ, Bilke S, Walker RL, et al. The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. Cancer Res. 2013; 73:4372–82. [PubMed: 23856246]
11. Ikediobi O, Davies H, Bignell G, Edkins S, Stevens C, O'Meara S, et al. Mutation analysis of twenty-four known cancer genes in the NCI-60 cell line set. Mol Cancer Ther. 2006; 5:2606–12. [PubMed: 17088437]
12. Szakacs G, Annereau J, Lababidi S, Shankavaram U, Arciello A, Bussey K, et al. Predicting drug sensitivity and resistance: profiling ABC transporter genes in cancer cells. Cancer Cell. 2004; 6:129–37. [PubMed: 15324696]
13. Liu H, D'Andrade P, Fulmer-Smentek S, Lorenzi P, Kohn KW, Weinstein JN, et al. mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. Mol Cancer Ther. 2010; 9:1080–91. [PubMed: 20442302]
14. Reinhold WC, Varma S, Sousa F, Sunshine M, Abaan OD, Davis SR, et al. NCI-60 Whole exome sequencing and pharmacological CellMiner analyses. PLoS One. 2014; 9:e101670. [PubMed: 25032700]
15. Nishizuka S, Charboneau L, Young L, Major S, Reinhold W, Waltham M, et al. Proteomic profiling of the NCI60 cancer cell lines using new high-density 'reverse-phase' lysate microarrays. Proc Natl Acad Sci U S A. 2003; 100:14229–34. [PubMed: 14623978]
16. AbMiner. Bethesda (MD): National Cancer Institute; [database on the Internet][cited 2015 Apr 13]. Available from: http://discover.nci.nih.gov/abminer/
17. Major SM, Nishizuka S, Morita D, Rowland R, Sunshine M, Shankavaram U, et al. AbMiner: a bioinformatic resource on available monoclonal antibodies and corresponding gene identifiers for genomic, proteomic, and immunologic studies. BMC bioinformatics. 2006; 7:192. [PubMed: 16600027]
18. Zoppoli G, Regairaz M, Leo E, Reinhold WC, Varma S, Ballestrero A, et al. Putative DNA/RNA helicase Schlafen-11 (SLFN11) sensitizes cancer cells to DNA-damaging agents. Proc Natl Acad Sci U S A. 2012; 109:15030–5. [PubMed: 22927417]
19. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012; 483:570–5. [PubMed: 22460902]
20. Gmeiner WH, Reinhold WC, Pommier Y. Genome-wide mRNA and microRNA profiling of the NCI 60 cell-line screen and comparison of FdUMP[10] with fluorouracil, floxuridine, and topoisomerase 1 poisons. Mol Cancer Ther. 2010; 9:3105–14. [PubMed: 21159603]
21. 1000 Genomes. Cambridgeshire (UK): The European Bioinformatics Institute; 2008–11. [homepage on the Internet][cited 2015 Apr 15]. Available from: http://www.1000genomes.org/
22. NHLBI Exome Sequencing Project (ESP). Seattle (WA): University of Washington; 2011. [database on the Internet][cited 2015 Apr 15]. Available from: http://evs.gs.washington.edu/EVS/
23. Ruffini F, Tentori L, Dorio AS, Arcelli D, D'Amati G, D'Atri S, et al. Platelet-derived growth factor C and calpain-3 are modulators of human melanoma cell invasiveness. Oncology Rep. 2013; 30:2887–96.

24. Filimon A, Zurac SA, Milac AL, Sima LE, Petrescu SM, Negroiu G. Value of dopachrome tautomerase detection in the assessment of melanocytic tumors. Melanoma Res. 2014; 24:219–36. [PubMed: 24709887]

25. Stadler S, Weina K, Gebhardt C, Utikal J. New therapeutic options for advanced non-resectable malignant melanoma. Advances in medical sciences. 2014; 60:83–8. [PubMed: 25596540]

26. Funt SA, Page DB, Wolchok JD, Postow MA. CTLA-4 antibodies: new directions, new combinations. Oncology (Williston Park). 2014; 28(Suppl 3):6–14. [PubMed: 25387681]

27. Forloni M, Dogra SK, Dong Y, Conte D Jr, Ou J, Zhu LJ, et al. miR-146a promotes the initiation and progression of melanoma by activating Notch signaling. eLife. 2014; 3:e01460. [PubMed: 24550252]

28. Boyle GM, Woods SL, Bonazzi VF, Stark MS, Hacker E, Aoude LG, et al. Melanoma cell invasiveness is regulated by miR-211 suppression of the BRN2 transcription factor. Pigment Cell Melanoma Res. 2011; 24:525–37. [PubMed: 21435193]

29. Patrawala S, Puzanov I. Vemurafenib (RG67204, PLX4032): a potent, selective BRAF kinase inhibitor. Future Oncol. 2012; 8:509–23. [PubMed: 22646766]

30. Robert C, Dummer R, Gutzmer R, Lorigan P, Kim KB, Nyakas M, et al. Selumetinib plus dacarbazine versus placebo plus dacarbazine as first-line treatment for BRAF-mutant metastatic melanoma: a phase 2 double-blind randomised study. Lancet Oncol. 2013; 14:733–40. [PubMed: 23735514]

31. Uluer ET, Aydemir I, Inan S, Ozbilgin K, Vatansever HS. Effects of 5-fluorouracil and gemcitabine on a breast cancer cell line (MCF-7) via the JAK/STAT pathway. Acta Histochem. 2012; 114:641–6. [PubMed: 22172707]

32. Quinlan MP, Settleman J. Isoform-specific ras functions in development and cancer. Future Oncol. 2009; 5:105–16. [PubMed: 19243303]

33. Lim HJ, Crowe P, Yang JL. Current clinical regulation of PI3K/PTEN/Akt/mTOR signalling in treatment of human cancer. J Cancer Res Clin Oncol. 2015; 141:671–89. [PubMed: 25146530]

34. Hollander MC, Blumenthal GM, Dennis PA. PTEN loss in the continuum of common cancers, rare syndromes and mouse models. Nat Rev Cancer. 2011; 11:289–301. [PubMed: 21430697]

35. Rodriguez-Escudero I, Oliver MD, Andres-Pons A, Molina M, Cid VJ, Pulido R. A comprehensive functional analysis of PTEN mutations: implications in tumor- and autism-related syndromes. Human Mol Genet. 2011; 20:4132–42. [PubMed: 21828076]

36. Janardhanan R, Banik NL, Ray SK. N-Myc down regulation induced differentiation, early cell cycle exit, and apoptosis in human malignant neuroblastoma cells having wild type or mutant p53. Biochem Pharmacol. 2009; 78:1105–14. [PubMed: 19540207]

37. Eddy J, Vallur AC, Varma S, Liu H, Reinhold WC, Pommier Y, et al. G4 motifs correlate with promoter-proximal transcriptional pausing in human genes. Nucleic Acids Res. 2011; 39:4975–83. [PubMed: 21371997]

38. Reinhold WC, Mergny JL, Liu H, Ryan M, Pfister TD, Kinders R, et al. Exon array analyses across the NCI-60 reveal potential regulation of TOP1 by transcription pausing at guanosine quartets in the first intron. Cancer Res. 2010; 70:2191–203. [PubMed: 20215517]

39. Jung H, Wang H, Patel R, Phillips B, Reinhold W, Cohen D, et al. miR-375 regulation of CIP2A controls oral cancer cell proliferation and survival. Mol Biol Cell. 2013; 24:1638–48. [PubMed: 23552692]

40. Zoppoli G, Douarre C, Dalla Rosa I, Liu H, Reinhold W, Pommier Y. Coordinated regulation of mitochondrial topoisomerase IB with mitochondrial nuclear encoded genes and MYC. Nucleic Acids Res. 2011; 39:6620–32. [PubMed: 21531700]

41. Giovinazzi S, Morozov VM, Summers MK, Reinhold WC, Ishov AM. USP7 and Daxx regulate mitosis progression and taxane sensitivity by affecting stability of Aurora-A kinase. Cell Death Differ. 2013; 20:721–31. [PubMed: 23348568]

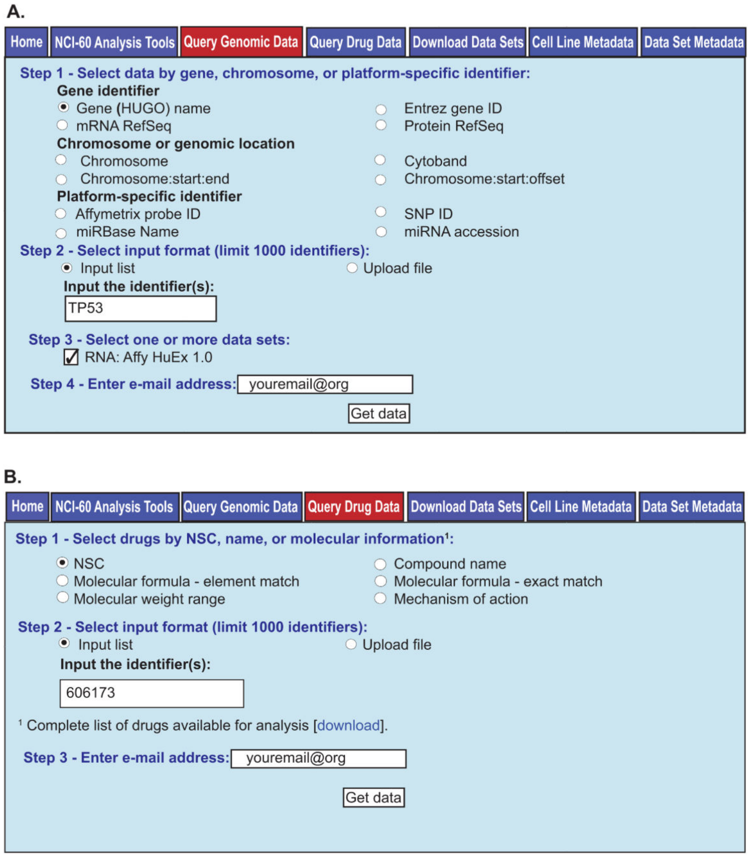
42. Kohn KW, Zeeberg BR, Reinhold WC, Sunshine M, Luna A, Pommier Y. Gene expression profiles of the NCI-60 human tumor cell lines define molecular interaction networks governing cell migration processes. PLoS One. 2012; 7:e35716. [PubMed: 22570691]

43. Reinhold WC, Erliandri I, Liu H, Zoppoli G, Pommier Y, Larionov V. Identification of a predominant co-regulation among kinetochore genes, prospective regulatory elements, and association with genomic instability. PLoS One. 2011; 6:e25991. [PubMed: 22016797]

44. Ruan X, Kocher JP, Pommier Y, Liu H, Reinhold WC. Mass homozygotes accumulation in the NCI-60 cancer cell lines as compared to HapMap trios, and relation to fragile site location. PLoS One. 2012; 7:e31628. [PubMed: 22347499]

45. Cunningham L, Finckbeiner S, Hyde RK, Southall N, Marugan J, Yedavalli VR, et al. Identification of benzodiazepine Ro5-3335 as an inhibitor of CBF leukemia through quantitative high throughput screen against RUNX1-CBFbeta interaction. Proc Natl Acad Sci U S A. 2012; 109:14592–7. [PubMed: 22912405]

46. Kohn KW, Zeeberg BM, Reinhold WC, Pommier Y. Gene expression correlations in human cancer cell lines define molecular interaction networks for epithelial phenotype. PLoS One. 2014; 9:e99269. [PubMed: 24940735]



A.

Step 1 - Select data by gene, chromosome, or platform-specific identifier:

Gene identifier

● Gene (HUGO) name ○ Entrez gene ID
○ mRNA RefSeq ○ Protein RefSeq

Chromosome or genomic location

○ Chromosome ○ Cytoband
○ Chromosome:start:end ○ Chromosome:start:offset

Platform-specific identifier

○ Affymetrix probe ID ○ SNP ID
○ miRBase Name ○ miRNA accession

Step 2 - Select input format (limit 1000 identifiers):

● Input list ○ Upload file

Input the identifier(s): TP53

Step 3 - Select one or more data sets:

☑ RNA: Affy HuEx 1.0

Step 4 - Enter e-mail address: youremail@org

B.

Step 1 - Select drugs by NSC, name, or molecular information[1]:

● NSC ○ Compound name
○ Molecular formula - element match ○ Molecular formula - exact match
○ Molecular weight range ○ Mechanism of action

Step 2 - Select input format (limit 1000 identifiers):

● Input list ○ Upload file

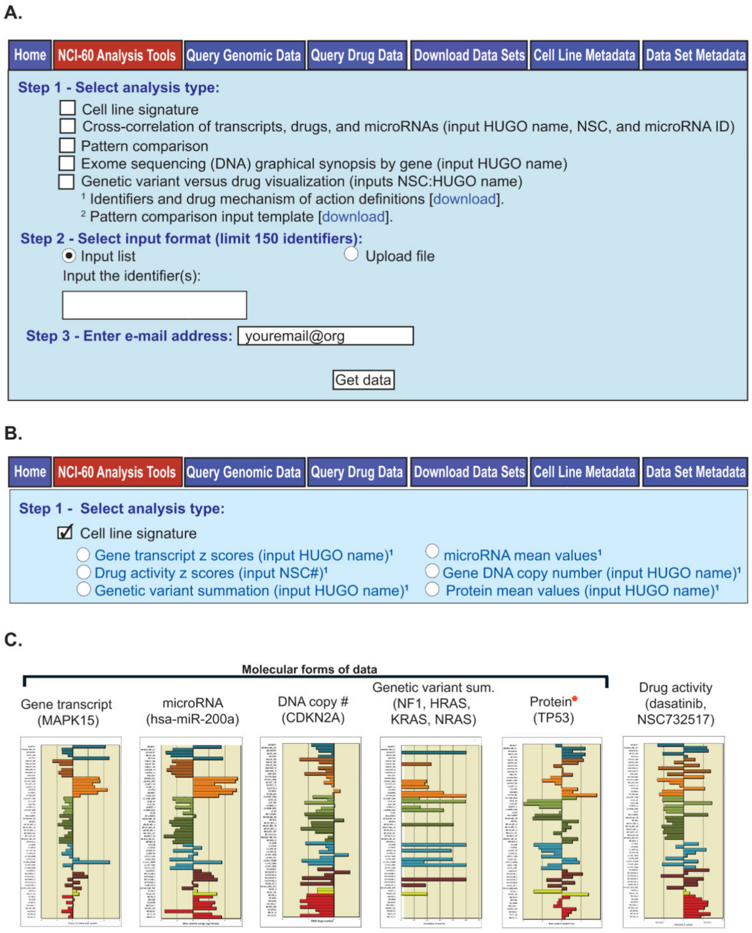
Input the identifier(s): 606173

[1] Complete list of drugs available for analysis [download].

Step 3 - Enter e-mail address: youremail@org

Figure 1.

The "Query Genomic Data" and "Query Drug Data" tools. **A.** The "Query Genomic Data" tool. To access this tool, click on the "Query Genomic Data" tab (with red fill). In "Step 1", the type of identifier to be used is selected. One may choose from the options shown, with the choice being dependent on the data set to be selected. In "Step 2", the identifiers are entered. In "Step 3", the data set of interest is selected. Currently there are 17 data sets, with only one being shown in the figure due to size constraints. In "Step 4", the user enters their e-mail address, and clicks "Get data". **B.** The "Query Drug Data" tool. To access this tool, click on the "Query Drug Data" tab (with red fill). In "Step 1", the type of identifier to be used is selected from the options shown. In "Step 2", the identifiers are entered. In "Step 3", the user enters their e-mail address, and clicks "Get data".



Step 1 - Select analysis type:

- □ Cell line signature
- □ Cross-correlation of transcripts, drugs, and microRNAs (input HUGO name, NSC, and microRNA ID)
- □ Pattern comparison
- □ Exome sequencing (DNA) graphical synopsis by gene (input HUGO name)
- □ Genetic variant versus drug visualization (inputs NSC:HUGO name)

[1] Identifiers and drug mechanism of action definitions [download].
[2] Pattern comparison input template [download].

Step 2 - Select input format (limit 150 identifiers):

● Input list ○ Upload file

Input the identifier(s):

Step 3 - Enter e-mail address: youremail@org

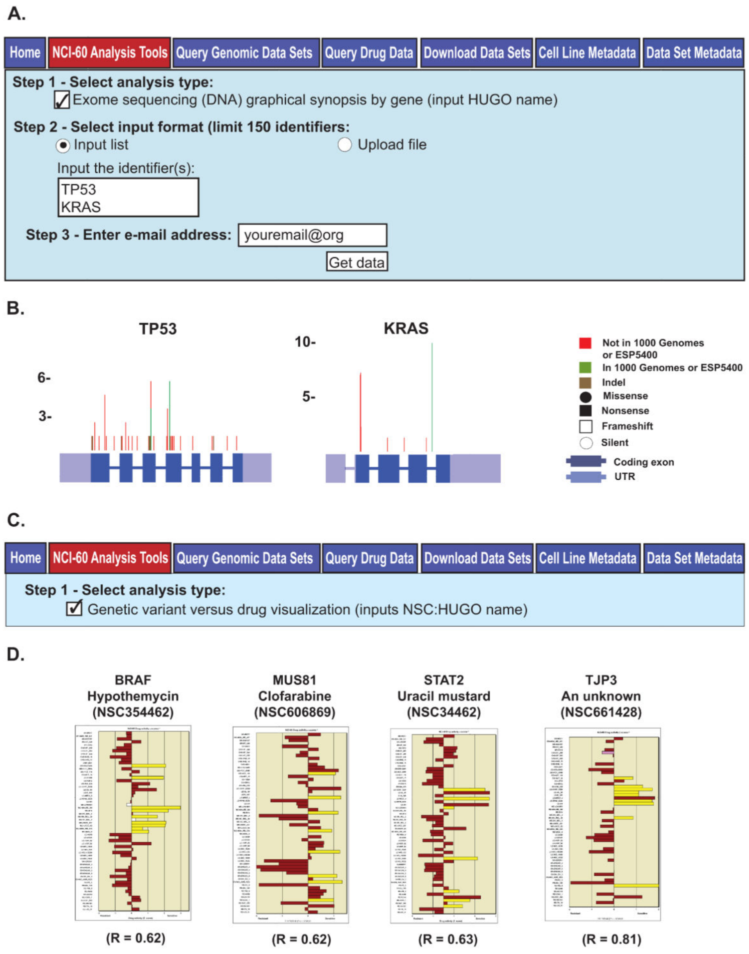
Step 1 - Select analysis type:

☑ Cell line signature

- ○ Gene transcript z scores (input HUGO name)[1]
- ○ Drug activity z scores (input NSC#)[1]
- ○ Genetic variant summation (input HUGO name)[1]
- ○ microRNA mean values[1]
- ○ Gene DNA copy number (input HUGO name)[1]
- ○ Protein mean values (input HUGO name)[1]

Figure 2.

The “NCI-60 Analysis Tools” and “Cell line signature”. **A.** The “NCI-60 Analysis Tools”. To access this tool, click on the “NCI-60 Analysis Tools” tab (with red fill). In “Step 1”, the type of analysis to be done is selected from the options shown. The identifiers used in these tools are available from the “Identifiers and drug mechanism of action definitions” download. In “Step 2”, the user selects whether to input the identifiers as a list or as an uploaded file, and the identifiers are entered or uploaded. In “Step 3”, the user enters their e-mail address, and clicks “Get data”. **B.** The “Cell line signature” tool. To access this tool, click on the “NCI-60 Analysis Tools” tab, followed by the “Cell line signature” check box. Select from the six forms of signatures using the radio buttons. Identifier input and data receipt is done as in Fig. 2A. **C.** Bar graph examples for the five forms of molecular data, plus the compound activities available currently. In all bar-plot outputs, the x-axis indicates higher levels to the right, and lower levels to the left. The y-axis is the 60 cell lines, organized by tissue of origin. The red star next to “Protein” indicates it is a new output (introduced here).



Step 1 - Select analysis type:
☑ Exome sequencing (DNA) graphical synopsis by gene (input HUGO name)

Step 2 - Select input format (limit 150 identifiers:
● Input list ○ Upload file

Input the identifier(s):
TP53
KRAS

Step 3 - Enter e-mail address: youremail@org

Step 1 - Select analysis type:
☑ Genetic variant versus drug visualization (inputs NSC:HUGO name)

Figure 3.

The "Exome sequencing (DNA) graphical synopsis by gene" and "Genetic variant versus drug visualization" tools. **A**. The "Exome sequencing (DNA) graphical synopsis by gene" tool. To access this tool, click on the "NCI-60 Analysis Tools" tab, followed by the "Exome sequencing (DNA) graphical synopsis by gene" check box. Use gene HUGO names as input, to a maximum of 150. Data receipt is done as in Fig. 2A. **B.** The composite images for the TP53 and KRAS inputs. The gene exons, UTRs and genetic variant locations, number and types are as indicated. **C.** The "Genetic variant versus drug visualization" tool. To access this tool, click on the "NCI-60 Analysis Tools" tab, followed by the "Genetic variant versus drug visualization" check box. Select input as NSC:HUGO name pairs, to a maximum of 150 (pairs). Data receipt is done as in Fig. 2A. **D.** The bar graph output for four NSC:HUGO name pairs are shown. The x-axis is the drug activity z score. The y axis is the cell lines. Brown filled bars indicate that the cell line has no variants that contribute to a statistically significant shift in drug activity. The yellow filled bars indicate that the cell line has variants that contribute to a statistically significant Pearson correlation to the selected drug. "R" is the correlation value between the presence of variants and shift in drug activity.

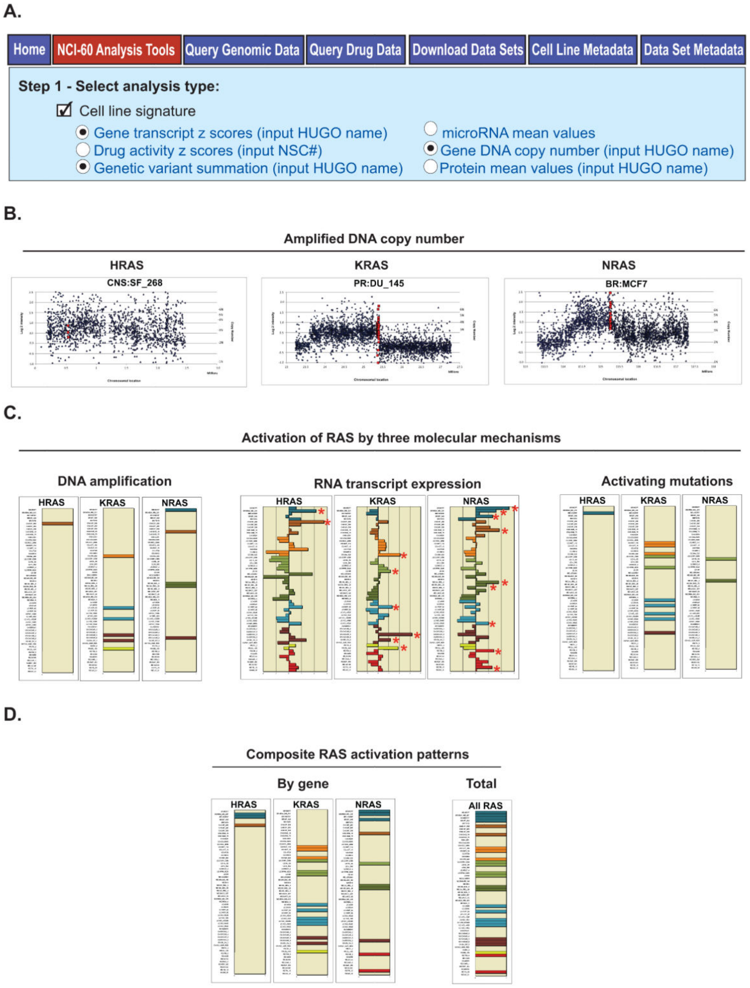
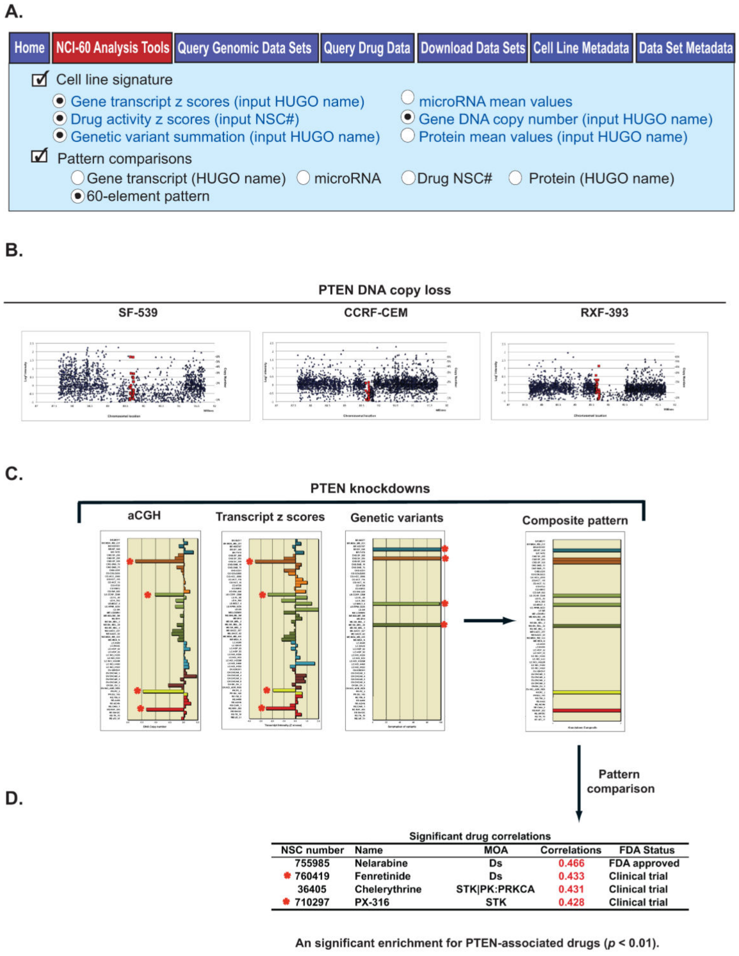


Figure 4.

A composite analysis of RAS activation. **A.** All data used in this analysis is from the “Cell line signature” tool”. Access, identifier input, and data receipt is as described for Fig. 2. Data from the three indicated signatures “Gene transcript z scores”, “Gene DNA copy number”, and “Genetic variant summation”, were selected one at a time. **B.** Examples of cell lines with amplified DNA copy number for H, K, and NRAS taken from the “Gene DNA copy number” tool (“Graphical Output” worksheet). The x-axis is the chromosomal location. The y-axis is dual labeled for both log2 intensity of the probes, and estimated DNA copy number. The dark blue points are the flanking probe intensities. The red points are the gene probe intensities. **C.** Three molecular indications of RAS activation. DNA amplification occurs for one, six, and five cell lines for H, K, and NRAS, respectively. For the bar graphs, the colored bars indicate amplification, and the y-axis indicates the cell lines. RNA transcript expression is up-regulated in two, six, and nine cell lines for H, K, and NRAS, respectively, as indicated by the red stars. For the bar graphs, the x-axis is relative expression as indicated by z score. The y-axis is the cell lines. Activating mutations occur in one, eleven, and two cell lines, respectively. For the bar graphs, the colored bars indicate activating mutations, and the y-axis indicates the cell lines. **D.** Composite activation patterns using the DNA amplification, RNA up-regulation, and activating mutations (from Fig. 4C). Indicators of activation occur for 3, 14, and 10 cell lines for H, K and NRAS, respectively. Taken in total, “RAS” in the generic sense is activated in 26 of the cell lines. For the bar graphs, the colored bars indicate RAS activation, and the y-axis indicates the cell lines. The bar graphs in “C” for “DNA amplification” and “Activating mutations”, and in “D” were generated to illustrate the point, and were not generated directly by CellMiner.



☑ Cell line signature
● Gene transcript z scores (input HUGO name)
● Drug activity z scores (input NSC#)
● Genetic variant summation (input HUGO name)
○ microRNA mean values
● Gene DNA copy number (input HUGO name)
○ Protein mean values (input HUGO name)
☑ Pattern comparisons
○ Gene transcript (HUGO name) ○ microRNA ○ Drug NSC# ○ Protein (HUGO name)
● 60-element pattern

Significant drug correlations

NSC number	Name	MOA	Correlations	FDA Status
755985	Nelarabine	Ds	0.466	FDA approved
✱ 760419	Fenretinide	Ds	0.433	Clinical trial
36405	Chelerythrine	STK|PK:PRKCA	0.431	Clinical trial
✱ 710297	PX-316	STK	0.428	Clinical trial

An significant enrichment for PTEN-associated drugs ($p < 0.01$).

Figure 5.

A composite analysis of PTEN knockdown and its link to pharmacology. **A.** All data used in this analysis is from the “Cell line signature” and “Pattern comparisons” tools. Access, identifier input, and data receipt is as described for Fig. 2 and Supplementary Fig. S1C and D. Data from the four indicated signatures “Gene transcript z scores”, “Gene DNA copy number”, “Genetic variant summation”, and “Drug activity z scores” were selected one at a time, using a “60 element pattern” input. **B.** Examples of cell lines with DNA copy number loss for PTEN, taken from the “Gene DNA copy number” tool (“Graphical Output” worksheet). The x-axis is the chromosomal location. The y-axis is dual labeled for both log2 intensity of the probes, and estimated DNA copy number. The blue points are the flanking probe intensities. The red points are the probe intensities that fall within the gene. **C.** Three molecular indications of PTEN knockdown. DNA loss as measured by aCGH occurs for four cell lines. For the bar graph, the x-axis indicates the DNA copy number, and the y-axis indicates the cell lines. RNA transcript expression is down-regulated in four cell lines. For the bar graphs, the x-axis is relative expression as indicated by z score. The y-axis is the cell lines. Deactivating mutations occur in four cell lines. For the bar graph, the colored bars indicate deactivating mutations, and the y-axis indicates the cell lines. In all three of these bar plots, the red stars indicate those cell lines with indication of PTEN knockdown. The composite (knockdown) pattern, derived from the three forms of molecular data, indicate knockdown occurs in eight cell lines. For this bar graph, the colored bars indicate PTEN knockdown, and the y-axis indicates the cell lines. The bar graphs for “Genetic variants”, and in the “Composite pattern” were generated to illustrate the point, and were not generated directly by CellMiner. **D.** Input of the composite pattern from Fig. 5C to “Pattern comparisons” identifies four FDA-approved or clinical trial drugs with significant correlation. MOA is mechanism of action. The red stars indicate that two of these have prior connection to PTEN, either through literature or pathway.